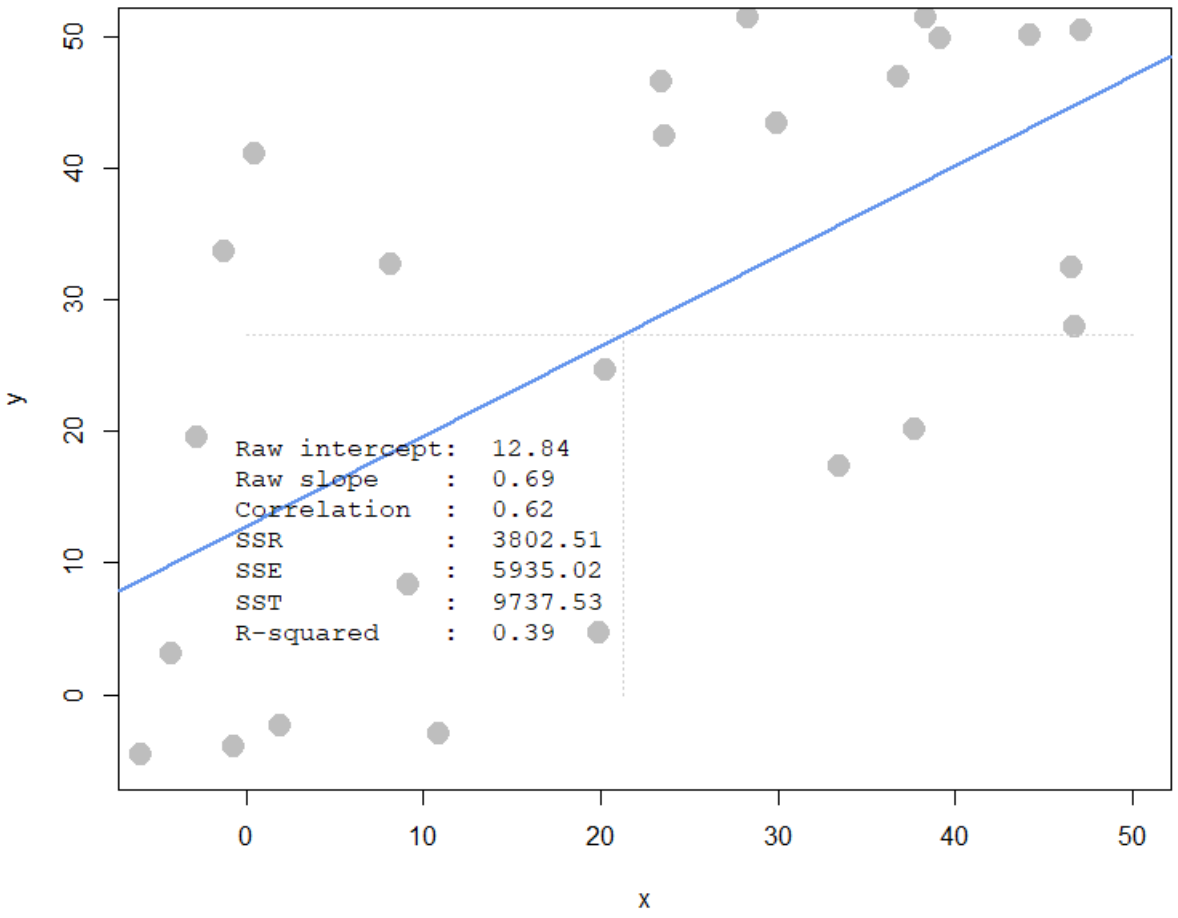
BACS HW11

106070020

2021年5月8日

Question 1

(a)(i)



Scenario 2

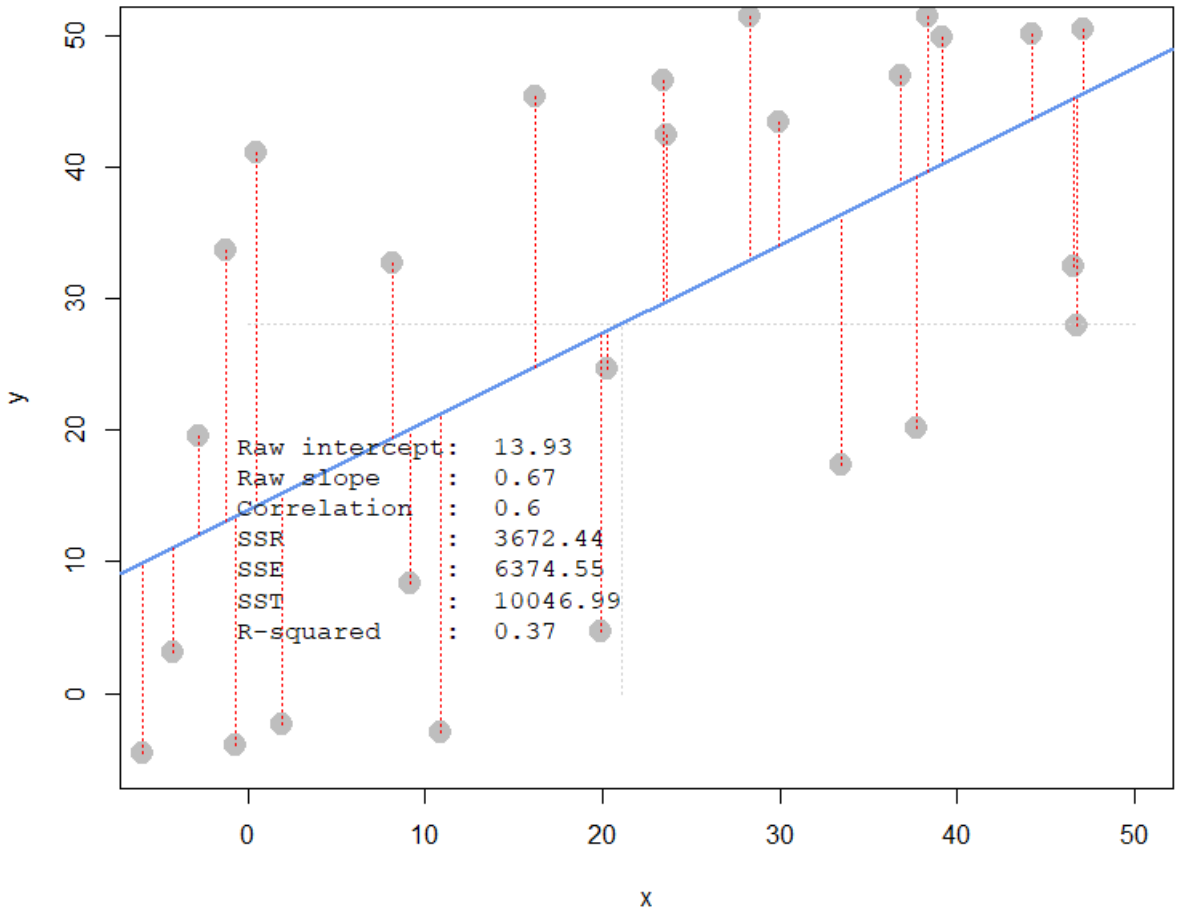
(a)(ii)

```
#(a)(ii)
pts<-read.csv("C:/Users/eva/Desktop/作業 上課資料(清大)/大四下/BACS/HW11 BACS/pts.csv")
regr <- lm(y ~ x, data=pts)
summary(regr)
```

```
##
## Call:
## lm(formula = y ~ x, data = pts)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -24.242 -16.699   5.673  12.398  26.888
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  13.9295     4.9725   2.801  0.00990 **
## x             0.6726     0.1809   3.718  0.00107 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16.3 on 24 degrees of freedom
## Multiple R-squared:  0.3655, Adjusted R-squared:  0.3391
## F-statistic: 13.83 on 1 and 24 DF,  p-value: 0.00107
```

(a)(iii)

```
#(a)(iii)
y_hat <- regr$fitted.values
# segments(pts$x, pts$y, pts$x, y_hat, col="red", lty="dotted")
```



segments

(a)(iv)

```
#(a)(iv)
SSE<-sum((pts$y-y_hat)^2)
SSE
```

```
## [1] 6374.554
```

```
SSR<-sum((y_hat-mean(pts$y))^2)
SSR
```

```
## [1] 3672.44
```

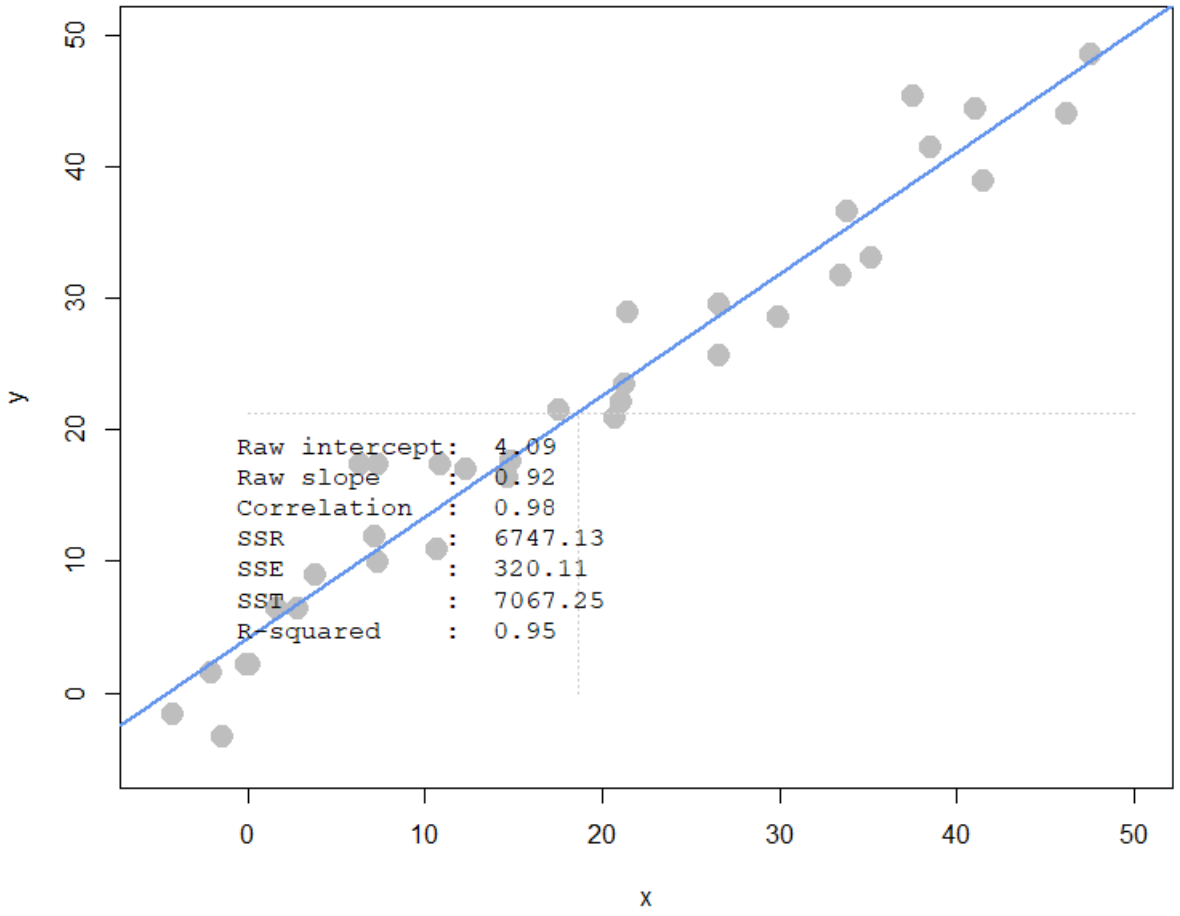
```
SST<-SSE+SSR
SST
```

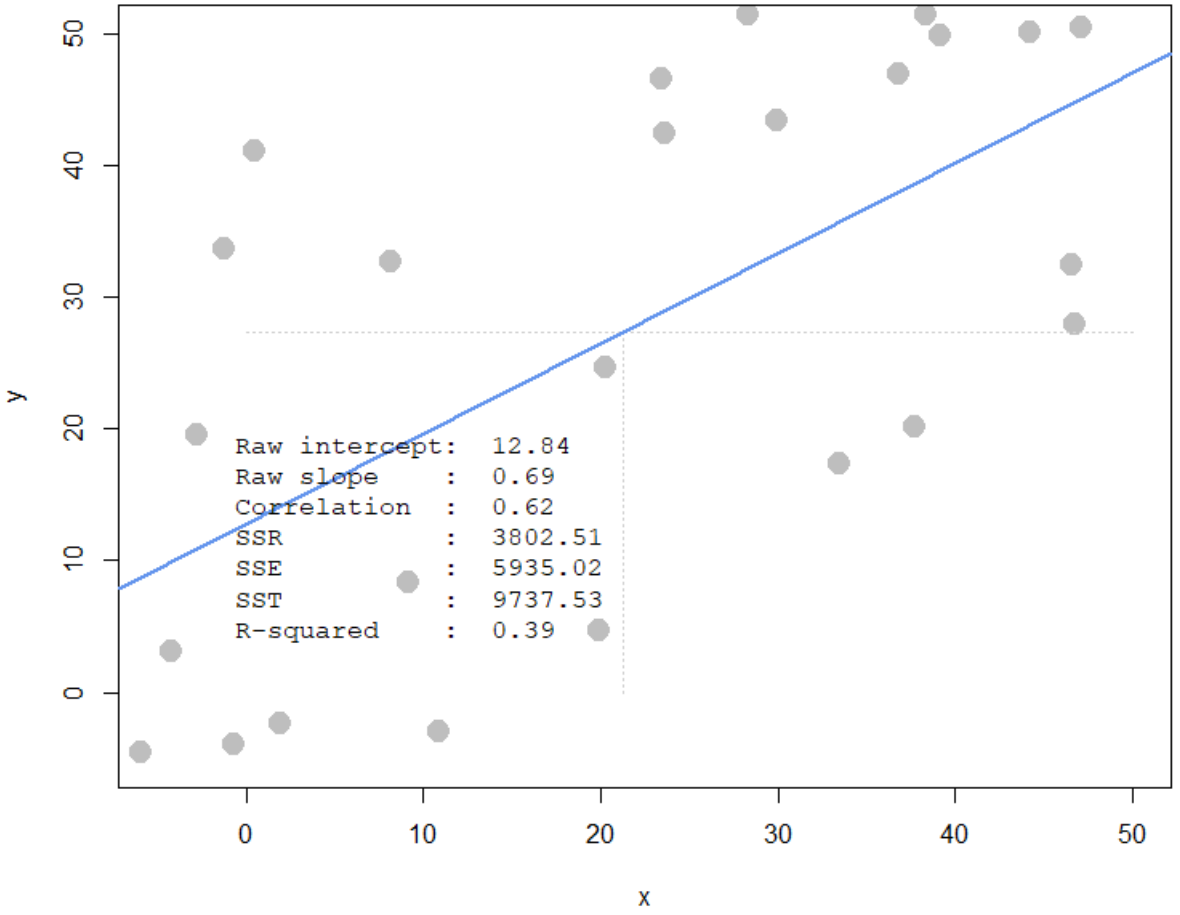
```
## [1] 10046.99
```

```
R_sqrt<-SSR/SST
R_sqrt
```

```
## [1] 0.3655263
```

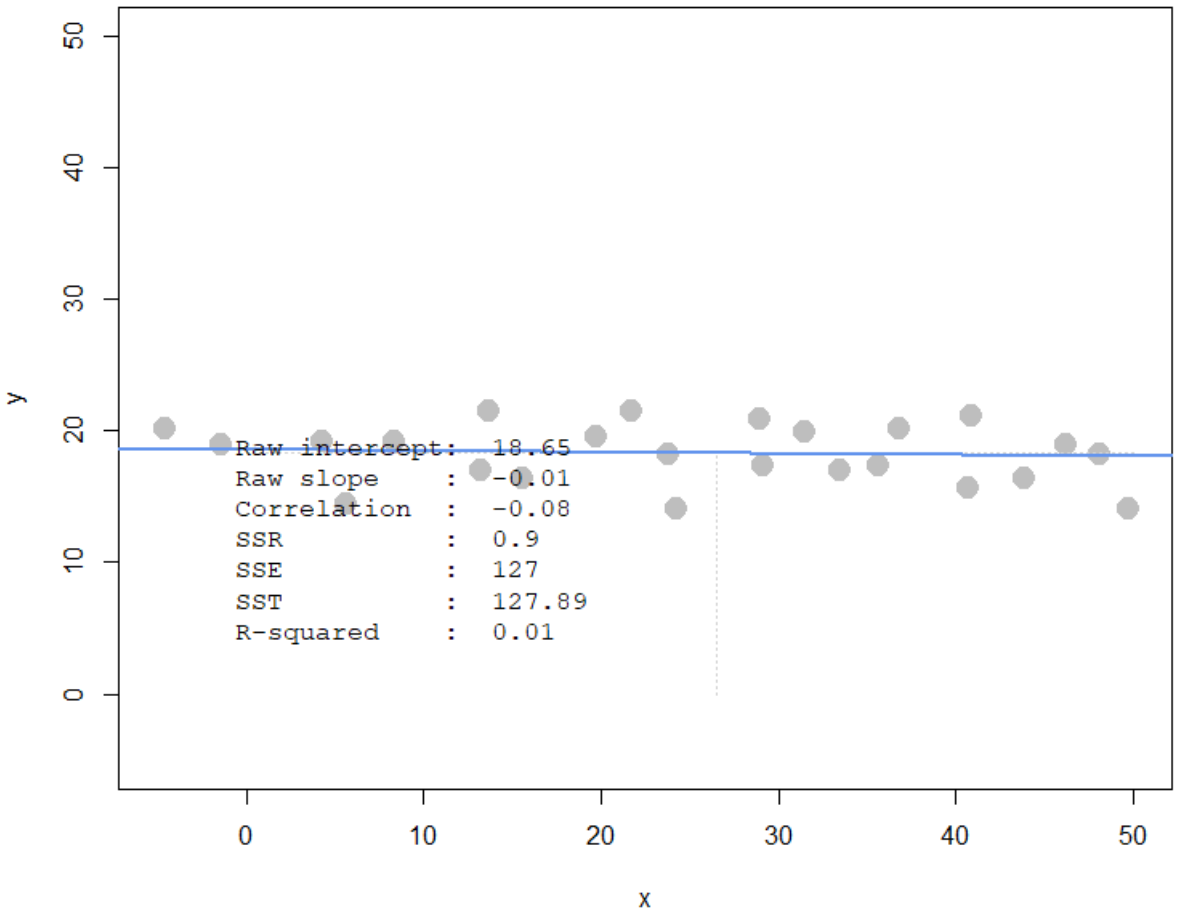
(b) Comparing scenarios 1 and 2, which do we expect to have a stronger R2

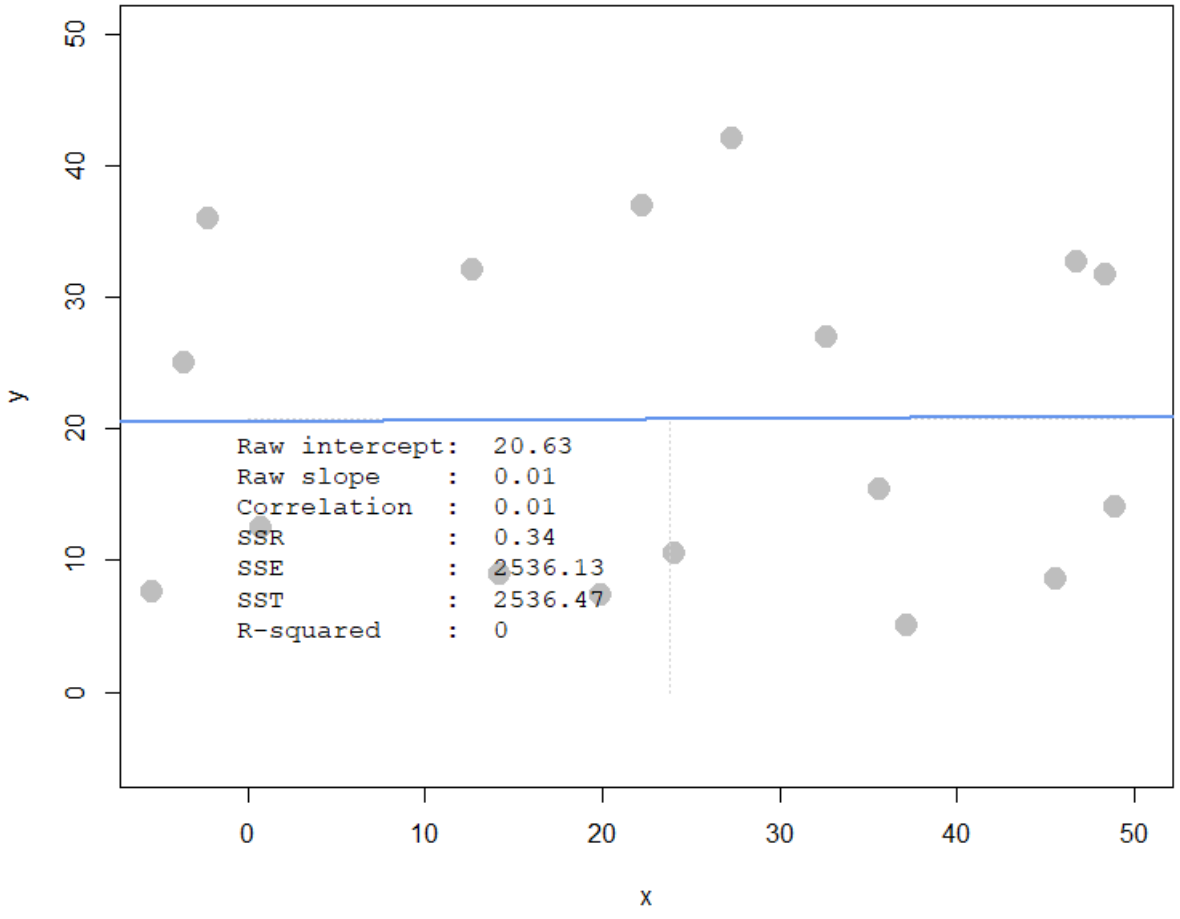

Raw intercept: 4.09
Raw slope : 0.92
Correlation : 0.98
SSR : 6747.13
SSE : 320.11
SST : 7067.25
R-squared : 0.95
x
y


Raw intercept: 12.84
Raw slope : 0.69
Correlation : 0.62
SSR : 3802.51
SSE : 5935.02
SST : 9737.53
R-squared : 0.39
x
y

> For the simple linear regression, the R^2 is the square of the correlation coefficient. By comparing scenarios 1 and 2, we should expect scenarios 1 to have a stronger R2, as scenarios 1 has larger correlation coefficient.

(c)Comparing scenarios 3 and 4, which do we expect to have a stronger R2 ?


Raw intercept: 18.65
Raw slope : -0.01
Correlation : -0.08
SSR : 0.9
SSE : 127
SST : 127.89
R-squared : 0.01
y
x
0
10
20
30
40
50


Raw intercept: 20.63
Raw slope : 0.01
Correlation : 0.01
SSR : 0.34
SSE : 2536.13
SST : 2536.47
R-squared : 0
y
x
0
10
20
30
40
50

> For the simple linear regression, the R^2 is the square of the correlation coefficient. By comparing scenarios 3 and 4, we should expect they have similar R2, as scenarios they have similar correlation correlation coefficient, which are both close to 0.

(d)Comparing scenarios 1 and 2, which do we expect has bigger/smaller SSE, SSR, and SST?

Scenerios 1 will have smaller SSE, as the data points are close to the regression line. Scenerios 2 will have less SSR. scenarios 1 and 2 will have similar SST.

(e)Comparing scenarios 3 and 4, which do we expect has bigger/smaller SSE, SSR, and SST?

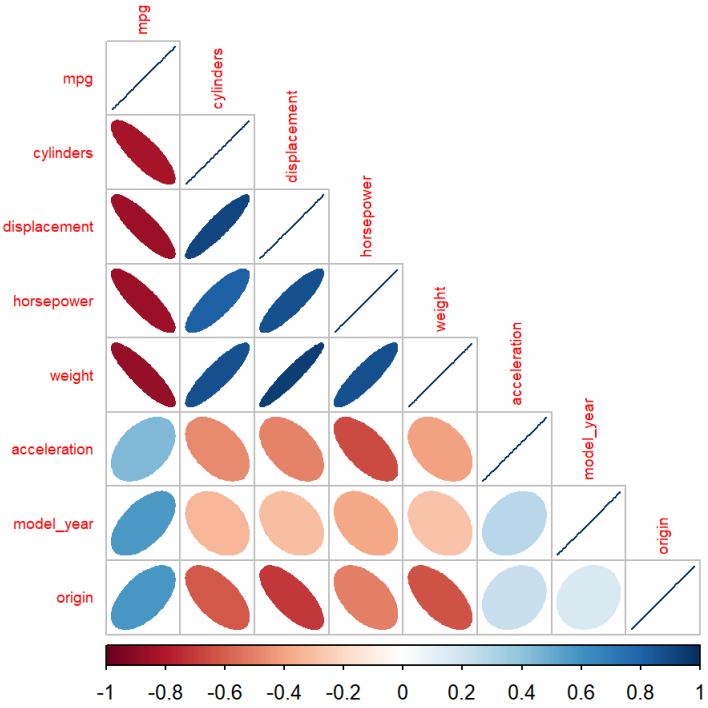
Scenerios 3 and 4 will have similar SSE, as the data points are close to the regression line. Scenerios 4 will have less SSR. scenarios 4 will have larger SST.

Problem 2

(a)(i)Visualize the data in any way you feel relevant (report only relevant/interesting ones)

```
#Q2(a)(i)
library(corrplot)
auto <- read.table("C:/Users/eva/Desktop/作業 上課資料(清大)/大四下/BACS/HW11 BACS/auto-data.txt", header=FALSE, na.stri
ngs = "?")
names(auto) <- c("mpg", "cylinders", "displacement", "horsepower", "weight",
                 "acceleration", "model_year", "origin", "car_name")

auto1<-na.omit(auto)
options(repr.plot.width = 14, repr.plot.height = 8) # => bigger plots for the following
cor_features <- cor(auto1[,1:8], method='spearman')
corrplot::corrplot(cor_features, tl.cex=0.6, type='lower', method="ellipse")
```



(a)(ii)Report a correlation table of all variables, rounding to two decimal places

```
#(ii)
round(cor(auto[,1:8], use="pairwise.complete.obs"),2)
```

```
##                mpg cylinders displacement horsepower weight acceleration
## mpg           1.00     -0.78        -0.80      -0.78  -0.83         0.42
## cylinders    -0.78      1.00         0.95       0.84   0.90        -0.51
## displacement -0.80      0.95         1.00       0.90   0.93        -0.54
## horsepower   -0.78      0.84         0.90       1.00   0.86        -0.69
## weight       -0.83      0.90         0.93       0.86   1.00        -0.42
## acceleration  0.42     -0.51        -0.54      -0.69  -0.42         1.00
## model_year    0.58     -0.35        -0.37      -0.42  -0.31         0.29
## origin        0.56     -0.56        -0.61      -0.46  -0.58         0.21
##               model_year origin
## mpg                 0.58   0.56
## cylinders          -0.35  -0.56
## displacement       -0.37  -0.61
## horsepower         -0.42  -0.46
## weight             -0.31  -0.58
## acceleration        0.29   0.21
## model_year          1.00   0.18
## origin              0.18   1.00
```

(a)(iii)From the visualizations and correlations, which variables seem to relate to mpg?

> Cylinders, displacement, horsepower and weight has the great negative correlation to the mpg. The other variables have positive correlation to mpg.

(a)(iv) Which relationships might not be linear?

> The relationship between every parameter and origin may not be linear, as origin is factor rather than data point.

(a)(v)Are there any pairs of independent variables that are highly correlated (r > 0.7)

> Yes, cylinders and mpg, displacement and mpg, horsepower and mpg, weight and mpg, cylinders and displacement, horsepower and cylinders, weight and cylinders, displacement and horsepower, weight and displacement and weight and horsepower are highly correlated.(r>0.7)

(b)(i)Which independent variables have a ‘significant’ relationship with mpg at 1% significance?

```
auto$origin<-as.factor(auto$origin)
summary(lm(mpg ~ cylinders+displacement+horsepower+weight+acceleration+model_year+origin, data = auto))
```

```
##
## Call:
## lm(formula = mpg ~ cylinders + displacement + horsepower + weight +
##     acceleration + model_year + origin, data = auto)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.0095 -2.0785 -0.0982  1.9856 13.3608
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -1.795e+01  4.677e+00  -3.839 0.000145 ***
## cylinders    -4.897e-01  3.212e-01  -1.524 0.128215
## displacement  2.398e-02  7.653e-03   3.133 0.001863 **
## horsepower   -1.818e-02  1.371e-02  -1.326 0.185488
## weight       -6.710e-03  6.551e-04 -10.243  < 2e-16 ***
## acceleration  7.910e-02  9.822e-02   0.805 0.421101
## model_year    7.770e-01  5.178e-02  15.005  < 2e-16 ***
## origin2       2.630e+00  5.664e-01   4.643 4.72e-06 ***
## origin3       2.853e+00  5.527e-01   5.162 3.93e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.307 on 383 degrees of freedom
##   (6 observations deleted due to missingness)
## Multiple R-squared:  0.8242, Adjusted R-squared:  0.8205
## F-statistic: 224.5 on 8 and 383 DF,  p-value: < 2.2e-16
```

> According to the p-value, displacement, weight, model_year, origin2, and origin3 have significant relationship with mpg at 1% significance.

(b)(ii)Looking at the coefficients, is it possible to determine which independent variables are the most effective at increasing mpg? If so, which ones, and if not, why not? (hint: units!)

> Origin3 is the most effective at increasing mpg, as it has the largest coefficient when mpg increase one unit.

(c)(i) Create fully standardized regression results: are these slopes easier to compare?

```
auto_std<-data.frame(scale(auto[,1:7]))
origin<-auto$origin
auto_std<-cbind(auto_std, origin)
summary(lm(mpg ~ cylinders+displacement+horsepower+weight+acceleration+model_year+origin, data = auto_std))
```

```
##
## Call:
## lm(formula = mpg ~ cylinders + displacement + horsepower + weight +
##     acceleration + model_year + origin, data = auto_std)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.15270 -0.26593 -0.01257  0.25404  1.70942
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.13323    0.03174  -4.198 3.35e-05 ***
## cylinders    -0.10658    0.06991  -1.524  0.12821
## displacement  0.31989    0.10210   3.133  0.00186 **
## horsepower   -0.08955    0.06751  -1.326  0.18549
## weight       -0.72705    0.07098 -10.243  < 2e-16 ***
## acceleration  0.02791    0.03465   0.805  0.42110
## model_year    0.36760    0.02450  15.005  < 2e-16 ***
## origin2       0.33649    0.07247   4.643 4.72e-06 ***
## origin3       0.36505    0.07072   5.162 3.93e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.423 on 383 degrees of freedom
##   (6 observations deleted due to missingness)
## Multiple R-squared:  0.8242, Adjusted R-squared:  0.8205
## F-statistic: 224.5 on 8 and 383 DF,  p-value: < 2.2e-16
```

These slopes are easier to compare, while the origins are the factor, which cannot do the standardization.

(c)(ii)Regress mpg over each nonsignificant independent variable, individually. Which ones become significant when we regress mpg over them individually?

```
summary(lm(mpg~cylinders, data = auto_std))
```

```
##
## Call:
## lm(formula = mpg ~ cylinders, data = auto_std)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.82455 -0.43297 -0.08288  0.32674  2.29046
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.834e-15  3.169e-02    0.00        1
## cylinders   -7.754e-01  3.173e-02  -24.43   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6323 on 396 degrees of freedom
## Multiple R-squared:  0.6012, Adjusted R-squared:  0.6002
## F-statistic: 597.1 on 1 and 396 DF,  p-value: < 2.2e-16
```

```
summary(lm(mpg~horsepower, data = auto_std))
```

```
##
## Call:
## lm(formula = mpg ~ horsepower, data = auto_std)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.73632 -0.41699 -0.04395  0.35351  2.16531
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.008784   0.031701  -0.277    0.782
## horsepower  -0.777334   0.031742 -24.489   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6277 on 390 degrees of freedom
##   (6 observations deleted due to missingness)
## Multiple R-squared:  0.6059, Adjusted R-squared:  0.6049
## F-statistic: 599.7 on 1 and 390 DF,  p-value: < 2.2e-16
```

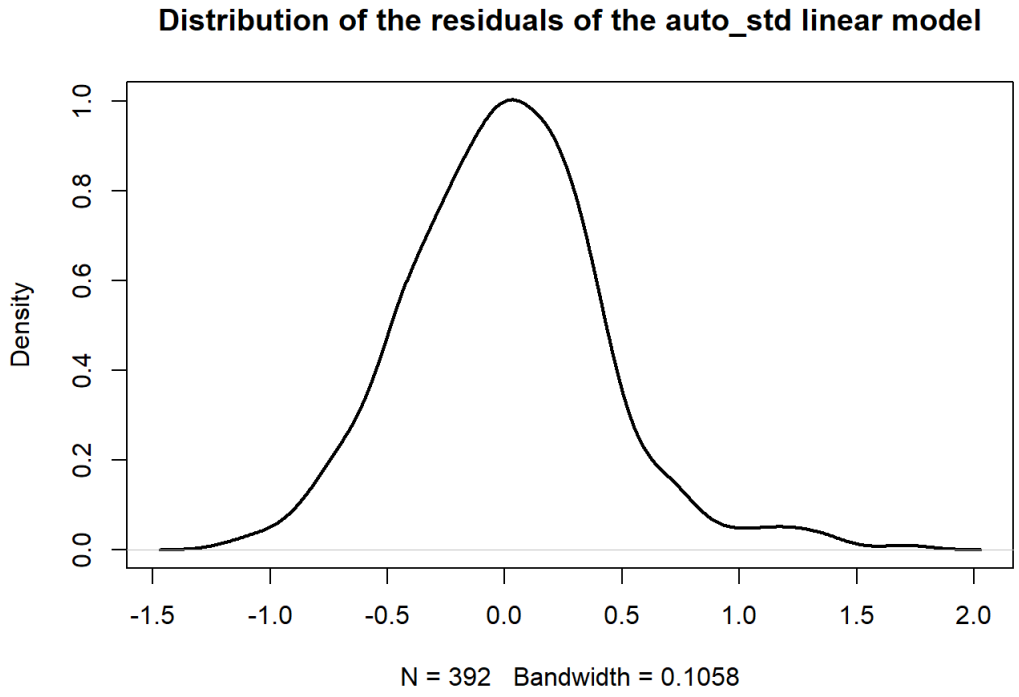
```
summary(lm(mpg~acceleration, data = auto_std))
```

```
##
## Call:
## lm(formula = mpg ~ acceleration, data = auto_std)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.3039 -0.7210 -0.1589  0.6087  2.9672
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.004e-16  4.554e-02   0.000        1
## acceleration 4.203e-01  4.560e-02   9.217   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9085 on 396 degrees of freedom
## Multiple R-squared:  0.1766, Adjusted R-squared:  0.1746
## F-statistic: 84.96 on 1 and 396 DF,  p-value: < 2.2e-16
```

> By regressing mpg over them individually, all nonsignificant independent variables become significant.

(c)(iii)Plot the density of the residuals: are they normally distributed and centered around zero?

```
regr_a<-summary(lm(mpg ~ cylinders+displacement+horsepower+weight+acceleration+model_year+origin, data = auto_std))
plot(density(regr_a$residuals), main='Distribution of the residuals of the auto_std linear model', lwd=2)
```




```
shapiro.test(regr_a$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  regr_a$residuals
## W = 0.98243, p-value = 0.0001061
```

The residuals are centered around zero. However, they are not normally distributed. The p-value of Shapiro-Wilk normality test is 0.0001061, which is less than 0.05, so we should reject H0(The residuals follow normal distribution).