Correlation and Simple Linear Regression
Objectives

• Explain when to use correlational techniques to answer research questions or test hypothesis
• Report a correlation coefficient in terms of statistical significance and meaningfulness
• Explain the difference between correlation and simple linear regression
• Use correlation and simple linear regression to test a null hypothesis
• Perform some basic interpretation of output from SPSS
What is Correlation?

• Correlation is the simplest way to test the relationship between 2 groups of numbers.
• Notice with correlation you must be looking at real numbers and not data that is nominal or ordinal in scale. There are some non-parametric tests that can be used with ordinal data but you will not learn about those this week.
• Numbers can be correlated directly – that is if one set of numbers goes up (or down) the other does as well, or indirectly – that is they go in opposite directions, one goes up and one goes down.
Types of Research Questions Correlation
Addresses

Correlational techniques are used:
– to study relationships
– in exploratory studies to see if any relationship exists between your samples
– to test a hypothesis about a relationship
– to test the assumption of independence between samples or variables
– to look for interactions between variables
How are numbers correlated?

• Again, you can have a direct relationship between numbers or an indirect one.
• A direct relationship is when both sets either go up together or down together. It is represented by a positive sign in front of the coefficient.
• An indirect relationship is where the numbers move in opposite directions, when one goes up the other goes down. It is represented by a negative sign in front of the coefficient.
• Correlation coefficients can range from -1.00 to +1.00. At both -1.00 and +1.00 the numbers are perfectly correlated with each other.
• At 0 the numbers are totally non-correlated, that is there is no relationship between them.
Examples of Correlation Coefficients

- Values of \( r \) can range from -1 to 1 with 0 representing no correlation and +/-1 representing perfect correlation.
- The +/- reflects the direction of the correlation.
• In reality, it is very unlikely you will ever see a correlation coefficient of 0 or 1.00. Most will range from .30 to .60 and can be positive or negative.

• The line formed in a positive correlation will go up and to the right, while the line formed with a negative correlation will go up and to the left.
Assumptions

- As with all statistical tests, the sample must represent the population.
- The variables being correlated must each continuous and approximately Normal.
- The relationship between $X_1$ and $X_2$ must be linear – at least approximate a straight line.
- The upper plot is not linear so would not meet the assumption of linearity, while the bottom one would.
Interpreting Correlation Coefficients

• In previous slides you saw that with correlation you can form a line from two set of data (or points).

• The correlation coefficient (r) represents the relationship between your two variables in terms of the line that you formed.

• The correlation coefficient and the square associated with it (r^2) allow you to judge the strength if the relationship.

• In order to evaluate the correlation between variables, it is important to know both this strength or magnitude and the significance of the correlation.
The Strength of the Correlation Coefficient

- How large is large enough to demonstrate a relationship?
- It depends on what you are measuring – for example tests designed to measure the same thing should be highly correlated while variables occurring together in nature rarely are.
- Relationships with \((r)\) equal to 0.26 or higher are worth pursuing further.

<table>
<thead>
<tr>
<th>Range of Coefficient</th>
<th>Strength (in either positive or negative direction)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.00 to 0.25</td>
<td>Little if any</td>
</tr>
<tr>
<td>0.26 to 0.49</td>
<td>Low</td>
</tr>
<tr>
<td>0.50 to 0.69</td>
<td>Moderate</td>
</tr>
<tr>
<td>0.70 to 0.89</td>
<td>High</td>
</tr>
<tr>
<td>0.90 to 1.00</td>
<td>Very high</td>
</tr>
</tbody>
</table>
Significance of Correlations

- As with the other statistical tests we have studies, the significance calculated for the correlation tells you how reliable it is.
- The level of significance must meet the same criteria as will the other statistical tests.
- However, unlike with chi-square, with correlation you are able to judge the strength as well as the significance of the relationship. This is the value of the correlation coefficient $r$.
- If the significance is greater than 0.05 you cannot reject the null hypothesis of no relationship between variables or samples.
- The significance of a correlation coefficient of a particular magnitude will change depending on the size of the sample from which it was computed. The larger the sample the more likely the result is significant.
- Keep in mind however, that a correlation coefficient of 0.86 with a significance (p value) of 0.04 suggests a stronger relationship than a correlation coefficient of 0.43 with a significance of 0.04.
Meaningfulness of \((r)\)

- As mentioned before, when you take the square of the correlation coefficient you get the coefficient of correlation \((r^2)\).
- This is often used to describe the meaningfulness of \(r\).
- It is a measure of the amount of variance the two variables share – think overlap.
- The more overlap there is the more information the two variables share.
- To account for half the variability of the independent variable the correlation coefficient between it and the dependent variable must be around \(0.7 (0.7 \times 0.7 = 0.49)\).

The amount of overlap between the two variables
If we determine that the \(r\) is \(0.20\), then the \(r^2\) is \(0.04\) or \(4\%\) so \(4\%\) of the variability of B (the independent variable) is accounted by A (the dependent one) – That leaves \(96\%\) unaccounted for so you have not explained much of B using A.
Simple Linear Regression and Correlation

- As you have seen, correlation forms a line between two sets of points and assigns a value based on how strong the relationship appears.
- Simple linear regression does the same thing but takes it a step further. In simple linear regression, you make an equation for the line that was formed and each part of that equation means something to a statistician.
Comparison between Correlation and Simple Linear Regression

**Correlation**
- There is no direction to the relationship so you do not need to have a dependent and an independent variable.
- The interpretation depends on the correlation coefficient, the significance, and the coefficient of determination.

**Simple Linear Regression**
- The equation of a line must have a direction so you must have a dependent and an independent variable.
- The interpretation depends on the coefficient of the independent variable, the significance of that coefficient, and the overall model fit or coefficient of determination.
Regression Analysis

- **Regression analysis** is a statistical technique that forms a line between a dependent variable (or outcome variable) and one or more independent variables. The goal is often to try to predict the behavior of the dependent variable based on the values of the independent one.

- **Some Needed Definitions:**
  - Dependent Variable
  - Independent Variable
  - Linear Relationship
  - Scatter Plot
  - Correlation
  - Model
Definition of Dependent Variable

• The **dependent variable**, also known as the response or outcome variable. This is the variable we are attempting to predict using the other variable or variables we have in our data.

• Example: If we are attempting to predict birth weight given characteristics of the mother and father, siblings, and the pregnancy, then birth weight is our dependent variable.

• It is usually represented by the letter *y* in models.

• There is only one dependent variable in both Simple and Multiple Linear Regression.

• The dependent variable MUST BE CONTINUOUS to use linear regression!!!!!
Definition of Independent Variable

- The **independent variable** is also known as the predictor variable and the explanatory variable. This is the variable or variables you think might have a linear relationship to your outcome variable.

- Example: In the previous example, while birth weight is the dependent variables, all of the other variables we have, i.e. – characteristics of the mother, father, siblings, and pregnancy are independent variables.

- In Simple Linear Regression there is only one independent variable, while in Multiple Linear Regression there are more than one.

- The independent variable is represented by the letter x. When there is more than one, a subscript number is added, such as x1, x2, x3…
Linear regression attempts to “fit” the data to a line (establish a linear relationship) in order to predict the behavior of the dependent variable.
Definition of Scatter Plots

Using an X and a Y axis, we plot the data points for our x variable(s), the independent ones, and our y variable, dependent. Each data point represents a case. The best line that fits the plotted points represents the regression function of y on x.
Definition of Correlation

As introduced in the beginning of these slides, this corresponds to the degree to which the x and y variables change together. The higher the correlation the better the fit between the variables.

In this scatter plot we could assume that the correlation is fairly close to +1.00; that is, as the x value increases, so does the y value.
Definition of Model

A representation of the linear relationship between the independent and dependent variables. Usually expressed as $y = a + bx$, where $a$ is the intercept and $b$ is the coefficient for the variable $x$ which represents the slope.

In this example the dependent variable ranges from 20 to 60, while the independent one ranges from 0 to 30. The intercept ($a$) would be where the fitted line crosses the y axis, while the slope ($b$) is the change in $y$ divided by the change in $x$.

An approximate model for these points might be $y = 20 + 1.3x$.

If you then know $x$ is 20, you could predict $y$ is 46, by plugging in the number 20 for $x$ in this equation.
Testing the “Fit” of the Model

• There are two numbers you need to consider when deciding whether the line formed (or the model) is a good one.

• There is a coefficient associated with the independent variable and it has a significance (p value) assigned to it. If the p value is greater than 0.05 then the independent variable does not have a significant effect on the dependent one.
Example using Simple Linear Regression

- An example: The equation of the line representing the relationship between estriol levels in pregnant women and birth weight was determined to be $y = 21.52 + 0.608x$. So for a given estriol level you could predict a birth weight. If the estriol level was 18 mg/24 hr, the birth weight predicted using that equation would be $y = 21.52 + (0.608*18) = 32.46 = 3246$ g.

- Since it is unlikely that this relationship produces a perfect line, there is some error involved in the above estimate. The error $(e)$ is normally distributed with a mean of 0 and a variance of $\sigma^2$. If the line were perfect then the variance ($\sigma^2$) would also equal zero.

- SO what could we say about the relationship? Well, certainly as estriol levels increase birth weight also increases – this may imply a relationship between estriol levels and birth weight.
Does Estriol Influence Birth Weight?

- You do not know what p-value was assigned to the coefficient of the estriol levels in the previous example.
- You do know the coefficient is 0.608. This means that each mg/24 hour increase in estriol leads to a little more than \( \frac{1}{2} \) gram increase in birth weight.
- If you assume the p-value for this is 0.05 or less then you can also assume that this is a significant influence.
- But how well does it predict birth weight? Just because the relationship between estriol and birth weight is significant, does not mean it is sufficient to predict the variance observed in birth weight. To determine that you should interpret the model fit or the r² of the model. The larger the r² is the better the fit and the more likely you could use this model to predict birth weight.
Testing a Null Hypothesis using Correlation and Linear Regression

In the example on the previous slide your null hypotheses could have been:

• There is no relationship between estriol levels and birth weight. This null hypothesis can be tested using correlation. You can reject the null hypothesis based on the significance and also judge how strong the relationship is using the correlation coefficient (r).

• Estriol levels during pregnancy have no influence on birth weight. Since this is directional you need to use Linear Regression to test this null hypothesis. You can reject the null hypothesis based on the significance of the coefficient and associated p-value for the independent variable.
This is a scatterplot of x and y. You could superimpose a line of best fit on this as the points are somewhat linear.

This output tells you the model fit. The r of .816 suggests that the correlation is strong. The r^2 of .667 means that approximately 67% of the variance of the dependent variable is explained by the independent one.

<table>
<thead>
<tr>
<th>Model</th>
<th>R</th>
<th>R Square</th>
<th>Adjusted R Square</th>
<th>Std. Error of the Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.816a</td>
<td>.667</td>
<td>.629</td>
<td>2.01881</td>
</tr>
</tbody>
</table>

This output tells you that the coefficient of the independent variable is .816. This is significant at .002. It is also significant based on the confidence interval. You could reject your null hypothesis based on this output.

The model represented here is: y = 3 + .500 x