亂填問卷與年齡、填卷時間之間的關係

什麼樣的人會亂填問卷？

問卷調查雖然能帶來大量的結果可供分析，但另一方面，常常也產生為數不小的廢卷(空白卷、不完整問卷或含有大量相同回答的問卷)，這也凸顯資料清理的重要性。恰巧，這次的問卷結果橫跨了許多不同的年齡層，並且也含有許多廢卷的情形產生。因此我希望透過此次的分析，藉由較客觀與能夠量化的年齡、填卷時間、性別等等要素來評估其與亂填問卷之間的關聯性。

在開始分析之前，先定義此次的廢卷分類，主要分為兩類：

- 亂填的卷：所有有答的題目皆回答相同答案者。
- 空白卷：除了個人資料，完全未作答者。

資料描述、EDA

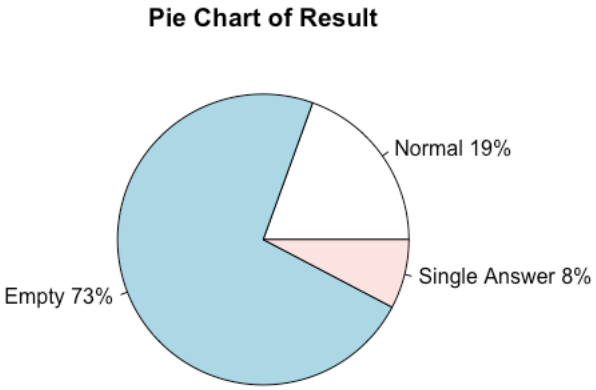
- Resposive Variable

此次的反應變數為”是否為廢卷“做區分的binary variable。資料來源由下面幾題的作答結果取得：

2-1. 如題 1.您選擇【健康風險預告】，在現在/未來生活中，你最擔心哪 2 個麻煩？ 備註：65 歲以上的填答者，問卷的時間副詞為現在；65 歲以下的填答者，問卷的時間副詞為未來	2-2. .這樣解決這個麻煩，您喜歡嗎？(填答題 2-1.選擇的項目)	非常不喜歡	不喜歡	普通	喜歡	非常喜歡
□ 我常忘記量血壓什麼的，資料不完整就沒用，健康手環也會忘記戴。(health_risk1)	手機定時提醒量測 (h_r1_s1)	□	□	□	□	□
	日常配戴首飾就有量測功能 (h_r1_s2)	□	□	□	□	□
	社區內有護理師幫忙量測、紀錄 (h_r1_s3)	□	□	□	□	□
	有專人打電話提醒量測 (h_r1_s4)	□	□	□	□	□
□ 量測的結果是數字和曲線圖，看不出我的健康狀況好不好。(health_risk2)	健康量測結果用顏色表達警示 (h_r2_s1)	□	□	□	□	□
	健康量測結果用圖片顯示身體狀況 (h_r2_s2)	□	□	□	□	□
	健康量測結果有專家幫忙分析解說 (h_r2_s3)	□	□	□	□	□
□ 就算測量到健康數字異常，若沒有人知道，就不會來關心我。(health_risk3)	健康量測設備會把數據傳給家人 (h_r3_s1)	□	□	□	□	□
	健康量測設備可選擇要分享的量測結果 (h_r3_s2)	□	□	□	□	□
	健康量測設備可提醒家人關心長輩 (h_r3_s3)	□	□	□	□	□
	有護理師定期關心健康狀況 (h_r3_s4)	□	□	□	□	□
□ 運動過量或動作不對就會受傷，我都不敢亂做運動。(health_risk4)	專家在旁指導正確運動姿勢 (h_r4_s1)	□	□	□	□	□
	有專家可請教合適的運動組合 (h_r4_s2)	□	□	□	□	□
	健康設備有運動過量的警告 (h_r4_s3)	□	□	□	□	□
	健康設備可分析個人最適運動量 (h_r4_s4)	□	□	□	□	□
□ 看到自己的健康數字異常，卻不知道怎麼做才能改善健康。(health_risk5)	健康量測設備，有作息改善建議 (h_r5_s1)	□	□	□	□	□
	健康量測設備，有飲食改善建議 (h_r5_s2)	□	□	□	□	□
	健康量測設備，有運動改善建議 (h_r5_s3)	□	□	□	□	□
	健康量測設備，有不當飲食作息警示 (h_r5_s4)	□	□	□	□	□
□ 身體突然不舒服，我不會自救，也無法通知別人來救我。(health_risk6)	健康偵測配備要給予自救建議 (h_r6_s1)	□	□	□	□	□
	健康偵測配備能發出警報請路人協助 (h_r6_s2)	□	□	□	□	□
	健康偵測配備有定位系統直接通報醫院 (h_r6_s3)	□	□	□	□	□
	健康偵測配備有定位系統自動通報家人 (h_r6_s4)	□	□	□	□	□

11-1. 如題 1.您選擇【學習數位科技】，在現在/未來生活中，你最擔心哪 2 個麻煩？ 備註：65 歲以上的填答者，問卷的時間副詞為現在；65 歲以下的填答者，問卷的時間副詞為未來	11-2. 這樣解決這個麻煩，您喜歡嗎？(請填答題 11-1. 選擇的項目)					
□ 我不太會操作 3C 產品，孩子被我問得不耐煩，嫌我學很慢又記不住。(tech_learn1)	電話客服提供諮詢 (t_l1_s1)	□	□	□	□	□
	數位裝置內建隨身數位助教 (t_l1_s2)	□	□	□	□	□
	影片教學操作步驟 (t_l1_s3)	□	□	□	□	□
□ 我不敢亂點亂按手機，怕按錯出問題。(tech_learn2)	回到初始畫面的按鍵 (t_l2_s1)	□	□	□	□	□
	電話客服提供諮詢 (t_l2_s2)	□	□	□	□	□
□ 政府和廠商有些消息只公告在網路上，我哪會知道。(tech_learn3)	鄰里長幫忙通知 (t_l3_s1)	□	□	□	□	□
	手機簡訊通知 (t_l3_s2)	□	□	□	□	□
	郵寄紙本通知 (t_l3_s3)	□	□	□	□	□
	手機中有數位佈告欄 (t_l3_s4)	□	□	□	□	□
□ 網路消息傳來傳去，都不知道是真是假。(tech_learn4)	政府主動刪除假消息 (t_l4_s1)	□	□	□	□	□
	政府主動通知訊息為假 (t_l4_s2)	□	□	□	□	□
	政府在假消息後加正確訊息連結 (t_l4_s3)	□	□	□	□	□
	可訂閱經過檢驗的資訊 (t_l4_s4)	□	□	□	□	□
□ 上網查資料好難，要打什麼字才找得到？(tech_learn5)	專人協助查詢資訊服務 (t_l5_s1)	□	□	□	□	□
	語音搜尋比對服務 (t_l5_s2)	□	□	□	□	□

上面幾題的作答結果如果符合亂填或空白卷的條件，則視為1, 否則視為0，最後取得一個length為68324的numeric list作為反應變數(取名為IsBad)。而中間統計的結果如下圖：



然而，看完資料的分佈後，由於空白的sample比例太大了，因此我另外做了一個只有正常卷和亂填卷的變數版本與模型來與之比較(取名為IsSingleAnswer)。

- Explanatory Variable
    - age: 問卷原有的年齡項目，以numeric的方式做變數。
    - sec: 問卷結果的填寫時間項目，以numeric的方式做變數，並以秒為單位。
    - language: 問卷原有的語言選擇，為categorical的形式，1為國語，2為台語，3為客語。
    - gender: 問卷原有的性別欄位，一樣是categorical的形式，1為男性，2為女性。

  選擇以上變數的主要理由有以下三點：

  1. 這些欄位並沒有空缺或遺漏的狀況。
  2. 其中有些變數較為客觀(如填寫時間)，比較不會有亂填的情形。
  3. 個人希望能以較能量化的numeric數據作分析。

模型建構

對於IsBad我的模型主要有以下2種，ageSQ為age的平方項：

```
model1 <- glm(formula = IsBad ~ language + sec + gender + ageSQ + age,
family = "binomial")
model2 <- glm(formula = IsBad ~ sec + ageSQ + age, family = "binomial")
```

對於IsSingleAnswer我的模型有以下2種，並且資料在regression之前都事先剔除空白卷的那幾筆資料：

```
model3 <- glm(formula = IsSingleAnswer ~ language + sec + gender + ageSQ
+ age, family = "binomial")
model4 <- glm(formula = IsSingleAnswer ~ sec + ageSQ + age, family =
"binomial")
```

變數選擇與模型評估

以下是model1的summary：

```
Call:
glm(formula = IsBad ~ language + sec + gender + ageSQ +
    age, family = "binomial")

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9264   0.6363   0.6613   0.6736   0.8343

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  9.332e+00  7.246e+01   0.129  0.89753
language1    4.082e-01  5.221e-01   0.782  0.43426
language2    4.610e-01  5.238e-01   0.880  0.37885
language3    1.457e-01  5.817e-01   0.251  0.80218
sec         -1.525e-04  5.331e-05  -2.861  0.00422 **
gender1     -8.167e+00  7.246e+01  -0.113  0.91026
gender2     -8.155e+00  7.246e+01  -0.113  0.91039
ageSQ        2.562e-03  1.173e-03   2.184  0.02893 *
age         -3.032e-02  1.809e-02  -1.676  0.09374 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 29888  on 30000  degrees of freedom
Residual deviance: 29870  on 29992  degrees of freedom
AIC: 29888

Number of Fisher Scoring iterations: 8
```

以下是model2的summary：

```
Call:
glm(formula = IsBad ~ sec + ageSQ + age, family = "binomial")
```

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9135   0.6381   0.6615   0.6734   0.7434

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.584e+00  8.166e-02  19.399  < 2e-16 ***
sec         -1.544e-04  5.307e-05  -2.910  0.00361 **
ageSQ        2.558e-03  1.171e-03   2.185  0.02890 *
age         -2.994e-02  1.807e-02  -1.657  0.09747 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 29888  on 30000  degrees of freedom
Residual deviance: 29874  on 29997  degrees of freedom
AIC: 29882

Number of Fisher Scoring iterations: 4
```

anova 比較結果：

```
Analysis of Deviance Table

Model 1: IsBad[2:30002] ~ language + sec + gender + ageSQ + age
Model 2: IsBad[2:30002] ~ sec + ageSQ + age
  Resid. Df Resid. Dev Df Deviance
1     29992      29870
2     29997      29874 -5  -3.3574
```

從比較結果與p-value不難看出，去掉categorical variable後的模型殘差並沒有太大的差異，且categorical variable的p-value都非常的大，可見這些variable對模型並沒有顯著的影響。因此我會傾向於用model2來做進一步的詮釋。

以下是model3的summary：

```
Call:
glm(formula = IsSingleAnswer ~ language + sec + gender + ageSQ + age,
family = "binomial")

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.2219  -0.8833  -0.8312   1.4404   1.9043

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.2587036  0.7466641   0.346  0.72898
language1   -0.2722610  0.7410967  -0.367  0.71334
```

```
language2   -0.1233392  0.7438383  -0.166  0.86830
language3   -0.0271994  0.8184796  -0.033  0.97349
sec         -0.0006001  0.0000756  -7.938 2.05e-15 ***
gender2     -0.0603774  0.0457240  -1.320  0.18668
ageSQ        0.0046451  0.0017115   2.714  0.00665 **
age         -0.0622911  0.0266902  -2.334  0.01960 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 12530  on 9999  degrees of freedom
Residual deviance: 12450  on 9992  degrees of freedom
AIC: 12466

Number of Fisher Scoring iterations: 4
```

以下是model4的summary：

```
Call:
glm(formula = IsSingleAnswer ~ sec + ageSQ + age, family = "binomial")

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.1741  -0.8839  -0.8347   1.4448   1.9020

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.0258377  0.1192665  -0.217  0.82849
sec         -0.0006173  0.0000754  -8.188 2.66e-16 ***
ageSQ        0.0048453  0.0017074   2.838  0.00454 **
age         -0.0636331  0.0266382  -2.389  0.01690 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 12530  on 9999  degrees of freedom
Residual deviance: 12457  on 9996  degrees of freedom
AIC: 12465

Number of Fisher Scoring iterations: 4
```

anova 比較結果：

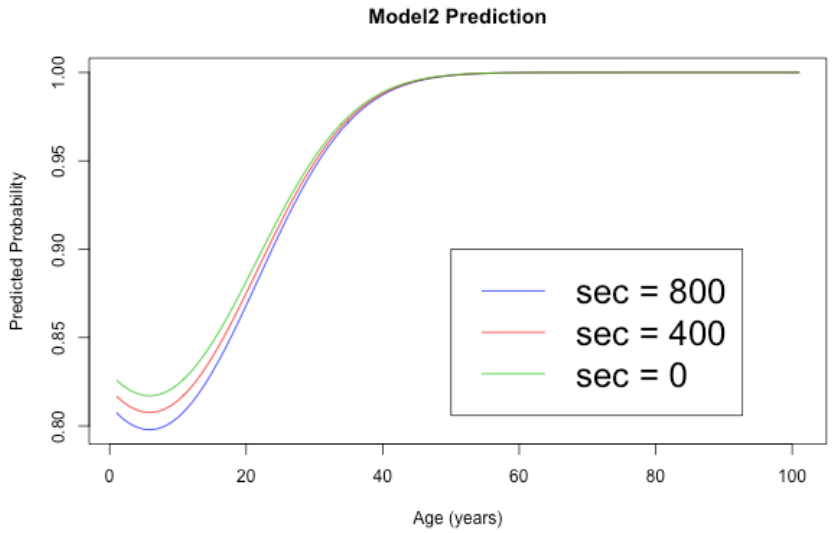
```
Analysis of Deviance Table

Model 1: IsSingleAnswer ~ language + sec + gender + ageSQ + age
Model 2: IsSingleAnswer~ sec + ageSQ + age
  Resid. Df Resid. Dev Df Deviance
1      9992      12450
2      9996      12457 -4  -7.1838
```

同樣地，從比較結果與p-value可以看出，去掉categorical variable後的模型殘差並沒有太大的變化，且categorical variable的p-value一樣都非常大。可以推斷這些variable對模型的影響並不顯著。因此我會選擇model4來做進一步的詮釋。

結果詮釋

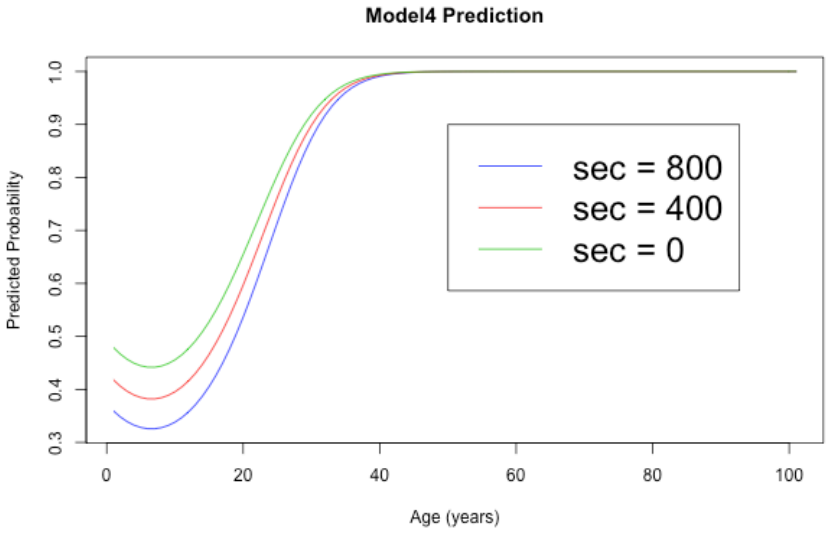
填卷時間與年齡對廢卷(空白卷、亂填的卷)的影響



模型: $logit(\hat{\pi}) = 1.584 + sec * -0.0001544 + age^2 * 0.002558 + age * -0.02994$

1. 由於資料本身便含有大量空白卷，所以模型預測的廢卷機率偏高。
2. 填表時間愈長，是廢卷的機率就愈低，可以由圖形與係數正負號判斷得到。
3. 廢卷機率對age而言是個二次函式，配方後約為 $0.002558(age - 5.85)^2$，可知在6歲之後年齡越高亂填或交空白卷的機率越高，6歲前則是年齡越小機率越高。
4. 45歲以後模型漸趨收斂，代表45歲以上的人亂填問卷或交空白卷的機率很大。

填卷時間與年齡對廢卷(僅含亂填的卷不含空白卷)的影響



模型: $logit(\hat{\pi}) = -0.0258377 + sec * -0.0006173 + age^2 * 0.0048453 + age * -0.0636331$

1. 這個模型用的資料是剔除空白卷的資料，可以發現去除空白卷後，是廢卷的機率較低。
2. 填表時間愈長，是廢卷的機率就愈低，可以由圖形與係數正負號判斷得到。
3. 廢卷機率對age而言是個二次函式，配方後約為 $0.0048453(age - 6.56648)^2$，可知在6.5歲之後年齡越高亂填的機率越高，6.5歲前則是年齡越小亂填的機率越大。
4. 40歲以後模型漸趨收斂，代表40歲以上的人亂填問卷的機率很大。