1. Question1

(a) Obtain and interpret the coefficient of determination R^2.

```
#Q1
GMAT = c(560,540,520,580,520,620,660,630,550,550,600,537)
GPA = c(3.20,3.44,3.70,3.10,3.00,4.00,3.38,3.83,2.67,2.75,2.33,3.75)

lm.1 = lm(GPA~GMAT)
summary(lm.1)
coef(lm.1)
```

```
> summary(lm.1)

Call:
lm(formula = GPA ~ GMAT)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.98608 -0.25048 -0.04539  0.47659  0.64531 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 2.157611   2.014430   1.071    0.309
GMAT        0.001931   0.003510   0.550    0.594

Residual standard error: 0.5326 on 10 degrees of freedom
Multiple R-squared:  0.02937,	Adjusted R-squared:  -0.06769 
F-statistic: 0.3026 on 1 and 10 DF,  p-value: 0.5943
```

R^2 is 0.02937.

R^2 is a statistical measure of how close the data are to the fitted regression line. It is also known as the coefficient of determination.

In this question, R^2 equals 2.9% means that the model explains 2.9% of the variability of the response data around its mean, which is not so good.

(b) Calculate the fitted value for the second person.

```
> predict(lm.1, data.frame(GMAT=540))
       1 
3.200232 
```

The fitted value for the second person is 3.20.

(c) Test whether GMAT is an important predictor variable (use significant level 0.05).

As we can see, p-value for GMAT is 0.594, which is far higher than the significance level 0.05. So GMAT is not an important predictor variable.

2. **Question2**

(a) **Which answer is correct, and why?**

iii is correct.

Y = 50 + 20*X1 + 0.07*X2 + 35*X3 + 0.01*X4 - 10*X5

When IQ and GPA are fixed, X1, X2 and X4 are fixed. When GPA is high enough, and the only variable, Gender, equals 0, Y gets higher. So for a fixed value of IQ and GPA, males earn more on average than females provided that the GPA is high enough.

(b) **Predict the salary of female with IQ of 110 and a GPA of 4.0.**

```
> Salary = 50 + 20*GPA + 0.07*IQ + 35*Gender + 0.01*GPA*IQ - 10*GPA*Gender
> Salary
[1] 137.1
```

The salary of female with IQ of 110 and a GPA of 4.0 will be 137.1 (in thousands).

(c) **True or false: Since the coefficient for the GPA/IQ interaction term is very small, there is very little evidence of an interaction effect. Justify your answer.**

False. A small coefficient doesn't even mean the interaction effect is small, since it is very sensitive to the units of the two variables.

3. **Question3**

(a) **Using the rnorm() function, create a vector, X, containing 100 observations drawn from a N(0, 1) distribution. This represents a feature, X.**

(b) **Using the rnorm() function, create a vector, e, containing 100 observations drawn from a N(0, 0.25) distribution i.e. a normal distribution with mean zero and variance 0.25.**

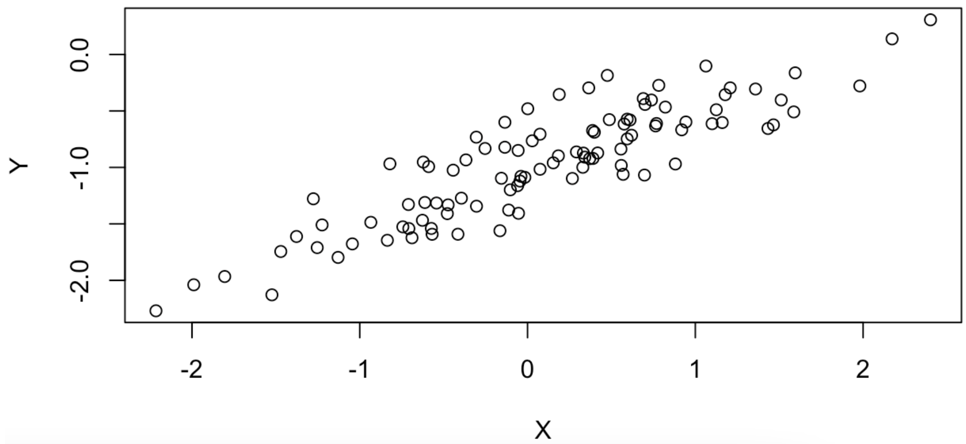
```
#Q3
set.seed(1)
X = rnorm(100, mean = 0, sd = 1)
e = rnorm(100, mean = 0, sd = 0.25)
Y = -1 + 0.5*x +e
```

(c) **Using X and e, generate a vector y according to the model Y = -1 + 0.5*X + e**

What is the length of the vector y? What are the values of β_0 and β_1 in this linear model?

The length of the vector Y is 100. β_0 is -1, β_1 is 0.5.

(d) Create a scatterplot displaying the relationship between x and y. Comment on what you observe.



Each students have different results but the graph approximately shows linear relationship.

(e) Fit a least squares linear model to predict y using x. Comment on the model obtained. How do $\widehat{\beta_0}$ and $\widehat{\beta_1}$ compare to β_0 and β_1

```
lm.3 = lm(Y~X)
summary(lm.3)
confint(lm.3)
```

```
> summary(lm.3)

Call:
lm(formula = Y ~ X)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.46921 -0.15344 -0.03487  0.13485  0.58654 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -1.00942    0.02425  -41.63   <2e-16 ***
X            0.49973    0.02693   18.56   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2407 on 98 degrees of freedom
Multiple R-squared:  0.7784,    Adjusted R-squared:  0.7762 
F-statistic: 344.3 on 1 and 98 DF,  p-value: < 2.2e-16
```

$\widehat{\beta_0}$ is -1.00942, $\widehat{\beta_1}$ is 0.49973. Both of them are very close to β_0 and β_1.

(f) Now fit a polynomial regression model that predicts y using x and x 2 . Is there evidence that the quadratic term improves the model fit? Explain your answer.

```
X_2 = X*X
lm.3.1 = lm(Y~X+X_2)
summary(lm.3.1)
```

```
> summary(lm.3.1)

Call:
lm(formula = Y ~ X + X_2)

Residuals:
    Min      1Q  Median      3Q     Max 
-0.4913 -0.1563 -0.0322  0.1451  0.5675 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -0.98582    0.02941 -33.516   <2e-16 ***
X            0.50429    0.02700  18.680   <2e-16 ***
X_2         -0.02973    0.02119  -1.403    0.164    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2395 on 97 degrees of freedom
Multiple R-squared:  0.7828,	Adjusted R-squared:  0.7784 
F-statistic: 174.8 on 2 and 97 DF,  p-value: < 2.2e-16
```

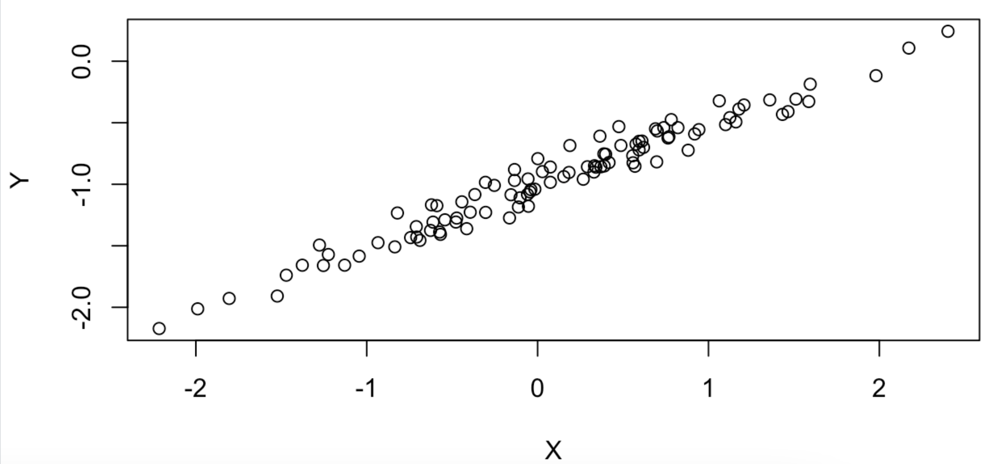
The quadratic term doesn’t improve the model fit because R^2 approximately keeps the same. X_2 is not a significant factor because the p-value for X_2 is 0.164, which is larger than 0.05.

(g) Repeat (a)-(f) after modifying the data generation process in such a way that there is less noise in the data. The model (1) should remain the same. You can do this by decreasing the variance of the normal distribution used to generate the error term in (b). Describe your results.

Let e = N(0, 0.1).

(g-a) to (g-c) remains the same.

(g-d):



(g-e):

```
> summary(lm.3)

Call:
lm(formula = Y ~ X)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.18768 -0.06138 -0.01395  0.05394  0.23462 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -1.003769   0.009699  -103.5   <2e-16 ***
X            0.499894   0.010773    46.4   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.09628 on 98 degrees of freedom
Multiple R-squared:  0.9565,	Adjusted R-squared:  0.956 
F-statistic:  2153 on 1 and 98 DF,  p-value: < 2.2e-16
```

$\widehat{\beta_0}$ is -1.003769, $\widehat{\beta_1}$ is 0.499894. Both of them are very close to β_0 and β_1.

(g-f):

```
> lm.3.1 = lm(Y~X+X_2)
> summary(lm.3.1)

Call:
lm(formula = Y ~ X + X_2)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.19650 -0.06254 -0.01288  0.05803  0.22700 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -0.994328   0.011766 -84.512   <2e-16 ***
X            0.501716   0.010798  46.463   <2e-16 ***
X_2         -0.011892   0.008477  -1.403    0.164    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.0958 on 97 degrees of freedom
Multiple R-squared:  0.9573,	Adjusted R-squared:  0.9565 
F-statistic:  1088 on 2 and 97 DF,  p-value: < 2.2e-16
```

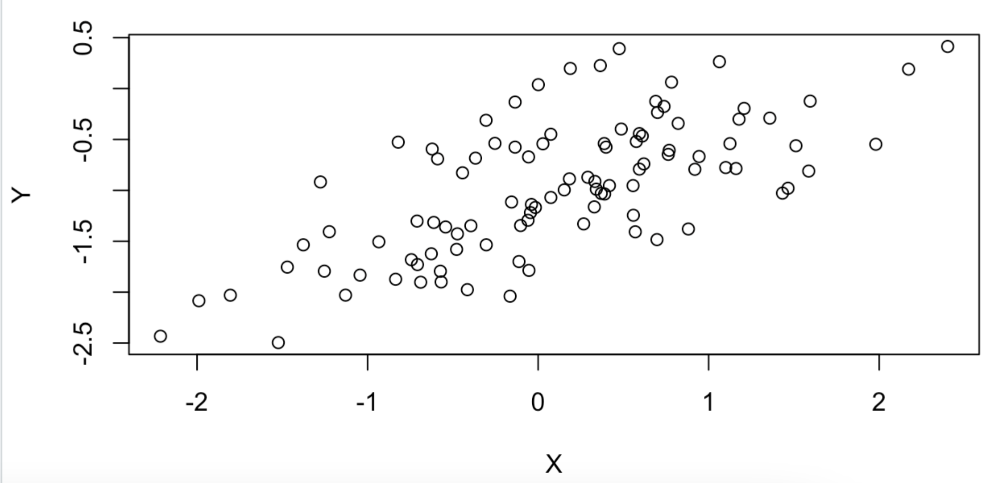
The quadratic term doesn't improve the model fit because R^2 approximately keeps the same. X_2 is not a significant factor because the p-value for X_2 is 0.164, which is larger than 0.05.

(h) Repeat (a)-(f) after modifying the data generation process in such a way that there is more noise in the data. The model (1) should remain the same. You can do this by increasing the variance of the normal distribution used to generate the error term in (b). Describe your results.

Let e = N(0, 0.5).

(h-a) to (g-c) remains the same.

(h-d):



(h-e):

```
> summary(lm.3)

Call:
lm(formula = Y ~ X)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.93842 -0.30688 -0.06975  0.26970  1.17309 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -1.01885    0.04849 -21.010  < 2e-16 ***
X            0.49947    0.05386   9.273 4.58e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4814 on 98 degrees of freedom
Multiple R-squared:  0.4674,	Adjusted R-squared:  0.4619 
F-statistic: 85.99 on 1 and 98 DF,  p-value: 4.583e-15
```

$\widehat{\beta_0}$ is -1.01885, $\widehat{\beta_1}$ is 0.49947. Both of them are very close to β_0 and β_1.

(h-f):

```
> summary(lm.3.1)

Call:
lm(formula = Y ~ X + X_2)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.98252 -0.31270 -0.06441  0.29014  1.13500 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -0.97164    0.05883 -16.517  < 2e-16 ***
X            0.50858    0.05399   9.420  2.4e-15 ***
X_2         -0.05946    0.04238  -1.403    0.164    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.479 on 97 degrees of freedom
Multiple R-squared:  0.4779,	Adjusted R-squared:  0.4672 
F-statistic:  44.4 on 2 and 97 DF,  p-value: 2.038e-14
```

The quadratic term doesn't improve the model fit because R^2 approximately keeps the same. X_2 is not a significant factor because the p-value for X_2 is 0.164, which is larger than 0.05.

(i) What are the confidence intervals for β0 and β1 based on the original data set, the noisier data set, and the less noisy data set? Comment on your results.

Original data set:

β_0's 95% confidence interval: [-1.0575402, -0.9613061]

β_1's 95% confidence interval: [0.4462897, 0.5531801]

```
> confint(lm.3)
                 2.5 %     97.5 %
(Intercept) -1.0575402 -0.9613061
X            0.4462897  0.5531801
```

Less noisy data set:

β_0's 95% confidence interval: [-1.0230161, -0.9845224]

β_1's 95% confidence interval: [0.4785159, 0.5212720]

```
> confint(lm.3)
                 2.5 %     97.5 %
(Intercept) -1.0230161 -0.9845224
X            0.4785159  0.5212720
```

Noisier data set:

β_0's 95% confidence interval: [-1.1150804, -0.9226122]

β_1's 95% confidence interval: [0.3925794, 0.6063602]

```
> confint(lm.3)
                 2.5 %     97.5 %
(Intercept) -1.1150804 -0.9226122
X            0.3925794  0.6063602
```

The noisier data set causes a wider confidence interval, and the less noisy data set causes a narrower confidence interval.