**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.

**Solution:** First we need to find the sample standard deviation where $x$ is the Quiz 1 scores and $y$ is the Quiz 13 scores. In order to find the sample standard deviation, we first need to calculate the sample average ($\overline{x}$). In this case, we will get $\overline{x} = 8$. The average of the Quiz 13 scores also gives us $\overline{y} = 8$. To keep track of the calculation, the following chart is helpful.

| | $x_i - \overline{x}$ | $y_i - \overline{y}$ |
|---|---|---|
| A | -1 | 1 |
| B | 4 | 3 |
| C | -2 | -3 |
| D | 3 | 2 |
| E | -4 | -3 |

It is **very important** that you keep track of whether each entry is positive or negative (for calculating covariance). Then we can get $\sigma_x^2$ by averaging the squares of the first column.

$$\sigma_x = \sqrt{\frac{1}{5}\left[(-1)^2 + 4^2 + (-2)^2 + 3^2 + (-4)^2\right]} = \sqrt{\frac{46}{5}} \approx 3.03$$

A similar calculation will give us that $\sigma_y = \sqrt{\frac{32}{5}} \approx 2.53$.

To get the sample covaraince, we average the **product** of each row.

$$\operatorname{cov}(x, y) = \frac{1}{5}\left[(-1)\cdot 1 + 4 \cdot 3 + (-2)\cdot(-3) + 3 \cdot 2 + (-4)\cdot(-3)\right] = 7$$

Finally we can simply calculate the correlation coefficient with the formula:

$$r = \frac{\operatorname{cov}(x, y)}{\sigma_x \sigma_y} \approx \frac{7}{3.03 \cdot 2.53} \approx 0.913$$

This confirms what we thought with our initial picture. There is a fairly strong *positive* linear relationship between scores on Quiz 1 and scores on Quiz 13.

First we calculate the slope $a$ of the line of best fit.

$$a = r\frac{\sigma_y}{\sigma_x}.$$

For our example,

$$a \approx 0.913 \cdot \frac{2.53}{3.03} \approx 0.762$$

Finally, we can calculate the $y$-intercept in our line of best fit by noting that $\overline{y} = a\overline{x} + b$ and solving for $b$, since we know the other 3 variables. Here we get $8 = 0.762 \cdot 8 + b$, so $b \approx 1.901$. So our line of best fit is $y = 0.762x + 1.901$. So if someone get 9 on Quiz 1, we would guess a score of about 8.76 on Quiz 12. Our large $r$-value gives us a pretty high confidence that this is an accurate guess.

## Problems

3. True **FALSE** The line of best fit always exists.

**Solution:** If there is only one point or all the points have the same $x$ value, then the line of best fit will not exist.

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

**Solution:** There is always a line between two points with different $x$ values. But, if the line is horizontal the $r$ coefficient will not always exist.

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 $Cov(10X, 10Y) = Cov(X, Y)$.

**Solution:** We have $Cov(10X, 10Y) = 10 \cdot 10Cov(X, Y)$.

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.

**Solution:** Note that $\overline{x} = 110$ and $\overline{y} = 105$. Then we can make the small chart.

| | $x_i - \overline{x}$ | $y_i - \overline{y}$ |
|---|---|---|
| A | 10 | -5 |
| B | -15 | 5 |
| C | 5 | 25 |
| D | 0 | -25 |

So we can calculate:

$$\begin{aligned}
\sigma_x &= \sqrt{\frac{1}{4}(100 + 225 + 25 + 0)} = \sqrt{\frac{350}{4}} \approx 9.35 \\
\sigma_y &= \sqrt{\frac{1}{4}(25 + 25 + 625 + 625)} = \sqrt{\frac{1300}{4}} \approx 18.03 \\
\operatorname{cov}(x, y) &= \frac{1}{4}(-50 - 75 + 125 + 0) = \frac{0}{4} = 0 \\
r &= \frac{2.5}{9.35 \cdot 18.03} \approx 0.015
\end{aligned}$$

We can calculate the line of best fit to get $y = 105$. However, with an $r$ value 0, we should generally expect Antibody A and Antibody B to be not correlated, so we shouldn't use this line to try to make predictions.

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 $\hat{b} = \bar{y} - a\bar{x}$.

**Solution:** We calculate $L(\theta|(x_1, y_1), \ldots, (x_n, y_n))$ below

$$\begin{aligned}
L(\theta|(x_1, y_1), \ldots, (x_n, y_n)) &= P((x_1, y_1), \ldots, (x_n, y_n)|b = \theta) \\
&= \prod_i \frac{1}{\sigma\sqrt{2\pi}} e^{-(y_i - (ax_i + \theta))^2/2\sigma^2} \\
&= \frac{1}{\sigma^n \sqrt{2\pi}^n} e^{-\sum(y_i - (ax_i + \theta))^2/2\sigma^2}
\end{aligned}$$

Now taking the log gets rid of the exponent and taking the derivative and setting it equal to 0 gives

$$\begin{aligned} 0 &= -\sum \frac{\partial}{\partial \theta} \frac{(y_i - ax_i - \theta)^2}{2\sigma^2} \\ &= \sum \frac{2(y_i - ax_i - \theta)}{2\sigma^2} \end{aligned}$$

So $\sum(y_i - ax_i - \theta) = \sum(y_i - ax_i) - n\theta = 0$ and so

$$\theta = \hat{b} = \frac{1}{n}\sum(y_i - ax_i) = \bar{y} - a\bar{x}.$$

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 $a$.

**Solution:** We calculate $L(\theta|(x_1, y_1), \ldots, (x_n, y_n))$ below

$$\begin{aligned} L(\theta|(x_1, y_1), \ldots, (x_n, y_n)) &= P((x_1, y_1), \ldots, (x_n, y_n)|a = \theta) \\ &= \prod_i \frac{1}{\sigma\sqrt{2\pi}} e^{-(y_i - (\theta x_i + (\bar{y} - \theta\bar{x})))^2/2\sigma^2} \\ &= \frac{1}{\sigma^n \sqrt{2\pi}^n} e^{-\sum((y_i - \bar{y}) + \theta(\bar{x} - x_i))^2/2\sigma^2} \end{aligned}$$

Now taking the log gets rid of the exponent and taking the derivative and setting it equal to 0 gives

$$\begin{aligned} 0 &= -\sum \frac{\partial}{\partial \theta} \frac{((y_i - \bar{y}) + \theta(\bar{x} - x_i))^2}{2\sigma^2} \\ &= \sum \frac{2(\bar{x} - x_i)((y_i - \bar{y}) + \theta(\bar{x} - x_i))}{2\sigma^2} \end{aligned}$$

Simplifying gets the result.