利用 Logistic Regression Model 解釋變數跟憂鬱症間的關係

回歸分析期末報告 B082040005 高念慈

一、動機

2020 年造成人類失能(disability)前十名的疾病，第一名為憂鬱症；人類社會整體疾病負擔(Global burden of Disease)第二名的疾病；聯合國世界衛生組織（WHO）將憂鬱症列為 2020 年需重視的疾病第二名，僅次於心血管疾病，並估計憂鬱症將成為 2030 年全球疾病負擔的主要原因，上述種種都在告訴我們憂鬱症是該被好好重視的疾病，但根據報告，高所得國家有 70%患有精神病的人能獲得治療，在低所得國家只有 12%，所以如果能利用模型預測是否有憂鬱的現象，就能初步篩選，減輕醫療資源不足的問題，把資源留給真正需要的人。

二、資料介紹

資料來源: https://www.kaggle.com/datasets/diegobabativa/depression

此資料來自 Kaggle，原始數據來自 Reference [1]，是 Busara 中心在 2015 年在肯亞西部維多利亞湖附近的 Siaya 農村進行的一項研究，背景源自每年有 130 萬肯亞人患有未經治療的重度抑鬱症(MDD，通常稱為憂鬱症)，而撒哈拉以南的非洲是世界上該病患病率最高的地區；然而，肯亞的心理健康治療缺乏資源和污名化，在肯亞，每百萬人中只有兩名經過認證的精神科醫生，城市地區以外的設施很少，人們也不太可能知道或訪問它們。

目的為對潛在病例進行定位可以使稀缺的資源到達最需要的人手中，改善或挽救無數人的生命，此須注意評判標準所使用的儀器是一種流行病學而非臨床憂鬱症測量方法；換句話說，它高度提示憂鬱症的存在，但不等同於診斷，診斷只能由有執照的臨床醫生做出。

變數介紹:

資料源頭 Reference [1]有著 1143 筆樣本，75 個變數，但多數變數並沒有詳細或直觀的介紹，故在此選擇使用 Kaggle 中的變數名稱對映資料源頭的 23 個變數，這是因為我發現 Kaggle 中的數據跟源頭數據相比有明顯錯誤(跟錢有關的變數小數點幾乎都錯)，所以取源頭資料對應 Kaggle 的變數作為分析的資料集。

1143 筆樣本，23 個變數

1. “surveyid”: int. 調查編號
2. “village”: int. 村莊編號
3. “femaleres”: int. 女受訪者為 1，男受訪者為 0
4. “age”: num. 受訪者年齡(歲)

5. “married”: int. 已婚為 1，未婚為 0
6. “children”: int. 受訪者家孩子數量(個)
7. “edu”: int. 受訪者接受幾年教育(年)
8. “hh_totalmembers”: int. 家庭總成員數(個)
9. “asset_livestock”: num. 家畜總價值(美元)
10. “asset_durable”: num. 耐用品總價值(美元)
11. “asset_savings”: num. 存款(美元)
12. “cons_social”: num. 每月生活開銷(美元)
13. “cons_other”: num. 其他支出(美元)
14. “ent_wagelabor”: int. 僱傭勞動創造的主要收入，是為 1，否為 0
15. “ent_ownfarm”: int. 主要收入來源為自己農場，是為 1，否為 0
16. “ent_business”: int. 農業非主要收入來源，是為 1，否為 0
17. “ent_nonagbusiness”: int. 非農業業主，是為 1，否為 0
18. “ent_farmrevenue”: num. 每月農場收入(美元)
19. “ent_farmexpenses”: num. 每月農場支出(美元)
20. “labor_primary”: int. 臨時工或受僱為主要收入，是為 1，否為 0
21. “durable_investment”: num. 長期投資金額(美元)
22. “nondurable_investment”: num. 短期投資金額(美元)
23. “depressed”: int. 達到中度憂鬱症的流行病學閾值，憂鬱為 1，否為 0

變數“depressed”為我們的目標變數。

三、EDA/資料前處理

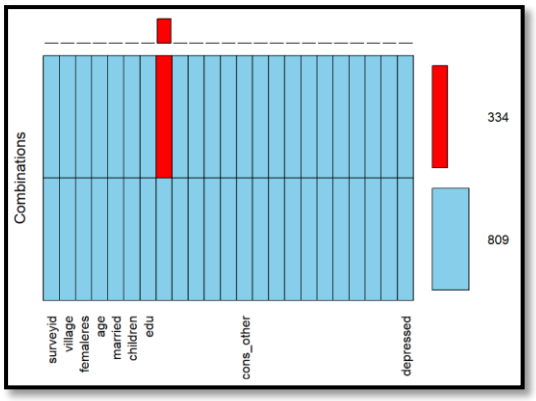
由(圖一)能看到整體資料在“hh_totalmembers”有 334 個缺失值，再看了(圖二) 得知 NA 值跟是否憂鬱比例差不多，可以推測此 NA 值有無對解釋目標影響不大，較不會有缺填的受訪者憂鬱傾向高的問題，但 Logistic Regression Model 是無法處理缺失值的，幸運的是，在處理完目標變數不平衡的問題後，會發現含缺失值的樣本已被自動捨去(圖三)。

目標變數不平衡(圖四)，處理前有 950 筆被檢測為 0，193 筆被檢測為 1，明顯有不平衡的問題，為了避免建出來的模型在未來預測上會有偏頗某一現象的問題，也避免補過多目標變數為 1 會讓模型侷限在這些固定特徵，在此採用了隨機過採樣和隨機欠採樣的組合，讓最後資料呈現 574 筆無症狀對上 569 筆診斷為憂鬱，總和一樣為 1143 筆(圖五)。

(圖六)則是可以觀察出目標變數幾乎對所有變數都沒有顯著線性相關，這點有點出乎我意料之外，留在最後探討；各變數間也只有少數幾個高度相關，如: 長期投資跟家畜價值、小孩數量跟家庭總成員、 主要收入為自有農場跟主要收入為僱傭等，推測會高度相關的原因為變數問題有重疊或排斥現象。

(圖七)、(圖八)則可以看出處理前後資料的分布並沒有明顯改變，受訪者大部分為 30 歲上下、已婚、婦女居多、家庭成員落在 2 至 9 為不等，這兩張圖所

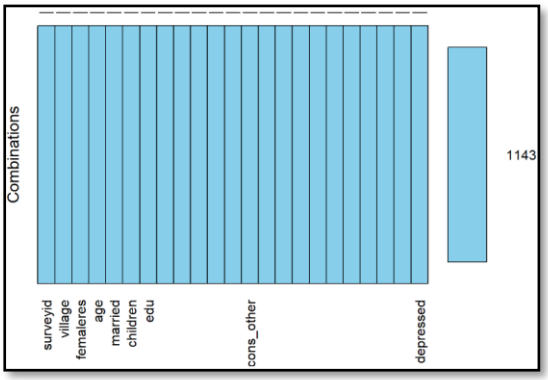
選取的變數為第四部份經過變數挑選後留下的 8 個變數加目標變數。



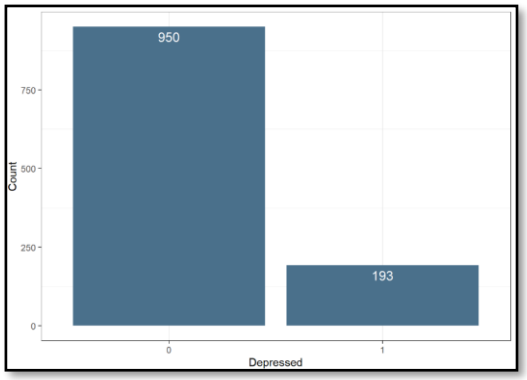
(圖一)

	0	1
FALSE	0.8417800	0.1582200
TRUE	0.8053892	0.1946108

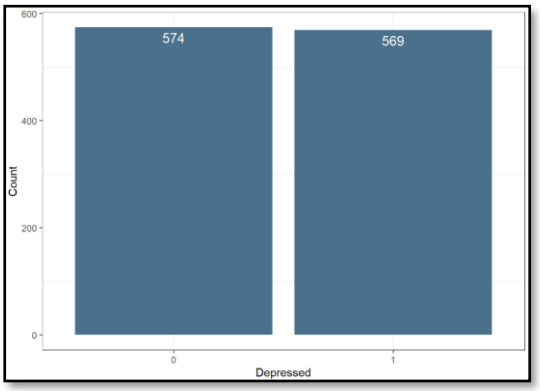
(圖二)



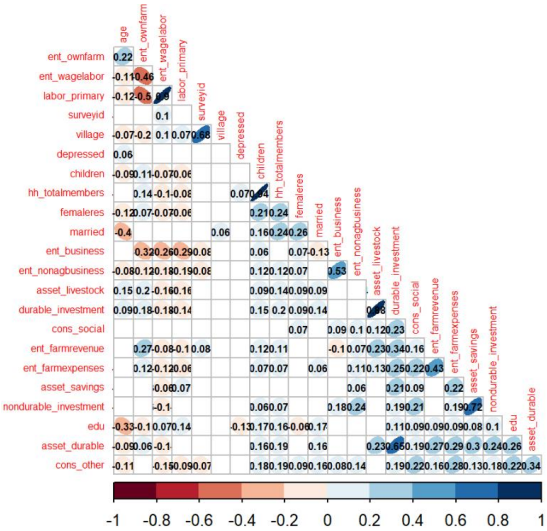
(圖三)



(圖四)



(圖五)



(圖六)


age
married
edu
_totalmembe
sset_saving
cons_social
ent_ownfarm
_farmexpen
depressed

(圖七) 處理前


age
married
edu
_totalmembe
sset_saving
cons_social
ent_ownfarm
_farmexpen
depressed

(圖八) 處理後

```
##
## Call:
## glm(formula = depressed ~ ., family = "binomial", data = balanced_sample)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7386  -1.1363  -0.4572   1.1536   1.5775
##
## Coefficients:
##                            Estimate Std. Error z value Pr(>|z|)
## (Intercept)            -1.720e-01  4.736e-01  -0.363 0.716520
## surveyid               -4.943e-05  2.000e-04  -0.247 0.804815
## village                 1.162e-03  1.299e-03   0.895 0.371007
## femaleres               5.714e-02  2.825e-01   0.202 0.839706
## age                     7.679e-03  5.786e-03   1.327 0.184426
## married                 2.103e-01  1.895e-01   1.110 0.267185
## children               -5.814e-02  1.079e-01  -0.539 0.590045
## edu                    -8.993e-02  2.406e-02  -3.737 0.000186 ***
## hh_totalmembers         1.399e-01  9.823e-02   1.424 0.154464
## asset_livestock        -3.901e-04  1.483e-03  -0.263 0.792527
## asset_durable          -1.263e-03  1.767e-03  -0.714 0.474963
## asset_savings          -1.315e-03  1.680e-03  -0.783 0.433707
## cons_social            -1.263e-02  9.051e-03  -1.396 0.162865
## cons_other              2.119e-04  2.688e-03   0.079 0.937169
## ent_wagelabor           1.703e-01  3.482e-01   0.489 0.624796
## ent_ownfarm            -2.977e-01  1.963e-01  -1.516 0.129405
## ent_business           -6.810e-03  2.489e-01  -0.027 0.978175
## ent_nonagbusiness       1.382e-01  1.594e-01   0.867 0.385951
## ent_farmrevenue        -1.156e-02  9.673e-03  -1.195 0.232098
## ent_farmexpenses        4.840e-02  1.975e-02   2.451 0.014262 *
## labor_primary          -1.262e-01  3.580e-01  -0.352 0.724495
## durable_investment      6.270e-04  1.443e-03   0.435 0.663906
## nondurable_investment -5.990e-04  7.782e-04  -0.770 0.441395
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1584.5  on 1142  degrees of freedom
## Residual deviance: 1530.1  on 1120  degrees of freedom
## AIC: 1576.1
##
## Number of Fisher Scoring iterations: 4
```

(圖九)

```
##            surveyid             village           femaleres
##            2.077374            2.099667            1.196954
##                 age             married            children
##            1.573181            1.544702           10.938885
##                 edu      hh_totalmembers      asset_livestock
##            1.310685           11.501110           52.481649
##       asset_durable       asset_savings         cons_social
##           16.849609            1.527300            1.266017
##          cons_other        ent_wagelabor         ent_ownfarm
##            1.330586            6.489656            2.424397
##        ent_business    ent_nonagbusiness      ent_farmrevenue
##            2.199140            1.608990            1.616358
##    ent_farmexpenses        labor_primary   durable_investment
##            1.409350            7.445057           84.424932
## nondurable_investment
##            1.631759
```

(圖十)

```
##            age        married            edu hh_totalmembers
##       1.388604       1.291804       1.179170       1.122454
##  asset_savings    cons_social   ent_ownfarm ent_farmexpenses
##       1.029293       1.078071       1.130663       1.110714
```

(圖十二)

```
##                          age                      married
##                     7.526356                    25.846225
##                          edu               hh_totalmembers
##                     7.127453                    13.838227
##                 asset_savings                  cons_social
##                    28.277737                    39.455420
##                   ent_ownfarm              ent_farmexpenses
##                    35.256985                     1.411386
##                  age:married               age:cons_social
##                    14.554829                    18.380328
##               age:ent_ownfarm       married:hh_totalmembers
##                    15.667648                    15.878554
##           married:ent_ownfarm           edu:hh_totalmembers
##                    14.013077                    17.084487
##             edu:asset_savings               edu:cons_social
##                    67.117523                    16.435688
##    hh_totalmembers:cons_social     asset_savings:cons_social
##                    14.548174                    24.636129
##      asset_savings:ent_ownfarm asset_savings:ent_farmexpenses
##                     2.220980                    19.125979
##        cons_social:ent_ownfarm
##                     2.748023
```

(圖十四)

```
##
## Call:
## glm(formula = depressed ~ age + married + edu + hh_totalmembers +
##     asset_savings + cons_social + ent_ownfarm + ent_farmexpenses,
##     family = "binomial", data = balanced_sample)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7166  -1.1399  -0.5116   1.1656   1.5795
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -0.028583   0.360248  -0.079  0.93676
## age               0.008978   0.005422   1.656  0.09777 .
## married           0.278228   0.172804   1.610  0.10738
## edu              -0.101840   0.022779  -4.471 7.79e-06 ***
## hh_totalmembers   0.092912   0.030638   3.033  0.00243 **
## asset_savings    -0.002055   0.001323  -1.554  0.12016
## cons_social      -0.012008   0.008397  -1.430  0.15267
## ent_ownfarm      -0.383226   0.133717  -2.866  0.00416 **
## ent_farmexpenses  0.037947   0.017285   2.195  0.02814 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1584.5  on 1142  degrees of freedom
## Residual deviance: 1536.3  on 1134  degrees of freedom
## AIC: 1554.3
##
## Number of Fisher Scoring iterations: 4
```

(圖十一)

```
## glm(formula = depressed ~ age + married + edu + hh_totalmembers +
##     asset_savings + cons_social + ent_ownfarm + ent_farmexpenses +
##     age:married + age:cons_social + age:ent_ownfarm + married:hh_totalmembers +
##     married:ent_ownfarm + edu:hh_totalmembers + edu:asset_savings +
##     edu:cons_social + hh_totalmembers:cons_social + asset_savings:cons_social +
##     asset_savings:ent_ownfarm + asset_savings:ent_farmexpenses +
##     cons_social:ent_ownfarm, family = "binomial", data = balanced_sample)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2734  -1.1198  -0.1006   1.0776   2.4099
##
## Coefficients:
##                                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)                    -1.2771471  0.9820854  -1.300 0.193449
## age                             0.0230152  0.0132014   1.743 0.081267 .
## married                         3.5884983  0.8576431   4.184 2.86e-05 ***
## edu                            -0.1945225  0.0592294  -3.284 0.001023 **
## hh_totalmembers                 0.3249754  0.1126404   2.885 0.003913 **
## asset_savings                   0.0083992  0.0083220   1.009 0.312838
## cons_social                    -0.1086059  0.0603720  -1.799 0.072027 .
## ent_ownfarm                    -4.4851411  0.7790912  -5.757 8.57e-09 ***
## ent_farmexpenses                0.0601565  0.0200263   3.004 0.002666 **
## age:married                    -0.0689624  0.0154208  -4.472 7.75e-06 ***
## age:cons_social                 0.0041701  0.0011134   3.745 0.000180 ***
## age:ent_ownfarm                 0.0520171  0.0122443   4.248 2.15e-05 ***
## married:hh_totalmembers        -0.3227938  0.0935478  -3.451 0.000559 ***
## married:ent_ownfarm             2.8164454  0.5060153   5.566 2.61e-08 ***
## edu:hh_totalmembers             0.0178462  0.0108925   1.638 0.101340
## edu:asset_savings              -0.0024896  0.0011357  -2.192 0.028367 *
## edu:cons_social                 0.0066449  0.0040536   1.639 0.101161
## hh_totalmembers:cons_social    -0.0220332  0.0062796  -3.509 0.000450 ***
## asset_savings:cons_social       0.0018915  0.0004271   4.429 9.46e-06 ***
## asset_savings:ent_ownfarm       0.0099590  0.0047619   2.091 0.036495 *
## asset_savings:ent_farmexpenses -0.0008424  0.0003469  -2.428 0.015176 *
## cons_social:ent_ownfarm        -0.0409958  0.0241497  -1.698 0.089589 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1584.5  on 1142  degrees of freedom
## Residual deviance: 1450.5  on 1121  degrees of freedom
## AIC: 1494.5
##
## Number of Fisher Scoring iterations: 5
```

(圖十三)

四、選取模型

目標為分類問題，故採取邏輯斯回歸，在此主要建了三個模型。

模型一:

因為資料不大，變數幾乎都跟錢有關，所以先建一個 full model 看是否如預期會顯著相關，結果如(圖九)，沒想到 23 個變數只有教育年限(edu)跟每月農場支出(ent_farmexpenses)有顯著相關，不只相關的變數少，就連唯二相關的也

只有一個是跟錢相關，再看了共線性後(圖十)還是覺得很神奇，可能是有未被發現的關係跟考量未被納入，於是模型二考慮挑選變數看結果會不會改善。

模型二:

對模型一使用了雙向逐步回歸算法以 AIC 作為標準選擇模型，最後得到的結果為(圖十一)，只剩 8 個變數，模型變得更簡潔易讀，顯著的變數也增加到 4 個，在共線性上(圖十二)也是表現不錯，但因為模型二只剩 8 個變數，雖然模型變好很多，但刪了 14 個變數，於是模型三考慮了交互項的影響。

模型三:

先拿模型二加交互向的結果，再做一次雙向逐步回歸算法也是以 AIC 作為標準選出最後的模型(圖十三)，可以看到選出了 21 個變數，顯著的變數增加到 15 個，其中一般變數 8 個中顯著的佔 5 個；交互項 13 個中顯著的佔 10 個，雖然比模型一好，但共線性的問題嚴重許多(圖十四)，猜測這可能在後續解釋變數上會出現問題。

五、分析結果及解釋

模型：

- General Linear Model(GLM): $y = Link(f(x)) = f(x) = b_0 + b_1x_1 + b_2x_2 + ..$
- Logistic Regression: $log(odd) = log(\frac{p}{1-p}) = b_0 + b_1x_1 + b_2x_2 + ... = f(x)$
- $odd = Exp(log(odd)) = \frac{p}{1-p}$ ， where $p = Pr[y = 1]$
- $Pr[y = 1] = prob = \frac{odd}{1+odd} = logistic(f(x)) = \frac{1}{1+\exp(-f(x))}$
- Logistic Function: $Logistic(F_x) = \frac{1}{1+Exp(-F_x)} = \frac{Exp(F_x)}{1+Exp(F_x)}$

係數：

- $Odd_0 = Exp(b_0 + b_1x_1 + ...)$
- $Odd_1 = Exp[b_0 + b_1(x_1 + 1) + ...] = Exp(b_0 + b_1x_1 + b_1 + ...) = Exp(b_0 + b_1x_1 + ...) \times Exp(b_1)$
- $Odd_1 = Odd_0 \times Exp(b_1)$
- $\frac{Odd_1}{Odd_0} = Exp(b_1)$ (勝率比)
- 係數的指數是 $y = 1$ 的 勝率 增加的倍數
- 係數的指數就是勝率比；也就是說，xi 每增加 1 單位，$y = 1$ 的勝率會變成原來的 $Exp(bi)$ 倍

主要是對表現校好的模型二及三解釋，

模型二顯著變數解釋(圖十五):

1. edu : -0.101840

每多讀一年書，得憂鬱症的勝率會變成原來的 exp(0.092912048)=0.903174 倍。

讀越多書，未來工作選擇多，生存可能比較容易，壓力下降。
2. hh_totalmembers : 0.092912048
家裡每多一位成員，得憂鬱症的勝率會變成原來的 exp(0.092912048)=1.0974 倍。
在肯亞有收入都不一定養得起自己了，多個人多張嘴多負擔，壓力上升。
3. ent_ownfarm : -0.383226020
主要收入來源來自自己的農場，得憂鬱症的勝率會變成原來的 exp(-0.383226020) = 0.6816588 倍。
有自己的農場推測有房地產，既有資產還能自給自足，壓力下降。
4. ent_farmexpenses : 0.037946947
每月農場支出每增加 1 美元，得憂鬱症的勝率會變成原來的 exp(0.037946947) = 1.038676 倍。
雖然此支出是為了讓農場賺錢，但花錢在成本上的感受可能還是會使人難過。

模型三顯著變數解釋(圖十六):
1. married：3.5884982710
有結婚的，得憂鬱症的勝率會變成原來的 exp(3.5884982710)=36.1797 倍。
一次增加許多家庭成員需要照顧，壓力很大，肯亞還是一夫多妻制。
(有 3 個交互項，影響有正有負)
2. edu：-0.1945225049
每多讀一年書，得憂鬱症的勝率會變成原來的 exp(-0.1945225049)=0.823228 倍。
讀越多書，未來工作選擇多，生存可能比較容易，壓力下降。
(有 3 個交互項，影響有正有負)
3. hh_totalmembers：0.3249753628
家裡每多一位成員，得憂鬱症的勝率會變成原來的 exp(0.3249754)=1.383997 倍。
在肯亞有收入都不一定養得起自己了，多個人多張嘴多負擔，壓力上升。
(有 3 個交互項，影響有正有負)
4. ent_ownfarm：-4.4851411297
主要收入來源來自自己的農場，得憂鬱症的勝率會變成原來的 exp(-4.48514113) = 0.0112753 倍。
有自己的農場推測有房地產，既有資產還能自給自足，壓力下降。
(有 4 個交互項，影響有正有負)
5. ent_farmexpenses：0.0601564682
每月農場支出每增加 1 美元，得憂鬱症的勝率會變成原來的 exp(0.0601564682) = 1.062003 倍。
雖然此支出是為了讓農場賺錢，但花錢在成本上的感受可能還是會使人難過。
(有 1 個交互項，影響負的，壓力會下降)
6. age:married：-0.0689624466
此人已婚且年紀每增加一歲，得憂鬱症的勝率會變成原來的 exp(-0.0689624466)

=0.9333617 倍。
生活慢慢步入正軌、熟悉、看開了，也許成為被後輩照顧的那個人了，下降。
同時須考慮 age 跟 married 每增加一單位的影響的影響。
7. age:cons_social：0.0041701012
年紀每增加一歲且生活開銷增加一美元，得憂鬱症的勝率會變成原來的 exp(0.0041701012) = 1.004179 倍。
年齡增加、支出上升，多少會焦慮，上升。
同時須考慮 age 跟 cons_social 每增加一單位的影響的影響。
8. age:ent_ownfarm：0.0520171433
年紀每增加一歲且主要收入來自自有農場，得憂鬱症的勝率會變成原來的 exp(0.0520171433) = 1.053394 倍。
年齡增加、可能體力不如從前，做的少賺得少，生活負擔大。
同時須考慮 age 跟 ent_ownfarm 每增加一單位的影響的影響。
9. married:hh_totalmembers：-0.3227938198
此人已婚且家庭成員每增加一位，得憂鬱症的勝率會變成原來的 exp(-0.3227938198) = 0.7241231 倍。
此時的成員增加應該為小孩(喜悅可能大於生活重擔)或新老婆(勞動力)，
在模型一雖然 children 這個變數不顯著，但在解釋上他是能讓憂鬱大幅下降的變數，加上需同時考慮原變數的影響，結果為下降。
同時須考慮 married 跟 hh_totalmembers 每增加一單位的影響的影響。
10. married:ent_ownfarm：2.8164453933
此人已婚且主要收入來自自有農場，得憂鬱症的勝率會變成原來的 exp(2.8164453933) = 16.71732 倍。
此數據集 30 歲婦女居多，猜測可能較會有財產上紛爭。
同時須考慮 ent_ownfarm 跟 married 每增加一單位的影響的影響。
11. edu:asset_savings：-0.0024896224
每多讀一年書且存款增加一美元，得憂鬱症的勝率會變成原來的 exp(-0.0024896) = 0.9975135 倍。
好上加好，降低，同時須考慮 edu 跟 asset_savings 每增加一單位的影響的影響
12. hh_totalmembers:cons_social：-0.0220332318
多一位成員且開銷增加一美元，得憂鬱症的勝率會變成原來的 exp(-0.02203323) = 0.9782077 倍。
同時須考慮 hh_totalmembers 跟 cons_social 每增加一單位的影響的影響。
13. asset_savings:cons_social：0.0018915498
開銷增加一美元且存款增加一美元，得憂鬱症的勝率會變成原來的 exp(0.0018915498) = 1.001893 倍。
同時須考慮 asset_savings 跟 cons_social 每增加一單位的影響的影響。
14. asset_savings:ent_ownfarm：0.0099589621

主要收入來自自有農場且存款增加一美元，得憂鬱症的勝率會變成原來的 exp(0.0099589621) = 1.010009 倍。
同時須考慮 asset_savings 跟 ent_ownfarm 每增加一單位的影響的影響。
15. asset_savings:ent_farmexpenses：-0.0008423672
每月農場支出每增加 1 美元且存款增加一美元，得憂鬱症的勝率會變成原來的 exp(-0.0008423672) = 0.999158 倍。
資金流動+存款，感覺正負相消，勝率比也很接近 1。
同時須考慮 asset_savings 跟 ent_farmexpenses 每增加一單位的影響的影響。

	crude OR(95%CI)	adj. OR(95%CI)
age (cont. var.)	1.0095 (1.0005,1.0185)	1.009 (0.9984,1.0198)
married: 1 vs 0	1.14 (0.85,1.53)	1.32 (0.94,1.85)
edu (cont. var.)	0.91 (0.87,0.95)	0.9 (0.86,0.94)
hh_totalmembers (cont. var.)	1.07 (1.01,1.13)	1.1 (1.03,1.17)
asset_savings (cont. var.)	0.998 (0.9953,1.0006)	0.9979 (0.9954,1.0005)
cons_social (cont. var.)	0.99 (0.97,1)	0.99 (0.97,1)
ent_ownfarm: 1 vs 0	0.86 (0.68,1.1)	0.68 (0.52,0.89)
ent_farmexpenses (cont. var.)	1.02 (0.99,1.05)	1.04 (1,1.07)

(圖十五)，勝率(OR)比

	OR	lower95ci	upper95ci	Pr(>|Z|)
age	1.0232821	0.997144906	1.05010431	8.126670e-02
married	36.1797030	6.736555499	194.30863581	2.862483e-05
edu	0.8232277	0.732999662	0.92456218	1.022647e-03
hh_totalmembers	1.3839965	1.109827162	1.72589617	3.913265e-03
asset_savings	1.0084346	0.992119696	1.02501779	3.128382e-01
cons_social	0.8970839	0.796974368	1.00976839	7.202726e-02
ent_ownfarm	0.0112753	0.002448862	0.05191485	8.567845e-09
ent_farmexpenses	1.0620027	1.021125791	1.10451597	2.665600e-03
age:married	0.9333617	0.905573600	0.96200256	7.748010e-06
age:cons_social	1.0041788	1.001989877	1.00637252	1.800959e-04
age:ent_ownfarm	1.0533938	1.028414903	1.07897940	2.154353e-05
married:hh_totalmembers	0.7241231	0.602815433	0.86984223	5.593939e-04
married:ent_ownfarm	16.7173214	6.200770599	45.07001692	2.607594e-08
edu:hh_totalmembers	1.0180063	0.996503428	1.03997326	1.013397e-01
edu:asset_savings	0.9975135	0.995295577	0.99973631	2.836672e-02
edu:cons_social	1.0066670	0.998700802	1.01469673	1.011610e-01
hh_totalmembers:cons_social	0.9782077	0.966242031	0.99032160	4.502568e-04
asset_savings:cons_social	1.0018933	1.001055060	1.00273232	9.462263e-06
asset_savings:ent_ownfarm	1.0100087	1.000625944	1.01947947	3.649489e-02
asset_savings:ent_farmexpenses	0.9991580	0.998478847	0.99983759	1.517576e-02
cons_social:ent_ownfarm	0.9598332	0.915460165	1.00635693	8.958946e-02

(圖十六)，勝率(OR)比

六、結論

交互項讓有些變數在解釋上變得不太直覺，依解釋性跟模型複雜度，模型二會是個好選擇，因為模型四有共線性問題，我猜測這可能是導致變數解釋不直覺甚至是相反的原因，也可能是背景知識了解不夠，不同地區的風俗民情、個性大相逕庭等因素所導致，在這筆資料中可以推測，錢，在不同地方，不一定是使人憂鬱的主要原因，在肯亞這個農村大部分人都普遍貧窮、也很少階級、種族上的不公，貧富差距、歧視造成的壓力可能偏少，這些都有可能是造成最初模型超出我預期的因素。

Reference

- https://zindi.africa/competitions/busara-mental-health-prediction-challenge/data
- https://depressytrouble.tw/index.php/portfolio/rule_the_world/
- https://www.storm.mg/article/4384549
- https://www.rdocumentation.org/packages/VIM/versions/6.2.2/topics/aggr
- https://www.rdocumentation.org/packages/ROSE/versions/0.0-4/topics/ovun.sample
- Stepwise regression in R with both direction - Cross Validated (stackexchange.com)
- https://www.zhihu.com/question/403333784