II. Descriptive Statistics
D. Linear Correlation and Regression

In this section
- Linear Correlation
- Cause and Effect
- Linear Regression

1. Linear Correlation

Quantifying Linear Correlation
The Pearson product-moment correlation coefficient, written as $r$, can describe a linear relationship between two quantitative variables. It is important to notice that when looking at this value, it only indicates the linear relationship. You could have another kind of relationship present in the data. The $r$ value indicates both the strength of the linear association and its direction.

You will not need to calculate an $r$ value by hand, but in case you are interested the formula is below:

$$SSX = \sum (x - \bar{x})^2 = \sum x^2 - \frac{(\sum x)^2}{n}$$

$$SSY = \sum (y - \bar{y})^2 = \sum y^2 - \frac{(\sum y)^2}{n}$$

$$SXY = \sum (x - \bar{x})(y - \bar{y}) = \sum xy - \frac{(\sum x)(\sum y)}{n}$$

$$r = \frac{SXY}{\sqrt{SSX} \sqrt{SSY}}$$

What you do need to be able to do is interpret what an $r$ value tells you. We will start with the direction of the relationship.

Positive $r$ suggests large values of X and Y occur together and that small values of X and Y occur together. This means that the slope of the line that best fits the points is positive. An example would be Experience and Salary. People with lower levels of experience tend to have lower salaries and people with more experience tend to have higher salaries.
Negative $r$ suggests large values of one variable will tend to occur with small values of the other variable. This means that the slope of the line that best fits the points is negative. An example would be Weight of a car and Gas mileage. Light cars tend to have higher gas mileage and heavier cars tend to have lower gas mileage.

So the sign of the $r$ value tells us the direction of the relationship. Strength of the relationship is measured by the actual value of the number. By the term strength, we mean how close are the points to a line? The closer the points are to a line, the stronger the relationship. Because of the setup of $r$, the maximum value for $r$ (in terms of absolute value) is 1. Below are some useful things to keep in mind.

$-1 \leq r \leq 1$

- If $r = 1$ then there is perfect positive linear correlation – all data are exactly on a line with positive slope
- If $r = -1$ then there is perfect negative linear correlation – all data are exactly on a line with negative slope
- If $r = 0$ then there is no linear relationship (keep in mind there could be another type of correlation)

The stronger the linear relationship, the larger $|r|$ (the closer to one this value will be).

Generally, we will say there is a strong relationship if $|r| \geq .75$

**Example**

Looking back to our example from the last section when introducing scatterplots we had $X =$ Dosage of Drug and $Y =$ Reduction in Blood Pressure, what do you think the $r$ value will be? Remember based on the scatterplot that the points had a strong positive linear relationship. Not perfect but pretty close meaning the $r$ value should be close to 1. If you calculate this value, you will get $r = .99728$. This should seem reasonable as it supports what we identified in the graph.

Another measure, you will sometimes see reported is the R-squared value. It is common for computer software to give you an R-squared value instead of $r$. This value represents the percent of variation in $Y$ explained by the model. It measures the strength of the relationship and in the linear case is simply calculated by squaring the $r$ value. The higher R-squared is, the better the model.

$0\% \leq R^2 \leq 100\%$

**Example**

For the Drug example $r = .99728$

$R^2 = (.99728)^2 = .995 = 99.5\%$
2. **Cause and Effect**

“**Causal**” Research – When the objective is to determine if a variable causes a certain behavior (whether there is a cause and effect relationship between variables)

Note: it is never possible to prove causality just based on the relationship between two variables

**Example**
There is a strong statistical correlation over the months of the year between ice cream consumption and the number of assaults in the U.S. The $r$ value for this data is above .9.

Does this mean ice cream manufacturers are responsible for crime?

No! The correlation occurs statistically because the hot temperatures of summer increases both ice cream consumption and assaults (High values occur at the same time and low values occur at the same time)

Thus, correlation does NOT imply causation. This is one of the biggest mistakes that I see in the interpretation of a correlation. You should always keep in mind that other factors besides cause and effect can create an observed correlation.

To establish whether two variables are causally related you must establish all of the following:

1) **Time order** - the cause must have occurred before the effect
2) **Co-variation (statistical association)** – the correlation coefficient and graph must show a strong relationship between the dependent and independent variable
3) **Rationale** - there must be a logical and compelling explanation for why one variable causes the other
4) **Non-spuriousness** - it must be established that the independent variable X, and only X, was the cause of changes in the dependent variable Y; rival explanations must be ruled out

The first three of these can be easily established in many cases. It is the fourth criteria which is hard and can rarely be truly shown. To help identify a relationship as cause and effect, a study should be performed many times. The study should yield the same results every time it is conducted. Given that the outside variables will differ from situation to situation, this helps rule out rival explanations.

“Causal” research is very complex and the researcher can rarely be completely certain that there are not other factors influencing a relationship.
3. **Linear Regression**

**Deterministic View** – This is the idea that Y is caused by X or that once X has happened, Y will follow. In this situation, the exact value of Y is known.

The deterministic view is studied in a typical algebra class. However, a deterministic view when applied to the behavior of many variables is not possible.

**Regression** – A technique used to predict variables (typically difficult to measure variables) based on a set of other variables (typically easier to measure variables).

**Linear Regression** – Used to predict the value of Y (the response variable), based on X (the explanatory variable) using a linear equation.

**Example**
Predict reaction time based on blood alcohol level. Reaction time is difficult to measure so instead we predict it with blood alcohol level which is easy to measure.

The linear regression model expresses Y as a function of X plus random error.

Random error reflects variation in Y values. Keep in mind we are going to measure X, so assuming we get a good measure there is no error in the X variable. However, when we go to use X to predict Y, the prediction will not be exact. Therefore, there is error in the Y variable. Graphically this error is represented by the vertical distance between the points and the line.

The linear regression model is:

\[ Y = b_0 + b_1 x \]

Where \( b_0 \) is the y-intercept and \( b_1 \) is the slope

The above formula is the same format as what you should be used to from an algebra class. However, the way we denote the relationship is different. It is important you become familiar with this notation.

In order to use linear regression, we must first make sure the model is reasonable. The scatter plot and \( r \) should indicate a strong relationship. If the model is not reasonable, do not fit a line. It may still be possible to do regression with a more complicated model. However, if there is no relationship between the variables then regression cannot be used. In this class we will not worry about more complicated models, but you should understand that a simple linear model is just one of the many options available.
When using a linear regression model, we need the line that is the “best” fit for our data. Since our purpose will be to predict, we will want to pick the line that will minimize the error in the prediction. To accomplish this we will use the method of least-squares.

**Method of Least-Squares** – says that the sum of the squares of the vertical distances from the points to the line is minimized. Remember it is the vertical distance that represents the error.

To calculate the “best” fit line you can use the following formulas. You do not have to do this by hand in this class. I show you the formulas in case you are interested.

\[ b_1 = \frac{S_{XY}}{S_{XX}} = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sum (x - \bar{x})^2} = \frac{\sum xy - \frac{(\sum x)(\sum y)}{n}}{\sum x^2 - \frac{(\sum x)^2}{n}} \]

\[ b_0 = \bar{y} - b_1 \bar{x} \]

**Example**
At the beginning of this section when looking at correlation for the Dosage of drug and Reduction in blood pressure example we identified \( r = .99728 \) which indicates a high positive linear correlation. This fact along with the scatterplot supports the use of linear regression in this case.

With the above formulas, you can calculate \( b_1 = 0.118 \) and \( b_0 = -3.4 \).

Therefore, the regression model in this case is:
\[ y = b_0 + b_1 x \]
\[ y = -3.4 + 0.118x \]

As I stated earlier, you will not have to calculate the formula by hand. Instead, I will provide computer output and you need to be able to answer questions based on the output. The computer output (a regression plot) for this example follows.
In this output, the equation and the R-squared value will be given. If you look above the graph, you will see this information. Notice the R-squared value for this example is exactly what we stated previously in this section of material. You need to be able to get the \( r \) value based on the R-squared that is given in the output. All you have to do is take the square root of the R-squared value. The thing you have to be careful of is the direction of the relationship. Remember that if the slope of the line is positive then \( r \) is positive and if the slope is negative then \( r \) is negative. Therefore, you must look at the slope in order to decide if \( r \) is positive or negative.

In terms of the equation, you need to be able to use it for prediction. This is a pretty direct process as we will always be predicting \( Y \) based on \( X \). Therefore, you will plug in for \( X \) and solve for \( Y \).

For our example, predict the Reduction in Blood Pressure if 250 is the Dosage of Drug.

\[
y = -3.4 + 0.118x
\]

\[
y = -3.4 + 0.118(250)
\]

\[
y = 26.1
\]
Example
Consider the following data and software output, which give the weight (in thousands of pounds), X, and gasoline mileage (miles per gallon), Y, for ten different automobiles.

<table>
<thead>
<tr>
<th>X</th>
<th>2.5</th>
<th>3.0</th>
<th>4.0</th>
<th>3.5</th>
<th>2.7</th>
<th>4.5</th>
<th>3.8</th>
<th>2.9</th>
<th>5.0</th>
<th>2.2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y</td>
<td>40</td>
<td>43</td>
<td>30</td>
<td>35</td>
<td>42</td>
<td>19</td>
<td>32</td>
<td>39</td>
<td>15</td>
<td>44</td>
</tr>
</tbody>
</table>

Regression Plot

\[ Y = 70.1058 - 10.6175X \]

\[ S = 3.02406 \quad R-Sq = 92.0 \% \quad R-Sq(adj) = 91.0 \% \]

1. Calculate \( r \).

2. Based on \( r \) and the scatterplot is linear regression justified in this case.

3. Predict the gas mileage for a car weighing 4.2 (4,200) pounds.

Answers

1. \[ r = -\sqrt{R^2} = -\sqrt{.92} = -.959 \]
   This is negative because of the negative slope

2. Yes linear regression is justified, \(|r| \geq .75\) and the points are spread reasonably about the line on the scatterplot

3. \[ Y = 70.1058 - 10.6175(4.2) = 25.512 \]
Cautions with regression

There are two common mistakes with regression. You must be aware of the problems with extrapolation and extreme values.

**Interpolation** – predicting Y values for X values that are within the range of the scatter plot (this is what regression should be used for)

**Extrapolation** – predicting Y values for X values beyond the range of the observations (this should not be done with a basic regression model, it is a complex problem)

If our X variable ranges from 100 to 500 as it does in the Dosage of drug and Reduction in blood pressure example then it is reasonable to predict within that range. However, if you try and predict for an X of 1000 then you have no data indicating that this relationship holds at that value. It is quite possible that the relationship changes beyond the range of the data. There is no way to know this without collecting data consistent with the X values you want to predict.

The least-squares line and the $r$ value can be affected greatly by extreme data points. In order to illustrate this we will look at some computer output.

![Regression Plot](image)

Calculate the $r$ value for the above data.

$$r = \sqrt{0} = 0$$

With an $r$ of 0, we know that there is no linear relationship between $X$ and $Y$. 
Calculate $r$ for the above data.

$$r = \sqrt{.779} = .883$$

With an $r$ of .883, which is bigger than the criteria of .75 it seems like we have a strong relationship. With further investigation via the scatterplot, you will see that all of the data is in the bottom left of the graph except one data point which is extreme. What I actually did was take the data from the previous graph with an $r$ value of 0 and add one extreme value. Notice the extreme value makes the other data points appear close together. They also appear numerically close since the one value is so extreme. Therefore, the $r$ value is high because the points are close to the line. In this case linear regression is not justified. If you have an extreme value in a plot like in this case, you should remove the extreme value and see if the relationship still exists. In this case it does not so linear regression will not work for this data.