**STAT452: Lab 5**

**Simple Linear Regression**

Terence Henriod

STAT 452

4/16/2013

**Introduction**

There are many situations where it is advantageous to predict the outcome of some process, given values of some related variable. This is because it can be difficult or impossible to measure the outcome variable, but easier to measure the related variable. There are various techniques for predicting outcomes using previously collected data. Among these, and the focus of the following report, is the technique of linear regression.

Linear Regression predicts outcomes through use of a linear prediction line, with simple intercept and slope variables. When one predictor variable is utilized, the technique is known as *Simple Linear Regression*. A simple linear regression takes the following form:

[1]

where the set {Yi , Xi} are observations from previously collected data, β0 and β1 are regression coefficients that define the regression model, and εi are the error values for the “true” regression model against the data.

In terms of an estimated linear regression model, we use predicted values for the outcome variable, and estimated regression coefficients, and we use a residual term rather than an error one:

[2]

It is important to distinguish residuals *e* and errors ε. Errors are found from the difference between the true regression and the data; the residuals are the difference from the estimated model and the data, in fact the residuals are found by:

[3]

while error calculations would use the values of the “true” coefficients. Care should be taken to avoid confusing the two.

Linear Regression is such a powerful tool because (1) it results in an intuitive model (linear models are easy to understand), and (2) because linear regression is robust due to the fact that its coefficients are estimated so that the squared sum of residuals is minimized:

[4]

These errors (residuals) don’t necessarily need to be independent or distributed in any particular manner in order to perform a regression, but if we wish to perform any inference on the results of the regression (i.e. finding confidence/prediction intervals, performing hypothesis testing on coefficients, etc.) the assumption that the errors ε are all independent, identically distributed random variables belonging to a distribution N(0, σ2) should hold (although they do not necessarily have to be independent). When the errors of a model share the same variance, then the model is known as *homoskedastic*; otherwise the model is called *heteroskedastic*.

In order to assess the goodness of fit for the model we rely on values such as R2 and the three sum of squares measures, where:

[5]

[6]

[7]

[8]

Where the ESS is the error between the estimated model and the data, the TSS is the error between the line and the data, and the RSS is the amount of error reduced by using the estimated model as opposed to the average outcome value for predictions. R2 ends up being a measure similar to the RSS, but as a proportion, so it is a more universal measure.

The data set that will be used is a Minitab sample data set regarding Bears and their measurements. The reason this data was collected was for the creation of tables that would aid in the determination of various bear measurements from measurements that are easier to obtain while in the field. This is the sort of exercise that will be performed using linear regression. Primarily of interest for this lab assignment will be bear weights and lengths, as we will be attempting to predict the weight of a bear given its length. As alluded to previously, it would be easier to measure a bear’s length in the field than its weight, hence the utility of a tool like linear regression. The length is measured in inches, and the weight will be measured in pounds.

Also to be considered are the bears’ head, neck, and chest measurements, all also measured in inches, which will be superficially investigated to determine if these factors would be viable candidates for alternate predictors, or possibly additional factors for a multiple regression.

**Assignments**

1. Use scatterplots to analyze the association between the bears’ lengths and weights. Choose a data transformation that will make the association approximately linear and homoskedastic.



Figure 1: From an initial scatterplot of bear Length and Weight, we can see that the association between the two is not readily linear. In order to apply linear regression in a meaningful way, we need to transform the data so as to achieve homoskedasticity.



Figure 2: A probability plot comparing the bear lengths to a Normal Distribution. There appears to be a sufficient fit.



Figure 3: A probability plot comparing bear weights to a Normal Distribution. This is not a satisfactory fit.

Based on the probability plots, it appears that the bear weights are out of order, and need to be transformed to induce homoskedasticity of the data.

One of the most common data transformations used is to take the Log10 of the data. I will start by applying this transformation to the bear Weights, and evaluate the results.



Figure 4: When plotting the Log10 of the Weight of the bears against the Length of the bears, the trend appears to be much more linear, and the vertical spread of the points appears to be greatly reduced, suggesting that homoskedasticity has been achieved.

It appears that the data transformation that should be used when estimating the regression model is to apply Log10(Weight), while leaving Length as it is.

1. Perform linear regression analysis with the bear’s weight as response (dependent variable) and the bear’s length as predictor (independent variable) using the data transformation found in (1):
   1. Find the fitted regression line

The regression line, as evaluated by Minitab, is:

[9]

It should be noted that the predicted values of bear weight will be in terms of Log10 Weight, and to be useful need to be converted. So, in more intuitive terms, the model is:

[10]



Figure 5: A graph representing the Linear Regression Fit of the data and appropriate transformations, along with a 95% Confidence Interval (red dashed lines) and a 95% Prediction Interval (green dashed lines).



Figure 6: The results of the linear regression plotted against a logarithmic scale for bear weight. Again, the confidence interval is outlined by red dashed lines, and the green dashed lines indicate the prediction interval. While the scale may not be as intuitive, the linear relation is easily seen.

* 1. Check the residuals’ mean, homoskedasticity, and normality



Figure 6: A collection of plots summarizing the residual of the linear regression reported above. These measures all indicate favorable results.

The 4-in-1 figure gives encouraging results: the probability plot shows that the residuals fit a Normal distribution quite well with a P-Value of 0.454 (not shown), the histogram displays a trend that is similar to a normal curve, and the remaining two plots display relative symmetry about their central measures (vertically), displaying a vertical spread indicative of homoskedasticity. Additionally, the mean of the residuals is 1.7915 E -15, which is very close to zero, indicating that the residuals likely belong to a N(0, σ2) distribution.

* 1. Find the coefficient of determination, regression sum of squares, error sum of squares, total sum of squares, and error sample standard deviation; discuss and interpret the values

|  |  |
| --- | --- |
| Important Regression Summary Statistics | |
| Coefficient of Determination (R2) | 0.8790 |
| Regression Sum of Squares (RSS) | 8.6110 |
| Error Sum of Squares (ESS) | 1.1806 |
| Total Sum of Squares (TSS) | 9.7916 |
| Error Sample Standard Deviation | 0.0915 |

Table 1: Values of interest for the regression.

The first thing to notice from the data in Table 1 is that R2 is a large, healthy value, indicating that there is a strong correlation between the outcome values and the values of the predictor variable (see Figure 4 for a visual sense of this). Alternately, this is also the proportion of Y-variance explained by the regression. This indicates that a good fit has been achieved by the model.

Next we see that the Error Sum of Squares is relatively small compared to the Total Sum of Squares, indicating a relatively tight fit of the model to the data compared to simply trying to predict the outcome (weight) by simply looking at the mean weight. The Regression Sum of Squares indicates how much better the model fairs in predicting bear weight from length with the regression line, as opposed to using a line with no slope at the level of the average bear weight. Going back to R2, we can say that R2 represents the reduction in error of the model as a proportion of the reduction in error divided by the error if only the mean was used for prediction.

Finally, the Error Sample Standard Deviation (which is really more of a measure of the residual standard deviation), is relatively small, further supporting the assertion that the model makes good predictions without large and variable error distributions. We this also completes our definition of the distribution of our errors (residuals): N(1.7915 E -15, 0.09152).

* 1. Find the standard deviations and P-values for the fitted regression coefficients, decide whether the coefficients are significant.

As a reminder, when performing hypothesis testing on regression coefficients, the hypotheses used when testing β1:

[11]

so a rejection of H0 means that we doubt that the value given by Minitab is not statistically significant, and we choose to accept the value given for β1 as valid; for β0, we test its value against 0, although we could test it against other values as well:

[12]

where a rejection of H0 means that we will proceed believing that β0 is statistically significantly different from zero.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Coefficient | Value | Standard  Deviation | P-Value | Conclusion |
| Β0 | 0.59699 | 0.05090 | 0.000 | Reject H0 |
| Β1 | 0.02633 | 0.00082 | 0.000 | Reject H0 |

Table 2: Statistics for the regression coefficients.

As can be seen in the table, both coefficients had very small P-Values, so it is highly unlikely that their values are not valid or significant.

* 1. Discuss unusual observations

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Observation | Length | Log10(Weight) | Fit | SE Fit | Residual | Std.  Residual | Code |
| 3 | 57 | 1.86923 | 2.09782 | 0.00842 | -0.22859 | -2.51 | R |
| 26 | 37 | 1.53148 | 1.57122 | 0.02136 | -0.03974 | -0.45 | X |
| 33 | 36 | 1.41497 | 1.54489 | 0.02212 | -0.12991 | -1.46 | X |
| 45 | 40 | 1.60206 | 1.65021 | 0.01908 | -0.04815 | -0.54 | X |
| 57 | 40 | 1.81291 | 1.65021 | 0.01908 | 0.1627 | 1.82 | X |
| 58 | 64 | 2.55145 | 2.28214 | 0.00797 | 0.26931 | 2.95 | R |
| 59 | 65 | 2.49969 | 2.30847 | 0.00824 | 0.19122 | 2.10 | R |
| 72 | 50 | 2.17026 | 1.91351 | 0.01202 | 0.25675 | 2.83 | R |
| 114 | 66.5 | 2.16435 | 2.34796 | 0.00877 | -0.18361 | -2.02 | R |
| 116 | 83 | 2.5977 | 2.78241 | 0.0194 | -0.18472 | -2.07 | X, R |
| 122 | 43.5 | 1.4624 | 1.74236 | 0.01648 | -0.27997 | -3.11 | R |

Table 3: A display of unusual observations, as identified by Minitab. The code R indicates that the observation has a large standard residual, and X means that the observation has a value that gives it large leverage.

In general, many of these “unusual observations” are picked out by Minitab because Minitab acts conservatively (carefully) when searching for such values.

Many of the values that had “large leverage” seem to be ones with small Length values, perhaps these present an issue because they belong to shorter, likely younger, bears, which may have a different relation between weight and length than fully mature bears, but these observations are few and not extreme, and therefore not likely to significantly detract from the goodness of the model fit. Also these large leverage observations come from the short bears, whom there were relatively few of, which is likely the reason for their large leverage. This could be remedied by adding more data for short bears.

As for the large residual observations, these were likely just simple, rare outliers. If we look at the regression plots, these are likely the points outside the confidence interval. Given that the confidence is a 95% one, it is reasonable to say that these points are just the 5% that will fall outside the confidence interval, and these points fall just outside, so they are not likely a reason for concern; Minitab was simply being concerned.

1. Find the CI and PI for the bear’s weight at the mean bear’s length

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Interval Estimation for Mean of Bear Length = 61.2825 inches | | | | |
|  | In terms of Log10(Weight) | | In terms of Weight  (lb) | |
|  | Lower Limit | Upper Limit | Lower Limit | Upper Limit |
| 95% Confidence Interval | 2.19546 | 2.22571 | 156.84114 | 168.15508 |
| 95% Prediction Interval | 2.02905 | 2.39212 | 106.91780 | 246.67208 |

Table 4: 95% Confidence and Prediction Intervals for the weight of bears given the mean length of the bears.

1. Perform correlation analysis of the regression residuals with the bears’ head length, head width, neck girth, and chest girth measurements; discuss and interpret the results



Figure 7: Scatterplots of all the secondary variables of interest as predictors against the residuals of the estimated model.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Head Length | Head Width | Neck Girth | Chest Girth |
| Residuals | 0.136 | 0.210 | 0.315 | 0.352 |

Table 5: The Pearson correlation coefficient estimates between the residuals of the estimated model and the secondary interest variables.

This brief correlation analysis indicates that since none of these secondary predictor candidates appears to have a significant correlation with the residuals of the estimated model, addition of these variables to a multiple regression model that already included length would not result in any substantial improvement of predictions.

**Conclusion**

Simple Linear Regression was applied to the Minitab sample data set BEARS.MTW. Using this data set, a regression line was found in order to use the length of a bear as a predictor for the weight of a bear. The results were sound: a model with statistically significant (non-zero) coefficients was found, that displayed homoskedasticity, and fared significantly better than the mean bear weight for predicting outcomes.

The data did need to be transformed: a Log10 transformation did need to be applied to the bears’ weight to produce a sound, linear model.

The model produced was as follows:

[9]

or in more intuitive terms, the model is:

[10]

The residuals did exhibit a nearly Normal(0, σ2) distribution, thus maintaining the assumptions of a linear model and allowing us to perform inference on the results of the regression.

Unusual observations were minimal and likely the result of Minitab’s conservatism when identifying such values, and did not detract significantly from the model.

Assigned Confidence and Prediction Intervals were estimated, given the mean bear’s length, and these results can be viewed in Table 4.

When checking for other variables that might make good additional predictors by comparing other variables with the residuals, it was seen that all of the secondary candidates (head length, head width, neck girth, and chest girth) fared poorly, and would not likely benefit a model that already included bear length as a predictor.