Lecture 5: Correlation and Linear Regression

3.5. (Pearson) correlation coefficient

The correlation coefficient measures the strength of the linear relationship between two variables.

- The correlation is always between -1 and 1.
- Points that fall on a straight line with positive slope have a correlation of 1.
- Points that fall on a straight line with negative slope have a correlation of -1.
- Points that are not linearly related have a correlation of 0.
- The farther the correlation is from 0, the stronger the linear relationship.
- The correlation *does not change* if we change units of measurement.

See Figure 3 on page 105.

Given a bivariate data sat of size n,

$$(x_1, y_1),\ (x_2, y_2),\ \ldots,\ (x_n, y_n),$$

the sample covariance $s_{x,y}$ is defined by

$$s_{x,y} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}).$$

Note that if $x_i = y_i$ for all $i = 1, \ldots, n$, then $s_{x,y} = s_x^2$.

The sample correlation coefficient r is defined by

$$r = \frac{s_{x,y}}{s_x\, s_y},$$

where s_x is the sample standard deviation of $x_1, \ldots, x_n$, i.e.

$$s_x = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}}.$$

To simplify calculation, we often use the following alternative formula:

$$r = \frac{\mathcal{S}_{x,y}}{\sqrt{\mathcal{S}_{x,x}}\,\sqrt{\mathcal{S}_{y,y}}},$$

where

$$\mathcal{S}_{x,y} = \sum_{i=1}^{n} x_i y_i - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n},$$

$$\mathcal{S}_{x,x} = \sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}$$

and

$$\mathcal{S}_{y,y} = \sum_{i=1}^{n} y_i^2 - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}.$$

Example: See page 107.

Causation; Lurking variables

Go to an elementary school and measure two variables for each child: Shoe size and Reading Level.

- You will find a positive correlation; as shoe size increases, reading level tends to increase.
- Should we buy our children bigger shoes?
  - No, the two variables both are positively associated with Age.
  - Age is called a **lurking variable**.

Remember: An observed correlation between two variables may be *spurious*. That is, the correlation may be caused by the influence of a *lurking variable*.

3.6. Prediction: Linear Regression

Objective: Assume two variables x and y are related: when x changes, the value of y also changes. Given a data set

$$(x_1, y_1),\ (x_2, y_2),\ \ldots,\ (x_n, y_n)$$

and a value x_{n+1}, can we predict the value of y_{n+1}.

In this context, x is called the *input variable* or predictor, and y is called the *output variable* or response.

Examples:

- Having known the price change history of IBM stock, can we predict its price for tomorrow?
- Based on your first quiz, predict you final score.
- Survey consumers' need for certain product, make a recommendation for the number of items to be produced.

Method: Linear regression (fitting a straight line to the data).

Question: Why do we only consider *linear* relationships? (Remember that correlation measures the strength and direction of the linear association between variables.)

- Linear relationships are easy to understand and analyze.
- Linear relationships are common.
- Variables with nonlinear relationships can sometimes be transformed so that the relationships are linear. (See Lab 4 for an example.)
- Nonlinear relationships can sometimes be closely approximated by linear relationships.

Recall: A straight line is determined by two constants: its intercept and slope. In its equation

$$y = \beta_1 x + \beta_0,$$

β_0 is the intercept of this line with the y-axis and β_1 represents the slope of the line.

Finding the "best-fitting" line

- **Idea:** Draw a line that seems to fit well and then find its equation.
- **Problems:**

- Different people will come up with different "best" lines. How do we pick the best?
- It's very hard for large datasets.
- It doesn't generalize to relationships between more than two variables.

- For these and other reasons, we look for the "least squares" line.

- The least squares line minimizes the sum of squared deviations from the data.

Steps for finding the regression line:

i . Plotting a scatter diagram to see whether a linear relation exists. If it does, go to the next step.

ii . Using the data to estimate β_0 and β_1. This can be done by using the least square method:

$$\text{Slope } \widehat{\beta}_1 = \frac{\mathcal{S}_{x,y}}{\mathcal{S}_{x,x}}$$
$$\text{Intercept } \widehat{\beta}_0 = \overline{y} - \widehat{\beta}_1 \overline{x}.$$

iii . The fitted regression line is

$$\widehat{y} = \widehat{\beta}_1 x + \widehat{\beta}_0.$$

Predicted values For a given value of the x-variable, we compute the predicted value by plugging the value into the least squares line equation.

Example 7. See page 117.

Example 8. Exercise 3.44.