SCHOOL OF MATHEMATICS AND STATISTICS

STAT 393 **Assignment 2** **Due: 11:59pm Thursday 11 August 2022**

Linear Models

This assignment is worth 10%

Note: *For any part of the assignment that requires using R, include R code and output. Make sure that your answers are clearly refered to the appropriate place in the R output. My suggestion is as follows:*

- *copy and paste R code and output into a word document, or*
- *create the pdf that comes from knitting the file using R Markdown.*

You could print the document and write your answers. Alternatively, you could type your answers in the document.

1. Low birth weight, defined as birth weight less than 2500 grams, is an outcome that has been of concern to physicians. This is due to the fact that infant mortality rates and birth defect rates are very high for low birth weight babies. A woman's behavior during pregnancy (e.g., smoking habits) can greatly alter the chances of carrying the baby to term and, consequently, of delivering a baby of normal birth weight.

   (a) Consider the following variables:
   - y: baby's weight (kg),
   - *smoke*: smoking status during pregnancy (0 for no; 1 for yes),
   - *race*: with three categories (1, 2, and 3),
   - *weight*: weight of mother at last menstrual period (kg).

   Generate the data using the R code (`a2_Q1_data.R`), available on Blackboard. **Note**: run the code without changing anything. For example, if the order of lines changes, the dataset becomes different. Attach the dataset.

   (b) Let y_i be the baby's weight, $smoke_i$ be the smoking status, $weight_i$ be the weight of mother at last mentrual period, and $race_i$ be the race of the ith mother. Consider the following models:

   M1: $y_i = \alpha + \beta\, smoke_i + \gamma\, weight_i + \varepsilon_i$
   M2: $y_i = \alpha + \beta\, smoke_i + \gamma\, weight_i + \tau_{race_i} + \varepsilon_i$

   where $\varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$. Fit models M1 and M2, respectively, using `lm()` in R, with the constraint $\tau_1 = 0$. Run `summary()`. For each of the models, give coefficients estimates, coefficients standard errors, p-values, and residual standard error. **Note:** Don't forget to setup *race* as a factor in R using `babydata$race <- factor(race)`.

   (c) Consider the model M2 from part (b). Show the model matrix X in R. Write down the vector of parameters $\boldsymbol{\beta}$ in the matrix form $\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$.

   (d) Consider the model M2 from part (b). Use the matrix method to reproduce the following results given by `summary()`:
   i. coefficients estimates,
   ii. residual standard error, and
   iii. coefficients standard errors.

(e) Consider the model M2 from part (b). State in words the interpretation of

i. $\hat{\beta}$

ii. $\hat{\gamma}$

iii. $\hat{\tau}_2$

iv. $\hat{\tau}_3$

2. Consider a linear model $y_i = \beta x_i + \varepsilon_i$, $i = 1, \ldots, n$, with $\varepsilon_i \sim N(0, \sigma^2)$ for a given n-pairs of values $(x_i, y_i), i = 1, \ldots, n$. Show that the MLE of β has the form

$$\hat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}$$

SCHOOL OF MATHEMATICS AND STATISTICS

STAT 393 **Assignment 2**

Linear Models
SOLUTIONS

1.

Assignment 2 Q1 solutions

Ivy Liu

11/08/2022

1 (a) Generate the data.

```
set.seed(300123456)
smoke <- rbinom(30,1,0.3)
race <- sample(1:3,size=30,replace=TRUE, prob=c(0.5,0.2,0.3))
weight <- rnorm(30,67,10)
e <- rnorm(30,0,0.6)
mu <- 4-0.8*smoke-0.007*weight+0.5*ifelse(race==1,1,0)+0.2*ifelse(race==2,1,0)-0.7*ifelse(race==3,1,0)
y <- mu+e
babydata <- data.frame(y,smoke,race,weight)
babydata
```

```
##            y smoke race   weight
## 1  2.5067959     0    3 49.75558
## 2  4.5040646     0    2 55.13305
## 3  4.0925889     0    1 66.34796
## 4  2.8732643     1    1 76.33683
## 5  2.7207261     1    2 63.22431
## 6  4.2707540     0    1 72.09313
## 7  3.4306684     0    3 60.46014
## 8  0.2448980     1    3 77.17965
## 9  3.9441334     0    3 54.95945
## 10 2.8309891     0    3 70.87376
## 11 3.0759469     1    1 69.92665
## 12 4.2271277     0    1 58.43894
## 13 3.9643385     1    1 76.53096
## 14 2.6543878     1    1 95.79329
## 15 3.8754072     0    1 68.24418
## 16 2.5535185     1    3 69.11276
## 17 3.6057961     0    1 65.45454
## 18 4.4308606     0    1 67.53852
## 19 4.9265515     0    1 69.10808
## 20 3.4351882     0    3 86.96658
## 21 3.2007138     1    1 73.53993
## 22 4.6278990     0    1 71.35932
## 23 4.0509876     0    2 53.33690
## 24 2.3696767     0    3 65.51677
## 25 0.9508626     1    3 65.37405
## 26 2.1278282     1    1 68.50423
## 27 3.0724919     1    1 70.20624
## 28 3.4897733     0    1 86.27051
## 29 3.0561780     1    1 59.29946
## 30 3.1468707     0    1 66.95694
```

1 (b) Fit model M1: $y_i = \alpha + \beta\ smoke_i + \gamma\ weight_i + \varepsilon_i$.

```
babydata$race <- factor(race)
out1<-lm(y~smoke+weight,data=babydata,x=T)
summary(out1)
```

```
##
## Call:
## lm(formula = y ~ smoke + weight, data = babydata, x = T)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.3011 -0.3440  0.1887  0.5314  1.4189
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.703492   1.130257   3.277  0.00289 **
## smoke       -1.229128   0.338855  -3.627  0.00118 **
## weight       0.000928   0.016831   0.055  0.95643
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8674 on 27 degrees of freedom
## Multiple R-squared:  0.3467, Adjusted R-squared:  0.2983
## F-statistic: 7.165 on 2 and 27 DF,  p-value: 0.00319
```

$\hat{\alpha} = 3.7035$, s.e $(\hat{\alpha})$=1.1303, p-value $= 0.00289$

$\hat{\beta} = -1.2291$, s.e $(\hat{\beta})$=0.3389, p-value $= 0.00118$

$\hat{\gamma} = 0.00093$, s.e $(\hat{\gamma})$=0.0168, p-value $= 0.9564$

$\hat{\sigma} = 0.8674$

Fit model M2: $y_i = \alpha + \beta\ smoke_i + \gamma\ weight_i + \tau_{race_i} + \varepsilon_i$.

```
out2<-lm(y~smoke+weight+ race,data=babydata,x=T)
summary(out2)
```

```
##
## Call:
## lm(formula = y ~ smoke + weight + race, data = babydata, x = T)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.30282 -0.45257  0.00443  0.32994  1.12079
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.588025   0.958509   4.787 6.48e-05 ***
## smoke       -1.297670   0.249703  -5.197 2.24e-05 ***
## weight      -0.005838   0.013664  -0.427    0.673
## race2       -0.062746   0.438072  -0.143    0.887
## race3       -1.292043   0.266807  -4.843 5.60e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6368 on 25 degrees of freedom
```

```
## Multiple R-squared:  0.6739, Adjusted R-squared:  0.6218
## F-statistic: 12.92 on 4 and 25 DF,  p-value: 7.771e-06
```

$\hat{\alpha} = 4.5880$, s.e ($\hat{\alpha}$)=0.9585, p-value = 0.00006

$\hat{\beta} = -1.2977$, s.e ($\hat{\beta}$)=0.2497, p-value = 0.0000224

$\hat{\gamma} = -0.005838$, s.e ($\hat{\gamma}$)=0.01366, p-value = 0.673

$\hat{\tau}_2 = -0.0627$, s.e ($\hat{\tau}_2$)=0.438, p-value =0.887

$\hat{\tau}_3 = -1.2920$, s.e ($\hat{\tau}_3$)=0.2668, p-value =0.000056

$\hat{\sigma} = 0.6368$

1 (c) Consider the model M2 from part (b). Show the model matrix X in R. Write down the vector of parameters $\boldsymbol{\beta}$ in the matrix form $\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$.

```
out2$x
```

```
##    (Intercept) smoke   weight race2 race3
## 1            1     0 49.75558     0     1
## 2            1     0 55.13305     1     0
## 3            1     0 66.34796     0     0
## 4            1     1 76.33683     0     0
## 5            1     1 63.22431     1     0
## 6            1     0 72.09313     0     0
## 7            1     0 60.46014     0     1
## 8            1     1 77.17965     0     1
## 9            1     0 54.95945     0     1
## 10           1     0 70.87376     0     1
## 11           1     1 69.92665     0     0
## 12           1     0 58.43894     0     0
## 13           1     1 76.53096     0     0
## 14           1     1 95.79329     0     0
## 15           1     0 68.24418     0     0
## 16           1     1 69.11276     0     1
## 17           1     0 65.45454     0     0
## 18           1     0 67.53852     0     0
## 19           1     0 69.10808     0     0
## 20           1     0 86.96658     0     1
## 21           1     1 73.53993     0     0
## 22           1     0 71.35932     0     0
## 23           1     0 53.33690     1     0
## 24           1     0 65.51677     0     1
## 25           1     1 65.37405     0     1
## 26           1     1 68.50423     0     0
## 27           1     1 70.20624     0     0
## 28           1     0 86.27051     0     0
## 29           1     1 59.29946     0     0
## 30           1     0 66.95694     0     0
## attr(,"assign")
## [1] 0 1 2 3 3
## attr(,"contrasts")
## attr(,"contrasts")$race
## [1] "contr.treatment"
```

$$\boldsymbol{\beta} = \begin{pmatrix} \alpha \\ \beta \\ \gamma \\ \tau_2 \\ \tau_3 \end{pmatrix}$$

1 (d) Consider the model M2 from part (b). Use the matrix method to reproduce the following results given by `summary()`: (i) coefficients estimates

```
y <- matrix(babydata$y,nrow=30,ncol=1)
x <- matrix(out2$x,nrow=30)
xtx <- t(x)%*%x
xtx
```

```
##          [,1]     [,2]        [,3]     [,4]     [,5]
## [1,]   30.000  12.0000   2053.8427   3.0000   9.0000
## [2,]   12.000  12.0000    865.0284   1.0000   3.0000
## [3,] 2053.843 865.0284 143527.3351 171.6943 600.1987
## [4,]    3.000   1.0000    171.6943   3.0000   0.0000
## [5,]    9.000   3.0000    600.1987   0.0000   9.0000
```

```
inv.xtx <- solve(xtx)

xty <- t(x)%*%y
xty
```

```
##            [,1]
## [1,]   98.26129
## [2,]   30.49515
## [3,] 6676.35968
## [4,]   11.27578
## [5,]   22.26673
```

```
beta <- inv.xtx%*%xty
beta
```

```
##             [,1]
## [1,]  4.58802530
## [2,] -1.29766991
## [3,] -0.00583823
## [4,] -0.06274564
## [5,] -1.29204319
```

```
out2$coef
```

```
## (Intercept)       smoke      weight       race2       race3
##  4.58802530 -1.29766991 -0.00583823 -0.06274564 -1.29204319
```

The coefficients estimates are $\hat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T \mathbf{y}$. In `R`, the notation `xtx` $= (X^T X)$, `inv.xtx` $= (X^T X)^{-1}$, `xty` $= X^T \mathbf{y}$, and `beta` $= \hat{\boldsymbol{\beta}}$.

The matrix method produces the same result as in `summary()`.

(ii) The residual standard error

```
id <- diag(rep(1,30))
mse <- (t(y)%*%(id-x%*%inv.xtx%*%t(x))%*%y)/(30-5)
mse
```

```
##           [,1]
## [1,] 0.4055266
```

```
se <- mse^0.5
se
```

```
##           [,1]
## [1,] 0.6368097
```

$\hat{\sigma} = \sqrt{MSE} = \sqrt{\mathbf{y}^T(I - X(X^TX)^{-1}X^T)\mathbf{y}/(n-p)}$. In R, the notation `id` $= I$, `mse` $= MSE$, and `se` $= \hat{\sigma}$.

The matrix method produces the same result as in `summary()`.

(iii) coefficients standard errors

```
cov <- inv.xtx*as.numeric(mse)
cov
```

```
##             [,1]         [,2]          [,3]         [,4]          [,5]
## [1,]  0.91873983  0.043298431 -0.0128539970 -0.197520114 -0.0759556554
## [2,]  0.04329843  0.062351776 -0.0009970640 -0.007018963  0.0024105979
## [3,] -0.01285400 -0.000997064  0.0001867066  0.002500868  0.0007351227
## [4,] -0.19752011 -0.007018963  0.0025008681  0.191907075  0.0330800031
## [5,] -0.07595566  0.002410598  0.0007351227  0.033080003  0.0711862211
```

```
diag(cov)^0.5
```

```
## [1] 0.95850917 0.24970338 0.01366406 0.43807200 0.26680746
```

```
coef(summary(out2))[,2]
```

```
## (Intercept)       smoke      weight       race2       race3
##  0.95850917  0.24970338  0.01366406  0.43807200  0.26680746
```

The estimated variance and covariance matrix for $\hat{\boldsymbol{\beta}}$ is $\hat{\text{Var}}(\hat{\boldsymbol{\beta}}) = \hat{\sigma}^2(X^TX)^{-1}$. In R, the notation `cov` $= \hat{\sigma}^2(X^TX)^{-1}$, and `diag(cov)^0.5` $= \sqrt{\hat{\sigma}^2(X^TX)^{-1}_{\ell\ell}}$ (square root of each diagonal elements in $\hat{\sigma}^2(X^TX)^{-1}$).

The matrix method produces the same result as in `summary()`.

1 (e) Consider the model M2 from part (b). State in words the interpretation of:

(i) $\hat{\beta} = -1.2977$

Given the weight of mother at last menstrual period, and race at a fixed level, the estimated average baby birth weight is 1.29771kg lighter for a smoker mother than a non-smoker mother.

(ii) $\hat{\gamma} = -0.0058$

Given the smoking status, and race at a fixed level, the estimated average baby birth weight decreases by 0.0058kg for every 1kg weight increase of mother at last menstrual period. Note: Because the p-value $= 0.67$ (> 0.05), the baby birth weight doesn't depend on the weight of mother at last menstrual period.

(iii) $\hat{\tau}_2 = -0.0627$

Given the smoking status and the weight of mother at last menstrual period, the estimated average baby birth weight is 0.0627kg lighter for a mother with race 2 than a mother with race 1.

(iv) $\hat{\tau}_3 = -1.292$

Given the smoking status and the weight of mother at last menstrual period, the estimated average baby birth weight is 1.292kg lighter for a mother with race 3 than a mother with race 1.

2. Write the model $y_i = \beta x_i + \varepsilon_i$ in the matrix form $\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$. We have

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad X = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}, \quad \text{and} \quad \boldsymbol{\beta} = \beta.$$

Note: This model doesn't have an intercept. Thus, the regression line passes the origin (0, 0).

Following the above matrix representation of the linear model, we have

$$X^\top X = (x_1, x_2, \ldots, x_n) \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = \sum_{i=1}^{n} x_i^2.$$

and

$$X^\top \mathbf{y} = (x_1, x_2, \ldots, x_n) \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \sum_{i=1}^{n} x_i y_i.$$

Therefore,

$$\widehat{\boldsymbol{\beta}} = \left(X^\top X\right)^{-1} X^\top \mathbf{y} = \left(\sum_{i=1}^{n} x_i^2\right)^{-1} \sum_{i=1}^{n} x_i y_i = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}.$$