M03 Homework

Matt Scheffel

2022-09-19

Part 1

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6     v purrr   0.3.4
## v tibble  3.1.8     v dplyr   1.0.9
## v tidyr   1.2.0     v stringr 1.4.1
## v readr   2.1.2     v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

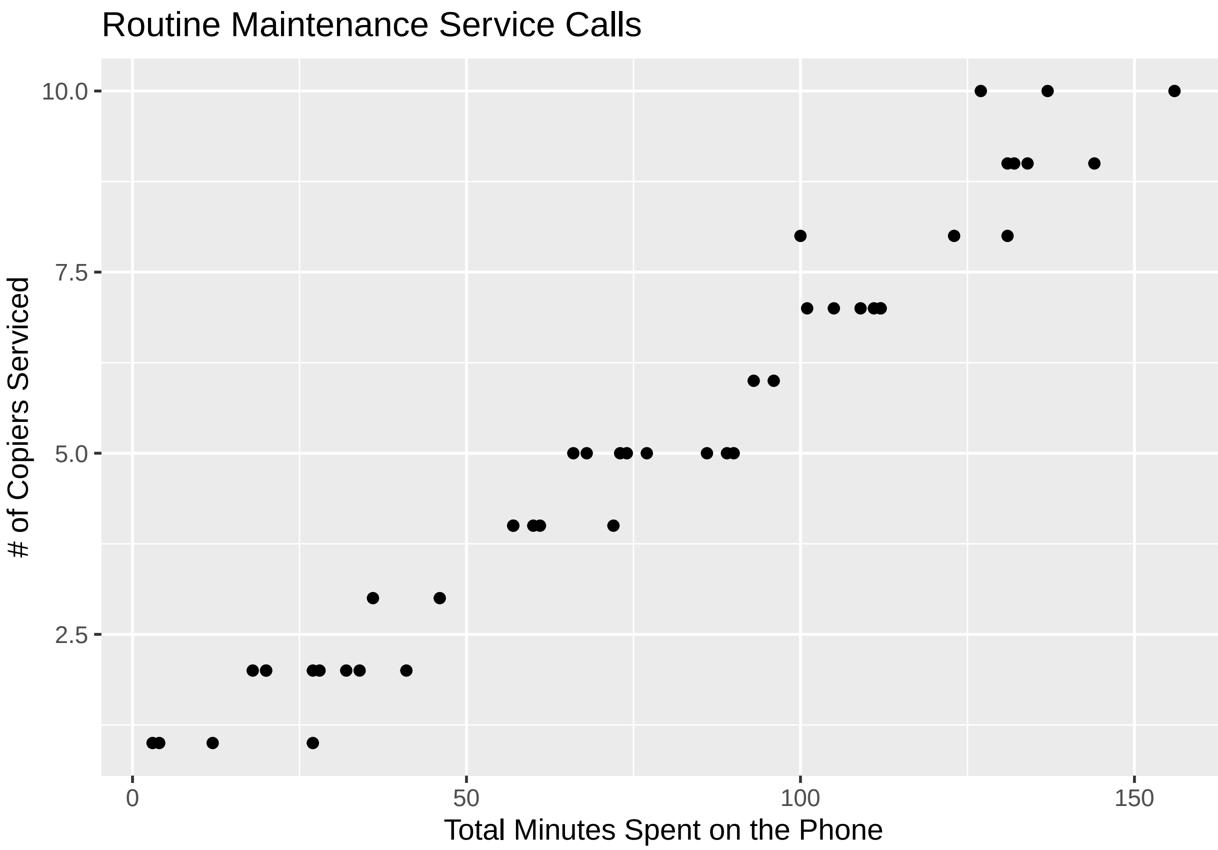
```
copier = read.table("copier.txt", header = TRUE)
```

1A

The response variable in this analysis is the number of minutes spent on the phone. The predictor in this analysis is the number of copiers serviced.

1B

```
#create scatter plot
ggplot(copier, aes(x=Minutes,y=Serviced))+
geom_point()+
labs(x="Total Minutes Spent on the Phone",
y="# of Copiers Serviced",
title="Routine Maintenance Service Calls")
```



The relationship between the 2 variables (the number of copiers serviced and the time spent by the service person) seems to be a strong positive linear relationship. The more copiers a person services, the more time they spend on the phone.

1C

```
copierlm<-lm(Minutes~Serviced, data=copier)
summary(copierlm)
```

```
##
## Call:
## lm(formula = Minutes ~ Serviced, data = copier)
##
## 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
```

B^1 is = 15.0352. It is found under the Serviced Estimate. B^0 is = -0.5802. It is found under the (Intercept) Estimate. R^2 is = 0.9565. It is found in the section where it reads "R-squared." I used the Adjusted R-squared. o^2 is = (8.914)^2. It is square of the residual std. error.

1D

Contextual interpretation: B0 does not make sense, as a negative value would indicate that a person has negative amount of calls. The minimum should be 0.

1E

```
anova.copier<-anova(copierlm)
anova.copier
```

```
## 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
```

F-statistic: 968.66

Null and alternative hypotheses: H0 : B1 = 0; Ha : B1 /= 0

Relevant conclusion: The p-value is less than 0.05 so we reject the null hypothesis. The data supports the hypothesis that there is a linear relationship between number of machines serviced and minutes spent on the phone.

Part 2

2A

See picture attached.

2B

See picture attached.

2C

Variance is = 38

2D

R^2 = 0.9733

2E

$F0 = 145.6$

Relevant conclusion: The p-value is less than 0.05 so we reject the null hypothesis.

Part 3

See picture attached.

$$R^2 = \frac{SSR}{SST}$$ | $SSR = \sum(\hat{y}_i - \bar{y})^2$ on

#2 a)

$\hat{y} = 20 + 0.8x$, $n = 6$

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

$\boxed{e_i = y_i - \hat{y}_i}$

2b) mean = (70 + 75 + 80 + 80 + 85 + 90) / 6 = 480/6 = $\boxed{80}$

$SSR: n\sum(y_i - \bar{y})^2$

$SSR^1 = 6\sum(75 - 84)^2 = 486$
$SSR^2 = 6\sum(82 - 84)^2 = 24$
$SSR^3 = 6\sum(80 - 84)^2 = 48$
$SSR^4 = 6\sum(86 - 84)^2 = 24$
$SSR^5 = 6\sum(90 - 84)^2 = 216$
$SSR^6 = 6\sum(91 - 84)^2 = 294$

$\sum = 1092$

$SSE = \sum(\hat{y} - y_i)^2$

$SSE^1 = \sum(76 - 75)^2 = 1$
$SSE^2 = \sum(80 - 82)^2 = 4$
$SSE^3 = \sum(84 - 80)^2 = 16$
$SSE^4 = \sum(84 - 86)^2 = 4$
$SSE^5 = \sum(88 - 90)^2 = 4$
$SSE^6 = \sum(92 - 91)^2 = 1$

$\sum = 30$

$R^2 = \frac{SSR}{SSt} = \frac{1500}{1650} = 0.85$

$SST = SSR + SSE$

$\rightarrow SST = 1092 + 30 = 1122$

ANOVA Table

	DF	SS	MS	F-stat	p-value
Regression	1	1092	1122	149.6	0.0099
Residual	4	30	7.5	**	**
Total	5	1122	**	**	**

$\boxed{\begin{matrix} k = 2 \\ n = 6 \end{matrix}} \rightarrow$ Regression $df = k - 1 = 2 - 1 = 1$

Residual $df = n - k = 6 - 2 = 4$

Total $df = n - 1 = 6 - 1 = 5$

$\text{Reg. MS} = \frac{SST}{df_{reg}} = \frac{1122}{1} = 1122$

$\text{Res. MS} = \frac{SSE}{df_{res}} = \frac{30}{4} = 7.5$

$\text{F-stat} = \frac{MS_{reg}}{MS_{res}} = \frac{1122}{7.5} = 149.6$

$R^2 = 1 - \frac{SSE}{SST} = 1 - \frac{30}{1122} = 0.9733$

$$\hat{y} = B_0 + B_1 x$$

$$\hat{y} = -0.5802 + 15.0352x$$

ANOVA F-test

$$F_0 = \frac{SS_R / 1}{SS_{res}/(n-2)} = \frac{1092}{30/(4)} = \boxed{145.6}$$

$F_0 = 145.6$

$F^* = 149.6$

$$\sum_{i=1}^{n} \hat{y}_i e_i = 0. \qquad (9)$$

Hint: Deriving the partial derivatives of the SS_{res}, (3), with respect to $\hat{\beta}_1$ and $\hat{\beta}_0$ will be useful.

Also, give a one-sentence interpretation of what the equalities (6) to (9) mean.

MacBook Air

2c) $s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1} = \cancel{\frac{1092}{5}} \cancel{= 245.4}$

$\sum \frac{(75-84)^2}{5} = 16.2$

$\sum \frac{(82-84)^2}{5} = 0.8$

$\sum \frac{(80-84)^2}{5} = 3.2$

$\sum \frac{(86-84)^2}{5} = 0.8$

$\sum \frac{(90-84)^2}{5} = 7.2$

$\sum \frac{(91-84)^2}{5} = 9.8$

$\sum \boxed{38} \div 6 = 6.333$

$81 + 4 + 16 + 4 + 36 + 49$

3) $\hat{B}_1 = \dfrac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$

& $\hat{B}_0 = \bar{y} - \hat{B}_1 \bar{x}$

$$\hat{B}_0 + \sum_{i=1} \frac{(x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \bar{y}$$

$$B_0 + \frac{\sum \left(x_i - \sum \frac{x_i}{n}\right)}{\sum \left(x_i - \sum \frac{x_i}{n}\right)^2} = \sum \frac{y_i}{n}$$

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

$$\to \frac{1}{n} \sum_{i=1}^{n} y_i = \frac{1}{n} \sum \hat{y}_i + \frac{1}{n} \sum \hat{e}_i = \frac{1}{n} \sum \hat{y}_i + 0 = \frac{1}{n} \sum \hat{y}_i$$

The sum of the observed values = 0.