ISE789/OR791
Homework 2

1. Let $y_i = \beta_0 + \beta_1(x_i - \bar{x}) + \epsilon_i$ for $i = 1, \cdots, n$. Using the formula $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$, show that

$$\begin{aligned} \hat{\beta}_0 &= \bar{y} \\ \hat{\beta}_1 &= \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}. \end{aligned}$$

   Now using the formula $var(\hat{\boldsymbol{\beta}}) = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$, show that

$$\begin{aligned} var(\hat{\beta}_0) &= \frac{\sigma^2}{n} \\ var(\hat{\beta}_1) &= \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}. \end{aligned}$$

2. Prove the following: (i) $var(e_i) = \sigma^2 - var(\hat{y}_i)$. (ii) $var(e_i) = \sigma^2(1 - \boldsymbol{x}_i'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_i)$, where $\boldsymbol{x}_i$ is the ith row of the $\boldsymbol{X}$ matrix.

3. Suppose we have n data from the model $y = \boldsymbol{x}'\boldsymbol{\beta} + \epsilon$, where the error satisfies the Gauss-Markov assumptions. Suppose, further, that we wish to predict the $(n+1)$st observation y_{n+1} at $\boldsymbol{x}_{n+1}$. The predictor based on the least squares estimate of $\boldsymbol{\beta}$ is given by $\hat{y}_{n+1} = \boldsymbol{x}_{n+1}'\hat{\boldsymbol{\beta}}$.

   (a) Show that $E(\hat{y}_{n+1} - y_{n+1}) = 0$.

   (b) Suppose $\tilde{y}_{n+1} = \boldsymbol{a}'\boldsymbol{y}$ is another predictor of y_{n+1} such that $E(\tilde{y}_{n+1} - y_{n+1}) = 0$. Show that $\boldsymbol{a}$ must satisfy $\boldsymbol{a}'\boldsymbol{X} = \boldsymbol{x}_{n+1}'$.

   (c) Show that $var(\hat{y}_{n+1}) \leq var(\tilde{y}_{n+1})$.

4. The dataset *teengamb* (available in the R library *faraway*) concerns a study of teenage gambling in Britain. Fit a regression model with the expenditure on gambling as the response and the sex (coded as male=0 and female=1), status, income, and verbal score as predictors. Present the output and answer the following questions.

(a) What percentage of variation in the response is explained by these predictors?

(b) Which observation has the largest (positive) residual? Give the case number.

(c) Compute the mean and median of the residuals.

(d) Compute the correlation of the residuals with the fitted values.

(e) Compute the correlation of the residuals with the income.

(f) For all other predictors held constant, what would be the difference in predicted expenditure on gambling for a male compared to a female.

(g) Which variables are statistically significant?

(h) Predict the amount that a male with average (given these data) status, income and verbal score would gamble along with an appropriate 95% CI. Repeat the prediction for a male with maximal values (for this data) of status, income, and verbal score. Which CI is wider and why is this expected?

(i) Fit a model with just income as a predictor and use an F-test to compare it to the full model.

Remarks:
1) Questions (f) and (h) will need to use the knowledge to be introduced in next class
2) Please find the dataset in the attachment. You may also find the data in R library *faraway*
3) Please choose the programming language you are comfortable with and submit your code along with your solution

ISE 789 / OR 791

Home work - 2

1.

Given :

$$y_i = \beta_0 + \beta_1 (x_i - \bar{x}) + \varepsilon_i \text{ for } i = 1 \ldots\ldots, n \quad — ①$$

$$\hat{\beta} = (X^T X)^{-1} X^T y \quad — ②$$

To prove :

$$\hat{\beta}_0 = y.$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{i=n} (x_i - \bar{x})^2}$$

Solution :

We can write ① as. $\underset{Y}{(y_i - \bar{y})} = \underset{\beta}{\hat{\beta}_1} \underset{X}{(x_i - \bar{x})}$

$\hat{\beta}_1 = \sum_{i=1}^{n} \left[((x_i - \bar{x})^T (x_i - \bar{x}))^{-1} (x_i - \bar{x})^T (y_i - \bar{y}) \right]$ — putting. $X = x_i - \bar{x}$ and $y = (y_i - \bar{y})$ into ②.

$$\hat{\beta}_1 = \sum_{i=1}^{n} \left[((x_i - \bar{x})^2)^{-1} (x_i - \bar{x})(y_i - \bar{y}) \right]$$

$$\boxed{\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{i=n} (x_i - \bar{x})^2}}$$

Putting $\hat{\beta}_1$ value in eqn. ①.

$$\beta_0 = y_i - \beta_1 (x_i - \bar{x}) - \varepsilon_i$$

$$\hat{\beta}_0 = \sum_{i=1}^{n} \left[(y_i - \bar{y}) - \hat{\beta}_1 (x_i - \bar{x}) \right]$$

2

For $\hat{\beta}_0$ $\quad X = \begin{bmatrix} 1 \\ 1 \\ 1 \\ \vdots \\ n \end{bmatrix}_{n \times 1}$ as intercept at $x = 1$

So, $$\hat{\beta}_0 = \sum_{i=1}^{n} \left(\begin{bmatrix} 1 \\ \vdots \\ n \end{bmatrix}^T_{n \times 1} \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} \right)^{-1} \left(\begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}^T y \right)$$

$$= \sum_{i=1}^{n} (n)^{-1} (y)$$

$$= \frac{\sum_{i=1}^{n} y}{n} = \bar{y}.$$

$$\boxed{\hat{\beta}_0 = \bar{y}}$$

$$Var(\hat{\beta}) = \sigma^2 (X^T X)^{-1}$$

$$Var(\hat{\beta}_0) = \sum_{i=1}^{n} \sigma^2 \left([\vdots]^T [\vdots] \right)^{-1} = \sigma^2 (n)^{-1} = \boxed{\frac{\sigma^2}{n}}$$

$$Var(\hat{\beta}_1) = \sigma^2 \sum_{i=}^{n} \left((x_i - \bar{x})^T (x_i - \bar{x}) \right)^{-1}$$

$$\boxed{Var(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}}$$

2) To prove:

$$Var(e_i) = \sigma^2 - var(y_i)$$

$$Var(e_i) = var(y_i - \hat{y}_i) = Var(x_i\beta + \varepsilon_i - x_i\hat{\beta})$$

$$= \cancel{} \; var(\underbrace{x_i\beta}_{\text{constant}}) + var(\underbrace{\varepsilon_i}_{\sigma^2}) - var(x_i\hat{\beta})$$

$$= 0 + \sigma^2 - var(x_i\hat{\beta}) = \sigma^2 - var(\hat{y}_i)$$

$$\boxed{Var\, e_i = \sigma^2 - var(x_i \hat{\beta})} \quad \text{—①.}$$

To prove: $Var(e_i) = \sigma^2 (1 - x_i^T (X^T X)^{-1} x_i)$

From ① $Var\, e_i = \sigma^2 - var(x_i\hat{\beta}) = \sigma^2 - \cancel{x}\, x_i^T Var(\hat{\beta}_i)$

$$= \sigma^2 - x_i^T \sigma^2 (X^T X)^{-1} x_i$$

$$= \sigma^2 - \sigma^2 x_i^T (X^T X)^{-1} x_i$$

$$\boxed{Var\, e_i = \sigma^2 (1 - x_i^T (X^T X)^{-1} x_i)}$$

3

we have n data from model $y = x^T \beta + \epsilon$.

error is in Gauss - Markov assumptions.

we wish to predict $(n+1)$st obs

y_{n+1} at x_{n+1}.

and $\hat{y}_{n+1} = x^T_{n+1} \hat{\beta}$.

(a). $E(\hat{y}_{n+1} - y_{n+1}) = 0$.

$E(\hat{y}_{n+1} - y_{n+1}) = E(X^T_{n+1} \hat{\beta} - X^T_{n+1} \beta - \epsilon)$

$= E(X^T_{n+1} \hat{\beta}) - X^T_{n+1} \beta - E(\epsilon)$

$= X^T_{n+1} E(\hat{\beta}) - X^T_{n+1} \beta - 0$

$= x^T_{n+1} \beta - x^T_{n+1} \beta$ (as $\hat{\beta}$ is unbiased estimator)

$= 0$.

(b). Suppose $\tilde{y}_{n+1} = a'y$ another predictor of y_{n+1}

such that $E(\tilde{y}_{n+1} - y_{n+1}) = 0$.

$E(\tilde{y}_{n+1} - y_{n+1}) = E(a'y - X^T_{n+1} \beta - \epsilon) = 0$

$\Rightarrow E(a'(X\beta + \epsilon) - X^T_{n+1} \beta - \epsilon) = 0$

$\Rightarrow E(a'X\beta - X^T_{n+1} \beta + a'\epsilon - \epsilon) = 0$

$\Rightarrow$ ~~a~~ (

$$\Rightarrow \; a'X\beta - x^T_{n+1}\beta + E[(a'-1)\varepsilon] = 0$$

$$\Rightarrow \; a'X\beta - x^T_{n+1}\beta + a'-1\, \underbrace{E[\varepsilon]}_{\to 0} = 0$$

$$\Rightarrow \; \boxed{a'X\beta = x^T_{n+1}}$$

(a) ~~Var~~ $Var(\hat{y}_{n+1}) \leq$ ~~Var~~ $Var(\tilde{y}_{n+1})$

$$Var(\hat{y}_{n+1}) = Var(x'_{n+1}\beta)$$

$$= \sigma^2((x'_{n+1})^T(X^TX)^{-1}(x'_{n+1}))$$

$$= \sigma^2(x_{n+1}(X^TX)^{-1}(x'_{n+1})$$

$$var(\tilde{y}_{n+1}) = var(a'y) = var(a'y).$$

$$=(a')'\, var(x\beta)a'$$

$$= ax'\, var(\hat{\beta})xa'$$

$$= a\, x'\sigma^2(X^TX)^{-1}xa'$$

$$= \sigma^2\, a\, X'(X'X)^{-1}Xa'$$

Since $a'x = x'_{n+1}$ then

$$var(\hat{y}_{n+1}) = \sigma^2(x_{n+1})(X^TX)^{-1}(x'_{n+1})$$

$\therefore \ var(\tilde{y}_{n+1}) = var(\tilde{y}_{n+1})$

otherwise $\quad var(\hat{y}_{n+1}) < var(\tilde{y}_{n+1})$

$$\boxed{\therefore \ var(\hat{y}_{n+1}) \leq var(\tilde{y}_{n+1})}$$

hw_2.R

4. The dataset teengamb (available in the R library faraway) concerns a study of teenage gambling in Britain. Fit a regression model with the expenditure on gambling as the response and the sex (coded as male=0 and female=1), status, income, and verbal score as predictors. Present the output and answer the following questions.

(a) What percentage of variation in the response is explained by these predictors?

(b) Which observation has the largest (positive) residual? Give the case number.

(c) Compute the mean and median of the residuals.

(d) Compute the correlation of the residuals with the fitted values.

(e) Compute the correlation of the residuals with the income.

(f) For all other predictors held constant, what would be the difference in predicted expenditure on gambling for a male compared to a female.

(g) Which variables are statistically significant?

(h) Predict the amount that a male with average (given these data) status, income and verbal score would gamble along with an appropriate 95% CI. Repeat the prediction for a male with maximal values (for this data) of status, income, and verbal score. Which CI is wider and why is this expected?

(i) Fit a model with just income as a predictor and use an F-test to compare it to the full model.

Solution:

deepa

Tue Sep 24 20:50:26 2019

```
## Accessing the library Faraway to access the data

library(faraway)

## Warning: package 'faraway' was built under R version 3.5.3

##Showing the data and doing categorical distribution of sex variable

data(teengamb)
teengamb$sex <- factor(teengamb$sex)
attach(teengamb)
teengamb[1:3,]

##   sex status income verbal gamble
## 1   1     51    2.0      8      0
## 2   1     28    2.5      8      0
## 3   1     37    2.0      6      0

## Fitting linear model on gamble data on sex, status, income, verbal variabl
es

gamb.lm <- lm(gamble ~ sex+status+income+verbal)

## Showing statistics of the variable
summary(gamb.lm)

##
## Call:
## lm(formula = gamble ~ sex + status + income + verbal)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -51.082 -11.320  -1.451   9.452  94.252
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  22.55565   17.19680   1.312   0.1968
## sex1        -22.11833    8.21111  -2.694   0.0101 *
## status        0.05223    0.28111   0.186   0.8535
## income        4.96198    1.02539   4.839 1.79e-05 ***
## verbal       -2.95949    2.17215  -1.362   0.1803
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 22.69 on 42 degrees of freedom
```

```
## Multiple R-squared:  0.5267, Adjusted R-squared:  0.4816
## F-statistic: 11.69 on 4 and 42 DF,  p-value: 1.815e-06
```

A. Since, R squared is 0.5267 so, 52.67% of the response is explained by these predictors

```
gamb.lm$residuals
```

```
##           1           2           3           4           5           6
##  10.6507430   9.3711318   5.4630298 -17.4957487  29.5194692  -2.9846919
##           7           8           9          10          11          12
##  -7.0242994 -12.3060734   6.8496267 -10.3329505   1.5934936  -3.0958161
##          13          14          15          16          17          18
##   0.1172839   9.5331344   2.8488167  17.2107726 -25.2627227 -27.7998544
##          19          20          21          22          23          24
##  13.1446553 -15.9510624 -16.0041386  -9.5801478 -27.2711657  94.2522174
##          25          26          27          28          29          30
##   0.6993361  -9.1670510 -25.8747696  -8.7455549  -6.8803097 -19.8090866
##          31          32          33          34          35          36
##  10.8793766  15.0599340  11.7462296  -3.5932770 -14.4016736  45.6051264
##          37          38          39          40          41          42
##  20.5472529  11.2429290 -51.0824078   8.8669438  -1.4513921  -3.8361619
##          43          44          45          46          47
##  -4.3831786 -14.8940753   5.4506347   1.4092321   7.1662399
```

```
max(gamb.lm$residuals)
```

```
## [1] 94.25222
```

B. 24^{th} observation has maximum residual of 94.2522174

```
mean(gamb.lm$residuals)
```

```
## [1] -3.065293e-17
```

```
median(gamb.lm$residuals)
```

```
## [1] -1.451392
```

C. Mean of the residual is approximately 0 while median is -1.451392.

```
fitted_value <- gamble - gamb.lm$residuals
cor(gamb.lm$residuals, fitted_value)
```

```
## [1] -1.070659e-16
```

D. Correlation between residuals and fitted value is approximately zero.

```
cor(gamb.lm$residuals, income)

## [1] -7.242382e-17

## E. Correlation of residuals and income is almost 0

## F. Keeping everything constant, since coefficient is -22.11833 thus averag
e female teen spend $ 22.1183 less than male teen.

## G. Since P-value of only income and sex is less than 5% so, only these are
significantly important.

## H.

male=data.frame(sex=0, status=mean(teengamb$status), income=mean(teengamb$inc
ome), verbal=mean(teengamb$verbal))
```

```
predict(gamb.lm,male, se.fit=FALSE, interval='confidence')
       fit      lwr      upr
1 28.24252 18.78277 37.70227
```

male_max=data.frame(sex=0, status=max(teengamb$status), income=max(teengamb$inco
me), verbal=max(teengamb$verbal))

```
predict(gamb.lm,male_max, se.fit=FALSE, interval='confidence')
       fit      lwr      upr
1 71.30794 42.23237 100.3835
```

Confidence interval of maximum values are wide because values of predictor is far away fro
m regression line.

I

```
gamb_income.lm <- lm(gamble ~ income)
summary(gamb_income.lm)

Call:
lm(formula = gamble ~ income)

Residuals:
    Min      1Q  Median      3Q     Max
-46.020 -11.874  -3.757  11.934 107.120

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   -6.325      6.030  -1.049      0.3
income         5.520      1.036   5.330 3.05e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 24.95 on 45 degrees of freedom
Multiple R-squared:  0.387,	Adjusted R-squared:  0.3734
F-statistic: 28.41 on 1 and 45 DF,  p-value: 3.045e-06
```

F-Test for full model is 11.69 on 4 variables and F-Test for income is 28.41 on 1 variable so, income model is working better than the full model.