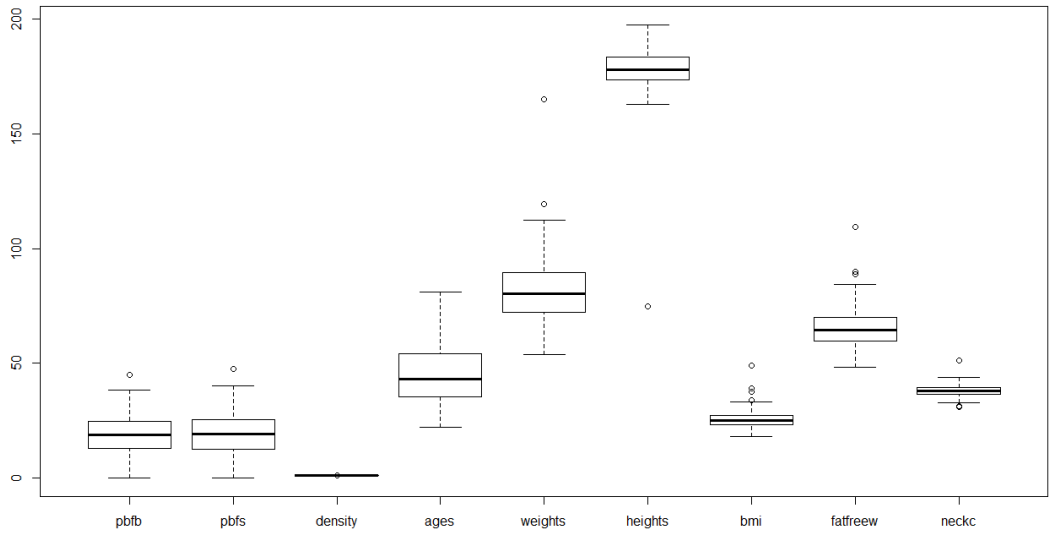
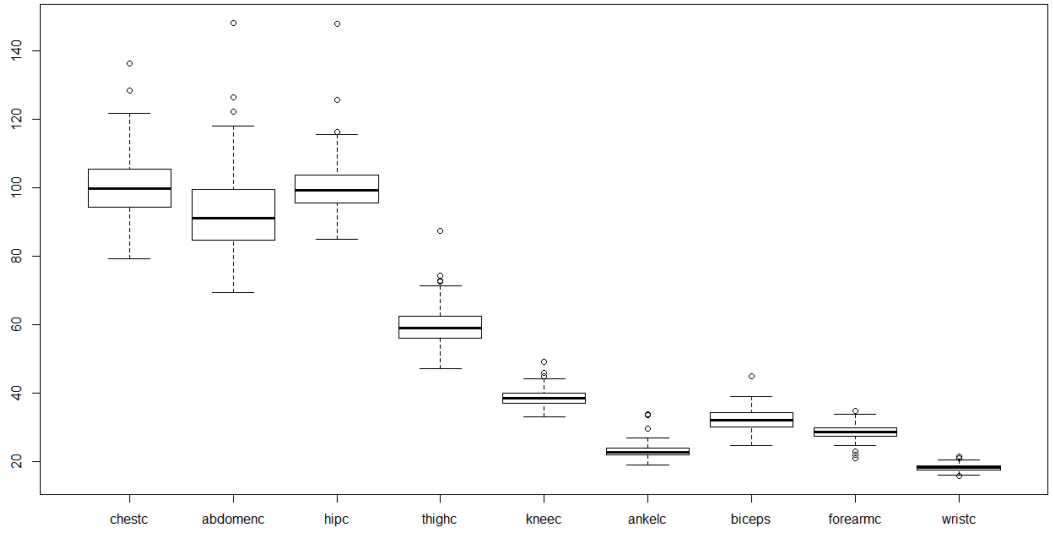
期末生統報告

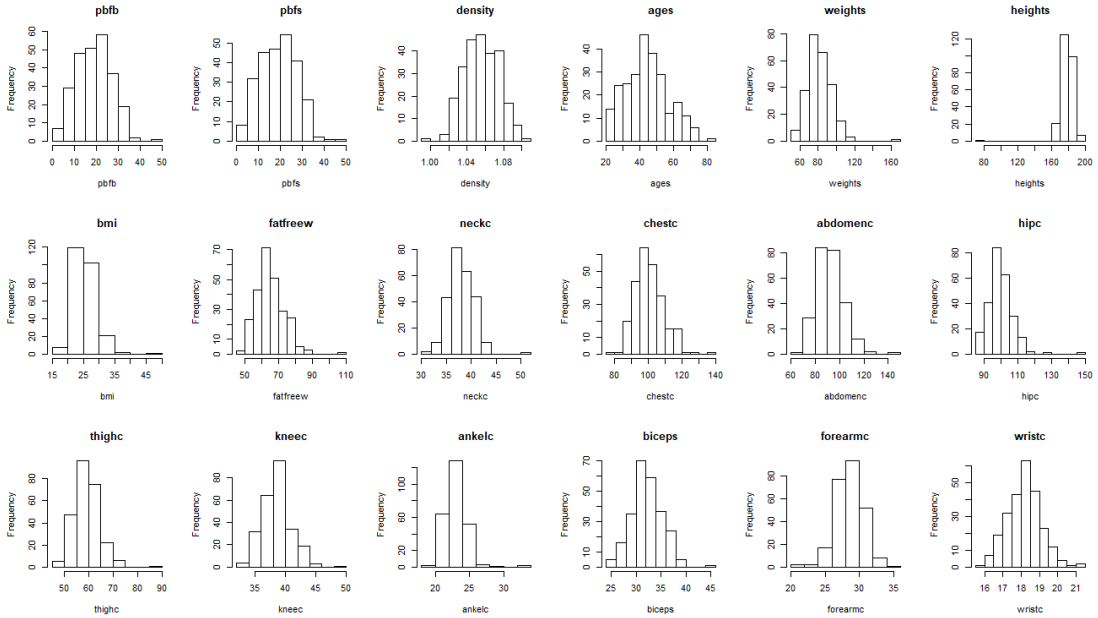
1. Use descriptive statistics to explore the data and identify potential outliers or dubious data entries (30%).

在分析這筆資料以前，首先我們先對這筆資料檢查看看是否有遺失值，從檢測結果發現，這筆資料不存在遺失值，因此我們接著對這筆資料中的每一個變數去計算平均數、標準差、最小值、中位數、最大值，並且畫出盒鬚圖以及直方圖來查看這些變數的分布情形以及資料的合理性。

Variables	mean	std	maximum	median	minimum
pbfb	18.938	7.751	45.1	19	0
pbfs	19.151	8.369	47.5	19.2	0
density	1.056	0.019	1.109	1.055	0.995
ages	44.885	12.602	81	43	22
weights	81.332	13.358	165.1	80.2	53.9
heights	178.181	9.304	197.5	177.8	74.9
bmi	25.437	3.648	48.9	25.05	18.1
fatfreew	65.322	8.29	109.3	64.35	48.1
neckc	37.992	2.431	51.2	38	31.1
chestc	100.824	8.43	136.2	99.65	79.3
abdomenc	92.556	10.783	148.1	90.95	69.4
hipc	99.905	7.164	147.7	99.3	85
thighc	59.406	5.25	87.3	59	47.2
kneec	38.59	2.412	49.1	38.5	33
ankelc	23.102	1.695	33.9	22.8	19.1
biceps	32.273	3.021	45	32.05	24.8
forearmc	28.664	2.021	34.9	28.7	21
wristc	18.23	0.934	21.4	18.3	15.8


200
150
100
50
0
pbfb
pbfs
density
ages
weights
heights
bmi
fatfreew
neckc


140
120
100
80
60
40
20
chestc
abdomenc
hipc
thighc
kneec
ankelc
biceps
forearmc
wristc



我們認為有關身體周長、年齡、身高、體重以及 BMI 這些變數所收集到的資料應該都要大於 0。除此之外，一些以百分比為單位的變數，其所收集到的資料值應該要介於 0 到 100 之間。然而從上面的圖表結果中，我們可以發現在盒鬚圖中，pbfb 以及 pbfs 這兩個變數有出現 0 的觀測值，回去對照資料發現這筆資料是 id 為 182 的觀測值，因為不確定這種變數出現這樣的值是否合理，為了保守起見，我們選擇將此筆資料拿掉。而在底下多個直方圖中，我們有看到多個變數同時都有出現一筆非常大的觀測值，在經過比對以後，我們發現這筆觀測值的 id 為 39，為 outlier，為了能夠更精準的預測其他人的資料，我們決定將此筆資料給刪除。而在完成上述的統計圖檢查以後，因為我們知道 BMI 以及 fatfreew 的計算公式，因此我們接下來便對這兩個變數去檢驗資料的品質。而從檢查結果中我們發現 id 為 42, 163, 221 這 3 筆資料的 BMI 以及 fatfreew 資料和我們使用公式所計算出來的結果存在著差距，而在此我們懷疑有資料輸入錯誤的情形發生，因此保險起見我們將此 3 筆 data 也刪除。總結來說，在資料檢測這個階段裏我們刪除了 id 為 39, 42, 163, 182, 221 共 5 筆觀測值。

2. Use the 16 body measurements to come up with a prediction model for the percentage of body fat estimated by Siri's method. As these16 body measurements are all taken from the same patients, they are very likely to be highly correlated. Multicollinearity can be a problem in your prediction model and explain how you deal with it (30%)?

在檢查完資料後，接著便對這筆資料下去建模，我們觀察了這筆資料裏頭的變數，並且發現有些變數在實際生活上的取得並不容易，這些變數包含 density 以及 fatfreew，考慮到模型的有用性，我們決定不把這兩個變數納入

我們變數選擇的候選變數。在建模的一開始，我們先把所有資料分成 200 筆的訓練資料集以及 47 筆的測試資料集，接著利用訓練資料集來建模，並利用測試資料集來評估我們的模型。首先，我們將所有變數(除了前面提到的兩個變數外)丟到模型內，並去觀察這個模型的表現。

Model 1:

```
Residuals:
    Min      1Q  Median      3Q     Max
-10.7418  -2.9772  -0.3264   3.2701   9.3870

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -48.199812  66.326110  -0.727   0.4683
ages          0.067310   0.036613   1.838   0.0676 .
weights      -0.353630   0.406528  -0.870   0.3855
heights       0.238318   0.366410   0.650   0.5162
bmi           1.475895   1.293666   1.141   0.2554
neckc        -0.231123   0.267340  -0.865   0.3884
chestc       -0.194224   0.133166  -1.459   0.1464
abdomenc      0.869611   0.105637   8.232 3.23e-14 ***
hipc         -0.314506   0.180816  -1.739   0.0836 .
thighc        0.126358   0.166354   0.760   0.4485
kneec         0.001378   0.274032   0.005   0.9960
ankelc        0.172417   0.239567   0.720   0.4726
biceps        0.235238   0.192341   1.223   0.2229
forearmc      0.156040   0.248166   0.629   0.5303
wristc       -1.706498   0.596278  -2.862   0.0047 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1
‘ ’ 1

Residual standard error: 4.325 on 185 degrees of freedom
Multiple R-squared:  0.7394,    Adjusted R-squared:  0.719
7
F-statistic:  37.5 on 14 and 185 DF,  p-value: < 2.2e-16
```

```
VIF:
     ages    weights    heights        bmi     neckc     ches
tc   abdomenc
  2.357609 257.922759  57.679525 193.504173   3.662350  11.
869026  12.134976
     hipc     thighc      kneec     ankelc     biceps   forea
rmc     wristc
 14.332657   7.082162   4.353123   1.741239   3.208515   2.4
86343   3.147729
```

從結果中我們看到了雖然整體的模型顯著，但有很多變數都是不顯著的，再加上該模型裡有很多變數的 VIF 是超過 5 的，因此我們從 VIF 超過 5 的那些變數中挑選一個與 Y 最相關的變數(abdomenc)留在模型內來當作我們下一個模型。

Model 2: (僅列出 VIF 的部分)

```
VIF:
ages abdomenc    neckc   biceps    kneec forearmc   wristc
 ankelc
1.379563 3.334894 3.257146 2.864172 2.848450 2.314363 2.82
4018 1.609885
```

而從這次的結果中我們發現雖然 VIF 的問題獲得緩解，但這個 model 中仍然存在著與 abdomenc 有著高相關的變數(neckc, kneec)，為了盡量移除共線性問題，我們接著將這兩個變數給移除。

Model 3:

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -16.10395    6.71073  -2.400 0.017358 *
ages          0.09575    0.02841   3.370 0.000907 ***
abdomenc      0.72815    0.04676  15.572  < 2e-16 ***
biceps        0.13117    0.18271   0.718 0.473667
forearmc      0.14802    0.23919   0.619 0.536764
wristc       -2.53676    0.52986  -4.788 3.35e-06 ***
ankelc        0.06533    0.22583   0.289 0.772663
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1
‘ ’ 1
```

```
Residual standard error: 4.42 on 193 degrees of freedom
Multiple R-squared:  0.7161,    Adjusted R-squared:  0.707
3
F-statistic: 81.13 on 6 and 193 DF,  p-value: < 2.2e-16
```

而從這裡可以發現雖然整體的模型是顯著的，但模型中出現了許多不顯著的變數，於是我們嘗試移除不顯著的變數來看整體模型的表現。

Model 4:

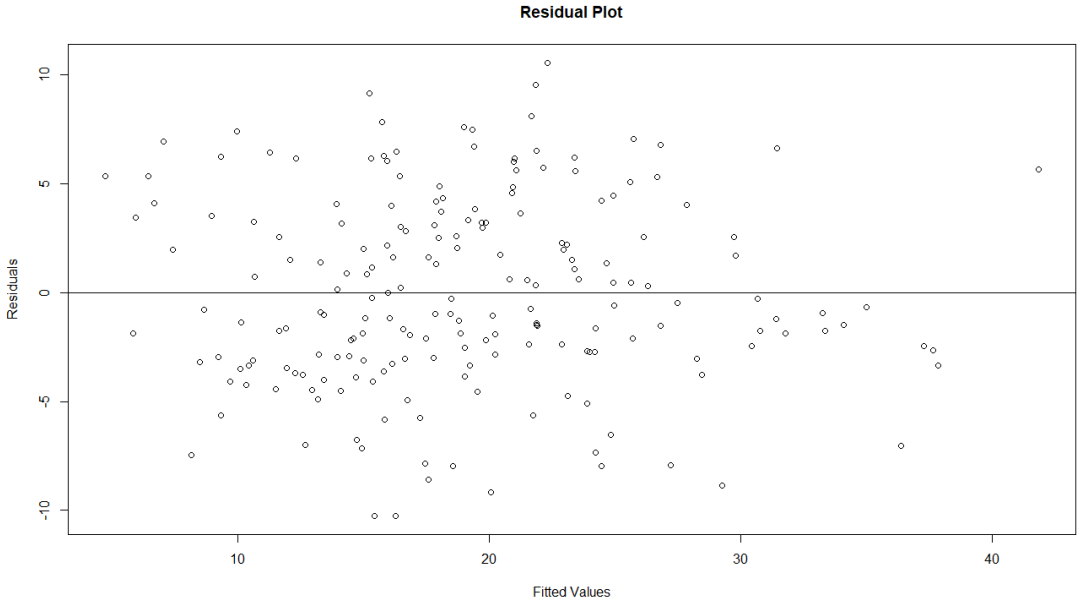
```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -14.25685    6.36501  -2.240  0.02622 *
ages          0.07972    0.02527   3.155  0.00186 **
abdomenc      0.75787    0.03934  19.265  < 2e-16 ***
wristc       -2.20163    0.43780  -5.029 1.11e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1
‘ ’ 1

Residual standard error: 4.406 on 196 degrees of freedom
Multiple R-squared:  0.7136,    Adjusted R-squared:  0.709
2
F-statistic: 162.7 on 3 and 196 DF,  p-value: < 2.2e-16
```

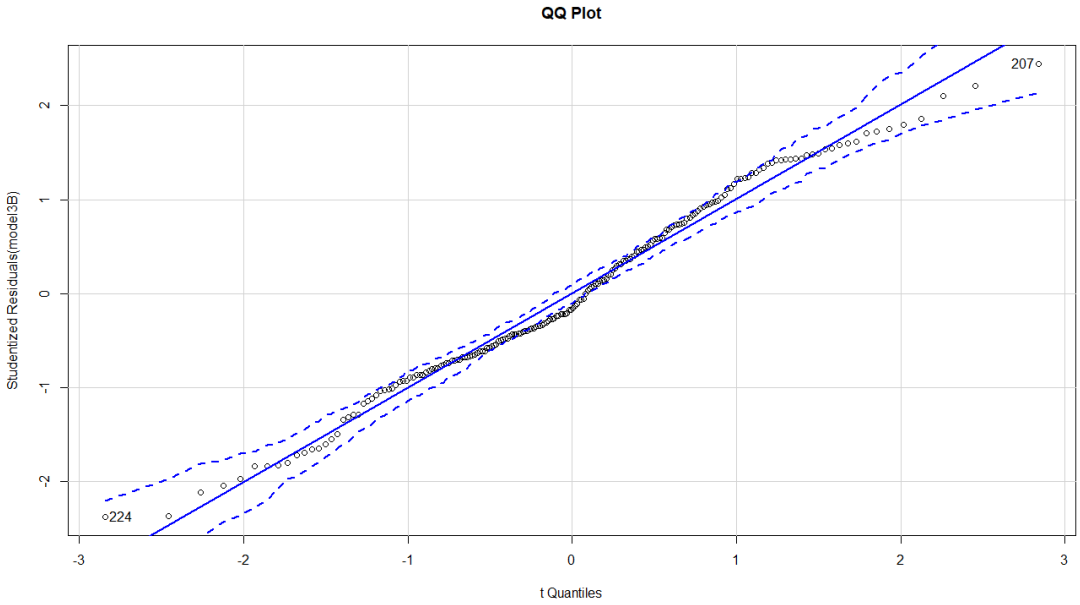
從上述的結果中可以觀察到，在移除不顯著的變數後 adjusted R-squared 反而略微提升，而且與上一個模型相比，此模型裡變數的顯著性並沒有發生太大的改變，再加上這個模型在我們的測試資料集的表現也是不差的(R-squared: 0.7776)，因此我們決定拿此模型來當我們最後篩選出來的模型。

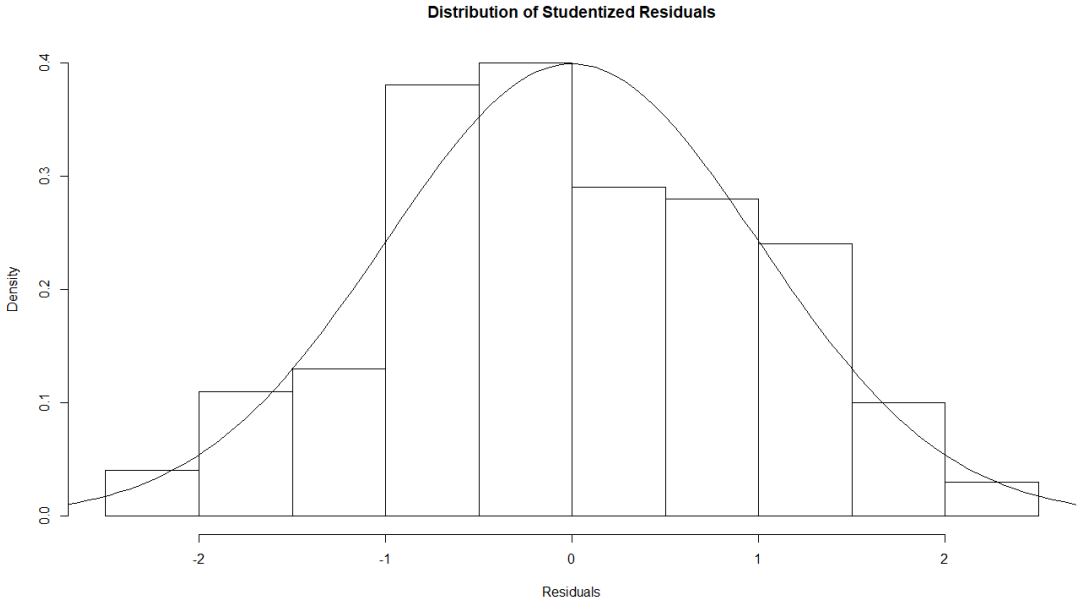
$$\text{pbfs} = -14.25685 + 0.07972 * \text{ages} + 0.75787 * \text{abdomenc} - 2.20163 * \text{wristc}$$

3. Please describe the rationale of your model building strategy (20%) and include an evaluation of your model using regression diagnostic tools such as, but not limited to, residual plots (20%)

   在做完模型的選擇以後，我們接著便使用殘差圖、QQ plot 以及直方圖來檢測此筆資料是否適用該回歸模型。


Residual Plot
Residuals
Fitted Values
10
5
0
-5
-10
10
20
30
40


QQ Plot
Studentized Residuals(model3B)
t Quantiles
207
224
2
1
0
-1
-2
-3
-2
-1
0
1
2
3



從殘差圖中可以看到，各個資料點很均勻地散布在 0 那條線的附近，這代表著此資料符合殘差期望值為 0、殘差變異數同質性以及 Y 和 X 之間的關係為線性的基本假設，而從 QQ plot 和直方圖中可以明顯看到資料點很貼近 QQ plot 中藍色的斜直線，這代表著此資料服從常態假設，證明該模型適用於此資料。