Chapter 8: Introduction to linear regression Homework 8

Alexis Mekueko

10/24/2020

Github link: https://github.com/asmozo24/DATA606_Chapter8_Homework8

Web linl: https://rpubs.com/amekueko/684046

R Packages

```
library(tidyverse) #loading all library needed for this assignment
library(openintro)
#head(fastfood)
#library(readxl)
library(data.table)
#library(DT)
library(knitr)

#library(readr)
#library(plyr)
#library(dplyr)
library(stringr)
#library(XML)
#library(RCurl)
#library(jsonlite)
#library(httr)

#library(maps)
#library(dice)
# #library(VennDiagram)
# #library(help = "dice")
#ibrary(DBI)
#library(dbplyr)

# library(rstudioapi)
# library(RJDBC)
# library(odbc)
# library(RSQLite)
# #library(rvest)

#library(readtext)
#library(ggpubr)
#library(fitdistrplus)
```

```
library(plyr)
library(pdftools)
library(plotrix)
library(gplots)
library(tibble)
#library(moments)
#library(qualityTools)
#library(normalp)
#library(utils)
#library(MASS)
#library(qqplotr)
#library(stats)
library(statsr)
```

```
## Warning: package 'statsr' was built under R version 4.0.3
```

```
#(DATA606)
```

Github Link: https://github.com/asmozo24/DATA606_Chapter8_Homework8

Web link: https://rpubs.com/amekueko/681421

Chapter 8 - Introduction to Linear Regression

Nutrition at Starbucks, Part I.(8.22, p. 326) The scatterplot below shows the relationship between the number of calories and amount of carbohydrates (in grams) Starbucks food menu items contain. SinceStarbucks only lists the number of calories on the display items, we are interested in predicting the amountof carbs a menu item has based on its calorie content.

a)Describe the relationship between number of calories and amount of carbohydrates (in grams) thatStarbucks food menu items contain.

Answer: there is likely a linear relationship between the number of calories and amount of carbohydrates (in grams) Starbucks food menu items contain.the fitted line showed y = ax + b

(b) In this scenario, what are the explanatory and response variables?

Answer: response variables = Carbs or amount of carbohydrates (in grams), explanatory = Calories or number of calories

(c) Why might we want to fit a regression line to these data?

Answer: this linear relationship carries some error and we might want to fit a regression line to measure the corelation between the two variables. In order words, we to predict the values of the dependent/response variable (Carbs or amount of carbohydrates (in grams)) using the value from explanatory/independant variable. Can Calories being a predictor for Carbs?

(d) Do these data meet the conditions required for fitting a least squares line?

Answer: we can say yes, they meet the conditions for fitting a least squares line. we see a normal probability plot of the residuals on the histogram plot, next , the residual plot looks scattered, there is no particular pattern, we could appreciate more if there were some vertical lines to show constant points on each side of line the residuals = 0.

Body measurements, Part I.(8.13, p. 316) Researchers studying anthropometry collected body girth-measurements and skeletal diameter measurements, as well as age, weight, height and gender for 507 physicallyactive individuals.19 The scatterplot below shows the relationship between height and shoulder girth (overdeltoid muscles), both measured in centimeters.

(a) Describe the relationship between shoulder girth and height.

Answer: we see a positive (uphill) linear relationship between shoulder girth and heigt. (b)How would the relationship change if shoulder girth was measured in inches while the units of heightremained in centimeters? Answer: we don't think the relationship would change. The change from cm to inches would change the scale on x-axis because this conversion applies to all the shoulder girth values.

Body measurements, Part III.(8.24, p. 326) Exercise above introduces data on shoulder girth and heightof a group of individuals. The mean shoulder girth is 107.20 cm with a standard deviation of 10.37 cm. Themean height is 171.14 cm with a standard deviation of 9.41 cm. The correlation between height and shouldergirth is 0.67.

(a) Write the equation of the regression line for predicting height.

Answer: y = ax + b ; y = height , a = c_correlation* (height_standardD/shoulder_standardD), x = shoulder girth, b = mean_height - a*mean_shoulder_girth. y = 0.6079749x + 105.9651

(b) Interpret the slope and the intercept in this context. Answer: for each centimeter increase in shoulder girth, there is an increase of 0.6079749 in height.

(c) Calculate R-square of the regression line for predicting height from shoulder girth, and interpret it in the context of the application. Answer: R_squared = 0.4489, meaning there is about 44.9% of the predictor that explain the variation in the response variable around the mean.

(d)A randomly selected student from your class has a shoulder girth of 100 cm. Predict the height of this student using the model. Answer: y = 0.6079749*100 + 105.9651 , height = 166.76

(e)The student from part (d) is 160 cm tall. Calculate the residual, and explain what this residual means. Residual = Observed value - Fitted value. Linear Residual = 160 - 166.76 Residual = 6.76 , meaning we overestimated the height of the student,

(f) A one year old has a shoulder girth of 56 cm. Would it be appropriate to use this linear model to predict the height of this child? Answer: in part I, shoulder girth starts at 80 cm to 140cm , so the one year old height does not fit in the data. In this case , we would not be able to find out if we overestimate or understimate the kid height based on the linear model.

```
# assigning value for the equation of the regression line for predicting height
mean_height = 171.14
mean_shoulder_girth = 107.2
height_standardD = 9.41
```

```r
shoulder_standardD = 10.37
c_correlation = 0.67

# slope equaiton
a = c_correlation* (height_standardD/shoulder_standardD)
a
```

```
## [1] 0.6079749
```

```r
# intercept equation
b = mean_height - a*mean_shoulder_girth
b
```

```
## [1] 105.9651
```

```r
R_squared = c_correlation^2
R_squared
```

```
## [1] 0.4489
```

```r
 y = 0.6079749*100 + 105.9651
 y
```

```
## [1] 166.7626
```

```r
0.357/4.034
```

```
## [1] 0.08849777
```

Cats, Part I.(8.26, p. 327) The following regression output is for predicting the heart weight (in g) of catsfrom their body weight (in kg). The coefficients are estimated using a dataset of 144 domestic cats.

Cats, Part I.(8.26, p. 327) The following regression output is for predicting the heart weight (in g) of catsfrom their body weight (in kg). The coefficients are estimated using a dataset of 144 domestic cats. x = independent variable = body weigt (kg), y = heart weight(g)

	Estimate	Std. Error	t value	Pr(>|t|)

(Intercept) -0.357 0.692 -0.515 0.607 body wt 4.034 0.250 16.119 0.000 s= 1.452 R_squared = 64.66% R_squared adjusted = 64.41%

(a) Write out the linear model.

Answer: y = ax + b , a = 4.034, y = 4.034x -0.357

(b) Interpret the intercept. Answer: none of the 144 cats has 0.0 kg weight which is more realistic. the negative intercept just mean that if there were one cat with that wasn't born, its heart will still in inception. or in other word if there was one cat with nearly 0.0 g heart weight, his body weight would have been as small as 88.5 g which is about a size of one duck egg.

(c) Interpret the slope. Answer: for every 1.0kg of body weight a cat grows, its heart weight will increase by 4.034g

(d) Interpret R-squared. Answer: R_squared = 64.66%, meaning there is about 64.66% of the cat body weight that explain the variation in its heart weight around the mean.

(e) Calculate the correlation coefficient. c_correlation = 0.8041144

```
R_squared = 0.6466
c_correlation = sqrt (R_squared)
c_correlation
```

```
## [1] 0.8041144
```

```
# # replicate the data for linear regression plot
# x1 = 5
# variables = 30
# cat <- matrix(ncol = variables, nrow = x1)
#
# for ( i in -1:x1)
# {
#   y1 = 4.034*i -0.357
#
#
# }
#
# cats <- data.frame(cat)
# ggplot(cats, aes(x=x1, y=y1)) + geom_point() +
#   geom_abline()
#
# #lm function in R to fit the linear model (a.k.a. regression line).
# lm1 <- lm(heart ~ body)
# #statistical summary
# summary(lm1)
# # gives the model coefficient, intercept and slope
# coef(lm1)
```

Rate my professor.(8.44, p. 340) Many college courses conclude by giving students the opportunity toevaluate the course and the instructor anonymously. However, the use of these student evaluations as anindicator of course quality and teaching effectiveness is often criticized because these measures may reflect theinfluence of non-teaching related characteristics, such as the physical appearance of the instructor. Researchersat University of Texas, Austin collected data on teaching evaluation score (higher score means better) andstandardized beauty score (a score of 0 means average, negative score means below average, and a positivescore means above average) for a sample of 463 professors. The scatterplot below shows the relationshipbetween these variables, and also provided is a regression output for predicting teaching evaluation scorefrom beauty score.

	Estimate	Std. Error	t value	Pr(>|t|)

(Intercept) 4.010 0.0255 157.21 0.0000 beauty 0.0322 4.13 0.000

(a)Given that the average standardized beauty score is -0.0883 and average teaching evaluation scoreis 3.9983, calculate the slope. Alternatively, the slope may be computed using just the information provided in the model summary table.

Answer: slope = 0.1325028 (b)Do these data provide convincing evidence that the slope of the relationship between teaching evaluation and beauty is positive? Explain your reasoning. Answer: yes, we can observe that the teaching evaluation increases as beauty increases.

(c)List the conditions required for linear regression and check if each one is satisfied for this model basedon the following diagnostic plots Answer: Linearity:there is no any apparent pattern in the residuals plot, we can say there is likely a linear relationship between beaty and teaching evaluation Nearly normal residuals: based on normal probability plot and the histogram show a nearly normal residuals distribution Constant variability: based on the residuals vs. fitted plot, we can see that the points are about constant on each side of the line residuals = 0, thus the constant variability condition is met.

```
slope = (3.9983 - 4.010)/(-0.0883)
slope
```

```
## [1] 0.1325028
```