**Name: Brad Hall**
**ISDS 7024**
**Homework 3**

**1)** Which of the following methods is least useful for detecting non-linearity in a regression model?

**(a)** Added variable plot
**(b)** Scatterplot with Spline Smoothing
**(c)** Leverage plot
**(d)** Residuals versus fitted values plot

*Solution:* **(c) Leverage plot**

**2)** If only the constant variance assumption is violated, the fitted values produced by the regression model will still be accurate, however, hypothesis tests may provide incorrect results. (True or False)

*Solution:* **True**.

**3)** The mean squared error of a regression model will be incorrect if the distribution of residuals is not normal. (True or false)

*Solution:* **False**.

---

Use `dataset18.xlsx` to answer the following questions.

Run the model $Y = X$. Use **diagnostic plots** to determine whether or not the model satisfies the assumptions. For each assumption, enter Yes if the assumption is violated, or No if it is not.

**4)** Linearity

*Solution:* **No.** Linearity is not violated since the $p$-value for the linear regression test is less than .0001 (way below the threshold of 0.05).

**5)** Constant variance

*Solution:* **Yes.** Constant variance is violated. This is because the $p$-value is $< .05$, which rejects the null hypothesis that states the variance is constant.

**(a)** What is the Chi-Square test statistic for the Breusch-Pagan Test?

*Solution:* $\text{LM} = (1/2)\text{SSR} = (1/2) * 831.558 = 415.7795$

**(b)** What is the $p$-value for the Breusch-Pagan Test?

*Solution:* `CHISQ.DIST.RT(LM, df)` $=$ `CHISQ.DIST.RT(415.7795, 1)` $= 0$

**6)** Normal distribution of the residuals

*Solution:* **Yes.** The assumption that the residuals are normally distribution is violated. This is because the Shapiro-Wilk test provides a $p$-value less than .0001, which rejects the null hypothesis that states the residuals are normally distributed.

**7)** Independence

*Solution:* **No.** This assumption is not violated since there is only one predictor in the model.

**8)** No influential cases

*Solution:* **Yes.** This assumption is violated by part (c) of this exercise.

**(a)** An observation is considered influential if its Cook's D is greater than **1**. (Round to 4 decimal places.)

**(b)** An observation has high leverage if its leverage value is greater than $\underline{3 * (k+1)/n = 3 * (1+1)/214 = \mathbf{0.0280}}$. (Round to 4 decimal places.)

**(c)** Do any observations have a Cook's D value greater than the threshold defined in part a above?

*Solution:* **Yes.**

- If YES for part c, which observation has the largest Cook's D? Put N/A if NO.

  *Solution:* **Observation 171**

- If YES for part c, for the observation with the largest Cook's D, what is the leverage value? (Round to 4 decimal places.) Put N/A if NO.

  *Solution:* **0.0711**

- If YES for part c, for the observation with the largest Cook's D, what is its externally studentized residual value? (Round to 3 decimal places.) Put N/A if NO.

  *Solution:* **12.922**

---

Use `Houseprices.xlsx` to answer the following questions.

**Part 1:** Run the model Sprice = Sqft. The use numerical methods (e.g., standardized values or statistical tests) to determine whether or not the assumptions listed below are violated. For each assumption, enter Yes if it is violated, or No if it is not.

**9)** Linearity

*Solution:* **No.** Linearity is not violated since the $p$-value for the linear regression test is .0003 (way below the threshold of 0.05).

**10)** Constant variance

*Solution:* **No.** Constant variance is not violated. This is because the $p$-value is $> .05$, which fails to reject the null hypothesis that states the variance is constant.

**(a)** What is the Chi-Square test statistic for the Breusch-Pagan Test?

*Solution:* $\text{LM} = (1/2)\text{SSR} = (1/2) * 0.430059 = 0.2150295$

**(b)** What is the $p$-value for the Breusch-Pagan Test?

*Solution:* `CHISQ.DIST.RT(LM, df)` = `CHISQ.DIST.RT(0.2150295, 1)` = 0.6429.

**11)** Normal distribution of the residuals

*Solution:* **Yes.** The assumption that the residuals are normally distribution is violated. This is because the Shapiro-Wilk test provides a $p$-value of 0.0049, which is less than 0.05 and thus leads us to reject the null hypothesis that states the residuals are normally distributed.

**12)** No influential cases

*Solution:* **No.** This assumption is not violated since there are no observations that have a Cook's D value greater than 1.

**Part 2:** Run the model SPrice = Sqft, then answer the questions below. (Truncate to 3 decimal places.)

**13)** What is the largest leverage value?

*Solution:* **0.198**

**14)** What is the largest (in absolute value) studentized deleted residual [i.e., externally Studentized residual]?

*Solution:* **1.927**

**15)** What is the largest Cook's distance?

*Solution:* **0.293**

**Part 3:** Consider the following predictors of Sprice: Sqft, Bdrms, Age, d1, d2.

**16)** Given that your present model is Sprice = Sqft, which variable would you add to the model next in order to increase $R^2$ as much as possible? Use the added variable plot.

**(a)** Bdrms

**(b)** Age

**(c)** d1

**(d)** d2

*Solution:* **(c) d1**

**17)** What is the $R^2$ for the model Sprice $=$ Sqft $+$ (the variable you chose)? (Truncate to 2 decimal places.)

*Solution:* $R^2 = 0.94$