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

$$Y = 10{,}000 + 150*X1 + 25*X1^2 + 60*X2 \qquad \text{(M1)}.$$

(1) (2 points) Is this a "linear" regression model, why or why not?
Yes this is a linear regression model. The term "linear" means linear in the parameters, not the predictor variables.

(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.
Let x1 denote a fixed value for X1. A 1 unit increase in X1 from x1 to (x1+1) will increase Y by 150 + 25*((x1+1)^2 – x1^2) = 150 + 25*(2*x1 + 1), while a 1 unit decrease in X1 from x1 to (x1-1) will decrease Y by 150 + 25*(x1^2 – (x1-1)^2) = 150 + 25*(2*x1 + 1). For X2 a 1 unit increase will increase Y by 60, and a 1 unit decrease will decrease Y by 60.

(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.
Note that this regression model has p=3 predictor variables plus an intercept so p+1 = 4.
The sample size is n=50 and hence (n-p-1) = 46.

R-squared = SSR / SST = 750/1000 = 0.7500
Adjusted R-squared = 1 – [SSE / (n-p-1)] / [SST / (n-1)]
= 1 – (250/46) / (1000/49) = 0.7337

b. (2 points) Compute the estimate of the Mean Square Error (MSE).
MSE = sigma^2 = SSE / (n-p-1) = 250/46 = 5.4348

c. (4 points) State the hypothesis and compute the test statistic for the overall F-test.
H0: b1 = b2 = b3 = 0
H1: bj != 0 for some j in {1,2,3}

F = [SSR / p] / [SSE / (n-p-1)] = [750/3] / [250/46] = 46

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

$$Y = 9{,}750 + 145*X1 + 75*X2 \qquad \text{(M2)}.$$

(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.
Note that this regression model has p=2 predictor variables plus an intercept so p+1 = 3. The sample size is n=50 and hence (n-p-1) = 47.

R-squared = SSR / SST = 725/1000 = 0.7250
Adjusted R-squared = 1 – [SSE / (n-p-1)] / [SST / (n-1)]
= 1 – (275/47) / (1000/49) = 0.7133

b. (2 points) Compute the estimate of the Mean Square Error (MSE).
MSE = sigma^2 = SSE / (n-p-1) = 275/47 = 5.8511

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

H0: b1 = b2 = 0
H1: bj != 0 for some j in {1,2}

F = [SSR / p] / [SSE / (n-p-1)] = [725/2] / [275/47] = 61.9546

(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?
The model M1 is said to *nest* model M0 if M1 contains all of the parameters (predictor variables) from M0. For example the model Y = b0 + b1*X1 + b2*X2 nests the model Y = b0 + b1*X1.

b. (2 points) Does M1 nest M2 or does M2 nest M1?
In our problem 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?
M1 has an adjusted R-squared value of 0.7337 while M2 has an adjusted R-squared value of 0.7133. Hence we prefer M1 to M2.

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*.
In this problem set M1 is the *full model (FM)* and M2 is the *reduced model (RM)*. For notational purposes we will say that the full model has *p* predictor variables and that the reduced model has *k* predictor variables. The full model always nests the reduced model, and the full model always fits better than the reduced model in an absolute measure such as the R-squared or the likelihood value. Keep in mind that both the R-squared and the likelihood value are monotonically increasing in the number of model parameters.

Let's let b3 (=25) be the generic representation of the coefficient for the X1^2 term. The null hypothesis that we test is H0: b3 = 0.

The F-test for nested models is defined by the test statistic

$$F = [(SSE(RM) - SSE(FM)) / (p-k)] / [SSE(FM) / (n-p-1)]$$
$$= [(275-250) / 1] / [250 / 46] = 25 / [250 / 46] = 4.6000$$

From Table A.4 we have F(1,40,0.05)=4.08 and F(1,60,0.05)=4.00. Since 4.6 is greater than both of these critical values, we can safely reject H0 and conclude that b3 is not statistically equal to zero. Hence, we prefer M1 to M2. Notice that the F-test for nested models provided the same conclusion as the adjusted R-squared metric.

(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^s is simply the MSE, hence sigma^2 = SSE / (n-p-1).

b. (6 points) What are two diagnostic checks of model goodness-of-fit that we perform in order to assess this distributional assumption?
We asses this distributional assumption by: (1) plotting a histogram or a QQ-Plot to assess the normality of the residuals, and (2) plotting the residuals versus the predictor variables to assess the homoscedasticity of the residuals.