Homework4

Dima Mikhaylow

9/27/2021

1. Questions based on “Copier.txt” dataset.

```
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
```

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:

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

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.

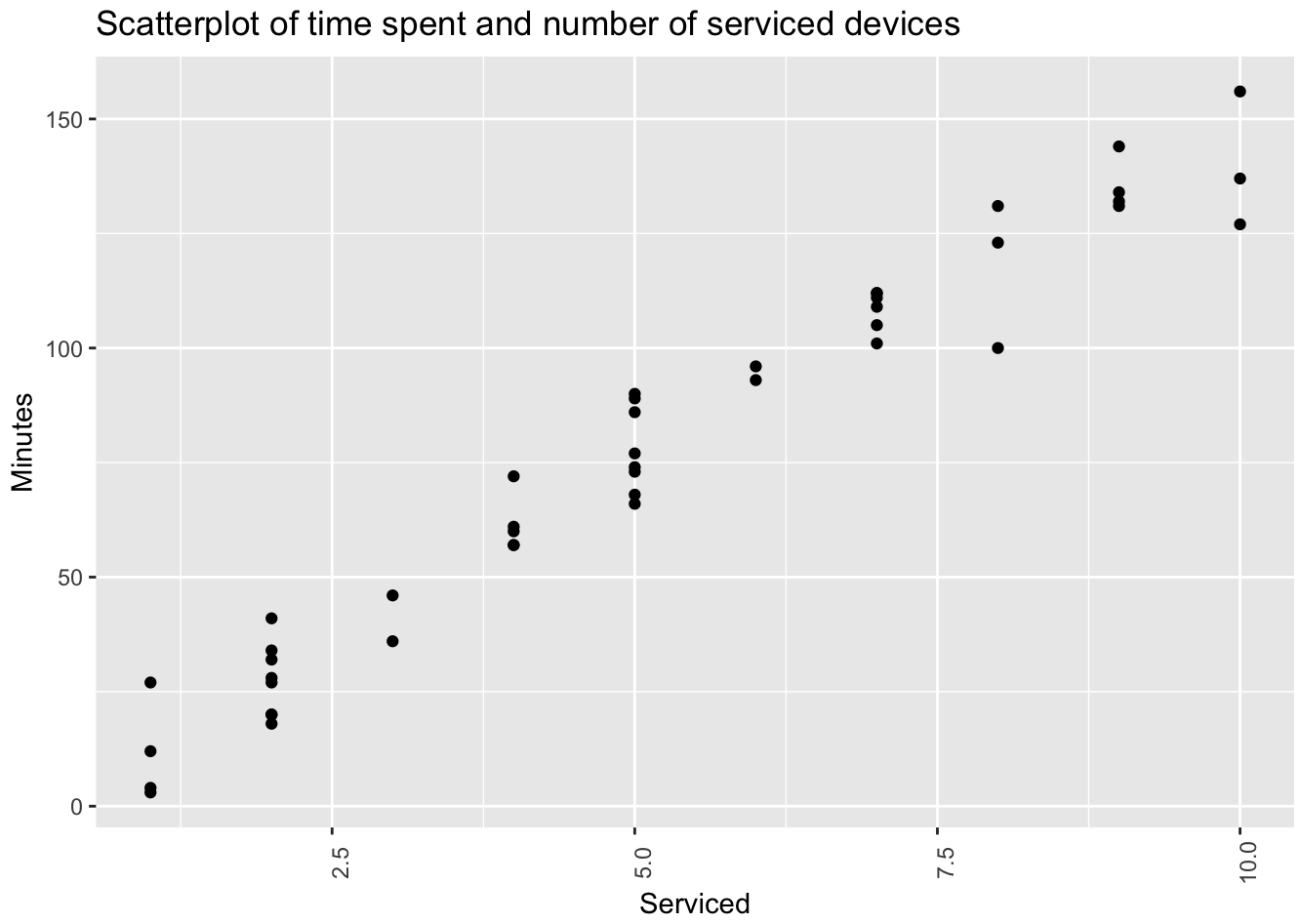
```
library(tidyverse)
```

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



Comment: there seems to be a linear relationship between the variables, as the time spent tends to increase with the number of serviced devices.

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$Minutes, Data$Serviced)
```

```
## [1] 0.978517
```

Comment: correlation of 0.97 suggests a possibility of a linear relationship between the variables, not a proof though. Positive sign indicates that they both increase at the same time. Absolute value close to 1 suggests that the linear relationship, if present, could be strong. Also, correlation is only a measure of association and is of little use in prediction.

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

Comment: no, it can't, because it may confuse other nonlinear relationships for a linear relationship.

d. Obtain the 95% confidence interval for the slope, B1

First, check the slope coefficient, B1 = 15.03

```
summary(result)$coefficients
```

```
##               Estimate Std. Error    t value     Pr(>|t|)
## (Intercept) -0.5801567  2.8039411 -0.2069076 8.370587e-01
## Serviced    15.0352480  0.4830872 31.1232581 4.009032e-31
```

Second, compute 95% CI:

```
confint(result, level = 0.95)
```

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

This is prediction interval for a response for a given x problem. Use `interval="prediction"`:

```
new_data <- data.frame(Serviced=5)
predict(result, new_data, level=0.95, interval="prediction")
```

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

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

This should be stored as a first element of residuals vector in the result object:

```
summary(result)$residuals[1]
```

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

Comment: in this case, the residual is equal to a difference between the actual value (y_i) of the first observation and the predicted value (y_hat). For the first observation, predicted value is much larger than the actual observed value and this is why the residual is negative 9.49. In other words, regression line goes above the first observation.

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

```
mean(result$residuals)
```

```
## [1] -2.612204e-16
```

Comment: the average value of the all the residuals seems to be very small, but it is not surprising as, in theory, it should be zero. It is not exactly zero probably due to possible floating-point numrical errors.

2. Calculating the missing values for the transfers problem.

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

Comment: the model aims to predict number of broken from number of transfers. This requires checking if B1 (slope) associated with transfers is statistically different from zero. H0 will be B1=0 and Ha: B1 != 0 (not zero). If H0 is true than the value of F statistics, i.e Regression Sum of Squares divided by Mean Squared Error, will be close to 1. How close to 1 is 160/2.2 = 72.7? It is very much different from 1, so H0 is likely false, and the slope B1 is in fact different from zero and thus there is a liner relationship between the variables of interest.

```
1 - pt(160/2.2, 8)
```

```
## [1] 7.11653e-13
```

Alternatively, t-test for the slop shoud give the same result as F statistic. Check if B1 is not different from zero: t_stat = (B1_hat-0)/se_of_B1_hat. Statistic is very large compared to critical region at alpha = 5%:

```
((4-0)/0.469)
```

```
## [1] 8.528785
```

```
#pt(-((4-0)/0.469), 8)*2
```

Critical t:

```
qt(1-0.05/2, 8)
```

```
## [1] 2.306004
```

Comment, H0 will be rejected at a very high level of statistical significance. Rejecting H0 indicates the the data support Ha and thus B1 is not zero. Accept that there is a linear relationship between the variables of interest.

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

Comment: interval for the unknown population slope can be calculated using statistics +/-(multiplier * s.e. of statistics). The lower bound is given by:

```
4 - qt(1-0.05/2, 8) * 0.469
```

```
## [1] 2.918484
```

And the upper bound is given by:

```
4 + qt(1-0.05/2, 8) * 0.469
```

```
## [1] 5.081516
```

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).

Comment: no transfers suggests x_i=0, so one needs to focus on the intercept, is it different from 9? Therefore, H0: B0 = 9 the mean number of broken ampules is not different from 9, and Ha: B0 != 9, the mean number is actually different. To test this, t_stat = (B0_hat - B0)/ se_B0_hat

```
(10.2-9) / 0.6633
```

```
## [1] 1.809136
```

Compared to critical t:

```
qt(1-0.05/2, 8)
```

```
## [1] 2.306004
```

Conclusion: H0 can't be rejected and therefore the mean number of broken ampulse is not different from 9.

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.

Confidence interval

```
x_i = 2
x_bar = 1
S_xx = 10
mu_hat = 10.2 + 4*x_i
alpha = 0.05
n = 10
t = qt(alpha/2, n-2)
se = sqrt(2.2*(1/n+(x_i-x_bar)**2/S_xx))
upper = mu_hat + t * se
lower = mu_hat - t * se

cat("Confidence interval for the mean when transfers equal to 2: [", upper, ",", lower, "]")
```

```
## Confidence interval for the mean when transfers equal to 2: [ 16.67037 , 19.72963 ]
```

Prediction interval

```
x_i = 2
x_bar = 1
S_xx = 10
mu_hat = 10.2 + 4*x_i
alpha = 0.05
n = 10
t = qt(alpha/2, n-2)
se = sqrt(2.2*(1 + 1/n+(x_i-x_bar)**2/S_xx))
upper = mu_hat + t * se
lower = mu_hat - t * se

cat("Prediction interval for transfers equal to 2: [", upper, ",", lower, "]")
```

```
## Prediction interval for transfers equal to 2: [ 14.45319 , 21.94681 ]
```

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

This relates to how the width of the confidence interval changes as the value of the predictor variable moves farther away from the sample mean of the predictor? The interval will become smaller.

f. What is the value of the F statistic for the ANOVA table?

This can be calculated from Mean Squares:

```
160/2.2
```

```
## [1] 72.72727
```

Alternatively, B1 t statistic squared should be equal to F statistics:

```
t = 4/0.469
t**2
```

```
## [1] 72.74017
```

g. Calculate the value of R2, and interpret this value in context.

```
160/(160+17.6)
```

```
## [1] 0.9009009
```

Comment: R2 is close to one. It implies that the regression is probably good hence more than 90% of variance in y was explained by the regression.

---

3. Population slope problem.

a. Describe how the straight line would look in a plot of y versus x.

I think it will look like a straight line parallel to x.

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.

Generally, we are interested in the slope because it shows how Y changes with X. Slope of 0 means that there is nothing in X that can help explain Y, or in other words there is no relationship between between X and Y. Nonzero slope will allow X to contribute some explanation into Y, and if this contribution or slope is significantly different from a random chance, X and Y may be deemed linearly related.