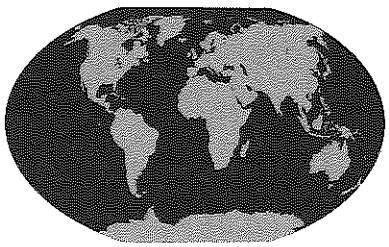
4. PREDICTING OUTCOMES USING LINEAR REGRESSION

R symbols, operators, and functions introduced in this chapter: `lm()` and `log()`.

We have already seen how we can analyze data to estimate causal effects and to infer population characteristics. Another goal of data analysis in the social sciences is to make predictions. In this chapter, we learn how to summarize with a line the relationship between the outcome variable of interest and another variable called a predictor (a process known as fitting a linear regression model). We then use this summary line to estimate the most likely value of the outcome, given a specific value of the predictor. As an illustration, we analyze data from 170 countries to predict GDP growth based on changes in night-time light emissions.

4.1 GDP AND NIGHT-TIME LIGHT EMISSIONS

Based on J. Vernon Henderson, Adam Storeygard, and David N. Weil, "Measuring Economic Growth from Outer Space," *American Economic Review* 102, no. 2 (2012): 994–1028.



To assess a country's economic activity, we often want to measure its gross domestic product (GDP). The GDP of a country is the monetary value of goods produced and services provided in that country during a specific period of time. The data required to construct GDP measures, however, may be either unreliable or hard to collect consistently, especially in developing countries. Consequently, we need good ways of predicting GDP using other observed variables.

In recent years, a group of social scientists noticed that changes in night-time light emissions, as measured from satellites circling the earth, were highly correlated with economic activity. As economic activity increases, so does use of electricity at night. As a result, change in a country's night-time light emissions as measured from space might be a good predictor of that country's GDP growth. In this chapter, we explore this connection and predict GDP growth using night-time light emission changes over time. We begin, though, with a simpler example. To practice fitting linear models and interpreting the results, we start by predicting a country's GDP at one point in time using a prior value of GDP.

4.2 PREDICTORS, OBSERVED VS. PREDICTED OUTCOMES, AND PREDICTION ERRORS

In the social sciences, we are often unable to observe the value of a particular variable of interest, Y, either because it hasn't occurred yet or because it is difficult to measure. In these situations, we typically observe the values of other variables that, if correlated with Y, can be used to predict Y. On the basis of these other variables, we can make an educated guess about what the value of Y is currently or will likely be at a different point in time, on average.

When analyzing data for the purpose of making predictions, we refer to the variable or variables that we use to make predictions as the predictor(s) and to the variable of interest that we want to predict as the outcome variable.

When making predictions, we distinguish between two types of variables:

- the predictor(s) (X): variable(s) that we use as the basis for our predictions
- the outcome variable (Y): variable that we are trying to predict based on the values of the predictor(s).

For example, if we are interested in predicting GDP using prior GDP, then GDP is the outcome variable, and prior GDP is the predictor. If we are interested in predicting GDP growth using the change in night-time light emissions, then GDP growth is the outcome variable, and the change in night-time light emissions is the predictor.

TIP: Predictors are also known as independent variables, and outcome variables as dependent variables.

In mathematical notation, we represent the predictor as X and the outcome variable as Y. Although we use the same mathematical notation as when estimating causal effects, the relationship between the X and Y variables here is not necessarily causal.

As we will see in detail later, to make good predictions, we choose predictors that are highly correlated with the outcome variable of interest. In other words, we choose predictors that have a strong linear association with the outcome variable. (Note that, when we speak of a "high degree of correlation," we mean that the correlation coefficient is high in absolute terms, regardless of its sign.) As discussed in chapter 3, correlation does not necessarily imply causation. Just because two variables are highly correlated with each other does not necessarily mean that changes in one variable cause changes in the other. When analyzing data for predictive purposes, then, we do not assume that there is a causal relationship between X and Y; we simply rely on a high degree of correlation between them and use one variable to estimate the value of the other.

RECALL: The correlation coefficient ranges from -1 to 1 and summarizes the direction and strength of the linear association between two variables. The closer the correlation coefficient is in absolute value to 1, the stronger the linear association between the two variables (that is, the closer the observations are to the line of best fit).

Making predictions is a two-step process. Once we have identified our X and Y variables, we need to understand how these two variables relate to each other. Our first step, then, is to analyze a dataset that contains both variables and summarize the relationship between X and Y with a mathematical model. We call this process "model fitting" because it consists of fitting to the data a model that characterizes how X is related to Y, on average.

The predicted outcomes, $\widehat{Y}$, are the values of Y we predict based on (i) the fitted model that summarizes the relationship between X and Y in a dataset where we observe both X and Y for each observation and (ii) the observed values of X.

Later, once we are in a situation where we cannot observe Y but we observe X, we use the fitted model to predict specific average values of the outcome variable for each observed value of the predictor. We refer to our predictions of Y as the predicted outcomes, and we denote them as $\widehat{Y}$ (pronounced Y-hat).

1. FIT A MODEL

- we observe both X and Y
- we summarize the relationship between the average Y and X with a model

2. MAKE PREDICTIONS

- we observe X but not Y
- we compute $\widehat{Y}$ by plugging the observed value of X into the fitted model

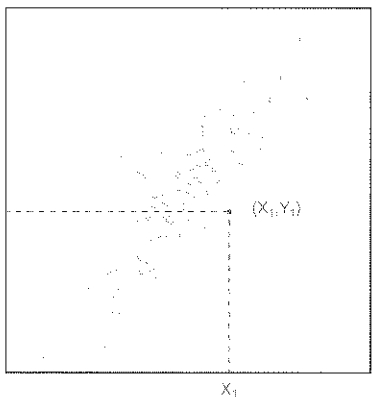
The prediction error, $\widehat{\epsilon}$, also known as a residual, measures how far our prediction is from the observed value; it is the difference between the observed outcome and the predicted outcome.

When making predictions, we aim to be as accurate as possible. In other words, we aim to minimize the prediction errors (also known as residuals). These are defined as the difference between the observed outcomes and the predicted outcomes and are denoted by $\widehat{\epsilon}$ (the Greek letter epsilon with a "hat" on top).

Note that to differentiate between observed and predicted variables, we often refer to Y as the *observed* outcome—and not just the outcome—to distinguish it more clearly from the *predicted* outcome $\widehat{Y}$.

4.3 SUMMARIZING THE RELATIONSHIP BETWEEN TWO VARIABLES WITH A LINE

When fitting a model for predictive purposes, we could use many different mathematical functions. In this book, we always summarize the relationship between X and Y with a line and, in particular, the line of best fit.

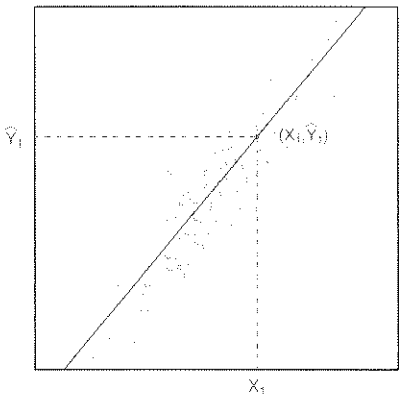


Let's get a sense of how this works using a hypothetical example. Suppose that the scatter plot of the X and Y variables (in the dataset where we can observe both) is as shown in the margin. As in all scatter plots, every dot represents a particular observation of X and Y. In this case, each dot is located based on the value of the predictor and the value of the observed outcome for a given observation. In the figure in the margin, we highlight, as an example, the dot representing the first observation of this imaginary dataset: (X_1, Y_1).

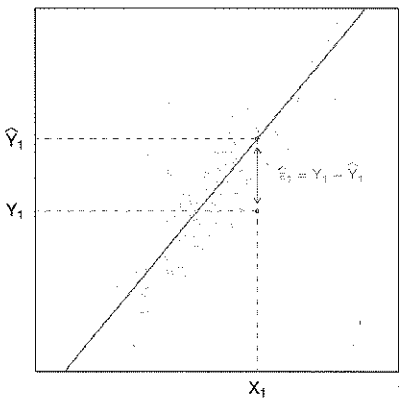
By looking at the scatter plot of X and Y, we get a general sense of how Y relates to X. In this case, given the observed upward slope of the data cloud, we conclude that high values of Y are likely to be associated with high values of X, and low values of Y are likely to be associated with low values of X. While this

is helpful information for predicting Y using X, it would be even better if we could summarize the relationship with a mathematical formula so that for each value of X, we could compute a predicted value of Y.

For example, we can summarize the relationship between X and Y with a line. In the top figure in the margin, in addition to the scatter plot of X and Y, we have plotted such a line, which we call the fitted line. Now, for every value of X, we can find a predicted Y ($\widehat{Y}$), by finding the value of X we are interested in on the x-axis, going up to the fitted line, and finding the height of the corresponding point on the line. For example, if we were interested in the value of X in the first observation in the dataset (X_1), based on the fitted line drawn on the plot, we would predict a Y equal to $\widehat{Y_1}$.



By looking at the scatter plot with the line, we get a sense of the prediction errors this fitted model would produce. If we use the line to compute the predicted outcomes for every observation, then we can measure the prediction errors ($\widehat{\epsilon}$) as the difference between the observed outcomes (Y) and the predicted outcomes ($\widehat{Y}$). ($\widehat{\epsilon_i}=Y_i-\widehat{Y}_i$.) Note that for each observation, this difference is equivalent to the vertical distance between the dot and the fitted line. See, for example, the bottom figure in the margin, where we show the prediction error of the first observation. In general, the closer the dots are to the fitted line, the smaller the prediction errors, and the farther the dots are from the line, the larger the prediction errors. To make the best possible predictions, then, we always summarize the relationship between X and Y with the line of best fit, which is the line closest to the data. (In subsection 4.3.4, we will explain the precise method used to choose this line.)



4.3.1 THE LINEAR REGRESSION MODEL

Now let's introduce some mathematical notation. The linear model, also known as the linear regression model, is defined as:

$$Y_i = \alpha + \beta X_i + \epsilon_i$$

where:

- Y_i is the outcome for observation i
- α (the Greek letter alpha) is the intercept coefficient
- β (the Greek letter beta) is the slope coefficient
- X_i is the value of the predictor for observation i
- ϵ_i (pronounced epsilon sub i) is the error for observation i.

This is the theoretical model that we assume reflects the true relationship between X and Y. If we knew the values of the

TIP: In statistics, we use Greek letters to represent quantities we do not know, such as α, β, and ϵ_i. The two coefficients, α and β, are not subscripted by i because they do not vary by observation. They are constants and not variables.

coefficients (α and β), as well as the values of the errors for each observation (ϵ_i), we could use this formula to compute the outcomes for each observation (Y_i) based on the observed values of the predictors (X_i). (By plugging the values of α, β, X_i, and ϵ_i into the formula above, we would compute Y_i.)

Unfortunately, we do not know the values of α, β, and ϵ_i. We have to estimate them based on data. We start by estimating the intercept (α) and the slope (β), the two coefficients that define the line. This is equivalent to fitting a line to the data, that is, finding the line that best summarizes the relationship between X and Y.

TIP: You might have seen the equation of a line written as $Y = mX + b$ where m is the slope and b the intercept. If so, it may be helpful for you to think that $\widehat{\alpha}$ is the b and $\widehat{\beta}$ is the m of the familiar model.

The formula of the line we fit to the data is:

$$\widehat{Y}_i = \widehat{\alpha} + \widehat{\beta} X_i$$

where:

- $\widehat{Y}_i$ (pronounced Y-hat sub i) is the predicted outcome for observation i
- $\widehat{\alpha}$ (pronounced alpha-hat) is the estimated intercept coefficient
- $\widehat{\beta}$ (pronounced beta-hat) is the estimated slope coefficient
- X_i is the value of the predictor for observation i.

Note that in this formula, Y, α, and β have a "hat" on top. This indicates that they are estimates or approximations. In addition, this formula no longer includes the errors (ϵ_i), which means that the resulting outcomes do not necessarily equal the true values of Y (Y_i); they equal the predicted values of Y ($\widehat{Y}_i$). In other words, for every value of X, this formula provides the corresponding value of Y *on the fitted line* (instead of on the observed data point). Note that the value of $\widehat{Y}$ produced by a fitted model is an *average* predicted value; it is the average predicted value of Y associated with a particular value of X. Indeed, predicted outcomes ($\widehat{Y}$) are equivalent to average outcomes ($\overline{Y}$).

The difference between the observed values of Y and the predicted values of Y are the estimated errors or residuals:

$$\widehat{\epsilon}_i = Y_i - \widehat{Y}_i$$

where:

- $\widehat{\epsilon}_i$ is the estimated error, or residual, for observation i
- Y_i is the observed outcome for observation i
- $\widehat{Y}_i$ is the predicted outcome for observation i.

These are the prediction errors that we try to minimize by using the line that best fits the data.

To recap, to make predictions using the linear regression model, we start by analyzing a dataset that contains both X and Y for each observation. We summarize the relationship between them with the line of best fit, which is the line with the smallest prediction errors possible. Fitting this line involves estimating the two coefficients that define any line: the intercept ($\widehat{\alpha}$) and the slope ($\widehat{\beta}$). Once we have fitted the line, we can use it to obtain the most likely average value of Y based on the observed value of X.

1. FIT A LINEAR REGRESSION MODEL

- we observe both X and Y
- we find the line that best summarizes the relationship between them; we estimate the intercept ($\widehat{\alpha}$) and slope ($\widehat{\beta}$) of the line with the smallest prediction errors possible

2. MAKE PREDICTIONS

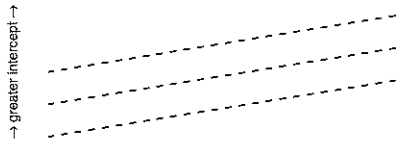
- we observe X but not Y
- we compute $\widehat{Y}$ by plugging the observed value of X into the fitted linear regression model:

$$\widehat{Y} = \widehat{\alpha} + \widehat{\beta}X$$

Let's take a moment now to understand what the two coefficients of a line measure and how to interpret them.

4.3.2 THE INTERCEPT COEFFICIENT

Generally speaking, the intercept of a line specifies the vertical location of the line. See, for example, the lines in the margin, which have different intercepts but the same slope. Increasing and decreasing the intercept moves the line up and down.



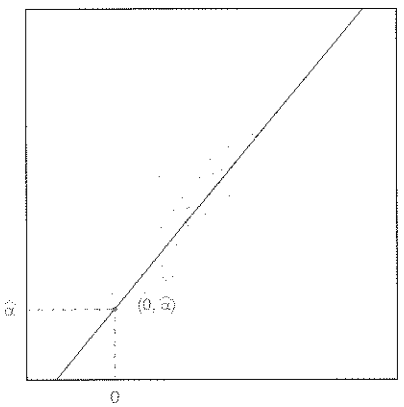
Specifically, the intercept ($\widehat{\alpha}$) is the value of $\widehat{Y}$ when X=0.

The intercept ($\widehat{\alpha}$) is the $\widehat{Y}$ when X=0.

Indeed, as we can see below, if in the fitted linear model, we plug in X=0, then $\widehat{Y}$ equals $\widehat{\alpha}$. So, $\widehat{\alpha}$ is the $\widehat{Y}$ when X=0.

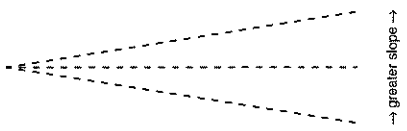
$$\widehat{Y} = \widehat{\alpha} + \widehat{\beta}\times 0 = \widehat{\alpha} \quad \text{(if } X=0\text{)}$$

This definition of the intercept is helpful. We can use it to figure out the value of $\widehat{\alpha}$ of any line on a graph. We just need to find X=0 on the x-axis, go up to the fitted line, and then find the height of the corresponding point. The value of $\widehat{Y}$ at the point on the fitted line where X=0 is the value of the intercept of the line. (See figure in the margin.) Note that the y-axis is not always drawn at X=0, and therefore, the intercept is *not* necessarily the value of $\widehat{Y}$ at the point where the line crosses the y-axis.



We can also use the definition above to help us substantively interpret the value of $\widehat{\alpha}$. In predictive models, we interpret the intercept as the predicted outcome, $\widehat{Y}$, when the predictor X equals zero. (We will see concrete examples soon.)

4.3.3 THE SLOPE COEFFICIENT



The slope ($\widehat{\beta}$) is $\triangle\widehat{Y}$ divided by $\triangle X$ between two points on the line.

Generally speaking, the slope of a line specifies the angle, or steepness of the line. See, for example, the lines in the margin, which have different slopes but the same intercept. The top line has a positive slope, the middle line has a slope of zero, and the bottom line has a negative slope.

Specifically, the slope ($\widehat{\beta}$) is the change in $\widehat{Y}$ divided by the change in X between two points on the line, commonly known as "rise over run":

$$\widehat{\beta} = \frac{\text{rise}}{\text{run}} = \frac{\triangle\widehat{Y}}{\triangle X} = \frac{\widehat{Y}_{\text{final}} - \widehat{Y}_{\text{initial}}}{X_{\text{final}} - X_{\text{initial}}}$$

where $\triangle$ (the Greek letter Delta) represents change, and thus, $\triangle\widehat{Y}$ is the change in $\widehat{Y}$ and $\triangle X$ is the change in X.

TIP: The change in a variable between two points (initial and final) is equivalent to the difference between the value of the variable at the final point and the value of the variable at the initial point. Examples:

$$\triangle\widehat{Y} = \widehat{Y}_{\text{final}} - \widehat{Y}_{\text{initial}}$$
$$\triangle X = X_{\text{final}} - X_{\text{initial}}$$

For example, see figure 4.1, which shows the change in $\widehat{Y}$ ($\triangle\widehat{Y}$) and the change in X ($\triangle X$) associated with two points on the line.

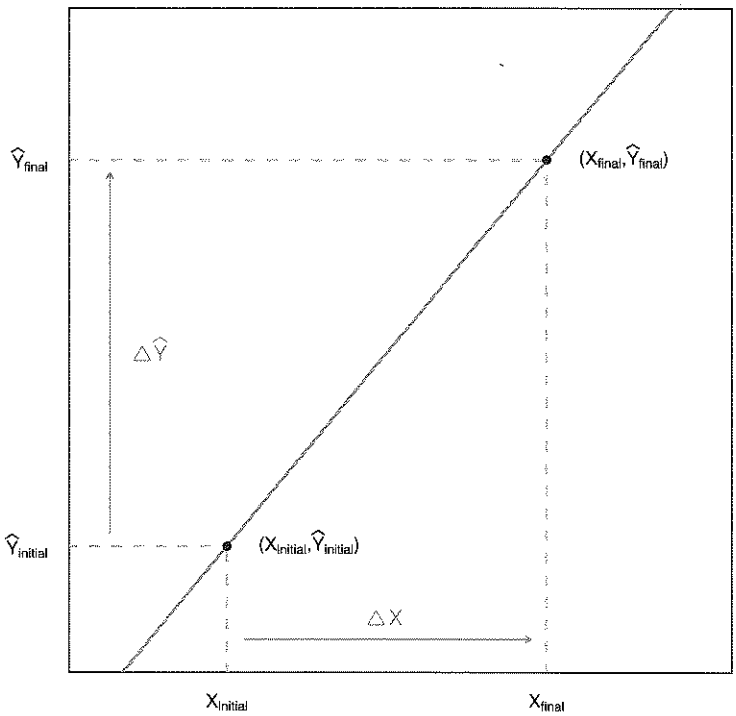


FIGURE 4.1. The slope ($\widehat{\beta}$) can be computed as "rise over run," where rise is the change in $\widehat{Y}$ and run is the change in X between two points on the line.

Substantively speaking, we can interpret the value of the slope as the change in $\widehat{Y}$ associated with a one-unit increase in X. In mathematical notation, when $\triangle X=1$, $\widehat{\beta}=\triangle\widehat{Y}$:

$$\widehat{\beta}=\frac{\triangle\widehat{Y}}{1}=\triangle\widehat{Y} \quad (\text{if } \triangle X=1)$$

The slope ($\widehat{\beta}$) represents the $\triangle\widehat{Y}$ associated with a one-unit increase in X.

In predictive models, then, we interpret the slope as the predicted change in the outcome, $\triangle\widehat{Y}$, associated with a one-unit increase in the predictor X. Since $\widehat{\beta}$ measures a *change* in $\widehat{Y}$, we interpret it as an increase when positive, a decrease when negative, and as no change when zero.

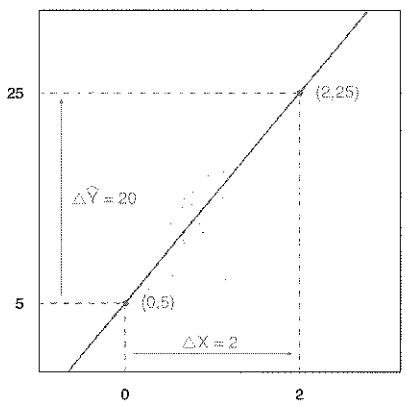
THE FITTED LINE IS:

$$\widehat{Y}=\widehat{\alpha}+\widehat{\beta}X$$

where:

- $\widehat{\alpha}$ is the estimated intercept coefficient, which can be interpreted as the $\widehat{Y}$ when $X=0$
- $\widehat{\beta}$ is the estimated slope coefficient, which can be interpreted as the $\triangle\widehat{Y}$ associated with $\triangle X=1$.

Before moving on to learning how to find the line of best fit, let's practice figuring out the specific formula of a line by looking at its depiction in a graph. (See figure in the margin.)



We start by finding the values of two points on the line:

- the point that corresponds to $X=0$
- the point that corresponds to a higher value of X than 0.

In the figure in the margin, these two points are (0,5) and (2,25). Given the values of these two points, we can conclude that:

- the intercept coefficient ($\widehat{\alpha}$) equals 5 because that is the value of $\widehat{Y}$ when $X=0$ (see the point (0,5) on the line)
- the slope coefficient ($\widehat{\beta}$) equals 10 because that is the value of $\triangle\widehat{Y}/\triangle X$ between the two points on the line:

$$\widehat{\beta}=\frac{\triangle\widehat{Y}}{\triangle X}=\frac{25-5}{2-0}=\frac{20}{2}=10$$

This particular fitted line is then: $\widehat{Y}=5+10X$.

We can check that the two points shown in the figure on the previous page—(0,5) and (2,25)—belong to the line $\widehat{Y} = 5 + 10X$. For each point, we plug the value of X into the formula of the line and find the corresponding $\widehat{Y}$:

$$\widehat{Y} = 5 + 10\times0 = 5 \qquad \text{(if } X=0\text{)}$$
$$\widehat{Y} = 5 + 10\times2 = 25 \qquad \text{(if } X=2\text{)}$$

The math above confirms that these two points are indeed on the line $\widehat{Y} = 5 + 10X$.

4.3.4 THE LEAST SQUARES METHOD

We could draw an infinite number of lines on a scatter plot, but some lines do a better job than others at summarizing the relationship between X and Y. For example, of the three lines shown in figure 4.2, we can agree that the last one does the best job of depicting how Y relates to X. (Intuitively, we know that the line of best fit should be as close to the dots as possible.)

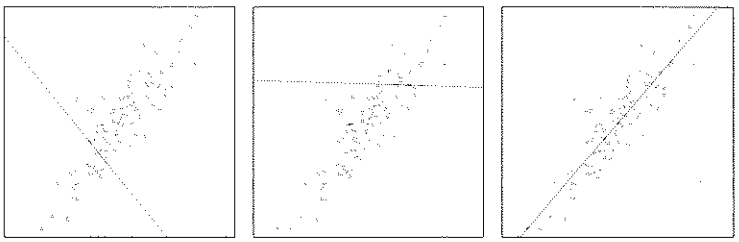


FIGURE 4.2. Three lines that we could draw on the scatter plot of X and Y out of the infinite number of possible lines.

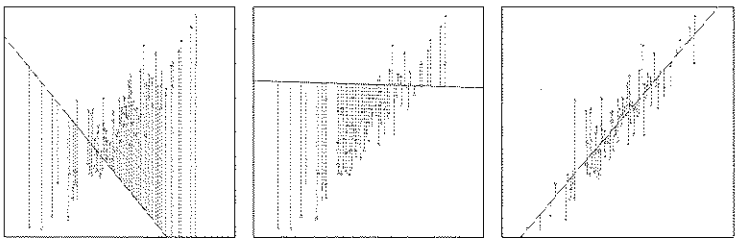


FIGURE 4.3. The fit of the line improves from left to right. The last line best summarizes the relationship between X and Y; it is closest to the observations, which means that it produces the smallest prediction errors (shown as red dashed lines).

How do we choose the line that best summarizes the relationship between X and Y? Given that we want our predictions to be as accurate as possible, generally speaking, we choose the line that reduces the prediction errors ($\hat{\epsilon}$), that is, the vertical distance between each dot and the fitted line. As we can observe in figure 4.3, the line on the right produces the smallest prediction errors (shown as red dashed lines). Therefore, we would choose this line over the other two to summarize the relationship between X and Y.

Formally, to choose the line of best fit, we use the "least squares" method, which identifies the line that minimizes the "sum of the squared residuals," known as SSR. (Recall that residuals is a different name for prediction errors; this method minimizes the sum of the squared prediction errors.)

$$\text{SSR} = \sum_{i=1}^{n} \hat{\epsilon}_i^2$$

Why do we want to minimize the sum of the *squared* residuals rather than the sum of the residuals? Because in the minimization process we want to avoid having positive prediction errors cancel out negative prediction errors. By squaring the residuals, we convert them all to positive numbers. (This procedure for choosing the line of best fit is called the "least squares" method because it *minimizes* the sum of the *squared* residuals.)

In practice, we do not undertake this minimization process ourselves. Instead, we rely on R to make the necessary computations. In the next section, we will go over a simple example and learn how to ask R to estimate the two coefficients of the line that minimizes the sum of the squared residuals. In other words, we will learn how to use R and the least squares method to find the line of best fit.

4.4 PREDICTING GDP USING PRIOR GDP

The code for this chapter's analysis can be found in the "Prediction.R" file. The dataset we analyze is provided in the file "countries.csv", and table 4.1 shows the names and descriptions of the variables included.

TABLE 4.1. Description of the variables in the countries dataset, where the unit of observation is countries.

variable	description
country	name of the country
gdp	country's GDP from 2005 to 2006 (in trillions of local currency units)
prior_gdp	country's GDP from 1992 to 1993 (in trillions of local currency units)
light	country's average level of night-time light emissions from 2005 to 2006 (in units on a scale from 0 to 63, where 0 is complete darkness and 63 is extremely bright light)
prior_light	country's average level of night-time light emissions from 1992 to 1993 (in units on a scale from 0 to 63, where 0 is complete darkness and 63 is extremely bright light)

RECALL: If the DSS folder is saved directly on your Desktop, to set the working directory, you must run setwd("~/Desktop/DSS") if you have a Mac and setwd("C:/*user*/Desktop/DSS") if you have a Windows computer (where *user* is your own username). If the DSS folder is saved elsewhere, please see subsection 1.7.1 for instructions on how to set the working directory.

As always, we begin by reading and storing the data (assuming we have already set the working directory):

```
co <- read.csv("countries.csv") # reads and stores data
```

To get a sense of the dataset, we look at the first few observations:

```
head(co) # shows first observations
```

	country	gdp	prior_gdp	light	prior_light
1	USA	11.107	7.373	4.227	4.482
2	Japan	543.017	464.168	11.926	11.808
3	Germany	2.152	1.793	10.573	9.699
4	China	16.558	4.901	1.451	0.735
5	UK	1.098	0.754	11.856	13.392
6	France	1.582	1.208	8.513	6.909

Based on table 4.1 and the output above, we learn that each observation in the dataset represents a country, and that the dataset contains five variables:

- *country* is a character variable that identifies the country.
- *gdp* and *prior_gdp* are each country's GDP at two different points in time, 13 years apart, from 2005 to 2006 and from 1992 to 1993. They are measured in trillions of local currency units (that is, in trillions of dollars in the case of the United States, trillions of yen in the case of Japan, trillions of euros in the case of Germany, and so on).
- *light* and *prior_light* are each country's average night-time light emissions at two different points in time, 13 years apart, from 2005 to 2006 and from 1992 to 1993. They are measured on a scale from 0 to 63, where 0 represents no light and 63 is extremely bright light.

We interpret the first observation as representing the United States, where GDP was $11 trillion from 2005 to 2006 and $7 trillion from 1992 to 1993, and average night-time light emissions were 4.2 units from 2005 to 2006 and 4.5 units from 1992 to 1993 (as measured on a scale from 0 to 63).

To find the total number of observations in the dataset, we run:

```
dim(co) # provides dimensions of dataframe: rows, columns
## [1] 170   5
```

The dataset contains information about 170 countries.

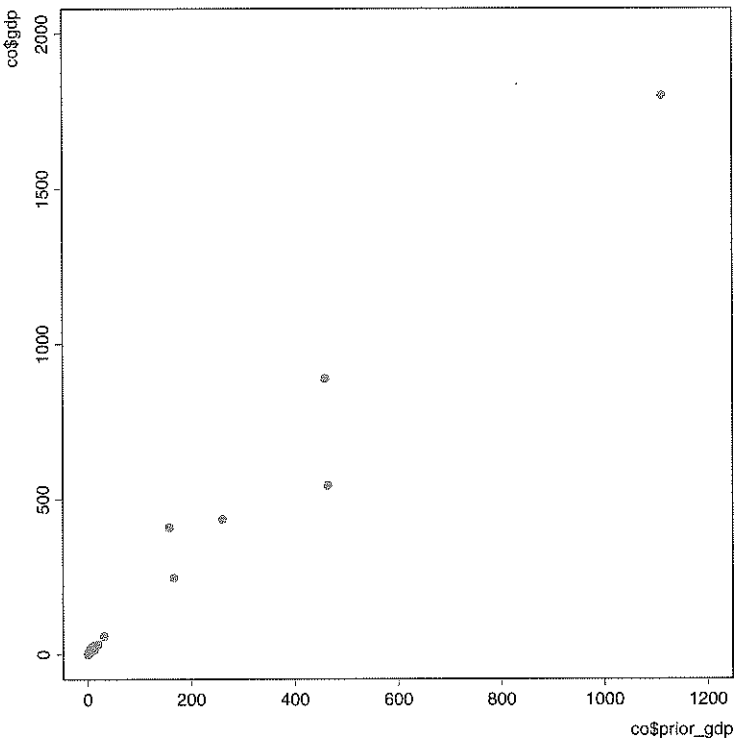
4.4.1 RELATIONSHIP BETWEEN GDP AND PRIOR GDP

To get a sense of the relationship between a country's GDP at two points in time, we analyze how the two measures of GDP that we have in the dataset, *gdp* and *prior_gdp*, relate to each other. Since these two variables were measured 13 years apart, our conclusions refer to the relationship between a country's GDP at one point in time and its GDP about 13 years prior.

We start the analysis by creating a scatter plot using the function `plot()` to visualize the relationship between the two variables of interest. Note that we always plot the predictor on the x-axis and the outcome variable on the y-axis. In this case, to visualize the relationship between *gdp* and *prior_gdp*, we run:

```
plot(x=co$prior_gdp, y=co$gdp) # creates scatter plot
```

RECALL: `plot()` creates the scatter plot of two variables. Examples: `plot(data$x_var, data$y_var)`, `plot(x=data$x_var, y=data$y_var)`, or `plot(y=data$y_var, x=data$x_var)`. Also, if R gives you the error message "Error in plot.new(): figure margins too large", try making the lower-right window larger and then re-run the code that creates the plot.



Looking at the scatter plot, we observe a positive association between the two variables. Higher values of prior GDP tend to be associated with higher values of GDP. In addition, we notice that the relationship between the two variables appears to be strongly linear. To further investigate the direction and strength of the linear association, we can compute the correlation coefficient using the function `cor()`:

RECALL: The correlation coefficient ranges from -1 to 1 and summarizes the direction and strength of the linear association between two variables. In R, the function `cor()` calculates the correlation coefficient between two variables. Example: `cor(data$variable1, data$variable2)`.

```
cor(co$gdp, co$prior_gdp) # computes correlation
## [1] 0.9903451
```

The correlation coefficient between the two variables turns out to be 0.99, which confirms what we noticed in the scatter plot above.

TIP: When writing the model for the first time, it is helpful to (i) emphasize that the variables may take different values for each observation by adding the subscript i, and (ii) specify what each observation, i, represents. In this case, the unit of observation, i, is countries.

Now that we have a general sense of how the two variables relate to each other, we can fit a linear model to summarize their relationship. This is the model we will use later to make predictions. Since our outcome of interest is *gdp* and our predictor is *prior_gdp*, the line we want to fit is:

$$\widehat{gdp_i} = \widehat{\alpha} + \widehat{\beta}\ prior\_gdp_i \quad (i\text{=countries})$$

where:

- $\widehat{gdp_i}$ is the average predicted GDP from 2005 to 2006 among countries in which the value of *prior_gdp* equals $prior\_gdp_i$
- $prior\_gdp_i$ is the GDP of country i from 1992 to 1993.

Once we estimate $\widehat{\alpha}$ and $\widehat{\beta}$, we will be able to plug into the formula above any value of *prior_gdp* and get a $\widehat{gdp}$ in return.

`lm()` fits a linear model. It requires a formula of the type `Y ~ X`, where *Y* identifies the *Y* variable and *X* identifies the *X* variable. To specify the object where the dataframe is stored, we can either use the `$` character in the code identifying each variable or set the optional argument `data`. Examples: `lm(data$y_var ~ data$x_var)` or `lm(y_var ~ x_var, data=data)`.

To estimate the coefficients of the linear model using the least squares method in R, we use the function `lm()`, which stands for "linear model." This function requires that we specify as the main argument a formula of the type `Y ~ X`, where *Y* identifies the outcome variable and *X* identifies the predictor. To fit a line to summarize the relationship between GDP and prior GDP, we run:

```
lm(co$gdp ~ co$prior_gdp) # fits linear model
##
## Call:
## lm(formula = co$gdp ~ co$prior_gdp)
##
## Coefficients:
## (Intercept)  co$prior_gdp
##      0.7161        1.6131
```

Note that since the variables in the model should always come from the same dataframe, there is an alternative way of specifying the `lm()` function. Instead of using the `$` character for each variable, we can use the optional argument `data` and set it to equal the name of the object where the dataframe containing all the variables is stored. For example, `lm(gdp ~ prior_gdp, data=co)` produces the same output as the code above.

As we can see in the output of the function `lm()` above, the estimated intercept ($\widehat{\alpha}$) is 0.72, and the estimated slope ($\widehat{\beta}$), the coefficient for the variable *prior_gdp*, is 1.61.

The fitted linear model is then:

$$\widehat{gdp} = 0.72 + 1.61\ prior\_gdp$$

How should we interpret $\widehat{\alpha}$=0.72? The value of $\widehat{\alpha}$ equals the $\widehat{Y}$ when X=0. Here, since Y is GDP and X is prior GDP (both measured in trillions of local currency units), we interpret the estimated intercept coefficient as indicating that when prior GDP is 0 trillion local currency units, we predict that GDP is 0.72 trillion local currency units, on average. (Note that the interpretation of the intercept does not always make substantive sense, especially when the range of observed values of the predictor does not include zero. This is a good example. It does not make sense for a country to have a prior GDP of 0 trillion local currency units. When we make predictions beyond the observed range of data, we make the strong assumption that the relationship between X and Y continues to hold. This is called "extrapolation," and it may lead to nonsensical predictions.)

How should we interpret $\widehat{\beta}$=1.61? The value of $\widehat{\beta}$ equals the $\Delta\widehat{Y}$ associated with ΔX=1. Here, since the Y is GDP and the X is prior GDP (both measured in trillions of local currency units), we interpret the estimated slope coefficient as indicating that an increase in prior GDP of 1 trillion local currency units is associated with a predicted increase in GDP of 1.61 trillion local currency units, on average.

To make it easier to work with the fitted model, we may want to store it as an object using the assignment operator <-. (Here, we chose the name *fit*, but we could have chosen another name.)

```
fit <- lm(gdp ~ prior_gdp, data=co) # stores fitted model
```

For example, now we can easily add the fitted line to the scatter plot by using the function abline(). As we saw in the previous chapter, this function adds a straight line to the most recently created graph. There, we saw how to draw horizontal and vertical lines. Here, we learn that this function will draw the fitted line when we specify as the main argument the object that contains the output of the fitted model. Go ahead and run:

```
abline( fit ) # adds line to scatter plot
```

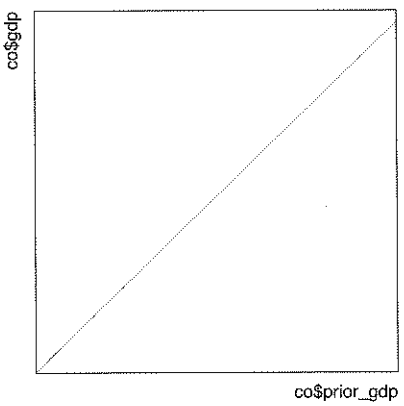
Remember, that R will give you an error message if you run this piece of code without having first created a graph. If you run all the code provided in this section, in sequence, you should see the figure shown in the margin.

Now that we have fitted a line to summarize the relationship between our two variables of interest (also known as fitting a linear regression model), we can use the fitted model to make predictions.

TIP: In what units of measurement are the two estimated coefficients, $\widehat{\alpha}$ and $\widehat{\beta}$?
- If Y is non-binary, both $\widehat{\alpha}$ and $\widehat{\beta}$ are in the same unit of measurement as Y.
- If Y is binary, $\widehat{\alpha}$ is in percentages, and $\widehat{\beta}$ is in percentage points (after multiplying both outputs by 100).

Here, since *gdp* is non-binary and measured in trillions of local currency units, both $\widehat{\alpha}$ and $\widehat{\beta}$ are measured in trillions of local currency units.

abline() adds the fitted line to the most recently created graph when we specify as the main argument the object that contains the output of the lm() function. Example: fit <- lm(Y ~ X) and then abline(fit).



TIP: In what units of measurement are our predictions, $\widehat{Y}$ and $\triangle\widehat{Y}$?

- If Y is non-binary, both $\widehat{Y}$ and $\triangle\widehat{Y}$ are in the same unit of measurement as Y.
- If Y is binary, $\widehat{Y}$ is in percentages and $\triangle\widehat{Y}$ is in percentage points (after multiplying both outputs by 100).

Here, since *gdp* is non-binary and measured in trillions of local currency units, both $\widehat{Y}$ and $\triangle\widehat{Y}$ are measured in trillions of local currency units.

Generally speaking, there are two types of predictions we may be interested in making. First, we may want to predict the average value of the outcome variable given a value of the predictor. When this is the case, we use the formula of the fitted line directly.

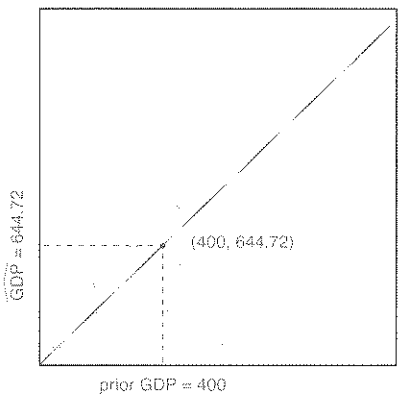
TO COMPUTE $\widehat{Y}$ BASED ON X: We plug the value of X into the fitted linear model and calculate $\widehat{Y}$.

$$\widehat{Y} = \widehat{\alpha} + \widehat{\beta}X$$

For example, suppose that we want to know the current GDP of a country, and for some reason we cannot measure it. But we do know that 13 years ago, the country's GDP was 400 trillion local currency units. What would our best guess be for current GDP, given the relationship between GDP and prior GDP that we estimated above? To predict the value of a country's current GDP based on the value of that country's GDP 13 years prior, we plug the value of prior GDP into the fitted linear model:

$$\begin{aligned}\widehat{gdp} &= 0.72 + 1.61\ prior\_gdp \\ &= 0.72 + 1.61 \times 400 = 644.72\end{aligned}$$

Based on the fitted line, we predict that the country has a current GDP of about 644.72 trillion local currency units. (See figure in the margin to visualize how we would arrive at the same conclusion using the fitted line drawn in the scatter plot.)



Second, we may want to predict the average change in the outcome variable associated with a change in the value of the predictor. When this is the case, we use the formula that computes the change in the predicted outcome, shown below.

TO COMPUTE $\triangle\widehat{Y}$ ASSOCIATED WITH $\triangle X$: We plug the value of $\triangle X$ into the formula below and calculate $\triangle\widehat{Y}$.

$$\triangle\widehat{Y} = \widehat{\beta}\,\triangle X$$

TIP: We arrive at this formula by using the definition of either (a) the slope coefficient or (b) the change in the predicted outcome between two points (initial and final).

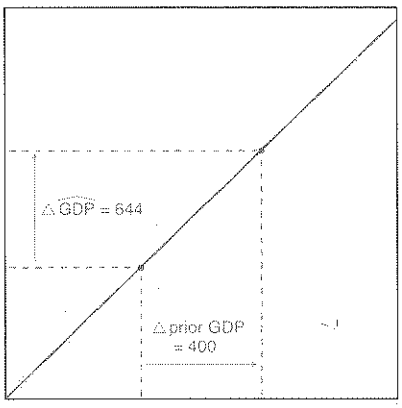
(a) Since $\widehat{\beta} = \triangle\widehat{Y}/\triangle X$, then $\triangle\widehat{Y} = \widehat{\beta}\,\triangle X$

(b) $$\begin{aligned}\triangle\widehat{Y} &= \widehat{Y}_{final} - \widehat{Y}_{initial} \\ &= (\widehat{\alpha} + \widehat{\beta}\,X_{final}) - (\widehat{\alpha} + \widehat{\beta}\,X_{initial}) \\ &= \widehat{\beta}\,(X_{final} - X_{initial}) = \widehat{\beta}\,\triangle X\end{aligned}$$

For example, imagine that we want to predict the change in GDP associated with an increase in prior GDP of 400 trillion local currency units. To make the calculations here, we start with the formula of change in the predicted GDP and plug in the value of change in prior GDP:

$$\widehat{\triangle gdp} = 1.61 \triangle prior\_gdp$$
$$= 1.61 \times 400 = 644$$

We predict that an increase in GDP 13 years ago of 400 trillion in local currency would likely be associated with an increase in current GDP of about 644 trillion local currency units. (Again, see figure in the margin to visualize how we would arrive at the same conclusion using the fitted line drawn in the scatter plot.)



4.4.2 WITH NATURAL LOGARITHM TRANSFORMATIONS

In the previous subsection, we saw how to fit a line using our two variables of interest, *gdp* and *prior_gdp*, without any transformations. To improve the fit of the line, there are times when we might want to transform one or both of our variables of interest. As we will soon see, these transformations affect how we interpret the coefficients.

When a variable contains a handful of either extremely large or extremely small values, the distribution of the variable will be skewed. (Recall that a distribution is considered skewed when it is not symmetric because one of its tails is longer than the other.) Under these circumstances, it is often a good idea to transform the variable by taking its natural logarithm. This transformation will make the variable of interest more normally distributed and, in turn, improve the fit of the line to the data. In the example at hand, we will transform both variables of interest by taking the natural logarithm, and then we will re-fit the line.

TIP: The natural logarithm is the inverse of the exponential function. The base of the natural logarithm is the constant e, known as Euler's number, which is approximately 2.7183. The natural logarithm of X, $\log(X)$, is the power to which e would have to be raised to equal X (if $X=e^Y$, then $\log(X)=Y$).

In R, the function to compute a natural logarithm is `log()`. To calculate the natural logarithm of each of the values inside a variable, we specify the code identifying the variable as the main argument. Then, to save the results as a new variable, we can use the assignment operator `<-`. To store this new variable inside the existing dataframe, we use the `$` character. Returning to the running example, to create the log-transformed GDP variables, *log_gdp* and *log_prior_gdp*, we run:

```
## create log-transformed GDP variables
co$log_gdp <- log(co$gdp)  # gdp
co$log_prior_gdp <- log(co$prior_gdp)  # prior gdp
```

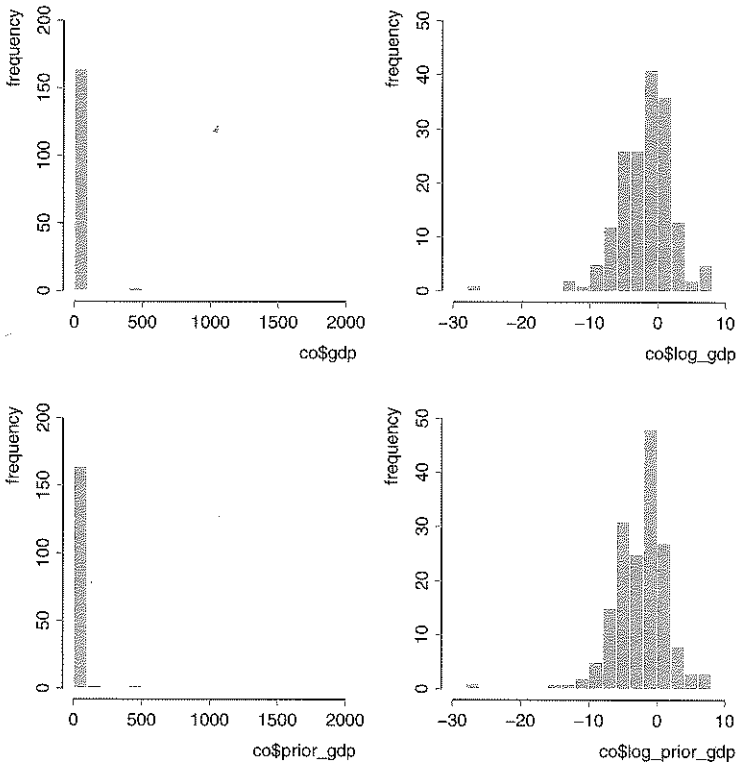
`log()` computes the natural logarithm of the argument specified inside the parentheses. Example: `log(10)`.

To check that the new variables were created correctly, we could look at the first few observations of the dataframe *co*. If you run `head(co)` again, you should see that the value of the first observation of *log_gdp* is 2.4 (since gdp_1=11.1 and log(11.1)=2.4), and the value of the first observation of *log_prior_gdp* is approximately 2 (since $prior\_gdp_1$=7.4 and log(7.4)=2).

RECALL: hist() creates the histogram of a variable. Example: hist(*data$variable*).

To visualize how the transformation affected the distribution of our two variables of interest, we can create the histograms of the original and log-transformed variables by running:

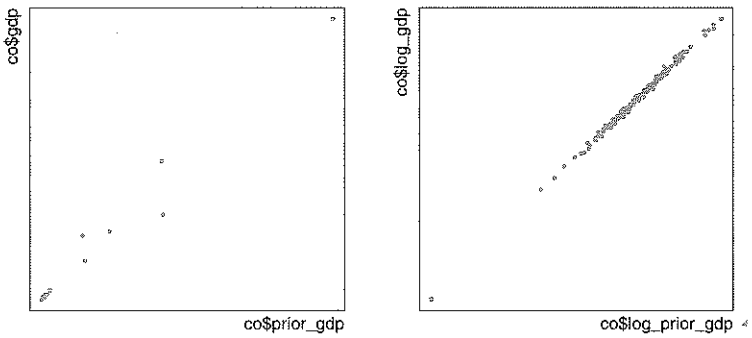
```
## create histograms
hist(co$gdp) # gdp
hist(co$log_gdp) # log-transformed gdp
hist(co$prior_gdp) # prior gdp
hist(co$log_prior_gdp) # log-transformed prior gdp
```



As we can see in the histograms on the left, the two original measures of GDP contained a handful of extremely large values, which skewed their distributions. (In both cases, the tail on the right is longer than the tail on the left). While most observations had values below 200 trillion local currency units, there were a few outliers. For example, Indonesia had a GDP of more than 1,100 trillion rupiahs in 1993 and of almost 1,800 trillion rupiahs in 2006. As we can see in the histograms on the right, the distributions become more symmetrical and bell-shaped once we log-transform the variables.

Now, we can visualize how the transformation of the variables affected the fit of the line by creating the scatter plots between the original variables and between the log-transformed variables:

```
## create scatter plots
plot(x=co$prior_gdp, y=co$gdp) # original
plot(x=co$log_prior_gdp, y=co$log_gdp) # log-transformed
```



Comparing the two scatter plots, we clearly see that the natural logarithm transformation makes the relationship between the two variables of interest more linear. To confirm this, we can compute the correlation coefficient between the log-transformed variables:

```
cor(co$log_gdp, co$log_prior_gdp) # computes correlation
## [1] 0.9982696
```

Indeed, the new correlation coefficient is even closer to 1 than it was before the logarithmic transformation (0.998 vs. 0.990).

Now that we have a sense of how the two log-transformed variables relate to each other, we can fit the following linear model to summarize their relationship:

$$\widehat{log\_gdp}_i = \widehat{\alpha} + \widehat{\beta}\ log\_prior\_gdp_i \quad (i\text{=countries})$$

where:

- $\widehat{log\_gdp}_i$ is the average predicted natural logarithm of GDP from 2005 to 2006 among countries in which the value of *log_prior_gdp* equals $log\_prior\_gdp_i$
- $log\_prior\_gdp_i$ is the natural logarithm of the GDP of country i from 1992 to 1993.

To estimate the coefficients of this new line of best fit, we use the `lm()` function again and run:

```
lm(log_gdp ~ log_prior_gdp, data=co) # fits linear model
##
## Call:
## lm(formula = log_gdp ~ log_prior_gdp, data=co)
##
## Coefficients:
##  (Intercept)  log_prior_gdp
##       0.4859         1.0105
```

RECALL: `lm()` fits a linear model. It requires a formula of the type `Y ~ X`. To specify the object where the dataframe is stored, we can use the optional argument `data` or the `$` character. Examples: `lm(y_var ~ x_var, data=data)` or `lm(data$y_var ~ data$x_var)`.

Using the estimated coefficients provided above, we can write the new fitted linear model as follows:

$$\widehat{log\_gdp} = 0.49 + 1.01\ log\_prior\_gdp$$

The fitted log-log linear model is a fitted linear model in which both Y and X have been log-transformed:

$$\widehat{\log(Y)} = \widehat{\alpha} + \widehat{\beta}\log(X)$$

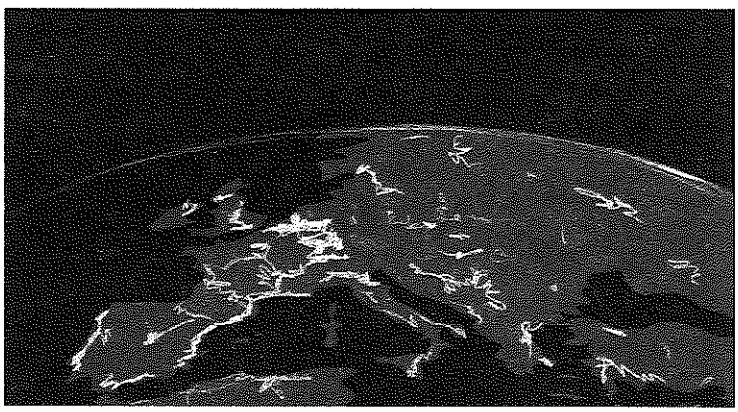
In this model, we interpret $\widehat{\beta}$ as the predicted percentage change in the outcome associated with an increase in the predictor of 1 percent.

This type of model, in which both the outcome and the predictor have been log-transformed, is called the log-log linear model. While we could interpret the coefficients the same way as in the normal linear model, in practice, we use an approximation to avoid dealing with the logarithms, especially when interpreting $\widehat{\beta}$.

As shown in the appendix near the end of this chapter, we interpret $\widehat{\beta}$ as the predicted *percentage* change in the outcome associated with an increase in the predictor of 1 *percent*. Since here $\widehat{\beta}$=1.01, an increase of prior GDP of 1% is associated with a predicted increase in GDP of 1.01%, on average. Note that in this interpretation of $\widehat{\beta}$, both the change in X and the change in $\widehat{Y}$ are measured in *percentages*, instead of in units, as is the case in the standard linear model. In other words, in the log-log model, we estimate change in relative rather than absolute terms.

4.5 PREDICTING GDP GROWTH USING NIGHT-TIME LIGHT EMISSIONS

Let's figure out how to fit a model to predict changes in GDP using changes in night-time light emissions. As mentioned earlier, being able to predict GDP growth using night-time light emissions would be quite useful. In remote areas of the world, where measuring GDP is difficult, measures of night-time light emissions are readily available through satellite imagery.



We start the analysis by creating the two variables whose relationship we want to understand. In this model, our outcome of interest is the percentage change in GDP between two points in time, which is defined as:

$$gdp\_change = \frac{gdp - prior\_gdp}{prior\_gdp} \times 100$$

As we saw in chapter 1, R understands arithmetic operators such as +, -, *, and /. Thus, to create this variable, we can run:

```
## create GDP percentage change variable
co$gdp_change <-
  ((co$gdp - co$prior_gdp) / co$prior_gdp) * 100
```

Our predictor is the percentage change in night-time light emissions over the same period of time, which is defined as:

$$light\_change = \frac{light - prior\_light}{prior\_light} \times 100$$

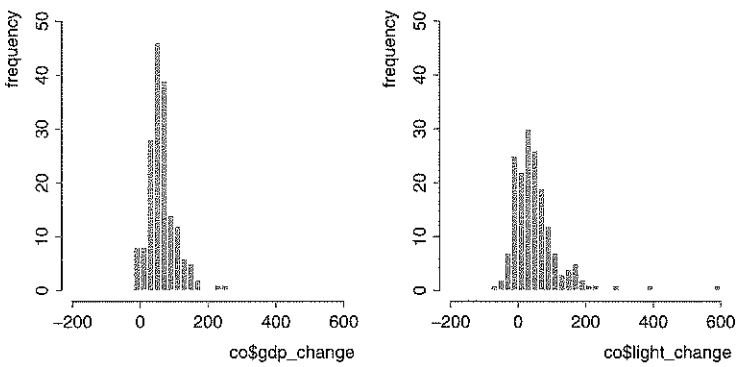
To create this variable, we run:

```
## create light percentage change variable
co$light_change <-
  ((co$light - co$prior_light) / co$prior_light) * 100
```

We could check that the new variables were created correctly by looking at the first few observations of the dataframe *co*. If you run `head(co)` again, you should see that the value of the first observation of *gdp_change* is approximately 51, and the value of the first observation of *light_change* is about -6. Since both new variables measure change as a percentage, we interpret the first number as indicating that the GDP of that country grew by 51% in the 13-year period under study; we interpret the second number as indicating that the night-time light emissions in that same country declined by 6% during the same time period.

To get a better sense of the contents of *gdp_change* and *light_change*, we can create their histograms by running:

```
## create histograms
hist(co$gdp_change) # of percentage change in gdp
hist(co$light_change) # of percentage change in light
```



TIP: Do not confuse percentage change with percentage-point change. The percentage change is defined as the change relative to the baseline:

$$\frac{Y_{final} - Y_{initial}}{Y_{initial}} \times 100$$

In contrast, the percentage-point change is defined as the difference between the final and initial values when these values are measured in percentages:

$$Y_{final} - Y_{initial} \quad \text{(both measured in \%)}$$

For example, if the voter turnout rate increased from 50% to 60%, the percentage change would be:

$$\frac{60\% - 50\%}{50\%} \times 100 = 20\%$$

And, the percentage-point change:

$$60\% - 50\% = 10 \text{ p.p.}$$

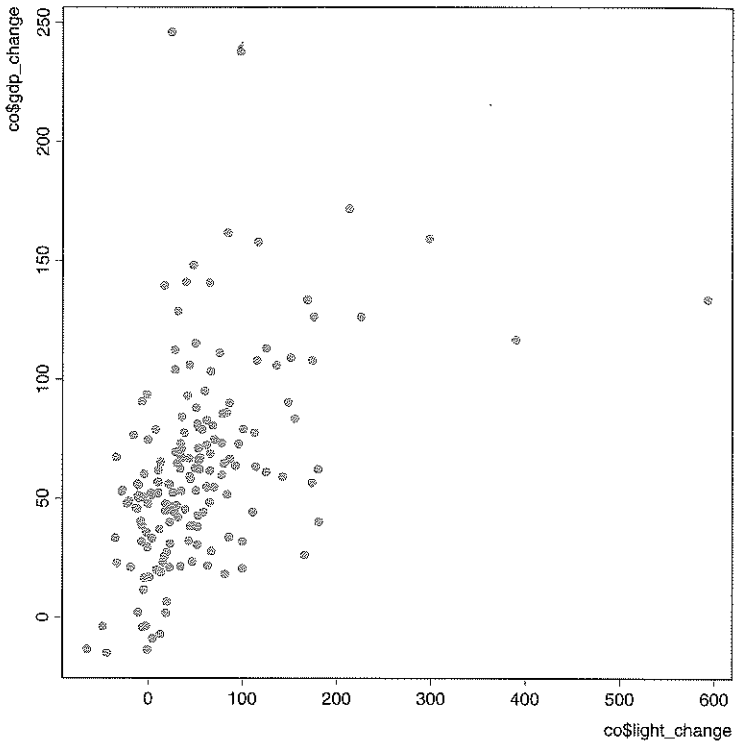
We could describe this change as an increase of either (a) 20 percent or (b) 10 percentage points.

Here we observe that both variables are more or less normally distributed and that while almost all countries saw their GDP grow by between 0 and 200% over the 13-year period, a fair number of countries saw their night-time light emissions either grow by more than 200% or actually decline.

RECALL: We always plot the predictor on the x-axis and the outcome variable on the y-axis.

Now that we have constructed and learned how to interpret our two variables of interest, we can create their scatter plot to get a sense of how they relate to each other:

```
## create scatter plot
plot(x=co$light_change, y=co$gdp_change)
```



As expected, looking at the scatter plot, we can see that higher values of night-time light change tend to be associated with higher values of GDP change. In other words, increases in a country's night-time light emissions are usually accompanied by increases in that country's GDP. The relationship appears to be only moderately linear, however. To confirm this, we compute the correlation coefficient:

```
cor(co$gdp_change, co$light_change) # computes correlation
## [1] 0.4577672
```

The correlation between the two variables is 0.46, which is consistent with what we saw in the scatter plot above.

To predict GDP growth using the change in night-time light emissions, we are interested in the following linear model:

$$\widehat{gdp\_change}_i = \widehat{\alpha} + \widehat{\beta}\ light\_change_i \quad (i\text{=countries})$$

where:

- $\widehat{gdp\_change}_i$ is the average predicted percentage change in GDP from 1992–1993 to 2005–2006 among countries in which the value of *light_change* equals $light\_change_i$
- $light\_change_i$ is the percentage change in night-time light emissions experienced by country i from 1992–1993 to 2005–2006.

To estimate the coefficients of the linear model, we can use the function `lm()` and run:

```
lm(gdp_change ~ light_change, data=co) # fits linear model
##
## Call:
## lm(formula = gdp_change ~ light_change, data = co)
##
## Coefficients:
## (Intercept)  light_change
##     49.8202        0.2546
```

Based on the estimated coefficients above, we write the fitted model as:

$$\widehat{gdp\_change} = 49.82 + 0.25\ light\_change$$

Now we can use the fitted model to make predictions. Imagine, for example, that we want to know a country's GDP growth over a period of 13 years but do not have the data necessary to measure it. Suppose also that we observe that night-time light emissions increased by 20% in that country over the same period of time. What would be our best guess for its GDP growth? To compute this prediction, we plug into the fitted linear model a *light_change* equal to 20:

$$\begin{aligned}\widehat{gdp\_change} &= 49.82 + 0.25\ light\_change \\ &= 49.82 + 0.25 \times 20 = 54.82\end{aligned}$$

Based on the fitted model, we predict that the country's GDP grew by an average of about 55% during the 13-year period.

4.6 MEASURING HOW WELL THE MODEL FITS THE DATA WITH THE COEFFICIENT OF DETERMINATION, R^2

R^2, also known as the **coefficient of determination**, ranges from 0 to 1 and measures the proportion of the variation of the outcome variable explained by the model. The higher the R^2, the better the model fits the data.

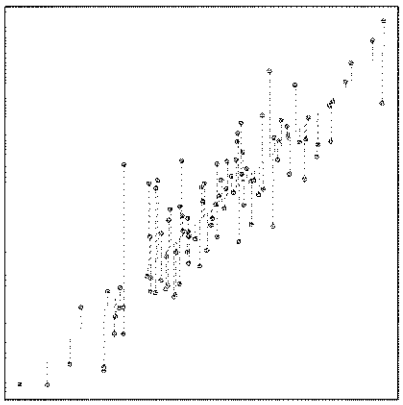
Whenever we use a model to make predictions, we want to know how well the model fits the data because a poor fit can lead to inaccurate predictions. For this purpose, we use a statistic called **coefficient of determination**, or R^2 (pronounced r-squared). The value of R^2 ranges from 0 to 1 and represents the proportion of the variation of Y explained by the model. For example, we interpret an R^2 of 0.8 as indicating that the model explains 80% of the variation of Y ($0.8 \times 100 = 80\%$). Therefore, the higher the R^2, the better the model fits the data.

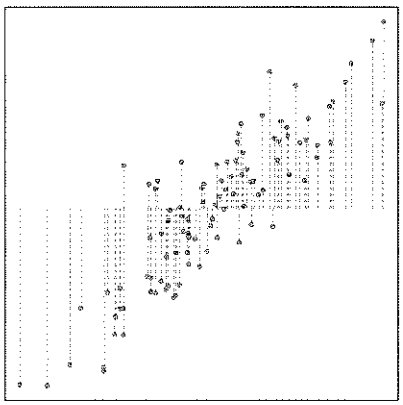
FORMULA IN DETAIL

In mathematical terms, R^2 is defined as:

$$R^2 = 1 - \frac{SSR}{TSS} = 1 - \frac{\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2}{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}$$

where:





- SSR stands for the "sum of the squared residuals" and measures the variation of Y *not* explained by the model. This is what we minimize by using the least squares method when choosing the line of best fit. More precisely:

$$SSR = \sum_{i=1}^{n} \widehat{\epsilon}_i^2 = \sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2$$

  In other words, SSR sums the squared distances between the dots and the line of best fit (shown as dashed red lines in the top figure in the margin).

- TSS stands for "total sum of squares" and measures the total variation of Y, explained and unexplained. This is the numerator of the variance of Y, which, as we saw in chapter 3, is a measure of the spread of the variable. More precisely:

$$TSS = \sum_{i=1}^{n}(Y_i - \overline{Y})^2$$

  In other words, TSS sums the squared distances between the dots and the mean of Y (shown as dashed red lines in the bottom figure in the margin).

Given the definitions above, we can interpret SSR/TSS as the proportion of the variation of Y *not* explained by the model. Therefore, $1-(SSR/TSS)$ is the proportion of the variation of Y that is explained by the model.

At one extreme, when the model perfectly fits the data, the model will produce no residuals, SSR will equal 0, and R^2 will equal 1. At the other extreme, when the model does not explain any of the variation of the outcome variable, SSR will equal TSS, and R^2 will equal 0. Most situations fall somewhere in between.

When we use a simple linear model, that is, a linear model with only one X variable, as is the case in this chapter, R^2 is also equivalent to the correlation between X and Y squared:

$$R^2 = cor(X, Y)^2$$

Given this definition of R^2, it becomes clear that the higher the correlation between X and Y (in absolute terms), the better the model fits the data. As the linear association between X and Y becomes stronger (for example, moving from the first scatter plot in figure 4.4 to the second one), the prediction errors in the model (the vertical distance between the dots and the line) become smaller, and the proportion of the variation of Y explained by the model (the value of R^2) increases.

TIP: Linear models with only one X variable are known as simple linear regression models (or just simple linear models) to differentiate them from multiple linear regression models, which use more than one X variable. Linear models with only one X variable are also known as bivariate linear models because they estimate the relationship between two variables, X and Y ("bi" means two, and "variate" means variable).

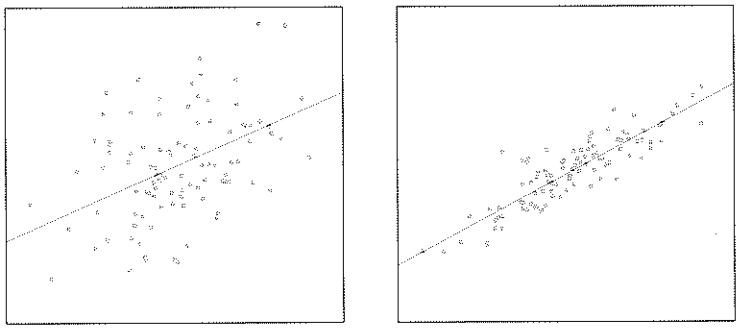


FIGURE 4.4. The higher the absolute value of the correlation between X and Y, the higher the R^2 and the better the model fits the data. For example, the correlation between the variables in the first plot is 0.48, and the R^2 of the model is 0.23 (0.48^2=0.23). By comparison, the correlation between the variables in the second plot is 0.88, and the R^2 of the model is 0.77 (0.88^2=0.77).

At one extreme, when the relationship between X and Y is perfectly linear (the correlation between X and Y equals either 1 or -1), the model explains 100% of the variation of Y (R^2=1^2=1 and R^2=$(-1)^2$=1). At the other extreme, when there is no linear relationship between X and Y (the correlation between X and Y equals 0), the model explains 0% of the variation of Y (R^2=0^2=0).

When building predictive models, then, we look for variables that are highly correlated with Y so that we can use them as predictors. The higher the correlation between X and Y (in absolute terms), the better the fitted linear model will usually be at predicting Y using X.

> PREDICTING OUTCOMES USING LINEAR REGRESSION: We look for X variables that are highly correlated with Y because the higher the correlation between X and Y (in absolute terms), the higher the R^2 and the better the fitted linear model will usually be at predicting Y using X.

4.6.1 HOW WELL DO THE THREE PREDICTIVE MODELS IN THIS CHAPTER FIT THE DATA?

Let's evaluate the three predictive models we fitted in this chapter. Figure 4.5 shows the three scatter plots with their fitted lines.

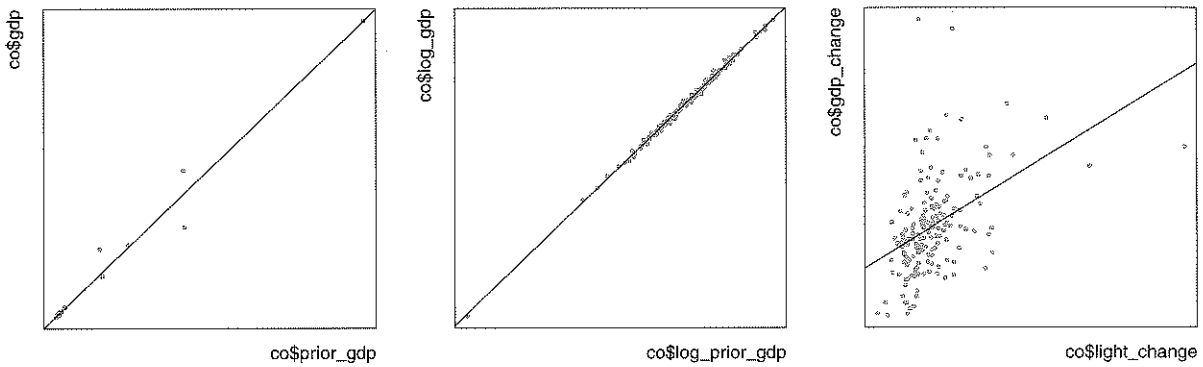


FIGURE 4.5. The first model predicts GDP using prior GDP. The second predicts the natural logarithm of GDP using the natural logarithm of prior GDP. The third predicts GDP growth using changes in night-time light emissions.

They were all simple linear models, so to compute R^2, we can square the correlation between the two variables of interest:

```
## compute R-squared for each predictive model
cor(co$gdp, co$prior_gdp)^2 # model 1
## [1] 0.9807834

cor(co$log_gdp, co$log_prior_gdp)^2 # model 2
## [1] 0.9965443

cor(co$gdp_change, co$light_change)^2 # model 3
## [1] 0.2095928
```

We can interpret the R^2 of the first model as indicating that the linear model that uses prior GDP as a predictor explains about 98% of the variation of GDP. If we compare the R^2 of the first model to that of the second (0.98 vs. approximately 1), we can see that the fit of the model improves ever so slightly by log-transforming both measures of GDP. In either case, the predictive models appear to fit the data remarkably well.

Finally, we can interpret the R^2 of the third model as indicating that the linear model that uses night-time light emission changes as a predictor explains about 21% of the variation in GDP growth. While this might appear to be a low R^2 at first, given how difficult it is to predict GDP growth, this model is quite good in relative terms. (Note that we should only compare R^2s between models that have the same outcome variable because some outcomes are intrinsically harder to predict than others.)

4.7 SUMMARY

This chapter introduced us to the linear regression model for making predictions. We learned how to fit a line to summarize the relationship between a predictor and an outcome variable. Then, we learned how to use the fitted line to (i) predict the average value of the outcome variable given a value of the predictor, and (ii) predict the average change in the outcome variable associated with a change in the value of the predictor. Along the way, we learned about prediction errors, the difference between observed and predicted outcomes, and how to interpret the two coefficients of a line: the intercept and the slope. We ended the chapter by learning how to compute and interpret R^2 to measure how well a model fits the data. In the next chapter, we will see how to use the linear regression model for the purpose of estimating causal effects.

4.8 APPENDIX: INTERPRETATION OF THE SLOPE IN THE LOG-LOG LINEAR MODEL

In the log-log linear model, both the outcome and the predictors have been log-transformed:

$$\widehat{\log(Y)} = \widehat{\alpha} + \widehat{\beta}\log(X)$$

RECALL: The slope refers to the change in the predicted outcome between two specific points on the line. In addition, the change in a variable between two points (initial and final) is equivalent to the difference between the value of the variable at the final point and the value of the variable at the initial point. For example:

$$\triangle\widehat{Y} = \widehat{Y}_{\text{final}} - \widehat{Y}_{\text{initial}}$$

Since we are interested in the interpretation of $\widehat{\beta}$, let's start with the formula for the change in the predicted outcome between two points on the line:

$$\begin{aligned}\widehat{\log(Y_{\text{final}})} - \widehat{\log(Y_{\text{initial}})} &= \left[\widehat{\alpha}+\widehat{\beta}\log(X_{\text{final}})\right] - \left[\widehat{\alpha}+\widehat{\beta}\log(X_{\text{initial}})\right] \\ &= \widehat{\alpha} - \widehat{\alpha} + \widehat{\beta}\log(X_{\text{final}}) - \widehat{\beta}\log(X_{\text{initial}}) \\ &= \widehat{\beta}\left[\log(X_{\text{final}}) - \log(X_{\text{initial}})\right]\end{aligned}$$

If we multiply both sides by 100, we arrive at:

$$\left[\widehat{\log(Y_{\text{final}})} - \widehat{\log(Y_{\text{initial}})}\right] \times 100 = \widehat{\beta}\left[\log(X_{\text{final}}) - \log(X_{\text{initial}})\right] \times 100$$

TIP: Based on the formula in detail below, we can make the following approximations:

$$\left[\widehat{\log(Y_{\text{final}})} - \widehat{\log(Y_{\text{initial}})}\right] \times 100 \approx \approx \frac{\triangle\widehat{Y}}{\widehat{Y}_{\text{initial}}} \times 100$$

$$\left[\log(X_{\text{final}}) - \log(X_{\text{initial}})\right] \times 100 \approx \approx \frac{\triangle X}{X_{\text{initial}}} \times 100$$

Now, if we use the approximations shown in the TIP in the margin, the formula becomes:

$$\frac{\triangle\widehat{Y}}{\widehat{Y}_{\text{initial}}} \times 100 \approx \widehat{\beta}\,\frac{\triangle X}{X_{\text{initial}}} \times 100$$

where:

- $\triangle\widehat{Y}/\widehat{Y}_{\text{initial}} \times 100$ is the predicted percentage change in the outcome variable
- $\widehat{\beta}$ is the estimated slope coefficient
- $\triangle X/X_{\text{initial}} \times 100$ is the percentage change in the predictor.

Given the formula above, if the predictor increases by 1 percent (that is, $\triangle X/X_{\text{initial}} \times 100 = 1$), then the outcome is predicted to increase by $\widehat{\beta}$ percent:

$$\frac{\triangle\widehat{Y}}{\widehat{Y}_{\text{initial}}} \times 100 \approx \widehat{\beta} \times 1 \approx \widehat{\beta}$$

Putting it all together, in the log-log model, the estimated slope coefficient $\widehat{\beta}$ is the predicted *percentage* change in the outcome associated with an increase in the predictor of 1 *percent*.

FORMULA IN DETAIL

The difference between the natural logarithms of two values in a variable is approximately equal to the percentage change between those two values in that variable, when the distance between the two values is relatively small. Here is the math:

$$\left[\log(X_{\text{final}}) - \log(X_{\text{initial}})\right] \times 100 =$$

$$= \left[\log(X_{\text{initial}}+\Delta X) - \log(X_{\text{initial}})\right] \times 100 \qquad \text{because } X_{\text{final}} = X_{\text{initial}} + \Delta X$$

$$= \left[\log\left(X_{\text{initial}} + X_{\text{initial}}\frac{\Delta X}{X_{\text{initial}}}\right) - \log(X_{\text{initial}})\right] \times 100 \qquad \text{because } \frac{X_{\text{initial}}}{X_{\text{initial}}} = 1$$

$$= \left[\log\left(X_{\text{initial}}\left(1 + \frac{\Delta X}{X_{\text{initial}}}\right)\right) - \log(X_{\text{initial}})\right] \times 100$$

$$= \left[\log(X_{\text{initial}}) + \log\left(1 + \frac{\Delta X}{X_{\text{initial}}}\right) - \log(X_{\text{initial}})\right] \times 100 \qquad \text{because } \log(A\times B) = \log(A) + \log(B)$$

$$= \log\left(1 + \frac{\Delta X}{X_{\text{initial}}}\right) \times 100$$

$$\approx \frac{\Delta X}{X_{\text{initial}}} \times 100 \qquad \text{because } \log(1+A) \approx A \text{ when A is small}$$

4.9 CHEATSHEETS

4.9.1 CONCEPTS AND NOTATION

concept/notation	description	example(s)
predictor (X)	variable that we use as the basis for our predictions; predictors are also known as independent variables	when trying to predict a country's current GDP based on prior GDP, the predictor is prior GDP
outcome variable (Y)	variable that we are trying to predict based on the values of the predictor(s); outcome variables are also known as dependent variables	when trying to predict a country's current GDP based on prior GDP, the outcome variable is current GDP
predicted outcomes ($\widehat{Y}$)	pronounced Y-hat; values of Y we predict based on (i) the fitted model that summarizes the relationship between X and Y, and (ii) the observed values of X	(see computing $\widehat{Y}$ based on X below)
observed outcomes (Y)	observed values of Y, in contrast with predicted values of Y, which are estimated, not observed	(see prediction errors below)
prediction errors ($\widehat{\epsilon}$)	pronounced epsilon-hat; also known as residuals; difference between the observed outcomes and the predicted outcomes: $\widehat{\epsilon}_i = Y_i - \widehat{Y}_i$ for each observation, this difference is equivalent to the vertical distance between the dot and the fitted line	if the observed outcome equals 5 and the predicted outcome equals 3, the prediction error equals 2 $\widehat{\epsilon}_i = 5 - 3 = 2$
linear model	also known as simple linear regression model, simple linear model, and bivariate linear model; theoretical model that we assume reflects the true relationship between X and Y $Y_i = \alpha + \beta X_i + \epsilon_i$ where: - Y_i is the outcome for observation i - α is the intercept coefficient - β is the slope coefficient - X_i is the value of the predictor for observation i - ϵ_i is the error for observation i	$Y_i = 2 - 3X_i + \epsilon_i$
fitted linear model	also known as fitted simple linear regression model and fitted simple linear model; line fitted to the data to summarize the relationship between X and Y $\widehat{Y}_i = \widehat{\alpha} + \widehat{\beta} X_i$ where: - $\widehat{Y}_i$ is the predicted outcome for observation i - $\widehat{\alpha}$ is the estimated intercept coefficient - $\widehat{\beta}$ is the estimated slope coefficient - X_i is the value of the predictor for observation i	$\widehat{Y}_i = 2 - 3X_i$

continues on next page...

4.9.1 CONCEPTS AND NOTATION (CONTINUED)

concept/notation	description	example(s)
estimated intercept ($\widehat{\alpha}$)	pronounced alpha-hat; estimated coefficient of the fitted line that specifies the vertical location of the line it is the $\widehat{Y}$ when X=0 unit of measurement of $\widehat{\alpha}$: - if Y is non-binary: in the same unit of measurement as Y - if Y is binary: in percentages (after multiplying the result by 100)	if $\widehat{Y} = 2 - 3X$: the estimated intercept, $\widehat{\alpha}$, is 2 when X equals 0, we predict that Y will equal 2 units, on average
estimated slope ($\widehat{\beta}$)	pronounced beta-hat; estimated coefficient of the fitted line that specifies the angle, or steepness of the line; it equals the change in the predicted outcome divided by the change in the predictor between two points on the line ("rise over run") it is the $\triangle\widehat{Y}$ associated with $\triangle X$=1 interpret as: - an average increase in Y if positive - an average decrease in Y if negative - no average change in Y if zero unit of measurement of $\widehat{\beta}$: - if Y is non-binary: in the same unit of measurement as Y - if Y is binary: in percentage points (after multiplying the result by 100)	if $\widehat{Y} = 2 - 3X$: the estimated slope, $\widehat{\beta}$, is -3 when X increases by 1, we predict an associated decrease in Y of 3 units, on average
computing $\widehat{Y}$ based on X	plug the value of X into the fitted linear model: $\widehat{Y} = \widehat{\alpha} + \widehat{\beta}X$ unit of measurement of $\widehat{Y}$: - if Y is non-binary: in the same unit of measurement as Y - if Y is binary: in percentages (after multiplying the result by 100)	if $\widehat{Y} = 2 - 3X$ and X=2: $\widehat{Y} = 2 - 3\times2 = -4$ when X equals 2, we predict that Y will equal -4 units, on average
computing $\triangle\widehat{Y}$ associated with $\triangle X$	plug the value of $\triangle X$ into the formula below: $\triangle\widehat{Y} = \widehat{\beta}\triangle X$ interpret as: - an average increase in Y if positive - an average decrease in Y if negative - no average change in Y if zero unit of measurement of $\triangle\widehat{Y}$: - if Y is non-binary: in the same unit of measurement as Y - if Y is binary: in percentage points (after multiplying the result by 100)	if $\widehat{Y} = 2 - 3X$ and $\triangle X$=2: $\triangle\widehat{Y} = -3\times2 = -6$ when X increases by 2, we predict an associated decrease in Y of 6 units, on average

continues on next page...

4.9.1 CONCEPTS AND NOTATION (CONTINUED)

concept/notation	description	example(s)
fitted log-log linear model	fitted linear model in which both Y and X have been log-transformed; in this model, we interpret the slope coefficient as the predicted percentage change in the outcome associated with an increase in the predictor of 1 percent $\widehat{\log(Y)} = \widehat{\alpha} + \widehat{\beta}\log(X)$	if $\widehat{\log(Y)} = 2 + 3\log(X)$: the estimated slope, $\widehat{\beta}$, is 3 when X increases by 1%, we predict an associated increase in Y of 3%, on average
R^2 or coefficient of determination	pronounced r-squared; statistic that measures the proportion of the variation of the outcome variable explained by the model it ranges from 0 to 1 the higher the R^2, the better the model fits the data in the simple linear model: $R^2 = cor(X, Y)^2$ when building predictive models, we look for X variables that are highly correlated with Y because the higher the correlation between X and Y (in absolute terms), the higher the R^2 and the better the fitted linear model will usually be at predicting Y using X	if the R^2 of a model equals 0.80, it means that 80% of the variation of the outcome variable is explained by the model

4.9.2 R FUNCTIONS

function	description	required argument(s)	example(s)
lm()	fits a linear model	formula of the type $Y \sim X$, where Y identifies the outcome variable and X identifies the X variable optional argument data: specifies the object where the dataframe is stored; alternative to using $ for each variable	## both of these pieces of code fit the same linear model: lm(*data*$*y_var* ~ *data*$*x_var*) lm(*y_var* ~ *x_var*, data=*data*)

abline()	adds a straight line to the most recently created graph	to add the fitted line, we specify as the main argument the object that contains the output of the lm() function; (for other uses, see page 97)	fit <- lm(*y_var* ~ *x_var*, data=*data*)
			# stores fitted line into an object named fit
			abline(fit)
			# adds the fitted line to the most recently created graph

log()	computes the natural logarithm	what we want to compute the natural logarithm of	log(10)