**Hypothesis Testing on the Coefficients in a Linear Regression Line**

**Chung Nguyen**

 Scientists or statisticians are often interested in the relationships between one variable with the other variables that are present within their data. “Regression analysis is the statistical method for investigating such relationship” (Yan and Su 1). Regression analysis is a common method in the world of mathematics, dating back to about two hundred years. There are three types of regression, simple linear regression, multiple linear regression, and nonlinear regression. However, the focus of this research paper will lean towards simple linear regression.
 The standard notation that is used in simple linear regression is $y= β\_{0}+ β\_{1}x+ ε$. β0, or beta sub zero, represents the predicted y-intercept of the data set where as β1 generally depicts the slope of the data set with respect to the variable x. $ε$, or epsilon, is the random error and is usually “difficult to discover since it changes for each observation Y” (Draper and Smith 11). Simplifying the standard notation in some situations is a good way to avoid slight confusion. The notation now takes the form of $y= β\_{0}+ β\_{1}x$. Generally, $β\_{1}$ and $β\_{0}$ are called the *parameters* of a linear regression model (Draper and Smith 10).
 There are many methods of actually testing the coefficients in simple linear regression equation. However, one must understand how to apply the idea of simple linear regression before testing the coefficients. One of the common methods to achieve this is the least squares estimation. The principle of the least squares estimation with reference to the simple linear regression model is to discover “estimates for b0 and b1 such that the sums of the squared distance from the actual response yi and predicted $\hat{y}\_{i }= β\_{0}+ β\_{1}x\_{i} $ reaches the minimum along all possible choices of regression coefficients $β\_{1}$ and $β\_{0}$” (Yan and Su 10). This method is also called the sum of squares of deviations from the predicted line based on a data set.
 Ŷ denotes the predicted value for Y for a given X when $β\_{1}$ and $β\_{0}$ are determined. Suppose there is an *n* set of observations in the form of (X1, Y1 ), (X2 , Y2 ), (X3 , Y3 ), …, (Xn , Yn ). The standard notation for these observations is $y\_{i}= β\_{0}+ β\_{1}x\_{i}+ ε\_{i}$. Generally, the equation that denotes the sum of squares is shown as $s^{2}= \sum\_{i=1}^{n}(Y\_{i}- \overbar{Y})^{2} $. This equation represents “the sum of the square of deviations from all the observations, *Yi* , from their mean, $\overbar{Y}$” (Weibull.com). The sum of squares could also be abbreviated as SST, which stands for the total sum of squares. Using this idea, one could estimate the variance of a population using the sample variance calculated through $s^{2}= \frac{SS\_{T}}{n-1}$ (Weibull.com).
 The denominator in the sample variance is called degrees of freedom. The number of degrees of freedom represents the number of parameters which may be independently varied based upon the data set. The number of degrees of freedom associated with the sample variance is therefore *n – 1*. Another name for sample variance is the *mean square*.
 When one is applying these ideas to fit a regression model from their observations, they are attempting to explain the variation of those observations. If the model had passed through all of the observation points when plotted, then the model is described as a perfect model. However, in most cases, the regression model will not be a perfect model. In these cases, there are *regression sum of squares*, which is abbreviated as *SSR*. In a perfect model, the regression sum of squares, or *SSR*, is equal to the total sum of squares, or SST , since the model had explained all of the observed variance. A visual take of this perfect model might look something similar to this graph.
 
 The calculation of the sum of squares due to regression is similar to that of the total sum of squares. Rather than using the actual observed data, the predicted value is used within the parameters of the equation, $SS\_{R}= \sum\_{i=1}^{n}(\hat{Y\_{i}}- \overbar{Y})^{2}.$ “The number of degrees of freedom associated with SSR is one” (Weibull.com). To put these ideas visually, here is a graph of all the components to a least squares estimation taken from a video by ZedStatistics.
 
 The unexplained variance is denoted as ei which could be calculated by taking the sum of the squares of each observed Y subtracted by the predicted Y value from the regression model, $SS\_{E}=\sum\_{i=1}^{n}e\_{i}^{2}= \sum\_{i=1}^{n}(Y\_{i}-\hat{Y}\_{i})^{2}$. The total variability could now be calculated by adding the sum of squares due to residuals or errors with the sum of squares due to regression, $SS\_{T}=SS\_{R}+SS\_{E}$. This equation is also known as the analysis of variance, or ANOVA. ANOVA consists of degrees of freedom, sum of squares, and the mean squares. This is a chart that displays the deviations for the three sums of squares.

 The three sums of squares could be useful in another statistical technique called hypothesis testing. Hypothesis testing provides evidence in either the rejection or acceptance of a hypothesis that was made. Generally, hypothesis testing includes an initial hypothesis, which is also referred to as a null hypothesis, and a hypothesis that is the opposite of the initial hypothesis, which is usually referred to as an alternative hypothesis. The null hypothesis is often noted as H0 where as the alternative hypothesis is noted as H1 or Ha.
 For example, if one were to conduct an experiment who’s goal was to prove that during the summer, temperatures in southern Massachusetts are generally higher than temperatures in eastern Massachusetts. Then the null hypothesis of this experiment could be that H0 = Higher temperatures in southern Massachusetts. Whereas the alternative hypothesis is that H1 = lower temperatures in southern Massachusetts or higher temperatures in eastern Massachusetts.
 To prove the null hypothesis, data is then collected and analyzed. Using the example above, I have made up a set of data measured in degrees Fahrenheit to help us decide whether or not to reject the null hypothesis. Using excel, I graphed these temperature measurements against each other to try and show some kind of relationship between them.

Shown in (a) is the perfect regression model of an observed data set. The linear regression model met all of the plotted observed data, hence explaining all of the variance from the data set.

Shown in(b) is the more common case when dealing with simple linear regression models. The scatter plot shows that this regression model is not perfect like that of (a). However, it is close to perfect since the regression model is only a prediction of the relationship between the variables.

(Weibull.com)

**----------** = $\overbar{Y}$ or The mean of the observed data set
**--------** = $\hat{Y}$ or the predicted Y values with respect to X
**--------** = SSR , the sum of squares due to residual or the explained variance
**--------** = SSE , The residual sum of squares, the error sum of squares, or the unexplained variance

(a) shows the mean of observed response values with the observed data plotted and the deviations for SST
(b) shows the fitted regression line and the deviations for SSR
(c) shows the deviations for SSE(Weibull.com)

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Southern MA Temp (F) | 75 | 80 | 81 | 82 | 80 | 76 | 74 | 71 | 75 |
| Eastern MA Temp (F) | 76 | 77 | 74 | 72 | 70 | 75 | 70 | 76 | 80 |

 The simple linear regression representation of this graph is set to y = -0.2697x + 95.24, where 95.24 is the predicted y-intercept and the slope of this linear regression is negative 0.2697. The slope of this line is also called the regression coefficient. The R2 value of this graph represents the coefficient of determination. The coefficient of determination is a measurement to show how well this linear regression model fit into the data set. It is also possible to calculate the coefficient of determination with the following equation; $R^{2}= \frac{SS\_{R}}{SS\_{T}}$ , the coefficient of determination is equal to the residuals sum of squares divided by the total sum of squares (Pertrucelli, Nandram, and Chen, 387).
 A coefficient of 0.0931 from our example is very low. It shows that the linear regression model is not a good fit for this data set. A coefficient of 1.0 marks a perfect fit to a data set, as described earlier. With just using simple linear regression, the R2 value was too low and thus is not strong enough to allow one to reject the null hypothesis.
 Using these statistical ideas, one could actually carry out a hypothesis testing on the coefficients of a simple linear regression model. There are many different ways to test this idea, such as the analysis of variance, or ANOVA, t tests, z tests, and the F test based upon the F distribution. However, I will mainly focus on the t tests and the analysis of variance. Generally, t tests are applied when the number of observations is less than thirty. T tests are done based upon the t-distribution rather than the normal distribution. Using the t test, one could perform a hypothesis testing on the coefficients of the simple linear regression model.
 The formal definition of a t distribution is that it is used to make an inference on the mean of a population when the variance is unknown. Hypothesis tests are then conducted on the mean value with the variance unknown. Usually, the null hypothesis would state that the regression coefficient, β1, is equal to some constant, β1,0. The alternative hypothesis would state that the regression coefficient is not equal to that constant. The standard equation to conduct t tests is that $T\_{0}=\frac{\hat{β}\_{1}-β\_{1,0}}{se(\hat{β}\_{1})}$ where $\hat{β}\_{1}$is the predicted regression coefficient, $β\_{1,0}$ is the constant in the null hypothesis, $se(\hat{β}\_{1})$ is the standard error of the predicted regression coefficient, and $T\_{0}$ is the test statistic.
 Referring to the terms described earlier, $\hat{β}\_{1}$ is the least square estimate and the test statistic has a number of degrees of freedoms based on the number of n observations subtracted by two. The standard error estimates the standard deviation of the sample mean based on the population mean. To calculate the standard error in a t test, the following equation is applied, $se\left(\hat{β}\_{1}\right)= \sqrt{\frac{\frac{\sum\_{i=1}^{n}e\_{i}^{2}}{n-2}}{\sum\_{i=1}^{n}(Y\_{i}-\overbar{Y)}^{2}}}$ . The standard error is made up of the the sum of squared residuals, degrees of freedom, and the total variance or the total sum of squares.
 In a t test, the null hypothesis is only rejected if the test statistic is not within the range of
± t1-α/2,n-2 , which is based upon the areas under the t distribution. These areas are put into a chart usually called a T Table or a T Test Table; see Appendix for the entire chart. The null hypothesis could also be rejected by using another method involving p-values.
 “P-value quantifies how consistent the observed value of the test statistic is with this distribution model, and hence with H0” (Pertruccelli, Nandram, and Chen, 291). The p-value provides evidence against the null hypothesis “and in favor of Ha as does the observed value of the test statistic” (291).
 The t test could also test the intercept, β0, of the simple linear regression model as well. However, there are different steps to accomplish this task. To test the intercept of the linear regression model using a t test, the test statistic equation will not be using the regression coefficient, but replacing it with the intercept, β0. The test statistic in this case is, $T\_{0}=\frac{\hat{β}\_{0}-β\_{0,0}}{se(\hat{β}\_{0})}$ (Weibull.com). The standard error, however, is calculated slightly different with the equation equal to.
 In a t test, if the constant that the regression coefficient is being compared to is equal to zero, then the hypothesis tests for significance of regression. If H0: β1=0 is not rejected, then there are no linear relationship that exists between the variables X and Y. For example, within this chart are four graphs that show how models fit with specific data sets (Weibull.com).


 On page ten is an example of a data set that shows the values sold in ice cream based upon a given day’s temperatures. I have taken this example off of a video and had used the t test method to apply hypothesis testing to the regression coefficients of the data set. The summary output table was generated using Microsoft Excel’s data analysis tool to solve the calculations that involves the equations mentioned before. Within the summary output, just about everything needed to proceed with hypothesis testing is printed, including the coefficient of determination, standard errors, number of observations, degrees of freedom, sum of squares, mean squares, F statistic for F test, T statistic for t test, and the p-value.
 The null hypothesis for this example was that there was no relationship between temperature and sales, which means that β1 = 0 and the alternative hypothesis is set so that β1≠0. From here, I calculated the test statistic using the equation $T\_{0}=\frac{\hat{β}\_{1}-β\_{1,0}}{se(\hat{β}\_{1})}$ resulting in a test statistic of 7.183. To determine the rejection of the null hypothesis, I used both the critical value method and the p value method.
 In order to proceed with the critical value method, a value must be found for the following, ±$t\_{1- \frac{α}{2}, n-2}$ . In this particular example, we must find the value of $t\_{0.975,18}$ using the t table chart within the Appendix. A t value of 2.101 is then compared to the test statistic. If the test statistic is not within the range of negative 2.101 and positive 2.101, then the null hypothesis is rejected.
 Using the p value method does not require ±$t\_{1- \frac{α}{2}, n-2}$ . However, we must use the equation where the p-value = 2P( t - |computed t|) = 2P( t > 7.183). The p-value could be found using the summary output generated by Microsoft Excel, under “P-value”. In conclusion to the example, if the null hypothesis was true, then there is a one in a million chance that one would end up with a regression coefficient of at least 0.433, which is not very likely at all. Therefore, this is classified as strong evidence against the null hypothesis using the t test method for a hypothesis testing on regression coefficients in a simple linear regression model.

Figure (a) and (b) shows that no linear models will fit the observed data. Figure (b) is a case where the relationship between the variables X and Y is not linear.

Figure (c) and (d) shows the rejection of the null hypothesis where β1 = 0, depicting a linear relationship between the variables X and Y. However, the linear model is more sufficient in the case of figure (c).

Example taken from a video (ProfTDub). Work Cited

1. Petruccelli, Joseph D., Balgobin Nandram, and Minghui Chen. *Applied Statistics for Engineers and Scientists*. Upper Saddle River, NJ: Prentice Hall, 1999. Print.
2. Draper, Norman Richard., and Harry Smith. *Applied Regression Analysis*. Second ed. New York: Wiley, 1981. Print.
3. Yan, Xin, and Xiaogang Su. *Linear Regression Analysis: Theory and Computing*. Singapore ; Hackensack, NJ: World Scientific, 2009. Print.
4. Headquarters, Worldwide. "Hypothesis Tests in Simple Linear Regression." *Hypothesis Tests in Simple Linear Regression*. Relia Soft, 2008. Web. 01 May 2013. <http://www.weibull.com/DOEWeb/hypothesis\_tests\_in\_simple\_linear\_regression.htm>.
5. ZedStatistics. "Regression 1 : What Is Regression? | SSE, SSR, SST | R-squared | Errors (Îµ vs. E)." *YouTube*. YouTube, 21 Nov. 2011. Web. 29 Apr. 2013. <http://www.youtube.com/watch?v=aq8VU5KLmkY>.
6. ProfTDub. "Doing a T-Test on the Slope Coefficient in Simple Regression." *YouTube*. YouTube, 19 Oct. 2010. Web. 29 Apr. 2013. <http://www.youtube.com/watch?v=7phcXl2uiYc>.

Appendix
T Table (Draper and Smith, 532):
