STAT 2509B4
Assignment 2

Krystian Wojcicki, 101001444

Winter 2020

Indicate whether or not each of the following models can be treated as an multiple linear regression (MLR) model:

(i) $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \epsilon$, can be treated as MLR

(ii) $y = (e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2^2})\epsilon$, cannot be treated as MLR

(iii) $y = \beta_0 + \beta_1 x_1 + \beta_2 e^{x_1} + \epsilon$, can be treated as MLR

(iv) $y = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \beta_3 x_1^3 + \beta_4 x_2 + \epsilon$, can be treated as MLR

(v) $y = \beta_0 e^{\beta_1 x_1 + \beta_2 x_2} + \epsilon$, cannot be treated as MLR

(a) **State all the assumptions that are necessary for the statistical inference under the MLR model.:**
Model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon, n =$ TODO
Assumptions

(1) X_1, X_2 are observed without error

(2) ϵ's are independently distributed

(3) ϵ's have common mean 0 in other words $E(\epsilon) = 0$ for all X_1, X_2.

(4) ϵ's have common/constant variance σ^2 meaning $Var(\epsilon) = \sigma^2$ for all X_1, X_2

(5) $\epsilon \sim N(0, \sigma^2)$ for any value of X_1, X_2

(b) **Use matrices to compute the least-squares estimates of the population parameters β_0, β_1 and β_2, and obtain the fitted least-squares regression line:**

$$\underline{\text{Hint:}}\ \boldsymbol{X}^\intercal \boldsymbol{X} = \begin{bmatrix} 25.00 & 114.00 & 86.46 \\ 114.00 & 552.00 & 404.07 \\ 86.46 & 404.07 & 304.5062 \end{bmatrix}, \boldsymbol{X}^\intercal \boldsymbol{Y} = \begin{bmatrix} 2380.00 \\ 11072.00 \\ 8306.16 \end{bmatrix},$$

$$\left(\boldsymbol{X}^\intercal \boldsymbol{X}\right)^{-1} \approx \begin{bmatrix} 2{,}3963567 & 0.11058177 & -0.8271483 \\ 0.1105818 & 0.06834592 & -0.1220909 \\ -0.8271483 & -0.12209090 & 0.4001512 \end{bmatrix},$$

$$\boldsymbol{Y}^\intercal \boldsymbol{Y} = \sum_{i=1}^n y_i^2 = 228230, \text{ and } \sum_{i=1}^n y_i = 2380.$$

$$(X^T X)^{-1} X^T Y = \begin{bmatrix} 2.3963567 & 0.11058177 & -0.8271483 \\ 0.1105818 & 0.06834592 & -0.1220909 \\ -0.8271483 & -0.12209090 & 0.400151 \end{bmatrix} * \begin{bmatrix} 2380.00 \\ 11072.00 \\ 8306.16 \end{bmatrix} = \begin{bmatrix} 57.2642 \\ 5.80416 \\ 3.31483 \end{bmatrix} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = \hat{\beta}$$

(c) **Set up the ANOVA table and test for significance of the model at the significance level of $\alpha = 0.05$**

$TSS = Y^T Y - \frac{(\sum_{i=1}^n y_i)^2}{n} = \sum_{i=1}^n y_i^2 - \frac{(\sum_{i=1}^n y_i)^2}{n} = 228230 - \frac{2380^2}{25} = 1654$

$$SSR = \hat{\beta}^T (X^T Y) - \frac{(\sum_{i=1}^n y_i)^2}{n} = \begin{bmatrix} 57.2642 & 5.80416 & 3.31483 \end{bmatrix} * \begin{bmatrix} 2380.00 \\ 11072.00 \\ 8306.16 \end{bmatrix} - \frac{2380^2}{25} = 1509.9638728$$

$SSE = TSS - SSR = 1654 - 1509.9638727 = 144.0361272$ $MSR = \frac{SSR}{k} = \frac{1509.9638728}{2} = 754.9819364$
$MSE = \frac{SSE}{n-(k+1)} = \frac{144.0361272}{22} = 6.547$ $F = \frac{MSR}{MSE} = \frac{754.9819364}{6.547} = 115.317$

Source	d.f	SS	MS	F
Regression	2	1509.9638728	754.9819364	115.317
Error	22	144.0361272	6.547	
Total	24	1654		

$H_0 : \beta_1 = \beta_2 = 0$
H_a : at least one of the $\beta's \neq 0$
$\alpha = 0.05$

test-statistics: $F = \frac{MSR}{MSE} = 115.317$

Rejection region, we rejected H_0 if $F > F_{(k,n-(k+1)),\alpha} = F_{2,22;0.05} = 3.4434$

Since $F = 115.317 > 3.4434$, we reject H_0 and conclude that at a 5% level of significance there is evidence to say there is a linear relationship between age, weight and the systolic BP.

(d) **Test whether age (x1) contributes to explaining (or predicting) the systolic blood pressure (y) under the MLR model. Use t-test with $\alpha = 0.05$.**

$H_0 : \beta_1 = 0$ $H_a : \beta_1 \neq 0$ $\alpha =\rightarrow \alpha/2 =$

test statistics: $t = \frac{\hat{\beta_1}}{\sqrt{v_{11} MSE}} = \frac{5.80416}{\sqrt{0.0683459 * 6.547}} = 8.677$

Rejection region, we reject H_0 if $|t| > t_{n-(k+1),\alpha/2} = 2.07383$.

Since $t = 8.677 > 2.07$, we reject H_0 and conclud that at a 5% level of significance there is evidence to say that the x_1 term contributes to the model.

(e) **Find the values of the coefficient of determination, r^2, and the adjusted r^2. Interpret their meanings in this problem**

$r^2 = \frac{SSR}{TSS} = \frac{1509.9638727}{1654} = 0.9129 = 91.29\%$

In other words approximately 91.29% of the total variation in the data is explined by the regression line. The rest is due to error.

$r^2_{adj} = 1 - \frac{SSE/n-(k+1)}{TSS/n-1} = 1 - \frac{MSE}{TSS/n-1} = \frac{6.547}{1654/24} =$

(f) **Run SAS to verify your above results:**

$H_0 : \beta_1 = 0$ $H_a : \beta_1 \neq 0$ $\alpha =$

full model: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$ reduced model: $y = \beta_0 + \beta_2 x_2 + \epsilon$

test statistics: $F_{part} = \frac{[SSR_f - SSR_r]/[df_{SSR_f} - df_{SSR_r}]}{SSE_f / df_{SEE_f}} =$

(g) **Run SAS to verify your answers to the above questions. In addition, use the SAS output to answer subquestion (d) using the partial F-test with $\alpha = 0.05$.** See attatched SAS output

1. **A medical study was conducted to study the relationship between infants' systolic blood pressure and two explanatory variables, age (days) and weight (kg). The data for 25 infants are given below..:**

Age (x_1)	Treadmill time in minutes (x_2)	Systoli BP (y)
3	**2.61**	**80**
4	**2.67**	**90**
5	**2.98**	**96**
6	**3.98**	**102**
3	**2.87**	**81**
4	**3.41**	**96**
5	**3.49**	**99**
6	**4.03**	**110**
3	**3.41**	**88**
4	**2.81**	**90**
5	**3.24**	**100**
6	**3.75**	**102**
3	**3.18**	**86**
4	**3.13**	**93**
5	**3.98**	**101**
6	**4.55**	**103**
3	**3.41**	**86**
4	**3.35**	**91**
5	**3.75**	**100**
6	**3.83**	**105**
3	**3.18**	**84**
4	**3.52**	**91**
5	**3.49**	**95**
6	**3.81**	**104**
6	**4.03**	**107**

(a) **Draw a scatter plot (using SAS, see part (i)) to get an idea of the form of the relationship between the treadmill time (x) and 10-km running time (y). Does the scatter plot suggest an approximate linear relationship between the two variables?:** See SAS output attatched

(b) **State a simple linear regression (SLR) model for two variables and describe all assumptions that are necessary for statistical inference. :**
Model $y = \beta_0 + \beta_1 * x + \epsilon, n = 20$. Assumptions

(1) The random errors ϵ_i's are mutually independent.
(2) ϵ_i's are normally distributed
(3) ϵ_i's have common mean 0 in other words $E(\epsilon_i) = 0$ for all i.
(4) ϵ_i's have common variance σ^2 meaning $Var(\epsilon_i) = \sigma^2$ for all i
(5) x's are observed without error.

(c) **Find the least squares estimates of β_0 and β_1 in the SLR model. Find the least square fitted regression line.:**

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{\sum_{i=1}^n x_i y_i - \frac{\sum_{i=1}^n x_i \sum_{i=1}^n y_i}{n}}{\sum_{i=1}^n x_i^2 - \frac{(\sum_{i=1}^n x_i)^2}{n}} = \frac{7852.25 - \frac{195.1*812}{20}}{1940.05 - \frac{(195.1)^2}{20}} = -1.8673252 \simeq -1.87$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = \frac{\sum_{i=1}^n y_i}{n} - \hat{\beta}_1 \frac{\sum_{i=1}^n x_i}{n} = \frac{812}{20} + 1.87 * \frac{195.1}{20} = 58.815757 \simeq 58.82$$

Therefore the least square fitted regression is given by $\hat{y} = 58.82 - 1.87x$

(d) **Find s^2, an estimate of σ^2:**

$$s^2 = \frac{SSE}{n-2} = \frac{S_{yy} - \frac{S_{xy}^2}{S_{xx}}}{n-2} = \frac{(\sum_{i=1}^n y_i^2 - \frac{(\sum_{i=1}^n y_i)^2}{n}) - \frac{(\sum_{i=1}^n x_i y_i - \frac{\sum_{i=1}^n x_i \sum_{i=1}^n y_i}{n})^2}{(\sum_{i=1}^n x_i^2 - \frac{(\sum_{i=1}^n x_i)^2}{n})}}{n-2}$$

$$= \frac{(33175.2 - \frac{812^2}{20}) - \frac{(7852.25 - \frac{195.1*812}{20})^2}{1940.05 - \frac{195.1^2}{20}}}{18} = \frac{208 - 128.49}{18} = 4.41718627269 \simeq 4.42$$

Therefore $s = \sqrt{s^2} = \sqrt{4.42} = 2.10171032083 \simeq 2.10$

(e) **Use the t-test to test whether there is a significant linear relationship between 10-km running time and the treadmill time. Use $\alpha = 0.05$.:**

$H_0 : \beta_1 = 0, H_a : B_1 \neq 0$

$\alpha = 0.05 \rightarrow \alpha/2 = 0.025$

Since we are using a t-test, our test statistic is t and $t = \frac{\hat{\beta_1} - 0}{s/\sqrt{S_{xx}}} = \frac{-1.87}{2.10/\sqrt{36.8495}} = -5.3934 \simeq -5.39$

Rejection region, we reject H_0 if $|t| > t_{n-2;\alpha/2} = t_{18;0.025} = 2.101$

Since $|t| = |-5.39| = 5.39 > 2.101$, we reject H_0 and we can conclude that at $\alpha = 0.05$ or 5% level of significance there is evidence that there is a linear relationship between 10-km running time and the treadmill time.

(f) **Find a 95% confidence interval for β_1.:**

$1 - \alpha = 0.95 \rightarrow \alpha = 0.05 \rightarrow \alpha/2 = 0.025$

Therefore β_1's 95% confidence interval is

$(\hat{\beta_1} \pm t_{n-2;\alpha/2} \frac{s}{\sqrt{S_{xx}}}) = (-1.87 \pm 2.101 * \frac{2.10}{\sqrt{36.85}}) = (-2.59474163696, -1.13990876535) \simeq (-2.60, -1.14)$.

And we can be 95% confident that in repeated sampling the true value of β_1 would lie in the interval $(-2.60, -1.14)$.

(g) **Set up the ANOVA table and use it to test whether there is a significant linear relationship between 10-km running time and the treadmill time. Use $\alpha = 0.05$.:**

$TSS = S_{yy} = \sum_{i=1}^n y_i^2 - \frac{(\sum_{i=1}^n y_i)^2}{n} = 33175.2 - \frac{812^2}{20} = 208$

$SSR = \frac{S_{xy}^2}{S_{xx}} = \frac{(7852.25 - \frac{195.1*812}{20})^2}{1940.05 - \frac{195.1^2}{20}} = 128.49$

$SSE = TSS - SSR = 208 - 128.49 = 79.51$

$MSR = SSR/1 = 128.49$

$MSE = \frac{SSE}{n-2} = \frac{79.51}{18} = 4.42$

$F = \frac{MSR}{MSE} = \frac{128.49}{4.42} = 29.09$

Source	d.f	SS	MS	F
Regression	1	128.49	128.49	29.09
Error	18	79.51	4.42	
Total	19	208		

$H_0 : \beta_1 = 0, H_a : \beta_1 \neq 0$ With $\alpha = 0.05$.

Using F-test so statistic is $F = \frac{MSR}{MSE} = 29.09$

Rejection region, we reject H_0 if $F > F_{1,n-2;\alpha} = F_{1,18;0.05} = 4.41$.

Since $F = 29.09 > 4.41$ we can reject H_0 and conclude that at a 5% level of significance there is evidence of a linear relationship between the 10-km running time and the treadmill time.

(h) **Find the values of the coefficient of correlation, r, and the coefficient of determination,r^2, and interpret their meaning in this problem.:**

$r = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}} = \frac{7852.25 - \frac{195.1*812}{20}}{\sqrt{(1940.05 - \frac{195.1^2}{20}) * (33175.2 - \frac{812^2}{20})}} = -0.785966599565 \simeq -0.79$

Therefore the 10-km running time and the treadmill time are quite strongly negatively correlated with the strength of their relationship close to 78.60%.

$r^2 = \frac{SSR}{TSS} = \frac{128.49}{208} = 0.617740384615 \simeq 0.62$

Therefore approximately 61.77% of the total variation in the data can be explainded by the regression line and the remaining % is due to error. And conclusion that the model is a good fit to the data as $r^2 > 50\%$

(i) **Verify your results for (b) to (h) using SAS.** See SAS output attatched

2. **Refer to Question 1.**

   (a) **Find a 95% confidence interval for the mean value of the response variable (i.e. the 10-km running time) and a 95% prediction interval for an individual value of the response variable when the treadmill time is 9.5 minutes. What can you say about the widths of these two intervals.:**

   95% confidence interval for $E(y)$ when $x_p = 9.5$.

   $\hat{y} = 58.82 - 1.87(9.5) = 41.055$ and since $1 - \alpha \to 0.95 \to \alpha = 0.05 \to \alpha/2 = 0.025$

   Therefore $E(9.5)$ falls into the interval

   $(\hat{y} \pm t_{n-2;\alpha/2} * s * \sqrt{\frac{1}{n} + \frac{(x_p - \bar{x})^2}{S_{xx}}}) = (41.06 \pm 2.101 * 2.10 * \sqrt{\frac{1}{20} + \frac{(41.06 - 9.755)^2}{1940.05 - \frac{195.1^2}{20}}}) = (39.9953486251, 42.0646513749) \simeq$ $(40.00, 42.07)$. So we are 95% confident that after repeating sampling the mean value of the 10-km running time when the treadmill time is 9.5 minutes would fall in the interval $(40.00, 42.07)$.

   95% prediction interval for y when $x_p = 9.5$

   Therefore y falls into the interval

   $(\hat{y} \pm t_{n-2;\alpha/2} * s * \sqrt{1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{S_{xx}}}) = (41.06 \pm 2.101 * 2.10 * \sqrt{1 + \frac{1}{20} + \frac{(41.06 - 9.755)^2}{1940.05 - \frac{195.1^2}{20}}}) =$ $(36.4714602301, 45.5885397699) \simeq (36.47, 45.59)$. So we are 95% confident that after repeating sampling the vaue of the 10-km running time when the treadmill time is 9.5 minutes would fall in the interval $(36.47, 45.59)$.

   The P.I is wider than the C.I. this is expected as the variability in the error for predicting a single value is greater than the variability of error for the estimation of the mean or average value of y.

   (b) **Use SAS to answer subquestion 2(a) and compare your SAS results to your handcalculated results. (See Part (c) of the SAS example.)**

   See SAS output attatched.

3. **Perform a residual analysis to check the SLR model assumptions using SAS (see Part (b) of the SAS example). What can you conclude?**

```
footnote 'Krystian Wojcicki, 101001444';

Data Run_Time;
Input treadmill_time tenkm_time @@;
Cards;
  7.5 43.5 7.8 45.2 7.9 44.9 8.1 41.1 8.3 43.8 8.7 44.4 8.9 38.7 9.2 43.1 9.4 41.8
  9.8 43.7 10.1 39.5 10.3 38.2 10.5 43.9 10.7 37.1 10.8 37.7 10.9 39.2 11.2 35.7 11.5 37.2
  11.7 34.8 11.8 38.5
Run;

ods pdf file="example1-output.pdf";
ods graphics off;

Proc reg;
  Model tenkm_time=treadmill_time;
  Plot tenkm_time*treadmill_time;
Run;

Data Predict;
  Input treadmill_time tenkm_time;
  Cards;
    9.5 .
Run;
Data Combine;
    Set Run_Time Predict;
Run;
Proc Reg;
  Model tenkm_time=treadmill_time/CLM CLI;
Run;

Proc reg;
  Model tenkm_time=treadmill_time;
  Plot R.*P.;
  Plot R.*treadmill_time;
  Output out=res R=resids;
Run;
Proc Chart;
  vbar resids;
Run;

ods pdf close
```