Homework 2

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 the estimated slope? Is it significant at the α=0.05 level? Clearly write out the null and alternative hypotheses, observed *t*-statistic, *p*-value, and interpret the estimate and test results.

Ho: There is no linear relationship between classes under 20 students and alumni donations
H1: There is a linear relationship between classes under 20 students and alumni donations
~OR~
Ho: $\beta_1 = 0$
H1: $\beta_1 \neq 0$

Estimated Slope: 0.6577
T-statistic: 5.734
P-Value: 7.23e(-7)

Since the p-value is less then 0.05 we reject the null hypothesis that there is no correlation, therefore we conclude that there is a significant linear relation.

b. Repeat part a. above using the equivalent *F*-test.

Ho: There is no linear relationship between classes under 20 students and alumni donations
H1: There is a linear relationship between classes under 20 students and alumni donations
~OR~
Ho: $\beta_1 = 0$
H1: $\beta_1 \neq 0$

Estimated Slope: 0.6577
F-statistic: 32.88
P-Value: 7.228e(-7)

Since the p-value is less then 0.05 we reject the null hypothesis that there is no correlation, therefore we conclude that there is a significant linear relation.

c. What is the value of R^2? Please interpret.

$R^2 = 0.4168.$

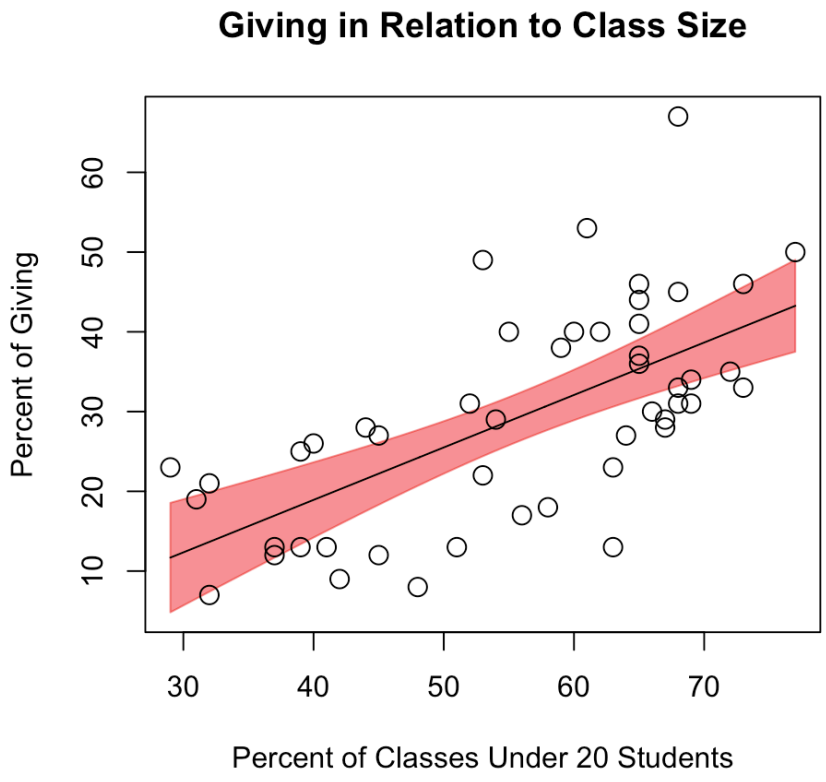
About 42% of the variation in Y is explained by its linear relationship with X.

d. What is the correlation coefficient r between X and Y? What is the relationship between r and R^2?

Correlation coefficient: 0.6456
$R^2 = 0.4168.$
If you square 0.6456 you get R^2
This is only the case in simple linear regression

e. Plot the training data (i.e., the data used to fit the model) with the fitted regression line and include a 95% (pointwise) confidence band for the mean responses. What do you observe about the confidence band at the point $(\bar{X}, \bar{Y})$? Is it narrower or wider compared to the rest?

The confidence band is narrower at the point $(\bar{X}, \bar{Y})$ compared to the rest.



Question 2

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

```
x <- rnorm(100,2,0.1)
y <- rnorm(100,10+5*x,0.5)
model_data<- data.frame(x,y)
```

- **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?**

Sample Size (n)	Error Variance	Intercept	Slope	Std. Error(B1)	MSE	Rsq	T value	P value
100	0.5	9.0218	5.5652	0.4155	0.2032934	0.6468	13.39	<2e-16
$\hat{Y} = 9.0218 + 5.5652\,X$								

alpha=0.05

Hypotheses: Ho: B1=0 ; H1: B1≠0

Since the p Value for the hypotheses<0.05, hence, we reject the hypothesis that slope(B1) is 0 at alpha=0.05 level. So, the predicted value increases by 5.5653 units for every unit increase in the predictor variable.

Mean Square Error: `> sigma(model)^2`

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

Sample Size (n)	Error Variance	Intercept	Slope	Std. Error(B1)	MSE	Rsq	T value	P value
100	1	8.0436	6.1303	0.8309	0.8131737	0.3571	7.7378	5.25e-11
$\hat{Y} = 8.0436 + 6.1303\,X$								

alpha=0.05

Hypotheses: Ho: B1=0 ; H1: B1≠0

Since the p Value for the hypotheses<0.05, hence, we reject the hypothesis that slope(B1) is 0 at alpha=0.05 level. So, the predicted value increases by 6.1303 units for every unit increase in the predictor variable.

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

Sample Size (n)	Error Variance	Intercept	Slope	Std. Error(B1)	MSE	Rsq	T value	P value
400	0.5	9.7466	5.1177	0.2490	0.2388269	0.5149	20.55	<2e-16
$\hat{Y} = 9.7466 + 5.1177\,X$								

Sample Size (n)	Error Variance	Intercept	Slope	Std. Error(B1)	MSE	Rsq	T value	P value
400	1	9.4933	5.2355	0.4979	.9553077	0.2174	10.514	<2e-16
$\hat{Y} = 9.4933 + 5.2355\,X$								

alpha=0.05

Hypotheses: Ho: B1=0 ; H1: B1≠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 slope(B1) is 0 at alpha=0.05 level. The predicted values increase by 5.1177 units for the first case and by 5.2355 units for the second case for every unit increase in the predictor variable.

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

- In all the cases, when the variance is increased for n =100, the std error increases from 0.4155 to 0.8309. The MSE is also increased. A similar trend is observed when n=400. Rsq values are also decreased.
- When the sample size is increased (refer table below) for the same error variance, the std. error is decreased. However, MSE is almost the same and is not much impacted by the sample size.

N	Error Variance	Std. Error	MSE
100	0.5	0.4155	0.2032934
400	0.5	0.2490	0.2388269

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

Ans. MSE = (∑e^2)/(n-2)

For both the models, the MSE is increased when the variance increases as it is directly impacted by the error variance as can be seen by the above formula.

Code: for n = 100, error variance = 0.5

```
library("investr")

set.seed(7052)

x <- rnorm(100,2,0.1)

y <- rnorm(100,10+5*x,0.5)

model_data<- data.frame(x,y)

model <- lm(y~x, data=model_data)

model

coef(model)

summary(model)

plotFit(model, lwd.fit = 2,col.fit = "red2", pch = 19)

# Mean Sq Error:

sigma(model)^2
```

Question 3

3. $E(\hat{\beta}_0) = E(\bar{y} - \hat{\beta}_1 \bar{x}) = E(\bar{y}) - \bar{x}E(\hat{\beta}_1)$

$= E\left(\frac{1}{n}\sum_{i=1}^{n} y_i\right) - \bar{x}\beta_1$

$= \frac{1}{n}\sum_{i=1}^{n} E(y_i) - \bar{x}\beta_1$

$= \frac{1}{n}\sum_{i=1}^{n} \beta_0 + \beta_1 x_i - \bar{x}\beta_1$

$= \frac{1}{n}\left[n\beta_0 + \beta_1 \sum_{i=1}^{n} x_i\right] - \bar{x}\beta_1$

$= \beta_0 + \bar{x}\beta_1 - \bar{x}\beta_1$

$= \beta_0$

Bias of $\hat{\beta}_0 = E(\hat{\beta}_0) - \beta_0 = 0$

$E(\hat{\beta}_1) = E\left(\sum_{i=1}^{n} w_i y_i\right) = \sum_{i=1}^{n} w_i(\beta_0 + \beta_1 x_i) = \beta_0 \sum_{i=1}^{n} w_i + \beta_1 \sum_{i=1}^{n} w_i x_i = \beta_1$

Bias of $\hat{\beta}_1 = E(\hat{\beta}_1) - \beta_1 = 0$

$\mathrm{Var}(\hat{\beta}_0) = \mathrm{Var}(\bar{y} - \hat{\beta}_1 \bar{x}) = \mathrm{Var}(\bar{y}) + (\bar{x})^2 \mathrm{Var}(\hat{\beta}_1) - 2\bar{x}\,\mathrm{Cov}(\bar{y}, \hat{\beta}_1)$ (Cov term $= 0$)

$= \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} y_i\right) + \left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)^2 \left(\frac{\sigma^2}{\sum (x_i - \bar{x})^2}\right)$

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

$= \sigma^2\left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right]$

$\mathrm{Var}(\hat{\beta}_1) = \mathrm{Var}\left(\sum_{i=1}^{n} w_i y_i\right) = \sum_{i=1}^{n} w_i^2 \mathrm{Var}(y_i) = \sigma^2 \sum_{i=1}^{n} w_i^2 = \sigma^2 / S_{xx} = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$

b. As the sample size n increases, $S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2$ will increase, so the variance of $\hat{\beta}_1$ will decrease toward zero. The variance of $\hat{\beta}_0$ will also decrease toward zero as n increases because of the $\frac{1}{n}$ and $\frac{1}{S_{xx}}$ terms.

As the error variance σ^2 increases, the variance of $\hat{\beta}_0$ and the variance of $\hat{\beta}_1$ will both increase.

c. $E(MSE) = E\left(\frac{SSE}{n-2}\right) = \frac{1}{n-2}E\left(\sum_{i=1}^{n} e_i^2\right) = \frac{\sigma^2(n-2)}{n-2} = \sigma^2$

So the model's MSE is unbiased.

The ML estimate of σ^2, $\hat{\sigma}^2 = \frac{1}{n}\sum_{n=1}^{n} e_i^2$, so $E(\hat{\sigma}^2_{ML}) = \frac{\sigma^2(n-2)}{n}$

$\text{Bias}(\hat{\sigma}^2_{ML}) = E(\hat{\sigma}^2_{ML}) - \sigma^2 = \sigma^2\left(\frac{n-2}{n}\right) - \sigma^2 = \frac{-2\sigma^2}{n}$

The ML estimate is biased.

The difference between the two estimates is the denominator. We use the model's MSE because it is unbiased. It uses $n-p$, where p is the number of estimated coefficients, as the degrees of freedom for SSE.