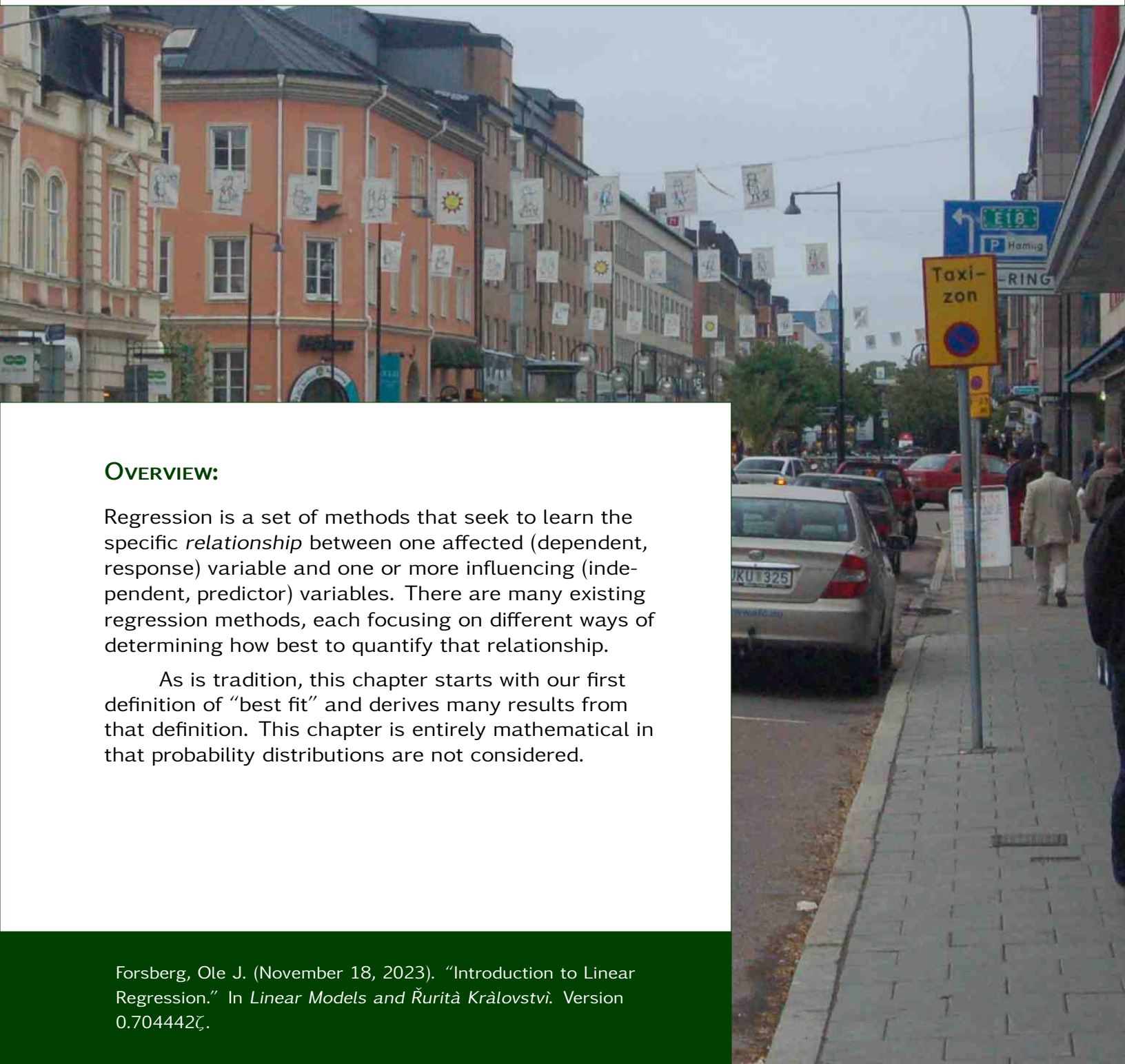
Chapter 2:

Introduction to Linear Regression



Overview:

Regression is a set of methods that seek to learn the specific *relationship* between one affected (dependent, response) variable and one or more influencing (independent, predictor) variables. There are many existing regression methods, each focusing on different ways of determining how best to quantify that relationship.

As is tradition, this chapter starts with our first definition of "best fit" and derives many results from that definition. This chapter is entirely mathematical in that probability distributions are not considered.

Forsberg, Ole J. (November 18, 2023). "Introduction to Linear Regression." In *Linear Models and Řuritȁ Krȁlovstvì*. Version 0.704442ζ.

Chapter Contents

❧ ❧ ❧

Let x and y be numeric variables. The linear relationship between x and y can be summarized by a line that "best" fits the observed data. That is, we will summarize the relationship between x and y using a linear equation:

$$y = \beta_0 + \beta_1 x + \varepsilon \tag{2.1}$$

What we mean by "best" determines where we go from here. In thinking about "best," it may help to see some sample data and the 'line of best fit' for it (Figure 2.1). There are at least three definitions of "best" that we can use:

1. Maximize the likelihood that the data were generated
2. Minimize the sum of the absolute value of the residuals
3. Minimize the sum of the square of the residuals

estimator

All three definitions are entirely legitimate — as are many other definitions. However, each leads to different estimation methods — and estimators. In a well-formed model, the substantive conclusions will rarely differ.

MLE

The first definition leads to maximum likelihood estimation, which will be covered in Chapter 9. It is an excellent technique that can be generalized to many more settings than can ordinary least squares. Its greatest strength, however, is that it makes use of the researcher's greater understanding of the data-generating process (Chapters 9 to 14).

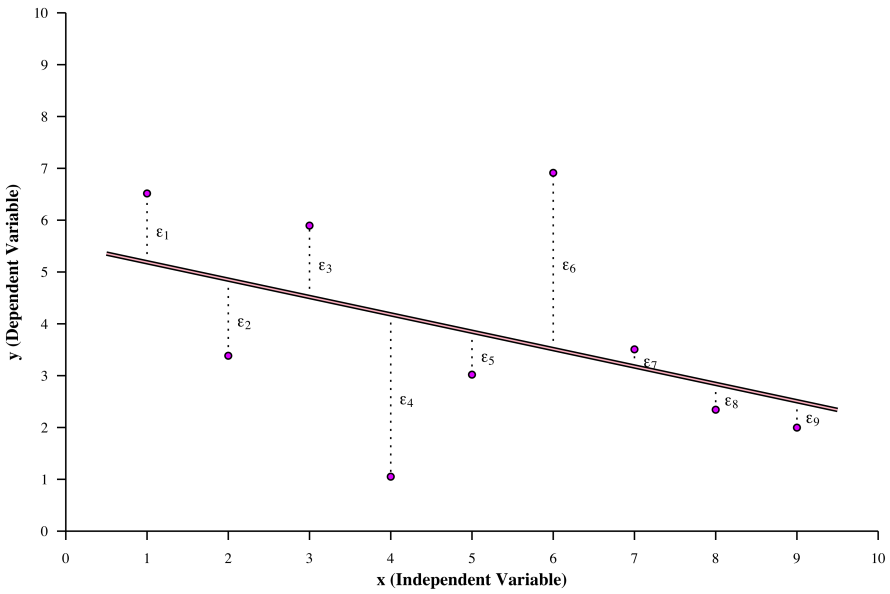


Figure 2.1: *Sample data and a line of best fit for that data. Also marked are the residuals, the difference between what was observed (dots) and what is predicted by the model (line). This particular line of best fit minimizes the sum of squared residuals.*

The second definition leads to a type of robust regression frequently termed median regression. This method is very helpful for times when there are outliers in the data that you cannot (should not) remove. The drawback to this method is that estimating the two parameters (β_0 and β_1) does not provide a closed-form solution. In other words, it requires a repetitive sequence of steps and can only approximate those estimates. It requires an approximation process that is computationally intensive. Because of this, median regression was little used until recently. The statistical theory behind it is not as well explored as other types. We will explore this in Chapter 8. **robust**

The most popular definition of "best," and the one that starts our journey, is the final definition. It leads to an estimation method called ordinary least squares (OLS). It is rather straight-forward to minimize a sum of squared values using differential calculus. One strength is that an equation results from this process — a closed-form solution with no need for iteration. This means that the process returns mathematically exact values. The drawback is that it is limited in the types of processes that can be modeled. **OLS**

We start exploring ordinary least squares immediately.

2.1: Scalar Representation

This section and the next show how our definition of "best" mathematically leads to specific results. That leading can be done by representing the regression problem in scalar or in matrix terms. At one level, there is no difference in the two representations. At another level, one representation may make proofs — and understandings — easier and more manifest.

And so, let us begin with the scalar representation of the regression problem. From experience, it seems to make more sense than starting with the matrix representation (Section 2.2).

Ordinary least squares estimation defines "best" as "having the lowest sum of squared errors." So, let us use this definition to obtain the OLS estimators of β_0 and β_1. In calculus, to optimize (maximize or minimize) a function, one takes its derivative(s) with respect to the parameter(s) of interest, set the resulting equations equal to 0, then solve the system of equations.[1]

The first step, as expected, is to form the objective function that we want to minimize. Since we seek to minimize the sum of squared errors, that Q is the sum of squared errors:

$$Q = \sum_{i=1}^{n} \varepsilon_i^2 \tag{2.2}$$

$$= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{2.3}$$

$$= \sum_{i=1}^{n} \Big(y_i - (\beta_0 + \beta_1 x_i) \Big)^2 \tag{2.4}$$

$$= \sum_{i=1}^{n} \Big(y_i - \beta_0 - \beta_1 x_i \Big)^2 \tag{2.5}$$

Now that we have the objective function, we take its derivative with respect to each parameter, set it equal to 0, and solve for that parameter.

[1] Also, one should perform the second derivative test to determine the type of optimization point found: minimum, maximum, and saddle point (neither).

Let us start with β_0:

$$\frac{\partial}{\partial \beta_0} Q = \sum_{i=1}^{n} -2\left(y_i - \beta_0 - \beta_1 x_i\right) \tag{2.6}$$

$$0 \overset{\text{set}}{=} \sum_{i=1}^{n} -2\left(y_i - b_0 - b_1 x_i\right) \tag{2.7}$$

$$= \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} b_0 - b_1 \sum_{i=1}^{n} x_i \tag{2.8}$$

$$= n\overline{y} - nb_0 - nb_1\overline{x} \tag{2.9}$$

This immediately leads to

$$b_0 = \overline{y} - b_1\overline{x} \tag{2.10}$$

This is the OLS estimator of β_0 in terms of b_1. Next, we take the derivative of Q with respect to the second parameter β_1:

$$\frac{\partial}{\partial \beta_1} Q = \sum_{i=1}^{n} -2x_i\left(y_i - \beta_0 - \beta_1 x_i\right) \tag{2.11}$$

$$0 \overset{\text{set}}{=} \sum_{i=1}^{n} -2x_i\left(y_i - b_0 - b_1 x_i\right) \tag{2.12}$$

$$= \sum_{i=1}^{n} x_i y_i - b_0 \sum_{i=1}^{n} x_i - b_1 \sum_{i=1}^{n} x_i^2 \tag{2.13}$$

$$= \sum_{i=1}^{n} x_i y_i - nb_0\overline{x} - b_1 \sum_{i=1}^{n} x_i^2 \tag{2.14}$$

Substituting our estimator b_0, we have

$$0 = \sum_{i=1}^{n} x_i y_i - \left(\overline{y} - b_1\overline{x}\right) n\overline{x} - b_1 \sum_{i=1}^{n} x_i^2 \tag{2.15}$$

$$= \sum_{i=1}^{n} x_i y_i - n\overline{x}\,\overline{y} + b_1 n\overline{x}^2 - b_1 \sum_{i=1}^{n} x_i^2 \tag{2.16}$$

$$b_1\left(\sum_{i=1}^{n} x_i^2 - n\overline{x}^2\right) = \sum_{i=1}^{n} x_i y_i - n\overline{x}\,\overline{y} \tag{2.17}$$

Finally, we have

$$b_1 = \frac{\sum_{i=1}^{n} x_i y_i - n\overline{x}\,\overline{y}}{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2} \tag{2.18}$$

Thus, the two OLS estimators of β_0 and β_1 are

$$\begin{aligned} b_0 &= \overline{y} - b_1\overline{x} \\ b_1 &= \frac{\sum_{i=1}^{n} x_i y_i - n\overline{x}\,\overline{y}}{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2} \end{aligned} \tag{2.19}$$

requirement

Note that this mathematical process had but one requirement:

$$\sum_{i=1}^{n} x_i^2 - n\overline{x}^2 \neq 0 \tag{2.20}$$

exercise

If that requirement is not met by the data, then the divisor of b_1 is zero, which leads to dividing by zero, armageddon, and a *really* bad hair day. However, note that $\sum_{i=1}^{n} x_i^2 - n\overline{x}^2$ is just $(n-1)s_x^2$. As such, the requirement is met when the variance of x is non-zero. In other words, we require that the independent variable varies.

> ***Note***: Technically, we also need to perform the second derivative test to show that these critical values constitute minimums instead of maximums or saddle points. I leave that as an exercise for you.

exercise

Also note that some sources will give the b_1 a different, yet equivalent, formula:

$$b_1 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \tag{2.21}$$

exercise

I leave it as an exercise to show that $\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})$ is equivalent to $\sum_{i=1}^{n} x_i y_i - n\overline{x}\,\overline{y}$ and that $\sum_{i=1}^{n}(x_i - \overline{x})^2$ is equivalent to $\sum_{i=1}^{n} x_i^2 - n\overline{x}^2$.

Finally, let me remind you that we will come across the denominator in many settings. Thus, we will symbolize it as S_{xx}:

$$S_{xx} = \sum_{i=1}^{n}(x_i - \overline{x})^2 \tag{2.22}$$

Thus, our OLS line of best fit is the line defined by the set of points $(x, \hat{y})$, where

$$\hat{y} = b_0 + b_1 x \tag{2.23}$$

Note that $\hat{y}$ is the expected (or predicted) value of Y (dependent variable), given that value of x (independent variable). In other words,

$$\hat{y}_i = \mathbb{E}[Y \mid x_i] \tag{2.24}$$

It is the conditional mean of Y given x_i; the expected value of Y, given this value of x_i; the mean of Y when the independent variable has value x_i.

❧ ❧ ❧

And this is as far as we can go without making additional assumptions/requirements. As such, it marks a great place for a toy example.

Example 2.1: Let us measure two variables on four subjects. Those two variables are x and y. For the first subject, the value of x is -2 and the value of y is 3. For the second subject the x and y values are 0 and 0. For the third subject, the values are 0 and 2. For the fourth subject, they are 2 and -1. This data is tabulated in Table 2.1

Given this information, let us calculate the ordinary least squares estimators of β_0 and β_1. •

x	y
-2	3
0	0
0	2
2	-1

Table 2.1: *Toy data to be used for toy Example 2.1 that could be about toys.*

Solution: First, the formulas for b_0 and b_1 require we calculate $\overline{x}$ and $\overline{y}$. They are 0 and 1, respectively. With that, we can use the formula for b_1 (Equation 2.19b):

$$b_1 = \frac{\sum_{i=1}^{n} x_i y_i - n\overline{x}\,\overline{y}}{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2} \tag{2.25}$$

$$= \frac{\Big(-2(3) + 0(0) + 0(2) + 2(-1)\Big) - 4(0)1}{\Big((-2)^2 + (0)^2 + (0)^2 + (2)^2\Big) - 4(0)^2} \tag{2.26}$$

$$= \frac{-8 - 0}{8 - 0} \tag{2.27}$$

$$= -1 \tag{2.28}$$

For the OLS estimator of the intercept, β_0, we have (Equation 2.19a):

$$b_0 = \overline{y} - b_1\overline{x} \tag{2.29}$$

$$= 1 - (-1)0 \tag{2.30}$$

$$= 1 \tag{2.31}$$

Thus, the OLS line of best fit is the line defined by the set of points $(x, \hat{y})$, where

$$\hat{y} = 1 - 1x \tag{2.32}$$

Figure 2.2 shows the points and the OLS line of best fit.

So, what does the equation *mean*? It means that the expected value of Y when $x = 0$ is 1, the y-intercept. It also means that for every one increase in the value of x, the expected (predicted) value of Y increases by -1, the value of the slope. ♦

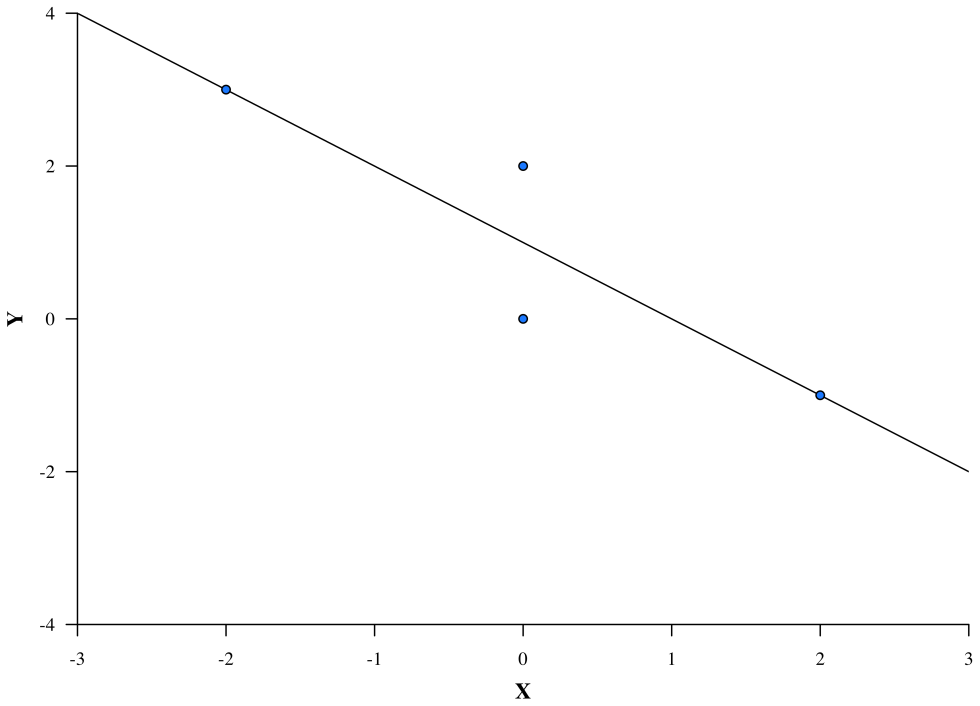


Figure 2.2: *Graphic of the data and the OLS line of best fit for the toy data of Example 2.1.*

Note: From a scientific standpoint, it does not make sense to meaningfully interpret the y-intercept when $x = 0$ is outside the observed range of the data (x-values). Models are best when you are trying to understand the relationship within the observed ranges of the independent variable(s). This is interpolation — "inter" from "within." **interpolation**

Trying to use the model to understand relationships outside the observed values of the independent variables is called "extrapolation," where 'extra' means 'outside.' Extrapolation is dangerous: all curves look linear at a small enough scale (remember Taylor's Theorem from Calculus). Thus, fitting the data with a line may be a good approximation in one scale, it may not make sense at a wider range, where the non-linearity of the relationship may become more pronounced.

Example 2.2: Let us measure two variables on four subjects. Those two variables are x and y. Sounds familiar? The difference here is that the independent variable is dichotomous.

Given the information in Table 2.2, let us calculate — and interpret — the ordinary least squares estimators of β_0 and β_1. •

dichotomous

Solution: The first thing to do is change our x-variable into a numeric variable. When the variable is dichotomous (has only two possible values), this is easy. Set one value to 0 and the other to 1. So, without loss of generality, let us follow the alphabet and replace `Female` with 0 and `Male` with 1. With this transformation, the x-values are now $\{1,0,0,1\}$ and we can use the same procedure as we used in Example 2.1.

First, the formulas for b_0 and b_1 require we calculate $\overline{x}$ and $\overline{y}$. They are 0.5 and 2, respectively. With that, we can use the formula for b_1 (Equation 2.19b):

$$b_1 = \frac{\sum_{i=1}^{n} x_i y_i - n\overline{x}\,\overline{y}}{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2} \tag{2.33}$$

$$= \frac{\Big(1(3)+0(1)+0(2)+1(2)\Big) - 4(0.5)2}{\Big((1)^2+(0)^2+(0)^2+(1)^2\Big) - 4(0.5)^2} \tag{2.34}$$

$$= \frac{5-4}{2-1} \tag{2.35}$$

$$= 1 \tag{2.36}$$

x	y
Male	3
Female	1
Female	2
Male	2

Table 2.2: *Data to be used for Example 2.2.*

For the OLS estimator of the intercept, β_0, we have (Equation 2.19a):

$$b_0 = \overline{y} - b_1 \overline{x} \tag{2.37}$$

$$= 2 - (1)0.5 \tag{2.38}$$

$$= 1.5 \tag{2.39}$$

Thus, the OLS line of best fit is the line defined by the set of points $(x, \hat{y})$, where

$$\hat{y} = 1.5 + 1\ x \tag{2.40}$$

Now, what does *this* equation mean? Remember that an x-value of 0 indicates we are discussing females. Thus, the expected value of Y for the females is $1.5 + (1)0 = 1.5$. The expected value of Y for the males is $1.5 + (1)1 = 2.5$.

Thus, the y-intercept is the predicted value for base level (females). The "slope" is the effect of gender (moving from female to male) on that y-intercept.

base level

♦

Note: You have seen an analysis of this type in your past introductory statistics course. This is just the two-sample t-procedure under the guise of linear models.

Note: Since we can compare the means of two group in the regression realm (Example 2.2), can we compare the means of more than two groups? In other words, can we extend linear models to ANOVA? The answer is Yes! In fact, ANOVA is built on a base of linear models, as we will see in the future (Example 2.5).

2.1.1 Results

Now that we have formulas for our estimators, we have several important mathematical results. The first result is that the line of best fit passes through the center of gravity.

Theorem 2.1. *The point* $(\overline{x},\ \overline{y})$*, the center of gravity, is on the OLS line of best fit.*

Proof. To see this just substitute $\overline{x}$ for x in the prediction equation and show that $\hat{y} = \overline{y}$.

From equation (2.23),

$$\hat{y} = b_0 + b_1 x \tag{2.41}$$

Substituting $\overline{x}$ for x gives

$$= b_0 + b_1 \overline{x} \tag{2.42}$$

Substituting the value of b_0 gives

$$= (\overline{y} - b_1 \overline{x}) + b_1 \overline{x} \tag{2.43}$$

Finally, simplification gives our result:

$$= \overline{y} \tag{2.44}$$

□

Example 2.3: From the previous example, we just need to show that the point $(\overline{x},\ \overline{y}) = (0.5,\ 2)$ is on the line. •

Solution: We have already shown that the line of best fit is $\hat{y} = 1.5 + x$. Substituting $\overline{x} = 0.5$ gives $\hat{y} = 1.5 + 0.5 = 2$. Note that 2 is also the value of $\overline{y}$. ♦

A second result is that the slope estimator b_1 is the ratio of the covariance between x and y to the variance of x. second

Theorem 2.2. $b_1 = \frac{\mathbb{C}ov[x,y]}{\mathbb{V}[X]} = \frac{s_{xy}}{s_x^2}$

Proof. To see this we substitute the formulas for the covariance and variance into this equation and quickly simplify:

$$b_1 = \frac{s_{xy}}{s_x^2} \tag{2.45}$$

$$= \frac{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2} \tag{2.46}$$

$$= \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \tag{2.47}$$

□

A third result is that the slope estimator can also be represented as third

$$b_1 = r_{xy}\frac{s_y}{s_x} \tag{2.48}$$

That is, the slope estimator is the correlation between the two variables times the ratio of their standard deviations. I leave this as an exercise for you to prove. exercise

A fourth result is that the slope estimator is zero if the y-values do not vary. I leave this as an exercise, as well. exercise

2.1.2 First Assumptions

This was fun. We were able to determine the correct formula for the line of best fit — given our particular definition of "best." Those equations lead to other equations. These are the *mathematical* results for our given sample.

Cool mule.

Until this point, we have only required variation in the independent variable. If we make three additional assumptions, we have additional results.

The first assumption is that they are realizations of a random variable (ε has a distribution). The second is that the expected value of the residuals is zero, $\mathbb{E}[\varepsilon] = 0$ (the measurements are not systematically biased). The third is that the residuals are independent and have a finite and constant variance, $\mathbb{V}[\varepsilon] = \sigma^2 < \infty$.

The above simple assumptions lead to several additional interesting results. Some are proven here, some are left as exercises.

Theorem 2.3. $\mathbb{E}[b_1] = \beta_1$

Proof. To prove this, we will start with the formula for b_1 and simplify until we obtain the results.

$$\mathbb{E}[b_1] = \mathbb{E}\left[\frac{\sum_{i=1}^n (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^n (x_i - \overline{x})^2}\right] \tag{2.49}$$

$$= \mathbb{E}\left[\frac{\sum_{i=1}^n (x_i - \overline{x}) y_i}{\sum_{i=1}^n (x_i - \overline{x})^2}\right] \tag{2.50}$$

$$= \frac{\sum_{i=1}^n (x_i - \overline{x}) \mathbb{E}[y_i]}{\sum_{i=1}^n (x_i - \overline{x})^2} \tag{2.51}$$

$$= \frac{\sum_{i=1}^n (x_i - \overline{x}) (\beta_0 + x_i \beta_1 + \varepsilon)}{\sum_{i=1}^n (x_i - \overline{x})^2} \tag{2.52}$$

$$= \frac{\sum_{i=1}^n (x_i - \overline{x}) \beta_0}{\sum_{i=1}^n (x_i - \overline{x})^2} + \frac{\sum_{i=1}^n (x_i - \overline{x}) x_i \beta_1}{\sum_{i=1}^n (x_i - \overline{x})^2} + \frac{\sum_{i=1}^n (x_i - \overline{x}) \varepsilon}{\sum_{i=1}^n (x_i - \overline{x})^2} \tag{2.53}$$

$$= \frac{\beta_0 \sum_{i=1}^n (x_i - \overline{x})}{\sum_{i=1}^n (x_i - \overline{x})^2} + \frac{\beta_1 \sum_{i=1}^n (x_i - \overline{x}) x_i}{\sum_{i=1}^n (x_i - \overline{x})^2} + \frac{\varepsilon \sum_{i=1}^n (x_i - \overline{x})}{\sum_{i=1}^n (x_i - \overline{x})^2} \tag{2.54}$$

$$= \beta_0 \frac{0}{\sum_{i=1}^{n}(x_i - \overline{x})^2} + \beta_1 \frac{\sum_{i=1}^{n}(x_i - \overline{x})(x_i - \overline{x})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} + \varepsilon \frac{0}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \quad (2.55)$$

$$= \beta_1 \quad (2.56)$$

Thus, the OLS estimator of β_1 is unbiased. This is a nice property. It means $\mathbb{E}[b_1] = \beta_1$. □

Note: Did you notice where we used $\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) = \sum_{i=1}^{n}(x_i - \overline{x})y_i$, $\sum_{i=1}^{n}(x_i - \overline{x})(x_i - \overline{x}) = \sum_{i=1}^{n}(x_i - \overline{x})x_i$, and $\sum_{i=1}^{n}(x_i - \overline{x}) = 0$? All three are just simple algebra.

Theorem 2.4. $\mathbb{E}[b_0] = \beta_0$

Theorem 2.5. $\mathbb{V}[b_1] = \sigma^2 / S_{xx}$

Theorem 2.6. $\mathbb{V}[b_0] = \sigma^2 \left(\frac{1}{n} + \frac{\overline{x}}{S_{xx}}\right)$

Theorem 2.7. $\mathbb{C}ov[b_0, b_1] = -\sigma^2 \frac{\overline{x}}{S_{xx}}$

Finally, if we define the mean square error as **MSE**

$$MSE = \frac{1}{n-2} \sum_{i=1}^{n} e_i^2 \quad (2.57)$$

then we also know

Theorem 2.8. $\mathbb{E}[MSE] = \sigma^2$

Note: In other words, this definition for MSE provides an *unbiased* estimator of the variance of the residuals. This is why we define it in this manner. **unbiased**

2.2: Matrix Representation

We learned a lot about our solution by exploring the scalar representation of the system of equations. We may be able to gain some additional insights by exploring its matrix representation. It may be helpful to revisit Appendix M at this point.

And so, let us begin with our matrix model.

$$\mathbf{Y} = \mathbf{XB} + \mathbf{E} \tag{2.58}$$

In this model, $\mathbf{Y}$ represents the response variable; $\mathbf{X}$, the predictor variable(s) prepended with a column of 1s; $\mathbf{B}$, the coefficient vector; and $\mathbf{E}$, the residuals. The dimensions are $n \times 1$ for $\mathbf{Y}$, $n \times p$ for $\mathbf{X}$, $p \times 1$ for $\mathbf{B}$, and $n \times 1$ for $\mathbf{E}$.

In this formulation, n is the sample size and p is the number of parameters that need to be estimated. Usually, this is one more than the number of independent variables.

Note that $\mathbf{X}$ is "the predictor variable(s) prepended with a column of 1s." What does this mean? Let our independent variable be the same as in Example 2.1, $\{-2, 0, 0, 2\}$. The $\mathbf{X}$ matrix is

$$\mathbf{X} = \begin{bmatrix} 1 & -2 \\ 1 & 0 \\ 1 & 0 \\ 1 & 2 \end{bmatrix} \tag{2.59}$$

Note that it is the values of the x-variable prepended by a column of 1s.

Again, we want to minimize the sum of squared errors. Again, we will create the objective function Q, take its derivative with respect to the parameter vector, $\mathbf{B}$, and solve:

$$\begin{aligned} Q &= \mathbf{E}'\mathbf{E} && (2.60)\\ &= (\mathbf{Y} - \mathbf{XB})'(\mathbf{Y} - \mathbf{XB}) && (2.61)\\ &= \mathbf{Y}'\mathbf{Y} - \mathbf{B}'\mathbf{X}'\mathbf{Y} - \mathbf{Y}'\mathbf{XB} + \mathbf{B}'\mathbf{X}'\mathbf{XB} && (2.62) \end{aligned}$$

Note that each of these terms is a 1×1 matrix, thus each is equal to its transpose. Using that on the third term and gathering the two like terms together gives our objective function.

$$Q = \mathbf{Y}'\mathbf{Y} - 2\mathbf{B}'\mathbf{X}'\mathbf{Y} + \mathbf{B}'\mathbf{X}'\mathbf{XB} \tag{2.63}$$

Now, taking the derivative with respect to $\mathbf{B}$ gives

$$\frac{d}{d\mathbf{B}}Q = -2\mathbf{X}'\mathbf{Y} + 2\mathbf{X}'\mathbf{X}\mathbf{B} \tag{2.64}$$

$$\mathbf{0} \stackrel{\text{set}}{=} -\mathbf{X}'\mathbf{Y} + \mathbf{X}'\mathbf{X}\mathbf{b} \tag{2.65}$$

$$\mathbf{X}'\mathbf{Y} = \mathbf{X}'\mathbf{X}\mathbf{b} \tag{2.66}$$

$$(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = \mathbf{b} \tag{2.67}$$

This formula is so important that I will repeat it here:

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \tag{2.68}$$

Note the switch between $\mathbf{B}$ and $\mathbf{b}$. The former concerns the population. It is a population parameter that we are trying to estimate.

$$\mathbf{B} := \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_{p-1} \end{bmatrix} \tag{2.69}$$

The latter concerns the sample. It is the estimator we are using to estimate the population parameter. **estimator**

$$\mathbf{b} := \begin{bmatrix} b_0 \\ b_1 \\ b_2 \\ \vdots \\ b_{p-1} \end{bmatrix} \tag{2.70}$$

Thus, the equation for our OLS regression line (plane, hyper-plane) is $\hat{\mathbf{Y}} = \mathbf{X}\mathbf{b}$.

2.2.1 Requirement

In performing these calculations, we made one assumption: $(\mathbf{X}'\mathbf{X})^{-1}$ exists. If it does not exist, then the last step in the process cannot be done. So, the first question to ask is: When does $(\mathbf{X}'\mathbf{X})^{-1}$ not exist?

It does not exist when $\det \mathbf{X}'\mathbf{X} = 0$.

So, when does $\det \mathbf{X}'\mathbf{X} = 0$? From linear algebra (and Appendix M) we know that this determinant is zero when the $\mathbf{X}$ matrix is not of full (column) rank , when rank $\mathbf{X} \neq p$. Equivalently, this happens when one column of the $\mathbf{X}$ matrix is a linear combination of the other columns.

full rank

When in the realm of multiple regression (more than one independent variable), this happens when one variable is a linear combination of the others. This condition is called "multicollinearity." When in the realm of simple linear regression (one independent variable), this happens when there is no variation in the x variable (it is a constant multiple of the columns of 1s).

multicollinearity

2.2.2 Assumptions

Before we continue, as before, let us make the three assumptions about our residuals. These are just the same non-parametric assumptions we made back in Section 2.1.2, but in matrix form. The first is that they are realizations of a random variable ($\mathbf{E}$ has a distribution). The second is that the expected value of the residuals is zero, $\mathbb{E}[\mathbf{E}] = \mathbf{0}$ (the measurements are not systematically biased). The third is that the residuals are independent and have a finite constant variance, $\mathbb{V}[\mathbf{E}] = \sigma^2\mathbf{I}$, with $\sigma^2 < \infty$.

2.2.3 Results

Again, we have several results from this.

Theorem 2.9. $\mathbb{E}[\mathbf{Y}] = \mathbf{XB}$

Proof. The proof of this proceeds from algebra.

$$\begin{aligned} \mathbb{E}[\mathbf{Y}] &= \mathbb{E}[\mathbf{XB} + \mathbf{E}] && (2.71)\\ &= \mathbb{E}[\mathbf{XB}] + \mathbb{E}[\mathbf{E}] && (2.72) \end{aligned}$$

One pervasive requirement is that the values of $\mathbf{X}$ are not random variables. That is, the *researcher* selected those particular x values. Since this is true,

$$\mathbb{E}[\mathbf{Y}] = \mathbf{X}\mathbb{E}[\mathbf{B}] + \mathbb{E}[\mathbf{E}] \quad (2.73)$$

Also, the values in the **B** matrix are population parameters. They, too, are not random variables. In fact, the only random variable on the right-hand side of that matrix equation is the zero-mean **E** matrix. Thus, we have

$$\mathbb{E}[\mathbf{Y}] = \mathbf{XB} + \mathbb{E}[\mathbf{E}] \tag{2.74}$$
$$= \mathbf{XB} \tag{2.75}$$

□

> ***Note***: The requirement that the independent variables are are not random allows us to easily calculate expected values, variances, and covariances. When designing experiments, this assumption is not problematic. When working with observational data, this becomes troublesome in terms of the mathematics.

Similarly, it is quite easy to prove $\mathbb{V}[\mathbf{Y} \mid \mathbf{XB}] = \sigma^2\mathbf{I}$. I leave that to you as an exercise. **exercise**

❧ ❧ ❧

Another result is that the two estimators are unbiased (their expected values equal the population parameter):

Theorem 2.10. *The OLS estimator* **b** *is unbiased for* **B**.

Proof. An estimator is unbiased for the parameter if its expected value equals the parameter. Thus, we need only show $\mathbb{E}[\mathbf{b}] = \mathbf{B}$.

$$\mathbb{E}[\mathbf{b}] = \mathbb{E}\left[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}\right] \tag{2.76}$$
$$= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbb{E}[\mathbf{Y}] \tag{2.77}$$
$$= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{XB} \tag{2.78}$$
$$= \mathbf{B} \tag{2.79}$$

□

A third result is that the two estimators are not necessarily independent.

Theorem 2.11. *The OLS estimators b_0 and b_1 are not necessarily independent.*

Proof. To see this, we calculate the covariance matrix of $\mathbf{b}$:

$$\begin{aligned}
\mathbb{V}[\mathbf{b}] &= \mathbb{V}\left[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}\right] && (2.80)\\
&= \left((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right)\mathbb{V}[\mathbf{Y}]\left((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right)' && (2.81)\\
&= \left((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right)\sigma^2\left((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right)' && (2.82)\\
&= \sigma^2\left((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right)\left((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right)' && (2.83)\\
&= \sigma^2\ (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\,\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} && (2.84)\\
&= \sigma^2\ (\mathbf{X}'\mathbf{X})^{-1} && (2.85)
\end{aligned}$$

Now, if this matrix is diagonal, then the estimators are independent.

To see that the two estimators are correlated, we just need to calculate the matrix $(\mathbf{X}'\mathbf{X})^{-1}$. In general, this is rather difficult to do by hand. However, if we restrict ourselves to simple linear regression, that inverse is rather straight-forward because $\mathbf{X}$ is

$$\mathbf{X} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & x_3 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \tag{2.86}$$

With that, we have

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} n & n\overline{x} \\ n\overline{x} & \sum_{i=1}^{n} x_i^2 \end{bmatrix} \tag{2.87}$$

The determinant of $\mathbf{X}'\mathbf{X}$ is

$$\det \mathbf{X}'\mathbf{X} = n\sum_{i=1}^{n} x_i^2 - n^2\overline{x}^2 = n\,S_{xx} \tag{2.88}$$

$$\tag{2.89}$$

Thus, the inverse is

$$(\mathbf{X}'\mathbf{X})^{-1} = \frac{1}{n\,S_{xx}}\begin{bmatrix} \sum_{i=1}^{n} x_i^2 & -n\overline{x} \\ -n\overline{x} & n \end{bmatrix} \tag{2.90}$$

Finally, the covariance matrix is

$$\mathbb{V}[\mathbf{b}] = \frac{\sigma^2}{n S_{xx}} \begin{bmatrix} \sum_{i=1}^{n} x_i^2 & -n\overline{x} \\ -n\overline{x} & n \end{bmatrix} \tag{2.91}$$

From this matrix, we see that the covariance between b_0 and b_1 is

$$\mathbb{C}ov[b_0, b_1] = -n\overline{x}\,\frac{\sigma^2}{n S_{xx}} = -\sigma^2 \frac{\overline{x}}{S_{xx}} \tag{2.92}$$

Thus, the OLS estimators are independent if and only if $\overline{x} = 0$. □

As an extension, note that the sign of the covariance is the opposite that of $\overline{x}$.

Finally, while this last results may seem just slightly interesting, it is the basis of the Working-Hotelling (1929) procedure, which we will see later in Section 3.4.

This last result also suggests why many disciplines tend to center their x-values (subtract off $\overline{x}$) before doing regression. It ensures that the two estimators are independent. **centering**

Example 2.4: Let us revisit Example 2.1 and show how to use the matrix representation to answer the same problem. •

Solution: The first step is to create the two matrices. The dependent variable matrix is

$$\mathbf{Y} = \begin{bmatrix} 3 \\ 0 \\ 2 \\ -1 \end{bmatrix} \tag{2.93}$$

The independent variable matrix, also called the "data matrix" and the "design matrix," is

$$\mathbf{X} = \begin{bmatrix} 1 & -2 \\ 1 & 0 \\ 1 & 0 \\ 1 & 2 \end{bmatrix} \tag{2.94}$$

Where did the column of 1s come from in **X**? Remember that the matrix equation is

$$\mathbf{Y} = \mathbf{XB} + \mathbf{E} \tag{2.95}$$

and that this is equivalent (in simple linear regression) to

$$y_i = \beta_0\ 1 + \beta_1\ x_i + \varepsilon_i \tag{2.96}$$

The 1s column in **X** is the multiplier of the β_0 in the **B** matrix. As long as you have a β_0 in your model, you need that column of 1s.

Now that we have the two matrices, we can calculate **b**.

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \tag{2.97}$$

$$= \left(\begin{bmatrix} 1 & -2 \\ 1 & 0 \\ 1 & 0 \\ 1 & 2 \end{bmatrix}' \begin{bmatrix} 1 & -2 \\ 1 & 0 \\ 1 & 0 \\ 1 & 2 \end{bmatrix}\right)^{-1} \begin{bmatrix} 1 & -2 \\ 1 & 0 \\ 1 & 0 \\ 1 & 2 \end{bmatrix}' \begin{bmatrix} 3 \\ 0 \\ 2 \\ -1 \end{bmatrix} \tag{2.98}$$

$$\begin{bmatrix} 1 & 1 & 1 & 1 \\ -2 & 0 & 0 & 2 \end{bmatrix} \begin{bmatrix} 1 & -2 \\ 1 & 0 \\ 1 & 0 \\ 1 & 2 \end{bmatrix} = \begin{bmatrix} 4 & 0 \\ 0 & 8 \end{bmatrix} \tag{2.99}$$

$$\Rightarrow \qquad (\mathbf{X}'\mathbf{X})^{-1} = \frac{1}{32}\begin{bmatrix} 8 & 0 \\ 0 & 4 \end{bmatrix} \tag{2.100}$$

$$\mathbf{X}'\mathbf{Y} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ -2 & 0 & 0 & 2 \end{bmatrix} \begin{bmatrix} 3 \\ 0 \\ 2 \\ -1 \end{bmatrix} \tag{2.101}$$

$$= \begin{bmatrix} 4 \\ -8 \end{bmatrix} \tag{2.102}$$

Thus, we have

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \tag{2.103}$$

$$= \frac{1}{32}\begin{bmatrix} 8 & 0 \\ 0 & 4 \end{bmatrix} \begin{bmatrix} 4 \\ -8 \end{bmatrix} \tag{2.104}$$

$$\mathbf{b} = \frac{1}{32}\begin{bmatrix} 32 \\ -32 \end{bmatrix} \tag{2.105}$$

And finally,

$$\mathbf{b} := \begin{bmatrix} b_0 \\ b_1 \end{bmatrix} = \begin{bmatrix} 1 \\ -1 \end{bmatrix} \tag{2.106}$$

Thus, we have $b_0 = 1$ and $b_1 = -1$. ♦

The conclusion is exactly the same, $\hat{y} = 1 - x$. The process is different. Here, this process is much easier for computers to perform, as they can do matrix multiplication (and inverting) with little problem. We have to spend a lot of extra effort to perform those operations.

Also, if we have more than one independent variable, we need to calculate the OLS estimator equations again; the ones in Equation (2.19) only hold for one independent variable. Using matrices, however, formula (2.68) holds for any number of independent variables.

2.3: Predictions and the Hat Matrix

Beyond modeling the relationship, one may also want to predict values of $\mathbf{Y}$ for a given value of $\mathbf{X}$. In matrix terms, this requires solving the equation $\hat{\mathbf{Y}} = \mathbf{Xb}$. But note the following:

$$\hat{\mathbf{Y}} = \mathbf{Xb} \tag{2.107}$$

$$= \mathbf{X}\ (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \tag{2.108}$$

Thus:

$$\hat{\mathbf{Y}} = \left(\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right)\mathbf{Y} \tag{2.109}$$

Note that the matrix $\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ "puts a hat" on the $\mathbf{Y}$ matrix. As such, it is called the "hat matrix," $\mathbf{H}$. Thus, we have simple matrix equations for the predictors and the residuals:

hat matrix

$$\hat{\mathbf{Y}} = \mathbf{HY} \tag{2.110}$$

$$\mathbf{E} = \mathbf{Y} - \hat{\mathbf{Y}} = (\mathbf{I} - \mathbf{H})\mathbf{Y} \tag{2.111}$$

Why is this important? It shows that the predictions and the residuals are orthogonal (see definition on page 435).

Theorem 2.12. *The matrices* $\mathbf{H}$ *and* $\mathbf{I}-\mathbf{H}$ *are orthogonal.*

Proof. To show orthogonality, we need to show that the inner product is zero:

$$\mathbf{H}'(\mathbf{I}-\mathbf{H}) = \mathbf{H}(\mathbf{I}-\mathbf{H}) \qquad = \mathbf{H}-\mathbf{H}\mathbf{H} \tag{2.112}$$
$$= \mathbf{H}-\mathbf{H} \tag{2.113}$$
$$= \mathbf{0} \tag{2.114}$$

□

In the proof, we used the fact that the hat matrix is symmetric idempotent. The next theorem shows this to be the case.

idempotent

Theorem 2.13. *The matrix* $\mathbf{H}$ *is symmetric idempotent.*

Proof. Let us start with showing $\mathbf{H}$ is symmetric.

$$\mathbf{H}' = \left(\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right)' \tag{2.115}$$

Recall from page 436 in Appendix M that $(\mathbf{AB})' = \mathbf{B}'\mathbf{A}'$. Thus

$$\left(\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right)' = \mathbf{X}''\left((\mathbf{X}'\mathbf{X})^{-1}\right)'\mathbf{X}' \tag{2.116}$$
$$= \mathbf{X}\left((\mathbf{X}'\mathbf{X})^{-1}\right)'\mathbf{X}' \tag{2.117}$$

I leave it as an exercise to show that $\mathbf{X}'\mathbf{X}$ is symmetric, and so is its inverse.

$$= \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \tag{2.118}$$
$$= \mathbf{H} \tag{2.119}$$

Now, let us show that $\mathbf{H}$ is idempotent.

$$\mathbf{H}\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\,\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \tag{2.120}$$
$$= \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \tag{2.121}$$
$$= \mathbf{H} \tag{2.122}$$

□

Since $\mathbf{H}$ is symmetric and idempotent, it is an orthogonal projection matrix that projects $\mathbf{Y}$-space onto the smaller $\hat{\mathbf{Y}}$-space (Appendix M.4). Because it is an orthogonal projection, $\hat{\mathbf{Y}}$ is as close to $\mathbf{Y}$ as possible in its subspace. That is, the errors are minimized. Figure 2.3 illustrates this.

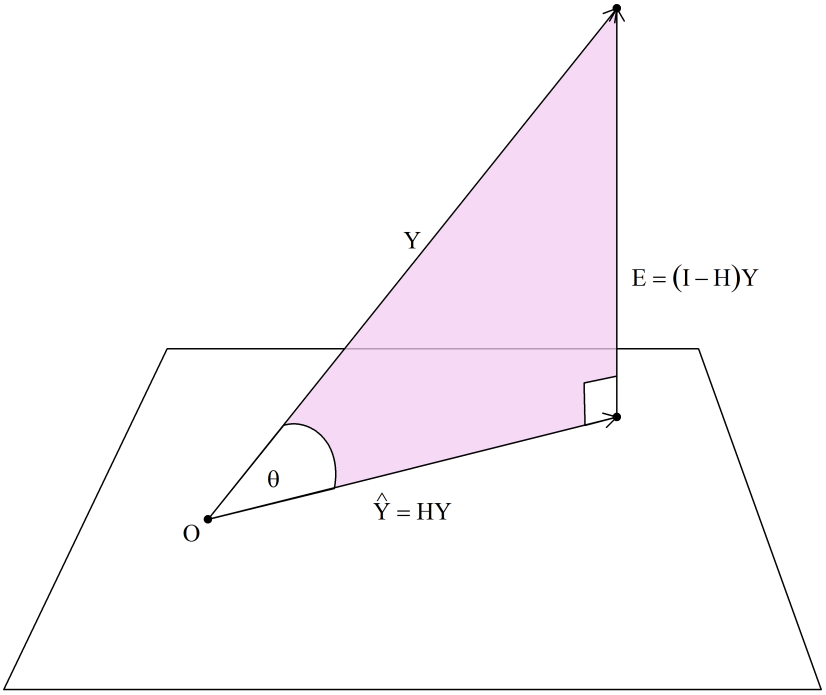


Figure 2.3: *A schematic illustrating that* $\hat{\mathbf{Y}}$ *is as close to* $\mathbf{Y}$ *as possible, while remaining in its subspace (represented by the plane). In other words, the* $\mathbf{Y}$ *matrix exists in an n-dimensional space. The solution,* $\hat{\mathbf{Y}}$*, is in a p-dimensional space, with* $n > p$*. Under the assumptions of ordinary least squares, the distance between* $\mathbf{Y}$ *and* $\hat{\mathbf{Y}}$ *(represented as the residuals,* $\mathbf{E}$*) is as small as possible if you define "distance" in terms of the Euclidean distance,* L_2*.*

Theorem 2.14. *The vectors* $\hat{\mathbf{Y}}$ *and* $\mathbf{E}$ *are orthogonal.*

cise

Proof. I leave this as an exercise. □

Since the predictions and residuals are orthogonal, we know the following is true by the Pythagorean Theorem:

$$\mathbf{Y}'\mathbf{Y} = \hat{\mathbf{Y}}'\hat{\mathbf{Y}} + \mathbf{E}'\mathbf{E} \tag{2.123}$$

Let us also prove this using matrices.

Theorem 2.15. $\mathbf{Y}'\mathbf{Y} = \hat{\mathbf{Y}}'\hat{\mathbf{Y}} + \mathbf{E}'\mathbf{E}$

Proof. Let us prove this without resorting to the Pythagorean Theorem. We know $\mathbf{Y} = \hat{\mathbf{Y}} + \mathbf{E}$. Thus,

$$\mathbf{Y}'\mathbf{Y} = \left(\hat{\mathbf{Y}} + \mathbf{E}\right)'\left(\hat{\mathbf{Y}} + \mathbf{E}\right) \tag{2.124}$$

$$= \hat{\mathbf{Y}}'\hat{\mathbf{Y}} + \mathbf{E}'\mathbf{E} + \hat{\mathbf{Y}}'\mathbf{E} + \mathbf{E}'\hat{\mathbf{Y}} \tag{2.125}$$

$$= \hat{\mathbf{Y}}'\hat{\mathbf{Y}} + \mathbf{E}'\mathbf{E} + (\mathbf{H}\mathbf{Y})'(\mathbf{I} - \mathbf{H})\mathbf{Y} + ((\mathbf{I} - \mathbf{H})\mathbf{Y})'\mathbf{H}\mathbf{Y} \tag{2.126}$$

$$= \hat{\mathbf{Y}}'\hat{\mathbf{Y}} + \mathbf{E}'\mathbf{E} + \mathbf{Y}'\mathbf{H}'(\mathbf{I} - \mathbf{H})\mathbf{Y} + \mathbf{Y}'(\mathbf{I} - \mathbf{H})'\mathbf{H}\mathbf{Y} \tag{2.127}$$

Remember that $\mathbf{H}$ and $\mathbf{I} - \mathbf{H}$ are symmetric. That gives us

$$= \hat{\mathbf{Y}}'\hat{\mathbf{Y}} + \mathbf{E}'\mathbf{E} + \mathbf{Y}'\mathbf{H}(\mathbf{I} - \mathbf{H})\mathbf{Y} + \mathbf{Y}'(\mathbf{I} - \mathbf{H})\mathbf{H}\mathbf{Y} \tag{2.128}$$

Finally, since $\mathbf{H}(\mathbf{I} - \mathbf{H}) = (\mathbf{I} - \mathbf{H})\mathbf{H} = \mathbf{0}$, we have

$$\mathbf{Y}'\mathbf{Y} = \hat{\mathbf{Y}}'\hat{\mathbf{Y}} + \mathbf{E}'\mathbf{E} + \mathbf{Y}'\mathbf{0}\mathbf{Y} + \mathbf{Y}'\mathbf{0}\mathbf{Y} \tag{2.129}$$

and

$$\mathbf{Y}'\mathbf{Y} = \hat{\mathbf{Y}}'\hat{\mathbf{Y}} + \mathbf{E}'\mathbf{E} \tag{2.130}$$

□

This will come in handy when we add probability distributions to our mathematics, thus creating statistics.

By the way, we also can show that the residuals and predicted values are uncorrelated by showing their covariance is zero.

Theorem 2.16. $\mathbb{C}ov\left[\hat{\mathbf{Y}}, \mathbf{E}\right] = 0.$

Proof.

$$\mathbb{C}ov\left[\hat{\mathbf{Y}}, \mathbf{E}\right] = \mathbb{C}ov\left[\mathbf{H}\mathbf{Y}, (\mathbf{I} - \mathbf{H})\mathbf{Y}\right] \tag{2.131}$$

$$= \mathbf{H}\,\mathbb{C}ov\left[\mathbf{Y}, \mathbf{Y}\right](\mathbf{I} - \mathbf{H})' \tag{2.132}$$

$$= \mathbf{H}\sigma^2(\mathbf{I} - \mathbf{H}) \tag{2.133}$$

$$= \sigma^2\mathbf{H}(\mathbf{I} - \mathbf{H}) \tag{2.134}$$

$$= \mathbf{0} \tag{2.135}$$

This result is not surprising given that the prediction and residual vectors are orthogonal. □

2.3.1 Consequences

In this section, we started with the matrix equation $\mathbf{Y} = \mathbf{XB} + \mathbf{E}$ and obtained the OLS estimator of $\mathbf{B}$. With that solution (and the requirement that $\mathbf{X}$ be full column rank), we have another result.

Theorem 2.17. $\mathbf{X}'\mathbf{E} = \mathbf{0}$

Proof.

$$\mathbf{Y} = \mathbf{XB} + \mathbf{E} \tag{2.136}$$

$$\mathbf{X}'\mathbf{Y} = \mathbf{X}'\mathbf{Xb} + \mathbf{X}'\mathbf{E} \tag{2.137}$$

$$\mathbf{X}'\mathbf{Y} = \mathbf{X}'\mathbf{X}\left((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}\right) + \mathbf{X}'\mathbf{E} \tag{2.138}$$

$$\mathbf{X}'\mathbf{Y} = \mathbf{X}'\mathbf{Y} + \mathbf{X}'\mathbf{E} \tag{2.139}$$

$$\mathbf{0} = \mathbf{X}'\mathbf{E} \tag{2.140}$$

What does this result mean? Recall that $\mathbf{X}'\mathbf{E}$ is a $p \times 1$ matrix. The first column of $\mathbf{X}$ is a column of 1s. Thus, the first element of $\mathbf{X}'\mathbf{E}$ is just the sum of the residuals. That means the residuals must sum to 0 when we use the OLS estimator.

The other elements in the $\mathbf{X}'\mathbf{E}$ matrix consist of the sum of the residuals times the values of each independent variable. This means that, under OLS, the residuals are necessarily linearly independent of each of the independent variables. It is a result of the mathematics used. **mathematics**

To see this in simple linear regression:

$$\mathbf{X}'\mathbf{E} = \begin{bmatrix} 1 & 1 & 1 & \cdots & 1 \\ x_1 & x_2 & x_3 & \cdots & x_n \end{bmatrix} \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ \vdots \\ e_n \end{bmatrix} \tag{2.141}$$

$$= \begin{bmatrix} \sum e_i \\ \sum x_i e_i \end{bmatrix} \tag{2.142}$$

This matrix is $\mathbf{0}$ only when all of its elements are also 0. Thus, we have $\sum e_i = 0$; the sum of the residuals in OLS is *mathematically* guaranteed to be zero.

We also have $\sum x_i e_i = 0$, which is equivalent to $\sum x_i e_i - n\overline{x}\,\overline{e}$ because $\overline{e} = 0$ and thus to $(n-1)\mathbb{C}ov[x, e]$. This covariance is zero if x and e are linearly independent. This means that the residuals arising from OLS estimation are *linearly* uncorrelated with the predictor variables. □

Note: Again, these are mathematical results from applying ordinary least squares. They are guaranteed simply because of the estimation method we selected. Had we chosen a different definition of "best fit," then this section may not hold.

Everything follows from our chosen definition of "best fit."

2.4: The PRE Measures

Now that we have reality (**Y**) and our errors (**E**), as pictured in Figure 2.3, we can create a measure of how well the model summarizes (fits) the data. In fact, we will create two of them! Both are "**p**roportional **r**eduction in **e**rror" measures; that is, they both are measures of how well the model reduces the unexplained variation in the dependent variable. The first is the venerable R^2 ("R-squared") measure. The second is the $\overline{R}^2$ ("adjusted R-squared") measure.

PRE

Both measure how much the model reduces the variation in the dependent variable. They differ in how that variation is measured. The R^2 measure uses the sum-of-squares; the $\overline{R}^2$, the variance.

2.4.1 R^2 Measure

The formula for the R^2 measure is

$$R^2 = 1 - \frac{SSE}{TSS} \tag{2.143}$$

Here, SSE is the sum of the squared errors using the model, and TSS is the sum of squares without using the model.

$$SSE := \sum (y_i - \hat{y})^2 \tag{2.144}$$

$$TSS := \sum (y_i - \overline{y})^2 \tag{2.145}$$

In this formula, $\hat{y}$ is the predicted value of Y for each value of x_i in the data according to the model; $\overline{y}$ is the predicted value of Y for each value of x_i in the data in the absence of the model.

Thus, the SSE is a measure of how much variation remains in the model — the residual (unexplained) variation after applying the model. The

TSS is a measure of the variation in the original data. It is called the residual variation after applying the "null model."[2]

> ***Note***: The R^2 measure only tells us how much of the variation in the dependent variable is described by the model (as compared to how much was there originally). It tells us nothing beyond that.
>
> For instance, an R^2 value of 0.04 tells us that the model explains only 4% of the variation in the dependent variable. There is a lot of variation left unexplained by the model. It does not mean that the model is "poor." An R^2 value of 0.98 tells us that the model explains 98% of the variation in the dependent variable, not that the model is "good."

Note that the formula for R^2 (2.143) is equivalent to

$$R^2 = 1 - \frac{\frac{1}{n-1}SSE}{\frac{1}{n-1}TSS} \tag{2.146}$$

Thus, the R^2 measure is how much the model reduces the unexplained *variance*, when variance is estimated using n in the denominator ($n-1$ degrees of freedom). This, we know, is a biased estimator of the population variance (the number of degrees of freedom is fewer than $n-1$).

2.4.2 $\overline{R}^2$ Measure

Where R^2 measures how much the model reduces the unexplained variance, when that variation is estimated using n in the denominator, $\overline{R}^2$ measures how much the model reduces the unexplained variance, when that variance is estimated using the appropriate degrees of freedom. In other words, the adjusted R^2 measure uses unbiased estimators of the variance to describe the proportional reduction in error.

$$\overline{R}^2 = 1 - \frac{\frac{1}{\nu_e}SSE}{\frac{1}{\nu_t}TSS} = 1 - \frac{\frac{1}{n-p}SSE}{\frac{1}{n-1}TSS} = 1 - \frac{(n-1)SSE}{(n-p)TSS} \tag{2.147}$$

[2]The "null model" always refers to the model with no independent variable. Thus, it is the model with only the y-intercept (here). The concept of the "null model" is extremely important in statistics, because it allows us to determine how much the model is a improvement over the "lack of" model.

Here, p is the number of parameters estimated in the model. For simple regression (one independent variable), $p = 2$, because we are calculating b_0 and b_1 from the data (to estimate β_0 and β_1). For the null model, $p = 1$, because we are only calculating $\overline{y}$ from the data (to estimate β_0).

Usually, one reports the R^2 value and uses the $\overline{R}^2$ to help with model selection (select the model with the larger $\overline{R}^2$). This, I believe, needs to change. It is the *adjusted* R-squared value that better estimates the proportion reduction in error. This is because the adjusted R-squared measure uses unbiased estimators of the variances.

fight the power!!!

One strength of the R^2 measure is that its range is from 0 to 1. The $\overline{R}^2$ measure can be less than 1. However, it is only less than 1 when the model describes very little of the variance.

2.4.3 Other PREs

These are the two most frequently used PRE measures. They are not the only ones, however. There is an entire class of PRE measures called "pseudo-R^2" measures. These are all genuine measures of how well the model helps reduce unexplained variation in the dependent variable. Their formulas tend to follow the structure of

$$\text{PRE} = 1 - \frac{\text{variation in the dependent variable with the model}}{\text{variation in the dependent variable without the model}}$$

In fact, some would use Figure 2.3 to create another PRE, this one based on the right triangle. Note that $\mathbf{Y}$ is the target we are trying to describe and $\hat{\mathbf{Y}}$ is where we landed. Thus, a PRE measure could be the angle $\theta = \angle\mathbf{YO\hat{Y}}$. This θ ranges between 0 and 90°, with 0 being an optimal fit and 90° being the worst fit.

PRE

As this measure is not intuitive as a measure of fit (larger is not better), we simply take its cosine and use $\cos\theta$. This value ranges between 0 and 1, with a 1 being the best fit ($\cos 0$) and a 0 being the worst ($\cos \pi/2$).

> ***Note***: In the future, you will be introduced to "pseudo-R^2 measures" for several different modeling schemes. Because these follow the same scheme as above, they have the same interpretation. They are measures of how well the model reduces the uncertainty in the dependent variable.

2.5: Multicollinearity and Categorical Independent Variables

So far, our independent variable was either numeric (Example 2.1) or dichotomous (Example 2.2). Let us now look at the interesting case of a discrete independent variable with three levels.

Example 2.5: His Majesty Rudolph II would like some input on his next five-year plan. The primary crop in Ruritania is corn. To help optimize the profits made by farmers, Rudolph wants to know if that crop should be changed to summer wheat or to soybeans.

To help him, let us model the relationship between farmer profit and crop in Ruritania. •

Solution: Collecting the data is not as difficult as it may seem at first. All three crops are currently grown in Ruritania. All we had to do was obtain

Crop	Profit per Acre
Wheat	722
Wheat	965
Wheat	940
Wheat	756
Corn	763
Corn	765
Corn	565
Corn	621
Soybean	566
Soybean	658
Soybean	540
Soybean	485

Table 2.3: *The data collected from Ruritania for Example 2.5. Note that this is the raw data with a categorical independent variable.*

a list of all farms and their primary crop and randomly select records from that. Table 2.3 provides our data.

Note that the response variable is numeric, and the predictor variable is categorical. How do we code that variable so that we can use the methods of this chapter (and this class)???

base level

In Example 2.2, it was easy to change our dichotomous variable into a numeric variable by selecting one level as the base level and measuring the other level from there. In other words, one level was given the value 0 (absence) and the other was given the value 1 (presence).

In this case, we have *three* levels in our independent variable. It does not seem to make sense to select one level to represent with 0 (absence), one to represent with 1 (presence), and one to represent with 2 (huh????).

One method that always works is to create a series of dichotomous indicator variables from the one nominal variable. Thus, since there are three levels here, we would create three new dichotomous variables: corn, soybeans, and wheat.

This change is presented in Table 2.4. Note that each of the three dichotomous variables is now numeric. Each value indicates absence (0) or presence (1) of that trait (crop). With this change, we can use the methods of this chapter to calculate the values of the OLS estimators β_0, β_1, β_2, and β_3... or can we?

To see why I ended that paragraph in an evil and foreboding voice, let us work through this using matrices.

Remember that the formula to calculate the OLS estimators is $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$. Here, $\mathbf{Y}$ is

$$\mathbf{Y} = \begin{bmatrix} 722 \\ 965 \\ 940 \\ 756 \\ 763 \\ 765 \\ 565 \\ 621 \\ 566 \\ 658 \\ 540 \\ 485 \end{bmatrix} \tag{2.148}$$

Wheat	Corn	Soybeans	Profit per Acre
1	0	0	722
1	0	0	965
1	0	0	940
1	0	0	756
0	1	0	763
0	1	0	765
0	1	0	565
0	1	0	621
0	0	1	566
0	0	1	658
0	0	1	540
0	0	1	485

Table 2.4: *Data to be used for Example 2.5. This table differs from Table 2.3 by taking the original* `Crop` *variable and replacing it with three indicator variables. This form allows us to more easily calculate the ordinary least squares estimators by hand.*

The design matrix, $\mathbf{X}$ is

$$\mathbf{X} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \end{bmatrix} \tag{2.149}$$

So far, so good!

At this point, can you see why this matrix is termed the "design" matrix? From it, one can deduce the experimental design that gave rise to the data.

design matrix

Next, let us calculate $\mathbf{X}'\mathbf{X}$:

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \end{bmatrix}' \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \end{bmatrix} \tag{2.150}$$

$$= \begin{bmatrix} 12 & 4 & 4 & 4 \\ 4 & 4 & 0 & 0 \\ 4 & 0 & 4 & 0 \\ 4 & 0 & 0 & 4 \end{bmatrix} \tag{2.151}$$

Fantastický!

Nice! That is a rather interesting matrix. From it, you can pick out the sample size ($n = 12$) and the sample sizes in each of the three levels ($n_i = 4$ in the diagonals). The next step is to calculate the inverse of this matrix.

At this point, it is *sooooooooo* much easier to use technology to perform this calculation. However, when you do, you will get a notification that the matrix is singular. This means two things. First: It cannot be inverted; its inverse does not exist. Second: One column is a linear combination of the others.

singular

Note that the first column is the sum of the other three columns. The columns are not linearly independent.

information

Note: From an information standpoint, if one column is a linear combination of the others, then that column is redundant. The model can be repeated without that information.

This is one of the very few places in statistics where throwing away information helps. It is rather ironic that it helps solely in terms of the mathematics.

So, what do we do? We drop one of the redundant columns. The one we drop determines how we interpret the results. **interpretation**

❧ ❧ ❧

Dropping the first column is appropriate. It leads to this design matrix:

$$\mathbf{X} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix} \tag{2.152}$$

This leads to

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} 4 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 4 \end{bmatrix} \tag{2.153}$$

$$\mathbf{X}'\mathbf{Y} = \begin{bmatrix} 3383 \\ 2714 \\ 2249 \end{bmatrix} \tag{2.154}$$

Finally, this leads to

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \tag{2.155}$$

$$= \begin{bmatrix} 845.75 \\ 678.50 \\ 562.25 \end{bmatrix} \tag{2.156}$$

Thus, from this decision, we have that the average profit for wheat is 845.75; for corn, 678.50; and for soybeans, 562.25.

This is called the "means model" because the returned values are the means in each group. **means model**

Dropping the second column is also appropriate (the second column corresponds to the wheat design). When doing so, the design matrix is

$$\mathbf{X} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{bmatrix} \tag{2.157}$$

Feel free to work through the calculation to obtain these estimates:

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \tag{2.158}$$

$$= \begin{bmatrix} 845.75 \\ -167.25 \\ -283.50 \end{bmatrix} \tag{2.159}$$

The interpretation here is that the average profit for wheat (the base category/dropped column) is 845.75. The effect of corn over wheat is −167.25, and the effect of soybeans over wheat is −283.50. In other words, the expected corn profit is −167.25 over the wheat profit, and the expected soybean profit is −283.50 over the wheat profit.

Note that we dropped the first data column. Thus, the first result is the expected value of the first variable and the other results are the effects of those levels *as compared to* the base category (wheat).

effects model

Because the estimate are the effects of the other levels as compared to the selected base level, this is called the "effects model."

Dropping the third column is appropriate, as well. When doing so, the design matrix is

$$\mathbf{X} = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{bmatrix} \tag{2.160}$$

Feel free to work through the calculation to obtain these estimates:

$$\mathbf{b} = \begin{bmatrix} 678.50 \\ 167.25 \\ -116.25 \end{bmatrix} \tag{2.161}$$

This interpretation is similar to the previous. The mean of the base category (corn) is 678.50 (first number). The effect of wheat over corn is 167.25. The effect of soybean over corn is -116.25.

Corn is the base category because the third column corresponds to the corn design. Note that this is *also* an effects model. The estimates are the effects in relation t the base category. ♦

If you look at these three sets of results, you will see a lot of commonalities. The one chosen depends on what you are trying to say about the relationship between the crop and the profit. Here, is it also very easy to move between the means model and the effects model.

Note, however, that we are only investigating expected values (averages) in this analysis. Should we also decide to include the uncertainties in our estimates (as we should), the two models are complementary. It is very difficult to move between the standard error in the means model and the standard error in the effects model. It is so much easier to have the computer perform that computation for you.

For the record, here is the code I used for fitting the means model:

```
X = matrix( c(1,0,0, 1,0,0, 1,0,0, 1,0,0,
    0,1,0, 0,1,0, 0,1,0, 0,1,0,
    0,0,1, 0,0,1, 0,0,1, 0,0,1 ),
    ncol=3, byrow=TRUE)
Y = matrix( c(722, 965, 940, 756, 763, 765,
    565, 621, 566, 658, 540, 485) )

solve(t(X)%*%X)
t(X)%*%Y
solve(t(X)%*%X) %*% t(X)%*%Y
```

Here is the code I used for the first effects model:

```
X = matrix( c(1,0,0, 1,0,0, 1,0,0, 1,0,0,
    1,1,0, 1,1,0, 1,1,0, 1,1,0,
    1,0,1, 1,0,1, 1,0,1, 1,0,1),
    ncol=3, byrow=TRUE)
Y = matrix( c(722, 965, 940, 756, 763, 765,
    565, 621, 566, 658, 540, 485) )

solve(t(X)%*%X)
t(X)%*%Y
solve(t(X)%*%X) %*% t(X)%*%Y
```

Note that the only change is in the line that defines the data matrix, **X**.

Finally, here is the code I used when dropping the third column.

```
X = matrix( c(1,1,0, 1,1,0, 1,1,0, 1,1,0,
    1,0,0, 1,0,0, 1,0,0, 1,0,0,
    1,0,1, 1,0,1, 1,0,1, 1,0,1),
    ncol=3, byrow=TRUE)
Y = matrix( c(722, 965, 940, 756, 763, 765,
    565, 621, 566, 658, 540, 485) )

solve(t(X)%*%X)
t(X)%*%Y
solve(t(X)%*%X) %*% t(X)%*%Y
```

Again, the only change is in the line that defines the data matrix, **X**.

2.6: Conclusion

This chapter started with defining what we can mean by "best." Because we decided to define "best" as "minimizing the sum of squared residuals," we were able to obtain closed-form solutions for the estimates.

From those equations, we were able to learn even more about our estimators — things not obvious from the definition. For instance, the estimators are only independent if $\overline{x} = 0$.

That was the entire purpose of this chapter: to see that our results arise from applying mathematics to our selected definition of "best." Had we chosen a different meaning, we may have arrived at different results.

In the next chapter, we will see what we can learn by taking the next step and applying statistics to the models. While the mathematics tells us the expected value... it is statistics that gives us an insight into the population based on our little sample.

2.7: End-of-Chapter Materials

Here are the expected materials to supplement the chapter. Since there is R code in this chapter, I am including an explanation of several helpful R functions.

2.7.1 R Functions

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

Mathematics:

%*% This multiplies two matrices in R. Thus, running the command `A%*%B` will return the matrix product **AB** (Section M.3.2). Be careful: `A*B` returns the Hadamard product (Section M.3), which is rarely what is needed.

c() This **c**ombines the several scalar values into a single vector of values.

matrix() This function creates a matrix from the given vector. The first slot belongs to the values in the matrix. After that is the number of rows (or columns) and whether you are entering the number by rows or by columns.

solve(m) This calculates the usual inverse of the provided matrix **m** (page 428).

t(m) This calculates the transpose of the provided matrix **m** (Section M.4).

2.7.2 Exercises

I left many things as exercises for you. Here they are. You should be able to prove any and all of them using your prior knowledge of mathematics (matrices and calculus).

1. Perform the second derivative test on b_0 and b_1 to show that these estimators are really minima.
2. Show that $\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})$ is equivalent to $\sum_{i=1}^{n} x_i y_i - n\overline{x}\,\overline{y}$ and that $\sum_{i=1}^{n}(x_i - \overline{x})^2$ is equivalent to $\sum_{i=1}^{n} x_i^2 - n\overline{x}^2$.
3. Prove $b_1 = r_{xy}\frac{s_y}{s_x}$.
4. Prove that the slope estimator b_1 is zero if the y-values do not vary.
5. Using the scalar form, show that $\mathbb{C}ov[b_0, b_1] = -\sigma^2\frac{\overline{x}}{S_{xx}}$.
6. Prove that $\mathbb{V}[\mathbf{Y} \mid \mathbf{XB}] = \sigma^2\mathbf{I}$.
7. Let $\mathbf{A}$ be any full column rank matrix. Prove that $\mathbf{A}'\mathbf{A}$ is symmetric. Prove that its inverse is symmetric.
8. Prove that the vectors $\hat{\mathbf{Y}}$ and $\mathbf{E}$ are orthogonal.