Section 4G – Quantitative Relationships: Correlation and Regression

Vocabulary

Correlation: Statistical analysis that determines if there is a relationship between two different quantitative variables.

Regression: Statistical analysis that involves finding the line or model that best fits a quantitative relationship, using the model to make predictions, and analyzing error in those predictions.

Explanatory Variable (x): Another name for the x-variable or independent variable in a correlation study.

Response Variable (y): Another name for the y-variable or dependent variable in a correlation study.

Correlation Coefficient (r): A statistic between −1 and +1 that measures the strength and direction of linear relationships between two quantitative variables.

R-squared (r^2): Also called the coefficient of determination. This statistic measures the percent of variability in the y-variable that can be explained by the linear relationship with the x-variable.

Residual $(y - \hat{y})$: The vertical distance between the regression line and a point in the scatterplot.

Standard Deviation of the Residual Errors (s_e): A statistic that measures how far points in a scatterplot are from the regression line on average and measures the average amount of prediction error.

Slope (b_1): The amount of increase or decrease in the y-variable for every one-unit increase in the x-variable.

Y-Intercept (b_0): The predicted y-value when the x-value is zero.

Regression Line $(\hat{y} = b_0 + b_1 x)$: Also called the line of best fit or the line of least squares. This line minimizes the vertical distances between it and all the points in the scatterplot.

Scatterplot: A graph for visualizing the relationship between two quantitative ordered pair variables. The ordered pairs (x, y) are plotted on the rectangular coordinate system.

Residual Plot: A graph that pairs the residuals with the x values. This graph should be evenly spread out and not fan shaped.

Histogram of the Residuals: A graph showing the shape of the residuals. This graph should be nearly normal and centered close to zero.

Introduction

Sometimes we want to know if two different quantitative variables are related to each other. This kind of relationship study is difficult because the units are different. We cannot directly compare the height of man in inches to his weight in pounds. Inches and pounds are completely different. Statisticians and mathematicians developed a type of analysis for this situation called “correlation and regression”. The idea is to let one variable be X and the other variable be Y. Then use ordered pair data to create a graph called a scatterplot and look for patterns. The most common is a linear pattern (correlation). If we see a linear pattern, we can also calculate the line that best fits the data and use this line to make predictions (regression).

Choosing your variables

It is important to determine which variable will be X and which variable will be Y. In statistics, we call the X-variable the “explanatory variable” or the “independent variable”. We call the Y-variable the “response variable” or “dependent variable”. How do we chose? Here are a couple key questions to ask yourself.

- Does one variable respond more than the other does?
- Which variable is the focus of the study and the variable I might want to make predictions about?

Let us look at some examples.

Example: Year (time) and unemployment rates in U.S.

Ask yourself the following question. Does one of the variables responds more than the other? Does time fluctuate in response to the unemployment rate? That does not sound right. Time seems to go on no matter what happens with unemployment. Do you think unemployment might fluctuate in response to time? That seems more likely. So we should let the explanatory variable X be time (years) and let the response variable y be unemployment rate. Unemployment responds to time, but not the other way around.

Example: The unemployment rate in U.S. and the national debt in the U.S.

These variables respond to each other, so either variable could be the response variable Y. In this case, pick the response variable (Y) to be the one you are most interested in (focus of the study) or the variable you may want to make predictions about. If there is a relationship, then the Y-variable will be the variable you can make predictions about.

Suppose the focus of your study and the variable you want to predict is the national debt. Unemployment may just be one factor that may be related to the national debt. If that is the case, you should make the national debt your response variable Y. By default, that means that unemployment rate would be explanatory variable X.

Correlation Graphs and Statistics with StatKey

To study the relationship between two different quantitative variable, you will need ordered pair data. For example, we will need the height and weight of the same men, or the unemployment rate and national debt of the same countries. Decide which variable should be X and which variable should be Y. The computer will then make ordered pairs from your data (X , Y) and plot all the points on the rectangular coordinate system. This graph of all the ordered pairs is called a scatterplot.

Example

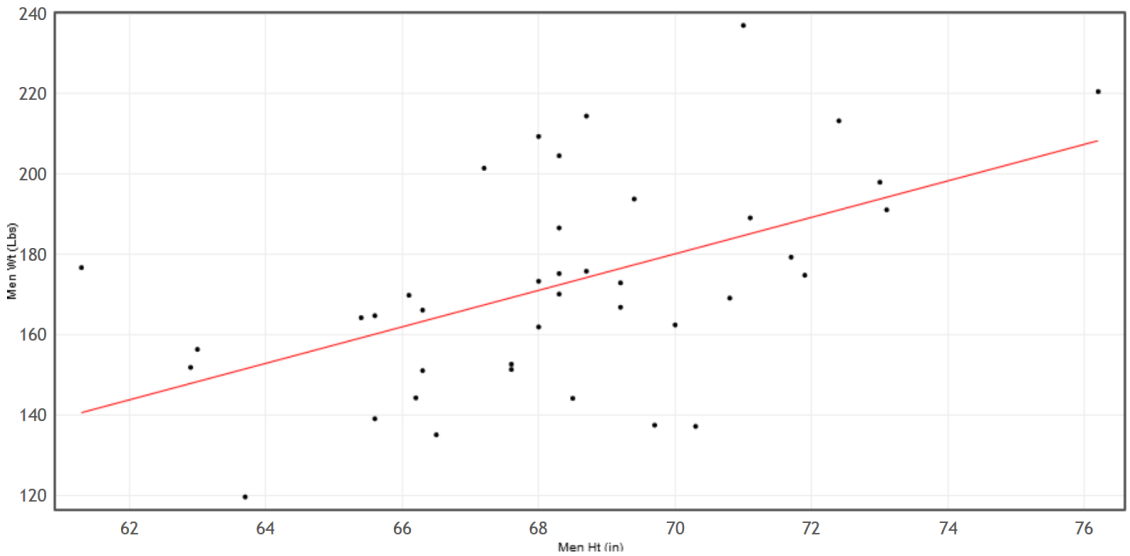
Suppose we want to study if the weights in pounds of the men in the health data is related to their heights in inches. I am most interested in predicting the weights of men from their heights so I will let the weight be the response variable Y and height be the explanatory variable X. Notice these are ordered pairs, since the heights and weights came from the same 40 men.

To put the data into StatKey, you will want to open a fresh excel spreadsheet and place the two data sets side by side. These two data sets are already next to each other in the health data, but in general, the data sets may not be. Copy the two columns of data together.

Go to www.lock5stat.com and click on “StatKey”. Under the “Descriptive Statistics and Graphs” menu, click on “two quantitative variables”. Under the “edit data” tab, paste the height and weight data into StatKey. The graph you see is the scatterplot. Notice StatKey has placed the heights on the horizontal x-axis and the weights on the vertical y-axis. If it is backward, simply click the “switch variables” button. It is also nice to check the “show regression line” box. The regression line is the line that best fits the points in the scatterplot. StatKey has also given us some statistics to help understand the relationship.

StatKey **Descriptive Statistics for Two Quantitative Variables**

Custom Dataset | Show Data Table | Edit Data | Upload File | Change Column(s)



Analyzing scatterplots is an important skill. In this graph, we see that the points seem to follow the linear pattern reasonably well and are reasonably close to the line. Shorter men on the left tend to have lower weights than taller men on the right. The line goes up from left to right. We call this a "positive linear relationship", or a "positive correlation". If the line goes down from left to right, we would call that a "negative linear relationship", or a "negative correlation".

Summary Statistics

Switch Variables

Statistic	Men Ht (in)	Men Wt (Lbs)
Mean	68.335	172.550
Standard Deviation	3.020	26.327
Sample Size	40	
Correlation	0.522	
Slope	4.553	
Intercept	-138.607	

Scatterplot Controls

☑ Show Regression Line

We see that StatKey has given us the mean and standard deviation of each data set (heights and weights). It has also given us the sample size (n) of 40. There were 40 ordered pairs (40 heights and 40 weights from the same 40 men). The number next to the word "Correlation" is 0.522. This is called the "correlation coefficient" (r) and is an important statistic in measuring the direction and strength of the linear relationship. Here are some general guidelines for understanding the correlation coefficient "r".

Correlation Coefficient (r)

The correlation coefficient (r) is a number between -1 and +1 that measures the strength and direction of correlation. The correlation coefficient is an extremely difficult calculation that is very time consuming. Like most statistics, it is better to use a computer program like StatKey or Statcato to calculate it.

If the r is negative, the regression line will go down from left to right. If you remember from algebra classes, this means the line has a negative slope. If the r is positive, the regression line will go up from left to right. This means the line has a positive slope. The closer r is to +1 or −1, the stronger the relationship. This means the points are very close to the line. The closer r is to zero, the weaker the relationship. The points are very far from the line. It is important to always look at the scatterplot with the r-value. Do not just look at an r-value without looking at the scatterplot. These are not strict rules, but general guidelines. A scatterplot with many points and a 0.7 r-value can mean something different from a scatterplot with only a few points and a 0.7 r-value.

- If r is close to +1 (like r = +0.893) ➔ Strong, Positive Correlation (line going up from left to right (positive slope) and the points in scatterplot are close to line) ,
(r ≈ +0.6, +0.7, +0.8 , +0.9 usually indicate pretty strong positive correlation)

- If r is close to −1 (like r = −0.916) ➔ Strong Negative Correlation (line going down from left to right (negative slope) and the points in the scatterplot are close to the line)
(r ≈ −0.6, −0.7, −0.8 , −0.9 usually indicate pretty strong negative correlation)

- If r close to zero (like +0.037 or −0.009) ➔ No linear correlation. Points in the scatterplot do not follow any linear pattern. There still could be a nonlinear curved pattern though.
(r ≈ ±0.1, ±0.0 usually indicate no linear correlation)

- If r ≈ ±0.2, ±0.3 usually indicate very weak linear correlation. There is some linear pattern but the points are very far from the regression line.

- If r ≈ ±0.4, ±0.5 usually indicate moderate linear correlation. There is a linear pattern and points are only moderately close to the regression line.

In the men's height and weight example, the r-value was +0.522. This tells us that there is a moderate positive linear relationship (or moderate positive correlation) between the height and weight of these men.

Important Note: Remember relationships or associations do not imply causation. Just because there is a positive linear relationship between the height and weight of these men, it does not give me the right to say that the height causes a man to have a certain weight. There are many confounding variables involved.

Correlation ≠ Causation

Coefficient of Determination (r^2)

If you square the r-value, you get the coefficient of determination. This statistic tells us the percentage of variability in the response variable (Y) that can be explained by the explanatory variable (X). In general, the higher the r^2 percentage, the stronger the relationship.

StatKey does not calculate r^2 for us, but it is not a difficult calculation. If we square the r-value, we get the following.

$r^2 = (0.522)^2 = 0.522 \times 0.522 \approx 0.272$ or 27.2%

So about 27.2% of the variability in the men's weights can be explained by the relationship with their heights.

Slope

The slope of the regression line is an important statistic in correlation and regression. It is a difficult calculation. If you are wondering how it is calculated, here is the formula the computer used.

Slope of the Regression Line = $\frac{Correlation\ Coefficient \times Standard\ Deviation\ of\ Y}{Standard\ Deviation\ of\ X}$

The slope is the amount of increase of decrease in Y for every 1-unit increase in X (per unit of X). If the slope is negative, than it is a "decrease" in Y and if the slope is positive, it is an "increase" in Y.

In this problem, StatKey gave us the slope as 4.553. Notice this is a positive slope so is indicating an increase in Y. The slope tells us that the weights of the men in the data set are increasing 4.553 pounds on average for every 1 inch taller they get. Another way to say that is that the weights are increasing on average 4.552 pounds per inch.

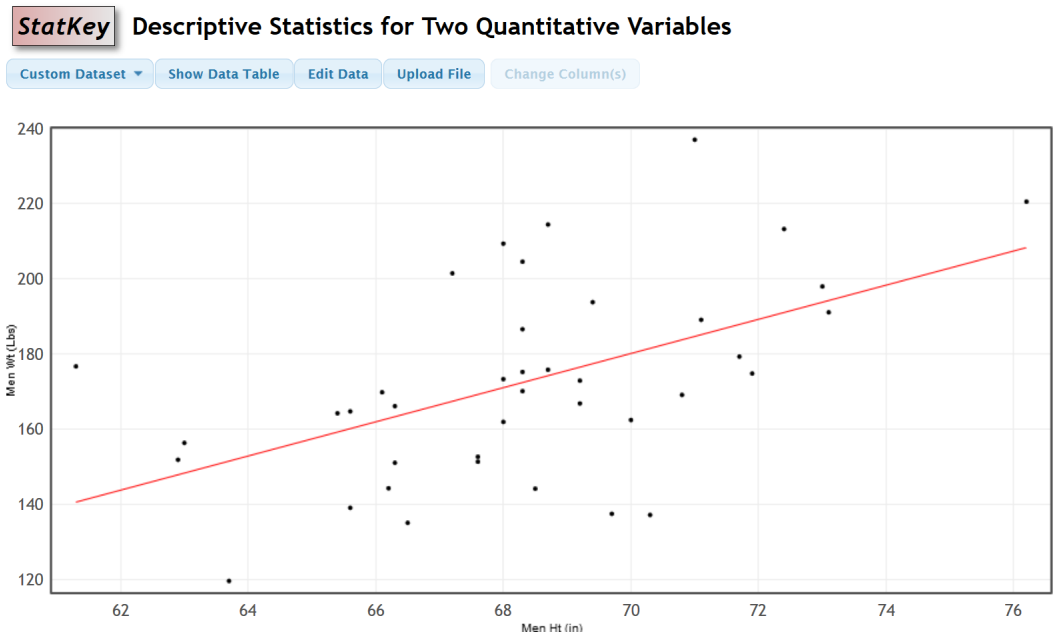
Y-intercept

The Y-intercept is another difficult calculation. In case you are wondering, here is the formula the computer used to calculate the Y-intercept. You must calculate the slope first, before you can find the Y-intercept.

Y-intercept of Regression Line = Mean of Y values $- (Slope \times Mean\ of\ X\ values)$

Y-intercepts can be difficult because they do not always make sense in context. The definition of a Y-intercept is the predicted Y-value when X is zero. StatKey calculated the Y-intercept for the height and weight data as −138.607. So by definition, the predicted average height of men that are zero inches tall is negative 138.607 pounds. That does not make sense.

In many situations (like heights of men), it is impossible for the X to be zero. Look at the scatterplot again for the height and weight data.



Notice that the points in the scatterplot have X values between about 61 inches and about 76 inches. This is called the scope of the X-values. The accuracy of this regression line is based on X values between about 61 and 76 inches. If we use this data to predict a man's weight from his height, we should only use heights in the scope (between 61 and 76). Going outside the scope is called extrapolation and can result in bad errors. So let us get back to the Y-intercept. The Y intercept is plugging in zero for X. Notice zero is not in the scope of the X values, so is an extrapolation. That means we will not expect the Y-intercept to make sense in this context. The number is correct and is important for the regression line accuracy, but a man cannot have a height of zero.

Some Y-intercepts do make sense in context. Suppose we are looking at the number of months a company has been in business (X) and their monthly revenue in thousands of dollars (Y). The Y-intercept may represent their starting capital at month zero or the amount of money the company had when they started their business.

Regression Line and Predictions

The regression line is also called the "line of best fit" or the "line of least squares". It minimizes the vertical distances between the points in the scatterplot and regression line itself. If there is correlation between the variables, then the regression line is also a prediction formula. If you plug in an X value into the equation for X, you can solve for Y and get a predicted Y value. The regression line is represented by the following formula.

$\hat{Y}$ = (Y-intercept) + (Slope) X

Plugging in our Y intercept (−138.607) and our slope (4.553), we get the following equation.

Regression Line for Heights and Weights of mean in the health data: $\hat{Y} = -138.607 + 4.553\ X$

The $\hat{Y}$ refers to the "predicted Y value" which can be very different from the actual Y values in the data set. You may also see computer programs put in the variable names for X and $\hat{Y}$.

Weights in pounds = −138.607 + 4.553 (Heights in inches)

We said already that there was a moderate correlation between the heights and weights of these men. So we should be able to use the formula to make a prediction.

Use the regression line equation to predict the average weights of men that are 73 inches tall. Remember Y represents weight and X represents height. Simply plug in 73 for X and solve for Y. Remember to follow the order of operations. Multiply the X value by the slope first, before you add it to the Y-intercept. Also, be aware of negative Y-intercepts and negative slopes.

$\hat{Y} = -138.607 + 4.553\ X$

$\hat{Y} = -138.607 + 4.553\ (73)$

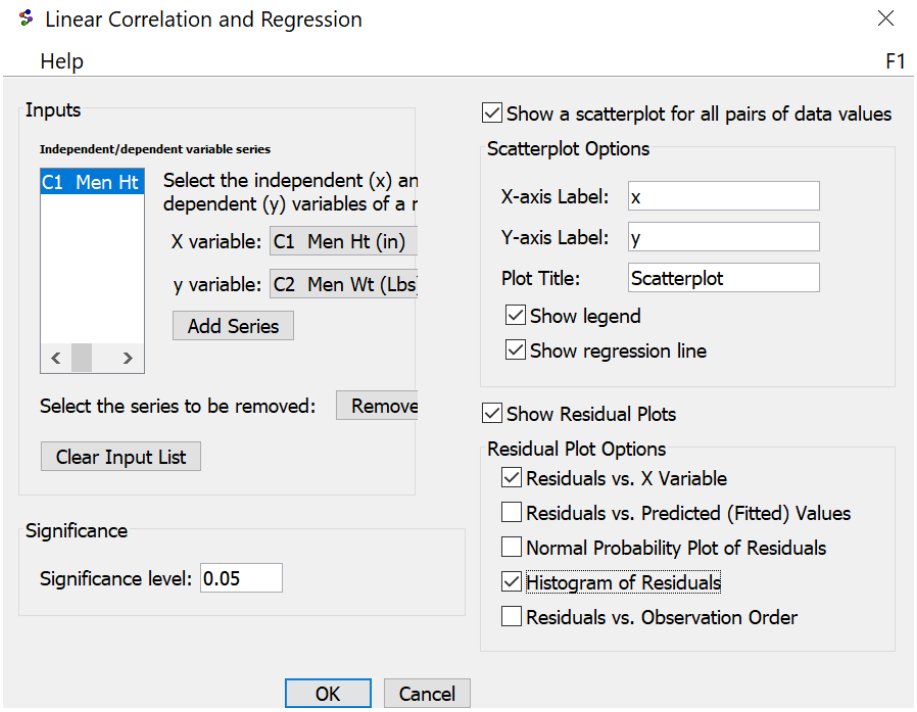
$\hat{Y} \approx -138.607 + 332.369$

$\hat{Y} \approx +193.762$

Therefore, we predict that the average weight of men that are 73 inches tall is about 193.8 pounds. Be careful of applying this prediction to all men. This data came from sample data and may not reflect the heights of all men on earth.

Calculating Correlation Graphs and Statistics with Statcato

We can also make scatterplots and calculate correlation statistics with Statcato. Copy and paste the men's height and weight data into two columns of Statcato. Go to the "statistics" menu, click on "correlation and regression" and then click on "linear". Click on the height to be the X-variable and the weight to be the Y-variable and then push "add series". Check the box that says "show scatterplot" and the box that says "show regression line". Statcato also has the capability of making residual plots. These are more advanced kinds of graphs that are studied in regression analysis. Check the box that says, "Show residual plots", the box that says "residuals vs x-variable", and the box that says "histogram of the residuals". Now push "OK".



Correlation and Regression: Significance level = 0.05

Series: C1 Men Ht (in),C2 Men Wt (Lbs)
x = C1 Men Ht (in)
y = C2 Men Wt (Lbs)
Sample size n = 40
Degrees of freedom = 38

Correlation:
H_0: ρ = 0 (no linear correlation)
H_1: ρ ≠ 0 (linear correlation)

	Test Statistic	Critical Value
r	0.5222	±0.3120
t	3.7750	±2.0244

p-Value = 0.0005

Regression:
Regression equation $Y = b_0 + b_1x$
b_0 = -138.6070
b_1 = 4.5534

Variation:
Explained variation = 7372.6464
Unexplained variation = 19659.0136
Total variation = 27031.66
Coefficient of determination r^2 = 0.2727
Standard error of estimate = 22.7452

Some of the information in this printout refers to the correlation hypothesis test that we will study in chapter five. Notice Statcato gave us the correlation coefficient r of 0.522 and the coefficient of determination r^2 = 0.2727 (27.27%). The slope is given as b_1 = 4.5534 and the Y-intercept is given as b_0 = −138.6070. Notice these are the same numbers as StatKey.

There is one statistic on the Statcato printout that was not on the StatKey printout that is important.

Standard error of estimate = 22.7452

This statistic is called the standard deviation of the residual errors (s_e). It measures the average vertical distance that points in the scatterplot are from the regression line. It also tells us the average prediction error for predictions made in the scope of the X-values. The units of the standard deviation of the residual errors is the same as the Y-variable (pounds). This statistic tells us the following.

The points in the scatterplot are 22.7452 pounds on average from the regression line.

If we use the regression line and the height of a man to predict the weight, our prediction could have an average error of 22.7452 pounds.

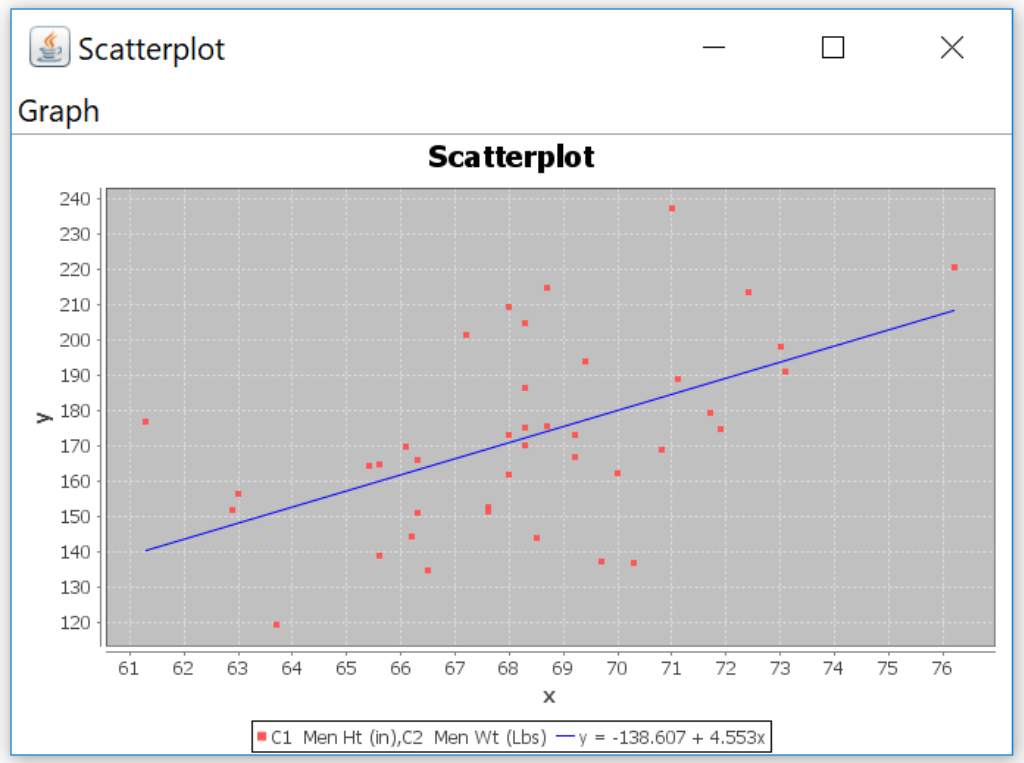
Remember the prediction we made earlier. We predicted that the average weight of men that are 73 inches tall is about 193.8 pounds. Well that prediction could be off by 22.7452 pounds on average.

A "residual" is the vertical distance that each point is from the regression line. Suppose a point has an ordered pair (X , Y). The point on the regression line with the same X value would have an ordered pair (X , $\hat{Y}$). To calculate a residual the computer subtracts the predicted $\hat{Y}$ value from the actual Y value of the point in the scatterplot. This gives the vertical distance that point is from the regression line.

Residual = $Y - \hat{Y}$

The standard deviation of the residual errors is an average of the residuals. The actual formula is shown below. Notice that we divide by n – 2 instead of n – 1 because there were two data sets. This again is called the degrees of freedom and will be discuss more in later chapters.

$$s_e = \sqrt{\frac{\sum(y - \hat{y})^2}{n-2}}$$



Notice Statcato also gave us a scatterplot of the data with the regression line drawn. The regression line formula is at the bottom of the graph.
