HW4

Hyun Suk (Max) Ryoo (hr2ee)

9/26/2021

1

For this question, you will use the dataset "Copier.txt" for this question. This is the same data set that you used in the last homework. 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. It is hypothesized that the total time spent by the service person can be predicted using the number of copiers serviced. Fit an appropriate linear regression and answer the following questions:

- Prework

```r
library(tidyverse)
```

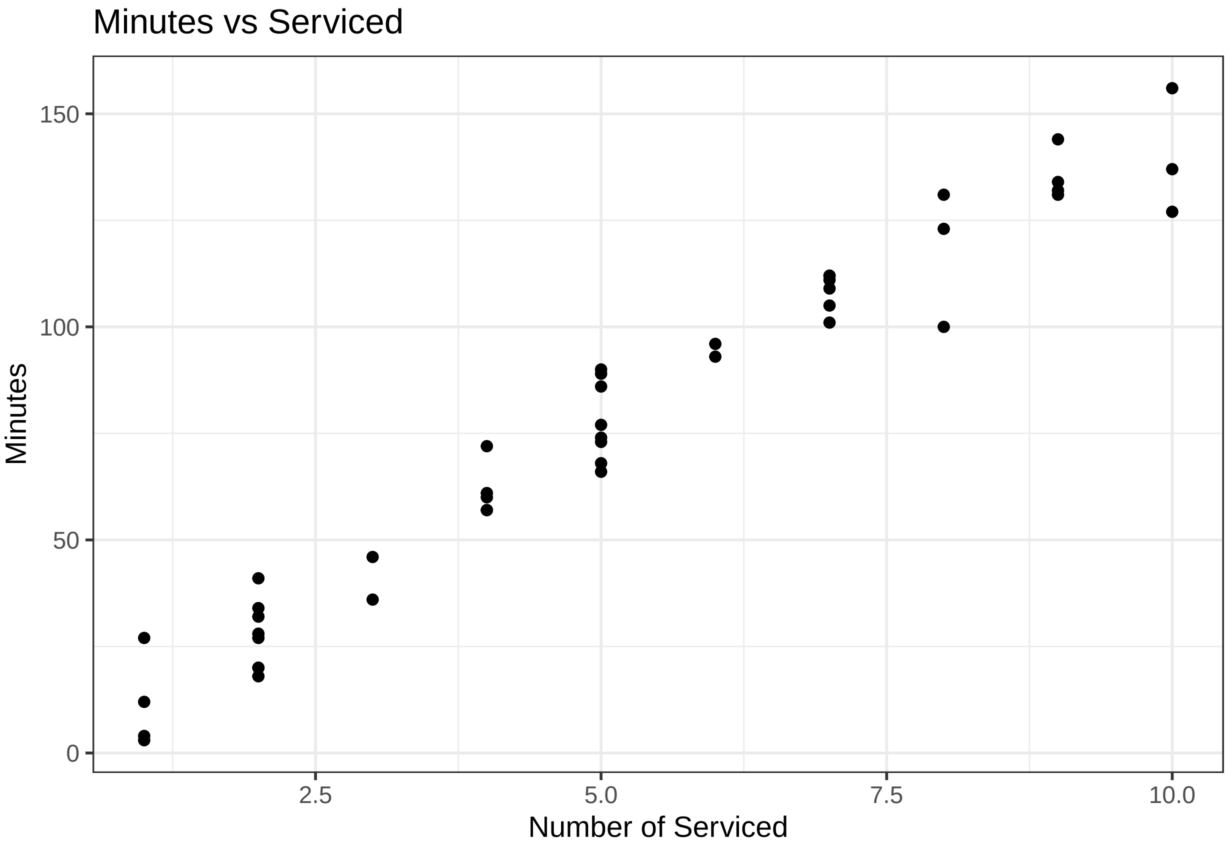
```
## Warning: package 'tidyverse' was built under R version 4.0.2

## -- 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()
```

```r
setwd("/Users/maxryoo/Documents/MSDS/STAT6021/Hw4")
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) Produce an appropriate scatterplot and comment on the relationship between the total time spent by the service person and the number of copiers serviced.

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



- (b) What is the correlation between the total time spent by the service person and the number of copiers serviced? Interpret this correlation contextually.

```
cor(data)
```

```
##           Minutes Serviced
## Minutes  1.000000 0.978517
## Serviced 0.978517 1.000000
```

The correlation is 0.978517. The correlation value shows that the two variables has a positive strong correlation.

- (c) Can the correlation found in part 1b be interpreted reliably? Briefly explain.

Yes. The correlation shows a high value as well as the scatter plot shows that it is a linear relationship.

- (d) Obtain the 95% confidence interval for the slope, β_1

```
result<-lm(Minutes~Serviced, data=data)
confint(result,level = 0.95)
```

```
##                  2.5 %    97.5 %
## (Intercept) -6.234843  5.074529
## Serviced    14.061010 16.009486
```

The confience interval for the slope, β_1 is (14.061010 16.009486)

- (e) Suppose a service person is sent to service 5 copiers. Obtain an appropriate 95% interval that predicts the total service time spent by the service person.

```
newdata <- data.frame(Serviced=5)
predict(result, newdata, interval="prediction")
```

```
##        fit      lwr      upr
## 1 74.59608 56.42133 92.77084
```

56.42133, 92.77084

- (f) What is the value of the residual for the first observation? Interpret this value contextually.

```
newdata <- data.frame(Serviced=data[1,]$Serviced)
pred = predict(result, newdata)
residual = data[1,]$Minutes - pred
residual
```

```
##         1
## -9.490339
```

The prediction from the lienar model was 9.490339 minutes above the actual observed minutes.

- (g) What is the average value of the all the residuals? Is this value surprising (or not)? Briefly explain.

```
all_data_serviced <- data.frame(Serviced= data$Serviced)
all_pred = predict(result, all_data_serviced)
sum(data$Minutes - all_pred)
```

```
## [1] 7.958079e-13
```

The sum residuals is close to zero. Theoretically it should be zero, but this could result from computer computation being only available for floating point numbers. However, the result shows that the value is close to zero, if not just zero. This result is not surprising at all.

2

(No R required) A substance used in biological and medical research is shipped by airfreight to users in cartons of 1000 ampules. The data consist of 10 shipments. The variables are number of times the carton was transferred from one aircraft to another during the shipment route (transfer), and the number of ampules found to be broken upon arrival (broken). We want to fit a simple linear regression. A simple linear regression model is fitted using R. The corresponding output from R is shown next, with some values missing.

- (a) Carry out a hypothesis test to assess if there is a linear relationship between the variables of interest.

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

In the analysis provided we know that the linear model has a slope of 4.0 and a Standard Error of 0.4690. To carry out a t test the degree of freedom is needed, which is $DF = n - 2 \rightarrow 10 - 2 \rightarrow 8$.

Therefore, the test statitics is defined as such $t = \frac{\beta_1}{SE} \rightarrow \frac{4}{0.4690} \rightarrow 8.52878464819$. This t-statistic shows that the probability of this relationship being just by chance is nearly 0. Therefore, we can reject the null hypothesis.

- (b) Calculate a 95% confidence interval that estimates the unknown value of the population slope.

The confidence interval on the slope β_1 is given by $\hat{\beta_1} - t_{\frac{\alpha}{2},n-2} \leq \hat{\beta_1} \leq \hat{\beta_1} + t_{\frac{\alpha}{2},n-2}$

$$\begin{aligned}
\hat{\beta_1} - t_{\frac{\alpha}{2},n-2} * SE(\hat{\beta_1}) \leq \hat{\beta_1} &\leq \hat{\beta_1} + t_{\frac{\alpha}{2},n-2} * SE(\hat{\beta_1}) \\
4 - t_{\frac{0.05}{2},8} * SE(\hat{\beta_1}) \leq \hat{\beta_1} &\leq 4 + t_{\frac{0.05}{2},8} * SE(\hat{\beta_1}) \\
4 - 2.306 * 0.4690 \leq \hat{\beta_1} &\leq 4 + 2.306 * 0.4690 \\
2.918486 \leq \hat{\beta_1} &\leq 5.081514
\end{aligned}$$

Thus the confidence interval is shown above for the population slope.

- (c) A consultant believes the mean number of broken ampules when no transfers are made is different from 9. Conduct an appropriate hypothesis test (state the hypotheses statements, calculate the test statistic, and write the corresponding conclusion in context, in response to his belief).

For this case we can state the following.

$$H_0 : \beta_0 = 9 H_A : \beta_0 \neq 9$$

We can utilize the test statistics of $\frac{\hat{\beta_0} - 9}{SE(\hat{\beta_0})} \rightarrow \frac{10.2-9}{0.6633} \rightarrow \frac{1.2}{0.66343} \rightarrow 1.80878163484$. The p-value for this test-statistic is 0.108 based on $2 * (1 - pt(1.809, 8))$. Therefore since this is greater than 0.05 we fail to reject the null hypothesis.

- (d) Calculate a 95% confidence interval for the mean number of broken ampules and a 95% prediction interval for the number of broken ampules when the number of transfers is 2.

The 95% CI for the mean number of broken ampules can be found like such for when the number of transfers is 2.

$$\begin{aligned}
\hat{\mu_0} - t_{0.975,8} SE_{res} \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum (x_i - \bar{x})^2}} \leq \hat{\mu_0} & & \leq \hat{\mu_0} + t_{0.975,8} SE_{res} \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum (x_i - \bar{x})^2}} \\
(10.2 + 4 * 2) - 2.306 * 1.483 \sqrt{\frac{1}{10} + \frac{(2-1)^2}{10}} \leq \hat{\mu_0} & & \leq (10.2 + 4 * 2) + 2.306 * 1.483 \sqrt{\frac{1}{10} + \frac{(2-1)^2}{10}} \\
18.2 - 2.306 * 1.483 * 0.4472136 \leq \hat{\mu_0} & & \leq 18.2 + 2.306 * 1.483 * 0.4472136 \\
16.67062 \leq \hat{\mu_0} & & \leq 19.72938
\end{aligned}$$

The 95% PI for the number of broken abroken ampules when the number of transfers is 2.

$$\begin{aligned}
\hat{y_0} - t_{0.975,8} SE_{res} \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum (x_i - \bar{x})^2}}, \hat{y_0} + t_{0.975,8} SE_{res} \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum (x_i - \bar{x})^2}} \\
18.2 - 2.306 * 1.483 \sqrt{1 + \frac{1}{10} + \frac{(2-1)^2}{10}}, 18.2 + 2.306 * 1.483 \sqrt{1 + \frac{1}{10} + \frac{(2-1)^2}{10}} \\
18.2 - 2.306 * 1.483 * 1.095445, 18.2 + 2.306 * 1.483 * 1.095445 \\
14.4538, 21.9462
\end{aligned}$$

- (e) What happens to the intervals from the previous part when the number of transfers is 1? (Describe what happens without calculating)

The confidence intervals will become more close with one another. The square root will yeild a smaller output and make the intervals closer to one another. It will get narrower since the transfer is closer to the man than previous.

- (f) What is the value of the F statistic for the ANOVA table?

F Statistics can be computed like such. $F = \frac{MS}{MS_{res}} \rightarrow \frac{160}{2.2} \rightarrow 72.72727$

- (g) Calculate the value of R^2, and interpret this value in context.

R^2 can be computed like such. $R^2 = \frac{SSR}{SST} \rightarrow \frac{160}{160+17.6} \rightarrow 0.9009009$

With the computed R^2 value we can say that about 90% of the variation in the nmumber of broken amupules is explained by the number of transfers.

3

(No R required) Suppose that the population slope for a straight-line relationship between y and x is 0.

- (a) Describe how the straight line would look in a plot of y versus x

The line will look like a straight horizantal line.

- (b) Explain why a slope that is equal to 0 would indicate that y and x are not linearly related, and why a slope that is not equal to 0 would indicate that y and x are linearly related.

If a slope is equal to 0, then the value of y (response) will be the same regardless of what the x value is. However, if a slope not equal to zero , the response (y) will be dependent upon the value of x and will change. The previous case of the y value being the same regardless of x means that they are not linearly related.