Spreadsheets in Education (eJSiE)

Volume 7 | Issue 3 Article 2

June 2014

Parametric Statistics and the General Linear Model

John A. Rochowicz Jr
Alvernia University, john.rochowicz@alvernia.edu

Parametric Statistics and the General Linear Model

Abstract
Too many students acquire statistical knowledge and techniques independent of each other. The purpose of this paper is to illustrate the many connections mathematically between parametric statistics and the General Linear Model. With these various connections students will see that parametric statistical analyses are essentially one technique the General Linear Model

Parametric inferential statistics include t-tests of hypothesis, Analysis of Variance (ANOVA), correlation and regression. The ability of analyzing data in a variety of ways and using the General Liner Model (GLM) as a mathematical tool that can be applied for making these inferences are described. This paper shows connections between t-tests, ANOVA's, correlations, and regression. The capabilities of EXCEL to do mathematics, apply the Data Analysis Toolpak, find p-values for these tests and perform matrix algebra operations are demonstrated. Numerous ways of achieving the same results are displayed.

Keywords
The General Linear Model, Mathematical Statistics, Matrix Algebra, Inferential Statistics and Spreadsheets

Parametric Statistics and the General Linear Model

John A Rochowicz Jr

Professor of Mathematics

Alvernia University

Abstract

Too many students acquire statistical knowledge and techniques independent of each other. The purpose of this paper is to illustrate the many connections mathematically between parametric statistics and the General Linear Model. With these various connections students will see that parametric statistical analyses are essentially one technique the General Linear Model

Parametric inferential statistics include t-tests of hypothesis, Analysis of Variance (ANOVA), correlation and regression. The ability of analyzing data in a variety of ways and using the General Liner Model (GLM) as a mathematical tool that can be applied for making these inferences are described. This paper shows connections between t-tests, ANOVA's, correlations, and regression. The capabilities of EXCEL to do mathematics, apply the Data Analysis Toolpak, find p-values for these tests and perform matrix algebra operations are demonstrated. Numerous ways of achieving the same results are displayed.

Keywords: The General Linear Model, Mathematical Statistics, Matrix Algebra, Inferential Statistics, and Spreadsheets

Introduction

Many students as well as researchers study and apply inferential statistics with no understanding that certain procedures are mathematically the same process. Connections and generalizations can be made as more statistical concepts are studied, compared and analyzed. The application of various statistical tests of hypothesis to inferences can be performed in a number of different ways. But all these tests: the t-test of hypothesis, the one way Analysis of Variance (ANOVA), correlation and regression analysis can all be analyzed applying one procedure called the General Linear Model (GLM). The General Linear Model enables the student or researcher to unify a number of inferential tests of hypothesis with one technique.

Students study many undergraduate and graduate statistics courses without the understanding that all parametric tests of hypothesis: t-tests, ANOVA's, correlations and regression are part of a more general process. When students mention they never know what test to use, they lack the ability to recognize that a specific inference can be conducted in various ways and that the General Linear Model can be applied for all of them.

The aim of this paper is to illustrate these various connections and how to apply the GLM to the analysis of a variety of statistical techniques. From many classroom experiences and textbooks in the social sciences, students learn parametric tests of hypothesis independent of each other. The unifying features and power of the GLM is invaluable to the student and the researcher.

Any statistical analysis involving linear regression, correlation, ANOVA's and t-tests of hypothesis can be easily conducted in EXCEL through classical calculations, the Data Analysis Toolpak and Matrix Algebra. An example illustrating these approaches follows. Although there are limitations to using EXCEL for some statistical applications, this software can illustrate with mathematics all the necessary computations for successful analysis.

Parametric Statistics

A statistic is a numerical quantity that comes from a random sample of the population. The goal of inferential statistics is to infer, predict or estimate the population parameter from these statistics. Various statistics for inference include among others, the mean, median, mode, range, variance, standard deviation, correlation coefficient and the coefficients beta weights for linear regression. In order to make inferences about a population parameter a researcher applies tests of hypothesis. There are many inferential tests of hypothesis. The most common inferential tests of hypothesis include: t-Tests of Hypothesis, One Way ANOVA's, Correlation Analysis, and Linear Regression.

In order to present valid results and findings, assumptions must be considered when conducting parametric tests of hypothesis. For purposes of this paper the assumptions for the example provided will be assumed verified.

Tests of Hypothesis

In conducting inferential statistics the following steps are usually followed: a) State the null and alternative hypothesis b) Calculate the test statistic. Is there enough evidence to reject null hypothesis and determine that the differences do not depend on chance. c) Set the significance level. d) Determine the observed significance level p-value. e) Decide to reject or accept the null hypothesis.

Data Analysis Toolpak is an add-in for EXCEL that can conduct many types of hypothesis tests. Illustrations of the Toolpak are provided for t-tests of hypothesis, one way ANOVA's, Correlation and Regression. In order to use Toolpak it must be installed. The techniques presented here through EXCEL are suitable for illustrating their relationships to the GLM.

Matrix Algebra for Tests of Hypothesis

If the researcher or learner uses the GLM to conduct inferential statistics knowledge about Linear Matrix Algebra is necessary. Linear Algebra [2] is a mathematical field that allows the researcher to analyze many variables using matrices. Applying the GLM process to inferential statistics is accomplished by using Linear Algebra. The main operations for matrices in the application of the GLM include the abilities to multiply, invert and transpose matrices. These abilities can be performed in EXCEL.

The steps for multiplying two matrices in EXCEL are found at http://office.microsoft.com/en-us/excel-help/mmult-function-HP010342697.aspx The syntax for multiplying 2 matrices is "=MMULT(array1,array2)" In order to get the completed calculated product do not use enter but "ctrl+shift+enter.

The steps for inverting a matrix are at http://office.microsoft.com/en-us/excel-help/minverse-HP005209179.aspx. The syntax for inverting a matrix is

"=MINVERSE(array)". In order to invert a matrix it must be square and to calculate the entire matrix array use "ctrl+shift+enter".

The steps for transposing a matrix are at http://office.microsoft.com/en-us/excel-help/transpose-function-HP010069834.aspx The syntax for transposing a matrix is "TRANSPOSE(array)". In order to get the correct transposed matrix again use "ctrl+shift+enter". Recall that the transpose of a matrix interchanges rows and columns in a matrix.

Various tests of hypothesis can be conducted with the General Linear Model using Matrix Algebra.

EXCEL, Statistics and Matrix Algebra

In Excel all computing formulas can be constructed and applied. EXCEL can calculate all test statistics, t, F, r, and beta coefficients for regression. With Excel p-values for all tests of hypothesis can also be found and from them decisions to reject or accept the null hypothesis are made. The availability of the EXCEL add-in Data Analysis Toolpak also allows the user to conduct: two independent samples t-tests, one-way ANOVA's, correlation, regression without any actual calculations using the classical formulas. The techniques are automated and emphasis is placed on understanding and interpretation. The capabilities of EXCEL to perform matrix operations enhance one's learning experiences. All parametric statistics can be analyzed with the general linear model by way of matrix algebra. The techniques necessary and available in EXCEL for performing matrix operations for statistical analyses are: matrix multiplication, matrix inversion, and matrix transposition.

The Example for Analysis

The following grades on a certain academic standardized calculus II test were collected for 2 groups of students:

Group A

56	65	69	78	72	82	81	87	89	81

Group B

78	88	89	91	84	87	92	94	92	90

Group A represents students taking calculus II right after high school, ages 18 to 25 years old while Group B represents students ages 26 to 30. Is there a significant difference in the mean grades for the 2 groups of students? Is there a relationship between the two types of calculus II students? Test at alpha = 0.05 This example is analyzed with the following techniques: a) t-test; b) 1way ANOVA; c) Correlation; d) Linear Regression; e) Data Analysis Toolpak; f) Matrix Algebra/Linear Models

Inferential Technique 1: t-Tests of Hypothesis

The t-test of hypothesis is an inferential test of hypothesis that determines whether there is a difference in the means of two independent or dependent sample means. The question of interest is: Is there a difference in the mean test grades for the two groups of students on a calculus II exam?

The t-test of hypothesis can be conducted by classical calculations. If the variances are known the t statistic is found by applying:

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - 0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Note that $\bar{x}_1 \; and \; \bar{x}_2$ are the sample means, $n_1 \; and \; n_2$ are the sample sizes and $\sigma_1^2 \; and \; \sigma_2^2$ are the variances for the populations.

Figure 1 illustrates EXCEL cell formulas for this t-test of hypothesis. In figure 2 the EXCEL results are shown. The t-statistic is -3.47804 and the p-value is 0.002684. In order to find the p-value use the command TDIST with a positive t-value.

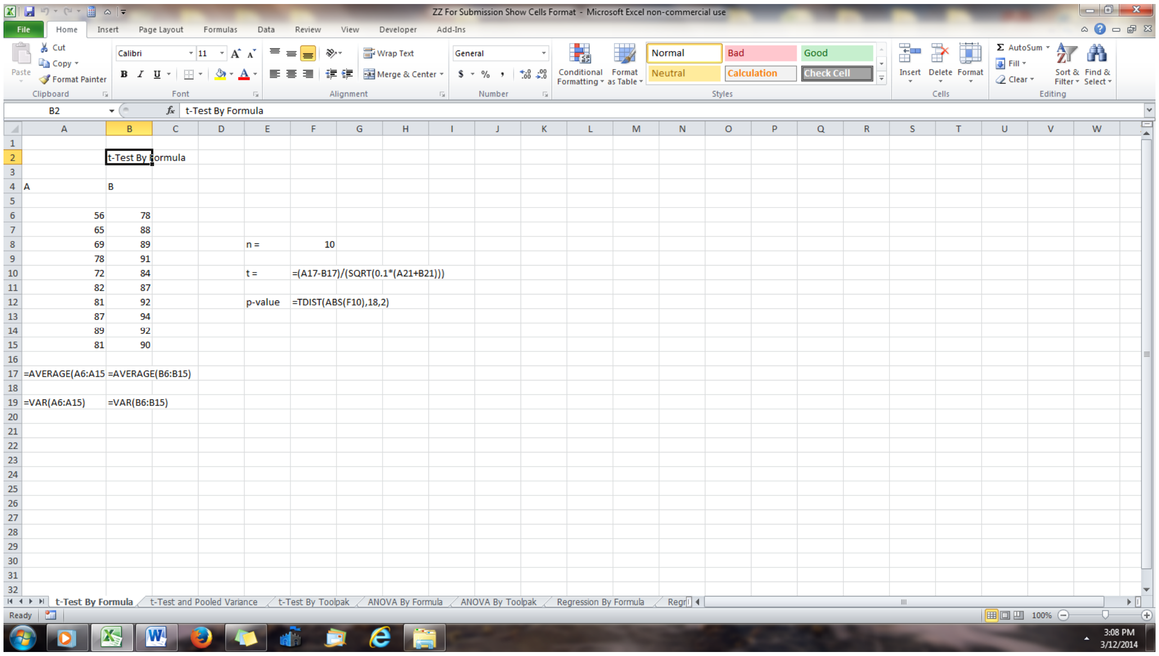

Figure 1: t-Test of Hypothesis with EXCEL Formulas

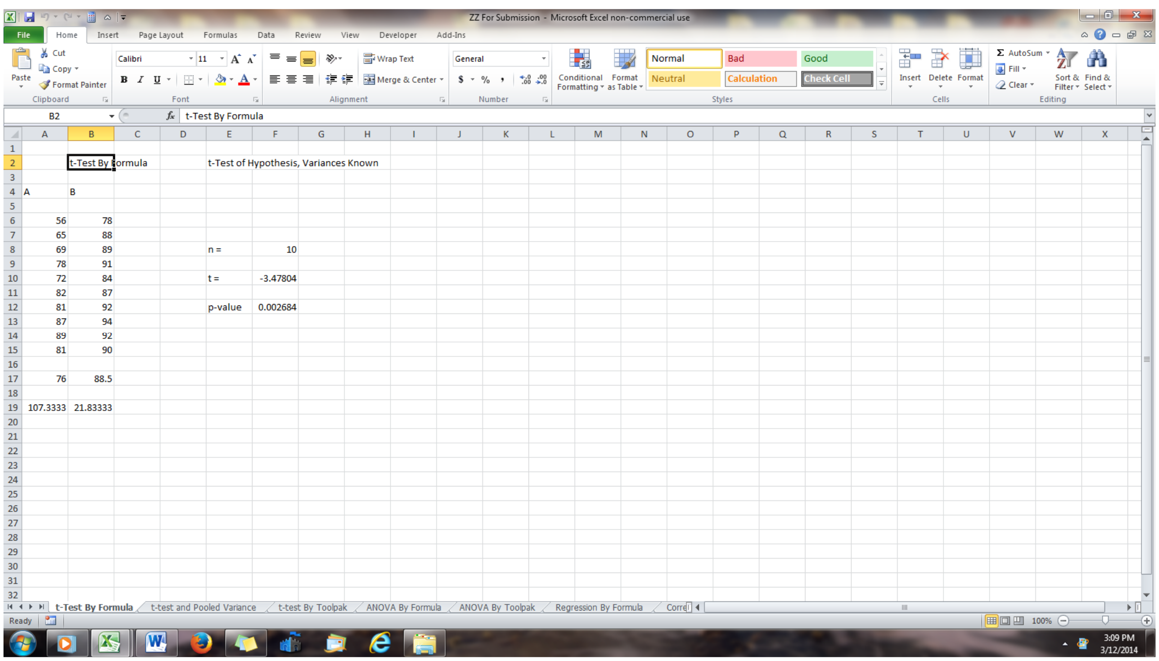

Figure 2: t-Test of Hypothesis with EXCEL Results

Inferential Technique 2: t-Test of Hypothesis, Variances Unknown

The t-test of hypothesis for two groups can be conducted if the variances are unknown by pooling the variances and calculating

$$s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2},$$

where n_1 and n_2 are the sample sizes and s_1^2 and s_2^2 are the sample variances for each group. The t statistic is then calculated by applying:

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - 0}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Figure 3 illustrates the EXCEL cell formulas for t-Test of Hypothesis, Variances Unknown. Figure 4 displays the results of this t-test. The t-statistic is -3.47804 and the p-value is 0.002684.

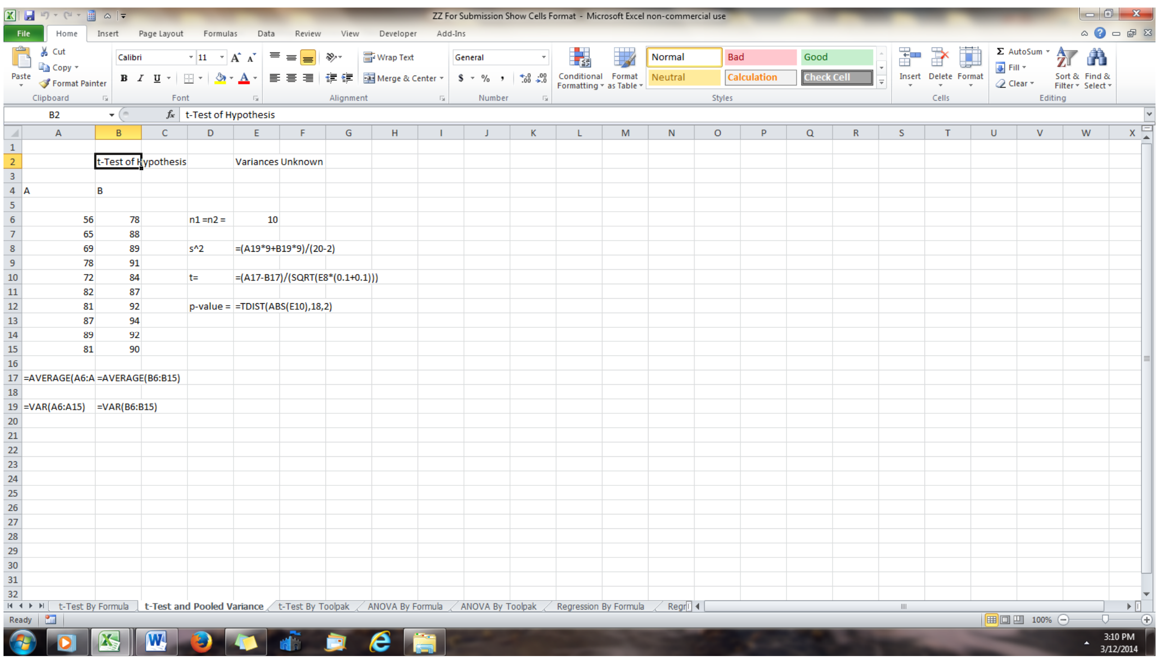

Figure 3: t-Test of Hypothesis with EXCEL Formulas, Variances Unknown

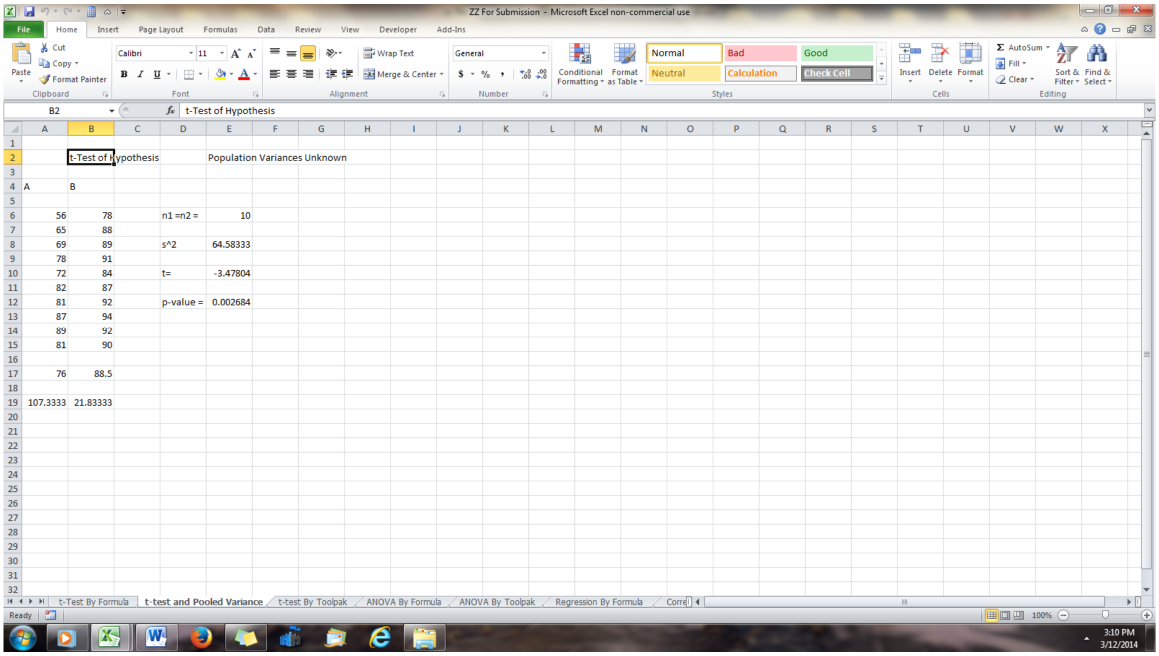

Figure 4: t-Test of Hypothesis with EXCEL Results, Variances Unknown

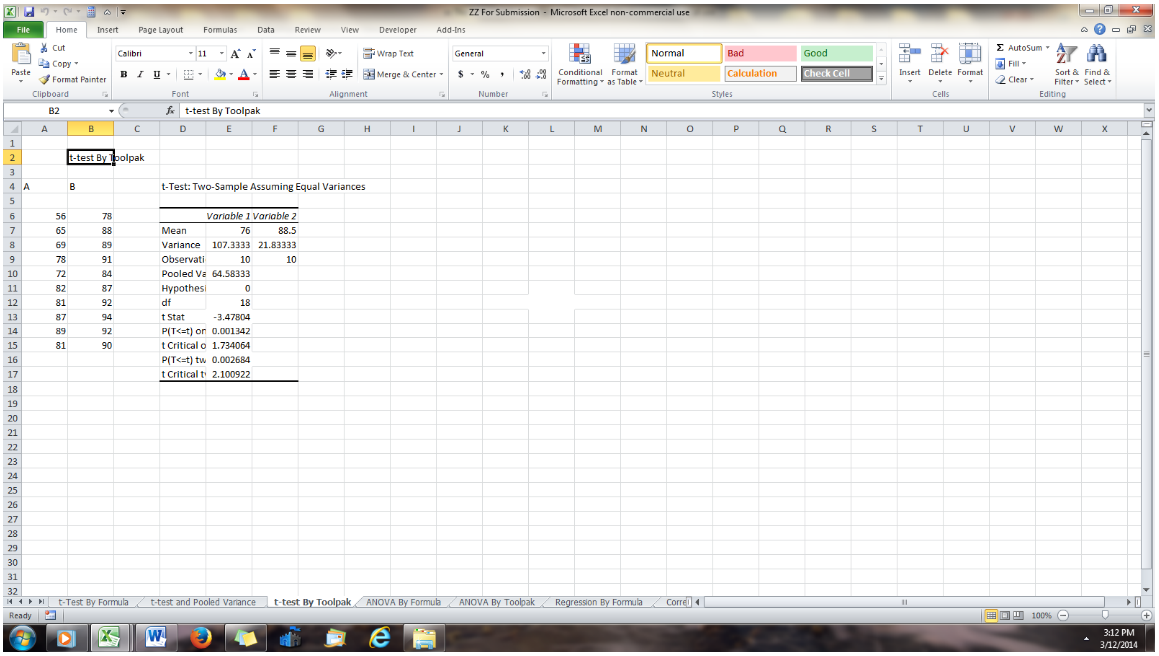

Figure 5: t-Test of Hypothesis with Data Analysis Toolpak, Variances Unknown

The results are same regardless of the technique. And since the p-value is less than or equal to alpha of 0.05 there is a significant difference between the mean grades for the different age groups. Using the Data Analysis Toolpak also displays the same t-statistics and p-value.

Inferential Technique 3: One-Way ANOVA

In a one-way ANOVA random samples are selected from 2 or more than 2 groups and the goal is to determine if there is a difference in population means for these groups. Consider the same example for the two groups of students. The ANOVA can be calculated [1] or analyzed by the Data Analysis Toolpak.

The calculations for the one-way ANOVA involve:

a) Find the sums of each variable: $\sum x_1, \sum x_2$.
b) Calculate the squared sums $(\sum x_1)^2$ and $(\sum x_2)^2$.
c) Calculate the means of each group.

The ANOVA process breaks into sections called the sum of squares between groups (SSB), the sum of squares within (error) groups (SSW) and the sums of squares for the total variance (SST). The formulas are:

SST = $\sum X^2 - \frac{(\sum X)^2}{N}$ where X represents all data elements and N represents the total number of observations. The degrees of freedom are N-1.

SSB = $\frac{(\sum x_1)^2}{n_1} + \frac{(\sum x_2)^2}{n_2} + \ldots\ldots \frac{(\sum x_n)^2}{n_n} - \frac{(\sum X)^2}{N}$ where X represents all data elements and N is the total number of observations. The degrees of freedom are number of groups minus 1.

SSW = SST-SSB

The mean squares are measures of SS terms divided by the degrees of freedom for each section of the ANOVA. MST is the variance of all the data elements in the example and all other values are calculated with the following formulas:

$$MSW = \frac{SSW}{dfw}; \; MSB = \frac{SSB}{dfb}; \text{ and } F = \frac{MSB}{MSE}$$

All these formulas are implemented in Figure 6. The degrees of freedom for numerator are the number of groups minus 1 (cell 15) and the degrees of freedom for the denominator are number of items minus 2 or 18 (cell I16). The total degrees of freedom are 20 minus 1 or 19 (cell I17).

The F statistic is 12.09677 and the p-value is 0.002684. In order to find the p-values with the one-way ANOVA the "=FDIST()" command must be used. The command uses the degrees of freedom for the numerator and the degrees of freedom for the denominator. In this example the value in cell K15 with 1 and 18 degrees of freedom respectively are used. Figure 7 displays the EXCEL results from these formulas and Figure 8 displays the results with the Data Analysis Toolpak.

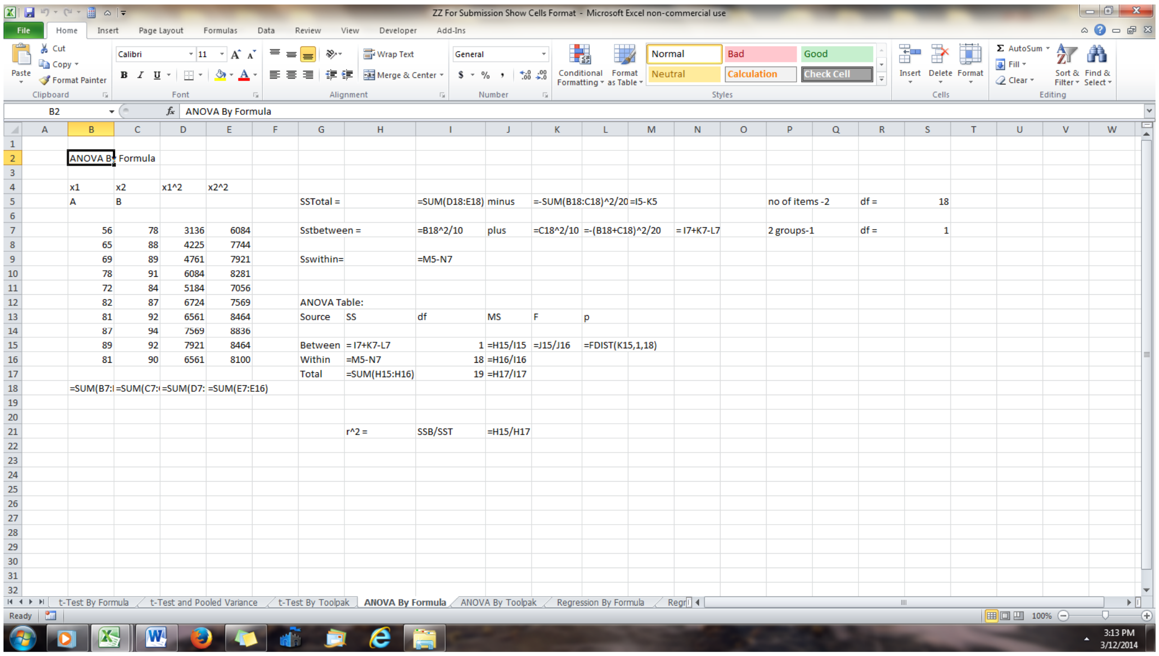

Figure 6: ANOVA with EXCEL Formulas

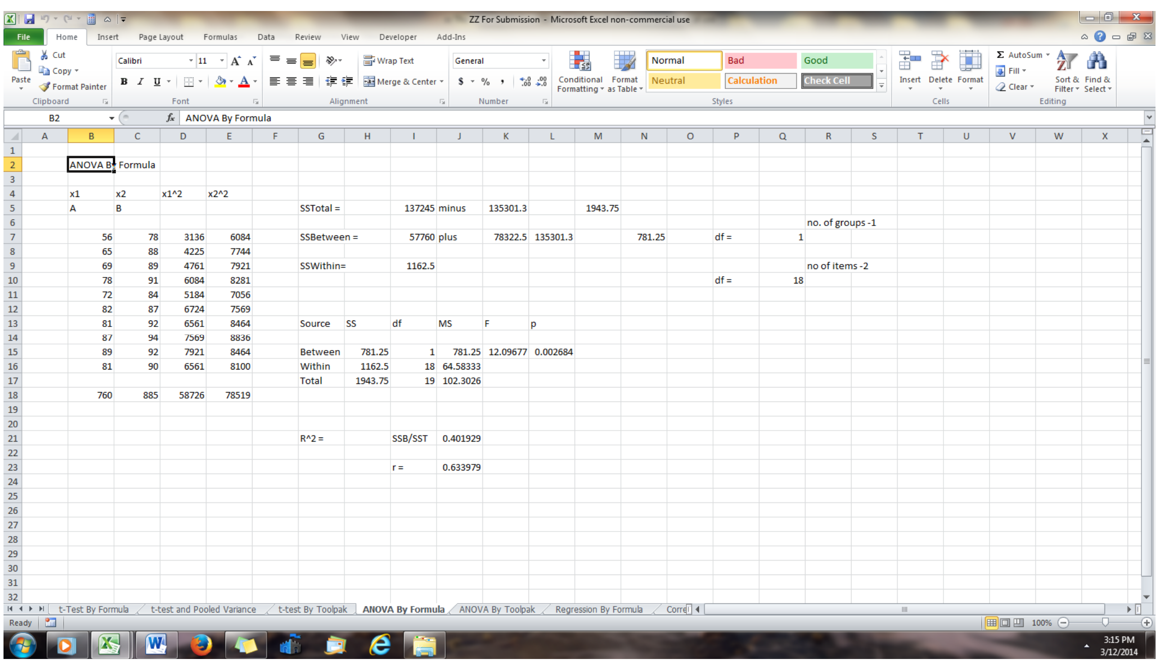

Figure 7: ANOVA with EXCEL Results

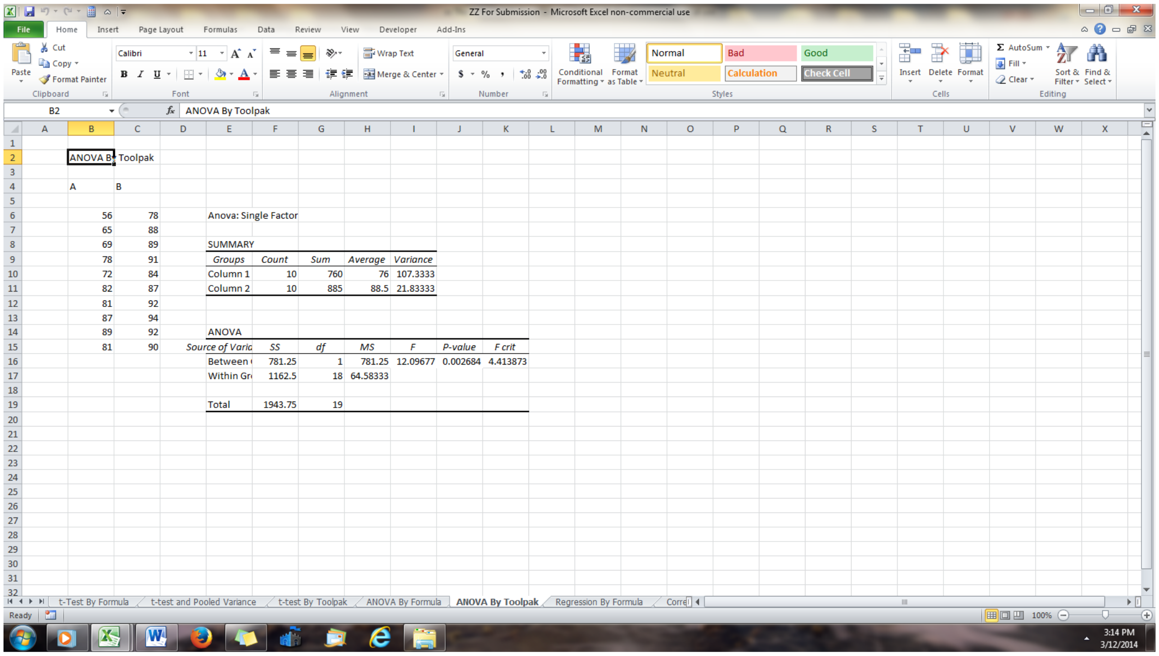

Figure 8: ANOVA with Data Analysis Toolpak

Again the results indicate that since the p-value is less than or equal to alpha of 0.05 there is a significant difference between the mean grades for the two groups. The p-value is the same as it was for the t-test.

The one-way analysis of variance or F-test of hypothesis is a test of hypothesis for determining if there is a difference in means between two or more groups. The researcher can accomplish the same results for an independent samples t-test as a one–way ANOVA as shown above. Note that $t^2 = F$

Inferential Technique 3: Linear Correlation and Regression

Consider the same example and analyze the relationship with the Pearson correlation. There are connections between t and r.

The t-statistic and r Pearson correlation are related by:

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

Solving for r gives:

$$r = \frac{t}{\sqrt{n-2+t^2}}$$

The r^2 (coefficient of determination) measures how data fit a linear relationship and is also determined by the ANOVA calculations above by:

$$r^2 = \frac{SSB}{SST}$$

The example can also be analyzed by applying linear correlation or regression. In order to do so the data has to be coded. Let x be 1 if the score is a member of the first group and a 0 if the score is in the second group. See Figure 9 for coding.

Considering the same example above, a researcher wants to determine if there is linear relationship between the two variables x and y. The goal is to determine if there is a correlation and or if there is a linear relationship. The approach is usually conducted by finding the correlation coefficient and the least squares linear equation or line of best fit [3].

The results are found by using the formulas of statistics [3]. In EXCEL these computations can be easily determined. The correlation is found by the calculating formulas: $\sum x, \sum y, \sum xy, \sum x^2, \sum y^2,$

In order to determine the strength of linearity the researcher calculates the Pearson correlation coefficient. This value is found by calculating

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}$$

where the $terms\ S_{xy}\ S_{xx}\ and\ S_{yy}$ are:

a) $S_{xx} = \sum x^2 - \frac{(\sum x)^2}{n}$

b) $S_{yy} = \sum y^2 - \frac{(\sum y)^2}{n}$

c) $S_{xy} = \sum xy - \frac{\sum x \sum y}{n}$

d) $r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}$

e) The beta coefficients are $b_0 = \frac{1}{n}(\sum y - b_1 \sum x)$ and $b_1 = \frac{S_{xy}}{S_{xx}}$

Figure 9 displays the EXCEL formulas for these calculations. Figure 10 displays the EXCEL results for Linear Correlation and Regression. Figure 11 shows the correlation results with Data Analysis Toolpak and figure 12 displays the linear regression analysis with the Toolpak.

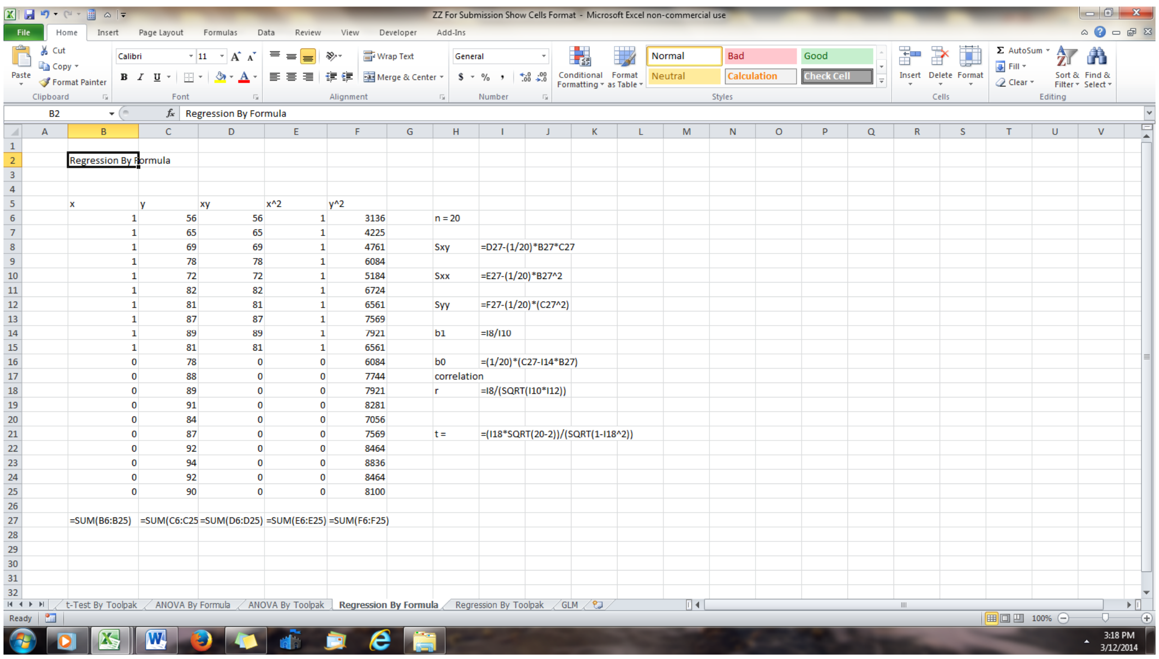

Figure 9: Correlation and Linear Regression with EXCEL Formulas

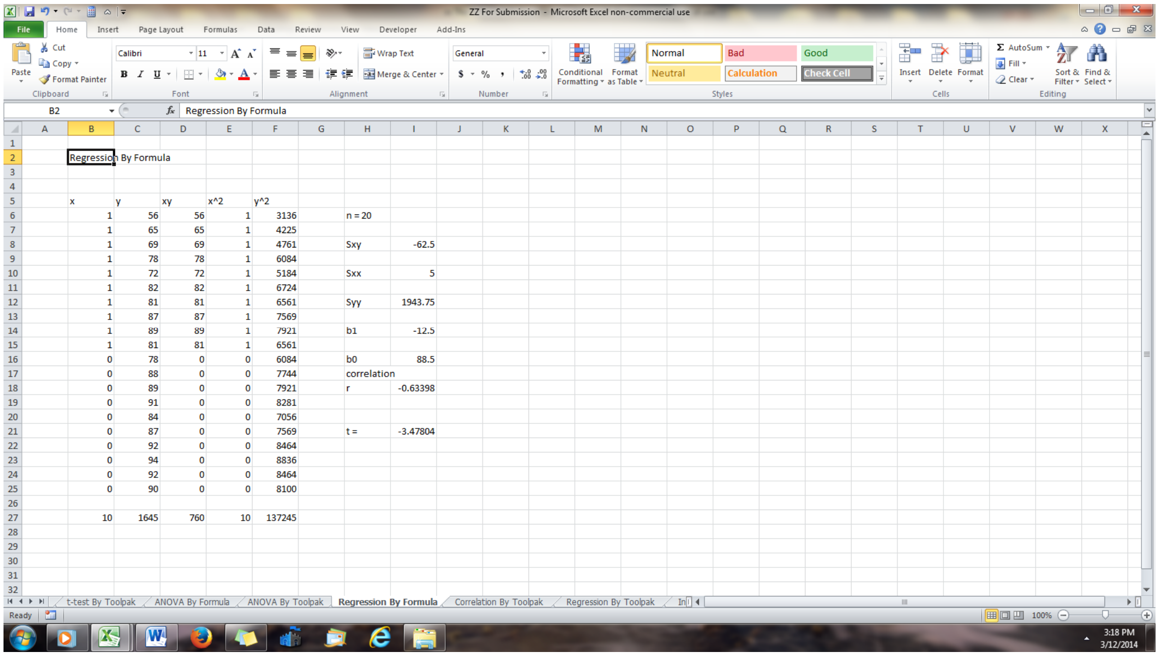

Figure 10: Correlation and Linear Regression with EXCEL Results

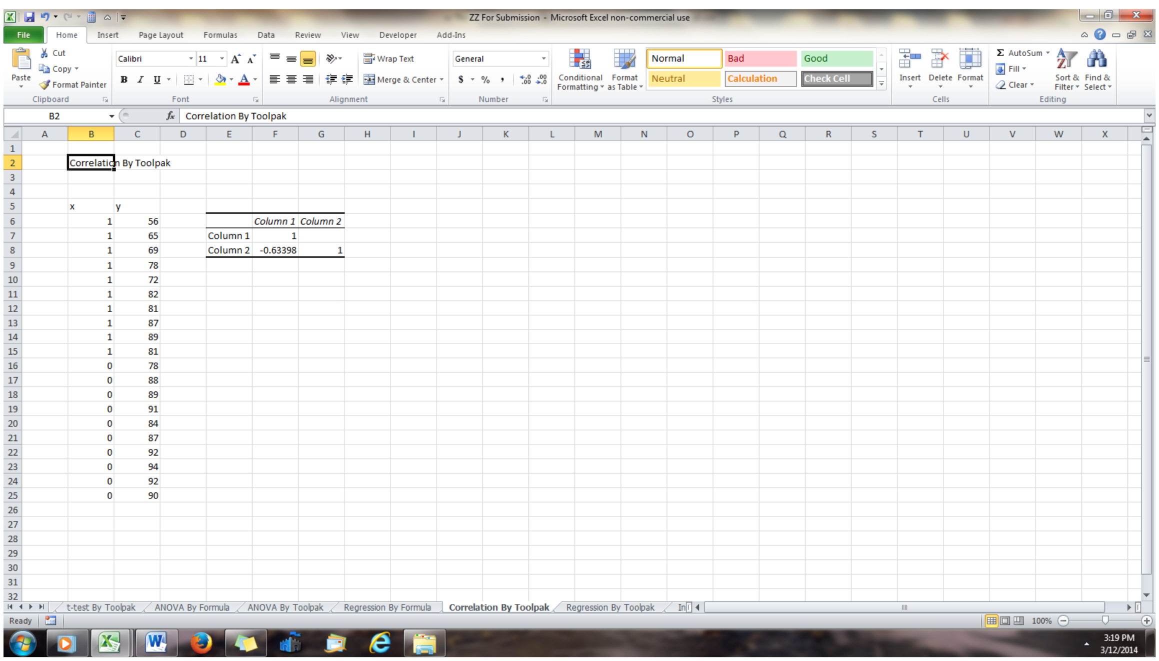

Figure 11: Correlation with Data Analysis Toolpak

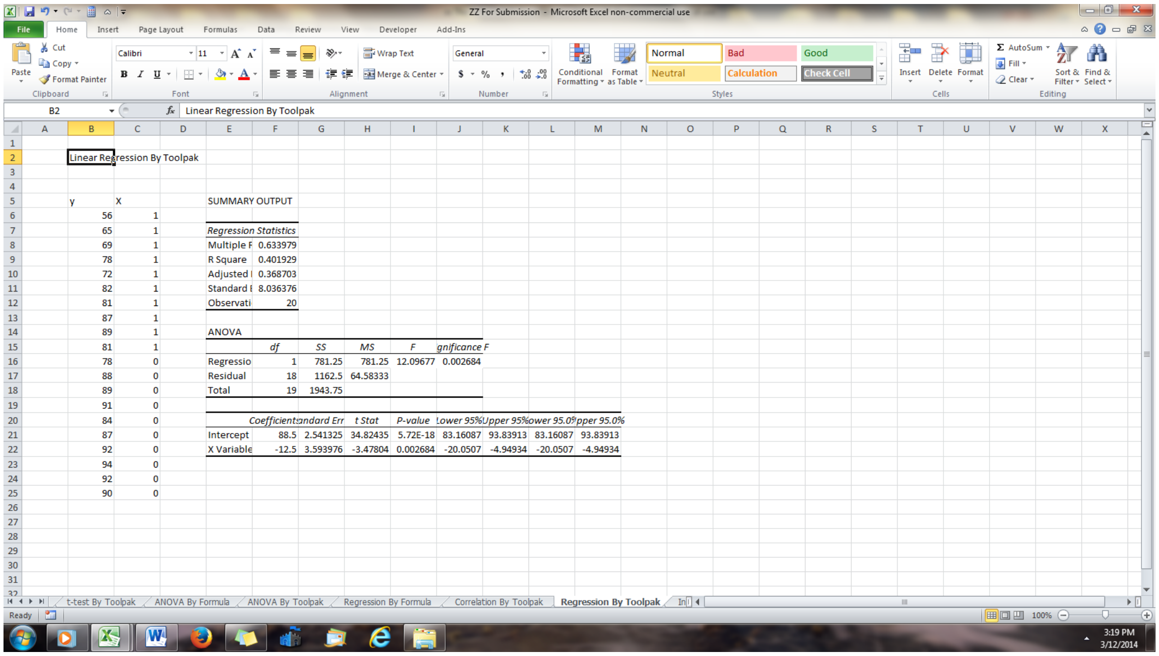

Figure 12: Linear Regression with Data Analysis Toolpak

The Pearson correlation is -0.63398. The values for the coefficients b_0 and b_1 are respectively 88.5 and -12.5. The t-statistic is calculated from the r shown above and is -3.47804. The p-value is the same as before 0.002684.

Inferential Technique 4: The GLM

The General Linear Model is the method that enables any researcher or learner to conduct parametric inferences by applying matrix algebra. In constructing these matrices the researcher sees the unifying of mathematics behind all parametric tests of hypothesis: t-tests, 1way ANOVAS, correlations and linear regression analysis. As shown above all the different tests of hypothesis presented yield the same results. Through the use of matrix algebra all these calculations are presented again. The value of GLM is illustrated on the same scores for the different age groups.

Matrix Algebra and the General Linear Model (GLM)

In studying statistics, researchers are interested in determining relationships between usually many variables, the design variables (independent) x's and the response variables (dependent) y's. Since this paper addresses linear relationships among numerous variables the use of matrix notation is appropriate. Matrices are tables of numbers or variables that allow the user to indicate how a statistical design can be analyzed

This equation **Y= Xb + e** represents the basic experimental designs described above, where **Y** is the matrix of response or dependent variables and **X** is the matrix of design or independent variables. The matrix notation is:

$$\begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_{11} \cdots & X_{1s} \\ \vdots & \vdots & \vdots \\ 1 & X_{n1} \cdots & X_{ns} \end{bmatrix} \begin{bmatrix} b_0 \\ \vdots \\ b_s \end{bmatrix} + \begin{bmatrix} e_0 \\ \vdots \\ e_n \end{bmatrix}$$

Consider the expression **Y = Xb** where **e** is negligible**.** The b's are the beta weights or coefficients of the independent variables x.

Solving for **b** in the expression using matrix algebra involves the following: First make **X** a square, symmetric matrix by multiplying both sides of the equation by the transpose of **X** or **X'**, that is **X'Y = X'Xb**

As a result a square, symmetric matrix is found. If **X'X** has an inverse, **(X'X)**$^{-1}$, then multiply both sides by this inverse matrix to get

$$(\mathbf{X'X})^{-1}\,\mathbf{X'Y} = (\mathbf{X'X})^{-1}\,(\mathbf{X'X})\mathbf{b}$$

Since a matrix multiplied by its inverse is the identity, this product is

$$(\mathbf{X'X})^{-1}\,\mathbf{X'Y} = \mathbf{Ib} = \mathbf{b}$$

The calculations [4] in parts for each matrix product are shown here:

$$\mathbf{X'X} = \begin{bmatrix} 1 & 1 \cdots & 1 \\ x1 & x2 \cdots xn \end{bmatrix} \begin{bmatrix} 1 & x1 \\ 1 & x2 \\ \vdots & \vdots \end{bmatrix} = \begin{bmatrix} n & \sum x \\ \sum x & \sum x^2 \end{bmatrix}$$

$$\mathbf{X'Y} = \begin{bmatrix} \sum y \\ \sum xy \end{bmatrix}$$

$$\boldsymbol{b} = (\boldsymbol{X'X})^{-1}\mathbf{X'Y}$$

$$\boldsymbol{b} = \begin{bmatrix} b_0 \\ b_1 \end{bmatrix}$$

$$SSE = S_{yy} - b_1 S_{xy} \text{ for Linear Regression}$$

$$\boldsymbol{SSE = Y'Y - b'X'Y} \textbf{ for GLM}$$

EXCEL Capabilities for the General Linear Model

Since the independent variable X is linear, a column of ones must be entered first. In the second column for X enter 1's if the score is in group 1 and a 0 otherwise. The dependent variable Y scores are entered as a 20 by 1 matrix and the independent variable X is entered as a 20 by 2 matrix. The transpose matrices X' and Y' are found by using "=TRANSPOSE(array)". The transposed matrices are 1 by 20 matrix and 2 by 20 matrix respectively. IN order to multiply matrices they have to be compatible. Multiplying a 20 by 1 matrix by a 1 by 20 produces a 1 by 1 matrix. The details are displayed in the figure. Similarly the products of the required matrices for the GLM are conducted. In order to multiply 2 matrices the command "=MMULT(array1, array2)" is applied. As shown in the figure the inverse of a matrix is also necessary. This is accomplished by using the command "=MINVERSE(array)." The results are displayed in the figure.

EXCEL Cell Formulas for the General Linear Model

The GLM formulas are implemented in EXCEL with the following cell formulas and details:

The transpose of the X matrix, X' is found by using the instructions provided above. The resulting cells display an array with the syntax "{=TRANSPOSE(D2:E21)}".

The transpose of the Y matrix, Y' is found similarly and the resulting array is shown with syntax "{=TRANSPOSE(C2:C21)}".

To find X'X use the multiplication of 2 matrices and get an array syntax "{=MMULT(H2:AA3,D2:E21)}" shown in cells J8 J9 K8 K9.

To find X'Y use the multiplication of 2 matrices and get an array syntax "{=MMULT(H2:AA3,C2:C21)}" in cells P8 and P9.

Finding the X'X inverse matrix applies the technique of finding inverses of matrices defined above. The array syntax "{=MINVERSE(J8:K9)}" is shown in cells J10, J12, K11 and K12.

Finding the Y'Y matrix again applies the multiplication of 2 matrices. The array syntax is "{=MMULT(H6:AA6,C2:C21)}"and shown in cell P11.

The coefficients $b_i's$ are shown in cells K16 and K17.

The SSE term is displayed in cell O18 and the variance s^2 term is calculated using SSE/df or SSE/18 and shown in cell I20. The square root of the variance is s and shown in cell I22. With these calculations the t-statistic is found using the formula:

$$t = \frac{b_1}{s\sqrt{\frac{1}{Sxx}}}$$

The excel formula is "=K16/(I22*SQRT(K12))" and is displayed in cell M20. The calculated results are shown in Figure 13. Again, the t statistic is -3.47804 and the p-value is 0.002684.

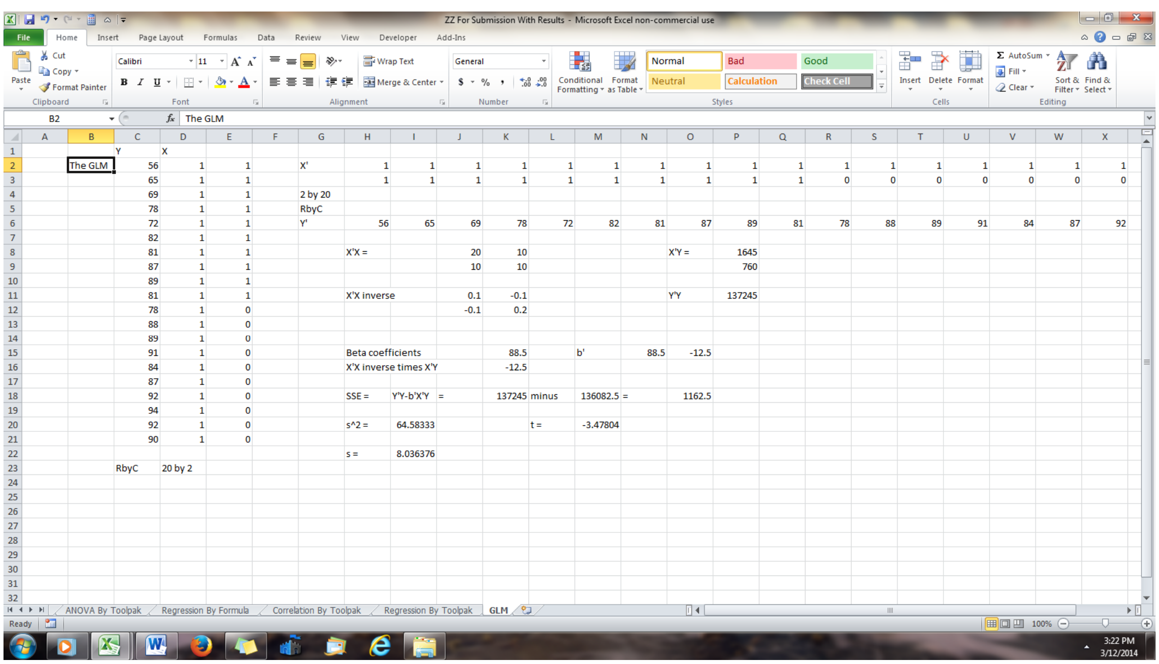

Figure 13: The General Linear Model with EXCEL Results

The goal in using the GLM is to find the beta coefficients that allow the data to fit the best linear model. All the figures in this paper illustrate the relationships between the t statistic, F statistic, correlation coefficient and the coefficients for the regression model.

Conclusions

Linear Models are not usually discussed in elementary statistics textbooks and courses. But there are many parametric statistics techniques that can be performed through the use of matrix algebra. The goal of the General Linear Model is to generate a linear combination of the x (independent) variables for one or more dependent variables y. The concept of linear models can be extended to multivariate where there are many independent and many dependent variables. There are some limitations in using EXCEL for statistical analysis, but the goal of this paper is to illustrate t-tests of hypothesis, one-way ANOVA's, correlation, regression and the relationships that are found using these various parametric statistical methods through matrix algebra. Many students and professors are not aware of the many possible connections.

Students and researchers implement various parametric tests of hypothesis without the knowledge that techniques such as t-tests, ANOVA's, correlation, and regression can all be analyzed through the application of the general linear model. Becoming aware of such possibilities provides the learner with the realization that there is really not one specific test of hypothesis that can answer any specific research question. Any parametric technique can be applied as long as the user can prepare the software for data analysis. When applying the GLM, interpretation and matrix understanding are necessary. Also the student or researcher should recognize the importance of inferential analysis in the determination of relationships that exist. The capabilities of EXCEL to apply computing formulas, use Data Analysis Toolpak, and calculate matrices are invaluable to statistical inference. Also through the application of software and especially spreadsheets concepts can be illustrated in many ways without reliance on calculations.

Learning the importance of the GLM to Parametric statistics becomes evident when students pursue specific situations and approach the situation with different techniques. Emphasis on "statistical literacy" is increased with the application of the GLM.

Inferential statistical analysis can be very complex. Technologies relieve the computational burden from the student. Focus is placed on interpretation and design.

Does it really matter which test of hypothesis is used? All the tests of hypothesis discussed can be analyzed in many ways. The general linear model generalizes many concepts into one main technique.

References

[1] Cohen, J. (1968). *Multiple Regression as a General Data-analytic System*. Psychological Bulletin, 70, 426-443.

[2] Leon, Steven J. ( ) *Linear Algebra, With Applications (6th Edition)*. Prentice Hall Publishing Co. Upper Saddle River, NJ.

[3] Mendenhall III, William, Beaver, Robert J and Beaver, Barbara M. (2013) *Introduction to Probability and Statistics* (14th Edition). Brooks/Cole Cengage Learning, Boston, MA

[4] Wackerly, Dennis D. Mendenhall III, William and Schaeffer, Richard L. (1996) *Mathematical Statistics with Applications 5th edition*. Wadsworth Publishing Co. Belmont, CA.