Lecture 25: Linear Regression Part II

Chapter 7.2-7.4

Quiz 9

http://www.nature.com/news/scientific-method-statistical-errors-1.14700

Question 1: What is p-hacking?
Answer 1: Data-dredging AKA "trying multiple things until you get the desired result"
http://simplystatistics.org/2013/08/26/statistics-meme-sad-p-value-bear/

Quiz 9

http://www.nature.com/news/scientific-method-statistical-errors-1.14700

Question 2: Say a scientist obtains a p-value of 0.01. An incorrect interpretation of this is that it is the probability of a "false alarm" (type I error)... If one wants to make a statement about this being a false alarm, what additional piece of information is required?
Answer 2: The plausibility of the hypothesis being tested for.

Superpopulation

What if your sample consists of the entire population. Who do the results generalize to?

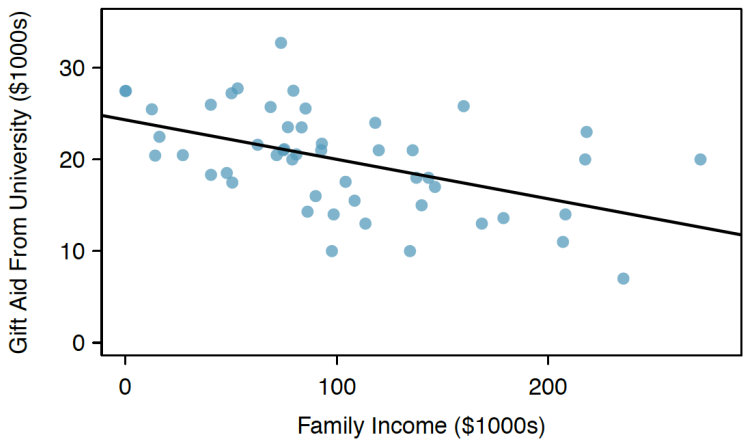
Example: Say we run a regression based everyone in this class.

You can view this class as a random sample from a hypothetical superpopulation of people:

- Reedies who would take MATH 141 from 2010-2018?
- Students at a liberal arts college who take an intro stats class?
- Parallel universe?

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



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:

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

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.

Here it is relevant since some families make no income, but the intercept may not make sense 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

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

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

Extrapolate with Care

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 \$1,000,000 (i.e. $x = 1000$) 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 can also be categorical.

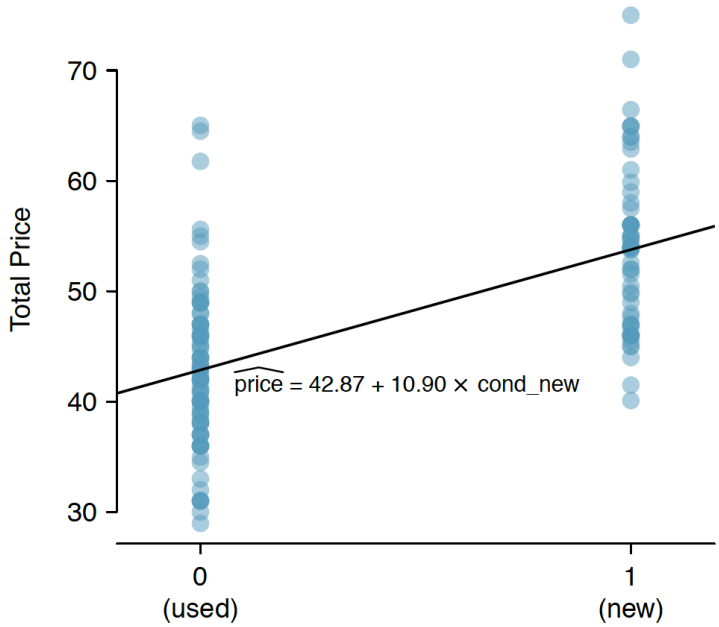
Ex: Ebay price for the video game Mario Kart. We convert the categorical x into a indicator variable cond_new which has 2 levels:

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

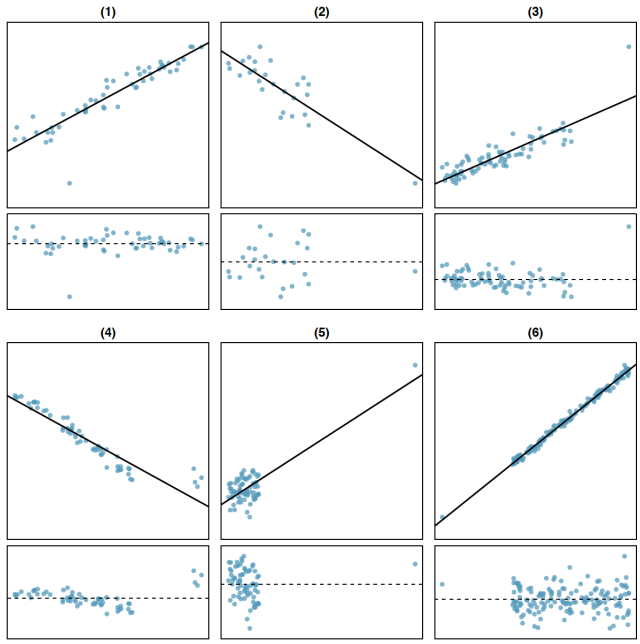
$$\widehat{\text{price}} = 42.87 + 10.90 \times \text{cond\_new}$$

So when the game is

- Used, 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.

Types of Outliers in Linear Regression



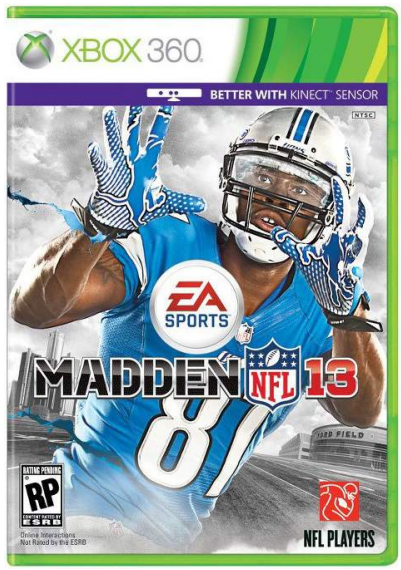
Types of Outliers in Linear Regression

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

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.

Concept: Regression to the Mean

The Madden Curse. Many NFL players who feature on the cover of the video game Madden end up having subpar subsequent years, leading many to believe there is a curse.



Concept: Regression to the Mean

Regression to the mean is the phenomenon that if a variable is extreme on its first measurement, it will tend to be closer to the average on its second measurement.

Madden is selecting players who had exceptional seasons the previous year: the exceptional performance by the players who appear on the cover is not sustainable.

So while it looks like a curse, it is just players reverting back to their "mean" level of performance.

Next Example

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

Next Time

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