# 第五組期末報告

組員:謝佳璇、陳芊霓

## 一.資料背景

在如今的社會中，「高血壓」十分盛行，已經不再是老年人才會罹患的了，那麼我們好奇「什麼樣的人罹患高血壓的機率會比較高？」，像是年齡越大罹患高血壓機率越大？或是性別會影響罹患高血壓機率嗎？那麼我們就好奇另一個問題「什麼樣的人平均去看高血壓次數會比較高？」，一樣用前面的年齡、性別來舉例，像是年齡越大的人每年平均看高血壓的次數會比較高？或是說男生每年平均看高血壓的次數會比較高？ 而這兩個問題就是我們這份報告的目的，由這筆資料去做分析後就能夠預測其他人罹患高血壓的機率、其他人每年平均看高血壓次數，所以我們為了達到目的，使用了 10000 個人的資料，裡面有受訪人的「性別」、「年齡」、「是否有高血壓」、「看診高血壓次數」、「看診高血壓診費」、「看診總次數」、「看診總花費」、「是否有糖尿病」。

## 二.目的

得到這兩個問題『「什麼樣的人罹患高血壓的機率會比較高？」、「什麼樣的人平均去看高血壓次數高？」』的關係式

## 三.資料圖表

首先我們做柱狀圖(histogram)，因為目的為判斷其是否為常態分配，而常態分配一定要是連續型態的資料，所以我們只做連續型態的柱狀圖。

註:final 中連續的資料為年齡、看診高血壓次數、看診高血壓診費、總看診次數、總看診花費為甚麼要找連續型態呢?

舉例來說，有無高血壓(圖 1)因為資料只有 0、1，為不連續型態，他的柱狀圖做出來是極端的，一定不會是常態分配，所以做不連續型態的柱狀圖沒有意義。

因為常態分配的柱狀圖要對稱，而年齡(圖 2)、看診高血壓次數(圖 3)、看診高血壓診費(圖 4)、總看診次數(圖 5)、總看診花費(圖 6)的柱狀圖都不對稱，所以他們都不是常態分配。

再來我們做 QQ plot(Quantile-Quantile Plot)，再次確認是否為常態分配，而常態分配一定要是連續型態的資料，所以我們只做連續型態的 QQ plot。

註:final 中連續的資料為年齡、看診高血壓次數、看診高血壓診費、總看診次數、總看診花費為甚麼要找連續型態呢?

舉例來說，有無高血壓(圖 7)因為資料只有 0、1，為不連續型態，他的 QQ plot 做出來是極端的，一定不會是常態分配，所以做不連續型態的 QQ plot 沒有意義。

因為常態分配的 QQ plot 要是左下到右上的斜對角線，而年齡(圖 8)、看診高血壓次數(圖 9)、看診高血壓診費(圖 10)、總看診次數(圖 11)、總看診花費(圖 12)的 QQ plot 中的圖形都不符合 45 度角，所以他們都不是常態分配。

最後我們要用盒鬚圖來看是否有離群點(outliers)，離群點就是資料中的異常值，代表跟資料有很大的誤差。
一樣是找連續型態的盒鬚圖就好，舉例來說高血壓(圖 13)的盒鬚圖，只分布在兩個點(其中一個點為離群點)，沒有參考價值。

我們得到年齡(圖 14)、看診高血壓次數(圖 15)、看診高血壓診費(圖 16)、總看診次數(圖 17)、總看診花費(圖 18)的盒鬚圖都有離群點。

接下來我們把年齡、看診高血壓次數、看診高血壓診費、看診總次數、看診總花費用性別、有無高血壓、有無糖尿病來分類，首先先看用性別分組的

## 3.1-1 性別 v.s. 年齡

年齡用性別分類的盒鬚圖(圖 19)可以看到，女生年齡沒有離群點，且大多數都落在 23.2 歲~53.6 歲之間，中間值為 37.9，最大值為 98.4 歲，最小值為 2.8 歲；男生年齡有離群點，且大多數都落在 20.1 歲~52.4 歲之間，中間值為 37.2，最大值為 101.7 歲，最小值為 2.8 歲，再用 t-test 我們得到 p-value=0.001089 < 0.05 ，拒絕掉 h0: mu1=mu2，意思為拒絕掉【男生平均年齡=女生平均年齡】，所以男生平均年齡和女生平均年齡在統計上有差異性，以及由 t-test 可以得知 95%信賴區間[0.5421159,2.1680903]，還得到男生平均年齡 37.65268，女生平均年齡 39.00779，女生平均年齡大於男生平均年齡。

## 3.1-2 性別 v.s. 看診高血壓次數

看診高血壓次數用性別分類由盒鬚圖(圖 20)可以看到因為大多數看診高血壓次數為零，所以男生看診高血壓次數、女生看診高血壓次數都有很多離群點，再用 t-test 我們得到 p-value=0.1472 > 0.05 ，沒有拒絕掉 h0: mu1=mu2，意思為沒有拒絕掉【男生平均看診高血壓次數=女生平均看診高血壓次數】，所以男生平均看診高血壓次數和女生平均看診高血壓次數在統計上沒有差異性，以及由 t-test 可以得知 95%信賴區間[-0.03987545, 0.26622883]，還得到男生平均看診高血壓次數 1.202961，女生平均看診高血壓次數 1.316138，可以看出女生的平均看診高血壓次數會高於男生的平均看診高血壓次數。

## 3.1-3 性別 v.s. 看診高血壓診費

看診高血壓診費用性別分類由盒鬚圖(圖 21)可以看到因為大多數診費為零，所以男生看診高血壓診費、女生看診高血壓診費都有很多離群點，再用 t-test 我們得到 p-value=0.2553 > 0.05 ，沒有拒絕掉 h0: mu1=mu2，意思為沒有拒絕掉【男生平均看診高血壓診費=女生平均看診高血壓診費】，所以男生平均看診高血壓診費和女生平均診高血壓診費在統計上沒有差異性，以及由 t-test 可以得知 95%信賴區間[-13.83872,52.11740]，還得到男生平均看診高血壓診費 271.5811，女生平均看診高血壓診費 290.7205，可以看出女生的平均看診高血壓診費會高於男生的平均看診高血壓診費。

## 3.1-4 性別 v.s.看診總次數

看診總次數用性別分類由盒鬚圖(圖 22)可以看到女生看診總次數有離群點，大多數的看診總次數落在 6 次~24 次之間，看診總次數最少 1 次，最多 200 次，中間值 13 次；男生看診總次數有離群點，大多數的看診總次數落在 4 次~19 次之間，看診總次數最少 1 次，最多 165 次，中間值 10 次，再用 t-test 我們得到 p-value < 8.046*10 的-16 次方 < 0.05 ，有拒絕掉 h0: mu1 =mu2，意思為有拒絕掉【男生平均看診總次數=女生平均看診總次數】，所以男生平均看診總次數和女生平均看診總次數在統計上有差異性，以及由 t-test 可以得知 95%信賴區間[1.819761, 3.112175]，還得到男生平均看診總次數 14.98293，女生平均看診總次數 17.44890，可以看出女生的平均看診總次數會高於男生的平均看診總次數。

## 3.1-5 性別 v.s.看診總花費

看診總花費用性別分類由盒鬚圖(圖 23)可以看到女生看診總花費有離群點，大多數的女生看診總花費都落在 1500 元~5506 元之間，看診總花費中間值是 3034 元，看診總花費最高的是 43066 元；男生看診總花費有離群點，大多數的男生看診總花費都落在 1110 元~4702 元之間，看診總花費中間值是 2437 元，看診總花費最高的是 42102 元，再用 t-test 我們得到 p-value < 2.29 6*10 的-9 次方 < 0.05 ，有拒絕掉 h0: mu1=mu2，意思為有拒絕掉【男生平均看診總花費=女生平均看診總花費】，所以男生平均看診總花費和女生平均看診總花費在統計上有差異性，以及由 t-test 可以得知 95%信賴區間[315.0278,622.1981]，還得到男生平均看診總花費 3656.18 0，女生平均看診總花費 4124.793，可以看出女生的平均看診總花費會高於男生的平均看診總花費。

## 3.2-1 有無糖尿病 v.s.年齡

年齡用有無糖尿病分組，由盒鬚圖(圖 24)可以看到無糖尿病年齡有離群點，沒有糖尿病的年齡大多落在 20.3 歲~50.7 歲之間，最小年齡 2.8 歲，最大年齡 101.7 歲，中間值年齡 35.7 歲；有糖尿病年齡有離群點，有糖尿病的年齡大多落在 52.55 歲~72.5 歲之間，最小年齡 6.6 歲，最大年齡 97.7 歲，中間值年齡 62.4 歲。再用 t-test 我們得到 p-value < 2.2*10 的-16 次方 < 0.05 ，有拒絕掉 h0: mu1=mu2，意思為有拒絕掉【無糖尿病的平均年齡=有糖尿病的平均年齡】，所以無糖尿病的平均年齡和有糖尿病的平均年齡在統計上有差異性，以及由 t-test 可以得知 95%信賴區間[-26.48860-24.15033]，還得到無糖尿病的平均年齡 36.70050，有糖尿病的平均年齡 62.01997，可以看出有糖尿病的平均年齡比沒有糖尿病的平均年齡高。

## 3.2-2 有無糖尿病 v.s.看診高血壓次數

看診高血壓次數用有無糖尿病分組，由盒鬚圖(圖 25)可以看到無糖尿病看診高血壓次數有離群點，沒有糖尿病去看診高血壓次數最多 44 次；有糖尿病看診高血壓次數有離群點，有糖尿病去看診高血壓次數大多落在 0 次到 13 次之間，最多 34 次，中間值是 4 次，再用 t-test 我們得到 p-value < 2.2*10 的-16 次方 < 0.05 ，有拒絕掉 h0: mu1=mu2，意思為有拒絕掉【無糖尿病

的平均看診高血壓次數=有糖尿病的平均看診高血壓次數】，所以無糖尿病的平均看診高血壓次數和有糖尿病的平均看診高血壓次數在統計上有差異性，以及由 t-test 可以得知 95%信賴區間[-6.343934-5.187151]，還得到無糖尿病的平均看診高血壓次數 0.8857632，有糖尿病的平均看診高血壓次數 6.6513057，可以看出有糖尿病的平均看診高血壓次數會高於沒有糖尿病的平均看診高血壓次數。

## 3.2-3 有無糖尿病 v.s.看診高血壓花費

看診高血壓花費用有無糖尿病分組，由盒鬚圖(圖 26)可以看到無糖尿病看診高血壓花費有離群點，看診高血壓花費最高為 9177 元；有糖尿病看診高血壓花費有離群點，大多數看診高血壓花費落在 0 元~2428 元之間，最多看診高血壓花費 7675 元，看診高血壓花費中間值 886 元，再用 t-test 我們得到 p-value < 2.2*10 的-16 次方 < 0.05 ，有拒絕掉 h0: mu1=mu2，意思為有拒絕掉【無糖尿病的平均看診高血壓診費=有糖尿病的平均看診高血壓診費】，所以無糖尿病的平均看診高血壓診費和有糖尿病的平均看診高血壓診費在統計上有差異性，以及由 t-test 可以得知 95%信賴區間[-1259.754,-1029.561]，還得到無糖尿病的平均看診高血壓診費 206.8958，有糖尿病的平均看診高血壓診費 1351.5530，可以看出有糖尿病的平均看診高血壓花費會高於沒有糖尿病的平均看診高血壓花費。

## 3.2-4 有無糖尿病 v.s.看診總次數

看診總次數用有無糖尿病分組，由盒鬚圖(圖 27)可以看到無糖尿病看診總次數有離群點，資料中無糖尿病看診總次數大多落在 5 次~20 次之間，最少 1 次，最多 183 次，中間值 10 次；有糖尿病看診總次數有離群點，有糖尿病看診總次數大多落在 18 次~44 次之間，最少 1 次，最多 200 次，中間值 28 次。再用 t-test 我們得到 p-value < 2.2*10 的-16 次方 < 0.05 ，有拒絕掉 h0: mu1=mu2，意思為有拒絕掉【無糖尿病的平均看診總次數=有糖尿病的平均看診總次數】，所以無糖尿病的平均看診總次數和有糖尿病的平均看診總次數在統計上有差異性，以及由 t-test 可以得知 95%信賴區間[-20.80536,-17.05225]，還得到無糖尿病的平均看診總次數 15.01744，有糖尿病的平均看診總次數 33.94624，可以看出有糖尿病的平均看診總次數會高於沒有糖尿病的平均看診總次數。

## 3.2-5 有無糖尿病 v.s.看診總花費

看診總花費用有無糖尿病分組，由盒鬚圖(圖 28)可以看到無糖尿病看診總花費有離群點，沒有糖尿病的人看診總花費大多落在 1210 元~4810 元，最少 0 元，最多 43066 元，總花費中間值 2580 元；有糖尿病看診總花費有離群點，有糖尿病的人看診總花費大多落在 3810 元~9321 元，最少 222 元，最多 40094 元，總花費中間值 6050 元。再用 t-test 我們得到 p-value < 2.2*10 的-16 次方 < 0.05 ，有拒絕掉 h0: mu1=mu2，意思為有拒絕掉【無糖尿病的平均看診總花費=有糖尿病的平均看診總花費】，所以無糖尿病的平均看診總花費和有糖尿病的平均看診總花費在統計上有差異性，以及由 t-test 可以得知 95%信賴區間[-4188.109,-3347.343]，還得到無糖

尿病的平均看診總花費 3651.627，有糖尿病的平均看診總花費 7419.353，可以看出有糖尿病的平均看診總花費會高於沒有糖尿病的平均看診總花費。

接下來看用有無高血壓分組的:

## 3.3-1 有無高血壓 v.s. 年齡

年齡用有無高血壓分組由盒鬚圖(圖 29)可以看到無高血壓年齡有離群點，資料中大多數沒有罹患高血壓的人年齡都落在 18.9 歲~47.3 歲之間，沒有高血壓的最小年齡是 2.8 歲，最大年齡是 98.4 歲，中間值年齡是 33.3 歲；有高血壓年齡有離群點，資料中大多數罹患高血壓的人年齡都落在 53.3 歲~74.1 歲之間，罹患高血壓的最小年齡是 20.5 歲，最大年齡是 101.7 歲，中間值年齡是 64.2 歲，再用 t-test 我們得到 p-value < 2.2*10 的-16 次方 < 0.05 ，有拒絕掉 h0: mu1=mu2，意思為有拒絕掉【無高血壓的平均年齡=有高血壓的平均年齡】，所以無高血壓的平均年齡和有高血壓的平均年齡在統計上有差異性，以及由 t-test 可以得知 95%信賴區間[-30.14392,-28.51544]，還得到無高血壓的平均年齡 34.28664，有高血壓的平均年齡 63.61632，可以看出有高血壓的平均年齡比沒有高血壓的平均年齡高。

## 3.3-2 有無高血壓 v.s. 看診高血壓次數

看診高血壓次數用有無高血壓分組，由盒鬚圖(圖 30)可以看到沒有高血壓沒有離群點，因為沒有高血壓的人看診高血壓次數一定為 0；有高血壓的人看診高血壓次數有離群點，資料中有高血壓的人看診高血壓次數大多數落在 4 次~13 次之間，最少看診高血壓 1 次，最多看診高血壓 44 次，看診高血壓次數中間值是 9.105 次，再用 t-test 我們得到 p-value < 2.2*10 的-16 次方 < 0.05 ，有拒絕掉 h0: mu1=mu2，意思為有拒絕掉【無高血壓的平均看診高血壓次數=有高血壓的平均看診高血壓次數】，所以無高血壓的平均看診高血壓次數和有高血壓的平均看診高血壓次數在統計上有差異性，以及由 t-test 可以得知 95%信賴區間[-9.434629,-8.776201]，還得到無高血壓平均看診高血壓次數 0.0，有高血壓平均看診高血壓次數 9.105415。

## 3.3-3 有無高血壓 v.s. 看診高血壓診費

看診高血壓診費用有無高血壓分組，由盒鬚圖(圖 31)可以看到沒有高血壓沒有離群點，因為沒有高血壓的人看診高血壓診費一定為 0；有高血壓的人看診高血壓診費有離群點，資料中有高血壓的人看診高血壓診費大多數落在 1065 元~2779 元之間，最少花費 0 元，最多花費 9177 元，看診高血壓花費中間值是 1985 元，再用 t-test 我們得到 p-value < 2.2*10 的-16 次方 < 0.05 ，有拒絕掉 h0: mu1=mu2，意思為有拒絕掉【無高血壓的平均看診高血壓診費=有高血壓的平均看診高血壓診費】，所以無高血壓的平均看診高血壓診費和有高血壓的平均看診高血壓診費在統計上有差異性，以及由 t-test 可以得知 95%信賴區間[-2097.794-1965.932]，還得到無高血壓平均看診高血壓診費 0.0，有高血壓平均看診高血壓診費 2031.863。

## 3.3-4 有無高血壓 v.s.看診總次數

看診總次數用有無高血壓分組，由盒鬚圖(圖 32)可以看到無高血壓的人看診總次數有離群點，資料中大多數沒有高血壓的人看診總次數落在 5 次~19 次之間，最少看診 1 次，最多看診 183 次，中間值次數 10 次；有高血壓年齡的人看診總次數有離群點，大多數有高血壓的人看診總次數落 16 次~40 次之間，最少看診 1 次，最多看診 200 次，中間值次數 26 次，再用 t-test 我們得到 p-value < 2.2*10 的-16 次方 < 0.05 ，有拒絕掉 h0: mu1=mu2，意思為有拒絕掉【無高血壓的平均看診總次數=有高血壓的平均看診總次數】，所以無高血壓的平均看診總次數和有高血壓的平均看診總次數在統計上有差異性，以及由 t-test 可以得知 95%信賴區間[-18.54553-16.10340]，還得到無高血壓平均看診總次數 13.85026，有高血壓平均看診總次數 31.17473，可以看出有高血壓的人平均看診總次數會比沒有高血壓的人平均看診總次數多。

## 3.3-5 有無高血壓 v.s.看診總花費

看診總花費用有無高血壓分組，由盒鬚圖(圖 33)可以看到無高血壓的人看診總花費有離群點，資料中沒有高血壓的人大多數看診總花費落在 1138 元~4470 元，最多總花費 43066 元，最小總花費 0 元，看診總花費中間值 2380 元；有高血壓年齡的人看診總花費有離群點，有高血壓的人大多數看診總花費落在 3644 元~9079 元，最多總花費 42102 元，最小總花費 220 元，看診總花費中間值 5690 元。再用 t-test 我們得到 p-value < 2.2*10 的-16 次方 < 0.05 ，有拒絕掉 h0: mu1=mu2，意思為有拒絕掉【無高血壓的平均看診總花費=有高血壓的平均看診總花費】，所以無高血壓的平均看診總花費和有高血壓的平均看診總花費在統計上有差異性，以及由 t-test 可以得知 95%信賴區間[-4016.550-3445.964]，還得到無高血壓平均看診總花費 3380.127，有高血壓平均看診總花費 7111.384，可以看出有高血壓的平均看診總花費比沒有高血壓的平均看診總花費多。

接下來要進一步，我們去判斷他們分別與高血壓的獨立關係，由不連續的開始，因為連續的要做切割。

## 4.1 不連續型態與高血壓的獨立關係

由 chisq.test 看到 p-value，觀察是否拒絕 H0，若 p-value >0.05：沒有拒絕 H0；
若 p-value <0.05，拒絕 H0；
首先是性別，p-value =0.6032，H0:有無高血壓與性別不相關，而這邊沒有拒絕 H0，所以有無高血壓與性別不相關。
再來是有無高血壓與有無糖尿病的關係，p-value < 2.2 乘以 10 的-16 次方，沒有拒絕 H0，所以有無高血壓與有無糖尿病不相關。

## 4.2 連續型態與高血壓的獨立關係

接下來是連續型態，我們將其做切割，但因為「看診高血壓次數」及「看診高血壓診費」是「看診總次數」與「看診總花費」在確定有高血壓的前提下做出的延伸，所以無法做 chisq.test ，資料無法收斂。所以剩下「年齡」、「看診總次數」、「看診總花費」：

### 4.2-1 年齡與高血壓的獨立關係

首先是年齡，將其切割為 4 層，第一個年齡層：[2.8,21.5]，有 2513 人；第二個年齡層：[21.5,37.6]，有 2494 人；第三個年齡層：[37.6,52.9]，有 2494 人；第四個年齡層：[52.9,102]，有 2499 人；接下來就可以利用 chisq.test 得到 p-value <2.2*10 的-16 次方，拒絕「H0：有無高血壓與年齡不相關」，所以有無高血壓與年齡相關；再由 prop.test 我們得到第一個年齡層患有高血壓的比率為 0.0003979308，第二個為 0.0152365678，第三個為 0.1178829190，第四個為 0.4209683874，並不難看出年齡層越往上，患有高血壓比率越大。

### 4.2-2 看診總次數與高血壓的獨立關係

再來是「看診總次數」，一樣切割為四層，第一個分層[1,5]，有 2643；第二個分層[5,11] ，有 2430 人；第三個分層[11,22] ，有 2576 人；第四個分層[22,200]，有 2351 人；p-value < 2.2*10 的-16 次方，拒絕 H0，所以有無高血壓與看診總次數相關；第一個分層患有高血壓的比率為 0.01740446，第二個為 0.05267490，第三個為 0.15916149，第四個為 0.34070608，並不難看出看診總次數越多，患有高血壓比率越高。

### 4.2-3 看診總花費與高血壓的獨立關係

最後是「看診總花費」，也是切割為 4 層，第一個花費層[0,1300]，有 2511 人；第二個花費層[1300,2750] ，有 2489 人；第三個花費層[2750,5120] ，有 2500 人；第四個花費層[5120,43100]，有 2500 人；p-value <2.2*10 的-16 次方，拒絕 H0，所以有無高血壓與看診總花費相關；第一個花費層患有高血壓比率為 0.01712465，第二個為 0.0618724，第三個為 0.166，第四個為 0.3092，並不難看出看診總花費越多，患有高血壓比率越高。

## 5.1 p-value 大小

我們得知【有無高血壓 v.s.性別】的 p-value 為 0.583，【有無高血壓 v.s.年齡】、【有無高血壓 v.s.看診總次數】、【有無高血壓 v.s.看診總花費】、【有無高血壓 v.s.有無糖尿病】

的 p-value 都小於 2.2*10 的 16 次方小於 0.05，所以要比較他們的 z-value 大小，z-value 越大則 p-value 越小，反之，z-value 越小則 p-value 越大；我們知道【有無高血壓 v.s.年齡】的 z-value 為 39.33、【有無高血壓 v.s.看診總次數】的 z-value 為 29.36、【有無高血壓 v.s.看診總花費】的 z-value 為 27.10、【有無高血壓 v.s.有無糖尿病】的 z-value 為 29.10，所以我們知道 p-value 大小為: 性別 > 總花費 > 有無糖尿病 > 看診總次數 > 年齡 。

## 5.2 Forward、Backward、Stepwise

知道 p-value 以後就可以做 Forward、Backward、Stepwise

### 5.2-1 Forward

先做 p-value 最小的【有無高血壓 v.s.年齡】，還可以繼續，再把 p-value 第二小的【有無高血壓 v.s.看診總次數】加進去，還可以繼續做，再把 p-value 第三小的【有無高血壓 v.s.有無糖尿病】加進去，還可以繼續做，再把 p-value 第四小的【有無高血壓 v.s.看診總花費】加進去，發現不能繼續做下去，所以 Forward 的結論是【有無高血壓 v.s.年齡+看診總次數+有無糖尿病】，bata 0 是-6.309045，bata 1 是 0.076670，bata 2 是 0.025983，bata 3 是 1.217404。我們可以得到方程式:

$$\pi(x) = e^{\beta 0+\beta 1+\beta 2+\beta 3}/(1+e^{\beta 0+\beta 1+\beta 2+\beta 3})$$

$$\pi(x) = \frac{e^{-6.309045+0.07667*x1+0.025983*x2+1.217404*x3}}{(1+e^{-6.309045+0.07667*x1+0.025983*x2+1.217404*x3})}$$

我們假設【年齡 30 歲、看診總次數 5 次、沒有糖尿病】為標準，得到高血壓的機率是 0.02025223，接下來我們改變標準資料的年齡，變成【年齡 60 歲、看診總次數 5 次、沒有糖尿病】，得到高血壓的機率是 0.1709471，接下來改變標準資料的看診總次數，【年齡 30 歲、看診總次數 10 次、沒有糖尿病】，得到高血壓的機率是 0.02299725，接下來改變標準資料的是否有糖尿病，【年齡 30 歲、看診總次數 5 次、有糖尿病】，得到高血壓的機率是 0.0652760。
由比較標準資料得出結論，年齡越大得高血壓的機率越高，看診總次數越多得高血壓的機率越大，有糖尿病得高血壓的機率也會比沒有糖尿病的機率高。

### 5.2-2 Backward

Backward 是全部相加後若是有 p-value > 0.05 的狀況出現，要將最大的 p-value 拿掉。
先做全部相加【有無高血壓 v.s.年齡+看診總次數+有無糖尿病+看診總花費+性別】，【有無高血壓 v.s.看診總次數】 > 0.05，接下來做【有無高血壓 v.s.年齡+有無糖尿病+看診總花費+性別】都小於 0.05，所以 Backward 的結論是【有無高血壓 v.s.年齡+有無糖尿病+看診總花費+性別】，bata 0 是-6.506，bata 1 是 0.07826，bata 2 是 1.252，bata 3 是 0.0001123，bata 4 是 0.2196。我們可以得到方程式:

$$\pi(x) = \frac{e^{-6.506+0.07826*x1+1.252*x2+0.0001123*x3+0.2196*x4}}{(1+e^{-6.506+0.07826*x1+1.252*x2+0.0001123*x3+0.2196*x4})}$$

我們假設【年齡 30 歲、沒有糖尿病、看診總花費 1000 元、女性】為標準，得到高血壓的機率是 0.01719318，接下來我們改變標準資料的年齡，變成【年齡 60 歲、沒有糖尿病、看診總花費 1000 元、女性】，得到高血壓的機率是 0.1547136，接下來改變標準資料的有無糖尿病，【年齡 30 歲、有糖尿病、看診總花費 1000 元、女性】，得到高血壓的機率是 0.0576547，接下來改變標準資料的看診總花費，變成【年齡 30 歲、沒有糖尿病、看診總花費 3000 元、女性】，得到高血壓的機率是 0.02143001，接下來改變標準資料的性別，【年齡 30 歲、沒有糖尿病、看診總花費 1000 元、男性】，得到高血壓的機率是 0.02132541，由比較標準資料得出結論，年齡越大得高血壓的機率越高，有糖尿病得高血壓的機率也會比沒有糖尿病的機率高，看診總花費越多得高血壓的機率越大，男生得高血壓的機率會比女生高。

## 5.2-3 Stepwise

Stepwise 是把 Forward 不能做的拿掉以後，再加入還沒加進去過的資料中 p-value 最小的。所以我們從 Forward 的結論可以知道【有無高血壓 v.s.看診總花費】不能繼續做下去，所以我們把它拿出來變成【有無高血壓 v.s.年齡+看診總次數+有無糖尿病】，再加上性別，發現可以做，所以 Stepwise 結論是【有無高血壓 v.s.年齡+看診總次數+有無糖尿病+性別】，
bata 0 是-6.441314，bata 1 是 0.076797，bata 2 是 0.026602，bata 3 是 1.207539，
bata 4 是 0.235294。我們可以得到方程式:

$$\pi(x)=\frac{e^{-6.441314+0.076797*x1+0.026602*x2+1.207539*x3+0.235294*x4}}{(1+e^{-6.441314+0.076797*x1+0.026602*x2+1.207539*x3+0.235294*x4})}$$

我們假設【年齡 30 歲、看診總次數 5 次、沒有糖尿病、女性】為標準，得到高血壓的機率是 0.01790876，接下來我們改變標準資料的年齡，變成【年齡 60 歲、看診總次數 5 次、沒有糖尿病、女性】，得到高血壓的機率是 0.1544021，接下來我們改變標準資料的看診總次數，【年齡 30 歲、看診總次數 10 次、沒有糖尿病、女性】，得到高血壓的機率是 0.02040451，接下來我們改變標準資料的有無糖尿病，【年齡 30 歲、看診總次數 5 次、有糖尿病、女性】，得到高血壓的機率是 0.05749456，接下來我們改變標準資料的性別，【年齡 30 歲、看診總次數 5 次、沒有糖尿病、男性】，得到高血壓的機率是 0.02255247，由比較標準資料得出結論，年齡越大得高血壓的機率越高，看診總次數越多得高血壓的機率越大，有糖尿病得高血壓的機率也會比沒有糖尿病的機率高，男生得高血壓的機率會比女生高。

最後我們要由資料統整去預測「什麼樣的人一年平均看診高血壓次數」，例如出現了一位「沒有糖尿病看診總次數 10 次且看診高血壓花費$500 元看診總花費$5000 的 30 歲女性」，我們可以由先前的統計資料去預測出來她一年平均看診高血壓的次數。

## 6.1 p-value 大小

要得到估計模型前，首先要比較「看診高血壓次數」與我們要得到這些條件（性別、年齡、是否有高血壓、看診高血壓花費、、看診總次數、看診總花費、是否有糖尿病）的 p-value 大小，而後才可以做「Forward selection 」、「Backward selection 」、「Stepwise selection」。這邊小注意一下，因為目的是「看診高血壓次數」，所以前提是「要有高血壓」，所以我們做【看診高血壓次數 v.s. 是否有高血壓】是沒有意義的，但我們還是做了，而結果的模型也確實沒有「是否有高血壓」這項，這部分後續會再提。

【看診高血壓次數 v.s. 性別】的 p-value 為 4.77*10 的-7 次方；【看診高血壓次數 v.s. 是否有高血壓】的 p-value 為 0.893；而【看診高血壓次數 v.s. 年齡】、【看診高血壓次數 v.s. 看診高血壓花費】、【看診高血壓次數 v.s. 看診總次數】、【看診高血壓次數 v.s. 看診總花費】、【看診高血壓次數 v.s. 是否有糖尿病】的 p-value 都小於 2.2*10 的-16 次方 < 0.05，因此我們要判斷 z-value 大小，z-value 越大則 p-value 越小；【看診高血壓次數 v.s. 年齡】的 z-value 為 130.82，【看診高血壓次數 v.s. 看診高血壓花費】的 z-value 為 251.16 ，【看診高血壓次數 v.s. 看診總次數】的 z-value 為 132.22，【看診高血壓次數 v.s. 看診總花費】的 z-value 為 113.81，【看診高血壓次數 v.s. 是否有糖尿病】的 z-value 為 107.5。

所以 p- value 的比大小為：是否有高血壓 > 性別 > 是否有糖尿病 > 看診總花費 > 年齡 > 看診總次數 > 看診高血壓花費

## 6.2 Forward、Backward、Stepwise

### 6.2-1 Forward

先做 p-value 最小的【看診高血壓次數 v.s. 看診高血壓花費】，還可以繼續，再把 p-value 第二小的【看診高血壓次數 v.s. 看診總次數】放進去，還可以繼續，再把 p-value 第三小的【看診高血壓次數 v.s. 年齡】放進去，還可以繼續，再把 p-value 第四小的【看診高血壓次數 v.s. 看診總花費】放進去，還可以繼續，再把 p-value 第五小的【看診高血壓次數 v.s. 是否有糖尿病】放進去，還可以繼續，再把 p-value 第六小的【看診高血壓次數 v.s. 性別】放進去，還可以繼續，最後是把 p-value 第七小也就是最大的【看診高血壓次數 v.s. 是否有高血壓】放進去，出現了 p-value > 0.05 也就是沒有拒絕掉的狀況出現，所以並不能再繼續做下去。

beta 0 是 -2.443 ，beta 1 是 -0.0006663，beta 2 是 0.02156，beta 3 是 0.03708，beta 4 是 -0.0001257，beta 5 是 0.4615，beta 6 是 0.1501，得到估計模型:

$$y = e^{-2.443+0.0006663*x1+0.02156*x2+0.03708*x3-0.0001257*x4+0.4615*x5+0.1501*x6}$$

## 6.2-2, 6.2-3 Backward, Stepwise

而 Backward 及 Stepwise 的結論都一樣，因為Backward 是全部相加後若是有 p-value > 0.05 的狀況出現，要將最大的 p-value 拿掉，而這個便是【看診高血壓次數 v.s.是否有高血壓】，然後就能得到估計模型了，而這個模型與 Forward 一樣。

Stepwise 是將剛剛 Forward 因【看診高血壓次數 v.s.是否有高血壓】停止，將這個移出去後繼續加入 p-value 第八小的，但已經沒有第八小了，所以 Forward 的結論就是 Stepwise 的結論。

$$y = e^{-2.443+0.0006663*x1+0.02156*x2+0.03708*x3-0.0001257*x4+0.4615*x5+0.1501*x6}$$

我們用最前面提過的【沒有糖尿病看診總次數 10 次且看診高血壓花費$500 元看診總花費$5000 元的 30 歲女性】來估計她每年平均看診高血壓次數，由模型得到她平均看診高血壓次數為 0.1253494；那我們將她定為標準資料，照 beta 順序整理為【看診高血壓花費$500 元、看診總次數 10 次、年齡 30 歲、看診總花費$5000 元、沒有糖尿病、女性】，首先改變標準資料的「看診高血壓花費」為$50 元，其平均看診高血壓次數為 0.1691761；第二個是改變標準資料的「看診總次數」為 5 次，其平均看診高血壓次數為 0.1125396；第三個是改變標準資料的「年齡」為 60 歲，其平均看診高血壓次數為 0.381296；第四個是改變標準資料的「看診總花費」為$10000 元，其平均看診高血壓次數為 0.06686029；第五個是改變標準資料的「是否有糖尿病」改成有，其平均看診高血壓次數為 0.1988608；最後一個是改變「性別」為男性，其平均看診高血壓次數為 0.1456498。

經由與標準資料的比較過後，我們得到「看診高血壓花費越低，平均看高血壓次數越高」、「看診總次數越低，平均看高血壓次數越低」、「年齡越大，平均看高血壓次數越高」、「看診總花費越高，平均看高血壓次數越低」、「患有糖尿病，平均看高血壓次數越高」、「男性，平均看高血壓次數越高」。

## 7.結論

我們從這 10000 個樣本中，得到當年齡越大得到高血壓的機率越高，看診總次數越多得到高血壓的機率越大，看診總花費越多得到高血壓的機率越大，男生得到高血壓的機率會比女生高，有糖尿病得高血壓的機率也會比沒有糖尿病的機率高；也得到「看診高血壓花費越低，平均看高血壓次數越高」、「看診總次數越低，平均看高血壓次數越低」、「年齡越大，平均看高血壓次數越高」、「看診總花費越高，平均看高血壓次數越低」、「患有糖尿病，平均看高血壓次數越高」、「男性，平均看高血壓次數越高」。

圖1

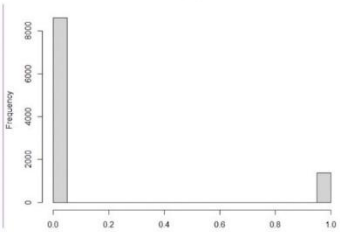



圖2

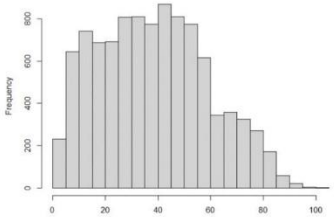



圖3

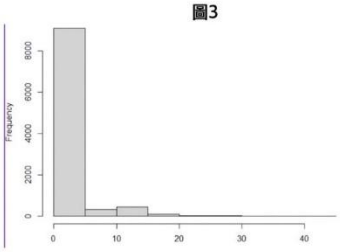



圖4

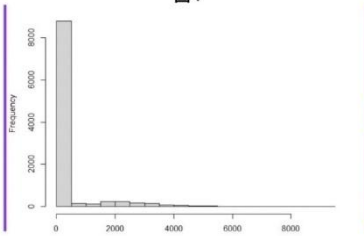



圖5

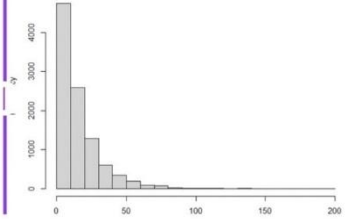



圖6

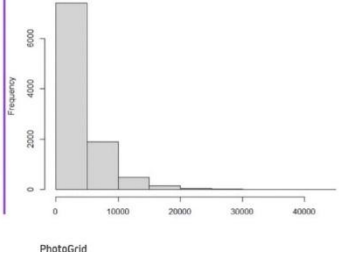



圖7

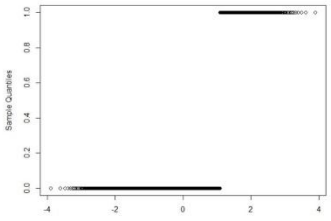



圖8

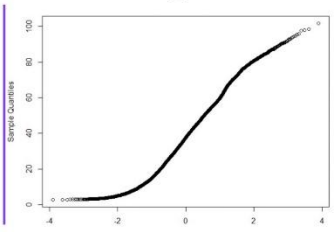



圖9

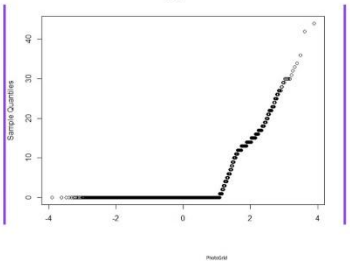



圖10

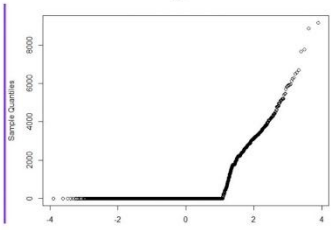



圖11

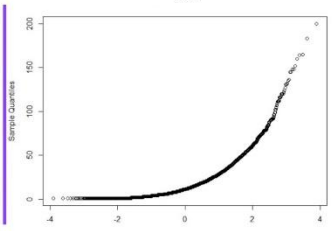



圖12

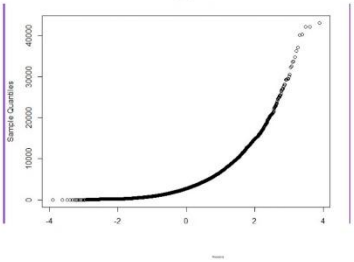



圖13

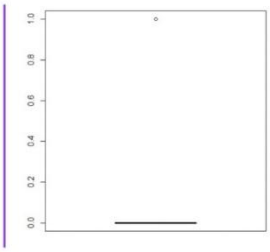



圖14

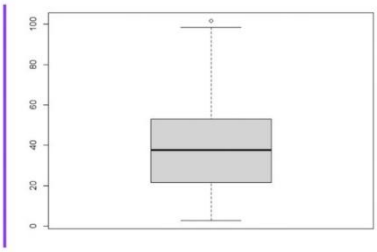



圖15

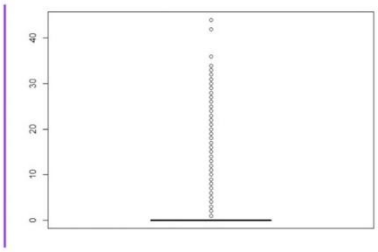


圖16

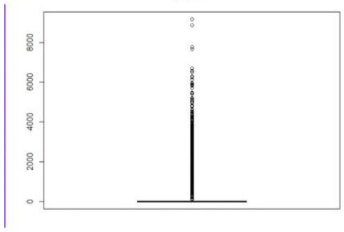


圖17

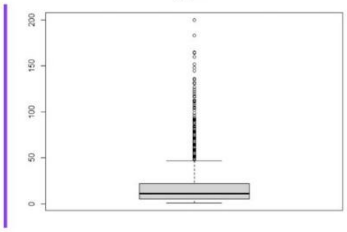


圖18

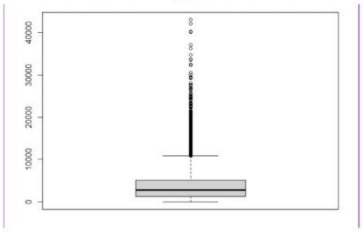


圖19

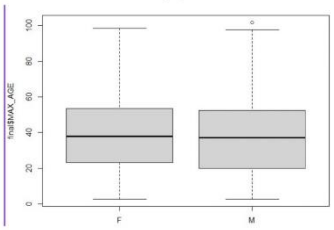



圖20

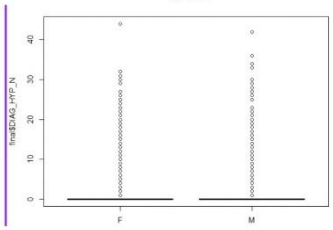



圖21

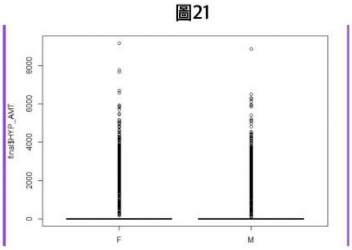



圖22

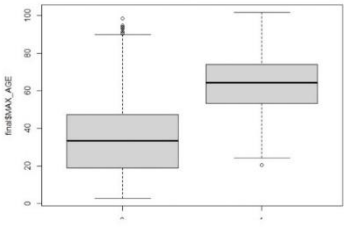



圖23

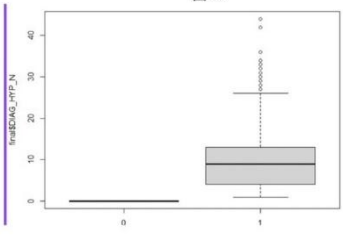



圖24

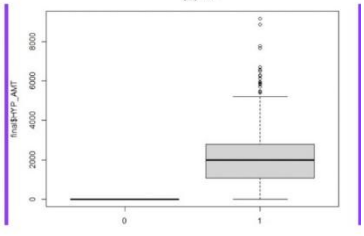



圖25

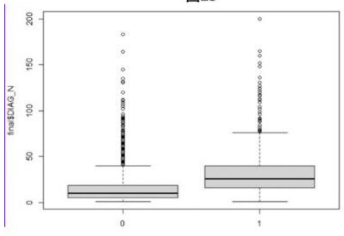



圖26

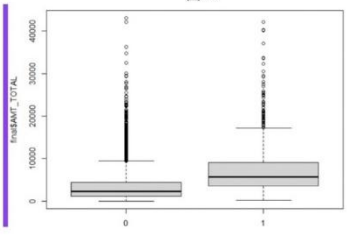



圖27

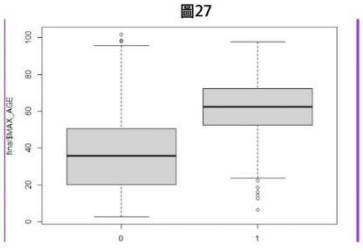



圖28

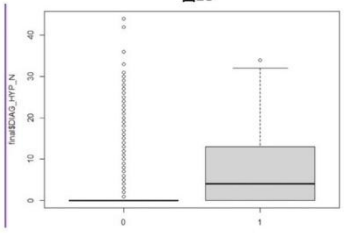



圖29

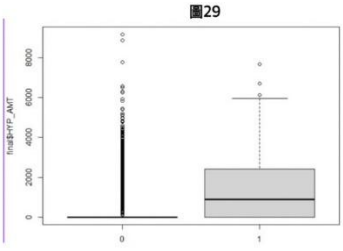



圖30

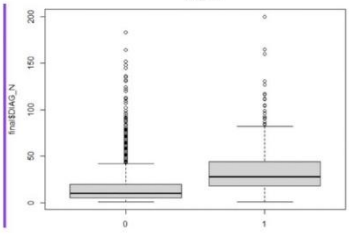


圖31

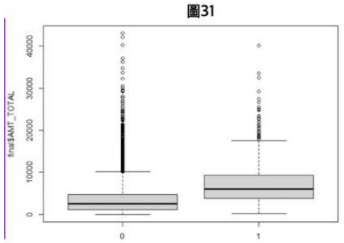

final$AMT_TOTAL
0
10000
20000
30000
40000
0
1

圖32

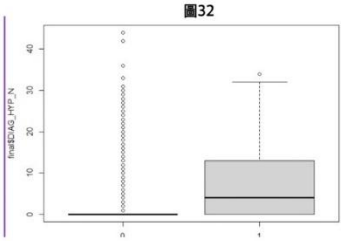
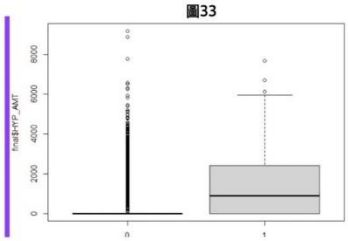

final$CHAS_HYP_N
0
10
20
30
40
0
1

圖33


final$HYP_AMT
0
2000
4000
6000
8000
0
1