ACT04 - Data Mining

Team 7 | Captain: Nadine Rose | Members: Kyle Scott, Lakshya Rathore, Bailey LaRea

February 21, 2022

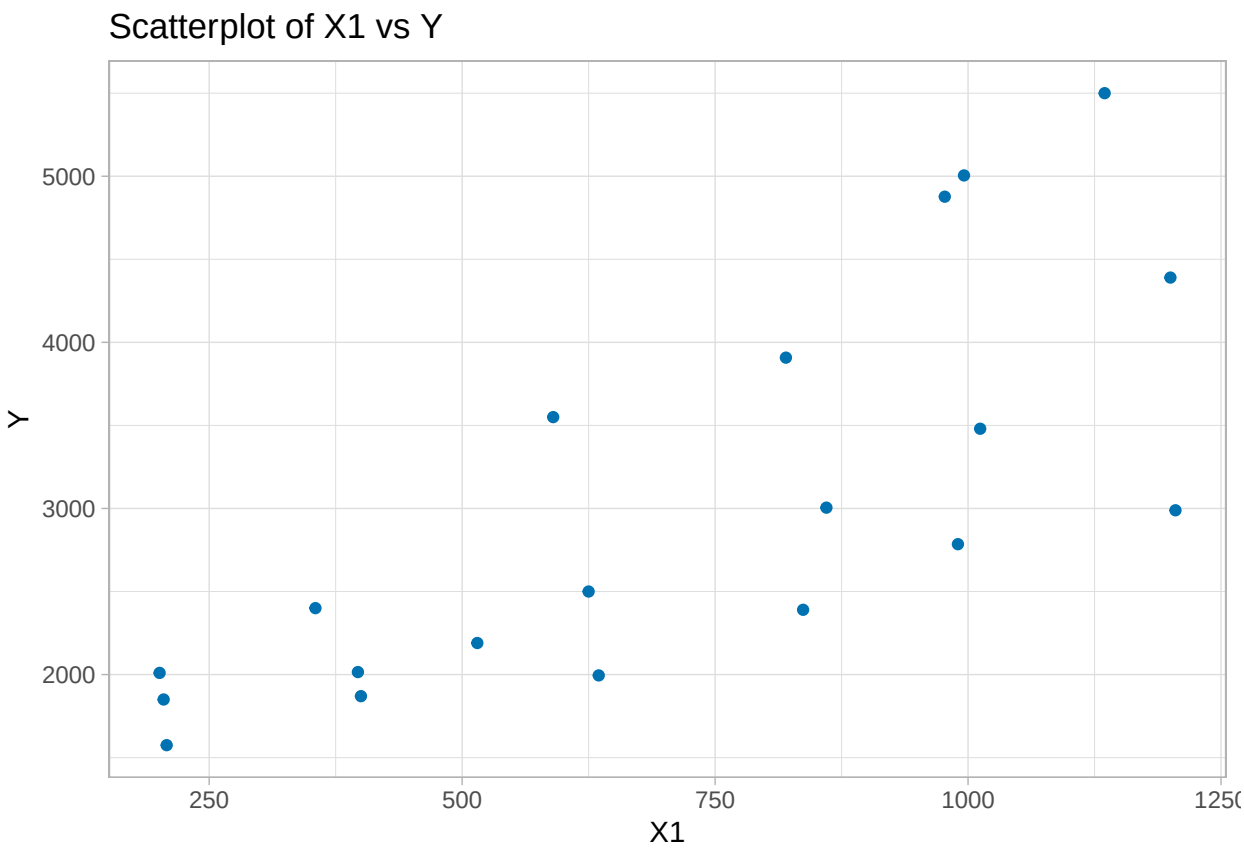
Problem 1

1. Read in ACT04 dataset

```
library(readr)

act04 <- read_csv("ACT_04_Data.csv")
```

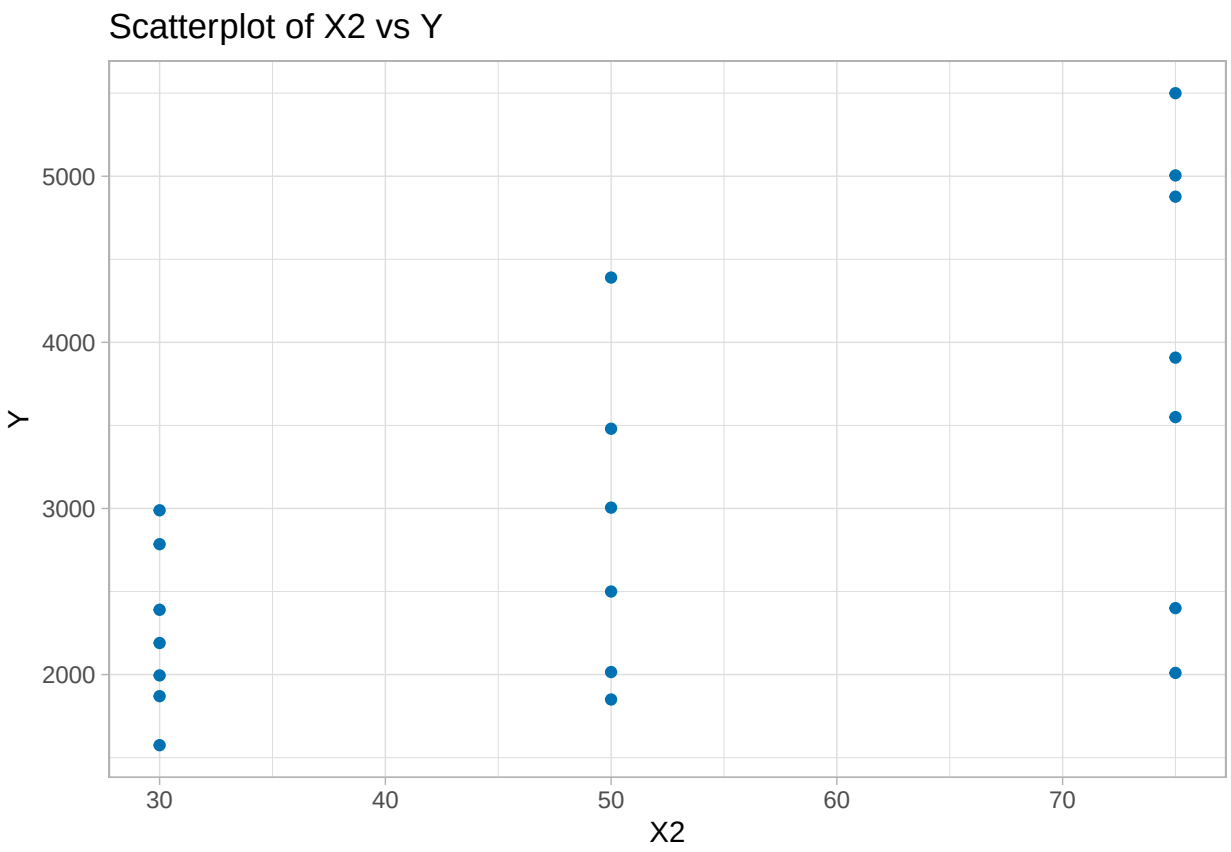
2. Produce a scatterplot of Y and X1

```
library(ggplot2)


# create ggplot object

plt1 <- ggplot(data = act04, aes(x  = X1, y = Y)) +

          geom_point(color = '#0072B2') +

          ggtitle("Scatterplot of X1 vs Y") +

          theme_light()


# print ggplot

plt1
```



3. Produce a scatterplot of Y and X2

```
library(ggplot2)


# create ggplot object
plt2 <- ggplot(data = act04, aes(x  = X2, y = Y)) +
          geom_point(color = '#0072B2') +
          ggtitle("Scatterplot of X2 vs Y") +
          theme_light()


# print ggplot
plt2
```



4. Build a Regression Model with 2 predictors (X1 and X2)

```
Model1 <- lm(Y ~ X1 + X2, data = act04)

m1sum <- summary(Model1)

mse1 <- format(mean(m1sum$residuals^2), nsmall = 3)

rsq1 <- format(m1sum$r.squared, nsmall = 3)
```

5. Build a Regression Model with 3 predictors (X1, X2 and X12)

```
Model2 <- lm(Y ~ X1 + X2 + X12, data = act04)

m2sum <- summary(Model2)

mse2 <- format(mean(m2sum$residuals^2), nsmall = 3)
```

```
rsq2 <- format(m2sum$r.squared, nsmall = 3)
```

6. Build a Regression Model with all 5 predictors

```
Model3 <- lm(Y ~ ., data = act04)

m3sum <- summary(Model3)

mse3 <- format(mean(m3sum$residuals^2), nsmall = 3)

rsq3 <- format(m3sum$r.squared, nsmall = 3)
```

7. Build a Regression Model with 4 predictors (X1, X2, X12 and X1SQ)

```
Model4 <- lm(Y ~ . - X2SQ, data = act04)

m4sum <- summary(Model4)

mse4 <- format(mean(m4sum$residuals^2), nsmall = 3)

rsq4 <- m4sum$r.squared
```

Problem 1 Final Table

Model	R^2	MSE
I	0.8997396	132110.792
II	0.9782077	28715.138
III	0.98615	18249.860
IV	0.9833072	21995.735

Problem 2

1.

```
qnorm(.0125)
```

```
## [1] -2.241403
```

```
qnorm(.99)
```

```
## [1] 2.326348
```

2.

```
qt(0.0125, 333)
```

```
## [1] -2.251584
```

```
qt(.99, 345)
```

```
## [1] 2.337205
```

3.

```
qchisq(.025, 125)
```

```
## [1] 95.94573
```

```
qchisq(.975, 245)
```

```
## [1] 290.2478
```

4.

```
qf(.01, 12, 250)
```

```
## [1] 0.2937983
```

```
qf(.99, 24, 500)
```

```
## [1] 1.828539
```

Problem 3 - Textbook

3.

(a) We know, X_1 = GPA, X_2 = IQ, X_3 = Level (1 for College and 0 for High School), X_4 = Interaction between GPA and IQ, and X_5 = Interaction between GPA and Level.

Equation:

$$y = 50 + 20X_1 + 0.07X_2 + 35X_3 + 0.01X_4 - 10X_5 \tag{1}$$

Since it is given that X_1(GPA) and X_2(IQ) are constants so our first three terms and the interaction of IQ and GPA i.e. X_4 would be treated as constants.

The equation then becomes:

$$y = constant + 35X_3 - 10X_5 \tag{2}$$

For high school graduates the value of $X_3 = 0$ and for college graduates $X_3 = 1$. So, for high school graduates(Level = 0) the equation becomes:

$$y1 = c + 35 * 0 - 10 * GPA * 0 \tag{3}$$

and for college graduates (Level = 1) the equation becomes:

$$y2 = c + 35 * 1 - 10 * GPA * 1 \tag{4}$$

We see that the value of y2 decreases with increase in GPA and becomes less than y1 if the GPA>3.5. So, the correct answer is: option iii: For a fixed value of IQ and GPA, high school graduates earn more, on average, than college graduates provided that the GPA is high enough.

(b)

$$y = 50 + 20X_1 + 0.07X_2 + 35X_3 + 0.01X_4 - 10X_5 \tag{5}$$

($X_3 = 1$ for college graduate), IQ(X_2)=110, GPA(X_1)=4.0

$$y = 50 + 20 * 4 + 0.07 * 110 + 35 * 1 + 0.01(110 * 4) - 10(4 * 1) = 137.1 \tag{6}$$

(c) False. We must evaluate the p-value of the regression coefficient to determine if there is a statistical significance of the interaction effect.

4.

(a) The cubic model will have more freedom to fit the training data closely than the linear model, despite the knowledge of linearity in the data. Therefore, the training data will have a lower RSS in the cubic model than in the linear model.

(b) Once we create our model, even though the cubic model will accurately follow the trend of the training data, the model is prone to overfitting whereas the linear model is not. This means that the RSS for the testing data will be lower in the linear model because of the overfitting in the cubic model and the knowledge that our data is actually linear.

(c) Even though we now know that the data is not linear, the cubic model will always have a lower RSS within the training data due to the additional freedom in the fitting.

(d) Because we don't know enough about the fit of the data (not knowing how linear), there is no way to tell which model will produce a lower RSS for the testing data. If the data were still close to linear, we would be able to say that the linear model would produce a lower RSS, but we can't say this due to our lack of knowledge on the fit.

5.

Based on our two given equations, we can substitute for $\hat{\beta}$:

$$\hat{y}_i = x_i \times \frac{\sum_{i'=1}^{n} (x_{i'} y_{i'})}{\sum_{j=1}^{n} x_j^2}$$

$$\hat{y}_i = \sum_{i'=1}^{n} \frac{(x_{i'} y_{i'}) \times x_i}{\sum_{j=1}^{n} x_j^2}$$

$$\hat{y}_i = \sum_{i'=1}^{n} \left(\frac{x_i x_{i'}}{\sum_{j=1}^{n} x_j^2} \times y_{i'} \right)$$

Knowing that $\hat{y_i} = \sum_{i'=1}^{n} a_{i'} y_{i'}$, we can use the previous equation to see that:

$$a_{i'} = \frac{x_i x_{i'}}{\sum_{j=1}^{n} x_j^2}$$

6.

To begin with, we know the equation for the least squares regression line is as follows:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

Equation (3.4) from the book also gives us the following:

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

If we substitute $\hat{\beta}_0$ from equation (3.4) into our regression line, we get the following:

$$\hat{y} = (\bar{y} - \hat{\beta}_1 \bar{x}) + \hat{\beta}_1 x$$

Now, for the point $(\bar{x}, \bar{y})$ to pass through the line, we know that it must be the case that $x = \bar{x}$ and $\hat{y} = \bar{y}$. Therefore, by substituting again, we get:

$$\bar{y} = \bar{y} - \hat{\beta}_1 x + \hat{\beta}_1 x$$
$$\bar{y} = \bar{y}$$
$$0 = 0$$

7.

We are told we can assume that $\bar{x} = \bar{y} = 0$. Therefore:

$$RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$
$$TSS = \sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2$$

$$R^2 = 1 - RSS/TSS$$
$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i)^2}$$

We know $\hat{y}_i = \hat{\beta}_1 x_i$, therefore we can begin to rewrite as follows:

$$R^2 = 1 - \frac{\sum(y_i - (\sum x_i y_i / \sum x_i^2) x_i)^2}{\sum(y_i)^2}$$
$$R^2 = \frac{2(\sum x_i y_i)^2 / \sum x_i^2 - \sum(x_i y_i)^2 / \sum x_i^2}{\sum y_i^2}$$

Remembering that $\bar{x} = \bar{y} = 0$ and $Cor(X,Y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$ we can see that the following is an equilvalency:

$$R^2 = Cor(X,Y)^2 = \frac{\sum_{i=1}^{n}(x_i y_i)^2}{\sum_{i=1}^{n}(x_i)^2 \sum_{i=1}^{n}(y_i)^2}$$