TEXAS A&M UNIVERSITY

Department of Statistics

STAT 608 - Regression Analysis
Homework III

Salih Kilicli

June 20, 2019

Question 1: **(Interpretation of a statistical model)**

A packaging company obtained data on the size (X_1) of a lot and the cost (Y) of assembling the lot. A scatter plot of the data suggested a broken straight line regression with a break point at lot size 250. The following linear model was formulated:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + e$$

where $X_2 = 1$ or 0 depending on whether the lot size was ≥ 250 or < 250 and where $X_3 = X_1 X_2$. Which one of the following hypotheses is equivalent to the statement: The two regression lines have the same slope?

(a) $H_0 : \beta_0 = 0$ (b) $H_0 : \beta_1 = 0$ (c) $H_0 : \beta_2 = 0$ (d) $H_0 : \beta_3 = 0$

Substantiate your answer by exhibiting appropriate algebraic manipulations.

Solution: **(d) Let's assume $X_2 = 1$ first, then the regression line is:**

$$Y = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) X_1 + e$$

Clearly, in that case slope is $(\beta_1 + \beta_3)$ whereas, when $X_2 = 0$, we have:

$$Y = \beta_0 + \beta_1 X_1 + e$$

with slope being β_1 since $X_3 = 0$, as well. Therefore, the slopes are equal in both case whenever $\beta_3 = 0$.

Question 2: (Chapter 5, Section 5.2) You have a "black box" that will calculate the residual sum of squares

$$RSS = (\mathbf{Y} - \mathbf{X}\hat{\beta})'(\mathbf{Y} - \mathbf{X}\hat{\beta})$$

(and nothing else) for **any** standard linear model $Y = X\beta + e$ that you specify. Suppose you have n observations, two predictor variables, u_1 and u_2, and that you want to test the hypothesis $H_0 : \beta_1 = \beta_2$ in the model

$$Y_j = \gamma_0 + u_{j1}\beta_1 + u_{j2}\beta_2 + \epsilon_j \ ; \ j = 1, ..., n$$

Describe how you would use the "black box" to accomplish this. (NOTE: The "Analysis of variance approach ..." with $p = 2$ on pages $135 - 136$ is a special case of this in which the common null hypothesis value of β_1 and β_2 is specified as zero. The null hypothesis in the homework question does **not** specify what the common value is.

Solution : **Let's assume that we have a "black box" that only calculates the RSS for a given model. Notice for $H_0 = \beta_1 - \beta_2 = 0$ (null) hypothesis, we have the model (Model-1):**

$$Y_j = \gamma_0 + \beta_1(u_{j1} + u_{j2}) + \epsilon_j \ ; \ j = 1, ..., n$$

and, let's say we have calculated the RSS(1) by the "black box". Similarly, for the $H_A = \beta_1 - \beta_2 \neq 0$ (alternative) hypothesis, we have the model (Model-2):

$$Y_j = \gamma_0 + \beta_1 u_{j1} + \beta_2 u_{j2} + \epsilon_j \ ; \ j = 1, ..., n$$

and let's say we also calculated the RSS(2) using "black box". Then, we can apply a partial F-test and find the p-value of the test to see if there is any evidence to reject the null hypothesis.

$$F = \frac{RSS(1) - RSS(2)/(df_1 - df_2)}{RSS(2)/df_2} = \frac{RSS(1) - RSS(2)}{RSS(2)/(n-3)}$$

since $df_1 = n-2$ and $df_2 = n-3$. If the p-value of the F-test shows significance, then Model-2 is better, i.e., we reject H_0. Otherwise, (if p-value shows insignificance) we fail to reject H_0 , and we adopt Model-1.

Question 3: (Chapter 5, Section 5.2) The output below that was obtained from fitting a three - covariate linear model,

$$Y_i = \beta_0 + \sum_{j=1}^{3} \beta_j x_{ij} + e_i$$

to observed responses:

coveriate	estimated coefficient	standard error	significance
x$_1$	$\hat{\beta}_1$	0.5	0.02
x$_2$	1.0	0.25	0.00
x$_3$	0.4	0.4	0.33

$RSS = 100$; $n = 30$; $R^2 = 0.9$ where *df* is short for "degrees of freedom", est. coef. denotes the least squares estimates of β_j , $j = 1, 2, 3$, and sig. denotes the p-value (two-sided) of the t-statistic for testing the hypothesis that the coefficient of the respective covariate is zero.

(3.1) Find an unbiased estimate of the error variance, σ^2.
(3.2) Find a possible numerical value of $\hat{\beta}_1$.
(3.3) Find a 95% confidence interval for β_2, the coefficient of X_2 in the model.
(3.4) Find the regression sum of squares, SSreg, for these data.
(3.5) Test at the 5% level the hypothesis $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$

Solution: **(3.1)**

$$S^2 = \frac{RSS}{n - (3 + 1)} = \frac{100}{26} \approx 3.846$$

(3.2) For given $p = 0.02$ **value, using t-table for 26 degrees of freedom we see that;** $T = \frac{\hat{\beta}_1 - 0}{se(\hat{\beta}_1)} = \frac{\hat{\beta}_1}{\frac{1}{2}} \approx 2.479$**, implying** $\hat{\beta}_1 \approx 1.2395$ **is a possible numerical value.**

(3.3) A 95% **confidence interval for** β_2 **is** $\beta_2 \in (1 \mp 0.25 * 2.056) = (0.486, 1.514)$**, since**

$$T_2 = \frac{\hat{\beta}_2 - \beta_2}{se(\hat{\beta}_2)} = \frac{1 - \beta_2}{0.25} \sim t_{0.025,26} = 2.056$$

(3.4) Notice, $SSReg = SST - RSS$ **and** $R^2 = 1 - \frac{RSS}{SST}$**.**

Therefore, $0.9 = 1 - \frac{100}{SST}$ **implies** $SST = 1000$**, and** $SSReg = 1000 - 100 = 900$

(3.5) Let's apply an F-test, then we get:

$$F = \frac{SSReg/3}{RSS/(30 - 3 - 1)} = \frac{900/3}{100/(26)} = 78 > 2.975 = F_{0.05,(3,26)}$$

Therefore, H_0 **is rejected.**

Question 4: (Chapter 6, Sections 6.1.1 and 6.1.2) In the standard linear model $\mathbf{Y} = \mathbf{X}\beta + \mathbf{e}$ with X fixed the vector of residuals $\hat{e}$ can be expressed compactly as

$$\hat{\mathbf{e}} = (\mathbf{I} - \mathbf{H})\mathbf{y}$$

where $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ is the **hat** matrix. The covariance matrix of the residuals is therefore equal to $(\mathbf{I} - \mathbf{H})\boldsymbol{\Sigma}(\mathbf{I} - \mathbf{H})'$, where $\mathbf{H}$ is the **covariance matrix** of the error vector $\mathbf{e}$. Assume that the errors are independent and identically distributed with variance σ^2.

(4.1) Show that H is idempotent.
(4.2) Show that the covariance matrix of the residuals reduces to $(I - H)\sigma^2$.

Solution: **[4.1]**

$$\begin{aligned} H^2 = HH = X(X'X)^{-1}X'(X(X'X)^{-1}X) &= X[\cancelto{I}{(X'X)^{-1}X'X}](X'X)^{-1}X' \\ = X[I](X'X)^{-1}X' = (X'X)^{-1}X &= X(X'X)^{-1}X' = H \end{aligned}$$

Therefore $H^2 = H$, i.e., H is an idempotent matrix.

[4.2] First, notice that H is also symmetric matrix, i.e. $H' = H$, since:

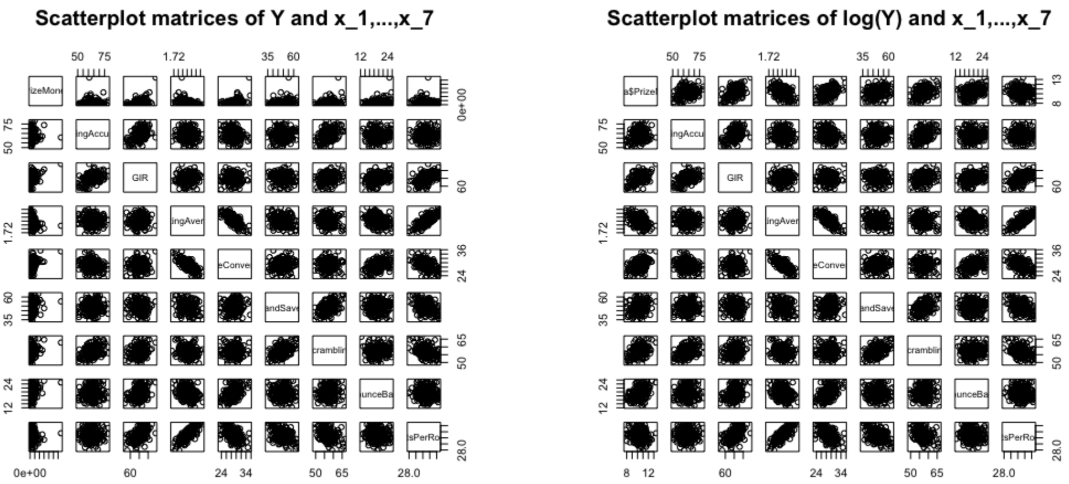
$$H' = (X(X'X)^{-1}X')' = (X')'((X'X)')^{-1}X' = XX(X'X)^{-1}X' = H$$

Similarly, $(I - H)' = I' - H' = I - H$. Now, using these two properties of H, we can show;

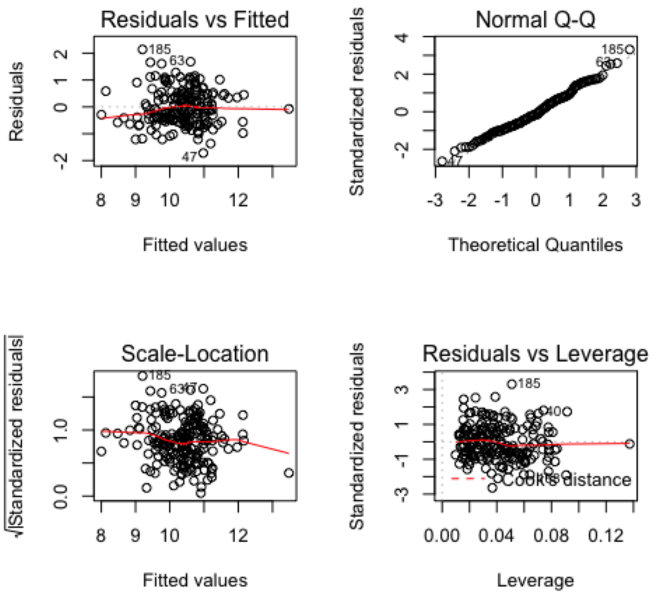
$$\begin{aligned} cov(\hat{e}) &= cov(Y - \hat{Y}) = cov(Y - HY) = cov((I - H)Y) \\ &= (I - H)\sigma^2 I(I - H)' = (I - H)(I - H)'\sigma^2 = (I - H)(I - H)\sigma^2 \\ &= (I^2 - IH - HI + H^2)\sigma^2 = (I - 2H + H)\sigma^2 = (I - H)\sigma^2. \end{aligned}$$

Question 5: Work **Exercise 5** on page 224 of our textbook. (Chapter 6)

Solution: **(a)Looking at the scatter plots given below, a log(Y) transformation helps reducing the skew in Y. All pairs appear Gaussian and so the transformation will likely lead to a good fit. A residual analysis after the fit might be helpful to confirm this approach's validity. Figure below is the comparison of scatter-plot matrices of Y vs $log(Y)$ against x_i's where $i = 1, 2, ..., 7$.**

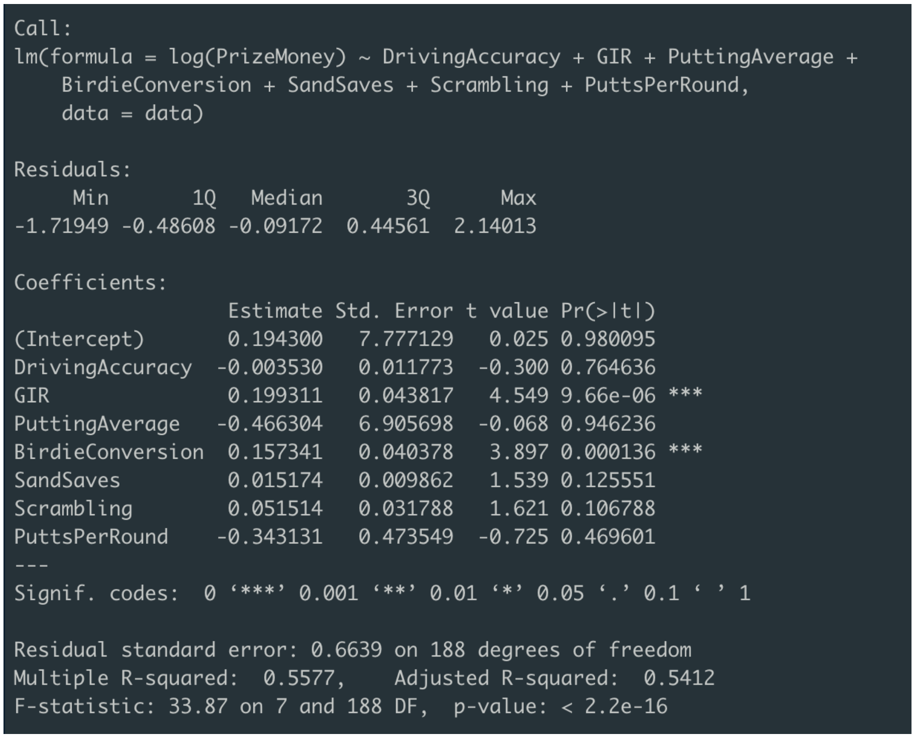


(b) The fit looks appropriate. None of the assumptions have been violated according to the plots. (Errors are approximately normally distributed with 0 mean and (almost) constant variance.)



(c)There is no data point with a large Cook's distance based on the Residual vs Leverage plot. So there are no ”bad” leverage points. However, data point 185 has a unusual standardized residual of (3.3090) for a data set with 196 observations. The next largest residual, corresponding to data point 47, is large (2.6) but arises with the expected probability for this data set. Also, data point 178 have a high leverage value, but it corresponds to Tiger Woods (the best golfer of the world, and an exceptionally good golf player). It may be helpful to see whether the parameter estimates vary more if that point was removed.

(d) The model summary below shows that overall the model is significant with F = 33.87 and a p-value that is almost zero. Nevertheless, only 2 out of 7 predictors are significant. R^2 value also might be improved. Even though, the model doesn't have an outlier, taking care of high leverage points might improve the model.



```
Call:
lm(formula = log(PrizeMoney) ~ DrivingAccuracy + GIR + PuttingAverage +
    BirdieConversion + SandSaves + Scrambling + PuttsPerRound,
    data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-1.71949 -0.48608 -0.09172  0.44561  2.14013

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       0.194300   7.777129   0.025 0.980095
DrivingAccuracy  -0.003530   0.011773  -0.300 0.764636
GIR               0.199311   0.043817   4.549 9.66e-06 ***
PuttingAverage   -0.466304   6.905698  -0.068 0.946236
BirdieConversion  0.157341   0.040378   3.897 0.000136 ***
SandSaves         0.015174   0.009862   1.539 0.125551
Scrambling        0.051514   0.031788   1.621 0.106788
PuttsPerRound    -0.343131   0.473549  -0.725 0.469601
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6639 on 188 degrees of freedom
Multiple R-squared:  0.5577,	Adjusted R-squared:  0.5412
F-statistic: 33.87 on 7 and 188 DF,  p-value: < 2.2e-16
```

(e) Removing all the non-significant predictors at once is not a good idea since correlations between predictors could hide the relationships between the response (Y) and other predictors.