Statistical-Methods-2-Homework-3

2023-02-01

Problem 1)

a)

Here, since without loss of generality we can assume the first K observations are from treatment group 1. The first 3 observations would be from treament group J=1. This would be true for the next 2 treament groups, with a total of 9 observations. Also, each indicator variable can be represented as 1 or 0, so:

$$\boldsymbol{X} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix}$$

b)

We can represent the mean response of the jth treament group, because our response variables $Y_1, ..., Y_9$ can have a represented mean of the jth treament group as they are consecutive. So the mean response for the 1st treatment group would $ = (Y_1 + Y_2 + Y_3)/3$ and so on for the other treatment groups. To verify this we can use the formula to find the beta coeffecients:

```
X <- cbind(c(1,1,1,0,0,0,0,0,0),c(0,0,0,1,1,1,0,0,0),c(0,0,0,0,0,0,1,1,1))

before_y <- solve(t(X)%*%X)%*%t(X)
before_y
```

```
##           [,1]      [,2]      [,3]      [,4]      [,5]      [,6]      [,7]
## [1,] 0.3333333 0.3333333 0.3333333 0.0000000 0.0000000 0.0000000 0.0000000
## [2,] 0.0000000 0.0000000 0.0000000 0.3333333 0.3333333 0.3333333 0.0000000
## [3,] 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000 0.3333333
##           [,8]      [,9]
## [1,] 0.0000000 0.0000000
## [2,] 0.0000000 0.0000000
## [3,] 0.3333333 0.3333333
```

Which multiplied with $\boldsymbol{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_9 \end{pmatrix}$ would equal:

$$\hat{\boldsymbol{\beta}} = \begin{pmatrix} \frac{Y_1+Y_2+Y_3}{3} \\ \frac{Y_4+Y_5+Y_6}{3} \\ \frac{Y_7+Y_8+Y_9}{3} \end{pmatrix}$$

Hence, we verified the that b hats of each treatment group j is equal to the mean response of the jth treatment group.

c)

To do this, we find the variance - covariance matrix:

```
vcov <- solve(t(X)%*%X)
vcov
```

```
##           [,1]      [,2]      [,3]
## [1,] 0.3333333 0.0000000 0.0000000
## [2,] 0.0000000 0.3333333 0.0000000
## [3,] 0.0000000 0.0000000 0.3333333
```

No matter the σ^2, the constants in the diagonal of the matrix should hold the same values, therefore interpreting this matrix we can clearly see that the variance are the same for the coefficients, and that there is no covariance (cov = 0) between them therefore are uncorrelated.

Problem 2)

a) and b)

We can determine residuals by: $Y_i - \hat{Y}_i$ and this can be extended to $\boldsymbol{e}_{n\times 1} = \boldsymbol{Y} - \hat{\boldsymbol{Y}}$

$$\sum \boldsymbol{e}_{n\times 1} = \sum \boldsymbol{Y} - \hat{\boldsymbol{Y}} = \sum Y - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y} = 0$$

Which is sum of observed responses, minus the sum of predicted response. Given that prediction is based on the responses, the difference will be 0. This is a generalization from p explanatory variables but it can can be simplified with p = 1 for OLS.

c)

No, typically this is randomness around the normal distribution with mean of 0. Although our expected value might zero for this sum, that doesn't mean it will always be 0. This is inherent with a property of a random variable, as it can vary.

Problem 3)

Here we can use the unbiased estimate of sigma.

a)

```
library("faraway")
X <-as.matrix(cbind(intercept=rep(1,47),teengamb[,-5]))
Y <-teengamb[,5]
I <- diag(Y)
H <- X%*%solve(t(X)%*%X)%*%t(X)
n <- dim(teengamb)[1]

ete <- t(Y-H%*%Y)%*%(Y-H%*%Y)
sigma <- sqrt(ete/(n-(4+1)))
sigma
```

```
##          [,1]
## [1,] 22.69034
```

Our residual standard error is 22.69, which in this context means our predicted vs actual values vary with about a standard deviation of 22.69.

b)

```
mod <- lm(teengamb$gamble ~ teengamb$sex +teengamb$status+teengamb$income+teengamb$verbal)
print(sigma(mod))
```

```
## [1] 22.69034
```

c)

```
vcov <- 22.69034^2*solve(t(X)%*%X)
vcov
```

```
##             intercept        sex      status      income       verbal
## intercept 295.730010 -72.731716 -2.39537913 -9.88839123 -15.18412146
## sex       -72.731716  67.422393  1.27368094  2.46514486  -3.54089828
## status     -2.395379   1.273681  0.07902369  0.09665740  -0.32175471
## income     -9.888391   2.465145  0.09665740  1.05142925  -0.05420869
## verbal    -15.184121  -3.540898 -0.32175471 -0.05420869   4.71823645
```

```
vcov(mod)
```

```
##                 (Intercept) teengamb$sex teengamb$status teengamb$income
## (Intercept)      295.730049   -72.731725     -2.39537944     -9.88839252
## teengamb$sex     -72.731725    67.422402      1.27368110      2.46514518
## teengamb$status   -2.395379     1.273681      0.07902370      0.09665741
## teengamb$income   -9.888393     2.465145      0.09665741      1.05142939
## teengamb$verbal  -15.184123    -3.540899     -0.32175476     -0.05420870
##                 teengamb$verbal
## (Intercept)         -15.1841234
## teengamb$sex         -3.5408987
## teengamb$status      -0.3217548
## teengamb$income      -0.0542087
## teengamb$verbal       4.7182371
```

Problem 4)

a)

```
library("Sleuth3")
data <- case1002
```

Consider the model below:

Energy $= \beta_0 + \beta_1$Mass$+\beta_2$Non Echolocating bat$+\beta_3$Non Echolocating Bird$+\beta_4$Mass$*$Non Echolocating Bat$+\beta_5$Mass$*$Echolo

Where Energy is predicted in-flight energy expenditure (in watts), Mass is body mass (in grams), and we split up each type into its own category along with the interaction. So in this case, echolocating bat would be the case when the other 2 types are = 0, so in a sense its apart of the intercept.

b)

```
mod2 <- lm(Energy ~ Mass + Type + Mass:Type, data = data)
summary(mod2)
```

```
##
## Call:
## lm(formula = Energy ~ Mass + Type + Mass:Type, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.0486 -2.2709 -0.0822  0.9937 12.4601
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     0.49398    3.19470   0.155    0.879
## Mass                            0.08964    0.06804   1.317    0.209
## Typenon-echolocating bats      10.73340    7.06170   1.520    0.151
## Typenon-echolocating birds      2.82276    4.26463   0.662    0.519
## Mass:Typenon-echolocating bats -0.04959    0.06904  -0.718    0.484
## Mass:Typenon-echolocating birds -0.02186    0.06866  -0.318    0.755
##
## Residual standard error: 5.041 on 14 degrees of freedom
## Multiple R-squared:  0.9045, Adjusted R-squared:  0.8703
## F-statistic:  26.5 on 5 and 14 DF,  p-value: 1.136e-06
```

To find the mean energy expenditure for non-echolocating bats when mass is held at zero, we plug in the values with the estimated coefficients. Here, non-echolocating bats = 1 and Mass = 0, so:

```
mod2$coefficients[1] + mod2$coefficients[2]*0+mod2$coefficients[3]*1 + mod2$coefficients[5]*0*1
```

```
## (Intercept)
##    11.22738
```

To find the mean energy expenditure for non-echolocating birds when mass is held at zero, we plug in the values with the estimated coefficients. Here, non-echolocating birds = 1 and Mass = 0, so:

```
mod2$coefficients[1] + mod2$coefficients[2]*0+mod2$coefficients[4]*1 + mod2$coefficients[6]*0*1
```

```
## (Intercept)
##    3.316742
```

To find the mean energy expenditure for echolocating bats, we simply don't indicate for the other 2 types (non echolocating) and we also hold mass at zero. So in this case, it's actually the intercept!

```
mod2$coefficients[1] + mod2$coefficients[2]*0
```

```
## (Intercept)
##    0.493983
```

I will choose β_5 which is the interaction between mass and non echolocating bird type. This coeffecient gives the difference in the change in the mean of energy expenditure when mass is increased by one unit and when the type is non echo locating bird. (ie = 1), compared to when mass is increased by one unit when the type is NOT non echo locating bird (ie = 0). That difference would be -.02186.