PSTAT 126 Quiz #2 Solutions

Question 1(Worth 10 points total)**:** Consider the mtcars dataset built into R. Use R to compute a Multiple Linear Regression model with disp as the response variable and mpg, wt, and hp as the predictor variables.

Is the overall regression considered significant (1 point)? How can you tell (2 points)? Explain what we mean in the context of this course (PSTAT 126) when we say that the overall regression model is or is not significant (2 points).

Are each of the variables mpg, wt, and hp considered individually, significant for the regression? Which ones (2 points)? How can you tell (1 point)? Explain what we mean in the context of this course when we say that an individual predictor variable is or is not significant for the regression (2 points).

(Please enter your entire answer directly into the online answer field provided with this quiz question below. There is no need for any complicated mathematical equations or formulas to be included in your answer.)

Answer: The corresponding R linear regression output summary report is the following:

```
model34=lm(disp ~ mpg + wt + hp, data = mtcars)

summary(model34)

Residuals:

    Min      1Q  Median      3Q     Max

-81.859 -24.070   1.339  35.430  99.870

Coefficients:

             Estimate Std. Error t value Pr(>|t|)

(Intercept) -118.3237   131.8139  -0.898  0.37702

mpg           -0.3123     3.4497  -0.091  0.92851

wt            80.9014    17.8080   4.543 9.67e-05 ***

hp             0.6479     0.2004   3.233  0.00313 **

---

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 48.18 on 28 degrees of freedom

Multiple R-squared:  0.8635,  Adjusted R-squared:  0.8489
```

F-statistic: 59.05 on 3 and 28 DF, p-value: 3.152e-12

What is meant by saying that the overall regression is significant is that the following Null Hypothesis H_0 is rejected in favor of the Alternative Hypothesis H_1, according to the associated hypothesis test p-value:

H_0: $\beta_1 = \beta_2 = \beta_3 = 0$, H_1: NOT ($\beta_1 = \beta_2 = \beta_3 = 0$),

where the 3 beta parameters correspond to the predictors mpg, wt, and hp. The low p-value for the F-test (last line of the summary report), which is much, much lower than the generally accepted threshold of 0.05, indicates that the regression overall is significant.

For the variables mpg, wt, and hp, each considered individually, significance of the regression, for each one, depends on the results of each hypothesis test

$$H_0: \beta_m = 0,\ H_1: \beta_m \neq 0,$$

For the values $m = 1,2,3,$ and the t-test associated to each variable is the relevant test. The p-values are in the center of the report (and at right) and show that wt and hp are individually significant but mpg is not.

Question 2 (Worth 10 points total)**:** Again as in Question 1, consider the mtcars dataset built into R and a corresponding regression model with disp as the response variable and mpg, wt, and hp as the predictor variables. How can you evaluate the proportion of the variance in the response variable that is explained from the predictor variables in this model? What numerical metric(s) can you use that measure this (and that is/are implemented in R) (3 points)? Can you think of more than one? What is (are) the actual numerical value(s) for this metric (or metrics) for the regression model as specified (2 points)? Do you get the same numerical value for both (1 point)? If not, why not (4 points)? Explain.

(Please enter your entire answer directly into the online answer field provided with this quiz question below. There is no need for any complicated mathematical formulas or equations in your answer. You are free to use if needed any key results stated in for example the course slides without proving them, but you should explicitly cite them if you use them -- saying where you got the result from.)

Answer: The measures you can use are the Coefficient of Determination (Multiple R^2 or just "R^2") and the Adjusted Coefficient of Determination R_a^2, and for the model as specified the respective values are Multiple R-squared: 0.8635, Adjusted R-squared: 0.8489 (see the report in the Answer to Q1). So, one does not get the same value for both. To see why, recall the definition of R_a^2:

$$R_a^2 = 1 - (1 - R^2)\frac{N-1}{N-M-1}.$$

In this case, $N = 32$ (there are 32 samples in the dataset), $M = 3,$ and $R^2 = 0.8635$. So, directly from the numerical definition above we see that we get $R_a^2 = 0.8489,$ which is different from $R^2 = 0.8635$.

Question 3 (Worth 25 points total)**:** Here in Question 3, you should answer by writing one and only of the three responses "Always true", "Always false", or "Sometimes true and sometimes false", depending on which one of these you believe to be correct, for each of the five statements (3a) through (3e) below. Please also provide a short one or two sentence reason for why you are giving the response you are for

each of the five statements in (3a) through (3e).

3a) (5 points) In Simple Linear or Multiple Linear Regression as we have defined them in this course (in the course slides), each of the residuals is normally-distributed provided that all of the model's error (noise) terms are normally-distributed.

3b) (5 points) For any given number *N* of data samples, the estimators $\hat{\beta}_m$ approximate (as measured by mean squared error) the values β_m, respectively, as closely as or more closely than any other linear, unbiased estimators in the context of Multiple Regression.

3d) (5 points) The following holds in Simple Linear Regression:

$$\sum_{n=1}^{N}(y_n - \overline{y})^2 < \sum_{n=1}^{N}(y_n - \hat{y})^2.$$

3d) (5 points) The statement, "the p-value is 0.003", is equivalent to the statement, "there is a 0.3% probability that the null hypothesis is true".

3e) (5 points) A low p-value in a hypothesis test implies, technically speaking, that the probability that the null hypothesis holds is small.

(Please enter your entire answer for each of (3a)-(3e) directly into the online answer field provided with this quiz question below. There is no need for any complicated mathematical formulas or equations in your answers. You are free to use if needed any key results stated in for example the course slides without proving them, but you should explicitly cite them if you use them -- saying where you got the result from.)

Answer: 3a) **Always true**. For the reasoning, see the top of Slide 20 of the course slides ("Normality of the Residuals") for the case of SLR (or HW #2, Problem 2). The general case for Multiple Regression is analogous, using (44) on Slide 38 ("Least-Squares Solution of Multiple Regression (cont'd)").

3b) **Always true**. The reason is due to the Gauss-Markov Theorem. See HW #2, Problem 3.

3c) **Always false**. The reason follows easily from the "DoV" equation, Equation (34) on Slide 31 ("Coefficient of Determination (R^2)").

3d) **Always false**. The p-value is not interpreted as a probability that the null hypothesis may be true. The null hypothesis itself is either true, or it is not true. In computing the p-value, it is assumed that the null hypothesis is true, so the p-value cannot indicate the probability that the null hypothesis is true. The p-value is the probability of observing a value of the test statistic that is as or more extreme than what was observed in the sample, assuming that the null hypothesis is true. That is, what a low p-value means is that, under the assumption of the Null Hypothesis H_0, the probability of seeing data at least as extreme as the test statistic is small.

3e) **Always false**. The p-value is not interpreted as a probability that the null hypothesis may be true. The null hypothesis itself is either true, or it is not true. In computing the p-value, it is assumed that the null hypothesis is true, so the p-value cannot indicate the probability that the null hypothesis is true. The p-value is the probability of observing a value of the test statistic that is as or more extreme than what was observed in the sample, assuming that the null hypothesis is true. That is, what a low p-value means is that, under the assumption of the Null Hypothesis H_0,

the probability of seeing data at least as extreme as the test statistic is small.