一、第四週課堂練習

簡單線性迴歸模型分析、複迴歸模型分析。

二、個人/成員：

A1093325 黃紹瑜 資訊管理學系

三、議題規劃

(一) 簡單線性迴歸模型分析：利用謀殺數量預測與意圖謀殺數量。分析年份對試圖謀殺數量的影響，並利用 ANOVA 和線性回歸模型進行相關性鑑定。

(二) 複迴歸模型分析：利用紐約市公寓評價建立模型去預測對 ValuePerSq 預測力最高。

四、問題定義

(一) 如何使用 ANOVA（變異數分析）鑑定年份對試圖謀殺數量的影響及結果的可視化呈現。

(二) 如何使用簡單線性回歸模型來探索年份和試圖謀殺數量之間的關係？

(三) 如何繪製年份與試圖謀殺數量之間的關係圖，並包括信賴區間。

(四) 如何進行線性回歸模型的係數分析和可視化呈現？

(五) 如何建立複迴歸模型並對其資料進行分析。

五、程式碼設計 和 執行結果

(一) 資料集介紹（以下只列出部分表頭說明）：

1. data_crimes：由 kaggle 中找到的印度於 2001 至 2013 的犯罪紀錄。
   - (1). state: The State or Union Territory where the crime was reported.
   - (2). district: The district within the State/UT where the crime was reported.
   - (3). year: The year when the crime was reported.
   - (4). murder: Number of reported cases of murder.
   - (5). attempt_to_murder: Number of reported cases of attempted murder.
2. housing：紐約市開放資料(NYC Open Data)的紐約市公寓評價資料
   - (1). SqFt: SqFt of the house.
   - (2). Value: Value of the house.
   - (3). ValuePerSqFt: The price of each SqFt.
   - (4). Boro: The district of the house.

(二) 程式碼：

```
library(readr)
library(plyr)
library(cowplot)
library(coefplot)

# 讀取資料
data_crimes <- read_csv("/Users/shaoyu/Desktop/🏫/2 進階 R/dataset/crimes.csv")
# 更改欄位名稱
names(data_crimes) <- c("state","state","year","murder",
                        "attempt_to_murder","culpable_homicide",
                        "rape","custodial_rape","other_rape",
```

```r
                "kidnapping","kidnapping_girls","kidnapping_others",
                "dacoity","intend_dacoity","robbery",
                "burglary","theft","auto_theft","other_theft",
                "riots","breach_of_trust","cheating","counterfeit",
                "arson","hurt","dowry_death","assult","insult",
                "cruelty_husband","import_girls","death_negligence","other",
                "total")

# 檢視資料前幾行
head(data_crimes)

#___________________________________________________________________________#

## 簡單線性迴歸模型
# 用謀殺值(murder)預測試圖謀殺值(attempt_to_murder)
ggplot(data_crimes, aes(x=murder, y=attempt_to_murder)) + geom_point() +
    geom_smooth(method="lm") + labs(x="murder", y="attempt_to_murder")

# 簡單線性回歸模型
# 結果：當 murder 每增加 1 次，預期 attempt_to_murder 增加 0.9019 次
murderLM <- lm(attempt_to_murder ~ murder, data = data_crimes)
murderLM

# 用 summary 檢驗 model 契合度
# 結果：估計值是顯著的
summary(murderLM)

#___________________________________________________________________________#

## 用 ANOVA 鑑定簡單線性回歸模型
# -1 代表去掉截距
# 結果：顯著的
data_crimesAnova <- aov(attempt_to_murder ~ year - 1, data = data_crimes)
# 顯示 ANOVA 結果摘要
# 結果：顯著的
summary(data_crimesAnova)

# 簡單線性回歸模型
data_crimesLM <- lm(attempt_to_murder ~ year - 1, data = data_crimes)
# 顯示模型摘要
# 結果：顯著的
summary(data_crimesLM)

## 看每年的試圖謀殺數量
# 更改 year 的類別成 factor
class(data_crimes$year)
data_crimes$year <- factor(data_crimes$year)

# 將資料照 year 進行分組
# 對每個年份計算了 attempt_to_murder 變數的平均值、標準差、觀測數、90% 的學生化範圍以及該範圍的上下限
crime <- ddply(data_crimes, "year", summarize,
               m.mean=mean(attempt_to_murder), m.sd=sd(attempt_to_murder),
               Length=NROW(attempt_to_murder), # 計算每個年份的觀測數
               tfrac=qt(p=.90, df=Length-1), # 計算 90%的學生化範圍的臨界值
               Lower=m.mean - tfrac*m.sd/sqrt(Length), # 計算 90%信賴區間的下限
               Upper=m.mean + tfrac*m.sd/sqrt(Length) # 計算 90%信賴區間的上限
)

# 顯示 crime 資料框
crime

# 產生 LM 估計值資訊
crimeInfo <- summary(data_crimesLM)
crimeInfo

## 計算信賴區間
# 法一：使用 as.data.frame() 將其轉換為資料框
# 信賴區間的計算是在後續的程式碼中進行的。
crimeCoef <- as.data.frame(crimeInfo$coefficients[, 1:2])
# 法二：使用 within() 在 crimeCoef 資料框中添加了 Lower 和 Upper 兩列，分別表示係數的下限和上限
crimeCoef <- within(crimeCoef, {
  Lower <- Estimate - qt(p=0.90, df=crimeInfo$df[2]) * `Std. Error`
  Upper <- Estimate + qt(p=0.90, df=crimeInfo$df[2]) * `Std. Error`
  crimes <- rownames(crimeCoef)
})
crimeCoef

# Anova(by anova calculated manually)
# 繪製 ANOVA 結果的圖表
anova_plot <- ggplot(crime, aes(x = m.mean, y = year)) + geom_point() +
  geom_errorbarh(aes(xmin = Lower, xmax = Upper), height = 0.3) +
  ggtitle("Crime by year calculated manually")

# lm(by regression model)
# 繪製線性回歸模型的圖表
lm_plot <- ggplot(crimeCoef, aes(x = Estimate, y = year_variables)) +
  geom_point() +
  geom_errorbarh(aes(xmin = Lower, xmax = Upper), height = 0.3) +
  ggtitle("Crime by year calculated from regression model")

# 整合兩個圖表
plot_grid(anova_plot, lm_plot, align = "h")

#____________________________________________________________________________#

## 多元(複)迴歸模型分析
```

```r
housing <- read.table("http://www.jaredlander.com/data/housing.csv",
                      sep = ",", header = TRUE,
                      stringsAsFactors = FALSE)
#修改欄位名稱
names(housing) <- c("Neighborhood", "Class", "Units", "YearBuilt",
                    "SqFt", "Income", "IncomePerSqFt", "Expense",
                    "ExpensePerSqFt", "NetIncome", "Value",
                    "ValuePerSqFt", "Boro")

head(housing)

# 畫出資料圖表
ggplot(housing, aes(x=ValuePerSqFt)) +
  geom_histogram(binwidth=10) + labs(x="Value per Square Foot")

# 依 Boro 做分區上色
ggplot(housing, aes(x=ValuePerSqFt, fill=Boro)) +
  geom_histogram(binwidth=10) + labs(x="Value per Square Foot")

# 依 Boro 分開圖表
ggplot(housing, aes(x=ValuePerSqFt, fill=Boro)) +
  geom_histogram(binwidth=10) + labs(x="Value per Square Foot") +
  facet_wrap(~Boro)

# 面積直方圖
histogram1 <- ggplot(housing, aes(x=SqFt)) + geom_histogram()
# 單位個數直方圖
histogram2 <- ggplot(housing, aes(x=Units)) + geom_histogram()
# 面積直方圖，移除個數多於 1000 的數據
histogram3 <- ggplot(housing[housing$Units < 1000, ],aes(x=SqFt)) + geom_histogram()
# 單位個數直方圖，移除個數多於 1000 的數據
histogram4 <- ggplot(housing[housing$Units < 1000, ],aes(x=Units)) + geom_histogram()

# 合併圖表以便比較與查看
plot_grid(histogram1, histogram2, histogram3, histogram4, labels = c("SqFT Histogram", "Units Histogram",
                                                                     "SqFT Histogram(-Units < 1000)", "Units Histogra(-Units < 1000)"), align = "h")


# 每平方呎價格對面積散佈圖
scatter1 <- ggplot(housing, aes(x = SqFt, y = ValuePerSqFt)) + geom_point()
# 每平方呎價格對單位個數散佈圖
scatter2 <- ggplot(housing, aes(x = Units, y = ValuePerSqFt)) + geom_point()
# 每平方呎價格對面積散佈圖，移除個數多於 1000 的數據
scatter3 <- ggplot(housing[housing$Units < 1000, ], aes(x = SqFt, y = ValuePerSqFt)) + geom_point()
# 每平方呎價格對單位個數散佈圖，移除個數多於 1000 的數據
```

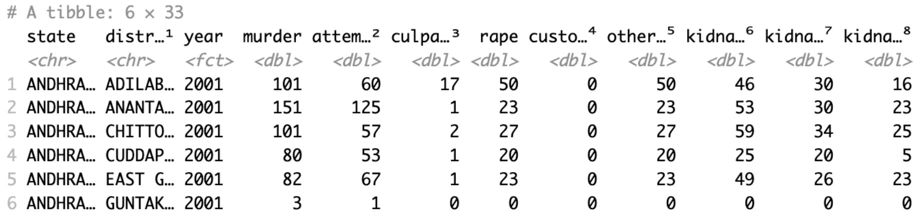
```
scatter4 <- ggplot(housing[housing$Units < 1000, ], aes(x = Units, y = ValuePerSqFt)) + geom_point()

# 合併圖表以便比較與查看
plot_grid(scatter1, scatter2, scatter3, scatter4, labels = c("SqFT Scatter", "Units Scatter",
                                              "SqFT Scatter(-Units < 1000)", "Units Scatter(-Units < 1000)"), align = "h")

#用 sum 計算有多少種建築物要被移除
sum(housing$Units >= 1000)

# 重畫散佈圖
housing <- housing[housing$Units < 1000, ]

## 繪製 valuePerSqFt 對 SqFt 的散佈圖，取 log 對建模也許有幫助
# 結果：有明顯集群，故對建模有幫助
# 房屋面積(SqFt)與每平方英尺的價值(ValuePerSqFt)散佈圖
scatter1_log <- ggplot(housing, aes(x=SqFt, y=ValuePerSqFt)) + geom_point()
# 房屋面積取 log (適合觀察面積的大範圍變化)
scatter2_log <- ggplot(housing, aes(x=log(SqFt), y=ValuePerSqFt)) + geom_point()
# 每平方英尺的價值取 log（適合觀察價值的大範圍變化）
scatter3_log <- ggplot(housing, aes(x=SqFt, y=log(ValuePerSqFt))) + geom_point()
# 皆取 log
scatter4_log <- ggplot(housing, aes(x=log(SqFt), y=log(ValuePerSqFt))) + geom_point()

# 合併圖表以便比較與查看
plot_grid(scatter1_log, scatter2_log, scatter3_log, scatter4_log, labels = c("Normal Scatter", "log(SqFT) Scatter",
                                              "log(ValuePerSqFt) Scatter", "Both logged Scatter"), align = "h")


## 繪製 valuePerSqFt 對 Units 的散佈圖，取 log 對建模不一定有幫助
# 結果：無明顯集群，故對建模無幫助
scatter5_log <- ggplot(housing, aes(x=Units, y=ValuePerSqFt)) + geom_point()
scatter6_log <- ggplot(housing, aes(x=log(Units), y=ValuePerSqFt)) + geom_point()
scatter7_log <- ggplot(housing, aes(x=Units, y=log(ValuePerSqFt))) + geom_point()
scatter8_log <- ggplot(housing, aes(x=log(Units), y=log(ValuePerSqFt))) + geom_point()

# 合併圖表以便比較與查看
plot_grid(scatter5_log, scatter6_log, scatter7_log, scatter8_log, labels = c("Normal Scatter", "log(Units) Scatter",
                                              "log(ValuePerSqFt) Scatter", "Both logged Scatter"), align = "h")


# 用 lm 建模（用於瞭解 Units、SqFt 和 Boro 對 ValuePerSqFt 的關係）
```

```
house1 <- lm(ValuePerSqFt ~ Units + SqFt + Boro, data = housing)
# 用 summary 顯示模型資訊
summary(house1)

## 迴歸模型方法
# 法一: 由 house1 提取係數做迴歸模型
house1$coefficients
# 法二
coef(house1)
# 法三
coefficients(house1)

# 繪製線性迴歸模型的係數圖
# 結果: 曼哈頓建築對每平方呎有顯著影響, SqFt 和 Units 對價格影響只有一點
coefplot(house1)

## 建立交互作用模型
# * -> 顯示個別變數及交互作用項
# : -> 只顯示交互作用
house2 <- lm(ValuePerSqFt ~ Units * SqFt + Boro, data = housing)
house3 <- lm(ValuePerSqFt ~ Units:SqFt + Boro, data = housing)
house2$coefficients
house3$coefficients
coefplot(house2)
coefplot(house3)


# 三個變數之間的交互作用
house4 <- lm(ValuePerSqFt ~ SqFt * Units * Income, housing)
house4$coefficients
house5 <- lm(ValuePerSqFt ~ Class * Boro, housing)
house5$coefficients

# 限制 x 軸的範圍
c1 <- coefplot(house1, sort='mag') + scale_x_continuous(limits=c(-.25, .1))
c2 <- coefplot(house1, sort='mag') + scale_x_continuous(limits=c(-.0005, .0005))

# 合併圖表以便比較與查看
plot_grid(c1, c2, align = "h")

# 用 scale() 放大進一步分析
house1.b <- lm(ValuePerSqFt ~ scale(Units) + scale(SqFt) + Boro,
               data=housing)
coefplot(house1.b, sort='mag')

# 三個變數之間的交互作用
house6 <- lm(ValuePerSqFt ~ I(SqFt/Units) + Boro, housing)
house6$coefficients
house7 <- lm(ValuePerSqFt ~ (Units + SqFt)^2, housing)
```

```r
house7$coefficients
house8 <- lm(ValuePerSqFt ~ Units * SqFt, housing)
identical(house7$coefficients, house8$coefficients)
house9 <- lm(ValuePerSqFt ~ I(Units + SqFt)^2, housing)
house9$coefficients

# 將這幾個模型的係數畫成圖表
# 在模型中 Manhattan 價值都是最高的
multiplot(house1, house2, house3)

## 檢視回歸模型的預測力
housingNew <- read.table("http://www.jaredlander.com/data/housingNew.csv",
                         sep = ",", header = TRUE,
                         stringsAsFactors = FALSE)

# 呼叫 predict() 來完成
housePredict <- predict(house1, newdata = housingNew, se.fit = TRUE,
                        interval = "prediction", level = .95)

# 結果： Brooklyn, Manhattan 對 ValuePerSq 預測力最高
head(housePredict$fit)
head(housePredict$se.fit)
```

(三) 執行結果：

1. head(data_crimes)



```
# A tibble: 6 × 33
  state     distr…¹ year  murder attem…² culpa…³  rape custo…⁴ other…⁵ kidna…⁶ kidna…⁷ kidna…⁸
  <chr>     <chr>   <fct>  <dbl>   <dbl>   <dbl> <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>
1 ANDHRA… ADILAB… 2001     101      60      17    50       0      50      46      30      16
2 ANDHRA… ANANTA… 2001     151     125       1    23       0      23      53      30      23
3 ANDHRA… CHITTO… 2001     101      57       2    27       0      27      59      34      25
4 ANDHRA… CUDDAP… 2001      80      53       1    20       0      20      25      20       5
5 ANDHRA… EAST G… 2001      82      67       1    23       0      23      49      26      23
6 ANDHRA… GUNTAK… 2001       3       1       0     0       0       0       0       0       0
```

圖一、data_crimes 資料集展示

2. ggplot(data_crimes, aes(x=murder, y=attempt_to_murder)) + geom_point() + geom_smooth(method="lm") + labs(x="murder", y="attempt_to_murder")

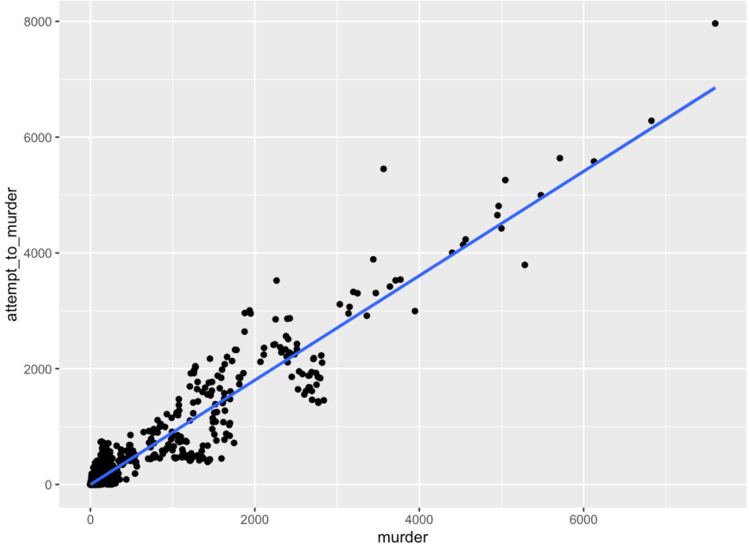
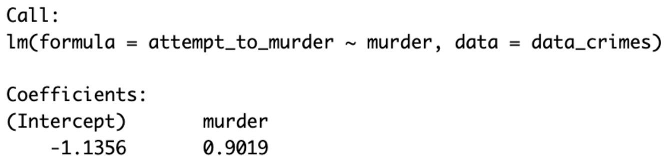


圖 二、謀殺與試圖謀殺值簡單線性迴歸模型

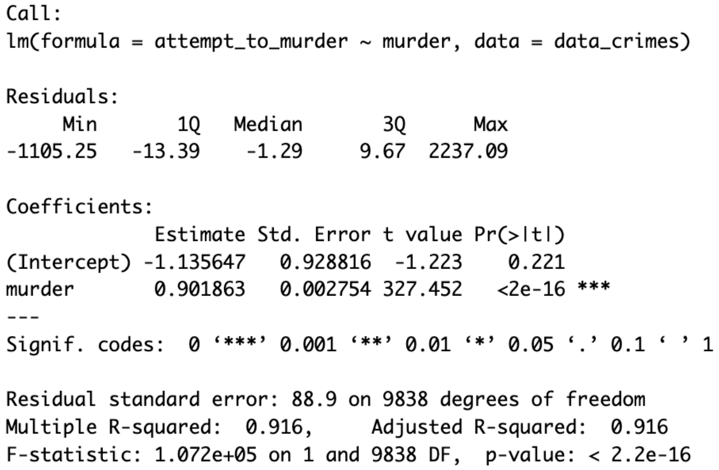
3. murderLM（當 murder 每增加 1 次，預期 attempt_to_murder 增加 0.9019 次）



```
Call:
lm(formula = attempt_to_murder ~ murder, data = data_crimes)

Coefficients:
(Intercept)       murder
    -1.1356       0.9019
```

圖 三、簡單線性回歸模型

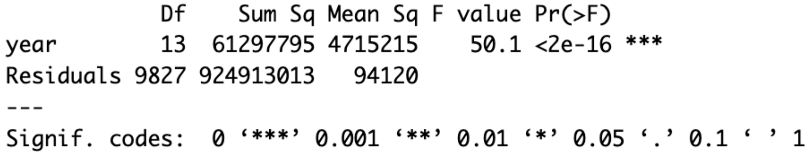
4. summary(murderLM)（檢驗 model 契合度）



```
Call:
lm(formula = attempt_to_murder ~ murder, data = data_crimes)

Residuals:
     Min       1Q   Median       3Q      Max 
-1105.25   -13.39    -1.29     9.67  2237.09 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -1.135647   0.928816  -1.223    0.221    
murder       0.901863   0.002754 327.452   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 88.9 on 9838 degrees of freedom
Multiple R-squared:  0.916,	Adjusted R-squared:  0.916 
F-statistic: 1.072e+05 on 1 and 9838 DF,  p-value: < 2.2e-16
```

圖 四、線性回歸模型的摘要統計

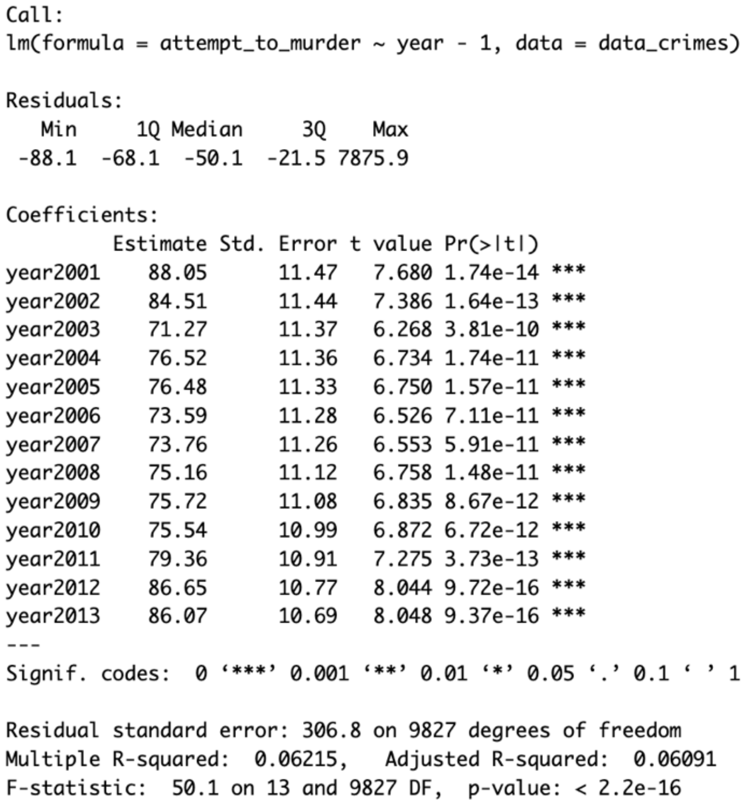
5. summary(data_crimesAnova)



```
              Df    Sum Sq Mean Sq F value Pr(>F)    
year          13  61297795 4715215    50.1 <2e-16 ***
Residuals   9827 924913013   94120                   
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

圖 五、ANOVA 結果摘要

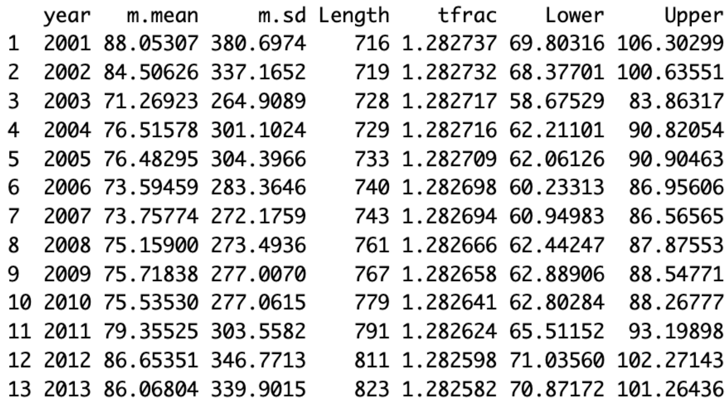
6. summary(data_crimesLM)



```
Call:
lm(formula = attempt_to_murder ~ year - 1, data = data_crimes)

Residuals:
   Min     1Q Median     3Q    Max 
 -88.1  -68.1  -50.1  -21.5 7875.9 

Coefficients:
         Estimate Std. Error t value Pr(>|t|)    
year2001    88.05      11.47   7.680 1.74e-14 ***
year2002    84.51      11.44   7.386 1.64e-13 ***
year2003    71.27      11.37   6.268 3.81e-10 ***
year2004    76.52      11.36   6.734 1.74e-11 ***
year2005    76.48      11.33   6.750 1.57e-11 ***
year2006    73.59      11.28   6.526 7.11e-11 ***
year2007    73.76      11.26   6.553 5.91e-11 ***
year2008    75.16      11.12   6.758 1.48e-11 ***
year2009    75.72      11.08   6.835 8.67e-12 ***
year2010    75.54      10.99   6.872 6.72e-12 ***
year2011    79.36      10.91   7.275 3.73e-13 ***
year2012    86.65      10.77   8.044 9.72e-16 ***
year2013    86.07      10.69   8.048 9.37e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 306.8 on 9827 degrees of freedom
Multiple R-squared:  0.06215,	Adjusted R-squared:  0.06091 
F-statistic:  50.1 on 13 and 9827 DF,  p-value: < 2.2e-16
```

圖 六、線性回歸模型摘要

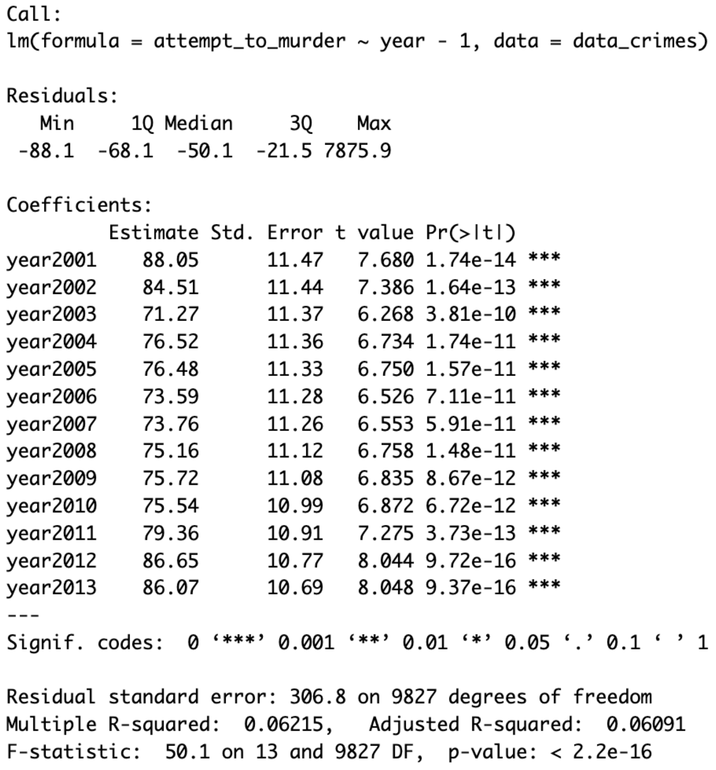
7. crime



```
   year  m.mean     m.sd Length    tfrac    Lower     Upper
1  2001 88.05307 380.6974    716 1.282737 69.80316 106.30299
2  2002 84.50626 337.1652    719 1.282732 68.37701 100.63551
3  2003 71.26923 264.9089    728 1.282717 58.67529  83.86317
4  2004 76.51578 301.1024    729 1.282716 62.21101  90.82054
5  2005 76.48295 304.3966    733 1.282709 62.06126  90.90463
6  2006 73.59459 283.3646    740 1.282698 60.23313  86.95606
7  2007 73.75774 272.1759    743 1.282694 60.94983  86.56565
8  2008 75.15900 273.4936    761 1.282666 62.44247  87.87553
9  2009 75.71838 277.0070    767 1.282658 62.88906  88.54771
10 2010 75.53530 277.0615    779 1.282641 62.80284  88.26777
11 2011 79.35525 303.5582    791 1.282624 65.51152  93.19898
12 2012 86.65351 346.7713    811 1.282598 71.03560 102.27143
13 2013 86.06804 339.9015    823 1.282582 70.87172 101.26436
```

圖 七、將資料按照 year 進行分組並計算之結果

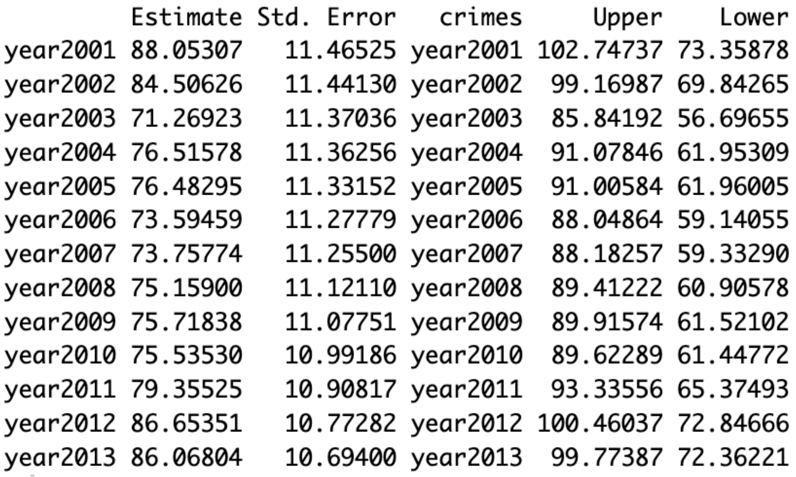
8. crimeInfo



```
Call:
lm(formula = attempt_to_murder ~ year - 1, data = data_crimes)

Residuals:
   Min     1Q Median     3Q    Max 
 -88.1  -68.1  -50.1  -21.5 7875.9 

Coefficients:
         Estimate Std. Error t value Pr(>|t|)    
year2001    88.05      11.47   7.680 1.74e-14 ***
year2002    84.51      11.44   7.386 1.64e-13 ***
year2003    71.27      11.37   6.268 3.81e-10 ***
year2004    76.52      11.36   6.734 1.74e-11 ***
year2005    76.48      11.33   6.750 1.57e-11 ***
year2006    73.59      11.28   6.526 7.11e-11 ***
year2007    73.76      11.26   6.553 5.91e-11 ***
year2008    75.16      11.12   6.758 1.48e-11 ***
year2009    75.72      11.08   6.835 8.67e-12 ***
year2010    75.54      10.99   6.872 6.72e-12 ***
year2011    79.36      10.91   7.275 3.73e-13 ***
year2012    86.65      10.77   8.044 9.72e-16 ***
year2013    86.07      10.69   8.048 9.37e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 306.8 on 9827 degrees of freedom
Multiple R-squared:  0.06215,	Adjusted R-squared:  0.06091 
F-statistic:  50.1 on 13 and 9827 DF,  p-value: < 2.2e-16
```

圖 八、data_crimesLM 估計值資訊

9. crimeCoef



```
         Estimate Std. Error   crimes     Upper    Lower
year2001 88.05307   11.46525 year2001 102.74737 73.35878
year2002 84.50626   11.44130 year2002  99.16987 69.84265
year2003 71.26923   11.37036 year2003  85.84192 56.69655
year2004 76.51578   11.36256 year2004  91.07846 61.95309
year2005 76.48295   11.33152 year2005  91.00584 61.96005
year2006 73.59459   11.27779 year2006  88.04864 59.14055
year2007 73.75774   11.25500 year2007  88.18257 59.33290
year2008 75.15900   11.12110 year2008  89.41222 60.90578
year2009 75.71838   11.07751 year2009  89.91574 61.52102
year2010 75.53530   10.99186 year2010  89.62289 61.44772
year2011 79.35525   10.90817 year2011  93.33556 65.37493
year2012 86.65351   10.77282 year2012 100.46037 72.84666
year2013 86.06804   10.69400 year2013  99.77387 72.36221
```

圖 九、線性回歸模型的係數之信賴區間

10. plot_grid(anova_plot, lm_plot, align = "h")

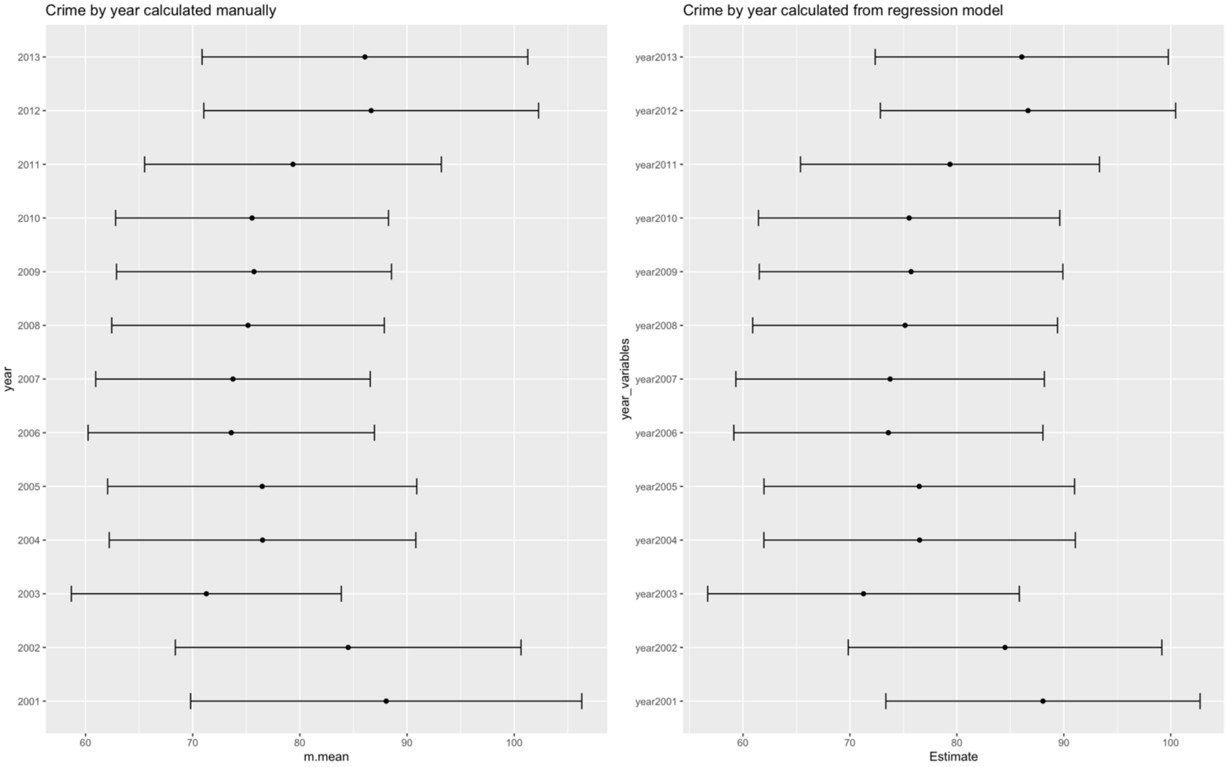
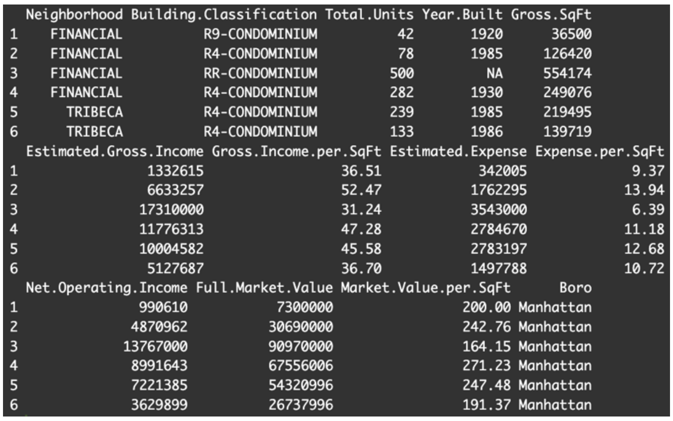


圖 十、Anova（左）和 Linear Regression（右）圖

11. head(housing)



```
  Neighborhood Building.Classification Total.Units Year.Built Gross.SqFt
1    FINANCIAL          R9-CONDOMINIUM          42       1920      36500
2    FINANCIAL          R4-CONDOMINIUM          78       1985     126420
3    FINANCIAL          RR-CONDOMINIUM         500         NA     554174
4    FINANCIAL          R4-CONDOMINIUM         282       1930     249076
5      TRIBECA          R4-CONDOMINIUM         239       1985     219495
6      TRIBECA          R4-CONDOMINIUM         133       1986     139719
  Estimated.Gross.Income Gross.Income.per.SqFt Estimated.Expense Expense.per.SqFt
1                1332615                 36.51            342005             9.37
2                6633257                 52.47           1762295            13.94
3               17310000                 31.24           3543000             6.39
4               11776313                 47.28           2784670            11.18
5               10004582                 45.58           2783197            12.68
6                5127687                 36.70           1497788            10.72
  Net.Operating.Income Full.Market.Value Market.Value.per.SqFt      Boro
1               990610           7300000                200.00 Manhattan
2              4870962          30690000                242.76 Manhattan
3             13767000          90970000                164.15 Manhattan
4              8991643          67556006                271.23 Manhattan
5              7221385          54320996                247.48 Manhattan
6              3629899          26737996                191.37 Manhattan
```

圖 十一、顯示 housing 資料集

12. ggplot(housing, aes(x=ValuePerSqFt)) +
  geom_histogram(binwidth=10) + labs(x="Value per Square Foot")

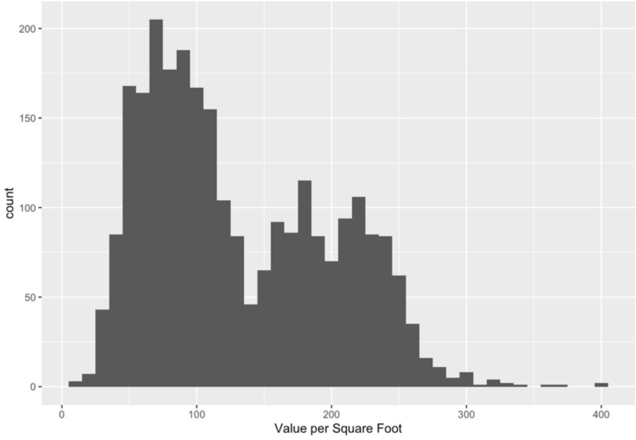


圖 十二、每平方呎數量直方圖

13.

```
ggplot(housing, aes(x=ValuePerSqFt, fill=Boro)) +
geom_histogram(binwidth=10) + labs(x="Value per Square Foot")
```

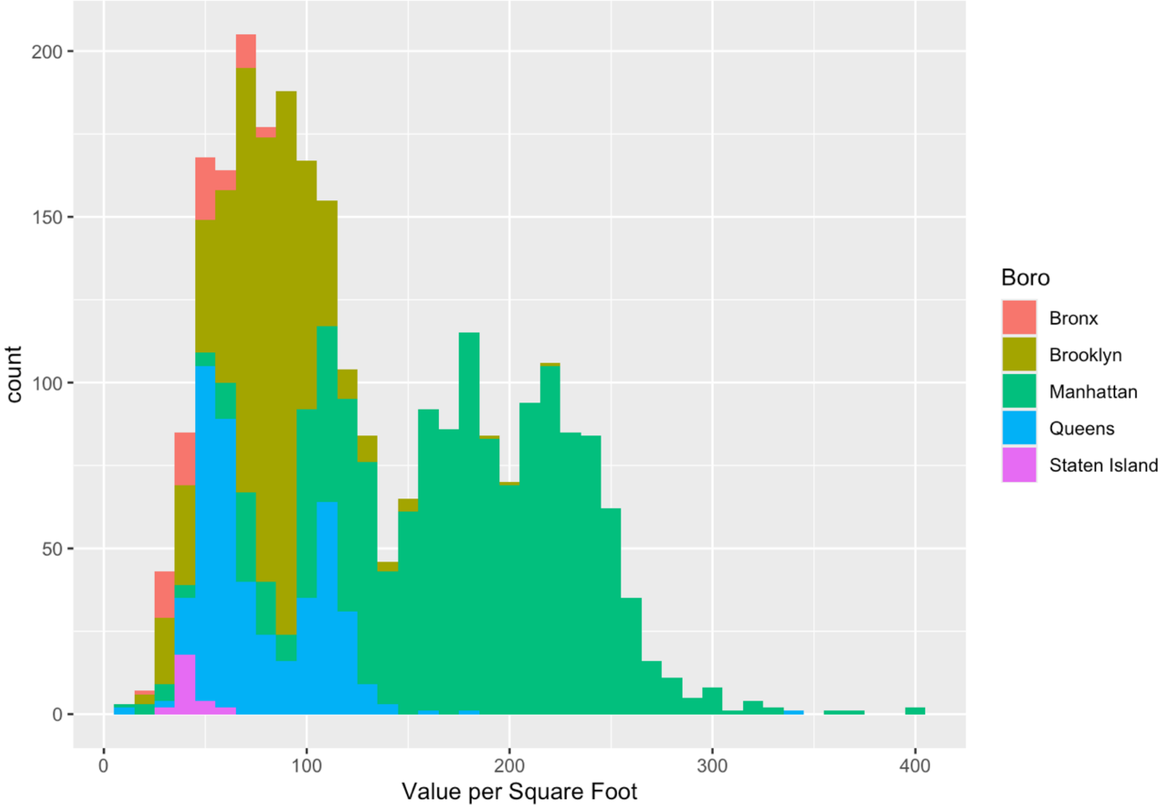


圖 十三、依 Boro 做分區上色 - 每平方呎數量直方圖

14.

```
ggplot(housing, aes(x=ValuePerSqFt, fill=Boro)) +
  geom_histogram(binwidth=10) + labs(x="Value per Square Foot") + facet_wrap(~Boro)
```

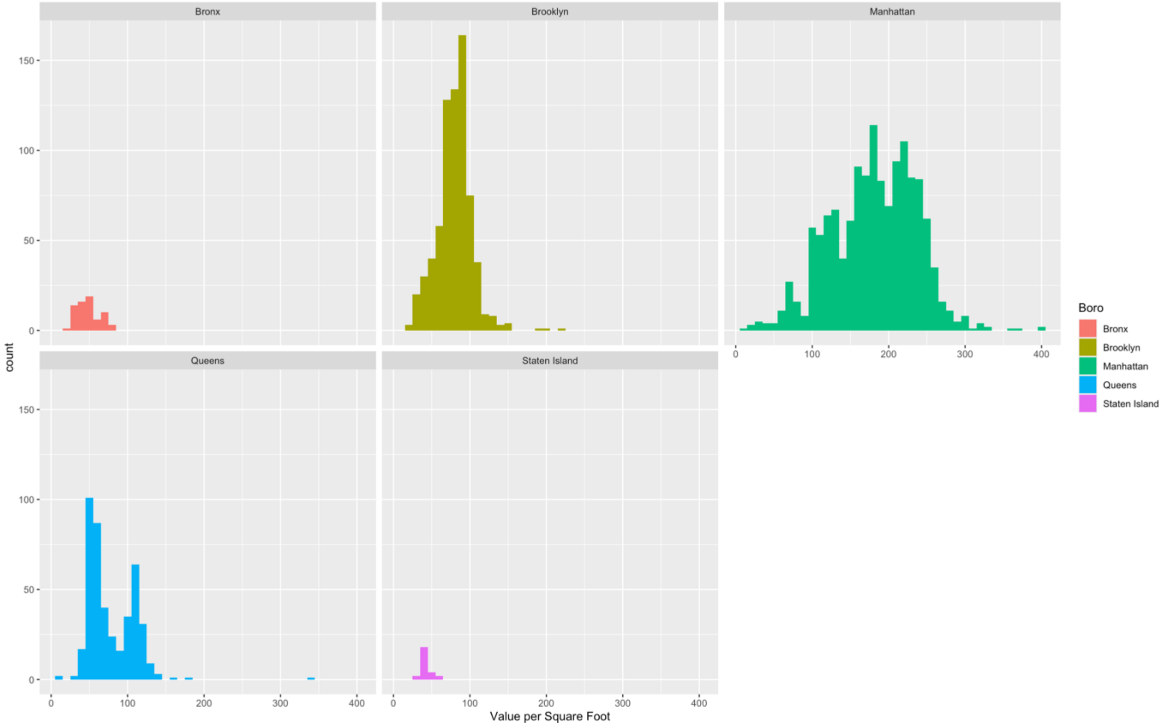


圖 十四、依 Boro 分開圖表 - 每平方呎數量直方圖

15.

```
plot_grid(histogram1, histogram2, histogram3, histogram4, labels = c("SqFT Histogram",
"Units Histogram", SqFT Histogram(-Units < 1000)", "Units Histogra(-Units < 1000)"),
align = "h")
```

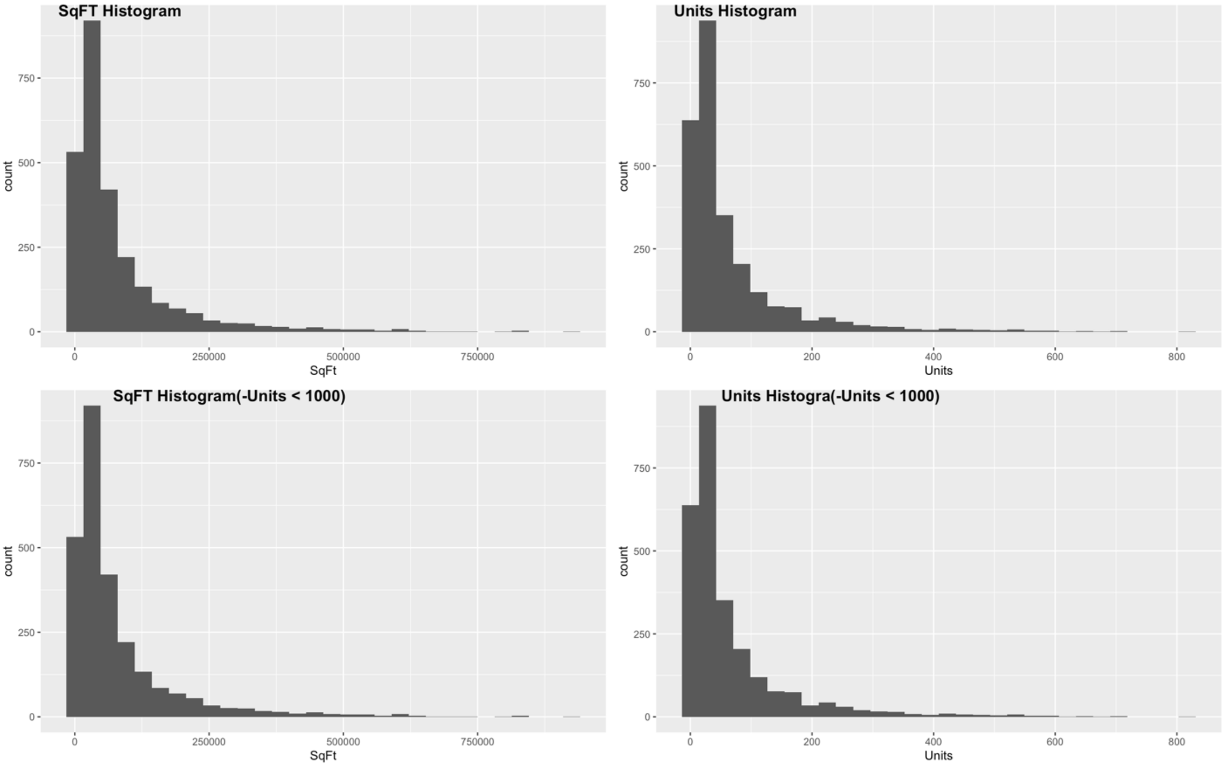


圖 十五、直方圖比較

16. plot_grid(scatter1, scatter2, scatter3, scatter4, labels = c("SqFT Scatter", "Units Scatter”, "SqFT Scatter(-Units < 1000)", "Units Scatter(-Units < 1000)"), align = "h")

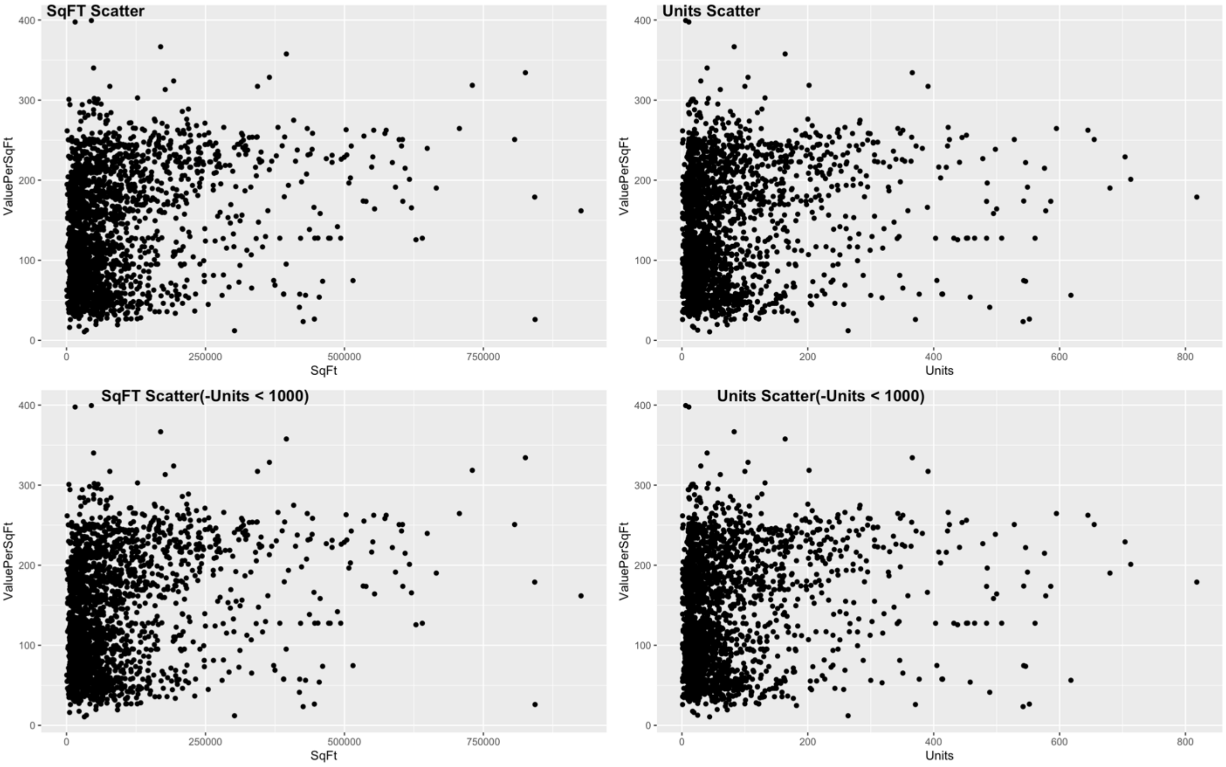


圖 十六、散佈圖比較

17. plot_grid(scatter1_log, scatter2_log, scatter3_log, scatter4_log, labels = c("Normal Scatter", "log(ValuePerSqFt) Scatter", "log(Units) Scatter", "Both logged Scatter"), align = "h")

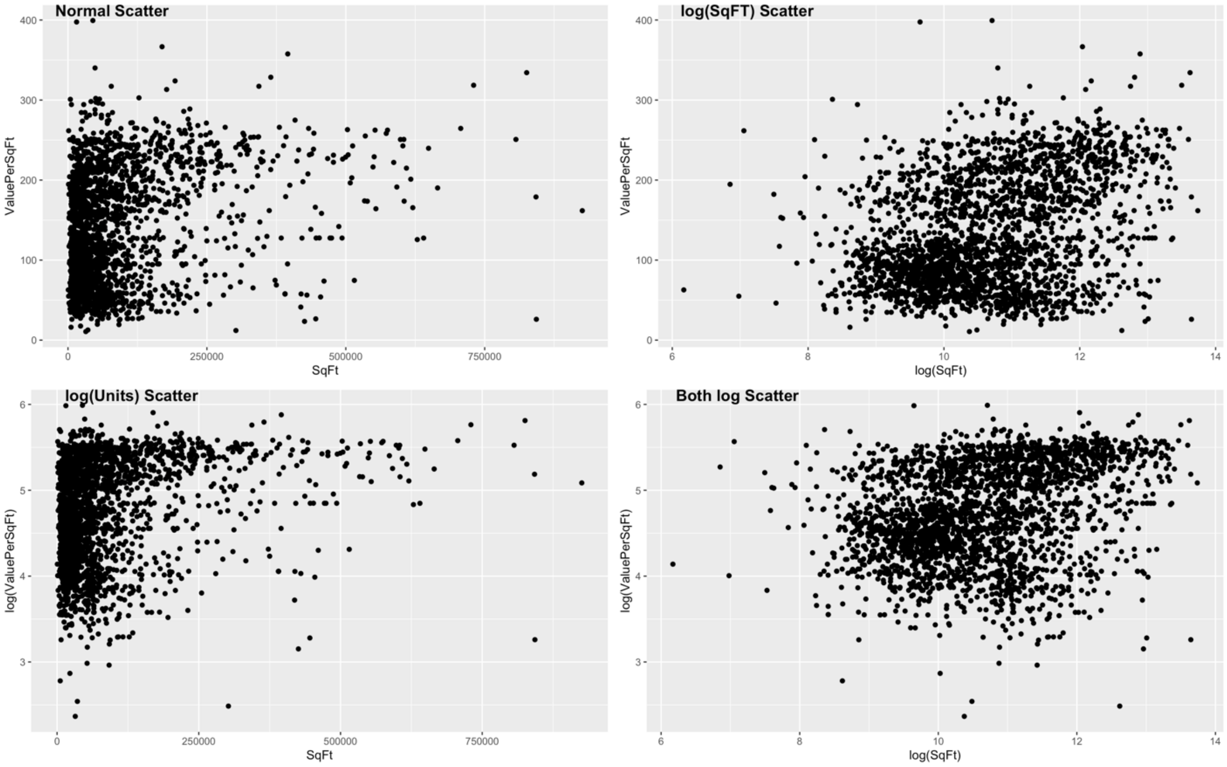


圖 十七、valuePerSqFt 對 SqFt 的散佈圖取 log 比較

18. plot_grid(scatter5_log, scatter6_log, scatter7_log, scatter8_log, labels = c("Normal Scatter", "log(Units) Scatter", "log(ValuePerSqFt) Scatter", "Both logged Scatter"), align = "h")

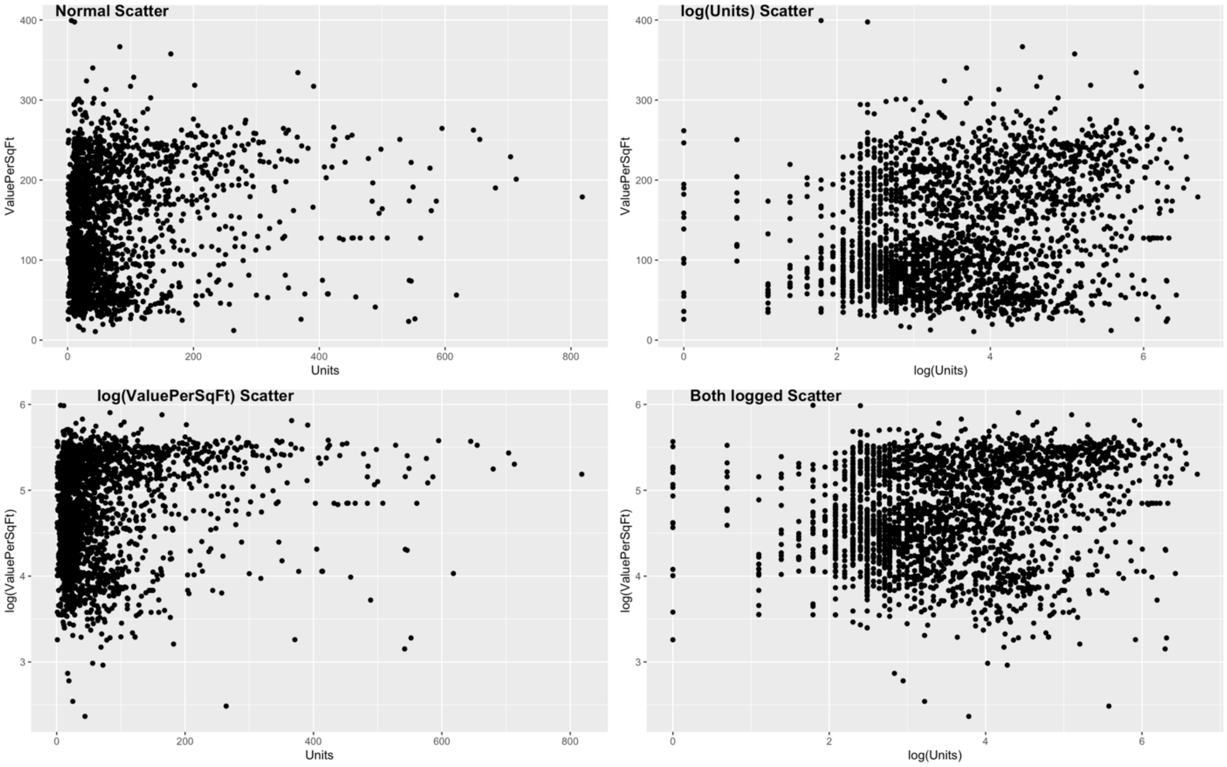
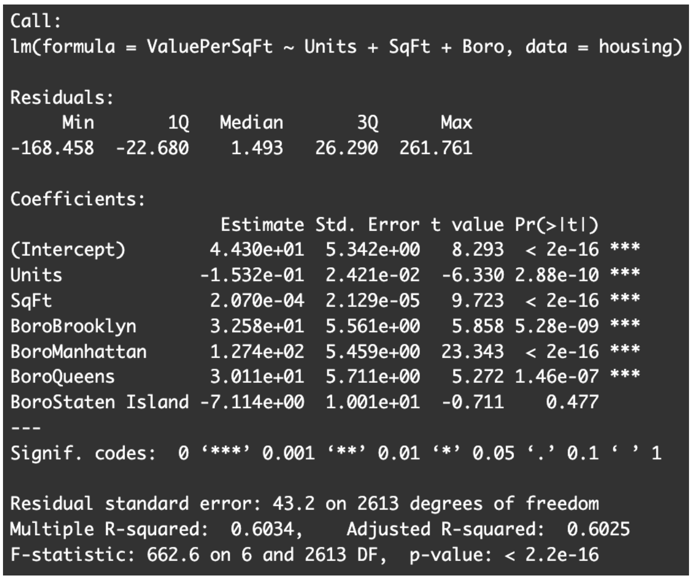


圖 十八、valuePerSqFt 對 Unit 的散佈圖取 log 比較

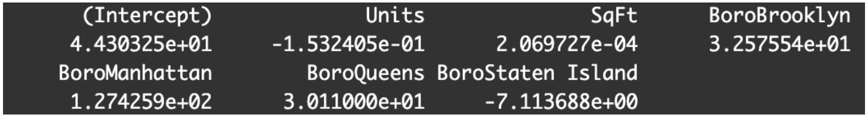
19. summary(house1)



```
Call:
lm(formula = ValuePerSqFt ~ Units + SqFt + Boro, data = housing)

Residuals:
     Min       1Q   Median       3Q      Max 
-168.458  -22.680    1.493   26.290  261.761 

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)    
(Intercept)       4.430e+01  5.342e+00   8.293  < 2e-16 ***
Units            -1.532e-01  2.421e-02  -6.330 2.88e-10 ***
SqFt              2.070e-04  2.129e-05   9.723  < 2e-16 ***
BoroBrooklyn      3.258e+01  5.561e+00   5.858 5.28e-09 ***
BoroManhattan     1.274e+02  5.459e+00  23.343  < 2e-16 ***
BoroQueens        3.011e+01  5.711e+00   5.272 1.46e-07 ***
BoroStaten Island -7.114e+00  1.001e+01  -0.711    0.477    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 43.2 on 2613 degrees of freedom
Multiple R-squared:  0.6034,	Adjusted R-squared:  0.6025 
F-statistic: 662.6 on 6 and 2613 DF,  p-value: < 2.2e-16
```

圖 十九、Units、SqFt 和 Boro 對 ValuePerSqFt 的關係資訊

20. house1$coefficients、coef(house1)、coefficients(house1)



```
      (Intercept)             Units              SqFt      BoroBrooklyn 
     4.430325e+01     -1.532405e-01      2.069727e-04      3.257554e+01 
    BoroManhattan        BoroQueens BoroStaten Island 
     1.274259e+02      3.011000e+01     -7.113688e+00 
```

圖 二十、回歸模型結果

21. coefplot(house1)

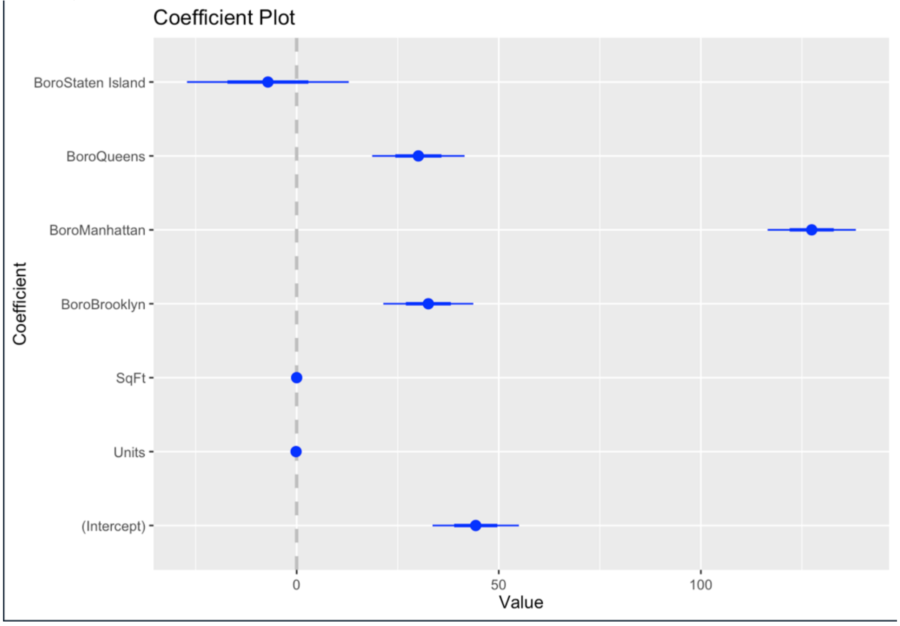
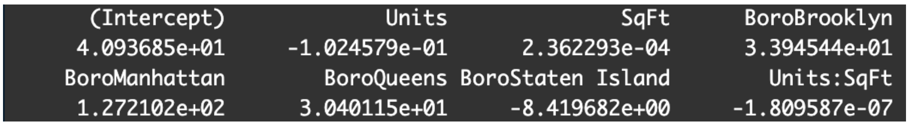


圖 二十一、線性迴歸模型的係數圖

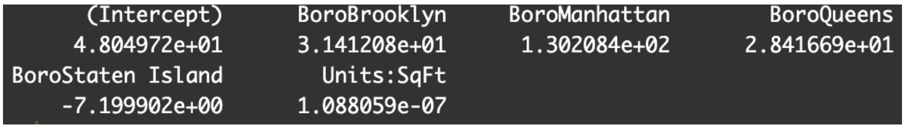
22. house2$coefficients



```
      (Intercept)             Units              SqFt      BoroBrooklyn 
     4.093685e+01     -1.024579e-01      2.362293e-04      3.394544e+01 
    BoroManhattan        BoroQueens BoroStaten Island        Units:SqFt 
     1.272102e+02      3.040115e+01     -8.419682e+00     -1.809587e-07 
```

圖 二十二、顯示個別變數及交互作用項

23. house3$coefficients



```
      (Intercept)      BoroBrooklyn     BoroManhattan        BoroQueens 
     4.804972e+01      3.141208e+01      1.302084e+02      2.841669e+01 
BoroStaten Island        Units:SqFt 
    -7.199902e+00      1.088059e-07 
```

圖 二十三、顯示交互作用

24. coefplot(house2)

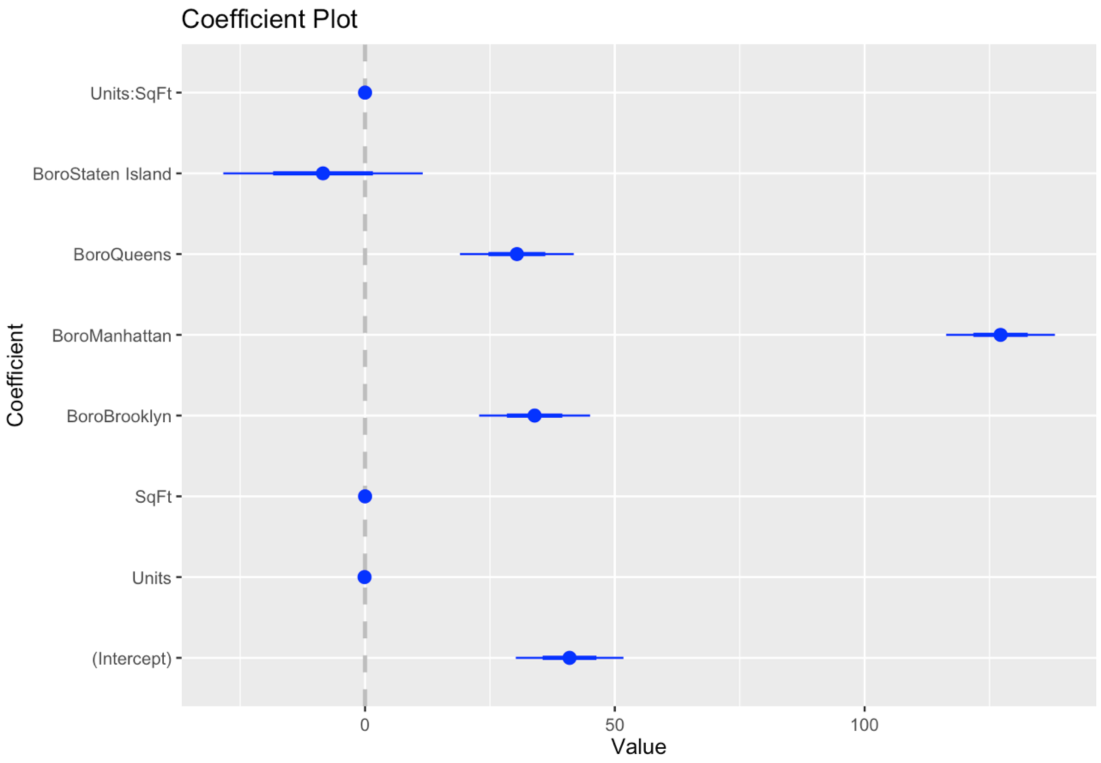


圖 二十四、顯示個別變數及交互作用項線性回歸圖

25. coefplot(house3)

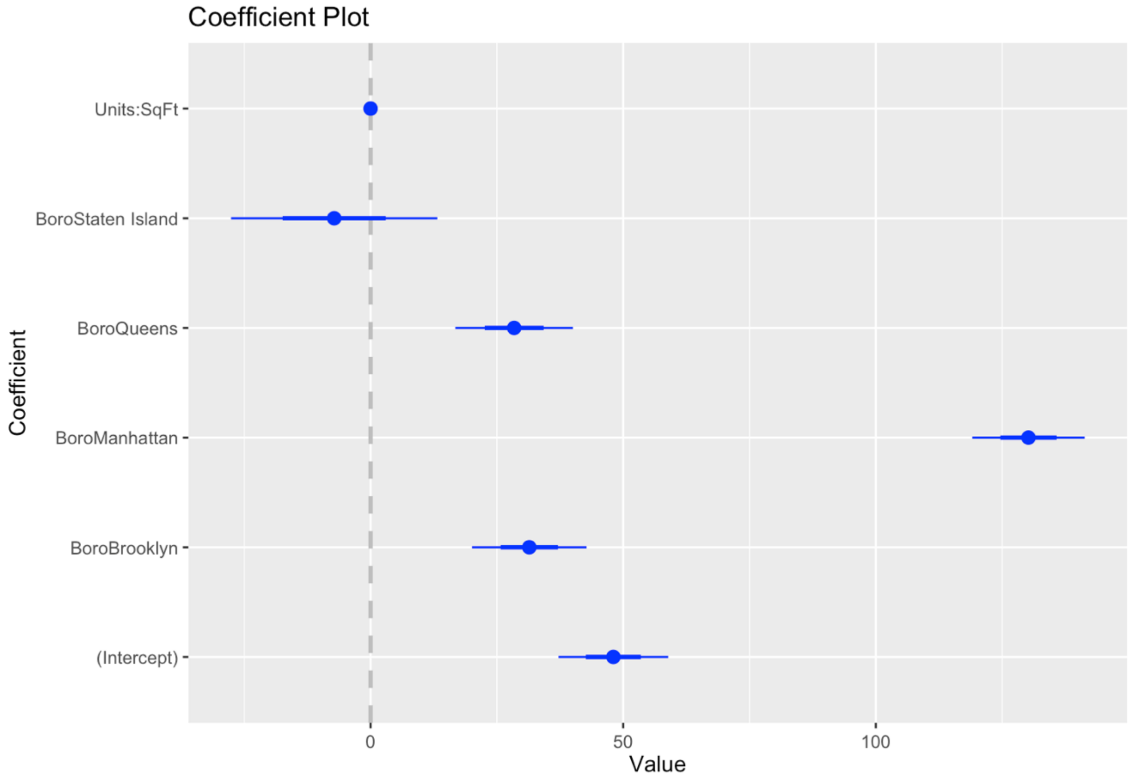
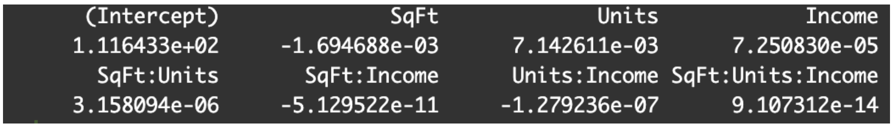


圖 二十五、顯示交互作用線性回歸圖

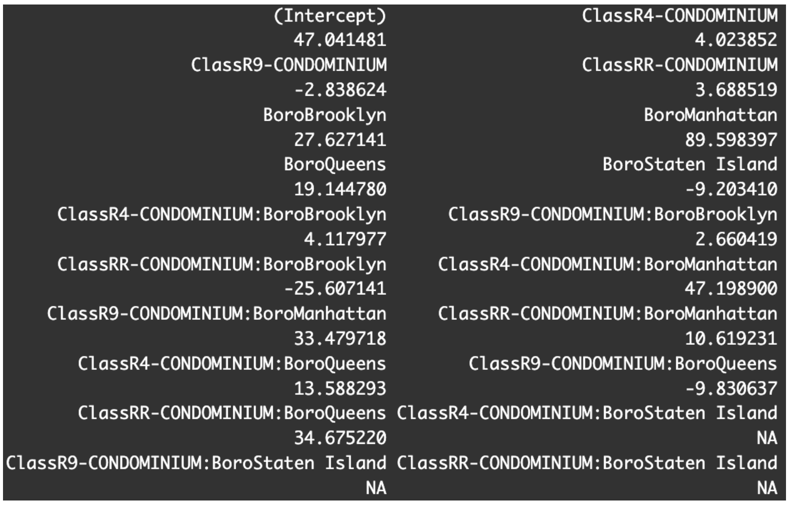
26. house4$coefficients



```
    (Intercept)            SqFt           Units          Income
   1.116433e+02   -1.694688e-03    7.142611e-03    7.250830e-05
     SqFt:Units     SqFt:Income    Units:Income SqFt:Units:Income
   3.158094e-06   -5.129522e-11   -1.279236e-07    9.107312e-14
```

圖 二十六、SqFt、Units、Income 的交互作用

27. house5$coefficients



```
                          (Intercept)                  ClassR4-CONDOMINIUM
                            47.041481                             4.023852
                  ClassR9-CONDOMINIUM                  ClassRR-CONDOMINIUM
                            -2.838624                             3.688519
                         BoroBrooklyn                        BoroManhattan
                            27.627141                            89.598397
                           BoroQueens                    BoroStaten Island
                            19.144780                            -9.203410
     ClassR4-CONDOMINIUM:BoroBrooklyn     ClassR9-CONDOMINIUM:BoroBrooklyn
                             4.117977                             2.660419
     ClassRR-CONDOMINIUM:BoroBrooklyn    ClassR4-CONDOMINIUM:BoroManhattan
                           -25.607141                            47.198900
    ClassR9-CONDOMINIUM:BoroManhattan    ClassRR-CONDOMINIUM:BoroManhattan
                            33.479718                            10.619231
       ClassR4-CONDOMINIUM:BoroQueens       ClassR9-CONDOMINIUM:BoroQueens
                            13.588293                            -9.830637
       ClassRR-CONDOMINIUM:BoroQueens ClassR4-CONDOMINIUM:BoroStaten Island
                            34.675220                                   NA
ClassR9-CONDOMINIUM:BoroStaten Island ClassRR-CONDOMINIUM:BoroStaten Island
                                   NA                                   NA
```

圖 二十七、Class、Boro 的交互作用

28. plot_grid(c1, c2, align = "h")

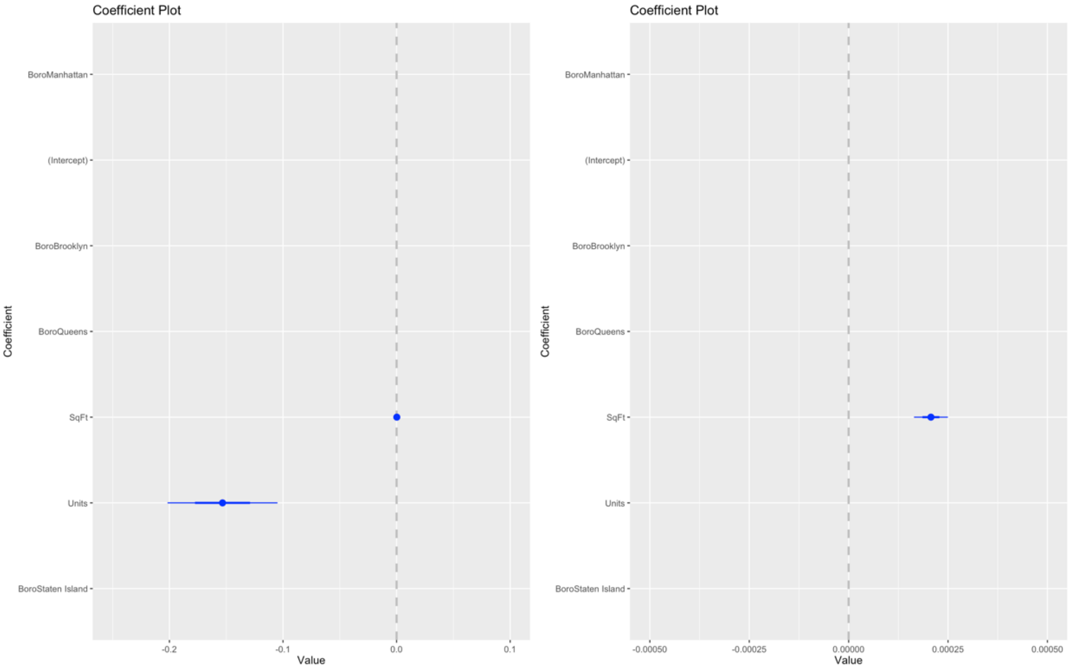


圖 二十八、線性回歸分析限制 X 範圍在-0.25 ~ 0.1（左）、-0.0005 ~ 0.0005（右）

29. coefplot(house1.b, sort='mag')

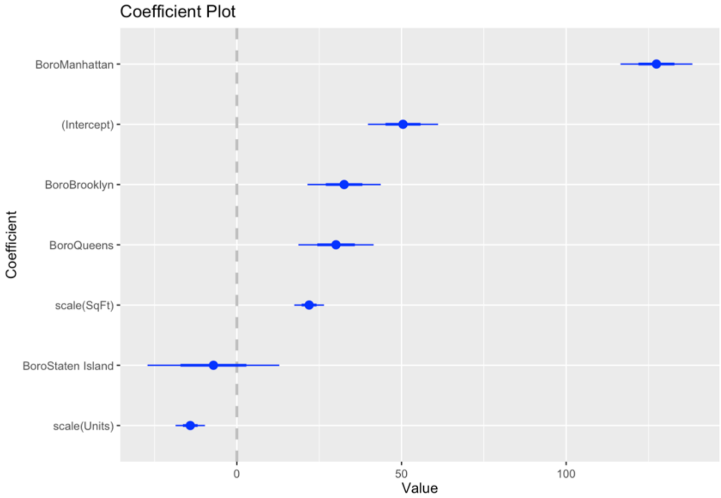


圖 二十九、線性回歸分析(用 scale() 放大)

30. multiplot(house1, house2, house3)

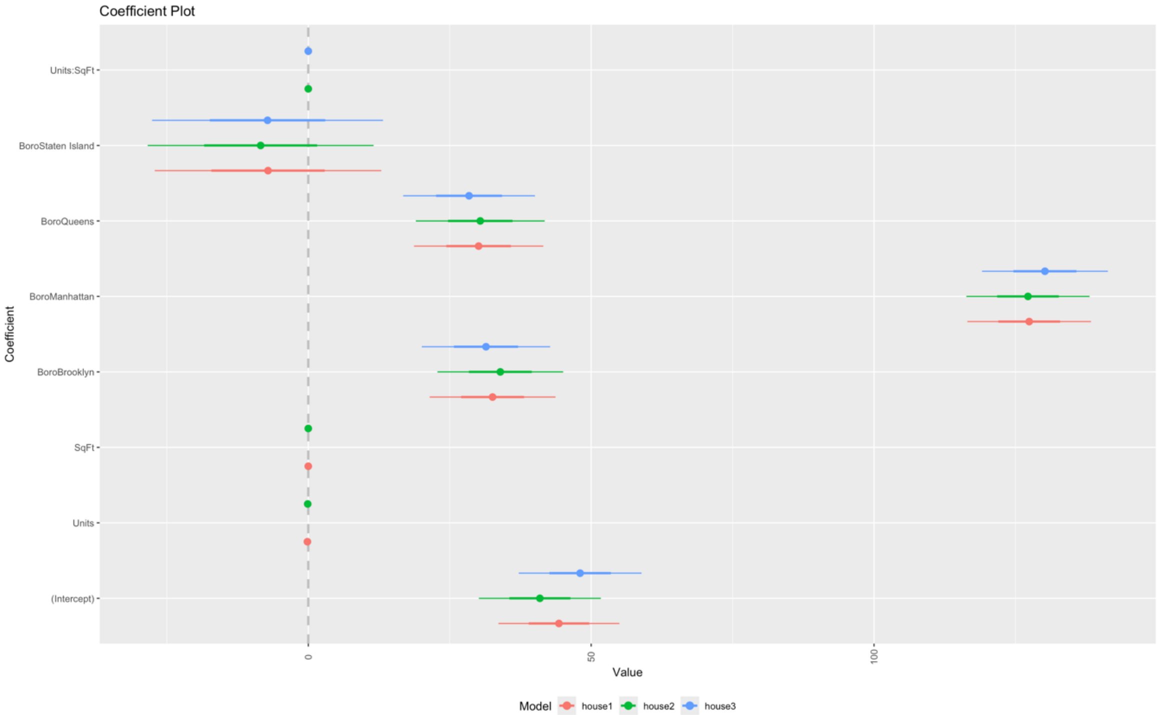
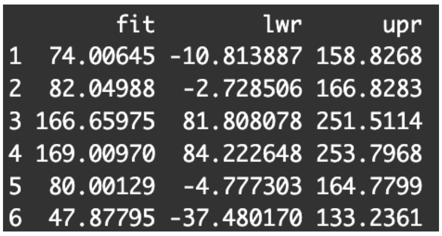


圖 三十、將以上的模型的係數畫成圖表

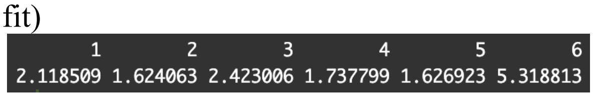
31. head(housePredict$fit)



```
        fit        lwr      upr
1  74.00645 -10.813887 158.8268
2  82.04988  -2.728506 166.8283
3 166.65975  81.808078 251.5114
4 169.00970  84.222648 253.7968
5  80.00129  -4.777303 164.7799
6  47.87795 -37.480170 133.2361
```

圖 三十一、

32. head(housePredict$se.fit)



```
       1        2        3        4        5        6 
2.118509 1.624063 2.423006 1.737799 1.626923 5.318813 
```

圖 三十二、

六、意涵詮釋

(一) 學習簡單與多元迴歸分析。

(二) 如何將圖表做合併。

(三) 更改變數類別成 factor。

(四) 第 25~103 行（用謀殺數預測試圖謀殺數）：藉由簡單線性回歸分析可以得知 murder 每增加 1 次，預期 attempt_to_murder 增加 0.9019 次，且估計值是顯著的。在 ANOVA 分析中顯示出所有的年份都有重疊的部分，平均數是差不多的，因此無法確定是否有顯著的差異。在 LM 線性回歸模型中也有重疊部分，故亦無法確定是否有顯著的差異。

(五) 第 107~264 行：由直方圖看出有離群值，得知 Brooklyn, Manhattan, Queens 個別形成一個峰。接著檢視面積和單位個數的直方圖與散佈圖，得知面積和單位個數這兩個元素很重要。藉由複迴歸模型得知面積和單位個數對價格只有一點影響。最後再藉由交互作用可以得知在 Units、SqFt、Boro 三個係數之 LM、個別變數及交互作用項、交互作用項模型下 Manhattan 的價值皆最高。

(六) 最後用剛做好的模型去預測新資料得知 Brooklyn, Manhattan 對 ValuePerSq 預測力最高。

七、參考說明

1. R 语言 cowplot 介绍——把不同的图像拼接到一起
https://blog.csdn.net/xspyzm/article/details/104345261