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: Income, weight, score on test, rainfall, life expectancy, blood sugar, temperature
Categorical: Gender, race, religion, college graduate, science major.

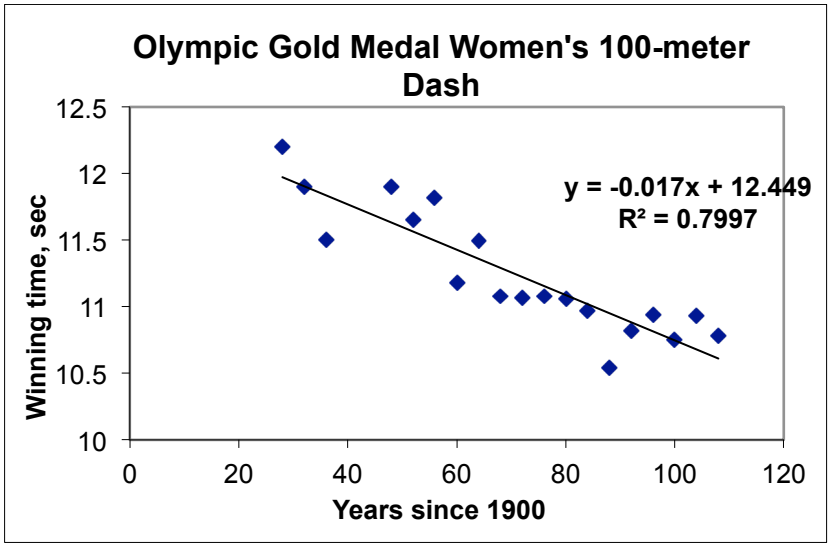
To analyze the association between two variables:

- **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?

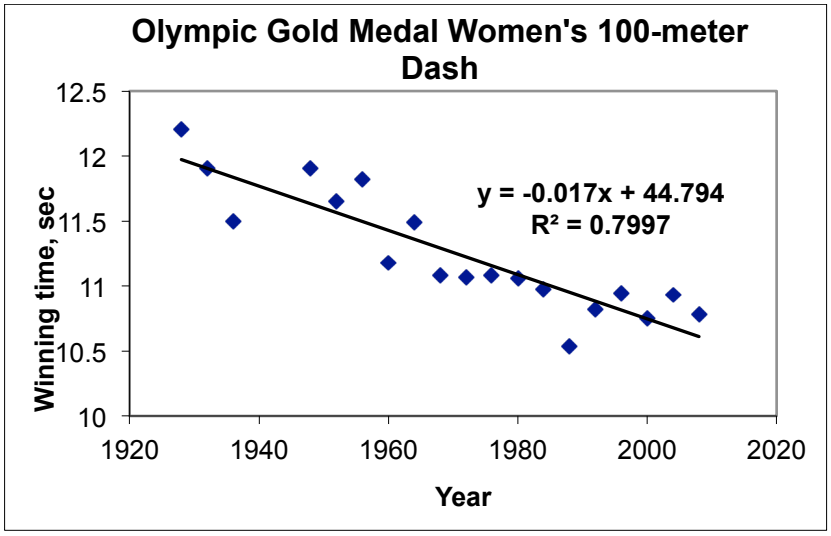


Units: Here y is in seconds and x is in years.

Slope: Negative because the time taken for the 100-meter dash decreases as runners get faster. The time decreases on average by 0.017, or about 0.02, of a second per year. That is a bit less than a tenth of a second per Olympic.

Constant: The 12.449 seconds represents what the winning time would have been in 1900 if the trend were extrapolated back that far. It is highly unlikely that this is a realistic estimate, because there is no reason the trend would extend backward.

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



The data is the same as before except that the horizontal axis is now year, The slope and the correlation coefficient are the same as in the previous case, because the line slopes at the same rates and is equally "scrunched" around the line.

The constant is different because the vertical axis is in a different place.

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

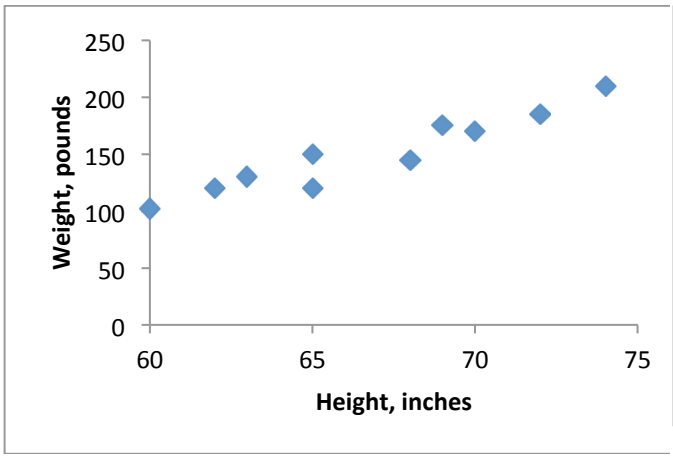
The x in the two equations are different. Let's call the second one t, where $t = 1900 + x$. Then the second equation is

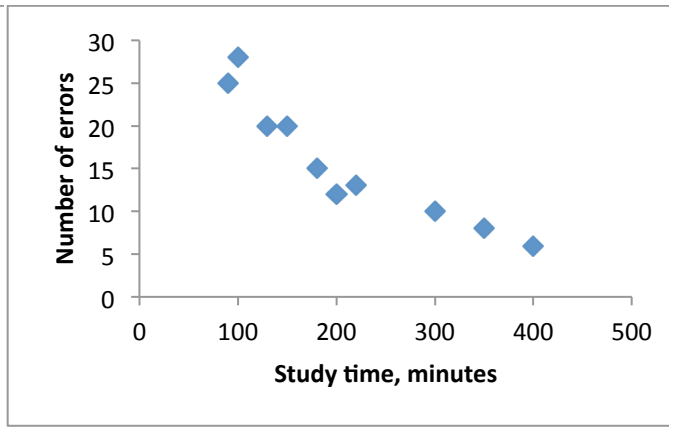
$$y = -0.017t + 44.794 = -0.017(1900 + x) + 44.794$$

$$y = -0.017(1900) - 0.017x + 44.794 = -0.017x + 44.794 - 32.3$$

$$y = -0.017x + 12.494.$$

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





The regression equation has slope approximately

$$\text{Slope} = \frac{200 - 100}{75 - 60} \approx 7.$$

The equation is therefore $y = 7x + b$ and substituting $x = 60$ and $y = 100$ gives $100 = 7 \cdot 60 + b$ so $b = -320$. The equation is

$$y = 7x - 320.$$

The regression equation has slope approximately

$$\text{Slope} = \frac{25 - 5}{100 - 400} = 0.07.$$

In this case we can read the vertical intercept off the graph as the vertical axis is shown. The intercept is about 30, so the equation is

$$y = -0.07x + 30.$$

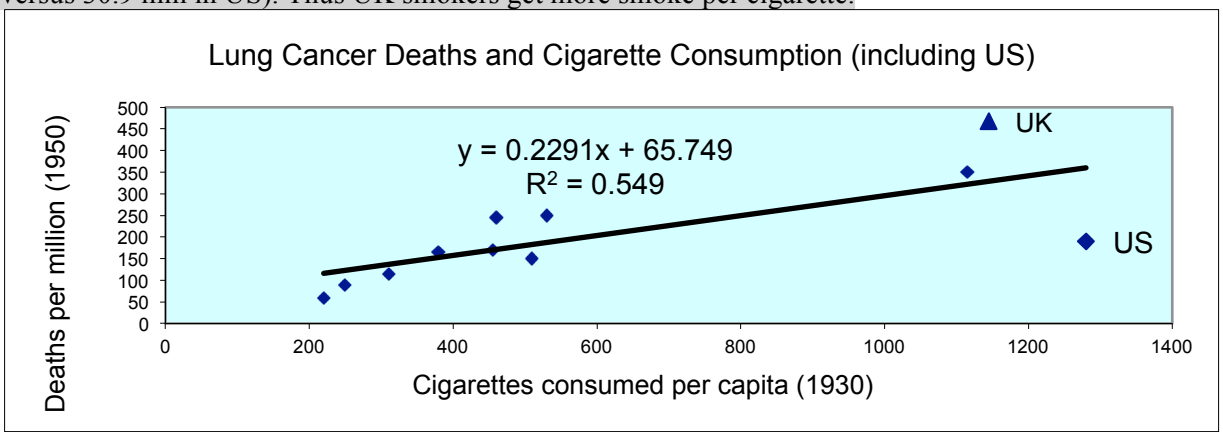
[1] http://luna.cas.usf.edu/~mbrannic/files/resmeth/lecture/corr_n_reg.html

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

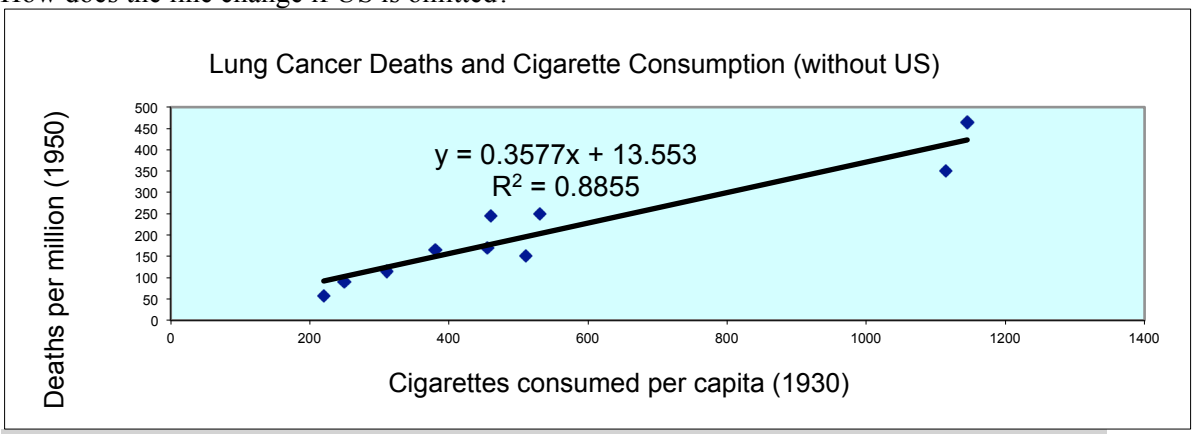
Example: Smoking and Lung Cancer, lagged by 20 years. [2]

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? In UK, cigarettes are heavily taxed, so they were smoked down to the butt (leaving 18.7 mm butts in UK versus 30.9 mm in US). Thus UK smokers get more smoke per cigarette.



How does the line change if US is omitted?



Leaving out the US makes the regression line steeper and increases the correlation coefficient. The line and the correlation are *sensitive* to outliers.

[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	Total
Died	63	17	80
Survived	1975	875	2850
Total	2038	892	2930

Ex: What does the 17 tell us? There were 17 patients in Hospital B who died.
What does the $63 + 17 = 80$ tell us? There were 80 patients who died overall.
What does the 2038 tell us? There were 2038 patients in Hospital A.
How many patients were in the study? 2930

Ex: Find the **joint distribution** that shows the proportions of the overall total (2924) in each cell:
For example $63/2930 = 0.022 = 2.2\%$ tells us that 2.2% of the patients were in Hospital A and died.

Joint distribution	Hospital A	Hospital B	Total
Died	0.022	0.006	0.027
Survived	0.674	0.299	0.973
Total	0.696	0.304	1

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).
There are two marginal distributions:
Marginal Distribution for Hospitals: Bottom row: 69.6% of the operations were in Hospital A, while 30.4% were in Hospital B.
Marginal Distribution for Deaths: Right column: 2.7% of the patients died, while 97.3% survived.

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.
Thus, for hospital A, we have $63/2038 = 0.031$, and so on.
Notice that the calculation can also be done with proportions: $0.022/0.696 = 0.031$

	Hospital A: *Conditional Distribution*	Hospital B: *Conditional Distribution*
Died	0.031	0.019
Survived	0.969	0.981
Total	1.000	1.000

Thus, the death rate in Hospital A was 3.1% while in Hospital B it was 1.9%—much lower.

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	0.010	0.011
Survived	0.990	0.989
Total	1	1

Poor Prognosis	
Hospital A	Hospital B
0.045	0.046
0.955	0.954
1	1

For patients with a good prognosis; Hospital A has a 1.0% death rate, where as Hospital B has a 1.1% death rate. For patients with a poor prognosis; Hospital A has a 4.5% death rate, where as Hospital B has a 4.6% death rate.

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.
Why? Because Hospital A has a larger proportion of patients with a poor prognosis

$$\text{Proportion Poor Prognosis in Hospital A} = \frac{1209}{829 + 1209} = 0.593 = 59.3\%.$$

$$\text{Proportion Poor Prognosis in Hospital B} = \frac{195}{697 + 195} = 0.219 = 21.9\%$$