

# STAT 2509B4

## Assignment 1

Krystian Wojcicki, 101001444

Winter 2020

- (a) **State the MLR model and all assumptions which are necessary for the statistical inference:**

Model  $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$ ,  $n = \text{TODO}$

Assumptions

- (1)  $X_1, X_2$  are observed without error
- (2)  $\epsilon$ 's are independently distributed
- (3)  $\epsilon$ 's have common mean 0 in other words  $E(\epsilon) = 0$  for all  $X_1, X_2$ .
- (4)  $\epsilon$ 's have common/constant variance  $\sigma^2$  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 estimates of the population parameters  $\beta_0, \beta_1, \beta_2$  and hence obtain the fitted least squares prediction line:**

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

- (c) **Set up the ANOVA table:**

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

$$SSR = \hat{\beta}^T (X^T Y) - \frac{(\sum_{i=1}^n y_i)^2}{n} =$$

$$SSE = TSS - SSR = MSR = \frac{SSR}{k} = MSE = \frac{SSE}{n-(k+1)} = F = \frac{MSR}{MSE}$$

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

$$H_0 : \beta_1 = \beta_2 = 0$$

$$H_a : \text{at least one of the } \beta' \text{'s} \neq 0$$

$$\alpha = \text{TODO}$$

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

$$\text{Rejection region, we rejected } H_0 \text{ if } F > F_{(k, n-(k+1)), \alpha}$$

- (d) **Test whether x term:**

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

$$\text{test statistics: } t = \frac{\hat{\beta}_1}{\sqrt{v_{11} MSE}}$$

(e) Find the values of:

$$r^2 = \frac{SSR}{TSS} =$$

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

(f) Run SAS to verify your above results:

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

$$\text{full model: } y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon \quad \text{reduced model: } y = \beta_0 + \beta_2 x_2 + \epsilon$$

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

1. Athletes are constantly seeking measures of the degree of their cardiovascular fitness prior to a major race. Athletes want to know when their training is at a level that will produce a peak performance. One such measure of fitness is the time to exhaustion from running on a treadmill at a specified angle and speed. The important question is then “Does this measure of cardiovascular fitness translate into performance in a 10-km running race?” Twenty experienced distance runners who professed to be at top condition were evaluated on the treadmill and then had their times recorded in a 10-km race. The data are given in the table below.:

Treadmill time in minutes (x)	10-km time in minutes (y)
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

- (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 attached
- (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  $\epsilon_i$ 's are mutually independent.
  - (2)  $\epsilon_i$ 's are normally distributed
  - (3)  $\epsilon_i$ 's have common mean 0 in other words  $E(\epsilon_i) = 0$  for all  $i$ .
  - (4)  $\epsilon_i$ 's have common variance  $\sigma^2$  meaning  $Var(\epsilon_i) = \sigma^2$  for all  $i$

(5)  $x$ 's are observed without error.

- (c) **Find the least squares estimates of  $\beta_0$  and  $\beta_1$  in the SLR model. Find the least square fitted regression line.:**

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

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

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

- (d) **Find  $s^2$ , an estimate of  $\sigma^2$ :**

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

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

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  $\alpha = 0.05$ :**

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

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

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  $\beta_1$ :**

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

Therefore  $\beta_1$ '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  $\beta_1$  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  $\alpha = 0.05$ :**

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

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

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

$$MSR = SSR/1 = 128.49$$

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

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

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.:**

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

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 explained 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 attached

## 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  $x_p = 9.5$ .

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

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 value 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 attached.

## 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
```