13.3: Fitting a Line by Least Squares Regression

Fitting linear models by eye is open to criticism since it is based on an individual preference. In this section, we use least squares regression as a more rigorous approach.

This section considers family income and gift aid data from a random sample of fifty students in the 2011 freshman class of Elmhurst College in Illinois. Gift aid is financial aid that is a gift, as opposed to a loan. A scatterplot of the data is shown in Figure 13.3.1 along with two linear fits. The lines follow a negative trend in the data; students who have higher family incomes tended to have lower gift aid from the university.

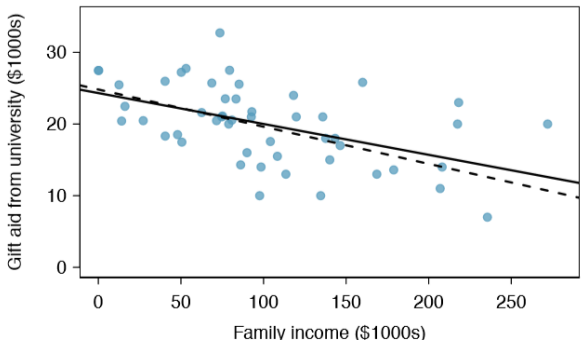

Figure 13.3.1: Gift aid and family income for a random sample of 50 freshman students from Elmhufirst College. Two lines are fit to the data, the solid line being the least squares line. These data were sampled from a table of data for all freshman from the 2011 class at Elmhurst College that accompanied an article titled What Students Really Pay to Go to College published online by The Chronicle of Higher Education: chronicle.com/article/What-Students-Really-Pay-to-Go/131435

Exercise 13.3.1

Is the correlation positive or negative in Figure 13.3.1?[6]

Solution

[6]*Larger family incomes are associated with lower amounts of aid, so the correlation will be negative. Using a computer, the correlation can be computed: -0.499.*

An Objective Measure for Finding the Best Line

We begin by thinking about what we mean by "best". Mathematically, we want a line that has small residuals. Perhaps our criterion could minimize the sum of the residual magnitudes:

$$|e_1| + |e_2| + \cdots + |e_n| \tag{13.3.1}$$

which we could accomplish with a computer program. The resulting dashed line shown in Figure 13.3.1 demonstrates this fit can be quite reasonable. However, a more common practice is to choose the line that minimizes the sum of the *squared* residuals:

$$e_1^2 + e_2^2 + \cdots + e_n^2 \tag{13.3.2}$$

The line that minimizes this *least squares criterion* is represented as the solid line in Figure 13.3.1. This is commonly called the *least squares line*. The following are three possible reasons to choose Criterion 13.3.2 over Criterion 13.3.1:

1. It is the most commonly used method.
2. Computing the line based on Criterion 13.3.2 is much easier by hand and in most statistical software.
3. In many applications, a residual twice as large as another residual is more than twice as bad. For example, being off by 4 is usually more than twice as bad as being off by squaring the residuals accounts for this discrepancy.

The first two reasons are largely for tradition and convenience; the last reason explains why Criterion 13.3.2 is typically most helpful.

There are applications where Criterion 13.3.1 may be more useful, and there are plenty of other criteria we might consider. However, this book only applies the least squares criterion.

Conditions for the Least Squares Line

When fitting a least squares line, we generally require

- **Linearity**. The data should show a linear trend. If there is a nonlinear trend (e.g. left panel of Figure 13.3.2), an advanced regression method from another book or later course should be applied.
- **Nearly normal residuals**. Generally the residuals must be nearly normal. When this condition is found to be unreasonable, it is usually because of outliers or concerns about influential points, which we will discuss in greater depth in Section 7.3. An example of non-normal residuals is shown in the second panel of Figure 13.3.2
- **Constant variability**. The variability of points around the least squares line remains roughly constant. An example of non-constant variability is shown in the third panel of Figure 13.3.2

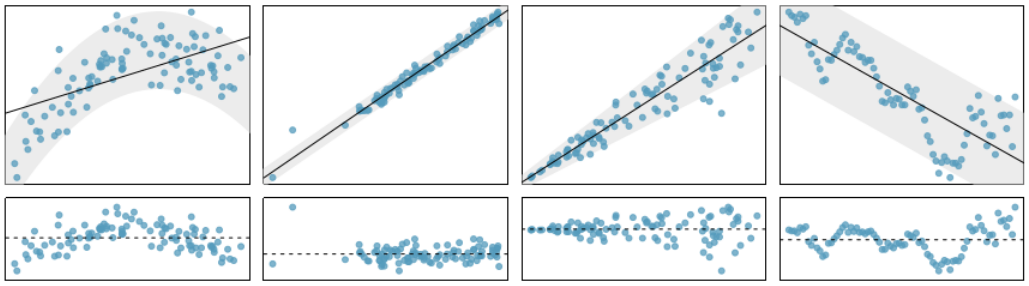

Figure 13.3.2: Four examples showing when the methods in this chapter are insufficient to apply to the data. In the left panel, a straight line does not t the data. In the second panel, there are outliers; two points on the left are relatively distant from the rest of the data, and one of these points is very far away from the line. In the third panel, the variability of the data around the line increases with larger values of x. In the last panel, a time series data set is shown, where successive observations are highly correlated.

Be cautious about applying regression to data collected sequentially in what is called a time series. Such data may have an underlying structure that should be considered in a model and analysis. There are other instances where correlations within the data are important. This topic will be further discussed in Chapter 8.

Exercise 13.3.2

Should we have concerns about applying least squares regression to the Elmhurst data in Figure 13.3.1?

Solution

The trend appears to be linear, the data fall around the line with no obvious outliers, the variance is roughly constant. These are also not time series observations. Least squares regression can be applied to these data.

Finding the Least Squares Line

For the Elmhurst data, we could write the equation of the least squares regression line as

$$\hat{aid} = \beta_0 + \beta_1 \times \text{family income} \tag{13.3.3}$$

Here the equation is set up to predict gift aid based on a student's family income, which would be useful to students considering Elmhurst. These two values, β_0 and β_1, are the parameters of the regression line.

As in Chapters 4-6, the parameters are estimated using observed data. In practice, this estimation is done using a computer in the same way that other estimates, like a sample mean, can be estimated using a computer or calculator. However, we can also find the parameter estimates by applying two properties of the least squares line:

• The slope of the least squares line can be estimated by

$$b_1 = \frac{s_y}{s_x} R \tag{13.3.4}$$

where R is the correlation between the two variables, and s_x and s_y are the sample standard deviations of the explanatory variable and response, respectively.

• If $\bar{x}$ is the mean of the horizontal variable (from the data) and $\bar{y}$ is the mean of the vertical variable, then the point $(\bar{x}, \bar{y})$ is on the least squares line.

We use b_0 and b_1 to represent the point estimates of the parameters β_0 and β_1.

Exercise 13.3.3

Table 7.14 shows the sample means for the family income and gift aid as \$101,800 and \$19,940, respectively. Plot the point (101.8, 19.94) on Figure 13.3.1 on page 324 to verify it falls on the least squares line (the solid line).[9]

Table 7.14: Summary statistics for family income and gift aid.

	family income, in \$1000s ("x")	gift aid, in \$1000s ("y")
mean sd	$\bar{x}$ = 101.8 s_x = 63.2	$\bar{y}$ = 19.94 s_y = 5.46
		R = -0.499

[9] *If you need help finding this location, draw a straight line up from the x-value of 100 (or thereabout). Then draw a horizontal line at 20 (or thereabout). These lines should intersect on the least squares line.*

Exercise 13.3.4

sing the summary statistics in Table 7.14, compute the slope for the regression line of gift aid against family income.

Hint:

Apply Equation 13.3.4 with the summary statistics from Table 7.14 to compute the slope:

$$b_1 = \frac{s_y}{s_x} R = \frac{5.46}{63.2}(-0.499) = -0.0431 \tag{13.3.5}$$

You might recall the point-slope form of a line from math class (another common form is slope-intercept). Given the slope of a line and a point on the line, (x_0, y_0), the equation for the line can be written as

$$y - y_0 = \text{slope} \times (x - x_0) \tag{13.3.6}$$

A common exercise to become more familiar with foundations of least squares regression is to use basic summary statistics and point-slope form to produce the least squares line.

TIP: Identifying the least squares line from summary statistics

To identify the least squares line from summary statistics:

- Estimate the slope parameter, b_1, using Equation 13.3.4.
- Noting that the point $(\bar{x}, \bar{y})$ is on the least squares line, use $x_0 = \bar{x}$ and $y_0 = \bar{y}$ along with the slope b_1 in the point-slope equation:

$$y - \bar{y} = b_1(x - \bar{x}) \tag{13.3.7}$$

- Simplify the equation.

Example 13.3.1

Using the point (101.8, 19.94) from the sample means and the slope estimate $b_1 = -0.0431$ from Exercise 7.14, and the least-squares line for predicting aid based on family income.

Solution

Apply the point-slope equation using (101.8, 19.94) and the slope $b_1 = -0.0431$:

$$y - y_0 = b_1(x - x_0) \tag{13.3.8}$$

$$y - 19.94 = -0.0431(x - 101.8) \tag{13.3.9}$$

Expanding the right side and then adding 19.94 to each side, the equation simplifies:

$$\hat{aid} = 24.3 - 0.0431 \times \text{ family income} \tag{13.3.10}$$

Here we have replaced y with $\hat{aid}$ and x with $family_{income}$ to put the equation in context.

We mentioned earlier that a computer is usually used to compute the least squares line. A summary table based on computer output is shown in Table 7.15 for the Elmhurst data. The first column of numbers provides estimates for b0 and b1, respectively. Compare these to the result from Example 7.16.

Table 7.15: Summary of least squares t for the Elmhurst data. Compare the parameter estimates in the rst column to the results of Example 7.16.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	24.3193	1.2915	18.83	0.0000
family_income	-0.0431	0.0108	-3.98	0.0002

Example 13.3.2

Examine the second, third, and fourth columns in Table 7.15. Can you guess what they represent?

Solution

We'll describe the meaning of the columns using the second row, which corresponds to β_1. The first column provides the point estimate for β_1, as we calculated in an earlier example: -0.0431. The second column is a standard error for this point estimate: 0.0108. The third column is a t test statistic for the null hypothesis that $\beta_1 = \beta_0 : T = -3.98$. The last column is the p-value for the t test statistic for the null hypothesis $\beta_1 = 0$ and a two-sided alternative hypothesis: 0.0002. We will get into more of these details in Section 7.4.

Example 13.3.3

Suppose a high school senior is considering Elmhurst College. Can she simply use the linear equation that we have estimated to calculate her nancial aid from the university?

Solution

She may use it as an estimate, though some qualifiers on this approach are important. First, the data all come from one freshman class, and the way aid is determined by the university may change from year to year. Second, the equation will provide an imperfect estimate. While the linear equation is good at capturing the trend in the data, no individual student's aid will be perfectly predicted.

Interpreting Regression Line Parameter Estimates

Interpreting parameters in a regression model is often one of the most important steps in the analysis.

Example 13.3.2

The slope and intercept estimates for the Elmhurst data are -0.0431 and 24.3. What do these numbers really mean?

Solution

Interpreting the slope parameter is helpful in almost any application. For each additional \$1,000 of family income, we would expect a student to receive a net difference of $\$1,000 \times (-0.0431) = -\$43.10$in aid on average, i.e. \$43.10 less. Note that a higher family income corresponds to less aid because the coefficient of family income is negative in the model. We must be cautious in this interpretation: while there is a real association, we cannot interpret a causal connection between the variables because these data are observational. That is, increasing a student's family income may not cause the student's aid to drop. (It would be reasonable to contact the college and ask if the relationship is causal, i.e. if Elmhurst College's aid decisions are partially based on students' family income.)

The estimated intercept $b_0 = 24.3$ (in \$1000s) describes the average aid if a student's family had no income. The meaning of the intercept is relevant to this application since the family income for some students at Elmhurst is \$0. In other applications, the intercept may have little or no practical value if there are no observations where x is near zero.

Interpreting parameters estimated by least squares

The slope describes the estimated difference in the y variable if the explanatory variable x for a case happened to be one unit larger. The intercept describes the average outcome of y if x = 0 and the linear model is valid all the way to x = 0, which in many applications is not the case.

Extrapolation is Treacherous

When those blizzards hit the East Coast this winter, it proved to my satisfaction that global warming was a fraud. That snow was freezing cold. But in an alarming trend, temperatures this spring have risen. Consider this: On February 6th it was 10 degrees. Today it hit almost 80. At this rate, by August it will be 220 degrees. So clearly folks the climate debate rages on.

(13.3.11)

[11] *http://www.colbertnation.com/the-col...videos/269929/*

Linear models can be used to approximate the relationship between two variables. However, these models have real limitations. Linear regression is simply a modeling framework. The truth is almost always much more complex than our simple line. For example, we do not know how the data outside of our limited window will behave.

Example 13.3.3

Use the model $\hat{aid} = 24.3 - 0.0431$ family income to estimate the aid of another freshman student whose family had income of \$1 million.

Recall that the units of family income are in \$1000s, so we want to calculate the aid for family income = 1000:

$$24.3 - 0.0431 \times \text{family income} = 24.3 - 0.0431 \times 1000 = -18.8 \qquad (13.3.12)$$

The model predicts this student will have -\$18,800 in aid (!). Elmhurst College cannot (or at least does not) require any students to pay extra on top of tuition to attend.

Applying a model estimate to values outside of the realm of the original data is called **extrapolation**. Generally, a linear model is only an approximation of the real relationship between two variables. If we extrapolate, we are making an unreliable bet that the approximate linear relationship will be valid in places where it has not been analyzed.

Using R^2 to describe the strength of a fit

We evaluated the strength of the linear relationship between two variables earlier using the correlation, R. However, it is more common to explain the strength of a linear t using R^2, called **R-squared**. If provided with a linear model, we might like to describe how closely the data cluster around the linear fit.

The R^2 of a linear model describes the amount of variation in the response that is explained by the least squares line. For example, consider the Elmhurst data, shown in Figure 7.16. The variance of the response variable, aid received, is $s^2_{aid} = 29.8$. However, if we apply our least squares line, then this model reduces our uncertainty in predicting

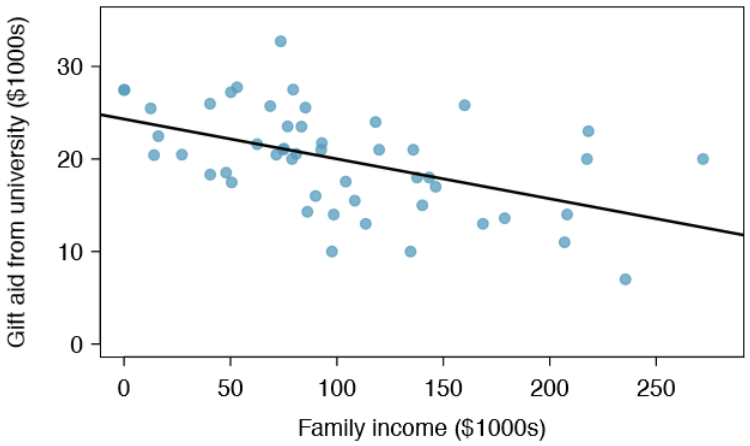

Figure 7.16: Gift aid and family income for a random sample of 50 freshman students from Elmhurst College, shown with the least squares regression line.

aid using a student's family income. The variability in the residuals describes how much variation remains after using the model: $s^2_{RES} = 22.4$. In short, there was a reduction of

$$\frac{s^2_{aid} - s^2_{RES}}{s^2_{GPA}} = \frac{29.9 - 22.4}{29.9} = \frac{7.5}{29.9} = 0.25 \tag{13.3.13}$$

or about 25% in the data's variation by using information about family income for predicting aid using a linear model. This corresponds exactly to the R-squared value:

$$R = -0.499 R^2 = 0.25 \tag{13.3.14}$$

Exercise 13.3.5

If a linear model has a very strong negative relationship with a correlation of -0.97, how much of the variation in the response is explained by the explanatory variable?[12]

Categorical Predictors with two Levels

Categorical variables are also useful in predicting outcomes. Here we consider a categorical predictor with two levels (recall that a level is the same as a category). We'll consider Ebay auctions for a video game, Mario Kart for the Nintendo Wii, where both the total price of the auction and the condition of the game were recorded.[13] Here we want to predict total price based on game condition, which takes values used and new. A plot of the auction data is shown in Figure 7.17.

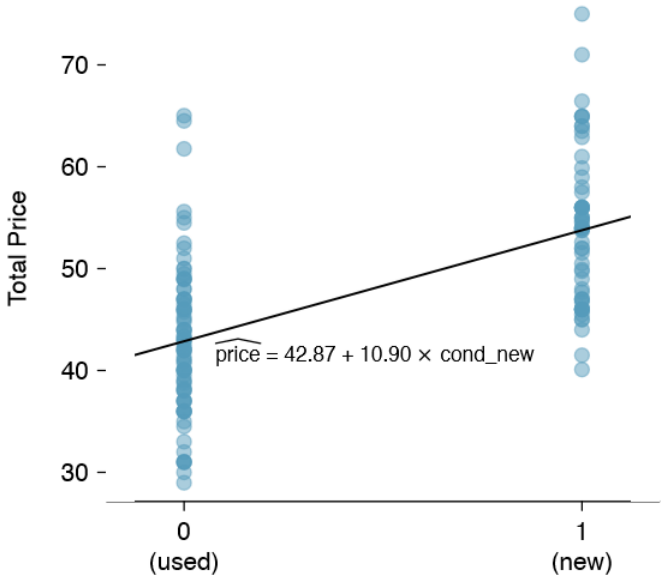

Figure 7.17: Total auction prices for the video game Mario Kart, divided into used (x = 0) and new (x = 1) condition games. The least squares regression line is also shown.

To incorporate the game condition variable into a regression equation, we must convert the categories into a numerical form. We will do so using an indicator variable called cond new, which takes value 1 when the game is new and 0 when the game is used. Using this indicator variable, the linear model may be written as

$$\hat{price} = \beta_0 + \beta_1 \times \text{cond new} \tag{13.3.15}$$

[12]*About* $R^2 = (-0.97)^2 = 0.94$ *or 94% of the variation is explained by the linear model.*

[13]*These data were collected in Fall 2009 and may be found at openintro.org.*

Table 7.18: Least squares regression summary for the nal auction price against the condition of the game.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	42.87	0.81	52.67	0.0000
cond_new	10.90	1.26	8.66	0.0000

The fitted model is summarized in Table 7.18, and the model with its parameter estimates is given as

$$\hat{price} = 42.87 + 10.90 \times \text{cond new} \tag{13.3.16}$$

For categorical predictors with just two levels, the linearity assumption will always be satis ed. However, we must evaluate whether the residuals in each group are approximately normal and have approximately equal variance. As can be seen in Figure 7.17, both of these conditions are reasonably satis ed by the auction data.

Example 7.22 Interpret the two parameters estimated in the model for the price of Mario Kart in eBay auctions.

The intercept is the estimated price when cond new takes value 0, i.e. when the game is in used condition. That is, the average selling price of a used version of the game is $42.87.

The slope indicates that, on average, new games sell for about $10.90 more than used games.

TIP: Interpreting model estimates for categorical predictors.

The estimated intercept is the value of the response variable for the first category (i.e. the category corresponding to an indicator value of 0). The estimated slope is the average change in the response variable between the two categories.

We'll elaborate further on this Ebay auction data in Chapter 8, where we examine the influence of many predictor variables simultaneously using multiple regression. In multiple regression, we will consider the association of auction price with regard to

each variable while controlling for the influence of other variables. This is especially important since some of the predictors are associated. For example, auctions with games in new condition also often came with more accessories.
