SMI Assignment 4

Andrew Martin

October 6, 2017

1. Writing design matrices
   Given the multiple regression model

$$M : Y = X\beta + \epsilon$$

   Where $Y_i \sim N(\eta_i, \sigma^2)$ indep for $i = 1, 2, \ldots, n$ and $\eta = X\beta$.

   (a) Write dimensions of Y, X, β and ϵ

Solution

Y will have dimension 35x1,
X will have dimension 35x4,
β will have dimension 4x1,
and,
ϵ will have dimension 35x1.

...

   (b) Write down β in full, and the first four rows of X and y
   Solution

$$\beta \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix}$$

$$y = \begin{bmatrix} 12.8 \\ 9.4 \\ 14 \\ 15.6 \\ \vdots \end{bmatrix}$$

$$X = \begin{bmatrix} 1 & 1 & 1590 & 15 \\ 1 & 0 & 968 & 11 \\ 1 & 2 & 732 & 12 \\ 1 & 3 & 780 & 13 \\ \vdots & \vdots & \vdots & \vdots \end{bmatrix}$$

...

2. Linear Transformations of the design matrix

   (a) Show that the columns of X^* are also linearly independent. (prove by contradiction)

   Solution

   Assume columns of X^* are not linearly independent, i.e. $BX^* = 0$ has non-trivial solutions.

$$
\begin{aligned}
BX^* &= 0 \text{ for non-zero some matrix of combinations B} \\
X^* &= B^{-1}0 \\
X^* &= 0 \\
AX &= 0 \\
X &= A^{-1}0 \\
X &= 0 \\
\implies &\text{X is zero - contradiction as X is non-zero.}
\end{aligned}
$$

...

   (b) Show that $X^*((X^*)^T X^*)^{-1}(X^*)^T = X(X^T X)^{-1}X^T$

   Solution

$$
\begin{aligned}
X^*((X^*)^T X^*)^{-1}(X^*)^T &= XA((XA)^T XA)^{-1}(XA)^T \\
&= XA(A^T X^T XA)^{-1}A^T X^T \\
&= XAA^{-1}X^{-1}X^{T^{-1}}A^{T^{-1}}A^T X^T \\
&= XIX^{-1}X^{T^{-1}}IX^T \\
&= XX^{-1}X^{T^{-1}}X^T \\
&= X(X^{-1}X^{T^{-1}})X^T \\
&= X(X^T X)^{-1}X^T
\end{aligned}
$$

...

   (c) Consider two alternative models

$$M : Y = X\beta + \epsilon \text{ and } M^* : Y = X^*\beta^* + \epsilon$$

   Show that $\hat{\eta^*} = \hat{\eta}$ i.e. the vector of fitted values is the same.

Solution

$$\begin{aligned}\hat{\eta}^* &= X^*\hat{\beta}^* \\ &= X^*\left((X^{*^T}X^*)^{-1}(X^*)^T y\right) \\ &= X(X^TX)^{-1}X^Ty \text{ using part b.} \\ &= X\hat{\beta} \\ &= \hat{\eta}\end{aligned}$$

...

3. Matrix calculations in R

   (a) Write down the design matrix, X, and the vector of observed values y, and enter them into R.

   Solution

   The R code outputted:

```
> X
     X0 X1 X2 X3
[1,]  1 −3  5 −1
[2,]  1 −2  0  1
[3,]  1 −1 −3 1
[4,]  1  0 −4  0
[5,]  1  1 −3 −1
[6,]  1  2  0 −1
[7,]  1  3  5  1
> Y
[1] 1 0 0 1 2 3 3
```

...

   (b) Use direct matrix calculations in R to find the LSE given by

$$\hat{\beta} = (X^TX)^{-1}X^Ty$$

   Solution

```
> betahat
           [,1]
X0  1.4285714
X1  0.5000000
X2  0.1190476
X3 −0.5000000
```

...

(c) Continuing to use R for your calculations, find the predicted value of Y when $x_1 = 1, \quad x_2 = -3, \quad x_3 = -1$

Solution

```
> Ypredict
          [,1]
[1,] 2.071429
```

...

(d) Test the null hypothesis that X_3 has no effect on Y, i.e. test $H_0 : \beta_3 = 0$ as follows:

i. The test statistic takes the form:

$$T = \frac{\lambda^T \hat{\beta} - 0}{s_e \sqrt{\lambda^T (X^T X)^{-1} \lambda}} \text{ where T}\sim t_{n-p} \text{ if } H_0 \text{ is true.}$$

In this case, write down λ, n, and p

Solution

$$\lambda = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}$$

$$n = 7$$

$$p = 4$$

...

ii. Calculate the observed value of the test statistic for this sample.
Recall $s_e^2 = \frac{1}{n-p}||y - X\hat{\beta}||^2 = \frac{1}{n-p}(y - X\hat{\beta})^T(y - X\hat{\beta})$.

Solution

```
> teststat
          [,1]
[1,] −13.74773
```

...

iii. Calculate the P-value, and hence state whether you reject or retain H_0 at significance level $\alpha = 0.05$

Solution

```
> pval
             [,1]
[1,] 8.973452e-05
```

Since the p-value is significantly lower than α **...**

iv. Find a 95% confidence interval for the expected value of Y given $x_1 = 1, \quad x_2 = -3, \quad x_3 = -1$ i.e. a 95% confidence interval for

$$\lambda^T \beta = \beta_0 + \beta_1 - 3\beta_2 - \beta_3$$

Where $\lambda^T = (1, 1, -3, -1)$.

Solution

```
> CI
[1] 1.880739 2.262119
```

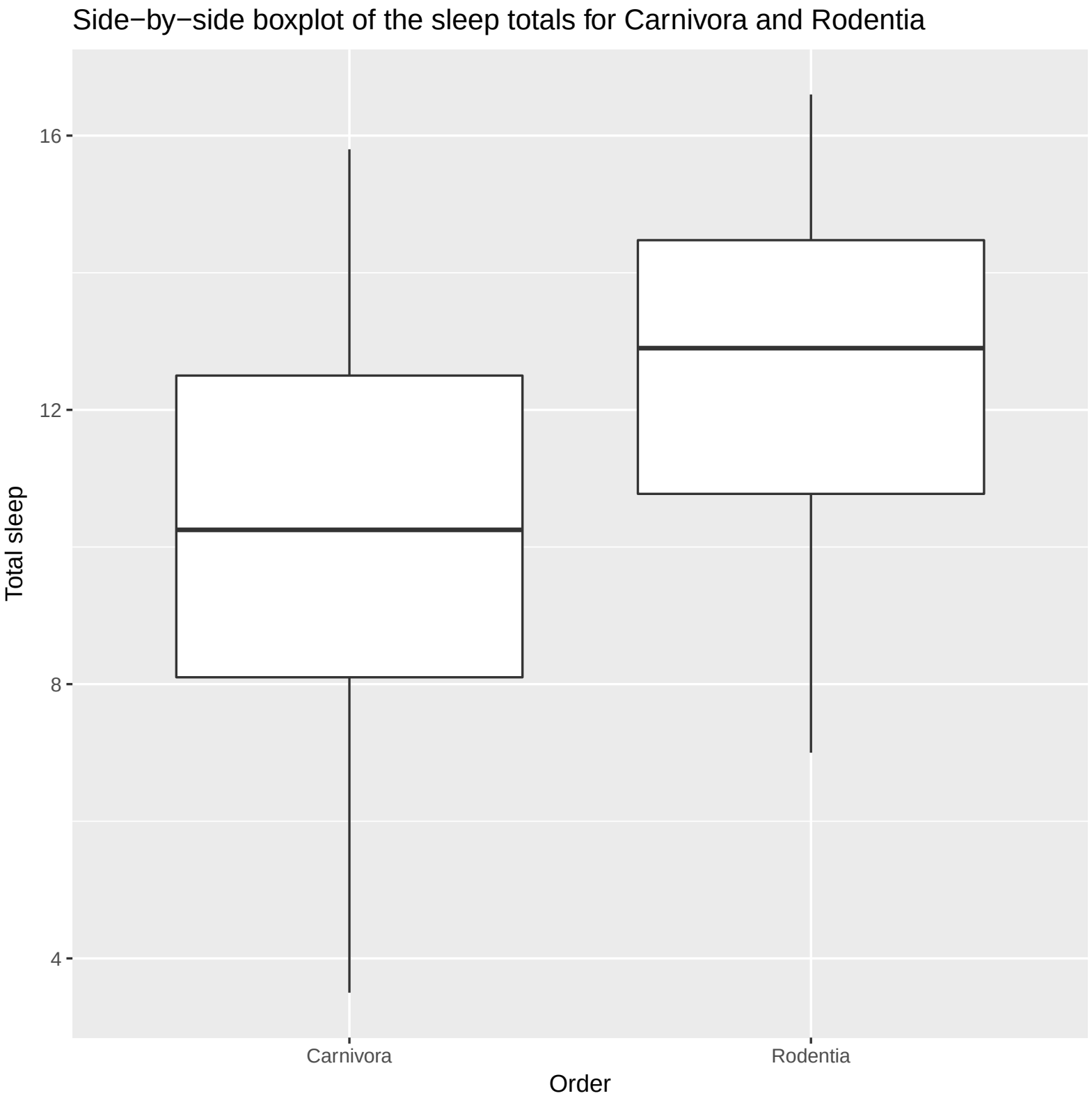
...

4. Rats versus cats question

   (a) Read in the data

   (b) Produce and include side-by-side boxplots of the sleep total for Carnivora and Rodentia. Describe the distributions

   Solution


Side−by−side boxplot of the sleep totals for Carnivora and Rodentia
16
12
8
4
Total sleep
Carnivora
Rodentia
Order

...

(c) Decide if a pooled two-sample t-test can be used. Give reasons

Solution

A pooled two-sample t-test could be used as the data appears to have similar, or the same variance. **...**

(d) Perform a two-sample t-test. For full marks include:

i. null and alternative hypotheses
ii. value of test statistic
iii. the distribution of the test statistic if the null hypothesis is true
iv. P-value
v. and conclusion

Solution

$$\begin{bmatrix} H_0 : Carnivora_\mu = Rodentia_\mu \\ H_a : Carnivora_\mu \neq Rodentia_\mu \end{bmatrix}$$

Where μ represents the population mean sleep total of that group.

If the null is true, the test-statistic will be t-distributed.

From the R code, $t = -2.0004$

The test statistic is t-distributed if the null hypothesis is true.

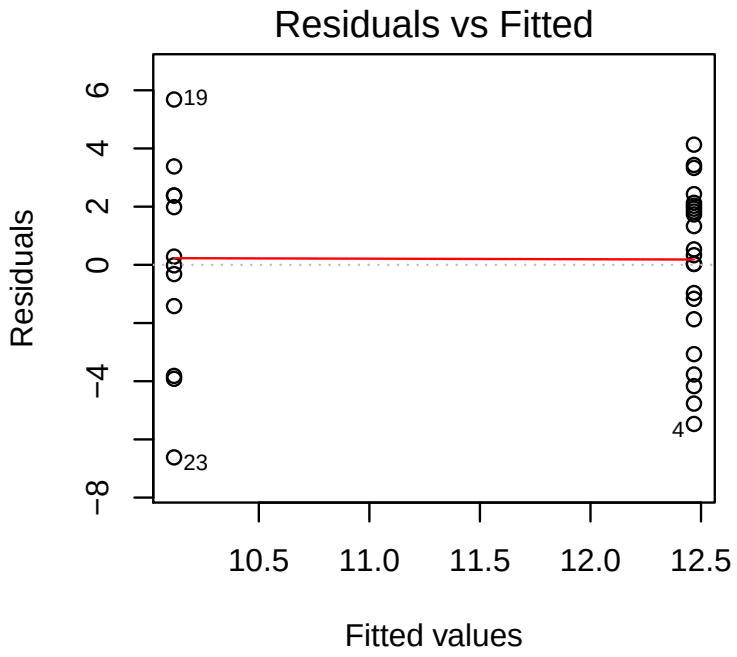
The P-value is $Pvalue = 0.06005 > 0.05$

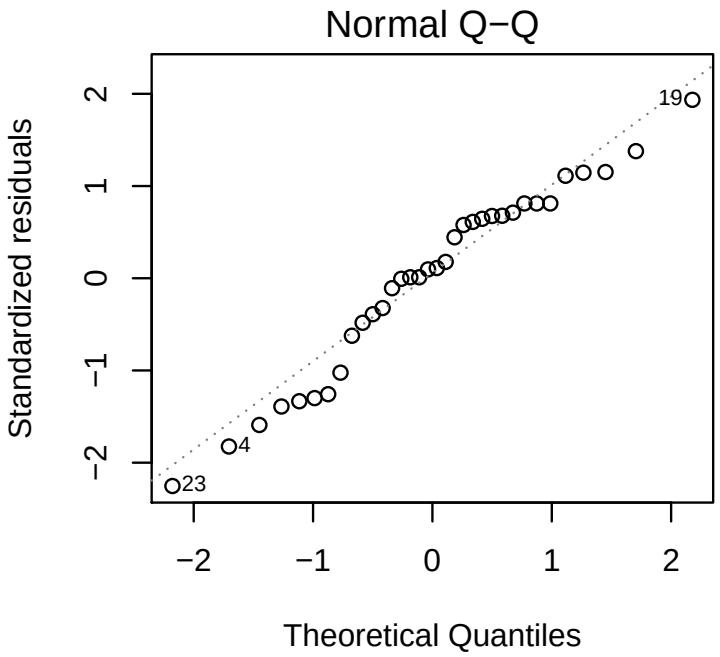
As the P-value was above the significance used $\alpha = 0.05$ the null-hypothesis is accepted, i.e. there is evidence to suggest the means of the two groups are the same. **...**

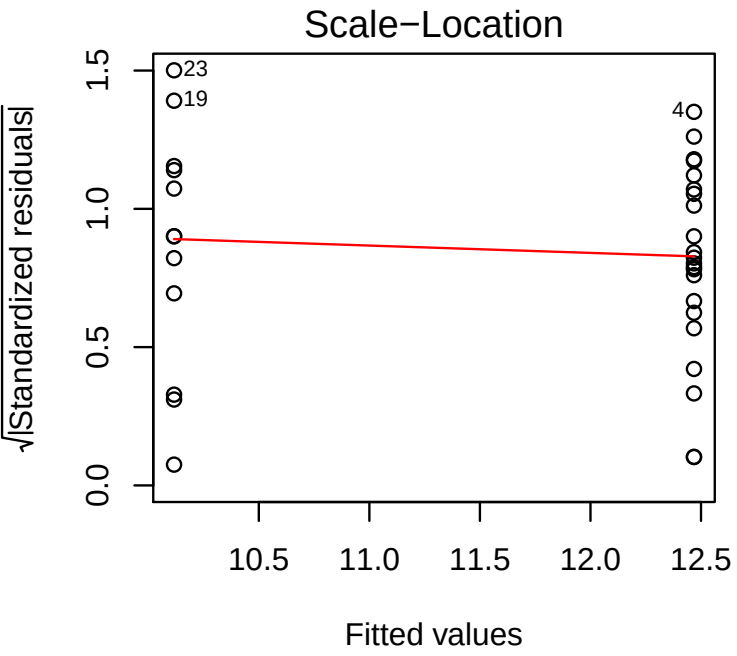
(e) Check the assumptions including appropriate plots if necessary.

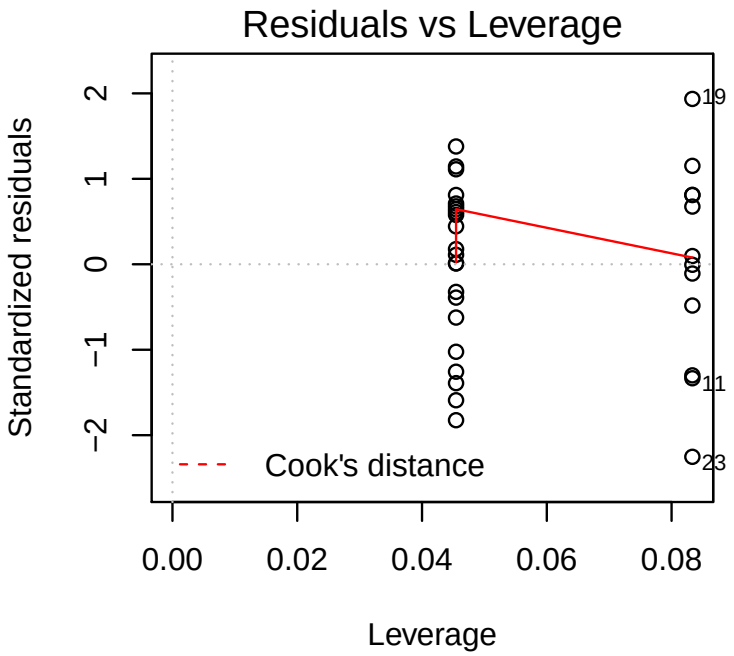
Solution

Using `plot()` in R the following plots are generated:


Residuals vs Fitted
Residuals
Fitted values
19
23
4
10.5
11.0
11.5
12.0
12.5
−8
−4
0
2
4
6


Normal Q−Q
Standardized residuals
Theoretical Quantiles
19
4
23
−2
−1
0
1
2


Scale−Location
√|Standardized residuals|
Fitted values
23
19
4
10.5
11.0
11.5
12.0
12.5
0.0
0.5
1.0
1.5


Residuals vs Leverage
Standardized residuals
Leverage
Cook's distance
19
11
23
0.00
0.02
0.04
0.06
0.08
−2
−1
0
1
2

The assumptions made are: homoscedasticity, normality, linearity and independence.

i. homoscedasticity - the variance appears to be equal
ii. normality - the residuals vs fitted shows the data is reasonably normal.
iii. linearity - the normal q-q plot shows a linear trend in the data.
iv. independence - design assumption - assume to be true

...

The code used in this assignment task is below:

```
library(tidyverse)
setwd("D:/Documents/Uni/Smi")


##Q3
X0 = c(1,1,1,1,1,1,1)
X1=c(-3,-2,-1,0,1,2,3)
X2=c(5,0,-3,-4,-3,0,5)
X3=c(-1,1,1,0,-1,-1,1)
X =cbind(X0,X1,X2,X3)
Y=c(1,0,0,1,2,3,3)
XTXinv = solve(t(X)%*%X)
betahat = XTXinv%*%t(X)%*%Y

Ypredict = c(1,1,-3,-1) %*% betahat

##Hypothesis testing
lambda = c(0,0,0,1)
n=7
p=4
stderror = (1/(n-p)) * t(Y-X%*%betahat)%*%(Y-X%*%betahat)
teststat =t(lambda)%*%betahat/(sqrt(stderror*t(lambda)%*%XTXinv%*%lambda))
pval = dt(teststat,n-p)
#reject Ho as p-val < 0.05
lambda2 = c(1,1,-3,-1)
teststat2 =t(lambda2)%*%betahat/(sqrt(stderror*t(lambda2)%*%XTXinv%*%lambda2))
tval= qt(0.975,n-p)
CI = t(lambda2)%*%betahat+c(-1,1)*tval*(sqrt(stderror*t(lambda2)%*%XTXinv%*%lambda2
## Q4 ----
##read in msleep
data(msleep)
summary(msleep)
```

```
##generate boxplots
#this will speed things up
temp = msleep%>%
   filter (order %in% c("Carnivora","Rodentia"))

pdf(file="BoxplotSleepTotal.pdf")
ggplot(data=temp,aes(x=order,y=sleep_total)) + geom_boxplot() + labs(title="Side-by-side_b
dev.off()
#pooled could be used as the bulk of the data seems to have similar spread, however there is a sl


t.test(sleep_total~order,data=temp)


#Assumption Checking
msleeplin = lm(sleep_total~order,data=temp)

summary(msleeplin)
library(broom)
tidy(msleeplin)
#PDF stuff to make graphs into a pdf
pdf(file="Graphs.pdf")
tmp = par(mfrow = c(2,2))
plot(msleeplin)
dev.off()
par(tmp)
```