EWMBA 296 (Fall 2015)

Section 4

GSI: Fenella Carpena

November 12, 2015

Agenda for Today

- Regression: Big Picture
- Linear Regression Model
  - Basic Concepts
  - Interpreting b_0 and b_1
  - Measures of Fit
  - Pitfalls to guard against when using r^2
- Incorporating Non-Linear Patterns

Regressions: Big Picture

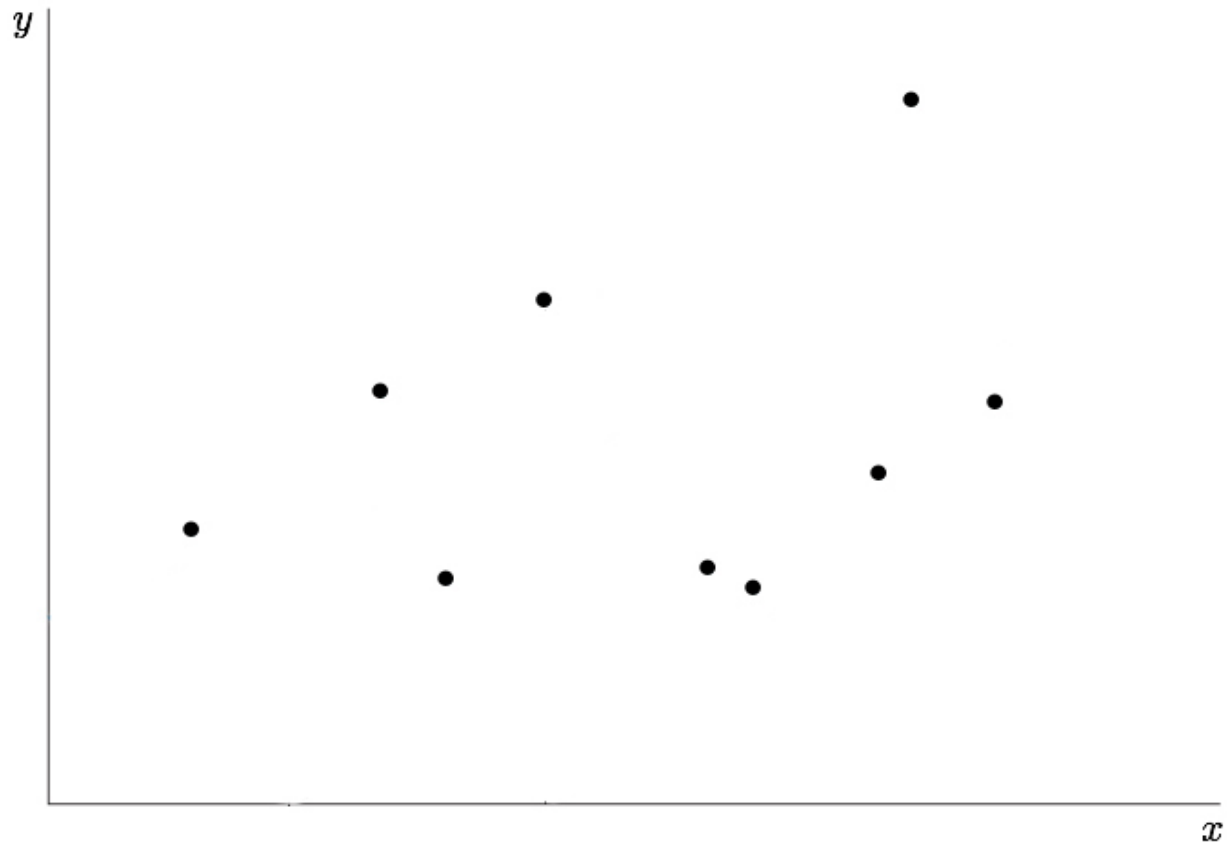
- **Regression analysis** is a statistical technique to estimate relationships among variables. Used for **predictive modeling** or **causal inference.**
- **Predictive modeling** is the process of applying a statistical model to *predict new or future observations*.
  - Example: Santa Cruz Police Department fits statistical models to historical crime data, tok make projections about areas/times at highest risk for future crimes.
- With **causal inference**, the purpose is to infer whether one variable has a *causal effect* on another variable. We typically use **experiments** to produce data that reveal causal relationships. Without experiments, finding causal relationships is often difficult because of **lurking variables.**
  - Example: Does eating meat cause colon cancer?
- In this class, we will focus on the **linear regression model**.

Linear Regression Model: Basic Concepts

- A **linear regression model** is a model that posits a linear relationship between a **response variable** y, and the **explanatory variable** x.
- With a scatterplot of data, what we are trying to do is to draw a line that "best fits" the data.

Linear Regression Model: Basic Concepts

- To find the b_0 and b_1 that "best" fit the data, we use the **least squares**.
- **Least squares** is a type of regression analysis that picks the line that minimizes the sum of squared residuals.



Linear Regression Model: Basic Concepts

- Using least squares, we obtain the following equations for b_0 and b_1 (very important!)

$$b_1 = r\frac{s_y}{s_x} \quad \text{or equivalently} \quad b_1 = \frac{cov(x,y)}{s_x^2}$$

$$b_0 = \bar{y} - b_1\bar{x}$$

- Important properties of the least squares regression line:
  - The line always passes through $(\bar{x}, \bar{y})$
  - Sample mean of the residuals is zero, $\bar{e} = \frac{1}{n}\sum_{i=1}^{n} e_i = 0$

Interpreting b_0 and b_1 in $\widehat{y} = b_0 + b_1 \cdot x$

- b_0 is the value of the regression line when $x = 0$
  - In some applications, has meaningful economic interpretation. For example (as in lecture):

$$\widehat{Price} = 15 + 2697 \cdot Weight$$

  - In other applications, has no real-world meaning. For example:

$$\widehat{BloodLoss} = 552.44 - 130 \cdot Height$$

- b_1 is the change in y *associated with* a one unit change in x
  - Important: It is incorrect to say b_1 is the change in y *caused by* a change in x
- We should be careful when making **extrapolations** (i.e., making predictions/statements beyond the scope of what is observed in the data), since they are less reliable.

Exercise 2.1 from Section Notes 4

A regression of *wage* (hourly wage, measured in 1976 dollars per hour) and *educ* (years of schooling) using data from a random sample of 526 American workers yields the following:

$$\widehat{wage} = -0.90 + 0.54 \cdot educ$$

(a) Interpret the intercept of this regression.
(b) It turns out that all workers in the data have at least 8 years of education. Does this help reconcile your answer in part (a)?

Exercise 2.1 from Section Notes 4

A regression of *wage* (hourly wage, measured in 1976 dollars per hour) and *educ* (years of schooling) using data from a random sample of 526 American workers yields the following:

$$\widehat{wage} = -0.90 + 0.54 \cdot educ$$

(c) Interpret the slope of this regression.
(d) Suppose a worker obtains 4 more years of schooling. What is the regression's prediction for the change in the worker's hourly wage?

Exercise 2.1 from Section Notes 4

A regression of *wage* (hourly wage, measured in 1976 dollars per hour) and *educ* (years of schooling) using data from a random sample of 526 American workers yields the following:

$$\widehat{wage} = -0.90 + 0.54 \cdot educ$$

(e) Suppose that the sample average wage in the data is \$8 per hour. What is the sample average years of schooling?
(f) Does this regression indicate that higher education causes higher earnings? Explain.

Linear Regression Model: Measures of Fit

- Having estimated a linear regression, you might wonder how well that regression line describes the data. Two measures we discussed in class that helps assess the goodness-of-fit are the r^2 and the **standard error of the regression (SER).**
- The r^2 is the fraction of the sample variance of y that is explained by the fitted line.
  - $r^2 = [corr(x, y)]^2$; it is unit free.
  - r^2 ranges between 0 and 1.
  - r^2 close to 1 means that x is good at predicting y (data points are closer to the line)
- The SER, denoted s_e, is the standard deviation of the residuals e.
  - It measures the spread of the data points around a regression line; units are the same as y.
  - Formula: $s_e = \sqrt{\frac{e_1^2+e_2^2+\ldots+e_n^2}{n-2}}$
- A high r^2 means that SER is low (and vice versa).

Exercise 2.3 from Section Notes 4

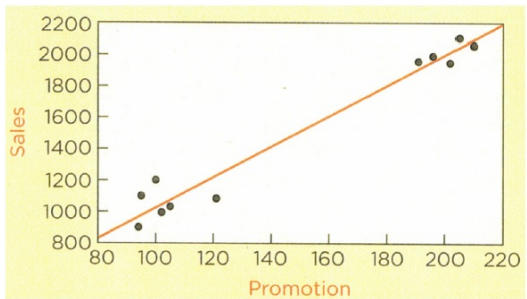
Sketch a hypothetical scatterplot of data for an estimated regression with $r^2 = 0.9$. Sketch a hypothetical scatterplot of data for a regression with $r^2 = 0.5$.

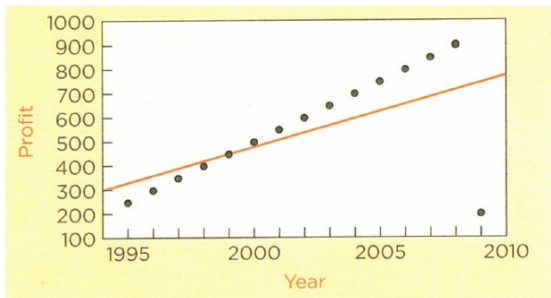
Pitfalls to guard against when using r^2

- A high r^2 does not mean that the regression indicates a causal relationship between x and y.

- A high r^2 does not mean there are no lurking variables (and vice versa).

- A high r^2 does not mean that a linear regression is the appropriate model to use.

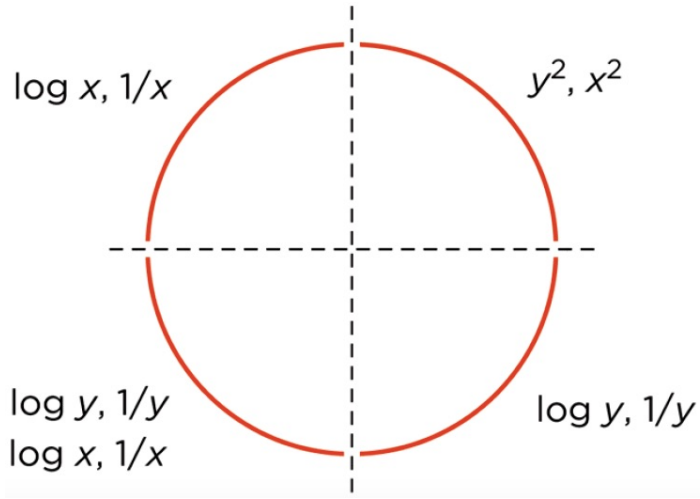




- You cannot compare r^2 between two regressions that do not use the same response variable and/or data.

Incorporating Non-Linear Patterns

- So far, we've focused on estimating a *linear model* between x and y. But in some applications, this linear relationship will not capture the curvature of the data.
- Fortunately, **we can incorporate these non-linear patterns into the linear regression model by transforming the x and/or y variables.**
- How to choose a transformation? No hard and fast rules, an iterative process in practice.
- Tukey's "Bulging Rule" provides suggestions on how to transform variables.



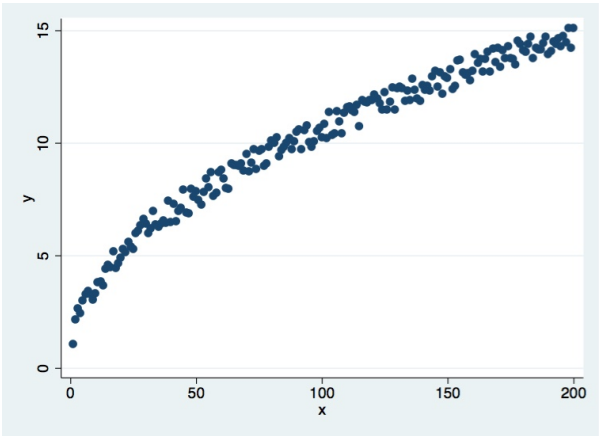
Incorporating Non-Linear Patterns

- One of the most common transformations we will encounter is the natural log.
- It is often used because it allows us to examine percentage changes.
- Note that the interpretation of b_0 and b_1 depends on whether x, y or both are in logarithms.

Case	Regression	Interpretation
Log-Log	$\widehat{ln(y)} = b_0 + b_1 \cdot ln(x)$	A 1% change in x is associated with a b_1% change in y, so b_1 is the **elasticity** of y with respect to x.
Linear-Log	$\widehat{y} = b_0 + b_1 \cdot ln(x)$	A 1% change in x is associated with a change in y of $0.01 \cdot b_1$
Log-Linear	$\widehat{ln(y)} = b_0 + b_1 \cdot x$	A change in x by one unit is associated with a $100 \cdot b_1$% change in y.

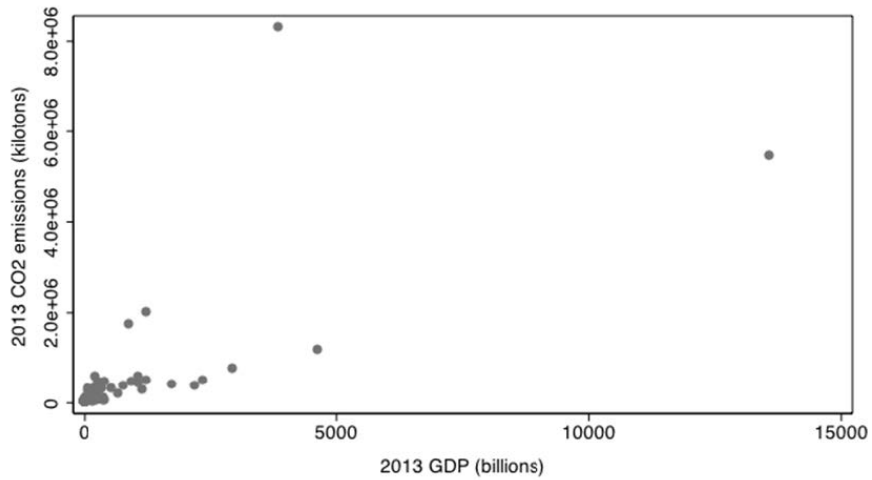
Exercise 2.5 from Section Notes 4

Consider the scatterplot of y and x below. Explain what transformation you would use, and what regression you would estimate to model this pattern. Can you think of two variables that might have an economic relationship shaped like this?



Quiz 5 2014, Question 1

In your new job as an economic consultant, you've been asked by a client to summarize the relationship between CO_2 emissions and economic growth. You obtain a 2013 data from the World Bank on CO_2 emissions (kilotons) and GDP (billions). You produce the following scatterplot. Describe two specific advantages of using a log-log transformation to analyze this dataset.



Quiz 5 2014, Question 2

After performing a log-log transformation, you run a regression of the two variables and obtain the following results. Write down the equation of the fitted line and interpret the slope in words.
Extra Question: Interpret the r^2 of this regression.

Regression statistics	
R-squared	0.8960
s_e	0.8164
Observations	178

	Coefficient	Standard error	t	P>|t|
Constant	6.2506	0.1008	61.98	0.000
Log(GDP, in billions)	1.0011	0.0257	38.95	0.000

Quiz 5 2014, Question 3 and 4

What is the correlation between log(GDP) and log(CO_2 emissions)?
Your analyst tells you that the standard deviation of log(GDP) is 2.3870. What is the standard deviation of log(CO_2 emissions)?

Regression statistics	
R-squared	0.8960
s_e	0.8164
Observations	178

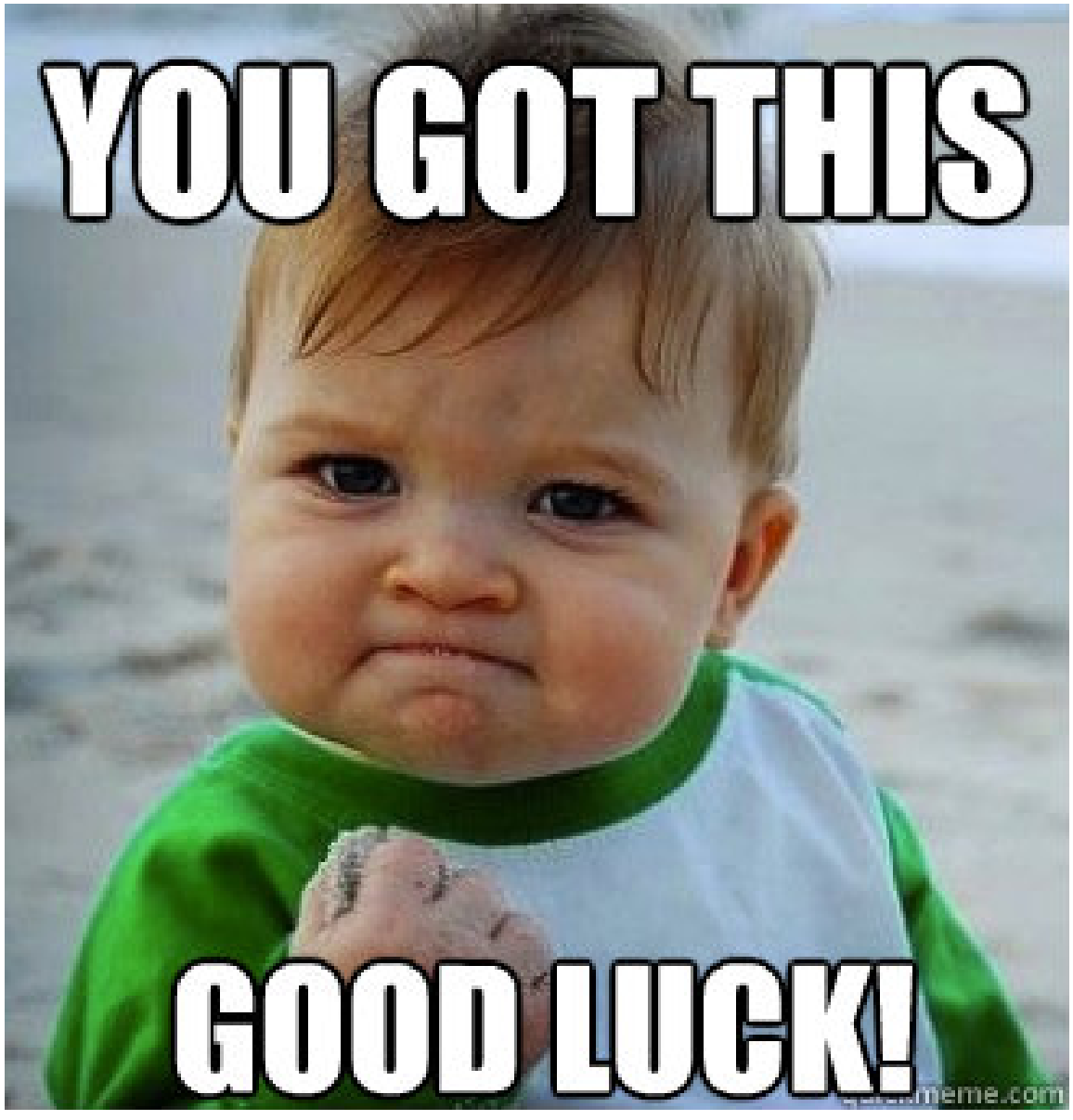
	Coefficient	Standard error	t	P>|t|
Constant	6.2506	0.1008	61.98	0.000
Log(GDP, in billions)	1.0011	0.0257	38.95	0.000

Quiz 5 2014, Question 6

Mark each statement TRUE or FALSE.

Statement 1: "High r-squared means you have small residuals."

Statement 2: "High r-squared means that you do not need to check residuals to determine whether a linear model fits the data."


YOU GOT THIS
GOOD LUCK!
meme.com