**Lab 6:**

**Multiple Linear Regression**

Terence Henriod

STAT 452

4/30/2013

**Introduction**

Previously, we discussed Linear Regression in general, and Simple Regression. In this lab, we will be discussing and practicing *Multiple Linear Regression*. Multiple Regression predicts outcomes through use of a linear prediction line, with an intercept and multiple slope variables. A multiple linear regression takes the following form, with predictors X1 through Xp:

[1]

where the set {Yi , X1, …, Xp} are observations from previously collected data, β0 and βj are regression coefficients that define the regression model, and εi are the error values for the “true” regression model against the data.

It should be noted that a multiple regression introduces the possibility for both a predictor and non-linear transformations of that predictor, as well as interaction effects between predictors to enter the model, for example:

[2]

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:

[3]

Again, as with simple regression, the coefficients of multiple regression are found through the application of mean-square error minimization:

[4]

and again, as with simple regression, 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 a multiple regression 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. Multiple regression is different from simple regression, however, in that an adjusted R2 measure is used, where the calculation for R2 (adj.) uses a different number of degrees of freedom, giving us a more sound assessment of the goodness of fit of the model:

[9]

where p is the number of parameters used in the model, n is the number of observations, and R2 is calculated as in [5]. The adjusted R2 prevents us from thinking a model performs better than it does simply because it has more parameters than a comparable model.

F-testing is sometimes used to help identify the best statistical model. For our purposes, we will simply be using an F-test (performed by Minitab) to evaluate the significance of our model parameters. The hypothesis tested is stated as:

[10]

.

There are issues that apply to multiple regression that do not arise in simple regression: *confounding*, *collinearity*, and *instrument variables*.

*Confounding* is when a predictor is correlated with a response variable, so the two appear to be related, but in reality they are not related, when, in fact, the relationship is explained by a third variable. This can be problematic (1) because we attribute the effects of change in a response variable to an unrelated [false] predictor, and (2) because the inclusion of a confound in a model can add bias, reducing the utility of the model compared to if the true predictor were used.

*Collinearity* (sometimes referred to as *multi-collinearity*) occurs when two (or more) predictors are highly linearly correlated, so when used in a regression model they both appear to be insignificant, because it is difficult to choose which one to use in terms of explaining the response variable. This occurs because if we use one of the collinear predictors to predict the response, the other is unnecessary for the model, and vice versa. Collinear variables can be identified by (1) their high correlation to one another, (2) having a high Variance Inflation Factor in a model, and (3) when collinear predictors that were significant to regression individually suddenly become insignificant when used together. In the case of collinearity, only the predictor that improves the model most individually should be used.

*Instrument Variables* are variables that are not correlated to the response variable, but are related to another predictor. Use of an instrument variable enhances a model because it can help to explain variability of the predictor it is related to, that is, the instrument helps to accommodate for the fact that the predictor it is related to is not equally effective across all observations.

Methods for finding the best statistical model given a certain data set include *Stepwise Regression* and *Best Subset Regression*.

*Stepwise Regression* is performed by either *Forward Selection*, where the model starts with no variables and variables are added one at a time to test their enhancement of the model until the model cannot be improved, or by *Backward Elimination*, where one starts with the fully saturated model, and eliminates variables one at a time until the model cannot be improved. In either method, variables are not immediately included/eliminated; rather at each step (number of variables included) each variable takes a turn being included/eliminated, and the variable with the most dramatic effect is selected/eliminated before moving on to the next step. It is common that a certain level of effect (selection criterion) is determined for a variable to be added/deleted, so as not to produce an overly complex model that has little added benefit over a simpler model.

*Best Subset Regression* is a Minitab command that when used causes Minitab to estimate models for n predictors, starting at n = 1, and stopping at n = all predictors used. Minitab will display the best and second best predictor set for every level of predictors n. This allows one to have a little more insight and freedom with the model selection process. Also presented is the Mallow’s Cp score. This score allows for the evaluation of the fit of a model constructed using ordinary least squares (as used in the types of regressions we have performed in this class). The Cp value is used to ideally prevent overfitting, or using too many predictors in a model to reduce the residual sum of squares, which often provides little benefit. Mallow’s Cp will not be discussed here, but in short, Mallow’s Cp is used to ideally find a model with the best combination of a low Cp and a low number of included predictors, and this should guide our selection of a model when using stepwise or subset approaches.

The data set that will be used was provided by the instructor, and is simply patterned data randomly generated such that collinearity, confounding, and an instrument variable are all present in the set, but a useful regression may still be produced. The number of observations in the data set is n = 500.

**Assignments**

1. **Perform simple linear regression of *Y* on each of *Xi*, discuss the results.**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Simple Linear Regressions of Y on Various Predictors Xi | | | | |
|  | X1 | X2 | X3 | X4 |
| Constant β0  (P-Value) | 0.05204  (0.333) | 0.05052  (0.347) | 0.02455  (0.691) | -0.01603  (0.715) |
| Slope β1  (P-Value) | 0.47844  (0.000) | 0.46917  (0.000) | 0.02812  (0.513) | 0.98784  (0.000) |
| s of Residuals | 1.19894 | 1.19976 | 1.38116 | 0.978864 |
| R2 | 0.247 | 0.246 | 0.001 | 0.498 |
| RSS | 234.96 | 233.97 | 0.818 | 473.63 |
| ESS | 715.85 | 716.83 | 949.988 | 477.17 |
| TSS | 950.81 | 950.81 | 950.806 | 950.81 |

Table 1: Simple regressions of Y on all Xi. These results will help us in determining the best model given this data.



Figure 1: The residual plot for the regression of Y on X1. It shows that the residuals are likely Normal and homoscedastic given the symmetry seen relative to the vertical axis.



Figure 2: The residual plot for the regression of Y on X2. It shows that the residuals are likely Normal and homoscedastic given the symmetry seen relative to the vertical axis.



Figure 3: The residual plot for the regression of Y on X3. It shows that the residuals are likely Normal and homoscedastic given the symmetry seen relative to the vertical axis.



Figure 4: The residual plot for the regression of Y on X4. It shows that the residuals are likely Normal and homoscedastic given the symmetry seen relative to the vertical axis.

Inference would be valid for all of these simple regressions, except possibly the regression of X3, which was not significant. Note that although none of the intercept parameters were significant, this is of little issue because the hypothesis that the P-Value refers to is that the coefficient is significantly different than zero. The values of these are so close to zero already that it wouldn’t matter much if they were actually zero, as well as the fact that we have no physical laws/evidence specifying that these constants should be different than zero. This same assertion can be assumed for all future models presented in this report.

1. **Perform correlation analysis among the predictors, discuss the results.**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Correlation Analysis Between Response Y and Predictors Xi | | | | |
|  | Y | X1 | X2 | X3 | X4 |
| Y |  | 0.497 | 0.496 | 0.029 | 0.706 |
| X1 | 0.000 |  | 0.991 | 0.531 | -0.024 |
| X2 | 0.000 | 0.000 |  | 0.521 | -0.017 |
| X3 | 0.000 | 0.000 | 0.000 |  | -0.019 |
| X4 | 0.000 | 0.599 | 0.700 | 0.665 |  |

Table 2: A correlation analysis of the predictors and response variable. Correlation values (r) are indicated above the diagonal are indicated above the diagonal, and the P-Values indicating the significance of the r value are below the diagonal. P-Values have been darkened because they are invalid and may not be used because they violate the assumption that the P-Value applies to a bivariate Normal sample (X4 is not Normal), and therefore are invalid for hypothesis testing. Evidence of Normality for the other samples is indicated by the following probability plots.



Figure 5: A probability plot displaying that Y is likely Normal, which is to be expected because we will see that the predictors used either are Normal, or nearly Normal.



Figure 6: Probability plots verifying the Normality of X1 and X2.



Figure 7: Probability plots that verify the Normality of X3, but show that X4 is not quite Normal.

Y was also used in the correlation analysis to aid in the identification of variables that might cause problems such as confounding or collinearity, or for possible instrument variables..

Because of the Normality of most of these variables, we can accept the P-Values presented by Minitab as valid, and all of the legitimate P-Values do confirm that the values of the correlation coefficients presented are significantly different from zero.

Based on the high correlation between X1 and X2, we should consider these variables as candidates for collinearity. X3 does not correlate well with Y, but does correlate somewhat with X1 and X2, so it should be put on a watch list as a possible instrument variable.

1. **Find the best multiple linear regression model for predicting Y using *Xi.***

Using (Minitab’s) Stepwise Regression:

[12]

|  |  |  |  |
| --- | --- | --- | --- |
| Predictor | Value of Coefficient | P-Value | VIF |
| Constant (β0) | 0.01998 | 0.427 | - |
| X1 | 0.65797 | 0.000 | 1.393 |
| X3 | -0.30649 | 0.000 | 1.392 |
| X4 | 1.00172 | 0.000 | 1.001 |

Table 3: The values of the coefficients estimated for the best model fit by Minitab using stepwise regression (and best subset regression).

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| s | Adjusted R2 | F Score  (P-Value) | RSS | ESS | TSS |
| 0.561362 | 0.835 | 840.40 (0.000) | 794.50 | 156.30 | 950.81 |

Table 4: More values pertaining to the goodness of fit for the best estimated model.



Figure 8: The residual plot for the regression of Y on X1, X3, and X4. It shows that the residuals are likely Normal and homoscedastic given the symmetry seen relative to the vertical axis, verifying that the assumptions necessary for performing inference can be made.

Using (Minitab’s) Best Subset Regression:

As it turns out, this method provides us with the same model as Stepwise Regression did, so the information about the regression will not be repeated here.

Of interest here is the Mallow’s Cp score of 3.9. This suggests that approximately 4 regressors should be used, which was near the number of parameters actually used. This was the lowest Cp score observed, and its value was the closest to the number of parameters used in a model to produce such a score.

1. **Identify the variables with (i) collinearity, (ii) confounding, and (iii) predictors uncorrelated with response but useful for regression (instruments). We do have each case in this data set.**
   1. Collinearity

The case of collinearity is found immediately when building a multiple regression model. Simple regressions indicate that X1 and X2 are similarly strong predictors, and when taking the first step in stepwise regression, suddenly both predictors become insignificant. Also, note the large Variance Inflation Factor values, which when larger than 5 give reason to believe collinearity exists. This is not surprising, as this pair was already under suspicion for collinearity after performing the correlation analysis.

|  |  |  |  |
| --- | --- | --- | --- |
| Predictor | Value of Coefficient | P-Value | VIF |
| Constant (β0) | 0.05157 | 0.337 | - |
| X1 | 0.2949 | 0.289 | 55.053 |
| X2 | 0.1820 | 0.505 | 55.053 |

Table 5: The Minitab output of a regression of Y on X1 and X2, demonstrating evidence of the collinearity of X1 and X2.

Because X1 has a lower corresponding P-Value, it is kept in future regression steps, but it was also observed that models containing X1 fared better than those containing X2 in terms of R2 values. Further, the Minitab Regression procedures selected X1.

* 1. Confounding

The confounding variable was the most difficult to identify (as it should be). X2 was identified as the confounding variable based on the criterion that a predictor that is significant in simple regression but is not significant in multiple regression is likely a confound, as seen in Table 6. Admittedly, this was particularly difficult to find, given that the ideal model was found using forward inclusion, so the saturated model was not considered formally, only in trying to detect the confound. Further, X2 had already been tagged as collinear based on correlation, VIF, and multiple regression criteria previously considered above.

|  |  |  |  |
| --- | --- | --- | --- |
| Predictor | Value of Coefficient | P-Value | VIF |
| Constant (β0) | 0.02026 | 0.421 | - |
| X1 | 0.7799 | 0.000 | 56.012 |
| X2 | -0.1205 | 0.347 | 55.249 |
| X3 | -0.30724 | 0.000 | 1.394 |
| X4 | 1.00280 | 0.000 | 1.003 |

Table 6: The Minitab output of the fully saturated model, which supports the claim that X2 may be a confounding variable.

* 1. Instrument(s)

It was observed that X3 acts as an instrument for X1. X3 does not appear to be correlated with our response Y, but it is correlated with X1. When X3 is used as a predictor, in models where X1 is not present, it is not significant (i.e. in a simple regression or in a model with only X3 and X4), but in models where X1 is present, X3 becomes a significant parameter, and the model’s quality improves significantly in terms of R2.

**Conclusion**

Using the patterned (but still random) data provided by the instructor, *Multiple Linear Regression* analysis was performed. It was determined that X1 and X2 were collinear, X2 was also a confounding variable, and X3 turned out to be an instrument variable for X1 by helping to better explain the variability of X1. The best model for predicting Y that was produced was produced by both stepwise and best subset methods, and is as follows:

[13]

This model boasts highly significant coefficients (P-Value = 0.000), except for the intercept, which is not significantly different from zero. The model is a good fit, with an adjusted R2 = 0.835, and does not contain an overwhelming number of parameters.