HW6

（一）匯入資料

```r
## read my flie
setwd("/Users/cindychen/Desktop/數據科學概論/HW/")
data <- read_csv("insurance.csv")
str(data)
```

```
## spec_tbl_df [1,338 x 7] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ age     : num [1:1338] 19 18 28 33 32 31 46 37 37 60 ...
##  $ sex     : chr [1:1338] "female" "male" "male" "male" ...
##  $ bmi     : num [1:1338] 27.9 33.8 33 22.7 28.9 ...
##  $ children: num [1:1338] 0 1 3 0 0 0 1 3 2 0 ...
##  $ smoker  : chr [1:1338] "yes" "no" "no" "no" ...
##  $ region  : chr [1:1338] "southwest" "southeast" "southeast" "northwest" ...
##  $ charges : num [1:1338] 16885 1726 4449 21984 3867 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   age = col_double(),
##   ..   sex = col_character(),
##   ..   bmi = col_double(),
##   ..   children = col_double(),
##   ..   smoker = col_character(),
##   ..   region = col_character(),
##   ..   charges = col_double()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

(二）從資料中隨機取樣

```r
## sample from the data
set.seed(122)
taken <- sample(1:nrow(data),500)
test1 <- data[taken,]
```

（三）利用取樣過後的資料做分析

單變量迴歸分析

1.
將歲數與費用的變數做單變量的回歸。並將回歸分兩次跑，分別為沒有 β_0 與有 β_0，從 summary 可以看出，在有截距項的回歸中（model1），β_0 並不顯著，且再去除掉截距項後的回歸模型 r-square 與調整後 r-square 都大幅提升，所以取除掉 β_0 的模型較適合此回歸。

```r
## make regression of age and charges
model1 <- lm(charges~age,test1)
```

```r
model1_without <- lm(charges~age -1 , test1)
summary(model1)
```

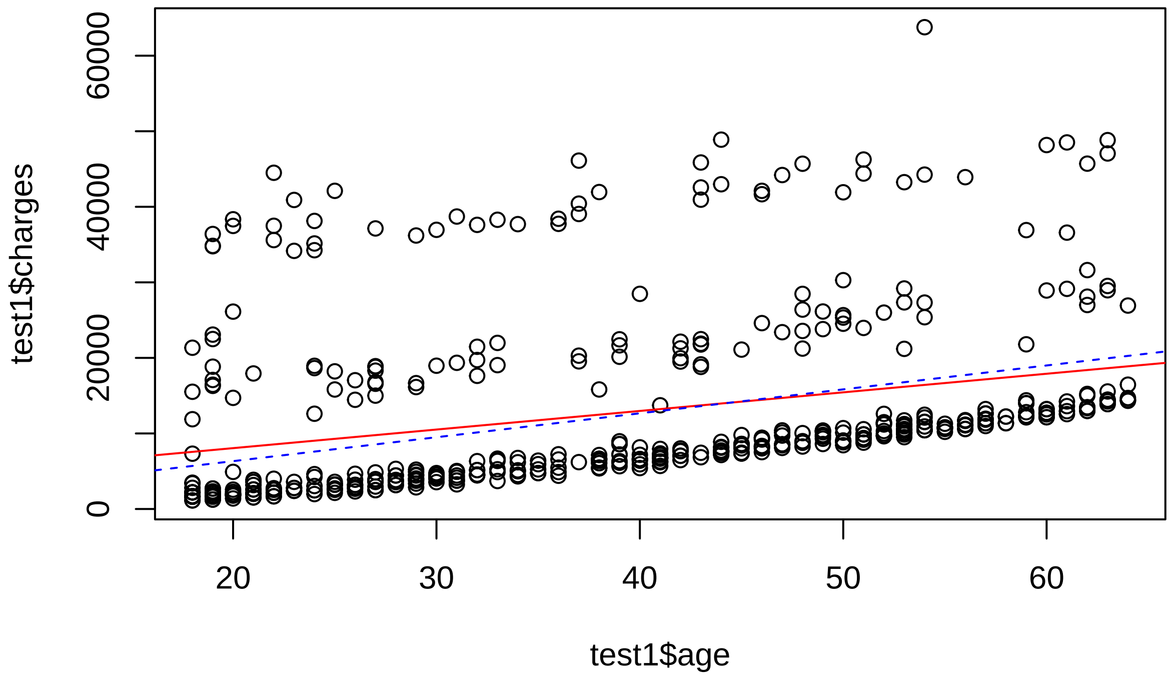
```
##
## Call:
## lm(formula = charges ~ age, data = test1)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -7591  -6275  -5494   4471  47347
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3133.17    1479.20   2.118   0.0347 *
## age           246.12      35.35   6.962 1.06e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11090 on 498 degrees of freedom
## Multiple R-squared:  0.08869,    Adjusted R-squared:  0.08686
## F-statistic: 48.47 on 1 and 498 DF,  p-value: 1.064e-11
```

```r
summary(model1_without)
```

```
##
## Call:
## lm(formula = charges ~ age - 1, data = test1)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -7390  -6130  -4983   5055  46671
##
## Coefficients:
##     Estimate Std. Error t value Pr(>|t|)
## age    316.7       11.9   26.61   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11130 on 499 degrees of freedom
## Multiple R-squared:  0.5867, Adjusted R-squared:  0.5858
## F-statistic: 708.2 on 1 and 499 DF,  p-value: < 2.2e-16
```

將圖形畫出來後可以更明顯的看出，紅色線條為有 β_0 的模型，藍色線段為沒有 β_0 的模型，第二條線較適合這模型的點的分佈，雖然有些點的位置較高，推測可能是有不同變數影響，所以才會造成分佈有些許的不同的狀況。

```r
plot(test1$age,test1$charges)
abline(model1 ,lty = 1 ,col = "red")
abline(model1_without,lty = 2, col = "blue")
```



2.

將 bmi 值與費用做回歸分析。同樣將回歸分兩次跑，分別是有 β_0 的模型與去除掉 β_0 的模型，將兩個模型的數據比較，可以發現有 β_0 的模型的截距項較不具顯著性，且上一個的狀況一樣，有 β_0 的模型的 r-square 與調整後 r-square 都呈現較低的結果，所以可以從此推斷取除掉 β_0 的模型較適合這個回歸。

```r
## make regression of bmi and charges
model2 <- lm(charges~bmi, test1)
model2_without <- lm(charges~bmi-1, test1)
summary(model2)
```

```
##
## Call:
## lm(formula = charges ~ bmi, data = test1)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -16234  -7734  -3418   4890  43140
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1434.74    2494.14  -0.575    0.565
## bmi           465.40      79.68   5.841 9.37e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11240 on 498 degrees of freedom
## Multiple R-squared:  0.06412,    Adjusted R-squared:  0.06224
## F-statistic: 34.12 on 1 and 498 DF,  p-value: 9.371e-09
```

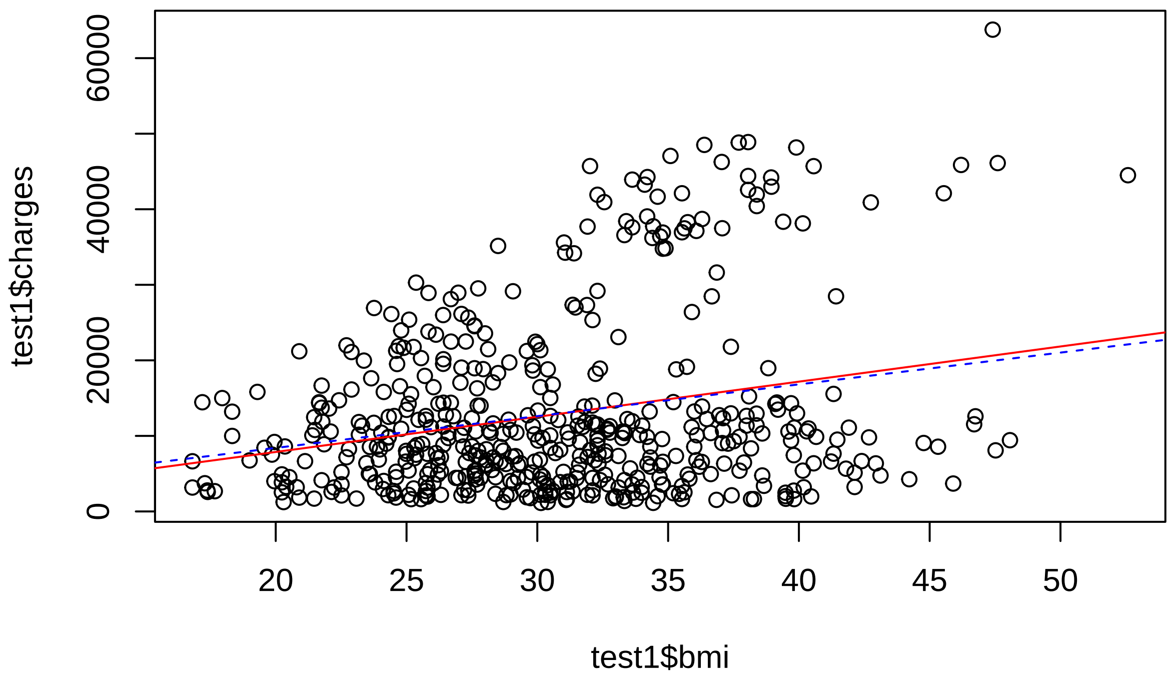
```r
summary(model2_without)
```

```
##
## Call:
## lm(formula = charges ~ bmi - 1, data = test1)
##
```

```
## Residuals:
##    Min     1Q Median     3Q    Max
## -15608  -7869  -3502   4654  43834
##
## Coefficients:
##     Estimate Std. Error t value Pr(>|t|)
## bmi   420.51      16.05    26.2   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11240 on 499 degrees of freedom
## Multiple R-squared:  0.579,  Adjusted R-squared:  0.5782
## F-statistic: 686.3 on 1 and 499 DF,  p-value: < 2.2e-16
```

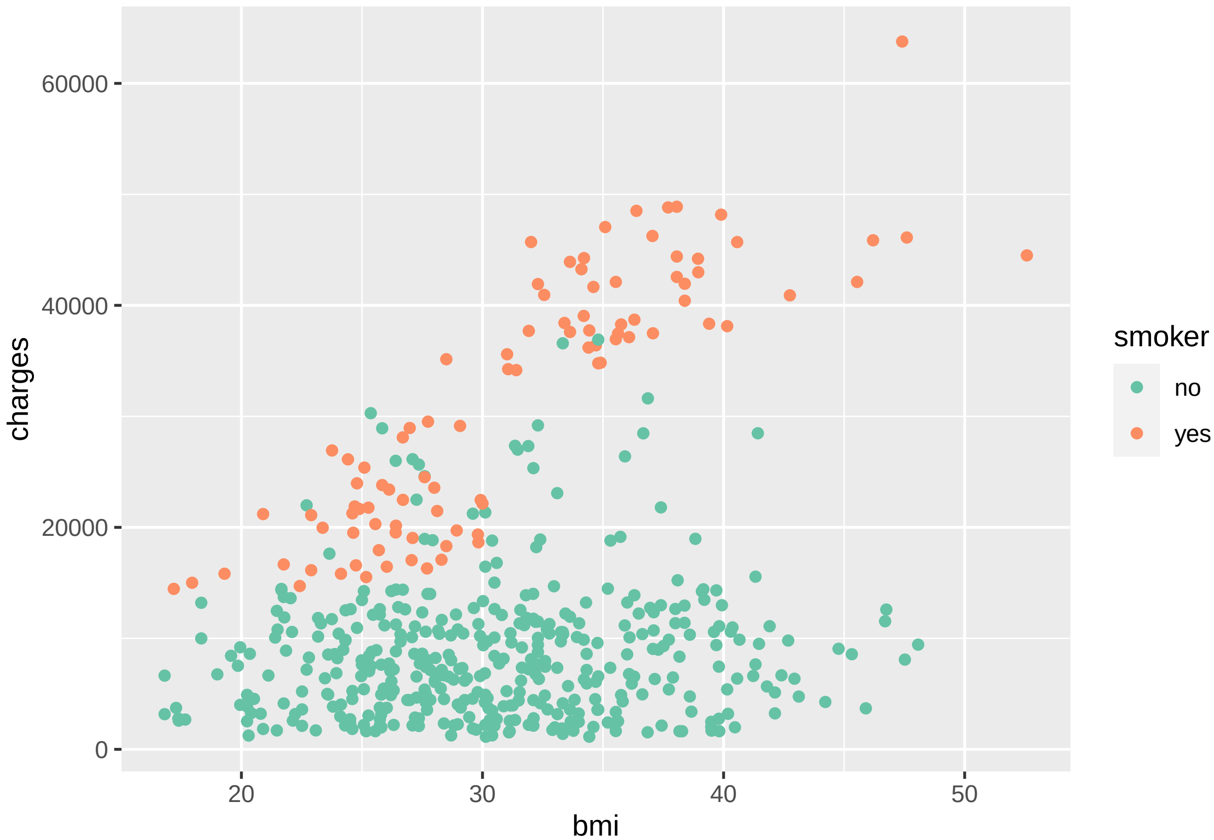
從圖形來判斷，紅色線條為有 β_0 的模型，藍色線條為去除掉 β_0 的模型，可以看出兩著的線條有些許的差異，但因為點的分佈較不明確，所以很難利用圖形推斷哪一個模型較為適切。利用圖形的散佈情況，我們可以發現他呈現兩種不同的分佈，有些點都較貼平水平線，有些則會呈現上升的趨勢。

```
plot(test1$bmi , test1$charges)
abline(model2 ,lty = 1 ,col = "red")
abline(model2_without,lty = 2, col = "blue")
```



因次，我們可以將其他因素帶入並觀察圖形的狀況與點的散佈狀況。
首先，先將有沒有抽菸這個變因帶入，可以發現，點的散佈情況會跟有沒有抽菸有所關聯，有抽菸的民眾保險收費會隨著 bmi 的增加而跟著增加，而沒抽菸的民眾，bmi 的高低對保險費用的收取較無影響。

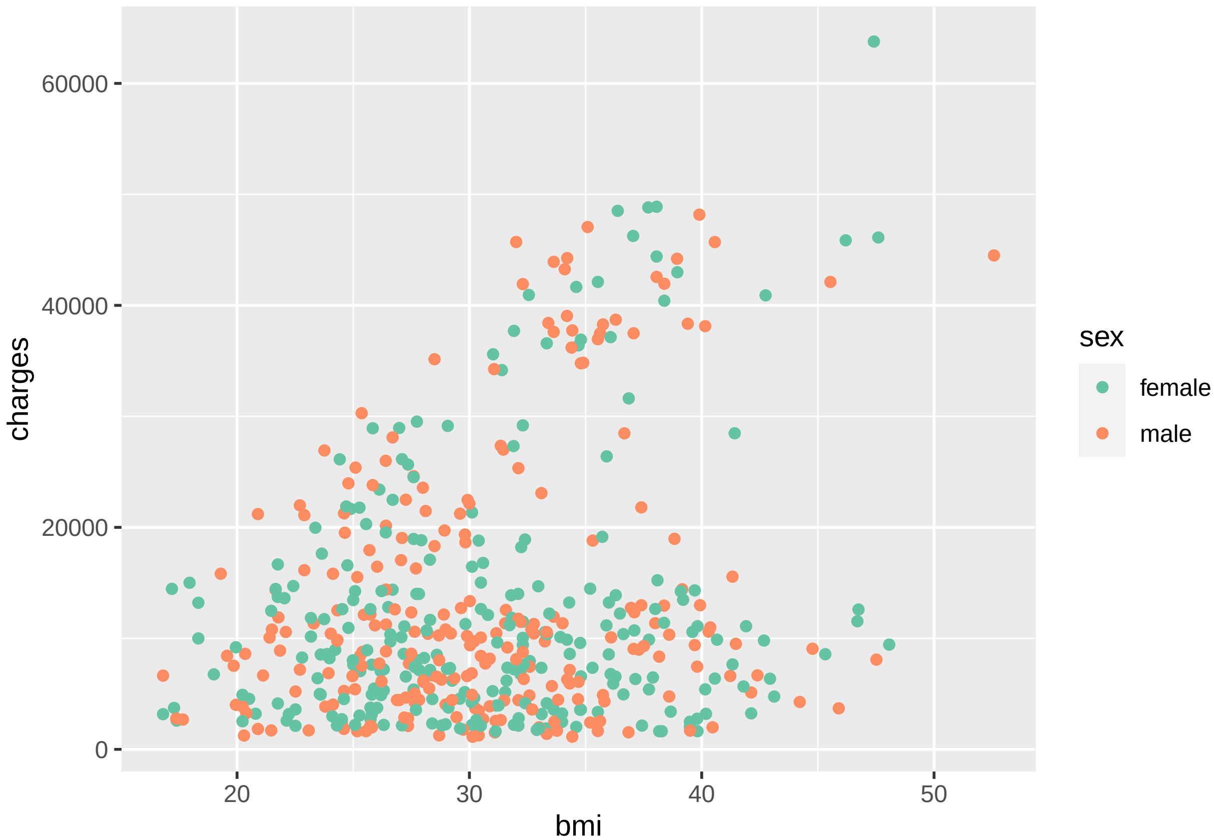
```
ggplot(test1 , aes(x = bmi, y = charges ))+
  geom_point(aes(colour = smoker))+
  scale_color_brewer(palette = "Set2")
```



再

利用性別當作變因進去看是否有關聯，從點的散佈情況可以看出，其較不受性別因素影響。

```
ggplot(test1 , aes(x = bmi, y = charges ))+
  geom_point(aes(colour = sex))+
  scale_color_brewer(palette = "Set2")
```



利用有沒有抽菸這個變數將 data 分割成兩半。

```
## and split the dataset by variable smoker
NS <- split(data,data$smoker)
```

將 bmi 與費用的回歸再跑一次，分為有 β_0 與沒有 β_0 的模型，並將預測的 model 放入圖形中檢視，在沒有抽菸的 data 中，沒有 β_0 的模型 r-square 與調整後 r-square 都較高，與上面的回歸不同的是，這次有 β_0 的模型截距項的顯著性較上面為高，且 r-square 也較上面模型跑出來的結果為高。

```
## run the regression line on two different dataset above
## With nonsmoker
model3_1 <- lm(charges~bmi,NS$no)
model3_1W <- lm(charges~bmi-1,NS$no)
summary(model3_1)
```

```
##
## Call:
## lm(formula = charges ~ bmi, data = NS$no)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -9144  -4360  -1009   2922  28131
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5879.42     947.48   6.205 7.81e-10 ***
## bmi            83.35      30.33   2.748  0.00609 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 5975 on 1062 degrees of freedom
## Multiple R-squared:  0.007062,   Adjusted R-squared:  0.006127
## F-statistic: 7.553 on 1 and 1062 DF,  p-value: 0.006091
```

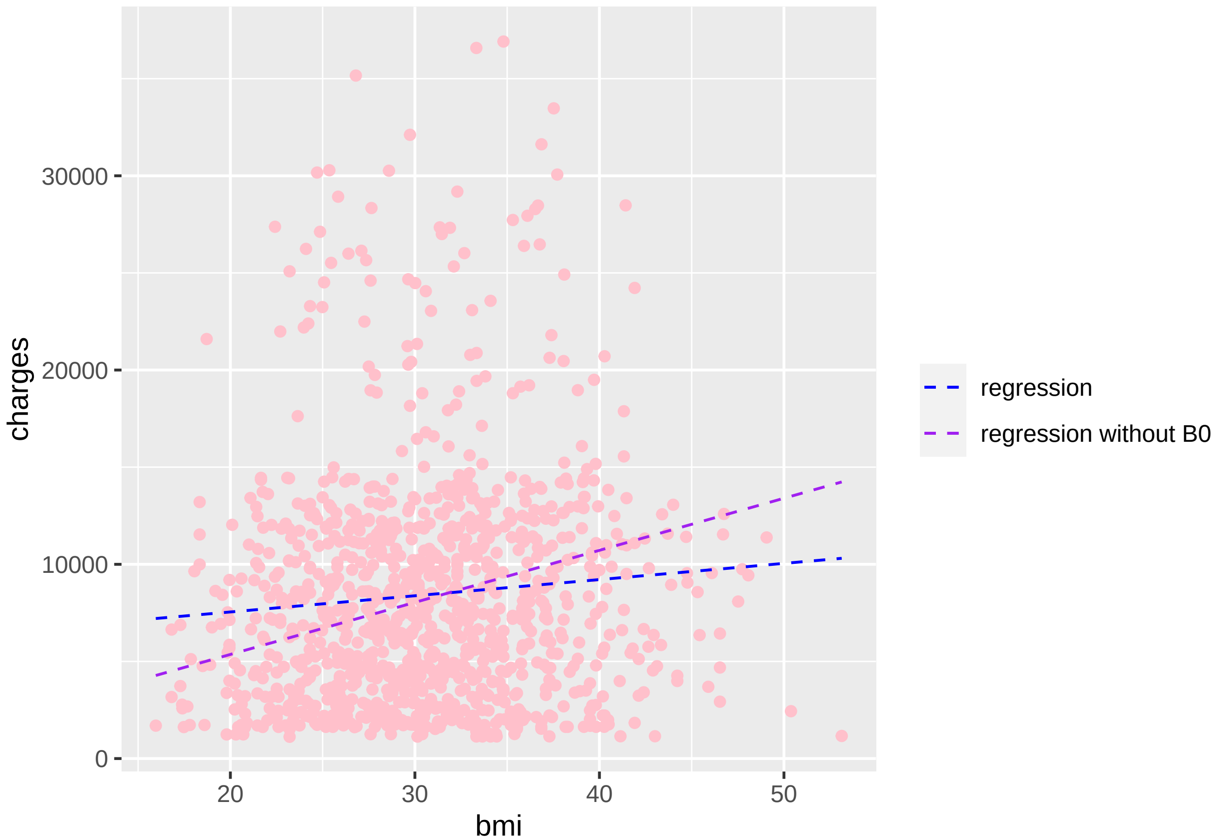
```
summary(model3_1W)
```

```
##
## Call:
## lm(formula = charges ~ bmi - 1, data = NS$no)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13075.0  -3965.0   -803.2   3008.1  27977.9
##
## Coefficients:
##     Estimate Std. Error t value Pr(>|t|)
## bmi  267.994      5.966   44.92   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6080 on 1063 degrees of freedom
## Multiple R-squared:  0.6549, Adjusted R-squared:  0.6546
## F-statistic:  2018 on 1 and 1063 DF,  p-value: < 2.2e-16
```

將預測後的模型帶入圖形中，可以發現兩者的差異。

```
predicted_bmi <- data.frame(charges = predict(model3_1,NS$no),
                            bmi = NS$no$bmi)
predicted_bmiw <- data.frame(charges = predict(model3_1W,NS$no),
                             bmi = NS$no$bmi)

ggplot(NS$no,aes(x = bmi,y = charges))+
  geom_point(col = "pink")+
  geom_line(data = predicted_bmi , aes(x = bmi, y = charges,color = "regression") , lty = 2)+
  geom_line(data = predicted_bmiw , aes(x = bmi, y = charges ,color = "regression without B0" ), lty = 2
  scale_colour_manual("", breaks = c("regression","regression without B0"),
                      values = c("blue","purple"))
```




```
##residuals(model3_1)
##residuals(model3_1W)
```

換成有抽菸的資料跑兩種回歸，在有 β_0 與沒有 β_0 的模型下，變數都具有顯著性，但是在去除掉截距項的模型，其 r-square 與調整後 r-square 較高。

```
## with smoker
model3_2 <- lm(charges~bmi,NS$yes)
model3_2W <- lm(charges~bmi-1,NS$yes)
summary(model3_2)
```

```
##
## Call:
## lm(formula = charges ~ bmi, data = NS$yes)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -19768.0  -4487.9     34.4   3263.9  31055.9
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -13186.58    2052.88  -6.423 5.93e-10 ***
## bmi           1473.11      65.48  22.496  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6837 on 272 degrees of freedom
## Multiple R-squared:  0.6504, Adjusted R-squared:  0.6491
```

```
## F-statistic: 506.1 on 1 and 272 DF,  p-value: < 2.2e-16
```

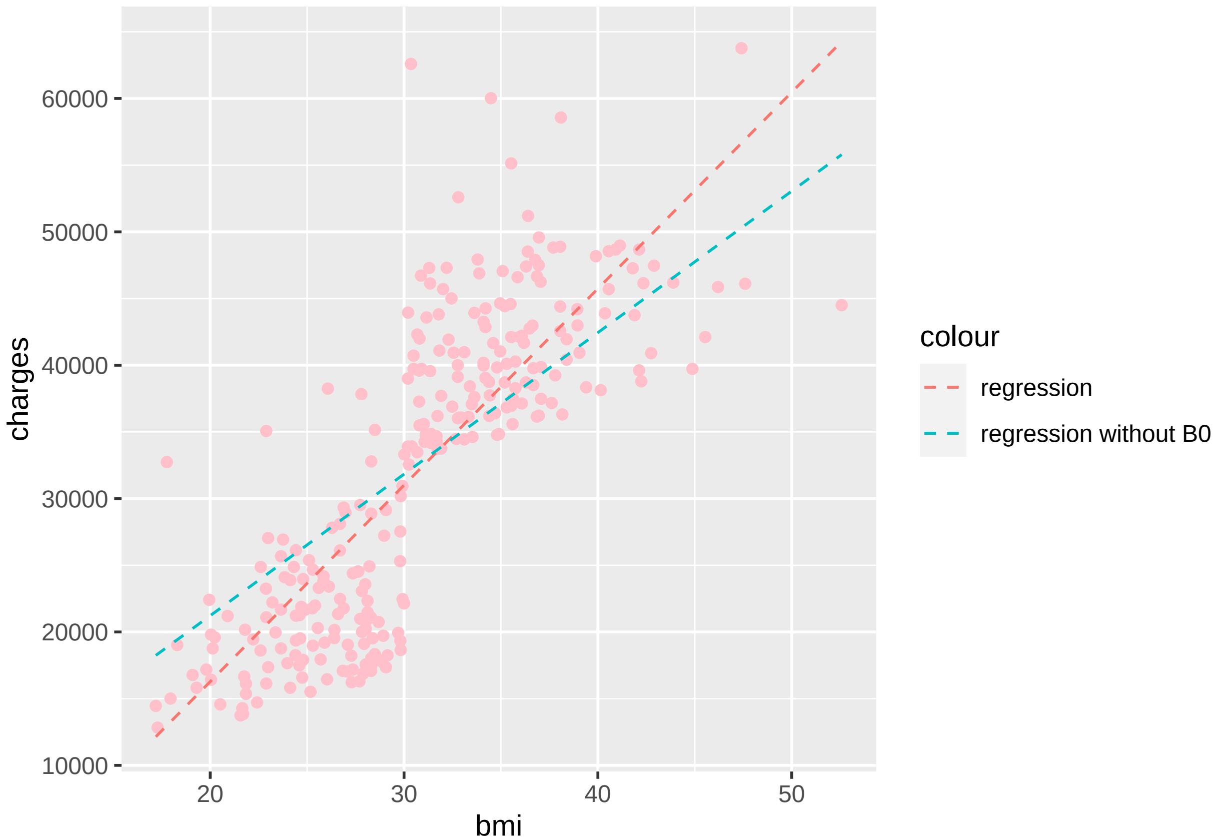
```
summary(model3_2W)
```

```
##
## Call:
## lm(formula = charges ~ bmi - 1, data = NS$yes)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13492.8  -5983.4   -560.5   3936.2  30378.6
##
## Coefficients:
##     Estimate Std. Error t value Pr(>|t|)
## bmi  1061.08      14.11   75.19   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7323 on 273 degrees of freedom
## Multiple R-squared:  0.9539, Adjusted R-squared:  0.9538
## F-statistic:  5653 on 1 and 273 DF,  p-value: < 2.2e-16
```

將兩者的預測放入圖形中，可以發現兩者的差異。

```
predicted_2bmi <- data.frame(charges = predict(model3_2,NS$yes),
                             bmi = NS$yes$bmi)
predicted_2bmiw <- data.frame(charges = predict(model3_2W,NS$yes),
                              bmi = NS$yes$bmi)

ggplot(NS$yes,aes(x = bmi,y = charges))+
  geom_point(col = "pink")+
  geom_line(data = predicted_2bmi , aes(x = bmi, y = charges,color = "regression") , lty = 2)+
  geom_line(data = predicted_2bmiw , aes(x = bmi, y = charges,color = "regression without B0" ), lty = 1
```




```
scale_colour_manual("", breaks = c("regression","regression without B0"),
                    values = c("blue","purple"))
```

```
## <ggproto object: Class ScaleDiscrete, Scale, gg>
##     aesthetics: colour
##     axis_order: function
##     break_info: function
##     break_positions: function
##     breaks: regression regression without B0
##     call: call
##     clone: function
##     dimension: function
##     drop: TRUE
##     expand: waiver
##     get_breaks: function
##     get_breaks_minor: function
##     get_labels: function
##     get_limits: function
##     guide: legend
##     is_discrete: function
##     is_empty: function
##     labels: waiver
##     limits: NULL
##     make_sec_title: function
##     make_title: function
##     map: function
##     map_df: function
```

```
##     n.breaks.cache: NULL
##     na.translate: TRUE
##     na.value: grey50
##     name:
##     palette: function
##     palette.cache: NULL
##     position: left
##     range: <ggproto object: Class RangeDiscrete, Range, gg>
##         range: NULL
##         reset: function
##         train: function
##         super:  <ggproto object: Class RangeDiscrete, Range, gg>
##     rescale: function
##     reset: function
##     scale_name: manual
##     train: function
##     train_df: function
##     transform: function
##     transform_df: function
##     super:  <ggproto object: Class ScaleDiscrete, Scale, gg>
```

```
## residuals(model3_2)
## residuals(model3_2W)
```

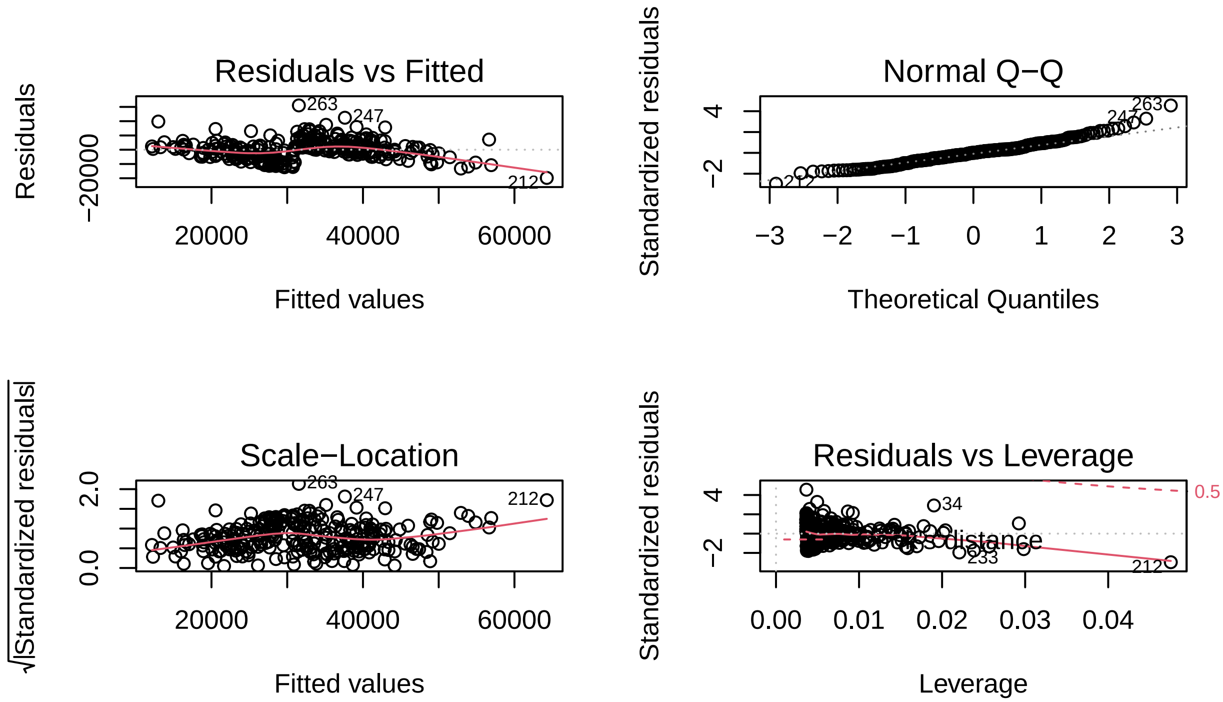
Kolmogorov-Smirnov test

利用 Kolmogorov-Smirnov test 檢查 residuals 的分佈是不是呈現常態分佈，從 test 中可以看出 p-value 小於 0，所以我們可以拒絕 H0 的假設，也就是 residuals 不呈常態分佈。
利用 plot 後的圖判斷，可以從 residuals vs fitted 圖看出，點的散佈呈現一種莫名其妙的圖形，且他的標準化後 residuals 有到 10000 多，所以可以知道他可能不呈常態分佈。

```
## check the residual of model3_2 is normal distribution with mean 0 and variance 1
ks.test(residuals(model3_2),"pnorm")
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  residuals(model3_2)
## D = 0.5073, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

```
par(mfrow = c(2,2))
plot(model3_2)
```



將 bmi 值取平方與三次方並放入回歸中觀察，可以發現，將平方項與三次方項放入後，r-square 與調整後 r-square 都提高了，且兩個變數都具有顯著性。

```r
## run the regression of bmi and charges with square bmi and cube bmi
b2 <- NS$yes$bmi^2
model3_3 <- lm(charges~bmi+b2,NS$yes)
b3 <- NS$yes$bmi^3
model3_4 <- lm(charges~bmi+b2+b3,NS$yes)
summary(model3_3)
```

```
##
## Call:
## lm(formula = charges ~ bmi + b2, data = NS$yes)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13086.0  -4103.1   -229.8   3886.4  30134.3
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -35946.608   7487.448  -4.801 2.61e-06 ***
## bmi           2970.914    478.826   6.205 2.04e-09 ***
## b2             -23.642      7.489  -3.157  0.00178 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6727 on 271 degrees of freedom
## Multiple R-squared:  0.6628, Adjusted R-squared:  0.6603
## F-statistic: 266.4 on 2 and 271 DF,  p-value: < 2.2e-16
```

```r
summary(model3_4)
```

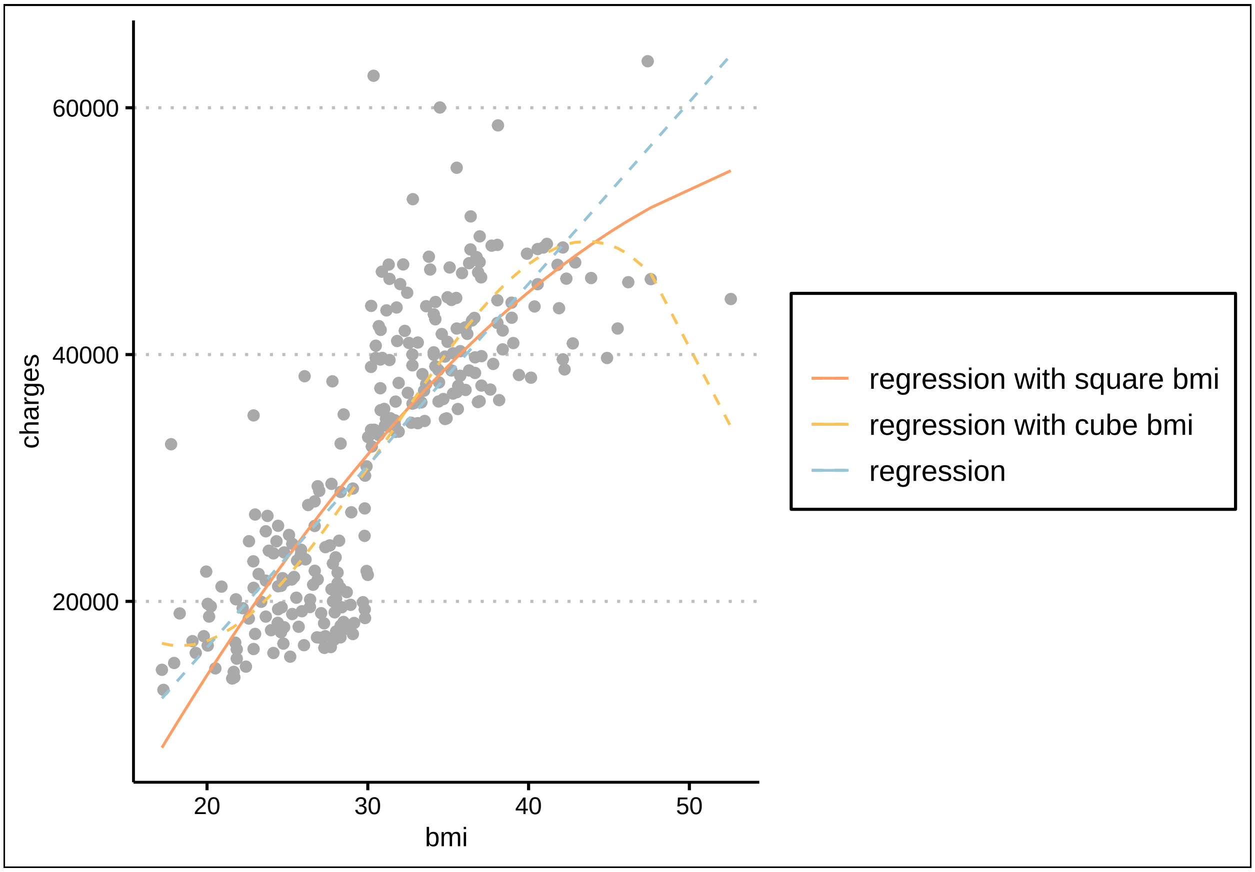
```
##
```

```
## Call:
## lm(formula = charges ~ bmi + b2 + b3, data = NS$yes)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11871.0  -3916.7   -347.8   3341.7  31049.8
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 92887.1193 23536.5063   3.947 0.000101 ***
## bmi         -9698.1111  2253.2233  -4.304 2.35e-05 ***
## b2            375.4257    69.8869   5.372 1.68e-07 ***
## b3             -4.0363     0.7032  -5.740 2.54e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6362 on 270 degrees of freedom
## Multiple R-squared:  0.6995, Adjusted R-squared:  0.6961
## F-statistic: 209.5 on 3 and 270 DF,  p-value: < 2.2e-16
```

將預測的模型放入圖形中觀察。

```
predict_3bmi <- data.frame(charges = predict(model3_3,NS$yes),
                           bmi = NS$yes$bmi)
predict_4bmi <- data.frame(charges = predict(model3_4,NS$yes),
                           bmi = NS$yes$bmi)

ggplot(NS$yes,aes(x = bmi , y = charges))+
  geom_point(col = "darkgrey")+
  geom_line(data = predict_3bmi,aes(x = bmi,y = charges,col = "regression with square bmi"))+
  geom_line(data = predict_4bmi,aes(x = bmi,y = charges,col = "regression with cube bmi"),lty =2)+
  geom_line(data = predicted_2bmi , aes(x = bmi, y = charges,color = "regression") , lty = 2)+
  scale_color_manual("",breaks = c("regression with square bmi","regression with cube bmi","regression"
                     values = c("#FC9F66","#FAC357","#97C5D8"))+
  theme_clean()
```



分別觀察四個模型 SSE，可以發現有加入立方項的回歸模型，其 SSE 值為最低。

```
## see the regression line which is more fit
list("Sum of square error with intercept" = sum(residuals(model3_2)^2),
     "Sum of square error without intercept" = sum(residuals(model3_2W)^2),
     "Sum of square error with quadratic" = sum(residuals(model3_3)^2),
     "Sum of square error with cube" = sum(residuals(model3_4)^2))
```

```
## $`Sum of square error with intercept`
## [1] 12713013435
##
## $`Sum of square error without intercept`
## [1] 14641490644
##
## $`Sum of square error with quadratic`
## [1] 12262105051
##
## $`Sum of square error with cube`
## [1] 10928620322
```

```
## change the variable of smoker into 0 and 1 (nonsmoker and smoker)
datacopy <- data
datacopy$Idsmoker <- ifelse(data$smoker == "yes" ,1, 0)
datacopy <- datacopy[,c(1,2,3,4,8,6,7)]
```

接著觀察小孩與費用的回歸模型，在沒有 β_0 的模型下，小孩變數的顯著性較高，且 r-square 與調整後 r-square 較高又

```r
## make regression of charges with children
model4 <- lm(charges~children,datacopy)
model4w <- lm(charges~children-1,datacopy)
summary(model4)
```

```
##
## Call:
## lm(formula = charges ~ children, data = datacopy)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -11585  -8759  -4071   3468  51248
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  12522.5      446.5  28.049   <2e-16 ***
## children       683.1      274.2   2.491   0.0129 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12090 on 1336 degrees of freedom
## Multiple R-squared:  0.004624,   Adjusted R-squared:  0.003879
## F-statistic: 6.206 on 1 and 1336 DF,  p-value: 0.01285
```

```r
summary(model4w)
```

```
##
## Call:
## lm(formula = charges ~ children - 1, data = datacopy)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -24588  -1655   3210  12874  63770
##
## Coefficients:
##          Estimate Std. Error t value Pr(>|t|)
## children   5855.2      255.7    22.9   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15230 on 1337 degrees of freedom
## Multiple R-squared:  0.2817, Adjusted R-squared:  0.2811
## F-statistic: 524.3 on 1 and 1337 DF,  p-value: < 2.2e-16
```

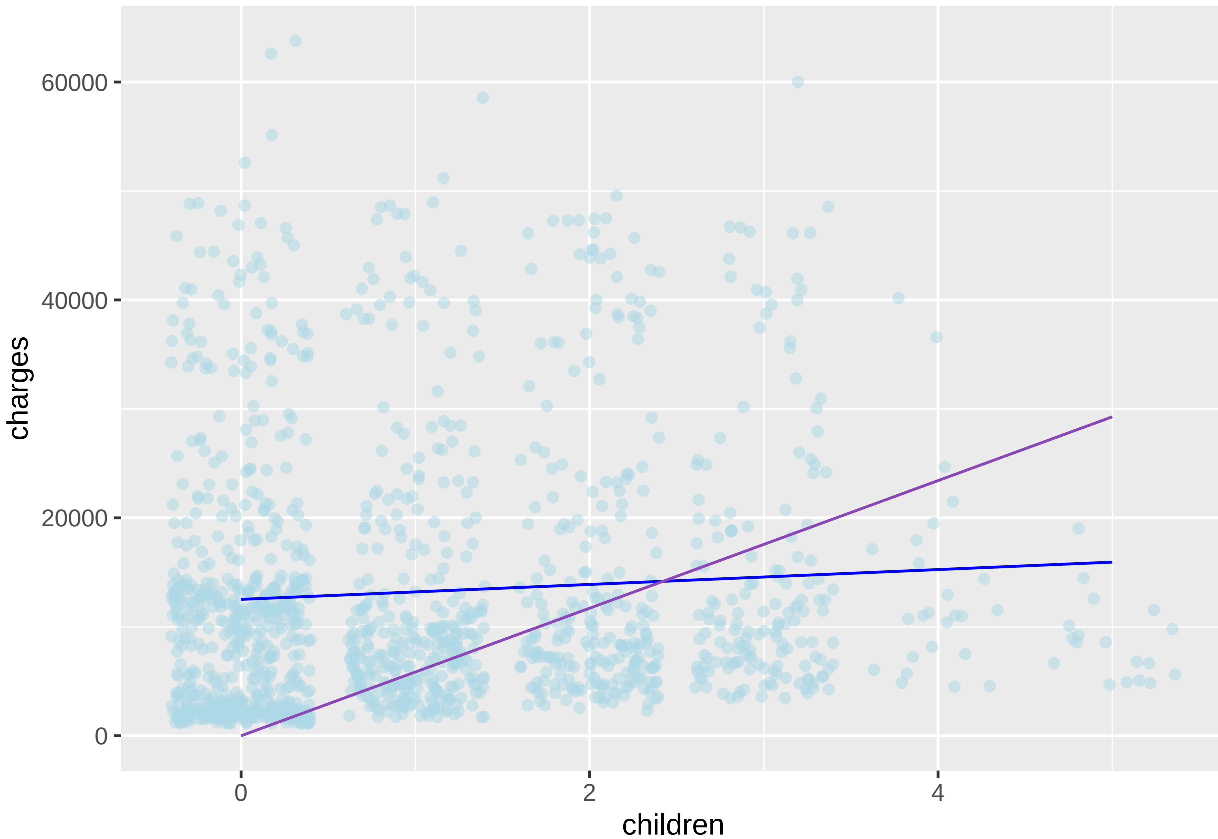
將預測的模型放入圖形中觀察。

```r
predict_1ch <- data.frame(charges = predict(model4,datacopy),
                          children = datacopy$children)

predict_1chw <- data.frame(charges = predict(model4w,datacopy),
                           children = datacopy$children)

ggplot(datacopy,aes(x=children,y = charges))+
  geom_jitter(alpha = 0.5,col = "lightblue")+
  geom_line(data = predict_1ch,aes(x = children , y = charges),col = "blue")+
```

```
  geom_line(data = predict_1chw,aes(x = children , y = charges),col = "#8B47B5")
```



接著利用 children 數當作分類標準，觀察在不同小孩數下的平均值，可以發現平均呈現一種鐘型的分佈，在小孩數為 2 或 3 時收取的費用會最高。

```
by(datacopy$charges,datacopy$children,mean)
```

```
## datacopy$children: 0
## [1] 12365.98
## ------------------------------------------------------------
## datacopy$children: 1
## [1] 12731.17
## ------------------------------------------------------------
## datacopy$children: 2
## [1] 15073.56
## ------------------------------------------------------------
## datacopy$children: 3
## [1] 15355.32
## ------------------------------------------------------------
## datacopy$children: 4
## [1] 13850.66
## ------------------------------------------------------------
## datacopy$children: 5
## [1] 8786.035
```

觀察地區與費用的回歸。

```
## make regression of charges with region
model5 <- lm(charges~region,datacopy)
```

```r
summary(model5)
```

```
##
## Call:
## lm(formula = charges ~ region, data = datacopy)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -13614  -8463  -3793   3385  49035
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      13406.4      671.3  19.971   <2e-16 ***
## regionnorthwest   -988.8      948.6  -1.042    0.297
## regionsoutheast   1329.0      922.9   1.440    0.150
## regionsouthwest  -1059.4      948.6  -1.117    0.264
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12080 on 1334 degrees of freedom
## Multiple R-squared:  0.006634,	Adjusted R-squared:  0.0044
## F-statistic:  2.97 on 3 and 1334 DF,  p-value: 0.03089
```

```r
## model.matrix(model5)
```

利用地區為分類標準，觀察費用各地區費用收取的平均。

```r
by(datacopy$charges,datacopy$region,mean)
```

```
## datacopy$region: northeast
## [1] 13406.38
## ------------------------------------------------------------
## datacopy$region: northwest
## [1] 12417.58
## ------------------------------------------------------------
## datacopy$region: southeast
## [1] 14735.41
## ------------------------------------------------------------
## datacopy$region: southwest
## [1] 12346.94
```

多變量回歸

利用多變量回歸觀察，可以發現某些變數不具有顯著性，所以將它排除並在觀察一次。

```r
## multiple regression
full_model <- lm(charges~. , datacopy)
summary(full_model)
```

```
##
## Call:
## lm(formula = charges ~ ., data = datacopy)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11304.9  -2848.1   -982.1   1393.9  29992.8
```

```
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -11938.5      987.8 -12.086  < 2e-16 ***
## age                  256.9       11.9  21.587  < 2e-16 ***
## sexmale             -131.3      332.9  -0.394 0.693348
## bmi                  339.2       28.6  11.860  < 2e-16 ***
## children             475.5      137.8   3.451 0.000577 ***
## Idsmoker           23848.5      413.1  57.723  < 2e-16 ***
## regionnorthwest     -353.0      476.3  -0.741 0.458769
## regionsoutheast    -1035.0      478.7  -2.162 0.030782 *
## regionsouthwest     -960.0      477.9  -2.009 0.044765 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6062 on 1329 degrees of freedom
## Multiple R-squared:  0.7509, Adjusted R-squared:  0.7494
## F-statistic: 500.8 on 8 and 1329 DF,  p-value: < 2.2e-16
```

去除掉性別這項變數後在觀察一次多變量回歸的模型，可以發現調整後 r-square 稍微的提高了，但兩個模型的變化並不大。

```
## the variable of sex is not significant in the last model
## so we exclude it and see the regression again
model_wsex <- lm(charges~.-sex , datacopy)
summary(model_wsex)
```

```
##
## Call:
## lm(formula = charges ~ . - sex, data = datacopy)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11367.2  -2835.4   -979.7   1361.9  29935.5
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -11990.27     978.76 -12.250  < 2e-16 ***
## age                  256.97      11.89  21.610  < 2e-16 ***
## bmi                  338.66      28.56  11.858  < 2e-16 ***
## children             474.57     137.74   3.445 0.000588 ***
## Idsmoker           23836.30     411.86  57.875  < 2e-16 ***
## regionnorthwest     -352.18     476.12  -0.740 0.459618
## regionsoutheast    -1034.36     478.54  -2.162 0.030834 *
## regionsouthwest     -959.37     477.78  -2.008 0.044846 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6060 on 1330 degrees of freedom
## Multiple R-squared:  0.7509, Adjusted R-squared:  0.7496
## F-statistic: 572.7 on 7 and 1330 DF,  p-value: < 2.2e-16
```

```
## predict(model_wsex)
```

Log-linear Regression

將 charges 變數取 log 值，並帶入上面的模型觀察，利用這個模型有兩個變數的顯著性提高了，r-square 與調整後 r-square 也提高了。

```
## using log-linear regression
log_model <- lm(log(charges)~.-sex , datacopy)
summary(log_model)
```

```
##
## Call:
## lm(formula = log(charges) ~ . - sex, data = datacopy)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.10302 -0.19707 -0.05206  0.06564  2.15091
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     7.0008478  0.0719853  97.254  < 2e-16 ***
## age             0.0346490  0.0008746  39.618  < 2e-16 ***
## bmi             0.0130711  0.0021004   6.223 6.52e-10 ***
## children        0.1013204  0.0101304  10.002  < 2e-16 ***
## Idsmoker        1.5472965  0.0302910  51.081  < 2e-16 ***
## regionnorthwest -0.0633386  0.0350174  -1.809 0.070712 .
## regionsoutheast -0.1568166  0.0351952  -4.456 9.07e-06 ***
## regionsouthwest -0.1285638  0.0351393  -3.659 0.000263 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4457 on 1330 degrees of freedom
## Multiple R-squared:  0.7663, Adjusted R-squared:  0.765
## F-statistic: 622.9 on 7 and 1330 DF,  p-value: < 2.2e-16
```

Locally Weighted Regression

利用 Locally Weighted Regression 製作費用與 bmi 的回歸模型，並將其與 log-linear 的回歸模型比較。

```
## loess regression
lossreg <- loess(charges ~ bmi, NS$yes)
logreg <- lm(charges~ log(bmi),NS$yes)

ggplot(NS$yes,aes(x = bmi , y = charges))+
  geom_point(col = "pink")+
  geom_line(data = data.frame(charges = predict(lossreg , NS$yes) , bmi = NS$yes$bmi),
            aes(x = bmi , y = charges, colour = "Locally weighted regression" ))+
  geom_line(data = data.frame(charges = predict(logreg , NS$yes), bmi = NS$yes$bmi),
            aes(x = bmi , y = charges, colour = "Log-linear regression"))+
  scale_color_manual("",breaks = c("Locally weighted regression","Log-linear regression"),
                     values = c("red","#7E1037"))+
  theme(legend.position = "bottom")
```

利用 Locally Weighted Regression 製作費用與 age 的回歸模型，並將其與 log-linear 的回歸模型比較。

```
lossreg_l <- loess(charges ~ age,NS$yes )
logreg_l <- lm(charges ~ log(age) , NS$yes)

ggplot(NS$yes, aes(x = age , y = charges))+
  geom_point(col = "#88CDF6")+
  geom_line(data = data.frame(charges = predict(lossreg_l,NS$yes), age = NS$yes$age),
            aes(x = age, y = charges , colour = "Locally weighted regression"))+
  geom_line(data = data.frame(charges = predict(logreg_l,NS$yes), age = NS$yes$age),
            aes(x = age , y = charges , colour = "Log-linear regression"))+
  scale_color_manual("" , breaks = c("Locally weighted regression","Log-linear regression"),
                     values = c("#73CD88","#23503A"))+
  theme(legend.position = "bottom")
```

Kernel Regression

我利用 kernel Regression 跑出來的圖形可能較不適合這個模型。

```
## kernel regression
model.np <- npreg(charges ~ bmi , ckertype = "gaussian" , ckerorder = 2 , data = NS$yes)
```

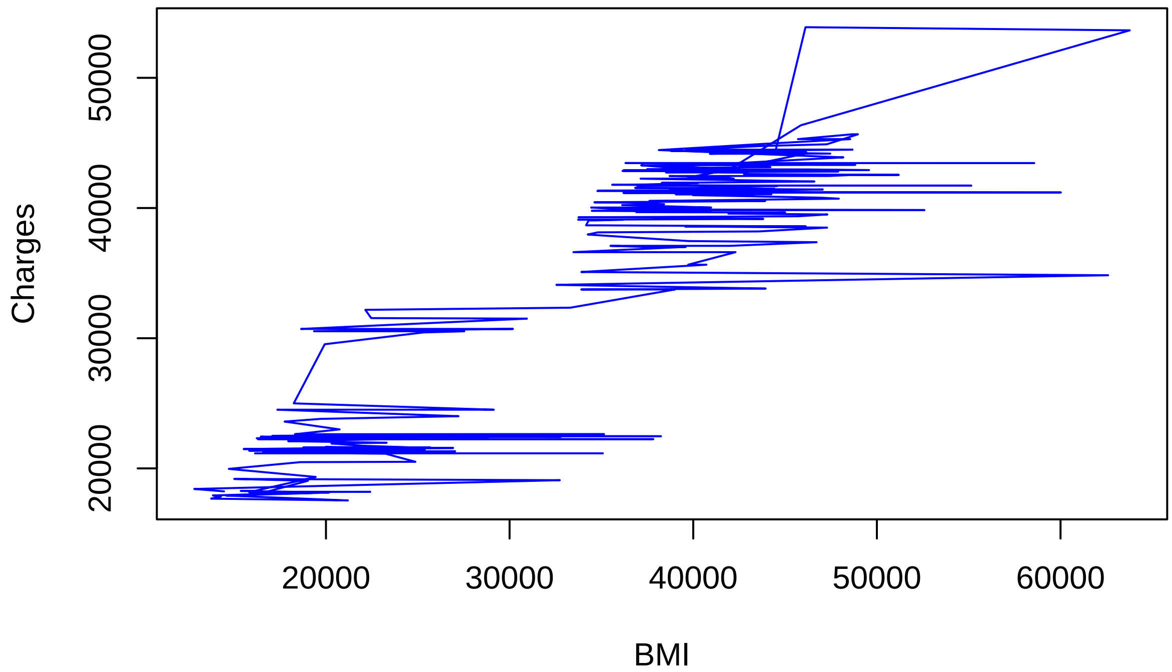
```
## Multistart 1 of 1 |Multistart 1 of 1 |Multistart 1 of 1 |Multistart 1 of 1 /Multistart 1 of 1 |
Multistart 1 of 1 |
```

```
model.np
```

```
##
## Regression Data: 274 training points, in 1 variable(s)
##                      bmi
## Bandwidth(s): 0.76139
##
## Kernel Regression Estimator: Local-Constant
## Bandwidth Type: Fixed
##
## Continuous Kernel Type: Second-Order Gaussian
## No. Continuous Explanatory Vars.: 1
```

```
series <- seq(10,20)
fit <- predict(model.np , series)

plot(NS$yes$charges[order(NS$yes$bmi)],fit$fit[order(NS$yes$bmi)],col = "blue", type = "l",
     xlab = "BMI" , ylab = "Charges")
```



Decision Tree

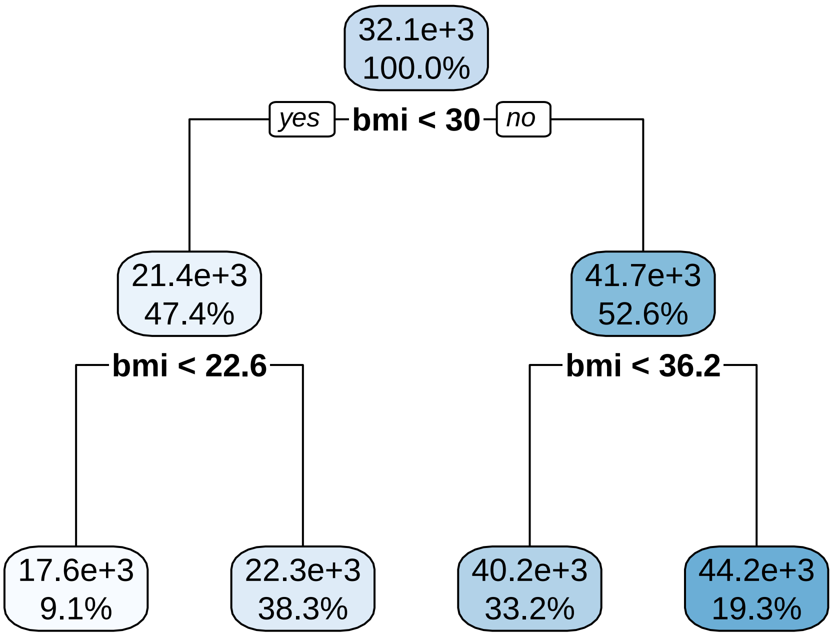
製作決策樹利用 bmi 值。

```
## regression tree
regTree <- rpart(charges ~ bmi,data = NS$yes)
regTree
```

```
## n= 274
##
## node), split, n, deviance, yval
##       * denotes terminal node
##
## 1) root 274 36365600000 32050.23
##   2) bmi< 30.01 130  3286655000 21369.22
##     4) bmi< 22.605 25   398685900 17581.07 *
##     5) bmi>=22.605 105  2443798000 22271.17 *
##   3) bmi>=30.01 144  4859010000 41692.81
##     6) bmi< 36.245 91  2796574000 40203.57 *
##     7) bmi>=36.245 53  1514085000 44249.81 *
```

```
## plot(regTree,uniform = T)
## text(regTree,digits = 6)

rpart.plot(regTree ,digits = 3, type = 2 , roundint = F)
```



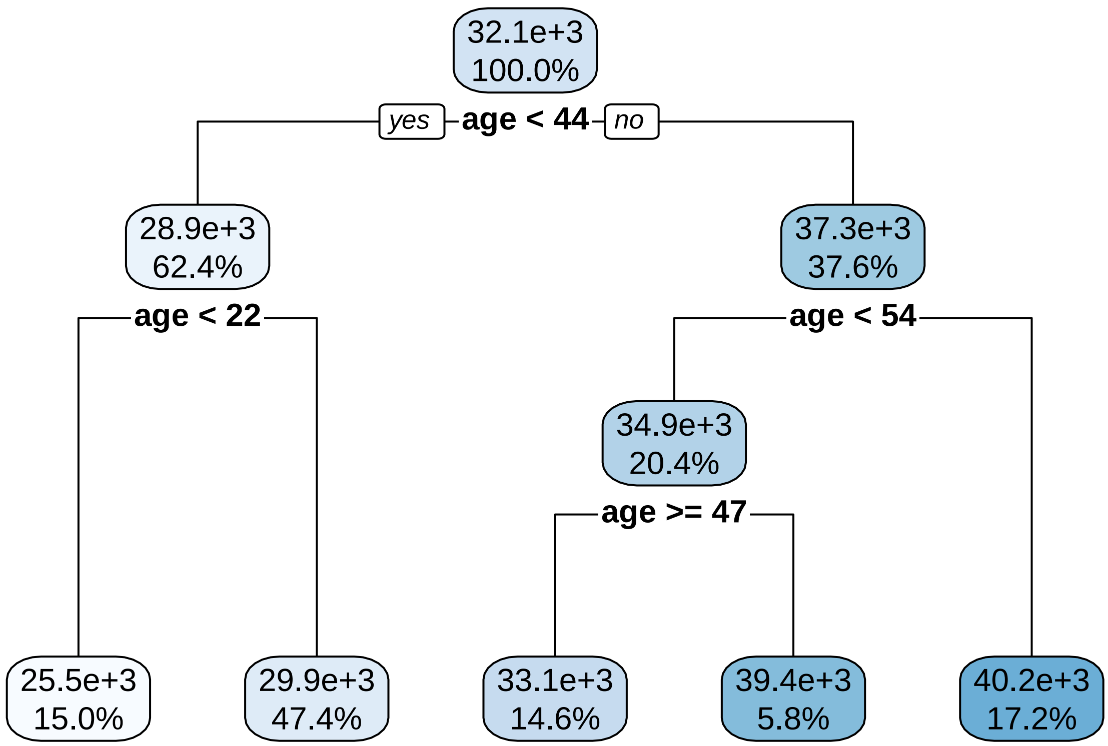
製作決策樹利用 age 值。

```
regTree_1 <- rpart(charges ~ age,data = NS$yes)
regTree_1
```

```
## n= 274
##
## node), split, n, deviance, yval
##       * denotes terminal node
##
##  1) root 274 36365600000 32050.23
##    2) age< 43.5 171 19372800000 28866.81
##      4) age< 21.5 41  3814390000 25516.52 *
##      5) age>=21.5 130 14953070000 29923.44 *
##    3) age>=43.5 103 12382830000 37335.33
##      6) age< 53.5 56  6971783000 34898.74
##       12) age>=46.5 40  4707243000 33113.53 *
##       13) age< 46.5 16  1818360000 39361.78 *
##      7) age>=53.5 47  4682440000 40238.50 *
```

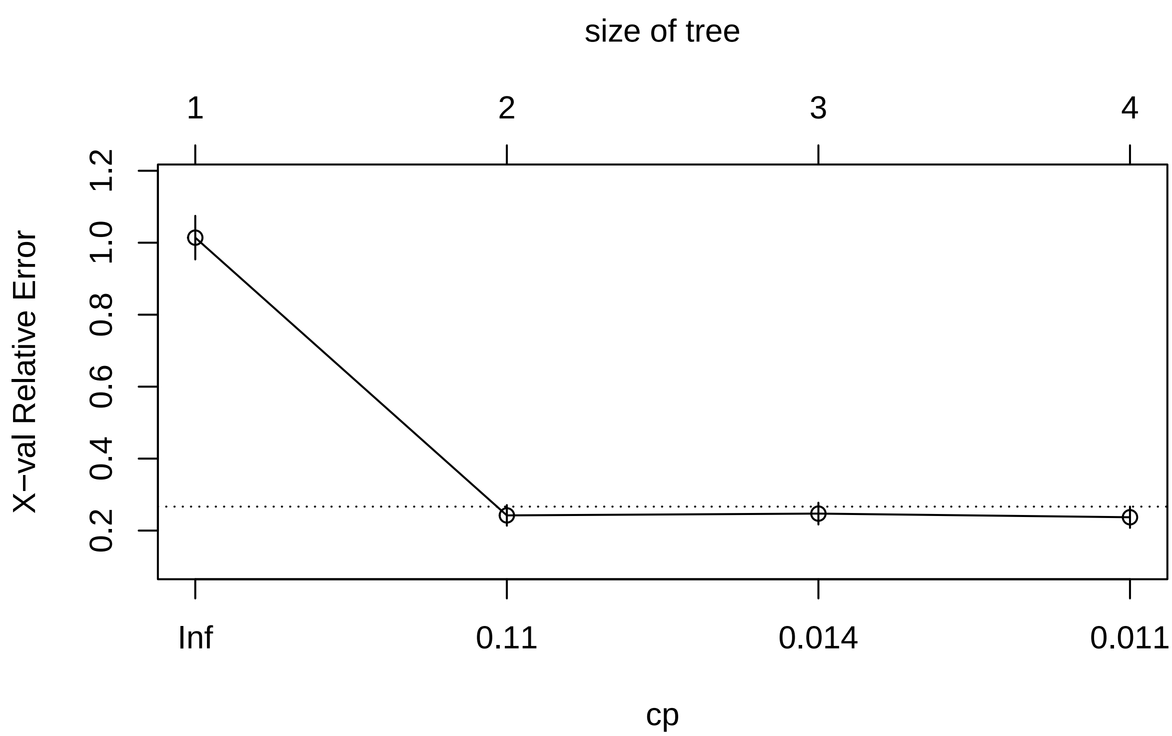
```
## plot(regTree_1,uniform = T)
## text(regTree_1,digits = 6)

rpart.plot(regTree_1, digits = 3)
```



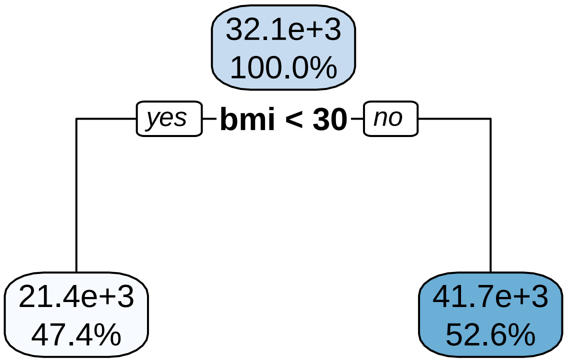
利用 cp 值去調整決策樹的長度與大小。

```
## regression tree - prune tree
plotcp(regTree)
```

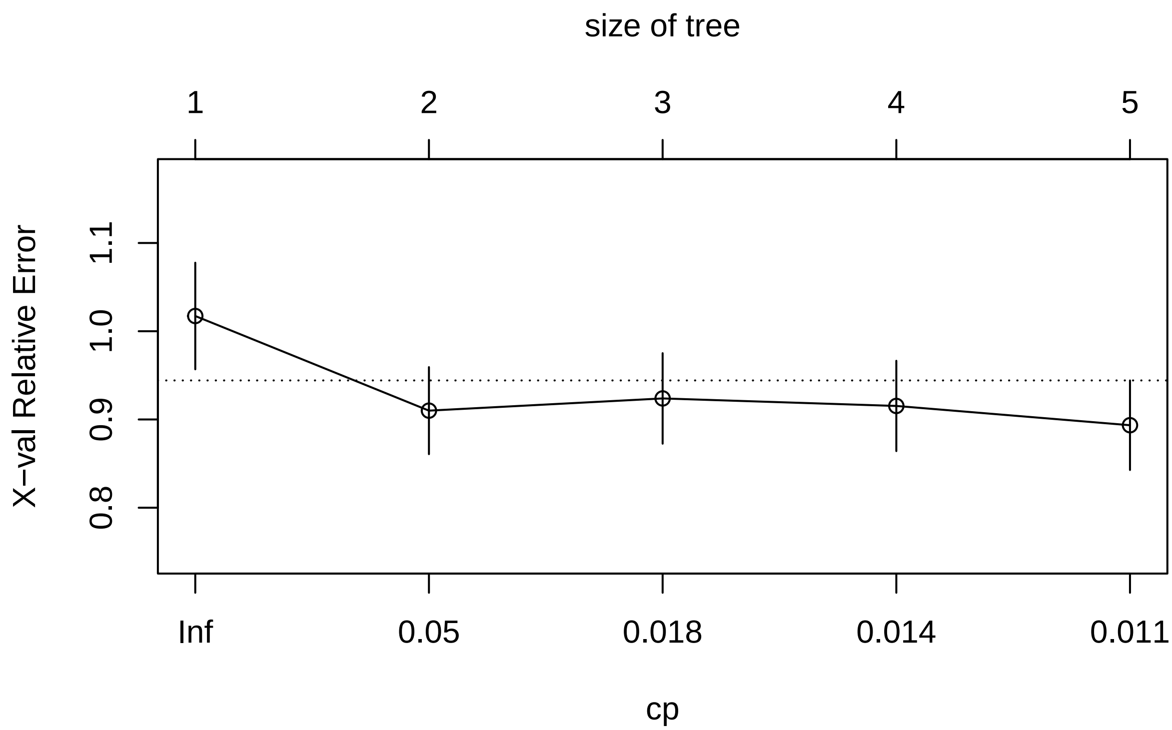



```
regTree_prune <- prune(regTree , cp =0.11)
## plot(regTree_prune,uniform = T)
## text(regTree_prune,digits =6)
```

```r
rpart.plot(regTree_prune, digits = 3)
```

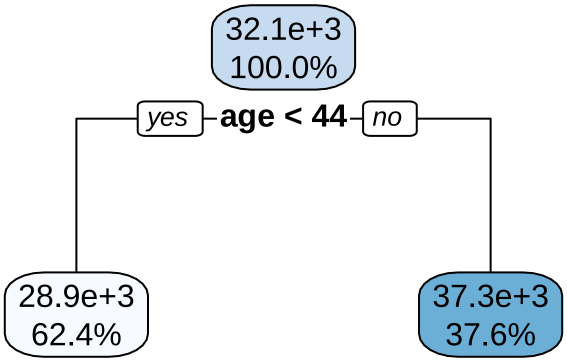



```r
plotcp(regTree_1)
```



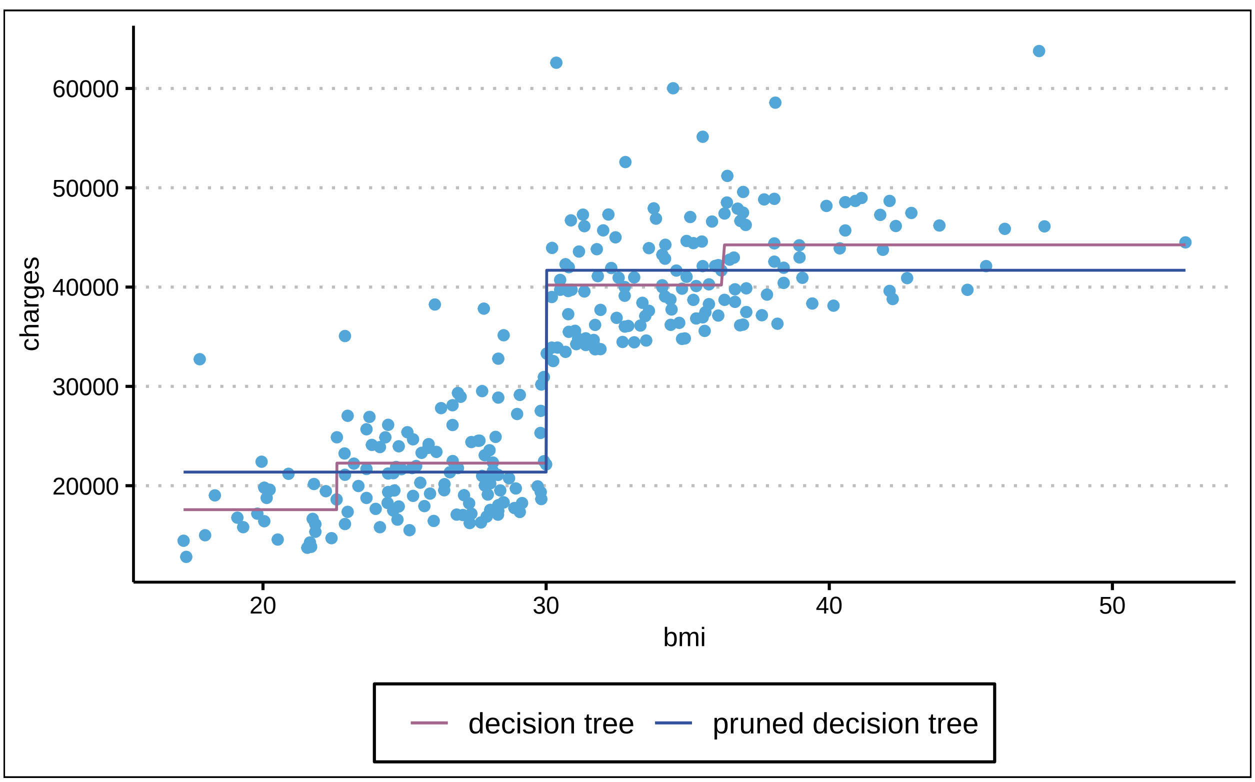

```r
regTree_prune1 <- prune(regTree_1 , cp =0.05)
## plot(regTree_prune1,uniform = T)
## text(regTree_prune1,digits =6)

rpart.plot(regTree_prune1, digits = 3)
```



比較取用較適 cp 值後的決策樹，在散佈圖中的呈現。

```
## compare before and after pruning tree
NS$yes %>%
  mutate(pred = predict(regTree,NS$yes)) %>%
  mutate(pred2 = predict(regTree_prune,NS$yes)) %>%
  ggplot(aes(x = bmi , y = charges))+
  geom_point(col = "#53A7D8")+
  geom_line(aes(y = pred , colour = "decision tree"))+
  geom_line(aes(y = pred2 , colour = "pruned decision tree"))+
  scale_color_manual("",breaks = c("decision tree", "pruned decision tree"),
                     values = c("#A5678E","#33539E"))+
  theme_clean()+
  theme(legend.position = "bottom")
```




```
NS$yes %>%
  mutate(pred = predict(regTree_1,NS$yes)) %>%
  mutate(pred2 = predict(regTree_prune1,NS$yes)) %>%
```

```r
ggplot(aes(x = age , y = charges))+
geom_point(col = "#FFDD94")+
geom_line(aes(y = pred , colour = "decision tree"))+
geom_line(aes(y = pred2 , colour = "pruned decision tree"))+
scale_color_manual("",breaks = c("decision tree", "pruned decision tree"),
                   values = c("#FA897B","#CCABDB"))+
theme_clean()+
theme(legend.position = "bottom")
```