Homework3

Dima Mikhaylow

9/21/2021

Linear regression based on `copier.txt` data

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

- Response variable is “Services” and predictor is “Minutes”. In other words, input X is time, i.e number of minutes spent, and the output Y is an estimate of how many devices coud be serviced.

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?

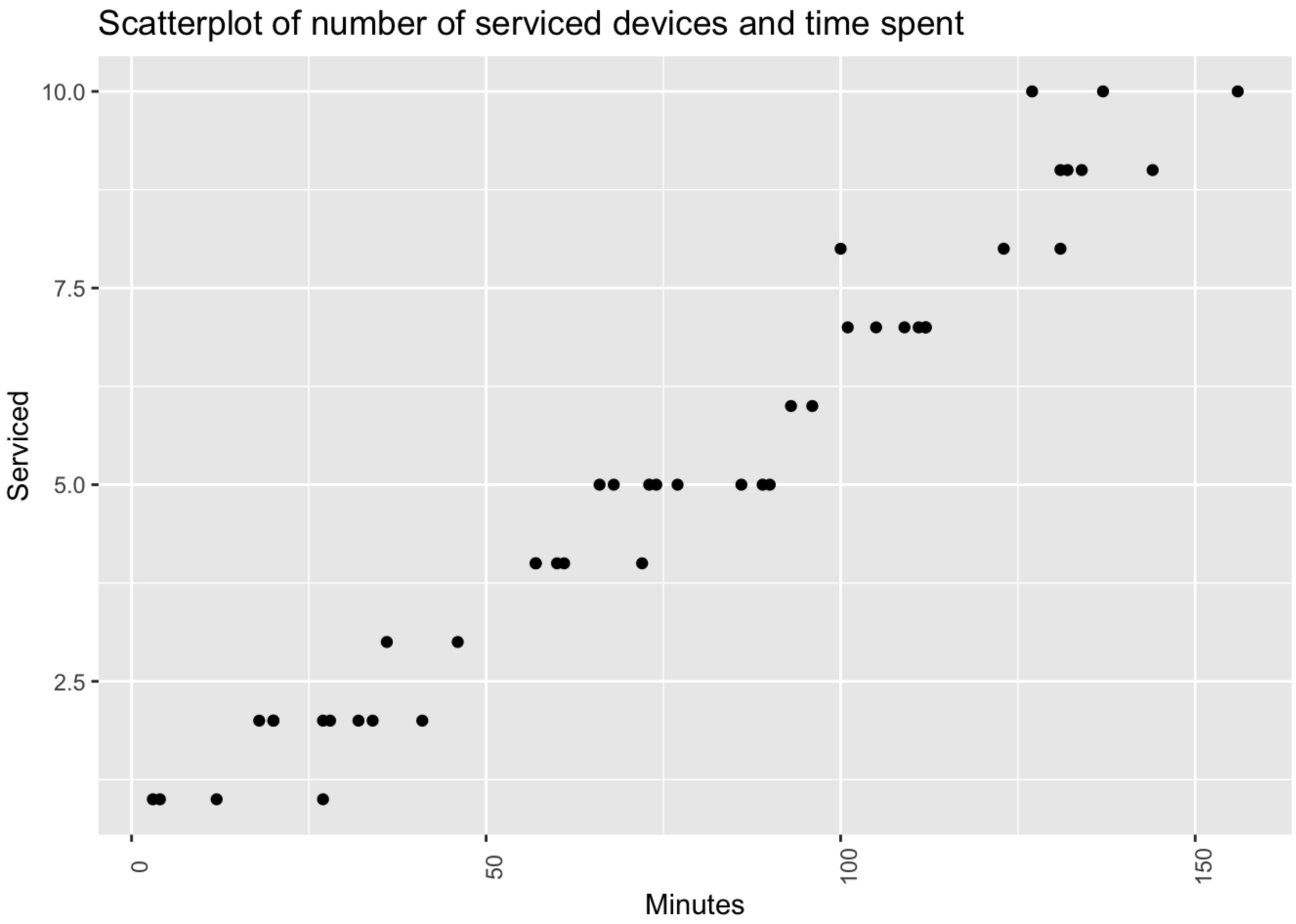
```
library(tidyverse)
```

```
## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.1 ──
```

```
## ✓ ggplot2 3.3.5     ✓ purrr   0.3.4
## ✓ tibble  3.1.4     ✓ dplyr   1.0.7
## ✓ tidyr   1.1.3     ✓ stringr 1.4.0
## ✓ readr   2.0.1     ✓ forcats 0.5.1
```

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

```
ggplot(Data, aes(x=Minutes, y=Serviced))+
  geom_point()+
  theme(axis.text.x=element_text(angle=90))+
  labs(x="Minutes",
       y="Serviced",
       title="Scatterplot of number of serviced devices and time spent")
```



Notes:

- Relationship appears to be linear.

c. Use the `lm()` function to fit a linear regression for the two variables.

```
result <- lm(Serviced ~ Minutes, data=Data)
summary(result)
```

```
##
## Call:
## lm(formula = Serviced ~ Minutes, data = Data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.98570 -0.36780 -0.03733  0.40328  1.65802
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.254192   0.178413   1.425    0.161
## Minutes     0.063683   0.002046  31.123   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5801 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
```

Notes:

- B0 (Intercept) is 0.25
- B1 (Slope for Minutes) is 0.06
- R-squared in 0.96
- Residual std. error is 0.58

d. Interpret the values of B1_hat and B0_hat contextually. Does the value of B0_hat make sense in this context?

Notes:

- B1_hat is positive and significant, it can be interpreted as an increase in number devices Serviced per unit of time spent.
- B0_hat is small and positive, but it probably does not have contextual meaning as there are no zero predictor observations.

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(result)
anova.tab
```

```
## Analysis of Variance Table
##
## Response: Serviced
##           Df Sum Sq Mean Sq F value    Pr(>F)
## Minutes    1 325.97  325.97  968.66 < 2.2e-16 ***
## Residuals 43  14.47    0.34
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Notes

F statistic is 968.7 and statistically significant. ANOVA F-test null hypothesis (H0: B1=0) is that slope is zero, hence no relationship. Alternative is that the slope is not zero, hence some relationship. Conclusion is that we have to reject H0 and accept alternative, hence the slope B1 is probably not zero.