Assignment #4: Problem Set for Ordinary Least Squares Regression (50 points)

This assignment will be made available in both pdf and Microsoft docx format. Answers should be typed into the docx file, saved, and converted into pdf format for submission into Blackboard. **Color your answers in green so that they can be easily distinguished from the questions themselves.**

Throughout this assignment keep all decimals to four places, i.e. X.xxxx.

Any computations that involve "the log function", denoted by log(x), are always meant to mean the natural log function (which will show as ln() on a calculator). The only time that you should ever use a log function other than the natural logarithm is if you are given a specific base.

When stating the null and alternate hypotheses in any statistical test in PREDICT 410, we should always state these hypotheses in terms of the model parameters, i.e. the model coefficients denoted by the betas.

Model 1: Let's consider the regression model, which we will refer to as Model 1, given by

Y = 10,000 + 150*X1 + 25*X1^2 + 60*X2 (M1).

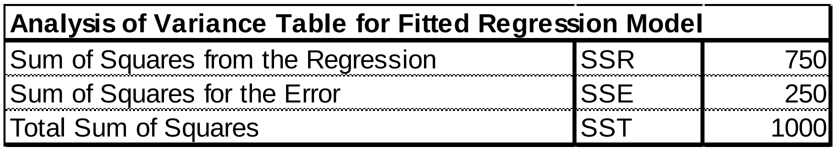
(1) (2 points) Is this a "linear" regression model, why or why not?

Yes, M1 is a linear regression model since the regression parameters enter the equation linearly.

(2) (4 points) How do we interpret this model? Hint: how does a one unit change in X1 or X2 affect the estimated value for Y? State the interpretation for both X1 and X2.

The model can be interpreted that the change in Y corresponds to a one unit change in Xj (where j=0,1) when all other predictors are held constant. For every one unit change in X1 this will affect Y by (150 + 25*X1) when holding X2 fixed. For every one unit change in X2 this will affect Y by 60 when holding X1 fixed.

(3) Consider the Analysis of Variance (ANOVA) table from fitting this model to a sample of 50 observations.



Analysis of Variance Table for Fitted Regression Model		
Sum of Squares from the Regression	SSR	750
Sum of Squares for the Error	SSE	250
Total Sum of Squares	SST	1000

a. (4 points) Compute the R-squared and adjusted R-squared values for this regression model.

$R^2 = \frac{SSR}{SST} = \frac{750}{1000} = 0.75$

R_{adj}^2 = 1 – $(\frac{n-1}{n-k-1})$ $(1 - \ R^2)$

$R_{adj}^2 = 1 - (\frac{50 - 1)}{50 - 2 - 1)}\ (1 - 0.75)$

$R_{adj}^2 =$ = 0.7394

b. (2 points) Compute the estimate of the Mean Square Error (MSE).

MSE = $\frac{SSE}{n-k-1}$ (where n = # of observations, k = degrees of freedom)

MSE = $\frac{250}{(50-2-1)} = \frac{250}{47} = 5.3192$

c. (4 points) Perform the overall F-test for this model, i.e. state the null and alternate hypothesis, compute the test statistic for the overall F-test, and make a decision to "reject " or "fail to reject" the null hypothesis. Test the statistical significance of the overall F-test using a critical value for alpha=0.05 from Table A.4 on page 376 in *Regression Analysis By Example*.

$H_0 : B_1 = \ B_2 = 0$
H_a: At least one of the coefficients is not equal to 0

$F \ = \ \frac{MSR}{MSE}$ = $\frac{\frac{SSR}{k}}{5.3192}$ = $\frac{\frac{750}{2}}{5.3192}$ = 70.4993

$F_{0.05,2,48} = 3.23$

$F > F_{0.05,2,48}$ Therefore the null hypothesis is rejected.

<u>Model 2:</u> Now let's consider an alternate regression model, which we will refer to as Model 2, given by

Y = 9,750 + 145*X1 + 75*X2 (M2).

(4) Consider the ANOVA table from fitting this model to the same sample of 50 observations that we used to fit M1.

Analysis of Variance Table for Fitted Regression Model		
Sum of Squares from the Regression	SSR	725
Sum of Squares for the Error	SSE	275
Total Sum of Squares	SST	1000

a. (4 points) Compute the R-squared and adjusted R-squared values for this regression model.

$$R^2 = \frac{SSR}{SST} = \frac{725}{1000} = 0.725$$

$$R^2_{adj} = 1 - \left(\frac{n-1}{n-k-1}\right)(1 - R^2)$$

$$R^2_{adj} = 1 - \left(\frac{50-1}{50-2-1}\right)(1 - 0.725) = 0.7133$$

b. (4 points) State the hypothesis and compute the test statistic for the overall F-test.

$H_0: B_1 = B_2 = 0$

H_a: At least one of the coefficients is not equal to 0

$$F = \frac{MSR}{MSE} = \frac{\frac{SSR}{k}}{\frac{SSE}{n-k-1}} = \frac{\frac{725}{2}}{\frac{275}{(50-2-1)}} = 61.95$$

$F_{0.05,2,48} = 3.23$

$F > F_{0.05,2,48}$ Therefore the null hypothesis is rejected.

(5) Now let's consider M1 and M2 as a pair of models. We want to decide which model we should use as our final model. Here are some concepts to help us make that decision.

a. (2 points) What is the definition of a nested model?

A nested model is a model that can be obtained from a larger model. This is what is also meant by a reduced model where the number of parameters is a subset of the full model.

b. (2 points) Does M1 nest M2 or does M2 nest M1?

To determine if M1 nests M2 or if M2 nests M1 we need to determine which model includes the full set of predictor variables. If M1 can be derived from M2 as a subset of the variables, then

M2 nests M1. If M2 has a subset of the variables that exist in M1, then M1 nests M2. One other method would be review the Sum of Squares Error (SSE). The full model, because of its additional predictors, cannot increase SSE. Given that M2 can be derived from M1 and the SSE for M1 is less than M2, M1 nests M2.

c. (2 points) Based on any of the metrics or statistics that you have computed in Questions #3 and #4, which model should we prefer (M1 or M2) and why?

Based on the F-statistic, both M1 and M2 are statistically significant. Normally the model with the lowest SSE resulting in a higher goodness of fit (R-square) would be preferred and on that basis alone M1 would be preferred.

d. (10 points) Perform a F-test for nested models and determine if we should choose M1 or M2. State the hypothesis that we will be testing, compute the test statistic, and test the statistical significance using a critical value for alpha=0.05 from Table A.4 on page 376 in *Regression Analysis By Example*.

$H_0: Reduced\ model\ is\ adequate$

$H_a: Full\ model\ is\ adquate$

$$F = \frac{\frac{[SSE(RM)-SSE(FM)]}{(p+1-k)}}{\frac{SSE(FM)}{(n-p-1)}} = \frac{\frac{[275-250]}{(3+1-3)}}{\frac{250}{(50-3-1)}} = \frac{25}{\frac{250}{(46)}} = 4.60$$

Where:

n = number of observations
k = parameters in reduced model
p = parameters in full model

The null hypothesis H_0 is rejected if:

$$F \geq F_{(p+1-k, n-p-1;\ \alpha)}$$

(numerator degrees of freedom = 1, denominator degrees of freedom = 46)

$F_{(1,46:\ p<0.05)} = 4.08$ (when denominator degrees of freedom = 40)

$F_{observed} > F_{critical}$ (4.60 > 4.08), therefore the null hypothesis is rejected. M1 (full model) is the chosen model.

(6) In Ordinary Least Squares (OLS) Regression we assume that the response variable is normally distributed with mean XB and variance sigma^2, i.e. Y ~ N(XB, sigma^2).

a. (2 points) How do we estimate sigma^2?

Sigma-squared can be estimated by: $\sigma^2\ hat = \frac{SSE}{n-p-1}$

SSE = sum of squared error, n = number of observations, p = number of predictors

b. (6 points) What are two diagnostic checks of model goodness-of-fit that we perform in order to assess this distributional assumption?

A t-Test and F-Test are two diagnostics for testing the OLS assumptions. The t-Test is used to assess the statistical significance of each regression model variable. The F-Test can be used to assess the statistical significance of the entire model.