Inference for linear regression

13 March 2019

PLSC 309

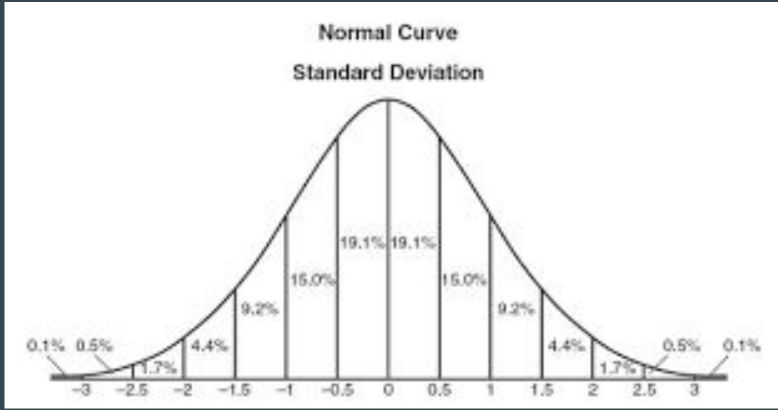
New office hours starting next week

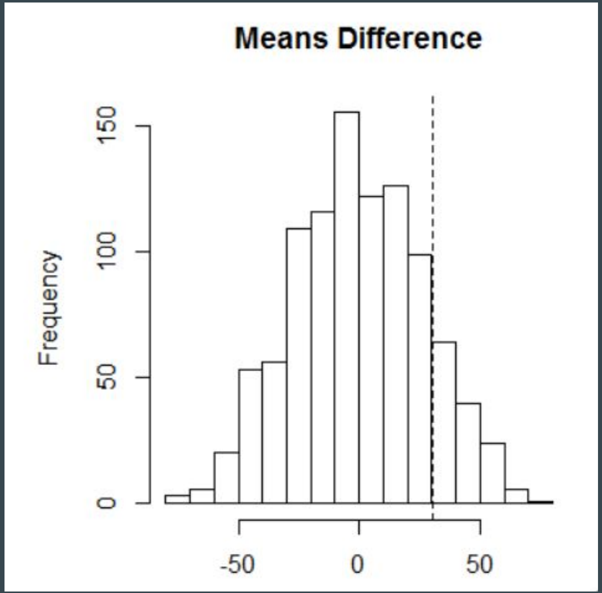
- Monday and Friday from 9-12 (Sparks B001)
- No more office hours on Wednesday!

Assumptions

- If you haven't noticed already, we make A LOT of assumptions in statistics
- But what exactly is an assumption?
- An assumption is a statement we make *without any evidence*
  - In statistics, evidence means data
  - An assumption is a statement we make *without any data*
- This is in contrast to *empirical estimation*, where we use data to make an informed guess

Null hypothesis: assumptions vs empirics

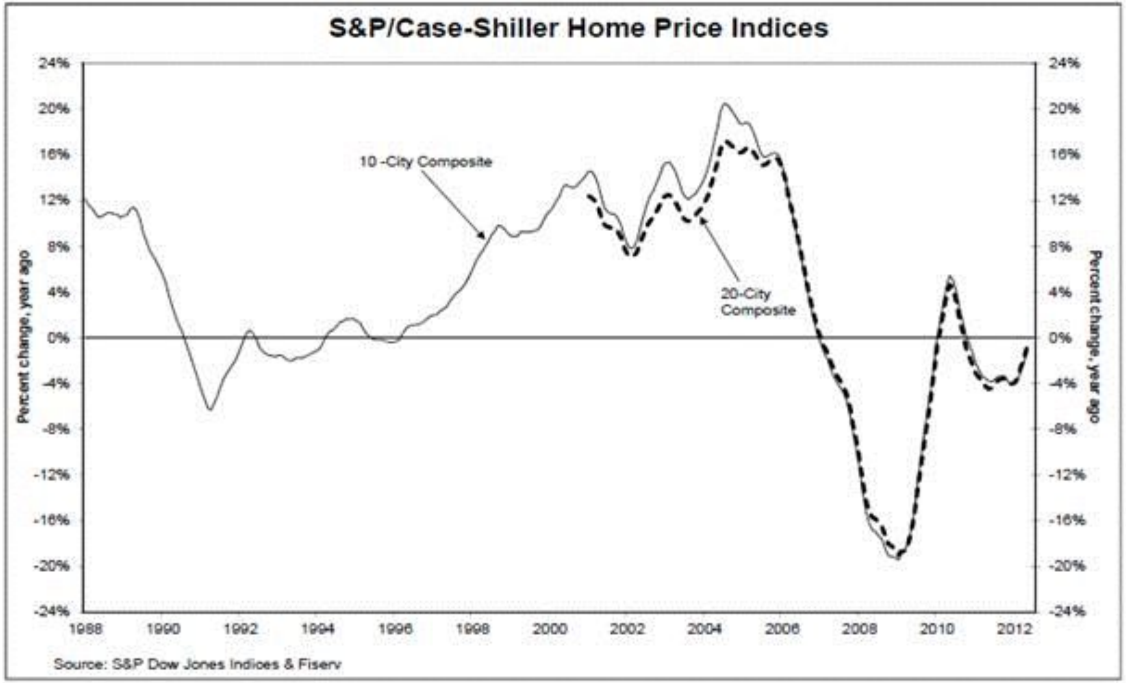




Do assumptions really matter?

- Let’s take the 2008 financial crisis as an example
- Mortgage-backed securities are collections of mortgages, bundled together, and traded on an open market
- Each security is rated by a credit agency
  - AAA+ rating is essentially good as cash
- When rating the credit agency assumes that all mortgages in the security are *independent from one another*

Do assumptions really matter?



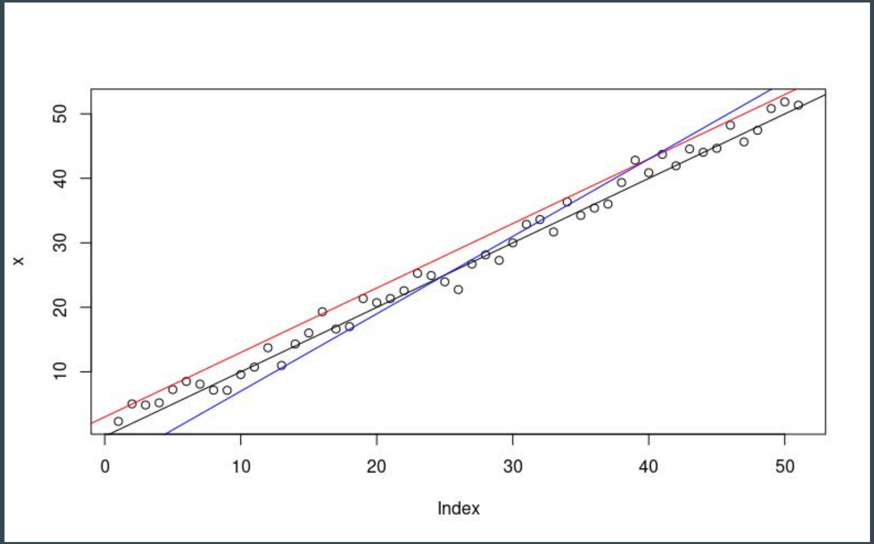
The chart above depicts the annual returns of the 10-City and the 20-City Composite Home Price Indices. In May 2012, both Composites were up by 2.2% month-over-month, and posted annual returns of -1.0% and -0.7%, respectively.

Review

- We want to build a statistical model that predicts Y from X (we call this $\hat{Y}$)
- We assume that the function that connects X to Y is linear (straight line)
- Find the straight line with the slope and y-intercept that best fits our data
  - Slope: β
  - Y-intercept: α

Review: Ordinary Least Squares

- Find the line (i.e slope and intercept) that minimizes the squared differences ($\hat{Y}$-Y)
- $\hat{Y}$-Y are known as errors or residuals



Calculating model error

- The best square fit wants to find the smallest residuals in either direction
- We can do this by first squaring each of our residuals
  - $e_1^2, e_2^2, e_3^2, \ldots$
- And then summing the residuals
  - $e_1^2 + e_2^2 + e_3^2 + \ldots$
- This is known as the *sum of least squares*
- A regression model calculated with the sum of least squares is known as *Ordinary Least Squares (OLS) Regression*

Assumptions for OLS

There are four assumptions that we have to make for linear regression:

1. Linearity / additivity
2. Our residuals must be:
   a. Independent
   b. Homoscedastic (constant variance)
   c. Normally distributed

Where do these assumptions come from?

- We are not just estimating *any* relationship, f(x) between **X** and Y
- We are estimating a very specific relationship
    - Linear / additive
- Why linear relationships?
    - The math is easy
    - Central Limit Theorem implies a linear relationship
- Because we are relying on the C.L.T., *i.i.d.* is a fundamental assumption
    - Independent
    - Identically distributed

Guessing our model parameters

- If we could compare all the possible slopes and intercepts
- We know how to tell what line is the best fit...
- ...the one that minimizes the sum of our errors
    - $e_i = \hat{Y}_i - Y_i$
    - $\sum(e_i^2)$
- So how do we guess the model parameters?

Key insight: linear models are additive

- Additivity means that if we add or subtract an X variable from the model, the parameters stay the same
- In other words, if we have a model: $Y = \beta_1 X_1 + \beta_2 X_2$
- Then β_1 will stay the same for a new model: $Y = \beta_1 X_1$
- This means that *we can calculate each slope separately*

Correlation

- The correlation represents the strength of linear relationships between two variables
- Ranges from -1 to 1
  - 1 = perfect positive linear relationship
  - -1 = perfect negative relationship
  - 0 = no linear relationship whatsoever

How to calculate correlation

$$R = \frac{1}{n-1} \sum_{i=1}^{n} \frac{x_i - \bar{x}}{s_x} \frac{y_i - \bar{y}}{s_y}$$

An average of deviations from the mean, scaled by their standard deviation

- Greater standard deviations = smaller correlation
- Greater deviations from average *for the same observation* = larger correlations

Calculating β

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

1. We need three pieces of information
   - R = correlation between X and Y
   - S_y = standard deviation of Y
   - S_x = standard deviation of X
2. Calculate correlation coefficient
3. Multiply by st. dev of Y / st. dev of X

Calculating β

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

- Correlation represents the linear relationship between X and Y
  - Can be positive or negative
- This is adjusted by the change in Y over the change in X
- If there is a very strong linear relationship and X accounts for a lot of variation in Y, there will be a large slope

Calculating α

- Now that we have our slope parameters we can estimate the intercept
- This involves subtracting the mean of our X variables from the mean of Y
- $\alpha = \text{mean}(Y) - \beta * \text{mean}(X)$

P-values for β

- Say we're estimating a model and we find a β of 1 for a variable X
- Is X related or unrelated to Y?
- This sounds like a null hypothesis test!
- $H_0: \beta = 0$
- $H_A: \beta \neq 0$

P-values for β

- We can find a Z-score
- PE = β - 0
- Z = PE / SE

P-values for β

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-6.7142	5.4567	-1.23	0.2300
unemp	-1.0010	0.8717	-1.15	0.2617
				$df = 25$

- PE = -1.0010 - 0 = -1.0010
- Z = -1.0010 / 5.4567 = -0.183
- $-2 < Z < 2$
- We can not reject H_0

P-values for β

	Estimate	Std. Error	t value
(Intercept)	24.3193	1.2915	18.83
family_income	-0.0431	0.0108	-3.98

- PE = -0.0431 - 0 = -0.0431
- Z = -0.0431/ 0.0108 = -3.99
- -Z < -2
- We reject H_0 and accept H_A

Confidence intervals for β

- We construct confidence intervals for β just like we do for difference in means
- PE $\pm$ Z * SE
  - PE (point estimate): your estimated value for β
  - Z: whichever value corresponds to your level of confidence
  - SE: standard error of β

Confidence intervals for β

	Estimate	Std. Error	t value
(Intercept)	24.3193	1.2915	18.83
family_income	-0.0431	0.0108	-3.98

- PE: -0.0431
- Z: 2 (95% confidence)
- SE: 0.0108
- CI = PE ± Z * SE
- CI = -0.0431 ± (2 *0.0108) = (-0.0647, -.0.0215)
- CI does not overlap 0, so we reject H_0 and accept H_A

Interpreting β

- The slope of the line measures the relationship between X and $\hat{Y}$
  - For a one unit change in X, what will be the change in $\hat{Y}$?
- It *does not* mean the effect of X on Y
  - This is a causal statement
- Correlation does not imply causation

Interpreting β: example

- We are interested in explaining voter turnout
- We'll use the following explanatory variables
  - X_1 = Age
  - X_2 = Income
  - X_3 = Political Party
  - X_4 = Education
- Say the estimate for β_2 is 3, and you find a p-value < .001
- This means that income has a strong positive effect on voter turnout, right?
  - Let's evaluate this statement

Problems with causality

In a linear regression, there are two threats to causality:

1. Endogeneity
   a. You argue that X causes Y, when really Y causes X
2. Omitted variable bias
   a. You argue that X causes Y, but really Z causes both X and Y

Endogeneity

- When you say X causes Y, but Y causes X
- Also called “reverse causality”
- Use of umbrellas positively correlated with rain; do not cause rain
- For example...
  - GDP is negatively related to war, but war causes drops in GDP
  - Education is positively related to income, but wealthier people are better able to afford schools

Omitted variable bias

- When you say X causes Y, but really Z causes both X and Y
- Also called “confounding”
- You have a regression to predict price of a use car by mileage, but you are omitting the car’s age
- For example...
    - Income and education are both predicted by parent’s wealth
    - School performance and graduation rates both affected by neighborhood poverty

Violations of causality

- We are interested in explaining voter turnout
- We'll use the following explanatory variables
  - X_1 = Age
  - X_2 = Income
  - X_3 = Political Party
  - X_4 = Education
- Not much endogeneity
- Significant problems with omitted variable bias
- But even if you do not have a causal interpretation, can you trust these results?

Violations of additivity

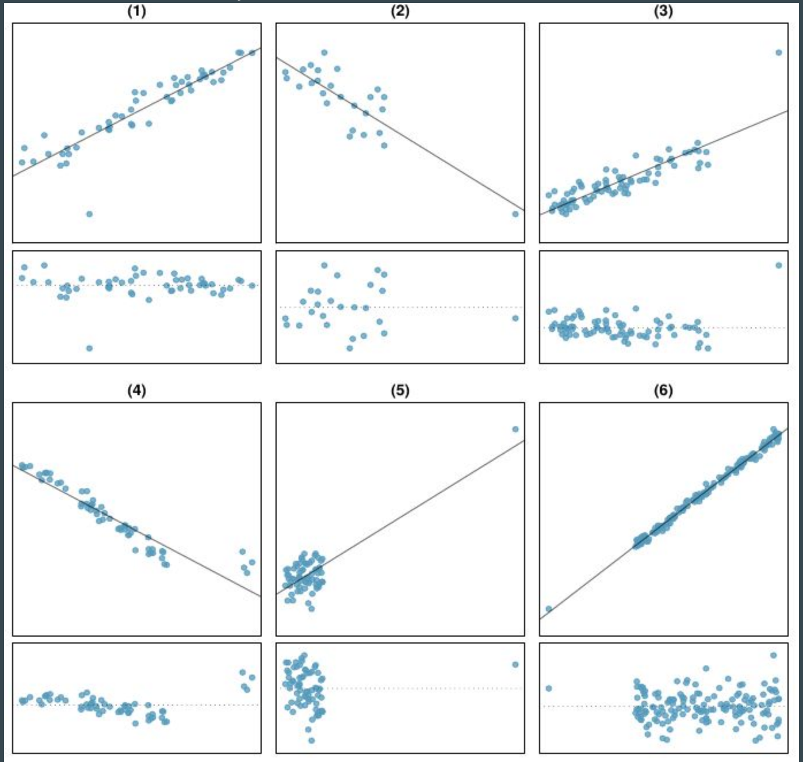
- Remember, linear regression learns the best *linear* model, f(x), for our outcome, Y
  - Linear = additive
- If a model is additive, one explanatory variable must not correlate or affect another variable
- To put it differently, all explanatory variables must be independent of one another

Violations of additivity

- We are interested in explaining voter turnout
- We'll use the following explanatory variables
  - X_1 = Age
  - X_2 = Income
  - X_3 = Political Party
  - X_4 = Education
- Education, income, political party, and age, are all hopelessly intertwined
- This means we cannot separately interpret each coefficient

Outliers

- An outlier is a point that is very different from the rest of the data



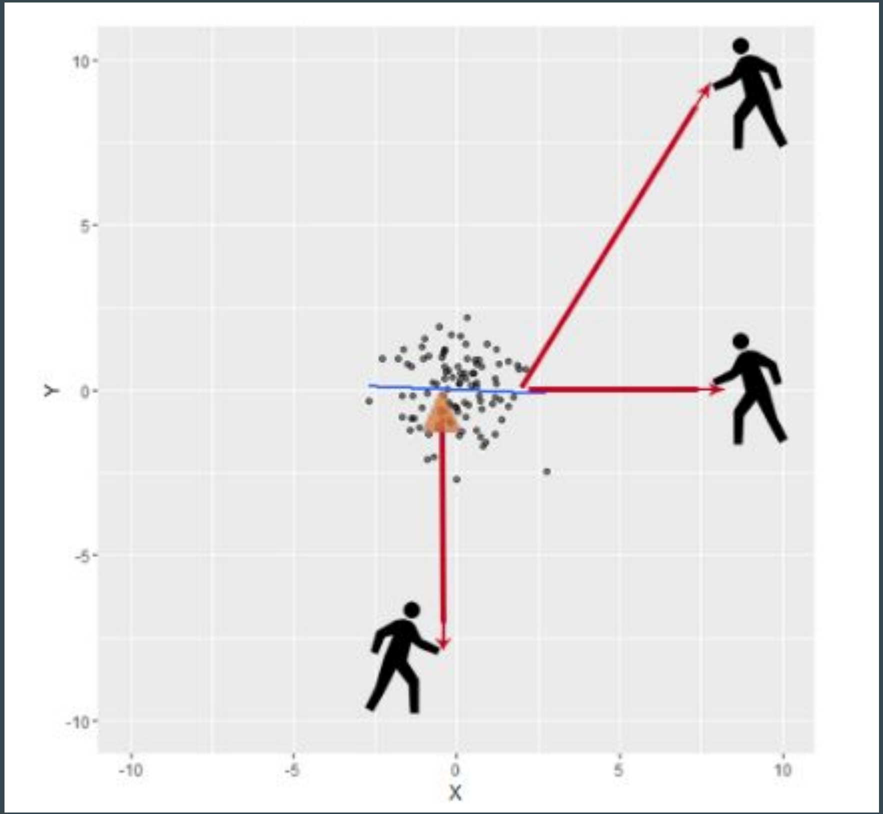
Types of outliers

- An outlier with *high leverage* is one where X_i is very different from mean(X)
  - The further away x_i is to mean(X), the *more leverage it has*
- Leverage measures the *potential* an observation has to distort or pull our best linear fit
- However, high leverage observations *do not necessarily* distort our best fit
- Outliers that pull our best fit away from its value if that outlier hadn't been there are *influential* outliers
  - Outliers that are *parallel or perpendicular* to the fitted line have low influence

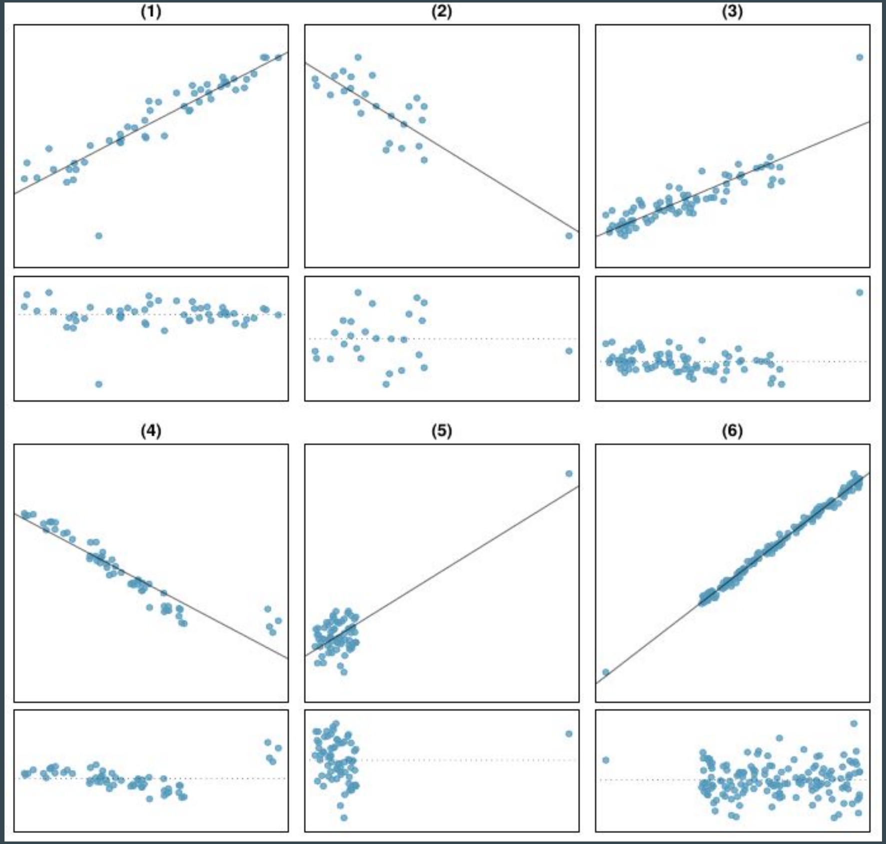
Types of outliers (succinctly)

- *Leverage* depends on how far the outlier is from the mean value
- *Influence* depends on how much the outlier affects the best linear fit

Influence vs leverage



Influence vs leverage



Review

- We learned how to calculate important quantities for linear regression
  - $\beta = R * (\sigma_Y / \sigma_{X)})$
  - α = mean(Y) - β * mean(X)
- We learned how to calculate P-values and CIs for β
  - H_0: $\beta = 0$
  - H_A: $\beta \neq 0$
  - Usual Z-score formula

Review

- We learned about problems with interpreting β
    - Endogeneity
    - Omitted variable bias
    - Failure of additivity assumptions
- We also learned about outliers and their potential impact on our model specification
    - Leverage
    - Influence