Assignment1

2023-03-16

## QUESTION 1

### (a)

knitr::include\_graphics(“Q1a.png”)

### (b)

### (c)

x <- c(3, 4, 5)  
x <- matrix(x, ncol = 1)  
y <- c(4, 9, 14)  
ones <- matrix(rep(1, length(x)), ncol = 1)  
x <- cbind(ones, x)  
  
xtx <- t(x) %\*% x  
xtx\_inv <- solve(xtx)  
  
Beta <- xtx\_inv %\*% t(x) %\*% y  
Beta

## [,1]  
## [1,] -11  
## [2,] 5

error <- y - x %\*% Beta  
error

## [,1]  
## [1,] -1.598721e-14  
## [2,] -1.598721e-14  
## [3,] -1.598721e-14

### (d)

#create a variable of X with values  
X <- c(3, 4, 5)  
#create a variable of Y with values  
Y <- c(4, 9, 14)  
linearmodel <- lm(Y ~ X)  
print(linearmodel)

##   
## Call:  
## lm(formula = Y ~ X)  
##   
## Coefficients:  
## (Intercept) X   
## -11 5

summary(linearmodel)

## Warning in summary.lm(linearmodel): essentially perfect fit: summary may be  
## unreliable

##   
## Call:  
## lm(formula = Y ~ X)  
##   
## Residuals:  
## 1 2 3   
## -3.82e-16 7.64e-16 -3.82e-16   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -1.100e+01 2.701e-15 -4.073e+15 <2e-16 \*\*\*  
## X 5.000e+00 6.616e-16 7.557e+15 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 9.357e-16 on 1 degrees of freedom  
## Multiple R-squared: 1, Adjusted R-squared: 1   
## F-statistic: 5.711e+31 on 1 and 1 DF, p-value: < 2.2e-16

## QUESTION 2

### (a)

If the explanatory variable x takes the values {2, 2, 2}, the computation of coefficients βˆ may lead to errors. The below methods are used:  
• If coefficients βˆ are computed by hand, the matrix (X ′X) is not invertible. This means that the coefficients cannot be determined using this method. • If coefficients βˆ are estimated using the function lm in R, the computation will result in an error message saying “Lapack routine dgesv: system is exactly singular: U[2,2] = 0”. This error occurs again due to the mean of x. In either way, it is impossible to calculate the coefficients βˆ for the given x and y.

### (b)

When X takes the values {2, 2, 2} and Y takes the values {4, 9, 14}, the estimation of the OLS estimator fails due to: • Statistical reason: the calculation of (X’X)-1 is required. 1. X has three identical elements, and its mean is 2, which is equal to each element in X. 2. X’X is [3 6 6 12]. a. The rank of X’X is 1. The second row is a scalar multiple of the first row, indicating that X’X matrix has only one linearly independent row. b. X’X has three identical rows, so the determinant of X’X is 0, indicating that it is a singular matrix. 3. Because X ′X is singular, (X ′X) −1 cannot be computed.

• Geometric: slope is undefined, intercept is 9 From the resulting plot of three observations on the graph, it is apparent that the values of X have no variation in the x-axis. It implies that there is no unique line that can fit data points. Therefore, no coefficients can be estimated to represent the relationship between X and Y.

## QUESTION 3

### (a)

# get working directory  
getwd()

## [1] "/Users/gezhongya/Desktop/Desktop - Ge’s MacBook Pro (2)/STAT448\_Big\_Data/Assignment"

#model creation  
library(readr)  
happy <- read\_csv("happy.csv")

## New names:  
## Rows: 498 Columns: 3  
## ── Column specification  
## ──────────────────────────────────────────────────────── Delimiter: "," dbl  
## (3): ...1, income, happiness  
## ℹ Use `spec()` to retrieve the full column specification for this data. ℹ  
## Specify the column types or set `show\_col\_types = FALSE` to quiet this message.  
## • `` -> `...1`

# show the first few rows  
head(happy)

## # A tibble: 6 × 3  
## ...1 income happiness  
## <dbl> <dbl> <dbl>  
## 1 1 3.86 2.31  
## 2 2 4.98 3.43  
## 3 3 4.92 4.60  
## 4 4 3.21 2.79  
## 5 5 7.20 5.60  
## 6 6 3.73 2.46

names(happy)

## [1] "...1" "income" "happiness"

#fit a linear model, predicet happiness from income  
happy.lm <- lm(happiness ~ income, data = happy)  
summary(happy.lm)

##   
## Call:  
## lm(formula = happiness ~ income, data = happy)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.02479 -0.48526 0.04078 0.45898 2.37805   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.20427 0.08884 2.299 0.0219 \*   
## income 0.71383 0.01854 38.505 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.7181 on 496 degrees of freedom  
## Multiple R-squared: 0.7493, Adjusted R-squared: 0.7488   
## F-statistic: 1483 on 1 and 496 DF, p-value: < 2.2e-16

names(happy.lm)

## [1] "coefficients" "residuals" "effects" "rank"   
## [5] "fitted.values" "assign" "qr" "df.residual"   
## [9] "xlevels" "call" "terms" "model"

# Coefficient statistics in the model  
coefficient = coef(happy.lm)  
coefficient

## (Intercept) income   
## 0.2042704 0.7138255

equation <- paste0("y = ", round(coefficient[2],3), "\*x +", round(coefficient[1],3))  
paste("The regression equation is ",equation)

## [1] "The regression equation is y = 0.714\*x +0.204"

### (b)

### (c)

### (d)

### (e)

library(ggplot2)  
attach(happy.lm)  
  
# Add the fitted values as a new column in the dataframe  
head(happy)

## # A tibble: 6 × 3  
## ...1 income happiness  
## <dbl> <dbl> <dbl>  
## 1 1 3.86 2.31  
## 2 2 4.98 3.43  
## 3 3 4.92 4.60  
## 4 4 3.21 2.79  
## 5 5 7.20 5.60  
## 6 6 3.73 2.46

p1 <- ggplot(happy.lm) +  
 geom\_point(aes(x = fitted.values, y = residuals, colour = "red")) +  
 labs(  
 title = "Regression fits of Happiness",  
 subtitle = "(1973-74)",  
 caption = "Data from the 1974 Motor Trend US magazine.",  
 tag = "Figure 1",  
 x = "Happiness",  
 y = "Model Fitted Values",  
 colour = "Gears"  
 )  
plot(p1)



#p1 + theme\_gray() # the default

### (f)

### (g)

You can also embed plots, for example:



Note that the echo = FALSE parameter was added to the code chunk to prevent printing of the R code that generated the plot.