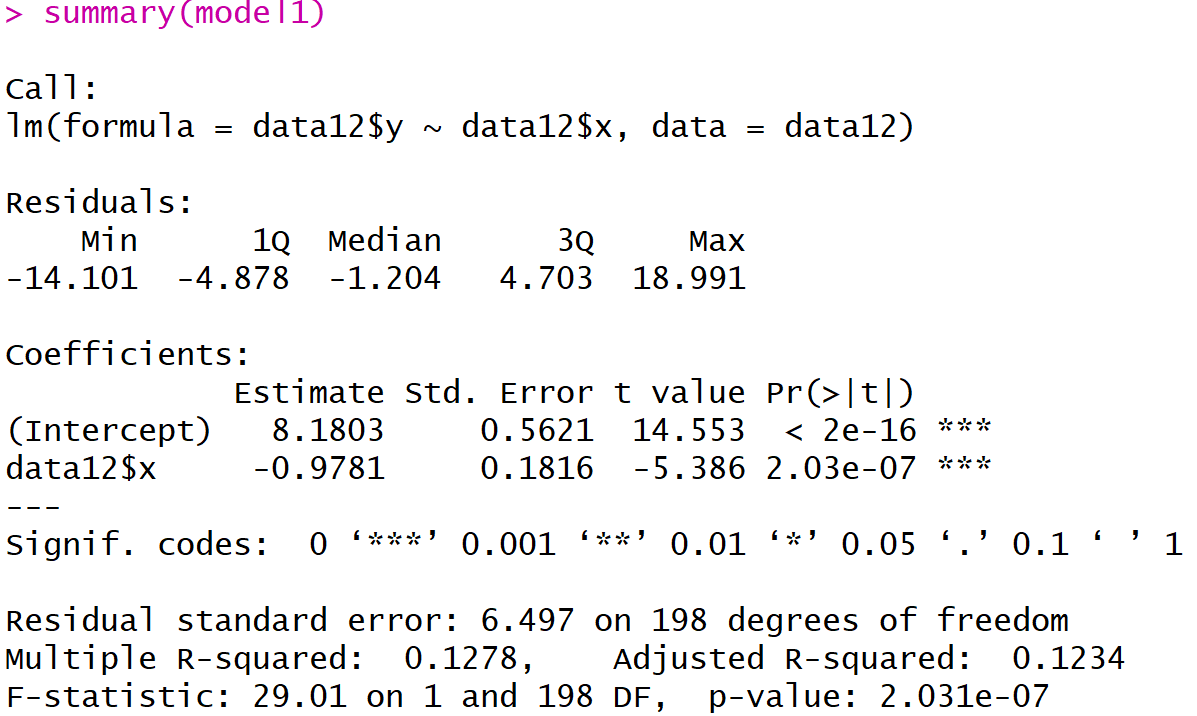


**1. Fit the model Y = β0+β1X +ε to answer “if X is useful in explaining Y ”. Report the result and make conclusion.**



Model: Y= β0+β1X +ε,

Y: y

X: x

β0: intercept in Y

β1: x對y的斜率

ε: 模型的誤差

H0: β1=0，意即x對y無影響

設定α =0.05

如上圖報表中紅框所示，β1的p-value=2.03e-07 < α (0.05)，拒絕H0，x對y有統計顯著的影響，因此解釋變數X可有效解釋Y。

**2. Fit the model Y = β0 + β1X + β2X2 + ε.**

Model2: Y= β0+β1X+ β2X2 +ε,

Y: y

X: x

β0: intercept in Y

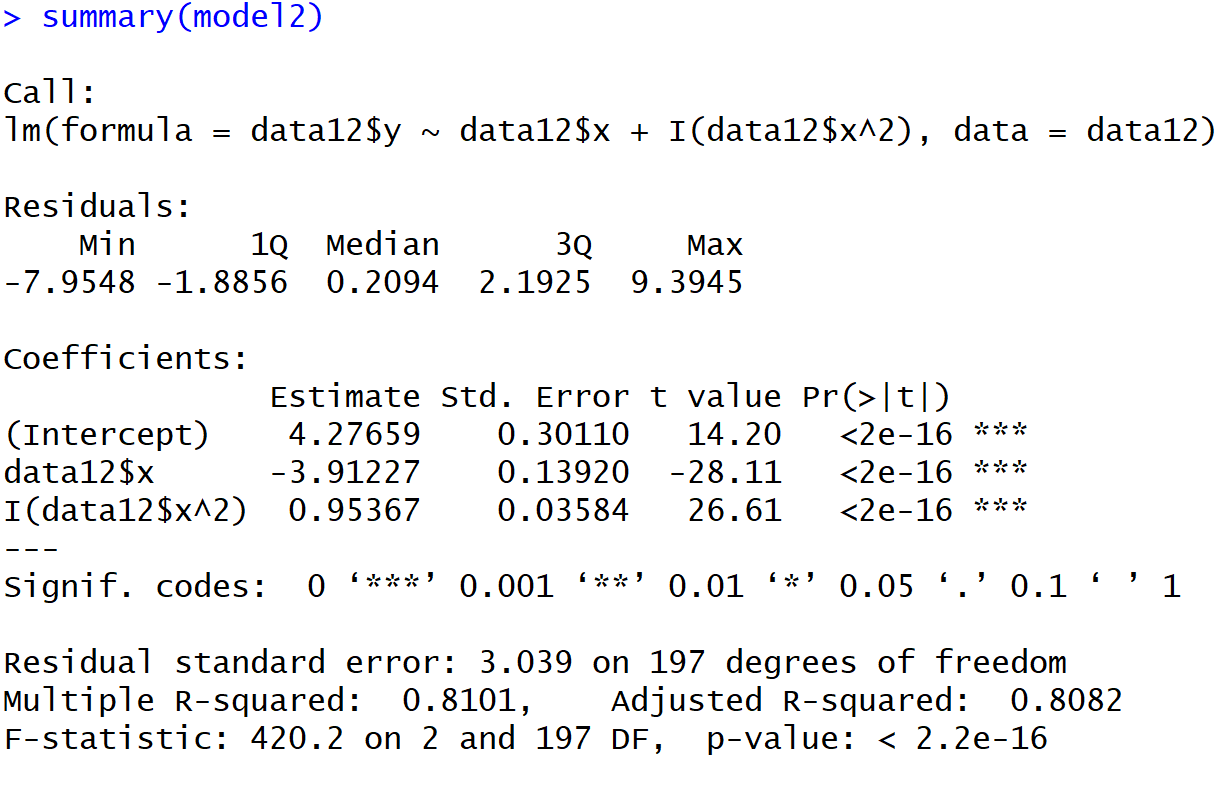
β1: x對y的效果

β2: x與x的交互作用對y的效果

ε: 模型的誤差

**(a) Report the LSE of β2 by fitting**

**(i) multiple linear regression for Y on (X, X2) and**



β2=0.95367

β1=-3.91227

**(ii) simple linear regression for Y on “the residual of X after being explained by X2”.**

**Show that the two estimates of β1 are the same.**

Model3: X = β0+β1X2 +ε,

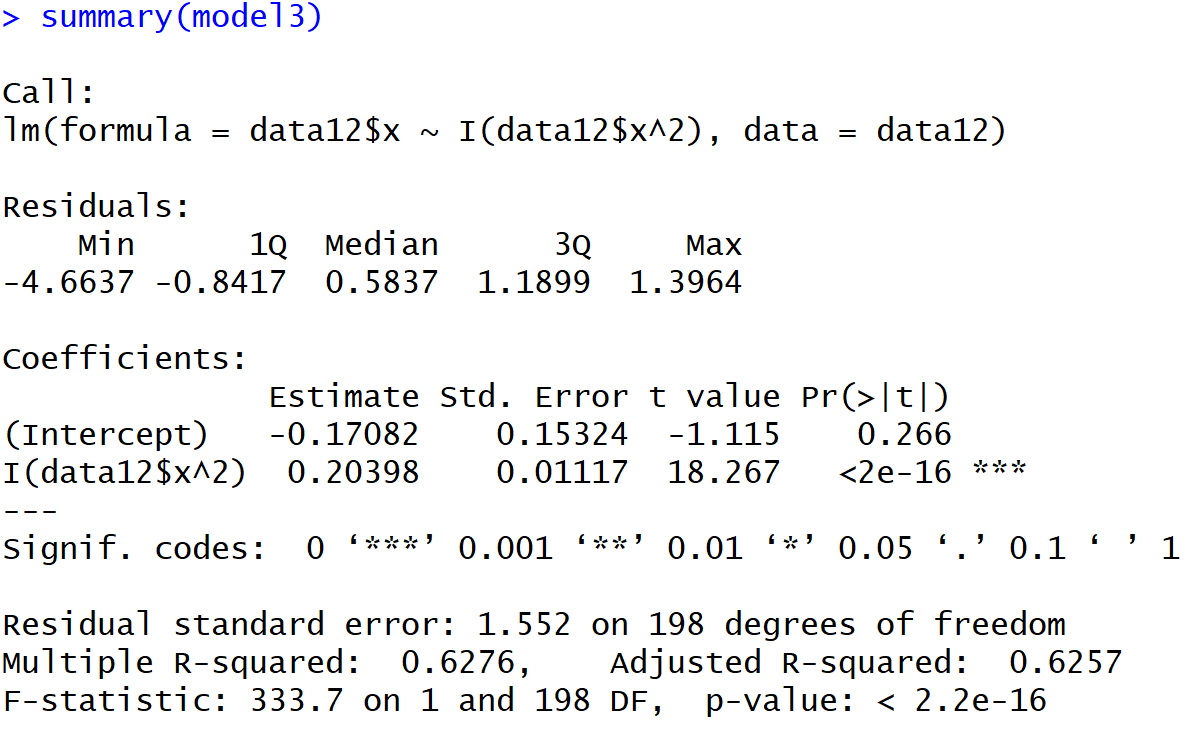
Y: x

X: x2

β0: intercept in X

β1: x2對x的效果

ε: 模型的誤差



由model3可知the residual of X after being explained by X2

Model3\_re: Y = β0+β1d+ε,

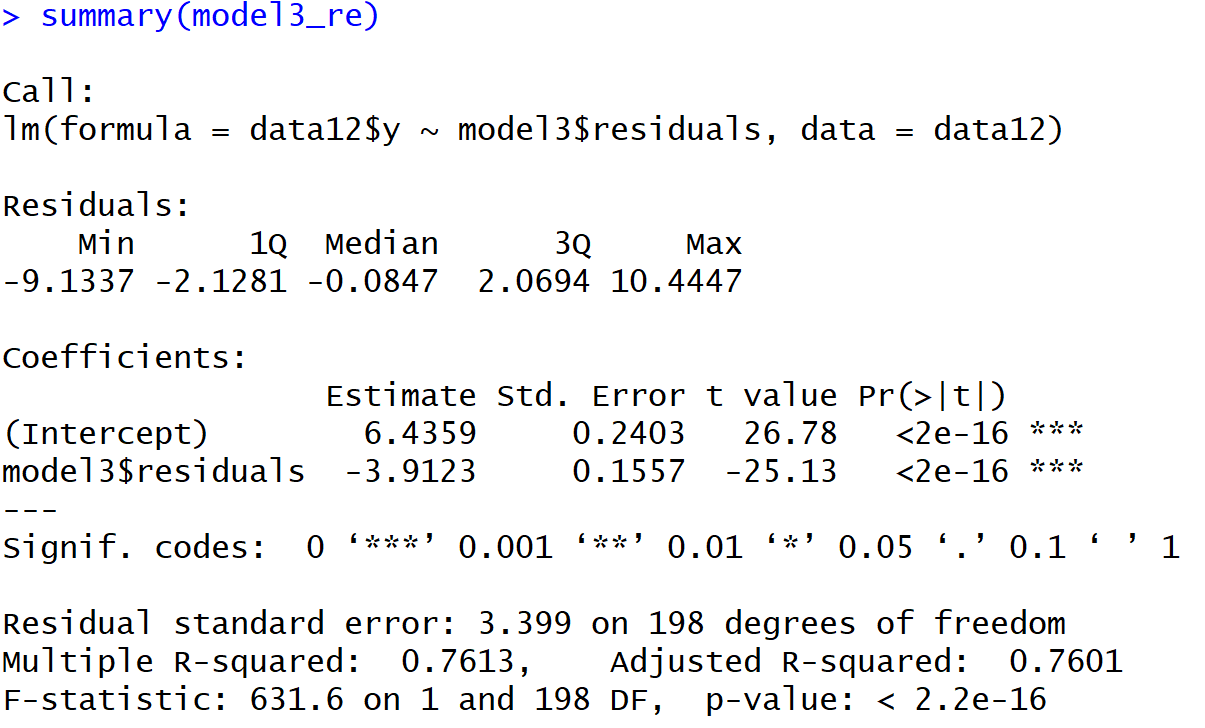
Y: y

d: the residual of X after being explained by X2

β0: intercept in X

β1: the residual of X after being explained by X2對y的效果

ε: 模型的誤差



β1=-3.9123

故可知model2與model3\_re的β1都相同(-3.9123)。

**(b) Answer “if X is useful in explaining Y ”. Report the result and make conclusion.**

對model2進行F-test

H0: β1=β2=0

α=0.05

p-value<2.2e-16，拒絕H0，β1與β2統計顯著地不均為0，X可有效解釋Y。

**(c) What is the predicted value of Y0 at X0 = −2.1?**



Y0=16.69804 at X0 = −2.1

95%C.I. : (10.63548, 22.76059)

**(d) Suggest a suitable procedure to select the model (from Problem-1 and Problem-2)**

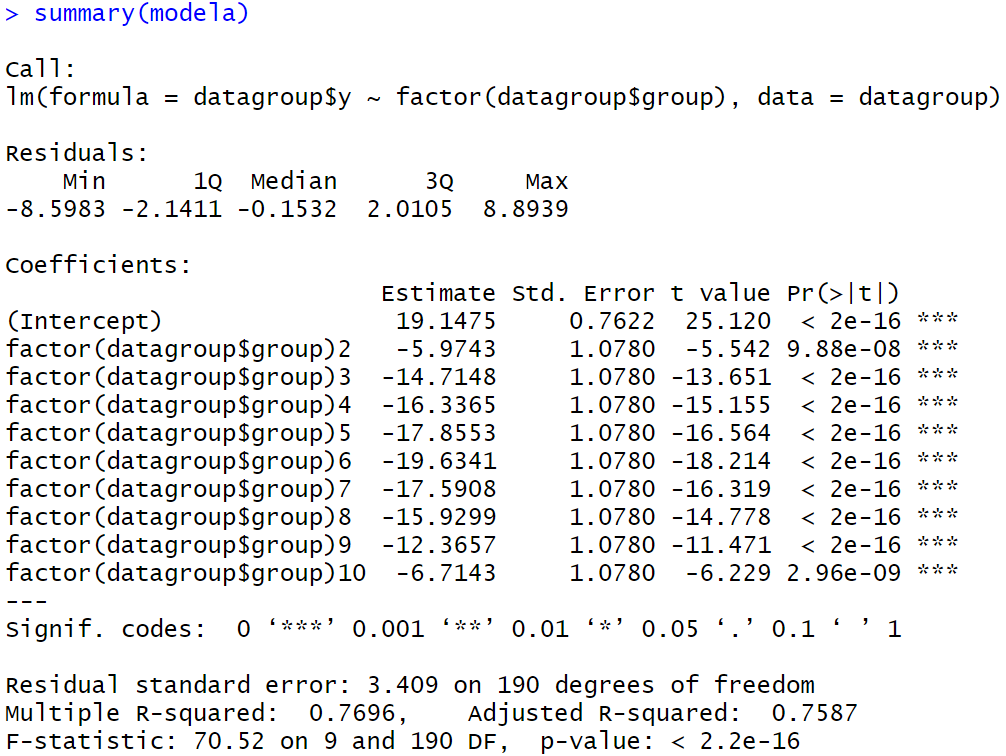
**that best explains the data.**

由problem1之model1 Adujusted R-squared:0.1234，與problem2之model2 Adujusted R-squared:0.8082比較後，後者的Adujusted R-squared較高(0.1234<0.8082)，可知problem2的模型較能解釋x與y的關係。

**3. Dividing the range of X into 10 regions [a0, a1], [a1, a2], . . . , [a9, a10] with equal length.**

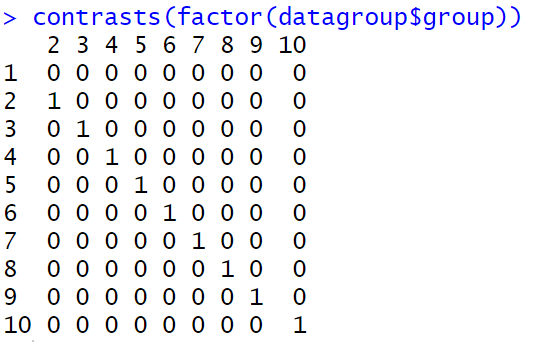
**Let µj = E(Y |X ∈ [aj−1, aj ]) for j = 1, . . . , 10. Another idea to answer “if X is**

**useful in explaining Y ” is to test “H0 : µ1 = µ2 = · · · = µ10”.**



**(a) How to use ANOVA to answer the question? State the model and the corresponding null hypothesis.**

ANOVA檢定是用於比較兩組以上的資料，各組間平均數是否均相等。



refernce <- datagroup$group=1

β0=µ1

β1=µ2-µ1

.

.

Β9=µ10-µ1

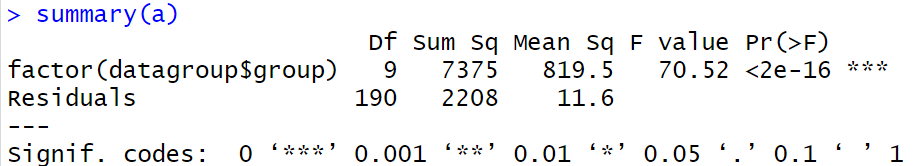
Model:Y= β0+β1X1 +...+β9X9 +ε,

H0 : µ1 = µ2 = · · · = µ10

* H0 : β1 = β2 = · · · = β9=0

**(b) Report the testing result and make conclusion. Does the conclusion support your**

**result in Problem-2 of HW2?**



p-value<2e-16，遠小於α=0.05，故拒絕H0 : µ1 = µ2 = · · · = µ10，

意即X能解釋Y。

這題的結果與 Problem-2 of HW2的結論相同，皆為X能解釋Y。

**4. Both the models in Problem-1 and Problem-3 contain only one covariate X. What is**

**the difference between them?**

同樣一個解釋變數X下，在problem1的模型為簡單線性回歸，當資料分布非線性分布時，模型對數據解釋力較差，problem3的模型是先把數據依照X大小排序後平分成十段，再分別做線性回歸，這樣不論整個資料有無呈線性分布，做出的模型都較接近資料真正分布以及趨勢。

**5. Considering the methods in Problem-1, Problem-2, Problem-3, and your model in**

**Problem-4 of HW2.**

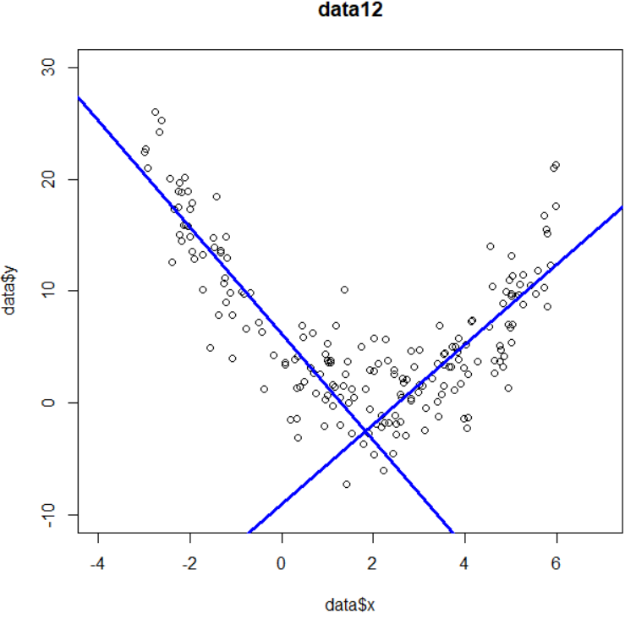
1. **Do they give the same result when answering “if X is useful in explainingY ”? Give reason to support your observation.**

藉由觀察檢定結果，發現無論是哪一題的X對Y都具備解釋力。

然而這幾個模型解釋資料點分布的能力並不相同，各模型adj- R square如下：

problem-1(model1):0.1234

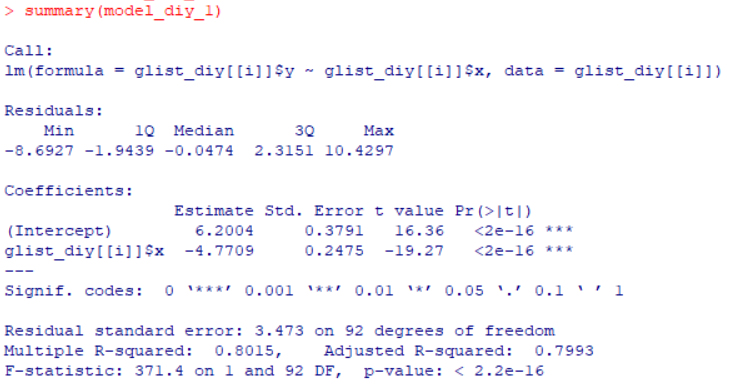
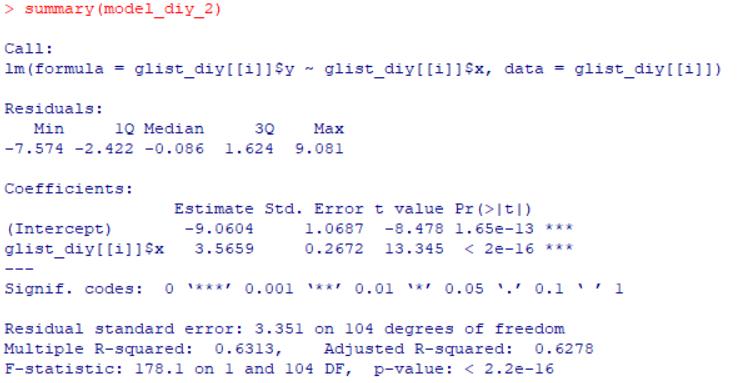
Problem-4 of HW2

problem-2(model2):0.8082

problem-3(model3\_re):0.7587

problem-4(1)( model\_diy\_1):0.7993

problem-4(2)( model\_diy\_2):0.6278

1. **Which model do you prefer to use? Why?**

我們會較傾向使用Problem-2。

選擇模型的首要考量是解釋能力的好壞，而這可以從 Adjusted R-squared的值得知。在這些模型中，Problem-2的Adjusted R-squared最大，代表它的解釋能力最佳，因此優先選擇Problem-2的模型。並且Problem-2只用一個式子，就可以描述資料點的分布趨勢，不用像Problem-3切割成十個區間，在code部分也不會需要一長串才能完成。

Code:

data12<-read.csv(file.choose())

#Fit the model

model1<-lm(data12$y~data12$x,data12)

summary(model2)

##2

#Fit the model

model2<-lm(data12$y~data12$x+I(data12$x^2),data12)

summary(model2)

model3<-lm(data12$x~I(data12$x^2),data12)

summary(model3)

model3\_re<-lm(data12$y~model3$residuals,data12)

summary(model3\_re)

model2$coefficients[1]+model2$coefficients[2]\*(-2.1)+model2$coefficients[3]\*(-2.1)^2

#Predict Y0

model2<-lm(y~x+I(x^2),data12)

pre\_data <- data.frame(x=-2.1)

predict(model2, pre\_data, interval = "prediction", level =0.95)

##3

datax<-data12[order(data12$x),]

for (i in c(1:10)) { assign(paste0("g",i),datax[(i\*20-19):(i\*20),]) }

g1$group<-1

g2$group<-2

g3$group<-3

g4$group<-4

g5$group<-5

g6$group<-6

g7$group<-7

g8$group<-8

g9$group<-9

g10$group<-10

datagroup<-rbind(g1,g2,g3,g4,g5,g6,g7,g8,g9,g10)

modela<-lm(datagroup$y~factor(datagroup$group),datagroup)

summary(modela)

a<-aov(modela)

summary(a)

contrasts(factor(datagroup$group))