Chapter 9

Logistic Regression

NOTE: This supplement to the first edition is being released online as supplemental Chapter 9. Its pagination corresponds to the printed first edition text. The hyperlinks to sections and equations do not work, but the references are correct. This supplement includes a set of exercises and solutions to the odd-numbered exercises. The data used in this chapter is in the supplement to the oibiostat data package and can be downloaded executing the following command in the R console:

```
devtools::install_github("OI-Biostat/oi_biostat_data", ref = "supplements")
```

Logistic regression is used to explore relationships between a response variable with two possible values (e.g., yes/no, success/failure, 0/1, etc.) and one or more predictor variables. The **logistic regression** model estimates the odds of an outcome given a predictor, and the odds ratio (OR) associated with change in the value of a predictor; in certain cases, the model also estimates the probability of an outcome given a predictor.

For labs, slides, and other resources, please visit
www.openintro.org/book/biostat

9.1 Chapter overview

This chapter focuses on traditional methods of inference for logistic regression that are commonly used in epidemiology and public health, with emphases on both inference for model parameters and prediction. The interpretation and use of logistic models for inference is contained in the three sections following this overview; these sections contain the core material used in many applications in medical research, epidemiology and public health.

The last two sections describe methods for assessing the fit of a logistic model, both for inference and for prediction, and present an extended case study using logistic regression to modify and evaluate a triage system in hospital emergency departments. The section on assessing fit is longer than most sections in the book, but the material is necessary for understanding the behavior of the proposed triage system. Since model predictions are sometimes used as diagnostic tools, it is particularly important to understand the strengths and weaknesses of a model, as well as methods for estimating error rates.

Logistic regression relies heavily on software. Even the simplest models cannot be fit by hand; direct formulas for parameter estimates and standard errors do not exist. Consistent with earlier chapters, the treatment here emphasizes interpretation of both models and computer output for estimated models. For students interested in working directly with data the chapter labs contain R-based exercises that illustrate how to fit and interpret models to data.

Logistic regression has also become an important tool in data exploration and detecting patterns in data and is now widely used in machine learning. There is not space here to explore those ideas, but Chapter 9 of *OpenIntro Statistics, 4th ed.* examines building a logistic regression explanatory model for possible bias in the review of resumes submitted for a listed job opening. That material can serve as an introduction to data exploration with logistic regression.

9.2 Introduction to simple logistic regression

9.2.1 The model for simple logistic regression

Hyperuricemia is the presence of abnormally high levels of uric acid in the blood, a condition than can lead to kidney stones and gout; hyperuricemia may also be responsible for chronic kidney disease, cardiovascular disease, and other metabolic disorders. According to current criteria, men are diagnosed as having hyperuricemia if a uric acid measurement is at least as high as 416μmol/L. The cutoff for women is 360μmol/L. Research suggests that risk of hyperuricemia is correlated with the consumption of red meat, seafood, and beans. Hyperuricemia is more common in high-income countries and economically developing countries with Western diets (characterized by high daily intake of saturated fats, animal protein, sodium, and refined sugars). The prevalence of hyperuricemia ranges between 15% and 25% in Asian countries.

Hyperuricemia is present without symptoms approximately 30% of the time, so it would be useful to identify clinical measurements indicative of hyperuricemia; i.e., measurements signaling that a patient should have their uric acid level tested.

Wang, et al.[1] report a cross-sectional study examining the association of hyperuricemia with dietary magnesium in 5,168 participants in China. The study measured several other possible predictors, including body mass index (BMI,measured in kg/m^2.). Some literature has suggested that BMI has a strong association with hyperuricemia in various populations. This section explores that relationship in a random sample of 500 participants from the Zeng study. The full dataset (`hyperuricemia`) and the random sample (`hyperuricemia.samp`) are in the data package `oibiostat`.

Figure 9.1 shows the presence of hyperuricemia on the y-axis and BMI on the x-axis. The light blue dots represent (x_i, y_i) pairs for each individual in the sample of 500, where x_i is an individual's BMI and y_i equals 0 (hyperuricemia absent) or 1 (hyperuricemia present). A small amount of random noise has been added to the y-values (referred to as "**jittering**") to make it easier to see where the points are most densely clustered.

The blue dots at $y = 0$ cluster between BMI values of about 17 to 30, while the dots at $y = 1$ are most concentrated around BMI 23 to 28, confirming that hyperuricemia is associated with larger values of BMI. It still difficult, however, to see enough details to judge the strength of association from this plot. For example, while this plot clearly implies that an individual with BMI lower than 22 is unlikely to have hyperuricemia (since practically all points with BMI less than 22 have $y = 0$), it is not clear how to judge the risk of hyperuricemia for individuals with moderate values of BMI, since points with BMI around 25 exist at both $y = 0$ and $y = 1$.

Computing summary measures can provide further insight about the association between hyperuricemia and BMI.

[1] Chao Zeng et al. "Association between low serum magnesium concentration and hyperuricemia". In: *Magnesium research* 28.2 (2015), pp. 56–63.

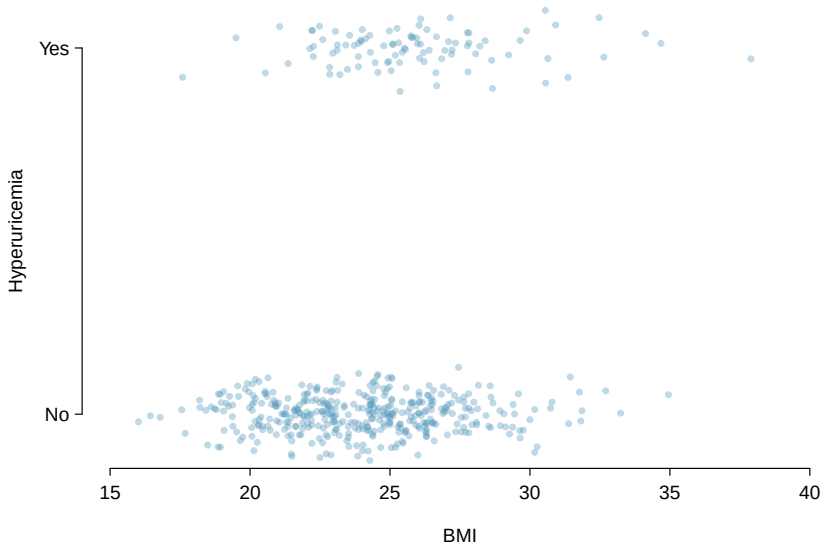


Figure 9.1: Presence of hyperuricemia versus BMI. For each case of $y_i = 1$ if hyperuricemia is present (labeled Yes on the vertical axis) and $y_i = 0$ if hyperuricemia is absent (labeled No). The y values have been jittered. The mean BMI in each group is marked by an "x".

EXAMPLE 9.1

The World Health Organization (WHO) labels BMI ≥ 30 as obese and $25 \leq$ BMI < 30 as overweight or pre-obese.[a] In the sample of 500 participants from the Zeng study, 204 individuals had BMI ≥ 25. Of these individuals, 57 had hyperuricemia. Compute the probability and odds that a study participant with BMI ≥ 25 has hyperuricemia.

(E)

Among these 204 participants, if 57 had hyperuricemia then the estimated conditional probability of hyperuricemia in this group is $57/204 = 0.279$. Odds as a summary measure for binary data are discussed in Section 8.5.3. Briefly, the odds of an outcome is the ratio of the number of times an outcome occurs divided by the number of times it does not; thus, the odds of hyperuricemia in these 204 study participants equals $57/(204 - 57) = 0.388$.

[a] See Section 9.7 for a discussion on the use of these cut-points in Asian populations

In the sample of 500 individuals, 95 were hyperuricemic and 405 were not, so the estimated probability and odds of hyperuricemia based on the sample of 500 are $95/500 = 0.190$ and $95/405 = 0.235$, respectively. An individual drawn at random from the entire study sample has a lower probability of being hyperuricemic than an individual drawn at random from the participants with BMI ≥ 25: probability 0.235 versus 0.279. Thus, these data suggest an association between BMI and hyperuricemia; specifically, that larger BMI is associated with increased risk of hyperuricemia. This is consistent with the trend visible in Figure 9.1.

Figure 9.2 shows the prevalence of hyperuricemia by quintile of BMI. Quintiles divide the study sample into five groups of equal size, so each row of the table has 100 observations. With increasing BMI quintile, the estimated probability and odds of hyperuricemia increase. In the lowest quintile, in which average BMI is 20.08, the probability of hyperuricemia is 0.05 and the odds of hyperuricemia are 0.053. In the highest quintile, in which average BMI is 28.92, the probability and odds of hyperuricemia are larger: 0.32 and 0.471, respectively.

Probabilities and odds are not identical but they provide similar information. Odds and probabilities increase or decrease together, and one can be calculated from the other. If p is the probability of an event, $p/(1-p)$ are the odds. Algebra can be used to show that

$p = \text{odds}/(1 + \text{odds})$.

BMI Quintile	Mean BMI	HU Absent	HU Present	Est. Probability	Est. Odds
1	20.08	95	5	0.05	0.053
2	22.55	85	15	0.15	0.176
3	24.32	82	18	0.18	0.220
4	25.84	75	25	0.25	0.333
5	28.92	68	32	0.32	0.471

Figure 9.2: Hyperuricemia (HU) by quintiles of BMI. Each row provides information within a specific BMI quintile: average BMI, number of individuals with and without hyperuricemia, and the estimated probability and estimated odds of hyperuricemia.

GUIDED PRACTICE 9.2

G Using the algebraic relationship between probability and odds, show that if the probability of hyperuricemia is 0.05, the odds of hyperuricemia are 0.053. Additionally, show that if the odds of hyperuricemia are 0.471 then the probability equals 0.32.[2]

Dividing the study sample into smaller groups and computing summary measures will provide more detail about how risk of hyperuricemia varies with individual BMI values. The dark blue circles in Figure 9.3 represent information obtained from grouping individuals into 2^{nd}-percentiles, just as Figure 9.2 groups individuals by every 20^{th} percentile; i.e., the 500 cases have been split into 50 groups of 10 cases per group. Each dark blue circle has x-value equal to the mean BMI within the group and y-value equal to the proportion of individuals with hyperuricemia within the group. The dark blue circles more clearly demonstrate that larger BMI tends to be associated with increased estimated probability of hyperuricemia than the light blue circles representing the observed data.

The green line in Figure 9.3 is the least squares line for a model predicting hyperuricemia from BMI. Since the mean of a binary variable taking on values 0 and 1 equals the estimated probability of the variable taking on the value 1 (i.e., the proportion of times that $y = 1$), the linear model estimates the probability of hyperuricemia at each value of BMI. While the line mostly fits the data reasonably well, it shows a lack of fit at the smallest BMI values where it predicts probabilities less than 0.

The least squares line in Figure 9.3 is based on the model

$$\begin{aligned} E(Y_i) &= P(Y_i = 1) \\ &= \beta_0 + \beta_1(\text{bmi}), \end{aligned}$$

where Y_i has value 1 when hyperuricemia is present and 0 otherwise. The green line drops below 0 for the smaller values of BMI because the linear model does not restrict predicted values to lie between 0 and 1. The red curve, which shows a model-based estimate of the probability of hyperuricemia using logistic regression, is a better fit to the data across the range of BMI values.

Suppose E is an event, x is a predictor, and $p_E(x)$ is the conditional probability of E given x. The odds of E given x are $p_E(x)/(1 - p_E(x))$. The simple logistic regression model for the odds of E given x is a linear model on the log(odds) scale. Just as in least squares linear regression, the right-hand side of the model is a linear combination of parameters (the intercept and slope) and x:

$$\log\left(\frac{p_E(x)}{1 - p_E(x)}\right) = \beta_0 + \beta_1 x, \tag{9.3}$$

[2] If $p = 0.05$, compute the odds as $p/(1 - p) = 0.05/(1 - 0.05) = 0.053$. If the odds are 0.471, compute the probability as $\text{odds}/(1 + \text{odds}) = 0.471/(1 + 0.471) = 0.32$.

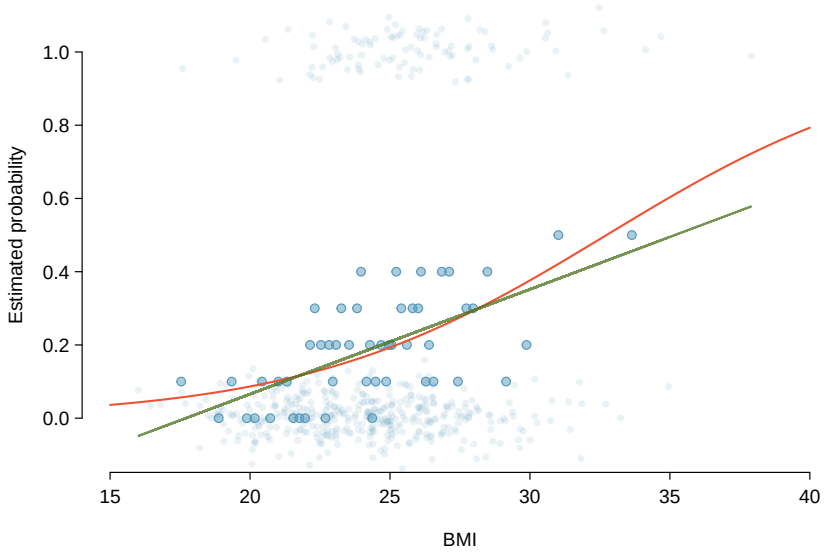


Figure 9.3: Estimated probability of hyperuricemia versus BMI. The small light blue dots show observed (x_i, y_i) pairs. Each large blue dot represents the proportion of individuals with hyperuricemia in each 2^{nd}-percentile; i.e., each group when the sample is divided into 50 groups based on BMI. The green line is the least squares model for hyperuricemia versus BMI. The red curve is a logistic model for hyperuricemia versus BMI.

or, equivalently,

$$\log(\text{odds}_E(x)) = \beta_0 + \beta_1 x. \tag{9.4}$$

Exponentiating both sides of Equation 9.4 yields

$$\begin{aligned} \text{odds}_E(x)) &= \exp(\beta_0 + \beta_1 x) \\ &= \exp(\beta_0)\exp(\beta_1 x). \end{aligned} \tag{9.5}$$

If Y is a binary variable with value 1 when E occurs and 0 otherwise, Equation 9.5 is a model for the odds that $Y = 1$, given x.

Probabilities can be estimated using the relationship

$$\begin{aligned} p_E(x) &= \frac{\text{odds}_E(x)}{1 + \text{odds}_E(x)} \\ &= \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)}. \end{aligned} \tag{9.6}$$

Software used to estimate logistic regression usually provides estimates for log(odds) in the form of Equation 9.3, and the conversion to odds or probabilities is done with either a separate step in the program or by hand.

GUIDED PRACTICE 9.7

Suppose the logistic regression model for an event E is given by

$$\begin{aligned}\log(\text{odds}_E(x)) &= \beta_0 + \beta_1 x \\ &= 0.5 - 0.75x.\end{aligned}$$

Calculate the odds and probability of E when $x = 1.0$. [3]

Computer algorithms that estimate the parameters in logistic regression use the method of **maximum likelihood**. Since a logistic regression model can be converted to a model for the probability of an event E given set of predictors, these probabilities can be used to write an algebraic expression for the probability of a set of observed responses given the predictors (details shown in more advanced courses). This expression is called the likelihood of the data; the method of maximum likelihood selects estimates for β_0 and β_1 that make the likelihood as large as possible.

The estimated logistic regression model shown in the red curve in Figure 9.3 is explored in Section 9.2.3.

The log in Equation 9.3 is $\log_e$, the natural logarithm function. Since the natural log is used often in statistics, the subscript e is usually omitted. The transformation $\log(\frac{p}{1-p})$ has its own name, the **logit function**.[4]

[3] The log(odds) are $0.5 - 0.75(1) = -0.25$, so the odds and probability are, respectively, $\exp(-0.25) = 0.779$ and $0.779/(1 + 0.779) = 0.438$.

[4] Specifically, $\text{logit}(p) = \log(\frac{p}{1-p})$.

9.2.2 Interpreting model parameters

Figure 9.4 shows the relationship between probability and the value of a predictor x for four different models of the form specified by Equation 9.6. The model coefficients (β_0, β_1) are (-3.0, 0.6) for the solid line, (-3.0, 0.8) for the dashed line, (3.0, -0.6) for the dotted line, and (-0.4, 0.0) for the horizontal line.

The model parameter β_1 determines the relationship between predicted probabilities and values of the predictor x. The solid and dashed lines show a positive association; when $\beta_1 > 0$, probabilities increase with increasing values of the predictor x. Since odds and probabilities increase together, positive values of β_1 indicate that the odds of an event increase with increasing values of x. A larger positive value for β_1 indicates of a stronger positive association. The dashed line, which has a larger β_1 than the solid line, shows a steeper incline in the center of the graph. Probabilities change more rapidly with changing values of x. The dotted line shows a negative association; when $\beta_1 < 0$, probabilities and odds decrease with increasing values of x. Probability starts out near 1 when x is small, then decreases to near 0 once x increases to 10. The horizontal line with $\beta_1 = 0$ shows no association between the event and values of x.

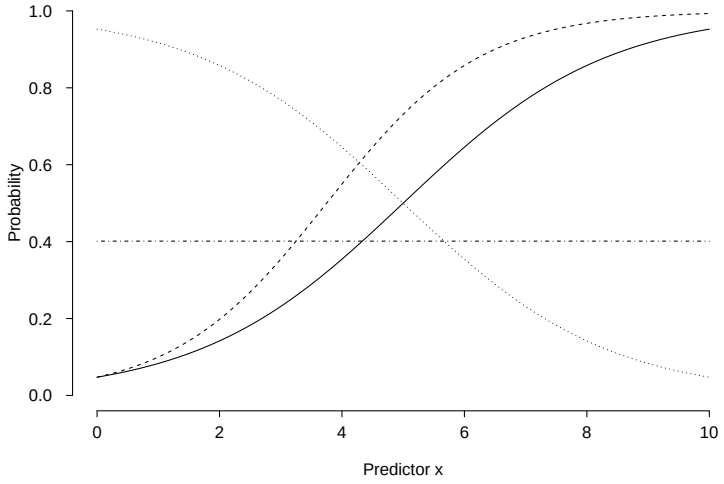


Figure 9.4: Probability versus a predictor x for four models of the form specified by Equation 9.6. The model coefficients (β_0, β_1) are (-3.0, 0.6) for the solid line, (-3.0, 0.8) for the dashed line, (3.0, -0.6) for the dotted line, and (-0.4, 0.0) for the horizontal line.

9.2.3 Hyperuricemia and BMI

If E is hyperuricemia and x = bmi, the logistic regression model for the association between hyperuricemia and BMI is

$$\log\left[\frac{p(E|\text{bmi})}{1-p(E|\text{bmi})}\right] = \beta_0 + \beta_1(\text{bmi}),$$

or, equivalently,

$$\log(\text{odds}_E(\text{bmi})) = \beta_0 + \beta_1(\text{bmi}). \tag{9.8}$$

Figure 9.5 shows the result of using R to estimate the coefficients in Equation 9.8. The 'Intercept' is the estimate b_0 of β_0 and the term labeled 'bmi' is the estimate b_1 of β_1.

Intercept	bmi
-6.054	0.185

Figure 9.5: Estimated logistic regression coefficients for the association of hyperuricemia with BMI.

Expressed algebraically, the estimated model is

$$\log(\widehat{\text{odds}}_{\text{E}}(\text{bmi})) = -6.054 + 0.185(\text{bmi}). \tag{9.9}$$

The red curve in Figure 9.3 is drawn using this estimated model after log(odds) were converted to probabilities, just as in Guided Practice 9.7. For example, a member of the study population with BMI 30.0 has an estimated log(odds) of hyperuricemia of $-6.054 + (0.185)(30) = -0.504$. To compute the odds, exponentiate the estimated log odds: $\exp(\log(\widehat{\text{odds}})) = \exp(-0.504) = 0.604$. Then, convert from odds to probability: the predicted probability of hyperuricemia for an individual with BMI 30.0 is $0.604/(1 + 0.604) = 0.377$. If these data represent a random sample from a large population, about 38% of individuals with BMI = 30 are predicted to have hyperuricemia.

Just as with 2×2 tables, probabilities can be estimated with logistic regression in either cross-sectional studies or studies with exposure based sampling; the hyperuricemia study was a cross-sectional study, so probabilities can be estimated using the estimated model. This issue is discussed in detail for 2×2 tables in Section 8.6.6 in the web supplement and is part of the assumptions for logistic regression listed in Section 9.3.

The coefficient 0.185 has an interpretation similar to a slope in linear regression: every one unit change in BMI is associated with an additive increase of 0.185 in the log odds of hyperuricemia.

EXAMPLE 9.10

Suppose two members of the study population have BMI values 30.0 and 33.2. What is the estimated relative odds for hyperuricemia (i.e., the odds ratio), comparing the individual with BMI = 33.2 to the one with BMI = 30.0?

When BMI = 33.2, the estimated log odds of hyperuricemia are

$$\log[\widehat{\text{odds}}_{E}(\text{bmi} = 33.2)] = -6.054 + (0.185)(33.2) = 0.088,$$

(E)

and the estimated odds of hyperuricemia are $\exp(0.088) = 1.092$. The estimated odds of hyperuricemia for an individual with BMI 30.0 are 0.604 (calculated earlier).

The estimated OR comparing these two individuals is $1.092/0.604 = 1.808$. The odds of hyperuricemia are estimated to be 1.8 times as large for an individual with BMI 33.2 versus an individual with BMI 30.0. This model is consistent with the data in Figure 9.2 and suggests there is indeed a strong association between BMI and the odds of hyperuricemia, as others have found. The tools of inference discussed in Section 9.3 will show that this association is stronger than would be expected by chance alone under the assumption the null hypothesis of no association is true.

Odds ratios can be calculated directly from the coefficients in the model. Since the model for logistic regression is

$$\log(\text{odds}(x)) = \beta_0 + \beta_1 x,$$

the difference in log odds for two values x_1 and x_2 is

$$\log[\text{odds}(x_2)] - \log[\text{odds}(x_1)] = \beta_1(x_2 - x_1).$$

The relationship

$$\log(b) - \log(a) = \log(b/a)$$

implies that

$$\log\left[\frac{\text{odds}(x_2)}{\text{odds}(x_1)}\right] = \beta_1(x_2 - x_1)$$

and

$$\frac{\text{odds}(x_2)}{\text{odds}(x_1)} = \exp[\beta_1(x_2 - x_1)]. \tag{9.11}$$

Suppose two members of a population have BMI values given by $x_1 = \text{bmi1}$ and $x_2 = \text{bmi2}$. The estimated odds ratio comparing these two individuals is

$$\begin{aligned}\widehat{\text{OR}} &= \text{odds(bmi1)/odds(bmi2)} \\ &= \exp[0.185(\text{bmi2} - \text{bmi1})].\end{aligned}$$

If two values of BMI differ by 1, the odds ratio (OR) will be $e^{0.185} = 1.20$. For every one unit increase in `bmi`, the odds changes by a factor of 1.20. When calculating a change in odds using the model coefficients, the intercept plays no role, just as in similar calculations in linear regression. More generally, in the model in Equation 9.3, β_1 and $\exp(\beta_1)$ are, respectively, the difference in log(odds) and the OR between two cases when x changes by 1 unit.

GUIDED PRACTICE 9.12

G Suppose two members of the study population have a BMI of 26 and 28, respectively. Calculate the odds of hyperuricemia for each of them using model 9.9. Calculate the relative odds (i.e., odds ratio) for an individual with BMI 28 compared to BMI 26. [5]

GUIDED PRACTICE 9.13

G Calculate the relative odds of hyperuricemia for the two individuals with BMI 26 and 28 by using the coefficients in the logistic regression model directly, i.e., without calculating the individual odds. [6]

The model can also be used to estimate prevalence ratios as discussed in Section 8.6.1.

[5] The odds of hyperuricemia for the two individuals are $\exp[-6.054+(0.185)(26)] = 0.288$ and $\exp[-6.054+(0.185)(28)] = 0.417$. The relative odds are $0.417/0.288 = 1.45$.

[6] Using the model coefficient, the relative odds is $\exp[(2)(0.185)] = 1.45$.

EXAMPLE 9.14

What is the estimated prevalence (i.e. probability) of hyperuricemia for two individuals with BMI 30.0 and 33.2? What is the estimated prevalence ratio for hyperuricemia, comparing the individual with BMI = 33.2 to the one with BMI = 30.0?

As mentioned earlier, the hyperuricemia data were collected in a cross-sectional study, so probabilities can be estimated (estimated probabilities were used to construct the red curve in Figure 9.3).

For these two individuals, the estimated probabilities of hyperuricemia are

(E)

$$\hat{p}_E(33.2) = \frac{1.092}{1+1.092} = 0.522$$

and

$$\hat{p}_E(30.0) = \frac{0.604}{1+0.604} = 0.377.$$

The prevalence ratio, comparing the participant with BMI = 33.2 to the one with BMI = 30.0 is 0.522/0.377 = 1.38; the prevalence of hyperuricemia for individual with BMI = 33.2 is estimated to be almost 1.4 times (40% larger) that of the individual with the lower BMI. Using the language of Section 8.6, the relative risk of hyperuricemia for an individual with a BMI of 33.2 vs 30.0 is approximately 1.4.

9.2.4 Checking model fit, hyperuricemia and BMI

This section describes a graphical method for checking the fit of a logistic model with a single continuous predictor, such as BMI. Methods for checking fit that use the inferential properties of logistic regression are discussed in Section 9.5.

Figure 9.6 shows values of the outcome variable $Y = 0$ (no hyperuricemia) or $Y = 1$ (hyperuricemia) plotted against model predicted probabilities. It is the analogue of plotting observed versus predicted values in linear regression, but because all the observed values are clustered at 0 or 1, it is less useful as a diagnostic than in linear regression. As noted earlier, close inspection of the plot indicates that larger predicted probabilities tend to have a increased frequency of $Y = 1$, but the trend is subtle.

Grouping observations reduces the variability in a plot and can sometimes be helpful in checking a model. Figure 9.22 shows the same plot as in Figure 9.6, but with the addition of summary statistics computed within 10 equally sized buckets of size 50. Each group is formed based on the predicted probability of hyperuricemia. For instance, the left-most point represents the group consisting of the 50 cases with the smallest predicted probabilities of hyperuricemia based on the model, which range between 0.043 to 0.091. Within this group, 2 individuals (a proportion of 2/50 = 0.04) were hyperuricemic and the average predicted probability was 0.076, so the point is at (0.076, 0.040). The vertical lines show 95% confidence intervals for each estimated proportion. If the logistic regression is a good fit, the estimated proportions and average predicted probabilities should be similar in each decile; the dashed line $y = x$ shows the extent to which the observed proportions and predicted probabilities agree. Since all of the confidence intervals touch the dashed line, the model seems to fit reasonably well.

With larger datasets, it is possible to obtain a clearer picture of the fit by increasing the number of buckets and/or the number of observations in each bucket.

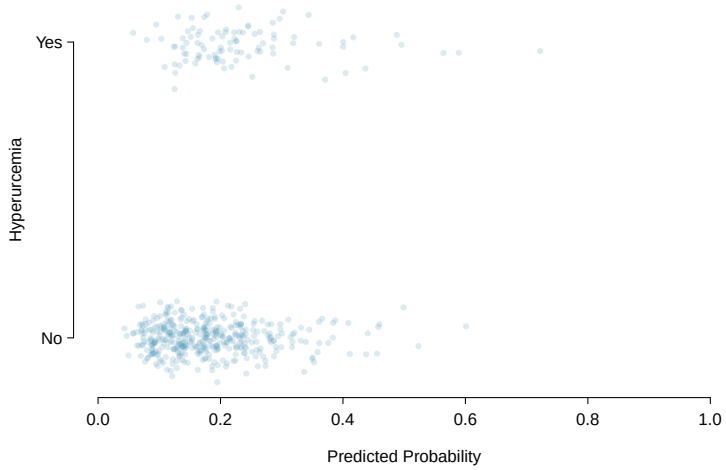


Figure 9.6: Predicted probabilities versus observed values of hyperuricemia.

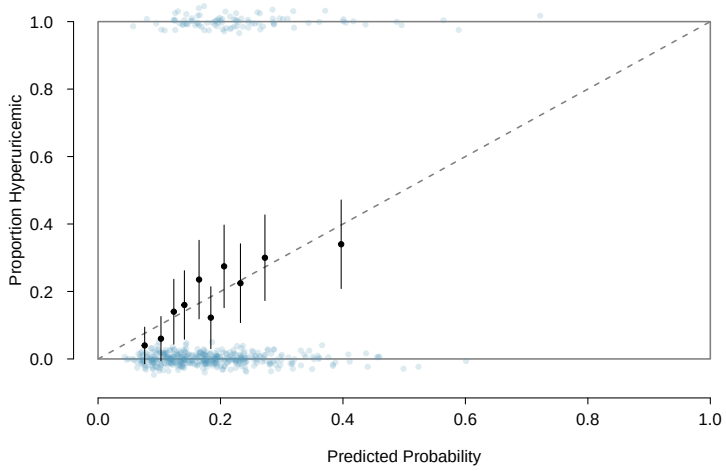


Figure 9.7: Predicted probabilities versus observed proportions, with data grouped according into 10 equal sized buckets of predicted probabilities. The light blue dots at $y = 0$ and $y = 1$ denote observed values of hyperuricemia (0 = "No", 1 = "Yes") plotted against predicted probabilities.

Figure 9.22 is a type of **calibration plot** discussed in more detail in Section 9.5.

9.3 Inference for simple logistic regression

How strong is the evidence for the association between BMI and hyperuricemia?

All models estimated from data have inherent uncertainty in the estimated parameters. The standard errors of estimated parameters are a reminder to pay attention to the margin of error of statistical estimates. Just as in linear regression, standard errors are used to calculate test statistics and confidence intervals.

Confidence intervals and tests for parameters in simple logistic regression will be valid if the assumptions behind the model are met, at least approximately.

ASSUMPTIONS FOR SIMPLE LOGISTIC REGRESSION

Let E be an event and Y a binary response variable that is 1 if E has occurred and 0 if not. Let X be a predictor thought to be related to the occurrence of E. A sample of observations $(x_1, y_1), (x_2, y_2) \ldots (x_n, y_n)$ can be used to estimate the log(odds) of the occurrence of E (equivalently that $Y = 1$) given $X = x$ using model 9.3 under the following conditions:

1. The logistic transformation is thought to be a reasonable model for the dependence of conditional probability or odds for the response variable given the predictor.
2. The observations are independent pairs, i.e., no single pair depends on any of the others.
3. If the sample was drawn using exposure-based or cross-sectional sampling, the conditional odds and probability of E given x can be estimated using relationships 9.5 and 9.6. These estimates can be used to estimate odds and prevalence ratios.
4. If the data are from a case-control study (i.e., outcome-based sampling) in which the sampling did not depend on exposure, conditional odds can be estimated but conditional probabilities cannot. Odds ratios can be estimated from the model, but prevalence ratios cannot.

Assumption 1 is more difficult to check than the usual linearity assumption in linear regression, but for continuous predictors such as BMI, scatterplots such as Figure 9.3 or Figure 9.6 can be helpful. Other diagnostic plots can be found in more advanced texts. For binary predictors, the model is generally reasonable.

Assumptions 2 - 4 depend on the study design. Assumption 2 is the standard assumption of independent observations. Assumptions 3 and 4 are analogous to the connection between study design and parameters that can be estimated in an analysis of 2×2 tables, where the usual calculation of risk ratio leads to a biased estimate in case-control studies. The formula for transforming an odds to a probability in a logistic model can be calculated but leads to incorrect estimates of probabilities. Section 8.6.6 contains a discussion of this issue in 2×2 tables.

In the logistic model given by Equation 9.3, a test of the null hypothesis $\beta_1 = 0$ is a test of no association between the predictor x and the odds or the probability of E; i.e., a test of the null hypothesis that x provides no information for predicting E.

As with all statistical models, tests and intervals are based on the sampling distributions of estimated parameters.

SAMPLING DISTRIBUTIONS OF ESTIMATED COEFFICIENTS

Suppose

$$\widehat{\log(\text{odds}_E(x))} = b_0 + b_1 x$$

is an estimated logistic regression model from a dataset with n observations on the outcome E and predictor x. The standardized statistic

$$\frac{b_1 - \beta_1}{\text{s.e.}(b_1)}$$

has a standard normal (z) distribution in moderate to large sample sizes.
Consequently, under the hypothesis $H_0 : \beta_1 = 0$, the statistic

$$\frac{b_1}{\text{s.e.}(b_1)}$$

has a standard normal (z) distribution in moderate to large sample sizes.

The sampling distribution for the estimated regression coefficient b_1 does not depend on the sample size n, unlike the t-based sampling distribution for a regression coefficient in linear regression, where the degrees of freedom depends on the sample size. One useful guideline for an adequately-sized sample is that there should be at least 10 cases in the dataset with the less frequent yes/no outcome.

TESTING A HYPOTHESIS ABOUT A LOGISTIC REGRESSION COEFFICIENT

A test of the two-sided hypothesis

$$H_0 : \beta_1 = 0 \text{ vs. } H_A : \beta_1 \neq 0$$

is rejected with significance level α when

$$\frac{|b_1|}{\text{s.e.}(b_1)} > z^\star,$$

where $z^\star$ is the point on a z-distribution with area $(1 - \alpha/2)$ in the left tail.

For one-sided tests, $z^\star$ is the point on a z-distribution with area $(1 - \alpha)$ in the left tail. A one-sided test of H_0 against $H_A : \beta_1 > 0$ rejects when the standardized coefficient $b_1/\text{s.e.}(b_1)$ is greater than $z^\star$; a one-sided test of H_0 against $H_A : \beta_1 < 0$ rejects when the standardized coefficient is less than $-z^\star$.

CONFIDENCE INTERVALS FOR A LOGISTIC REGRESSION COEFFICIENT

A two-sided $100(1 - \alpha)\%$ confidence interval for the model coefficient β_1 is

$$b_1 \pm [\text{s.e.}(b_1) \times z^\star].$$

All statistical software packages provide standard errors (s.e.) of coefficients, and most provide the z statistic and its p-value directly. The estimate b_0 has a sampling distribution as well, but since the coefficient is often not scientifically meaningful, tests and intervals for β_0 are not discussed here.

Inference for the association of BMI with hyperuricemia can be based on the more complete

table of output from R shown in Figure 9.8 (output has been rounded to two or three significant digits for readability). The assumptions for logistic regression seem reasonable for this example. Figure 9.3 suggests that the probability of hyperuricemia follows a logistic function as BMI increases, and assumptions 2 and 3 are satisfied since this was a cohort study with independent data from the participants. In the sample of 500, 95 were hyperuricemic and 405 were not, so the sample size is sufficient.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-6.054	0.947	-6.39	< 0.001
bmi	0.185	0.037	4.99	< 0.001

Figure 9.8: Logistic regression with response variable hyperuricemia and predictor BMI.

The inferential results show that the positive association between BMI and log(odds) (and consequently the odds) of hyperuricemia is statistically significant ($p < 0.001$, z statistic 4.99); i.e., the observed association is larger than would be expected by chance if there were no population level association. The data in the Zeng study support the increased prevalence of hyperuricemia with increasing BMI found in other studies and populations.

As always, p-values and parameter estimates in should be interpreted with care, but there are issues that arise in observational studies. Estimates of association should not be given a causal interpretation, and even estimates of association may be subject to confounding. It is common in observational studies to examine more than one association, leading to the possibility of inflated type I error from multiple testing. The hyperuricemia study was primarily intended to study the association between hyperuricemia and dietary magnesium, not hyperuricemia and BMI. The analysis presented here is not one planned by the study team.

Confidence intervals for estimated parameters are more informative than z statistics and p-values and are the preferred method for conveying inferential results. However, confidence intervals are subject to the same sources of bias and lack of generalizability as test statistics and should also be interpreted with caution.

Confidence intervals for β_1 in logistic regression are on the log(odds) scale and not easily interpreted. Exponentiating the lower and upper bounds of a confidence interval for β_1 yields a confidence interval for $\exp(\beta_1)$ on the odds scale.

In the hyperuricemia example, the 95% confidence interval for the coefficient of BMI on the odds scale:

$$0.185 \pm (1.96)(0.037) \longrightarrow (0.113, 0.258) \longrightarrow (e^{0.113}, e^{0.258}) = (1.119, 1.294).$$

These data suggest that with 95% confidence, an increase of 1 unit BMI is associated with a larger odds of hyperuricemia by a factor of 1.1 to 1.3.

EXAMPLE 9.15

Calculate and interpret a 95% confidence interval for the odds ratio of hyperuricemia comparing two individuals with BMI 33 and 30.

First compute a confidence interval for $3\beta_1$, then exponentiate the endpoints of the interval to convert to the odds scale. The estimated log odds ratio for participants whose BMI differ by 3 is $3b_1 = (3)(0.185) = 0.555$. The standard error for $3b_1$ can be computed based on rules for linear transformations of random variables. Since $\text{Var}(aX) = a^2\text{Var}(X)$ (where a is a constant and X is a random variable), $\text{SD}(3b_1) = (3)\text{SD}(b_1) = (3)(0.037) = 0.111$. Thus, the 95% confidence interval for the OR for two individuals with BMI values that differ by 3 is calculated as

$$0.555 \pm (1.96)(0.111) \longrightarrow (0.337, 0.773) \longrightarrow (1.401, 2.165).$$

(E) Since computing a confidence interval for $a\beta_1$ on the log(odds) scale involves multiplying both b_1 and its standard error by a factor of a, the confidence interval for $a\beta_1$ can be obtained by simply multiplying both endpoints of the confidence interval for β_1 by a:

$$((0.113)(3), (0.258)(3)) = (0.339, 0.774).$$

This interval differs slightly from the one computed previously only due to rounding of the original confidence interval bounds. If no rounding is done in the intermediate calculations, the confidence interval on the odds scale is $(1.401, 2.165)$.

These data suggest that with 95% confidence, the odds ratio of hyperuricemia for participants with a BMI of 33 versus 30 is between 1.40 and 2.17. The individual with BMI larger by 3 units has a odds of hyperuricemia that may be from 1.40 to 2.17 times higher. This confidence interval depends only on the difference in the values of BMI, so it applies to any two values of BMI that differ by 3.

GUIDED PRACTICE 9.16

(G) Calculate a 99% confidence interval for the odds ratio of hyperuricemia comparing two individuals with BMI 29 and 31. [7]

The above examples illustrate confidence intervals for the slope parameter. Confidence intervals for (predicted) odds and probabilities are more difficult and not discussed in this text. Since odds are estimated using $\exp(b_0 + b_1\text{bmi})$, the standard error for the estimate uses the sampling distribution of each of the estimated coefficients and the their correlation, something that is not covered in this chapter. The same is true for estimates of probabilities.

9.3.1 The connection between logistic regression and the χ^2 test

Tuberculosis (TB) is a communicable disease that is among the top 10 causes of death worldwide; it is the leading cause of death from a single infectious agent. [8] Despite the virulent nature of the disease, it is often treatable. If the disease is diagnosed early and treated with

[7] The estimate and standard error for $2(\beta_1)$ are, respectively, $(2)(0.185) = 0.370$ and $(2)(0.037) = 0.074$. For a 99% interval $z^\star = 2.58$ so the interval is calculated as

$$0.370 \pm (2.58)(0.074) \longrightarrow (0.179, 0.561) \longrightarrow (1.196, 1.752).$$

[8] World Health Organization et al. "Global tuberculosis report 2019. 2020". In: *Geneva: World Health Organization* (2020).

effective antibiotics for six months, it can be cured, preventing further infections in others. Unfortunately, many patients are not able to complete the six to eight month course of TB therapy, leading to further spread of the disease. Treatment interruptions and premature endings are particular problems in low and middle income countries.

The World Health Organization (WHO) and other health care organizations have used the term *treatment default* in TB to denote a treatment interruption of at least two months, and nearly all published papers use that term. This chapter uses the more descriptive term *two-month interruption* for the premature ending of treatment. When the context is clear, this is shortened to *interruption*.

A 2015 cross-sectional study by Lackey, et. al.[9] examined patient characteristics associated with interrupted treatment in a section of Lima, Peru where the incidence of TB was 213 cases per 100,000 persons at the time the study was conducted. For comparison, the incidence of TB in the United States is approximately 2.5 cases per 100,000.[10] The study enrolled 1,294 participants and reported results based on data from 1,233 participants for whom there were no missing data on outcome and patient characteristics. Figure 1 in the Lackey article describes the criteria for exclusions that led to the data from 1,233 participants used in their analysis. **Complete case analysis** is the term used to refer to an analysis using only the cases without any missing observations; while this is often not the best way to adjust for missing data, alternative methods are beyond the scope of this text. The dataset `tb.interruption` in the `oibiostat` package contains data on 1,293 of the 1,294 all the participants enrolled; data from one participant whose treatment was stopped prematurely by the clinical team was dropped before the dataset was posted by the study team.

[9] Brian Lackey et al. "Patient characteristics associated with tuberculosis treatment default: a cohort study in a high-incidence area of Lima, Peru". In: *PLoS One* 10.6 (2015), e0128541.

[10] **https://www.cdc.gov/tb/statistics/default.htm**.

EXAMPLE 9.17

Figure 9.9 shows a logistic regression model estimating the association of a two-month treatment interruption among participants who had completed a secondary school education. (Decimals from the output have been rounded to 3 significant figures for readability.) Interruption (the variable `two.mo.interruption` in the dataset) is a binary variable coded `0` for individuals who completed therapy and 1 for those who did not. The predictor `education` is a factor variable, with levels "Yes" and "No" for participants with and without secondary school education, respectively. Among the 1,233 cases in the dataset, 127 (10.3%) experienced a treatment interruption and 719 had at least a secondary school education. Compute the odds ratio for ending TB therapy prematurely, comparing participants with a secondary school education to those without, along with a 95% confidence interval for the odds ratio.

The assumptions for the logistic model are reasonable in this example. The participants were sampled independently, the predictor is binary, and there are more than 10 cases with either outcome. The coefficient of `educationYes` indicates that participants with secondary school education have a log(odds) that is reduced additively by 0.785 compared to those without secondary school education. The odds ratio comparing someone with secondary school education to someone without is $e^{-0.785} = 0.456$. The odds of a premature treatment interruption among participants with a secondary school education are 0.456 times the odds of those without a secondary education. The odds are reduced by more than 50%.

Because the z statistic has value -4.12, the evidence for an association is strong ($p < 0.001$). A 95% confidence interval for the odds ratio can be calculated by first calculating the corresponding interval for the log(OR) and exponentiating. The 95% confidence interval for the log(OR) is

$$-0.785 \pm (1.96)(0.191) \longrightarrow (-1.159, -0.411).$$

The confidence interval for the odds ratio is

$$(e^{-1.158}, e^{-0.411}) = (0.314, 0.663).$$

Individuals with secondary school education have a lower relative odds of treatment interruption than those without; with 95% confidence, the odds of interruption may be from 0.314 to 0.663 times lower in individuals with a secondary education. This is sometimes phrased as an odds that is 34% to 69% ((100 – 66.3)% to (100 – 31.4%)) lower.

Confidence intervals for odds ratios can also be calculated using the methods in Section 8.6.4, although answers may differ slightly because of the different formulas.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.767	0.125	-14.14	0.000
educationYes	-0.785	0.191	-4.12	0.000

Figure 9.9: Estimated logistic regression, the association of two-month treatment interruption with secondary school education.

The association between treatment interruption and secondary school education in the logistic regression model is evident in a 2×2 table (Figure 9.10). Among 514 participants without a secondary school education, 75/514 = 14.6% experienced a treatment interruption, while 52/719 = 7.2% participants with a secondary education had an interruption.

	No Sec. Edu.	Sec. Edu.	Sum
No interruption	439	667	1106
Interruption	75	52	127
Sum	514	719	1233

Figure 9.10: Two-month treatment interruption by secondary school education.

GUIDED PRACTICE 9.18

G Using Figure 9.10, compute the odds ratio for treatment interruption comparing participants without and with a secondary school education and show that it is the same as the odds ratio calculated in the logistic regression, 0.456. [11]

The χ^2 value for the table (16.8 with one degree of freedom) is highly statistically significant ($p < 0.001$) as is the z statistic in the logistic regression in Figure 9.10. In the setting of a 2×2 table, logistic regression produces the same summary statistic for an association as a direct analysis of the table; this is analogous to how linear regression with a binary predictor provides the same results as a two-sample t-test.

Associations in observational studies should never be interpreted as causal effects and this example underscores that principle. Increasing access to secondary education in hopes of increasing successful completion of TB treatment may not change outcome; members of the population likely have many characteristics that enabled them to have access to both a secondary education and adequate health care.

[11] For participants without a secondary school education, the odds of treatment interruption are 75/439 = 0.171. For patients with at least a secondary school education, the corresponding odds are 52/667 = 0.078. The relative odds, or odds ratio, comparing those with a secondary school education to those without is 0.078/0.171 = 0.456.

9.4 Multiple logistic regression

9.4.1 Models with two predictors

The next sections introduce multiple logistic regression using examples with two predictors and categorical predictors with more than two levels. The more abstract discussion of the general logistic regression model and methods for inference for its parameters are deferred to Section 9.4.4.

Women are generally less likely to experience hyperuricemia than men for reasons that are not completely understood, but may be due to increased levels of estrogen.[12] Figure 9.11 shows that is the case in these data, where the estimated OR for hyperuricemia, comparing females to males is (34/213)/(61/192) = 0.5025. In these data, the odds of hyperuricemia in females is half what it is in males. Does the relationship between hyperuricemia and BMI in Figure 9.8 change when sex is added to the model?

	No	Yes	Sum
male	192	61	253
female	213	34	247
Sum	405	95	500

Figure 9.11: Table showing the association between hyperuricemia (No, Yes) and sex in the random sample of 500 participants from the hyperuricemia data

Let E denote hyperuricemia, and

$$p_E(\text{bmi}, \text{sex}) = P(E|\text{bmi}, \text{sex}).$$

The two-variable model used to answer this question is

$$\log\left[\frac{p_E(\text{bmi}, \text{sex})}{1 - p_E(\text{bmi}, \text{sex})}\right] = \beta_0 + \beta_1 \text{bmi} + \beta_2 \text{sex}. \tag{9.19}$$

The sample size guidelines for logistic regression outlined in Section 9.4.4 specify that the number of predictors in a model (including the intercept) should be no larger than 10% of the smaller of the number of successes or failures. There are 95 cases in the dataset with hyperuricemia (the smaller number of the two outcomes), so a model with 2 predictors meets the sample size guideline. The estimated model is shown in Figure 9.12. The factor sex is coded "male" (the baseline category) or "female", and the units of BMI are kg/m^2. The estimated regression indicates that BMI remains strongly associated with hyperuricemia after adjusting for sex.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-5.503	0.982	-5.61 0.000	
bmi	0.171	0.038	4.56	0.000
sexfemale	-0.480	0.245	-1.96	0.050

Figure 9.12: Logistic regression with response variable hyperuricemia predictors BMI and sex.

[12] Victoria L Halperin Kuhns and Owen M Woodward. "Sex differences in urate handling". In: *International journal of molecular sciences* 21.12 (2020), p. 4269.

The algebraic form of the estimated model is

$$\log(\widehat{\text{odds}}_E) = -5.503 + (0.171)\text{bmi} - (0.480)\text{sexfemale}. \tag{9.20}$$

A great deal can be learned about the interpretation of logistic regression from even this simple model. The coefficient of BMI can be used to calculate the estimated change in odds associated with a change in BMI as long as the sex variable remains constant, i.e., for participants of the same sex.

EXAMPLE 9.21

Calculate the OR for two individuals of the same sex but with BMI values of 28 and 29.

The estimated model coefficients can be used to calculate the difference in log(odds) for one unit change in BMI using the same steps that led to Equation 9.11. When the variable sex does not change, the difference in log odds for two values of `bmi` given by x_1 and x_2 is

E

$$[b_0 + b_1(x_2) + b_2(\text{sex})] - [b_0 + b_1(x_1) + b_2(\text{sex})] \tag{9.22}$$

$$= b_1(x_2 - x_1). \tag{9.23}$$

For a one unit change in BMI the difference in log odds is the $b_1 = 0.171$, and the odds ratio is

$$\text{OR} = e^{0.171} = 1.186,$$

a roughly 20% increase in the odds of hyperuricemia associated with the larger BMI.

Confidence intervals are calculated using standard errors just as in single variable logistic regression.

GUIDED PRACTICE 9.24

G Calculate a 95% confidence interval for the odds ratio of hyperuricemia associated with a three unit increase in BMI for two individuals of the same sex. [13]

GUIDED PRACTICE 9.25

G Does the intercept have scientific meaning in this model? [14]

Since the hyperuricemia study had a cross-sectional design, the probability of hyperuricemia for values of the predictors can be estimated from the model, as discussed later in Section 9.4.4.

[13] A 95% confidence interval for the change in log(odds) for a 1 unit change in BMI is $0.171 \pm (1.96)(0.038) = (0.097, 0.246)$. The confidence interval for a three unit change can be calculated by multiplying the lower and upper bounds by 3: $[(3)(0.097), (3)(0.246)] = (0.291, 0.738)$. The corresponding interval for the OR is $(e^{0.290}, e^{0.736}) = (1.338, 2.092)$.

[14] No. The intercept is the log(odds) for an individual with baseline category "male" but BMI = 0.

EXAMPLE 9.26

Calculate the estimated probability of hyperuricemia for a female with BMI 28.

The log(odds) are

$$-5.503 + (0.171)(28) - 0.480 = -1.195,$$

(E)

so the odds are $e^{-1.195} = 0.303$. The estimated probability of hyperuricemia is

$$\exp\left[\frac{0.303}{1+0.303}\right] = 0.232.$$

A female with BMI 28 has an estimated chance of 23% of being hyperuricemic.

The OR for hyperuricemia comparing males to females is the same, for any value of BMI as long as BMI is held constant. When both predictors change, the full model must be used to calculate odds ratios.

EXAMPLE 9.27

What is the OR for hyperuricemia, comparing a woman with BMI 32 to a male with BMI 30?

The log(odds) of hyperuricemia for a woman with BMI 32 is

$$-5.503 + (0.171)(32) - 0.480 = -0.511,$$

so the corresponding odds are $e^{-0.511} = 0.600$.

(E)

For the male with BMI 30, the log(odds) are

$$-5.503 + (0.171)(30) = -0.373,$$

so the odds of hyperuricemia are 0.689. The OR comparing the female to the male is 0.600/0.689 = 0.871.

A woman whose BMI is 2kg/m^2 larger than a male still has a lower estimated odds of hyperuricemia.

In the model for hyperuricemia the change in log odds when when one predictor changes does not depend on the value of the other predictor. The same is not true for estimated probabilities.

EXAMPLE 9.28

For males, use the estimated probabilities of hyperuricemia for individuals with BMI 28 and BMI 30 to calculate estimated prevalence differences and risk ratios. Repeat the calculation for females.

(E) For a male with BMI 28 the estimated log odds and odds of hyperuricemia are $-5.503+(0.171)(28) = -0.715$ and $e^{-0.715} = 0.489$. The estimated prevalence (probability) of hyperuricemia is $0.489/(1 + 0.489) = 0.328$. The estimated odds of hyperuricemia for a male with BMI 30 were calculated in Example 9.27 and are 0.689, so the estimated prevalence is 0.408. The estimated prevalence difference and ratio risk are, respectively, $0.408 - 0.328 = 0.080$ and $0.408/0.328 = 1.244$.

The prevalence difference and risk ratio for females are calculated similarly and are, respectively, 0.066 and 1.290. The prevalence differences and ratios associated with a change in BMI from 28 to 30 are different for males than for females, and must be calculated using all the coefficients in the model. This result is another reason why an estimated OR from a logistic regression should not be interpreted as a risk ratio.

In a model that includes sex, the log(OR) for hyperuricemia for a one unit change in BMI for participants of the same sex is 0.171, slightly attenuated toward 0 from the earlier log(OR) of 0.185 in the model with only BMI. In these data, males tend to have larger BMI (25 vs 23.6kg/m^2) and have double the odds hyperuricemia than females, so the estimated association in the model with BMI alone is influenced by the males with larger BMI. Adding sex to the model separates the sex and BMI associations, at least within the assumptions of the logistic model.

9.4.2 Modeling a possible interaction

A regression model is called an **additive model** in the predictors when the change in association between a response and predictor does not depend on values of the other predictors. The logistic model in Equation 9.20 is additive in the predictors BMI and sex for the log odds of hyperuricemia; the difference in log(odds) for two values of BMI does not depend on sex. What is the evidence that the association between BMI and hyperuricemia might differ for males and females?

When an association may differ between categories of another predictor, such as sex, it is common in the epidemiological literature to call that predictor a potential **effect modifier**, and the phenomenon is called **effect modification**. This section does not use that terminology for reasons explained later and instead uses the more statistically descriptive term **interaction**.

In regression models interactions are usually explored by including an interaction term. Section 7.7 discusses modeling an interaction in linear regression. In these data, a two variable model with an interaction term in the logistic model is

$$\log(\text{odds}_E) = \beta_0 + \beta_1 \text{bmi} + \beta_2 \text{sex} + \beta_3 (\text{bmi} \times \text{sex}). \tag{9.29}$$

The last term is the product of `bmi` and `sex`.

The interaction term (`bmi×sex`) allows the slope coefficient for `bmi` to depend on `sex`. For the reference sex category "male" the coefficient of `bmi` is β_1; for the category "female" the slope of `bmi` is $\beta_1 + \beta_3$. Confidence intervals for β_3 or a test of the null hypothesis $\beta_3 = 0$ can be used to assess the evidence against the hypothesis that the log(odds) for the relationship between hyperuricemia and BMI does not depend on sex.

The number of hyperuricemic events (95) is sufficient to add another parameter, and Figure 9.13 shows the estimated model. Equation 9.30 shows the algebraic form of the this model.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-5.006	1.264	-3.96	0.000
bmi	0.152	0.049	3.12	0.002
sexfemale	-1.652	1.947	-0.85	0.396
bmi:sexfemale	0.046	0.077	0.61	0.544

Figure 9.13: Logistic regression with interaction: response variable hyperuricemia, predictors BMI and sex.

$$\begin{aligned}\widehat{\log}\left[\text{odds}_E(\text{bmi, sex})\right] &= b_0 + b_1\text{bmi} + b_2\text{sexfemale} + b_3(\text{bmi})(\text{sexfemale}) \\ &= -5.006 + (0.152)\text{bmi} - (1.652)\text{sexfemale} + (0.046)(\text{bmi})(\text{sexfemale}). \quad (9.30)\end{aligned}$$

The evidence for the interaction term is weak ($p = 0.544$). The observed difference in association between the log odds of hyperuricemia and BMI between males and females is not inconsistent with what would be expected if there were actually no population-level difference in association. In this model, there is no support for the hypothesis that the relationship between hyperuricemia and BMI differs by sex, and the interaction term should be left out of the exploratory model. Exercise 9.25 explores the interpretation of a model with an interaction term.

A data analyst starting with the interaction model might mistakenly conclude that neither `sex` nor the interaction of `sex` with `bmi` should be retained. Analyses should always begin without interaction terms and add them only when there is a reason to look more closely at the relationship between a response and a predictor across the levels of another variable.

This chapter avoids the use of the terms effect modifier and effect modification in observational studies. The term "effect" implies a causal link that cannot be established in an observational study with the methods described in this text. It is common, though, in applications to label the non-interaction terms as **main effects** and interaction terms as **interaction effects**. The terminology can be a useful abbreviation as long as no causal association is meant or inferred.

9.4.3 Categorical predictors with more than two levels

When spawning, a female horseshoe crab migrates to shore with a male attached to her spine to lay clusters of eggs in the sand. Additional male crabs may join the pair and fertilize the eggs as well, presumably increasing genetic diversity of the offspring. The additional male crabs are called satellites. The data used here originally appeared in Brockman[15] and can be found at the website for *Categorical Data Analysis, 3rd ed.*[16] and in the R package `glmbb`. The dataset contains information on 173 female crabs, 111 with at least one male satellite.

This section examines the association between the odds of the event E that a female has one or more satellites and her carapace (shell) width and color. Let the variable y denote whether a female has one or more satellites ($y = 1$) or none ($y = 0$), `width` gives the carapace width in centimeters and the levels of the factor variable `color` are "Light", "MedLight" (for medium light), "MedDark" (for medium dark), and "Dark", denoting increasingly dark colors. The predictor `color` is an ordinal categorical variable, but since methods that take advantage of ordinal variables in contingency tables and logistic regression are beyond the scope of this text, the analyses in this section treat color as a standard unordered categorical variable.

The contingency table in Figure 9.14 shows the association between color and the presence of at least one satellite. The estimated odds vary by color; the odds of dark females having at least

[15] H Jane Brockmann. "Satellite male groups in horseshoe crabs, Limulus polyphemus". In: *Ethology* 102.1 (1996), 1–21.
[16] Alan Agresti. *Categorical data analysis, 3rd ed.* Vol. 792. John Wiley & Sons, 2013.

one satellite are 7/15 = 0.467, while the odds for a female with medium light color are 69/26 = 2.654. The OR, comparing medium light to dark, is 2.654/0.467 = 5.683; the odds of medium light female crab having at least one satellite are between 5 and 6 times larger than for a dark female.

The conditions given in Section 8.3.2 for the validity of a χ^2 test are met in the table (just barely, see Exercise 9.21); the χ^2 statistic has value 14.08 on 3 degrees of freedom, $p = 0.003$. The extension of Fisher's exact test to a 4×2 table yields the same p-value, so the table provides evidence that in these data, color and having more than one satellite are not independent.

Color	$y = 0$	$y = 1$	Sum
Dark	15	7	22
MedDark	18	26	44
MedLight	26	69	95
Light	3	9	12
Sum	62	111	173

Figure 9.14: Absence ($y = 0$) or presence ($y = 1$) of at least one satellite versus color of a female horseshoe crab.

The interpretation of logistic regression with a categorical predictor with four levels is the same as that for a predictor with 2 levels described in Section 9.3.1 – odds ratios calculated from the 4×2 table will match those computed from the regression coefficients. Figure 9.15 shows the estimated regression with the predictor `color`, with the color "Dark" set as the reference category. The less frequent response category $y = 0$ has 62 observations and the model has 4 parameters including the intercept, 2 fewer than the maximum 6 the guidelines suggest, so estimates and inference should be reliable.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.762	0.458	-1.67	0.096
colorMedDark	1.130	0.551	2.05	0.040
colorMedLight	1.738	0.512	3.39	0.001
colorLight	1.861	0.809	2.30	0.021

Figure 9.15: Logistic regression with horseshoe crab data, response variable presence of male satellites, predictor variable `color`.

The algebraic form of the model is

$$\log[\text{odds}_E(\text{color})] = -0.762 + (1.130)\texttt{colorMedDark} + (1.738)\texttt{colorMedLight} + (1.861)\texttt{colorLight}. \tag{9.31}$$

Since the reference category is "Dark", the log(odds) of a dark female having at least one satellite is the intercept term −0.762, with corresponding odds $e^{-0.762} = 0.467$, the same value when using the table in Figure 9.14. This is one instance where the intercept term is meaningful. More generally, when there are no other predictors in a model with a categorical predictor, the intercept term is the log(odds) of the outcome for the reference category. Using Equation 9.31, the log(odds) for the color "MedLight" is $-0.762 + 1.738 = 0.976$, with corresponding odds $e^{0.976} = 2.654$. The OR comparing "MedLight" to "Dark" 2.654/0.467 = 5.683, also agreeing with the OR calculated from Figure 9.14. When comparing a category against the reference, ORs can be calculated directly. The coefficient for "MedLight" is the difference in log(odds) between "MedLight" and the reference category "Dark", so the OR comparing the two categories is $e^{1.738} = 5.686$. The small difference between this OR and the one calculated from Figure 9.14 is due to the rounding of the coefficients from the logistic model.

The pattern of the coefficients is consistent with what is known about horseshoe crabs – the

log(odds) and hence odds and probability of having satellites increase with lighter colors of the female carapace.

Calculating ORs for two categories that do not include "Dark" can be done with the model coefficients. The log(odds) for the category "Light" is $-0.762 + 1.861 = 1.099$. The difference in log(odds), comparing "Light" to "MedLight" is $1.099 - 0.976 = 0.123$, so the OR is $e^{0.123} = 1.131$. This odds ratio can also be calculated directly from model coefficients. Suppose b_0 is the intercept, and let b_3 and b_4 denote the coefficients of the categories "MedLight" and "Light", respectively. The difference in log(odds) for the two categories is

$$\begin{aligned}(b_0 + b_4) - (b_0 + b_3) &= b_4 - b_3 \\ &= 1.861 - 1.738 \\ &= 0.123.\end{aligned}$$

Since the coefficient for the intercept cancels in the subtraction, the odds ratio comparing "MedLight" to "Light" is $\exp(b_4 - b_3) = \exp(0.123) = 0.131$. This argument easily generalizes to any two categories when predictors have more than 4 levels.

The calculation of a confidence interval for the OR comparing two categories that are not the reference category is a more difficult calculation, since it requires the standard error of the difference of two estimated log(OR)s, a topic not covered here.

Since the χ^2 test based on Figure 9.14 and the deviance based test for the model are both used to test the null hypothesis of no relationship between the response and the predictor, both should yield approximately the same statistic and p-value. The null and residual deviances for the model are 225.76 and 212.06. The difference 13.7 yields $p = 0.003$ for a χ^2 with 3 degrees of freedom. Both approaches support the conclusion that, when other factors are not accounted for, color is associated with the tendency for a female crab to have at least one satellite. (The two χ^2 values are slightly different because they are calculated using different formulas.)

The p-values of the coefficients are used to test the null hypothesis that the difference in log(odds) between a category and the reference "Dark" is 0, i.e. that the two log(odds) are equal. They cannot be used to test the importance of a particular color, and since the colors are levels of the single predictor `color`, one level cannot be retained and the others dropped. In these data, the p-values for the coefficients may that all colors are associated with an increase in the odds of satellites compared to "Dark", but no adjustment has been made for multiple testing. Using a Bonferroni correction as in ANOVA (Section 5.5.3) and multiplying all p-values by 3 suggests that only "MedLight" crabs have significantly larger odds of satellites compared to "Dark".

9.4.4 Inference for multiple logistic regression

This section discusses the principles used for inference in multiple logistic regression,putting some of the model features discussed earlier in a general context. In the multiple regression model, E is an event (e.g., a TB treatment interruption, or presence of hyperuricemia) that may be associated with p predictors $X_1, \ldots, X_p$. Let $x = (x_1, \ldots, x_p)$ and $p_E(x)$ the conditional probability

$$p_E(x) = p_E(x_1, x_2, \cdots, x_p) = P(E|x_1, x_2, \cdots, x_p).$$

In the multiple logistic regression model,

$$\log\left[\frac{p_E(x)}{1 - p_E(x)}\right] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p,$$

or, equivalently,

$$\log[\text{odds}_E(x)] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p.$$

The model is sometimes written in terms of the log odds of a binary response variable Y that takes on the value 1 if the event E occurs and 0 otherwise:

$$\log\left[\frac{P(Y=1|x)}{P(Y=0|x)}\right] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p.$$

In statistical terms, Y is the indicator variable for the event E.

The coefficient of a predictor is the change in the conditional log(odds) of E associated with a one unit change of that predictor, if the values of the other variables in the model do not change. The argument showing that the change in log(odds) for a variable depends on only its coefficient and not on the intercept or the values of the other variables is similar to that used in deriving Equations 9.11 and 9.23. Suppose for simplicity that the logistic regression is the two variable model

$$\log[\text{odds}_E(x)] = \beta_0 + \beta_1 x_1 + \beta_2 x_2.$$

If x_1 changes from x_1^a to x_1^b the change in log odds will be

$$(\beta_0 + \beta_1 x_1^a + \beta_2 x_2) - (\beta_0 + \beta_1 x_1^b + \beta_2 x_2) = \beta_1(x_1^a - x_1^b),$$

as long as x_2 remains constant. The resulting OR, $\exp[\beta_1(x_1^a - x_1^b)]$, does not depend on the value of either β_0 or x_2. When x_1 changes by one unit ($x_1^a - x_1^b = 1$), the coefficient β_1 is the additive change in log(odds) and e^{β_1} is multiplicative change in the odds for a one unit change in x_1. Equivalently, β_1 and e^{β_1} are, respectively, the log(OR) and (OR) for a one unit change in x_1. This same derivation applies to any variable in models with more than two variables.

The conditional odds of E are

$$\frac{p_E(x)}{1 - p_E(x)} = \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p), \tag{9.32}$$

and using the relationship between odds and probabilities,

$$p_E(x) = \frac{\exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p)}{1 + \exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p)}. \tag{9.33}$$

The assumptions for inference with multiple logistic regression are similar to those for simple logistic regression: (1), the transformation $\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$ is a reasonable model for the log odds of E; (2), the set of response and predictor variables for each case are independent of those in the other cases; (3), log(odds), odds and probabilities can all be estimated when the data are a random sample in an exposure-based or cross-sectional design; and (4), log(odds) and odds can be estimated in case-control studies but probabilities cannot.

The first assumption is usually the most difficult to justify without some of the diagnostic tools discussed in Section 9.5. The other three all depend on the study design, just as in simple logistic regression.

Hypothesis tests and confidence intervals are based on the approximate normal sampling distributions of the estimates for the coefficients.

SAMPLING DISTRIBUTIONS OF ESTIMATED COEFFICIENTS IN MULTIPLE LOGISTIC REGRESSION

Let E be an event and suppose

$$\log(\widehat{\text{odds}_E}(x)) = b_0 + b_1 x + \cdots + b_p x_p$$

is an estimated logistic regression model from a dataset with n cases. For a coefficient b_k with standard error s.e.(b_k), the statistic

$$\frac{b_k - \beta_k}{\text{s.e.}(b_k)}$$

has approximately a standard normal (z) distribution in moderate to large sample sizes. Consequently, under the hypothesis $H_0 : \beta_k = 0$, the statistic

$$\frac{b_k}{\text{s.e.}(b_k)}$$

has an approximate standard normal (z) distribution.

There is no clear dividing line between a sample size that is adequate and one that is not, and there have been many suggested guidelines. The guideline used here is based on the smaller number of outcomes in the two values of the response variable. If N is the number of observations in this category, the number of parameters (including the intercept) should be no larger than $N/10$.[17] Using this rule, for instance, in a dataset with 40 successes and 50 failures, a logistic regression should have no more than $(40/10) = 4$ parameters, including the intercept.

The sampling distribution can be used for tests and confidence intervals.

TESTING A HYPOTHESIS ABOUT A LOGISTIC REGRESSION COEFFICIENT

A test of the two-sided hypothesis

$$H_0 : \beta_k = 0 \text{ vs. } H_A : \beta_k \neq 0$$

is rejected with significance level α when

$$\frac{|b_k|}{\text{s.e.}(b_k)} > z^\star,$$

where $z^\star$ is the point on a z-distribution with area $(1 - \alpha/2)$ in the left tail.

For one-sided tests, $z^\star$ is the point on a z-distribution with area $(1 - \alpha)$ in the left tail. A one-sided test of H_0 against $H_A : \beta_1 > 0$ rejects when the standardized coefficient is greater than $z^\star$; a one-sided test of H_0 against $H_A : \beta_1 < 0$ rejects when the standardized coefficient is less than $-z^\star$.

CONFIDENCE INTERVALS FOR A LOGISTIC REGRESSION COEFFICIENT

A two-sided $100(1 - \alpha)\%$ confidence interval for the model coefficient β_k is

$$b_k \pm [\text{s.e.}(b_k) \times z^\star].$$

All statistical software packages provide standard errors (s.e.) of coefficients, and most

[17] Peter Peduzzi et al. "A simulation study of the number of events per variable in logistic regression analysis". In: *Journal of clinical epidemiology* 49.12 (1996), pp. 1373–1379.

provide the z statistic and its p-value directly.

The selection of variables to include in a regression model depends on many factors, including the intent of the analysis and the statistical precision of estimated coefficients. The selection rarely depends only on a significance test, but assessing the strength of evidence of the association between a variable or set of variables and a response is a good place to start the process, and the deviance statistic is a useful statistic. An analysis often begins by assessing whether a model is useful at all. A logistic regression model may not useful for estimating odds ratios or probabilities if a model with predictors is not significantly better than a model with only the intercept term, that is, if there is not strong evidence against the hypothesis that coefficients of the predictors are all 0. A test of the null hypothesis that all model coefficients are 0 uses a statistic called the deviance. Multiple logistic regression models are estimated by the method of maximum likelihood, the same approach that is used for simple logistic regression, and the deviance is a function of the maximized likelihood function. Its mathematical definition is beyond the scope of this book; it is enough to know that the deviance decreases as the fit of a model improves.

THE DEVIANCE STATISTIC FOR OVERALL MODEL FIT

In logistic regression, the **residual deviance** is a measure of the fit of an estimated model and **null deviance** is a measure of fit of a model with only an intercept term. A test of the hypothesis $H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$ versus the alternative that at least one coefficient is not zero can be based on the statistic

$$D = \text{null deviance} - \text{residual deviance}.$$

If the conditions for logistic regression are met, D has approximately a χ^2 distribution with p degrees of freedom under H_0. A level α test of H_0 is rejected if D is in the right tail with area α of a χ^2 distribution with p degrees of freedom.

The statistic D will be small when the residual deviance for the current model is close to the deviance of a model without any predictors; the current model is unlikely to be useful. Large values of D mean that the residual deviance for the current model is much smaller than the deviance for a model with no predictors and, consequently, provides a useful summary of the data. The statistic D uses a different metric than the overall F-statistic in least squares regression, but it serves the same purpose.

In the model for hyperuricemia with predictors `sex` and `bmi`, both coefficients have small p-values, so it is reasonable to expect that model including the two variables is better than a model with only an intercept, and the deviance statistic confirms that. The software R reports that the null and residual deviances are 486.22 and 455.27, respectively. The difference, 30.95, yields $p < 0.001$ from a χ^2 with 2 degrees of freedom.

The deviance statistic can also be used to compare two nested models, i.e., models where the parameters in one are a subset of those in the second. Nested models most commonly occur when examining the evidence for keeping a set of variables as part of a larger model.

THE DEVIANCE STATISTIC FOR COMPARING TWO NESTED MODELS

Let

$$\log[\text{odds}_E(x)] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p \tag{9.34}$$

be the usual multiple logistic regression model for the association between an event E and potential predictors $x_1, x_2, \ldots, x_p$, and let D_p be the residual deviance for the model.
Suppose the nested model

$$\log[\text{odds}_E(x)] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k \tag{9.35}$$

is based on only the first k of the p predictors, where $k < p$, and let D_k be the residual deviance for the smaller, nested model. The hypothesis that the $p-k$ predictors $x_{k+1}, x_{k+2}, \ldots, x_p$ may not be needed in the model is equivalent to the null hypothesis $H_0 : \beta_{k+1} = \beta_{k+2} = \cdots = \beta_p = 0$.

If the conditions for logistic regression are met, then under H_0, $D_k - D_p$ has a χ^2 distribution with $p-k$ degrees of freedom. The hypothesis H_0 is rejected at level α if $D_k - D_p$ is in the right tail with area α of a χ^2 distribution with $p-k$ degrees of freedom.

The coefficients in H_0 can, of course, can be any subset of the p variables in the full model and need not be adjacent in the variable listing.

The residual deviance always decreases when variables are added to a model, just as R^2 always increases in linear regression. Adding variables simply because the deviance is decreasing can lead to overfitting. Section 7.3.2 describes an adjusted R^2 that 'penalizes' R^2 by a factor that depends on the number of parameters. The Akaike Information Criterion, or AIC, plays a similar role as the deviance.

THE AKAIKE INFORMATION CRITERION (AIC) FOR COMPARING TWO NESTED MODELS

The Akaike Information Criterion (AIC) for a model with p predictors is given by

$$\text{AIC}_p = D_p + 2(p+1).$$

Let D_p and D_k be the residual deviances for the larger and smaller (nested) models, respectively, and let AIC_p and AIC_k be the respective values of AIC.
The deviance D_p will necessarily be smaller than D_k, but the larger model may not be worth the added complexity if $\text{AIC}_p \geq \text{AIC}_k$.

In the two variable model for hyperuricemia, the evidence for the value of sex as a predictor is weaker than for `bmi` but still relatively strong, with a p-value of 0.05. Should it be kept in the simple model for hyperuricemia using `bmi` alone? The deviance statistics for the two models are $D_{\text{bmi}} = 459.54$ and $D_{\text{bmi,sex}} = 455.26$. As expected, $D_{\text{bmi,sex}} < D_{\text{bmi}}$. The AIC for the two variable model is

$$\begin{aligned} \text{AIC}_{\text{bmi, sex}} &= 455.26 + 2(3) && (9.36)\\ &= 461.26, && (9.37) \end{aligned}$$

while the AIC for the one variable model

$$\begin{aligned} \text{AIC}_{\text{bmi}} &= 459.54 + 2(2) && (9.38) \\ &= 463.26. && (9.39) \end{aligned}$$

The AIC for the larger model is still smaller than that for the smaller model even after accounting for the number of parameters, so it seems reasonable to leave sex in an explanatory model for hyperuricemia. AIC is also an indirect measure of how well a model predicts future observations and is discussed in that context in subsection on estimating discrimination.

Selecting a model often involves a balance between the goal of an analysis and the use of AIC or other in automated model selection methods. An extended discussion of model selection is beyond the scope of this text. The analyses later in this chapter use AIC informally along with the context of the analysis to examine the value of nested models.

9.5 Assessing model adequacy

Exploring model diagnostics is an important part of any analysis and should not be overlooked. This section discusses diagnostics commonly used with logistic regression, using the TB and hyperuricemia data as examples.

The first step in model checking should assess how well the model matches the data. Section 9.5.1 discusses two goodness-of-fit statistics, a traditional χ^2 statistic when all predictors are categorical, and the more general **Hosmer-Lemeshow statistic** that allows continuous predictors. These two statistics may be sufficient when the goal of an analysis is an explanatory model to estimate associations between a response and predictors.

Logistic regression is often used to predict binary outcomes (might this person be hyperuricemic?) or to build a classification model that groups members of a population into categories (which patients admitted to a hospital emergency room should be given high priority for care?). After checking model fit, it is important to use some of the methods in Section 9.5.2 to check the accuracy of predictions. The **Brier score** is a summary statistic used to estimate how well predicted probabilities match observed outcomes. **Calibration plots** provide more detail than Brier scores; they provide a graphical diagnostic for the match between predicted probabilities and outcomes. When a logistic model will be used to classify individuals into two subgroups of a population (typically, those with or without an undiagnosed condition) false negative rates and false positive rates estimate the probabilities of incorrectly classifying an individual with (false negative) or without (false positive) the condition. Receiving operator characteristic curves (ROC curves) show graphically how classification errors depend on the prediction rule.

When statistics and graphics for checking the accuracy of predictions are calculated using the data on which the model for the model fit, estimated errors are called apparent error rates and may not accurately reflect errors when the model is used in new data. Section 9.5.3 explores the use for estimating . The use of a (also called a validation dataset) is explored in the case study in Section 9.6.

9.5.1 Goodness-of-Fit Statistics

Goodness-of-fit statistics typically assess how well estimates from a model match the observed data, similar to the use of a χ^2 test for the fit of a distribution discussed in Section 8.4. The deviance statistic used in Section 9.4.1 is sometimes called a goodness-of-fit statistic, but it assesses whether a model is better than no model at all (i.e., "better than nothing"). A significant deviance statistic can be useful in deciding whether to examine a model more closely, but it does not imply that the model adequately reflects the data. The use of the deviance to compare nested models, as in Section 9.4.3, should also not be viewed as a goodness-of-fit statistic. It provides guidance on whether a smaller model is adequate compared to a larger model, but does not test the fit of either.

This section uses the TB dataset to illustrate goodness-of-fit when all predictors are categorical and the hyperuricemia data to illustrate the other methods.

The χ^2 goodness-of-fit statistic with categorical predictors

The simplest setting for assessing fit is one in which all predictors are categorical. Each combination of predictor values yields a unique profile or pattern into which cases can be grouped, and the observed numbers of responses within a profile can be compared with the

expected number calculated from the model. Pearson residuals are standardized differences between observed and expected, and a χ^2 test is based on the sum of squared residuals. The approach is illustrated using the TB interruption dataset.

Treatment for multidrug-resistant tuberculosis (MDR-TB) lasts longer than standard therapy and may lead to a higher frequency of treatment interruptions. The dataset `tb` contains the predictor `mdr.tb` indicating whether a study participant was receiving the longer course of treatment. Figure 9.16 shows an estimated logistic regression model with predictors `education` and `mdr.tb`.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.839	0.129	-14.225	< 0.000
educationYes	-0.793	0.191	-4.144	< 0.000
mdr.tbYes	0.861	0.300	2.869	0.004

Figure 9.16: Logistic regression, response variable two-month interruption, predictors `education` and `mdr.tb`.

The data suggest that education and treatment for MDR-TB may be important predictors of interruption.

Each of the predictors has two values, so each participant falls into 1 of 4 profiles. The rows of Figure 9.17 show summary statistics for the 4 profiles, defined by the values of `education` and `mdr.tb`, using 1 to denote the level "Yes" and 0 for "No". The figure is an abbreviated version of a table produced by the function `dx` in the R package `LogisticDx`.

Profile	`educationYes`	`mdr.tbYes`	Observed responses	Predicted probability	n	Predicted responses	Pearson residual r
1	1	0	44	0.067	671	45.020	-0.157
2	0	0	67	0.137	481	65.980	0.135
3	1	1	8	0.145	48	6.980	0.418
4	0	1	8	0.273	33	9.020	-0.398

Figure 9.17: Summary statistics for the 4 profiles in the TB dataset defined by education and treatment for MDR-TB

The first column labels the 4 profiles, and columns 2 – 3 show the values of the predictors. The remaining columns contain the following data for each profile:

- *Observed responses:* The observed number of participants with treatment interruptions.
- *Predicted Probability*: The predicted probability of a treatment interruption from the model. Since the participants in a profile all have the same values for the predictor, there is a single predicted probability for a profile.
- *n:* The number of participants who match the profile.
- *Predicted responses*: The predicted number of participants with treatment interruptions from the model, calculated below.
- *Pearson residual:* The Pearson residual r, a measure of the discrepancy between the observed and predicted number of treatment interruptions. The definition of the Pearson residual is given below.

The predicted probability of a treatment interruption can calculated directly from the model;

the predicted probability $\hat{p}_1$ for profile 1 is

$$\begin{aligned}\hat{p}_1 &= \frac{\exp(-1.89 - 0.793[1] + 0.861[0])}{1 + \exp(-1.89 - 0.793[1] + 0.861[0])} \\ &= 0.06709.\end{aligned}$$

The value in the table (0.067) has been rounded from the more precise 0.06709. Using the more precise value, the predicted number of responses for profile 1 is

$$\begin{aligned}(n_1)(\hat{p}_1) &= (671)(0.06709) \\ &= 45.02.\end{aligned}$$

The Pearson residual r is a standardized version of the observed - predicted number of responses, using the formula for the standard error of a binomial variable. For profile 1,

$$\begin{aligned}r_1 &= \frac{44 - 45.02}{\sqrt{n_1[\hat{p}_1][1 - \hat{p}_1]}} \\ &= \frac{-1.02}{\sqrt{671[0.06709][0.93290]}} \\ &= -1.57.\end{aligned}$$

The residual is small because the predicted value 45.02 is close to the observed number of responses 44.

A χ^2 goodness of fit is based on $\sum_i r_i^2$, with degrees of freedom equal to the number of profiles minus the total number of parameters (including the intercept). For the TB data, the χ^2 statistic χ^2 is

$$\begin{aligned}\chi^2 &= \sum_{i=1}^{4} r_i^2 \\ &= (-0.157)^2 + (0.135)^2 + (0.418)^2 + (-.398)^2 \\ &= 0.376.\end{aligned}$$

Since there are 4 profiles and 3 parameters in the model, the p-value is $P(\chi^2_{1df} > 0.376) = 0.540$.

The logistic model with predictors `education` and `mdr.tb` fits the data reasonably well – the observed and expected numbers of responses are similar, and the goodness-of-fit test is non-significant. However, even when a model seems to fit data, it is not necessarily the best model. The TB dataset contains additional predictors not examined here that may provide a better model for predicted probabilities.

The χ^2 goodness-of-fit test discussed above cannot be used when some profiles have a small number of observations or when one or more predictors are continuous. Profiles may have only one case if a continuous predictor has different values for each case, causing the number of profiles to be the number of cases. The validity of the test depends on the number of observations within each profile being reasonably large, just as in the usual χ^2 goodness-of-fit test. While it might be tempting to create a smaller number of profiles by combining categories of some categorical variables, creating profiles post hoc may also violate the assumptions for the test. In fact, even when all predictors are categorical but there are a large number of profiles, some with small numbers of observations, the χ^2 test may not be reliable.

The Hosmer-Lemeshow goodness-of-fit test

When the data cannot be grouped into profiles, Hosmer and Lemeshow have proposed a goodness-of-fit statistic that uses groupings according to predicted probabilities. The test is described in more detail in Hosmer, Lemeshow and Sturdivant[18] and is outlined here, using the logistic model for the association of hyperuricemia and BMI in Figure 9.5.

1. Let n be the number of cases in the dataset, x_i be the set of predictor values for case i, $i = 1, \ldots, n$, and E the event of interest (e.g., hyperuricemia). Calculate the model-based predicted probabilities $\hat{p}_i$ for each case and sort the probabilities in increasing order.

2. Group the observations into g groups. Hosmer and Lemeshow recommend $g = 10$ equally sized groups with boundaries based on the deciles of the sorted predicted probabilities. The rows of Figure 9.18 show the groups; the first group contains the 500/10 = 50 cases with predicted probabilities between 0.0434 and 0.0913; the second group contains the 50 cases with predicted probabilities larger than 0.0913 but no larger than 0.1144, etc. For instance, the case with bmi = 17.68 has an estimated probability of hyperuricemia E given by

$$\begin{aligned} p_E(17.68) &= \frac{\text{odds}_E(17.68)}{1 + \text{odds}_E(17.68)} \\ &= \frac{\exp(-6.05 + 0.185(17.68))}{1 + \exp(-6.05 + 0.185(17.68))} \\ &= 0.058, \end{aligned}$$

   so this observation is part of group 1.

3. For each of the g groups, record the observed numbers of individuals without and with the event (o_0 and o_1, respectively), and compute the expected counts for each category ($\hat{e}_0$ and $\hat{e}_1$). The expected count $\hat{e}_0 = \sum_{o_i=0}(1 - \hat{p}_i)$, where the sum is over cases within a group, and $\hat{e}_1 = \sum_{o_i=1} \hat{p}_i$. The first row in Figure 9.18 shows that in the smallest 10% of the predicted probabilities, 48 individuals did not experience hyperuricemia, while 2 did. The corresponding expected counts were 46.2 and 3.8.

4. Calculate the test statistic $\hat{C}$ and its significance level:

$$\hat{C} = \sum_{k=1}^{g} \left[\frac{(o_{0k} - \hat{e}_{0k})^2}{\hat{e}_{0k}} + \frac{(o_{1k} - \hat{e}_{1k})^2}{\hat{e}_{1k}} \right].$$

   Hosmer and Lemeshow argued that the statistic has an approximate χ^2 distribution with $g - 2$ degrees of freedom. For the hyperuricemia data, $\hat{C} = 7.62$ on $10 - 2 = 8$ degrees of freedom, so $p = 0.47$ and a null hypothesis of an adequately fitting model is not rejected. A non-significant goodness-of-fit statistic does not imply that a model fits very well, of course; it only demonstrates that there is not substantial evidence of a poor fit to the data.

The Hosmer-Lemeshow statistic extends naturally to models with more than one predictor since it depends on predictors only through predicted probabilities. In the hyperuricemia data with predictors BMI and sex, the steps in calculating the entries for both a summary table and the goodness-of-fit statistic are the same, except that the predicted probabilities are calculated using BMI and sex.

Figure 9.19 shows a table summarizing the fit of the Hosmer-Lemeshow statistic for the model using BMI and sex. Just as in Figure 9.18, the study sample has been grouped according to deciles of the estimated probabilities. The observed counts for both the absence and presence of

[18] David W Hosmer Jr et al. *Applied logistic regression, 3rd ed.* John Wiley & Sons, 2013.

	Probability ranges	o_0	$\hat{e}_0$	o_1	$\hat{e}_1$
1	[0.0434,0.0913]	48	46.2	2	3.8
2	(0.0913,0.114]	47	44.9	3	5.1
3	(0.114,0.133]	43	43.8	7	6.2
4	(0.133,0.152]	42	42.9	8	7.1
5	(0.152,0.174]	39	42.6	12	8.4
6	(0.174,0.193]	43	40.0	6	9.0
7	(0.193,0.221]	37	40.5	14	10.5
8	(0.221,0.245]	38	37.6	11	11.4
9	(0.245,0.299]	35	36.4	15	13.6
10	(0.299,0.722]	33	30.1	17	19.9

Figure 9.18: Hosmer-Lemeshow goodness-of-fit table for the logistic regression with response hyperuricemia and predictor BMI.

hyperuricemia match the predicted counts reasonably well. The value of the $\hat{C}$ is 4.1 on 8 degrees of freedom, $p = 0.80$. The statistic provides no evidence that the two variable model fits poorly.

	Probability Ranges	o_0	$\hat{e}_0$	o_1	$\hat{e}_1$
1	[0.0376,0.0843]	47	46.6	3	3.4
2	(0.0843,0.108]	46	45.2	4	4.8
3	(0.108,0.126]	45	44.2	5	5.8
4	(0.126,0.145]	43	43.2	7	6.8
5	(0.145,0.168]	41	42.2	9	7.8
6	(0.168,0.195]	39	40.9	11	9.1
7	(0.195,0.225]	43	39.5	7	10.5
8	(0.225,0.261]	35	38.0	15	12.0
9	(0.261,0.323]	34	35.5	16	14.5
10	(0.323,0.624]	32	29.8	18	20.2

Figure 9.19: Hosmer-Lemeshow goodness of fit table for the logistic regression with response variable hyperuricemia and predictors BMI and sex.

The hyperuricemia example highlights an important aspect of testing model fit. The Hosmer-Lemeshow tests suggest that neither the one nor two variable model fits poorly. The AIC statistics for the models with and without sex in Equations 9.37 and 9.39 suggested that adding the predictor sex to the model with BMI may be worth the small increase in model complexity, especially because measuring and recording sex for each participant is relatively easy. Even though the model with BMI alone does not fail a goodness-of-fit test it may not be the better model.

The Hosmer-Lemeshow test has some weaknesses, and several alternatives have been proposed, all with their own advantages and disadvantages. Grouping cases by deciles of probabilities has no theoretical justification, a χ^2 distribution with $g-2$ degrees of freedom does not always provide a good approximation to the sampling distribution, and the test has been shown to have low power in some situations. These shortcomings of the test, however, are largely about the statistical properties of the test statistic. It is important to keep in mind that statistical tests for goodness-of-ft have limited value generally. A statistical test for goodness-of-fit will reject the null hypothesis of adequate fit only when there is strong evidence of lack of fit. Many models fit poorly but not so badly that a goodness-of-fit statistic is significant. The table associated with the Hosmer-Lemeshow statistic is at least as valuable as its p-value, since it may show regions of the data where the fit is either adequate or particularly poor. Users of the test should pay more attention to the table than to the p-value.

Advanced texts explore a wider range of alternative goodness-of-fit statistics that are beyond

the level of this text, such as those described in Section 5.2 of Hosmer, et. al[19] and Section 10.5 of Harrell.[20]

9.5.2 Estimating the accuracy of predictions

The Brier score

The **Brier score** B estimates prediction accuracy by comparing the predicted probabilities of the outcome to observed values:

$$B = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{p}_i)^2,$$

where $\hat{p}_i$ and y_i are the predicted probabilities and observed response, and n is the sample size. Like mean square prediction error in linear regression, the Brier score assesses fit by estimating the squared distance between observed and predicted values.

An observation y_i can take on only two values, 0 or 1, and $\hat{p}_i$ will be a number in the interval $(0, 1)$ since predicted probabilities are never exactly 0 or 1. When $y_i = 1$ and $\hat{p}_i$ is close to 1 or when $y_i = 0$ and $\hat{p}_i$ is close to 0, $\hat{p}_i$ is an accurate predictor for case i and the contribution to the Brier score will be close to 0. When the reverse happens ($\hat{p}_i$ is very different from y_i) the contribution to the Brier score will be close to 1. A Brier score close to 0 indicates that predictions are generally accurate; if it is close to 1, predictions are generally poor. When evaluating prediction accuracy, a low Brier score indicates a good prediction model.

There is no universal definition of a good Brier score, but a simple example helps. If all predicted probabilities are 0.50 (essentially, coin flips), the contribution of each case to the Brier score will always be 0.25, since $y_i - \hat{p}_i$ is always 0.5. So a Brier score of 0.25 is no better than guessing an outcome with probability 0.5. In most cases, investigators want a Brier score smaller than 0.20 or 0.15. In the hyperuricemia data, the Brier score for the model with predictor BMI (Equation 9.9) is 0.1459, suggesting reasonably accurate predictions overall; the Brier score when both BMI and sex are used is 0.1447, a small improvement that is consistent with the relatively small decrease in the AIC when sex is added to the model. The two variable model seems to be better, but not by much.

As will be seen in the methods for evaluating discrimination discussed later, a model may make reasonably accurate predictions overall, but be a poor predictor in some subsets of cases.

There are analogues to R^2 from linear models not covered here and can be found in more advanced texts, such as Agresti[21] and Hosmer, Lemeshow and Sturdivant.[22]

Calibration plots

Calibration plots are a visual display of the match between predicted probabilities and observed outcomes. Figure 9.20 shows calibration plots for the logistic models for hyperuricemia with the single predictor BMI (blue) and with predictors BMI and sex (green). Because the outcome is binary, the agreement between predicted probabilities and outcomes is difficult to see in a scatterplot of observed versus predicted values, so calibration plots typically add a best fitting smooth curve, using `loess` or a similar function in R. The largest values on the horizontal axis for the two curves are different to avoid extrapolation; the largest predicted probability is 0.722 for the model with BMI alone and 0.624 for the model that adds sex. If a model is well-calibrated , the smooth curve should lie close to the 45-degree line $y = x$ (the dotted line in the curve). The figure

[19] David W Hosmer Jr et al. *Applied logistic regression, 3rd ed.* John Wiley & Sons, 2013.
[20] Frank E Harrell et al. *Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis*. Vol. 3. Springer, 2015.
[21] Alan Agresti. *Categorical data analysis, 3rd ed.* Vol. 792. John Wiley & Sons, 2013.
[22] David W Hosmer Jr et al. *Applied logistic regression, 3rd ed.* John Wiley & Sons, 2013.

shows that both models are reasonably well-calibrated for predicted probabilities between 0.0 and 0.3, less so for probabilities larger than 0.4, where the data are sparse. The model including BMI and sex is closer to the 45-degree line for predicted probabilities less than 0.60 than the model with BMI alone. The behavior of the blue curve at the right edge of the plot is likely due to the "edge effects" of loess when data are sparse for large or small values on the horizontal axis.

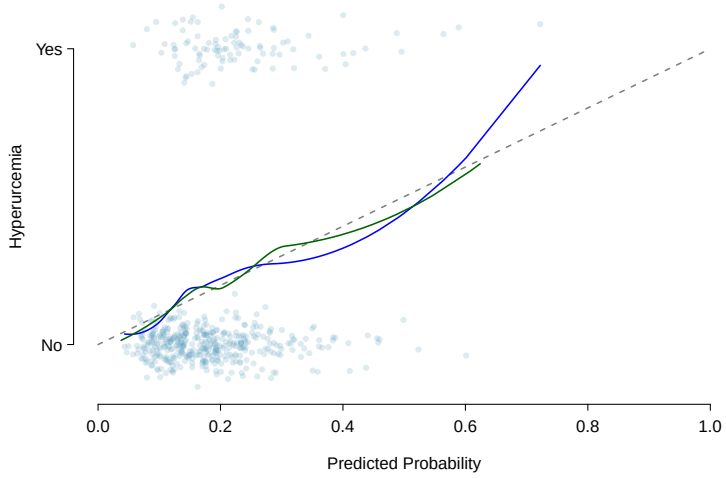


Figure 9.20: Calibration plots, logistic model for the association between hyperuricemia and the predictor BMI (blue) and the predictors BMI and sex (green). The light blue jittered dots at $y = 0$ and $y = 1$ denote observed values of hyperuricemia (0 = "No", 1 = "Yes") plotted against predicted probabilities. The smooth curves are drawn using the R function `loess` on the scatterplots of actual predicted versus observed probabilities for the two models.

As noted earlier, Figure 9.22 can be viewed as a calibration plot. Instead of fitting a smooth curve to the scatterplot of observed values and predicted probabilities, the agreement between outcomes and predicted probabilities is shown by examining the match between predicted probabilities and observed proportions of outcomes in buckets of the data.

Calibration plots are valuable, but their appearance depends on decisions made by the data analyst. The choice of buckets when comparing proportions to predicted probabilities is arbitrary, and the choice of parameters in the estimated loess curve can affect the appearance of the curve.

Estimating discrimination

Predicted probabilities from a logistic model can be used to group cases into two groups – those predicted to have versus not have the outcome of interest. A naive but often used approach is to predict that a case will have the outcome if the predicted probability is 0.50 or greater, and to predict the outcome will not happen otherwise. The value 0.50 is called a threshold value for predicting an outcome. Any value between 0 and 1 can be used as a **threshold probability**, and 0.50 may not always be the best one. A good model is reasonably successful at discriminating between cases likely versus not likely to have an event.

Suppose y is an observed binary outcome, and $\hat{y}$ its predicted value. If $\hat{p}_i$ is the predicted probability for case i in the data set, the naive prediction rule is

$$\hat{y}_i = 1 \text{ if } \hat{p}_i \geq 0.50, \text{ and}$$
$$\hat{y}_i = 0 \text{ if } \hat{p}_i < 0.50.$$

If this rule were applied to the hyperuricemia data using the model in Figure 9.12, a patient

would be predicted to be hyperuricemic if the predicted probability based on BMI and sex was 0.50 or larger. Figure 9.21 shows observed versus predicted hyperuricemia using 0.50 as a threshold.

The number of cases with correct predictions is the total number of cases where the predicted and observed are both "No" or both "Yes", or the sum of the diagonal elements, 402 + 4 = 406. The prediction rule is correct 406/500 = 81.6% of the time and incorrect 18.4% of the time. The total error rate for the prediction rule is 0.184.

	Observed		
Predicted	No	Yes	Sum
No	402	91	493
Yes	3	4	7
Sum	405	95	500

Figure 9.21: Observed versus predicted hyperuricemia, threshold value 0.50, logistic model with predictors BMI and sex.

The error rates among cases with or without the outcome can be very different from the total error rate. The false negative rate, or FNR, of a prediction rule is the proportion of times cases with the outcome are predicted not to have it; it is an estimated conditional probability. Figure 9.21 shows that among the 95 cases that were hyperuricemic, 91 were predicted to be free of hyperuricemia, a false negative rate of 91/95 = 0.958. The false positive rate , or FPR, is the proportion of times cases without the outcome are predicted to have it. For the prediction rule that uses a threshold of 0.50, the false positive rate is 3/405 = 0.007.

If BMI and sex were used to screen for the possibility of hyperuricemia in a population similar to the study population, the large false negative rate indicates that it would never be used in practice. More than 95% of patients with undiagnosed hyperuricemia would be falsely predicted not to have the condition.

When sex was added to `bmi` in the model for hyperuricemia, the AIC for the two variable model was slightly smaller than for the single variable model (461.26 vs. 463.26), suggesting that the two variable model might provide more accurate predictions. Figure 9.22 shows observed versus predicted numbers of cases of hyperuricemia in the model with `bmi` alone. Comparing it with Figure 9.21 shows the small differences between the model predictions. The model with sex predicts an additional false negative case, and one fewer false positive.

	Observed		
	0	1	Sum
FALSE	403	92	495
TRUE	2	3	5
Sum	405	95	500

Figure 9.22: Predicted versus observed instances of hyperuricemia, threshold value 0.50, logistic model with predictor BMI

The error rates of a prediction rule change when the threshold value changes. Increasing the threshold will lead to both more cases correctly being predicted as having the outcome (more true positive results) and more cases incorrectly being predicted as having the outcome (more false positive results). Since there is an increase in true positives, the FNR decreases; since there is an increase in false positives, the FPR increases.

Figure 9.23 shows how the FPR and FNR change with the threshold value for the prediction rule. Choosing a threshold is not a statistical problem; it involves assessing which of the two error rates should be kept small, and that will depend on the clinical situation. In settings where it is important to avoid missing cases, it is reasonable to prioritize keeping the false negative rate small. However, it may be the case that for some conditions, the intervention that follows after a

positive result has serious side effects; this would be a justification for keeping false positive rates small.

Figure 9.23 shows that the FPR and FNR are approximately 0.40 at a threshold of approximately $\hat{p} = 0.20$. Any threshold value that yields an FNR lower than 0.40 will lead to an FPR larger than 0.40; correspondingly, reducing the FPR by changing the threshold will increase the FNR. The figure reinforces the conclusion that the predictors BMI and sex do not provide enough information to accurately predict hyperuricemia, even though the calibration plot in Figure 9.20 indicates that the model is a good fit to the data. There is more variability in the outcome than is captured by the model. This is analogous to linear regression where residual plots indicate that a model is a reasonable fit to data but the R^2 is low.

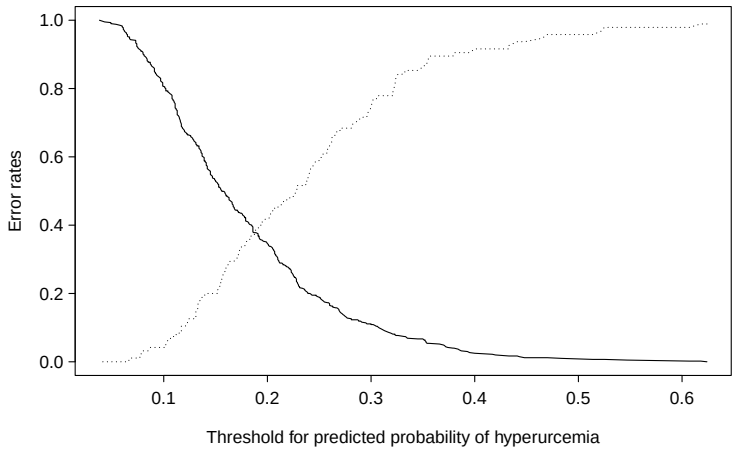


Figure 9.23: Estimated false positive (solid line) and false negative (dotted line) probabilities of hyperuricemia as a function of estimated cutoff value for the predicted probability of hyperuricemia. Predicted probabilities are from the logistic model for the odds of hyperuricemia as a function of BMI and sex.

A receiver operating characteristic (ROC) curve is another graphic that shows how a binary classification rule behaves as its threshold value changes. The ROC curve plots the true positive rate (TPR) on the vertical axis against the false positive rate (FPR) on the horizontal axis at each threshold setting for the predicted probability of the outcome. An ROC curve shows directly that increases in the true positive rate can only be achieved by increasing the false positive rate.

Figure 9.24 shows the ROC curve for the model for hyperuricemia based on BMI and sex. When the FPR is approximately 0.40, the TPR is approximately 0.60. Figure 9.23 shows that this corresponds to a threshold value of 0.20.

If a prediction rule is used as a diagnostic test in a clinical setting (e.g., one which predicts whether someone has a disease), the TPR is the sensitivity of the test, and the FNR is 1 - specificity. ROC curves are widely used in evaluating diagnostic tests, and are often defined equivalently as plotting sensitivity against 1 - specificity.

The 45-degree line on an ROC plot is used to distinguish between tests which may have some value vs. those which are worse than guessing. Let D and D^C indicate the presence or absence of a disease D respectively, and + and − indicate a positive or negative test. For points on the 45-degree line, sensitivity equals 1 - specificity, so

$$\begin{aligned} P(+|\mathrm{D}) &= 1 - P(-|\mathrm{D}^C) \\ &= P(+|\mathrm{D}^C). \end{aligned}$$

The likelihood of a positive test is the same whether or not disease is present. Bayes' rule can be

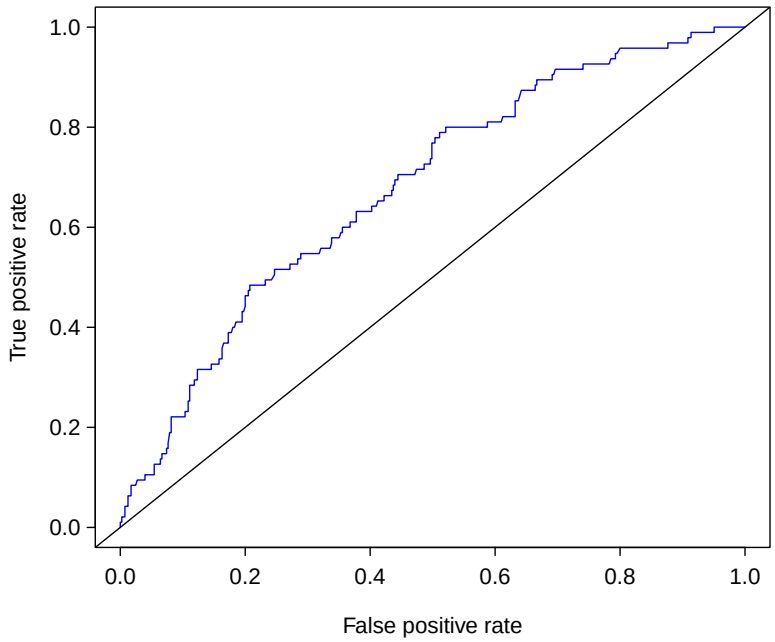


Figure 9.24: Receiver operating curve (ROC) for predicting hyperuricemia from the logistic model for the odds of hyperuricemia as a function of BMI and sex. The ROC curve is in blue; the black line is the is 45-degree line $y = x$.

used to show in this case that

$$\frac{P(\mathrm{D}|+)}{P(\mathrm{D}^C|+)} = \frac{P(D)}{P(\mathrm{D}^C)}.$$

In other words, the odds of disease given a positive test is the same as the odds of disease in the population in the absence of the test. Diagnostic tests in which the ROC curve lies on the 45-degree line are no better than guessing the presence of disease using the overall prevalence.

The same algebra can be used to show that tests with ROC curves above the 45-degree line, the odds of disease given a positive test are larger than the odds of disease without the test, i.e., the diagnostic test is better than guessing based on the prevalence. For tests with ROC curves below the 45-degree line the odds of disease given a positive test are lower than the odds of disease in the population; the diagnostic test is worse than guessing based on the population prevalence.

The area under an ROC curve (labeled **AUC**, **AUC-ROC** or the **c-statistic**) is 0.5 when the curve is the 45-degree line, larger than 0.5 when the curve lies above the 45-degree line (the test is better than guessing) test and smaller than 0.5 when the curve lies below the 45-degree line (the test is worse than guessing). It is possible to show

$$\mathrm{AUC} = \frac{P(+|\mathrm{D})}{P(+|\mathrm{D}^C)}.$$

A randomly selected member of the population without the disease is less likely to test positive by a factor of the value of AUC than a member selected from the population with the disease. More simply, the diagnostic test performs better in the population with than without the disease.

Software can be used to calculate estimates and confidence intervals for AUC for a given ROC. The analyses in this and the next section use the R package `cvAUC`.

How much better than random guessing is a prediction rule for hyperuricemia based on BMI and sex? The estimated AUC for Figure 9.24 is 0.678 with 95% confidence interval (0.620, 0.726). With 95% confidence the model has a estimated chance of 62% to 73% of correctly distinguishing

between an individual with versus without hyperuricemia. There is no single definition of a good AUC, but there are guidelines that may be useful in some settings. For biomedical data, Hosmer, et. al.,[23] recommend the guidelines in Figure 9.25.

AUC under ROC curve	Suggested interpretation
0.50	No discrimination, no better than random guessing
(0.50, 0.70)	Poor discrimination
[0.70, 0.80)	Acceptable discrimination
[0.80, 0.90)	Excellent discrimination
[0.90, 1.00)	Outstanding discrimination

Figure 9.25: Hosmer, Lemeshow and Sturdivant suggested guidelines for interpreting the area under the ROC curve (AUC).

Using these guidelines, the AUC for the model for hyperuricemia with predictors BMI and sex discriminates poorly between cases with and without hyperuricemia – adding one more piece of information that the model would not be useful in a clinical setting.

Figure 9.26 shows four example ROC curves corresponding to hypothetical models with increasing ability to discriminate: the 45-degree line, AUC = 0.5, random guessing; the green curve, AUC = 0.667, poor discrimination; blue curve, AUC = 0.785, acceptable discrimination; red curve, AUC = 0.874, excellent discrimination. ROC curves will be used to compare models for triage in an emergency department in the case study in Section 9.6.

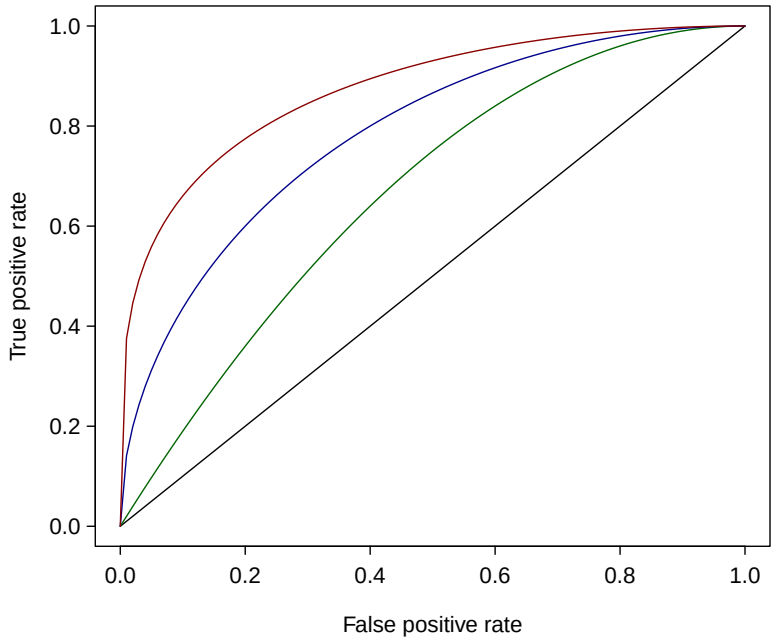


Figure 9.26: Example Receiver ROC curves for hypothetical prediction rules. The black line is the is 45-degree line $y = x$.

9.5.3 Out-of-sample accuracy of predictions

When predicting outcomes is the goal of an analysis, the data used to estimate a prediction model is usually called the training data . When the prediction error from a model such as a Brier

[23] David W Hosmer Jr et al. *Applied logistic regression, 3rd ed.* John Wiley & Sons, 2013.

score or a false negative rate is estimated using the training data, the estimate is called an in-sample prediction error or an apparent prediction error. Methods such as maximum likelihood (used to estimate a logistic regression model) choose parameter estimates that are well matched to the data, so in-sample prediction error is generally smaller than the prediction error in new data where the relationships between outcome and predictors may be slightly different. Out-of-sample prediction error characterizes the behavior of a model when fit to new data. Out-of-sample prediction error can be estimated in a new dataset, usually called test data or validation data or using cross-validation when a validation dataset is not available. The use of a validation dataset is illustrated in Section 9.6.5; this section outlines cross-validation.

Cross-validation

Cross-validation estimates out-of-sample prediction error by repeatedly resampling from the training data to create a collection of paired training and test datasets. In *k*-**fold cross-validation** the data are randomly divided into k non-overlapping, approximately equal sized subsets, called folds; typically $k = 5$ or 10. Each fold is used as training data to re-estimate a model, then a prediction error (e.g., a Brier score or a false negative rate) is estimated by applying the re-estimated model to the data not in the fold, i.e., the data held out from the fold. The process produces k estimates of prediction error, which are then averaged. When the fold sizes are identical, a simple average can be used since each estimate of prediction error is based on the same amount of data. Figure 9.27 shows a graphical representation of 5-fold cross-validation.

The randomly chosen subsets i.e., the folds, use training datasets that may reflect different associations between the response and predictors, so even though cross-validation uses the training data its estimates of error rates are less subject to the bias of in-sample estimates of error.

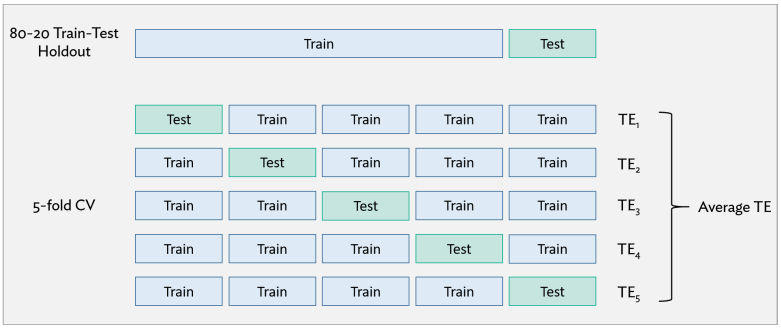


Figure 9.27: 5-fold cross-validation.

Cross-validated Brier scores for the hyperuricemia data can be calculated using the R program `cv.glm` in the package `boot` or using the code in the lab for this chapter. For the model with just BMI, the 5-fold cross-validated Brier score is 0.1484, larger than the apparent Brier score 0.1459. The cross-validated Brier score for the model with BMI and sex is 0.1474. Adding sex to the model with BMI reduces the Brier score by only 0.001.

Cross-validation can be used for more complicated estimates, such as a calibration curve, but the principle is the same. For each fold, a model is re-estimated and the calibration curve is constructed using the data held out from the fold. For 5-fold calibration, the resulting graph might show the 5 calibration curves as well as a curve constructed using the average value at each point on the 5 estimated curves.

Cross-validation has some drawbacks – the best choice of k for a given dataset is not always clear, the test sets are relatively small, and unlike using an external dataset, estimates of error rates do not generalize to populations that might differ in important ways from the study sample. However, validation datasets large enough to be useful are rarely available. Cross-validation has some strengths: it is available in software such as R, does not require an additional dataset, allows

for relatively large training sets, and averaging the k estimated prediction error rates mitigates the small size of the test sets. When a validation dataset is not available, cross-validation is the preferred method of estimating out-of-sample prediction error.

As noted earlier, prediction from statistical models is something that should be done with great care, especially in a clinical situation where prediction may be a diagnostic tool leading to an intervention. In this setting is important to examine a statistical model from several perspectives.

What do these methods for assessing a model tell us about the relatively simple model for predicting hyperuricemia from BMI and sex along with the initial look at significance levels for predictors and changes in AIC earlier in the chapter? Although the significant p-values for BMI and sex did not have the interpretation as tests of predefined hypotheses, they suggested a potentially important association between the predictors and hyperuricemia. While comparing AIC values did not provide a clear answer as to the value of adding sex as a predictor, the calibration plots suggested that the two-variable model better fit the data than the model with only BMI. Brier scores indicate that the two predictors may provide acceptably accurate predictions overall, but predictions within subsets of individuals either with or without underlying hyperuricemia were not always accurate, even after adjusting the threshold probability for predicting presence or absence. In summary, a logistic regression with BMI and sex fits a model for the log(odds) of hyperuricemia reasonably well, but not well enough to be used as a diagnostic tool.

The hyperuricemia data is useful for exploring how logistic regression might describe the association between an outcome and predictors, but it is a simple example that does not reflect the complexity of many clinical situations. The next section presents a case study on improving a triage strategy in an hospital emergency department based on a published paper.

9.6 Case study: Triage in an emergency department

9.6.1 Introduction and background

Most hospital emergency departments triage arriving patients so that the most severely compromised are given higher priority. It is an especially valuable process when the case load is high, since waiting time to treatment is an important factor in outcome. This section presents a case study developing a logistic regression model for triaging patients using data from Kristensen, et al.,[24] a cohort study conducted in the Emergency Department (ED) of the Nordsjælland University Hospital in Denmark. In the paper, the study team proposed a revision to an ED triage algorithm based on predictions for the probability of death within 30 days from admission (30-day mortality). The study used a primary cohort of 6,249 participants to model alternative triage algorithms and a validation cohort of 6,383 individuals to evaluate the models.

At the time the study was published, the hospital used the Danish Emergency Process Triage (DEPT) algorithm, a 5-level system ranking patients based on vital signs and presenting conditions (listed in the Kristensen paper) that assigns color codes for the predicted 30-day mortality probability. Let p be the probability of a patient dying within 30 days from admission to the ED. The color codes correspond to the following values of p: "red", $p > 0.25$; "orange", $0.10 < p \leq 0.25$; "yellow", $0.01 \leq p \leq 0.10$; "green", $p < 0.01$; and "blue", minor conditions for which the patient should not be admitted to the ED. The analysis in this section uses the term target probabilities for the probability ranges associated with each color. Patients in category "blue" are not included either in the published analysis or the one presented here, making the triage classification a 4-level. The colors for target probabilities in the 4 remaining categories can be thought of as risk categories for a death within 30 days of admission: high risk ("red"), moderately high risk ("orange"), moderately low risk ("yellow"), and low risk ("green").

Based on prior studies, the Kristensen team conjectured that revising DEPT using the results of routine biochemical screening normally done in an ED (albumin, creatinine, c-reactive protein, hemoglobin, lactate dehydrogenase, leukocyte count, potassium, and sodium) would improve the algorithm compared to the previous scoring based on vital signs and presenting conditions. The analysis in the Kristensen paper showed that was indeed the case.

This section examines a simpler modification of DEPT – adding the demographic variables age and sex to the existing color rankings – for several reasons. A more complete analysis might use a logistic regression that adds age and sex to the original variables used to create DEPT but those variables were not available for this case study. Readers of this text are unlikely to be familiar with the definitions of the biochemical measurements and their clinical implications. The Nordsjælland group used transformations of these measurements to model increased risk of death for abnormally low or high values of the biochemical measurements, and the transformations used are beyond the scope of this text. The steps used to build and test models that add only age and sex to DEPT are similar to those examining more predictors. Finally, while the triage system augmented by age and sex does not improve DEPT as much as the model in the Kristensen paper, it does surprisingly well. It may not be a useful tool in an ED, but it is more than sufficient as an example to study risk classification. Readers interested in the full analysis should be able to read the Kristensen paper after mastering the material in this section.

Since the goal of this analysis is a potential prediction model, statistical significance levels are less relevant than model fit and predictive accuracy. In most cases, point estimates, standard

[24] Michael Kristensen et al. "Routine blood tests are associated with short term mortality and can improve emergency department triage: a cohort study of> 12,000 patients". In: *Scandinavian Journal of Trauma, Resuscitation and Emergency Medicine* 25.1 (2017), pp. 1–8.

errors, and confidence intervals are provided for summary statistics and model coefficients instead of p-values. Model fit is assessed with Hosmer-Lemeshow statistics and calibration curves; predictive accuracy is estimated using Brier scores and ROC curves.

The full data for both cohorts are contained the data package `oibiostat`, as `DanishEDPrimaryCohort` and `DanishEDValidationCohort`. These datasets have also been posted by the authors (`DOI:10.5061/dryad.m2bq5`). The datasets in `oibiostat` have been re-formatted for readability.

9.6.2 Examining the data

The training dataset used here is based on data from 6,203 participants from the original cohort of 6,279 for whom there were no missing values for the DEPT score, 30-day mortality, age and sex. Since this excludes only 1.2% (76/6279) of the cohort, there is little chance of bias caused by a complete case analysis.

Of the 6,203 participants, 325 (5.2%) died within 30 days from admission to the ED. Figure 9.28 shows the association of the original DEPT scoring with 30-day mortality. The scoring was based on prior studies of ED outcomes, so, as expected, the χ^2 test of for the null hypothesis of independence shows strong evidence of an association ($\chi^2 = 131$, on 3 df right tail area < 0.001). The scoring identifies clusters of cases with decreasing risk of dying within 30 days; the estimated probabilities of death decrease monotonically from 49/273 = 0.179 to 51/1972 = 0.026 as the categories change from "red" (highest risk) to "green" (lowest risk).

	Died within 30 days		
Triage classification	No	Yes	Sum
red	224	49	273
orange	1462	114	1576
yellow	2271	111	2382
green	1921	51	1972
Sum	5878	325	6203

Figure 9.28: Association of DEPT triage classification with 30-day mortality.

Figure 9.29 shows, however, that the observed proportion of deaths falls outside the predicted range for three of the four categories: "red", "orange" and "green". Since the observed proportion of deaths is less than the lower bound in the high risk categories "red" and "orange", too many low risk patients would be classified into those categories. The reverse happens with the low risk category "green"; the observed proportion of deaths is larger than the upper bound for the target probabilities. Too many higher risk patients would be classified as low risk.

	Likelihood of death within 30 days	
Triage classification	DEPT target probabilities	Observed proportion
red	(0.25, 1.00]	0.180
orange	(0.10, 0.25]	0.073
yellow	[0.01, 0.10]	0.047
green	[0.00, 0.01)	0.026

Figure 9.29: DEPT target probabilities versus observed proportion of death within 30 days for the DEPT color categories. The target probabilities are the ranges of 30-day mortality probabilities that define the color scores.

Figure 9.30 shows the left-skewed age distribution, with mean 59.6 and median 63 and

minimum and maximum ages 16 and 108. The maximum age of 108 is striking, and the histogram in shows that there are several elders in the study sample at least 100 years old. These cases have been left in the dataset for the initial analysis, but are re-examined during the modeling process.

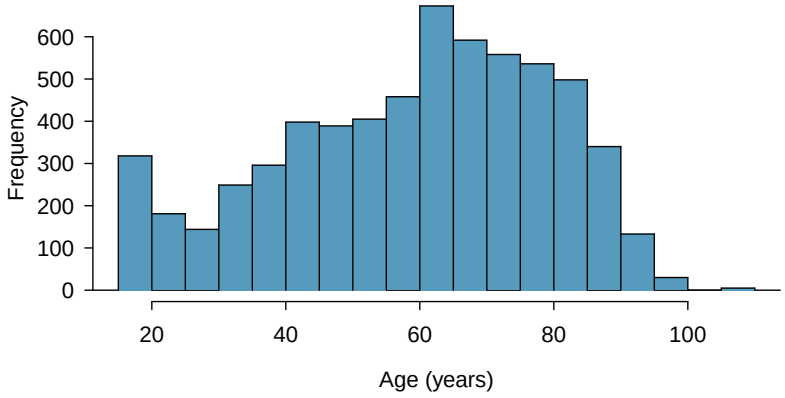


Figure 9.30: Histogram of age in the Danish ED data.

Figure 9.31 shows the association of age and 30-day mortality, with age grouped by quartile. The table shows (unsurprisingly) that the estimated probability of death increases monotonically from 6/1537 = 0.004 in the youngest quartile to 184/1542 = 0.119 in the oldest. The odds ratio comparing 30-day mortality in the oldest to youngest quartile is (184/1358)/(6/1531) = 34.5, 95% confidence interval (15.5, 95.5). The association of age with outcome is clearly important.

	Died within 30 days		
Age	No	Yes	Sum
(16,45]	1531	6	1537
(45,63]	1587	45	1632
(63,75]	1353	90	1443
(75,108]	1358	184	1542
Sum	5829	325	6154

Figure 9.31: Age quartile versus 30-day mortality

In these data, women are less likely than men to die within 30 days of admission to the ED. In Figure 9.32 the estimated probabilities for 30-day mortality for women and men are, respectively, 157/3217 = 4.9% and 168/2986 = 5.6%. The OR for 30-day mortality comparing men to women is (168/2818)/(157/3060) = 1.16, with 95% confidence interval (0.92, 1.46). The association between sex and outcome does not appear to be a strong one.

	Died within 30 days		
Sex	No	Yes	Sum
female	3060	157	3217
male	2818	168	2986
Sum	5878	325	6203

Figure 9.32: Association of sex with 30-day mortality.

9.6.3 Modeling the relationship between 30-day mortality and DEPT triage score, age and sex.

The initial analysis of the training data uses logistic regression with response variable 30-day mortality and predictors DEPT triage score, age and sex. Even though sex is not by itself strongly associated with 30-day mortality, it is included in the initial model to explore its relationship with outcome after adjusting for age and DEPT score. The conditions for a logistic regression are met in this dataset: the cases are independent; 369 individuals died within 30 days after admission so a model can have up to 33 parameters (369/11 = 33.5); and since the study did not gather data using outcome-based sampling, probabilities can be estimated from logistic regression. The calibration plots and goodness-of-fit statistics shown later support the assumption that a logistic model is reasonable for estimating the association between 30-day mortality and the predictors DEPT, age and sex.

Figure 9.33 shows the initial model. The columns labeled 2.5% and 97.5% are, respectively, the lower and upper bounds of 95% confidence intervals for the coefficients. Except for sex, all of the confidence intervals suggest substantial associations between the predictors and outcome – the upper bounds of the confidence intervals are substantially smaller than 0. The confidence interval for the coefficient of sex suggests a more moderate association but sill has a confidence interval that sex does not include 0.

	Estimate	Std. Error	2.5%	97.5%
(Intercept)	-5.589	0.367	-6.310	-4.869
triageorange	-1.244	0.198	-1.632	-0.856
triageyellow	-1.545	0.197	-1.932	-1.159
triagegreen	-2.122	0.223	-2.560	-1.685
age	0.059	0.004	0.050	0.067
sexmale	0.274	0.120	0.039	0.508

Figure 9.33: Logistic regression with response 30-day mortality and predictors DEPT triage, age and sex.

The role of the predictor sex warrants a closer look, for several reasons. In most countries, women outlive men, and that is true in the country where these data were collected. According to the Norwegian Institute of Public Health, the life expectancy Norwegian women in 2016 was 84.2 years versus 80.6 for men. This suggests that as Norwegians age, women are more robust than their male counterparts, suggesting that 30-day mortality rates for elders in an ED may be different for females than males. In fact, Figure 9.34 shows that the association between sex and 30-day mortality is very different within age groups. In the age group 16-45, the overall proportion of deaths within 30 days is low $((5+1)/(791+746) = 0.0039$, but the relative risk of death comparing males to females is $(1/746)/(5/791) = 0.212$. Males in this age group are approximately 80% less likely to die than females. In contrast, in the highest age category, the relative risk of death comparing males to females is 1.646. Males in this age category are approximately 65% more likely to die within 30 days. There appears to be an age-sex interaction in the risk of death within 30 days of admission.

A logistic model for 30-day mortality with the addition of an age-sex interaction is shown in Figure 9.35. None of the coefficient confidence intervals cover 0.

AIC statistics can be used to examine the potential predictive value of adding predictors to the base model with only the DEPT score in the sequence of models that add age (M_1), then sex (M_2), then the interaction age−sex (M_3). Figure 9.36 shows the deviance (D), number of predictors (p) and AIC ($D + 2(p+1)$) statistics for each of the 3 models. The values of the AIC statistics continue to decrease as parameters are added to the model so all of these variables will be retained.

		Died within 30 days	
Age category	Sex	Yes	No
(16,45]	female	786	5
	male	745	1
(45,63]	female	778	21
	male	809	24
(63,75]	female	650	46
	male	703	44
(75,108]	female	818	85
	male	540	99

Figure 9.34: 30-day mortality by sex, within each of the 4 age categories.

	Estimate	Std. Error	2.5%	97.5%
(Intercept)	-4.514	0.444	-5.383	-3.645
triageorange	-1.253	0.199	-1.643	-0.863
triageyellow	-1.564	0.198	-1.953	-1.175
triagegreen	-2.155	0.225	-2.595	-1.714
age	0.045	0.006	0.034	0.056
sexmale	-1.983	0.655	-3.266	-0.699
age:sexmale	0.030	0.009	0.013	0.047

Figure 9.35: Logistic regression with response 30-day mortality and predictors triage, age, sex and an age-sex interaction.

How well does model M_3 fit the data? The Hosmer-Lemeshow statistic does not provide evidence for a lack of fit ($\chi^2 = 9.7$ on 8 df, right tail area 0.3). Calibration curves, however, suggest something else.

Figure 9.37 shows calibration curves using the two methods discussed earlier – computing average predicted probabilities with observed proportions of outcomes in buckets of the data (the plotted black points), and fitting a smooth curve to the scatterplot of observed outcomes versus predicted probabilities (the solid blue curve).

The black points with vertical lines provide a view similar to Figure 9.22 used to show the fit of the model for the association between hyperuricemia and BMI. The estimated proportion is plotted against the average probability in each bucket, and the scatter above and below the dashed line $y = x$ shows the extent to which the observed proportions and predicted probabilities agree. Unlike Figure 9.22, the large size of this data set has been exploited by adjusting the buckets adaptively to place more buckets in regions with a high density of predicted probabilities, so that each bucket contains approximately 1% of the predicted probabilities. The solid blue line uses the R function `loess` to fit a smooth curve to predicted probabilities versus observed events.

Taken together, the two curves show that model predictions are reasonably accurate when the predicted probabilities are less then 0.2, but the smooth curve shows predicted probabilities larger than 0.2 are less accurate. The downward slope of the smooth blue curve indicates that observed outcomes happen less frequently that the model predicts. Age is the only predictor that is not categorical so large predicted probabilities may be caused by outliers in age.

In the training dataset there are 5 cases older than 100 years or older, and none died within 30 days. It is possible that these elderly cases are different from the rest of the population in important ways. There are two general approaches that might be used here – adapt the model using a transformation of the predictor age, or drop the cases 100 or older from the analysis and note that the subsequent model applies only to patients less than 100 years old. The analysis here uses the latter approach.

Figure 9.38 uses the same calculations as for Figure 9.37, but with the model re-estimated using the dataset restricted to patients less that 100 years old. The calibration plot shows the

Model	Deviance	No. Predictors (p)	AIC
M_1	2195.5	4	2205.5
M_2	2190.2	5	2202.2
M_3	2177.9	6	2192.9

Figure 9.36: Deviance and AIC statistics for the sequence of models M_1 - M_3

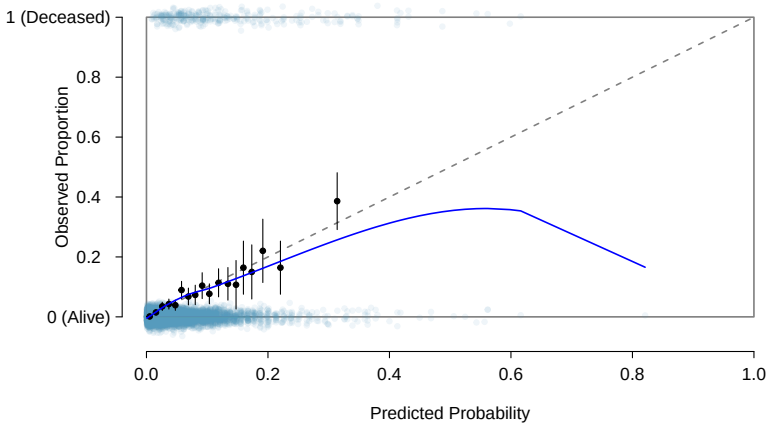


Figure 9.37: Predicted probabilities versus observed proportions, with data grouped adaptively into buckets of predicted probabilities (black points), and a smooth curve fit to the scatterplot of observed outcomes versus predicted probabilities. The light blue dots at $y = 0$ and $y = 1$ denote observed values of 30-day mortality (0 = "No", 1 = "Yes") plotted against predicted probabilities.

model fits the restricted dataset better than the full dataset. The Hosmer-Lemeshow goodness-of-fit statistic again shows no evidence of lack of fit, with $\chi^2 = 0.4$ on 8 df, right tail area = 0.40.

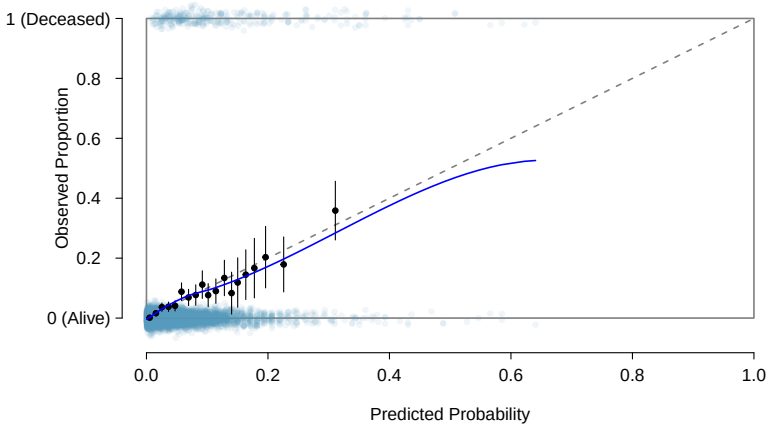


Figure 9.38: A figure with the same interpretation as Figure 9.37, but based on the dataset with cases removed whose age was ≥ 100 years.

Figure 9.39 shows the model coefficients using the age-restricted dataset. This is the model that will be used for a revised triage score in the next section.

	Estimate	Std. Error	2.5%	97.5%
(Intercept)	-4.518	0.445	-5.390	-3.646
triageorange	-1.301	0.199	1.692	-0.910
triageyellow	-1.611	0.199	-2.000	-1.221
triagegreen	-2.179	0.225	-2.619	-1.739
age	0.045	0.006	0.034	0.056
sexmale	-2.151	0.671	-3.465	-0.837
age:sexmale	0.033	0.009	0.015	0.050

Figure 9.39: Logistic regression with response 30-day mortality and predictors triage, age, sex and an age-sex interaction; dataset restricted to patients < 100 years old

9.6.4 Triaging patients with a modified score

The model in Figure 9.39 can be used to create a revised triage scoring system in the training data, using the same probability cutoff values as the DEPT classification, but now applied to individuals younger than 100. Using the risk descriptors rather instead of colors, individuals with a predicted probability larger than 0.25 are labeled high-risk, between 0.10 and 0.25, moderately high risk, between 0.01 and 0.10, moderately low risk, and less than 0.01, low risk.

Unless otherwise stated, all tables and figures in this section used the age-restricted dataset. Since this dataset differs slightly from the full dataset explored in Section 9.6.2, summary tables of predictors and outcome may differ from earlier tables.

Figures 9.40 and 9.41 compare the behavior of the old and new scoring, using the risk descriptors for both DEPT and the new scores. Figure 9.40 compares the behavior of the DEPT triage classification with the modified version. The second column shows the target ranges of 30-day mortality probabilities. The third and fourth columns show estimated 30-day mortality probabilities when participants are assigned a risk score using the DEPT or revised classification, respectively. The last column contains a 10-fold cross-validated estimate of the 30-day mortality probabilities using the revised classification. With the revised score, all 30-day mortality proportions now fall within the predicted ranges, as opposed to the DEPT scoring where 3 of 4 categories fell outside the predicted range. The 10-fold cross-validated estimates of mortality probabilities in the last column are based on the assumption that the model that adds age, sex and an age-sex interaction is fixed. The coefficients are re-estimated in each fold, and the mortality probabilities are estimated using the cases held-out of the fold. The values in the table show the estimates after averaging over the 10 folds. These out-of-sample estimates are generally consistent with the in-sample estimates in column 4.

The improvement in prediction accuracy of the revised score is the result of it placing fewer low risk patients in the high risk categories. The two-way table in Figure 9.41 shows that the revised triage classification places fewer patients than DEPT in the two highest risk categories. Only 115 of the 271 cases labeled high risk in DEPT are designated high risk; the remaining 156 are redistributed to lower risk categories. That also happens in the moderately high risk classification in DEPT; of the 1576 originally in that category, 797 are coded moderately high risk and the majority of the remaining cases are regrouped into lower risk categories.

The observed and cross-validated proportions in Figure 9.40 for the model using DEPT, age and sex is well-calibrated but the DEPT score based on the original model is not, at least in the training data.

The estimated Brier score for the DEPT classification in the training cohort is 0.049. Using 10-fold cross validation, the estimated and cross-validated Brier score for the revised triage is 0.046, a small improvement.

Section 9.6.5 uses an external dataset to check the predictions of the revised classification system.

Risk Category	Target Probability	DEPT Classification	Revised Classification	Cross-Validated Revised Classification
High	(0.25, 1.00]	0.181	0.409	0.373
Moderately High	(0.10, 0.25]	0.072	0.122	0.124
Moderately Low	[0.01, 0.10]	0.047	0.046	0.047
Low	[0.00, 0.01)	0.026	0.001	0.001

Figure 9.40: Estimated probability of 30-day mortality by risk category and triage score in the age-restricted training data. The last column shows a 10-fold cross-validated estimate of 30-day mortality using the revised score.

		Revised Predicted Risk				
		High	Moderately High	Moderately Low	Low	Sum
DEPT Predicted Risk	High	78	88	87	18	271
	Moderately High	29	399	971	177	1576
	Moderately Low	8	263	1631	480	2382
	Low	0	47	1198	724	1969
	Sum	115	797	3887	1399	6198

Figure 9.41: DEPT versus revised risk category in the age restricted dataset based on the model which adds age, sex and an age-sex interaction to the DEPT classification.

9.6.5 Evaluating the revised triage score

The team for the Danish study made available both a training dataset (their primary cohort) used in deriving their revision to the DEPT score and a test dataset (their validation cohort). The training and test data are based on cohorts treated in the emergency department during 2010 and 2013, respectively. The 2013 cohort consists of 6,383 individuals treated in the Nordsjælland University Hospital ED. The test dataset for this analysis consists of 6,224 participants with no missing values on DEPT score, age or sex, are not coded with DEPT score "blue" (no intervention needed) and are less than 100 years old. There were 249 deaths within 30 days after admission, a proportion of 249/6224 = 0.040, lower than the 325/6198 = 0.052 proportion in the age-restricted training data.

When a model estimated in a training dataset is evaluated in test data, the model and its coefficients are not re-estimated. Predictions, estimated Brier scores, etc., are calculated for the test data using the model estimated in the training set.

An ED triage scoring system performs adequately if, on average, it classifies patients into correct risk groups. The two right-most columns in Figure 9.40 show that in the training sample the DEPT score modified with the addition of age, sex and an age-sex interaction groups patients into categories with 30-day mortality proportions all in the target ranges. Figure 9.42 shows similar information for the test data.

Generally, a model is expected to perform less well in a test versus a training dataset. In this case, however, the Brier score in the test dataset is 0.036 compared to 0.046 in the training data. Evidently, the model including DEPT score, age and sex is more accurate in predicting probabilities in the test data than in the training data.

Along with the low Brier score, calibration curves help explain why the new score provides accurate predictions for 30-day mortality. Figure 9.43 shows the calibration curves that have been used earlier to evaluate model fit. Predicted probabilities and observed proportions are close.

The ROC curves discussed in Section 9.5.2 can be used to quantify the improvement in prediction from adding age, sex and the age-sex interaction to the DEPT scoring system. Figure 9.44 shows ROC curves for the DEPT and modified triage systems in the test data. The blue ROC curves correspond to the model that adds age, sex and an age-sex interaction to DEPT,

Risk Category	Target Probability	DEPT Classification	Revised Classification
High	$(0.25, 1.00]$	0.342	0.320
Moderately High	$(0.10, 0.25]$	0.060	0.125
Moderately Low	$[0.01, 0.10]$	0.038	0.034
Low	$[0.00, 0.01)$	0.020	0.001

Figure 9.42: Estimated probability of 30-day mortality by risk category and triage score in the age-restricted training data. Revised scoring uses coefficients from the model fit to the age-restricted training data with predictors age, sex and an age-sex interaction.

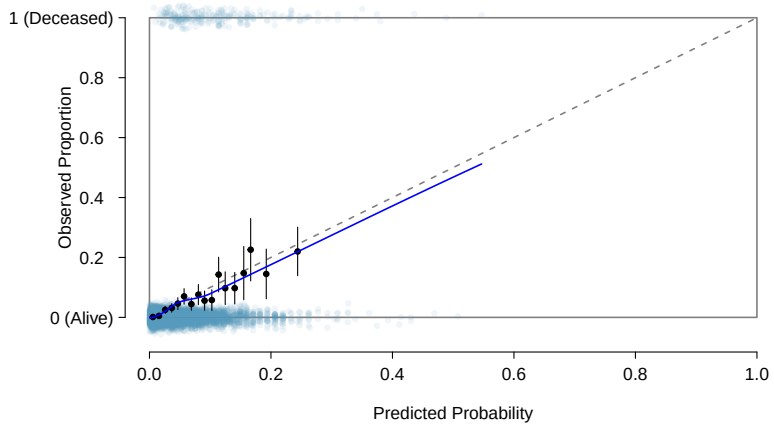


Figure 9.43: Calibration plot for the DEPT, age and age-sex interaction model showing predicted probabilities versus observed proportions in the age-restricted test dataset.

the green to DEPT alone. The area under the ROC curve (AUC) for the expanded model applied to the test data is 0.802 (95% confidence interval $(0.79, 0.825)$). The corresponding value for the DEPT classification alone is 0.632 (confidence interval $(0.590, 0.674)$). The DEPT classification has an estimated 63% chance of distinguishing between an individual who will survive at least 30 days after admission to an ED versus dying; the expanded model has an estimated 80% chance to make that distinction.[25]

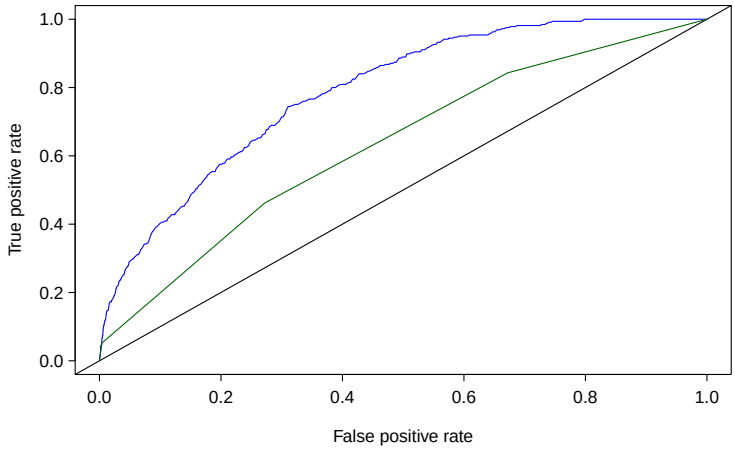


Figure 9.44: Receiver operating curves (ROC) for predicting 30-day mortality, green for the DEPT triage system, blue for the model that adds age and sex to DEPT. The black line is the is 45-degree line $y = x$.

[25]The AUC estimates and confidence intervals were calculation using the function `AUC` in the R package `cvAUC`.

9.6.6 Summary

The ED triage model explored here study illustrates the steps used in building and evaluating a prediction model. The analysis makes compromises, however, that limit its practical value. A more complete modification of the DEPT scoring should start by using the predictors used to construct DEPT, add age and sex to those predictors and construct a model based on the full set of predictors. The analysis here has not used any of the biochemical variables used in the Kristensen paper, many of which are strongly associated with 30-day mortality. The analysis dropped individuals 100 years old or older, rather than exploring transformations of the age variable that might have led to better fit with the full dataset and to a model that could be applied to all age groups. Incorporating an age-sex interaction led to a better fitting model, but it is possible that a model without the interaction term would be easier to interpret and still have made reasonably accurate predictions.

A full analysis would explore some questions raised by the analysis here. What is the reason for the surprisingly low mortality among the oldest members of the study population? Are they truly more hardy, or are very old patients referred to the ED for different, perhaps less serious conditions and still coded as high-risk? What might explain the age-sex interaction?

Despite the caveats, the analysis illustrates aspects that readers should look for in similar analyses. The source of the data should be clearly articulated; steps in model building should be clearly explained; model evaluation should incorporate diagnostic plots whenever possible and not rely on only numerical measures of fit. The predictive ability in models for binary outcome data should always include ROC curves and estimates and confidence intervals for AUC.

9.7 Notes

This and other chapters use body mass index (BMI) in several examples and analyses because it is widely available and has been measured and recorded in many studies of human populations. Despite its widespread use, BMI is now increasingly questioned because of its potential bias when applied to certain populations.

BMI was first proposed in 1832 by Adolphe Quetelet based on data from Caucasian western Europeans and was originally known as the Quetelet Index. The category labels for BMI were set in 1995 by an expert panel sponsored by WHO but their applicability in Asian and other populations have been questioned and studied. Several large studies have confirmed that high and low values of BMI in Asians confer an elevated risk of death just as in European populations (for example Lin, et al.[26]) but did not find evidence that the WHO cutpoints should be adjusted for Chinese populations. Nevertheless, we chose not to use the current WHO categories when analyzing the hyperuricemia dataset. In general, labeled categories associated with cutpoints in a continuous predictor should be interpreted with caution. Chapter 2 of Wiggins and Jones[27] describe some of Quetelet's work on BMI and his place in the history of statistics.

The examples using the TB dataset were chosen to illustrate concepts in logistic regression rather than a detailed analysis and examine only the two predictors level of education and presence of MDR TB. For readers interested in a more in-depth look at the data, the paper by Lackey referenced in Section 9.3.1 uses logistic regression to examine the association between TB treatment interruption a many more predictors.

[26] Wen-Yuan Lin et al. "Body mass index and all-cause mortality in a large Chinese cohort". In: *Cmaj* 183.6 (2011), E329–E336.

[27] Wiggins C and Jones ML. How Data Happened. A History from the Age of Reason to the Age of Algorithms (2023). WW Norton and Company.

9.8 Exercises

9.8.1 Introduction to simple logistic regression

9.1 Odds and probabilities. Suppose an experiment consists of rolling a fair six-sided die once.

(a) What are the odds of rolling a six?

(b) What are the odds of rolling an even number?

(c) Explain to someone who has not taken statistics the interpretation of the odds versus the probability of rolling an even number.

9.2 Diabetes. In the United States, approximately 9% of the population have diabetes.

(a) What are the odds that a randomly selected member of the US population has diabetes?

(b) Suppose that in a primary care clinic, the prevalence of diabetes among the patients seen in the clinic is 12%. What is the probability that a randomly selected patient in the clinic has diabetes?

(c) If in a particular population the probability of diabetes is twice what it is in the general population, does the odds of diabetes double?

9.3 Hyperuricemia and BMI, Part I. The fourth quintile of BMI in Figure 9.2 ranges from 25.02 to 26.64 meters per $(\text{kg})^2$ and has median value 25.93.

(a) Calculate the estimated conditional odds and probability of hyperuricemia for the value `bmi` = 25.93 using the model shown in Figure 9.5.

(b) Does the conditional probability of hyperuricemia calculated for the fourth quintile in Figure 9.2 lie above or below the value estimated in part (a)?

9.4 Interpreting model parameters, Part I. The curve with the solid line in Figure 9.4 corresponds to $\beta_0 = -3.0$ and $\beta_1 = 0.6$.

(a) Using the formula for the curve, calculate the odds ratio for E comparing $x = 6$ to $x = 4$.

(b) Using this curve, calculate the relative risk of the event E comparing the value of the predictor $x = 6$ versus $x = 4$.

(c) What role does the intercept play in the two calculations in (a) and (b)?

9.5 Interpreting model parameters, Part II. The curve with the dotted line in Figure 9.4 corresponds to $\beta_0 = 3.0$ and $\beta_1 = -0.6$.

(a) Using the formula for this curve, calculate the odds ratio for E comparing $x = 6$ to $x = 4$.

(b) Calculate the relative risk of the event E comparing the value of the predictor $x = 6$ versus $x = 4$.

(c) What role does the intercept play in the two calculations in (a) and (b)?

9.6 CPR and survival to discharge, Part I. Suppose a logistic regression model is used to estimate the association of the odds of surviving to discharge and the number of minutes cardiopulmonary resuscitation (CPR) was given to patients admitted to an emergency room following cardiac arrest. The response variable is survival to hospital discharge and the predictor is length of CPR in minutes. In the model the coefficient of CPR time is -0.065.

(a) Is increased time of CPR associated with a increase or decrease in the chance of survival to discharge?

(b) What is OR for survival to discharge comparing someone given CPR for 10 versus someone requiring 20 minutes of CPR time?

(c) In three sentences, describe your answers to parts (a) and (b) to someone who has not studied statistics.

9.7 CPR and survival to discharge, Part II. Suppose in the model for CPR and survival to discharge the coefficient of the intercept is 1.44.

(a) What are the odds of survival to discharge for someone requiring 10 minutes of CPR?
(b) Check your answer to Part I(b) by calculating the odds of survival to discharge for someone requiring 20 minutes of CPR and using it and the answer to (a) above to calculate the the OR for 10 versus 20 minutes of CPR.
(c) Calculate the estimated probabilities of survival to discharge for 10 and 20 minutes of CPR.
(d) What is the relative risk of survival to discharge, comparing 10 versus 20 minutes of CPR.
(e) Explain the distinction between the estimated OR and RR to someone who has not taken statistics.

9.8 Hyperuricemia and dietary magnesium, Part I. The investigators who studied hyperuricemia in China also measured daily dietary intake of magnesium. The logistic regression model for the association between hyperuricemia (the response variable) and dietary magnesium (the predictor, measured in units of 1 gram) is given in the table below.

Intercept	Magnesium (per gram)
-1.46	0.033

(a) Write the algebraic form of the logistic regression model for the association of hyperuricemia and dietary magnesium.
(b) Is dietary magnesium positively or negatively associated with hyperuricemia?
(c) What are the predicted odds of hyperuricemia for someone with 0.5 grams magnesium/day in their diet?
(d) By what factor will predicted odds change if a person with 0.5gm of dietary magnesium reduces their intake by 50%?
(e) What is the predicted probability of hyperuricemia for someone with 0.5gm magnesium in their daily diet?
(f) By what factor will predicted probability change if a person with 0.5gm of dietary magnesium reduces their intake by 50%?

9.9 Hyperuricemia and age. The logistic regression model for the association between hyperuricemia (the response variable) and age (the predictor, measured years) is given in the table below.

Intercept	Age (per year)
-1.089	-0.007

(a) Write the algebraic form of the logistic regression model for the association of hyperuricemia and age.
(b) Is increasing age associated with an increase or decrease in the odds of hyperuricemia?
(c) What are the predicted odds of hyperuricemia for a 50 year old from this population?
(d) By what factor will predicted odds differ between someone who is 30 and someone who is 50 years old?
(e) What is the predicted probability of hyperuricemia for a 50 year old?
(f) What is the relative risk of hyperuricemia, comparing a 50 year old to a 30 year old?

9.8.2 Inference for Simple Logistic Regression

9.10 Logistic Regression short answer, Part I. For the true/false questions, provide a reason for your answer. The short answer questions can usually be answered in 2 - 3 sentences.

(a) True or false: Equation 9.6 can always be used to estimate probabilities after fitting a logistic regression.
(b) True or false: Using the results of a logistic regression, the odds ratio for two cases with numerical predictor values 100 and 110 will be the same for two different cases with predictor values 20 and 30.
(c) In your own words, explain the concepts of the odds of an event.
(d) Suppose in a dataset, a binary outcome is a response variable and there is a single numerical predictor. True or false: if both linear and logistic regression models are fit to the data, the estimated slopes will have the same interpretation.

9.11 Logistic Regression short answer, Part II. For the true/false questions, provide a reason for your answer. The short answer questions can usually be answered in 2 - 3 sentences.

(a) True or false: Since the sampling distributions of the estimated parameters in a logistic regression do not depend on sample size, logistic regression can be fit to arbitrarily small data sets.

(b) Suppose a logistic regression has been fit to a dataset and the estimated slope parameter for the log(odds) is 0.750. Are increasing values of the predictor associated with increased or decreased risk of the outcome?

(c) Suppose the dataset was gathered in a prospective study with exposure based sampling. Is the information in part (b) sufficient to estimate the probability of the outcome, given a value of the exposure variable?

(d) If the standard error of the estimate in part (b) is 0.650, does the study provide strong evidence for the association of the predictor with outcome?

9.12 TB treatment interruption and sex of the participant. The two-way table in Figure 9.45 shows the relationship between the occurrence of a two-month TB treatment interruption and the recorded sex of the study participant in the TB data. Figure 9.46 contains the result of a logistic regression fit to the participant level data with response two-month interruption and predictor sex, with "female" as the reference category.

	Treatment Interruption		
	No	Yes	Sum
Female	461	29	490
Male	645	98	743
Sum	1106	127	1233

Figure 9.45: Sex versus a two-month treatment interruption

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-2.766	0.191	-14.449	0.000
sexMale	0.882	0.220	4.008	0.000

Figure 9.46: Logistic regression, response two month interruption, predictor recorded sex of participant

(a) Show that the odds ratios for an interruption, comparing males to females, is the same in using the data in the table and the logistic regression.

(b) Show that a 95% confidence interval for the OR in part (a) can be calculated using either the logistic regression output or the methods outlined Section 8.6.4 for ORs calculated in two-way tables.

9.13 Hyperuricemia and dietary magnesium, Part II. The table below shows more detail about the logistic regression model for hyperuricemia and dietary magnesium.

	Estimate	Std. Error
(Intercept)	-1.462	0.229
magnesium.intake.gm	0.033	0.526

(a) What is the value of the z-statistic used to test the null hypothesis of no association between hyperuricemia and dietary magnesium?

(b) Do the data show a statistically significant association between hyperuricemia and dietary magnesium?

(c) Construct a 95% confidence interval for the coefficient of dietary magnesium. What is the interpretation of the interval?

(d) Construct a 95% confidence interval for the odds ratio comparing individuals with 0.75gm versus 0.25gm of daily dietary magnesium.

9.14 Hyperuricemia and age, Part II. The table below shows additional details of the logistic regression model for the association between hyperuricemia and age.

	Estimate	Std. Error
(Intercept)	-1.089	0.817
age	-0.007	0.015

(a) What is the value of the z-statistic for testing the null hypothesis of no association between hyperuricemia and age?

(b) Do the data show a statistically significant relationship between hyperuricemia and age?

(c) Construct and interpret a 95% confidence interval for the coefficient of age.

(d) Find a 95% confidence interval for the odds ratio for hyperuricemia comparing a 75 versus a 50 year old individual.

9.15 Rare events . Public health research often involves the study of the association between an exposure and rare events. Radiation of certain wavelengths, called ionizing radiation, may have sufficient energy to damage DNA in a way that may lead to cancer. Radon is a form of ionizing radiation that is found in many homes and is known to cause lung cancer. It is produced from a natural breakdown of uranium in soil, rock and water. Radon is measured in in picocuries per liter, (pCI/L), and the US Environmental Protection Agency considers an average exposure of 4 pCI/L a safe level for adults.

Suppose a team is studying the possibility that pediatric leukemia may be associated with a low dose of radon exposure during pregnancy. In 10,000 randomly selected homes in a metropolitan area, the team records radon levels (in picocuries per liter, pCI/L) and whether or not a woman in the home is pregnant. One year later the team records whether or not the recorded pregnancies led to a successful birth and, if so, the health status of the infant.

(a) Suppose 1,500 of the women in the homes successfully delivered infants (all singleton births) and of those infants, 0.25% of the infants were diagnosed with a from of leukemia. Does the team have sufficient data to study the association of the dose of radon and the diagnosis of leukemia in an infant using logistic regression?

(b) Assume that the estimated proportions of successful pregnancies and a subsequent diagnosis of leukemia in an infant are accurate in this metropolitan area. What is the minimum number of homes the team should sample to reliably use logistic regression to study a dose-response relationship between infant leukemia and radon?

(c) Suggest a way that the data from the original study be used to calculate a larger sample size that would be more likely to yield enough events to use logistic regression in this setting, and calculate that sample size using your suggestion.

9.16 Challenger disaster, Part I. On January 28, 1986, a routine launch was anticipated for the Challenger space shuttle. Seventy-three seconds into the flight, disaster happened: the shuttle broke apart, killing all seven crew members on board. An investigation into the cause of the disaster focused on a critical seal called an O-ring, and it is believed that damage to these O-rings during a shuttle launch may be related to the ambient temperature during the launch. The table below summarizes observational data on O-rings for 23 shuttle missions, where the mission order is based on the temperature at the time of the launch. *Temp* gives the temperature in Fahrenheit, *Damaged* represents the number of damaged O- rings, and *Undamaged* represents the number of O-rings that were not damaged.

Shuttle Mission	1	2	3	4	5	6	7	8	9	10	11	12
Temperature	53	57	58	63	66	67	67	67	68	69	70	70
Damaged	5	1	1	1	0	0	0	0	0	0	1	0
Undamaged	1	5	5	5	6	6	6	6	6	6	5	6

Shuttle Mission	13	14	15	16	17	18	19	20	21	22	23
Temperature	70	70	72	73	75	75	76	76	78	79	81
Damaged	1	0	0	0	0	1	0	0	0	0	0
Undamaged	5	6	6	6	6	5	6	6	6	6	6

(a) Each column of the table above represents a different shuttle mission. Examine these data and describe what you observe with respect to the relationship between temperatures and damaged O-rings.

(b) Failures have been coded as 1 for a damaged O-ring and 0 for an undamaged O-ring, and a logistic regression model was fit to these data. A summary of this model is given below. Describe the key components of this summary table in words.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	11.6630	3.2963	3.54	0.0004
Temperature	-0.2162	0.0532	-4.07	0.0000

(c) Write out the logistic model using the point estimates of the model parameters.

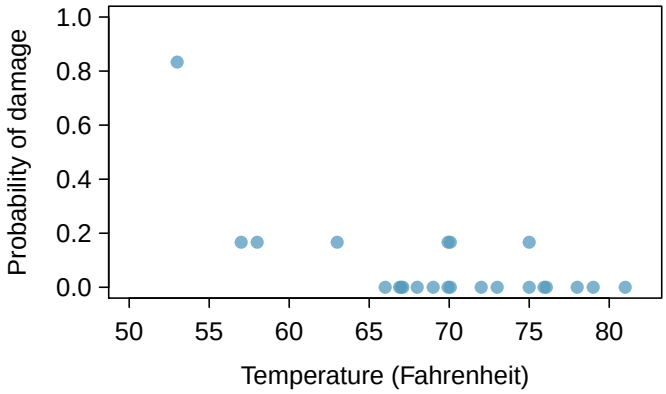
(d) Based on the model, do you think concerns regarding O-rings are justified? Explain.

9.17 Hyperuricemia and BMI, Part II.

(a) Use the entries in Figure 9.8 to calculate a 95% confidence interval for the odds ratio for hyperuricemia comparing two individuals with BMI 27 and 23.

(b) Ignoring issues of multiple testing, can the interval be used to support the claim that the data show that a BMI of 27 puts someone at significantly higher risk of hyperuricemia that someone with a BMI of 23?

9.18 Challenger disaster, Part II. Exercise 9.16 introduced us to O-rings that were identified as a plausible explanation for the breakup of the Challenger space shuttle 73 seconds into takeoff in 1986. The investigation found that the ambient temperature at the time of the shuttle launch was closely related to the damage of O-rings, which are a critical component of the shuttle. See this earlier exercise if you would like to browse the original data.



(a) The data provided in the previous exercise are shown in the plot. The logistic model fit to these data may be written as

$$\log\left(\frac{\hat{p}}{1-\hat{p}}\right) = 11.6630 - 0.2162 \times Temperature$$

where $\hat{p}$ is the model-estimated probability that an O-ring will become damaged. Use the model to calculate the probability that an O-ring will become damaged at each of the following ambient temperatures: 51, 53, and 55 degrees Fahrenheit. The model-estimated probabilities for several additional ambient temperatures are provided below, where subscripts indicate the temperature:

$$\hat{p}_{57} = 0.341 \qquad \hat{p}_{59} = 0.251 \qquad \hat{p}_{61} = 0.179 \qquad \hat{p}_{63} = 0.124$$
$$\hat{p}_{65} = 0.084 \qquad \hat{p}_{67} = 0.056 \qquad \hat{p}_{69} = 0.037 \qquad \hat{p}_{71} = 0.024$$

(b) Add the model-estimated probabilities from part (a) on the plot, then connect these dots using a smooth curve to represent the model-estimated probabilities.

(c) Describe any concerns you may have regarding applying logistic regression in this application, and note any assumptions that are required to accept the model's validity.

9.8.3 Multiple logistic regression

9.19 Risk of fracture.

Osteoporosis a bone disease characterized by decreasing bone mineral density and bone mass and is associated with a higher risk of fractures (broken bones) after falls. The Global Longitudinal Study of

Osteoporosis in Women (GLOW) collected data on over 60,000 women over 55 years of age diagnosed with osteoporosis. This exercise uses data from the study provided in Hosmer, Lemeshow and Sturdivant,[28] which contains additional information about the study. Briefly, the study followed the participants during the study period, recording potential predictors of fracture at enrollment and the first occurrence of a fracture during the follow-up period. The GLOW data can be found in the R package APLORE3.

The data in this exercise contains information from 500 participants and includes the occurrence of a fracture and selected possible risk factors. This sample was drawn by Hosmer, Lemeshow and Sturdivant from the full dataset by oversampling participants with fractures and under-sampling those without fractures, since only approximately 4% of the participants experienced fractures.

Figure 9.47 shows a logistic regression model with response variable whether or not the participant experienced a fracture during the study and two predictor variables: an indicator of whether a prior fracture was present and age in years.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-4.214	0.848	-4.97	0.000
priorfracYes	0.839	0.234	3.58	0.000
age	0.041	0.012	3.38	0.001

Figure 9.47: Logistic Regression for fracture with predictors age and presence of prior fracture

(a) Use the model to estimate the odds ratio for the occurrence of a fracture among 60 year old women, comparing women with a prior fracture to those without a prior fracture.

(b) Estimate the odds ratio for the occurrence of a fracture among women who have not had a prior fracture, comparing 75 year old women to those who are 65 years old.

(c) Does the design of the study and this sample of 500 participants support the estimates of ORs in parts (a) and (b)? Justify your answer.

(d) Can the data from the study be used to estimate prevalence differences and ratios in parts (a) and (b)? Justify your answer.

9.20 HIV test status and TB treatment interruption.

Show that the conditions for a χ^2 test are met for the data displayed in Figure 9.10.

9.21 Female horseshoe crabs, color and satellites.

Show that the conditions for a χ^2 test are met for the data displayed in Figure 9.14.

9.22 Emergency room outcomes in Denmark, Part I.

An important problem in emergency medicine is the prioritization of high-risk patients. Traditional triage algorithms classify patients into categories based on vital signs (such as heart rate and level of consciousness) in addition to the patient's reason for seeking medical care: "red" (life-threatening), "orange" (seriously ill), "yellow" (ill), "green" (needs assessment), and "blue" (minor complaints). A study in Denmark[29] studied the association of triage score and other variables with 30 day mortality in a dataset of 12,661 individuals[30] treated in the Emergency Department (ED) of Nordsjælland University Hospital in Denmark.

The model in Figure 9.48 is the result of fitting a logistic regression with response variable 30-day mortality (0 = alive 30 days a after admission) and predictor triage score in a random sample of 1,000 cases from the 5,371 participants in the primary dataset used for initial model building. In this sample of 1,000, there were 62 deaths within 30 days from admission to the ED.

Individuals classified as category blue were not included in the study.

(a) What is the reference category in the regression?

[28] David W Hosmer Jr et al. *Applied logistic regression, 3rd ed.* John Wiley & Sons, 2013.

[29] Michael Kristensen et al. "Routine blood tests are associated with short term mortality and can improve emergency department triage: a cohort study of> 12,000 patients". In: *Scandinavian Journal of Trauma, Resuscitation and Emergency Medicine* 25.1 (2017), pp. 1–8.

[30] **Data available at DOI:10.5061/dryad.m2bq5**.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.551	0.416	-3.73	0.000
triageorange	-0.958	0.474	-2.02	0.043
triageyellow	-1.290	0.468	-2.76	0.006
triagegreen	-1.585	0.518	-3.06	0.002

Figure 9.48: Logistic regression with response 30 day mortality and predictor triage level, using a random sample of 1,000 cases from Danish ED study primary cohort.

(b) Is the pattern in the estimates of the coefficients consistent with traditional triage coding?

(c) Write the equation for the model.

(d) Does the intercept have a meaningful interpretation in this model? If so, what is its interpretation?

(e) What is the estimated OR and 95% confidence interval for 30 day mortality, comparing category "yellow" with "red"?

(f) What is the OR for 30 day mortality, comparing category "yellow" with "orange"?

9.23 Emergency room outcomes in Denmark, Part II.

Figure 9.49 is a contingency table showing the association between 30-day mortality and DEPT triage classification for the data used in Exercise 9.22.

	Died within 30 days		
	No	Yes	Sum
red	33	7	40
orange	258	21	279
yellow	394	23	417
green	253	11	264
Sum	938	62	1000

Figure 9.49: Contingency table of 30-day mortality by triage classification, Danish ED study, random sample of 1,000 participants

(a) Show that the table can be used to calculate the estimate of the intercept given in Figure 9.48.

(b) The data in the table can be used to estimate each of the coefficients in the logistic model. Show that it can be used to calculate the estimate of the coefficient for the triage category "green".

(c) Can the estimates of the standard errors in Figure 9.48 be calculated directly from the table?

9.24 Emergency room outcomes in Denmark, Part III.

The dataset used in Exercise 9.22 also contains the age and sex of the participants. Figure 9.50 shows the logistic model in which age in years and sex have been added to the traditional triage coding.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-5.560	0.880	-6.32	0.000
triageorange	-1.062	0.501	-2.12	0.034
triageyellow	-1.245	0.494	-2.52	0.012
triagegreen	-1.489	0.543	-2.74	0.006
age	0.058	0.010	5.89	0.000
sexmale	0.009	0.277	0.03	0.974

Figure 9.50: Logistic regression with response 30 day mortality and predictors triage level, age and sex, using a random sample of 1,000 cases from Danish ED study.

(a) Does the intercept in this model have a meaningful interpretation? If so, what is the interpretation?

(b) Is increasing age associated with an increase or decrease in the risk of 30 day mortality?

(c) The residual deviance for the models in Figures 9.50 and 9.48 are, respectively, 407.58 and 463.60. Conduct a test of the null hypothesis that the pair of variables age and sex do not add useful information to a model based on triage score alone.

(d) Based on the estimated model and your answers to the above, do you believe that both age and sex should be retained in the model? Explain your answer.

9.25 Interaction in logistic regression.

The interaction term in the model given in Equation 9.30 would not be retained in a model with BMI and sex, but it is instructive to explore the implications of an interaction when estimating ORs.

(a) Calculate the estimated OR for hyperuricemia for two males with BMI 33.2 vs 30.

(b) Repeat the calculation for two females.

(c) How do these estimates differ from the corresponding ORs when using the model without an interaction in Figure 9.12?

9.26 Color and width of female crabs.

Females with wider carapaces are known to attract more males. A logistic regression with carapace width as the only predictor confirms the association between the odds of one or more satellites and width – the estimated log(odds) are 0.497 with $p < 0.001$. When color is held constant, each centimeter of width increases the odds of having satellites by a factor of $e^{0.497} = 1.644$. How strong is the evidence that color is an important predictor in a model that adjusts for carapace width?

Figure 9.51 shows an estimated model with both width and color as predictors.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-12.715	2.762	-4.60	0.000
width	0.468	0.106	4.43	< 0.001
colorMedDark	1.106	0.592	1.87	0.062
colorMedLight	1.402	0.548	2.56	0.011
colorLight	1.330	0.853	1.56	0.119

Figure 9.51: Logistic regression with horseshoe crab data, response presence of male satellites, predictors `width` and `color`.

The residual deviances for the regression with just `width` and with `width` and `color` are, respectively, 194.45 and 187.46, respectively.

(a) Calculate the deviance statistic and its significance level for the nested model that includes just `width` compared to the larger model with both `width` and `color`.

(b) Calculate the AIC for the two models in part (a)

(c) What do you conclude?

Appendix A

End of chapter exercise solutions

1 Introduction to data

1.1 (a) Treatment: 10/43 = 0.23 → 23%.
(b) Control: 2/46 = 0.04 → 4%. (c) A higher percentage of patients in the treatment group were pain free 24 hours after receiving acupuncture. (d) It is possible that the observed difference between the two group percentages is due to chance.

1.3 (a) "Is there an association between air pollution exposure and preterm births?" (b) 143,196 births in Southern California between 1989 and 1993. (c) Measurements of carbon monoxide, nitrogen dioxide, ozone, and particulate matter less than $10\mu g/m^3$ (PM_{10}) collected at air-quality-monitoring stations as well as length of gestation. Continuous numerical variables.

1.5 (a) "Does explicitly telling children not to cheat affect their likelihood to cheat?". (b) 160 children between the ages of 5 and 15. (c) Four variables: (1) age (numerical, continuous), (2) sex (categorical), (3) whether they were an only child or not (categorical), (4) whether they cheated or not (categorical).

1.7 (a) Control: the group of 16 female birds that received no treatment. Treatment: the group of 16 female birds that were given supplementary diets.
(b) "Does egg coloration indicate the health of female collared flycatchers?"
(c) Darkness of blue color in female birds' eggs. Continuous numerical variable.

1.9 (a) Each row represents a participant.
(b) The response variable is colon cancer stage. The explanatory variables are the abundance levels of the five bacterial species.
(c) Colon cancer stage: ordinal categorical variable. Abundance levels of bacterial species: continuous numerical variable.

1.11 (a) The population of interest consists of babies born in Southern California. The sample consists of the 143,196 babies born between 1989 and 1993 in Southern California.
(b) Assuming that the sample is representative of the population of interest, the results of the study can be generalized to the population. The findings cannot be used to establish causal relationships because the study was an observational study, not an experiment.

1.13 (a) The population of interest consists of asthma patients who rely on medication for asthma treatment. The sample consists of the 600 asthma patients ages 18-69 who participated in the study.
(b) The sample may not be representative of the population because study participants were recruited, an example of a convenience sample. Thus, the results of the study may not be generalizable to the population. The findings can be used to establish causal relationships because the study is an experiment conducted with control, randomization, and a reasonably large sample size.

1.15 (a) Experiment.
(b) The experimental group consists of the chicks that received vitamin supplements. The control group consists of the chicks that did not receive vitamin supplements.
(c) Randomization ensures that there are not systematic differences between the control and treatment groups. Even if chicks may vary in ways that affect body mass and corticosterone levels, random allocation essentially evens out such differences, on average, between the two groups. This is essential for a causal interpretation of the results to be valid.

1.17 (a) Observational study.
(b) Answers may vary. One possible confounding variable is the wealth of a country. A wealthy country's citizens tend to have a higher life expectancy due to a higher quality of life, and the country tends to have a higher percentage of internet users because there is enough money for the required infrastructure and citizens can afford computers. Wealth of a country is associated with both estimated life expectancy and percentage of internet users. Omitting the confounder from the analysis distorts the relationship between the two variables, such that there may seem to be a direct relationship when there is none.

1.19 (a) Simple random sampling is reasonable if 500 students is a large enough sample size relative to the total student population of the university.
(b) Since student habits may vary by field of study, stratifying by field of study would be a reasonable decision.
(c) Students in the same class year may have more similar habits. Since clusters should be diverse with respect to the outcome of interest, this would not be a good approach.

1.21 (a) Non-responders may have a different response to this question, e.g. parents who returned the surveys likely don't have difficulty spending time with their children.
(b) It is unlikely that the women who were reached at the same address 3 years later are a random sample. These missing responders are probably renters (as opposed to homeowners) which means that they might be in a lower socio-economic class than the respondents.
(c) This is an observational study, not an experiment, so it is not advisable to draw conclusions about causal relationships. The relationship may be in the other direction; i.e., that these people go running precisely because they do not have joint problems. Additionally, the data are not even sufficient to provide evidence of an association between running and joint problems because data have only been collected from individuals who go running regularly. Instead, a sample of individuals should be collected that includes both people who do and do not regularly go running; the number of individuals in each group with joint problems can then be compared for evidence of an association.

1.23 The lead author's statements are not accurate because he or she drew conclusions about causation (that increased alcohol sales taxes lower rates of sexually transmitted infections) from an observational study. In addition, although the study observed that there was a decline in gonorrhea rate, the lead author generalized the observation to all sexually transmitted infections.

1.25 (a) Randomized controlled experiment. (b) Explanatory: treatment group (categorical, with 3 levels). Response variable: Psychological well-being. (c) No, because the participants were volunteers. (d) Yes, because it was an experiment. (e) The statement should say "evidence" instead of "proof".

1.27 (a) The two distributions have the same median since they have the same middle number when ordered from least to greatest. Distribution 2 has a higher IQR because its first and third quartiles are farther apart than in Distribution 2.
(b) Distribution 2 has a higher median since it has a higher middle number when ordered from least to greatest. Distribution 2 has a higher IQR because its first and third quartiles are farther apart than in Distribution 1.
(c) Distribution 2 has a higher median since all values in this distribution are higher than in Distribution 1. The two distributions have the same IQR since the distance between the first and third quartiles in each distribution is the same.
(d) Distribution 2 has a higher median since most values in this distribution are higher than those in Distribution 1. Distribution 2 has a higher IQR because its first and third quartiles are farther apart than those of Distribution 1.

1.29 (a) The distribution is bimodal, with peaks between 15-20 and 25-30. Values range from 0 to 65.
(b) The median AQI is about 30.
(c) I would expect the mean to be higher than the median, since there is some right skewing.

1.31 (a) The median is a much better measure of the typical amount earned by these 42 people. The mean is much higher than the income of 40 of the 42 people. This is because the mean is an arithmetic average and gets affected by the two extreme observations. The median does not get effected as much since it is robust to outliers. (b) The IQR is a much better measure of variability in the amounts earned by nearly all of the 42 people. The standard deviation gets affected greatly by the two high salaries, but the IQR is robust to these extreme observations.

1.33 (a) These data are categorical. They can be summarized numerically in either a frequency table or relative frequency table, and summarized graphically in a bar plot of either counts or proportions.
(b) The results of these studies cannot be generalized to the larger population. Individuals taking the survey represent a specific subset of the population that are conscious about dental health, since they are at the dentist's office for an appointment. Additionally, there may be response bias; even though the surveys are anonymous, it is likely that respondents will feel some pressure to give a "correct" answer in such a setting, and claim to floss more often than they actually do.

1.35 (a) Yes, there seems to be a positive association between lifespan and length of gestation. Generally, as gestation increases, so does life span.
(b) Positive association. Reversal of the plot axes does not change the nature of an association.

1.37 (a) 75% of the countries have an adolescent fertility rate less than or equal to 75.73 births per 1,000 adolescents.
(b) It is likely that the observations are missing due to the Iraq War and general instability in the region during this time period. It is unlikely that the five-number summary would have been affected very much, even if the values were extreme; the median and IQR are robust estimates, and the dataset is relatively large, with data from 188 other countries.
(c) The median and IQR decreases each year, with Q1 and Q3 also decreasing.

1.39 (a) 4,371/8,474 = 0.56 → 56%
(b) 110/190 = 0.58 → 58%
(c) 27/633 = 0.04 → 4%
(d) 53/3,110 = 0.02 → 2%
(e) Relative risk: $\frac{27/633}{53/3,110} = 2.50$. Yes, since the relative risk is greater than 1. A relative risk of 2.50 indicates that individuals with high trait anger are 2.5 times more likely to experience a CHD event than individuals with low trait anger.
(f) Side-by-side boxplots, since blood cholesterol level is a numerical variable and anger group is categorical.

2 Probability

2.1 (a) False. These are independent trials.

(b) False. There are red face cards.

(c) True. A card cannot be both a face card and an ace.

2.3 (a) $\frac{1}{4}$.

Solution 1: A colorblind male has genotype X^-Y. He must have inherited X^- from his mother (probability of $\frac{1}{2}$) and Y from his father (probability of $\frac{1}{2}$). Since these are two independent events, the probability of both occuring is $(\frac{1}{2})(\frac{1}{2}) = \frac{1}{4}$.

Solution 2: Determine the possibilities using a Punnett square. There are 4 equally likely possibilities, one of which is a colorblind male. Thus, the probability is $\frac{1}{4}$.

	X^+	Y
X^+	X^+X^+	X^+Y
X^-	X^+X^-	X^-Y

(b) True. An offspring of this couple cannot be both female and colorblind.

2.5 (a) 0.25. Let H represent the event of being a high school graduate and F represent the event of being a woman. $P(H) = P(H \text{ and } W) + P(H \text{ and } W^C) = P(H|W)P(W) + P(H|W^C)P(W^C) = (0.20)(0.50) + (0.30)(0.50) = 0.25$.

(b) 0.91.$(A^C) = P(A^C \text{ and } W) + P(A^C \text{ and } W^C) = (1 - 0.09) + (1 - 0.09) = 0.91$.

(c) 0.25. Let X represent the event of having at least a Bachelor's degree, where B represents the event of attaining at most a Bachelor's degree and G the event of attaining at most a graduate or professional degree. $P(X|W^C) = P(B|W^C) + P(G|W^C) = 0.16 + 0.09 = 0.25$.

(d) 0.26. $P(X|W) = P(B|W) + P(G|W) = 0.17 + 0.09 = 0.26$.

(e) 0.065. Let X_W be the event that a woman has at least a Bachelor's degree, and X_M be the event that a man has at least a Bachelor's degree. Assuming that the education levels of the husband and wife are independent, $P(X_W \text{ and } X_M) = P(X_W) \times P(X_M) = (0.25)(0.26) = 0.065$. This assumption is probably not reasonable, because people tend to marry someone with a comparable level of education.

2.7 (a) Let C represent the event that one urgent care center sees 300-449 patients in a week. Assuming that the number of patient visits are independent between urgent care centers in a given county for a given week, the probability that three random urgent care centers see 300-449 patients in a week is $[P(C)]^3 = (0.288)^3 = 0.024$. This assumption is not reasonable because a county is a small area with relatively few urgent care centers; if one urgent care center takes in more patients than usual during a given week, so might other urgent care centers in the same county (e.g., this could occur during flu season).

(b) 2.32×10^{-7}. Let D represent the event that one urgent care center sees 450 or more patients in a week. Assuming independence, the probability that 10 urgent care centers throughout a state all see 450 or more patients in a week is $[P(D)]^{10} = (0.217)^{10} = 2.32 \times 10^{-7}$. This assumption is reasonable because a state is a large area that contains many urgent care centers; the number of patients one urgent care center takes in is likely independent of the number of patients another urgent care center in the state takes in.

(c) No, it is not possible, because it is not reasonable to assume that the patient visits for a given week are independent of those for the following week.

2.9 (a) If the class is not graded on a curve, they are independent. If graded on a curve, then neither independent nor disjoint – unless the instructor will only give one A, which is a situation we will ignore in parts (b) and (c). (b) They are probably not independent: if you study together, your study habits would be related, which suggests your course performances are also related. (c) No. See the answer to part (a) when the course is not graded on a curve. More generally: if two things are unrelated (independent), then one occurring does not preclude the other from occurring.

2.11 (a) $0.60 + 0.20 - 0.18 = 0.62$

(b) $0.18/0.20 = 0.90$

(c) $0.11/0.33 = 0.33$

(d) No, because the answers to parts (c) and (d) are not equal. If global warming belief were independent of political party, then among liberal Democrats and conservative Republicans, there would be equal proportions of people who believe the earth is warming.

(e) $0.06/0.34 = 0.18$

2.13 (a) $375,264/436,968 = 0.859$

(b) $229,246/255,980 = 0.896$

(c) 0.896. This is equivalent to (b).

(d) $146,018/180,988 = 0.807$

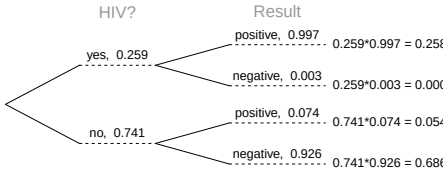
(e) $4,719/7,394 = 0.638$

(f) No, because the answers to (c) and (d) are not equal. If gender and seat belt usage were independent, then among males and females, there would be the same proportion of people who always wear seat belts.

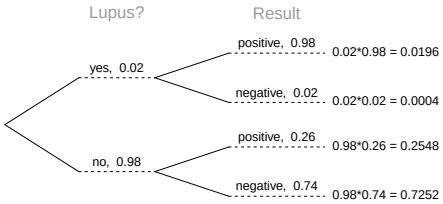
2.15 The PPV is 0.8248. The NPV is 0.9728.

$$P(D|T^+) = \frac{P(T^+|D)P(D)}{P(T^+|D)P(D)+P(T^+|D^C)P(D^C)} = \frac{(0.997)(0.0259)}{(0.997)(0.0259)+(1-0.926)(1-0.259)} = 0.8248.$$

$$P(D^C|T^-) = \frac{P(T^-|D^C)P(D^C)}{P(T^-|D^C)P(D^C)+P(T^-|D)P(D)} = \frac{(0.926)(1-0.259)}{(0.926)(1-0.259)+(1-0.997)(0.259)} = 0.9728.$$



2.17 0.0714. Even when a patient tests positive for lupus, there is only a 7.14% chance that he actually has lupus. House may be right.



2.19 (a) Let E represent the event of agreeing with the idea of evolution and D be the event of being a Democrat. From the problem statement, $P(E|D) = 0.67$. $P(E^C|D) = 1 - P(E|D) = 1 - 0.67 = 0.33$.

(b) Let I represent the event of being an independent. $P(E|I) = 0.65$, as stated in the problem.

(c) Let R represent the event of being a Republican. $P(E|R) = 1 - P(E^C|R) = 1 - 0.48 = 0.52$.

(d) 0.35. $P(R|E) = \frac{P(E \text{ and } R)}{P(E)} = \frac{P(R)P(E|R)}{P(E)} = \frac{(0.40)(0.52)}{0.60} = 0.35$.

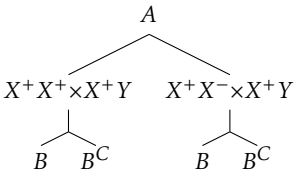
2.21 Mumps is the most likely disease state, since $P(B_3|A) = 0.563$, $P(B_1|A) = 0.023$, and $P(B_2|A) = .415$. $P(B_i|A) = \frac{P(A|B_i)P(B_i)}{P(A)}$. $P(A) = P(A \text{ and } B_1)+P(A \text{ and } B_2)+P(A \text{ and } B_3) = P(A|B_1)P(B_1)+P(A|B_2)P(B_2)+P(A|B_3)P(B_3)$.

2.23 (a) Let A be the event of knowing the answer and B be the event of answering it correctly. Assume that if a participant knows the correct answer, they answer correctly with probability 1: $P(B|A) = 1$. If they guess randomly, they have 1 out of m chances to answer correctly, thus $P(B|A^C) = 1/m$. $P(A|B) = \frac{1 \cdot p}{(1 \cdot p)+(\frac{1}{m} \cdot (1-p))} = \frac{p}{p+\frac{1-p}{m}}$.

(b) 0.524. Let A be the event of having an IQ over 150 and B be the event of receiving a score indicating an IQ over 150. From the problem statement, $P(B|A) = 1$ and $P(B|A^C) = 0.001$. $P(A^C|B) = \frac{0.001 \cdot (1-\frac{1}{1,100})}{(1 \cdot (\frac{1}{1,100}))+(0.001 \cdot (1-\frac{1}{1,100}))} = 0.524$.

2.25 (a) In descending order on the table, the PPV for each age group is 0.003, 0.064, 0.175, 0.270; the NPV for each age group is 0.999, 0.983, 0.948, 0.914.
(b) As prevalence of prostate cancer increases by age group, PPV also increases. However, with rising prevalence, NPV decreases.
(c) The probability that a man has prostate cancer, given a positive test, necessarily increases as the overall probability of having prostate cancer increases. If more men have the disease, the chance of a positive test result being a true positive increases (and the chances of the result being a false positive decreases). The decreasing NPV values follow similar logic: if more men have the disease, the chance of a negative test being a true negative decreases (and the chances of the result being a false negative increases).
(d) Lowering the cutoff for a positive test would result in more men testing positive, since men with PSA values 2.5 ng/ml to 4.1 ng/ml were not previously classified as testing positive. Since the sensitivity of a test is the proportion who test positive among those who have disease, and the number with disease does not change, the proportion will increase, except in the rare and unlikely situation where the additional positive tests are among only men without the disease.

2.27 (a) Frequency of X^+X^+: 0.863. Frequency of X^+X^-: 0.132. Frequency of X^-X^-: 0.005. Frequency of X^-Y: 0.07. Frequency of X^+Y: 0.93. From frequency of X^-X^-, frequency of X^- allele is $\sqrt{0.005} = 0.071$; thus, frequency of X^+ allele is $1 - 0.071 = 0.929$. Frequency of X^+Y is $1 - 0.093 = 0.07$.
(b) 0.033. Let A be the event that two parents are not colorblind, and B represent the event of having a colorblind child. On the tree, $\times$ represents a mating between two genotypes. $P(B|A) = [P(X^+X^+ \times X^+Y|A) \cdot P(B|X^+X^+ \times X^+Y)] + [P(X^+X^- \times X^+Y|A) \cdot P(B|X^+X^- \times X^+Y)] = (0.867)(0) + (0.133)(1/4) = 0.033$.



2.29 (a) Calculate $P(M \cap B)$, the probability a dog has a facial mask and a black coat. Note that the event M consists of having either a unilateral mask or a bilateral mask.

$$\begin{aligned} P(M \cap B) &= P(M_1 \cap B) + P(M_2 \cap B) \\ &= P(M_1|B)P(B) + P(M_2|B)P(B) \\ &= (0.25)(0.40) + (0.35)(0.40) \\ &= 0.24 \end{aligned}$$

The probability an Australian cattle dog has a facial mask and a black coat is 0.31.
(b) Calculate $P(M_2)$, the prevalence of bilateral masks. The event of having a bilateral mask can be partitioned into either having a bilateral mask and a red coat or having a bilateral mask and a black coat.

$$\begin{aligned} P(M_2) &= P(R \cap M_2) + P(B \cap M_2) \\ &= P(M_2|R)P(R) + P(M_2|B)P(B) \\ &= (0.10)(0.60) + (0.35)(0.40) \\ &= 0.20 \end{aligned}$$

The prevalence of bilateral masks in Australian cattle dogs is 0.20.
(c) Calculate $P(R|M_2)$, the probability of having a red coat given having a bilateral mask. Apply the definition of conditional probability.

$$\begin{aligned}P(R|M_2) &= \frac{P(R \cap M_2)}{P(M_2)} \\ &= \frac{P(M_2|R)P(R)}{P(M_2)} \\ &= \frac{(0.10)(0.60)}{0.20} \\ &= 0.30\end{aligned}$$

The probability of being a Red Heeler among Australian cattle dogs with bilateral facial masks is 0.30.

(d) The following new information has been introduced:

- $P(D_1|R, M_0) = P(D_1|R, M_1) = 0.15$, $P(D^C|R, M_0) = P(D^C|R, M_1) = 0.60$.
- $P(M_2 \cap D_2 \cap R) = 0.012$, $P(M_2 \cap D_1 \cap R) = 0.045$
- $P(M_2 \cap D_2 \cap B) = 0.012$, $P(M_2 \cap D_1 \cap B) = 0.045$
- $P(D_1|M_0, B) = P(D_1|M_1, B) = 0.05$, $P(D_2|M_0, B) = P(D_2|M_1, B) = 0.01$

i. Calculate $P(M_2 \cap D^C \cap R)$.

$$\begin{aligned}P(M_2 \cap D^C \cap R) &= P(D^C|M_2, R)P(M_2|R)P(R) \\ &= P(D^C|M_2, R)(0.10)(0.60)\end{aligned}$$

To calculate $P(D^C|M_2, R)$, first calculate $P(D_1|M_2, R)$ and $P(D_2|M_2, R)$ from the joint probabilities given in the problem, then apply the complement rule.

$$P(D_1|M_2, R) = \frac{P(M_2 \cap D_1 \cap R)}{P(M_2 \cap R)} = \frac{0.045}{(0.10)(0.60)} = 0.75$$

$$P(D_2|M_2, R) = \frac{P(M_2 \cap D_2 \cap R)}{P(M_2 \cap R)} = \frac{0.012}{(0.10)(0.60)} = 0.20$$

Back to the original question...

$$\begin{aligned}P(M_2 \cap D^C \cap R) &= P(D^C|M_2, R)P(M_2|R)P(R) \\ &= P(D^C|M_2, R)(0.10)(0.60) \\ &= [1 - (0.75 + 0.20)](0.10)(0.60) \\ &= (0.05)(0.10)(0.60) \\ &= 0.003\end{aligned}$$

The probability that an Australian cattle dog has a bilateral mask, no hearing deficits, and a red coat is 0.003.

ii. Calculate $P(D^C|M_2, B)$.

$$\begin{aligned}P(D^C|M_2, B) &= 1 - [P(D_1|M_2, B) + P(D_2|M_2, B)] \\ &= 1 - \left[\frac{P(D_1 \cap M_2 \cap B)}{P(M_2 \cap B)} + \frac{P(D_2 \cap M_2 \cap B)}{P(M_2 \cap B)}\right] \\ &= 1 - \left[\frac{0.045}{P(M_2|B)P(B)} + \frac{0.012}{P(M_2|B)P(B)}\right] \\ &= 1 - \left[\frac{0.045}{(0.35)(0.40)} + \frac{0.012}{(0.35)(0.40)}\right] \\ &= 0.593\end{aligned}$$

The proportion of bilaterally masked Blue Heelers without hearing deficits is 0.593.

iii. Calculate $P(D|R)$ and $P(D|B)$.

$$\begin{aligned}P(D|R) =& P(D \cap M_0|R) + P(D \cap M_1|R) + P(D \cap M_2|R)\\ =& [1 - P(D^C|R, M_0)](P(M_0|R) + [1 - P(D^C|R, M_1)](P(M_1|R) + [1 - P(D^C|M_2, R)](P(M_2|R)\\ =& (1 - 0.60)(0.50) + (1 - 0.60)(0.40) + (1 - 0.05)(0.10)\\ =& 0.455\end{aligned}$$

$$\begin{aligned}P(D|B) =& P(D \cap M_0|B) + P(D \cap M_1|B) + P(D \cap M_2|B)\\ =& [P(D_1|B, M_0) + P(D_2|B, M_0)](P(M_0|B) + [P(D_1|B, M_0) + P(D_2|B, M_0)](P(M_1|B)\\ &+ [1 - P(D^C|M_2, B)](P(M_2|B)\\ =& (0.05 + 0.01)(0.40) + (0.05 + 0.01)(0.25) + (1 - 0.593)(0.35)\\ =& 0.181\end{aligned}$$

The prevalence of deafness among Red Heelers is higher, at 0.455 versus 0.181 in Blue Heelers.

iv. Calculate $P(B|D^C)$.

$$\begin{aligned}P(B|D^C) =& \frac{P(B \cap D^C)}{P(D^C)}\\ =& \frac{P(D^C|B)P(B)}{P(D^C \cap B) + P(D^C \cap R)}\\ =& \frac{[1 - P(D|B)]P(B)}{[1 - P(D|B)]P(B) + [1 - P(D|R)]P(R)}\\ =& \frac{(1 - 0.181)(0.40)}{(1 - 0.181)(0.40) + (1 - 0.455)(0.60)}\\ =& 0.50\end{aligned}$$

The probability that a dog is a Blue Heeler given that it is known to have no hearing deficits is 0.50.

3 Distributions of random variables

3.1 (a) 13. (b) No, these 27 students are not a random sample from the university's student population. For example, it might be argued that the proportion of smokers among students who go to the gym at 9 am on a Saturday morning would be lower than the proportion of smokers in the university as a whole.

3.3 (a) The probability of drawing three hearts equals (13/52)(12/51)(11/50) = 0.0129, and the probability of drawing three black cards equals (26/52)(25/51)(24/50) = 0.1176; thus, the probability of any other draw is $1-0.0129-0.1176 = 0.8694$. $E(X) = 0.0129(50)+0.1176(25)+0.8694(0) = 3.589$. $Var(X) = 0.0129(50-3.589)^2 + 0.1176(25-3.589)^2 + 0.8694(0-3.589)^2 = 93.007$. $SD(X) = \sqrt{Var(X)} = 9.644$.

(b) Let Y represent the net profit/loss, where $Y = X-5$. $E(Y) = E(X-5) = E(X)-5 = -1.412$. Standard deviation does not change from a shift of the distribution; $SD(Y) = SD(X) = 9.644$.

(c) It is not advantageous to play, since the expected winnings are lower than \$5.

3.5 (a) 215 eggs. Let X represent the number of eggs laid by one gull. $E(X) = 0.25(1) + 0.40(2) + 0.30(3) + 0.05(4) = 2.15$. $E(100X) = 100E(X) = 215$.

(b) 85.29 eggs. $Var(X) = 0.25(1-2.15)^2+0.40(2-2.15)^2+0.30(3-2.15)^2+0.05(4-2.15)^2 = 0.7275$. $Var(100X) = 100^2 Var(X) = 7275 \rightarrow \sqrt{7275} = 85.29$.

3.7 (a) Binomial conditions are met: (1) Independent trials: In a random sample across the US, it is reasonable to assume that whether or not one 18-20 year old has consumed alcohol does not depend on whether or not another one has. (2) Fixed number of trials: $n = 10$. (3) Only two outcomes at each trial: Consumed or did not consume alcohol. (4) Probability of a success is the same for each trial: $p = 0.697$.
(b) Let X be the number of 18-20 year olds who have consumed alcohol; $X \sim \text{Bin}(10, 0.697)$. $P(X = 6) = 0.203$.
(c) Let Y be the number of 18-20 year olds who have not consumed alcohol; $Y \sim \text{Bin}(10, 1 - 0.697)$. $P(Y = 4) = P(X = 6) = 0.203$.
(d) $X \sim \text{Bin}(5, 0.697)$. $P(X \leq 2) = 0.167$.
(e) $X \sim \text{Bin}(5, 0.697)$. $P(X \geq 1) = 1 - P(X = 0) = 0.997$.

3.9 (a) $\mu 34.85$, $\sigma = 3.25$ (b) $Z = \frac{45 - 34.85}{3.25} = 3.12$. 45 is more than 3 standard deviations away from the mean, we can assume that it is an unusual observation. Therefore yes, we would be surprised. (c) Using the normal approximation, 0.0009. With 0.5 correction, 0.0015.

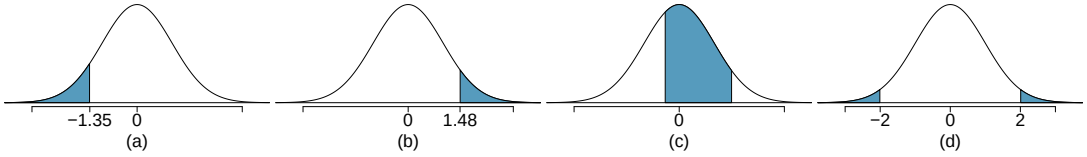
3.11 (a) Both O+ and O- individuals can donate blood to a Type O+ patient; $n = 15$, $p = 0.45$. $\mu = np = 6.75$. $\sigma = \sqrt{np(1-p)} = 1.93$.
(b) Only O- individuals can donate blood to a Type O- patient; $n = 15$, $p = 0.08$. $P(X \geq 3) = 0.113$.

3.13 0.132. Let X be the number of IV drug users who contract Hepatitis C within a month; $X \sim \text{Bin}(5, 0.30)$, $P(X = 3) = 0.132$.

3.15 (a) Let X represent the number of infected stocks in the sample; $X \sim \text{Bin}(250, 0.30)$. $P(X = 60) = 0.006$.
(b) $P(X \leq 60) = 0.021$.
(c) $P(X \geq 80) = 0.735$.
(D) 40% of 250 is 100. $P(X \leq 100) = 0.997$. Yes, this seems reasonable; it is essentially guaranteed that within a sample of 250, no more than 40% will be infected.

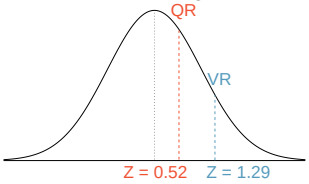
3.17 (a) (200)(0.12) = 24 cases of hyponatremia are expected during the marathon.
(b) Let X represent the number of cases of hyponatremia during the marathon. $P(X > 30) = 0.082$.

3.19 (a) 8.85%. (b) 6.94%. (c) 58.86%. (d) 4.56%.



3.21 (a) 0.005. (b) 0.911. (c) 0.954. (d) 1.036. (e) -0.842

3.23 (a) Verbal: $N(\mu = 151, \sigma = 7)$, Quant: $N(\mu = 153, \sigma = 7.67)$. $Z_{VR} = 1.29$, $Z_{QR} = 0.52$. She did better on the Verbal Reasoning section since her Z-score on that section was higher.



(b) $Perc_{VR} = 0.9007 \approx 90\%$, $Perc_{QR} = 0.6990 \approx 70\%$. 100% – 90% = 10% did better than her on VR, and 100% – 70% = 30% did better than her on QR.
(c) 159. (d) 147.

3.25 (a) 0.115. (b) The coldest 10% of days are colder than 70.59°F.

3.27 (a) 0.023. (b) 72.66 mg/dL.

3.29 (a) 82.4%. (b) About 38 years of age.

3.31 (a) $n = 50$, and $p = 0.70$. $\mu = np = 35$. $\sigma = \sqrt{np(1-p)} = 3.24$.
(b) Both np and $n(1-p)$ are greater than 10. Thus, it is valid to approximate the distribution as $X \sim N(35, 3.24)$, where X is the number of 18-20 year olds who have consumed alcohol. $P(X \geq 45) = 0.001$.

3.33 Let X represent the number of students who accept the offer; $X \sim \text{Bin}(2500, 0.70)$. This distribution can be approximated by a $N(1750, 22.91)$. The approximate probability that the school does not have enough dorm room spots equals $P(X \geq 1,786) = 0.06$.

3.35 The data appear to follow a normal distribution, since the points closely follow the line on the normal probability plot. There are some small deviations, but this is to be expected for such a small sample size.

3.37 (a) $P(X = 2) = \frac{\exp^{-2}(2^2)}{2!} = 0.271$. (b) $P(X \leq 2) = P(X = 0) + P(X = 1) + P(X = 2) = 0.677$. (c) $P(X \geq 3) = 1 - P(X \leq 2) = 0.323$.

3.39 (a) $\mu = \lambda = 75$, $\sigma = \sqrt{\lambda} = 8.66$. (b) $Z = -1.73$. Since 60 is within 2 standard deviations of the mean, it would not generally be considered unusual. Note that we often use this rule of thumb even when the normal model does not apply. (c) Using Poisson with $\lambda = 75$: 0.0402.

3.41 (a) The expected number of cases of osteosarcoma in NYC in a given year is 11.2. (b) Let X represent the number of osteosarcoma cases diagnosed. The probability that 15 or more cases will be diagnosed in a given year is the quantity $P(X \geq 15) = 1 - P(X < 15) = 1 - P(X \leq 14) = 0.161$. (c) First, calculate λ_B given that $n = 450,000$ for Brooklyn: 3.6. The probability of observing 10 or more cases in Brooklyn in a given year is the quantity $P(X_B \geq 10) = 1 - P(X_B < 10) = 1 - P(X_B \leq 9) = 0.004$. (d) No, he is not correct. The probability calculated in c) deals only with Brooklyn: the probability that there are 10 or more cases in Brooklyn for a single year. It does not say anything about cases in other boroughs. If we assume independence between boroughs, the probability that the official is referring to is:

$P(X = 0$ in other boroughs$) \times P(X \geq 10$ in Brooklyn$)$.

There is no reason to expect that $P(X = 0$ in other boroughs$)$ should equal 1, so this probability is different from the one in part c). (e) o, this probability is not equal to the probability calculated in part c). Over five years, there are five opportunities for the event of 10 or more cases in Brooklyn in a single year to occur. Let Y represent the event that in a single year, 10 or more cases of osteosarcoma are observed in Brooklyn. If we assume independence between years, then Y follows a binomial distribution with $n = 5$ and p of success as caculated in part c); $P(Y = 1) = 0.020$.

3.43 (a) λ for a population of 2,000,000 male births is 400. The probability of at most 380 newborn males with hemophilia is $P(X \leq 380)$, where $X \sim \text{Pois}(400)$: 0.165.
(b) $P(X \geq 450) = 0.0075$.
(c) The number of male births is (1/2)(1,500,000) = 750,000. The rate λ for one year is 150. Over 5 years, the rate λ is 750. The expected number of hemophilia births over 5 years is 750 and the standard deviation is $\sqrt{750} = 27.39$.

3.45 (a) On average, 2 women would need to be sampled in order to select a married woman ($\mu = 1/p = 2.123$), with standard deviation 1.544 ($\sigma = \sqrt{\frac{(1-p)}{p^2}}$).
(b) $\mu = 3.33$. $\sigma = 2.79$.
(c) Decreasing the probability increases both the mean and the standard deviation.

3.47 (a) Let X represent the number of stocks that must be sampled to find an infected stock; $X \sim \text{Geom}(0.30)$. $P(X \leq 5) = 0.832$.
(b) $P(X \leq 6) = 0.882$.
(c) $P(X \geq 3) = 1 - P(X \leq 2) = 0.49$.

3.49 (a) $0.875^2 \times 0.125 = 0.096$. (b) $\mu = 8$, $\sigma = 7.48$.

3.51 (a) 0.0804. (b) 0.0322. (c) 0.0193.

3.53 (a) 0.102, geometric with $p = 1994/14,604 = 0.137$.

(b) 0.854, binomial with $n = 10$, $p = 0.137$.

(c) 0.109, binomial with $n = 10$, $p = 0.137$.

(d) The mean and standard deviation of a negative binomial random variable with $r = 4$ and $p = 0.137$ are 29.30 and 13.61, respectively.

3.55 (a) $\mu = 2.05$; $\sigma^2 = 1.77$.

(b) Let X represent the number of soapy-taste detectors; $X \sim \text{HGeom}(1994, 14604 - 1994, 15)$. $P(X = 4) = 0.09435$.

(c) $P(X \leq 2) = 0.663$.

(d) 0.09437, from the binomial distribution. With a large sample size, sampling with replacement is highly unlikely to result in any particular individual being sampled again. In this case, the hypergeometric and binomial distributions will produce equal probabilities.

3.57 (a) The marginal distributions for X is obtained by summing across the two rows, and for Y by summing the columns. The marginal probabilities for $X = 0$ and $X = 1$ are 0.60 and 0.40, and for $Y = -1$ and $Y = 1$ are both 0.50; i.e., $p_X(0) = 0.60$, $p_X(1) = 0.40$, $p_Y(-1) = p_Y(1) = 0.50$ (b) The mean and variance of X are calculated using the formulas in Section 3.1.2 and 3.1.3 and are

$$\mu_X = (0)(0.60) + (1)(0.40) = 0.40$$
$$\sigma_X^2 = (0 - 0.40)^2(0.60) + (1 - 0.40)^2(0.40) = 0.24$$

The standard deviation of X is $\sqrt{0.24} = 0.49$. (c) The two standardized values of X are obtained by subtracting the mean of X from each value and dividing by the standard deviation. The two standardized values are -0.82 and 1.23. (d) The correlation between X and Y adds the 4 products of the standardized values, weighted by the values in the joint distribution:

$$\rho_{X,Y} = (-0.82)(-1)(0.20) + (-0.83)(1)(0.40) + (1.23)(-1)(0.30) + (1.23)(1)(0.10) = -.41$$

(e) No. The correlation between X and Y is not zero.

3.59 (a) Sum over the margins to calculate the marginal distributions.

$$p_Y(-1) = 0.25 \quad p_Y(0) = 0.20 \quad p_Y(1) = 0.55$$

$$p_X(-1) = 0.45 \quad p_X(0) = 0.20 \quad p_X(1) = 0.35$$

(b) The expected value of X is calculated as follows:

$$E(X) = \sum_i x_i P(X = x_i) = (-1)(0.45) + (0)(0.20) + (1)(0.35) = -0.10$$

(c) The variance of Y is calculated by first calculating $E(Y)$, then using that in the formula for a variance of a random variable.

$$E(Y) = \sum_i y_i P(Y = y_i) = (-1)(0.25) + (0)(0.20) + (1)(0.55) = 0.30$$

$$\text{Var}(Y) = \sum_i (y_i - E(Y))^2 P(Y = y_i) = (-1 - 0.30)^2(0.25) + (0 - 0.30)^2(0.20) + (1 - 0.30)^2(0.55) = 0.71$$

(d) $P(X = -1|Y = 0) = 0/0.20 = 0$; $P(X = 0|Y = 0) = 0.10/0.20 = 0.50$; $P(X = 1|Y = 0) = 0.10/0.20 = 0.5$.

3.61 (a) No. The new marginal distributions for the costs for the two members of the couple are shown in the following table. The values and the marginal distribution for the partner's cost do not change, so the expected value and standard deviation will not change. The previous values for the mean and standard deviation were \$980 and \$9.80.

	Partner Costs, Y		
Employee costs, X	**\$968**	**\$988**	**Marg. Dist.,** X
\$968	0.18	0.12	0.30
\$1,008	0.15	0.25	0.40
\$1,028	0.07	0.23	0.30
Marg. Dist., Y	0.40	0.60	1.00

(b) The expected value and standard deviation of the employee's costs are calculated as in Example 3.6, but using the new marginal distribution. The new values for the mean and standard deviation are \$1,002 and \$23.75. (c) The expected total cost is \$1,002 + \$980 = \$1,982. (d) The calculation correlation depends on the standardized costs for each member of the couple and the joint probabilities. The new standardized values for the employee costs are -1.43, 0.25, and 1.09; the corresponding values for the partner are -1.22 and 0.82. The correlation is the weighted sum of the 6 products, weighted by the joint probabilities: $\rho_{X,Y} = 0.29$. (e) The new variance for the total cost will be $(23.80)^2 + (9.80)^2 + (2)(23.8)(9.80)(0.29) = 796.00$ The new standard deviation is $\sqrt{796.00} = \$28.21$.

4 Foundations for inference

4.1 (a) $\overline{x} = 0.6052$.
(b) $s = 0.0131$.
(c) $Z_{0.63} = \frac{0.63-0.6052}{0.0131} = 1.893$. No, this level of BGC is within 2 SD of the mean.
(d) The standard error of the sample mean is given by $\frac{s}{\sqrt{n}} = \frac{0.0131}{\sqrt{70}} = 0.00157$.

4.3 (a) This is the sampling distribution of the sample mean.
(b) The sampling distribution will be normal and symmetric, centered around the theoretical population mean μ of the number of eggs laid by this hen species during a breeding period.
(c) The variability of the distribution is the standard error of the sample mean: $\frac{s}{\sqrt{n}} = \frac{18.2}{\sqrt{45}} = 2.71$.
(d) The variability of the new distribution will be greater than the variability of the original distribution. Conceptually, a smaller sample is less informative, which leads to a more uncertain estimate. This can be shown concretely with a calculation: $\frac{18.2}{\sqrt{10}} = 5.76$ is larger than 2.71.

4.5 (a) We are 95% confident that the mean number of hours that U.S. residents have to relax or pursue activities that they enjoy is between 3.53 and 3.83 hours.
(b) A larger margin of error with the same sample occurs with a higher confidence level (i.e., larger critical value).
(c) The margin of error of the new 95% confidence interval will be smaller, since a larger sample size results in a smaller standard error. (d) A 90% confidence interval will be smaller than the original 95% interval, since the critical value is smaller and results in a smaller margin of error. The interval will provide a more precise estimate, but have an associated lower confidence of capturing μ.

4.7 (a) False. Provided the data distribution is not very strongly skewed ($n = 64$ in this sample, so we can be slightly lenient with the skew), the distribution of the sample mean will be nearly normal, allowing for the normal approximation.
(b) False. Inference is made on the population parameter, not the point estimate. The point estimate is always in the confidence interval.
(c) True.
(d) False. The confidence interval is not about a sample mean.
(e) False. A wider interval is required to be more confident about capturing the parameter.
(f) True. The margin of error is half the width of the interval, and the sample mean is the midpoint of the interval.
(g) False. To halve the margin of error requires sampling $2^2 = 4$ times the number of people in the initial sample.

4.9 (a) i. False. There is a 5% chance that any 95% confidence interval does not contain the true population mean days out of the past 30 days that U.S. adults experienced poor mental health. ii. False. The population parameter μ is either inside or outside the interval; there is no probability associated with whether the fixed value μ is in a certain calculated interval. The randomness is associated with the interval (and the method for calculating it), not the parameter μ. Thus, it would not be reasonable to say there is a 95% chance that the particular interval (3.40, 4.24) contains μ; this interpretation is coherent with the statement in part iii. of this question. iii. True. This is the definition of what it means to be 95% confident. iv. True. The interval corresponds to a two-sided test, with $H_0 : \mu = 4.5$ days and $H_A : \mu \neq 4.5$ days and $\alpha = 1 - 0.95 = 0.05$. Since μ_0 of 4.5 days is outside the interval, the sample provides sufficient evidence to reject the null hypothesis and accept the alternative hypothesis. v. False. We can only be confident that 95% of the time, the entire interval calculated contains μ. It is not possible to make this statement about $\bar{x}$ or any other point within the interval. vi. False. The confidence interval is a statement about the population parameter μ, the mean days out of the past 30 days that all US adults experienced poor mental health. The sample mean $\bar{x}$ is a known quantity.
(b) The 90% confidence interval will be smaller than the 95% confidence interval. If we are less confident that an interval contains μ, this implies that the interval is less wide; if we are more confident, the interval is wider. Think about a theoretical "100%" confidence interval—to be 100% confident of capturing μ, then the range must be all possible numbers that μ could be. (c) (3.47, 4.17) days

4.11 (a) The null hypothesis is that New Yorkers sleep an average of 8 hours of night ($H_0 : \mu = 8$ hours). The alternative hypothesis is that New Yorkers sleep less than 8 hours a night on average ($H_A : \mu < 8$ hours).
(b) The null hypothesis is employees spend on average 15 minutes on non-business activities in a day ($H_0 : \mu = 15$ minutes). The alternative hypothesis is that employees spend on average more than 15 minutes on non-business activities in a day ($H_A : \mu > 15$ minutes).

4.13 Hypotheses are always made about the population parameter μ, not the sample mean $\bar{x}$. The correct value of μ_0 is 10 hours, as based on the previous evidence; both hypotheses should include μ_0. The correct hypotheses are $H_0 : \mu = 10$ hours and $H_A : \mu > 10$ hours.

4.15 (a) This claim is not supported by the confidence interval. 3 hours corresponds to a time of 180 minutes; there is evidence that the average waiting time is lower than 3 hours.
(b) 2.2 hours corresponds to 132 minutes, which is within the interval. It is plausible that μ is 132 minutes, since we are 95% confident that the interval (128 minutes, 147 minutes) contains the average wait time.
(c) Yes, the claim would be supported based on a 99% interval, since the 99% interval is wider than the 95% interval.

4.17 $H_0 : \mu = 130$ grams, $H_A : \mu \neq 130$ grams. Test the hypothesis by calculating the test statistic: $t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{130-134}{17/sqrt35} = 1.39$. This results in a p-value of 0.17. There is insufficient evidence to reject the null hypothesis. There is no evidence that the nutrition label does not provide an accurate measure of calories.

4.19 (a) The 95% confidence interval is $3,150 \pm (1.96 \times 250/\sqrt{50}) = (3080.7, 3219.3)$ grams.
(b) She will conduct a test of the null against the two-sided alternative $H_A : \mu \neq 3250$ grams. Calculate the test statistic: $t = \frac{\bar{x}-\mu_0}{s/\sqrt{n}} = \frac{3150-3250}{250/\sqrt{50}} = -2.83$. The p-value is 0.007. There is sufficient evidence to reject the null hypothesis and conclude that the mean birthweight of babies from inner-city teaching hospitals is lower than 3,260 grams.

4.21 (a) H_0: Anti-depressants do not help symptoms of fibromyalgia. H_A: Anti- depressants do treat symptoms of fibromyalgia. (b) Concluding that anti-depressants work for the treatment of fibromyalgia symptoms when they actually do not. (c) Concluding that anti-depressants do not work for the treatment of fibromyalgia symptoms when they actually do. (d) If she makes a Type 1 error, she will continue taking medication that does not actually treat her disorder. If she makes a Type 2 error, she will stop taking medication that could treat her disorder.

4.23 (a) The standard error is larger under scenario I; standard error is larger for smaller values of n.
(b) The margin of error is larger under scenario I; to be more confidence of capturing the population parameter requires a larger confidence interval.
(c) The p-value from a Z-statistic only depends on the value of the Z-statistic; the value is equal under the scenarios.
(d) The probability of making a Type II error and falsely rejecting the alternative is higher under scenario I; it is easier to reject the alternative with a high α.

5 Inference for numerical data

5.1 (a) $df = 6 - 1 = 5$, $t^\star_5 = 2.02$ (column with two tails of 0.10, row with $df = 5$). (b) $df = 21 - 1 = 20$, $t^\star_{20} = 2.53$ (column with two tails of 0.02, row with $df = 20$). (c) $df = 28$, $t^\star_{28} = 2.05$. (d) $df = 11$, $t^\star_{11} = 3.11$.

5.3 On a z-distribution, the cutoff value for the upper 5% of values is 1.96. A t-distribution has wider tails than a normal distribution but approaches the shape of a standard normal as degrees of freedom increases. Thus, 1.98 corresponds to the cutoff for a t-distribution with 100 degrees of freedom, 2.01 the cutoff for 50 degrees of freedom, and 2.23 the cutoff for 10 degrees of freedom.

5.5 The mean is the midpoint: $\bar{x} = 20$. Identify the margin of error: $ME = 1.015$, then use $t^\star_{35} = 2.03$ and $SE = s/\sqrt{n}$ in the formula for margin of error to identify $s = 3$.

5.7 (a) H_0: $\mu = 8$ (New Yorkers sleep 8 hrs per night on average.) H_A: $\mu \neq 8$ (New Yorkers sleep less or more than 8 hrs per night on average.) (b) Independence: The sample is random. The min/max suggest there are no concerning outliers. $T = -1.75$. $df = 25 - 1 = 24$. (c) p-value = 0.093. If in fact the true population mean of the amount New Yorkers sleep per night was 8 hours, the probability of getting a random sample of 25 New Yorkers where the average amount of sleep is 7.73 hours per night or less (or 8.27 hours or more) is 0.093. (d) Since p-value > 0.05, do not reject H_0. The data do not provide strong evidence that New Yorkers sleep more or less than 8 hours per night on average. (e) No, since the p-value is smaller than $1 - 0.90 = 0.10$.

5.9 T is either -2.09 or 2.09. Then $\bar{x}$ is one of the following:

$$-2.09 = \frac{\bar{x} - 60}{\frac{8}{\sqrt{20}}} \rightarrow \bar{x} = 56.26$$

$$2.09 = \frac{\bar{x} - 60}{\frac{8}{\sqrt{20}}} \rightarrow \bar{x} = 63.74$$

5.11 (a) We will conduct a 1-sample t-test. H_0: $\mu = 5$. H_A: $\mu \neq 5$. We'll use $\alpha = 0.05$. This is a random sample, so the observations are independent. To proceed, we assume the distribution of years of piano lessons is approximately normal. $SE = 2.2/\sqrt{20} = 0.4919$. The test statistic is $T = (4.6 - 5)/SE = -0.81$. $df = 20 - 1 = 19$. The one-tail area is about 0.21, so the p-value is about 0.42, which is bigger than $\alpha = 0.05$ and we do not reject H_0. That is, we do not have sufficiently strong evidence to reject the notion that the average is 5 years. (b) Using $SE = 0.4919$ and $t^\star_{df=19} = 2.093$, the confidence interval is (3.57, 5.63). We are 95% confident that the average number of years a child takes piano lessons in this city is 3.57 to 5.63 years. (c) They agree, since we did not reject the null hypothesis and the null value of 5 was in the t-interval.

5.13 If the sample is large, then the margin of error will be about $1.96 \times 100/\sqrt{n}$. We want this value to be less than 10, which leads to $n \geq 384.16$, meaning we need a sample size of at least 385 (round up for sample size calculations!).

5.15 (a) Since it's the same students at the beginning and the end of the semester, there is a pairing between the datasets; for a given student their beginning and end of semester grades are dependent. (b) Since the subjects were sampled randomly, each observation in the men's group does not have a special correspondence with exactly one observation in the other (women's) group. (c) Since it's the same subjects at the beginning and the end of the study, there is a pairing between the datasets; for a subject their beginning and end of semester artery thickness are dependent. (d) Since it's the same subjects at the beginning and the end of the study, there is a pairing between the datasets; for a subject their beginning and end of semester weights are dependent.

5.17 (a) For each observation in one data set, there is exactly one specially corresponding observation in the other data set for the same geographic location. The data are paired. (b) $H_0 : \mu_{\text{diff}} = 0$ (There is no difference in average number of days exceeding 90°F in 1948 and 2018 for NOAA stations.) $H_A : \mu_{\text{diff}} \neq 0$ (There is a difference.) (c) Locations were randomly sampled, so independence is reasonable. The sample size is at least 30, so we're just looking for particularly extreme outliers: none are present (the observation off left in the histogram would be considered a clear outlier, but not a particularly extreme one). Therefore, the conditions are satisfied. (d) $SE = 17.2/\sqrt{197} = 1.23$. $T = \frac{2.9-0}{1.23} = 2.36$ with degrees of freedom $df = 197 - 1 = 196$. This leads to a one-tail area of 0.0096 and a p-value of about 0.019. (e) Since the p-value is less than 0.05, we reject H_0. The data provide strong evidence that NOAA stations observed more 90°F days in 2018 than in 1948. (f) Type 1 Error, since we may have incorrectly rejected H_0. This error would mean that NOAA stations did not actually observe a decrease, but the sample we took just so happened to make it appear that this was the case. (g) No, since we rejected H_0, which had a null value of 0.

5.19 (a) $SE = 1.23$ and $t^\star = 1.65$. $2.9 \pm 1.65 \times 1.23 \rightarrow (0.87, 4.93)$.
(b) We are 90% confident that there was an increase of 0.87 to 4.93 in the average number of days that hit 90°F in 2018 relative to 1948 for NOAA stations.
(c) Yes, since the interval lies entirely above 0.

5.21 (a) Each of the 36 mothers is related to exactly one of the 36 fathers (and vice-versa), so there is a special correspondence between the mothers and fathers. (b) $H_0 : \mu_{diff} = 0$. $H_A : \mu_{diff} \neq 0$. Independence: random sample from less than 10% of population. Sample size of at least 30. The skew of the differences is, at worst, slight. $Z = 2.72 \rightarrow$ p-value = 0.0066. Since p-value < 0.05, reject H_0. The data provide strong evidence that the average IQ scores of mothers and fathers of gifted children are different, and the data indicate that mothers' scores are higher than fathers' scores for the parents of gifted children.

5.23 (a) Since $p < 0.05$, there is statistically significant evidence that the population difference in BGC is not 0. Since the observed mean BGC is higher in the food supplemented group, these data suggest that food supplemented birds have higher BGC on average than birds that are not food supplemented. (b) The 95% confidence interval is $\overline{d} \pm t^\star \frac{s_d}{\sqrt{n}}$. Since the mean of the differences is equal to the difference of the means, $\overline{d} = 1.70 - 0.586 = 1.114$. The test statistic is $t = \frac{\overline{d}}{s_d/\sqrt{n}}$, so the standard error ($s_d/\sqrt{n}$) can be solved for: $s_d/\sqrt{n} = \overline{d}/t = 1.114/2.64 = 0.422$. The critical t-value for a 95% confidence interval on a t-distribution with 16–1 = 15 degrees of freedom is 2.13. Thus, the 95% confidence interval is $1.114 \pm (2.13 \times 0.422) \rightarrow (0.215, 2.01)$ grams. With 95% confidence, the interval (0.215, 2.01) grams contains the population mean difference in egg mass between food supplemented birds and non supplemented birds.

5.25 (a) These data are paired. For example, the Friday the 13th in say, September 1991, would probably be more similar to the Friday the 6th in September 1991 than to Friday the 6th in another month or year.
(b) Let $\mu_{diff} = \mu_{sixth} - \mu_{thirteenth}$. $H_0 : \mu_{diff} = 0$. $H_A : \mu_{diff} \neq 0$.
(c) Independence: The months selected are not random. However, if we think these dates are roughly equivalent to a simple random sample of all such Friday 6th/13th date pairs, then independence is reasonable. To proceed, we must make this strong assumption, though we should note this assumption in any reported results. Normality: With fewer than 10 observations, we would need to see clear outliers to be concerned. There is a borderline outlier on the right of the histogram of the differences, so we would want to report this in formal analysis results.
(d) $T = 4.93$ for $df = 10 - 1 = 9 \rightarrow$ p-value = 0.001.
(e) Since p-value < 0.05, reject H_0. The data provide strong evidence that the average number of cars at the intersection is higher on Friday the 6^{th} than on Friday the 13^{th}. (We should exercise caution about generalizing the interpretation to all intersections or roads.)
(f) If the average number of cars passing the intersection actually was the same on Friday the 6^{th} and 13^{th}, then the probability that we would observe a test statistic so far from zero is less than 0.01.
(g) We might have made a Type 1 Error, i.e. incorrectly rejected the null hypothesis.

5.27 (a) $H_0 : \mu_{diff} = 0$. $H_A : \mu_{diff} \neq 0$. $T = -2.71$. $df = 5$. p-value = 0.042. Since p-value < 0.05, reject H_0. The data provide strong evidence that the average number of traffic accident related emergency room admissions are different between Friday the 6^{th} and Friday the 13^{th}. Furthermore, the data indicate that the direction of that difference is that accidents are lower on Friday the 6^{th} relative to Friday the 13^{th}.
(b) (-6.49, -0.17).
(c) This is an observational study, not an experiment, so we cannot so easily infer a causal intervention implied by this statement. It is true that there is a difference. However, for example, this does not mean that a responsible adult going out on Friday the 13^{th} has a higher chance of harm than on any other night.

5.29 (a) Chicken fed linseed weighed an average of 218.75 grams while those fed horsebean weighed an average of 160.20 grams. Both distributions are relatively symmetric with no apparent outliers. There is more variability in the weights of chicken fed linseed. (b) $H_0 : \mu_{ls} = \mu_{hb}$. $H_A : \mu_{ls} \neq \mu_{hb}$. We leave the conditions to you to consider. $T = 3.02$, $df = min(11, 9) = 9 \rightarrow 0.01 <$ p-value < 0.02. Since p-value < 0.05, reject H_0. The data provide strong evidence that there is a significant difference between the average weights of chickens that were fed linseed and horsebean. (c) Type 1 Error, since we rejected H_0. (d) Yes, since p-value > 0.01, we would have failed to reject H_0.

5.31 $H_0 : \mu_C = \mu_S$. $H_A : \mu_C \neq \mu_S$. $T = 3.27$, $df = 11 \rightarrow$ p-value < 0.01. Since p-value < 0.05, reject H_0. The data provide strong evidence that the average weight of chickens that were fed casein is different than the average weight of chickens that were fed soybean (with weights from casein being higher). Since this is a randomized experiment, the observed difference can be attributed to the diet.

5.33 $H_0 : \mu_T = \mu_C$. $H_A : \mu_T \neq \mu_C$. $T = 2.24$, $df = 21 \rightarrow 0.02 <$ p-value < 0.05. Since p-value < 0.05, reject H_0. The data provide strong evidence that the average food consumption by the patients in the treatment and control groups are different. Furthermore, the data indicate patients in the distracted eating (treatment) group consume more food than patients in the control group.

5.35 Let $\mu_{diff} = \mu_{pre} - \mu_{post}$. $H_0 : \mu_{diff} = 0$: Treatment has no effect. $H_A : \mu_{diff} \neq 0$: Treatment has an effect on P.D.T. scores, either positive or negative. Conditions: The subjects are randomly assigned to treatments, so independence within and between groups is satisfied. All three sample sizes are smaller than 30, so we look for clear outliers. There is a borderline outlier in the first treatment group. Since it is borderline, we will proceed, but we should report this caveat with any results. For all three groups: $df = 13$. $T_1 = 1.89 \rightarrow$ p-value $= 0.081$, $T_2 = 1.35 \rightarrow$ p-value $= 0.200$), $T_3 = -1.40 \rightarrow$ (p-value $= 0.185$). We do not reject the null hypothesis for any of these groups. As earlier noted, there is some uncertainty about if the method applied is reasonable for the first group.

5.37 Difference we care about: 40. Single tail of 90%: $1.28 \times SE$. Rejection region bounds: $\pm 1.96 \times SE$ (if 5% significance level). Setting $3.24 \times SE = 40$, subbing in $SE = \sqrt{\frac{94^2}{n} + \frac{94^2}{n}}$, and solving for the sample size n gives 116 plots of land for each fertilizer.

5.39 H_0: $\mu_1 = \mu_2 = \cdots = \mu_6$. H_A: The average weight varies across some (or all) groups. Independence: Chicks are randomly assigned to feed types (presumably kept separate from one another), therefore independence of observations is reasonable. Approx. normal: the distributions of weights within each feed type appear to be fairly symmetric. Constant variance: Based on the side-by-side box plots, the constant variance assumption appears to be reasonable. There are differences in the actual computed standard deviations, but these might be due to chance as these are quite small samples. $F_{5,65} = 15.36$ and the p-value is approximately 0. With such a small p-value, we reject H_0. The data provide convincing evidence that the average weight of chicks varies across some (or all) feed supplement groups.

5.41 (a) H_0: The population mean of MET for each group is equal to the others. H_A: At least one pair of means is different. (b) Independence: We don't have any information on how the data were collected, so we cannot assess independence. To proceed, we must assume the subjects in each group are independent. In practice, we would inquire for more details. Normality: The data are bound below by zero and the standard deviations are larger than the means, indicating very strong skew. However, since the sample sizes are extremely large, even extreme skew is acceptable. Constant variance: This condition is sufficiently met, as the standard deviations are reasonably consistent across groups. (c) See below, with the last column omitted:

	Df	Sum Sq	Mean Sq	F value
coffee	4	10508	2627	5.2
Residuals	50734	25564819	504	
Total	50738	25575327		

(d) Since p-value is very small, reject H_0. The data provide convincing evidence that the average MET differs between at least one pair of groups.

5.43 (a) H_0: Average GPA is the same for all majors. H_A: At least one pair of means are different. (b) Since p-value > 0.05, fail to reject H_0. The data do not provide convincing evidence of a difference between the average GPAs across three groups of majors. (c) The total degrees of freedom is $195 + 2 = 197$, so the sample size is $197 + 1 = 198$.

5.45 (a) False. As the number of groups increases, so does the number of comparisons and hence the modified significance level decreases. (b) True. (c) True. (d) False. We need observations to be independent regardless of sample size.

5.47 (a) H_0: Average score difference is the same for all treatments. H_A: At least one pair of means are different. (b) We should check conditions. If we look back to the earlier exercise, we will see that the patients were randomized, so independence is satisfied. There are some minor concerns about skew, especially with the third group, though this may be acceptable. The standard deviations across the groups are reasonably similar. Since the p-value is less than 0.05, reject H_0. The data provide convincing evidence of a difference between the average reduction in score among treatments. (c) We determined that at least two means are different in part (b), so we now conduct $K = 3 \times 2/2 = 3$ pairwise t-tests that each use $\alpha = 0.05/3 = 0.0167$ for a significance level. Use the following hypotheses for each pairwise test. H_0: The two means are equal. H_A: The two means are different. The sample sizes are equal and we use the pooled SD, so we can compute $SE = 3.7$ with the pooled $df = 39$. The p-value for Trmt 1 vs. Trmt 3 is the only one under 0.05: p-value = 0.035 (or 0.024 if using s_{pooled} in place of s_1 and s_3, though this won't affect the final conclusion). The p-value is larger than 0.05/3 = 1.67, so we do not have strong evidence to conclude that it is this particular pair of groups that are different. That is, we cannot identify if which particular pair of groups are actually different, even though we've rejected the notion that they are all the same!

6 Simple linear regression

6.1 (a) Strong relationship, but a straight line would not fit the data. (b) Strong relationship, and a linear fit would be reasonable. (c) Weak relationship, and trying a linear fit would be reasonable. (d) Moderate relationship, but a straight line would not fit the data. (e) Strong relationship, and a linear fit would be reasonable. (f) Weak relationship, and trying a linear fit would be reasonable.

6.3 (a) There is a moderate, positive, and linear relationship between shoulder girth and height. (b) Changing the units, even if just for one of the variables, will not change the form, direction or strength of the relationship between the two variables.

6.5 Over-estimate. Since the residual is calculated as $observed - predicted$, a negative residual means that the predicted value is higher than the observed value.

6.7 (a) $\widehat{murder} = -29.901 + 2.559 \times poverty\%$. (b) Expected murder rate in metropolitan areas with no poverty is -29. 901 per million. This is obviously not a meaningful value, it just serves to adjust the height of the regression line. (c) For each additional percentage increase in poverty, we expect murders per million to be higher on average by 2.559. (e) $\sqrt{0.7052} = 0.8398$.

6.9 (a) The slope of -1.26 indicates that on average, an increase in age of 1 year is associated with a lower RFFT score by 1.26 points. The intercept of 137.55 represents the predicted mean RFFT score for an individual of age 0 years; this does not have interpretive meaning since the RFFT cannot be reasonably administered to a newborn. (b) RFFT score differs on average by $10(-1.26) = 12.6$ points between an individual who is 60 years old versus 50 years old, with the older individual having the lower score. (c) According to the model, average RFFT score for a 70-year-old is $137.55 - 1.26(70) = 49.3$ points. (d) No, it is not valid to use the linear model to estimate RFFT score for a 20-year-old. As indicated in the plot, data are only available for individuals as young as about 40 years old.

6.11 (a) The residual plot will show randomly distributed residuals around 0. The variance is also approximately constant. (b) The residuals will show a fan shape, with higher variability for smaller x. There will also be many points on the right above the line. There is trouble with the model being fit here.

6.13 (a) The points with the lowest and highest values for height have relatively high leverage. They do not seem particularly influential because they are not outliers; the one with a low x-value has a low y-value and the one with a high x-value has a high y-value, which follows the positive trend visible in the data. (b) Yes, since the data show a linear trend, it is appropriate to use R^2 as a metric for describing the strength of the model fit. (c) Height explains about 72% of the observed variability in length.

6.15 There is an upwards trend. However, the variability is higher for higher calorie counts, and it looks like there might be two clusters of observations above and below the line on the right, so we should be cautious about fitting a linear model to these data.

6.17 (a) There is an outlier in the bottom right. Since it is far from the center of the data, it is a point with high leverage. It is also an influential point since, without that observation, the regression line would have a very different slope.
(b) There is an outlier in the bottom right. Since it is far from the center of the data, it is a point with high leverage. However, it does not appear to be affecting the line much, so it is not an influential point.
(c) The observation is in the center of the data (in the x-axis direction), so this point does *not* have high leverage. This means the point won't have much effect on the slope of the line and so is not an influential point.

6.19 (a) Linearity is satisfied; the data scatter about the horizontal line with no apparent pattern. The variability seems constant across the predicted length values. (b) The fish were randomly sampled from a river, so without additional details about the life cycle of the fish, it seems reasonable to assume the height and length of any one fish does not provide information about the height and length of another fish. This could be violated, if, for example, the fish in a river tend to be closely related and height and length are highly heritable.
(c) The residuals are approximately normally distributed, with some small deviations from normality in the tails. There are more outliers in both tails than expected under a normal distribution.

6.21 One possible equation is $\widehat{price} = 44.51 + 12.3(carat_{1.00})$, where the explanatory variable is a binary variable taking on value 1 if the diamond is 1 carat.

6.23 (a) The relationship is positive, moderate-to-strong, and linear. There are a few outliers but no points that appear to be influential.
(b) $\widehat{weight} = -105.0113 + 1.0176 \times height$.
Slope: For each additional centimeter in height, the model predicts the average weight to be 1.0176 additional kilograms (about 2.2 pounds).
Intercept: People who are 0 centimeters tall are expected to weigh - 105.0113 kilograms. This is obviously not possible. Here, the y- intercept serves only to adjust the height of the line and is meaningless by itself.
(c) H_0: The true slope coefficient of height is zero ($\beta_1 = 0$).
H_A: The true slope coefficient of height is different than zero ($\beta_1 \neq 0$).
The p-value for the two-sided alternative hypothesis ($\beta_1 \neq 0$) is incredibly small, so we reject H_0. The data provide convincing evidence that height and weight are positively correlated. The true slope parameter is indeed greater than 0.
(d) $R^2 = 0.72^2 = 0.52$. Approximately 52% of the variability in weight can be explained by the height of individuals.

6.25 (a) H_0: $\beta_1 = 0$. H_A: $\beta_1 \neq 0$. The p-value, as reported in the table, is incredibly small and is smaller than 0.05, so we reject H_0. The data provide convincing evidence that wives' and husbands' heights are positively correlated.
(b) $\widehat{height}_W = 43.5755 + 0.2863 \times height_H$.
(c) Slope: For each additional inch in husband's height, the average wife's height is expected to be an additional 0.2863 inches on average. Intercept: Men who are 0 inches tall are expected to have wives who are, on average, 43.5755 inches tall. The intercept here is meaningless, and it serves only to adjust the height of the line.
(d) The slope is positive, so r must also be positive. $r = \sqrt{0.09} = 0.30$.
(e) 63.33. Since R^2 is low, the prediction based on this regression model is not very reliable.
(f) No, we should avoid extrapolating.
(g) Yes, the p-value for the slope parameter is less than $\alpha = 0.05$. There is sufficient evidence to accept the alternative hypothesis, $H_A : \beta_1 \neq 0$. These data suggest that wife height and husband height are positively associated at the population level.
(h) No, a 95% confidence interval for β_1 would not be expected to contain the null value 0, since the p-value is less than 0.05.

6.27 (a) The point estimate and standard error are $b_1 = 0.9112$ and $SE = 0.0259$. We can compute a T-score: $T = (0.9112 - 1)/0.0259 = -3.43$. Using $df = 168$, the p-value is about 0.001, which is less than $\alpha = 0.05$. That is, the data provide strong evidence that the average difference between husbands' and wives' ages has actually changed over time. (b) $\widehat{age}_W = 1.5740 + 0.9112 \times age_H$. (c) Slope: For each additional year in husband's age, the model predicts an additional 0.9112 years in wife's age. This means that wives' ages tend to be lower for later ages, suggesting the average gap of husband and wife age is larger for older people. Intercept: Men who are 0 years old are expected to have wives who are on average 1.5740 years old. The intercept here is meaningless and serves only to adjust the height of the line. (d) $R = \sqrt{0.88} = 0.94$. The regression of wives' ages on husbands' ages has a positive slope, so the correlation coefficient will be positive. (e) $\widehat{age}_W = 1.5740 + 0.9112 \times 55 = 51.69$. Since R^2 is pretty high, the prediction based on this regression model is reliable. (f) No, we shouldn't use the same model to predict an 85 year old man's wife's age. This would require extrapolation. The scatterplot from an earlier exercise shows that husbands in this data set are approximately 20 to 65 years old. The regression model may not be reasonable outside of this range.

6.29 (a) Yes, since $p < 0.01$. $H_0 : \beta_1 = 0$, $H_A : \beta_1 \neq 0$, where β_1 represents the population average change in RFFT score associated with a change in 1 year of age. There is statistically significant evidence that age is negatively associated with RFFT score. (b) With 99% confidence, the interval (-1.49, -1.03) points contains the population average difference in RFFT score between individuals who differ in age by 1 year; the older individual is predicted to have a lower RFFT score.

6.31 (a) First, compute the standard error: s.e.$(E(age_{wife}|\widehat{age_{husband}} = 55)) = 3.95\sqrt{\frac{1}{170} + \frac{(55-42.92)^2}{(170-1)11.76^2}} = 0.435$. The critical value is $t^\star_{0.975,df=169} = 1.97$. Thus, the 95% confidence interval is $51.69 \pm (1.97)(0.435) = (50.83, 52.55)$ years. (b) First, compute the standard error: s.e.$(age_{wife}|\widehat{age_{husband}} = 55) = 3.95\sqrt{1 + \frac{1}{170} + \frac{(55-42.92)^2}{(170-1)11.76^2}} = 3.97$. The 95% prediction interval is $51.69 \pm (1.97)(3.97) = (43.85, 59.54)$ years. (c) For the approximate 95% confidence interval, use $s/\sqrt{n} = 3.95/\sqrt{170} = 0.303$ as the approximate standard error: (51.09, 52.29) years. For the approximate 95% prediction interval, use $s\sqrt{1 + 1/n} = 3.95\sqrt{1 + 1/170} = 4.25$ as the approximate standard error: (43.30, 60.09) years.

7 Multiple linear regression

7.1 Although the use of statins appeared to be associated with lower RFFT score when no adjustment was made for possible confounders, statin use is not significantly associated with RFFT score in a model that adjusts for age. After adjusting for age, the estimated difference in mean RFFT score between statin users and non-users is 0.85 points; there is a 74% chance of observing such a difference if there is no difference between mean RFFT score in the population of statin users and non-users.

7.3 (a) $\widehat{baby\_weight} = 123.57 - 8.96(smoke) - 1.98(parity)$ (b) A child born to a mother who smokes has a birth weight about 9 ounces less, on average, than one born to a mother who does not smoke, holding birth order constant. A child who is the first born has birth weight about 2 ounces less, on average, than one who is not first born, when comparing children whose mothers were either both smokers or both nonsmokers. The intercept represents the predicted mean birth weight for a child whose mother is not a smoker and who was not the first born. (c) The estimated difference in mean birth weight for two infants born to non-smoking mothers, where one is first born and the other is not, is -1.98. (d) This is the same value as in part (c). (e) $123.57 - 8.96(0) - 1.98(1) = 121.59$ ounces.

7.5 (a) $\widehat{baby\_weight} = -80.41 + 0.44 \times gestation - 3.33 \times parity - 0.01 \times age + 1.15 \times height + 0.05 \times weight - 8.40 \times smoke$. (b) $\beta_{gestation}$: The model predicts a 0.44 ounce increase in the birth weight of the baby for each additional day of pregnancy, all else held constant. β_{age}: The model predicts a 0.01 ounce decrease in the birth weight of the baby for each additional year in mother's age, all else held constant. (c) Parity might be correlated with one of the other variables in the model, which complicates model estimation. (d) $\widehat{baby\_weight} = 120.58$. $e = 120 - 120.58 = -0.58$. The model over-predicts this baby's birth weight. (e) $R^2 = 0.2504$. $R^2_{adj} = 0.2468$.

7.7 Nearly normal residuals: With so many observations in the data set, we look for particularly extreme outliers in the histogram and do not see any. Variability of residuals: The scatterplot of the residuals versus the fitted values does not show any overall structure. However, values that have very low or very high fitted values appear to also have somewhat larger outliers. In addition, the residuals do appear to have constant variability between the two parity and smoking status groups, though these items are relatively minor.
Independent residuals: The scatterplot of residuals versus the order of data collection shows a random scatter, suggesting that there is no apparent structures related to the order the data were collected.
Linear relationships between the response variable and numerical explanatory variables: The residuals vs. height and weight of mother are randomly distributed around 0. The residuals vs. length of gestation plot also does not show any clear or strong remaining structures, with the possible exception of very short or long gestations. The rest of the residuals do appear to be randomly distributed around 0.
All concerns raised here are relatively mild. There are some outliers, but there is so much data that the influence of such observations will be minor.

7.9 (b) True. (c) False. This would only be the case if the data was from an experiment and x_1 was one of the variables set by the researchers. (Multiple regression can be useful for forming hypotheses about causal relationships, but it offers zero guarantees.) (d) False. We should check normality like we would for inference for a single mean: we look for particularly extreme outliers if $n \geq 30$ or for clear outliers if $n < 30$.

7.11 (a) (-0.32, 0.16). We are 95% confident that male students on average have GPAs 0.32 points lower to 0.16 points higher than females when controlling for the other variables in the model. (b) Yes, since the p-value is larger than 0.05 in all cases (not including the intercept).

7.13 (a) $\widehat{eggs.laid} = -17.88 + 4.28(wolbachia) + 0.272(tibia)$ (b) An increase in *Wolbachia* density of one unit is associated with on average 4.28 more eggs laid over a lifetime, assuming body size is held constant. (c) In a multiple regression model adjusting for body size as a potential confounder, increase in *Wolbachia* density was significantly positively associated with realized fitness, measured as the number of eggs laid over a female's full lifetime ($p = 0.002$). These data are consistent with the scientific hypothesis that *Wolbachia* is beneficial for its host in nature. (d) (1.85, 7.05) eggs (e) As a group, the predictors *Wolbachia* density and tibia length are useful for predicting the number of eggs laid over a lifetime.

7.15 (a) Since the difference is taken in the direction (pre - post), a positive value for `trt.effect` indicates that the post-intervention score is lower than the pre-intervention score, which represents efficacy of the intervention. A negative value would represent a patient's deviant T scores increasing after the intervention. (b) Let Y be the change in MMPI score for a participant in this study, $X_{neutral}$ a variable with value 1 for participants assigned to the neutral tape and 0 otherwise, and $X_{therapeutic}$ a variable with value 1 for participants in the emotional neutral group and 0 otherwise. The population-level equation is $E(Y) = \beta_0 + \beta_{neutral}X_{neutral} + \beta_{therapeutic}X_{therapeutic}$. For these data, the estimated model equation is $\widehat{y} = -3.21 + 6.07X_{neutral} + 9.43X_{therapeutic}$. (c) The predicted difference scores $\widehat{y}$ for a patient receiving the neutral tape will be $\widehat{y} = b_0 + b_{neutral}X_{neutral} + b_{therapeutic}X_{therapeutic} = -3.21 + 6.07 + 0 = 2.86$. (d) Yes. The intercept is the average of the score difference for the group that did not hear a taped message. (e) The two slopes represent the change in average MMPI score difference from the average for the group that did not receive a tape. The `Absent` category is the reference group. (f) The p–value for the intercept corresponds to a test of the null hypothesis that the average difference score was 0 in the group that did not hear a taped message. The slope p-values correspond to tests of the null hypotheses of (on average) no change in difference scores between the intervention with no tape and each of the other two interventions.

7.17 (a) Let *pre* and *post* denote the pre- and post-intervention scores, respectively. The estimated equation for the model is $\widehat{post} = 28.41 + 0.66(pre) - 5.73X_{neutral} - 9.75X_{therapeutic}$. (b) Since the coefficient of the pre-intervention score is positive, post-intervention scores tend to increase as the pre-intervention score increases. (c) Yes. The t-statistic for the coefficient of `pre` is 4.05 and is statistically significant. (d) In this model, treatment is a factor variable with three levels and the intervention with no tape is the baseline treatment that does not appear in the model. For a participant with `pre` = 70 and no tape, the predicted value of `post` is $28.41 + 0.66(73) - 5.73(1) = 70.86$ (e) For a given value of `pre`, the coefficient of `treatmentNeutral` is the predicted change in `post` between an participant without a tape and one with the emotionally neutral tape. The model implies that `post` will be 5.7 points lower with the emotionally neutral tape. The evidence for a treatment effect of the emotionally neutral tape is weak; the coefficient is not statistically significant at $\alpha = 0.05$.

7.19 (a) $\widehat{post} = -17.58 + 1.28(pre) + 67.75(neutral) + 64.42(therapeutic) - 0.99(pre \times neutral) - 1.01(pre \times therapeutic)$ (b) The coefficient for `pre` is the predicted increase in post score associated with a 1 unit increase in pre-score for individuals in the absent arm, while the coefficients of the interaction terms for neutral and therapeutic represent the difference in association between pre and post scores for individuals in those groups. For example, an individual in the neutral group is expected to have a $1.28 - 0.99 = 0.29$ point increase in post score, on average, per 1 point increase in pre-score. The coefficients of the slopes for `neutral` and `therapeutic` are differences in intercept values relative to the intercept for the model, which is for the baseline group (absent). (c) Absent: $\widehat{post} = -17.58 + 1.28(pre)$ Neutral: $\widehat{post} = -17.58 + 67.75 + 1.28(pre) - 0.99(pre) = 50.17 + 0.29(pre)$ Therapeutic: $\widehat{post} = -17.58 + 64.42 + 1.28(pre) - 1.01(pre) = 46.84 + 0.27(pre)$ (d) These data suggest there is a statistically significant difference in association between pre- and post-intervention scores by treatment group relative to the group that did not receive any treatment. The coefficients of both interaction terms are statistically significant at $\alpha = 0.05$. Since the slopes are smaller than the slope for the treatment absent group, the data demonstrate that individuals in either treatment group show less increase in MMPI score than occurs when no treatment is applied.

7.21 (a) $\widehat{RFFT} = 140.20 - 13.97(Statin) - 1.31(Age) + 0.25(Statin \times Age)$ (b) The model intercept represents the predicted mean RFFT score for a statin non-user of age 0 years; the intercept does not have a meaningful interpretation. The slope coefficient for age represents the predicted change in RFFT score for a statin non-user; for non-users, a one year increase in age is associated with a 1.32 decrease in RFFT score. The slope coefficient for statin use represents the difference in intercept between the regression line for users and the regression line for non-users; the intercept for users is -13.97 points lower than that of non-users. The interaction term coefficient represents the difference in the magnitude of association between RFFT score and age between users and non-users; in users, the slope coefficient representing predicted change in RFFT score per 1 year change in age is higher by 0.25 points. (c) No, there is not evidence that the association between RFFT score and age differs by statin use. The p-value of the interaction coefficient is 0.32, which is higher than $\alpha = 0.05$.

7.23 Age should be the first variable removed from the model. It has the highest p-value, and its removal results in an adjusted R^2 of 0.255, which is higher than the current adjusted R^2.

7.25 (a) The strongest predictor of birth weight appears to be gestational age; these two variables show a strong positive association. Both parity and smoker status show a slight association with gestational age; the first born child tends to be a lower birth weight and children from mothers who smoke tend to have lower birth weight. While there does not appear to be an association between birth weight and age of the mother, there may be a slight positive association between both birth weight and height and birth weight and weight. All predictor variables with exception of age seem potentially useful for inclusion in an initial model. (b) Height and weight appear to be positively associated.

7.27 (a) The F-statistic for the model corresponds to a test of $H_0 : \beta_{neutral} = \beta_{therapeutic} = 0$. (b) The intercept coefficient is the estimated mean difference score for the no intervention group, and the estimated mean difference score for the other two groups can be calculated by adding each of the slope estimates to the intercept. (c) Under the null hypothesis that the two slope coefficients are 0, all three interventions would have the same mean difference in MMPI scores. This is the same as the null hypothesis for an ANOVA with three groups ($H_0 : \mu_1 = \mu_2 = \mu_3$), which states that all three population means are the same. (d) The assumptions for multiple regression and ANOVA are outlined in Sections 7.3.1 and 5.5, respectively. The assumptions for the two models are the same, though they may be phrased differently. The first assumption in multiple regression is linear change of the mean response variable when one predictor changes and the others do not change. Since each of the two predictor variables in this model can only change from 0 to 1, this assumption is simply that the means in the three groups are possibly different, which is true in ANOVA. The second assumption in regression is that the variance of the residuals is approximately constant. Since the predicted response for an intervention group is its mean, the constant variance assumption in regression is the equivalent assumption in ANOVA that the three groups have approximately constant variance. Both models assume that the observations are independent and that the residuals follow a normal distribution. This is a very long way of saying that the two models are identical!

8 Inference for categorical data

8.1 (a) False. Doesn't satisfy success-failure condition. (b) True. The success-failure condition is not satisfied. In most samples we would expect $\hat{p}$ to be close to 0.08, the true population proportion. While $\hat{p}$ can be much above 0.08, it is bound below by 0, suggesting it would take on a right skewed shape. Plotting the sampling distribution would confirm this suspicion. (c) False. $SE_{\hat{p}} = 0.0243$, and $\hat{p} = 0.12$ is only $\frac{0.12-0.08}{0.0243} = 1.65$ SEs away from the mean, which would not be considered unusual. (d) True. $\hat{p} = 0.12$ is 2.32 standard errors away from the mean, which is often considered unusual. (e) False. Decreases the SE by a factor of $1/\sqrt{2}$.

8.3 (a) False. A confidence interval is constructed to estimate the population proportion, not the sample proportion. (b) True. 95% CI: 82% ± 2%. (c) True. By the definition of the confidence level. (d) True. Quadrupling the sample size decreases the SE and ME by a factor of $1/\sqrt{4}$. (e) True. The 95% CI is entirely above 50%.

8.5 With a random sample, independence is satisfied. The success-failure condition is also satisfied. $ME = z^\star\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 1.96\sqrt{\frac{0.56\times0.44}{600}} = 0.0397 \approx 4\%$

8.7 (a) No. The sample only represents students who took the SAT, and this was also an online survey. (b) (0.5289, 0.5711). We are 90% confident that 53% to 57% of high school seniors who took the SAT are fairly certain that they will participate in a study abroad program in college. (c) 90% of such random samples would produce a 90% confidence interval that includes the true proportion. (d) Yes. The interval lies entirely above 50%.

8.9 (a) We want to check for a majority (or minority), so we use the following hypotheses:

$$H_0 : p = 0.5 \qquad\qquad H_A : p \neq 0.5$$

We have a sample proportion of $\hat{p} = 0.55$ and a sample size of $n = 617$ independents.
Since this is a random sample, independence is satisfied. The success-failure condition is also satisfied: 617×0.5 and $617\times(1-0.5)$ are both at least 10 (we use the null proportion $p_0 = 0.5$ for this check in a one-proportion hypothesis test).
Therefore, we can model $\hat{p}$ using a normal distribution with a standard error of

$$SE = \sqrt{\frac{p(1-p)}{n}} = 0.02$$

(We use the null proportion $p_0 = 0.5$ to compute the standard error for a one-proportion hypothesis test.) Next, we compute the test statistic:

$$Z = \frac{0.55 - 0.5}{0.02} = 2.5$$

This yields a one-tail area of 0.0062, and a p-value of $2 \times 0.0062 = 0.0124$.
Because the p-value is smaller than 0.05, we reject the null hypothesis. We have strong evidence that the support is different from 0.5, and since the data provide a point estimate above 0.5, we have strong evidence to support this claim by the TV pundit.
(b) No. Generally we expect a hypothesis test and a confidence interval to align, so we would expect the confidence interval to show a range of plausible values entirely above 0.5. However, if the confidence level is misaligned (e.g. a 99% confidence level and a $\alpha = 0.05$ significance level), then this is no longer generally true.

8.11 Since a sample proportion ($\hat{p} = 0.55$) is available, we use this for the sample size calculations. The margin of error for a 90% confidence interval is $1.65 \times SE = 1.65 \times \sqrt{\frac{p(1-p)}{n}}$. We want this to be less than 0.01, where we use $\hat{p}$ in place of p:

$$1.65 \times \sqrt{\frac{0.55(1-0.55)}{n}} \leq 0.01$$
$$1.65^2 \frac{0.55(1-0.55)}{0.01^2} \leq n$$

From this, we get that n must be at least 6739.

8.13 (a) $H_0 : p = 0.5$. $H_A : p \neq 0.5$. Independence (random sample) is satisfied, as is the success-failure conditions (using $p_0 = 0.5$, we expect 40 successes and 40 failures). $Z = 2.91 \rightarrow$ the one tail area is 0.0018, so the p-value is 0.0036. Since the p-value < 0.05, we reject the null hypothesis. Since we rejected H_0 and the point estimate suggests people are better than random guessing, we can conclude the rate of correctly identifying a soda for these people is significantly better than just by random guessing. (b) If in fact people cannot tell the difference between diet and regular soda and they were randomly guessing, the probability of getting a random sample of 80 people where 53 or more identify a soda correctly (or 53 or more identify a soda incorrectly) would be 0.0036.

8.15 (a) Yes, it is reasonable to use the normal approximation to the binomial distribution. The sample observations are independent and the expected numbers of successes and failures are greater than 10: $n\hat{p} = (100)(.15) = 15$ and $n(1-\hat{p}) = (100)(0.85) = 85$. (b) An approximate 95% confidence interval is $\hat{p} \pm 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \rightarrow (0.08, 0.22)$. (c) The interval does not support the claim. Since the interval does not contain 0.05, there is statistically significant evidence at $\alpha = 0.05$ that the proportion of young women in the neighborhood who use birth control is different than 0.05. The interval is above 0.05, which is indicative of evidence that more than 5% of young women in the neighborhood use birth control.

8.17 This is not a randomized experiment, and it is unclear whether people would be affected by the behavior of their peers. That is, independence may not hold. Additionally, there are only 5 interventions under the provocative scenario, so the success-failure condition does not hold. Even if we consider a hypothesis test where we pool the proportions, the success-failure condition will not be satisfied. Since one condition is questionable and the other is not satisfied, the difference in sample proportions will not follow a nearly normal distribution.

8.19 (a) Standard error:

$$SE = \sqrt{\frac{0.79(1-0.79)}{347} + \frac{0.55(1-0.55)}{617}} = 0.03$$

Using $z^\star = 1.96$, we get:

$$0.79 - 0.55 \pm 1.96 \times 0.03 \rightarrow (0.181, 0.299)$$

We are 95% confident that the proportion of Democrats who support the plan is 18.1% to 29.9% higher than the proportion of Independents who support the plan. (b) True.

8.21 (a) Test $H_0 : p_1 = p_2$ against $H_A : p_1 \neq p_2$, where p_1 represents the population proportion of clinical improvement in COVID-19 patients treated with remdesivir and p_2 represents the population proportion of clinical improvement in COVID-19 patients treated with placebo. Let $\alpha = 0.05$. The p-value is 0.328, which is greater than α; there is insufficient evidence to reject the null hypothesis of no difference. Even though the proportion of patients who experienced clinical improvement about 7% higher in the remdesivir group, this difference is not extreme enough to represent sufficient evidence that remdesivir is more effective than placebo. (b) The 95% confidence interval is (-0.067, 0.217); with 95% confidence, this interval captures the difference in population proportion of clinical mortality between COVID-19 patients treated with remdesivir and those treated with placebo. The interval contains 0, which is consistent with no statistically significant evidence of a difference. The interval reflects the lack of precision around the effect estimate that is characteristic of an insufficiently large sample size.

8.23 (a) False. The entire confidence interval is above 0. (b) True. (c) True. (d) True. (e) False. It is simply the negated and reordered values: (-0.06,-0.02).

8.25 Subscript $_C$ means control group. Subscript $_T$ means truck drivers. $H_0 : p_C = p_T$. $H_A : p_C \neq p_T$. Independence is satisfied (random samples), as is the success-failure condition, which we would check using the pooled proportion ($\hat{p}_{pool} = 70/495 = 0.141$). $Z = -1.65 \rightarrow$ p-value = 0.0989. Since the p-value is high (default to alpha = 0.05), we fail to reject H_0. The data do not provide strong evidence that the rates of sleep deprivation are different for non-transportation workers and truck drivers.

8.27 (a) False. The chi-square distribution has one parameter called degrees of freedom. (b) True. (c) True. (d) False. As the degrees of freedom increases, the shape of the chi-square distribution becomes more symmetric.

8.29 (a) Two-way table:

	Quit		
Treatment	Yes	No	Total
Patch + support group	40	110	150
Only patch	30	120	150
Total	70	230	300

(b-i) $E_{row_1,col_1} = \frac{(row\ 1\ total) \times (col\ 1\ total)}{table\ total} = 35$. This is lower than the observed value.
(b-ii) $E_{row_2,col_2} = \frac{(row\ 2\ total) \times (col\ 2\ total)}{table\ total} = 115$. This is lower than the observed value.

8.31 (a) H_0: There is no association between statin use and educational level. H_A: There is an association between statin use and educational level
(b) It is reasonable to assume the counts are independent. The smallest expected value in the table is 39.27, so the success-failure condition is reasonably met. (c) There is statistically significant evidence at $\alpha = 0.05$ of an association between educational level and statin use. Individuals with a higher educational level are less likely to be statin users.

8.33 (a)

	No Default	Default	Sum
Non-Diabetic	1053	127	1180
Diabetic	54	0	54
Sum	1107	127	1234

(b) $H_0 : p_1 = p_2$ versus $H_A : p_1 \neq p_2$, where p_1 represents the population proportion of treatment default in diabetics and p_2 represents the population proportion of treatment default in non-diabetics. (c) It is reasonable to assume the counts are independent. The smallest expected value is 5.56, which is not smaller than 5. (d) The χ^2 test statistic is 5.37, with 1 degree of freedom. The p-value of the test statistic is 0.02. There is sufficient evidence to conclude that the proportion of treatment default is higher in non-diabetics than in diabetics.

8.35 (a) One possible 2×2 contingency table:

	Mosquito Nets		
	No	**Yes**	**Total**
Malaria	30	22	52
No Malaria	70	78	148
Total	100	100	200

(b) Expected number of infected children among 100 families who did receive a net: $\frac{52 \times 100}{200} = 26$.

(c) The null hypothesis is H_0 : Using a mosquito net and being infected with malaria are not associated. The alternative is H_A : using a net and being infected with malaria are associated. The χ^2 statistic (1.66) has 1 degree of freedom and the table A3 can be used to show that $p > 0.10$. There is not statistically significant evidence of an association between malaria infection and use of a net in children.

(d) Because this is a prospective study, the relative risk can be calculated directly from the table. Let $p_{\text{No Nets}}$ be the probability that a child without a net will be infected with malaria: $\hat{p}_{\text{No Nets}} = \frac{30}{100} = 0.30$. Let p_{Nets} be the probability that a child with a net will be infected with malaria: $\hat{p}_{\text{Nets}} = \frac{22}{100} = 0.22$. The estimated relative risk: $\widehat{RR} = \frac{\hat{p}_{\text{No Nets}}}{\hat{p}_{\text{Nets}}} = \frac{0.30}{0.22} = 1.36$. The risk of malaria infection for children in the control group is 36% higher than risk for children in the treatment group.

8.37 (a) Under the null hypothesis of no association, the expected cell counts are 9.07 and 7.93 in the wait together and wait alone groups, respectively, for those considered "high anxiety" and 6.93 and 6.07 in the wait together and wait alone groups, respectively, for those considered "low anxiety". (b) Use the hypergeometric distribution with parameters $N = 30$, $m = 16$, and $n = 17$; calculate $P(X = 12)$. Consider the "successes" to be the individuals who wait together, and the "number sampled" to be the people randomized to the high-anxiety group. The probability of the observed set of results, assuming the marginal totals are fixed and the null hypothesis is true, is 0.0304. (c) More individuals than expected in the high-anxiety group were observed to wait together; thus, tables that are more extreme in the same direction also consist of those where more people in the high-anxiety group wait together than observed. These are tables in which 13, 14, 15, or 16 individuals in the high-anxiety group wait together.

	Wait Together	Wait Alone	Sum
High-Anxiety	13	4	17
Low-Anxiety	3	10	13
Sum	16	14	30

	Wait Together	Wait Alone	Sum
High-Anxiety	14	3	17
Low-Anxiety	2	11	13
Sum	16	14	30

	Wait Together	Wait Alone	Sum
High-Anxiety	15	2	17
Low-Anxiety	1	12	13
Sum	16	14	30

	Wait Together	Wait Alone	Sum
High-Anxiety	16	1	17
Low-Anxiety	0	13	13
Sum	16	14	30

(d) Let p_1 represent the population proportion of individuals waiting together in the high-anxiety group and p_2 represent the population proportion of individuals waiting together in the low-anxiety group. Test $H_0 : p_1 = p_2$ against $H_A : p_1 \neq p_2$. Let $\alpha = 0.05$. The two-sided p-value is 0.063. There is insufficient evidence to reject the null hypothesis; the data do not suggest there is an association between high anxiety and a person's desire to be in the company of others.

8.39 (a) H_0: The distribution of the format of the book used by the students follows the professor's predictions. H_A: The distribution of the format of the book used by the students does not follow the professor's predictions. (b) $E_{hard\ copy} = 126 \times 0.60 = 75.6$. $E_{print} = 126 \times 0.25 = 31.5$. $E_{online} = 126 \times 0.15 = 18.9$. (c) Independence: The sample is not random. However, if the professor has reason to believe that the proportions are stable from one term to the next and students are not affecting each other's study habits, independence is probably reasonable. Sample size: All expected counts are at least 5. (d) $\chi^2 = 2.32$, $df = 2$, p-value = 0.313. (e) Since the p-value is large, we fail to reject H_0. The data do not provide strong evidence indicating the professor's predictions were statistically inaccurate.

8.41 (a)

	CVD	No CVD
Age Onset $\leq$ 50 Years	15	25
Age Onset > 50 Years	5	55

(b) The odds of CVD for patients older than 50 years when diagnosed with diabetes is 5/55 = 0.09. The odds of CVD for the patients younger than 50 years at diabetes onset is 15/25 = 0.60. The relative odds (or odds ratio, OR) is 0.09/0.60 = 0.15.

(c) The odds of CVD for someone with late onset diabetes is less than 1/5 that of people with earlier onset diabetes. This can be explained by the fact that people with diabetes tend to build up plaque in their arteries; with early onset diabetes, plaque has longer time to accumulate, eventually causing CVD.

(d) $H_0 : OR = 1$.

(e) The chi-square test can be used to test H_0 as long as the conditions for the test have been met. The observations are likely independent; knowing one person's age of diabetes onset and CVD status is unlikely to provide information about another person's age of diabetes onset and CVD status. Under H_0, the expected cell count for the lower left cell is (60)(20)/100 = 12, which is bigger than 5; all other expected cell counts will be larger.

(f) Since the study is not a randomized experiment, it cannot demonstrate causality. It may be the case, for example, that CVD presence causes earlier onset of diabetes. The study only demonstrates an association between cardiovascular disease and diabetes.

8.43 (a) No. This is an example of outcome dependent sampling. Subjects were first identified according to presence or absence of the CNS disorder, then queried about use of the drug. It is only possible to estimate the probability that someone had used the drug, given they either did or did not have a CNS disorder.

(b) The appropriate measure of association is the odds ratio.

(c) The easiest way of calculating the OR for the table is the cross-product of the diagonal elements of the table: [(10)(4000)]/[(2000)(7)] = 2.86. Using the definition, it can be calculated as:

$$\hat{OR} = \frac{\frac{\hat{P}(\text{CNS}|\text{ Usage})}{1-\hat{P}(\text{CNS}|\text{ Usage})}}{\frac{\hat{P}(\text{CNS}|\text{ No Usage})}{1-\hat{P}(\text{CNS}|\text{ No Usage})}} = \frac{ad}{bc} = \frac{(10)(4000)}{(2000)(7)} = 2.86$$

(d) The odds ratio has the interpretation of the relative odds of presence of a CNS disorder, comparing people who have used the weight loss drug to those who have not. People who have used the weight loss drug have odds of CNS that are almost three times as large as those for people who have not used the drug.

(e) Fisher's exact test is better than the chi-square test. The independence assumption is met, but the expected cell count corresponding the presence of a CNS disorder and the use of the drug is 5.68, so not all the expected cell counts are less than 10.

8.45 (a) The p-value is 0.92; there is insufficient evidence to reject the null hypothesis of no association. These data are plausible with the null hypothesis that green tea consumption is independent of esophageal carcinoma. (b) Since the study uses outcome-dependent sampling, the odds ratio should be used as a measure of association rather than relative risk. The odds ratio of esophageal carcinoma, comparing green tea drinkers to non-drinkers, is 1.08; the odds of carcinoma for those who regularly drink green tea are 8% larger than the odds for those who never drink green tea.

8.47 (a) The prevalence difference is 0.15 – 0.10 = 0.05 and the prevalence ratio is 0.15/0.10 = 1.50. The absolute prevalence of disease in one group is 0.05 higher than in the other group. For instance, in a population

of 100,000 one would expect 10,000 cases in the first group 15,000 in the second group, and increase of 5,000 cases. If the prevalence is 1.50 times as large as that in the other group, the difference of 10,000 vs 15,000 cases in the hypothetical example represents 50% more cases. (b) The prevalence difference is 0.45 – 0.40 = 0.05 and the prevalence ratio is 0.45/0.40 = 1.125. In a population of 100,000, one would expect 40,000 cases in the lower prevalence group and 45,000 cases in the higher prevalence group, a difference of 5,000 cases. The difference of 5,000 cases is a 12.5% incease.

8.49 (a) The estimated odds that a male had a high salt diet are 7/53 = 0.132 and the estimated odds that a male had a low salt diet are 53/7 = 7.58. (b) Among the men where the recorded death was due to CVD, the odds of high salt diet are 5/30 = 0.167. The odds of low salt diet in the same group are 30/5 = 6. (c) The OR for a CVD related death, comparing a high to a low salt diet are (5/2)/(30/23) = 1.92. (d) The OR for a non CVD related death, comparing a high to a low salt diet are (2/5)/(23/30) = 0.522.

8.51 (a) Let $\hat{p}_1$ represent the observed proportion who experience the outcome of interest among those assigned to placebo and $\hat{p}_2$ the observed proportion who experience the outcome of interest among those assigned to tofacitinib; $\hat{p}_1 = 42/145 = 0.290$ and $\hat{p}_2 = 26/144 = 0.181$. The 95% CI is (−0.0527, −.2709). (b) Test $H_0 : p_1 = p_2$ against $H_A : p_1 \neq p_2$. Let $\alpha = 0.05$. With the z-test method, the z-statistic is 2.186. The two-sided p-value is $P|Z| \geq 2.186 = 0.0289$, which is smaller than 0.05. There is sufficient evidence to reject the null hypothesis; the evidence suggests that tofacitinib is an effective treatment compared to placebo. (c) The 95% CI for the risk ratio is (1.0422, 2.469). There is a larger risk of death or respiratory failure during the follow-up period for individuals on the placebo group than for individuals on tofactinib that could be as high as over twice the risk or as low as 1.04 times the risk.

8.53 (a) Given that the upper left cell has value 4 and that the margins are fixed, the other values in the table (going clockwise) are 1, 5, and 1. (b) The relative risk for response, comparing treatment to control, is (4/5)/(1/6) = 4.8. (c) There is only one table more extreme whose results favor treatment; the table in which all 5 individuals in the treatment group show a response. (d) The one-sided p-value consists of the probability of the observed table plus the probability of the table with a 5 in the upper left cell. Thus, the p-value is $\frac{\binom{5}{4}\binom{6}{1}}{\binom{11}{5}} + \frac{\binom{5}{5}\binom{6}{0}}{\binom{11}{5}} = 0.067$.

9 Logistic Regression

9.1 (a) Odds of rolling a six are (1/6)/(5/6) = 1/5. (b) Odds of rolling an even number are (3/6)/(3/6) = 1. (c) The probability of rolling an even number is 1/2; in a large number of rolls of the die an even number will appear approximately 50% of the time. The odds of rolling an even number is the ratio of the number times an even number appears to the number of times it does not. The odds are 1 because an even number shows up as often as it does not.

9.3 (a) The estimated conditional log odds are exp[−6.054 + (0.185)(25.93)] = 0.285. The estimated probability is 0.221. (b) Both the odds and the probability calculated from the model lie above the tabulated values in Figure 9.2.

9.5 (a) The odds ratio can be calculated directly and is exp[−(0.6)(6 − 4)] = exp[−(.6)(2)] = 0.301. (b) Relative risk is the ratio of the two probabilities, which depend on individual odds. The two odds are exp[3.0 − (0.6)(6)] = 0.549 and exp[3.0 − (0.6)(4)] = 1.822; the two probabilities are 0.549/(1.0 + 0.549) = 0.354 and 1.822/(1 + 1.822) = 0.646. The relative risk is 0.254/0.646 = 0.393. (c) The odds ratio does not depend on the intercept, but the probabilities and hence the relative risk does.

9.7 (a) The odds of survival are exp[1.44 − (0.065)(10)] = 2.203. (b) The odds of survival for someone requiring 20 minutes of CPR are exp[1.44 − (0.065)(20)] = 1.15. The OR is 2.203/1.15 = 1.916. (c) The two estimated probabilities for survival to discharge are 2.203/(1 + 2.203) = 0.688 and 1.15/(1 + 1.15) = 0.535. (d) The relative risk is given by $RR = 0.688/0.535 = 1.286$. (e) A relative risk of 1.286 means that patients requiring 10 minutes of CPR have a chance of surviving to discharge that is approximately 1.3 times that of patients requiring 20 minutes, or approximately 30% larger.

9.9 (a) The algebraic form of the model is

$$\log(\widehat{\text{odds}}_{\text{E}}(\text{Mg})) = -1.089 - 0.007(\text{age}),$$

where E is the event of being hyperuricemic. (b) Because the coefficient of age is negative, increasing age is associated with an decrease in the odds of hyperuricemia. (c) The predicted odds are exp[−1.089 − 0.007(50)] = 0.237. (d) The OR comparing a 50 to a 30 year old is exp[−0.007(50 − 30)] = 0.869. The odds of hyperuricemia in a 50 year old is 0.869 times that of a 30 year old; odds are decreased by 13%. (e) The predicted probability of hyperuricemia for a 50 year old is 0.237/(1 + 0.237) = 0.192. (f) The odds of hyperuricemia for 30 year old are exp[−1.089 − 0.007(30)] = 0.273, so the estimated probability is 0.273/(1 + 0.273) = 0.215. The risk ratio is 0.192/0.215 = 0.893. The probability will be decreased by approximately 11%.

9.11 (a) False. Logistic regression models should be fit only when there are at least 10 cases with the less frequent yes/no outcome. (b) Increased risk. Since the log(odds) is positive, increasing values of the predictor will be associated with increases in log(odds) and odds. Probabilities increase whenever odds do. (c) No. Estimated probabilities also depend on the intercept in a logistic regression. (d) No. The z-score for the estimate is 0.750/0.650 = 1.154, smaller than 1.96 for a traditional 0.05 level test.

9.13 (a) The z statistic is the estimate divided by its standard error, 0.033/0.526 = 0.063. (b) No, the data do not show a statistically significant association, since $p = 2P(Z > 0.063) = 0.950$. (c) A 95% confidence interval for the estimate is $0.033 \pm (1.96)(0.526) = (-0.998, 1.064)$. With 95% confidence, a 1gm change in magnesium is associated with a change in log odds from −0.998 to 1.064. (d) First construct the confidence interval on the log odds scale. The estimated model coefficient b_1 is the change in log odds corresponding to a one unit change in magnesium. When magnesium increases from 0.25gm to 0.75gm, the change in log odds will be $(0.75 - 0.25)b_1 = (0.50)(0.033) = 0.017$. On the log odds scale, the 95% interval for the change will be the confidence interval for $0.50b_1$. Since the standard error of $0.50b_1 = (0.50)(0.526) = 0.263$, the 95% interval for the change in log odds is $0.017 \pm 1.96(0.263) = (-0.499, 0.532)$. The 95% interval for the odds ratio is $\exp(-0.499), \exp(0.532) = 0.607, 1.702$. The confidence interval on the log odds scale could also have been calculated by multiplying the upper and lower bounds for the confidence interval for b_1 by 0.50.

9.15 (a) No. The number of recorded leukemia cases will be (1500)(0.0025) = 3.75, much less than the minimum of 10 events in the lower prevalence outcome. (b) Since 10/3.75 = 2.67, the number of surveyed homes

should be larger by at least a factor of 2.67, or $(10,000)(2.67) = 27,600$ homes. (c) It might be reasonable to use a 95% lower confidence bound for the proportion of observed leukemia cases, using the observed proportion $3.75/10,000 = 0.000375$. Rounding 3.75 to 4, the 95% confidence interval for a binomial proportion with 4 events in 10,000 observations is $(0.000128, 0.001010)$. Using 0.000128, 10 events would be expected in a sample size of $10/0.000128 = 78,125$ homes.

9.17 (a) The two individuals differ by 4 units of BMI, so find the 95% interval for $4\beta_1$. The estimate $4b_1 = 4(0.185) = 0.740$, with standard error $4(0.037) = 0.148$. The confidence interval is given by

$$0.740 \pm (1.96)(0.148) \longrightarrow (0.450, 1.03).$$

(b) No, the 95% confidence interval includes 1.

9.19 (a) The OR does not depend on the age the participant. From the data in the Figure 9.47, the log odds ratio comparing women with and without a prior fracture is 0.839, so the estimated OR = exp(0.839) = 2.314. A woman with a prior fracture has more than double the odds of experiencing a fracture during the study than one without a prior fracture. (b) The OR does not depend on whether or not the woman has experienced a prior fracture. There is a 10 year difference in the ages of the two women, so log(OR) = 10(0.041) = 0.140. The OR = exp(0.140) = 0.15. The older women has an estimated OR that is 15% larger than the younger woman. (c) The design of the study was exposure based, so in the full data of approximately 60,000 women both odds and prevalence ratios could have been estimated. Since the sample of 500 was outcome based, it still allows estimates of ORs. (d) Because the sample of 500 was outcome based, prevalence ratios cannot be estimated from the data.

9.21 The conditions for the χ^2 test are discussed in Section 8.3.2. The formulas for expected cell counts under the assumption of independence are given in Section 8.3.1. All of the cases in the dataset contribute independent data. In a two-way table with more than 4 cells, no more than 1/5 of the expected cell counts should be less than 5 and all expected cell counts should be greater than 1. Since there are 8 cells in the table the two lowest expected cell counts should be at least 5. There are no cell counts less than 1. Only one expected cell count is less than 5 (calculations not shown). The expected cell count for the number crabs with a light colored carapace and no satellites is 4.30. The conditions are satisfied.

9.23 (a) The estimate of the intercept is $log(7/33) = -1.551$. (b) The estimate for the triage category "green" is the log of the estimated OR comparing "green" to "red", $log((11/253)/(7/33)) = -1.585$. (c) Yes. Each of the log odds ratios comparing a category with "red" can be thought of as coming from a 2×2 table with the two rows consisting of counts from "red" and the category of interest. In those tables, the standard errors of the log(OR) can be calculated using the formulas in Section 8.6.4. For instance for the category "green" the standard error of the log(OR), the standard error is given by $\sqrt{(1/253 + 1/11 + 1/33 + 1/7)} = 0.518$. Because the calculations of standard errors in logistic regression and 2-way tables are different the standard errors from the two methods may differ slightly.

9.25 (a) For males, Equation 9.30 reduces to

$$\log(\text{odds}_E) = -5.006 + (0.152)\text{bmi},$$

since sex has the reference value "male". The difference between the two values of BMI is $33.2 - 30.0 = 3.2$ so the difference between the log(odds) is $(0.152)(3.2) = 0.486$. The estimated OR = $e^{0.486} = 1.626$. (b) For females, Equation 9.30 becomes

$$\begin{aligned}\log(\text{odds}_E) &= -5.006 + (0.152)\text{bmi} - 1.652 + (0.046)\text{bmi} \\ &= -(5.006 + 1.652) + (0.152 + 0.046)\text{bmi} \\ &= -6.658 + (0.198)\text{bmi}.\end{aligned}$$

The difference in estimated log odds is $(0.198)(3.2) = 0.634$, and the estimated OR is $e^{0.634} = 1.885$. Females with a BMI = 33.2 have an estimated odds of hyperuricemia that is approximately 1.88 times higher (88% higher) than females with BMI = 30, while for males the OR is 1.63 times higher (63% higher). (c) Using the model in Figure 9.12, the estimated OR for hyperuricemia for BMI = 33.2 versus 30 kg/m^2 is exp[(3.2)(0.171)] =

1.728, a value that is different from the two ORs calculated above.