NOTES ON CORRELATION AND REGRESSION

1. Correlation

- Correlation is a measure of association between two variables. The variables are not designated as dependent or independent.
- The two most popular correlation coefficients are: **Spearman's correlation coefficient rho** and **Pearson's product-moment correlation coefficient**. When calculating a correlation coefficient for ordinal data, select Spearman's technique. For interval or ratio-type data, use Pearson's technique.
- The value of a correlation coefficient can vary from minus one to plus one. A minus one indicates a perfect negative correlation, while a plus one indicates a perfect positive correlation. A correlation of zero means there is no relationship between the two variables. When there is a negative correlation between two variables, as the value of one variable increases, the value of the other variable decreases, and vice versa.
- The **standard error of a correlation coefficient** is used to determine the confidence intervals around a true correlation of zero. If your correlation coefficient falls outside of this range, then it is significantly different from zero. The standard error can be calculated for interval or ratio-type data (i.e., **only for Pearson's** product-moment correlation).
- The significance (probability) of the correlation coefficient is determined from the **t-statistic**. The probability of the t-statistic indicates whether the observed correlation coefficient occurred by chance if the true correlation is zero. In other words, it asks if the correlation is significantly different than zero. When the t-statistic is calculated for Spearman's rank-difference correlation coefficient, there must be at least 30 cases before the t-distribution can be used to determine the probability. If there are fewer than 30 cases, you must refer to a special table to find the probability of the correlation coefficient.

2. Regression

- **Simple regression** is used to examine the relationship between one dependent and one independent variable. After performing an analysis, the regression statistics can be used to predict the dependent variable when the independent variable is known.
- The **regression line** (known as the *least squares line*) is a plot of the expected value of the dependent variable for all values of the independent variable. Technically, it is the line that "minimizes the squared residuals". The regression line is the one that **best fits the data** on a scatterplot.
- Using the **regression equation**, the dependent variable may be predicted from the independent variable. The slope of the regression line (b) is defined as the rise divided by the run. The y intercept (a) is the point on the y axis where the regression line would intercept the y axis. The slope and y intercept are incorporated into the regression equation. The intercept is usually called the constant, and the slope is referred to as the coefficient. Since the regression model is usually not a perfect predictor, there is also an error term in the equation.
- In the regression equation, *y* is always the dependent variable and *x* is always the independent variable.
- Here is a way to mathematically describe a linear regression model:
  $y = a + bx + e$
- The significance of the slope of the regression line is determined from the **t-statistic**. It is the probability that the observed correlation coefficient occurred by chance if the true correlation is zero.
- Some researchers prefer to report the F-ratio instead of the t-statistic. The F-ratio is equal to the t-statistic squared.
- **The t-statistic is equal to the estimated coefficient divided by its standard error**.

- The t-statistic for the significance of the slope is essentially a test to determine if the regression model (equation) is usable. If the slope is significantly different than zero, then we can use the regression model to predict the dependent variable for any value of the independent variable.
- On the other hand, take an example where the slope is zero. It has no prediction ability because for every value of the independent variable, the prediction for the dependent variable would be the same. Knowing the value of the independent variable would not improve our ability to predict the dependent variable. Thus, **if the slope is not significantly different than zero, don't use the model to make predictions**.
- The **coefficient of determination (r-squared)** is the square of the correlation coefficient. Its value may vary from zero to one. It has the advantage over the correlation coefficient in that it may be interpreted directly as the **proportion of variance in the dependent variable that can be accounted for by the regression equation**. For example, an r-squared value of 0.49 means that 49% of the variance in the dependent variable can be explained by the regression equation. The other 51% is unexplained.
- The **standard error of the estimate for regression measures the amount of variability in the points around the regression line**. It is the standard deviation of the data points as they are distributed around the regression line. The standard error of the estimate can be used to develop confidence intervals around a prediction.

3. Example

A company wants to know if there is a significant relationship between its advertising expenditures and its sales volume. The independent variable is advertising budget and the dependent variable is sales volume. A lag time of one month will be used because sales are expected to lag behind actual advertising expenditures. Data was collected for a six month period. All figures are in thousands of dollars. Is there a significant relationship between advertising budget and sales volume?

Ad budget	Sales
4.2	27.1
6.1	30.4
3.9	25.0
5.7	29.7
7.3	40.1
5.9	28.8

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.89
R Square	**0.80**
Adjusted R Square	0.75
Standard Error	2.64
Observations	6

ANOVA

					Significance
	df	*SS*	*MS*	*F*	*F*
Regression	1	109.095341	109.09534	15.689839	0.016663128
Residual	**4**	27.81299234	6.9532481		
Total	5	136.9083333			

		Standard		
	Coefficients	*Error*	*t Stat*	*P-value*
Intercept	9.9	5.239	1.9	0.133
x	3.7	0.929	4.0	0.017

Model: ***y = 9.9 + 3.7x + e***

Standard error of the estimate = 2.64

t-test for the significance of the slope = 4.0

Degrees of freedom = 4

Two-tailed probability = .0149

r-squared = .80

You might make a statement in a report like this:

A simple linear regression was performed on six months of data to determine if there was a significant relationship between advertising expenditures and sales volume. The t-statistic for the slope was significant at the 0.05 critical alpha level, t(4)=4.0, p=0.017. Thus, we **reject the null hypothesis** and conclude that there was a **positive significant relationship** between advertising expenditures and sales volume. Furthermore, **80% of the variability in sales volume could be explained by advertising expenditures**.

4. Durbin-Watson statistic

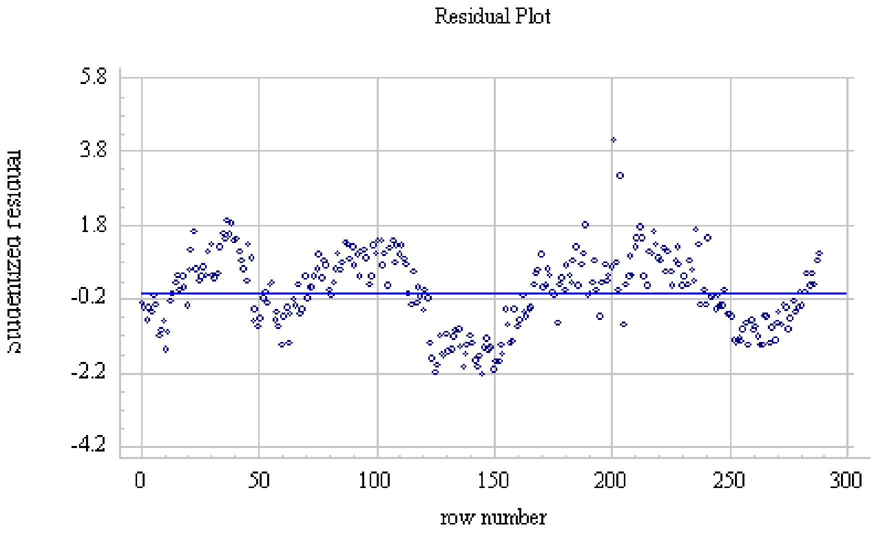
The DW statistic tests for the presence of significant autocorrelation--also known as **serial correlation--**at lag 1, and a "good" value for the DW statistic is something **close to 2.0**

$$DW = \frac{\sum_{2}^{n}(e_i - e_{i-1})^2}{\sum_{1}^{n}(e_i)^2}$$

Calculating DW in Excel

- Suppose residuals are in C1 through C50.
- Then use **=SUMXMY2(C2:C50,C1:C49)/SUMSQ(C1:C50)**
- The **SUMXMY2 function** sums the squares of X minus Y. Notice that two ranges lie in the parentheses following this function. The first range (here C2:C50) lists the X values and the second range (here C1:C49) lists the Y functions. So the first difference is C2-C1, the second is C3-C2, and the last will be C50-C49. There will be 49 differences which are squared and summed.

- The denominator employs the **SUMSQ functions** which squares the values in a range and then sums the squares. Here we square all 50 residuals, then sum the squares.
- As a very rough rule of thumb, you should be suspicious of a DW stat that is less than 1.4 or greater than 2.6.
- If the DW value is less than 1.4, there may be some indication of positive serial correlation. Plot the residuals versus row order to see if there is any pattern which can be seen.
- **Plotting residuals versus row number (i.e., versus time) is *always* a good idea** when you are dealing with time series data, and here is what the plot looks like in this case:



Yikes! There is a rather serious problem here: the **residuals clearly have a *very* strong pattern of positive autocorrelation**--notice the long runs of errors with the same sign.

5. Overall goodness of fit of the model

- The F test statistic (F) and its corresponding p-value (Significance F) indicate an overall goodness of fit for the model. A p-value of less than 1% (0.01) is considered highly significant.