MATH 6364: Statistical Methods

Mid term

Author: Md Salman Rahman

problem 1:

Solution:

Matreice $X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}_{n\times 2}$

and $x^T = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_n \end{bmatrix}_{2\times n}$

$$X^TX = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_n \end{bmatrix}_{2\times n} \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}_{n\times 2}$$

$$= \begin{bmatrix} 1\times1+1\times1+\cdots+1\times1 & x_1+x_2+\cdots+x_n \\ x_1+x_2+\cdots+x_n & x_1^2+x_2^2+\cdots+x_n^2 \end{bmatrix}_{2\times 2}$$

$$= \begin{bmatrix} n & \sum_i x_i \\ \sum_i x_i & \sum_i x_i^2 \end{bmatrix}_{2\times 2}$$

Inverse:

$$(X^T X)^{-1} = \frac{1}{n \sum_i x_i^2 - \left(\sum_i x_i\right)^2} \begin{bmatrix} \sum_i x_i^2 & -\sum_i x_i \\ -\sum_i x_i & n \end{bmatrix}$$

The linear regression model:

$$y = \beta_0 + \beta_1 x$$

$$\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x)^2$$

Estimates of β_0 and β_1:

The method of estimating β_0, β_1 by minimizing $S(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x)^2$ is referred to as method of least square. There are other principle available but for this problem I will use least square estimates.

$$\text{The error } \epsilon_i = y_i - \mu_i = y_i - \beta_0 - \beta_1 x_i \quad (i = 1, 2, \ldots n)$$

To obtain a line $\mu_i = \beta_0 + \beta_1 x_i$ that is closest to the point (x_i, y_i), the error ϵ_i should be as small as possible. Our aim is to minimize the function.

So, letus write,

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (y_i - \mu_i)^2$$

$$= \sum (y_i - \beta_0 - \beta_1 x_i)^2$$

Now, taking derivative with respect to β_0 and β_1 and setting the derivatives to zero.

$$\frac{\partial S(\beta_0, \beta_1)}{\partial \beta_0} = \frac{\partial}{\partial \beta_0}\left(\sum_{i=1}^{n}\left(y_i - \beta_0 - \beta_1 x_i\right)^2\right)$$

$$\Rightarrow (+2) \sum_{i=1}^{n} \left(y_i - \beta_0 - \beta_1 x_i\right)(-1) = 0 \quad \text{------} \quad (1)$$

And,
$$\frac{\partial S(\beta_0, \beta_1)}{\partial \beta_1} = \frac{\partial}{\partial \beta_1}\left(\sum_{i=1}^{n}\left(y_i - \beta_0 - \beta_1 x_i\right)^2\right)$$

$$\Rightarrow (+2) \sum_{i=1}^{n} \left(y_i - \beta_0 - \beta_1 x_i\right)(-x_i) = 0 \quad \text{------} \quad (2)$$

From equation (1) and (2) $\Rightarrow$

Normal equation
$$\begin{cases} n\beta_0 + \left(\sum x_i\right)\beta_1 = \sum y_i \\ \left(\sum x_i\right)\beta_0 + \left(\sum x_i^2\right)\beta_1 = \sum x_i y_i \end{cases}$$

Now solving,
$$n\beta_0 + \left(\sum x_i\right)\beta_1 = \sum y_i$$

$$\Rightarrow n\beta_0 = \sum y_i - \left(\sum x_i\right)\beta_1$$

$$\Rightarrow \beta_0 = \frac{\sum y_i}{n} - \beta_1 \frac{\sum x_i}{n}$$

$$\Rightarrow \beta_0 = \bar{y} - \beta_1 \bar{x}$$

where,
$$\bar{y} = \frac{\sum y_i}{n}$$

$$\bar{x} = \frac{\sum x_i}{n}$$

$\hat{\beta}_1$ can be written as

$$\hat{\beta}_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}}$$

because,

$$\frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

$$= \frac{\sum x_i y_i - \bar{y} \sum x_i - \bar{x} \sum y_i + n \bar{x} \bar{y}}{\sum x_i^2 - 2\bar{x} \sum x_i + n(\bar{x})^2}$$

$$= \frac{\sum x_i y_i - \frac{\sum y_i}{n} \sum x_i - \frac{\sum x_i}{n} \sum y_i + \cancel{n} \frac{\sum x_i}{\cancel{n}} \frac{\sum y_i}{n}}{\sum x_i^2 - 2 \frac{\sum x_i}{n} \sum x_i + \cancel{n} \frac{1}{\cancel{n^2}_{\,n}} \left(\sum x_i\right)^2}$$

$$= \frac{\sum x_i y_i - \frac{1}{n} \sum x_i \sum y_i}{\sum x_i^2 - \frac{1}{n} \left(\sum x_i\right)^2}$$ (this is we got in previous page)

So, $$\hat{\beta}_1 = \frac{\sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n}}{\sum x_i^2 - \frac{(\sum x_i)^2}{n}} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}}$$

and, $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$ where, $\bar{y} = \frac{\sum y_i}{n}$ & $\bar{x} = \frac{\sum x_i}{n}$

called least square estimates (LSE) of β_0 and β_1.

Finding the variance of β_0 and β_1:

In the previous page we find,

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

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

$$= \frac{\sum_{i=1}^{n}(x_i-\bar{x})\,y_i}{\sum_{i=1}^{n}(x_i-\bar{x})^2} \quad \left[\text{Note: } \sum(x_i-\bar{x})=0\right]$$

$$= \frac{1}{S_{xx}}\left[\sum_{i=1}^{n}(x_i-\bar{x})\,y_i\right] \quad \left[\because S_{xx}=\sum_{i=1}^{n}(x_i-\bar{x})^2\right]$$

$$\therefore \hat{\beta}_1 = \sum_{i=1}^{n} c_i y_i \quad \left[\text{where, } c_i = \frac{\sum(x_i-\bar{x})}{S_{xx}}\right]$$

$$\text{Var}(\hat{\beta}_1) = \text{Var}\left(\sum_{i=1}^{n} c_i y_i\right) = \sum_{i=1}^{n} c_i^2\, V(y_i)$$

$$= \sum_{i=1}^{n} c_i^2 \sigma^2$$

Since, the y_i's are independent and $V(y_i)=\sigma^2$ is constant.

$$\text{Var}(\hat{\beta}_1) = \frac{\sigma^2}{S_{xx}} \quad \left[\because \sum c_i^2 = \frac{1}{S_{xx}}\right]$$

From properties of c_i

we find, $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$

$$= \sum_{i=1}^{n} (y_i/n) - \bar{x} \sum_{i=1}^{n} c_i y_i$$

$$= \sum_{i=1}^{n} \left(\frac{1}{n} - \bar{x} c_i\right) y_i$$

$$= \sum_{i=1}^{n} K_i y_i$$

where, $K_i = \frac{1}{n} - \bar{x} c_i$

$$= \frac{1}{n} - \frac{\bar{x}(x_i - \bar{x})}{S_{xx}}$$

$$Var(\hat{\beta}_0) = Var\left(\sum_{i=1}^{n} K_i y_i\right)$$

$$= \sum_{i=1}^{n} K_i^2 \, Var(y_i)$$

$$= \sigma^2 \sum_{i=1}^{n} K_i^2$$

we have, $\sum_{i=1}^{n} K_i^2 = \sum \left[\frac{1}{n} - \frac{\bar{x}(x_i - \bar{x})}{S_{xx}}\right]^2$

$$= \sum \left[\frac{1}{n^2} - 2\frac{1}{n}\frac{\bar{x}(x_i - \bar{x})}{S_{xx}} + (\bar{x})^2 \left(\frac{x_i - \bar{x}}{S_{xx}}\right)^2\right]$$

$$= \sum \left[\frac{1}{n^2} + (\bar{x})^2 \left(\frac{x_i - \bar{x}}{S_{xx}}\right)^2\right]$$

$$\left[\because \sum (x_i - \bar{x}) = 0\right]$$

$$= n \times \frac{1}{n^2} + (\bar{x})^2 \frac{\sum (x_i - \bar{x})^2}{S_{xx}^2}$$

$$= \frac{1}{n} + (\bar{x})^2 \frac{S_{xx}}{S_{xx}^2}$$

$$= \left[\frac{1}{n} + \frac{(\bar{x})^2}{S_{xx}}\right]$$

$$Var(\hat{\beta}_0) = \sigma^2\left[\frac{1}{n} + \frac{(\bar{x})^2}{S_{xx}}\right]$$

Entries of the H (hat) matrix:

We can write (i,j)th element of the hat matrix H as

$$h_{ij} = x_i^T (X^T X)^{-1} x_j$$

earlier we find,

$$(X^T X)^{-1} = \frac{1}{n\sum_i x_i^2 - (\sum_i x_i)^2}\begin{bmatrix} \sum_i x_i^2 & -\sum x_i \\ -\sum x_i & n \end{bmatrix}$$

$$= \begin{bmatrix} \dfrac{\sum_i x_i^2}{n\sum_i x_i^2 - (\sum_i x_i)^2} & -\dfrac{\sum x_i}{n\sum x_i^2 - (\sum_i x_i)^2} \\ \dfrac{-\sum x_i}{n\sum_i x_i^2 - (\sum_i x_i)^2} & \dfrac{n}{n\sum_i x_i^2 - (\sum_i x_i)^2} \end{bmatrix}$$

---------- (1)

we can write this

$$(X^TX)^{-1} = \begin{pmatrix} \frac{1}{n} + \frac{\bar{x}^2}{\sum(x_i-\bar{x})^2} & -\frac{\bar{x}}{\sum(x_i-\bar{x})^2} \\ -\frac{\bar{x}}{\sum(x_i-\bar{x})^2} & \frac{1}{\sum(x_i-\bar{x})^2} \end{pmatrix} = A$$

We find this by resolving each element of matrix in eq (1), Let us proof this, by reverse approach:

element 11 of matrix $(X^TX)^{-1}$

$$= \frac{1}{n} + \frac{\bar{x}^2}{\sum(x_i-\bar{x})^2} = \frac{1}{n} + \frac{(\bar{x})^2}{\sum x_i^2 - 2\bar{x}\sum x_i + n(\bar{x})^2}$$

$$= \frac{1}{n} + \frac{\left(\frac{\sum x_i}{n}\right)^2}{\sum x_i^2 - 2\frac{\sum x_i}{n}\sum x_i + n\left(\frac{\sum x_i}{n}\right)^2}$$

$$= \frac{1}{n} + \frac{\left(\frac{\sum x_i}{n}\right)^2}{\sum x_i^2 - 2\frac{(\sum x_i)^2}{n} + \frac{(\sum x_i)^2}{n}}$$

$$= \frac{1}{n} + \frac{\left(\frac{\sum x_i}{n}\right)^2}{\sum x_i^2 - \frac{(\sum x_i)^2}{n}}$$

$$= \frac{1}{n} + \frac{(\sum x_i)^2}{\cancel{n^2}\, n} \times \frac{\cancel{n}}{n\sum x_i^2 - (\sum x_i)^2}$$

$$= \frac{n\sum x_i^2 - (\sum x_i)^2 + (\sum x_i)^2}{n\left(n\sum x_i^2 - (\sum x_i)^2\right)}$$

This is same as first element of matrix in eq (1) $\longleftarrow$

$$= \frac{\cancel{n}\sum x_i^2}{\cancel{n}\left(n\sum x_i^2 - (\sum x_i)^2\right)}$$

Now coming into second element, (A_{12}), and 3rd element (A_{21})
(both are same)

$$- \frac{\bar{x}}{\sum (x_i - \bar{x})^2}$$

$$= - \frac{\bar{x}}{\sum x_i^2 - 2\bar{x}\sum x_i + n(\bar{x})^2}$$

$$= - \frac{\frac{\sum x_i}{n}}{\sum x_i^2 - 2\frac{\sum x_i}{n}\sum x_i + n\left(\frac{\sum x_i}{n}\right)^2}$$

$$= - \frac{\frac{\sum x_i}{n}}{\frac{n\sum x_i^2 - 2(\sum x_i)^2 + (\sum x_i)^2}{n}}$$

$$= - \frac{\sum x_i}{n} \times \frac{n}{n\sum x_i^2 - (\sum x_i)^2}$$

$$= - \frac{\sum x_i}{n\sum x_i^2 - (\sum x_i)^2}$$

Now the last element (A_{22})

$$\frac{1}{\sum (x_i - \bar{x})^2}$$

$$= \frac{1}{\sum x_i^2 - 2\bar{x}\sum x_i + n(\bar{x})^2}$$

$$= \frac{1}{\sum x_i^2 - 2\frac{\sum x_i}{n}\sum x_i + n\left(\frac{\sum x_i}{n}\right)^2}$$

$$= \frac{1}{\frac{n\sum x_i^2 - 2(\sum x_i)^2 + (\sum x_i)^2}{n}}$$

$$= \frac{n}{n\sum x_i^2 - (\sum x_i)^2}$$

So, we can say that element of both matrix are

Same so,

$$(X^T X)^{-1} = \begin{pmatrix} \frac{1}{n} + \frac{\bar{x}^2}{\sum (x_i - \bar{x})^2} & -\frac{\bar{x}}{\sum (x_i - \bar{x})^2} \\ -\frac{\bar{x}}{\sum (x_i - \bar{x})^2} & \frac{1}{\sum (x_i - \bar{x})^2} \end{pmatrix}$$

So, $h_{ij} = x_i^T (X^T X)^{-1} x_j$

$$= \begin{pmatrix} 1 & x_i \end{pmatrix} \begin{pmatrix} \frac{1}{n} + \frac{\bar{x}^2}{\sum (x_i - \bar{x})^2} & -\frac{\bar{x}}{\sum (x_i - \bar{x})^2} \\ -\frac{\bar{x}}{\sum (x_i - \bar{x})^2} & \frac{1}{\sum (x_i - \bar{x})^2} \end{pmatrix} \times \begin{pmatrix} 1 \\ x_j \end{pmatrix}$$

$$= \left(\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum (x_i - \bar{x})^2} - \frac{\bar{x} x_i}{\sum (x_i - \bar{x})^2} \right) \quad \left(-\frac{\bar{x}}{\sum (x_i - \bar{x})^2} + \frac{x_i}{\sum (x_i - \bar{x})^2} \right) \right) \times \begin{pmatrix} 1 \\ x_j \end{pmatrix}$$

$$= \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum (x_i - \bar{x})^2} - \frac{\bar{x} x_i}{\sum (x_i - \bar{x})^2} - \frac{\bar{x} x_j}{\sum (x_i - \bar{x})^2} + \frac{x_i x_j}{\sum (x_i - \bar{x})^2} \right)$$

$$= \frac{1}{n} + \frac{\bar{x}^2 - \bar{x} x_i - \bar{x} x_j + x_i x_j}{\sum (x_i - \bar{x})^2}$$

$$= \frac{1}{n} + \frac{x_j (x_i - \bar{x}) - \bar{x} (x_i - \bar{x})}{\sum (x_i - \bar{x})^2}$$

So, $h_{ij} = \frac{1}{n} + \frac{(x_i - \bar{x})(x_j - \bar{x})}{\sum (x_i - \bar{x})^2}$ [Answer: showed]

Problem 2:

Solution:

Condition: intercept $\beta_0 = 0$

This is called Regression through the origin.

Here, the regression model:

$y = \beta_1 x$ (According to question & condition)

We will estimate the slope β_1 by least square estimates.

The error $\epsilon_i = y_i - \mu_i$
$= y_i - \beta_1 x_i \quad (i = 1, 2, ---- n)$

Our aim is to minimize the error,

So, $S(\beta_1) = \sum_{i=1}^{n} (y_i - \beta_1 x_i)^2$

Now, $\frac{\partial S(\beta_1)}{\partial \beta_1} = 2 \sum_{i=1}^{n} (y_i - \beta_1 x_i)(-x_i) = 0$

$\Rightarrow \sum_{i=1}^{n} \beta_1 x_i^2 - \sum_{i=1}^{n} x_i y_i = 0$

$\Rightarrow \beta_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i$

$\therefore \hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}$ Answer

[Showed]

problem 3:

Solution:

Given three models

(i) $E(y_i) = \beta_0 + \beta_1 x_{1i}$

(ii) $E(y_i) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}$

(iii) $E(y_i) = \beta_0 + \beta_2 x_{2i}$

where, x_{1i} = right leg strength

x_{2i} = left leg strength.

(a) - plot of residuals for these three models is given in next page.

* Note: In data cleaning step, I convert the response (y) into one unit feet

- plot aganist $\hat{y}$ is also shown in the next page.

Comment on three models:

Using R programming (Code is attached at the end of problem)

(i) For first model - Coefficient $\beta_0 = 14.9093$

$\beta_1 = 0.9027$

and Residual standard error = 16.58 on 11 DF (Degree of freedom)

Also for model (i),

$$\text{Adjusted } R^2 = 0.5926$$

$$\text{Multiple } R^2 = 0.6266$$ (multiple means multiple regression R^2)

$$\text{F Statistics} = 18.46$$

$$\text{p value} = 0.001264$$ (from regression model)

(ii) <u>For Second model:</u>

Coefficient of regression,

$$\beta_0 = 12.7483$$

$$\beta_1 = 0.7213$$

$$\beta_2 = 0.2012$$

And, Residual standard error = 17.25 on 10 DF

$$\text{Multiple } R^2 = 0.6328$$

$$\text{Adjusted } R^2 = 0.5594$$

$$\text{F statistics} = 8.617$$

$$\text{p value} = 0.006675$$

(iii) For third model:

Coefficient of Regression, $\beta_0 = 26.9112$

$\beta_1 = 0.8434$

Residual Standard error = 18.13

Multiple $R^2 = 0.5537$

Adjusted $R^2 = 0.5131$

F statistics = 13.65

P value = 0.003537

Comparison:

From the residual plot, we know that positive value for the residual (on Y axis of Residual vs Fitted values plot) means the prediction was too low, and negative means prediction is too high. 0 means (in the residuals) the guess was exactly correct.

From the above data & residual plot the three model look alike having some outlier. There is no noticeable normality among them. Hence analyzing all the information I will say model 2 is more appropriate having largest R^2 and better fitting.

3 (b) :

We will test the hypothesis for model 2 & model 3 as β_2 in involved in this two model:

Here Null Hypothesis : $H_0 : \beta_2 = 0$

Alternative " $H_1 : \beta_2 = 0$

(i) Model 2 :

From 3(a) we find (using R) [All the R code attach at the end of problem]

$\hat{\beta}_2 = 0.2012$ with

t value = 0.412

and p value = 0.689 > 0.05

So, there is not enough evidence to reject the null hypothesis $\beta_2 = 0$.

Also, we can verify the result using ANOVA Table. (more details in R code)

Source	DF	SS	MS	F	P
$\hat{\beta}_1$	1	5076.7	5076.7	17.0644	0.002041
$\hat{\beta}_2$	1	50.5	50.5	0.1696	0.689115
Residuals	10	2975	297.5		

Also we can see that the confidence interval

for β_2 is : $(-0.8871681, 1.289544)$

we see that this interval contains 0 inside it.

So, we can conclude, from P value, Anova table, and CI that, there is not enough evidence to reject null hypothesis for model 2.

(ii) Model 3 :

For model 3, $\hat{\beta}_2 = 0.8434$

t value $= 3.694$

p value $= 0.00354 < 0.05$

So, we can say that, the null hypothesis is rejected. $\hat{\beta}_2$ is statistically different from 0.

Verifying the result using ANOVA Table.

Source	DF	SS	MS	F	P
$\bar{\beta}_2$	1	4486.3	4486.3	13.648	0.003537
Residuals	11	3615.9	328.7		

Also, the confidence interval for β_2 is $(0.34090, 1.345807)$ and we see that CI doesn't contain 0.

From Anova table (with small p value), CI we can conclude that we reject the null hypothesis as there is not enough evidence to support it.

Final Conclusion:

we fit three model for problem 3(a) and analyze the result (All result are provided) previous, we find in problem (a) model seems most appropriate

And we test the hypothesis for β_2 in problem 3(b) and for model 3 the hypothesis is rejected and for model 2, we donot have enough evidence to reject the null hypothesis.

Problem 4:

Solution:

(a) Multiple Regression Model:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + \beta_5 x_{5i} + \varepsilon_i$$

The residual for this multiple regression is plotted in the next page:

— The value of Residual (given in next page)

— Also, For more easy visualization I provide a Residual Vs No of obs. plot.

Also, we find the coefficient,

$\hat{\beta}_0 = -6.5122$

$\hat{\beta}_1 = 1.9994$

$\hat{\beta}_2 = -3.6751$

$\hat{\beta}_3 = 2.5245$

$\hat{\beta}_4 = 5.1581$

$\hat{\beta}_5 = 14.4012$

(b) 95% Confidence intervals of the mean response

given in the next page, with lower & upper limit

95% prediction interval on a new observation

given in the next page, (for 22 specimen) problem 4 (a)

(i) New multiple regressor using regressor x_2, x_4, x_5:

$$y_i = \beta_0 + \beta_2 x_{2i} + \beta_4 x_{4i} + \beta_5 x_{5i} + \varepsilon_i$$

Using R [Code is attached at the end of problem]

the coefficient of regression:

$$\hat{\beta}_0 = -5.9804$$

$$\hat{\beta}_2 = -2.8898$$

$$\hat{\beta}_4 = 6.8793$$

$$\hat{\beta}_5 = 17.1090$$

s^2: From R code, Residual standard error, $= 0.6783$

$$s^2 = 0.4600908$$

Standard error of predictions:

Standard error, intercept $\hat{\beta}_0 = 0.6591$

Standard error, $\hat{\beta}_2 = 2.3465$

Standard error $\hat{\beta}_4 = 2.7828$

Standard error, $\hat{\beta}_5 = 3.0819$

[R code is given at the end]

95% confidence intervals on the mean response at the regressor location for the 22 specimen:

is shown in the next page.