Chapter 12:

Linear Regression

Previous chapters have dealt with a single independent variable that is categorical. This chapter continues testing simple hypotheses concerning the expected value. However, here, the independent variable is continuous. We also deal with multiple independent variables.

❧ ❧ ❧

The voters of Maine are being sent to the polls to vote on a constitutional referendum (ballot measure) that proposes to limit the definition of marriage to the union of one man and one woman. This was not the first time that Americans were sent to the polls to vote on this or a closely related issue. Given the information from previous votes, what is the estimated probability that this ballot measure will pass in Maine?

The t-test and its analysis of variance extension from previous chapters were suitable when the independent variables were all categorical. The categorical nature of those variables made it easy to group the records (trials) into groups of identical levels and compare the means. However, such techniques are unsuitable for continuous independent variables. In such cases, there is no way to separate the records (trials) into a useful number of cases from which we can meaningfully compare means.

ANOVA

level

There are a couple solutions. The first is to discretize the variables—turn the continuous variables into categorical variables and use the methods from the previous chapters. Unfortunately, this is an inefficient use of the data; we are discarding some rather important information. The second method takes advantage of the continuous aspect of the data. This second method is called regression.

discretize

regression

There are two primary classes of regression: linear and non-linear. The former models linear functions of the *coefficients* of the independent variables to illuminate their relationship with the dependent variable. The latter does the same with non-linear functions of the coefficients of the independent variables. The latter is beyond the scope of this text. This chapter focuses on the Classical Linear Model (CLM) in which data is fit using linear functions of parameters.

relationship

The former standard and conceptually most straight-forward method of fitting data using the classical linear model is to use the Ordinary Least Squares method. Before we cover modeling the relationship between two continuous variables, let us graph the variables to visualize the process.

OLS

12.1: Scatterplots

The typical way of graphing two continuous variables to see the relationship between the two is to use a bivariate scatterplot. Just from the scatterplot, one can see if the relationship is strong, is positive, and is linear. If the relationship is not linear, then alterations to the model should happen to take advantage of its shape. Ignoring the shape of the relationship invalidates the usefulness of linear regression.

bivariate

To illustrate this, let us create quadratic data:

```
set.seed(57)
x = runif(100) - 0.5
y = x^2 + rnorm(100, m=0, s=0.01)
```

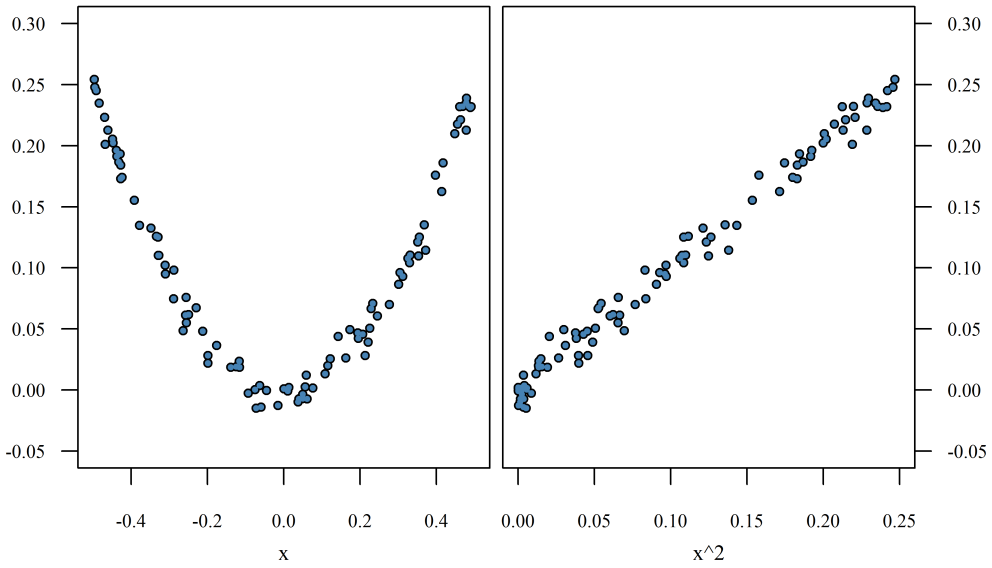

Figure 12.1: *Two scatterplots. The left is of y against x; the right, y against x^2. Note that the (linear) correlation between x and y is insignificant ($r = -0.03$). However, the correlation between x^2 and y is extremely high ($r = 0.99$).*

The first line sets the random number seed, allowing our results to match. The second line creates 100 x-values ranging from -0.5 to 0.5. Line three defines the response (dependent) variable, which is the square of the x-variable, plus some noise.

y **against** *x*

Plotting the raw data, y against x, shows the strong quadratic component (Figure 12.1, left). The correlation between x and y is insignificant, which indicates there is no detected linear relationship between the two variables: `cor.test(x,y)`. Unfortunately, as we can see, a strong relationship exists, one that can easily be detected from the graphic. **Plot your data first.**

How does one create a scatterplot in R? Before answering that question, how does one create a scatterplot by hand? Basically, there are 4 steps:

1. Draw the x and y axes, making sure they span enough distance to allow you to plot all data values.
2. Place the axis values along each axis.
3. Label each axis with a title.
4. Plot each point on the graphic.

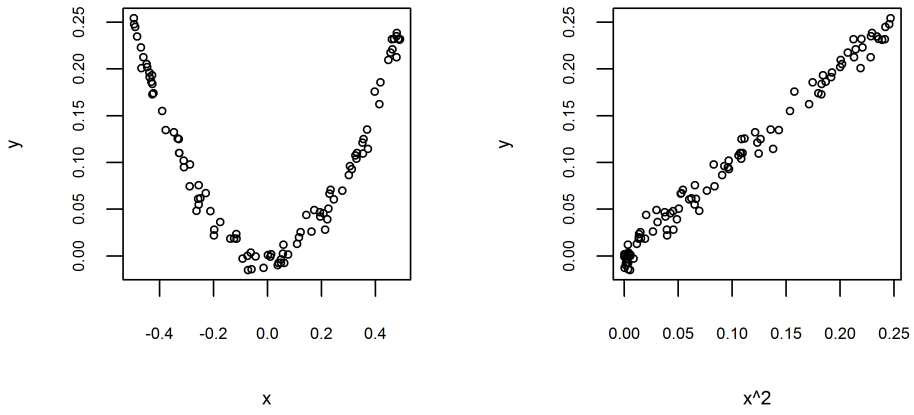

Figure 12.2: *Two default scatterplots. The left is of y against x; the right, y against x^2. Note that even the default scatterplots are very informative about the apparent relationship between the variables.*

In R, the code to get a bare plot is just one line:

```
plot(x,y)
```

This command draws the axes, automatically calculating a correct range. It writes the axis values and axis labels. It plots the points.

Jeremy Bentham

That code produces an excellent utilitarian graphic allowing you to visually determine if the relationship is linear or not. Using the data generated for Figure 12.1, the default scatterplot is shown in Figure 12.2. Note that it does differ from Figure 12.1, but only in the customizations made.

12.1.1 The Three Realms

While the default plot is a fully-functional scatterplot, there are things you can do to improve its aesthetics. Those things include the points plotted, the axis labels, and the x and y ranges. All of these can be customized in R.

To understand the model for how R graphs, think of a canvas. You start with a blank canvas. You then add a point to it. Then, another. Then, a line, perhaps. You started with emptiness and annotated it piece by piece.

In R, there are three realms to graphics code: The preamble, the plotting, and the annotation. Different graphing commands belong in different realms. The plotting section creates the canvas to your preamble specification and draws the plot. If you would like to add a line to the canvas, you draw it. If you would also like to add additional points, you draw them. These are added to the canvas, in order.

I used the following code to create the left graphic of Figure 12.1. The three realms are marked.

```
## Realm 1: Preamble
  par(cex=0.8, cex.lab=0.8,cex.axis=0.7)
  par(family="serif", las=1)
  par(mar=c(3,2,0,0)+0.2)

## Realm 2: Plotting
  plot(x,y, pch=20, xlim=c(-0.52,0.52), ylim=c(-0.05,0.3),
    xlab="",ylab="")

## Realm 3: Annotating
  title(xlab="x",line=2)
  points(x,y,col="steelblue", pch=20,cex=0.6)
```

The Preamble: The first realm of the graphic code is the preamble section. In this section, you specify aspects of the entire graphic. This includes scaling factors, backgrounds, and colors. Every line in this realm uses the versatile `par` command. If you read the help file on that command, you will see everything that can be set.

In the graphics for this book, I typically set four things: the scale, the font, the label orientation, and the margins. Let us examine each in turn. First, I set the scale using the first `par` command in the listing above. There are many aspects of the graphic you can scale. The parameter `cex` sets the base scale for the entire graphic; `cex.lab`, for the axis labels; `cex.axis`, the axis values. Values above 1 make the areas larger; below 1, smaller.

Second, I set the font family to serif; sans-serif is default. Options for the `family` parameter also include "sans" and "mono" with advanced options for Hershey fonts and for symbol fonts. The parameter `las=1` makes all axis values horizontal.

Finally, I set the margin using the `mar` parameter. The margin takes four values, which are the margins (in lines) for the bottom, left, top, and right sides (note that this is clockwise, starting at the x-axis). Since this pa-

rameter requires four values, they need to be collected using the `c()` function. Thus, the margins for this graphic will be 3.2 lines on the bottom, 2.2 lines on the left side, 0.2 lines on the top, and 0.2 lines on the right.

The Plotting: After you have set the global aspects of the graphic, you are ready to plot the base graphic. This realm contains a single command that starts a new plot. The two commands commonly found here are `plot` for a scatterplot and `boxplot` for a box-and-whiskers plot. These two create a new plot.

boxplot

Note that the commands in the preamble plotted nothing. They just told `R` what things should look like in the future. This section actually starts a new plot according to the rules created in the preamble.

In this example, I created a scatterplot using the `plot` command. With that command, there are several options I can use to customize the plot. Some of these, like `pch`, can also be set in the preamble to apply to all plots. Others, like `xlab`, must be placed here. Notice that I set both the x-axis label and the y-axis label to empty. I did this to allow me greater control of their placing, later.

The Annotations: The final realm is the annotation section. Now that you have a graphic plotted, you can add to it (annotate it).

In this section, you can set axis labels using the `title` command. You can add points to the original plot using the `points` command. You can add lines to the original plot using either `lines` or `abline`. You can add things to the axis values using either `axis` or `mtext`. You can add text to the main plot using `text`. You can do almost anything you should want.

A strength with using a scripting language to create your graphics is that you can create them using trial and error. If you made the margins too wide, change them and re-run the entire block of code again. If you want to change the color of the points, change the color and re-run the entire block of code. If you want to add the sample mean to the plot, but are not sure where, add it somewhere, then reposition it until it is in the right place.

Were you to use a menu-driven system, you would have to start from scratch each time you wanted to change something. Repeating the menu options and clicking the mouse takes a lot of time.

You can also be extremely precise with placement and even place things on the plot that are calculated. For instance, if I wanted to draw a

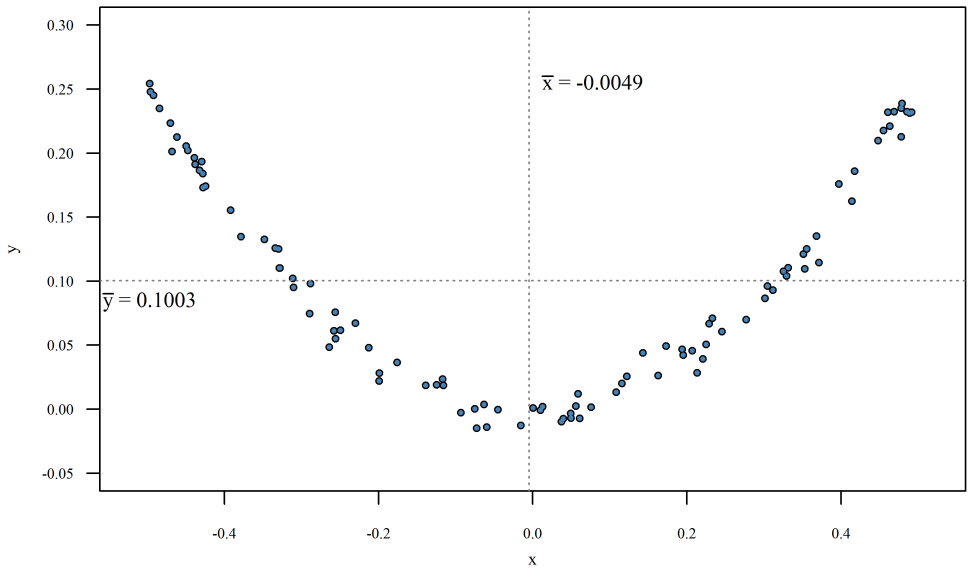

Figure 12.3: *An annotated scatterplot of y against x. Note that the sample means are added to this plot.*

vertical line at the sample mean of the x-values and a horizontal line at the sample mean of the y-values, I would add the following two lines:

```
abline(v=mean(x), col="grey50", lty=3)
abline(h=mean(y), col="grey50", lty=3)
```

The first draws a vertical line at $\overline{x}$. The second draws a horizontal line at $\overline{y}$. Both are grey and dotted (`lty=3`). I never calculated the means before these lines, nor do I care what they are. I just created the lines.

With a little more work, I could have written the values of the two means (see Figure 12.3). Since I am placing the text in the plotting region, the command is `text()`. It takes at least three things: the x- and y-value of where to place the text, and the text itself. The following lines place the sample means on the graphic. The optional `pos` parameter adjusts the text relative to the point stated. The positions correspond to the axis numbering. Thus, `pos=4` means the text is written to the right (towards the fourth axis) of the stated point; `pos=3` is above.

```
text(mean(x),max(y), round(mean(x),4), pos=4 )
text(min(x),mean(y), round(mean(y),4), pos=3 )
```

The following lines take it to the next level by including $\overline{x}$ and $\overline{y}$, as symbols, to the plot. This makes the graphic even more understandable by the average reader. While some of this is beyond the scope of this book, I encourage you to tinker with the lines to see what each part does.

```
text(mean(x),max(y), pos=4, substitute( paste( bar(x)," = "
    ,mn), list(mn=round(mean(x),4))))
text(min(x),mean(y), pos=1, substitute( paste( bar(y)," = "
    ,mn), list(mn=round(mean(y),4))))
```

If this looks interesting to you, I encourage you to read two books: *The R Book* (Crawley 2007) and *R Graphics* (Murrell 2011). The first deals with more than just graphics, while the second focuses on them. Both are good resources for creating graphics in R.

After adding these annotations, you will have a graphic much like that in Figure 12.3. Notice that it is now clear what the numbers mean; they are the sample means for each variable. The crossing point of the two dotted lines is the center of mass for the data. One thing that you should also add to the graphic is the measured correlation between the variables. I leave that as an exercise for you.

12.1.2 Graphics Conclusion

With a little bit of practice and a little bit of patience, one can make graphics in R look perfect. This section just provided an introduction to what R graphics can do. Figure 12.4 shows more of what one can do using R for graphics.

Beyond the two books mentioned in this section, there is also "Data Analysis and Graphics Using R: An Example-Based Approach," by Maindonald and Braun. This book has the advantage of weaving together analysis and graphics, as they should. Remember that graphics teach the researcher about the data. They also allow the researcher to tell the story of the data.

Story

In addition to print sources, there are many online sources for inspiration. Since websites tend to blink out of existence quickly, it would be a waste of space to list them. However, a web search for "graphics in R" should return several sites. Of course, a more precise search on "how do I make a scatterplot in R" may return better results.

Finally, I would like to leave you with Figure 12.4. This graphic illustrates several things. First, it shows the relationship between the number of Nobel Prizes won, adjusted for population, and the chocolate consumption in the country. Second, it uses the country's flag instead of a dot. This pro-

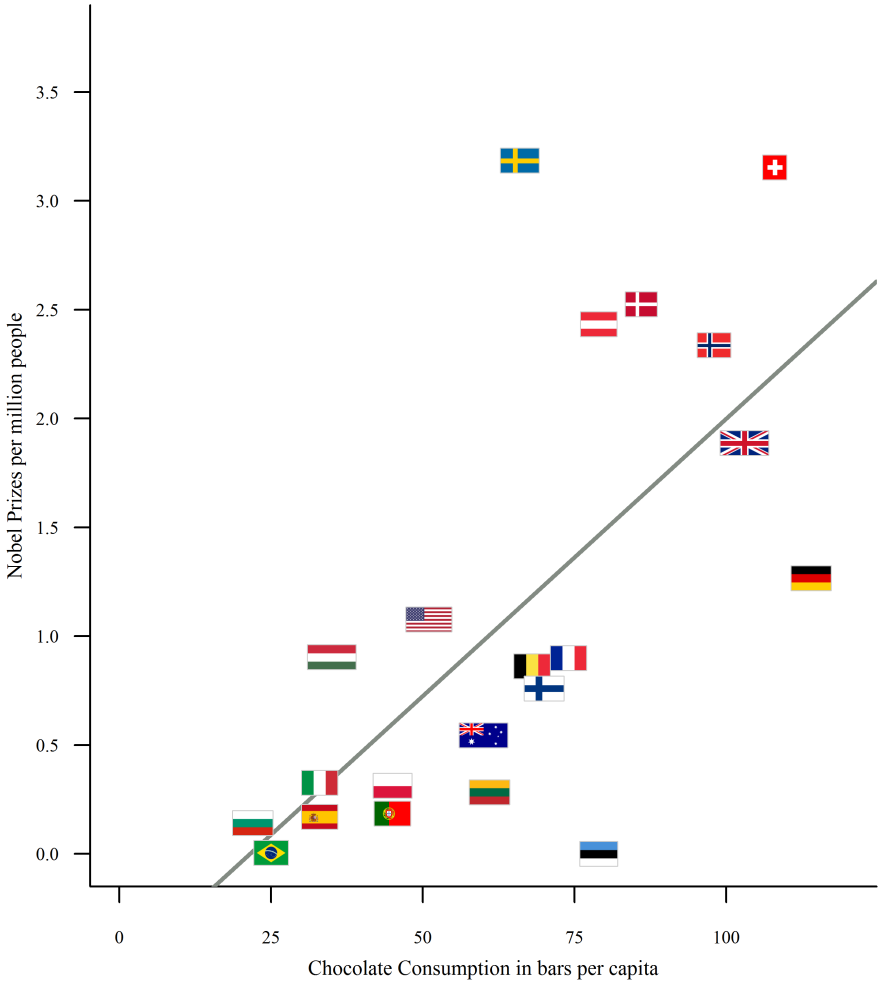

Figure 12.4: *Scatterplot of Nobel Prize rate against chocolate consumption, per capita. Note that the graphic also provides the identity of the State for each point. The State is identified by its flag.*

vides much more information about the relationship. It also suggests that there may by some bias in the results. How does it suggest this? Which countries are represented in the graphic? Do they tend to come from one region more than another? Why were these countries selected? Similarly: Why were other countries left out?

Third, the graphic contains the regression line (next section), which is the best estimate for the relationship between the two variables. Note that

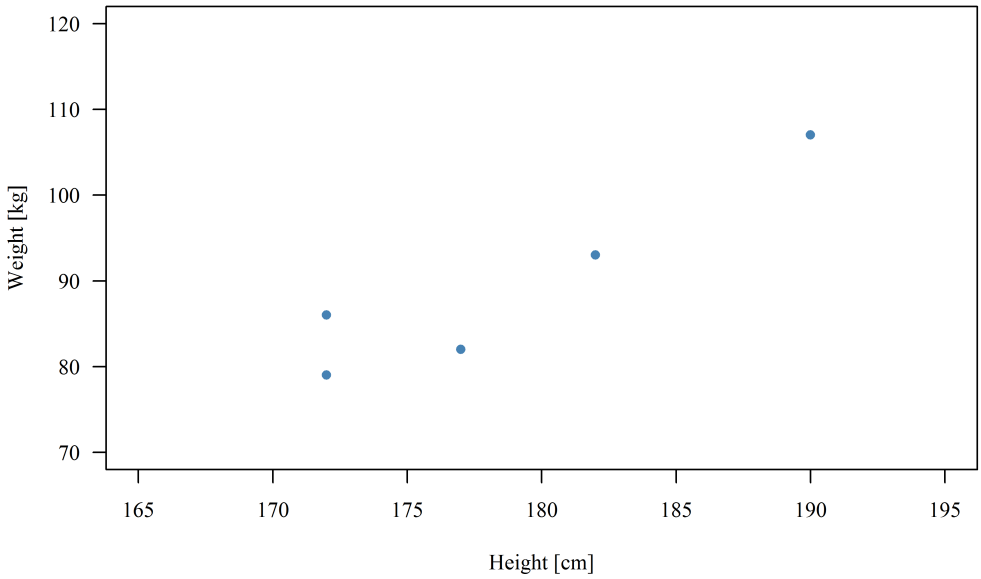

Figure 12.5: *Height and weights of five randomly-selected males.*

the relationship is positive: as chocolate consumption increases, the Nobel Prize rate also tends to increase. This is interesting, no?

12.2: The Method of Ordinary Least Squares

As with earlier statistical tests, the idea for the statistical test arose from a graphic of the data. Here, that graphic is the scatterplot of the last section. In a scatterplot, the marks (usually dots) represent the unit of analysis. The position of the dot on the graphic is due to the values of the two variables measured on the unit.

experimental unit

For instance, let us assume we would like to determine if height and weight are independent of each other for adult males. To test this, I randomly select five adult males from the sampled population. For each, I measure the height (in centimeters) and the weight (in kilograms). Thus, we have repeated measures on the units and we would like to determine the relationship between the two measurements.

random sample

While the data is given below, it is better to first graph the data. This allows you to see if the relationship is linear. It also allows you to formulate

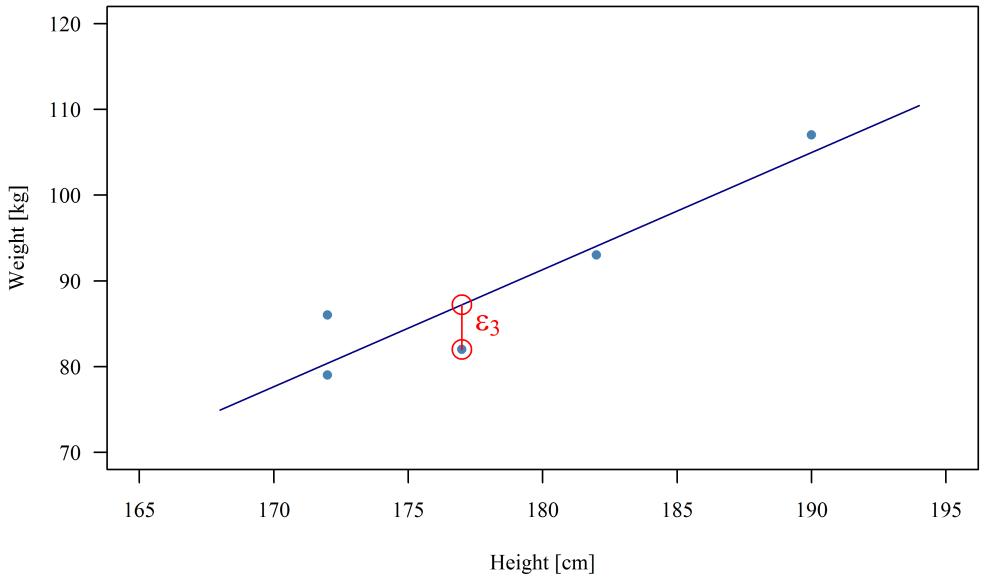

Figure 12.6: *Heights and weights of five randomly-selected males, with the line of best fit summarizing the data. The length of the vertical segment marked ε_3 is the estimation error of the third point.*

expectations on the relationship: Is it a strong relationship? Is it positive? Figure 12.5 is a scatterplot of the data.

Person	Height	Weight
1	172	79
2	172	86
3	177	82
4	182	93
5	190	107

The scatter plot of the data tells the story of the data: as the height increases, the weight also increases, although there are exceptions to this rule. To put numbers on that relationship, we will fit a line to the data. The purpose of this line is to best summarize the bivariate data. How do we do this?

The line of best fit is the line that comes closest to the data *as a whole*.

global

In Figure 12.6, the estimation error of the third point is marked ε_3. Thus, a line of best fit will minimize some function of those estimation errors. By

definition, the OLS line of best fit is the line that minimizes the sum of the square of the estimation errors.[1] Calculating the slope and y-intercept of the OLS line of best fit is straight-forward and follows from the definition of that line:

$$\varepsilon_i = \beta_0 + x_i\beta_1 - y_i \tag{12.1}$$

$$t = \sum_{i=1}^{n} \varepsilon_i^2 \tag{12.2}$$

All we have to do is substitute the five equations of Eqn 12.1 into Eqn 12.2, differentiate the resulting equation with respect to each of the two parameters (β_0, β_1), set the two equations equal to zero, and solve for the two parameters—the usual method for calculating minimums. To wit:

$$t = \sum_{i=1}^{n} (\beta_0 + x_i\beta_1 - y_i)^2$$

Differentiating with respect to β_0 gives

$$\begin{aligned}\frac{\partial}{\partial\beta_0} t &= \sum_{i=1}^{n} 2\,(\beta_0 + x_i\beta_1 - y_i) \\ &= 2n\beta_0 + 2n\overline{x}\beta_1 - 2n\overline{y}\end{aligned}$$

Differentiating the first equation with respect to β_1 gives

$$\begin{aligned}\frac{\partial}{\partial\beta_1} t &= \sum_{i=1}^{n} 2x_i\,(\beta_0 + x_i\beta_1 - y_i) \\ &= 2n\overline{x}\beta_0 + 2n\beta_1 \sum_{i=1}^{n} x_i^2 - 2\sum_{i=1}^{n} x_i y_i\end{aligned}$$

Now, we continue as usual, setting the two equations equal to zero and solving the system of equations for the two parameter estimates. Doing so for the first equation gives

$$\begin{aligned}0 &= 2n\hat{\beta}_0 + 2n\overline{x}\hat{\beta}_1 - 2n\overline{y} \\ \hat{\beta}_0 &= \overline{y} - \hat{\beta}_1\overline{x}\end{aligned}$$

[1] Other estimation methods may use different definitions of the line of best fit. For instance, the least absolute deviations (LAD) line of best fit minimizes the sum of the absolute values of the estimation errors.

The second gives

$$0 = 2n\overline{x}\beta_0 + 2n\beta_1 \sum_{i=1}^{n} x_i^2 - 2\sum_{i=1}^{n} x_i y_i$$

Substitution and algebra give

$$0 = 2n\overline{x}\left(\overline{y} - \hat{\beta}_1\overline{x}\right) + 2n\hat{\beta}_1 \sum_{i=1}^{n} x_i^2 - 2\sum_{i=1}^{n} x_i y_i$$

$$0 = n\overline{xy} - n\overline{x}^2\hat{\beta}_1 + 2n\hat{\beta}_1 \sum_{i=1}^{n} x_i^2 - 2\sum_{i=1}^{n} x_i y_i$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - n\overline{xy}}{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2}$$

Using the definitions of variance and covariance, we have

$$\hat{\beta}_1 = \frac{Cov(x,y)}{Var(x)}$$

This last formula can be written as $\hat{\beta}_1 = r_{xy}\left(\frac{s_y}{s_x}\right)$, where s_x and s_y are the standard deviations of x and y, respectively, and r_{xy} is the observed correlation between x and y.

Thus, for this example, we have

$$\begin{aligned}
\overline{x} &= 178.6 \\
\overline{y} &= 89.4 \\
\sum_{i=1}^{5} x_i y_i &= 80,150 \\
\sum_{i=1}^{5} x_i^2 &= 159,721
\end{aligned}$$

This leads to

$$\begin{aligned}
\hat{\beta}_1 &= \quad 1.3659 \\
\hat{\beta}_0 &= -154.5528
\end{aligned}$$

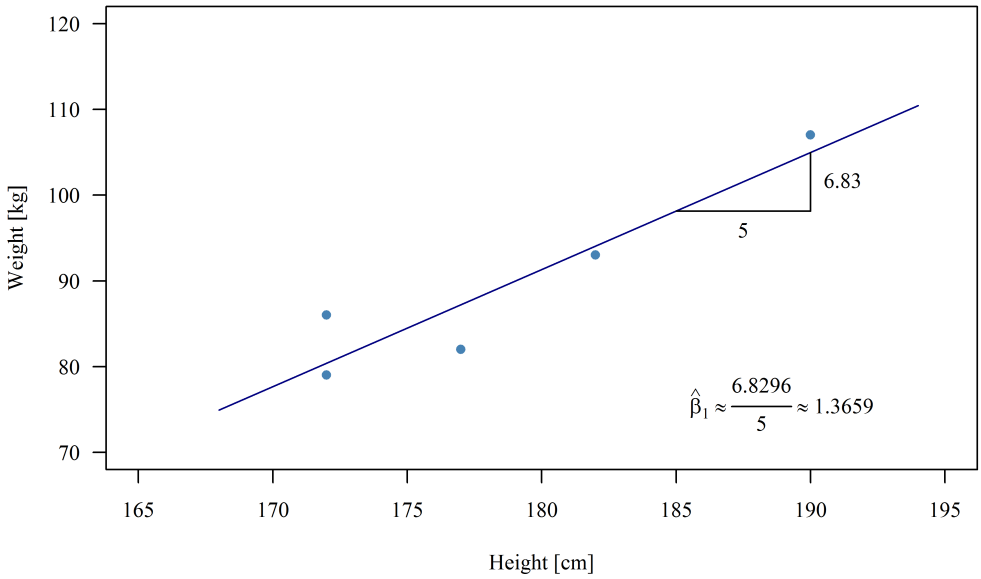

Figure 12.7: *Height and weights of five randomly-selected people, with the line of best fit summarizing the data. The effect estimate is the slope of the line of best fit, as shown in the figure.*

Finally, the OLS line of best fit is

$$\text{weight} = -154.5528 + 1.3659 \text{ height}$$

This line is plotted in Figure 12.6 and in Figure 12.7.

Now, among other things, this tells us that the expected weight of an 182cm tall adult male is $\mathbb{E}[Y \mid x = 182] = -154.5528 + 1.3659(182) = 94.04$ cm. We see that the actual value of y at $x = 182$ is 93, therefore the error is $e_4 = 93 - 94.04 = -1.04$ cm. While this particular error is rather small, especially when compared to the data values, this line is optimal; that is, it *globally* minimizes the (sum of squares of) errors.

effect size

The slope of this line is called the effect size, as it provides the effect on the dependent variable of increasing the value of this independent variable by *one unit*, with the values of all other variables held constant. As this regression line is a line, the slope is constant. Thus, the values of the other independent variables do not affect the effect of this independent variable.

ceteris paribus

The effect size is also the marginal effect of the independent variable on the dependent variable. Figure 12.7 illustrates the effect of an increase of 5cm on the expected weight. The expected increase is 6.83 kg. While this

effect is illustrated for an increase from 185 to 190cm, the effect is constant. Thus, the same weight increase would have happened had we examined a change from 170 to 175cm. Lines have constant slopes; linear models estimate constant effects.

CLM

This constant effect is neither a strength nor a weakness. It is something of which we must be mindful. Does the constant-effect model really represent the data-generating process? If so, it can be used. If not, it should not. The classical linear model describes some reality, not *all* reality.

12.2.1 Matrix representation*

Extending Equations 12.1 and 12.2 to handle multiple independent variables introduces an unnecessary level of complexity if we insist on using the algebraic notation above. We can write the entire system in matrix form (Younger 1979). In matrix form, we need to solve Equation 12.3 for **B**.

$$\mathbf{Y} = \mathbf{XB} + \mathbf{E} \tag{12.3}$$

design matrix

In the equation, **Y** is the vector of response values (values of the dependent variable), **X** is the design matrix (the values of the independent variables with the first column all 1s, corresponding to including the constant term in the regression equation), **E** is the vector of random errors (more on that later), and **B** is the vector of parameter values.

For our simple example, we have the following:

$$\begin{pmatrix} 93 \\ 82 \\ 107 \\ 79 \\ 86 \end{pmatrix} = \begin{pmatrix} 1 & 182 \\ 1 & 177 \\ 1 & 190 \\ 1 & 172 \\ 1 & 172 \end{pmatrix} \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} + \begin{pmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \\ e_5 \end{pmatrix}$$

To solve the matrix equation, we make certain assumptions about **E** to eliminate it from consideration (more on that later) and use some matrix algebra to get Equation 12.4, where **X**′ indicates the transpose of the matrix and $\mathbf{X}^{-1}$ indicates the inverse of the matrix.

$$\mathbf{B} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \tag{12.4}$$

Solving the matrix equations 12.3 and 12.4 gives us the coefficients (**B**) and the residuals (**E**).

$$\mathbf{B} = \begin{pmatrix} -154.552768 \\ 1.365917 \end{pmatrix} \qquad \mathbf{E} = \begin{pmatrix} -1.044118 \\ -5.214533 \\ 2.028547 \\ -1.384948 \\ 5.615052 \end{pmatrix}$$

The advantage to this form is that it looks the same no matter how many independent variables we include; the formulas in Equation 12.1, however, become unwieldy quickly. Not that any of this is important to you and your calculations. That the computer uses the matrix form is between the computer and its operating system.

Note: The *important* things for your understanding are the following:

1. We made two assumptions about the **X** matrix: There is some variation in the independent variables; and the independent variables are not linear combinations of each other.
2. We made one assumption about the **E** vector: The errors have a zero mean.
3. We made one assumption regarding the relationship between the dependent and independent variables: it is linear.

Under these assumptions, we can calculate an expected value for the dependent variable given values for the independent variables. This is called prediction.

12.2.2 Toward a Test

At this point, we have only created the line of best fit, which summarizes the data. Creating that line allows us to create estimates. It does not, however, allow us to determine if the line is *good* at summarizing the data. To do this, we need to have test statistics and distributions for those test statistics. Luckily, if we assume the residuals are Normally distributed, then we can use everything we know about the Normal distribution to create tests and test statistics.

prediction line

test statistics

And so, the rest of this section briefly covers the assumptions of ordinary least squares regression. This will allow us to determine how well the model fits the data and whether the effects are significantly different from zero.

First, let us make the assumption that we know the values of the independent variable *without* error. Second, let us make the assumption that the errors are independent and identically-distributed Normal with mean zero:

$$\varepsilon_i \overset{\text{iid}}{\sim} \mathcal{N}\left(0, \sigma^2\right) \tag{12.5}$$

With these assumptions, we have the distribution of the effect (slope) estimates:

$$\hat{\beta}_1 \overset{\text{iid}}{\sim} \mathcal{N}\left(\beta_1, \sigma^2_{\hat{\beta}_1}\right).$$

However, as was true with the z-test (§5.2), we do not know the population variance. As such, we need to estimate it from the data:

$$s^2_{\hat{\beta}_1} = \frac{\frac{1}{n-2}\sum_{i=1}^{n} e_i^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}.$$

This estimation introduces uncertainty. As before, this means our test statistic will have a t-distribution. Here, that test statistic is $\frac{\hat{\beta}_1 - \beta_1}{s_{\hat{\beta}_1}}$, and its distribution is a t-distribution with $n-2$ degrees of freedom.

Similarly, we have the distribution of the y-intercept as

$$\hat{\beta}_0 \overset{\text{iid}}{\sim} \mathcal{N}\left(\beta_0, \sigma^2_{\hat{\beta}_0}\right).$$

Again, we do not know the population variance. We estimate it using

$$s^2_{\hat{\beta}_0} = \frac{1}{n} s^2_{\hat{\beta}_1} \sum_{i=1}^{n} x_i^2.$$

This test statistic, $\frac{\hat{\beta}_0 - \beta_0}{s_{\hat{\beta}_0}}$ has a t-distribution with $n-2$ degrees of freedom.

> ***Note***: In general, the number of degrees of freedom in linear regression is n minus the number of parameters being estimated in the model, which is the number of independent variables plus one (for the intercept).

Statistical packages do not require you to perform these calculations. In `R`, you will fit the model using either the `lm` or the `glm` command, saving the

results into a variable. Displaying the results in the usual regression table format requires a `summary` command.

The regression table provides estimates for the intercept and for the effect of the independent variable(s). It also provides the standard errors, s_*, the test statistics, t_*, and the p-value of the null hypotheses $\beta_0 = 0$ and for $\beta_1 = 0$.

The regression table for the above example is **Regression Table**

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -154.5528    55.0678  -2.807   0.0675 .
x              1.3659     0.3081   4.433   0.0213 *
```

Note that, at the usual level of significance, the intercept is not significantly different from zero. However, there is significant evidence that height can be used to predict weight; each additional 1cm increase in height tends to result in a 1.3659kg increase in weight, on average. **p-value**

One can get confidence intervals using the `confint` command on the model. As usual, the confidence level can be adjusted using the `level` parameter in the `confint` command. Thus, to obtain 90% confidence intervals, type `confint(model, level=0.90)`. With this example, this results in **confidence interval**

```
                    5 %       95 %
(Intercept) -284.147300 -24.958237
x              0.640829   2.091005
```

Thus, we are 90% confident that the effect of height on weight is between 0.64 and 2.09 kg per cm. We are 5% certain that the actual effect of height on weight is greater than 2.09 kg/cm. We are also 5% certain that the actual effect of height on weight is less than 0.64 kg/cm, which includes negative effects (increased height corresponds to reduced weight).

12.3: Goodness of Fit

Thus far, we have calculated the line of best fit for the data and determined if the relationship(s) between the dependent variable and the independent variable(s) is statistically significant. We have not addressed the question of *how well* the model fits the data. There are two common methods of measuring the goodness of fit of the OLS model to the data. Both are measures **model fits the data**

of error (variation) reduction.[2] Their difference lies in how that variation is calculated.

12.3.1 R-squared Measure

The first way of defining the variation leads to the famous (or infamous) R^2 value. Let us define variation as the average squared distance from the data value to the predicted value; that is, $v = \frac{1}{n}\sum(y_i - \hat{y}_i)^2$.

Without the model, $\hat{y}_i = \overline{y}$. Thus, the variation without the model is $\frac{1}{n}\sum(y_i - \overline{y})^2$, which is close to MST. Recalling the notation of Section 7.2, this original variation is $\frac{1}{n}SST$.

original variation

With the model, the variation is $\frac{1}{n}\sum(y_i - \hat{y}_i)^2$, which is similar to MSE. Again, using the notation of Section 7.2, this *remaining* variation is $\frac{1}{n}SSE$. Thus, the model will have reduced the variation by

remaining variation

$$R^2 := 1 - \frac{\frac{1}{n}SSE}{\frac{1}{n}SST} = 1 - \frac{SSE}{SST} \tag{12.6}$$

Performing the calculation (or allowing R to do so), the R^2 for our model is $R^2 = 1 - \frac{65.8}{65.8+497.2} = 0.8676$. Thus, this model explains 86.76% of the original variation in weights.

The SSE and SST numbers are from the results of the `summary.aov` function:

```
            Df Sum Sq Mean Sq F value Pr(>F)
height       1  431.4   431.4   19.65 0.0213 *
Residuals    3   65.8    21.9
```

0 to 1

The R^2 value ranges between 0 and 1. When $R^2 = 1$, the model perfectly predicts the dependent variable based on the independent variable(s). When $R^2 = 0$, the model is no better than using $\overline{y}$ in predicting y for any value of x.

> ***Note***: One drawback to the R^2 value is that one can increase it simply by adding additional independent variables. This is a drawback because it encourages complex models—science prefers simpler models.

Occam's Razor

[2] The appropriate acronym is PRE, which stands for "Proportional Reduction in Error." PRE measures are measures of how much the model reduces the prediction error of the dependent variable.

12.3.2 Adjusted R-squared Measure

The strength of the R^2 measure is that it is a 'Proportional Reduction in Error' (PRE) measure; that is, we can conclude that this model reduces the unexplained error by 86.76%. This is its strength (and its limit). There is another PRE measure that is commonly used (although not entirely understood by its users).

The 'adjusted R-squared' (Equation 12.7) is a PRE measure for errors measured in terms of variance (Younger 1979). With that said, most people use it to determine whether a specified variable should be kept in the linear model, since it has the effect of adjusting for both the number of independent variables you are using (k) as well as the number of data points in the sample (n).

With the R^2 measure, we defined variation as $v = \frac{1}{n}\sum(y_i - \hat{y}_i)^2$. I had to call it *variation*, because it is not the variance, strictly speaking. The adjusted R^2 measure defines the variation *as* the variance. In other words, if we define MSE as the variance in the residuals and MST as the variance in the original data (see Section 7.2), the *adjusted* R^2 is defined as

$$\overline{R}^2 := 1 - \frac{MSE}{MST}$$

To show its relationship to the R^2 measure, substitution gives

$$\begin{aligned} 1 - \overline{R}^2 &= \frac{SSE/(n-k-1)}{SST/(n-1)} \\ &= \left(\frac{n-1}{n-k-1}\right)\frac{SSE}{SST} \end{aligned}$$

And, we finally have

$$1 - \overline{R}^2 = \left(\frac{n-1}{n-k-1}\right)(1 - R^2) \tag{12.7}$$

In the literature, this measure tends to only be used to determine if the addition or removal of an independent variable is supported.[3] Fortunately, it is quite valid as a PRE measure in its own right.

According to the model summary, $\overline{R}^2 = 0.8234$. We could have calculated this from Formula 12.7 and the fact that $n = 5$ and $k = 1$. Thus, using a

[3] To determine if the variable(s) should be included in the model, compare the adjusted R-squared values. Use the model with a higher adjusted R-squared.

different definition of variation, we can conclude that this model reduces the prediction error by 82.34%.

> ***Note***: There is one major drawback to the $\overline{R}^2$ measure. Where the R^2 value was bounded between 0 and 1, the $\overline{R}^2$ value is not. While it is still bounded above by 1, it *can* take on negative values. This may happen when R^2 is small, because $\frac{n-1}{n-k-1}$ is always greater than 1.

12.4: Maine and the Ballot Measure

To illustrate the process of model selection, let us revisit the framing question for this chapter, that of Maine's ballot measure:

Example 12.1: The voters of Maine are being sent to the polls to vote on a constitutional referendum (ballot measure) that proposes to limit the definition of marriage to the union of one man and one woman. This was not the first time that Americans were sent to the polls to vote on this or a closely related issue. Given the information from previous votes, what is the estimated probability that this ballot measure will pass in Maine?

Before attempting any analysis, there needs to be a search of the literature to inform us as to which variables should be present, and which directions those variables should affect the dependent variable. From that literature review, we hypothesize that the vote in favor of such ballot measures depends on three variables: age of the population, religiosity of the population, and whether the ballot measure also restricts civil unions. The effect direction for each is that States that are more religious should vote against single-sex marriage at a higher rate; Measures that also ban civil unions should have a harder time passing; Measures passed later should have a more difficult chance of passing, as the young tend to support single-sex marriage, and the elderly tend to oppose it.

Directional Hypotheses

With this theory and the resulting hypotheses, we can take our next step: Getting to know the data.

	Year Passed (post-2000)	Civil Ban	Religious Percent
Minimum	-2	0	51.00
Maximum	8	1	85.00
Median	4	1	67.50
Mean	4	0.5938	66.75
Variance	6.0650	0.2490	88.1935
Coefficient of Variation	0.5794	0.8404	0.1407

Table 12.1: *Descriptive statistics on the variables in the* `ssm` *dataset.*

12.4.1 Get to Know the Data

Before we begin trying to answer this question, we must get to know our data. There are several functions available to us to visualize the data: histogram, scatterplots, and quantile-quantile plots. In addition to visualizing the data, we should calculate several of the descriptive statistics for the variables of interest.

Variability: Since we have multiple independent variables, we should calculate both univariate and bivariate descriptive statistics. Table 12.1 provides the univariate descriptive statistics. The primary question to ask about the independent variables here is whether there is sufficient variation. The two measures we need to examine are the variance and the coefficient of variation. If both of these numbers are small, then there may be an issue.

Variation

In this data, the variance of the Civil Ban variable is quite small and potentially worrisome; however, its coefficient of variation (a scaled standard deviation, $c_v = \left| \frac{s}{\bar{x}} \right|$) indicates that there is no issue (the value is close to 1).[4] None of the three variables have small enough variation to cause us concern.

coefficient of variation

Relationships: After getting to know the variables individually, it is important to get to know the relationships between the variables. This can be done through correlation tests and bivariate scatter plots. Independent variables with strong correlations with the dependent variable should be considered for inclusion in the model. Independent variables with strong correlations with other independent variables should be of concern. Remember that one of the assumptions of OLS regression is that the independent variables are independent of each other. If independent variables are highly correlated, the statistical properties of the method weaken.

correlated

[4] As this is a dichotomous variable, the mean is the percent of the values equal to 1. Thus, there are about 60% of the values 1 and 40% of the values 0—more than sufficient variation.

	Year Passed	Civil Ban	Religious Percent
Year Passed	1.0000	0.1903	0.2399
Civil Ban Included	0.1903	1.0000	0.5146
Religious Percent	0.2399	0.5146	1.0000

Table 12.2: *The correlations between the variables in the ssm data. The correlation between Civil Ban and Percent Religious is statistically significant ($t = 3.2869; \nu = 30; p = 0.0026$). This is the sole statistically significant correlation.*

The six pairwise correlations are provided in Table 12.2. Of the three independent variables, only Civil Ban and Religious Percent have a statistically significant correlation ($t = 3.2869; \nu = 30; p = 0.0026$). Should the level of correlation be a concern? Perhaps. While their correlation is $r = 0.5146$, this corresponds to an R^2 value of just 0.2648. As such, the correlation may not be large enough to severely affect our coefficient estimates (see §13.1). Let us just remember this relationship for the future.

> ***Note***: The issue is actually more than just a statistics issue. If two independent variables are highly correlated with each other, it is logically impossible to determine *which* affects the dependent variable or how much of the effect to partition to each independent variable. Statistics is, however, able to tease out the independent relationships better than not. As a rule of thumb, if the correlation is greater than $r = 0.90$, there may be a serious logical issue.

12.4.2 Model the Data

The example asked us to determine the probability that the ballot measure will pass in Maine. Before we can answer that question, we need to model the proportion of the vote in favor of the ballot measure using our independent variables; that is, we need to be able to predict the proportion of the vote in favor of the ballot measure with the information we have.

prediction

Thus, the dependent variable will be `propWin` and the independent variables will be `yearPassed`, `civilBan`, and `religPct`. For now, let us assume a linear relationship between the independent variables and the dependent variable.

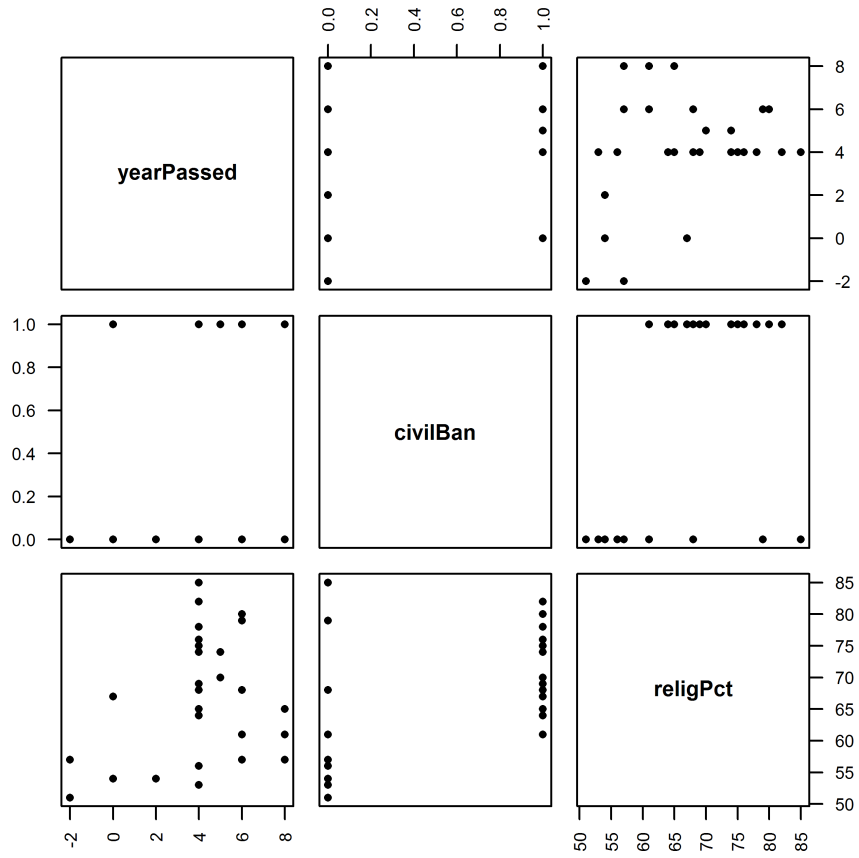

Figure 12.8: *Pairwise plots between the three independent variables. The correlation between* `civilBan` *and* `religPct` *was statistically significant according to the Pearson product-moment correlation test. This is evident in this graph, as well.*

MODEL SELECTION: Unless you have a lot of independent variables, I recommend you start with the interaction model. The interaction model includes the effects of each independent variable singly (main effects) as well as all possible combinations of those variables (interaction effects). **interaction model**

R uses the usual formula grammar (Table 12.3). Its use takes practice. For instance, if you wish to fit the model $y = \beta_0 + \beta_1 x + \varepsilon$, you would use `y ~ x`. **grammar**

~	Separates the dependent variable (left-hand side) and the independent variables (right-hand side)
+	Indicates the following variable is added to the formula
−	Indicates the following variable is removed from the formula
:	Indicates the following and the preceding variable are multiplied in the formula
*	Indicates the following and the preceding variable are crossed in the formula
∧	Includes the specified level of interactions.
I()	Replaces the formula grammar of what is in the parentheses with algebraic grammar.

Table 12.3: *The symbols and their meanings in the grammar of formulas.*

If you wish to fit the model $y = \beta_1 x + \varepsilon$, you would use either `y ~ x - 1` (my usual) or `y ~ x + 0`.

Some other examples include:

Algebraic form	Formula form
$y = \beta_0 + \beta_1 x1 + \beta_2 x2 + \beta_3 x1x2$	`y ~ x1*x2`
$y = \beta_0 + \beta_1 x1 + \beta_2 x2 + \beta_3 x1x2$	`y ~ (x1 + x2)^2`
$y = \beta_0 + \beta_1 x1x2$	`y ~ x1:x2`
$y = \beta_0 + \beta_1 x1^3 + \sin(x2)$	`y ~ I(x1^3) + I(sin(x2))`
$y = \beta_0 + \beta_1 x1 + \beta_2 x2 + \beta_3 x3 + \beta_4 x1x2 + \beta_5 x1x3 + \beta_6 x2x3 + \beta_7 x1x2x3$	`y ~ x1*x2*x3`

With this brief introduction to the grammar of formulas, we can return to our example. We have three independent variables; the formula to give a full interaction model is

asterisk

```
propWin ~ yearPassed * civilBan * religPct
```

As we will use this model a bit, we save the linear regression results into a variable. Thus, the two lines to run are

```
mod1 = lm(propWin ~ yearPassed * civilBan * religPct)
summary.aov(mod1)
```

These lines give output.

The following is the first, fourth, and fifth column of that output:

```
                                t value   Pr(>|t|)
(Intercept)                       1.148      0.262
yearPassed                       -0.901      0.377
civilBan                         -1.084      0.289
religPct                          1.557      0.133
yearPassed:civilBan               0.950      0.352
yearPassed:religPct               0.510      0.615
civilBan:religPct                 0.979      0.338
yearPassed:civilBan:religPct     -0.895      0.379
```

The line starting `yearPassed:civilBan:religPct` is the three-way interaction term. As it is the highest interaction, it is the only one we look at here. Note that it is not statistically significant ($p = 0.379$). Thus, removing that term will do two things. First, it will simplify the model. Second, it will not harm the model's predictive ability.

three-way interaction term

That second model can be written as either

```
mod2 = lm(propWin ~ yearPassed * civilBan * religPct -
   yearPassed:civilBan:religPct)
```

or as

```
mod2 = lm(propWin ~ (yearPassed + civilBan + religPct)^2)
```

The two formulas are equivalent.

formula grammar

Note that the `summary.aov(mod2)` command indicates that none of the three two-way interactions are statistically significant. Thus, these two-way interactions should be removed from the model. This leaves a model with no interactions—an additive model. Fitting the additive model and checking the statistical significance of the variables is as above

two-way interaction terms

additive model

```
mod3 = lm(propWin ~ yearPassed + civilBan + religPct)
summary.aov(mod3)
```

Note that all three variables are significant according to this output. Thus, this is our provisional model.

provisional model

	Estimate	Std. Error	t-value	p-value
Constant Term	0.1512	0.0659	2.293	0.0295
Year Passed (post 2000)	-0.0201	0.0036	-5.618	$\ll 0.0001$
Banned Civil Unions	-0.0373	0.0200	-1.868	0.0723
Percent Religious	0.0095	0.0011	8.801	$\ll 0.0001$

Table 12.4: *Results table for the regression of proportion support of a generic ballot outlawing same-sex marriage against the three included variables. The R^2 for the model is 0.7801; the $\overline{R}^2$, 0.7565. The p-values calculated are based on two-tailed test. The hypotheses were one-tailed hypotheses. As such, all three explanatory variables are statistically significant at the standard level of significance ($\alpha = 0.05$).*

THE ADDITIVE MODEL: That is, the equation we will use to fit the data is

$$\texttt{propWin} = \beta_0 + \beta_1(\texttt{yearPassed}) + \beta_2(\texttt{civilBan}) + \beta_3(\texttt{religPct}) + \varepsilon$$

If $\varepsilon \sim \mathcal{N}(0, \sigma^2)$, then we know

$$\mathbb{E}[\texttt{propWin}] = \beta_0 + \beta_1(\texttt{yearPassed}) + \beta_2(\texttt{civilBan}) + \beta_3(\texttt{religPct})$$

The regression table for model `mod3`, produced using `summary(mod3)`, is given in Table 12.4. Notice that all three variables of interest are statistically significant at the $\alpha = 0.05$ level.[5] Additionally, the model has an $\overline{R}^2$ of 0.7565, which is a great fit in most of the social sciences. The direction of the coefficients also agrees with theory.

directional hypothesis

research hypotheses

prediction line

Thus, the equation for the line of best fit is approximately

$$\begin{aligned}\mathbb{E}[\texttt{propWin}] = 0.1512 \\ &- 0.0201(\texttt{yearPassed}) \\ &- 0.0373(\texttt{civilBan}) \\ &+ 0.0095(\texttt{religPct})\end{aligned}$$

PREDICTING MAINE: According to this model, what is the expected vote in Maine? To answer this, we need information about the Maine ballot measure, specifically the value of the independent variables: `yearPassed` = 9,

[5] You may claim that the Civil Unions variable is not statistically significant at the $\alpha = 0.05$ level. However, these p-values are two-tailed p-values. Our hypotheses were all directional hypotheses (one-tailed). Thus, to get the one-tailed p-values just halve the two-tailed p-values. With that, all three independent variables are statistically significant.

civilBan = 0, religPct = 48. With this information, and under the assumption that the model is correct, we have our prediction that 42% of the Maine voters will vote in favor of this ballot measure.

Thankfully, R does not require us to do this calculation by hand. The R code for predicting the percent of Maine voters voting in favor of this ballot measure can be

```
MAINE = data.frame(yearPassed=9, civilBan=0, religPct=48)
predict(mod3, newdata=MAINE)
```

The first line was used to make the code more readable. It is also helpful to first define the variable MAINE if you are going to make predictions for Maine using several models.

If neither of these appeal to you and you wish to do this in one line, that line would be

```
predict(mod3, newdata=data.frame(yearPassed=9, civilBan=0,
    religPct=48))
```

Note the inclusion of the predict() function, which predicts the dependent variable value given values for each of the independent variables (read the help file on predict; we will use this function frequently).

predict

12.4.3 Checking the Assumptions

Before we can conclude that our prediction is good, we need to determine if our model violates any of the assumptions. However, let us skip this until next chapter. Let us pretend that this model passes the assumption tests. In reality, it does not, but Chapter 14 gives us the tools to fix the issues.

assumptions

12.4.4 Graphing the Results

Now that we have confidence in our model, we can use it to predict the effects of each of the three independent variables on the vote in favor of these ballot measures. There are three independent variables, so we cannot create a single graph that displays the results. However, as one of the variables is dichotomous, we can show the results in just two graphs (the number of continuous independent variables).

Both of these graphs will have the vote in favor as the dependent variable (vertical axis). One of the two graphs will have percent religious as the primary independent variable, whereas the other will have the year passed as the primary independent variable. The civil ban variable will be present in both graphs, signified by two separate curves (Figure 12.9).

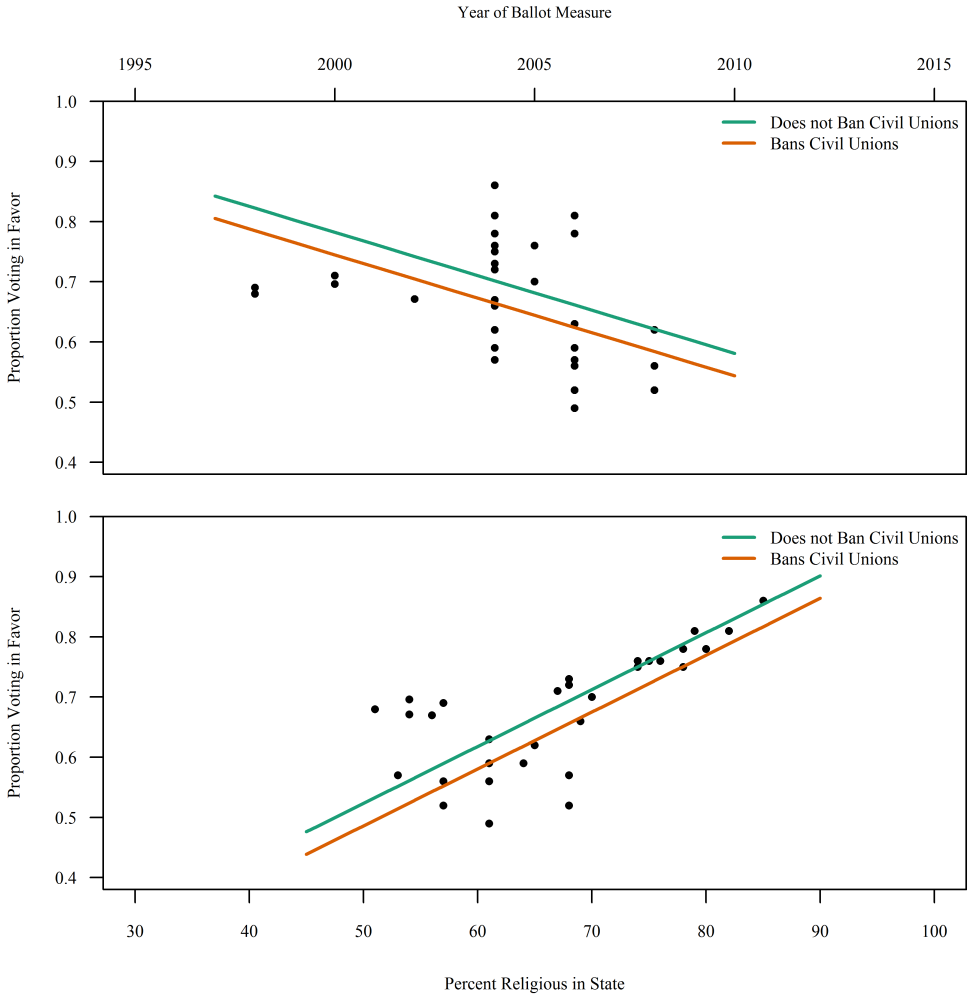

Figure 12.9: *Prediction graphs of our* `ssm` *model. These graphs contain two independent variables plotted against the dependent variable. Note that the effect of each of the independent variables is made manifest by these two graphs.*

The graphs illustrate the results of the model—this is their purpose. Although the graphs "illustrate the story," we must still "tell the story" of the graphics, including numbers from the prediction table (Table 12.4). The following paragraphs explain the graphics.

tell the story

Both graphics show that the effect of adding a civil union ban to the referendum tends to reduce the vote in favor of the refer-

endum. All things being equal, a ballot measure banning civil unions will have 3.7% *fewer* people vote for it than a like measure not banning civil unions ($s = 1.9988, t = -1.87, p = 0.0723$).

The top graphic illustrates the effect of passing time on the proportion of the vote in favor of these referenda: As the year increases by one, the proportion voting in favor of the referendum decreases by 2% on average ($s = 0.3577, t = -5.62, p \ll 0.0001$).

The bottom graphic shows the effect of religiosity on the ballot outcome: those states with higher levels of religiosity tend to vote in favor of these measures at a higher level than states with lower levels of religiosity. In fact, increasing the level of religiosity in the state by 1% will tend to increase the vote in favor of the ballot measure by 0.95% ($s = 0.1074, t = 8.80, p \ll 0.0001$).

Note the interweaving of the graphic discussion with concrete, numerical effects (and statistical significance in parentheses) from the prediction table. This combination aids the reader in interpreting the graphic(s) in terms of statistical language.

regression table

12.4.5 Answering the Question*

Thus, we have a prediction of 42% of the voters will support the ballot measure. However, this is *not* the answer to the original question, which asked about the *probability* of the ballot measure passing. From a modeling standpoint, this probability depends on the coefficient estimates, which are just estimates of the true population value, and the standard errors, which are measures of our certainty in those estimates.

point prediction

In the Ordinary Least Squares method, those parameter estimates are random variables, since they are *functions of the data*. In other words, if we re-ran human history, the estimated effect would be different, since reality would be different. Furthermore, as these are random variables, they have an associated distribution—the Normal distribution. In fact, the distribution of each parameter estimate is Normal, with expected value equal to the estimate and standard deviation equal to the standard error. Thus, for example, the effect of `yearPassed` is $\hat{\beta}_1 \sim \mathcal{N}(\mu = -0.0201, \sigma = 0.0036)$; of `civilBan`, $\hat{\beta}_2 \sim \mathcal{N}(\mu = -0.0373, \sigma = 0.0200)$; and of `pctRelig`, $\hat{\beta}_3 \sim \mathcal{N}(\mu = 0.0095, \sigma = 0.0011)$.

random variable

distribution of estimators

Let us leverage these facts to (virtually) re-run human history several times, get the parameter estimates for each history, and predict the outcome of the ballot measure in Maine. In other words, let us perform a Monte Carlo

Monte Carlo

analysis. The steps are the same as with any Monte Carlo analysis we have done (Kennedy 2008, and §1.4). The only difference is what we do within the loop. Here, we draw random numbers from the appropriate distribution and calculate the predicted vote.

Before you look at the following algorithm, write your own and compare it to the one below:

1. Initialize variables
2. Perform loop
   a) Draw from the four distributions
   b) Predict the Maine outcome
3. Calculate the number of times the ballot measure garnered more than 50% of the vote

One can also store the random numbers inside the loop and predict outside the loop. Also, if the statistical program allows it, you can avoid the loop and just draw all the numbers at once. This last has the advantage of being *very* fast.

It is also the method I use here, in the `R` script:

```
# Initialize variables
  outcome <- numeric()
  trials  <- 1000000

# Coefficient estimates
  b.intc <-  0.151221
  b.year <- -0.020095
  b.cban <- -0.037331
  b.rpct <-  0.009452

# Coefficient standard errors
  s.intc <-  0.065938
  s.year <-  0.003577
  s.cban <-  0.019988
  s.rpct <-  0.001074

# Distributions (the "loop")
  e.intc  <- rnorm(trials, m=b.intc, s=s.intc)
  e.year  <- rnorm(trials, m=b.year, s=s.year)
  e.cban  <- rnorm(trials, m=b.cban, s=s.cban)
  e.rpct  <- rnorm(trials, m=b.rpct, s=s.rpct)
  outcome <- e.intc + e.year*9 + e.cban*0 + e.rpct*48
```

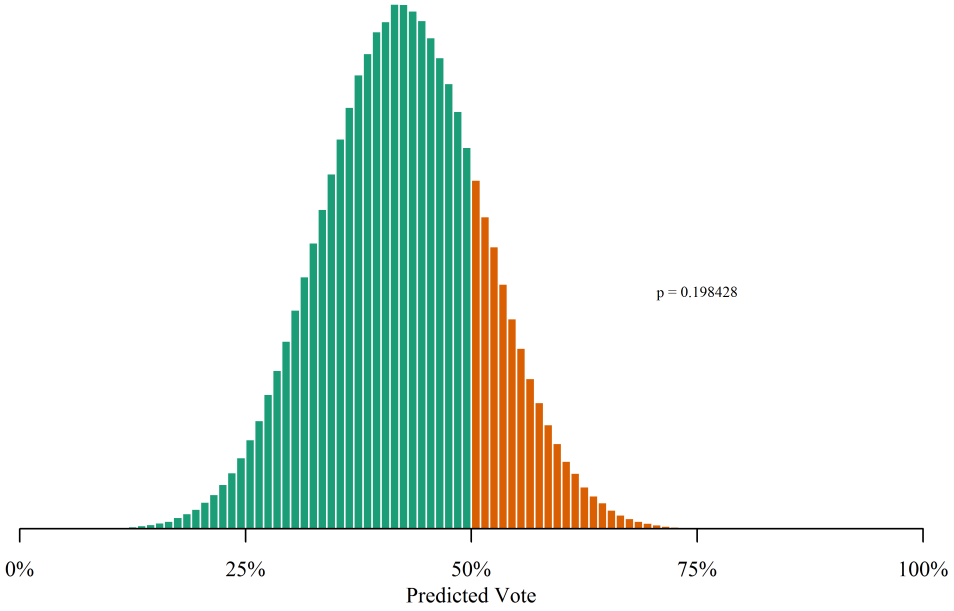

Figure 12.10: *Plot of the predicted vote outcomes from the Monte Carlo experiment described in the text. Note that, while the expected proportion of the vote in favor of the ballot measure is 42%, there is still a 19.86% chance of the ballot measure passing, given that our model is correct.*

At this point, the variable `outcome` holds the proportion of people voting in favor of the ballot measure in one million simulated elections. To answer the question, we just need to determine the proportion of those elections in which the `outcome` is greater than 0.50: `mean(outcome>0.50)` will work.

Of course the numbers are nice, but a histogram may tell a better story. The following code will give a histogram like that in Figure 12.10.

```
hist(outcome, main="", xlab="Proportion Vote for Ballot
   Measure", breaks=-1:99/100)
hist(outcome[outcome>0.50], main="", yaxt="n", breaks=-1:99
   /100, col=2, add=TRUE)
axis(1, at=0.50, labels="50%")
pp = length(which(outcome>0.50))/trials
text(0.75,trials/100, paste("p=",pp), cex=0.7)
```

The histogram of the Maine predictions is presented in Figure 12.10. Note that the expected outcome is still 42%, which we found above, but that there is a spread to that prediction the histogram makes manifest, which

confidence interval

the single prediction did not. In fact, prior to this analysis, we may have concluded that there was no possibility that the ballot measure would pass in Maine based on our model; now, we see that there is a 20% chance of the ballot measure passing.[6]

❧ ❧ ❧

Thus, we have the answer to our original question. Given that our model is correct, there is approximately a 20% chance that the ballot measure to define marriage in terms of one man and one woman will pass in Maine, with a point prediction of 42% in favor of the bill.

point prediction

12.5: Conclusion

In this chapter, we started our entry into what is arguably the most important part of statistical analysis in the real world: Regression—attempting to determine the relationship between continuous variables. The hypotheses we tested in this chapter were straight-forward, and the question we answered was rather interesting, if unorthodox.

We also mentioned the assumptions of ordinary least squares regression, which is the method used to estimate the parameters of our linear model. In the next chapter, we will cover those assumptions in some detail.

As you leave this chapter, please keep in mind three important things.

- First, remember that you must know your data before you can analyze it.
- Second, remember that you must test your assumptions before you can be satisfied with the model (Chapter 13).
- Finally, remember what the numbers in the regression table actually mean (for instance, Table 12.4).

All three of these are extremely important. If you forget any of these, your analyses will be incomplete at best and incorrect at worst.

[6] As with all statistical analysis, the *caveat* is that the model and the assumptions must be correct. In the next chapter, we cover the assumptions of OLS—the method used to estimate the model parameters.

12.6: End of Chapter Materials

12.6.1 R Functions

In this chapter, we were introduced to several R functions that will be useful in the future. These are listed here.

Statistics:

cor(x,y) This function determines the correlation between two vectors of data (of the same length). You may be able to use the shortcut to find the correlations between all of the variables in a data set by using `cor(data)`. Unfortunately, this shortcut ('bug') has been removed in recent versions of R.

cor.test(x,y) This function calculates the correlation between two vectors and performs a parametric test determining whether these two vectors are statistically correlated.

lm(formula) This function performs linear regression on the data, with the supplied formula. As there is much information contained in this function, you will want to save the results in a variable.

predict(model, newdata) As with almost all statistical packages, R has a predict function. It takes two parameters, the model, and a dataframe of the independent values from which you want to predict. If you omit `newdata`, then it will predict based on the independent variables of the data itself, which can be used to calculate residuals. The dataframe must list all independent variables with their associate new values. You can specify multiple new values for a single independent variable.

Graphing:

axis(n, label, at) This adds annotations `label` to the axis located on side `n` (1=x, 2=y) at the provided place, `at`.

lines(x,y) This graphing function adds lines to a currently-open plot, connecting the points described in the vectors `x` and `y`.

text(x,y,t) Adds text `t` to the active plot at position `x`, `y`.

12.6.2 Extensions

This section offers suggestions on things you can practice from this chapter. Save the scripts in your Chapter 12 folder.For each of the following problems, please save the associated `R` script in the chapter folder as `ext0x.R`, where `x` is the problem number.

Summary:

1. How is regression different from t-tests and the analysis of variance procedure of Chapter 7?

2. What is a PRE measure? How is R^2 a PRE measure? How is $\overline{R}^2$ a PRE measure? How can both be PRE measures if their formulas are different?

3. What is the difference between the classical linear model and ordinary least squares?

Data:

4. With the `ssm` datafile, what is the correlation between the year passed and whether the ballot measure banned civil unions? Is it a statistically significant correlation?

5. Note that the percent of the people in Iowa claiming to be religious is 46%. If, in the year 2015, the voters of Iowa are faced with a ballot measure defining marriage as one man plus one woman, but not restricting civil unions, what is the predicted vote in favor of the ballot measure?

6. A different, although related, question is "What is the probability that the ballot measure passes in Iowa?" Note that the previous question concerned the expected vote. This question emphasizes that the previous answer was just an estimate. Here, we need to run many simulated elections and determine the proportion of those simulated elections resulting in a win. Now, estimate the probability that this ballot measure passes in Iowa.

7. For the sake of education, let us assume that Mississippi considered putting such a ballot measure before the people in 1994, including a ban on recognition of civil unions. The percent of people claiming to be religious in Mississippi is 85%. According to the model, what would be

the expected proportion of the vote in favor of the referendum? What would have been the probability of it passing? What is wrong with these estimates?

8. The `crime` datafile has variables for violent crime rate and property crime rate in both 1990 and 2000 (`vcrime90`, `vcrime00`, `pcrime90`, and `pcrime00`). These four variables suggest six different bivariate regressions (one independent variable and one dependent variable). Perform those six regressions. Check the assumptions of ordinary least squares for each of the six regressions. Check the *logic* of each of the six regressions. Settle on a single regression model—the best of the group according to you. Create an appropriate graphic. Add a prediction line to that graphic. Thoroughly explain the results of your analysis in this problem.

Monte Carlo:

9. The `ssm` datafile also contains the variable `churchAttendance`. This variable is the percent of the population who claims to attend church at least once weekly. However, there are no measurements for Alaska and Hawaii, hence the `NA` values. Using the data, create a model and estimate the church attendance percent for Alaska and Hawaii. In the `ssm` data, replace the `NA`s with the values you estimated. Save this new dataset as `ssmx.csv`.

10. Now that you have a full dataset from Problem 9, use church attendance in lieu of state religiosity in a new model, called `model.x`. What is the expected proportion of the vote in favor of the ballot measure in Maine using this dataset?

11. From Problem 10, what is the *probability* that the ballot measure will pass in Maine? Why will your answer be slightly different from that of others?

12.6.3 Applied Research

This section offers some applied research works that are connected with the topics in this chapter.

- Christopher Achen. (1992) "Social psychology, demographic variables, and linear regression: Breaking the iron triangle in voting research." *Political Behavior* **14**(3): 195–211.
- Colleen L. Barry, Victoria L. Brescoll, Kelly D. Brownell, and Mark Schlesinger. (2009) "Obesity Metaphors: How Beliefs about the Causes of Obesity Affect Support for Public Policy." *The Milbank Quarterly* **87**(1): 7–47.
- Ko Maeda. (2010) "Factors behind the Historic Defeat of Japan's Liberal Democratic Party in 2009." *Asian Survey* **50**(5): 888–907.
- Nolan McCarty, Keith T. Poole, and Howard Rosenthal. (2009) "Does Gerrymandering Cause Polarization?" *American Journal of Political Science* **53**(3): 666–680.
- Brian Kelleher Richter, Krislert Samphantharak, and Jeffrey F. Timmons. (2009) "Lobbying and Taxes." *American Journal of Political Science* **53**(4): 893–909.
- Jaime E. Settle, Christopher T. Dawes, and James H. Fowler. (2009) "The Heritability of Partisan Attachment." *Political Research Quarterly* **62**(3): 601–13.
- Jennifer Stuber, Sandro Galea, and Bruce G. Link. (2009) "Stigma and Smoking: The Consequences of Our Good Intentions." *The Social Service Review* **83**(4): 585–609.

12.6.4 References and Additional Readings

This section provides a list of statistical works. Those works cited in the chapter are here. Also here are works that complement the chapter's topics.

- Michael J. Crawley. (2007) *The R Book*. J. Wiley and Sons.
- Rudolf J. Freund and William J. Wilson. (2003) *Statistical Methods*, Second Edition. Academic Press.
- Peter Kennedy. (2008) *A Guide to Econometrics*. Wiley-Blackwell.
- John Maindonald and W. John Braun. (2010) *Data Analysis and Graphics Using R: An Example-Based Approach*, Third Edition. Cambridge University Press.
- Paul Murrell. (2011) *R Graphics*, Second Edition. CRC Press.
- William Navidi. (2008) *Statistics for Engineers and Scientists*, Second Edition. McGraw-Hill.
- R. Lyman Ott and Michael Longnecker. (2010) *An Introduction to Statistical Methods and Data Analysis*, Sixth Edition. Brooks/Cole.
- Mary Sue Younger. (1979) *Handbook for Linear Regression*. Duxbury Press.