Lecture 11.1: Linear Regression Part II

Chapter 7.2-7.4

2014/04/14

Questions for Today: Example From Text

Questions for Today: Example From Text

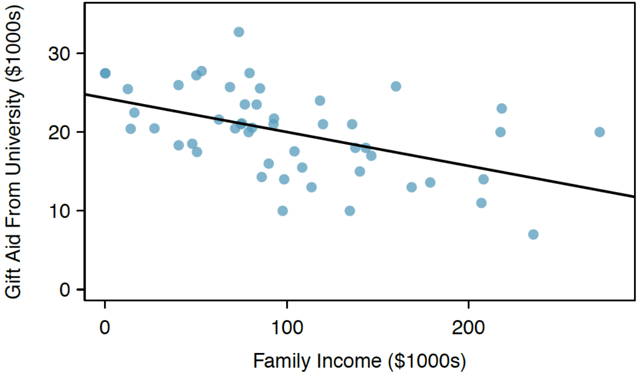
- Data: random sample of 50 students in the 2011 freshman class of Elmhurst College in Illinois.

Questions for Today: Example From Text

- Data: random sample of 50 students in the 2011 freshman class of Elmhurst College in Illinois.
- Explanatory variable: Family income

Questions for Today: Example From Text

- Data: random sample of 50 students in the 2011 freshman class of Elmhurst College in Illinois.
- Explanatory variable: Family income
- Outcome variable: Gift Aid (financial aid that is a gift, not a loan)



Questions for Today: Example From Text

Using these values,

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

they fit the least-squares line:

Questions for Today: Example From Text

Using these values,

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

they fit the least-squares line:

$$\begin{aligned} \widehat{y} &= b_0 + b_1 x \\ \text{i.e. } \widehat{\text{aid}} &= 24.3 - 0.0431 \times \text{family\_income} \end{aligned}$$

What do 24.3 and -0.0431 mean?

Point Estimates of Intercept

Point estimate of intercept b_0: 24.3 (in \$1000's) describes the average aid if the family had no income.

Point Estimates of Intercept

Point estimate of intercept b_0: 24.3 (in \$1000's) describes the average aid if the family had no income.

In this case it is relevant since some families make no income, but the intercept may have little or no practical value if there are no observations near $x = 0$.

Point Estimates of Slope

Point estimate of slope b_1: More interesting. It describes the relationship between x and y

Point Estimates of Slope

Point estimate of slope b_1: More interesting. It describes the relationship between x and y

For this example, for each additional \$1000 of family income, we would expect a student to receive a net difference of $\$1000 \times (-0.0431) = -\$43.10$ in aid on average.

Point Estimates of Slope

Point estimate of slope b_1: More interesting. It describes the relationship between x and y

For this example, for each additional \$1000 of family income, we would expect a student to receive a net difference of $\$1000 \times (-0.0431) = -\$43.10$ in aid on average.

Again, even though we've labeled aid as the outcome variable, we are not positing a causal relationship, just an association.

Extrapolate with Care

Definition of extrapolation: extend the application of a method or conclusion to an unknown situation by assuming that existing trends will continue or similar methods will be applicable.

Extrapolate with Care

Definition of extrapolation: extend the application of a method or conclusion to an unknown situation by assuming that existing trends will continue or similar methods will be applicable.

What would be the gift aid given to a family with one million dollars ($x = 1000$) in family income?

$$24.3 - 0.0431 \times 1000 = -18.8$$

The school will take $18,800 dollars away from you?

Categorical Predictor x With Two Levels

x need not just be a numerical value; it can also be categorical.

Categorical Predictor x With Two Levels

x need not just be a numerical value; it can also be categorical.

Ex: Ebay price for the video game Mario Kart. We convert the categorical x into a indicator variable cond_new which has 2 levels:

1. $x = 0$: game is used. This is the baseline level.
2. $x = 1$: game is new.

Categorical Predictor x With Two Levels

x need not just be a numerical value; it can also be categorical.

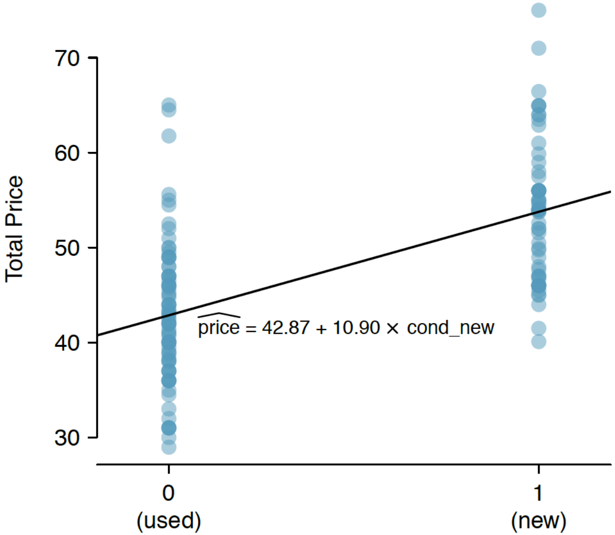
Ex: Ebay price for the video game Mario Kart. We convert the categorical x into a indicator variable cond_new which has 2 levels:

1. $x = 0$: game is used. This is the baseline level.
2. $x = 1$: game is new.

The linear model is thus

$$\widehat{\text{price}} = b_0 + b_1 \times \text{cond\_new}$$

Categorical Predictor x With Two Levels



Categorical Predictor x With Two Levels

The least-squares line is

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

Categorical Predictor x With Two Levels

The least-squares line is

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

So when the game is

- Old, we have $x = 0$, so the fitted value is $\$42.87 + \$0 = \$42.87$
- New, we have $x = 1$, so the fitted value is $\$42.87 + \$10.90 = \$53.77$

Categorical Predictor x With Two Levels

The least-squares line is

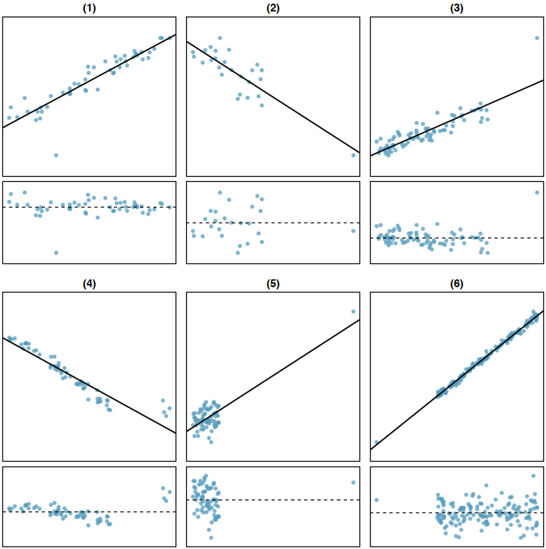
$$\widehat{\text{price}} = 42.87 + 10.90 \times \text{cond\_new}$$

So when the game is

- Old, we have $x = 0$, so the fitted value is $\$42.87 + \$0 = \$42.87$
- New, we have $x = 1$, so the fitted value is $\$42.87 + \$10.90 = \$53.77$

This can be generalized for predictor variables x with more than two levels, but this requires a different encoding of x.

Types of Outliers in Linear Regression



Types of Outliers in Linear Regression

Especially in cases 3 and 5, the outliers seem to be pulling the least-squares line towards them.

Types of Outliers in Linear Regression

Especially in cases 3 and 5, the outliers seem to be pulling the least-squares line towards them.

Points that fall horizontally away from the center of the cloud tend to pull harder on the line, so we call them points with high leverage, i.e. large influence.

Next Example

Are Higher Movie Budgets Associated with Higher IMDB Ratings for Movies Made from 1980-2005? Guesses?

Next Example

Are Higher Movie Budgets Associated with Higher IMDB Ratings for Movies Made from 1980-2005? Guesses?

But first, are the changes from

- 100 to 200
- 100,100 to 100,200

the same?

Next Example

We consider a $\log_{10}$ transformation:

x	y	$y-x$	$\frac{y}{x}$	$\log_{10}\left(\frac{y}{x}\right)$
100	200	100	2	0.301
100100	100200	100	1.000999	0.00004342

Next Example

We consider a $\log_{10}$ transformation:

x	y	$y - x$	$\frac{y}{x}$	$\log_{10}\left(\frac{y}{x}\right)$
100	200	100	2	0.301
100100	100200	100	1.000999	0.00004342
100000	200000	100000	2	0.301

Next Example

We consider a $\log_{10}$ transformation:

x	y	$y - x$	$\frac{y}{x}$	$\log_{10}\left(\frac{y}{x}\right)$
100	200	100	2	0.301
100100	100200	100	1.000999	0.00004342
100000	200000	100000	2	0.301

Recall that $\log_{10}\left(\frac{y}{x}\right) = \log_{10}(y) - \log_{10}(x)$

Next Example

We consider a $\log_{10}$ transformation:

x	y	$y - x$	$\frac{y}{x}$	$\log_{10}\left(\frac{y}{x}\right)$
100	200	100	2	0.301
100100	100200	100	1.000999	0.00004342
100000	200000	100000	2	0.301

Recall that $\log_{10}\left(\frac{y}{x}\right) = \log_{10}(y) - \log_{10}(x)$

So we are considering multiplicative changes, and not additive changes.

Next Time

Multiple Regression: As opposed to simple linear regression where there is only one predictor/explanatory variable x, we now consider many variables $x_1, x_2, \ldots$