STT 200 – Lecture 1, section 2,4
Recitation 6
(10/9/2012)

TA: Zhen (Alan) Zhang

zhangz19@stt.msu.edu
Office hour: (C500 WH) 1:45 – 2:45PM Tuesday
(office tel.: 432-3342)
Help-room: (A102 WH) 11:20AM-12:30PM, Monday, Friday

Class meet on Tuesday:
3:00 – 3:50PM A122 WH, Section ***02***
12:40 – 1:30PM A322 WH, Section ***04***

Overview

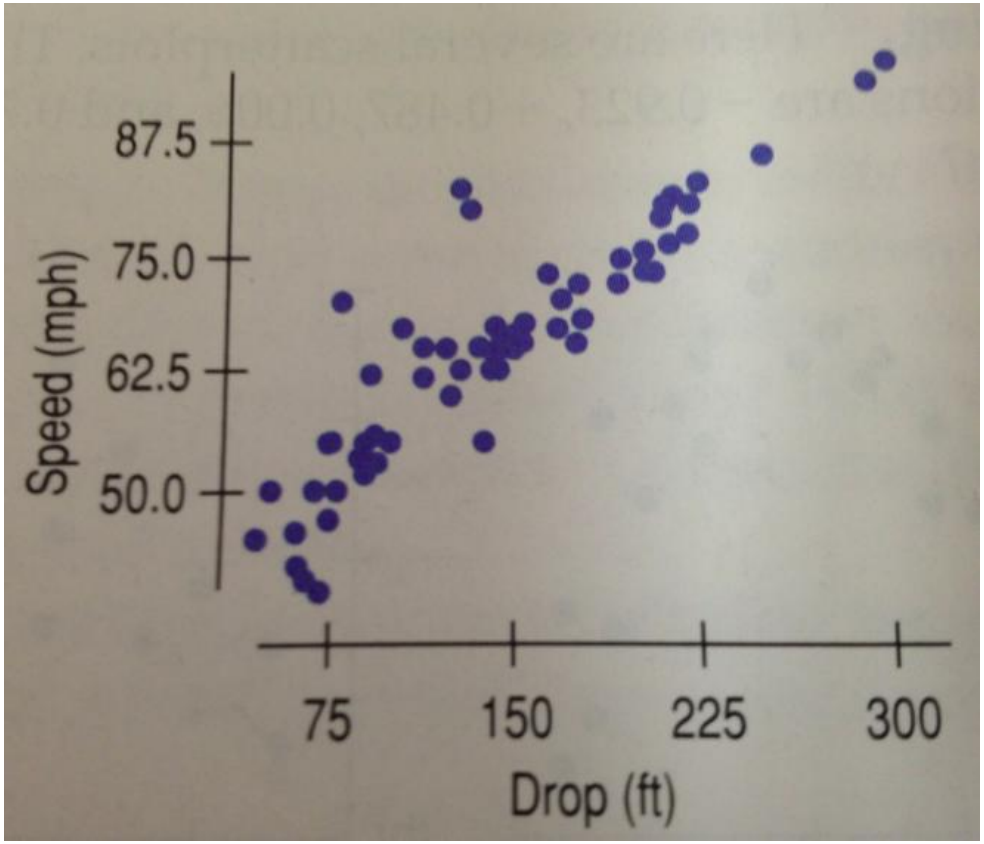
- We will discuss following problems:
  - Chapter 7 "*Scatterplots, Association, and Correlation*" (Page 188): #15, 16, 27, 32
  - Chapter 8 "*Liner Regression*" (Page 216): #11, 28

- All recitation PowerPoint slides available at here

- Chapter 7 (Page 188): #15:

  Scatterplot of top speed and

  largest drop for 75 roller coasters.



  - Appropriate to calculate the correlation? Explain.
  - Correlation = 0.91. Describe the association.

- Chapter 7 (Page 188): #15 (continued) :

  Scatterplot of top speed and

  largest drop for 75 roller coasters.

  - Appropriate to calculate the correlation? Explain.

  Ans.: Yes. It shows a linear form and no outliers.

  - Correlation = 0.91. Describe the association.

  Ans.: There is a strong, positive, linear association between drop and speed; the greater the coaster's initial drop, the higher the top speed.

  Tips: Association: Direction (positive? negative?), Form (Straight?), and Strength (strong? little?)

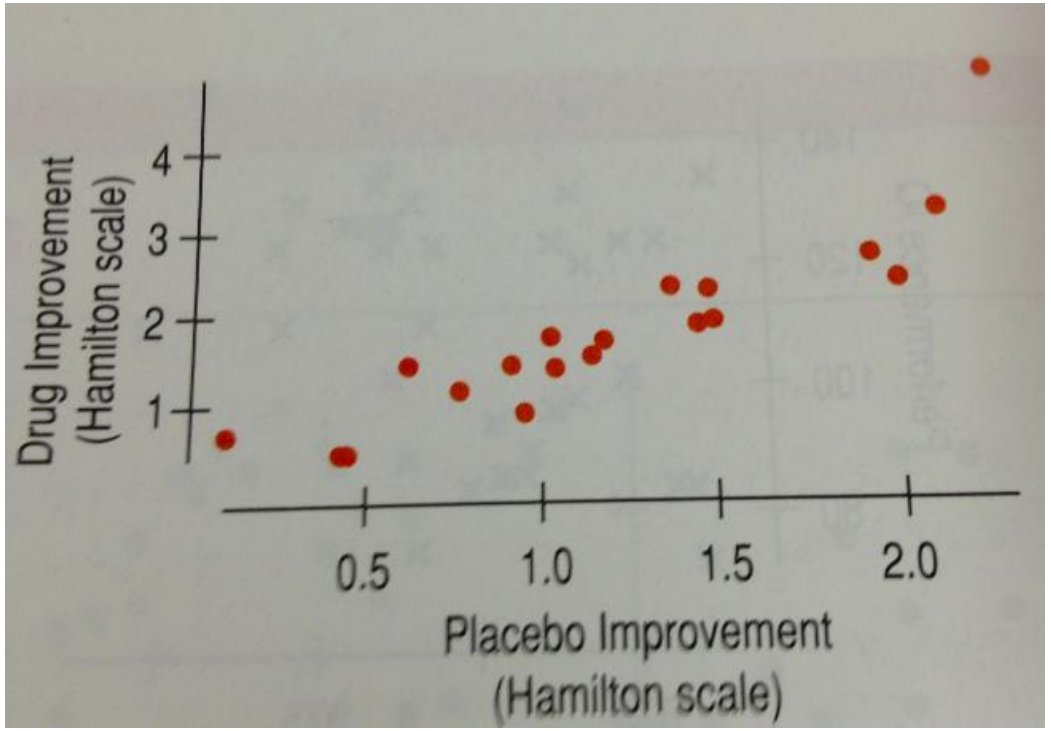
- Chapter 7 (Page 188): #16:

  Scatterplot comparing mean improvement levels for the antidepressants and placebos. Patient's depression levels were evaluated on the Hamilton scale, where larger numbers indicate greater improvements.

  - Appropriate to calculate the correlation? Explain.
  - Correlation

  = 0.898.

  Conclusions?



- Chapter 7 (Page 188): #16 (continued) :

  Hamilton Rating Scales for Depression (Wiki)

  "The **Hamilton Rating Scale for Depression** (**HRSD**), also known as the **Hamilton Depression Rating Scale** (**HDRS**) or abbreviated to **HAM-D**, is a multiple choice questionnaire that clinicians may use to rate the severity of a patient's major depression.[1] ……, The questionnaire, which is designed for adult patients and is in the public domain, rates the severity of symptoms observed in depression such as low mood, insomnia, agitation, anxiety and weight loss. ……, *A score of 0-7 is considered to be normal*, scores of 20 or higher indicate moderately severe depression and are usually required for entry into a clinical trial."

- Chapter 7 (Page 188): #16 (continued) :

  Scatterplot comparing mean improvement levels for the antidepressants and placebos.

  - Appropriate to calculate the correlation? Explain.

  Ans.: No, no units for the Hamilton Depression Rating Scale are given. These variables are not truly quantitative.

  Hints: any other reasons? E.g.: any outliers?

  - Correlation = 0.898. Conclusions?

  Ans.: Nothing. Correlation is not appropriate.

- Summary: Correlation Conditions (Page 173)
  - Quantitative Variables Condition
  - Straight Enough Condition
  - Outlier Condition

- Chapter 7 (Page 189): #27:

  Correlation between age and income r = 0.75 from 100 people.

  Justify:

  - When age increases, income increases as well.
  - The form of relationship between age and income is straight.
  - There are no outliers in the scatterplot of income vs. age.
  - Whether we measure age in years or months, the correlation will still be 0.75.

- Chapter 7 (Page 189): #27 (continued)

  Correlation between age and income r = 0.75 from 100 people.

  Justify:

  - When age increases, income increases as well.

  Ans.: No. Possible nonlinear relationship or outliers.

  - The form of relationship between age and income is straight.

  Ans.: No. We can't tell from the correlation coefficients alone.

  - There are no outliers in the scatterplot of income vs. age.

  Ans.: No. We can't tell from the correlation coefficients alone.

  - Whether we measure age in years or months, the correlation will still be 0.75.

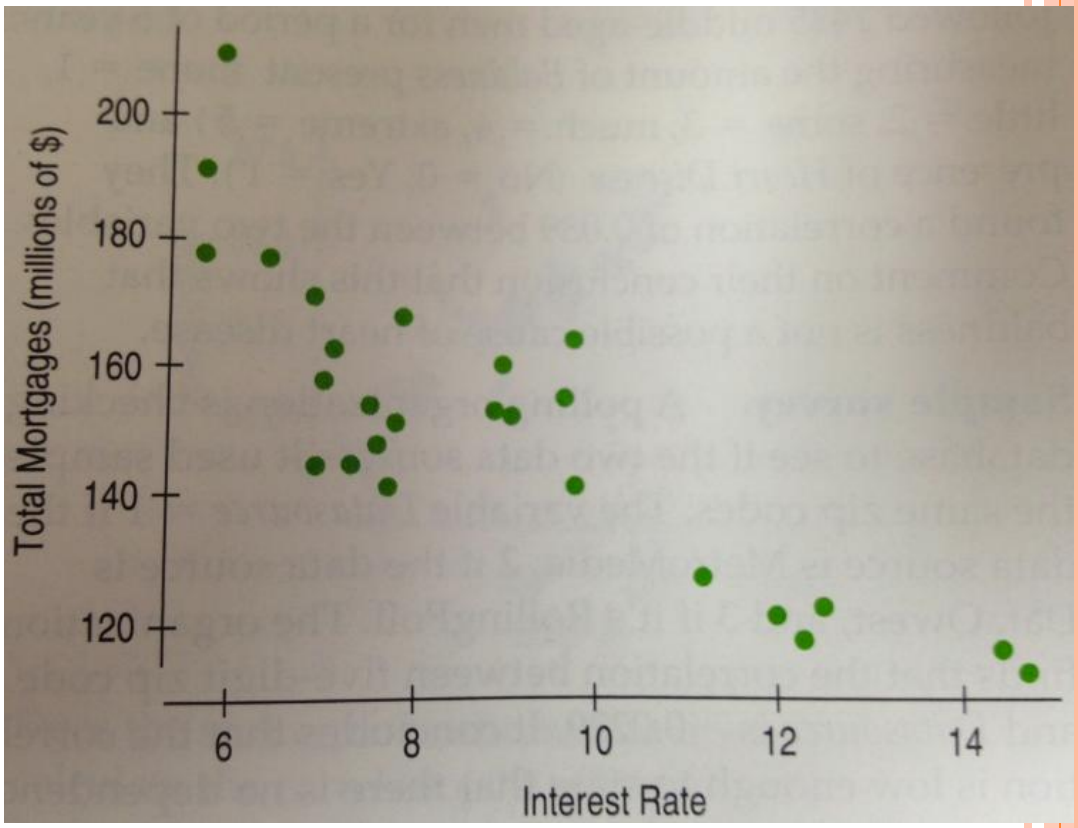
  Ans.: Yes. Correlation coefficients does not depends on the units.

Tips: $r = \frac{\sum z_x z_y}{n-1}$ **Pearson Correlation Coefficients, location, scale invariant, however sensitive to outliers.**

- Chapter 7 (Page 189): #32

  Scatterplot of total mortgages (T.M) vs. interest rate (I.R.). Corr. = -0.84.

  - Describe the relationship.
  - What if we standardize both variables?
  - What if we measure mortgages in thousands of dollars?
  - In another year, I.R.=11%, T.M.=$250 million, how Corr. Changes if add this year?
  - Rates lowered => more mortgages? Explain.



- Chapter 7 (Page 189): #32 (continued) :

Scatterplot of total mortgages (T.M) vs. interest rate (I.R.). Corr. = -0.84.

- Describe the relationship.

Ans.: The association is negative, quite strong, fairly straight, no outliers.

- What if we standardize both variables? *Ans.: No change.*
- What if we measure mortgages in thousands of dollars? *Ans.: No change.*
- In another year, I.R.=11%, T.M.=$250 million, how Corr. Changes if add this year? *Ans.: Weaken the correlation, closer to zero.*
- Rates lowered => more mortgages? Explain.

Ans.: No. We can only say that lower interest rates are associated with larger mortgage amounts, but we don't know why/ There may be other economic variables at work. i.e., the relationship may not be ***causal****.*

(Correlation can not imply Causality, there might be lurking variables.)

- Chapter 8 (Page 216): #11:

Regression equations. Fill in the missing information:

	$\bar{X}$	S_X	$\bar{y}$	S_y	r	$\hat{y} = b_0 + b_1 x$
a)	10	2	20	3	-0.5	
b)	2	0.06	7.2	1.2	-0.4	
c)	12	6			-0.8	$\hat{y} = 100 - 4x$
d)	2.5	1.2		100		$\hat{y} = -100 + 50x$

- Chapter 8 (Page 216): #11 (continued) :

Regression equations. Fill in the missing information:

	$\bar{x}$	S_x	$\bar{y}$	S_y	r	$\hat{y} = b_0 + b_1 x$
a)	10	2	20	3	0.5	$\hat{y} = 12.5 + 0.75x$
b)	2	0.06	7.2	1.2	-0.4	$\hat{y} = 23.2 - 8x$
c)	12	6	**152**	**30**	-0.8	$\hat{y} = 200 - 4x$
d)	2.5	1.2	**25**	100	**0.6**	$\hat{y} = -100 + 50x$

- *Answer: use the formulae:*

$$b_1 = r\frac{S_y}{S_x}$$

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

- Chapter 8 (Page 216): #11 (continued) :

the formulae:

$$b_1 = r\frac{S_y}{S_x}$$

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

From them you can also calculate any quantities given the rest, for example:

$$S_x = \frac{r\,S_y}{b_1}, \quad S_y = \frac{b_1\,S_x}{r}, \quad r = \frac{b_1\,S_x}{S_y},$$

$$\bar{x} = \frac{\bar{y} - b_0}{b_1}, \quad b_1 = \frac{\bar{y} - b_0}{\bar{x}}.$$

- Flexibly use the formula.
- **Never forget the signs! Particularly the sign of b_1.**

- Chapter 8 (Page 217): #28:

  Regression model for roller coasters:

$$\widehat{Duration} = 91.033 + 0.242\ Drop$$

  - Explain what the slope of the line says about how long a roller coaster ride may last and the height of the coaster.
  - A new roller coaster with drop = 200, predict rides last?
  - Another coaster with drop = 150, ride = 2 minutes. Longer or shorter than you'd expect? By how much? What's that called?

- Chapter 8 (Page 217): #28 (continued) :

Regression model for roller coasters:

$$\widehat{Duration} = 91.033 + 0.242\ Drop$$

- Explain what the slope of the line says about how long a roller coaster ride may last and the height of the coaster.

Ans.: On average, rides last about 0.242 seconds longer per foot of initial drop. (i.e., on average, drop increase by 1 foot, Duration will last about 0.242 seconds longer!)

- A new roller coaster with drop = 200, predict rides last?

Ans.: 91.033 + 0.242**200 = 139.433 seconds.*

- Another coaster with drop = 150, ride = 2 minutes. Longer or shorter than you'd expect? By how much? What's that called?

Ans.: 91.033 + 0.242**150 = 127.333 seconds > 2 minutes by 7.333 seconds*

Negative Residual. (Recall: Residual = Yobserved – Ypredict)

So Ypredict – Yobserved *should be "Negative residual".*

Thank you.