Checklist

1. Would you help me? If yes, unmute yourself, and let me know.
    1. Be my co-host, and admit people from the waiting room.
    2. If there is a question in the chat, please remind me.
    3. If you notice somebody raises their hand during lecture, please remind me.
2. Type your questions, if any, in the chat now.
3. I will set an alarm for 1hr at 12:45pm.
4. I will record to the cloud.

Unit 3: Basic regression

3. Linear regression condition

Stat 140 - 02

Mount Holyoke College

Fall 2020

Dr. Shan Shan

Slides posted at *http://sshanshans.github.io/stat140*

Outline

1. Announcements

2. Today: Linear regression condition

3. Main ideas
    1. Conditions for linear regression

- EA 02 is due tomorrow at 6pm EST
  - Group leader: submit your pdf on moodle
  - Everyone: fill out the peer evaluation form `https://forms.gle/HPXHtPru981vAWE16`
- Office hours:
  - Shan Shan today 2pm -3pm
  - Alex Friday 10am - 11am
- Post your team information for EA 03 on Piazza

Outline

1. Announcements

2. Today: Linear regression condition

3. Main ideas
    1. Conditions for linear regression

Tutorial exercise: 10 minutes

Let’s take a look at Exercise 10: I. Anscombe’s data

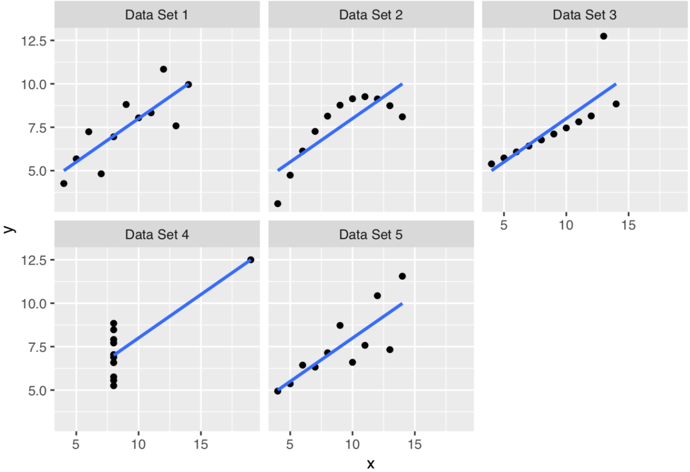
Anscombe's data

```
x1, y1     x2, y2     x3, y3     x4, y4     x5, y5
b0 = 3     b0 = 3     b0 = 3     b0 = 3     b0 = 3
b1 = 0.5   b1 = 0.5   b1 = 0.5   b1 = 0.5   b1 = 0.5
R2 = 67%   R2 = 67%   R2 = 67%   R2 = 67%   R2 = 67%
```

All 5 have essentially the same estimated intercept, slope, R^2!

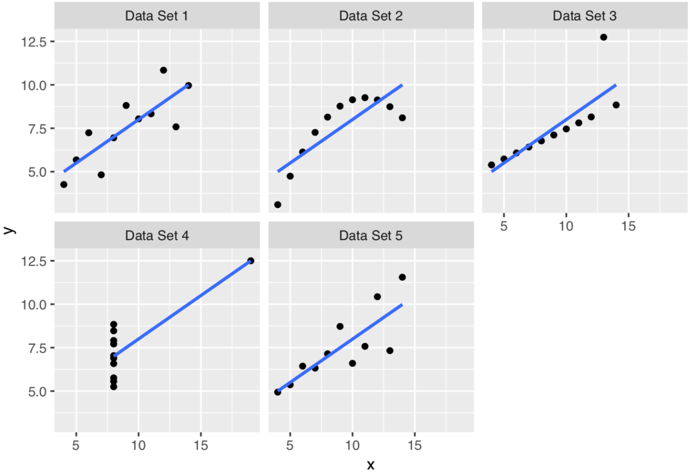
That means the five data sets should be pretty much the same right?

The scatterplots tell a different story.



Words of caution: always plot your data!

Is a linear model useful here?



Outline

1. Announcements

2. Today: Linear regression condition

3. Main ideas
    1. Conditions for linear regression

Outline

1. Announcements

2. Today: Linear regression condition

3. Main ideas
    1. Conditions for linear regression

The particular linear equation

$$\hat{y} = b_0 + b_1 x$$

that satisfies the least squares criterion is called the **least squares regression line**.

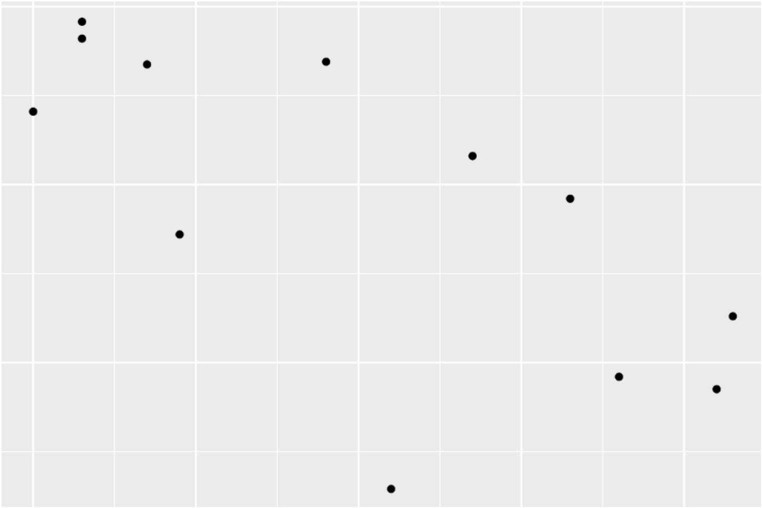
Casually, we often just call it the linear regression line,

Checking conditions

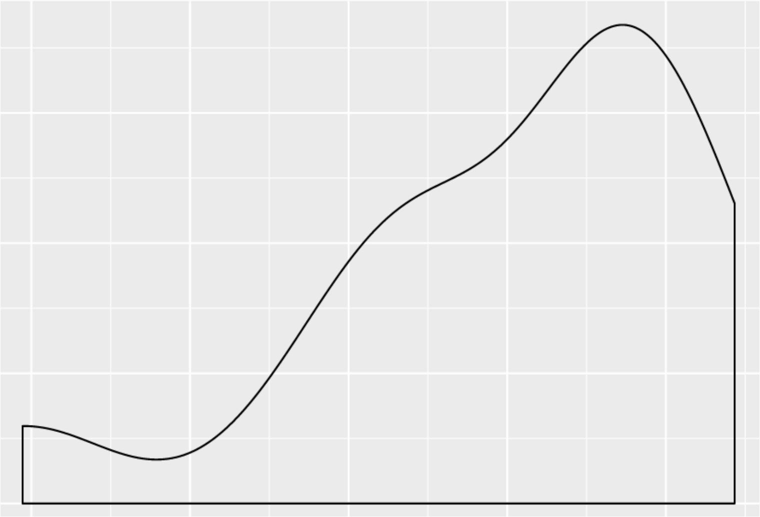
Be sure to check the conditions for linear regression before reporting or interpreting a linear model.

- From the scatterplot of y against x, check the
  - **Straight Enough Condition** Is the relationship between y and x straight enough to proceed with a linear regression model?
  - **Outlier Condition** Are there any outliers that might dramatically influence the fit of the least squares line?
  - **Does the Plot Thicken? Condition** Does the spread of the data around the generally straight relationship seem to be consistent for all values of x?

Does the plot thicken?



Does the plot thicken?



Checking conditions

Be sure to check the conditions for regression before reporting or interpreting a regression model.

- From the scatterplot of y against x, check the
  - **Straight Enough Condition** Is the relationship between y and x straight enough to proceed with a linear regression model?
  - **Outlier Condition** Are there any outliers that might dramatically influence the fit of the least squares line?
  - **Does the Plot Thicken? Condition** Does the spread of the data around the generally straight relationship seem to be consistent for all values of x?
- From the density of the residuals, check
  - **Symmetric around 0 Condition** Is the density plot of the residuals symmetric around 0.

Tutorial exercise: For the rest of the class

Let's work on Exercise 10 Wildfire.

- Part I: let's work together
- Part II: on your own

Post your answer to q3-6 of Part II on Piazza.
No need to submit anything on moodle.

Is a linear model appropriate for these data?

The relationship is straight enough. There does not seem to be any obvious outliers that will dramatically influence the fit of the least squares line. The residuals seem to have equal variability around the least squares line (The plot doesn't thicken). Stop.

Is a linear model appropriate for these data?

The relationship is straight enough. There does not seem to be any obvious outliers that will dramatically influence the fit of the least squares line. The residuals seem to have equal variability around the least squares line (The plot doesn't thicken). Stop.

The density plot of the residuals is symmetric around 0. So a linear model is indeed appropriate.

Interpret the slope in this context.

We expect the number of wildfires has been increasing by about 78 per year.

Is a linear model appropriate for these data?

The relationship is straight enough. There does not seem to be any obvious outliers that will dramatically influence the fit of the least squares line. The residuals seem to have equal variability around the least squares line (The plot doesn't thicken). Stop.

The density plot of the residuals is symmetric around 0. So a linear model is indeed appropriate.

Interpret the slope in this context.

We expect the number of wildfires has been increasing by about 78 per year.

Can we interpret the intercept? Why or why not?

Yes, the intercept estimates the number of wildfires in 1985 as about 75,609.

What does R^2 mean in this context?

Less than 1% of the variation in the number of wildfires can be accounted for by the linear model on Year. There is very little association between these variables — that is, that there has been little change in the number of wildfires during this period. The average number might provide as good a prediction as the model.