HW2 GR5205 - Simple Linear Regression Model

Mathieu Sauterey - UNI: mjs2364

1 octobre 2017

Problem 1 (Question 1.20 in Chapter 1)

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 below have been collected from 45 recent calls on users to perform routine preventive maintenance service; for each call, X is the number of copiers serviced and Y is the total number of minutes spent by the service person. Assume that first-order regression model (1.1) is appropriate.

i : 1 2 3 . . . 43 44 45 X_i : 2 4 3 . . . 2 4 5 Y_i : 20 60 46 . . . 27 61 77

(a)(10p) Obtain the estimated regression function.

First, we read the data using data.table function and store it in a data frame named *data*. We then rename the columns of the data frame accordingly.

```
data <- read.table("Homework_2_data.txt", header = FALSE, as.is =TRUE)
names(data) <- c("Minutes_spent","Number_of_copiers")
```

Considering a simple linear regression model of the form

$$\hat{Y_i} = \beta_0 + \beta_1 * X_i$$

, we compute the estimated regression coefficients β_0 and β_1 using the function lm.

```
lm1 <- lm(data$Minutes_spent ~ data$Number_of_copiers) #computes the estimated regression model
beta_0 <- lm1$coefficients[1]
beta_1 <- lm1$coefficients[2]
names(lm1$coefficients) <- c("Intercept","Slope")
lm1$coefficients[1] #grabs the intercept b0 of the estimated regression line
```

```
##  Intercept
## -0.5801567
```

```
lm1$coefficients[2] #grabs the slope b1 of the estimated regression line
```

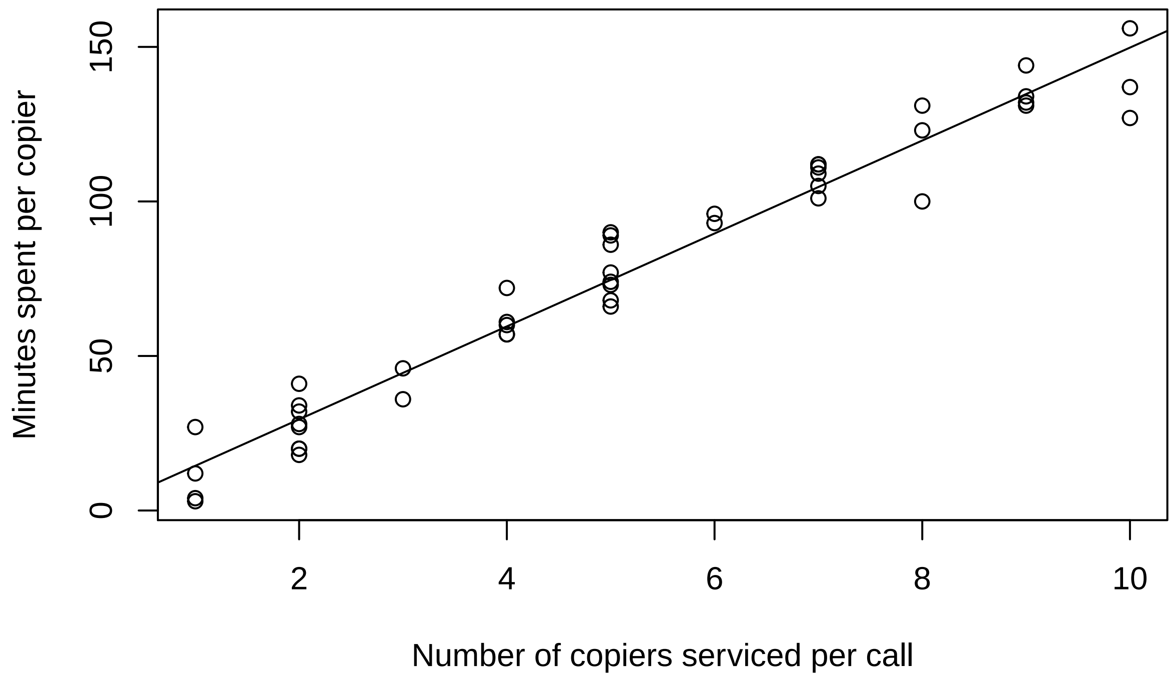
```
##    Slope
## 15.03525
```

We finally obtain the estimated regression function

$$\hat{Y_i} = -0.5801567 + 15.03525 * X_i$$

(b)(10p) Plot the estimated regression function and the data. How well does the estimated regression function fit the data?

```
plot(data$Number_of_copiers,data$Minutes_spent,xlab="Number of copiers serviced per call",
     ylab="Minutes spent per copier") #plots the data
abline(lm1$coefficients[1],lm1$coefficients[2]) #plots the regression line
```



We plot the data first and then plot the estimated regression function using abline function. The estimated regression function fits the data well.

(c)(10p) Interpret b0 in your estimated regression function. Does b0 provide any relevant information here? Explain.

b_0 is the intercept of the estimated regression line. It provides the fitted value of the model when $X_i = 0$, that it is to say, when the number of copiers serviced per call is equal to 0. In this situation however, such information is irrelevant because if 0 copiers are serviced then no calls are made and the time spent is exactly equal to 0. Furthermore, if $X_i = 0$ then $\hat{Y}_i = -0.5801567$ which makes no sense as time spent cannot be negative.

(d)(10p) Obtain a point estimate of the mean service time when X = 5 copiers are serviced.

Using the formula

$$\hat{Y}_i = -0.5801567 + 15.03525 * X_i$$

we can compute the point estimate of the mean service time when $X = 5$ are serviced.

```
Y_5 <- lm1$coefficients[1]+lm1$coefficients[2]*5
names(Y_5)="Estimated Y for X=5"
Y_5
```

```
## Estimated Y for X=5
##            74.59608
```

When $X = 5$ we estimate that the mean service time is $\hat{Y}_i = 74.6$.

Problem 2 (Question 2.5 in Chapter 2)

(a)(15p) Estimate the change in the mean service time when the number of copiers serviced increases by one. Use a 90 percent confidence interval. Interpret your confidence interval.

The confidence interval of the estimated slop b_1 is given as:

$$b_1 = \pm t(1 - \alpha/2; n - 2) * s\{b_1\}$$

In the above, $n = 45$ is the number of degrees of freedom, $\alpha = 0.1$ for a 90 percent confidence interval, and t() returns the percentile from the student t distribution.

We know that the standard deviation of b_1 is given by

$$s\{b_1\} = \sqrt{s^2\{b_1\}}$$

with

$$s^2\{b_1\} = \frac{MSE}{\sum_{i=1}^{45}((X_i - \bar{X})^2)} = \frac{\sum_{i=1}^{45}((Y_i - \hat{Y}_i)^2)}{(n-2)\sum_{i=1}^{45}((X_i - \bar{X})^2)}$$

```
MSE <- sum(lm1$residuals^2)/lm1$df.residual #computes the mean square error of the fitted model
mean_copiers = apply(data, 2, mean)[2] #computes the mean number of copiers serviced per call
Var_b1 <- MSE / sum((data$Number_of_copiers - mean_copiers)^2) #computes the variance of b1
Std_b1 <- sqrt(Var_b1) #computes the standard deviation of b1
t_score <- qt(c(0.95),43) # computes the 90% confidence student t distribution percentile
lower_b1 <- lm1$coefficients[2] - t_score*Std_b1 #calculates lower bound of b1
upper_b1 <- lm1$coefficients[2] + t_score*Std_b1 #calculates upper bound of b1
names(lower_b1)="Lower b1 bound"
names(upper_b1)="Upper b1 bound"
lower_b1
```

```
## Lower b1 bound
##       14.22314
```

```
upper_b1
```

```
## Upper b1 bound
##       15.84735
```

Therefore

$$14.22314 \leq b_1 \leq 15.84735$$

which means that with confidence 90% we estimate that the mean number of minutes spent increases by somewhere between 14.22 and 15.85 minutes for each additional copiers serviced per call.

(b)(15p) Conduct a t-test to determine whether or not there is a linear association between X and Y here; control the risk at 0.10. State the alternatives, decision rule, and conclusion. What is the P-value of your test?

We will conduct a two-sided test with $\alpha = 0.1$. We set the null hypothesis $H_0 : \beta_1 = 0$ and the alternative hypothesis $H_a : \beta_1 \neq 0$ First let's recall that the explicit test of the alternatives H_a is based on the test statistic $t^* = \frac{b_1}{(s\{b_1\})}$.

The decision rule with this test statistic for controlling the level of significance at α is:

If

$$|t^*| \leq t(1 - \alpha/2; n - 2) = 1.68107$$

then conclude H_0

If

$$|t^*| > t(1 - \alpha/2; n - 2) = 1.68107$$

then conclude H_a

```
t <- lm1$coefficients[2]/Std_b1 #computes the test statistic t* for beta_1=0
names(t)="Two-sided test statistic"
t
```

```
## Two-sided test statistic
##                 31.12326
```

t* is equal to 31.12326 which is larger than 1.68107 so we reject the null hyothesis $H_0 : \beta_1 = 0$ and conclude $H_a : \beta_1 \neq 0$.

The P-value of this test is:

```
p_value1 <- summary(lm1)$coefficients[2,4] #displays the P-value of beta_1 against null hypothesis H0=0
names(p_value1) <- "P-value of two-sided test"
p_value1
```

```
## P-value of two-sided test
##              4.009032e-31
```

Since the P-value is close to 0 which is less than the specified level of significance $\alpha = 0.1$ then it does make sense that we reject H_0 and conclude H_a.

(c)(15p) Are your results in parts (a) and (b) consistent? Explain.

Our results in (a) and (b) are consistent as (b) concludes H_a and rejects H_0 and we could have reached this result at once using the 90% confidence interval of b_1 computed in (a) because this interval does not include 0 (and is nowhere close to 0 neither).

(d)(15p) The manufacturer has suggested that the mean required time should not increase by more than 14 minutes for each additional copier that is serviced on a service call. Conduct a test to decide whether this standard is being satisfied by Tri-City. Control the risk of a Type I error at 0.05. State the alternatives, decision rule, and conclusion. What is the P-value of the test?

We will conduct a one-sided test with $\alpha = 0.05$. We set the null hypothesis $H_0 : \beta_1 \leq 14$ and the alternative hypothesis $H_a : \beta_1 > 14$ First let's recall that the explicit test of the alternatives H_a is based on the test statistic $t^* = \frac{b_1 - 14}{(s\{b_1\})}$.

The decision rule with this test statistic for controlling the level of significance at α is:

If

$$t^* \leq t(1-\alpha; n-2) = 1.68107$$

then conclude H_0

If

$$t^* > t(1-\alpha; n-2) = 1.68107$$

then conclude H_a

```
t2 <- (lm1$coefficients[2]-14)/Std_b1 #computes the test statistic t* for beta_1=0
names(t2) <- "Two-sided test statistic"
t2
```

```
## Two-sided test statistic
##                 2.142984
```

t* is equal to 2.142984 which is larger than 1.68107 so we reject the null hyothesis $H_0 : \beta_1 \leq 14$ and conclude $H_a : \beta_1 > 14$.

The P-value of this one-sided test is:

```
p_value2 <- 1-pt(t2,43)
names(p_value2) <- "P-value of one-sided test"
p_value2
```

```
## P-value of one-sided test
##                0.01890766
```

Since the P-value is 0.0189 which is less than the specified level of significance at $\alpha = 0.05$ then it does make sense that we reject H_0 and conclude H_a.