

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?

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”

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?

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?

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.

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:

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:

- ▶ Reemies who would take MATH 141 from 2010-2018?

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:

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

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:

- ▶ Reemies 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

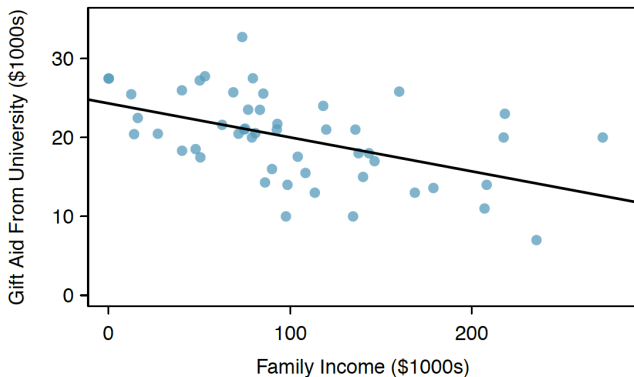
- ▶ Data: random sample of 50 students in the 2011 freshman class of Elmhurst College in Illinois.

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



Questions for Today: Example From Text

Using these values,

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

Point Estimates of Intercept

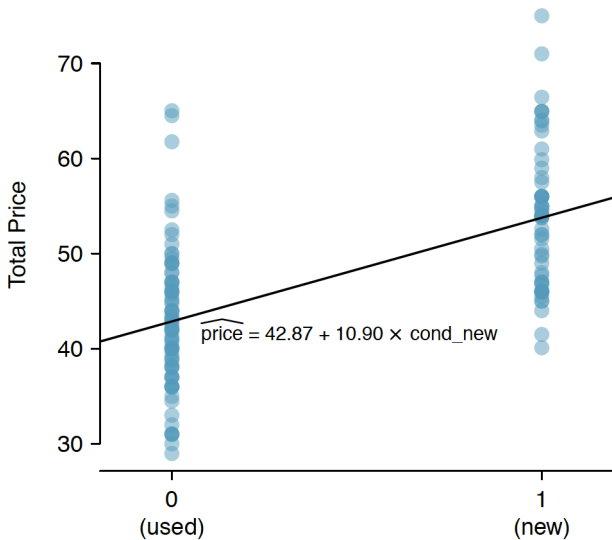
Point Estimates of Slope

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.

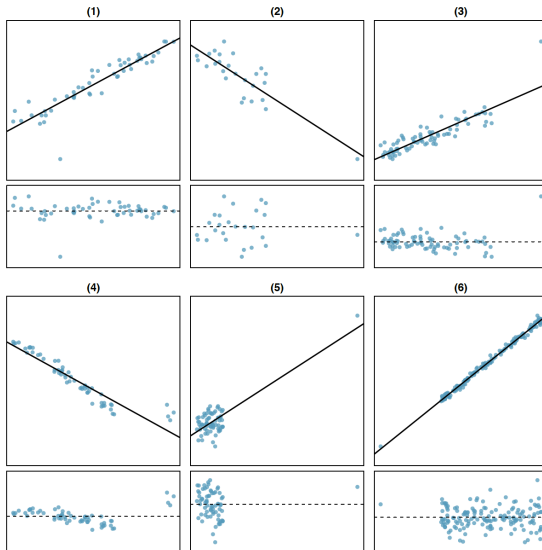
Categorical Predictor x With Two Levels

Categorical Predictor x With Two Levels



Categorical Predictor x With 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.

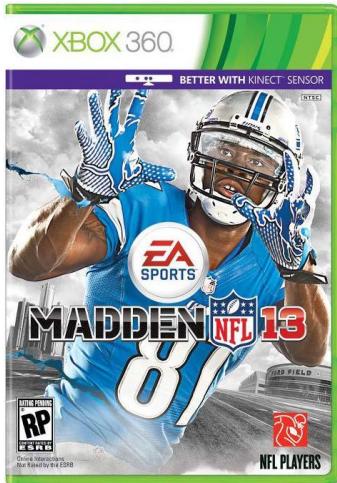
Types of Outliers in Linear Regression

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

Concept: Regression to the Mean

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**.

Concept: Regression to the Mean

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, \dots