Linear Model Assignment 4

鄭雅珊、邱繼賢、廖偉傑

Problem 1.

首先配適模型 $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \epsilon$

```
dat = read.table("http://www.stat.nthu.edu.tw/~swcheng/Teaching/stat5410/data/wastes.txt",
                 header = T,fileEncoding = "UTF-8-BOM", nrows = 20)
g <- lm(y~x1+x2+x3+x4+x5, data=dat)
summary(g)
```

```
##
## Call:
## lm(formula = y ~ x1 + x2 + x3 + x4 + x5, data = dat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.39447 -0.11847  0.00053  0.08313  0.56232
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.156e+00  9.135e-01  -2.360   0.0333 *
## x1          -9.012e-06  5.184e-04  -0.017   0.9864
## x2           1.316e-03  1.263e-03   1.041   0.3153
## x3           1.278e-04  7.690e-05   1.662   0.1188
## x4           7.899e-03  1.400e-02   0.564   0.5815
## x5           1.417e-04  7.375e-05   1.921   0.0754 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2618 on 14 degrees of freedom
## Multiple R-squared:  0.8107, Adjusted R-squared:  0.743
## F-statistic: 11.99 on 5 and 14 DF,  p-value: 0.0001184
```

最終得到模型為

$$\hat{y} = -2.156 - 9.012 \times 10^{-6} x_1 + 1.316 \times 10^{-3} x_2 + 1.278 \times 10^{-4} x_3 + 7.899 \times 10^{-3} x_4 + 1.417 \times 10^{-4} x_5$$

a.

$100(1-\alpha)\%$ confidence region for $A\beta$, A: $d \times p$ matrix, $\beta:\ p \times 1$ vector

General form: $(A\hat{\beta} - A\beta)^T (A(X^T X)^{-1} A^T)^{-1} (A\hat{\beta} - A\beta)/(d\hat{\sigma}^2) \leq F_{d,n-p}(\alpha)$

If $d = 1$,

$$\frac{(A\hat{\beta} - A\beta)^2}{((A(X^TX)^{-1}A^T)\hat{\sigma}^2} \le F_{1,n-p}(\alpha) \Rightarrow \frac{|A\hat{\beta} - A\beta|}{\sqrt{(A(X^TX)^{-1}A^T}\hat{\sigma}} \le t_{n-p}(\alpha/2)$$

$100(1-\alpha)\%$ Confidence interval (CI) for $A\beta$:

$$A\hat{\beta} \mp t_{n-p}(\alpha/2) \times \sqrt{(A(X^TX)^{-1}A^T)}\hat{\sigma} \equiv A\hat{\beta} \mp t_{n-p}(\alpha/2) \times s.e.(A\hat{\beta})$$

本題 $n = 20$，$p = 6$，$\alpha = 0.05$，$t_{14,0.025} = 2.145$。從 summary table 可直接獲得 $A\hat{\beta}$ 及 $s.e.(A\hat{\beta})$ 結果。

95% CI for β_3 $(A = (0,0,0,1,0,0),\ A\hat{\beta} = \hat{\beta}_3)$:

$$\hat{\beta}_3 \mp t_{14}(0.025) \times s.e.(\hat{\beta}_3) = 1.728 \times 10^{-4} \mp 2.145 \times 7.690 \times 10^{-5} = (-3.714 \times 10^{-5}, 2.927 \times 10^{-4})$$

95% CI for β_5 $(A = (0,0,0,0,0,1),\ A\hat{\beta} = \hat{\beta}_5)$:

$$\hat{\beta}_5 \mp t_{14}(0.025) \times s.e.(\hat{\beta}_5) = 1.417 \times 10^{-4} \mp 2.145 \times 7.375 \times 10^{-5} = (-1.652 \times 10^{-5}, 2.998 \times 10^{-4})$$

另外也可使用 $confint$ 指令求得 β_3 及 β_5 的 95% CI：

```
confint(g)[c(4,6),]
```

```
##              2.5 %       97.5 %
## x3 -3.713929e-05 0.0002927368
## x5 -1.652198e-05 0.0002998305
```

b.

承 a，本題 $A = (0,0,0,1,0,2),\ A\hat{\beta} = \hat{\beta}_3 + 2\hat{\beta}_5$。

```
A <- t(c(0,0,0,1,0,2))
x <- model.matrix(g)
xtxi <- solve(t(x)%*%x)
c(sqrt(A%*%xtxi%*%t(A)) * summary(g)$sigma)
```

```
## [1] 0.0001641751
```

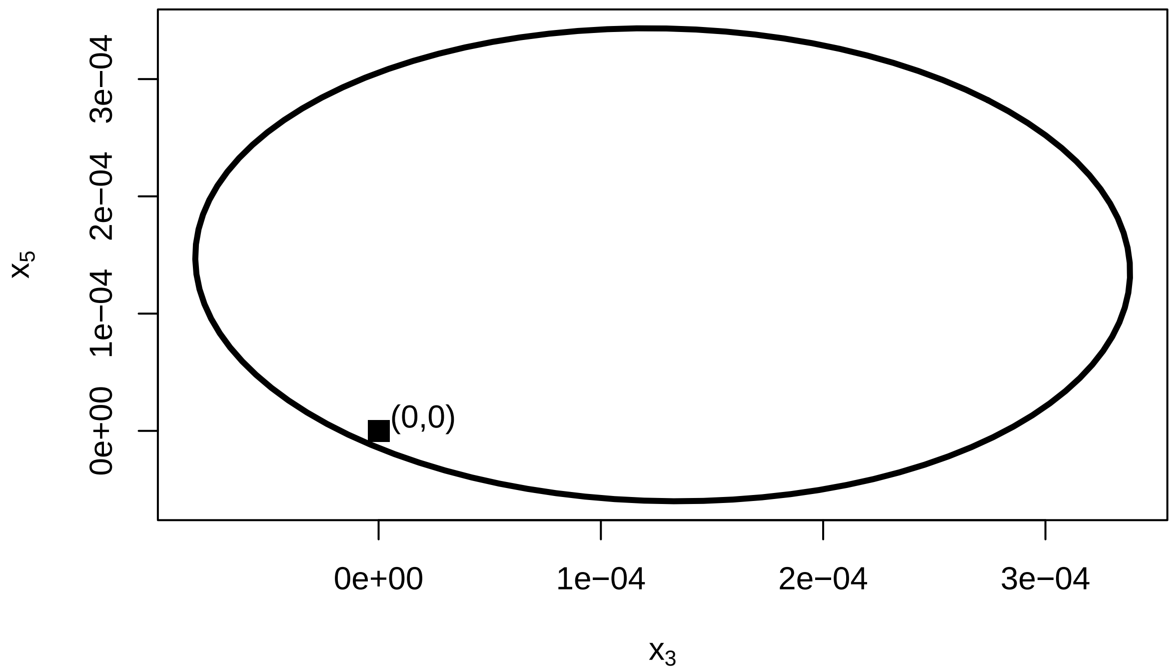
由以上結果得知 $\hat{\sigma}\sqrt{A(X^TX)^{-1}A^T} = 1.642 \times 10^{-4}$，因此 95% CI for $\beta_3 + 2\beta_5$:

$$\begin{aligned}\hat{\beta}_3 + 2\hat{\beta}_5 \mp t_{14}(0.025) \times \hat{\sigma}\sqrt{A(X^TX)^{-1}A^T} &= 1.278 \times 10^{-4} + 2 \times 1.417 \times 10^{-4} \mp 2.145 \times 1.642 \times 10^{-4} \\ &= (5.899 \times 10^{-5}, 7.632 \times 10^{-4})\end{aligned}$$

c.

繪製 β_3 及 β_5 的 95% confidence region 如下：

```
library(ellipse)
plot(ellipse(g, c(4,6)), lwd=3, type="l",xlab=expression(x[3]),ylab=expression(x[5]))
points(0, 0, cex=1.5, pch=15)
text(2e-5,1e-5,"(0,0)")
```



由 confidence region 和 hypothesis test 的對偶關係，可描述檢定問題如下：$H_0 : \beta_3 = \beta_5 = 0$, $H_1 :$ not H_0。從圖中可知原點落在橢圓內，因此沒有足夠證據拒絕 $H_0 : \beta_3 = \beta_5 = 0$。

d.

本題同等於檢定問題 $H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$, H_1 : at least one $\beta_i \neq 0, i = 1, 2, 3, 4, 5$ (overall F-test)。由第一頁的 summary report 可知 F-statistics $= 11.99$，自由度為 5 和 14，p-value=0.0001184。在顯著水準 $\alpha = 0.05$ 下，有充分證據可拒絕 H_0。亦即 95% 的 confidence region 不會包含原點 $(\beta_1, \beta_2, \beta_3, \beta_4, \beta_5) = (0, 0, 0, 0, 0)$ 在內。

e.

從資料敘述可知，可令 non-volatile solids 為 $x_6 = x_3 - x_4$，原始模型可改寫為

$$
\begin{aligned}
y &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3(x_4 + x_6) + \beta_4 x_4 + \beta_5 x_5 + \epsilon \\
&= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + (\beta_3 + \beta_4)x_4 + \beta_3 x_6 + \beta_5 x_5 + \epsilon
\end{aligned}
$$

檢定 non-volatile solids 對 response 是否有 linear effect 等價於檢定 $H_0 : \beta_3 = 0$ vs. $H_0 : \beta_3 \neq 0$

$$
\begin{aligned}
\Omega &: y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + (\beta_3 + \beta_4)x_4 + \beta_3 x_6 + \beta_5 x_5 + \epsilon \\
\omega &: y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_4 x_4 + \beta_5 x_5 + \epsilon
\end{aligned}
$$

又原始模型檢定 $H_0 : \beta_3 = 0$ vs. $H_0 : \beta_3 \neq 0$

$$
\begin{aligned}
\Omega_0 &: y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \epsilon \\
\omega_0 &: y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_4 x_4 + \beta_5 x_5 + \epsilon
\end{aligned}
$$

由於 x_3、x_4 和 x_6 之間存在線性關係，$\text{span}\{x_3, x_4\} = \text{span}\{x_4, x_6\}$，可知 $\Omega_0 = \Omega$，且 $\omega_0 = \omega$，因此，可直接套用 a 小題結果，由於 β_3 的 95% CI $(-3.714 \times 10^{-5}, 2.927 \times 10^{-4})$ 包含 0，因此不拒絕 $H_0 : \beta_3 = 0$，也就是沒有足夠證據說明 non-volatile solids 對 response 有 linear effect。

Problem 2.

匯入資料：

```
library(dplyr)
library(knitr)
data2 = read.table("http://www.stat.nthu.edu.tw/~swcheng/Teaching/stat5410/data/houseprices.txt",
                   header = T)
data2$NE = as.factor(data2$NE)
data2$Corner = as.factor(data2$Corner)
dim(data2)
```

```
## [1] 117   7
```

```
summary(data2)
```

```
##      Price           SQFT           Age           Features      NE      Corner
##  Min.   : 540   Min.   : 837   Min.   : 1.00   Min.   :0.00   0:39   0:95
##  1st Qu.: 780   1st Qu.:1280   1st Qu.: 5.75   1st Qu.:3.00   1:78   1:22
##  Median : 960   Median :1549   Median :13.00   Median :4.00
##  Mean   :1063   Mean   :1654   Mean   :14.97   Mean   :3.53
##  3rd Qu.:1200   3rd Qu.:1894   3rd Qu.:19.25   3rd Qu.:4.00
##  Max.   :2150   Max.   :3750   Max.   :53.00   Max.   :8.00
##                                NA's   :49
##       Tax
##  Min.   : 223.0
##  1st Qu.: 600.0
##  Median : 731.0
##  Mean   : 793.5
##  3rd Qu.: 919.0
##  Max.   :1765.0
##  NA's   :10
```

- 此筆資料共 117 筆觀測值，7 個變數 (1 response：*Price*, 6 predictors)
- *NE, Corner* 皆為 nominal variables，建模時需採用 dummy variable 方式
- *Age, Tax* 變數中皆有 NA 值，*Age* 中甚至多達 49 個

a.

Strategy (1) 將所有 *Age* 有缺失值的資料都去除：

```
data2.1 = data2 %>% filter(!is.na(Age))
dim(data2.1)
```

```
## [1] 68  7
```

```
summary(data2.1)
```

```
##      Price             SQFT           Age           Features       NE     Corner
##  Min.   : 580.0   Min.   : 970   Min.   : 1.00   Min.   :1.000   0:26   0:52
##  1st Qu.: 893.8   1st Qu.:1400   1st Qu.: 5.75   1st Qu.:3.000   1:42   1:16
##  Median :1049.5   Median :1690   Median :13.00   Median :4.000
##  Mean   :1163.9   Mean   :1746   Mean   :14.97   Mean   :3.971
##  3rd Qu.:1277.2   3rd Qu.:1922   3rd Qu.:19.25   3rd Qu.:4.000
##  Max.   :2150.0   Max.   :2931   Max.   :53.00   Max.   :8.000
##
##       Tax
##  Min.   : 398.0
##  1st Qu.: 673.5
##  Median : 810.0
##  Mean   : 872.7
##  3rd Qu.:1003.5
##  Max.   :1765.0
##  NA's   :2
```

資料僅剩下 68 筆，可能會因為資料數過少造成參數估計的 standard error 過大，進而影響變數對於模型的顯著性

```
summary(lm(Price~.,data2.1))
```

```
##
## Call:
## lm(formula = Price ~ ., data = data2.1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -411.40  -71.20   -6.15   65.79  581.73
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -33.3945   101.2900  -0.330 0.742801
## SQFT          0.3965     0.1022   3.880 0.000266 ***
## Age           0.2137     2.1439   0.100 0.920924
## Features     16.9495    19.5711   0.866 0.389972
## NE1         -34.1625    50.6735  -0.674 0.502836
## Corner1     -68.7001    52.9532  -1.297 0.199553
## Tax           0.5430     0.1703   3.189 0.002286 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 171.3 on 59 degrees of freedom
##   （因為不存在，2 個觀察量被刪除了）
## Multiple R-squared:  0.8371, Adjusted R-squared:  0.8205
## F-statistic: 50.52 on 6 and 59 DF,  p-value: < 2.2e-16
```

- 即使將變數 Age 加入模型中也並不顯著
- 變數 $Corner$ 呈現不顯著

Strategy (2) 直接將變數 Age 去除：

```
data2.2 = data2 %>% select(-Age)
dim(data2.2)
```

```
## [1] 117   6
```

```
summary(data2.2)
```

```
##      Price            SQFT          Features         NE      Corner       Tax
##  Min.   : 540   Min.   : 837   Min.   :0.00   0:39   0:95   Min.   : 223.0
##  1st Qu.: 780   1st Qu.:1280   1st Qu.:3.00   1:78   1:22   1st Qu.: 600.0
##  Median : 960   Median :1549   Median :4.00                 Median : 731.0
##  Mean   :1063   Mean   :1654   Mean   :3.53                 Mean   : 793.5
##  3rd Qu.:1200   3rd Qu.:1894   3rd Qu.:4.00                 3rd Qu.: 919.0
##  Max.   :2150   Max.   :3750   Max.   :8.00                 Max.   :1765.0
##                                                             NA's   :10
```

資料依舊維持 117 筆觀測值，雖然變數 Tax 中還是有 10 筆缺失值，但對於整筆資料的個數所佔比例並不高，在分析時可以直接刪除

```
summary(lm(Price~.,data2.2))
```

```
##
## Call:
## lm(formula = Price ~ ., data = data2.2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -544.22  -74.05  -15.03   68.34  615.26
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  77.6954    62.1231   1.251   0.2139
## SQFT          0.2666     0.0620   4.301 3.93e-05 ***
## Features     13.8581    13.5727   1.021   0.3097
## NE           -3.3995    36.4875  -0.093   0.9260
## Corner1     -89.1245    42.4246  -2.101   0.0382 *
## Tax           0.6627     0.1097   6.042 2.57e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 171.9 on 101 degrees of freedom
##   （因為不存在，10 個觀察量被刪除了）
## Multiple R-squared:  0.8091, Adjusted R-squared:  0.7996
## F-statistic:  85.6 on 5 and 101 DF,  p-value: < 2.2e-16
```

⇒ 此時變數 $Corner$ 呈現為顯著
因此在本資料選擇 **Strategy (2)** 是一個較好的方式，但不見得對所有資料都應採取此種策略。

b.

建構模型

$$Price \ = \ \beta_0 + \beta_1 \ SQFT + \beta_2 \ Features + \beta_3 \ NE + \beta_4 \ Corner + \beta_5 \ Taxes + \epsilon$$

```
data2.2 = na.omit(data2.2)
fit2.2 = lm(Price ~ ., data2.2)
summary(fit2.2)
```

```
##
## Call:
## lm(formula = Price ~ ., data = data2.2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -544.22  -74.05  -15.03   68.34  615.26
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  77.6954    62.1231   1.251   0.2139
## SQFT          0.2666     0.0620   4.301 3.93e-05 ***
## Features     13.8581    13.5727   1.021   0.3097
## NE1          -3.3995    36.4875  -0.093   0.9260
## Corner1     -89.1245    42.4246  -2.101   0.0382 *
## Tax           0.6627     0.1097   6.042 2.57e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 171.9 on 101 degrees of freedom
## Multiple R-squared:  0.8091, Adjusted R-squared:  0.7996
## F-statistic:  85.6 on 5 and 101 DF,  p-value: < 2.2e-16
```

計算 $(1-\alpha)100\%$ confidence interval for each coefficient β_i：

$$\hat{\beta}_i \ \pm \ t_{n-p}(\alpha/2) \times s.e.\left(\hat{\beta}_i\right) \ , \text{ where } \ s.e.\left(\hat{\beta}_i\right) \ = \ \hat{\sigma}\sqrt{(X^T X)^{-1}_{ii}}$$

各係數的 95% & 99% confidence interval 呈現如下：

```
kable(confint(fit2.2)[-1,],digits = 3)
```

	2.5 %	97.5 %
SQFT	0.144	0.390
Features	-13.067	40.783
NE1	-75.781	68.982
Corner1	-173.284	-4.965
Tax	0.445	0.880

```
kable(confint(fit2.2,level=0.99)[-1,],digits = 3)
```

	0.5 %	99.5 %
SQFT	0.104	0.429
Features	-21.776	49.492
NE1	-99.193	92.394
Corner1	-200.505	22.257
Tax	0.375	0.951

p-value(β_{Corner}) $= 0.0382 \Rightarrow 0.01 < 0.0382 < 0.05$：代表在 0.05 顯著水準下 β_{Corner} 和 0 有顯著差異，但在顯著水準 0.01 時則沒有；同樣可以觀察 β_{Corner} 在 95% 和 99% 下的 confidence intervals：前者不會包含 0，後者則有包含，這個和我們透過 p-value 所得結論一致。

c.

針對題目我們建構 future observation 的預測值及信賴區間，根據 b. 所建構模型和配適結果，可計算預測值

$$\hat{y} = x_0^T\hat{\beta} = \hat{\beta}_0 + \hat{\beta}_1\ SQFT + \hat{\beta}_2\ Features + \hat{\beta}_3\ NE + \hat{\beta}_4\ Corner + \hat{\beta}_5\ Taxes,$$

和信賴區間

$$x_0\hat{\beta} \pm t_{101}(\alpha/2) \times \hat{\sigma}\sqrt{x_0^T(X^TX)^{-1}x_0 + 1}.$$

```
x0 <- data.frame(SQFT=2500, Features=5, NE="1", Corner="1", Tax=1200)
new.pred2.2 <- t(as.numeric(c(1,x0)))%*%fit2.2$coefficients
d.m <- as.matrix(cbind(1,data2.2[,2:3],as.numeric(data2.2[,4])-1,
                       as.numeric(data2.2[,5])-1, data2.2[,6]))
cat("Prediction value: ",new.pred2.2,"\n")
```

```
## Prediction value:  1516.361
```

```
cat("Confidense interval: (",
    new.pred2.2+qt(0.025,101)*
      summary(fit2.2)$sigma*
      sqrt(t(as.numeric(c(1,x0)))%*%solve(t(d.m)%*%d.m)%*%as.numeric(c(1,x0))+1),", ",
    new.pred2.2+qt(0.975,101)*
      summary(fit2.2)$sigma*
      sqrt(t(as.numeric(c(1,x0)))%*%solve(t(d.m)%*%d.m)%*%as.numeric(c(1,x0))+1),")"
    ,"\n",sep="")
```

```
## Confidense interval: (1162.761, 1869.961)
```

亦可用 predict() 計算其信賴區間。

```
predict(fit2.2, newdata=x0,interval="predict")
```

```
##        fit      lwr      upr
## 1 1516.361 1162.761 1869.961
```

另外，針對滿足 SQFT=2500, Features=5, NE=1, Corner=1, Taxes=1200 條件的所有房子之平均房價亦可計算其估計平均值

$$\widehat{E(y)} = x_0^T\hat{\beta} = \hat{\beta}_0 + \hat{\beta}_1\ SQFT + \hat{\beta}_2\ Features + \hat{\beta}_3\ NE + \hat{\beta}_4\ Corner + \hat{\beta}_5\ Taxes,$$

和所夠成信賴區間

$$x_0\hat{\beta} \pm t_{101}(\alpha/2) \times \hat{\sigma}\sqrt{x_0^T(X^TX)^{-1}x_0}.$$

```
x0 <- data.frame(SQFT=2500, Features=5, NE="1", Corner="1", Tax=1200)
new.pred2.2 <- t(as.numeric(c(1,x0)))%*%fit2.2$coefficients
d.m <- as.matrix(cbind(1,data2.2[,2:3],as.numeric(data2.2[,4])-1,
                       as.numeric(data2.2[,5])-1, data2.2[,6]))
cat("Prediction value: ",new.pred2.2,"\n")
```

```
## Prediction value:  1516.361
```

```
cat("Confidense interval: (",
    new.pred2.2+qt(0.025,101)*
      summary(fit2.2)$sigma*
      sqrt(t(as.numeric(c(1,x0)))%*%solve(t(d.m)%*%d.m)%*%as.numeric(c(1,x0))),", ",
    new.pred2.2+qt(0.975,101)*
      summary(fit2.2)$sigma*
      sqrt(t(as.numeric(c(1,x0)))%*%solve(t(d.m)%*%d.m)%*%as.numeric(c(1,x0))),")"
    ,"\n",sep="")
```

```
## Confidense interval: (1422.713, 1610.009)
```

```
predict(fit2.2, newdata=x0,interval="confidence")
```

```
##        fit      lwr      upr
## 1 1516.361 1422.713 1610.009
```

d.

假設對於預測只有 SQFT=2500 的資訊時，是將 SQFT=2500 的所有房子當成一個 sub-population 並以此預測其 mean response 或其所產生的 future observation，公式表示下可視為 conditional expectation，表示為

$$\begin{aligned} E(y|SQFT=2500) =& \beta_0 + \beta_1(SQFT=2500) + \beta_2 E(Features|SQFT=2500) + \beta_3 E(NE|SQFT=2500) \\ &+ \beta_4 E(Corner|SQFT=2500) + \beta_5 E(Taxes|SQFT=2500), \end{aligned}$$

以下提供兩種做法：

1. 對於除 SQFT 以外之變數透過將 $E(x_p)$ 取代 $E(x_p|SQFT=2500)$ 代入來替代其他變數的資訊，意即將其他解釋變數帶入觀測資料算得的變數平均數、中位數或眾數（Feature 變數帶中位數，類別變數帶眾數，Tax 變數帶平均數）當作條件期望值估計式，而參數部分使用 b. 的參數估計結果帶入，算其預測值及預測區間

```
x1 <- data.frame(SQFT=2500, Features=median(data2.2$Features),
                 NE=c("0","1")[which.max(table(data2.2$NE))],
                 Corner=c("0","1")[which.max(table(data2.2$Corner))],
                 Tax=mean(data2.2$Tax))
predict(fit2.2, newdata=x1,interval="predict")
```

```
##       fit      lwr      upr
## 1 1322.22 962.7517 1681.688
```

此做法在所有解釋變數的 correlation 都低的時候, 是可行的, 但若共線性過強時, 則不適合。

2. 重新建立僅有 SQFT 變數之模型

$$Price \ = \ \beta_0 + \beta_1 \ SQFT + \epsilon,$$

雖然此模型中沒有明顯做 $E(x_p|SQFT = 2500)$ 的計算，但由 Lecture note p.5-14 的公式可知，此 fitted model 是有能力” 部分” 解釋其它變數所造成的影響，儘管僅有 SQFT 當解釋變數，因此較方法 1. 更合適。透過此方法計算其預測值及預測區間

```
fit2.3 = lm(Price ~ SQFT, data2.2)
predict(fit2.3, newdata=data.frame(SQFT=2500),interval="predict")
```

```
##        fit      lwr      upr
## 1 1584.563 1166.205 2002.921
```

針對此做法預測之結果，雖然因與 c. 所對應的 sub-populations 不同，所以無法比較預測值，但對於其預測區間長度，理論上應是 c. 預測區間長度較短，因 c. 所使用的資訊較多，預期應會有較為準確的結果。