4

Multiple Linear Regression Models

We've learned about several aspects of simple LRMs, in particular how to interpret slopes and intercepts, their source equations and assumptions, and some inferential issues. The model introduced in the last chapter considers only the LRM with a single explanatory variable. We'll now extend the model to include more than one explanatory variable, the *multiple LRM*. Many of the same issues exist, but a key difference between simple and multiple LRMs involves the interpretation of the coefficients.

Most simple LRMs are not interesting. In the last chapter, for instance, we assessed the association between average life satisfaction and opioid deaths among states in the U.S. and discovered a negative association. But you don't need to be a sociologist or epidemiologist to realize that other factors also relate to opioid deaths at an aggregate level. Some potential factors include economic conditions, crime rates, substance use rates, the age structure; we could go on. The point is that many potential explanatory variables might be useful for predicting an outcome variable. The selection of these variables should be guided by theoretical and conceptual models, yet limiting LRMs to one explanatory variable is often incomplete and unsatisfactory.

One reason to include other variables in a model—in addition to theoretical concerns—is that they may account for the association between one of the explanatory variables and the outcome variable. This is known as *confounding*, since we say that if variable x_2 accounts for the association between x_1 and y, x_2 is a *confounder* of their association.[1] For example, suppose we're interested in the association between the number of lighters purchased and the rate of lung disease across a sample of U.S. cities. Is there a positive association between these two variables? Probably, but is it fair to conclude that purchasing lighters "causes" or produces lung disease? No, because cigarette smoking is related to both purchasing lighters and lung disease. Cigarette smoking is thus a confounding variable, and the association between the number of lighters purchased and the rate of lung disease is called *spurious*. Smoking should be included in a regression model that predicts lung disease, especially if the model also includes the frequency of lighter purchases. (Smoking should always be in models predicting lung disease.)

[1] Chapter 6 provides a detailed discussion of confounding. Recall, moreover, that one of Mill's criteria for causation is that one should eliminate other explanations or factors that might account for the association between a presumed cause and effect (see Chapter 2, fn. 5). If our goal is to identify causal relationships among variables then surely we must consider a number of factors.

DOI: 10.1201/9781003162230-4

An Example of a Multiple LRM

Like simple LRMs, multiple LRMs are estimated with ordinary least squares (OLS) (see Chapter 3), but adding explanatory variables makes the coefficients and other features of the model more challenging to interpret. Before examining some of these features, let's estimate a multiple LRM in R and examine its output. We'll then learn how to interpret its slope coefficients. The *StateData2018.csv* dataset used in the previous chapter includes measures of the states' number of violent crimes per 100,000 residents (`ViolentCrimeRate`), which we shall treat as the outcome variable, percent of children living in poverty (`PerChildPoverty`), and median household income (`MedHHIncome`). Before estimating an LRM, let's ask R to create a subset of the *StateData2018* dataset that consists of these three variables. We then use this subset to examine the correlations among the variables to see, in a rough sense, the direction and strength of their associations. R has several functions that compute correlations; we'll use the `corr.test` function in the `psych` package since it provides correlations and their *p*-values.

R code

```
library(psych) # activate the psych package
sub.corr <- StateData2018[c("ViolentCrimeRate",
            "PerChildPoverty", "MedHHIncome")]
  # create a subset of the data

corr.test(sub.corr)  # request the correlation matrix
                       using psych's corr.test function
```

R output (annotated and abbreviated)

```
              # Pearson's r #
         Violent PerChildPov MedHHInc
Violent    1.00      0.49      -0.21
ChildPov   0.49      1.00      -0.76
MedHHInc  -0.21     -0.76       1.00
              # p-values[2] #
         Violent PerChildPov MedHHInc
Violent             0.0003    0.1471
ChildPov  0.0003               0.0000
MedHHInc  0.1471    0.0000
```

Violent crimes have a positive correlation with child poverty and a negative correlation with median household income. The latter correlation is not

[2] Two of the *p*-values are listed as zero in the R output. Yet, *p*-values are never zero; R has rounded down a small number. When a *p*-value is listed with several digits but appears as, say, 0.0000, it indicates that it is less than 0.0001. When presenting small *p*-values, they should be listed as $p < 0.001$ or with some similar designation.

below the standard p-value threshold of 0.05, though ($p = 0.15$). Child poverty and median household income also have a substantial negative correlation ($r = –0.76$, $p < 0.001$). If we hypothesize that child poverty and median household income are explanatory variables that predict violent crime rates, we have tentative evidence that the key variables are associated.

What does an LRM indicate about these associations? We'll begin with a simple LRM using the percent of children in poverty as an explanatory variable (see LRM4.1). The interpretation of the slope coefficient should be straightforward:

> Each 1% difference (or increase) in children below the poverty level across states is associated with 13.34 more violent crimes per 100,000 residents.

The p-value suggests that, if the population slope is zero, we'd expect to find a slope coefficient of 13.34 or one farther from zero less than one time out of every 1,000 samples (what are your thoughts on this?). The CI denotes that we may be 95% confident that the population slope representing the association between child poverty and violent crimes falls in the interval {6.49, 20.18}. The intercept indicates that the expected number (mean) of violent crimes per 100,000 is 122 when the percent of children in poverty is zero (is this a reasonable number?). In general, then, the evidence from the statistical model is compatible with a conceptual model asserting a positive association between child poverty and violent crimes.

R code

```
LRM4.1 <- lm(ViolentCrimeRate ~ PerChildPoverty,
             data=StateData2018)
summary(LRM4.1)
confint(LRM4.1))
```

R output (abbreviated)

```
Coefficients:
               Estimate  Std. Error  t value Pr(>|t|)
(Intercept)     122.244     59.556     2.053 0.045587 *
PerChildPoverty  13.335      3.406     3.915 0.000285 ***
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1

Residual standard error: 113.3 on 48 degrees of freedom
Multiple R-squared: 0.242, Adjusted R-squared: 0.2262
F-statistic: 15.33 on 1 and 48 DF, p-value: 0.000285

[CIs]                2.5%     97.5%
(Intercept)         2.498   241.990
PerChildPoverty     6.486    20.184
```

Let's now estimate a multiple LRM by adding median household income to the `lm` function (see LRM4.2). The interpretation of the slope coefficient is simpler if we transform the measurement scale of this variable to $1,000s.

R code

```
StateData2018$MedHHInc <- StateData2018$MedHHIncome/1000
 # transform its measurement scale to $1,000s
LRM4.2 <- lm(ViolentCrimeRate ~ PerChildPoverty +
             MedHHInc, data=StateData2018)
summary(LRM4.2)
confint(LRM4.2)
```

R output (abbreviated)

```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     -310.566   217.913   -1.425  0.160712
PerChildPoverty   21.180     5.038    4.204  0.000116 ***
MedHHInc           4.991     2.423    2.060  0.045008 *
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1

Residual standard error: 109.7 on 47 degrees of freedom
Multiple R-squared: 0.3048, Adjusted R-squared: 0.2752
F-statistic: 10.3 on 2 and 47 DF, p-value: 0.000195

[CIs]               2.5%    97.5%
(Intercept)     -748.950  127.819
PerChildPoverty   11.046   31.314
MedHHInc           0.116    9.866
```

The output lists two slope coefficients, one for each explanatory variable. What does a predictor effects plot (recall Figure 3.9) indicate about the model? Figure 4.1 represents the positive linear association between percent child poverty and predicted violent crimes after statistically *adjusting for* the effects of median household income in $1,000s. Recall that the confidence bands reflect uncertainty in the association.

R code for Figure 4.1

```
library(effects)
plot(predictorEffect("PerChildPoverty", LRM4.2,
     xlevels=60), main="", xlab="Percent child poverty",
     ylab="Predicted violent crimes per 100,000")
  # set median household income at its approximate mean
    of 60 (xlevels=60)
```

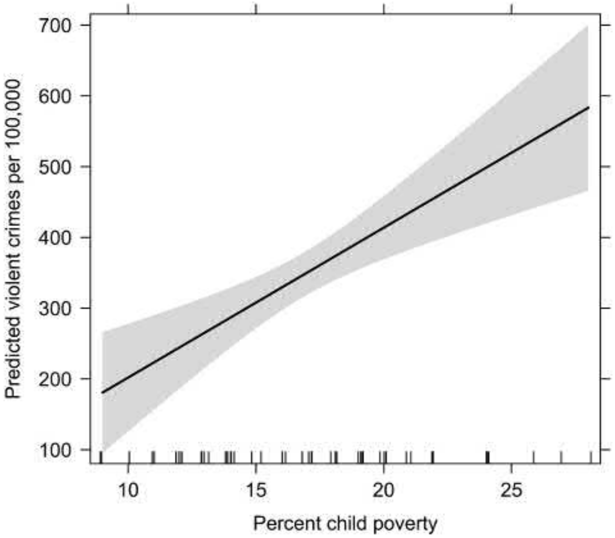

FIGURE 4.1
Predictor effects plot of violent crimes by percent child poverty from LRM4.2.

But what does it mean to claim that the association *adjusts for* the effects of the other explanatory variable? Let's first learn how to interpret the intercept and slope coefficients in the R output and then try to understand how a multiple LRM produces them.

We now have three variables in the model, so we may envision the associations in three dimensions, with the intercept represented as the point on the *y*-axis when the other two variables' axes are at zero (see Figure 4.2). The intercept is thus the expected value of the outcome variable when both of the explanatory variables are zero. If an imaginary state has no child poverty and zero median income, we expect its mean number of violent crimes to be −310.6 per 100,000. The intercept is meaningless, however, since there cannot be a negative number of violent crimes nor states that have no child poverty or zero median income.

The slope coefficients are interpreted in a familiar manner, with a phrase added to each statement. The child poverty slope coefficient is interpreted as:

> Statistically adjusting for the effects of median household income, each 1% difference (or increase) in children living below the poverty level across states is associated with 21.2 more violent crimes per 100,000 residents.

Similar to the interpretation for Figure 4.1, we use the phrase "statistically adjusting for." Some researchers employ alternative phrases such as

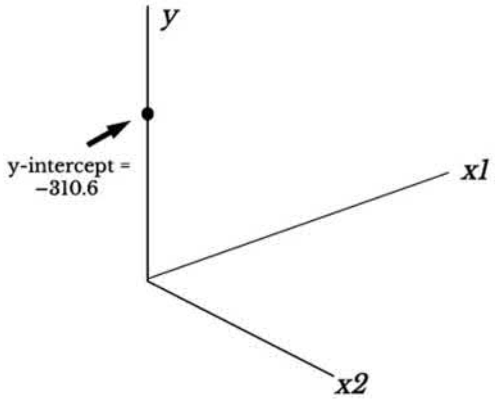

FIGURE 4.2
Illustration of an intercept from a multiple LRM.

"controlling for," "holding constant," or "partialling out." Because we are, presumably, partialling out the effects of a third variable, multiple linear regression coefficients are also called *partial regression coefficients*, *partial slope coefficients*, or *partial slopes*. The terms "controlling for" or "holding constant" may be misleading, however, because they imply that researchers have control over a variable. But "controlling" variables is rare in social and behavioral sciences, especially with observational data. Experimental designs are typically needed if researchers are to control the level of or exposure to a variable (see Chapter 2). For example, in the *StateData2018* dataset, we have no control over—and cannot change—a state's median household income. But, as we'll learn later in the chapter, claiming that we are holding another variable "constant" as we assess the linear association between an explanatory variable and the outcome variable offers a useful way to understand how to interpret an LRM. We'll typically use the phrase "adjusting for," though, when interpreting a model's slope coefficients.

Statistical adjustment can be a difficult concept, but one way to understand it is that we are estimating the slope of one explanatory variable on the outcome regardless of the level of the other explanatory variable (or by setting it at a specific value, such as in Figure 4.1—median household income is set at its mean). For instance, if we claim that each 1% difference in child poverty across states is associated with 21.2 more violent crimes, we're assuming that this occurs for any value of median household income, whether $50,000 or $70,000. Since statistical adjustment is such an essential topic, the following discussion provides four ways to understand it.

The first is designed for those with good spatial perception skills. It utilizes a three-dimensional graph to visualize the relationship among the three variables. R has several options for creating these graphs, including functions available in the packages `plotly` and `rgl`. The following function from the `plotly` package creates a dynamic three-dimensional graph that

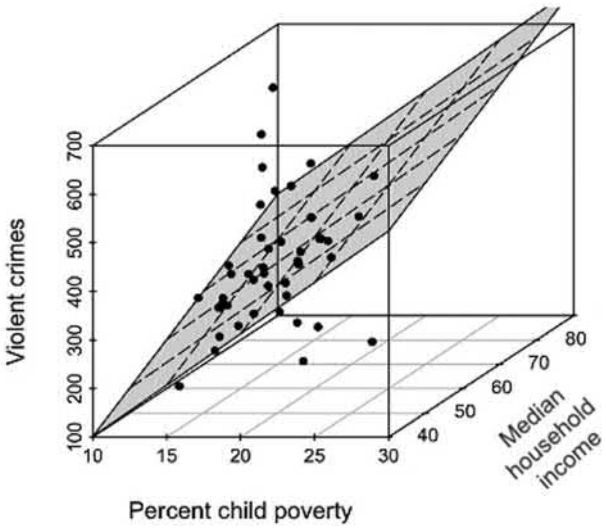

FIGURE 4.3
Three-dimensional representation of a multiple LRM slope.

may be rotated to examine the associations among the three variables from a variety of angles.

```
R code
library(plotly)
plot_ly(StateData2018, x=~PerChildPoverty, y=~MedHHInc,
        z=~ViolentCrimeRate, type="scatter3d",
        mode="markers")
```

The dataset is not particularly large (n = 50), but recognizing patterns in the associations is difficult. If the angle is just right, though, the positive association between child poverty and violent crimes is perceptible. Figure 4.3 offers an alternative depiction of the associations. The plane in gray represents the association between percent child poverty and violent crimes per 100,000 residents, adjusting for the effects of median household income. There's a slight tilt to the plane, but it shows that the regression surface is flat regardless of the level of median household income. A flat surface suggests that the association between percent child poverty and violent crimes does not vary by median household income. The association is also implied by Figure 4.1, though it is not as evident.

A second way to understand statistical adjustment assumes experience with calculus. Suppose y is a function of two variables, x and w, that we wish to include in a regression model. If w is held constant (e.g., $w = w_0$), then y is a function of a single variable x. Its derivative at a particular value of x is called the *partial derivative* of y concerning x, which is represented by

$\frac{\partial y}{\partial x}$ or $\frac{\partial f(x,w)}{\partial x}$ where $y = f(x, w)$. Partial derivatives, which reflect partial slope coefficients, offer a valuable way to understand statistical adjustment. But this book assumes no background in calculus, so we won't go into more detail. Many calculus textbooks and online tools include graphical depictions of partial derivatives that allow one to explore the notion of holding one variable constant while allowing another to vary, thus providing an illustration of multiple LRM coefficients.[3]

A third method involves computing residuals from two distinct LRMs and then using these residuals to compute the partial slope coefficient. As discussed in Chapter 3, residuals, which are computed as $\hat{\varepsilon}_i = (y_i - \hat{y}_i)$, gauge the vertical distance from the observed y values to the predicted values ($\hat{y}_i$) represented by the linear fit line (see Figure 3.7). Residuals also measure the variation that is left over in an outcome variable after accounting for the systematic part that is associated with an explanatory variable. Part of the remaining variation may be associated with another explanatory variable. This part is represented by the partial slope coefficient.

To illustrate how residuals can help us understand statistical adjustment, consider the following steps in R:

1. Estimate a simple LRM with violent crimes as the outcome variable and median household income as the explanatory variable (leave child poverty out of the model). Call this model `resid1`:

   ```
   resid1 <- lm(ViolentCrimeRate ~ MedHHInc, data =
                StateData2018)
   ```

   Save the residuals from this model in a new R object called `resid1a`:

   ```
   resid1a <- resid1$residuals
   ```

   The residuals (`resid1a`) measure the variability in violent crimes not accounted for by median household income. Review them using RStudio's `View` option or in the `Global Environment` window.

2. Estimate a second LRM with child poverty as the outcome variable and median household income as the explanatory variable. Call this model `resid2`:

   ```
   resid2 <- lm(PerChildPoverty ~ MedHHInc, data =
                StateData2018)
   ```

   Save the residuals from this model in a new object called `resid2a`:

   ```
   resid2a <- resid2$residuals
   ```

[3] See, for example, Robert P. Gilbert et al. (2020), *Multivariable Calculus with Mathematica*, Boca Raton, FL: Chapman and Hall/CRC Press.

These residuals (resid2a) measure the variability in child poverty that is not accounted for by median household income.

3. Estimate an LRM that uses the residuals from step 1 (resid1a) as the outcome variable and the residuals from step 2 (resid2a) as the explanatory variable:

   ```
   summary(lm(resid1a ~ resid2a))
   ```

   There's no reason to examine the residuals from this model.

The slope coefficient for resid2a from the third model is 21.18, which is the same number as the partial slope coefficient for child poverty from LRM4.2.[4] The slopes are identical because they represent the same thing: the shared variability between child poverty and violent crimes that does not involve median household income.

The final method that helps illustrate statistical adjustment is similar to the previous one, except it's visual. Figure 4.4 shows three overlapping circles labeled y, x_1, and x_2 because they represent three variables in an LRM. The total area of each circle represents the variable's dispersion, such as its sum of squares or variance. The overlapping areas symbolize their joint variability or covariance. The cross-hatched area is the overlap between y and x_1 that does not include the circle representing x_2. This area represents the joint

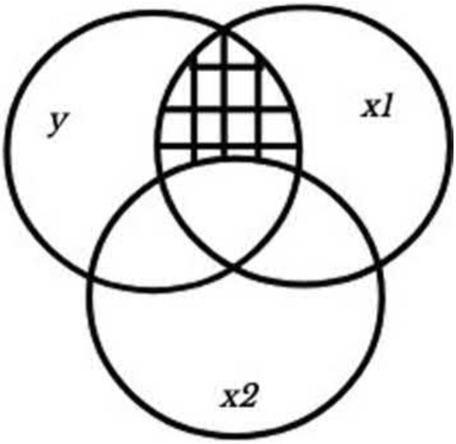

FIGURE 4.4
Overlapping variability of three variables to represent statistical adjustment.

[4] Don't get confused by R's use of scientific notation to represent the slope coefficient in the LRM that uses the residuals. The value 2.118e + 01 is 2.118 × 10 = 21.18. R uses scientific notation when a number has a large number of digits before or after the decimal place. If you don't care for scientific notation, you may turn it off using the function `options(scipen = 999)`.

variability of y and x_1 that is not accounted for by the variable x_2. It thus signifies the partial slope coefficient for x_1 from a multiple LRM.

As Figure 4.4 demonstrates, the explanatory variables, x_1 and x_2, do not need to be completely independent; they may covary. As mentioned in Chapter 3, some researchers use the term *independent variable* to describe explanatory variables in LRMs, even though this can be misleading. The explanatory variables are not independent of one another, but, in a multiple LRM, they are interpreted as *independently* predicting the outcome variable (though in some models we may examine their joint association with the outcome—see Chapter 11's discussion of interaction terms).

Multiple LRMs with more than two explanatory variables require no deeper level of understanding than what we now possess. We should continue to mention the additional explanatory variables when interpreting slope coefficients; for instance, "statistically adjusting for the effects of the other variables in the model, each one-unit increase/difference in variable x_1 is associated with a [partial slope, $\hat{\beta}_1$] unit increase/difference in variable y." But remember that a more precise interpretation uses the units of the variables, such as percentages, dollars, pounds, centimeters, and so forth (e.g., "each $1,000 difference in median household income is associated with …").

Comparing Slope Coefficients

One of the aims of multiple LRMs is to determine which explanatory variable is the best (linear) predictor of or has the strongest statistical association with the outcome. Because many explanatory variables are measured in different units, however, comparing the size of slope coefficients directly is rarely appropriate.[5] For example, in LRM4.2, median household income's slope coefficient is 4.99 and child poverty's is 21.18. Does child poverty have a stronger association with violent crimes? Is it a better predictor? Consider the generic interpretation of the slope coefficient: "a *one-unit* difference in x is associated with a $\hat{\beta}$-unit difference or increase/decrease in y." A one-unit difference is not the same for the two explanatory variables in LRM4.2, though, because one variable is measured in percentages and the other is measured in $1,000s. We should not attempt to compare coefficients based

[5] Some researchers also use p-values to compare the relative predictive strength of explanatory variables. If one variable's p-value is, say, below 0.05 and another's is above 0.05, the temptation is to consider the latter as "more significant" or more strongly related than the other to the outcome. Avoid this—it is not appropriate (see Andrew Gelman and Hal Stern (2006), "The Difference Between 'Significant' and 'Not Significant' Is Not Itself Statistically Significant," *American Statistician* 60(4): 328–331, and Adeline Lo et al. (2015), "Why Significant Variables Aren't Automatically Good Predictors," *Proceedings of the National Academy of Sciences* 112(45): 13892–13897).

on what R provides in its `summary` function unless the x variables are measured in the same manner (e.g., dollars, percentages).[6]

Researchers have developed several techniques to compare the predictive strength of explanatory variables that are measured in different units. Two common methods fall under the classification of *effect sizes*. An effect size measures the strength of the statistical association between two variables on a numeric scale, typically a standardized scale. Effect sizes for LRMs include correlations and standardized slope coefficients.[7]

The simplest effect size measure for LRMs utilizes a bivariate correlation matrix and identifies the largest Pearson's correlation between an explanatory variable and the outcome variable. Inspecting the correlation matrix at the beginning of the chapter, for instance, it appears that child poverty has a larger correlation with violent crimes than median household income (r = 0.49 vs. −0.21), so we might view it as a stronger predictor. Using bivariate correlations can be misleading, though, because they don't account for the associations among the explanatory variables (see Figure 4.4). Once these are considered, the association between the x and y variables might change. For instance, the Pearson's correlation between median household income and violent crimes is −0.21 with a p-value of 0.15. But the partial regression slope coefficient for this variable is 4.99 with a p-value of 0.045. Not only does the 95% CI for the slope coefficient not include zero {0.12, 9.87} and the p-value (0.045) fall below the threshold of 0.05, but, in contrast to the correlation, the slope is positive.

Some researchers prefer to use *standardized slope coefficients* as effect sizes to compare associations in LRMs. Their relationship to unstandardized slope coefficients—those provided by R's `summary` function—is shown in Equation 4.1.

$$\text{Standardized slope}\left(\hat{\beta}_k^*\right) = \hat{\beta}_k\left(\frac{s_{x_k}}{s_y}\right) \tag{4.1}$$

The term s_{xk} denotes the standard deviation of x for variable k. The s_y term is the standard deviation of y. Based on LRM4.2, the standardized slope coefficient for child poverty is computed in Equation 4.2.

$$21.18 \times \left(\frac{4.75}{128.82}\right) = 0.78 \tag{4.2}$$

[6] Chapter 5 provides an example of comparing slope coefficients from variables measured in the same way (see LRM5.3).

[7] The best-known effect size metric is called *Cohen's d*, which measures standardized differences in means (see Jacob Cohen (1988), *Statistical Power Analysis for the Behavioral Sciences*, 2nd Ed., Lawrence Erlbaum). Cohen's d is calculated by taking the difference between the means from two groups and dividing by the pooled standard deviation. A d value of $|0.5|$ or larger is considered a substantial effect size. This metric has been adapted for use with LRMs but is usually based on comparing the explained variance (R^2s) (see Chapter 5) from separate models.

R provides standardized slope coefficients with the `lm.beta` package. After re-estimating the original multiple LRM, the R code for LRM4.3.beta demonstrates how to request these coefficients.

R code

```
library(lm.beta)
LRM4.3 <- lm(ViolentCrimeRate ~ PerChildPoverty +
             MedHHInc, data=StateData2018)
LRM4.3.beta <- lm.beta(LRM4.3) # use the lm.beta
                                 function
print(LRM4.3.beta) # print the standardized coefficients
```

R output (abbreviated)

```
Standardized Coefficients:
  (Intercept) ChildPoverty  MedHHInc
     0.000       0.781        0.383
```

Whether computed by hand, calculator, or with the `lm.beta` function, the standardized slope coefficients are 0.78 (child poverty) and 0.38 (median household income).[8]

These coefficients, which are also called *beta weights*, are interpreted using standard deviations. Another way to understand them is to imagine transforming the explanatory variables and the outcome variable into *z*-scores, estimating the LRM, and inspecting the slope coefficients. They thus represent the association between variables in standard deviation units and are identical to the `Standardized Coefficients` produced by R's `lm.test` function. For example, the interpretation of median household income's beta weight in LRM4.3.beta is

> Statistically adjusting for the effects of child poverty, each one standard deviation difference (increase) in median household income is associated with 0.38 standard deviation-unit additional violent crimes per 100,000 residents.

Some researchers prefer to use and report beta weights because they argue that this allows direct comparisons of the slopes' magnitudes within the same LRM. They maintain that if one explanatory variable has a beta weight farther from zero than another (e.g., $\left|\hat{\beta}^{*}\right|(child\,poverty) > \left|\hat{\beta}^{*}\right|(median\,income)$), it also has a stronger association with the outcome variable. This assumes, though, that the distributions of the explanatory variables are similar. But one variable might be more skewed than another, so standard deviation

[8] Recall from Chapter 2 that the correlation is a standardized version of the covariance. In a simple LRM, the standardized slope coefficient is the Pearson's correlation between x and y. In a multiple LRM, they are not equivalent, however, at least not in simple form, but instead are called *partial correlations* ($cor(x, y \mid \mathbf{z})$). R's `ppcor` package computes partial correlations.

shifts in the two variables are not equivalent. (Are the two explanatory variables in LRM4.3 distributed similarly?) Beta weights also have no clear interpretation for indicator (binary) variables (see Chapter 7) and are therefore a limited way to compare the strength of the associations between particular explanatory variables and the outcome variable.[9]

Slope coefficients from LRMs may be compared with several other metrics. The R package `relaimpo` (*relative importance*) furnishes five different measures designed to identify the strongest predictor in an LRM.[10] The following example extends LRM4.3 by adding two explanatory variables—the state's unemployment rate and the percent of residents without health insurance—and then uses the package's `calc.relimp` function to compute the relative importance metrics (see LRM4.4 and the R code that follows).

R code

```
library(relaimpo)
LRM4.4 <- lm(ViolentCrimeRate ~ PerChildPoverty +
             MedHHInc + UnemployRate + PercentUninsured,
             data=StateData2018)
summary(LRM4.4)
X.compare <- calc.relimp(LRM4.4, type = c("lmg",
                         "first", "last", "betasq",
                         "pratt")) # requests five
                         metrics
X.compare       # outputs the five metrics
plot(X.compare) # provides bar graphs of the results
                  (results not shown)
```

R output (abbreviated)

```
                 Estimate  Std. Error t value Pr(>|t|)
(Intercept)      -398.488  219.041   -1.819   0.0755 .
PerChildPoverty    15.418    5.913    2.608   0.0123 *
MedHHInc            4.688    2.579    1.818   0.0758 .
UnemployRate       27.633   23.150    1.194   0.2389
PercentUninsured   10.313    4.722    2.184   0.0342 *
---
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1

Response variable: ViolentCrimeRate
Total response variance: 16594.39
```

[9] Standardized slope coefficients are helpful when the explanatory variables one wishes to compare are (a) continuous and (b) normally distributed (or distributed in a similar manner). In other situations, they should be used with caution.

[10] For information about this R package and the metrics it estimates, see Ulrike Gromping (2006), "Relative Importance for Linear Regression in R: The Package relaimpo," *Journal of Statistical Software* 17(1): 1–27.

```
Analysis based on 50 observations

4 Regressors:
PerChildPoverty MedHHInc UnemployRate PercentUninsured
Proportion of variance explained by model: 39.44%
Metrics are not normalized (rela=FALSE).

Relative importance metrics:
             lmg   last   first  betasq  pratt
PerChildPov 0.155  0.092  0.242   0.324  0.280
MedHHInc    0.039  0.044  0.043   0.130 -0.075
Unemploy    0.106  0.019  0.200   0.029  0.077
Uninsured   0.094  0.064  0.158   0.080  0.113
```

The metrics are based on decomposing how much of the variance in the outcome variable is attributable to the explanatory variables. The metric `first` is the least rigorous since it uses simple LRMs to compare the strength of the associations, similar to comparing Pearson's correlations. The metric `last` examines how much more of the variance in *y* is accounted for by each explanatory variable after the others are in an LRM. The metrics `lmg` and `pratt` are based on more complex approaches, whereas `betasq` is the beta weight squared (see LRM4.3.beta).[11]

The results favor child poverty as the strongest predictor of violent crimes per 100,000 residents, with the largest value in each of the relative importance metrics. The unemployment rate is the second largest value in two of the metrics (`lmg` and `first`), whereas percent uninsured is the second largest in two others (`last` and `pratt`). Median household income is the second largest value in the beta-squared (`betasq`) column, but since it's the squared value of its beta weight, it has limitations as a comparison tool.

We should now ask why child poverty is the strongest predictor of violent crimes. Is this what we hypothesized? If we assume a sample, have we taken into consideration sampling error when determining the strongest predictor? Examine the documentation for the `relaimpo` package to see if it provides CIs for its measures. CIs might be useful for making a more exhaustive determination about the explanatory variables' predictive power.

Another approach that helps us understand the practical—in addition to the statistical—significance of LRM results is to compare predicted means of the outcome for different levels of the explanatory variables. For example, what is the difference in the predicted means of violent crimes based on the results of LRM4.3? One tactic is to compare differences based on quantiles of the explanatory variables. We might choose, for instance, to compute predicted means at the 25th and 75th percentiles of the explanatory variables

[11] Gromping (2006), *op. cit.*, and Johnson and LeBreton (2004) recommend using `lmg` to compare slope coefficients ("History and Use of Relative Importance Indices in Organizational Research," *Organizational Research Methods* 7(3): 238–257).

TABLE 4.1

Predicted violent crimes per 100,000 population

	25th percentile	75th percentile	Raw difference	Percentage difference (%)
Percent child poverty	260	408	148	57
Median household income	304	377	73	24

and determine how many violent crimes are expected for each percentile. The 25th and 75th percentiles are 13% and 30% for child poverty and 53.1% and 67.7% for median household income (in $1,000s). Based on the following code that utilizes R's `predict` function, the predicted numbers of violent crimes for these percentiles are provided in Table 4.1.[12]

R code

```
LRM4.3a <- data.frame(PerChildPoverty=c(13, 30),
           MedHHInc=59.2)
predict(LRM4.3, LRM4.3a)
LRM4.3b <- data.frame(PerChildPoverty=16.5,
           MedHHInc=c(53.1, 67.7))
predict(LRM4.3, LRM4.3b)
```

The fourth and fifth columns suggest that differences in percent childhood poverty are associated with larger differences in predicted violent crimes. Violent crimes per 100,000 residents are expected to be 57% higher in states at the 75th percentile of child poverty compared to those at the 25th percentile. The corresponding difference based on median household income is 24%. This result reinforces the notion that median household income's association with violent crimes is weaker. But compare the results to what Figure 4.1 estimates for child poverty relative to what a similar graph indicates for median household income. Is there more uncertainty at lower or higher levels of the explanatory variables? Does this affect your interpretations?

Although comparing predicted means is helpful, the best approach for understanding the results of an LRM and its explanatory capabilities is to begin before estimation with a conceptual model based on previous research and theoretical concerns. After estimating the LRM, combine evidence from the coefficients (magnitudes, CIs, predictor effects plots), relative importance metrics, and percentage differences in predicted outcomes to reach sensible conclusions about the associations. These careful steps will allow a

[12] The R code sets one explanatory variable at its median—median household income → 59.2; percent child poverty → 16.5—and estimates the two predicted means based on the other. Chapter 7 provides another example of comparing predicted means.

reasonable determination of whether or not an LRM's results are compatible with a predetermined conceptual model or hypothesis.

Assumptions of Multiple LRMs

We now have a basic understanding of simple and multiple LRMs, including how to interpret their slope coefficients, intercepts, p-values, and CIs, as well as some of their practical implications. Let's next consider the assumptions of the model. We learned about some of these assumptions in Chapter 3. We'll revisit them, introduce a new one, and mention some special situations when they might not be satisfied. The following is a brief overview because several of the subsequent chapters examine the assumptions in detail.

1. *Independence*: the errors of prediction are independent of one another. This affects the bias and efficiency of the estimates. We can understand this assumption better now that we've examined a couple of LRMs. For example, when analyzing state-level data, we assume the errors of prediction are independent across states. But is this true? States that share borders are similar in many ways relative to states that are far apart. The errors in prediction are likely to be similar in adjacent states but different in states that are far away from one another. When we collect data over time, errors of prediction from those time points closer together are usually more alike than those farther apart. A simple way to predict whether the errors of prediction are not independent is when the units of observation in a dataset are also not independent. In this situation, the careful researcher will take steps to address the likely dependence of the errors. Chapter 8 provides more information about the independence assumption.
2. *Homoscedasticity* (*constant variance*): the variance of the errors is constant for all combinations of Xs. Homoscedasticity means "same scatter." Its antonym is *heteroscedasticity* ("different scatter"). This important assumption, when not satisfied, has implications for the efficiency of the LRM slope coefficients. We'll learn more about this critical issue in Chapter 9.
3. *Collinearity*: no combination of the Xs has a perfect association—they are not *perfectly collinear.* This assumption is not listed in Chapter 3 since the simple LRM includes only one explanatory variable. Revisit the overlapping circles in Figure 4.4 and you can visualize what collinearity implies. Suppose circles x_1 and x_2 overlap completely. Is it possible to estimate the covariance between x_1 and y exclusive of x_2? No, because no variability is left over in x_1 once we consider its

association with x_2. We'll discover in Chapter 10 that even a higher degree of collinearity among explanatory variables can lead to unusual results in LRMs.

4. *Normality*: the errors are a normally distributed random variable with a mean of zero. This statement includes two assumptions—normality and mean of zero—that are combined for convenience. The first part is considered a weak assumption since, even when contravened, the model performs fairly well, especially in large samples.[13] The second part—that the mean is zero—is important for estimating the correct intercept. As we learned earlier, though, the intercept in many models is of little use since its value often falls outside the range of the explanatory and outcome variables. Chapter 11 provides a detailed discussion of the normality assumption.[14]
5. *Linearity*: the mean value of Y for each specific combination of the Xs is a linear function of the Xs. In other words, the regression surface is assumed flat in three dimensions (see Figure 4.3). In simple LRMs we assume straight-line relationships in two dimensions, but multiple LRMs include two or more explanatory variables so we must move to higher dimensions. One way to understand this is to imagine three-dimensional space and then visualize the difference between a flat surface and a curved surface. We assume that the relationship between the Xs and Y is not curved. In Chapter 11, we'll learn about some tools for analyzing relationships that are not linear.

These assumptions may not be satisfied in several situations that are not uncommon in research applications, including the following:

1. *Specification error*: we assume that the covariance (or correlation) between each explanatory variable and the errors of prediction is zero, which is symbolized $\text{cov}(\mathbf{X}, \varepsilon_i) = 0$. A nonzero covariance suggests we've left something important out of the model and might reach the wrong conclusions. In Chapter 12, we'll learn that this problem involves whether we have *specified* the correct model, hence the term specification error. Its occurrence contravenes the independence assumption. Since this issue has such important implications for the way empirical models are developed and tested, it warrants

[13] A large sample for LRMs has about 100 or more observations, though the implications for the normality assumption also depend on the degree of non-normality and the number of explanatory variables (see Thomas Lumley et al. (2002), "The Importance of the Normality Assumption in Large Public Health Data Sets," *Annual Review of Public Health* 23(1): 151–169).

[14] As mentioned in Chapter 3, fn. 7, some researchers rearrange and combine aspects of some assumptions with the *iid* assumption. In this presentation, we'll adhere to the distinctions listed here.

its own chapter. Another type of specification error arises when nonlinear associations exist among the *X*s and *Y*, so we have not satisfied the linearity assumption (see Chapter 11).

2. *Measurement errors*: we presume that the *x*s and *y* are measured without error. Lowercase letters designate the explanatory and outcome variables because we are concerned primarily with sample measures (even though the same concerns apply to population measures). As the name implies, knowing whether we have accurate measures of the variables is crucial. When we ask people to record their family incomes, do they provide correct information? When we ask people whether they are happy, might some interpret this question differently than others? As we'll learn in Chapter 13, measurement error is a special problem since it involves the independence assumption in general and is a type of specification error in particular. Claiming that this problem is a common and substantial nuisance in social and behavioral sciences is not an exaggeration.
3. *Influential observations*: we should evaluate discrepant or extreme values and whether they affect the model. They can affect whether the model satisfies several of the assumptions, including linearity, normality, and homoscedasticity. Suppose we measure annual income and find the following values: $25,000, $35,000, $50,000, $75,000, and $33,000,000. As we'll see in Chapter 14, the last entry is labeled either a *high leverage point* (if income is an explanatory variable) or an *outlier* (if income is the outcome variable). High leverage points and outliers—which are known collectively as influential observations—can affect, sometimes in untoward ways, LRM results. The question when confronted with such values is why they have occurred. Did someone record a wrong number, perhaps by placing a decimal place in the wrong spot? Or is the value accurate? Does someone in the sample earn that much money per year? If the value is a coding error, there's an easy fix. But extreme values that are measured accurately tend to have a disproportionate effect on LRMs, so they require our attention.

Evaluating assumptions of LRMs and their derivative issues involves two concerns: (1) how to test whether the assumptions are satisfied and (2) what to do if they are not. Diagnostic tests, which are known collectively as *regression diagnostics*, are available for each of them. Methods to adjust the model are available if one or more assumptions are contravened. If these don't work, alternative regression models exist. Subsequent chapters discuss diagnostic tests and various solutions.

Some of these assumptions are stringent, but, one argument goes, the models are often saved by statistical theory's notable *Central Limit Theorem* (CLT). This occurs, for example, when the normality of errors assumption is not met, yet we have a large random sample. As you might recall from an introductory statistics book or course, the CLT states that, for relatively large samples, the sampling distribution of the mean of a variable is approximately normally distributed even if the distribution of the underlying variable is not normally distributed.[15] More formally, the CLT states:

> For random variables with finite sample variance, the sampling distribution of the standardized sample means approaches the standard normal distribution as the sample size approaches infinity.

The theorem concerns the sampling distribution of the mean, which, as we saw in Chapter 2, assumes taking many samples from the population. If these samples are drawn randomly, the distribution of means tends to approximate the normal distribution after about 30 samples. But if the underlying variable's distribution is highly skewed, it may take more samples to approximate the normal. Since intercepts and slope coefficients are related to means, they also follow particular normal-like distributions (such as the *t*-distribution). Given a large enough sample, we can thus infer that even if the errors in predicting the outcome variable are not normally distributed, the results of an LRM estimated with OLS tend to be unbiased. Learning how to use techniques appropriate for situations where assumptions like normality are not met is important, though. And this discussion should also reinforce the idea that using random samples or having control over explanatory variables is valuable.[16]

Let's assume, though, that the assumptions are met. In particular, if the assumptions regarding independence, homoscedasticity, collinearity, and linearity are satisfied, then, according to the Gauss–Markov theorem,[17] the OLS estimator offers the best linear unbiased estimator (BLUE) among the class of linear estimators: no other linear estimator hits the population target as often, on average, as the OLS estimator.[18]

[15] See Neil A. Weiss (1999), *Introductory Statistics*, 5th Ed., Reading, MA: Addison-Wesley, p.427, for a helpful review. Bernard W. Lindgren (1993), *Statistical Theory*, 4th Ed., New York: Chapman & Hall, p.140, provides a standard proof.

[16] Appendix B provides statistical simulations that examine several of the assumptions.

[17] See Lindgren (1993), *op. cit.*, p.510. Carl Friedrich Gauss proved an early version of the theorem in his 1823 work *Theoria combinationis Observationum Erroribus Minimis Obnoxiae* (*The Theory of the Combination of Errors in Observations*) (Göttingen, DE: Apud Henricum Dieterich). Russian mathematician Andrei Andreevich Markov provided another application and proof of the theorem in 1900.

[18] When the normality assumption is also satisfied, the OLS estimator is the most *efficient* unbiased estimator (see John Fox (2016), *Applied Regression Analysis and Generalized Linear Models*, 3rd Ed., Los Angeles: Sage). Its sampling variability is smaller than other estimators. This is also called the *minimum variance unbiased* property (S. D. Silvey (1975), *Statistical Inference*, Boca Raton, FL: CRC Press).

Some Important Characteristics of Multiple LRMs

As mentioned earlier, OLS is the main technique used to estimate LRMs. We've now learned that OLS has some nice features that make it especially useful in regression analysis. The OLS regression equation also provides the *linear combination* of the *x*s (recall the summation signs in the linear regression equation) that has the largest possible correlation with the *y* variable ($cor\left(y_i,\hat{y}_i\right)$). Since we hope the model explains as much of the variability in the outcome variable as possible with the explanatory variables, this is a beneficial property. We'll learn how to estimate this correlation in Chapter 5.

Chapter 3 notes that statistical software uses matrix routines to compute slope coefficients, standard errors, and other features of LRMs. For those familiar with vectors and matrices, think of the *y* values as a vector of observations and the *x* values as a matrix of observations, as illustrated in Equation 4.3.[19]

$$\mathbf{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \quad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & \dots & x_{1k} \\ 1 & x_{21} & \dots & x_{2k} \\ \vdots & \vdots & \dots & \vdots \\ 1 & x_{n1} & \dots & x_{nk} \end{bmatrix} \tag{4.3}$$

Expressing the explanatory and outcome variables this way leads to an abbreviated depiction of the multiple linear regression equation: $\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$. $\hat{\boldsymbol{\beta}}$ is a vector of the intercept (denoted in the ***X*** matrix with 1s) and slope coefficients for the explanatory variables. The ***X*** matrix is listed first in this equation because it is *postmultiplied* by the vector of slope coefficients.[20]

The matrix formula in Equation 4.4 estimates the vector of slope coefficients with ***X*** and ***Y***.

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{Y} \tag{4.4}$$

[19] Translate the vector and matrix into spreadsheet format, such as with R's `View` function, if it's not clear how this works, and you will understand the utility of representing data in matrix form.

[20] Readers who have experience with linear algebra may recognize the key role that matrix routines play in solving systems of linear equations. They should not be surprised, therefore, that estimation with OLS uses matrices. But other routines for solving systems of linear equations also exist. Perhaps the most common in regression modeling is maximum likelihood estimation (MLE), which iterates through a set of reasonable estimates and determines which set is most likely given the data, though it tends to be slower than OLS and may lead to biases with small samples or sparse data matrices. Recent advances that use "coordinated guesses" of sets of estimates may improve the efficiency and speed of solving these equations even beyond matrix manipulations, however (Richard Peng and Santosh Vempala (2021), "Solving Sparse Linear Systems Faster than Matrix Multiplication," retrieved from https://arxiv.org/abs/2007.10254).

The accent next to the matrix $\boldsymbol{X}$ denotes its *transpose* and the superscript –1 indicates the inverse of the product in parentheses. Using R with matrix routines and a small dataset facilitates a deeper understanding of LRMs.[21]

Matrix algebra is also useful for estimating several other features of multiple LRMs. For example, the standard errors of the coefficients are estimated by taking the square roots of the diagonal elements of the matrix shown in Equation 4.5.

$$\boldsymbol{V} = \left(\boldsymbol{X'X}\right)^{-1}\hat{\sigma}^2 \quad \text{where} \quad \hat{\sigma}^2 = \frac{\hat{\varepsilon}'\hat{\varepsilon}}{n-k} \tag{4.5}$$

The n refers to the sample size and k denotes the number of explanatory variables in the model. We'll learn more about $\hat{\sigma}^2$ in Chapter 5 when discussing goodness-of-fit statistics; in brief it is a measure of dispersion: it measures the amount of variability of the residuals around the regression line—the *residual variance* or the *mean square error* (MSE).[22]

Equation 4.6 also estimates the standard errors of multiple LRM slope coefficients.

$$se\left(\hat{\beta}_i\right) = \sqrt{\frac{\Sigma\left(y_i - \hat{y}_i\right)^2}{\Sigma\left(x_i - \bar{x}\right)^2\left(1 - R_i^2\right)\left(n - k - 1\right)}} \tag{4.6}$$

The R^2 in the equation is from an LRM—called an *auxiliary regression model*—with x_i as the outcome variable and all the other explanatory variables as predictors (e.g., $x_{1i} = \alpha + \beta_1 x_{2i} + \ldots + \beta_k x_{ki}$). $\left(1 - R_i^2\right)$ is the *tolerance*. Continuing with the discussion surrounding Equation 3.13 in Chapter 3, the standard error increases as (a) the R^2 from the auxiliary regression model increases, (b) the variability of x decreases, (c) the variability of y increases, or (d) the sample size decreases. Think about the practical implications these factors have for significance testing with LRMs.

Chapter Summary

This chapter provides an introduction to multiple LRMs—models with two or more explanatory variables. A key difference between simple and

[21] For more information about matrix routines in R, see Nick Fieller (2016), *Basics of Matrix Algebra for Statistics with R*, Boca Raton, FL: CRC Press.

[22] Faraway (2014), *op. cit.*, and David G. Kleinbaum et al. (1998), *Applied Regression Analysis and Other Multivariable Methods*, 3rd Ed., Pacific Grove, CA: Duxbury Press (Appendix B), provide perspicuous overviews of matrix routines used in LRMs. A more formidable treatment, but one worth the effort for understanding the role of matrix routines, is in James R. Schott (2016), *Matrix Analysis for Statistics*, 3rd Ed., Hoboken, NJ: Wiley.

multiple LRMs is in how we interpret the slope coefficients. The multiple LRM's slopes are "statistically adjusted" for the effects of the other explanatory variables. The same assumptions apply to both models, but multiple LRMs are also concerned with collinearity—statistical associations among explanatory variables. Understanding the assumptions is so important that several of the following chapters provide details about each. Before getting to this, though, we'll learn about some other features of LRMs in Chapters 5–7.

Chapter Exercises

The dataset called *TeenBirths.csv* consists of data from almost 3,000 counties in the U.S. Our objective is to estimate a multiple LRM with a set of variables and provide interpretations of the results. The variables in the dataset include

• `state`	State name
• `county`	County name
• `teen_birth_rate`	percentage of births to teenage mothers
• `per_uninsured`	percentage of residents with no health insurance
• `per_hsgrads`	percentage of adult residents who are high school graduates
• `per_child_poverty`	percentage of children living in poverty
• `per_singleparent`	percentage of children living in single parent households

After importing the dataset into R, complete the following exercises.

1. Compute the Pearson's correlations of the following variables: percentage of births to teenage mothers, percentage of children living in poverty, percentage of children living in single parent households, and the percentage of residents with no health insurance. Which variable appears to be the strongest predictor of the percentage of births to teenage mothers? Why?
2. Estimate a simple LRM with the percentage of births to teenage mothers as the outcome variable and percent of children living in poverty as the explanatory variable.
   a. Interpret the slope coefficient.

   b. Interpret the p-value associated with the slope coefficient. What are some of its limitations for judging the results of the model?

   c. Interpret the 95% CI associated with the slope coefficient.

3. Estimate a multiple LRM with the percentage of births to teenage mothers as the outcome variable and the following explanatory variables: the percentage of children living in poverty, the percentage of children living in single parent households, and the percentage of residents with no health insurance.

   a. Interpret the slope coefficient associated with the percentage of children living in poverty. How does this slope coefficient compare with the slope coefficient in exercise 2 (a)?

   b. Interpret the slope coefficient associated with the percentage of residents with no health insurance.

4. Compute the standardized regression coefficients (beta weights) from the LRM in exercise 3. What do they suggest about the relative strength of the associations between each explanatory variable and percentage of births to teenage mothers? What are some possible limitations of using these coefficients for judging the strengths of the associations?

5. Compute the relative importance measures from the LRM in exercise 3. What do the results suggest about the best predictor(s) of the percentage of births to teenage mothers?

6. Compute the predicted means of the percentage of births to teenage mothers for the following groups (set the other variables at their means):

   a. Counties at the 25th percentile of the percentage of children living in poverty.

   b. Counties at the 75th percentile of the percentage of children living in poverty.

   c. Counties at the 25th percentile of the percentage of residents with no health insurance.

   d. Counties at the 75th percentile of the percentage of residents with no health insurance.

   What do these predicted means suggest about the associations of the explanatory variables with the outcome variable?

7. *Challenge*: save the predicted values (R labels them `fitted.values`) from the LRM in exercise 3. Compute the Pearson's correlation between the predicted values and each explanatory variable. What do the correlations suggest about the model and its predictive capabilities? Does this information strengthen your conclusions in exercises 5 and 6? Why or why not?