Math 10B with Professor Stankova
Worksheet, Discussion #27; Thursday, 5/2/2019
GSI name: Roy Zhao

Linear Regression and Correlation

Concepts

1. We have
$$a = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x - \bar{x})^2}, b = \bar{y} - a\bar{x},$$
where $\bar{x} = \frac{1}{n}\sum_{i=1}^n x_i$ is the average of the x values and $\bar{y} = \frac{1}{n}\sum_{i=1}^n y_i$ is the average of the y values.

   The **correlation coefficient** of a set of points $\{(x_i, y_i)\}$ is given by
$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}\sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}.$$
Another way to represent that the correlation coefficient is the cosine of the angle between the two vectors $\vec{x} = (x_i - \bar{x})$ and $\vec{y} = (y_i - \bar{y})$. So, we can write
$$r = \frac{\vec{x} \circ \vec{y}}{|\vec{x}||\vec{y}|}.$$
It is always between -1 and 1 by Cauchy-Schwarz.

   Another way to write this is in terms of the sample covariance and sample standard deviation. They are defined as
$$cov(x, y) = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n}, \sigma_x = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n}}, \sigma_y = \sqrt{\frac{\sum (y_i - \bar{y})^2}{n}}.$$
Then another formula is
$$r = \frac{cov(x, y)}{\sigma_x \sigma_y}, a = r\frac{\sigma_y}{\sigma_x}.$$

Examples

2. Suppose you want to know whether performance on Quiz 1 is correlated with performance on Quiz 13. You randomly choose 5 students' quiz scores and get the following values.

Student	Quiz 1	Quiz 13
A	7	9
B	12	11
C	6	5
D	11	10
E	4	5

   Calculate the correlation coefficient r as well as the line of best fit.

Problems

3. True False The line of best fit always exists.

4. True False If you only have two data points with different x values, then the correlation coefficient r is either 1 or -1.

5. True False The correlation is always between -1 and 1 inclusive.

6. True False If the correlation between two sets of data is -1, then y is proportional to x^{-1}.

7. True False If we shift the data (by for instance adding 5 to all of the y values), then the correlation does not change.

8. True False For two random variables X, Y, we have $Cov(10X, 10Y) = Cov(X, Y)$.

9. Is there a relationship between the amount of antibody A and antibody B in a sick patient? You take antibody A and B counts per milliliter from 4 patients (in reality you will have a much, much larger sample size).

Patient	Antibody A	Antibody B
A	120	100
B	95	110
C	115	130
D	110	80

Calculate the correlation coefficient and line of best fit.

10. The formulas for the slope and y intercept of the line of best fit come from MLE. Suppose that error is normally distributed. This means that if we predict $y = ax_i + b$, then the probability of actually getting y_i follows the PDF

$$\frac{1}{\sigma\sqrt{2\pi}} e^{-(y_i - y)^2/2\sigma^2} = \frac{1}{\sigma\sqrt{2\pi}} e^{-(y_i - (ax_i + b))^2/2\sigma^2}.$$

Use MLE to show that $\hat{b} = \bar{y} - a\bar{x}$.

11. Now with $b = \bar{y} - a\bar{x}$, do MLE to show that $\hat{a} = r\frac{\sigma_y}{\sigma_x}$ the formula that we use for a.