

fp_step3

Zeyu Hou

2025-06-04

Question 2

2. Does family mental health history predict depression likelihood even when controlling for academic performance?

```
data <- read_csv("C:/Users/zeyul/Desktop/Files/text book/pstat 100/student_depression_dataset.csv")
```

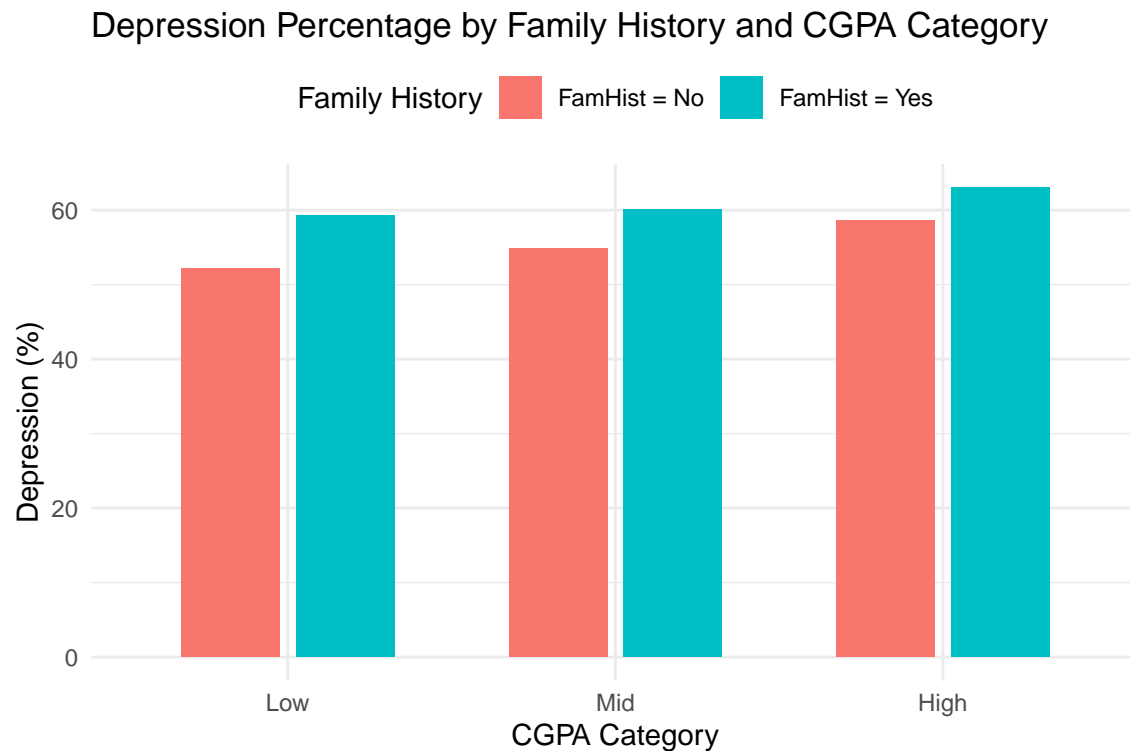
```
## Rows: 27901 Columns: 18
## -- Column specification -----
## Delimiter: ","
## chr (8): Gender, City, Profession, Sleep Duration, Dietary Habits, Degree, ...
## dbl (10): id, Age, Academic Pressure, Work Pressure, CGPA, Study Satisfactio...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
# Create summary_df
summary_df <- data %>%
  mutate(
    CGPA_Category = factor(case_when(
      CGPA < 6.0 ~ "Low",
      CGPA >= 6.0 & CGPA < 8.0 ~ "Mid",
      CGPA >= 8.0 ~ "High"
    ), levels = c("Low", "Mid", "High")),
    FamHist = ifelse(`Family History of Mental Illness` == "Yes", "FamHist = Yes", "FamHist = No"),
    Dep = ifelse(Depression == 1, 1, 0)
  ) %>%
  group_by(CGPA_Category, FamHist) %>%
  summarize(
    n = n(),
    depression_percent = mean(Dep)*100,
    .groups = "drop"
  )
summary_df
```

```
## # A tibble: 6 x 4
##   CGPA_Category FamHist      n depression_percent
##   <fct>         <chr>    <int>          <dbl>
## 1 Low          FamHist = No  2801          52.1
## 2 Low          FamHist = Yes  2611          59.2
```

```
## 3 Mid          FamHist=No    5063          54.8
## 4 Mid          FamHist=Yes   4791          60.1
## 5 High         FamHist=No    6534          58.6
## 6 High         FamHist=Yes   6101          63.1
```

```
# Plot
ggplot(summary_df, aes(x = CGPA_Category, y = depression_percent, fill = FamHist)) +
  geom_bar(stat = "identity", position = position_dodge(width = 0.7), width = 0.6) +
  labs(
    x = "CGPA Category",
    y = "Depression (%)",
    fill = "Family History",
    title = "Depression Percentage by Family History and CGPA Category"
  ) +
  theme_minimal() +
  theme(
    legend.position = "top",
  )
```



```
# Create df2 for Anova Analysis
df2 <- data %>%
  mutate(
    Depression = factor(Depression, levels = c(0,1)),
    FamHist = factor(ifelse(`Family History of Mental Illness` == "Yes", "Yes", "No")),
    CGPA_Category = factor(case_when(
      CGPA < 6.0 ~ "Low",
      CGPA >= 6.0 & CGPA < 8.0 ~ "Mid",
      CGPA >= 8.0 ~ "High"
    ))
```

```

    ), levels = c("Low", "Mid", "High"))
  ) %>%
  select(Depression, FamHist, CGPA_Category)

# Fit model
fit_null <- glm(
  Depression ~ CGPA_Category,
  data = df2,
  family = binomial
)
# Full model adding FamHist
fit_full <- glm(
  Depression ~ CGPA_Category + FamHist,
  data = df2,
  family = binomial
)

# Compare with ANOVA (likelihood-ratio test; "Chisq" tests FamHist's added value)
anova(fit_null, fit_full, test = "Chisq")

```

```

## Analysis of Deviance Table
##
## Model 1: Depression ~ CGPA_Category
## Model 2: Depression ~ CGPA_Category + FamHist
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1      27898      37809
## 2      27897      37729  1    80.008 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

Question 3

3. Are academic pressure thresholds associated with increased suicidal ideation? Could we use academic pressure to predict suicidal ideation?

```

# Fit linear vs. quadratic
data <- data %>%
  mutate(
    Suiside = ifelse(`Have you ever had suicidal thoughts ?` == "Yes", 1, 0),
    AcademicPressure = `Academic Pressure`,
  )
fit_lin <- glm(
  Suiside ~ AcademicPressure + Depression,
  data = data, family = binomial
)
data$AP <- data$AcademicPressure^2
fit_quad <- glm(
  Suiside ~ AcademicPressure + AP + Depression,
  data = data, family = binomial
)

# AIC comparison
print("AIC linear:")

```

```
## [1] "AIC linear:"
```

```
AIC(fit_lin)
```

```
## [1] 28063.13
```

```
print("AIC quad:")
```

```
## [1] "AIC quad:"
```

```
AIC(fit_quad)
```

```
## [1] 28048.83
```

```
# Create curve data for Suicide Thought Question with Answer No
```

```
newAP <- seq(1,10, by=0.1)
```

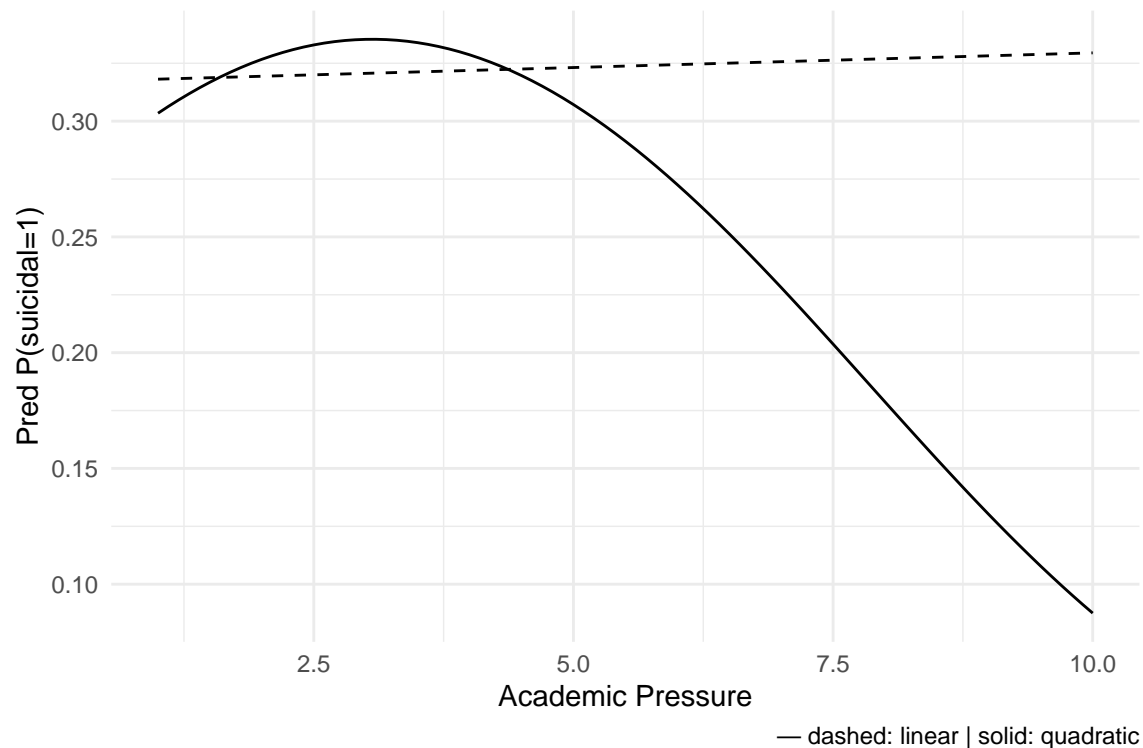
```
Suicide_No <- data.frame(  
  AcademicPressure = newAP,  
  AP = newAP^2,  
  Depression = 0  
)
```

```
Suicide_No$pred_lin <- predict(fit_lin, newdata = Suicide_No, type = "response")
```

```
Suicide_No$pred_quad <- predict(fit_quad, newdata = Suicide_No, type = "response")
```

```
# Plot for Suicide Thought Question with Answer No
```

```
ggplot(Suicide_No, aes(x=AcademicPressure)) +  
  geom_line(aes(y=pred_lin), linetype="dashed") +  
  geom_line(aes(y=pred_quad)) +  
  labs(x="Academic Pressure", y="Pred P(suicidal=1)",  
       caption="- dashed: linear | solid: quadratic") +  
  theme_minimal()
```



```
# Create curve data for Suicide Thought Question with Answer Yes
Suicide_Yes <- data.frame(
  AcademicPressure = newAP,
  AP = newAP^2,
  Depression = 1
)

Suicide_Yes$pred_lin <- predict(fit_lin, newdata = Suicide_Yes, type = "response")
Suicide_Yes$pred_quad <- predict(fit_quad, newdata = Suicide_Yes, type = "response")

# Plot for Suicide Thought Question with Answer Yes
ggplot(Suicide_Yes, aes(x=AcademicPressure)) +
  geom_line(aes(y=pred_lin), linetype="dashed") +
  geom_line(aes(y=pred_quad)) +
  labs(x="Academic Pressure", y="Pred P(suicidal=1)",
       caption="- dashed: linear | solid: quadratic") +
  theme_minimal()
```

