

Modelling for Dataset

Lim Jun Yi Keith

2025-11-03

```
# load libraries
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ----- tidyverse 2.0.0 --
## v dplyr      1.1.4      v readr      2.1.5
## v forcats    1.0.0      v stringr   1.5.1
## v ggplot2    3.5.2      v tibble    3.3.0
## v lubridate  1.9.4      v tidyr     1.3.1
## v purrr      1.1.0
## -- Conflicts ----- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(broom) #sweeps messy data into neat tables
library(ggplot2)
```

```
#use logistic regression as dependent variable is binary (depression)
#independent variables will be academic pressure,CGPA, study satisfaction,
#work/study hours -->
#Reasons for these IVs: which is the most important education-related
#factor that predicts depression? Knowing the most important factor
#can help inform solutions and mental health approaches.

#Financial stress also added as an IV.
```

Import data

```
#import data
df <- read_csv("student_depression_dataset.csv")
```

```
## Warning: One or more parsing issues, call 'problems()' on your data frame for details,
## e.g.:
##   dat <- vroom(...)
##   problems(dat)

## Rows: 27901 Columns: 18
## -- Column specification -----
```

```
## Delimiter: ","
## chr (8): Gender, City, Profession, Sleep Duration, Dietary Habits, Degree, ...
## dbl (10): id, Age, Academic Pressure, Work Pressure, CGPA, Study Satisfactio...
##
## i Use 'spec()' to retrieve the full column specification for this data.
## i Specify the column types or set 'show_col_types = FALSE' to quiet this message.
```

```
#new dataset, df1, to focus on variables being used
df1 <- df %>%
  rename("AP" = "Academic Pressure",
         "SS" = "Study Satisfaction",
         "W/S Hours" = "Work/Study Hours",
         "FS" = "Financial Stress") %>%
  select("AP", "SS", "CGPA", "W/S Hours", "FS", "Depression")
view(df1)
```

Model (logistics regression)

```
##
## Call:
## glm(formula = Depression ~ AP + SS + CGPA + 'W/S Hours' + FS,
##      family = binomial, data = df1)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -4.478648   0.110250 -40.623  < 2e-16 ***
## AP           0.828785   0.012401  66.832  < 2e-16 ***
## SS          -0.220092   0.011285 -19.504  < 2e-16 ***
## CGPA         0.051573   0.010324   4.995 5.87e-07 ***
## 'W/S Hours'  0.117091   0.004136  28.311  < 2e-16 ***
## FS           0.573264   0.011147  51.425  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##    Null deviance: 37855  on 27897  degrees of freedom
## Residual deviance: 26640  on 27892  degrees of freedom
## (3 observations deleted due to missingness)
## AIC: 26652
##
## Number of Fisher Scoring iterations: 5
```



```
## (Intercept)          AP          SS          CGPA 'W/S Hours'          FS
## 0.01134874  2.29053393  0.80244469  1.05292595  1.12422218  1.77404723
```

