Section 8.2: Least Squares Regression

STAT 1201
Introduction to Probability and Statistics

8.2 Least squares regression

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

8.2.1 Gift aid for freshman at Elmhurst College

This section considers family income and gift aid data from a random sample of fifty students in the freshman class of Elmhurst College in Illinois. Gift aid is financial aid that does not need to be paid back, as opposed to a loan. A scatterplot of the data is shown in Figure 8.11 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.

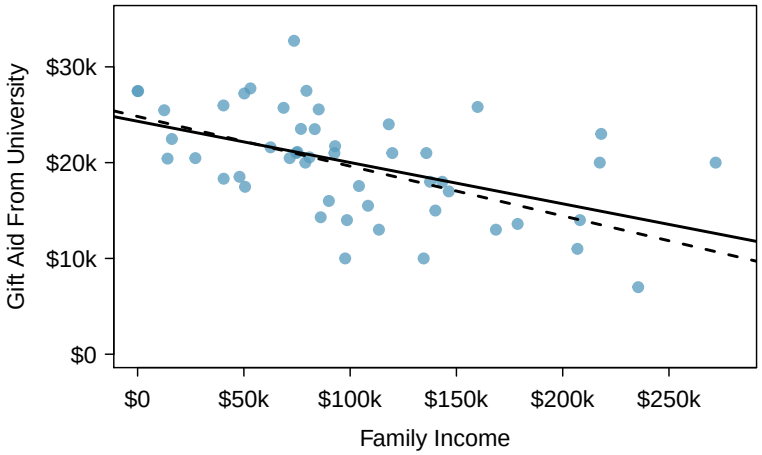

Figure 8.11: Gift aid and family income for a random sample of 50 freshman students from Elmhurst College. Two lines are fit to the data, the solid line being the *least squares line*.

G

GUIDED PRACTICE 8.7

Is the correlation positive or negative in Figure 8.11?[8]

8.2.2 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. The first option that may come to mind is to minimize the sum of the residual magnitudes:

$$|e_1| + |e_2| + \cdots + |e_n|$$

which we could accomplish with a computer program. The resulting dashed line shown in Figure 8.11 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:

$$e_1^2 + e_2^2 + \cdots + e_n^2$$

[8]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.

The line that minimizes this **least squares criterion** is represented as the solid line in Figure 8.11. This is commonly called the **least squares line**. The following are three possible reasons to choose this option instead of trying to minimize the sum of residual magnitudes without any squaring:

1. It is the most commonly used method.
2. Computing the least squares line is widely supported in 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 2. Squaring the residuals accounts for this discrepancy.

The first two reasons are largely for tradition and convenience; the last reason explains why the least squares criterion is typically most helpful.[9]

8.2.3 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 8.12), 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'll talk about more in Sections 8.3. An example of a residual that would be a potentially concern is shown in Figure 8.12, where one observation is clearly much further from the regression line than the others.

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 8.12, which represents the most common pattern observed when this condition fails: the variability of y is larger when x is larger.

Independent observations. Be cautious about applying regression to **time series** data, which are sequential observations in time such as a stock price each day. Such data may have an underlying structure that should be considered in a model and analysis. An example of a data set where successive observations are not independent is shown in the fourth panel of Figure 8.12. There are also other instances where correlations within the data are important, which is further discussed in Chapter 9.

GUIDED PRACTICE 8.8

G Should we have concerns about applying least squares regression to the Elmhurst data in Figure 8.11?[10]

[9] There are applications where the sum of residual magnitudes may be more useful, and there are plenty of other criteria we might consider. However, this book only applies the least squares criterion.

[10] 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.

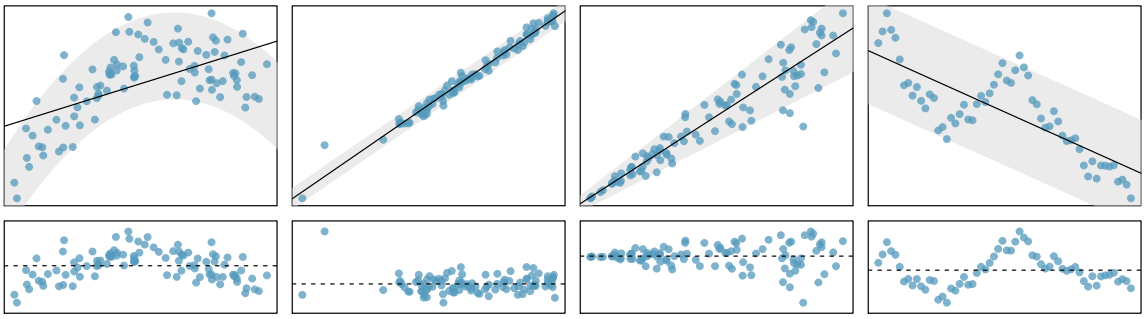

Figure 8.12: Four examples showing when the methods in this chapter are insufficient to apply to the data. First panel: linearity fails. Second panel: there are outliers, most especially one point that is very far away from the line. Third panel: the variability of the errors is related to the value of x. Fourth panel: a time series data set is shown, where successive observations are highly correlated.

8.2.4 Finding the least squares line

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

$$\widehat{aid} = \beta_0 + \beta_1 \times family\_income$$

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 5, 6, and 7, 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

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

  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 sample mean of the explanatory variable and $\bar{y}$ is the sample mean of the vertical variable, then the point $(\bar{x}, \bar{y})$ is on the least squares line.

  Figure 8.13 shows the sample means for the family income and gift aid as \$101,780 and \$19,940, respectively. We could plot the point $(101.8, 19.94)$ on Figure 8.11 on page 317 to verify it falls on the least squares line (the solid line).

Next, we formally find the point estimates b_0 and b_1 of the parameters β_0 and β_1.

	Family Income (x)	Gift Aid (y)
mean	$\bar{x} = \$101{,}780$	$\bar{y} = \$19{,}940$
sd	$s_x = \$63{,}200$	$s_y = \$5{,}460$
		$R = -0.499$

Figure 8.13: Summary statistics for family income and gift aid.

GUIDED PRACTICE 8.9

G Using the summary statistics in Figure 8.13, compute the slope for the regression line of gift aid against family income.[11]

You might recall the **point-slope** form of a line from math class, which we can use to find the model fit, including the estimate of b_0. 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

$$y - y_0 = slope \times (x - x_0)$$

IDENTIFYING THE LEAST SQUARES LINE FROM SUMMARY STATISTICS

To identify the least squares line from summary statistics:

- Estimate the slope parameter, $b_1 = (s_y/s_x)R$.
- Noting that the point $(\bar{x}, \bar{y})$ is on the least squares line, use $x_0 = \bar{x}$ and $y_0 = \bar{y}$ with the point-slope equation: $y - \bar{y} = b_1(x - \bar{x})$.
- Simplify the equation, which would reveal that $b_0 = \bar{y} - b_1\bar{x}$.

EXAMPLE 8.10

Using the point $(101780, 19940)$ from the sample means and the slope estimate $b_1 = -0.0431$ from Guided Practice 8.9, find the least-squares line for predicting aid based on family income.

Apply the point-slope equation using $(101.78, 19.94)$ and the slope $b_1 = -0.0431$:

E
$$y - y_0 = b_1(x - x_0)$$
$$y - 19{,}940 = -0.0431(x - 101{,}780)$$

Expanding the right side and then adding 19,940 to each side, the equation simplifies:

$$\widehat{aid} = 24{,}327 - 0.0431 \times family\_income$$

Here we have replaced y with $\widehat{aid}$ and x with *family_income* to put the equation in context. The final equation should always include a "hat" on the variable being predicted, whether it is a generic "y" or a named variable like "*aid*".

A computer is usually used to compute the least squares line, and a summary table generated using software for the Elmhurst regression line is shown in Figure 8.14. The first column of numbers provides estimates for b_0 and b_1, respectively. These results match those from Example 8.10 (with some minor rounding error).

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	24319.3	1291.5	18.83	<0.0001
family_income	-0.0431	0.0108	-3.98	0.0002

Figure 8.14: Summary of least squares fit for the Elmhurst data. Compare the parameter estimates in the first column to the results of Example 8.10.

[11] Compute the slope using the summary statistics from Figure 8.13:

$$b_1 = \frac{s_y}{s_x}R = \frac{5{,}460}{63{,}200}(-0.499) = -0.0431$$

EXAMPLE 8.11

Examine the second, third, and fourth columns in Figure 8.14. Can you guess what they represent? (If you have not reviewed any inference chapter yet, skip this example.)

(E)

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: $b_1 = -0.0431$. The second column is a standard error for this point estimate: $SE_{b_1} = 0.0108$. The third column is a t-test statistic for the null hypothesis that $\beta_1 = 0$: $T = -3.98$. 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 8.4.

EXAMPLE 8.12

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

(E)

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.

8.2.5 Interpreting regression model parameter estimates

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

EXAMPLE 8.13

The intercept and slope estimates for the Elmhurst data are $b_0 = 24{,}319$ and $b_1 = -0.0431$. What do these numbers really mean?

(E)

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 $b_0 = 24{,}319$ 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.

8.2.6 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 6^{th} 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.

Stephen Colbert
April 6th, 2010[12]

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 8.14

Use the model $\widehat{aid} = 24{,}319 - 0.0431 \times family\_income$ to estimate the aid of another freshman student whose family had income of \$1 million.

(E) We want to calculate the aid for $family\_income = 1{,}000{,}000$:

$$24{,}319 - 0.0431 \times family\_income = 24{,}319 - 0.0431 \times 1{,}000{,}000 = -18{,}781$$

The model predicts this student will have -\$18,781 in aid (!). However, Elmhurst College does not offer *negative aid* where they select some 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.

8.2.7 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 fit 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.

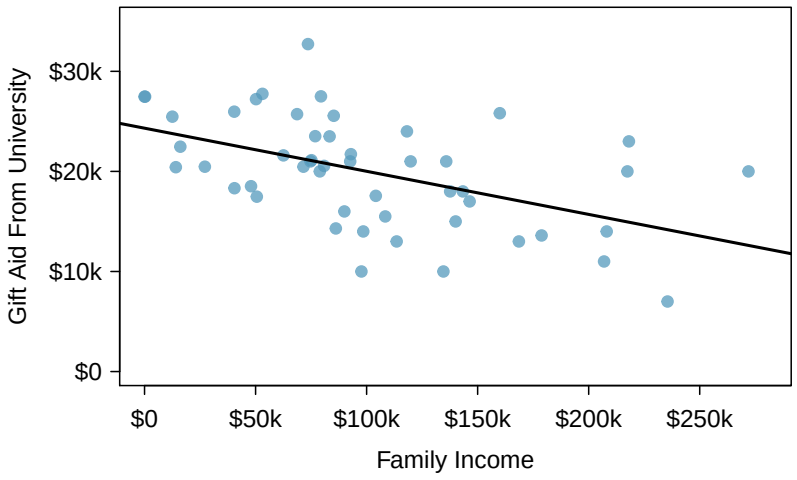

Figure 8.15: Gift aid and family income for a random sample of 50 freshman students from Elmhurst College, shown with the least squares regression line.

[12] www.cc.com/video-clips/l4nkoq

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 8.15. The variance of the response variable, aid received, is about $s^2_{aid} \approx 29.8$ million. However, if we apply our least squares line, then this model reduces our uncertainty in predicting aid using a student's family income. The variability in the residuals describes how much variation remains after using the model: $s^2_{RES} \approx 22.4$ million. In short, there was a reduction of

$$\frac{s^2_{aid} - s^2_{RES}}{s^2_{aid}} = \frac{29{,}800{,}000 - 22{,}400{,}000}{29{,}800{,}000} = \frac{7{,}500{,}000}{29{,}800{,}000} = 0.25$$

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:

$$R = -0.499 \qquad\qquad R^2 = 0.25$$

GUIDED PRACTICE 8.15

G 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?[13]

8.2.8 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. 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 8.16.

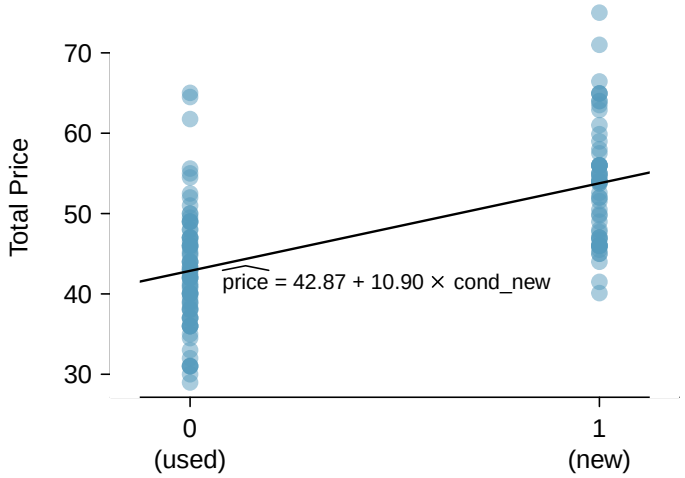

Figure 8.16: Total auction prices for the video game *Mario Kart*, divided into used ($x = 0$) and new ($x = 1$) condition games. The least squares regression line is also shown.

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

$$\widehat{price} = \beta_0 + \beta_1 \times \texttt{cond\_new}$$

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

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	42.87	0.81	52.67	<0.0001
cond_new	10.90	1.26	8.66	<0.0001

Figure 8.17: Least squares regression summary for the final auction price against the condition of the game.

The parameter estimates are given in Figure 8.17, and the model equation can be summarized as

$$\widehat{price} = 42.87 + 10.90 \times \texttt{cond\_new}$$

For categorical predictors with just two levels, the linearity assumption will always be satisfied. However, we must evaluate whether the residuals in each group are approximately normal and have approximately equal variance. As can be seen in Figure 8.16, both of these conditions are reasonably satisfied by the auction data.

EXAMPLE 8.16

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

E

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.

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 topic in Chapter 9, where we examine the influence of many predictor variables simultaneously using multiple regression.