Lecture 12: Correlation and Introduction to Regression

SW, Chapter 12; JMP Start Statistics Chapters 8 and 13

Suppose we select \( n = 10 \) persons from the population of college seniors who plan to take the MCAT exam. Each takes the test, is coached, and then retakes the exam. Let \( X_i \) be the pre-coaching score and let \( Y_i \) be the post-coaching score for the \( i^{th} \) individual, \( i = 1, 2, \cdots , n \). There are several questions of potential interest here, for example: Are \( Y \) and \( X \) related (associated), and how? Does coaching improve your MCAT score? Can we use the data to develop a mathematical model (formula) for predicting post-coaching scores from the pre-coaching scores? These questions can be addressed using correlation and regression models.

The correlation coefficient is a standard measure of association or relationship between two features \( Y \) and \( X \). Most scientists equate \( Y \) and \( X \) being correlated to mean that \( Y \) and \( X \) are associated, related, or dependent upon each other. However, correlation is only a measure of the strength of a linear relationship. For later reference, let \( \rho \) be the correlation between \( Y \) and \( X \) in the population and let \( r \) be the sample correlation. I define \( r \) below. The population correlation is defined analogously from population data.

Suppose each of \( n \) sampled individuals is measured on two quantitative characteristics called \( Y \) and \( X \). The data are pairs of observations \((X_1, Y_1), (X_2, Y_2), \cdots (X_n, Y_n)\), where \((X_i, Y_i)\) is the \((X, Y)\) pair for the \( i^{th} \) individual in the sample. The sample correlation between \( Y \) and \( X \), also called the Pearson product moment correlation coefficient, is

\[
r = \frac{S_{XY}}{S_X S_Y} = \frac{\sum_i (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_i (X_i - X)^2 \sum_i (Y_i - Y)^2},
\]

where

\[
S_{XY} = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{n - 1}
\]

is the sample covariance between \( Y \) and \( X \), and \( S_Y = \sqrt{\sum_i (Y_i - \bar{Y})^2/(n - 1)} \) and \( S_X = \sqrt{\sum_i (X_i - \bar{X})^2/(n - 1)} \) are the standard deviations for the \( Y \) and \( X \) samples. Here are eight important properties of \( r \):
1. $-1 \leq r \leq 1$.
2. If $Y_i$ tends to increase linearly with $X_i$ then $r > 0$.
3. If $Y_i$ tends to decrease linearly with $X_i$ then $r < 0$.
4. If there is a perfect linear relationship between $Y_i$ and $X_i$ with a positive slope then $r = +1$.
5. If there is a perfect linear relationship between $Y_i$ and $X_i$ with a negative slope then $r = -1$.
6. The closer the points $(X_i, Y_i)$ come to forming a straight line, the closer $r$ is to $\pm 1$.
7. The magnitude of $r$ is unchanged if either the $X$ or $Y$ sample is transformed linearly (i.e. feet to inches, pounds to kilograms, Celsius to Fahrenheit).
8. The correlation does not depend on which variable is called $Y$ and which is called $X$.

If $r$ is near $\pm 1$, then there is a strong linear relationship between $Y$ and $X$ in the sample. This suggests we might be able to accurately predict $Y$ from $X$ with a linear equation (i.e. linear regression). If $r$ is near 0, there is a weak linear relationship between $Y$ and $X$, which suggests that a linear equation provides little help for predicting $Y$ from $X$. The pictures below should help you develop a sense about the size of $r$.

Note that $r = 0$ does not imply that $Y$ and $X$ are not related in the sample. It only implies they are not linearly related. For example, in the last plot on the next page, $r = 0$ yet $Y_i = X_i^2$. 

![Diagram showing different correlation values](image-url)
Testing that $\rho = 0$

Suppose you want to test $H_0 : \rho = 0$ against $H_A : \rho \neq 0$, where $\rho$ is the population correlation between $Y$ and $X$. This test is usually interpreted as a test of no association, or relationship, between $Y$ and $X$ in the population. Keep in mind, however, that $\rho$ measures the strength of a linear relationship.

The standard test of $H_0 : \rho = 0$ is based on the magnitude of $r$. If we let

$$t_s = r \sqrt{\frac{n - 2}{1 - r^2}},$$

then the test rejects $H_0$ in favor of $H_A$ if $|t_s| \geq t_{crit}$, where $t_{crit}$ is the two-sided test critical value from a $t$-distribution with $df = n - 2$. The p-value for the test is the area under the $t$-curve outside $\pm t_s$ (i.e. two-tailed test p-value).

This test assumes that the data are a random sample from a bivariate normal population for $(X, Y)$. This assumption implies that all linear combinations of $X$ and $Y$, say $aX + bY$, are normal. In particular, the (marginal) population frequency curves for $X$ and $Y$ are normal. At a minimum, you should make boxplots of the $X$ and $Y$ samples to check marginal normality. For large-sized samples, a plot of $Y$ against $X$ should be roughly an
elliptical cloud, with the density of the points decreasing as the points move away from the center of the cloud.

**The Spearman Correlation Coefficient**

The Pearson correlation $r$ can be highly influenced by outliers in one or both samples. For example, $r \approx -1$ in the plot above. If you delete the one extreme case with the largest $X$ and smallest $Y$ value then $r \approx 0$. The two analyses are contradictory. The first analysis (ignoring the plot) suggests a strong linear relationship, whereas the second suggests the lack of a linear relationship. I will not strongly argue that you should (must?) delete the extreme case, but I am concerned about any conclusion that depends heavily on the presence of a single observation in the data set.

**Spearman’s rank correlation coefficient** $r_S$ is a sensible alternative to $r$ when normality is unreasonable or outliers are present. Most books give a computational formula for $r_S$. I will verbally describe how to compute $r_S$. First, order the $X_i$s and assign them ranks. Then do the same for the $Y_i$s and replace the original data pairs by the pairs of ranked values. The Spearman rank correlation is the Pearson correlation computed from the pairs of ranks.
The Spearman correlation $r_S$ estimates the **population rank correlation coefficient**, which is a measure of the strength of linear relationship between population ranks. The Spearman correlation, as with other rank based methods, is not sensitive to the presence of outliers in the data. In the plot above, $r_S \approx 0$ whether the unusual point is included or excluded from the analysis. In samples without unusual observations and a linear trend, you often find that $r_S \approx r$.

An important point to note is that the magnitude of the Spearman correlation does not change if either $X$ or $Y$ or both are transformed (monotonically). Thus, if $r_S$ is noticeably greater than $r$, a transformation of the data might provide a stronger linear relationship.

**Example**

Eight patients underwent a thyroid operation. Three variables were measured on each patient: weight in kg, time of operation in minutes, and blood loss in ml. The scientists were interested in the factors that influence blood loss. **JMP-IN** output for this example follows the next page. Kristina will show you how to use the **MULTIVARIATE** platform in **JMP-IN** to compute correlations.

<table>
<thead>
<tr>
<th>weight</th>
<th>time</th>
<th>blood loss</th>
</tr>
</thead>
<tbody>
<tr>
<td>44.3</td>
<td>105</td>
<td>503</td>
</tr>
<tr>
<td>40.6</td>
<td>80</td>
<td>490</td>
</tr>
<tr>
<td>69.0</td>
<td>86</td>
<td>471</td>
</tr>
<tr>
<td>43.7</td>
<td>112</td>
<td>505</td>
</tr>
<tr>
<td>50.3</td>
<td>109</td>
<td>482</td>
</tr>
<tr>
<td>60.2</td>
<td>96</td>
<td>513</td>
</tr>
<tr>
<td>35.4</td>
<td>120</td>
<td>464</td>
</tr>
</tbody>
</table>

**Comments:**

1. (Pearson correlations). Blood loss tends to decrease linearly as weight increases, so $r$ should be negative. The output gives $r = -.77$. There is not much of a linear relationship between blood loss and time, so $r$ should be close to 0. The output gives $r = -.11$. Similarly, weight and time have a weak negative correlation, $r = -.07$. 
2. The Pearson and Spearman correlations are fairly consistent here. Only the correlation between blood loss and weight is significant at the $\alpha = 0.05$ level (the p-values are given in the rightmost column).

3. Two other measures of association available in **JMP-IN** are Kendall’s $\tau$ (given here) and Hoeffding’s D (not given).
Simple Linear Regression

In linear regression, we are interested in developing a linear equation that best summarizes the relationship in a sample between the **response variable** $Y$ and the **predictor variable** (or **independent variable**) $X$. The equation is also used to predict $Y$ from $X$. The variables are not treated symmetrically in regression, but the appropriate choice for the response and predictor is usually apparent.

Linear Equation

If there is a perfect linear relationship between $Y$ and $X$ then $Y = \beta_0 + \beta_1X$ for some $\beta_0$ and $\beta_1$, where $\beta_0$ is the $Y$-intercept and $\beta_1$ is the slope of the line. Two plots of linear relationships are given below. The left plot has $\beta_0 = 5$ and $\beta_1 = 3$. The slope is positive, which indicates that $Y$ increases linearly when $X$ increases. The right plot has $\beta_0 = 7$ and $\beta_1 = -2$. The slope is negative, which indicates that $Y$ decreases linearly when $X$ increases.

![](image1.png)

Least Squares

Data rarely, if ever, fall on a straight line. However, a straight line will often describe the trend for a set of data. Given a data set $(X_i, Y_i), \ i = 1, ..., n$ with a **linear trend**, what linear equation “best” summarizes the observed relationship between $Y$ and $X$? There is no
universally accepted definition of “best”, but many researchers accept the **Least Squares** line (LS line) as a reasonable summary.

Mathematically, the LS line chooses the values of $\beta_0$ and $\beta_1$ that minimize

$$\sum_{i=1}^{n} \{Y_i - (\beta_0 + \beta_1 X_i)\}^2$$

over all possible choices of $\beta_0$ and $\beta_1$. These values can be obtained using calculus. Rather than worry about this calculation, note that the LS line makes the sum of squared deviations between the responses $Y_i$ and the line as small as possible, over all possible lines. The LS line typically goes through “the heart” of the data, and is often closely approximated by an eye-ball fit to the data.

The equation of the LS line is

$$\hat{Y} = b_0 + b_1 X$$

where the intercept $b_0$ satisfies

$$b_0 = \bar{Y} - b_1 \bar{X}$$

and the slope is

$$b_1 = \frac{\sum_i (Y_i - \bar{Y})(X_i - \bar{X})}{\sum_i (X_i - \bar{X})^2} = r \frac{S_Y}{S_X}.$$
As before, $r$ is the Pearson correlation between $Y$ and $X$, whereas $S_Y$ and $S_X$ are the sample standard deviations for the $Y$ and $X$ samples, respectively. The **sign of the slope** and the **sign of the correlation** are identical (i.e. + correlation implies + slope).

Special symbols $b_0$ and $b_1$ identify the LS intercept and slope to distinguish the LS line from the generic line $Y = \beta_0 + \beta_1 X$. You should think of $\hat{Y}$ as the **fitted value** at $X$, or the value of the LS line at $X$.

**JMP-IN Implementation**

A plot of the data, and a least squares fit of a line is carried out in **JMP-IN** using the **FIT Y by X** platform under **BASIC STATS**. Kristina will illustrate this in Thursday’s lab.

For the **thyroid operation data** with $Y =$ Blood loss in $ml$ and $X =$ Weight in $kg$, the LS line is $\hat{Y} = 552.44 - 1.30X$, or Predicted Blood Loss = $552.44 - 1.30$ Weight. For an 86$kg$ individual, the Predicted Blood Loss = $552.44 - 1.30*86 = 440.64ml$. See the **JMP-IN** output at the end of the handout.

The LS regression coefficients for this model are interpreted as follows. The intercept $b_0$ is the predicted blood loss for a 0 $kg$ individual. The intercept has no meaning here. The slope $b_1$ is the predicted increase in blood loss for each additional $kg$ of weight. The slope is -1.30, so the predicted decrease in blood loss is 1.30 $ml$ for each increase of 1 $kg$ in weight.

Any fitted linear relationship holds only approximately and does not necessarily extend outside the range of the data. In particular, nonsensical predicted blood losses of less than zero are obtained at very large weights outside the range of data.

**ANOVA Table for Regression**

The LS line minimizes
\[
\sum_{i=1}^{n} \{Y_i - (\beta_0 + \beta_1 X_i)\}^2
\]
over all choices for $\beta_0$ and $\beta_1$. Inserting the LS estimates $b_0$ and $b_1$ into this expression gives
\[
\text{Residual Sums of Squares} = \sum_{i=1}^{n} \{Y_i - (b_0 + b_1 X_i)\}^2.
\]
Several bits of notation are needed. Let

\[ \hat{Y}_i = b_0 + b_1 X_i \]

be the predicted or fitted \( Y \)–value for an \( X \)–value of \( X_i \) and let \( e_i = Y_i - \hat{Y}_i \). The fitted value \( \hat{Y}_i \) is the value of the LS line at \( X_i \) whereas the residual \( e_i \) is the distance that the observed response \( Y_i \) is from the LS line. Given this notation,

\[
\text{Residual Sums of Squares} = \text{Res SS} = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} e_i^2.
\]

Here is a picture to clarify matters:

![Diagram](image)

The Residual SS, or sum of squared residuals, is small if each \( \hat{Y}_i \) is close to \( Y_i \) (i.e. the line closely fits the data). It can be shown that

\[
\text{Total SS in } Y = \sum_{i=1}^{n} (Y_i - \bar{Y})^2 \geq \text{Res SS} \geq 0.
\]

Also define

\[
\text{Regression SS} = \text{Reg SS} = \text{Total SS} - \text{Res SS} = b_1 \sum_{i=1}^{n} (Y_i - \bar{Y})(X_i - \bar{X}).
\]

The Total SS measures the variability in the \( Y \)–sample. Note that

\[ 0 \leq \text{Regression SS} \leq \text{Total SS}. \]
The percentage of the variability in the $Y$– sample that is **explained by the linear relationship** between $Y$ and $X$ is

$$R^2 = \text{coefficient of determination} = \frac{\text{Reg SS}}{\text{Total SS}}.$$ 

Given the definitions of the Sums of Squares, we can show $0 \leq R^2 \leq 1$ and

$$R^2 = \text{square of Pearson correlation coefficient} = r^2.$$ 

To understand the interpretation of $R^2$, at least in two extreme cases, note that

Reg SS = Total SS $\iff$ Res SS = 0

$\iff$ all the data points fall on a straight line

$\iff$ all the variability in $Y$ is explained by the linear relationship with $X$

(which has variation)

$\iff$ $R^2 = 1$. (see the picture below)

Furthermore,

Reg SS = 0 $\iff$ Total SS = Res SS

$\iff$ $b_1 = 0$
\[ \Rightarrow \text{LS line is } \hat{Y} = \bar{Y} \]
\[ \Rightarrow \text{none of the variability in } Y \text{ is explained by a linear relationship} \]
\[ \Rightarrow R^2 = 0. \]

LS line with slope zero and intercept of average \( Y \).

Each Sum of Squares has a corresponding \( df \) (degrees of freedom). The Sums of Squares and \( df \) are arranged in an analysis of variance (ANOVA) table:

<table>
<thead>
<tr>
<th>Source</th>
<th>( df )</th>
<th>SS</th>
<th>MS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Residual</td>
<td>( n - 2 )</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>( n - 1 )</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The Total \( df \) is \( n - 1 \). The Residual \( df \) is \( n \) minus the number of parameters (2) estimated by the LS line. The Regression \( df \) is the number of predictor variables (1) in the model. A Mean Square is always equal to the Sum of Squares divided by the \( df \). SW use the following notation for the Residual MS: \( s^2_{Y|X} = Resid(SS)/(n - 2) \).
Brief Discussion of JMP-IN Output for Blood Loss Problem

1. Identify fitted line: Blood Loss = 552.44 - 1.30 Weight (i.e. $b_0 = 552.44$ and $b_1 = -1.30$).

2. Locate Analysis of Variance Table. More on this later.

3. Locate Parameter Estimates Table. More on this later.

4. Note that $R^2 = .5967 = (-.77247)^2 = r^2$. 