Homework 3

BANA 7052: Applied Linear Regression

Section 1: Wednesday 6:00 – 9:50

Group 5

Tess Newkold

Aabhaas Sethi

Anirudh Chekuri

James Brand

Question 1

A. What is your final estimated model?

Y = 39.6556 + 0.1662(x1) – 1.7021(x2) + error

B. What is the predicted alumni giving rate for an observation with $(X_1=50, X_2=10)$?
Predicted Giving Rate: 64.99

C. Test the statistical significance of the regression coefficient using *t*-tests; use α=0.05. Obtain the *t*-statistics and *p*-values, interpret the results, make a conclusion (i.e. reject or not reject), and explain why. **Note:** please explain what the null hypothesis is.
Ho: $\beta_1 = 0$
H1: $\beta_1 \neq 0$
AND
Ho: $\beta_2 = 0$
H1: $\beta_2 \neq 0$
t Values
β_1 =1.022 β_2=-3.850
p Values:
β_1=0.312 β_2=0.000371
Since the p-value is less than 0.05 in the student to Faculty ratio, β_2, there is a significant association between the student faculty predictor to the giving rate, therefore we reject the null hypothesis that there is no correlation between them. However, the p-value for β_1, classrooms under 20 students, is greater than 0.05 and not significant in the model. We fail to reject that there is no correlation between the predictor variable and the outcome variable.

D. What is the F statistic? Is it significant? Clearly write out the null hypothesis, F-statistic, and *p*-value and interpret the test results.

Ho: $\beta_1 = \beta_2 = 0$
H1: $\beta_1 \neq \beta_2 \neq 0$
F Value: 28.79
p Value: 8.869^(-9)
Since the p-value is much less than 0.05 it is highly significant, which means that at least one of the predictor variables is significantly related to the outcome variable.

E. What is the value of the coefficient of determination? Please interpret.

R^2 = 0.5613
About 56% of the variation in Y is explained by it's relationship with X1 and X2.

F. What is the correlation coefficient r_1 between X_1 and Y and the correlation coefficient r_2 between X_2 and Y? Do you see any relationship between r_1, r_2, and R^2?
r_1 = 0.6456
r_2 = -0.7423
There is no relationship between r_1, r_2, and R^2

Question 2

- **Generate data with $X \sim N(\mu=2, \sigma=0.1)$, sample size n=100, and error term $\epsilon \sim N(\mu=0, \sigma=0.5)$.**

```
library("investr")
set.seed(7052)
x1 <- rnorm(100,2,0.1)
x2<- rnorm(100, 0,0.4)
y <- rnorm(100,10+5*x1-2*x2,0.5)
model.data<- data.frame(x1,x2,y)
l.model <- lm(y~., data=model.data)
```

- **Fit a simple linear regression to the simulated data from part a. What is the estimated prediction equation? Report the estimated coefficients and their standard errors. Are they significant? Clearly write out the null and alternative hypotheses, observed *t*-statistic(s), *p*-value(s), and interpret the estimates and test results. What is fitted model's MSE?**

n	Error Variance	Intercept	B1	Std. Error (B1)	T value	P value	B2	Std. Error (B2)	T value	P value	MSE	Adj. Rsq
100	0.5	11.8320	4.1160	0.4298	9.577	1.1e-15	-1.8873	0.1294	-14.584	< 2e-16	0.21352	0.7313
$\hat{Y} = 11.8320 + 4.1160X1 - 1.8873X2$												

alpha=0.05

Hypotheses1: Ho: B1=0 ; H1: B1≠0 Hypotheses2: Ho: B2=0 ; H1: B2≠0

Since the p Value for both the hypotheses<0.05, hence, we reject the hypothesis that B1=0 and B2= 0 at alpha=0.05 level. So, all held constant, the predicted value increases by ~4.11 units for every unit increase in the predictor variable B1. And, all held constant, the predicted value decreases by ~1.88 units for every unit increase in the predictor variable B2.

For F test, Hypotheses: Ho:β1= β2=…..= βk ; H1: βj≠0 for at least one j. Ho is rejected as F-statistic: 135.7 , p-value: < 2.2e-16, which is <0.05.

- **Repeat part b), but re-simulate the data and change the error term to $\epsilon \sim N(0, \sigma=1)$**

n	Error Variance	Intercept	B1	Std. Error (B1)	T value	P value	B2	Std. Error (B2)	T value	P value	MSE	Adj. Rsq
100	1	13.66402	3.232016	0.8596	3.760	0.00029	-1.7746	0.2588	-6.857	6.57e-10	0.8540912	0.3494
$\hat{Y} = 13.66402 + 3.232016X1 - 1.7746X2$												

alpha=0.05

Hypotheses1: Ho: B1=0 ; H1: B1≠0

Hypotheses2: Ho: B2=0 ; H1: B2≠0

Since the p Value for both the hypotheses<0.05, hence, we reject the hypothesis that B1=0 and B2= 0 at alpha=0.05 level. So, all held constant, the predicted value increases by ~3.23 units for every unit increase in the predictor variable B1. And, all held constant, the predicted value decreases by ~1.77 units for every unit increase in the predictor variable B2. For F test, Hypotheses: Ho:β1= β2=…..= βk ; H1: βj≠0 for at least one j. Ho is rejected as F-statistic: 27.58, p-value: 3.292e-10, which is <0.05.

- **Repeat parts a)–c) using n=400. What do you conclude?**

n	Error Variance	Intercept	B1	Std. Error (B1)	T Value	P value	B2	Std. Error (B2)	T Value	P value	MSE	Adj. Rsq
400	0.5	10.6982	4.6623	0.2586	18.02	<2e-16	-1.9577	0.0650	-30.08	<2e-16	0.2576	0.75
400	1	11.3965	4.3247	0.5174	8.359	<1.08e-15	-1.9156	0.1302	-14.716	<2e-16	1.030758	0.4113
$\hat{Y} = 11.3965 + 4.3247X1 - 1.9156X2$												

Hypotheses1: Ho: B1=0 ; H1: B1≠0

Hypotheses2: Ho: B2=0 ; H1: B2≠0

The hypotheses tests give the same result as in the above cases as the p Values for the hypotheses<0.05, hence, we reject the hypothesis that B1 and B2 are 0 at alpha=0.05 level. All held constant(individually for B1 and B2), the predicted values increase by ~4.66 and decrease by ~1.95 units for the first case for every unit increase in B1 and B2 values. All held constant(individually for B1 and B2), the predicted values increase by ~4.32 and decrease by ~1.91 units for the second case for every unit increase in B1 and B2 values.

For F test (n=400, error var.=0.5), Hypotheses: Ho:β1= β2=…..= βk ; H1: βj≠0 for at least one j. Ho is rejected as F-statistic: 602.3, p-value: < 2.2e-16, which is <0.05. For F test (n=400, error var.=1), Hypotheses: Ho:β1= β2=…..= βk ; H1: βj≠0 for at least one j. Ho is rejected as F-statistic: 27.58, p-value: 3.292e-10, which is <0.05.

- **What is the effect on the model parameter estimates when error variance gets smaller? What is the effect when sample size gets bigger?**

- When the variance is increased for n =100 and n=400, the std error for B1 and B2 increases (refer tables above). Adjusted Rsq values are decreased as a result of increase in error variance.

- When the sample size is increased (refer table below) for the same error variance, the std. error is decreased.

N	Error Variance	Std. Error (B1)	Std. Error (B2)	MSE
100	0.5	0.429	0.129	0.21352
400	0.5	0.258	0.065	0.2576894

- **What about the MSE from each model?**

Ans. MSE = (∑e^2)/(n-2)

For both the models, the MSE may increase or decrease as it will try to approach the original error value of the model. For our model, the MSE increases from 0.21352 (n=100) to 0.25768 (n=400)

Question3

$$Y = X\beta + \varepsilon$$

a.

$$Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} \qquad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{bmatrix} \qquad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

response coefficients error terms

$$X = \begin{bmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1,p-1} \\ 1 & X_{21} & X_{22} & & X_{2,p-1} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & X_{n1} & X_{n2} & \cdots & X_{n,p-1} \end{bmatrix} \quad n \times p \text{ model matrix}$$

b.

$$\varepsilon \sim N(0_n, \sigma^2 I_n)$$

The errors are normally distributed with mean 0 and constant variance. This is the matrix version of saying the errors are identically and independently Normally distributed with constant variance.
We also assume that there is a linear relationship between the response and predictor variables, and that our predictor variables are not highly correlated with each other.

c.

$$X = \begin{bmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1,p-1} \\ 1 & X_{21} & X_{22} & & X_{2,p-1} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & X_{n1} & X_{n2} & \cdots & X_{n,p-1} \end{bmatrix} \quad n \times p \text{ model matrix}$$

d. $\hat{\beta} = (X^T X)^{-1} X^T Y$ $\Rightarrow$ vector of length p that estimates $\beta_0, \beta_1, \beta_2, \ldots, \beta_{p-1}$

e. $\hat{\beta}$ is unbiasid means $E(\hat{\beta}) = \beta$
The mean of the sampling distribution of $\hat{\beta}$ is the true vector β.