Stat 6021: Homework Set 3

Hyun Suk (Max) Ryoo (hr2ee)

Question 1

(R required) We will use the dataset "Copier.txt" for this question. The Tri-City Office Equipment Corporation sells an imported copier on a franchise basis and performs preventive maintenance and repair service on this copier. The data have been collected from 45 recent calls on users to perform routine preventive maintenance service; for each call, Serviced is the number of copiers serviced and Minutes is the total number of minutes spent by the service person.

Prework

```
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.0.2
```

```
## -- Attaching packages ---------------------------------------------------------------------------
```

```
## v ggplot2 3.3.2     v purrr   0.3.4
## v tibble  3.0.1     v dplyr   1.0.2
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.0
```

```
## Warning: package 'ggplot2' was built under R version 4.0.2
```

```
## Warning: package 'tidyr' was built under R version 4.0.2
```

```
## Warning: package 'readr' was built under R version 4.0.2
```

```
## Warning: package 'dplyr' was built under R version 4.0.2
```

```
## Warning: package 'stringr' was built under R version 4.0.2
```

```
## Warning: package 'forcats' was built under R version 4.0.2
```

```
## -- Conflicts ------------------------------------------------------------------------------------
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
setwd("/Users/maxryoo/Documents/MSDS/STAT6021/hw3")
data <- read.csv("copier.txt", sep="\t")
head(data)
```

```
##   Minutes Serviced
## 1      20        2
## 2      60        4
## 3      46        3
## 4      41        2
## 5      12        1
## 6     137       10
```

A

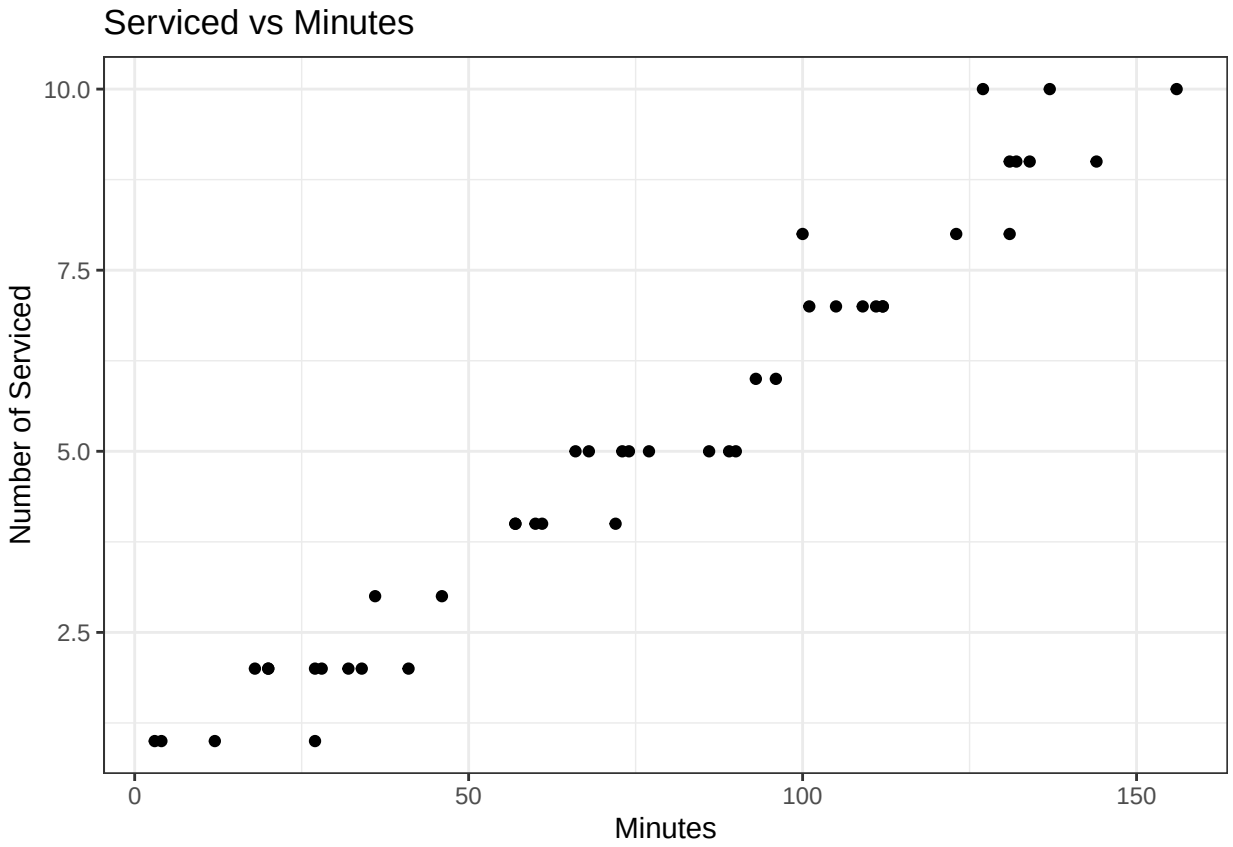
What is the response variable in this analysis? What is predictor in this analysis?

The response variable for this analysis will be the Serviced attribute. The predictor will be the minutes attribute. The analysis is based on the idea of how many people can the service person service given some amount of time.

B

Produce a scatterplot of the two variables. How would you describe the relationship between the number of copiers serviced and the time spent by the service person?

```
data %>%
  ggplot(aes(Minutes, Serviced)) +
  geom_point() +
  labs(
    title="Serviced vs Minutes",
    x="Minutes",
    y="Number of Serviced"
  ) +
  theme_bw()
```



Based on the visiual representation above, it casn be observed that the relationship is linear and has a strong posivite relationship. We know that it is a positive linear relationship since as the minutes increase the numebr of serviced users also increased. We can also observe that the strength of the relationship is strong because the points are very close to one another as minutes and serviced increased if there was a linear plot shown.

C

Use the lm() function to fit a linear regression for the two variables. Where are the values of $\hat{\beta}_1$, $\hat{\beta}_0$, R^2, and $\hat{\sigma}^2$ for this linear regression?

```
linear.model <- lm(Minutes~Serviced, data=data)
summary(linear.model)
```

```
##
## Call:
## lm(formula = Minutes ~ Serviced, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -22.7723  -3.7371   0.3334   6.3334  15.4039
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.5802     2.8039  -0.207    0.837
## Serviced     15.0352     0.4831  31.123   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.914 on 43 degrees of freedom
## Multiple R-squared:  0.9575, Adjusted R-squared:  0.9565
## F-statistic: 968.7 on 1 and 43 DF,  p-value: < 2.2e-16
```

From the above we can find the following values. $\hat{\beta}_1 = 15.0352$ $\hat{\beta}_0 = -0.5802$ $R^2 = 0.9575$ $\hat{\sigma}^2 = 8.914^2$

D

Interpret the values of $\hat{\beta}_1$ and $\hat{\beta}_0$ contextually. Does the value of $\hat{\beta}_0$ make sense in this context?

The value of $\hat{\beta}_1$ means that for every increase of minutes of the call that the service person spends there is an increase of 15.0352 completed services.

The value of $\hat{\beta}_0$ means that if a service person spent zero there will be -0.5802 services completed, which in the real world doesn't make much sense because you cannot service less than 0 calls.

E

Use the anova() function to produce the ANOVA table for this linear regression. What is the value of the ANOVA F statistic? What null and alternative hypotheses are being tested here? What is a relevant conclusion based on this ANOVA F statistic?

```
anova.tab<-anova(linear.model)
anova.tab
```

```
## Analysis of Variance Table
##
## Response: Minutes
##           Df Sum Sq Mean Sq F value    Pr(>F)
## Serviced   1  76960   76960  968.66 < 2.2e-16 ***
## Residuals 43   3416      79
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the table above we can see that the F statistic is 968.66. The null hypothesis is $H_0 : \beta_1 = 0$ and the alternative hypothesis is $H_A : \beta_1 \neq 0$ Based on the p-value also shown above we can assume that this sample

slope did not occur by chance and we reject the null hypothesis.

2

(Do not use R in this question) Suppose that for n = 6 students, we want to predict their scores on the second quiz using scores from the first quiz. The estimated regression line is $\hat{y} = 20 + 0.8x$

A

For each individual observation, calculate its predicted score on the second quiz $\hat{y}_i$ and the residual e_i. You may show your results in the table below.

x_i	70	75	80	80	85	90
y_i	75	82	80	86	90	91
$\hat{y}_i$	76	80	84	84	88	92
e_i	-1	2	-4	2	2	-1

B

Complete the ANOVA table for this dataset below. Note: Cells with *** in them are typically left blank.

	DF	SS	MS	F-Stat	p-value
Regression	1	$\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 = (76-84)^2 + ... + (92-84)^2 = 160$	160	$\frac{160}{7.5} = 21.33333$	0.0099
Residual	4	$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = (75-76)^2 + ... + (91-92)^2 = 30$	$\frac{30}{4} = 7.5$	NA	NA
Total	5	$\sum_{i=1}^{n}(y_i - \bar{y})^2 = (75-84)^2 + ... + (91-84)^2 = 190$	NA	NA	NA

C

Calculate the sample estimate of the variance σ^2 for the regression model.

The definition is as follows.

$$\begin{aligned}\sigma^2 &= \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-2} \\ &= \frac{(75-76)^2 + ... + (91-92)^2}{4} \\ &= \frac{30}{4} \\ &= 7.5\end{aligned}$$

D

What is the value of R^2 here?

$$\begin{aligned}R^2 &= 1 - \frac{SSE}{SST} \\ &= 1 - \frac{30}{190} \\ &= 1 - 0.1578947 \\ R^2 &= 0.8421053\end{aligned}$$

E

Carry out the ANOVA F test. What is an appropriate conclusion?

The null hypothesis is $H_0 : \beta_1 = 0$ and the alternative hypothesis is $H_A : \beta_1 \neq 0$. For this case our F-Statistic was 21.33333. With this F-Statisics the p-value showed to be 0.0099, which is lower than an alpha of 0.05 and even lower than 0.01. This means that this slope of the regression line did not occur by chance and thus we can reject the null hypothesis.

3

Using the following equations show the following equations hold.

- Given

$$\begin{aligned}
\hat{\beta}_1 &= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \\
\hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x} \\
SS_{res} &= \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \\
\hat{y}_i &= \hat{\beta}_0 + \hat{\beta}_1 x_i \\
e_i &= y_i - \hat{y}_i
\end{aligned}$$

- Show (Equation 6)

$$\begin{aligned}
\sum_{i=1}^{n} e_i &= 0 \\
\sum_{i=1}^{n} y_i - \hat{y}_i &= 0 \\
\sum_{i=1}^{n} y_i - \hat{\beta}_0 + \hat{\beta}_1 x_i &= 0 \\
\sum_{i=1}^{n} y_i - \bar{y} - \hat{\beta}_1 \bar{x} + \hat{\beta}_1 x_i &= 0 \\
\sum_{i=1}^{n} y_i - \bar{y} + \sum_{i=1}^{n} -\hat{\beta}_1 \bar{x} + \hat{\beta}_1 x_i &= 0 \\
0 + \sum_{i=1}^{n} -\hat{\beta}_1 \bar{x} + \hat{\beta}_1 x_i &= 0 \\
\sum_{i=1}^{n} -\hat{\beta}_1 \bar{x} + \hat{\beta}_1 x_i &= 0 \\
\hat{\beta}_1 \sum_{i=1}^{n} -\bar{x} + x_i &= 0 \\
\hat{\beta}_1 * 0 &= 0 \\
0 &= 0
\end{aligned}$$

The above shows that the sum of residuals is equal to 0

- Shown (Equation 7)

$$\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \hat{y}_i$$
$$\sum_{i=1}^{n} e_i + \hat{y}_i = \sum_{i=1}^{n} \hat{y}_i$$
$$\sum_{i=1}^{n} e_i + \sum_{i=1}^{n} \hat{y}_i = \sum_{i=1}^{n} \hat{y}_i$$
$$0 + \sum_{i=1}^{n} \hat{y}_i = \sum_{i=1}^{n} \hat{y}_i$$
$$0 = 0$$

The above states that the sum of observed values equals to the sum of the predicted values by the regression equation * Show (equation 8)

$$\sum_{i=1}^{n} x_i e_i = 0$$
$$\sum_{i=1}^{n} x_i * (y_i - \hat{y}_i) = 0$$
$$\sum_{i=1}^{n} x_i * (y_i - \hat{\beta}_0 + \hat{\beta}_1 x_i) = 0 \text{ (By taking partial derivative of } SS_{res} \text{ shown below)}$$

$$\begin{aligned} SS_{res} &= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \\ &= \sum_{i=1}^{n} (y_i - \hat{\beta}_0 + \hat{\beta}_1 x_i)^2 \\ \frac{\partial\ SS}{\partial \hat{\beta}_0} &= -2 \sum_{i=1}^{n} (y_i - \hat{\beta}_0 + \hat{\beta}_1 x_i)^2 \\ \frac{\partial\ SS}{\partial \hat{\beta}_1} &= -2 x_i \sum_{i=1}^{n} (y_i - \hat{\beta}_0 + \hat{\beta}_1 x_i)^2 \end{aligned}$$

The above shows that the sum of the residuals weighted by x_i is always 0.

- Show (equation 8)

$$\sum_{i=1}^{n} \hat{y}_i e_i = 0$$
$$\sum_{i=1}^{n} (\hat{\beta}_0 + \hat{\beta}_1 x_i) e_i = 0$$
$$\sum_{i=1}^{n} \hat{\beta}_0 e_i + \hat{\beta}_1 x_i e_i = 0$$
$$\hat{\beta}_0 \sum_{i=1}^{n} e_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i e_i = 0$$
$$0 + 0 = 0$$

The above shows that the residuals weighted by the corresponding fitted value is equal to zero.