Class 4: Two Variables Sections 2.3, 2.4, 2.5, 2.6 Math 263, Section 5, Deborah Hughes Hallett

1. TYPES OF VARIABLE: QUANTITATIVE AND CATEGORICAL VARIABLES

Quantitative:
Categorical:

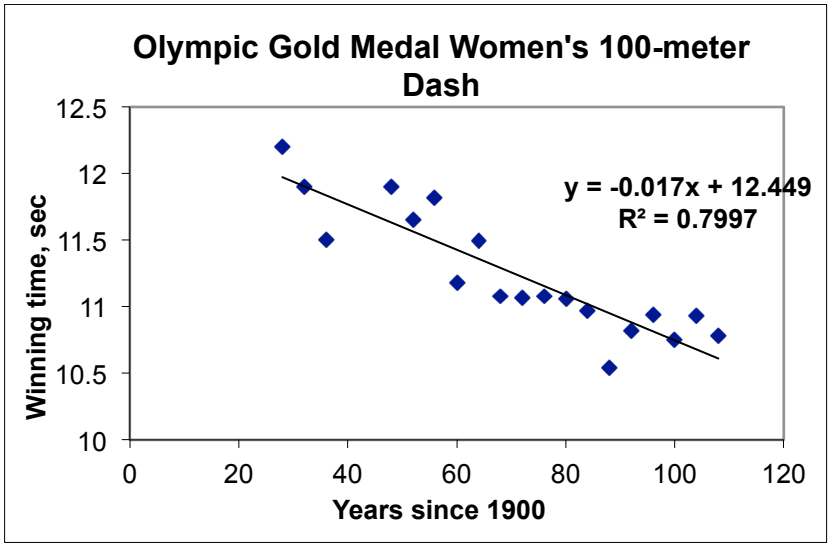
To analyze the association between two variables:

- **Quantitative variables**: Scatterplot. Correlation coefficient, and regression line
- **Categorical variables**: Two way table. Marginal and conditional distributions

Note: We can find correlation coefficients and regression lines *only* for quantitative variables.

2. TWO QUANTITATIVE VARIABLES: REGRESSION LINE (Ordinary Least Squares, OLS, Line)

Example: Interpret the coefficients in the regression line. What are their units?

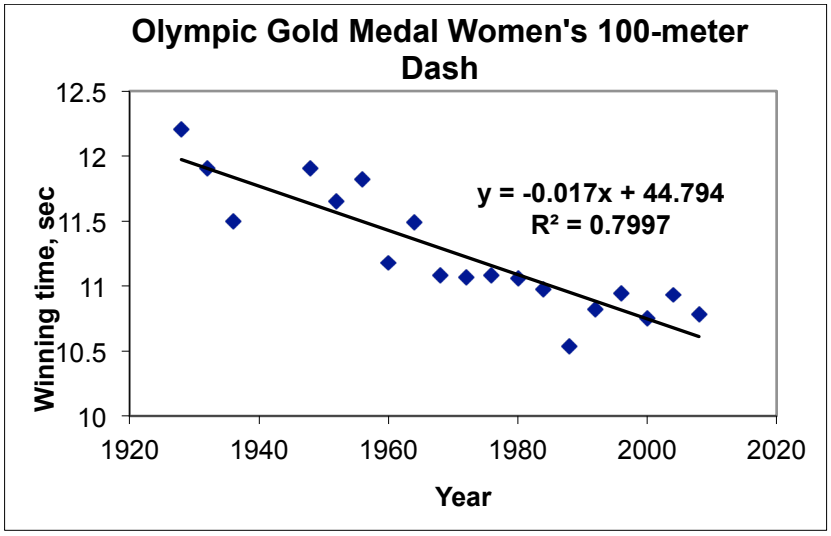


Units:

Slope:

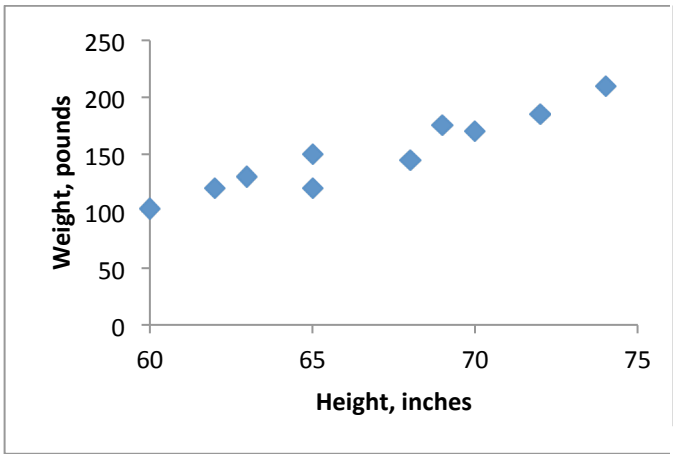
Constant:

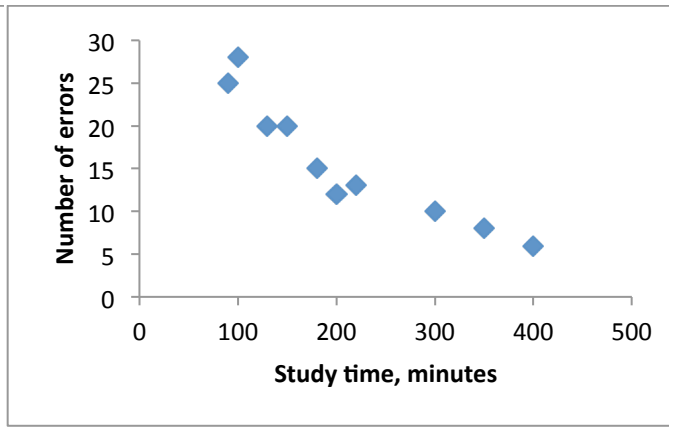
Example: How is the following graph similar to the previous one and how is it different? Why?



Example: Derive the exact relationship between the constants in the lines $y = -0.017x + 12.449$ and $y = -0.017x + 44.794$

Example: Estimate the regression equations by eye for the following two data sets.[1]





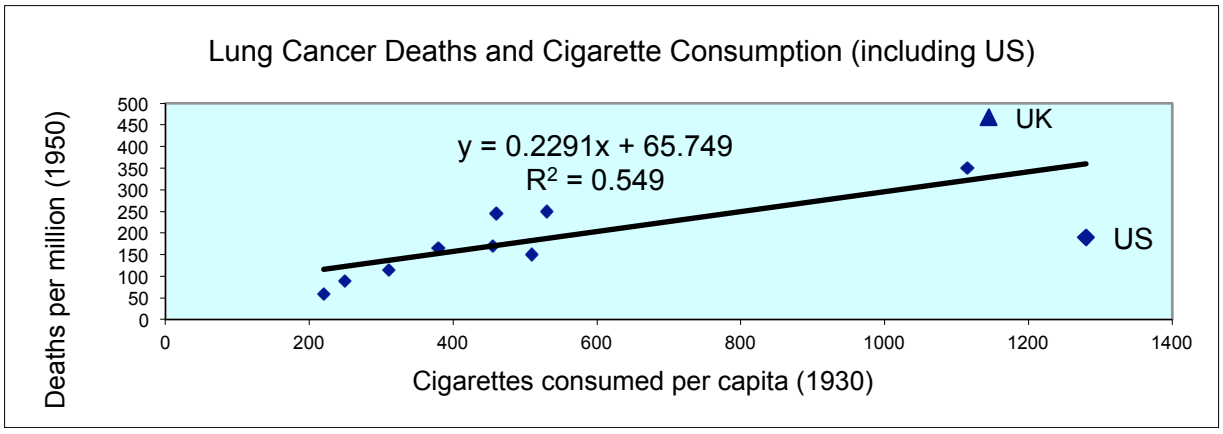
[1] http://luna.cas.usf.edu/~mbrannic/files/resmeth/lecture/corr_n_reg.html

3. EFFECT OF OUTLIERS ON THE REGRESSION LINE AND THE CORRELATION COEFFICIENT. ARE THEY ROBUST OR SENSITIVE?

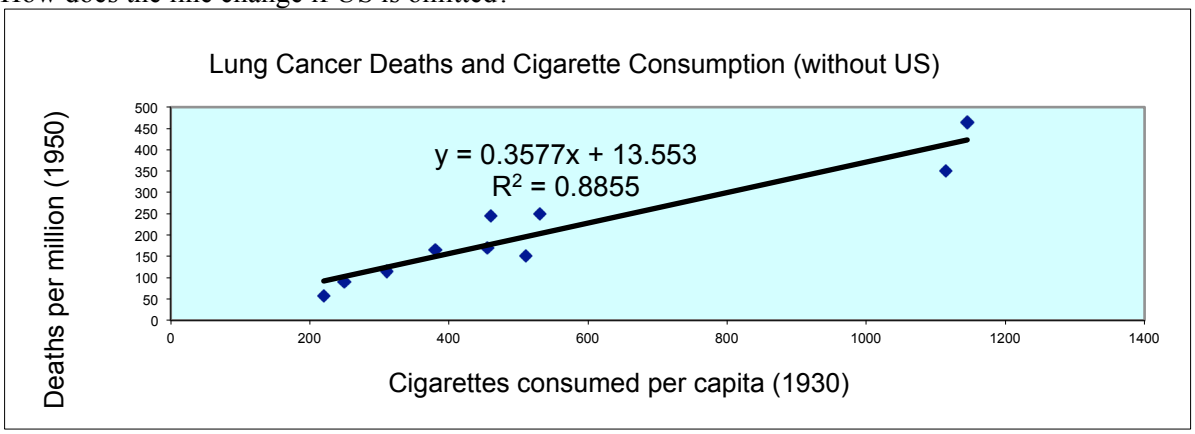
Example: Smoking and Lung Cancer, lagged by 20 years. [2]

Country	Cigarettes Consumed per capita (1930)	Lung Cancer deaths per million (1950)
Iceland	220	58
Norway	250	90
Sweden	310	115
Denmark	380	165
Australia	455	170
Holland	460	245
Canada	510	150
Switzerland	530	250
Finland	1115	350
Great Britain	1145	465
United States	1280	190

The US (heavy diamond below) is an outlier; possibly also UK (triangle). Why might they be so different?



How does the line change if US is omitted?



[2] From: *Data Analysis for Politics and Policy*, Edward Tufte, Prentice Hall, 1974

4. TWO CATEGORICAL VARIABLES: TWO WAY TABLES

Two hospitals perform the same surgical procedure and have the survival data shown in the two-way table:

Counts	Hospital A	Hospital B	
Died	63	17	
Survived	1975	875	

Ex: What does the 17 tell us?
What does the $63 + 17 = 80$ tell us?
What does the 2038 tell us?
How many patients were in the study?

Ex: Find the **joint distribution** that shows the proportions of the overall total (2924) in each cell:

Joint distribution	Hospital A	Hospital B	Total
Died			
Survived			
Total			

Ex: Interpret the **marginal distributions** in the edges of the table. The numbers in the bottom row show the proportions of patients in each of the hospitals (ignoring the outcome) and those in the right column show the proportions who died and survived (ignoring the hospital).

Is there an association between dying and the hospital in which the operation was done?
To decide, look at the **conditional distribution** of death and survival rates in each hospital separately. For each hospital, we look at what proportion died and survived of the patients in that hospital.

	Hospital A: *Conditional Distribution*	Hospital B: *Conditional Distribution*
Died		
Survived		
Total		

We can say:
- There *is* an association between death rate and hospital—Hospital B has a lower death rate.
- We do not know whether this difference is *significant*, or whether it could be just be result of random variation. (We will know how to determine this later in the semester.)
- We do not know what *causes* the difference. It could be that Hospital B is better than Hospital A; it could be there are *confounding variable*s.
- What are possible confounding variables

Suppose that the patients about to undergo this surgical procedure can be catalogued as having a "Good" or "Poor" prognosis and that the breakdown between hospitals is as follows:

	Good Prognosis	
	Hospital A	Hospital B
Died	8	8
Survived	821	689
Total	829	697

Poor Prognosis	
Hospital A	Hospital B
55	9
1154	186
1209	195

Ex: Find the conditional distribution for each prognosis and interpret it.

	Good Prognosis	
	Hospital A	Hospital B
Died		
Survived		
Total		

Poor Prognosis	
Hospital A	Hospital B

Ex: What can you conclude about the comparisons between Hospital A and Hospital B?
Hospital A has a much larger proportion of patients with a poor prognosis and who have a higher death rate. This means that overall, Hospital A has a higher death rate—even though in each category separately, Hospital A has a lower death rate.

Simpson's Paradox
Notice that Hospital A is better in both categories (Good and Poor prognosis), yet worse overall.