Chapter 8

Multiple and logistic regression

The principles of simple linear regression lay the foundation for more sophisticated regression methods used in a wide range of challenging settings. In Chapter 8, we explore multiple regression, which introduces the possibility of more than one predictor, and logistic regression, a technique for predicting categorical outcomes with two possible categories.

8.1 Introduction to multiple regression 

Multiple regression extends simple two-variable regression to the case that still has one response but many predictors (denoted x_1, x_2, x_3, ...). The method is motivated by scenarios where many variables may be simultaneously connected to an output.

We will consider Ebay auctions of a video game called *Mario Kart* for the Nintendo Wii. The outcome variable of interest is the total price of an auction, which is the highest bid plus the shipping cost. We will try to determine how total price is related to each characteristic in an auction while simultaneously controlling for other variables. For instance, all other characteristics held constant, are longer auctions associated with higher or lower prices? And, on average, how much more do buyers tend to pay for additional Wii wheels (plastic steering wheels that attach to the Wii controller) in auctions? Multiple regression will help us answer these and other questions.

The data set `mario_kart` includes results from 141 auctions.[1] Four observations from this data set are shown in Table 8.1, and descriptions for each variable are shown in Table 8.2. Notice that the condition and stock photo variables are indicator variables. For instance, the `cond_new` variable takes value 1 if the game up for auction is new and 0 if it is used. Using indicator variables in place of category names allows for these variables to be directly used in regression. See Section 7.2.7 for additional details. Multiple regression also allows for categorical variables with many levels, though we do not have any such variables in this analysis, and we save these details for a second or third course.

[1] Diez DM, Barr CD, Çetinkaya-Rundel M. 2015. `openintro`: OpenIntro data sets and supplement functions. github.com/OpenIntroOrg/openintro-r-package.

	price	cond_new	stock_photo	duration	wheels
1	51.55	1	1	3	1
2	37.04	0	1	7	1
⋮	⋮	⋮	⋮	⋮	⋮
140	38.76	0	0	7	0
141	54.51	1	1	1	2

Table 8.1: Four observations from the `mario_kart` data set.

variable	description
`price`	final auction price plus shipping costs, in US dollars
`cond_new`	a coded two-level categorical variable, which takes value `1` when the game is new and `0` if the game is used
`stock_photo`	a coded two-level categorical variable, which takes value `1` if the primary photo used in the auction was a stock photo and `0` if the photo was unique to that auction
`duration`	the length of the auction, in days, taking values from 1 to 10
`wheels`	the number of Wii wheels included with the auction (a *Wii wheel* is a plastic racing wheel that holds the Wii controller and is an optional but helpful accessory for playing Mario Kart)

Table 8.2: Variables and their descriptions for the `mario_kart` data set.

8.1.1 A single-variable model for the Mario Kart data

Let's fit a linear regression model with the game's condition as a predictor of auction price. The model may be written as

$$\widehat{price} = 42.87 + 10.90 \times cond\_new$$

Results of this model are shown in Table 8.3 and a scatterplot for price versus game condition is shown in Figure 8.4.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	42.8711	0.8140	52.67	0.0000
cond_new	10.8996	1.2583	8.66	0.0000
				$df = 139$

Table 8.3: Summary of a linear model for predicting auction price based on game condition.

⊙ **Guided Practice 8.1** Examine Figure 8.4. Does the linear model seem reasonable?[2]

[2]Yes. Constant variability, nearly normal residuals, and linearity all appear reasonable.

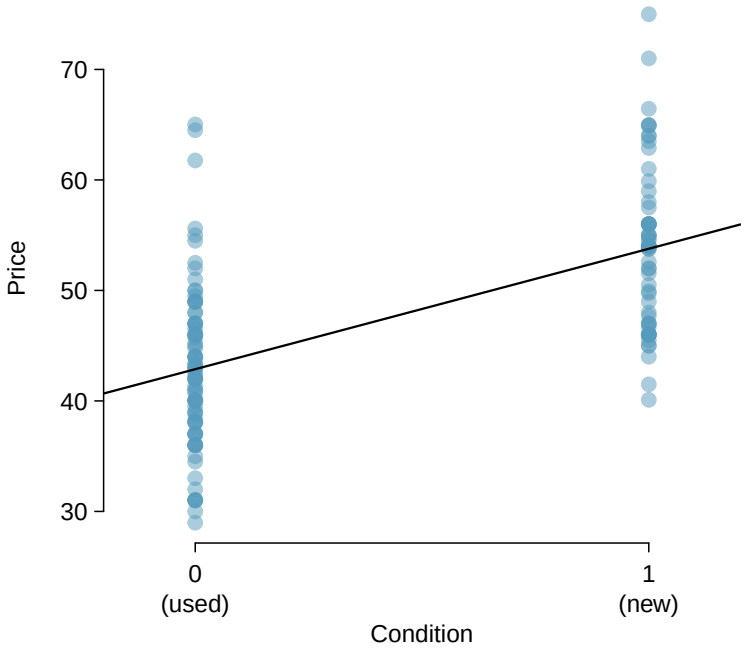


Figure 8.4: Scatterplot of the total auction price against the game's condition. The least squares line is also shown.

Example 8.2 Interpret the coefficient for the game's condition in the model. Is this coefficient significantly different from 0?

Note that `cond_new` is a two-level categorical variable that takes value 1 when the game is new and value 0 when the game is used. So 10.90 means that the model predicts an extra $10.90 for those games that are new versus those that are used. (See Section 7.2.7 for a review of the interpretation for two-level categorical predictor variables.) Examining the regression output in Table 8.3, we can see that the p-value for `cond_new` is very close to zero, indicating there is strong evidence that the coefficient is different from zero when using this simple one-variable model.

8.1.2 Including and assessing many variables in a model

Sometimes there are underlying structures or relationships between predictor variables. For instance, new games sold on Ebay tend to come with more Wii wheels, which may have led to higher prices for those auctions. We would like to fit a model that includes all potentially important variables simultaneously. This would help us evaluate the relationship between a predictor variable and the outcome while controlling for the potential influence of other variables. This is the strategy used in **multiple regression**. While we remain cautious about making any causal interpretations using multiple regression, such models are a common first step in providing evidence of a causal connection.

We want to construct a model that accounts for not only the game condition, as in Section 8.1.1, but simultaneously accounts for three other variables: `stock_photo`, `duration`, and `wheels`.

$$\begin{aligned}\widehat{\texttt{price}} &= \beta_0 + \beta_1 \times \texttt{cond\_new} + \beta_2 \times \texttt{stock\_photo} \\ &\quad + \beta_3 \times \texttt{duration} + \beta_4 \times \texttt{wheels} \\ \hat{y} &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 \end{aligned} \tag{8.3}$$

In this equation, y represents the total price, x_1 indicates whether the game is new, x_2 indicates whether a stock photo was used, x_3 is the duration of the auction, and x_4 is the number of Wii wheels included with the game. Just as with the single predictor case, a multiple regression model may be missing important components or it might not precisely represent the relationship between the outcome and the available explanatory variables. While no model is perfect, we wish to explore the possibility that this one may fit the data reasonably well.

We estimate the parameters β_0, β_1, ..., β_4 in the same way as we did in the case of a single predictor. We select b_0, b_1, ..., b_4 that minimize the sum of the squared residuals:

$$SSE = e_1^2 + e_2^2 + \cdots + e_{141}^2 = \sum_{i=1}^{141} e_i^2 = \sum_{i=1}^{141} (y_i - \hat{y}_i)^2 \tag{8.4}$$

Here there are 141 residuals, one for each observation. We typically use a computer to minimize the sum in Equation (8.4) and compute point estimates, as shown in the sample output in Table 8.5. Using this output, we identify the point estimates b_i of each β_i, just as we did in the one-predictor case.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	36.2110	1.5140	23.92	0.0000
cond_new	5.1306	1.0511	4.88	0.0000
stock_photo	1.0803	1.0568	1.02	0.3085
duration	-0.0268	0.1904	-0.14	0.8882
wheels	7.2852	0.5547	13.13	0.0000
				$df = 136$

Table 8.5: Output for the regression model where `price` is the outcome and `cond_new`, `stock_photo`, `duration`, and `wheels` are the predictors.

Multiple regression model

A multiple regression model is a linear model with many predictors. In general, we write the model as

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

when there are k predictors. We often estimate the β_i parameters using a computer.

⊙ **Guided Practice 8.5** Write out the model in Equation (8.3) using the point estimates from Table 8.5. How many predictors are there in this model?[3]

⊙ **Guided Practice 8.6** What does β_4, the coefficient of variable x_4 (Wii wheels), represent? What is the point estimate of β_4?[4]

⊙ **Guided Practice 8.7** Compute the residual of the first observation in Table 8.1 on page 373 using the equation identified in Guided Practice 8.5.[5]

● **Example 8.8** We estimated a coefficient for `cond_new` in Section 8.1.1 of $b_1 = 10.90$ with a standard error of $SE_{b_1} = 1.26$ when using simple linear regression. Why might there be a difference between that estimate and the one in the multiple regression setting?

If we examined the data carefully, we would see that some predictors are correlated. For instance, when we estimated the connection of the outcome `price` and predictor `cond_new` using simple linear regression, we were unable to control for other variables like the number of Wii wheels included in the auction. That model was biased by the confounding variable `wheels`. When we use both variables, this particular underlying and unintentional bias is reduced or eliminated (though bias from other confounding variables may still remain).

Example 8.8 describes a common issue in multiple regression: correlation among predictor variables. We say the two predictor variables are **collinear** (pronounced as *co-linear*) when they are correlated, and this collinearity complicates model estimation. While it is impossible to prevent collinearity from arising in observational data, experiments are usually designed to prevent predictors from being collinear.

⊙ **Guided Practice 8.9** The estimated value of the intercept is 36.21, and one might be tempted to make some interpretation of this coefficient, such as, it is the model's predicted price when each of the variables take value zero: the game is used, the primary image is not a stock photo, the auction duration is zero days, and there are no wheels included. Is there any value gained by making this interpretation?[6]

8.1.3 Adjusted R^2 as a better estimate of explained variance

We first used R^2 in Section 7.2 to determine the amount of variability in the response that was explained by the model:

$$R^2 = 1 - \frac{\text{variability in residuals}}{\text{variability in the outcome}} = 1 - \frac{Var(e_i)}{Var(y_i)}$$

where e_i represents the residuals of the model and y_i the outcomes. This equation remains valid in the multiple regression framework, but a small enhancement can often be even more informative.

[3] $\hat{y} = 36.21 + 5.13x_1 + 1.08x_2 - 0.03x_3 + 7.29x_4$, and there are $k = 4$ predictor variables.

[4] It is the average difference in auction price for each additional Wii wheel included when holding the other variables constant. The point estimate is $b_4 = 7.29$.

[5] $e_i = y_i - \hat{y}_i = 51.55 - 49.62 = 1.93$, where 49.62 was computed using the variables values from the observation and the equation identified in Guided Practice 8.5.

[6] Three of the variables (`cond_new`, `stock_photo`, and `wheels`) do take value 0, but the auction duration is always one or more days. If the auction is not up for any days, then no one can bid on it! That means the total auction price would always be zero for such an auction; the interpretation of the intercept in this setting is not insightful.

⊙ **Guided Practice 8.10** The variance of the residuals for the model given in Guided Practice 8.7 is 23.34, and the variance of the total price in all the auctions is 83.06. Calculate R^2 for this model.[7]

This strategy for estimating R^2 is acceptable when there is just a single variable. However, it becomes less helpful when there are many variables. The regular R^2 is a less estimate of the amount of variability explained by the model. To get a better estimate, we use the adjusted R^2.

Adjusted R^2 as a tool for model assessment

The **adjusted R^2** is computed as

$$R^2_{adj} = 1 - \frac{Var(e_i)/(n-k-1)}{Var(y_i)/(n-1)} = 1 - \frac{Var(e_i)}{Var(y_i)} \times \frac{n-1}{n-k-1}$$

where n is the number of cases used to fit the model and k is the number of predictor variables in the model.

Because k is never negative, the adjusted R^2 will be smaller – often times just a little smaller – than the unadjusted R^2. The reasoning behind the adjusted R^2 lies in the **degrees of freedom** associated with each variance.[8]

⊙ **Guided Practice 8.11** There were $n = 141$ auctions in the `mario_kart` data set and $k = 4$ predictor variables in the model. Use n, k, and the variances from Guided Practice 8.10 to calculate R^2_{adj} for the Mario Kart model.[9]

⊙ **Guided Practice 8.12** Suppose you added another predictor to the model, but the variance of the errors $Var(e_i)$ didn't go down. What would happen to the R^2? What would happen to the adjusted R^2?[10]

Adjusted R^2 could have been used in Chapter 7. However, when there is only $k = 1$ predictors, adjusted R^2 is very close to regular R^2, so this nuance isn't typically important when considering only one predictor.

[7] $R^2 = 1 - \frac{23.34}{83.06} = 0.719$.

[8] In multiple regression, the degrees of freedom associated with the variance of the estimate of the residuals is $n-k-1$, not $n-1$. For instance, if we were to make predictions for new data using our current model, we would find that the unadjusted R^2 is an overly optimistic estimate of the reduction in variance in the response, and using the degrees of freedom in the adjusted R^2 formula helps correct this bias.

[9] $R^2_{adj} = 1 - \frac{23.34}{83.06} \times \frac{141-1}{141-4-1} = 0.711$.

[10] The unadjusted R^2 would stay the same and the adjusted R^2 would go down.

8.2 Model selection

The best model is not always the most complicated. Sometimes including variables that are not evidently important can actually reduce the accuracy of predictions. In this section we discuss model selection strategies, which will help us eliminate variables from the model that are found to be less important.

In practice, the model that includes all available explanatory variables is often referred to as the **full model**. The full model may not be the best model, and if it isn't, we want to identify a smaller model that is preferable.

8.2.1 Identifying variables in the model that may not be helpful

Adjusted R^2 describes the strength of a model fit, and it is a useful tool for evaluating which predictors are adding value to the model, where *adding value* means they are (likely) improving the accuracy in predicting future outcomes.

Let's consider two models, which are shown in Tables 8.6 and 8.7. The first table summarizes the full model since it includes all predictors, while the second does not include the `duration` variable.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	36.2110	1.5140	23.92	0.0000
cond_new	5.1306	1.0511	4.88	0.0000
stock_photo	1.0803	1.0568	1.02	0.3085
duration	-0.0268	0.1904	-0.14	0.8882
wheels	7.2852	0.5547	13.13	0.0000
$R^2_{adj} = 0.7108$				$df = 136$

Table 8.6: The fit for the full regression model, including the adjusted R^2.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	36.0483	0.9745	36.99	0.0000
cond_new	5.1763	0.9961	5.20	0.0000
stock_photo	1.1177	1.0192	1.10	0.2747
wheels	7.2984	0.5448	13.40	0.0000
$R^2_{adj} = 0.7128$				$df = 137$

Table 8.7: The fit for the regression model for predictors `cond_new`, `stock_photo`, and `wheels`.

Example 8.13 Which of the two models is better?

We compare the adjusted R^2 of each model to determine which to choose. Since the first model has an R^2_{adj} smaller than the R^2_{adj} of the second model, we prefer the second model to the first.

Will the model without `duration` be better than the model with `duration`? We cannot know for sure, but based on the adjusted R^2, this is our best assessment.

8.2.2 Two model selection strategies

Two common strategies for adding or removing variables in a multiple regression model are called *backward elimination* and *forward selection*. These techniques are often referred to as **stepwise** model selection strategies, because they add or delete one variable at a time as they "step" through the candidate predictors.

Backward elimination starts with the model that includes all potential predictor variables. Variables are eliminated one-at-a-time from the model until we cannot improve the adjusted R^2. The strategy within each elimination step is to eliminate the variable that leads to the largest improvement in adjusted R^2.

Example 8.14 Results corresponding to the *full model* for the `mario_kart` data are shown in Table 8.6. How should we proceed under the backward elimination strategy?

Our baseline adjusted R^2 from the full model is $R^2_{adj} = 0.7108$, and we need to determine whether dropping a predictor will improve the adjusted R^2. To check, we fit four models that each drop a different predictor, and we record the adjusted R^2 from each:

Exclude ...	`cond_new`	`stock_photo`	`duration`	`wheels`
	$R^2_{adj} = 0.6626$	$R^2_{adj} = 0.7107$	$R^2_{adj} = 0.7128$	$R^2_{adj} = 0.3487$

The third model without `duration` has the highest adjusted R^2 of 0.7128, so we compare it to the adjusted R^2 for the full model. Because eliminating `duration` leads to a model with a higher adjusted R^2, we drop `duration` from the model.

Since we eliminated a predictor from the model in the first step, we see whether we should eliminate any additional predictors. Our baseline adjusted R^2 is now $R^2_{adj} = 0.7128$. We now fit three new models, which consider eliminating each of the three remaining predictors:

Exclude `duration` and ...	`cond_new`	`stock_photo`	`wheels`
	$R^2_{adj} = 0.6587$	$R^2_{adj} = 0.7124$	$R^2_{adj} = 0.3414$

None of these models lead to an improvement in adjusted R^2, so we do not eliminate any of the remaining predictors. That is, after backward elimination, we are left with the model that keeps `cond_new`, `stock_photos`, and `wheels`, which we can summarize using the coefficients from Table 8.7:

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + b_4 x_4$$

$$\widehat{price} = 36.05 + 5.18 \times \texttt{cond\_new} + 1.12 \times \texttt{stock\_photo} + 7.30 \times \texttt{wheels}$$

The **forward selection** strategy is the reverse of the backward elimination technique. Instead of eliminating variables one-at-a-time, we add variables one-at-a-time until we cannot find any variables that improve the model (as measured by adjusted R^2).

Example 8.15 Construct a model for the `mario_kart` data set using the forward selection strategy.

We start with the model that includes no variables. Then we fit each of the possible models with just one variable. That is, we fit the model including just `cond_new`, then the model including just `stock_photo`, then a model with just `duration`, and a model with just `wheels`. Each of the four models provides an adjusted R^2 value:

Add ...	`cond_new`	`stock_photo`	`duration`	`wheels`
	$R^2_{adj} = 0.3459$	$R^2_{adj} = 0.0332$	$R^2_{adj} = 0.1338$	$R^2_{adj} = 0.6390$

In this first step, we compare the adjusted R^2 against a baseline model that has no predictors. The no-predictors model always has $R^2_{adj} = 0$. The model with one predictor that has the largest adjusted R^2 is the model with the `wheels` predictor, and because this adjusted R^2 is larger than the adjusted R^2 from the model with no predictors ($R^2_{adj} = 0$), we will add this variable to our model.

We repeat the process again, this time considering 2-predictor models where one of the predictors is `wheels` and with a new baseline of $R^2_{adj} = 0.6390$:

Add `wheels` and ...	`cond_new`	`stock_photo`	`duration`
	$R^2_{adj} = 0.7124$	$R^2_{adj} = 0.6587$	$R^2_{adj} = 0.6528$

The best predictor in this stage, `cond_new`, has a higher adjusted R^2 (0.7124) than the baseline (0.6390), so we also add `cond_new` to the model.

Since we have again added a variable to the model, we continue and see whether it would be beneficial to add a third variable:

Add `wheels`, `cond_new`, and ...	`stock_photo`	`duration`
	$R^2_{adj} = 0.7128$	$R^2_{adj} = 0.7107$

The model adding `stock_photo` improved adjusted R^2 (0.7124 to 0.7128), so we add `stock_photo` to the model.

Because we have again added a predictor, we check whether adding the last variable, `duration`, will improve adjusted R^2. We compare the adjusted R^2 for the model with `duration` and the other three predictors (0.7108) to the model that only considers `wheels`, `cond_new`, and `stock_photo` (0.7128). Adding `duration` does not improve the adjusted R^2, so we do not add it to the model, and we have arrived at the same model that we identified from backward elimination.

Model selection strategies

Backward elimination begins with the largest model and eliminates variables one-by-one until we are satisfied that all remaining variables are important to the model. Forward selection starts with no variables included in the model, then it adds in variables according to their importance until no other important variables are found.

There is no guarantee that backward elimination and forward selection will arrive at the same final model. If both techniques are tried and they arrive at different models, we choose the model with the larger R^2_{adj}; other tie-break options exist but are beyond the scope of this book.

8.2.3 The p-value approach, an alternative to adjusted R^2

The p-value may be used as an alternative to adjusted R^2 for model selection.

In backward elimination, we would identify the predictor corresponding to the largest p-value. If the p-value is above the significance level, usually $\alpha = 0.05$, then we would drop that variable, refit the model, and repeat the process. If the largest p-value is less than $\alpha = 0.05$, then we would not eliminate any predictors and the current model would be our best-fitting model.

In forward selection with p-values, we reverse the process. We begin with a model that has no predictors, then we fit a model for each possible predictor, identifying the model where the corresponding predictor's p-value is smallest. If that p-value is smaller than $\alpha = 0.05$, we add it to the model and repeat the process, considering whether to add more variables one-at-a-time. When none of the remaining predictors can be added to the model and have a p-value less than 0.05, then we stop adding variables and the current model would be our best-fitting model.

⊙ **Guided Practice 8.16** Examine Table 8.7 on page 378, which considers the model including the `cond_new`, `stock_photo`, and `wheels` predictors. If we were using the p-value approach with backward elimination and we were considering this model, which of these three variables would be up for elimination? Would we drop that variable, or would we keep it in the model?[11]

While the adjusted R^2 and p-value approaches are similar, they sometimes lead to different models, with the adjusted R^2 approach tending to include more predictors in the final model. For example, if we had used the p-value approach with the auction data, we would not have included the `stock_photo` predictor in the final model.

When to use the adjusted R^2 and when to use the p-value approach

When the sole goal is to improve prediction accuracy, use adjusted R^2. This is commonly the case in machine learning applications.

When we care about understanding which variables are statistically significant predictors of the response, or if there is interest in producing a simpler model at the potential cost of a little prediction accuracy, then the p-value approach is preferred.

Regardless of whether you use adjusted R^2 or the p-value approach, or if you use the backward elimination of forward selection strategy, our job is not done after variable selection. We must still verify the model conditions are reasonable.

[11] The `stock_photo` predictor is up for elimination since it has the largest p-value. Additionally, since that p-value is larger than 0.05, we would in fact eliminate `stock_photo` from the model.

8.3 Checking model assumptions using graphs

Multiple regression methods using the model

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

generally depend on the following four assumptions:

1. the residuals of the model are nearly normal,
2. the variability of the residuals is nearly constant,
3. the residuals are independent, and
4. each variable is linearly related to the outcome.

Diagnostic plots can be used to check each of these assumptions. We will consider the model from the Mario Kart auction data, and check whether there are any notable concerns:

$$\widehat{price} = 36.05 + 5.18 \times \texttt{cond\_new} + 1.12 \times \texttt{stock\_photo} + 7.30 \times \texttt{wheels}$$

Normal probability plot. A normal probability plot of the residuals is shown in Figure 8.8. While the plot exhibits some minor irregularities, there are no outliers that might be cause for concern. In a normal probability plot for residuals, we tend to be most worried about residuals that appear to be outliers, since these indicate long tails in the distribution of residuals.

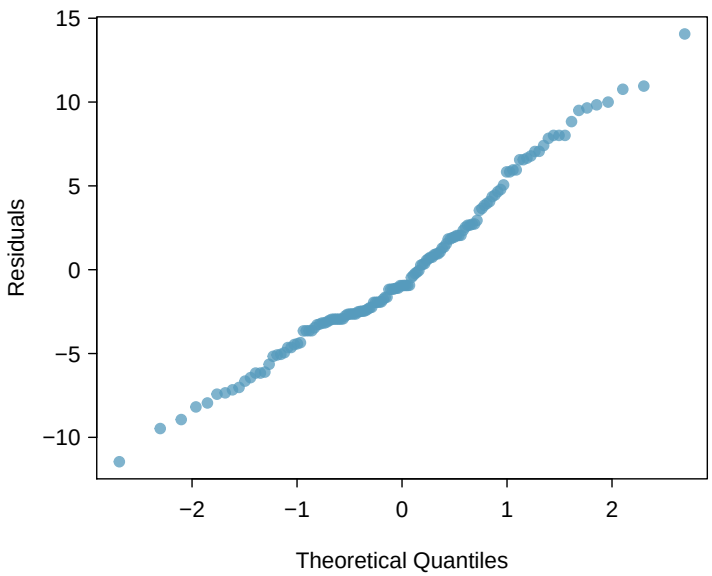


Figure 8.8: A normal probability plot of the residuals is helpful in identifying observations that might be outliers.

Absolute values of residuals against fitted values. A plot of the absolute value of the residuals against their corresponding fitted values ($\hat{y}_i$) is shown in Figure 8.9. This plot is helpful to check the condition that the variance of the residuals is approximately constant. We don't see any obvious deviations from constant variance in this example.

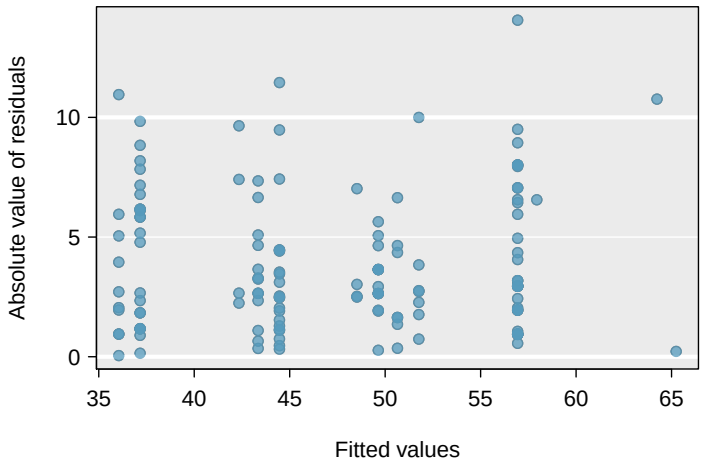


Figure 8.9: Comparing the absolute value of the residuals against the fitted values ($\hat{y}_i$) is helpful in identifying deviations from the constant variance assumption.

Residuals in order of their data collection. A plot of the residuals in the order their corresponding auctions were observed is shown in Figure 8.10. Such a plot is helpful in identifying any connection between cases that are close to one another, e.g. we could look for declining prices over time or if there was a time of the day when auctions tended to fetch a higher price. Here we see no structure that indicates a problem.[12]

Residuals against each predictor variable. We consider a plot of the residuals against the `cond_new` variable, the residuals against the `stock_photo` variable, and the residuals against the `wheels` variable. These plots are shown in Figure 8.11. For the two-level condition variable, we are guaranteed not to see any remaining trend, and instead we are checking that the variability doesn't fluctuate across groups, which it does not. However, looking at the stock photo variable, we find that there is some difference in the variability of the residuals in the two groups. Additionally, when we consider the residuals against the `wheels` variable, we see some possible structure. There appears to be curvature in the residuals, indicating the relationship is probably not linear.

It is necessary to summarize diagnostics for any model fit. If the diagnostics support the model assumptions, this would improve credibility in the findings. If the diagnostic assessment shows remaining underlying structure in the residuals, we should try to adjust the model to account for that structure. If we are unable to do so, we may still report the model but must also note its shortcomings. In the case of the auction data, we report that there appears to be non-constant variance in the stock photo variable and that there may be a nonlinear relationship between the total price and the number of wheels included for an auction. This information would be important to buyers and sellers who may review the analysis, and omitting this information could be a setback to the very people who the model might assist.

[12] An especially rigorous check would use **time series** methods. For instance, we could check whether consecutive residuals are correlated. Doing so with these residuals yields no statistically significant correlations.

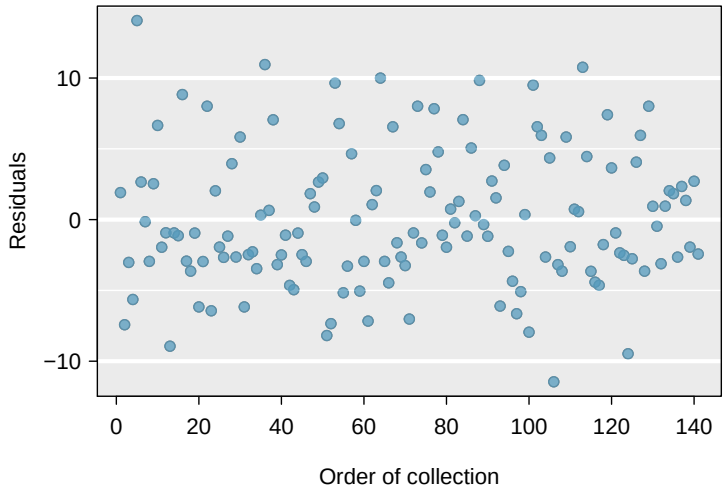


Figure 8.10: Plotting residuals in the order that their corresponding observations were collected helps identify connections between successive observations. If it seems that consecutive observations tend to be close to each other, this indicates the independence assumption of the observations would fail.

"All models are wrong, but some are useful" -George E.P. Box

The truth is that no model is perfect. However, even imperfect models can be useful. Reporting a flawed model can be reasonable so long as we are clear and report the model's shortcomings.

Caution: Don't report results when assumptions are grossly violated

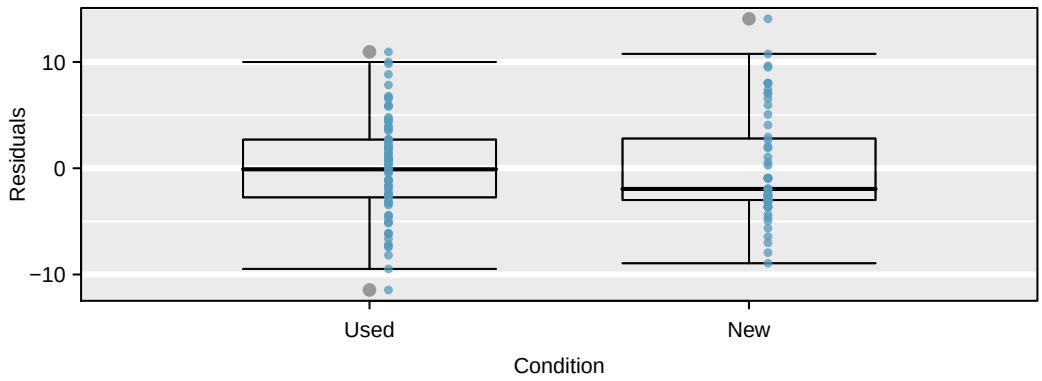
While there is a little leeway in model assumptions, don't go too far. If model assumptions are very clearly violated, consider a new model, even if it means learning more statistical methods or hiring someone who can help.

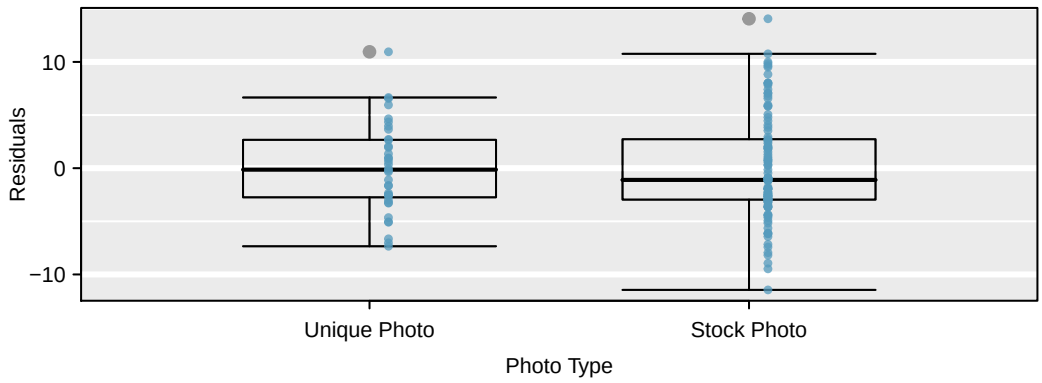
TIP: Confidence intervals in multiple regression

Confidence intervals for coefficients in multiple regression can be computed using the same formula as in the single predictor model:

$$b_i \pm t^\star_{df} SE_{b_i}$$

where $t^\star_{df}$ is the appropriate t-value corresponding to the confidence level and model degrees of freedom, $df = n - k - 1$.





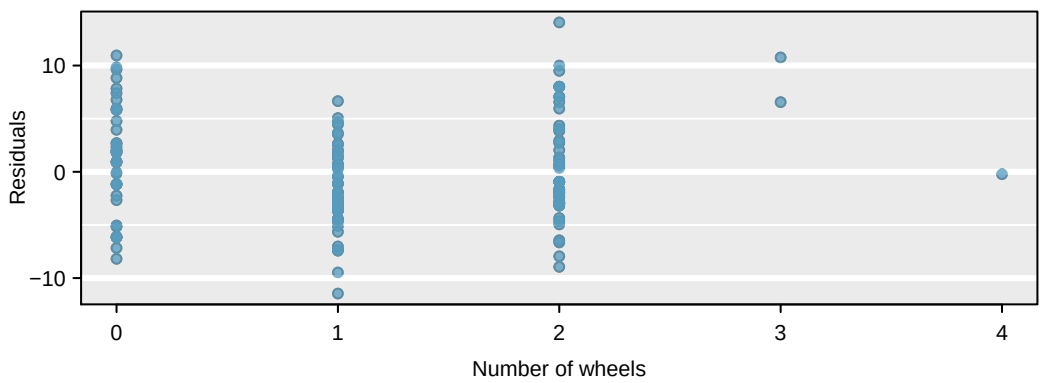


Figure 8.11: For the condition and stock photo variables, we check for differences in the distribution shape or variability of the residuals. In the case of the stock photos variable, we see a little less variability in the unique photo group than the stock photo group. For numerical predictors, we also check for trends or other structure. We see some slight bowing in the residuals against the `wheels` variable in the bottom plot.

8.4 Introduction to logistic regression

In this section we introduce **logistic regression** as a tool for building models when there is a categorical response variable with two levels. Logistic regression is a type of **generalized linear model** (GLM) for response variables where regular multiple regression does not work very well. In particular, the response variable in these settings often takes a form where residuals look completely different from the normal distribution.

GLMs can be thought of as a two-stage modeling approach. We first model the response variable using a probability distribution, such as the binomial or Poisson distribution. Second, we model the parameter of the distribution using a collection of predictors and a special form of multiple regression.

In Section 8.4 we will revisit the `email` data set from Chapter 1. These emails were collected from a single email account, and we will work on developing a basic spam filter using these data. The response variable, `spam`, has been encoded to take value 0 when a message is not spam and 1 when it is spam. Our task will be to build an appropriate model that classifies messages as spam or not spam using email characteristics coded as predictor variables. While this model will not be the same as those used in large-scale spam filters, it shares many of the same features.

8.4.1 Email data

The `email` data set was first presented in Chapter 1 with a relatively small number of variables. In fact, there are many more variables available that might be useful for classifying spam. Descriptions of these variables are presented in Table 8.12. The `spam` variable will be the outcome, and the other 10 variables will be the model predictors. While we have limited the predictors used in this section to be categorical variables (where many are represented as indicator variables), numerical predictors may also be used in logistic regression. See the footnote for an additional discussion on this topic.[13]

8.4.2 Modeling the probability of an event

TIP: Notation for a logistic regression model

The outcome variable for a GLM is denoted by Y_i, where the index i is used to represent observation i. In the email application, Y_i will be used to represent whether email i is spam ($Y_i = 1$) or not ($Y_i = 0$).

The predictor variables are represented as follows: $x_{1,i}$ is the value of variable 1 for observation i, $x_{2,i}$ is the value of variable 2 for observation i, and so on.

Logistic regression is a generalized linear model where the outcome is a two-level categorical variable. The outcome, Y_i, takes the value 1 (in our application, this represents a spam message) with probability p_i and the value 0 with probability $1 - p_i$. It is the probability p_i that we model in relation to the predictor variables.

[13] Recall from Chapter 7 that if outliers are present in predictor variables, the corresponding observations may be especially influential on the resulting model. This is the motivation for omitting the numerical variables, such as the number of characters and line breaks in emails, that we saw in Chapter 1. These variables exhibited extreme skew. We could resolve this issue by transforming these variables (e.g. using a log-transformation), but we will omit this further investigation for brevity.

variable	description
`spam`	Specifies whether the message was spam.
`to_multiple`	An indicator variable for if more than one person was listed in the *To* field of the email.
`cc`	An indicator for if someone was CCed on the email.
`attach`	An indicator for if there was an attachment, such as a document or image.
`dollar`	An indicator for if the word "dollar" or dollar symbol ($) appeared in the email.
`winner`	An indicator for if the word "winner" appeared in the email message.
`inherit`	An indicator for if the word "inherit" (or a variation, like "inheritance") appeared in the email.
`password`	An indicator for if the word "password" was present in the email.
`format`	Indicates if the email contained special formatting, such as bolding, tables, or links
`re_subj`	Indicates whether "Re:" was included at the start of the email subject.
`exclaim_subj`	Indicates whether any exclamation point was included in the email subject.

Table 8.12: Descriptions for 11 variables in the `email` data set. Notice that all of the variables are indicator variables, which take the value 1 if the specified characteristic is present and 0 otherwise.

The logistic regression model relates the probability an email is spam (p_i) to the predictors $x_{1,i}$, $x_{2,i}$, ..., $x_{k,i}$ through a framework much like that of multiple regression:

$$transformation(p_i) = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \cdots \beta_k x_{k,i} \tag{8.17}$$

We want to choose a transformation in Equation (8.17) that makes practical and mathematical sense. For example, we want a transformation that makes the range of possibilities on the left hand side of Equation (8.17) equal to the range of possibilities for the right hand side; if there was no transformation for this equation, the left hand side could only take values between 0 and 1, but the right hand side could take values outside of this range. A common transformation for p_i is the **logit transformation**, which may be written as

$$logit(p_i) = \log_e \left(\frac{p_i}{1 - p_i} \right)$$

The logit transformation is shown in Figure 8.13. Below, we rewrite Equation (8.17) using the logit transformation of p_i:

$$\log_e \left(\frac{p_i}{1 - p_i} \right) = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \cdots + \beta_k x_{k,i}$$

In our spam example, there are 10 predictor variables, so $k = 10$. This model isn't very intuitive, but it still has some resemblance to multiple regression, and we can fit this model using software. In fact, once we look at results from software, it will start to feel like we're back in multiple regression, even if the interpretation of the coefficients is more complex.

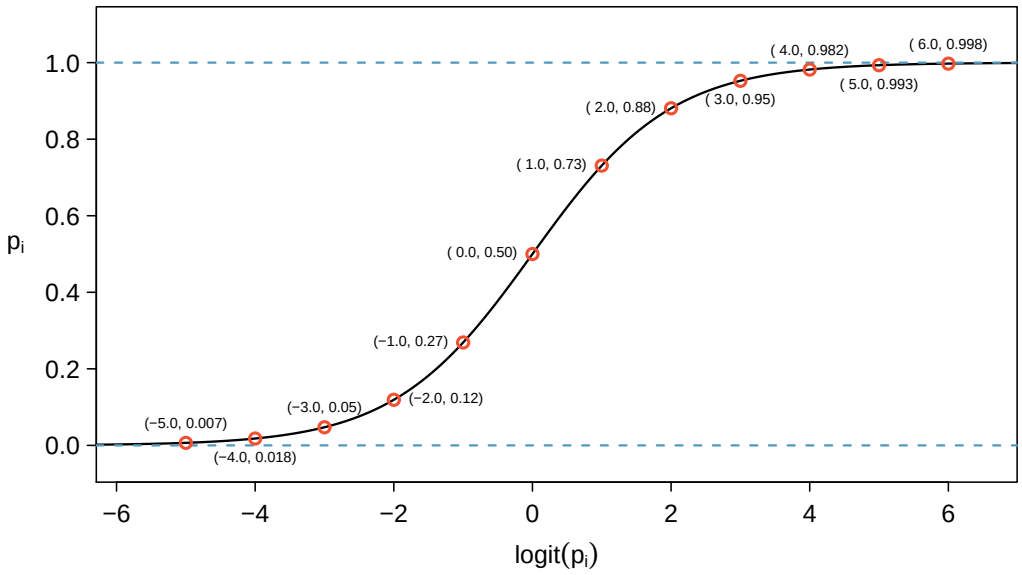


Figure 8.13: Values of p_i against values of $logit(p_i)$.

Example 8.18 Here we create a spam filter with a single predictor: `to_multiple`. This variable indicates whether more than one email address was listed in the *To* field of the email. The following logistic regression model was fit using statistical software:

$$\log\left(\frac{p_i}{1-p_i}\right) = -2.12 - 1.81 \times \texttt{to\_multiple}$$

If an email is randomly selected and it has just one address in the *To* field, what is the probability it is spam? What if more than one address is listed in the *To* field?

If there is only one email in the *To* field, then `to_multiple` takes value 0 and the right side of the model equation equals -2.12. Solving for p_i: $\frac{e^{-2.12}}{1+e^{-2.12}} = 0.11$. Just as we labeled a fitted value of y_i with a "hat" in single-variable and multiple regression, we will do the same for this probability: $\hat{p}_i = 0.11$.

If there is more than one address listed in the *To* field, then the right side of the model equation is $-2.12 - 1.81 \times 1 = -3.93$, which corresponds to a probability $\hat{p}_i = 0.02$.

Notice that we could examine -2.12 and -3.93 in Figure 8.13 to estimate the probability before formally calculating the value.

To convert from values on the regression-scale (e.g. -2.12 and -3.93 in Example 8.18), use the following formula, which is the result of solving for p_i in the regression model:

$$p_i = \frac{e^{\beta_0+\beta_1 x_{1,i}+\cdots+\beta_k x_{k,i}}}{1 \ + \ e^{\beta_0+\beta_1 x_{1,i}+\cdots+\beta_k x_{k,i}}}$$

As with most applied data problems, we substitute the point estimates for the parameters (the β_i) so that we may make use of this formula. In Example 8.18, the probabilities were calculated as

$$\frac{e^{-2.12}}{1 \ + \ e^{-2.12}} = 0.11 \qquad\qquad \frac{e^{-2.12-1.81}}{1 \ + \ e^{-2.12-1.81}} = 0.02$$

While the information about whether the email is addressed to multiple people is a helpful start in classifying email as spam or not, the probabilities of 11% and 2% are not dramatically different, and neither provides very strong evidence about which particular email messages are spam. To get more precise estimates, we'll need to include many more variables in the model.

We used statistical software to fit the logistic regression model with all ten predictors described in Table 8.12. Like multiple regression, the result may be presented in a summary table, which is shown in Table 8.14. The structure of this table is almost identical to that of multiple regression; the only notable difference is that the p-values are calculated using the normal distribution rather than the t-distribution.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.8362	0.0962	-8.69	0.0000
to_multiple	-2.8836	0.3121	-9.24	0.0000
winner	1.7038	0.3254	5.24	0.0000
format	-1.5902	0.1239	-12.84	0.0000
re_subj	-2.9082	0.3708	-7.84	0.0000
exclaim_subj	0.1355	0.2268	0.60	0.5503
cc	-0.4863	0.3054	-1.59	0.1113
attach	0.9790	0.2170	4.51	0.0000
dollar	-0.0582	0.1589	-0.37	0.7144
inherit	0.2093	0.3197	0.65	0.5127
password	-1.4929	0.5295	-2.82	0.0048

Table 8.14: Summary table for the full logistic regression model for the spam filter example.

Just like multiple regression, we could trim some variables from the model using the p-value. Using backward elimination with a p-value cutoff of 0.05 (start with the full model and trim the predictors with p-values greater than 0.05), we ultimately eliminate the `exclaim_subj`, `dollar`, `inherit`, and `cc` predictors. The remainder of this section will rely on this smaller model, which is summarized in Table 8.15.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.8595	0.0910	-9.44	0.0000
to_multiple	-2.8372	0.3092	-9.18	0.0000
winner	1.7370	0.3218	5.40	0.0000
format	-1.5569	0.1207	-12.90	0.0000
re_subj	-3.0482	0.3630	-8.40	0.0000
attach	0.8643	0.2042	4.23	0.0000
password	-1.4871	0.5290	-2.81	0.0049

Table 8.15: Summary table for the logistic regression model for the spam filter, where variable selection has been performed.

⊙ **Guided Practice 8.19** Examine the summary of the reduced model in Table 8.15, and in particular, examine the `to_multiple` row. Is the point estimate the same as we found before, -1.81, or is it different? Explain why this might be.[14]

Point estimates will generally change a little – and sometimes a lot – depending on which other variables are included in the model. This is usually due to colinearity in the predictor variables. We previously saw this in the Ebay auction example when we compared the coefficient of `cond_new` in a single-variable model and the corresponding coefficient in the multiple regression model that used three additional variables (see Sections 8.1.1 and 8.1.2).

● **Example 8.20** Spam filters are built to be automated, meaning a piece of software is written to collect information about emails as they arrive, and this information is put in the form of variables. These variables are then put into an algorithm that uses a statistical model, like the one we've fit, to classify the email. Suppose we write software for a spam filter using the reduced model shown in Table 8.15. If an incoming email has the word "winner" in it, will this raise or lower the model's calculated probability that the incoming email is spam?

The estimated coefficient of `winner` is positive (1.7370). A positive coefficient estimate in logistic regression, just like in multiple regression, corresponds to a positive association between the predictor and response variables when accounting for the other variables in the model. Since the response variable takes value 1 if an email is spam and 0 otherwise, the positive coefficient indicates that the presence of "winner" in an email raises the model probability that the message is spam.

● **Example 8.21** Suppose the same email from Example 8.20 was in HTML format, meaning the `format` variable took value 1. Does this characteristic increase or decrease the probability that the email is spam according to the model?

Since HTML corresponds to a value of 1 in the `format` variable and the coefficient of this variable is negative (-1.5569), this would lower the probability estimate returned from the model.

8.4.3 Practical decisions in the email application

Examples 8.20 and 8.21 highlight a key feature of logistic and multiple regression. In the spam filter example, some email characteristics will push an email's classification in the direction of spam while other characteristics will push it in the opposite direction.

If we were to implement a spam filter using the model we have fit, then each future email we analyze would fall into one of three categories based on the email's characteristics:

1. The email characteristics generally indicate the email is not spam, and so the resulting probability that the email is spam is quite low, say, under 0.05.
2. The characteristics generally indicate the email is spam, and so the resulting probability that the email is spam is quite large, say, over 0.95.
3. The characteristics roughly balance each other out in terms of evidence for and against the message being classified as spam. Its probability falls in the remaining range, meaning the email cannot be adequately classified as spam or not spam.

[14] The new estimate is different: -2.87. This new value represents the estimated coefficient when we are also accounting for other variables in the logistic regression model.

If we were managing an email service, we would have to think about what should be done in each of these three instances. In an email application, there are usually just two possibilities: filter the email out from the regular inbox and put it in a "spambox", or let the email go to the regular inbox.

⊙ **Guided Practice 8.22** The first and second scenarios are intuitive. If the evidence strongly suggests a message is not spam, send it to the inbox. If the evidence strongly suggests the message is spam, send it to the spambox. How should we handle emails in the third category?[15]

⊙ **Guided Practice 8.23** Suppose we apply the logistic model we have built as a spam filter and that 100 messages are placed in the spambox over 3 months. If we used the guidelines above for putting messages into the spambox, about how many legitimate (non-spam) messages would you expect to find among the 100 messages?[16]

Almost any classifier will have some error. In the spam filter guidelines above, we have decided that it is okay to allow up to 5% of the messages in the spambox to be real messages. If we wanted to make it a little harder to classify messages as spam, we could use a cutoff of 0.99. This would have two effects. Because it raises the standard for what can be classified as spam, it reduces the number of good emails that are classified as spam. However, it will also fail to correctly classify an increased fraction of spam messages. No matter the complexity and the confidence we might have in our model, these practical considerations are absolutely crucial to making a helpful spam filter. Without them, we could actually do more harm than good by using our statistical model.

8.4.4 Diagnostics for the email classifier

Logistic regression conditions

There are two key conditions for fitting a logistic regression model:

1. Each predictor x_i is linearly related to $\text{logit}(p_i)$ if all other predictors are held constant.
2. Each outcome Y_i is independent of the other outcomes.

The first condition of the logistic regression model is not easily checked without a fairly sizable amount of data. Luckily, we have 3,921 emails in our data set! Let's first visualize these data by plotting the true classification of the emails against the model's fitted probabilities, as shown in Figure 8.16. The vast majority of emails (spam or not) still have fitted probabilities below 0.5.

This may at first seem very discouraging: we have fit a logistic model to create a spam filter, but no emails have a fitted probability of being spam above 0.75. Don't despair; we will discuss ways to improve the model through the use of better variables in Section 8.4.5.

[15] In this particular application, we should err on the side of sending more mail to the inbox rather than mistakenly putting good messages in the spambox. So, in summary: emails in the first and last categories go to the regular inbox, and those in the second scenario go to the spambox.

[16] First, note that we proposed a cutoff for the predicted probability of 0.95 for spam. In a worst case scenario, all the messages in the spambox had the minimum probability equal to about 0.95. Thus, we should expect to find about 5 or fewer legitimate messages among the 100 messages placed in the spambox.

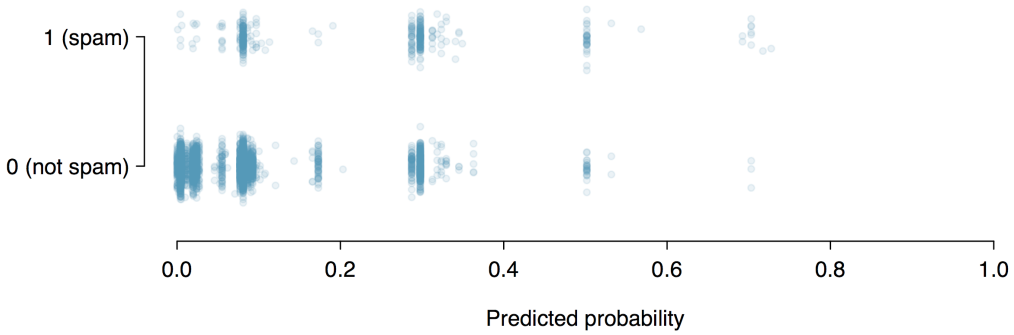


Figure 8.16: The predicted probability that each of the 3,912 emails is spam is classified by their grouping, spam or not. Noise (small, random vertical shifts) have been added to each point so that points with nearly identical values aren't plotted exactly on top of one another. This makes it possible to see more observations.

We'd like to assess the quality of our model. For example, we might ask: if we look at emails that we modeled as having a 10% chance of being spam, do we find about 10% of them actually are spam? To help us out, we'll borrow an advanced statistical method called **natural splines** that estimates the local probability over the region 0.00 to 0.75 (the largest predicted probability was 0.73, so we avoid extrapolating). All you need to know about natural splines to understand what we are doing is that they are used to fit flexible lines rather than straight lines.

The curve fit using natural splines is shown in Figure 8.17 as a solid black line. If the logistic model fits well, the curve should closely follow the dashed $y = x$ line. We have added shading to represent the confidence bound for the curved line to clarify what fluctuations might plausibly be due to chance. Even with this confidence bound, there are weaknesses in the first model assumption. The solid curve and its confidence bound dips below the dashed line from about 0.1 to 0.3, and then it drifts above the dashed line from about 0.35 to 0.55. These deviations indicate the model relating the parameter to the predictors does not closely resemble the true relationship.

We could evaluate the second logistic regression model assumption – independence of the outcomes – using the model residuals. The residuals for a logistic regression model are calculated the same way as with multiple regression: the observed outcome minus the expected outcome. For logistic regression, the expected value of the outcome is the fitted probability for the observation, and the residual may be written as

$$e_i = Y_i - \hat{p}_i$$

We could plot these residuals against a variety of variables or in their order of collection, as we did with the residuals in multiple regression. However, since the model will need to be revised to effectively classify spam and you have already seen similar residual plots in Section 8.3, we won't investigate the residuals here.

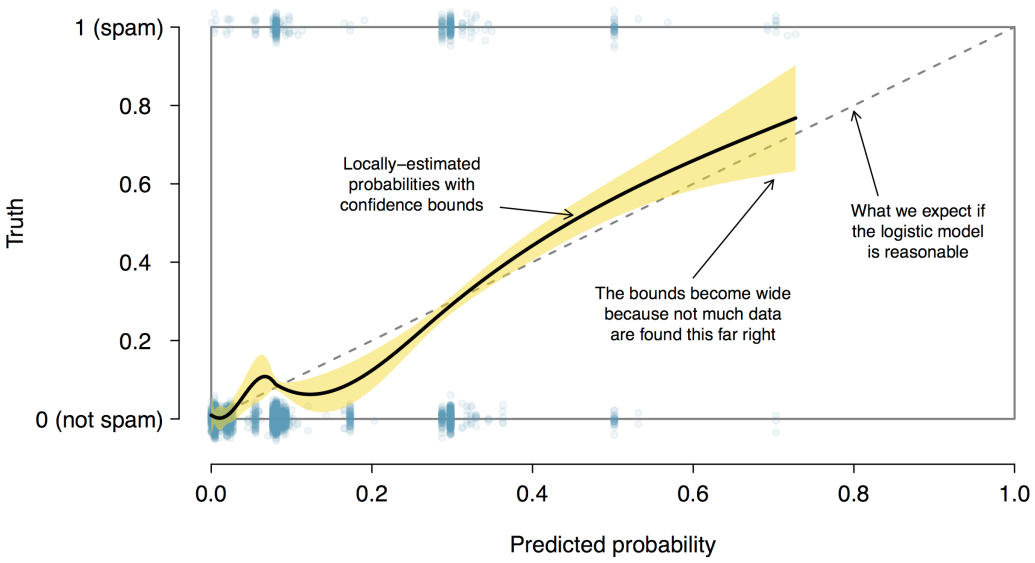


Figure 8.17: The solid black line provides the empirical estimate of the probability for observations based on their predicted probabilities (confidence bounds are also shown for this line), which is fit using natural splines. A small amount of noise was added to the observations in the plot to allow more observations to be seen.

8.4.5 Improving the set of variables for a spam filter

If we were building a spam filter for an email service that managed many accounts (e.g. Gmail or Hotmail), we would spend much more time thinking about additional variables that could be useful in classifying emails as spam or not. We also would use transformations or other techniques that would help us include strongly skewed numerical variables as predictors.

Take a few minutes to think about additional variables that might be useful in identifying spam. Below is a list of variables we think might be useful:

(1) An indicator variable could be used to represent whether there was prior two-way correspondence with a message's sender. For instance, if you sent a message to john@example.com and then John sent you an email, this variable would take value 1 for the email that John sent. If you had never sent John an email, then the variable would be set to 0.

(2) A second indicator variable could utilize an account's past spam flagging information. The variable could take value 1 if the sender of the message has previously sent messages flagged as spam.

(3) A third indicator variable could flag emails that contain links included in previous spam messages. If such a link is found, then set the variable to 1 for the email. Otherwise, set it to 0.

The variables described above take one of two approaches. Variable (1) is specially designed to capitalize on the fact that spam is rarely sent between individuals that have two-way

communication. Variables (2) and (3) are specially designed to flag common spammers or spam messages. While we would have to verify using the data that each of the variables is effective, these seem like promising ideas.

Table 8.18 shows a contingency table for spam and also for the new variable described in (1) above. If we look at the 1,090 emails where there was correspondence with the sender in the preceding 30 days, not one of these message was spam. This suggests variable (1) would be very effective at accurately classifying some messages as not spam. With this single variable, we would be able to send about 28% of messages through to the inbox with confidence that almost none are spam.

	prior correspondence		
	`no`	`yes`	Total
`spam`	367	0	367
`not spam`	2464	1090	3554
Total	2831	1090	3921

Table 8.18: A contingency table for `spam` and a new variable that represents whether there had been correspondence with the sender in the preceding 30 days.

The variables described in (2) and (3) would provide an excellent foundation for distinguishing messages coming from known spammers or messages that take a known form of spam. To utilize these variables, we would need to build databases: one holding email addresses of known spammers, and one holding URLs found in known spam messages. Our access to such information is limited, so we cannot implement these two variables in this textbook. However, if we were hired by an email service to build a spam filter, these would be important next steps.

In addition to finding more and better predictors, we would need to create a customized logistic regression model for each email account. This may sound like an intimidating task, but its complexity is not as daunting as it may at first seem. We'll save the details for a statistics course where computer programming plays a more central role.

For what is the extremely challenging task of classifying spam messages, we have made a lot of progress. We have seen that simple email variables, such as the format, inclusion of certain words, and other circumstantial characteristics, provide helpful information for spam classification. Many challenges remain, from better understanding logistic regression to carrying out the necessary computer programming, but completing such a task is very nearly within your reach.

8.5 Exercises

8.5.1 Introduction to multiple regression

8.1 Baby weights, Part I. The Child Health and Development Studies investigate a range of topics. One study considered all pregnancies between 1960 and 1967 among women in the Kaiser Foundation Health Plan in the San Francisco East Bay area. Here, we study the relationship between smoking and weight of the baby. The variable `smoke` is coded 1 if the mother is a smoker, and 0 if not. The summary table below shows the results of a linear regression model for predicting the average birth weight of babies, measured in ounces, based on the smoking status of the mother.[17]

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	123.05	0.65	189.60	0.0000
smoke	-8.94	1.03	-8.65	0.0000

The variability within the smokers and non-smokers are about equal and the distributions are symmetric. With these conditions satisfied, it is reasonable to apply the model. (Note that we don't need to check linearity since the predictor has only two levels.)

(a) Write the equation of the regression line.

(b) Interpret the slope in this context, and calculate the predicted birth weight of babies born to smoker and non-smoker mothers.

(c) Is there a statistically significant relationship between the average birth weight and smoking?

8.2 Baby weights, Part II. Exercise 8.1 introduces a data set on birth weight of babies. Another variable we consider is `parity`, which is 0 if the child is the first born, and 1 otherwise. The summary table below shows the results of a linear regression model for predicting the average birth weight of babies, measured in ounces, from `parity`.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	120.07	0.60	199.94	0.0000
parity	-1.93	1.19	-1.62	0.1052

(a) Write the equation of the regression line.

(b) Interpret the slope in this context, and calculate the predicted birth weight of first borns and others.

(c) Is there a statistically significant relationship between the average birth weight and parity?

[17] Child Health and Development Studies, Baby weights data set.

8.3 Baby weights, Part III. We considered the variables `smoke` and `parity`, one at a time, in modeling birth weights of babies in Exercises 8.1 and 8.2. A more realistic approach to modeling infant weights is to consider all possibly related variables at once. Other variables of interest include length of pregnancy in days (`gestation`), mother's age in years (`age`), mother's height in inches (`height`), and mother's pregnancy weight in pounds (`weight`). Below are three observations from this data set.

	bwt	gestation	parity	age	height	weight	smoke
1	120	284	0	27	62	100	0
2	113	282	0	33	64	135	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
1236	117	297	0	38	65	129	0

The summary table below shows the results of a regression model for predicting the average birth weight of babies based on all of the variables included in the data set.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-80.41	14.35	-5.60	0.0000
gestation	0.44	0.03	15.26	0.0000
parity	-3.33	1.13	-2.95	0.0033
age	-0.01	0.09	-0.10	0.9170
height	1.15	0.21	5.63	0.0000
weight	0.05	0.03	1.99	0.0471
smoke	-8.40	0.95	-8.81	0.0000

(a) Write the equation of the regression line that includes all of the variables.

(b) Interpret the slopes of `gestation` and `age` in this context.

(c) The coefficient for `parity` is different than in the linear model shown in Exercise 8.2. Why might there be a difference?

(d) Calculate the residual for the first observation in the data set.

(e) The variance of the residuals is 249.28, and the variance of the birth weights of all babies in the data set is 332.57. Calculate the R^2 and the adjusted R^2. Note that there are 1,236 observations in the data set.

8.4 Absenteeism, Part I. Researchers interested in the relationship between absenteeism from school and certain demographic characteristics of children collected data from 146 randomly sampled students in rural New South Wales, Australia, in a particular school year. Below are three observations from this data set.

	eth	sex	lrn	days
1	0	1	1	2
2	0	1	1	11
⋮	⋮	⋮	⋮	⋮
146	1	0	0	37

The summary table below shows the results of a linear regression model for predicting the average number of days absent based on ethnic background (`eth`: 0 - aboriginal, 1 - not aboriginal), sex (`sex`: 0 - female, 1 - male), and learner status (`lrn`: 0 - average learner, 1 - slow learner).[18]

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	18.93	2.57	7.37	0.0000
eth	-9.11	2.60	-3.51	0.0000
sex	3.10	2.64	1.18	0.2411
lrn	2.15	2.65	0.81	0.4177

(a) Write the equation of the regression line.

(b) Interpret each one of the slopes in this context.

(c) Calculate the residual for the first observation in the data set: a student who is aboriginal, male, a slow learner, and missed 2 days of school.

(d) The variance of the residuals is 240.57, and the variance of the number of absent days for all students in the data set is 264.17. Calculate the R^2 and the adjusted R^2. Note that there are 146 observations in the data set.

8.5 GPA. A survey of 55 Duke University students asked about their GPA, number of hours they study at night, number of nights they go out, and their gender. Summary output of the regression model is shown below. Note that male is coded as 1.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	3.45	0.35	9.85	0.00
studyweek	0.00	0.00	0.27	0.79
sleepnight	0.01	0.05	0.11	0.91
outnight	0.05	0.05	1.01	0.32
gender	-0.08	0.12	-0.68	0.50

(a) Calculate a 95% confidence interval for the coefficient of gender in the model, and interpret it in the context of the data.

(b) Would you expect a 95% confidence interval for the slope of the remaining variables to include 0? Explain

[18] W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S*. Fourth Edition. Data can also be found in the R MASS package. New York: Springer, 2002.

8.6 Cherry trees. Timber yield is approximately equal to the volume of a tree, however, this value is difficult to measure without first cutting the tree down. Instead, other variables, such as height and diameter, may be used to predict a tree's volume and yield. Researchers wanting to understand the relationship between these variables for black cherry trees collected data from 31 such trees in the Allegheny National Forest, Pennsylvania. Height is measured in feet, diameter in inches (at 54 inches above ground), and volume in cubic feet.[19]

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-57.99	8.64	-6.71	0.00
height	0.34	0.13	2.61	0.01
diameter	4.71	0.26	17.82	0.00

(a) Calculate a 95% confidence interval for the coefficient of height, and interpret it in the context of the data.

(b) One tree in this sample is 79 feet tall, has a diameter of 11.3 inches, and is 24.2 cubic feet in volume. Determine if the model overestimates or underestimates the volume of this tree, and by how much.

8.5.2 Model selection

8.7 Baby weights, Part IV. Exercise 8.3 considers a model that predicts a newborn's weight using several predictors (gestation length, parity, age of mother, height of mother, weight of mother, smoking status of mother). The table below shows the adjusted R-squared for the full model as well as adjusted R-squared values for all models we evaluate in the first step of the backwards elimination process.

	Model	Adjusted R^2
1	Full model	0.2541
2	No gestation	0.1031
3	No parity	0.2492
4	No age	0.2547
5	No height	0.2311
6	No weight	0.2536
7	No smoking status	0.2072

Which, if any, variable should be removed from the model first?

[19] D.J. Hand. *A handbook of small data sets.* Chapman & Hall/CRC, 1994.

8.8 Absenteeism, Part II. Exercise 8.4 considers a model that predicts the number of days absent using three predictors: ethnic background (`eth`), gender (`sex`), and learner status (`lrn`). The table below shows the adjusted R-squared for the model as well as adjusted R-squared values for all models we evaluate in the first step of the backwards elimination process.

	Model	Adjusted R^2
1	Full model	0.0701
2	No ethnicity	-0.0033
3	No sex	0.0676
4	No learner status	0.0723

Which, if any, variable should be removed from the model first?

8.9 Baby weights, Part V. Exercise 8.3 provides regression output for the full model (including all explanatory variables available in the data set) for predicting birth weight of babies. In this exercise we consider a forward-selection algorithm and add variables to the model one-at-a-time. The table below shows the p-value and adjusted R^2 of each model where we include only the corresponding predictor. Based on this table, which variable should be added to the model first?

variable	gestation	parity	age	height	weight	smoke
p-value	2.2×10^{-16}	0.1052	0.2375	2.97×10^{-12}	8.2×10^{-8}	2.2×10^{-16}
R^2_{adj}	0.1657	0.0013	0.0003	0.0386	0.0229	0.0569

8.10 Absenteeism, Part III. Exercise 8.4 provides regression output for the full model, including all explanatory variables available in the data set, for predicting the number of days absent from school. In this exercise we consider a forward-selection algorithm and add variables to the model one-at-a-time. The table below shows the p-value and adjusted R^2 of each model where we include only the corresponding predictor. Based on this table, which variable should be added to the model first?

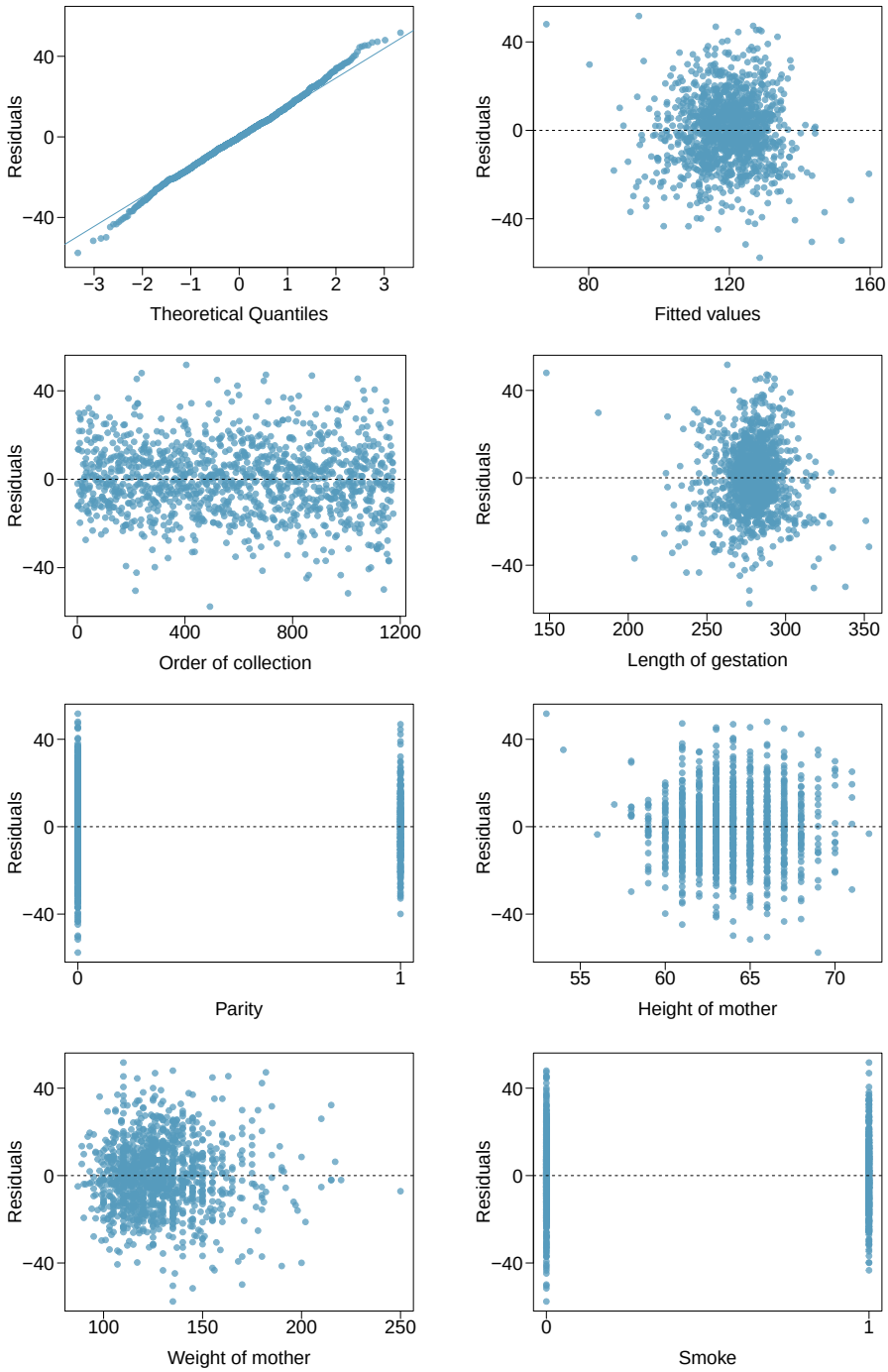
variable	ethnicity	sex	learner status
p-value	0.0007	0.3142	0.5870
R^2_{adj}	0.0714	0.0001	0

8.11 Movie lovers, Part I. Suppose a social scientist is interested in studying what makes audiences love or hate a movie. She collects a random sample of movies (genre, length, cast, director, budget, etc.) as well as a measure of the success of the movie (score on a film review aggregator website). If as part of her research she is interested in finding out which variables are significant predictors of movie success, what type of model selection method should she use?

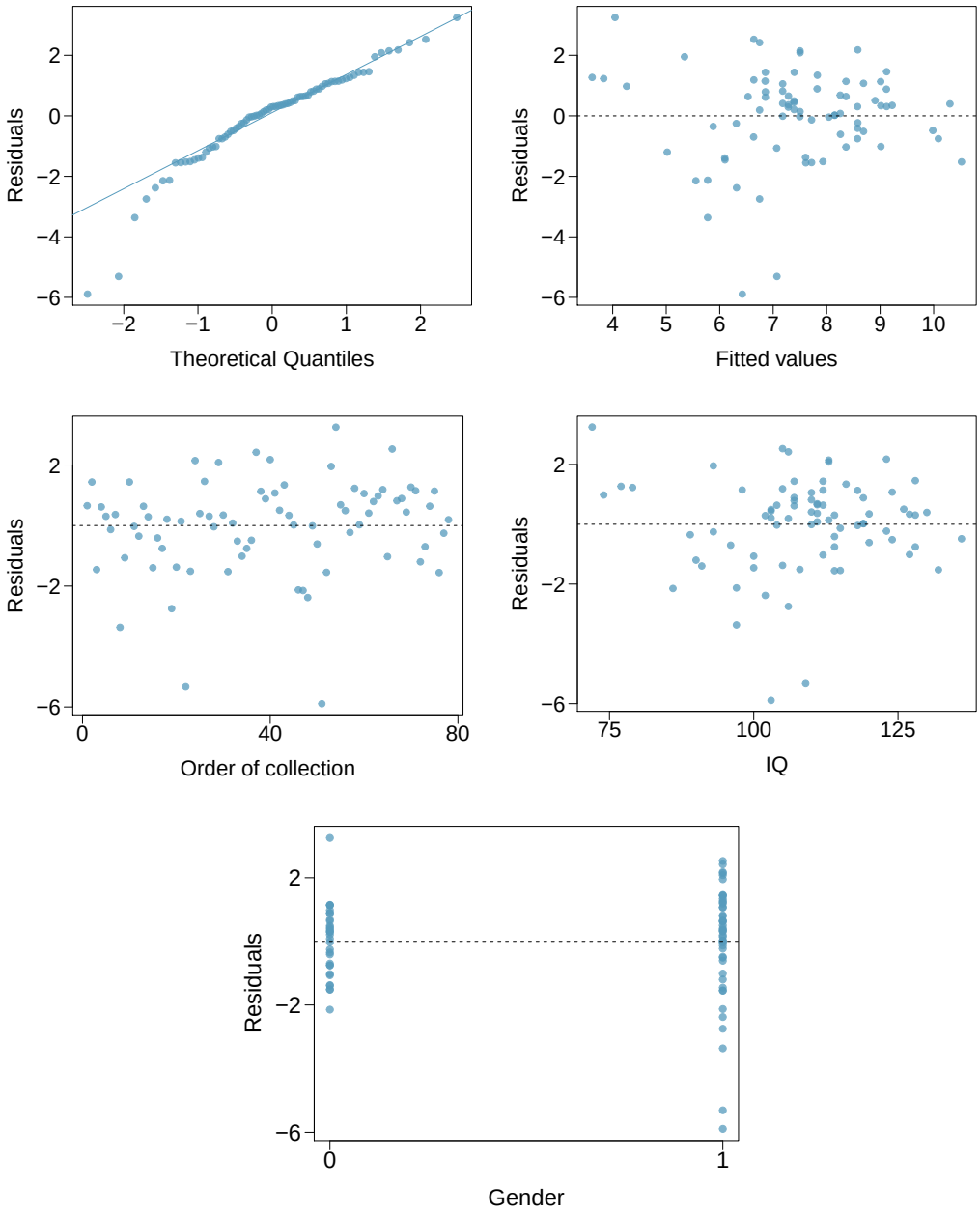
8.12 Movie lovers, Part II. Suppose an online media streaming company is interested in building a movie recommendation system. The website maintains data on the movies in their database (genre, length, cast, director, budget, etc.) and additionally collects data from their subscribers (demographic information, previously watched movies, how they rated previously watched movies, etc.). The recommendation system will be deemed successful if subscribers actually watch, and rate highly, the movies recommended to them. Should the company use the adjusted R^2 or the p-value approach in selecting variables for their recommendation system?

8.5.3 Checking model assumptions using graphs

8.13 Baby weights, Part V. Exercise 8.3 presents a regression model for predicting the average birth weight of babies based on length of gestation, parity, height, weight, and smoking status of the mother. Determine if the model assumptions are met using the plots below. If not, describe how to proceed with the analysis.



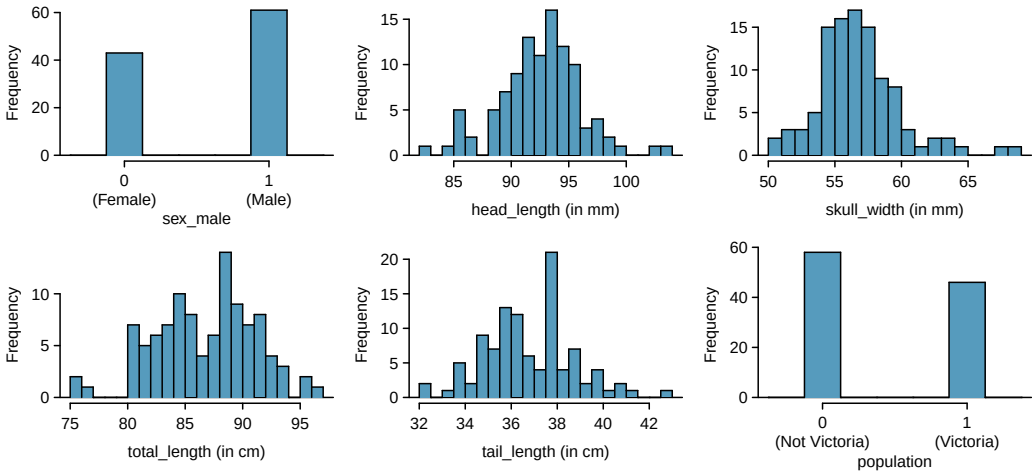
8.14 GPA and IQ. A regression model for predicting GPA from gender and IQ was fit, and both predictors were found to be statistically significant. Using the plots given below, determine if this regression model is appropriate for these data.



8.5.4 Introduction to logistic regression

8.15 Possum classification, Part I. The common brushtail possum of the Australia region is a bit cuter than its distant cousin, the American opossum (see Figure 7.5 on page 334). We consider 104 brushtail possums from two regions in Australia, where the possums may be considered a random sample from the population. The first region is Victoria, which is in the eastern half of Australia and traverses the southern coast. The second region consists of New South Wales and Queensland, which make up eastern and northeastern Australia.

We use logistic regression to differentiate between possums in these two regions. The outcome variable, called `population`, takes value 1 when a possum is from Victoria and 0 when it is from New South Wales or Queensland. We consider five predictors: `sex_male` (an indicator for a possum being male), `head_length`, `skull_width`, `total_length`, and `tail_length`. Each variable is summarized in a histogram. The full logistic regression model and a reduced model after variable selection are summarized in the table.



	Full Model				*Reduced Model*			
	Estimate	SE	Z	Pr(>|Z|)	Estimate	SE	Z	Pr(>|Z|)
(Intercept)	39.2349	11.5368	3.40	0.0007	33.5095	9.9053	3.38	0.0007
sex_male	-1.2376	0.6662	-1.86	0.0632	-1.4207	0.6457	-2.20	0.0278
head_length	-0.1601	0.1386	-1.16	0.2480				
skull_width	-0.2012	0.1327	-1.52	0.1294	-0.2787	0.1226	-2.27	0.0231
total_length	0.6488	0.1531	4.24	0.0000	0.5687	0.1322	4.30	0.0000
tail_length	-1.8708	0.3741	-5.00	0.0000	-1.8057	0.3599	-5.02	0.0000

(a) Examine each of the predictors. Are there any outliers that are likely to have a very large influence on the logistic regression model?

(b) The summary table for the full model indicates that at least one variable should be eliminated when using the p-value approach for variable selection: `head_length`. The second component of the table summarizes the reduced model following variable selection. Explain why the remaining estimates change between the two models.

8.16 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.

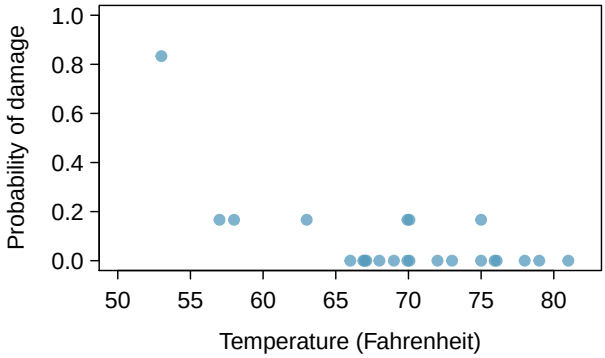
8.17 Possum classification, Part II. A logistic regression model was proposed for classifying common brushtail possums into their two regions in Exercise 8.15. The outcome variable took value 1 if the possum was from Victoria and 0 otherwise.

	Estimate	SE	Z	Pr(>|Z|)
(Intercept)	33.5095	9.9053	3.38	0.0007
sex_male	-1.4207	0.6457	-2.20	0.0278
skull_width	-0.2787	0.1226	-2.27	0.0231
total_length	0.5687	0.1322	4.30	0.0000
tail_length	-1.8057	0.3599	-5.02	0.0000

(a) Write out the form of the model. Also identify which of the variables are positively associated when controlling for other variables.

(b) Suppose we see a brushtail possum at a zoo in the US, and a sign says the possum had been captured in the wild in Australia, but it doesn't say which part of Australia. However, the sign does indicate that the possum is male, its skull is about 63 mm wide, its tail is 37 cm long, and its total length is 83 cm. What is the reduced model's computed probability that this possum is from Victoria? How confident are you in the model's accuracy of this probability calculation?

8.18 Challenger disaster, Part II. Exercise 8.16 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.

Appendix A

End of chapter exercise solutions

1 Introduction to data

1.1 (a) Treatment: $10/43 = 0.23 \rightarrow 23\%$. Control: $2/46 = 0.04 \rightarrow 4\%$. (b) There is a 19% difference between the pain reduction rates in the two groups. At first glance, it appears patients in the treatment group are more likely to experience pain reduction from the acupuncture treatment. (c) Answers may vary but should be sensible. Two possible answers: [1]Though the groups' difference is big, I'm skeptical the results show a real difference and think this might be due to chance. [2]The difference in these rates looks pretty big, so I suspect acupuncture is having a positive impact on pain.

1.3 (a) 143,196 eligible study subjects born in Southern California between 1989 and 1993. (b) 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. (c) "Is there an association between air pollution exposure and preterm births?"

1.5 (a) 160 children. (b) Age (numerical, continuous), sex (categorical), whether they were an only child or not (categorical), and whether they cheated or not (categorical). (c) Research question: "Does explicitly telling children not to cheat affect their likelihood to cheat?"

1.7 (a) $50 \times 3 = 150$. (b) Four continuous numerical variables: sepal length, sepal width, petal length, and petal width. (c) One categorical variable, species, with three levels: *setosa*, *versicolor*, and *virginica*.

1.9 (a) Population: all births, sample: 143,196 births between 1989 and 1993 in Southern California. (b) If births in this time span at the geography can be considered to be representative of all births, then the results are generalizable to the population of Southern California. However, since the study is observational the findings cannot be used to establish causal relationships.

1.11 (a) Population: all asthma patients aged 18-69 who rely on medication for asthma treatment. Sample: 600 such patients. (b) If the patients in this sample, who are likely not randomly sampled, can be considered to be representative of all asthma patients aged 18-69 who rely on medication for asthma treatment, then the results are generalizable to the population defined above. Additionally, since the study is experimental, the findings can be used to establish causal relationships.

1.13 (a) Observation. (b) Variable. (c) Sample statistic (mean). (d) Population parameter (mean).

1.15 (a) Explanatory: number of study hours per week. Response: GPA. (b) Somewhat weak positive relationship with data becoming more sparse as the number of study hours increases. One responded reported a GPA above 4.0, which is clearly a data error. There are a few respondents who reported unusually high study hours (60 and 70 hours/week). Variability in GPA is much higher for students who study less than those who study more, which might be due to the fact that there aren't many respondents who reported studying higher hours. (c) Observational. (d) Since observational, cannot infer causation.

1.17 (a) Observational. (b) Use stratified sampling to randomly sample a fixed number of students, say 10, from each section for a total sample size of 40 students.

1.19 (a) Positive, non-linear, somewhat strong. Countries in which a higher percentage of the population have access to the internet also tend to have higher average life expectancies, however rise in life expectancy trails off before around 80 years old. (b) Observational. (c) Wealth: countries with individuals who can widely afford the internet can probably also afford basic medical care. (Note: Answers may vary.)

1.21 (a) Simple random sampling is okay. In fact, it's rare for simple random sampling to not be a reasonable sampling method! (b) The student opinions may vary by field of study, so the stratifying by this variable makes sense and would be reasonable. (c) Students of similar ages are probably going to have more similar opinions, and we want clusters to be diverse with respect to the outcome of interest, so this would **not** be a good approach. (Additional thought: the clusters in this case may also have very different numbers of people, which can also create unexpected sample sizes.)

1.23 (a) The cases are 200 randomly sampled men and women. (b) The response variable is attitude towards a fictional microwave oven. (c) The explanatory variable is dispositional attitude. (d) Yes, the cases are sampled randomly. (e) This is an observational study since there is no random assignment to treatments. (f) No, we cannot establish a causal link between the explanatory and response variables since the study is observational. (g) Yes, the results of the study can be generalized to the population at large since the sample is random.

1.25 (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 status than the respondents. (c) There is no control group in this study, this is an observational study, and there may be confounding variables, e.g. these people may go running because they are generally healthier and/or do other exercises.

1.27 (a) Simple random sample. Non-response bias, if only those people who have strong opinions about the survey responds his sample may not be representative of the population. (b) Convenience sample. Under coverage bias, his sample may not be representative of the population since it consists only of his friends. It is also possible that the study will have non-response bias if some choose to not bring back the survey. (c) Convenience sample. This will have a similar issues to handing out surveys to friends. (d) Multi-stage sampling. If the classes are similar to each other with respect to student composition this approach should not introduce bias, other than potential non-response bias.

1.29 No, students were not randomly sampled (voluntary sample) and the sample only contains college students at a university in Ontario.

1.31 (a) Exam performance. (b) Light level: fluorescent overhead lighting, yellow overhead lighting, no overhead lighting (only desk lamps). (c) Sex: man, woman.

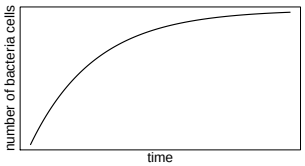
1.33 (a) Exam performance. (b) Light level (overhead lighting, yellow overhead lighting, no overhead lighting) and noise level (no noise, construction noise, and human chatter noise). (c) Since the researchers want to ensure equal gender representation, sex will be a blocking variable.

1.35 Need randomization and blinding. One possible outline: (1) Prepare two cups for each participant, one containing regular Coke and the other containing Diet Coke. Make sure the cups are identical and contain equal amounts of soda. Label the cups A (regular) and B (diet). (Be sure to randomize A and B for each trial!) (2) Give each participant the two cups, one cup at a time, in random order, and ask the participant to record a value that indicates how much she liked the beverage. Be sure that neither the participant nor the person handing out the cups knows the identity of the beverage to make this a double- blind experiment. (Answers may vary.)

1.37 (a) Experiment. (b) Treatment: 25 grams of chia seeds twice a day, control: placebo. (c) Yes, gender. (d) Yes, single blind since the patients were blinded to the treatment they received. (e) Since this is an experiment, we can make a causal statement. However, since the sample is not random, the causal statement cannot be generalized to the population at large.

1.39 (a) 1: linear. 3: nonlinear.
(b) 4: linear. (c) 2.

1.41



1.43 (a) Population mean, $\mu_{2007} = 52$; sample mean, $\bar{x}_{2008} = 58$. (b) Population mean, $\mu_{2001} = 3.37$; sample mean, $\bar{x}_{2012} = 3.59$.

1.45 Any 10 employees whose average number of days off is between the minimum and the mean number of days off for the entire workforce at this plant.

1.47 (a) Dist 2 has a higher mean since $20 > 13$, and a higher standard deviation since 20 is further from the rest of the data than 13. (b) Dist 1 has a higher mean since $-20 > -40$, and Dist 2 has a higher standard deviation since -40 is farther away from the rest of the data than -20. (c) Dist 2 has a higher mean since all values in this distribution are higher than those in Dist 1, but both distribution have the same standard deviation since they are equally variable around their respective means. (d) Both distributions have the same mean since they're both centered at 300, but Dist 2 has a higher standard deviation since the observations are farther from the mean than in Dist 1.

1.49 (a) Q1 $\approx$ 5, median $\approx$ 15, Q3 $\approx$ 35 (b) Since the distribution is right skewed, we would expect the mean to be higher than the median.

1.51 (a) About 30. (b) Since the distribution is right skewed the mean is higher than the median. (c) Q1: between 15 and 20, Q3: between 35 and 40, IQR: about 20. (d) Values that are considered to be unusually low or high lie more than 1.5×IQR away from the quartiles. Upper fence: Q3 + 1.5 × IQR = 37.5 + 1.5 × 20 = 67.5; Lower fence: Q1 - 1.5 × IQR = 17.5 + 1.5 × 20 = −12.5; The lowest AQI recorded is not lower than 5 and the highest AQI recorded is not higher than 65, which are both within the fences. Therefore none of the days in this sample would be considered to have an unusually low or high AQI.

1.53 The histogram shows that the distribution is bimodal, which is not apparent in the box plot. The box plot makes it easy to identify more precise values of observations outside of the whiskers.

1.55 (a) The distribution of number of pets per household is likely right skewed as there is a natural boundary at 0 and only a few people have many pets. Therefore the center would be best described by the median, and variability would be best described by the IQR. (b) The distribution of number of distance to work is likely right skewed as there is a natural boundary at 0 and only a few people live a very long distance from work. Therefore the center would be best described by the median, and variability would be best described by the IQR. (c) The distribution of heights of males is likely symmetric. Therefore the center would be best described by the mean, and variability would be best described by the standard deviation.

1.57 No, we would expect this distribution to be right skewed. There are two reasons for this: (1) there is a natural boundary at 0 (it is not possible to watch less than 0 hours of TV), (2) the standard deviation of the distribution is very large compared to the mean.

1.59 The statement "50% of Facebook users have over 100 friends" means that the median number of friends is 100, which is lower than the mean number of friends (190), which suggests a right skewed distribution for the number of friends of Facebook users.

1.61 (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.

1.63 (a) The distribution is unimodal and symmetric with a mean of about 25 minutes and a standard deviation of about 5 minutes. There does not appear to be any counties with unusually high or low mean travel times. Since the distribution is already unimodal and symmetric, a log transformation is not necessary. (b) Answers will vary. There are pockets of longer travel time around DC, Southeastern NY, Chicago, Minneapolis, Los Angeles, and many other big cities. There is also a large section of shorter average commute times that overlap with farmland in the Midwest. Many farmers' homes are adjacent to their farmland, so their commute would be brief, which may explain why the average commute time for these counties is relatively low.

1.65 (a) We see the order of the categories and the relative frequencies in the bar plot. (b) There are no features that are apparent in the pie chart but not in the bar plot. (c) We usually prefer to use a bar plot as we can also see the relative frequencies of the categories in this graph.

1.67 The vertical locations at which the ideological groups break into the Yes, No, and Not Sure categories differ, which indicates that likelihood of supporting the DREAM act varies by political ideology. This suggests that the two variables may be dependent.

1.69 (a) (i) False. Instead of comparing counts, we should compare percentages of people in each group who suffered cardiovascular problems. (ii) True. (iii) False. Association does not imply causation. We cannot infer a causal relationship based on an observational study. The difference from part (ii) is subtle. (iv) True.
(b) Proportion of all patients who had cardiovascular problems: $\frac{7,979}{227,571} \approx 0.035$
(c) The expected number of heart attacks in the rosiglitazone group, if having cardiovascular problems and treatment were independent, can be calculated as the number of patients in that group multiplied by the overall cardiovascular problem rate in the study: $67,593 * \frac{7,979}{227,571} \approx 2370$.
(d) (i) H_0: The treatment and cardiovascular problems are independent. They have no relationship, and the difference in incidence rates between the rosiglitazone and pioglitazone groups is due to chance. H_A: The treatment and cardiovascular problems are not independent. The difference in the incidence rates between the rosiglitazone and pioglitazone groups is not due to chance and rosiglitazone is associated with an increased risk of serious cardiovascular problems. (ii) A higher number of patients with cardiovascular problems than expected under the assumption of independence would provide support for the alternative hypothesis as this would suggest that rosiglitazone increases the risk of such problems. (iii) In the actual study, we observed 2,593 cardiovascular events in the rosiglitazone group. In the 1,000 simulations under the independence model, we observed somewhat less than 2,593 in every single simulation, which suggests that the actual results did not come from the independence model. That is, the variables do not appear to be independent, and we reject the independence model in favor of the alternative. The study's results provide convincing evidence that rosiglitazone is associated with an increased risk of cardiovascular problems.

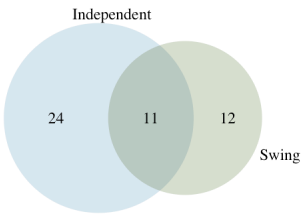
2 Probability

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) 10 tosses. Fewer tosses mean more variability in the sample fraction of heads, meaning there's a better chance of getting at least 60% heads. (b) 100 tosses. More flips means the observed proportion of heads would often be closer to the average, 0.50, and therefore also above 0.40. (c) 100 tosses. With more flips, the observed proportion of heads would often be closer to the average, 0.50. (d) 10 tosses. Fewer flips would increase variability in the fraction of tosses that are heads.

2.5 (a) $0.5^{10} = 0.00098$. (b) $0.5^{10} = 0.00098$. (c) $P(\text{at least one tails}) = 1 - P(\text{no tails}) = 1 - (0.5^{10}) \approx 1 - 0.001 = 0.999$.

2.7 (a) No, there are voters who are both independent and swing voters.
(b)



(c) Each Independent voter is either a swing voter or not. Since 35% of voters are Independents and 11% are both Independent and swing voters, the other 24% must not be swing voters. (d) 0.47. (e) 0.53. (f) P(Independent) × P(swing) = 0.35 × 0.23 = 0.08, which does not equal P(Independent and swing) = 0.11, so the events are dependent.

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) $0.16 + 0.09 = 0.25$. (b) $0.17 + 0.09 = 0.26$. (c) Assuming that the education level of the husband and wife are independent: $0.25 \times 0.26 = 0.065$. You might also notice we actually made a second assumption: that the decision to get married is unrelated to education level. (d) The husband/wife independence assumption is probably not reasonable, because people often marry another person with a comparable level of education. We will leave it to you to think about whether the second assumption noted in part (c) is reasonable.

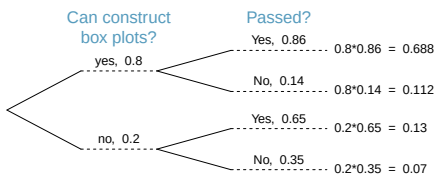
2.13 (a) Invalid. Sum is greater than 1. (b) Valid. Probabilities are between 0 and 1, and they sum to 1. In this class, every student gets a C. (c) Invalid. Sum is less than 1. (d) Invalid. There is a negative probability. (e) Valid. Probabilities are between 0 and 1, and they sum to 1. (f) Invalid. There is a negative probability.

2.15 (a) No, but we could if A and B are independent. (b-i) 0.21. (b-ii) 0.79. (b-iii) 0.3. (c) No, because $0.1 \neq 0.21$, where 0.21 was the value computed under independence from part (a). (d) 0.143.

2.17 (a) No, 0.18 of respondents fall into this combination. (b) $0.60 + 0.20 - 0.18 = 0.62$. (c) $0.18/0.20 = 0.9$. (d) $0.11/0.33 \approx 0.33$. (e) No, otherwise the answers to (c) and (d) would be the same. (f) $0.06/0.34 \approx 0.18$.

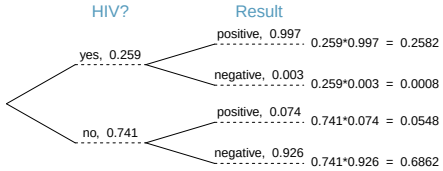
2.19 (a) No. There are 6 females who like Five Guys Burgers. (b) $162/248 = 0.65$. (c) $181/252 = 0.72$. (d) Under the assumption of a dating choices being independent of hamburger preference, which on the surface seems reasonable: $0.65 \times 0.72 = 0.468$. (e) $(252 + 6 - 1)/500 = 0.514$.

2.21 (a)

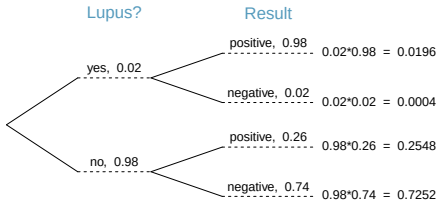


(b) 0.84

2.23 0.8247.



2.25 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.



2.27 (a) 0.3. (b) 0.3. (c) 0.3. (d) $0.3 \times 0.3 = 0.09$. (e) Yes, the population that is being sampled from is identical in each draw.

2.29 (a) $2/9 \approx 0.22$. (b) $3/9 \approx 0.33$. (c) $\frac{3}{10} \times \frac{2}{9} \approx 0.067$. (d) No, e.g. in this exercise, removing one chip meaningfully changes the probability of what might be drawn next.

2.31 $P(^1\text{leggings}, {}^2\text{jeans}, {}^3\text{jeans}) = \frac{5}{24} \times \frac{7}{23} \times \frac{6}{22} = 0.0173$. However, the person with leggings could have come 2nd or 3rd, and these each have this same probability, so $3 \times 0.0173 = 0.0519$.

2.33 (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.

2.35 (a) E(X) = 3.59. SD(X) = 9.64. (b) E(X) = -1.41. SD(X) = 9.64. (c) No, the expected net profit is negative, so on average you expect to lose money.

2.37 5% increase in value.

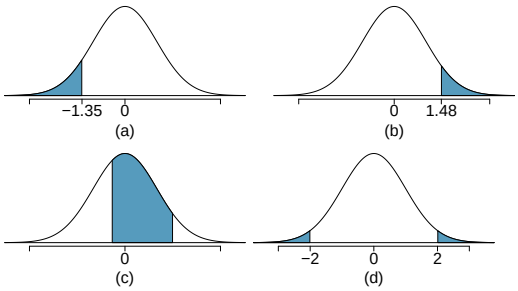
2.39 E = -0.0526. SD = 0.9986.

2.41 (a) E = \$3.90. SD = \$0.34. (b) E = \$27.30. SD = \$0.89.

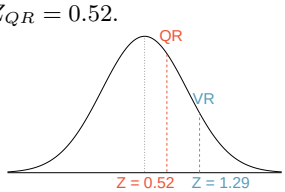
2.43 Approximate answers are OK. (a) $(29 + 32)/144 = 0.42$. (b) $21/144 = 0.15$. (c) $(26 + 12 + 15)/144 = 0.37$.

3 Distributions of random variables

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



3.3 (a) Verbal: $N(\mu = 151, \sigma = 7)$, Quant: $N(\mu = 153, \sigma = 7.67)$. (b) $Z_{VR} = 1.29$, $Z_{QR} = 0.52$.



(c) She scored 1.29 standard deviations above the mean on the Verbal Reasoning section and 0.52 standard deviations above the mean on the Quantitative Reasoning section. (d) She did better on the Verbal Reasoning section since her Z-score on that section was higher. (e) $Perc_{VR} = 0.9007 \approx 90\%$, $Perc_{QR} = 0.6990 \approx 70\%$. (f) $100\% - 90\% = 10\%$ did better than her on VR, and $100\% - 70\% = 30\%$ did better than her on QR. (g) We cannot compare the raw scores since they are on different scales. Comparing her percentile scores is more appropriate when comparing her performance to others. (h) Answer to part (b) would not change as Z-scores can be calculated for distributions that are not normal. However, we could not answer parts (d)-(f) since we cannot use the normal probability table to calculate probabilities and percentiles without a normal model.

3.5 (a) $Z = 0.84$, which corresponds to approximately 160 on QR. (b) $Z = -0.52$, which corresponds to approximately 147 on VR.

3.7 (a) $Z = 1.2 \rightarrow 0.1151$. (b) $Z = -1.28 \rightarrow 70.6°$F or colder.

3.9 (a) $N(25, 2.78)$. (b) $Z = 1.08 \rightarrow 0.1401$. (c) The answers are very close because only the units were changed. (The only reason why they differ at all is because 28°C is 82.4°F, not precisely 83°F.) (d) Since $IQR = Q3 - Q1$, we first need to find $Q3$ and $Q1$ and take the difference between the two. Remember that $Q3$ is the 75^{th} and $Q1$ is the 25^{th} percentile of a distribution. Q1 = 23.13, Q3 = 26.86, IQR = 26. 86 - 23.13 = 3.73.

3.11 (a) $Z = 0.67$. (b) $\mu = \$1650$, $x = \$1800$. (c) $0.67 = \frac{1800-1650}{\sigma} \rightarrow \sigma = \$223.88$.

3.13 $Z = 1.56 \rightarrow 0.0594$, i.e. 6%.

3.15 (a) $Z = 0.73 \rightarrow 0.2327$. (b) If you are bidding on only one auction and set a low maximum bid price, someone will probably outbid you. If you set a high maximum bid price, you may win the auction but pay more than is necessary. If bidding on more than one auction, and you set your maximum bid price very low, you probably won't win any of the auctions. However, if the maximum bid price is even modestly high, you are likely to win multiple auctions. (c) An answer roughly equal to the 10th percentile would be reasonable. Regrettably, no percentile cutoff point guarantees beyond any possible event that you win at least one auction. However, you may pick a higher percentile if you want to be more sure of winning an auction. (d) Answers will vary a little but should correspond to the answer in part (c). We use the 10^{th} percentile: $Z = -1.28 \rightarrow \$69.80$.

3.17 (a) 70% of the data are within 1 standard deviation of the mean, 95% are within 2 and 100% are within 3 standard deviations of the mean. Therefore, we can say that the data approximately follow the 68-95-99.7% Rule. (b) The distribution is unimodal and symmetric. The superimposed normal curve seems to approximate the distribution pretty well. The points on the normal probability plot also seem to follow a straight line. There is one possible outlier on the lower end that is apparent in both graphs, but it is not too extreme. We can say that the distribution is nearly normal.

3.19 (a) No. The cards are not independent. For example, if the first card is an ace of clubs, that implies the second card cannot be an ace of clubs. Additionally, there are many possible categories, which would need to be simplified. (b) No. There are six events under consideration. The Bernoulli distribution allows for only two events or categories. Note that rolling a die could be a Bernoulli trial if we simply to two events, e.g. rolling a 6 and not rolling a 6, though specifying such details would be necessary.

3.21 (a) $(1 - 0.471)^2 \times 0.471 = 0.1318$. (b) $0.471^3 = 0.1045$. (c) $\mu = 1/0.471 = 2.12$, $\sigma = \sqrt{2.38} = 1.54$. (d) $\mu = 1/0.30 = 3.33$, $\sigma = 2.79$. (e) When p is smaller, the event is rarer, meaning the expected number of trials before a success and the standard deviation of the waiting time are higher.

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

3.25 (a) Binomial conditions are met: (1) Independent trials: In a random sample, 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) 0.203. (c) 0.203. (d) 0.167. (e) 0.997.

3.27 (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.

3.29 Want to find the probability that there will be 1,786 or more enrollees. Using the normal approximation: 0.0582. With a 0.5 correction: 0.0559.

3.31 (a) $1 - 0.75^3 = 0.5781$. (b) 0.1406. (c) 0.4219. (d) $1 - 0.25^3 = 0.9844$.

3.33 (a) Geometric distribution: 0.109. (b) Binomial: 0.219. (c) Binomial: 0.137. (d) $1 - 0.875^6 = 0.551$. (e) Geometric: 0.084. (f) Using a binomial distribution with $n = 6$ and $p = 0.75$, we see that $\mu = 4.5$, $\sigma = 1.06$, and $Z = 2.36$. Since this is not within 2 SD, it may be considered unusual.

3.35 0 wins (-$3): 0.1458. 1 win (-$1): 0.3936. 2 wins (+$1): 0.3543. 3 wins (+$3): 0.1063.

3.37 (a) $\overset{Anna}{1/5} \times \overset{Ben}{1/4} \times \overset{Carl}{1/3} \times \overset{Damian}{1/2} \times \overset{Eddy}{1/1} = 1/5! = 1/120$. (b) Since the probabilities must add to 1, there must be $5! = 120$ possible orderings. (c) $8! = 40{,}320$.

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

3.41 (a) Negative binomial with $n = 4$ and $p = 0.55$, where a success is defined here as a female student. The negative binomial setting is appropriate since the last trial is fixed but the order of the first 3 trials is unknown. (b) 0.1838. (c) $\binom{3}{1} = 3$. (d) In the binomial model there are no restrictions on the outcome of the last trial. In the negative binomial model the last trial is fixed. Therefore we are interested in the number of ways of orderings of the other $k-1$ successes in the first $n-1$ trials.

3.43 (a) Poisson with $\lambda = 75$. (b) $\mu = \lambda = 75$, $\sigma = \sqrt{\lambda} = 8.66$. (c) $Z = -1.73$. 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. (d) Using Poisson with $\lambda = 75$: 0.0402.

4 Foundations for inference

4.1 (a) Mean. Each student reports a numerical value: a number of hours. (b) Mean. Each student reports a number, which is a percentage, and we can average over these percentages. (c) Proportion. Each student reports Yes or No, so this is a categorical variable and we use a proportion. (d) Mean. Each student reports a number, which is a percentage like in part (b). (e) Proportion. Each student reports whether or not s/he expects to get a job, so this is a categorical variable and we use a proportion.

4.3 (a) Mean: 13.65. Median: 14. (b) SD: 1.91. IQR: $15 - 13 = 2$. (c) $Z_{16} = 1.23$, which is not unusual since it is within 2 SD of the mean. $Z_{18} = 2.28$, which is generally considered unusual. (d) No. Point estimates that are based on samples only approximate the population parameter, and they vary from one sample to another. (e) We use the SE, which is $1.91/\sqrt{100} = 0.191$ for this sample's mean.

4.5 (a) We are building a distribution of sample statistics, in this case the sample mean. Such a distribution is called a sampling distribution. (b) Because we are dealing with the distribution of sample means, we need to check to see if the Central Limit Theorem applies. Our sample size is greater than 30, and we are told that random sampling is employed. With these conditions met, we expect that the distribution of the sample mean will be nearly normal and therefore symmetric. (c) Because we are dealing with a sampling distribution, we measure its variability with the standard error. $SE = 18.2/\sqrt{45} = 2.713$. (d) The sample means will be more variable with the smaller sample size.

4.7 Recall that the general formula is

$$\text{point estimate} \pm Z^\star \times SE$$

First, identify the three different values. The point estimate is 45%, $Z^\star = 1.96$ for a 95% confidence level, and $SE = 1.2\%$. Then, plug the values into the formula:

$$45\% \pm 1.96 \times 1.2\% \quad \rightarrow \quad (42.6\%, 47.4\%)$$

We are 95% confident that the proportion of US adults who live with one or more chronic conditions is between 42.6% and 47.4%.

4.9 (a) False. Confidence intervals provide a range of plausible values, and sometimes the truth is missed. A 95% confidence interval "misses" about 5% of the time. (b) True. Notice that the description focuses on the true population value. (c) True. If we examine the 95% confidence interval computed in Exercise 4.9, we can see that 50% is not included in this interval. This means that in a hypothesis test, we would reject the null hypothesis that the proportion is 0.5. (d) False. The standard error describes the uncertainty in the overall estimate from natural fluctuations due to randomness, not the uncertainty corresponding to individuals' responses.

4.11 (a) We are 95% confident that Americans spend an average of 1.38 to 1.92 hours per day relaxing or pursuing activities they enjoy. (b) Their confidence level must be higher as the width of the confidence interval increases as the confidence level increases. (c) The new margin of error will be smaller since as the sample size increases the standard error decreases, which will decrease the margin of error.

4.13 (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 sample mean will be nearly normal, allowing for the method normal approximation described. (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. To be more confident that we capture the parameter, we need a wider interval. Think about needing a bigger net to be more sure of catching a fish in a murky lake. (f) True. Optional explanation: This is true since the normal model was used to model the sample mean. The margin of error is half the width of the interval, and the sample mean is the midpoint of the interval. (g) False. In the calculation of the standard error, we divide the standard deviation by the square root of the sample size. To cut the SE (or margin of error) in half, we would need to sample $2^2 = 4$ times the number of people in the initial sample.

4.15 Independence: sample from $< 10\%$ of population, and it is a random sample. We can assume that the students in this sample are independent of each other with respect to number of exclusive relationships they have been in. Notice that there are no students who have had no exclusive relationships in the sample, which suggests some student responses are likely missing (perhaps only positive values were reported). The sample size is at least 30. The skew is strong, but the sample is very large so this is not a concern. 90% CI: (2.97, 3.43). We are 90% confident that undergraduate students have been in 2.97 to 3.43 exclusive relationships, on average.

4.17 (a) $H_0 : \mu = 8$ (On average, New Yorkers sleep 8 hours a night.)
$H_A : \mu < 8$ (On average, New Yorkers sleep less than 8 hours a night.)
(b) $H_0 : \mu = 15$ (The average amount of company time each employee spends not working is 15 minutes for March Madness.)
$H_A : \mu > 15$ (The average amount of company time each employee spends not working is greater than 15 minutes for March Madness.)

4.19 The hypotheses should be about the population mean (μ), not the sample mean. The null hypothesis should have an equal sign and the alternative hypothesis should be about the null hypothesized value, not the observed sample mean. Correction:

$$H_0 : \mu = 10 \; hours$$
$$H_A : \mu > 10 \; hours$$

The one-sided test indicates that we are only interested in showing that 10 is an underestimate. Here the interest is in only one direction, so a one-sided test seems most appropriate. If we would also be interested if the data showed strong evidence that 10 was an overestimate, then the test should be two-sided.

4.21 (a) This claim does is not supported since 3 hours (180 minutes) is not in the interval. (b) 2.2 hours (132 minutes) is in the 95% confidence interval, so we do not have evidence to say she is wrong. However, it would be more appropriate to use the point estimate of the sample. (c) A 99% confidence interval will be wider than a 95% confidence interval, meaning it would enclose this smaller interval. This means 132 minutes would be in the wider interval, and we would not reject her claim based on a 99% confidence level.

4.23 $H_0 : \mu = 130$. $H_A : \mu \neq 130$. $Z = 1.39 \rightarrow$ p-value $= 0.1646$, which is larger than $\alpha = 0.05$. The data do not provide convincing evidence that the true average calorie content in bags of potato chips is different than 130 calories.

4.25 (a) Independence: The sample is random and 64 patients would almost certainly make up less than 10% of the ER residents. The sample size is at least 30. No information is provided about the skew. In practice, we would ask to see the data to check this condition, but here we will make the assumption that the skew is not very strong. (b) $H_0 : \mu = 127$. $H_A : \mu \neq 127$. $Z = 2.15 \rightarrow$ p-value $= 0.0316$. Since the p-value is less than $\alpha = 0.05$, we reject H_0. The data provide convincing evidence that the average ER wait time has increased over the last year. (c) Yes, it would change. The p-value is greater than 0.01, meaning we would fail to reject H_0 at $\alpha = 0.01$.

4.27 $Z = 1.65 = \frac{\bar{x} - 30}{10/\sqrt{70}} \rightarrow \bar{x} = 31.97$.

4.29 (a) H_0: Anti-depressants do not help symptoms of Fibromyalgia. H_A: Anti- depressants do treat symptoms of Fibromyalgia. Remark: Diana might also have taken special note if her symptoms got much worse, so a more scientific approach would have been to use a two-sided test. If you proposed a two-sided approach, your answers in (b) and (c) will be different. (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.

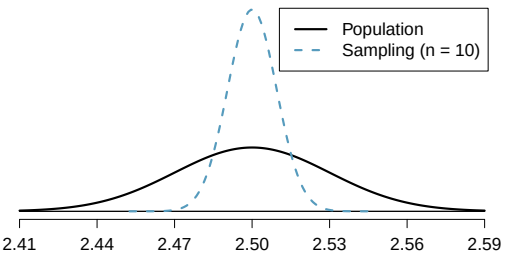
4.31 (a) Scenario I is higher. Recall that a sample mean based on less data tends to be less accurate and have larger standard errors. (b) Scenario I is higher. The higher the confidence level, the higher the corresponding margin of error. (c) They are equal. The sample size does not affect the calculation of the p- value for a given Z-score. (d) Scenario I is higher. If the null hypothesis is harder to reject (lower α), then we are more likely to make a Type 2 Error when the alternative hypothesis is true.

4.33 (a) The distribution is unimodal and strongly right skewed with a median between 5 and 10 years old. Ages range from 0 to slightly over 50 years old, and the middle 50% of the distribution is roughly between 5 and 15 years old. There are potential outliers on the higher end. (b) When the sample size is small, the sampling distribution is right skewed, just like the population distribution. As the sample size increases, the sampling distribution gets more unimodal, symmetric, and approaches normality. The variability also decreases. This is consistent with the Central Limit Theorem. (c) n = 5: $\mu_{\bar{x}} = 10.44$, $\sigma_{\bar{x}} = 4.11$; n = 30: $\mu_{\bar{x}} = 10.44$, $\sigma_{\bar{x}} = 1.68$; n = 100: $\mu_{\bar{x}} = 10.44$, $\sigma_{\bar{x}} = 0.92$. The centers of the sampling distributions shown in part (b) appear to be around 10. It is difficult to estimate the standard deviation for the sampling distribution when $n = 5$ from the histogram (since the distribution is somewhat skewed). If 1.68 is a plausible estimate for the standard deviation of the sampling distribution when $n = 30$, then using the 68-95-99.7% Rule, we would expect the values to range roughly between $10.44 \pm 3 * 1.68 = (5.4, 15.48)$, which seems to be the case. Similarly, when $n = 100$, we would expect the values to range roughly between $10.44 \pm 3 * 0.92 = (7.68, 13.2)$, which also seems to be the case.

4.35 (a) Right skewed. There is a long tail on the higher end of the distribution but a much shorter tail on the lower end. (b) Less than, as the median would be less than the mean in a right skewed distribution. (c) We should not. (d) Even though the population distribution is not normal, the conditions for inference are reasonably satisfied, with the possible exception of skew. If the skew isn't very strong (we should ask to see the data), then we can use the Central Limit Theorem to estimate this probability. For now, we'll assume the skew isn't very strong, though the description suggests it is at least moderate to strong. Use $N(1.3, SD_{\bar{x}} = 0.3/\sqrt{60})$: $Z = 2.58 \rightarrow 0.0049$. (e) It would decrease it by a factor of $1/\sqrt{2}$.

4.37 The centers are the same in each plot, and each data set is from a nearly normal distribution, though the histograms may not look very normal since each represents only 100 data points. The only way to tell which plot corresponds to which scenario is to examine the variability of each distribution. Plot B is the most variable, followed by Plot A, then Plot C. This means Plot B will correspond to the original data, Plot A to the sample means with size 5, and Plot C to the sample means with size 25.

4.39 (a) $Z = -3.33 \rightarrow 0.0004$. (b) The population SD is known and the data are nearly normal, so the sample mean will be nearly normal with distribution $N(\mu, \sigma/\sqrt{n})$, i.e. $N(2.5, 0.0095)$. (c) $Z = -10.54 \rightarrow \approx 0$. (d) See below:



(e) We could not estimate (a) without a nearly normal population distribution. We also could not estimate (c) since the sample size is not sufficient to yield a nearly normal sampling distribution if the population distribution is not nearly normal.

4.41 (a) We cannot use the normal model for this calculation, but we can use the histogram. About 500 songs are shown to be longer than 5 minutes, so the probability is about $500/3000 = 0.167$. (b) Two different answers are reasonable. $^{Option\ 1}$Since the population distribution is only slightly skewed to the right, even a small sample size will yield a nearly normal sampling distribution. We also know that the songs are sampled randomly and the sample size is less than 10% of the population, so the length of one song in the sample is independent of another. We are looking for the probability that the total length of 15 songs is more than 60 minutes, which means that the average song should last at least $60/15 = 4$ minutes. Using $SD_{\bar{x}} = 1.63/\sqrt{15}$, $Z = 1.31 \rightarrow 0.0951$. $^{Option\ 2}$Since the population distribution is not normal, a small sample size may not be sufficient to yield a nearly normal sampling distribution. Therefore, we cannot estimate the probability using the tools we have learned so far. (c) We can now be confident that the conditions are satisfied. $Z = 0.92 \rightarrow 0.1788$.

4.43 (a) $H_0 : \mu_{2009} = \mu_{2004}$. $H_A : \mu_{2009} \neq \mu_{2004}$. (b) $\bar{x}_{2009} - \bar{x}_{2004} = -3.6$ spam emails per day. (c) The null hypothesis was not rejected, and the data do not provide convincing evidence that the true average number of spam emails per day in years 2004 and 2009 are different. The observed difference is about what we might expect from sampling variability alone. (d) Yes, since the hypothesis of no difference was not rejected in part (c).

4.45 (a) $H_0 : p_{2009} = p_{2004}$. $H_A : p_{2009} \neq p_{2004}$. (b) -7%. (c) The null hypothesis was rejected. The data provide strong evidence that the true proportion of those who once a month or less frequently delete their spam email was higher in 2004 than in 2009. The difference is so large that it cannot easily be explained as being due to chance. (d) No, since the null difference, 0, was rejected in part (c).

4.47 True. If the sample size is large, then the standard error will be small, meaning even relatively small differences between the null value and point estimate can be statistically significant.

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 (a) between 0.025 and 0.05 (b) less than 0.005 (c) greater than 0.2 (d) between 0.01 and 0.025

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: $\mu < 8$ (New Yorkers sleep less than 8 hrs per night on average.) (b) Independence: The sample is random and from less than 10% of New Yorkers. The sample is small, so we will use a t-distribution. For this size sample, slight skew is acceptable, and the min/max suggest there is not much skew in the data. $T = -1.75$. $df = 25 - 1 = 24$. (c) $0.025 <$ p-value < 0.05. 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 hrs per night or less is between 0.025 and 0.05. (d) Since p-value < 0.05, reject H_0. The data provide strong evidence that New Yorkers sleep less than 8 hours per night on average. (e) No, as we rejected H_0.

5.9 $t^\star_{19}$ is 1.73 for a one-tail. We want the lower tail, so set -1.73 equal to the T-score, then solve for $\bar{x}$: 56.91.

5.11 (a) We will conduct a 1-sample t-test. H_0: $\mu = 5$. H_A: $\mu < 5$. 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 p-value is about 0.21, which is bigger than $\alpha = 0.05$, so we do not reject H_0. That is, we do not have sufficiently strong evidence to reject Georgianna's claim.
(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) Two-sided, we are evaluating a difference, not in a particular direction. (b) Paired, data are recorded in the same cities at two different time points. The temperature in a city at one point is not independent of the temperature in the same city at another time point. (c) t-test, sample is small and population standard deviation is unknown.

5.17 (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 student 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 student their beginning and end of semester weights are dependent.

5.19 (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 : $\mu_{diff} = 0$ (There is no difference in average daily high temperature between January 1, 1968 and January 1, 2008 in the continental US.) H_A : $\mu_{diff} > 0$ (Average daily high temperature in January 1, 1968 was lower than average daily high temperature in January, 2008 in the continental US.) If you chose a two-sided test, that would also be acceptable. If this is the case, note that your p-value will be a little bigger than what is reported here in part (d). (c) Locations are random and represent less than 10% of all possible locations in the US. The sample size is at least 30. We are not given the distribution to check the skew. In practice, we would ask to see the data to check this condition, but here we will move forward under the assumption that it is not strongly skewed. (d) $T_{50} \approx 1.60 \rightarrow 0.05 <$ p-value < 0.10. (e) Since the p-value $> \alpha$ (since not given use 0.05), fail to reject H_0. The data do not provide strong evidence of temperature warming in the continental US. However it should be noted that the p-value is very close to 0.05. (f) Type 2 Error, since we may have incorrectly failed to reject H_0. There may be an increase, but we were unable to detect it. (g) Yes, since we failed to reject H_0, which had a null value of 0.

5.21 (a) (-0.05, 2.25). (b) We are 90% confident that the average daily high on January 1, 2008 in the continental US was 0.05 degrees lower to 2.25 degrees higher than the average daily high on January 1, 1968. (c) No, since 0 is included in the interval.

5.23 (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. $T_{35} = 2.72 \rightarrow$ p-value $= 0.01$. 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.

5.25 No, he should not move forward with the test since the distributions of total personal income are very strongly skewed. When sample sizes are large, we can be a bit lenient with skew. However, such strong skew observed in this exercise would require somewhat large sample sizes, somewhat higher than 30.

5.27 (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 use the t-distribution to model the sample mean. The normal probability plot of the differences shows an approximately straight line. There isn't a clear reason why this distribution would be skewed, and since the normal quantile plot looks reasonable, we can mark this condition as reasonably satisfied. (d) $T = 4.94$ for $df = 10-1 = 9 \rightarrow$ p-value < 0.01. (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 might believe this intersection is representative of all roads, i.e. there is higher traffic on Friday the 6^{th} relative to Friday the 13^{th}. However, we should be cautious of the required assumption for such a generalization.) (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.

5.29 (a) $H_0 : \mu_{diff} = 0$. $H_A : \mu_{diff} \neq 0$. $T = -2.71$. $df = 5$. $0.02 <$ p-value < 0.05. 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.

5.31 (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.

5.33 $H_0 : \mu_C = \mu_S$. $H_A : \mu_C \neq \mu_S$. $T = 3.27$, $df = 11 \rightarrow$ 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.

5.35 $H_0 : \mu_T = \mu_C$. $H_A : \mu_T \neq \mu_C$. $T = 2.24$, $df = 21 \rightarrow 0.02 <$ 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.

5.37 Let $\mu_{diff} = \mu_{pre} - \mu_{post}$. $H_0 : \mu_{diff} = 0$: Treatment has no effect. $H_A : \mu_{diff} > 0$: Treatment is effective in reducing P.D.T. scores, the average pre-treatment score is higher than the average post-treatment score. Note that the reported values are pre minus post, so we are looking for a positive difference, which would correspond to a reduction in the P.D.T. score. Conditions are checked as follows. Independence: The subjects are randomly assigned to treatments, so the patients in each group are independent. All three sample sizes are smaller than 30, so we use t-tests. Distributions of differences are somewhat skewed. The sample sizes are small, so we cannot reliably relax this assumption. (We will proceed, but we would not report the results of this specific analysis, at least for treatment group 1.) For all three groups: $df = 13$. $T_1 = 1.89$ ($0.025 <$ p-value < 0.05), $T_2 = 1.35$ (p-value $= 0.10$), $T_3 = -1.40$ (p-value > 0.10). The only significant test reduction is found in Treatment 1, however, we had earlier noted that this result might not be reliable due to the skew in the distribution. Note that the calculation of the p-value for Treatment 3 was unnecessary: the sample mean indicated a increase in P.D.T. scores under this treatment (as opposed to a decrease, which was the result of interest). That is, we could tell without formally completing the hypothesis test that the p-value would be large for this treatment group.

5.39 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.

5.41 Alternative.

5.43 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.

5.45 (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. Approx. normal: 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.

5.47 (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 $197 + 1 = 198$.

5.49 (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.

5.51 (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 only for Trmt 1 vs. Trmt 3 may be statistically significant: $0.01 <$ p-value < 0.02. Since we cannot tell, we should use a computer to get the p-value, 0.015, which is statistically significant for the adjusted significance level. That is, we have identified Treatment 1 and Treatment 3 as having different effects. Checking the other two comparisons, the differences are not statistically significant.

6 Inference for categorical data

6.1 (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 $1/\sqrt{2}$.

6.3 (a) True. See the reasoning of 6.1(b). (b) True. We take the square root of the sample size in the SE formula. (c) True. The independence and success-failure conditions are satisfied. (d) True. The independence and success-failure conditions are satisfied.

6.5 (a) False. A confidence interval is constructed to estimate the population proportion, not the sample proportion. (b) True. 95% CI: 70% $\pm$ 8%. (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%.

6.7 With a random sample from $<$ 10% of the population, independence is satisfied. The success-failure condition is also satisfied. $ME = z^\star \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 1.96\sqrt{\frac{0.56\times 0.44}{600}} = 0.0397 \approx 4\%$

6.9 (a) Proportion of graduates from this university who found a job within one year of graduating. $\hat{p} = 348/400 = 0.87$. (b) This is a random sample from less than 10% of the population, so the observations are independent. Success-failure condition is satisfied: 348 successes, 52 failures, both well above 10. (c) (0.8371, 0.9029). We are 95% confident that approximately 84% to 90% of graduates from this university found a job within one year of completing their undergraduate degree. (d) 95% of such random samples would produce a 95% confidence interval that includes the true proportion of students at this university who found a job within one year of graduating from college. (e) (0.8267, 0.9133). Similar interpretation as before. (f) 99% CI is wider, as we are more confident that the true proportion is within the interval and so need to cover a wider range.

6.11 (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%.

6.13 (a) This is an appropriate setting for a hypothesis test. $H_0 : p = 0.50$. $H_A : p > 0.50$. Both independence and the success-failure condition are satisfied. $Z = 1.12 \rightarrow$ p-value $= 0.1314$. Since the p-value $> \alpha = 0.05$, we fail to reject H_0. The data do not provide strong evidence that more than half of all Independents oppose the public option plan. (b) Yes, since we did not reject H_0 in part (a).

6.15 (a) $H_0 : p = 0.38$. $H_A : p \neq 0.38$. Independence (random sample, $< 10\%$ of population) and the success-failure condition are satisfied. $Z = -20.5 \rightarrow$ p-value ≈ 0. Since the p-value is very small, we reject H_0. The data provide strong evidence that the proportion of Americans who only use their cell phones to access the internet is different than the Chinese proportion of 38%, and the data indicate that the proportion is lower in the US. (b) If in fact 38% of Americans used their cell phones as a primary access point to the internet, the probability of obtaining a random sample of 2,254 Americans where 17% or less or 59% or more use their only their cell phones to access the internet would be approximately 0. (c) (0.1545, 0.1855). We are 95% confident that approximately 15.5% to 18.6% of all Americans primarily use their cell phones to browse the internet.

6.17 (a) $H_0 : p = 0.5$. $H_A : p > 0.5$. Independence (random sample, $< 10\%$ of population) is satisfied, as is the success-failure conditions (using $p_0 = 0.5$, we expect 40 successes and 40 failures). $Z = 2.91 \rightarrow$ p-value $= 0.0018$. Since the p-value < 0.05, we reject the null hypothesis. The data provide strong evidence that 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 randomly guess, the probability of getting a random sample of 80 people where 53 or more identify a soda correctly would be 0.0018.

6.19 (a) Independence is satisfied (random sample from $< 10\%$ of the population), as is the success-failure condition (40 smokers, 160 non-smokers). The 95% CI: (0.145, 0.255). We are 95% confident that 14.5% to 25.5% of all students at this university smoke. (b) We want $z^\star SE$ to be no larger than 0.02 for a 95% confidence level. We use $z^\star = 1.96$ and plug in the point estimate $\hat{p} = 0.2$ within the SE formula: $1.96\sqrt{0.2(1-0.2)/n} \leq 0.02$. The sample size n should be at least 1,537.

6.21 The margin of error, which is computed as $z^\star SE$, must be smaller than 0.01 for a 90% confidence level. We use $z^\star = 1.65$ for a 90% confidence level, and we can use the point estimate $\hat{p} = 0.52$ in the formula for SE. $1.65\sqrt{0.52(1-0.52)/n} \leq 0.01$. Therefore, the sample size n must be at least 6,796.

6.23 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.

6.25 (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).

6.27 (a) (0.23, 0.33). We are 95% confident that the proportion of Democrats who support the plan is 23% to 33% higher than the proportion of Independents who do. (b) True.

6.29 (a) College grads: 23.7%. Non-college grads: 33.7%. (b) Let p_{CG} and p_{NCG} represent the proportion of college graduates and non-college graduates who responded "do not know". $H_0 : p_{CG} = p_{NCG}$. $H_A : p_{CG} \neq p_{NCG}$. Independence is satisfied (random sample, $< 10\%$ of the population), and the success-failure condition, which we would check using the pooled proportion ($\hat{p} = 235/827 = 0.284$), is also satisfied. $Z = -3.18 \rightarrow$ p-value $= 0.0014$. Since the p-value is very small, we reject H_0. The data provide strong evidence that the proportion of college graduates who do not have an opinion on this issue is different than that of non-college graduates. The data also indicate that fewer college grads say they "do not know" than non-college grads (i.e. the data indicate the direction after we reject H_0).

6.31 (a) College grads: 35.2%. Non-college grads: 33.9%. (b) Let p_{CG} and p_{NCG} represent the proportion of college graduates and non-college grads who support offshore drilling. $H_0 : p_{CG} = p_{NCG}$. $H_A : p_{CG} \neq p_{NCG}$. Independence is satisfied (random sample, $< 10\%$ of the population), and the success-failure condition, which we would check using the pooled proportion ($\hat{p} = 286/827 = 0.346$), is also satisfied. $Z = 0.39 \rightarrow$ p-value $= 0.6966$. Since the p-value $> \alpha$ (0.05), we fail to reject H_0. The data do not provide strong evidence of a difference between the proportions of college graduates and non-college graduates who support offshore drilling in California.

6.33 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, $< 10\%$ of the population), as is the success-failure condition, which we would check using the pooled proportion ($\hat{p} = 70/495 = 0.141$). $Z = -1.65 \rightarrow$ 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.

6.35 (a) Summary of the study:

		Virol. failure		
		Yes	No	Total
Treatment	Nevaripine	26	94	120
	Lopinavir	10	110	120
	Total	36	204	240

(b) $H_0 : p_N = p_L$. There is no difference in virologic failure rates between the Nevaripine and Lopinavir groups. $H_A : p_N \neq p_L$. There is some difference in virologic failure rates between the Nevaripine and Lopinavir groups. (c) Random assignment was used, so the observations in each group are independent. If the patients in the study are representative of those in the general population (something impossible to check with the given information), then we can also confidently generalize the findings to the population. The success-failure condition, which we would check using the pooled proportion ($\hat{p} = 36/240 = 0.15$), is satisfied. $Z = 2.89 \rightarrow$ p-value $= 0.0039$. Since the p-value is low, we reject H_0. There is strong evidence of a difference in virologic failure rates between the Nevaripine and Lopinavir groups do not appear to be independent.

6.37 No. The samples at the beginning and at the end of the semester are not independent since the survey is conducted on the same students.

6.39 (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.

6.41 (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 > 0.3. (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.

6.43 Use a chi-squared goodness of fit test. H_0: Each option is equally likely. H_A: Some options are preferred over others. Total sample size: 99. Expected counts: (1/3) * 99 = 33 for each option. These are all above 5, so conditions are satisfied. $df = 3 - 1 = 2$ and $\chi^2 = \frac{(43-33)^2}{33} + \frac{(21-33)^2}{33} + \frac{(35-33)^2}{33} = 7.52 \rightarrow 0.02 <$ p-value < 0.05. Since the p-value is less than 5%, we reject H_0. The data provide convincing evidence that some options are preferred over others.

6.45 (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.

6.47 H_0: The opinion of college grads and non-grads is not different on the topic of drilling for oil and natural gas off the coast of California. H_A: Opinions regarding the drilling for oil and natural gas off the coast of California has an association with earning a college degree.

$$E_{row\ 1,col\ 1} = 151.5 \quad E_{row\ 1,col\ 2} = 134.5$$
$$E_{row\ 2,col\ 1} = 162.1 \quad E_{row\ 2,col\ 2} = 143.9$$
$$E_{row\ 3,col\ 1} = 124.5 \quad E_{row\ 3,col\ 2} = 110.5$$

Independence: The samples are both random, unrelated, and from less than 10% of the population, so independence between observations is reasonable. Sample size: All expected counts are at least 5. $\chi^2 = 11.47$, $df = 2 \rightarrow 0.001 <$ p-value < 0.005. Since the p-value $< \alpha$, we reject H_0. There is strong evidence that there is an association between support for off-shore drilling and having a college degree.

6.49 (a) H_0: The age of Los Angeles residents is independent of shipping carrier preference variable. H_A: The age of Los Angeles residents is associated with the shipping carrier preference variable. (b) The conditions are not satisfied since some expected counts are below 5.

6.51 No. For a confidence interval, we check the success-failure condition using the data, and there are only 9 respondents who said bullying is no problem at all.

6.53 (a) $H_0 : p = 0.69$. $H_A : p \neq 0.69$. (b) $\hat{p} = \frac{17}{30} = 0.57$. (c) The success-failure condition is not satisfied; note that it is appropriate to use the null value ($p_0 = 0.69$) to compute the expected number of successes and failures. (d) Answers may vary. Each student can be represented with a card. Take 100 cards, 69 black cards representing those who follow the news about Egypt and 31 red cards representing those who do not. Shuffle the cards and draw with replacement (shuffling each time in between draws) 30 cards representing the 30 high school students. Calculate the proportion of black cards in this sample, $\hat{p}_{sim}$, i.e. the proportion of those who follow the news in the simulation. Repeat this many times (e.g. 10,000 times) and plot the resulting sample proportions. The p-value will be two times the proportion of simulations where $\hat{p}_{sim} \leq 0.57$. (Note: we would generally use a computer to perform these simulations.) (e) The p-value is about $0.001 + 0.005 + 0.020 + 0.035 + 0.075 = 0.136$, meaning the two-sided p-value is about 0.272. Your p-value may vary slightly since it is based on a visual estimate. Since the p-value is greater than 0.05, we fail to reject H_0. The data do not provide strong evidence that the proportion of high school students who followed the news about Egypt is different than the proportion of American adults who did.

6.55 The subscript $_{pr}$ corresponds to provocative and $_{con}$ to conservative. (a) $H_0 : p_{pr} = p_{con}$. $H_A : p_{pr} \neq p_{con}$. (b) -0.35. (c) The left tail for the p-value is calculated by adding up the two left bins: $0.005 + 0.015 = 0.02$. Doubling the one tail, the p-value is 0.04. (Students may have approximate results, and a small number of students may have a p-value of about 0.05.) Since the p-value is low, we reject H_0. The data provide strong evidence that people react differently under the two scenarios.

7 Introduction to linear regression

7.1 (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.

7.3 (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.

7.5 (a) Exam 2 since there is less of a scatter in the plot of final exam grade versus exam 2. Notice that the relationship between Exam 1 and the Final Exam appears to be slightly nonlinear. (b) Exam 2 and the final are relatively close to each other chronologically, or Exam 2 may be cumulative so has greater similarities in material to the final exam. Answers may vary for part (b).

7.7 (a) $r = -0.7 \rightarrow$ (4). (b) $r = 0.45 \rightarrow$ (3). (c) $r = 0.06 \rightarrow$ (1). (d) $r = 0.92 \rightarrow$ (2).

7.9 (a) True. (b) False, correlation is a measure of the linear association between any two numerical variables.

7.11 (a) The relationship is positive, weak, and possibly linear. However, there do appear to be some anomalous observations along the left where several students have the same height that is notably far from the cloud of the other points. Additionally, there are many students who appear not to have driven a car, and they are represented by a set of points along the bottom of the scatterplot. (b) There is no obvious explanation why simply being tall should lead a person to drive faster. However, one confounding factor is gender. Males tend to be taller than females on average, and personal experiences (anecdotal) may suggest they drive faster. If we were to follow-up on this suspicion, we would find that sociological studies confirm this suspicion. (c) Males are taller on average and they drive faster. The gender variable is indeed an important confounding variable.

7.13 (a) There is a somewhat weak, positive, possibly linear relationship between the distance traveled and travel time. There is clustering near the lower left corner that we should take special note of. (b) Changing the units will not change the form, direction or strength of the relationship between the two variables. If longer distances measured in miles are associated with longer travel time measured in minutes, longer distances measured in kilometers will be associated with longer travel time measured in hours. (c) Changing units doesn't affect correlation: $r = 0.636$.

7.15 (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.

7.17 In each part, we can write the husband ages as a linear function of the wife ages.
(a) $age_H = age_W + 3$.
(b) $age_H = age_W - 2$.
(c) $age_H = 2 \times age_W$.
Since the slopes are positive and these are perfect linear relationships, the correlation will be exactly 1 in all three parts. An alternative way to gain insight into this solution is to create a mock data set, e.g. 5 women aged 26, 27, 28, 29, and 30, then find the husband ages for each wife in each part and create a scatterplot.

7.19 Correlation: no units. Intercept: kg. Slope: kg/cm.

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

7.23 (a) There is a positive, very strong, linear association between the number of tourists and spending. (b) Explanatory: number of tourists (in thousands). Response: spending (in millions of US dollars). (c) We can predict spending for a given number of tourists using a regression line. This may be useful information for determining how much the country may want to spend in advertising abroad, or to forecast expected revenues from tourism. (d) Even though the relationship appears linear in the scatterplot, the residual plot actually shows a nonlinear relationship. This is not a contradiction: residual plots can show divergences from linearity that can be difficult to see in a scatterplot. A simple linear model is inadequate for modeling these data. It is also important to consider that these data are observed sequentially, which means there may be a hidden structure not evident in the current plots but that is important to consider.

7.25 (a) First calculate the slope: $b_1 = R \times s_y/s_x = 0.636 \times 113/99 = 0.726$. Next, make use of the fact that the regression line passes through the point $(\bar{x}, \bar{y})$: $\bar{y} = b_0 + b_1 \times \bar{x}$. Plug in $\bar{x}$, $\bar{y}$, and b_1, and solve for b_0: 51. Solution: $\widehat{travel\ time} = 51 + 0.726 \times distance$. (b) b_1: For each additional mile in distance, the model predicts an additional 0.726 minutes in travel time. b_0: When the distance traveled is 0 miles, the travel time is expected to be 51 minutes. It does not make sense to have a travel distance of 0 miles in this context. Here, the y-intercept serves only to adjust the height of the line and is meaningless by itself. (c) $R^2 = 0.636^2 = 0.40$. About 40% of the variability in travel time is accounted for by the model, i.e. explained by the distance traveled. (d) $\widehat{travel\ time} = 51 + 0.726 \times distance = 51 + 0.726 \times 103 \approx 126$ minutes. (Note: we should be cautious in our predictions with this model since we have not yet evaluated whether it is a well-fit model.) (e) $e_i = y_i - \hat{y}_i = 168 - 126 = 42$ minutes. A positive residual means that the model underestimates the travel time. (f) No, this calculation would require extrapolation.

7.27 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.

7.29 (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. (d) Poverty level explains 70.52% of the variability in murder rates in metropolitan areas. (e) $\sqrt{0.7052} = 0.8398$

7.31 (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.

7.33 (a) There is a negative, moderate-to-strong, somewhat linear relationship between percent of families who own their home and the percent of the population living in urban areas in 2010. There is one outlier: a state where 100% of the population is urban. The variability in the percent of homeownership also increases as we move from left to right in the plot. (b) The outlier is located in the bottom right corner, horizontally far from the center of the other points, so it is a point with high leverage. It is an influential point since excluding this point from the analysis would greatly affect the slope of the regression line.

7.35 (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 greater than zero ($\beta_1 > 0$). A two-sided test would also be acceptable for this application. The p-value for the two-sided alternative hypothesis ($\beta_1 \neq 0$) is incredibly small, so the p-value for the one-sided hypothesis will be even smaller. That is, 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.

7.37 (a) H_0: $\beta_1 = 0$. H_A: $\beta_1 > 0$. A two-sided test would also be acceptable for this application. The p-value, as reported in the table, is incredibly small. Thus, for a one-sided test, the p-value will also be incredibly small, and 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) 63.2612. Since R^2 is low, the prediction based on this regression model is not very reliable. (f) No, we should avoid extrapolating.

7.39 (a) $r = \sqrt{0.28} \approx -0.53$. We know the correlation is negative due to the negative association shown in the scatterplot. (b) The residuals appear to be fan shaped, indicating non-constant variance. Therefore a simple least squares fit is not appropriate for these data.

7.41 (a) $H_0 : \beta_1 = 0; H_A : \beta_1 \neq 0$ (b) The p-value for this test is approximately 0, therefore we reject H_0. The data provide convincing evidence that poverty percentage is a significant predictor of murder rate. (c) $n = 20, df = 18, T^*_{18} = 2.10$; $2.559 \pm 2.10 \times 0.390 = (1.74, 3.378)$; For each percentage point poverty is higher, murder rate is expected to be higher on average by 1.74 to 3.378 per million. (d) Yes, we rejected H_0 and the confidence interval does not include 0.

7.43 This is a one-sided test, so the p-value should be half of the p-value given in the regression table, which will be approximately 0. Therefore the data provide convincing evidence that poverty percentage is positively associated with murder rate.

8 Multiple and logistic regression

8.1 (a) $\widehat{baby\_weight} = 123.05 - 8.94 \times smoke$ (b) The estimated body weight of babies born to smoking mothers is 8.94 ounces lower than babies born to non-smoking mothers. Smoker: $123.05 - 8.94 \times 1 = 114.11$ ounces. Non-smoker: $123.05 - 8.94 \times 0 = 123.05$ ounces. (c) H_0: $\beta_1 = 0$. H_A: $\beta_1 \neq 0$. $T = -8.65$, and the p-value is approximately 0. Since the p-value is very small, we reject H_0. The data provide strong evidence that the true slope parameter is different than 0 and that there is an association between birth weight and smoking. Furthermore, having rejected H_0, we can conclude that smoking is associated with lower birth weights.

8.3 (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) $R^2 = 0.2504$. $R^2_{adj} = 0.2468$.

8.5 (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).

8.7 Remove age.

8.9 Based on the p-value alone, either gestation or smoke should be added to the model first. However, since the adjusted R^2 for the model with gestation is higher, it would be preferable to add gestation in the first step of the forward-selection algorithm. (Other explanations are possible. For instance, it would be reasonable to only use the adjusted R^2.)

8.11 She should use p-value selection since she is interested in finding out about significant predictors, not just optimizing predictions.

8.13 Nearly normal residuals: The normal probability plot shows a nearly normal distribution of the residuals, however, there are some minor irregularities at the tails. With a data set so large, these would not be a concern.
Constant 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.

8.15 (a) There are a few potential outliers, e.g. on the left in the `total_length` variable, but nothing that will be of serious concern in a data set this large. (b) When coefficient estimates are sensitive to which variables are included in the model, this typically indicates that some variables are collinear. For example, a possum's gender may be related to its head length, which would explain why the coefficient (and p-value) for `sex_male` changed when we removed the `head_length` variable. Likewise, a possum's skull width is likely to be related to its head length, probably even much more closely related than the head length was to gender.

8.17 (a) The logistic model relating $\hat{p}_i$ to the predictors may be written as $\log\left(\frac{\hat{p}_i}{1-\hat{p}_i}\right) = 33.5095 - 1.4207 \times sex\_male_i - 0.2787 \times skull\_width_i + 0.5687 \times total\_length_i - 1.8057 \times tail\_length_i$. Only `total_length` has a positive association with a possum being from Victoria. (b) $\hat{p} = 0.0062$. While the probability is very near zero, we have not run diagnostics on the model. We might also be a little skeptical that the model will remain accurate for a possum found in a US zoo. For example, perhaps the zoo selected a possum with specific characteristics but only looked in one region. On the other hand, it is encouraging that the possum was caught in the wild. (Answers regarding the reliability of the model probability will vary.)