Econometric Exam - Midterm 2
Professor: Kevin Foster

Student: Aqib Shahzad
Date: 11/20/2025

Answer (1):

Step 1: Understanding the Two Approaches

We have survey data where:

- Y = he_more_than_5yrs_than_her (whether the husband is more than 5 years older than the wife)
- Age = age of the household head
- We split by race: White (W) and Non-White (NW)

1.1: Separate Regressions

We run two completely separate regressions:

For White households:

$$Y_w = \beta_0 + \beta_1 Age_w + \epsilon_w$$

For Non-White households:

$$Y_{nw} = \alpha_0 + \alpha_1 Age_{nw} + \nu_{nw}$$

1.2: Single Regression with Interactions

We create a dummy variable:

- D = 1 if Non-White
- D = 0 if White

Then run one regression:

$$Y = \gamma 0 + \gamma 1 Age + \gamma 2 D + \gamma 3 (D \times Age) + \eta$$

Step 2: Writing Out What Each Model Actually Estimates

Let's see what each model predicts for different groups.

2.1: For Approach 2 (pooled model):

When D = 0 (White):

$$Y = \gamma 0 + \gamma 1 Age + \gamma 2(0) + \gamma 3(0) = \gamma 0 + \gamma 1 Age$$

When D = 1 (Non-White):

$$Y = \gamma 0 + \gamma 1 Age + \gamma 2(1) + \gamma 3(1 \times Age) = (\gamma 0 + \gamma 2) + (\gamma 1 + \gamma 3) Age$$

2.2: For Approach 1 (separate models):

White model: $Y = \beta 0 + \beta 1 Age$

Non-White model: $Y = \alpha 0 + \alpha 1 Age$

Step 3: Matching Up the Coefficients

By comparing the prediction equations

3.1: For White Individuals

Approach 1 says: $Y=\beta 0+\beta 1Age$

Approach 2 says: $Y=(\gamma 0+\gamma 2)+(\gamma 1+\gamma 3)Age$

Therefore: $\beta_0=\gamma_0$ *and* $\beta_1=\gamma_1$

3.2: For Non-White Individuals

Approach 1 says: $Y=\alpha 0+\alpha 1Age$

Approach 2 says: $Y=(\gamma 0+\gamma 2)+(\gamma 1+\gamma 3)Age$

Therefore: $\alpha 0=\gamma 0+\gamma 2$ *and* $\alpha 1=\gamma 1+\gamma 3$

Step 4: The Complete Relationship

We can express this in both directions

4.1: From Pooled Model (γ) to Separate Models (β, α):

$\beta 0=\gamma 0$

$\beta 1=\gamma 1$

$\alpha 0=\gamma 0+\gamma 2$

$\alpha 1=\gamma 1+\gamma 3$

4.2: From Separate Models (β, α) to Pooled Model (γ):

$\gamma 0=\beta 0$

$\gamma 1=\beta 1$

$\gamma 2=\alpha 0-\beta 0$

$\gamma 3=\alpha 1-\beta 1$

Step 5: Economic Interpretation

- γ_0 = Intercept for White group (same as β_0)
- γ_1 = Slope of Age for White group (same as β_1)
- γ_2 = Difference in intercepts: (Non-White intercept) - (White intercept)
- γ_3 = Difference in slopes: (Non-White Age effect) - (White Age effect)

Final Answer

The exact relationships between the coefficients are:

$\beta 0 = \gamma 0$

$\beta 1 = \gamma 1$

$\alpha 0 = \gamma 0 + \gamma 2$

$\alpha 1 = \gamma 1 + \gamma 3$

This means the single pooled regression with interactions gives us exactly the same fitted regression lines for each racial group as we would get by running separate regressions. The advantage of the pooled approach is that we can directly test whether the racial differences (γ_2 and γ_3) are statistically significant using standard t-tests.

Answer (2):

"I used the (d_HHP2020_24.Rdata) file shared by the Professor in Slack (General).

For Education Joint Test:

- F-statistic: 271.64
- p-value: < 2.2e-16 (essentially 0)
- Conclusion: We strongly reject the null hypothesis that all education coefficients are zero. Education is statistically significant in predicting mental health.

For Income Joint Test:

- F-statistic: 18,224
- p-value: < 2.2e-16 (essentially 0)
- Conclusion: We strongly reject the null hypothesis that the income coefficient is zero. Income is statistically significant in predicting mental health.

Relative Importance: Both are extremely significant, but income has a MUCH larger F-statistic (18,224 vs 272), suggesting income is the relatively more important predictor of mental health in this model.

The Codes and the Results for Question 2

```
file_path <- file.choose()

> load(file_path)

> ls()
[1] "d_HHP2020_24" "file_path"

> head(d_HHP2020_24)
  Age Gender    Education Mar_Stat income_midpoint  Race      Hispanic Number_people_HH
Number_kids_HH Number_adults_HH
1  34 female college grad  Married           62500 white not Hispanic                 4
2                2
2  65   male some college divorced           30000 white not Hispanic                 1
0                1
3  44 female college grad  Married          225000 other not Hispanic                 2
0                2
4  56   male some college divorced           12500 white not Hispanic                 2
0                2
5  57 female   adv degree    never           62500 white not Hispanic                 1
0                1
6  44 female   adv degree  Married          125000 white not Hispanic                 2
0                2
```

```
  private_health_ins public_health_ins                          work_kind
workloss DOWN ANXIOUS

1                  0                 0           employed by private co
no     1        4

2                  0                 0                             <NA>
no     4        3

3                  0                 0 employed by nonprofit or charity
no     1        1

4                  0                 0                             <NA> yes recent household loss
of work     4        4

5                  0                 0 employed by nonprofit or charity
no     2        2

6                  0                 0           employed by private co
no     2        3

  WORRY INTEREST YEAR Begin_Date K4SUM income_midpoint_factor

1     3        1   20 2020-04-23     9                  62500

2     4        4   20 2020-04-23    15                  30000

3     1        1   20 2020-04-23     4                 225000

4     4        4   20 2020-04-23    16                  12500

5     1        2   20 2020-04-23     7                  62500

6     2        2   20 2020-04-23     9                 125000

> table(d_HHP2020_24$Education)


       lt hs      some hs  high school some college    assoc deg college grad   adv degree

        6787        14934       122541       210698       103575       279400       246855

> ols_model1 <- lm(K4SUM ~ Education + income_midpoint + Age + Gender + Mar_Stat,

+                  data = d_HHP2020_24)

>

> ols_model1 <- lm(K4SUM ~ Education + income_midpoint + Age + Gender + Mar_Stat,

+                  data = d_HHP2020_24)

> summary(ols_model1)


Call:

lm(formula = K4SUM ~ Education + income_midpoint + Age + Gender +

    Mar_Stat, data = d_HHP2020_24)


Residuals:

    Min      1Q  Median      3Q     Max

-7.8262 -2.2801 -0.8621  1.5434 12.6157


Coefficients:

                           Estimate Std. Error  t value Pr(>|t|)

(Intercept)              1.039e+01  5.055e-02  205.473  < 2e-16 ***

Educationsome hs        -2.728e-01  5.755e-02   -4.741 2.13e-06 ***

Educationhigh school    -3.929e-01  4.904e-02   -8.011 1.14e-15 ***
```

```
Educationsome college -1.315e-01  4.856e-02   -2.709  0.00675 **
Educationassoc deg    -2.999e-01  4.920e-02   -6.096 1.09e-09 ***
Educationcollege grad -5.142e-01  4.851e-02  -10.599  < 2e-16 ***
Educationadv degree   -4.949e-01  4.870e-02  -10.163  < 2e-16 ***
income_midpoint       -8.758e-06  6.487e-08 -134.994  < 2e-16 ***
Age                   -5.133e-02  2.572e-04 -199.557  < 2e-16 ***
Genderfemale           4.974e-01  7.425e-03   66.992  < 2e-16 ***
Gendertrans            2.448e+00  7.923e-02   30.899  < 2e-16 ***
Genderother            1.703e+00  4.847e-02   35.139  < 2e-16 ***
Mar_Statwidowed        1.916e-01  1.685e-02   11.373  < 2e-16 ***
Mar_Statdivorced       5.247e-01  1.069e-02   49.075  < 2e-16 ***
Mar_Statseparated      1.154e+00  2.778e-02   41.527  < 2e-16 ***
Mar_Statnever          3.110e-01  1.040e-02   29.899  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.176 on 789784 degrees of freedom
  (194990 observations deleted due to missingness)
Multiple R-squared:  0.1178,  Adjusted R-squared:  0.1178
F-statistic:  7030 on 15 and 789784 DF,  p-value: < 2.2e-16

> linearHypothesis(ols_model1,
+                  c("Educationsome hs = 0",
+                    "Educationhigh school = 0",
+                    "Educationsome college = 0",
+                    "Educationassoc deg = 0",
+                    "Educationcollege grad = 0",
+                    "Educationadv degree = 0"))

Linear hypothesis test:

Educationsome hs = 0

Educationhigh school = 0

Educationsome college = 0

Educationassoc deg = 0

Educationcollege grad = 0

Educationadv degree = 0

Model 1: restricted model
Model 2: K4SUM ~ Education + income_midpoint + Age + Gender + Mar_Stat

  Res.Df     RSS Df Sum of Sq      F    Pr(>F)
1 789790 7982829
```

```
2 789784 7966389  6     16440 271.64 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> linearHypothesis(ols_model1, "income_midpoint = 0")

Linear hypothesis test:
income_midpoint = 0

Model 1: restricted model
Model 2: K4SUM ~ Education + income_midpoint + Age + Gender + Mar_Stat

  Res.Df     RSS Df Sum of Sq     F    Pr(>F)
1 789785 8150206
2 789784 7966389  1    183817 18224 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Answer 3:

Data Subset Analysis

1. Subset Selection and Rationale

Selected Subset: Working-age adults (25-65 years) with college degrees
Rationale: This aligns with the research focus mentioned in the background - examining how college education relates to mental health outcomes. By focusing on the college-educated working-age population, we can isolate the mental health patterns of a key demographic group that has invested in higher education, while controlling for life stage by restricting to prime working years (25-65).

2. Summary Statistics

The subset contains 351,473 working-age adults with college degrees after removing missing values.

Key Characteristics:

- Mental Health Prevalence: 23.03% of college-educated adults report significant mental health concerns (K4SUM > 8)
- Average Mental Health Score: 6.89 (on scale of 4-16)
- Average Age: 45.76 years
- Average Income: $125,228
- Gender Distribution: 57.68% female
- Education Breakdown:
    - College graduates: 189,883 (54.0%)
    - Advanced degree holders: 161,590 (46.0%)

Interpretation:

- `MentalHealth_01 = 1`: Individual reports K4SUM > 8, indicating more frequent mental health symptoms (equivalent to reporting symptoms "more than half the days" or "nearly every day" on average across the four measures)
- `MentalHealth_01 = 0`: Individual reports K4SUM ≤ 8, indicating less frequent mental health symptoms

This threshold (K4SUM > 8) represents a clinically meaningful cutoff where individuals experience mental health symptoms more persistently throughout the month.

Distribution:

- 0 (Better mental health): 270,528 cases (77.0%)
- 1 (Worse mental health): 80,945 cases (23.0%)

Codes and Results in R for Question 3

```
subset_data <- d_HHP2020_24 %>%
+    filter(Education %in% c("college grad", "adv degree"),
+           Age >= 25 & Age <= 65)
> nrow(subset_data)
[1] 388787
> subset_data$MentalHealth_01 <- ifelse(subset_data$K4SUM > 8, 1, 0)
> head(subset_data$MentalHealth_01)
[1] 1 0 0 1 0 0
> table(subset_data$MentalHealth_01)

     0      1
270528  80945
> summary_stats <- subset_data %>%
+    summarise(
+        N = n(),
+        MentalHealth_Rate = mean(MentalHealth_01),
+        Avg_K4SUM = mean(K4SUM, na.rm = TRUE),
+        Avg_Age = mean(Age, na.rm = TRUE),
+        Avg_Income = mean(income_midpoint, na.rm = TRUE),
+        Female_Pct = mean(Gender == "female", na.rm = TRUE)
+    )
> print(summary_stats)
       N MentalHealth_Rate Avg_K4SUM  Avg_Age Avg_Income Female_Pct
1 388787               NA  6.885112 45.65307   125206.4  0.5767631
> # Remove any missing values first
> subset_data_clean <- subset_data %>% filter(!is.na(K4SUM))
>
> summary_stats <- subset_data_clean %>%
+    summarise(
```

```
+         N = n(),
+         MentalHealth_Rate = mean(MentalHealth_01, na.rm = TRUE),
+         Avg_K4SUM = mean(K4SUM, na.rm = TRUE),
+         Avg_Age = mean(Age, na.rm = TRUE),
+         Avg_Income = mean(income_midpoint, na.rm = TRUE),
+         Female_Pct = mean(Gender == "female", na.rm = TRUE)
+     )
> print(summary_stats)
       N MentalHealth_Rate Avg_K4SUM  Avg_Age Avg_Income Female_Pct
1 351473         0.2303022  6.885112 45.76104   125227.6  0.5767982
> table(subset_data_clean$Education)

       lt hs      some hs  high school some college    assoc deg college grad   adv degree
          0            0            0            0            0       189883       161590
```

Answer (4):

Linear Probability Model Analysis

a) Predictor Variables and Exogeneity

Variables Chosen:

- Age: Continuous variable (25-65 years)
- Gender: Categorical (male, female, trans, other)
- income_midpoint: Continuous household income
- Education: Binary (college grad vs adv degree)
- Interaction Terms:
    - Age:Gender: Tests if age effects differ by gender
    - income_midpoint:Education: Tests if income effects differ by education level

Exogeneity Assessment:
The predictors are likely not fully exogenous. While age and gender are predetermined, income and education may suffer from:

- Simultaneity: Mental health affects earning capacity and educational attainment
- Omitted variable bias: Unobserved factors (family background, personality traits) likely affect both education/income and mental health
- Measurement error: Self-reported income and mental health may contain errors

b) Coefficient Plausibility and Significance

Plausibility:

- Age (-0.0032): Negative effect plausible - mental health issues may decrease with maturity
- Gender female (+0.085): Plausible - women often report higher mental health burdens
- Income (-1.111e-06): Plausible - higher income reduces financial stress
- Adv degree (-0.0169): Plausible - advanced education may provide coping resources

Statistical Significance:
All main effects are highly significant ($p < 2.2e-16$) except:

- Age:Gendertrans ($p = 0.570$) - not significant
- Most interaction effects are significant but small in magnitude

c) Joint Test of Education Coefficients

Null Hypothesis: All education-related coefficients = 0

$H0 : \beta$Educationadv degree $= \beta$income_midpoint:Educationadv degree $= 0$

Test Results:

- F-statistic: 16.807
- p-value: 5.027e-08

Conclusion: Strongly reject H_0 - education variables jointly have statistically significant effects on mental health.

d) Predicted Probabilities

Age	Gender	Income	Education	Predicted Probability
35	Female	$75,000	College Grad	35.3%
35	Female	$150,000	Adv Degree	26.5%
50	Male	$75,000	College Grad	24.6%
50	Male	$150,000	Adv Degree	15.8%

Patterns: Higher income and advanced degrees reduce predicted mental health problems; older individuals and males have lower probabilities.

e) Type I and Type II Errors

Using 0.5 classification threshold:

- Type I Errors (False Positives): 528 cases
- Type II Errors (False Negatives): 75,027 cases
- Total Misclassifications: 75,555 cases
- Type I Error Rate: 0.21%
- Type II Error Rate: 99.04%

Critical Issue: The model suffers from extremely high Type II error because it rarely predicts mental health problems (only when probability > 50%), while the actual prevalence is only 23%. This suggests the 0.5 threshold is inappropriate for this imbalanced classification problem. A lower threshold would better balance error rates.

Code & Result for Question 4 in R

```
# Create the linear probability model

> ols_model_01 <- lm(MentalHealth_01 ~ Age + Gender + income_midpoint + Education +

+                        Age:Gender + income_midpoint:Education,

+                    data = subset_data_clean)

>

> summary(ols_model_01)



Call:

lm(formula = MentalHealth_01 ~ Age + Gender + income_midpoint +
    Education + Age:Gender + income_midpoint:Education, data = subset_data_clean)



Residuals:

     Min       1Q   Median       3Q      Max

-0.67545 -0.26677 -0.18016 -0.05344  0.96927



Coefficients:

                                     Estimate Std. Error t value Pr(>|t|)

(Intercept)                          4.915e-01  5.085e-03  96.660  < 2e-16 ***

Age                                 -3.245e-03  1.006e-04 -32.267  < 2e-16 ***

Genderfemale                         8.506e-02  6.240e-03  13.632  < 2e-16 ***

Gendertrans                          3.001e-01  6.512e-02   4.608 4.06e-06 ***

Genderother                          3.706e-01  3.703e-02  10.006  < 2e-16 ***

income_midpoint                     -1.111e-06  1.507e-08 -73.674  < 2e-16 ***

Educationadv degree                 -1.686e-02  3.130e-03  -5.386 7.20e-08 ***

Age:Genderfemale                    -7.568e-04  1.321e-04  -5.729 1.01e-08 ***

Age:Gendertrans                     -9.853e-04  1.737e-03  -0.567  0.57045

Age:Genderother                     -3.666e-03  8.576e-04  -4.274 1.92e-05 ***

income_midpoint:Educationadv degree  8.164e-08  2.179e-08   3.746  0.00018 ***

---
```

```
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4116 on 325909 degrees of freedom
  (25553 observations deleted due to missingness)
Multiple R-squared:  0.05047, Adjusted R-squared:  0.05044
F-statistic:  1732 on 10 and 325909 DF,  p-value: < 2.2e-16

> # Check the range of fitted values (probabilities)
> fitted_values <- fitted(ols_model_01)
> summary(fitted_values)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.03073 0.16110 0.23504 0.23244 0.30078 0.67545
> # Test education main effect and interactions
> linearHypothesis(ols_model_01,
+                  c("Educationadv degree = 0",
+                    "income_midpoint:Educationadv degree = 0"))

Linear hypothesis test:
Educationadv degree = 0
income_midpoint:Educationadv degree = 0

Model 1: restricted model
Model 2: MentalHealth_01 ~ Age + Gender + income_midpoint + Education +
    Age:Gender + income_midpoint:Education

  Res.Df   RSS Df Sum of Sq      F    Pr(>F)
1 325911 55220
2 325909 55214  2    5.6946 16.807 5.027e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
> # Create example cases for prediction
> example_cases <- data.frame(
+    Age = c(35, 35, 50, 50),
+    Gender = c("female", "female", "male", "male"),
+    income_midpoint = c(75000, 150000, 75000, 150000),
+    Education = c("college grad", "adv degree", "college grad", "adv degree")
+ )
>
> # Get predictions
> predictions <- predict(ols_model_01, newdata = example_cases)
> example_cases$Predicted_Probability <- predictions
>
> print(example_cases)
  Age Gender income_midpoint    Education Predicted_Probability
1  35 female           75000 college grad             0.3532482
2  35 female          150000   adv degree             0.2653384
3  50   male           75000 college grad             0.2459962
4  50   male          150000   adv degree             0.1580863
> # Create predictions and classifications
> threshold <- 0.5  # Standard threshold for binary classification
> predicted_class <- ifelse(fitted_values > threshold, 1, 0)
> actual_class <- subset_data_clean$MentalHealth_01
>
> # Create confusion matrix
> confusion_matrix <- table(Actual = actual_class, Predicted = predicted_class)
Error in table(Actual = actual_class, Predicted = predicted_class) :
  all arguments must have the same length
> # Get the actual data used in the model (after removing NAs)
```

```
> model_data <- model.frame(ols_model_01)

> actual_class_model <- model_data$MentalHealth_01

> fitted_values_model <- fitted(ols_model_01)

>

> # Create classifications using 0.5 threshold

> predicted_class <- ifelse(fitted_values_model > 0.5, 1, 0)

>

> # Create confusion matrix

> confusion_matrix <- table(Actual = actual_class_model, Predicted = predicted_class)

> print(confusion_matrix)

      Predicted

Actual      0      1

     0 249634    528

     1  75027    731

>

> # Calculate error rates

> if(nrow(confusion_matrix) == 2 && ncol(confusion_matrix) == 2) {

+     type_I_error <- confusion_matrix[1, 2] / sum(confusion_matrix[1, ])  # False Positive Rate

+     type_II_error <- confusion_matrix[2, 1] / sum(confusion_matrix[2, ])  # False Negative Rate

+     total_errors <- confusion_matrix[1, 2] + confusion_matrix[2, 1]

+

+     cat("\nType I Errors (False Positives):", confusion_matrix[1, 2])

+     cat("\nType II Errors (False Negatives):", confusion_matrix[2, 1])

+     cat("\nTotal Misclassifications:", total_errors)

+     cat("\nType I Error Rate (False Positive Rate):", round(type_I_error, 4))

+     cat("\nType II Error Rate (False Negative Rate):", round(type_II_error, 4))

+ }



Type I Errors (False Positives): 528

Type II Errors (False Negatives): 75027
```

```
Total Misclassifications: 75555


Type I Error Rate (False Positive Rate): 0.0021

Type II Error Rate (False Negative Rate): 0.9904
```

Answer (5):

Logit Model Analysis

a) Predictor Variables and Exogeneity

Variables Chosen: (Same as OLS for comparability)

- Age, Gender, income_midpoint, Education
- Interaction Terms: **Age:Gender, income_midpoint:Education**

Exogeneity Assessment:
Same concerns as OLS - predictors are likely not fully exogenous due to:

- Simultaneity: Mental health affects income and educational attainment
- Omitted variable bias: Unobserved confounders affect both predictors and outcomes
- Self-report bias: In both mental health and income measures

b) Coefficient Plausibility and Significance

Plausibility: (All coefficients in log-odds)

- Age (-0.0216): Plausible - each year reduces log-odds of mental health issues
- Gender female (+0.285): Plausible - women have higher odds of mental health concerns
- Income (-6.501e-06): Plausible - higher income protective against mental health issues
- Adv degree (-0.0459): Plausible - advanced education provides protective resources

Statistical Significance:
All main effects highly significant ($p < 0.01$) except interactions:

- **Age:Genderfemale (p = 0.475) - not significant**
- **Age:Gendertrans (p = 0.476) - not significant**
- **Age:Genderother (p = 0.157) - not significant**
- **income_midpoint:Educationadv degree (p = 0.546) - not significant**

c) Joint Test of Education Coefficients

Null Hypothesis: All education-related coefficients = 0

Test Results:

- Chi-squared statistic: 18.093
- p-value: 0.0001178

Conclusion: Strongly reject H_0 - education variables jointly significant in predicting mental health.

d) Predicted Probabilities

Group	Age	Gender	Income	Education	Logit Probability
35F_75K_CG	**35**	**Female**	**$75,000**	**College Grad**	**35.8%**

35F_150K_AD	**35**	**Female**	**$150,000**	**Adv Degree**	**24.9%**
50M_75K_CG	**50**	**Male**	**$75,000**	**College Grad**	**22.9%**
50M_150K_AD	**50**	**Male**	**$150,000**	**Adv Degree**	**15.0%**

Patterns: Consistent with OLS - higher income, advanced degrees, older age, and male gender reduce mental health risk.

e) Type I and Type II Errors

Using 0.5 classification threshold:

- **Type I Errors (False Positives): 696 cases**
- **Type II Errors (False Negatives): 74,856 cases**
- **Total Misclassifications: 75,552 cases**
- **Type I Error Rate: 0.28%**
- **Type II Error Rate: 98.81%**

Critical Issue: Same as OLS - extremely high Type II error due to imbalanced classification and inappropriate 0.5 threshold.

f) Logit vs OLS Comparison

Similarities:

- Both models show same directional effects and significance patterns
- Predicted probabilities very similar (differences < 2 percentage points)
- Same high Type II error rates with 0.5 threshold

Codes and Results for Question 5 in R:

```
> # Create the logit model with same predictors as OLS
> logit_model <- glm(MentalHealth_01 ~ Age + Gender + income_midpoint + Education +
+                      Age:Gender + income_midpoint:Education,
+                    data = subset_data_clean,
+                    family = binomial(link = "logit"))
>
> summary(logit_model)


Call:
glm(formula = MentalHealth_01 ~ Age + Gender + income_midpoint +
    Education + Age:Gender + income_midpoint:Education, family = binomial(link = "logit"),
```

```
    data = subset_data_clean)



Coefficients:

                                    Estimate Std. Error z value Pr(>|z|)

(Intercept)                         3.551e-01  3.044e-02  11.667  < 2e-16 ***

Age                                -2.160e-02  6.302e-04 -34.273  < 2e-16 ***

Genderfemale                        2.846e-01  3.633e-02   7.835 4.70e-15 ***

Gendertrans                         9.240e-01  3.265e-01   2.830  0.00466 **

Genderother                         1.288e+00  1.892e-01   6.807 9.94e-12 ***

income_midpoint                    -6.501e-06  9.365e-08 -69.422  < 2e-16 ***

Educationadv degree                -4.586e-02  1.746e-02  -2.627  0.00861 **

Age:Genderfemale                    5.675e-04  7.939e-04   0.715  0.47476

Age:Gendertrans                     6.195e-03  8.690e-03   0.713  0.47594

Age:Genderother                    -6.258e-03  4.423e-03  -1.415  0.15708

income_midpoint:Educationadv degree  8.228e-08  1.363e-07   0.604  0.54604

---

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1



(Dispersion parameter for binomial family taken to be 1)



    Null deviance: 353436  on 325919  degrees of freedom

Residual deviance: 336623  on 325909  degrees of freedom

  (25553 observations deleted due to missingness)

AIC: 336645



Number of Fisher Scoring iterations: 4



> # Test education main effect and interactions

> linearHypothesis(logit_model,

+                  c("Educationadv degree = 0",
```

```
+                     "income_midpoint:Educationadv degree = 0"))



Linear hypothesis test:

Educationadv degree = 0

income_midpoint:Educationadv degree = 0



Model 1: restricted model

Model 2: MentalHealth_01 ~ Age + Gender + income_midpoint + Education +

    Age:Gender + income_midpoint:Education

  Res.Df Df  Chisq Pr(>Chisq)

1 325911

2 325909  2 18.093  0.0001178 ***

---

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> # Use the same example cases as before

> example_cases <- data.frame(

+    Age = c(35, 35, 50, 50),

+    Gender = c("female", "female", "male", "male"),

+    income_midpoint = c(75000, 150000, 75000, 150000),

+    Education = c("college grad", "adv degree", "college grad", "adv degree")

+ )

>

> # Get predicted probabilities from logit model

> logit_predictions <- predict(logit_model, newdata = example_cases, type = "response")

> example_cases$Logit_Probability <- logit_predictions

>

> print(example_cases)

  Age Gender income_midpoint    Education Logit_Probability

1  35 female           75000 college grad         0.3580086
```

```
2  35 female          150000   adv degree         0.2487780

3  50   male           75000 college grad         0.2292574

4  50   male          150000   adv degree         0.1501237

> # Get fitted probabilities from logit model

> logit_fitted <- predict(logit_model, type = "response")

>

> # Create classifications using 0.5 threshold

> logit_predicted_class <- ifelse(logit_fitted > 0.5, 1, 0)

> actual_class <- subset_data_clean$MentalHealth_01[!is.na(subset_data_clean$MentalHealth_01)]

>

> # Create confusion matrix for logit

> logit_confusion_matrix <- table(Actual = actual_class, Predicted = logit_predicted_class)
```

Show Traceback

Rerun with Debug

```
Error in table(Actual = actual_class, Predicted = logit_predicted_class) :

  all arguments must have the same length
```

```
> # Get the actual data used in the logit model

> logit_model_data <- model.frame(logit_model)

> logit_actual_class <- logit_model_data$MentalHealth_01

> logit_fitted_model <- predict(logit_model, type = "response")

>

> # Create classifications using 0.5 threshold

> logit_predicted_class <- ifelse(logit_fitted_model > 0.5, 1, 0)

>

> # Create confusion matrix for logit

> logit_confusion_matrix <- table(Actual = logit_actual_class, Predicted = logit_predicted_class)

> print(logit_confusion_matrix)

      Predicted
```

```
Actual      0      1
     0 249466    696
     1  74856    902
>
> # Calculate error rates
> if(nrow(logit_confusion_matrix) == 2 && ncol(logit_confusion_matrix) == 2) {
+    logit_type_I_error <- logit_confusion_matrix[1, 2] / sum(logit_confusion_matrix[1, ])
+    logit_type_II_error <- logit_confusion_matrix[2, 1] / sum(logit_confusion_matrix[2, ])
+    logit_total_errors <- logit_confusion_matrix[1, 2] + logit_confusion_matrix[2, 1]
+
+    cat("\nLogit Model - Type I Errors (False Positives):", logit_confusion_matrix[1, 2])
+    cat("\nLogit Model - Type II Errors (False Negatives):", logit_confusion_matrix[2, 1])
+    cat("\nLogit Model - Total Misclassifications:", logit_total_errors)
+    cat("\nLogit Model - Type I Error Rate:", round(logit_type_I_error, 4))
+    cat("\nLogit Model - Type II Error Rate:", round(logit_type_II_error, 4))
+ }

Logit Model - Type I Errors (False Positives): 696
Logit Model - Type II Errors (False Negatives): 74856
Logit Model - Total Misclassifications: 75552
Logit Model - Type I Error Rate: 0.0028
Logit Model - Type II Error Rate: 0.9881
> # Compare AIC/BIC and other metrics
> cat("MODEL COMPARISON:\n")
MODEL COMPARISON:
> cat("==================\n")
==================
>
> # OLS metrics
> ols_r2 <- summary(ols_model_01)$r.squared
```

```
> cat("OLS R-squared:", round(ols_r2, 4), "\n")

OLS R-squared: 0.0505

>

> # Logit metrics

> logit_aic <- AIC(logit_model)

> logit_null_deviance <- logit_model$null.deviance

> logit_residual_deviance <- logit_model$deviance

> logit_pseudo_r2 <- 1 - (logit_residual_deviance / logit_null_deviance)

>

> cat("Logit AIC:", round(logit_aic, 1), "\n")

Logit AIC: 336645.4

> cat("Logit Pseudo R-squared:", round(logit_pseudo_r2, 4), "\n")

Logit Pseudo R-squared: 0.0476

> cat("Logit Null Deviance:", round(logit_null_deviance, 1), "\n")

Logit Null Deviance: 353435.5

> cat("Logit Residual Deviance:", round(logit_residual_deviance, 1), "\n")

Logit Residual Deviance: 336623.4

>

> # Compare predictions

> cat("\nPredicted Probability Comparison:\n")

Predicted Probability Comparison:

> comparison_df <- data.frame(

+     Group = c("35F_75K_CG", "35F_150K_AD", "50M_75K_CG", "50M_150K_AD"),

+     OLS = c(0.3532, 0.2653, 0.2460, 0.1581),

+     Logit = logit_predictions

+ )

> print(comparison_df)

         Group    OLS     Logit

1  35F_75K_CG 0.3532 0.3580086
```

```
2 35F_150K_AD 0.2653 0.2487780


3  50M_75K_CG 0.2460 0.2292574

4 50M_150K_AD 0.1581 0.1501237
```

Answer (6):

Additional Model Comparison

Models Estimated:

1. OLS (Linear Probability Model) - Baseline
2. Logit (Logistic Regression) - Binary outcome
3. Probit (Probit Regression) - Binary outcome
4. Poisson (Poisson Regression) - Count outcome (K4SUM)

Model Comparisons:

a) Predictive Performance:

- Poisson has highest pseudo-R^2 (0.0844) but uses count data (K4SUM)
- OLS has highest R^2 (0.0505) for binary outcome
- Logit and Probit have nearly identical pseudo-R^2 (~0.0475)

b) Model Fit (AIC):

- Logit: 336,645 (best for binary outcomes)
- Probit: 336,696 (very similar to Logit)
- OLS: 346,295 (higher AIC - worse fit)
- Poisson: 1,612,699 (not comparable - different scale)

c) Predicted Values Comparison:

Group	OLS	Logit	Probit	Poisson (Count)
35F_75K_CG	**35.3%**	**35.8%**	**35.7%**	**8.05**
35F_150K_AD	**26.5%**	**24.9%**	**25.3%**	**7.17**
50M_75K_CG	**24.6%**	**22.9%**	**23.2%**	**6.93**
50M_150K_AD	**15.8%**	**15.0%**	**15.1%**	**6.16**

All models show consistent patterns: Higher income, advanced degrees, older age, and male gender reduce mental health burden.

d) Classification Performance (0.5 threshold):

- Type I Error: OLS (0.21%), Logit (0.28%), Probit (0.21%)
- Type II Error: OLS (99.04%), Logit (98.81%), Probit (99.02%)
- All suffer from extreme Type II error due to imbalanced data

e) Poisson Specific Findings:

- Overdispersion parameter: ~1.245 (moderate overdispersion)
- Significant interaction: income_midpoint:Educationadv degree (p = 5.26e-07)
- Interpretation: Each unit increase in predictors multiplicatively affects expected K4SUM count

Strengths and Weaknesses Assessment:

OLS (Linear Probability):
Strengths: Simple interpretation, consistent coefficients
Weaknesses: Can predict outside [0,1], homoskedasticity violation, highest AIC
Logit:
Strengths: Proper probability bounds, best AIC for binary models, odds-ratio interpretation
Weaknesses: Interpretation less intuitive than OLS
Probit:
Strengths: Proper probability bounds, normal distribution assumption
Weaknesses: Coefficients harder to interpret (z-scores), slightly worse AIC than Logit
Poisson:
Strengths: Uses full count information, highest explanatory power, appropriate for count data
Weaknesses: Overdispersion present, AIC not comparable, different scale

For binary classification: Logit is preferred due to best AIC and proper probability bounds, though all binary models give similar predictions.

For model richness: Poisson is theoretically superior as it uses the full count data rather than dichotomizing, but requires addressing overdispersion (e.g., with negative binomial).

Practical recommendation: Use Logit for binary decision-making and Poisson/Negative Binomial for understanding the full spectrum of mental health severity. The consistent results across all models provide robust evidence that income, education, age, and gender are important predictors of mental health outcomes.

Codes and Results for Question 6 in R (next page)

```
# Create the probit model with same predictors
> probit_model <- glm(MentalHealth_01 ~ Age + Gender + income_midpoint + Education +
+                         Age:Gender + income_midpoint:Education,
+                     data = subset_data_clean,
+                     family = binomial(link = "probit"))
>
> summary(probit_model)

Call:
glm(formula = MentalHealth_01 ~ Age + Gender + income_midpoint +
    Education + Age:Gender + income_midpoint:Education, family = binomial(link = "probit"),
    data = subset_data_clean)

Coefficients:
                                    Estimate Std. Error z value Pr(>|z|)
(Intercept)                        1.700e-01  1.762e-02   9.648  < 2e-16 ***
Age                               -1.246e-02  3.584e-04 -34.764  < 2e-16 ***
Genderfemale                       1.847e-01  2.122e-02   8.706  < 2e-16 ***
Gendertrans                        6.117e-01  2.017e-01   3.033  0.00242 **
Genderother                        8.293e-01  1.160e-01   7.149 8.77e-13 ***
income_midpoint                   -3.738e-06  5.332e-08 -70.113  < 2e-16 ***
Educationadv degree               -3.161e-02  1.036e-02  -3.051  0.00228 **
Age:Genderfemale                  -1.096e-04  4.582e-04  -0.239  0.81099
Age:Gendertrans                    2.597e-03  5.372e-03   0.483  0.62879
Age:Genderother                   -5.009e-03  2.703e-03  -1.853  0.06384 .
income_midpoint:Educationadv degree  8.810e-08  7.738e-08   1.139  0.25489
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 353436  on 325919  degrees of freedom
Residual deviance: 336674  on 325909  degrees of freedom
  (25553 observations deleted due to missingness)
AIC: 336696

Number of Fisher Scoring iterations: 4

> # Create Poisson model using the count version K4SUM
> poisson_model <- glm(K4SUM ~ Age + Gender + income_midpoint + Education +
+                          Age:Gender + income_midpoint:Education,
+                      data = subset_data_clean,
+                      family = poisson(link = "log"))
>
> summary(poisson_model)

Call:
glm(formula = K4SUM ~ Age + Gender + income_midpoint + Education +
    Age:Gender + income_midpoint:Education, family = poisson(link = "log"),
    data = subset_data_clean)

Coefficients:
                                     Estimate Std. Error  z value Pr(>|z|)
(Intercept)                         2.310e+00  4.710e-03  490.378  < 2e-16 ***
Age                                -5.239e-03  9.524e-05  -55.007  < 2e-16 ***
Genderfemale                        6.654e-02  5.711e-03   11.652  < 2e-16 ***
Gendertrans                         2.838e-01  5.101e-02    5.564 2.64e-08 ***
Genderother                         2.844e-01  2.976e-02    9.557  < 2e-16 ***
income_midpoint                    -1.498e-06  1.412e-08 -106.139  < 2e-16 ***
Educationadv degree                -1.965e-02  2.803e-03   -7.009 2.39e-12 ***
Age:Genderfemale                    1.508e-04  1.226e-04    1.230    0.219
Age:Gendertrans                    -5.292e-05  1.383e-03   -0.038    0.969
Age:Genderother                    -8.766e-04  7.043e-04   -1.245    0.213
income_midpoint:Educationadv degree  1.025e-07  2.043e-08    5.017 5.26e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 442507  on 325919  degrees of freedom
Residual deviance: 405143  on 325909  degrees of freedom
  (25553 observations deleted due to missingness)
AIC: 1612699

Number of Fisher Scoring iterations: 4

> # Get predictions for the same example cases
> example_cases <- data.frame(
+   Age = c(35, 35, 50, 50),
+   Gender = c("female", "female", "male", "male"),
+   income_midpoint = c(75000, 150000, 75000, 150000),
+   Education = c("college grad", "adv degree", "college grad", "adv degree")
+ )
>
> # OLS predictions (probabilities)
> ols_pred <- predict(ols_model_01, newdata = example_cases)
>
```

```
> # Logit predictions (probabilities)
> logit_pred <- predict(logit_model, newdata = example_cases, type = "response")
>
> # Probit predictions (probabilities)
> probit_pred <- predict(probit_model, newdata = example_cases, type = "response")
>
> # Poisson predictions (expected K4SUM counts)
> poisson_pred <- predict(poisson_model, newdata = example_cases, type = "response")
>
> # Create comparison table
> comparison <- data.frame(
+   Group = c("35F_75K_CG", "35F_150K_AD", "50M_75K_CG", "50M_150K_AD"),
+   OLS_Prob = round(ols_pred, 4),
+   Logit_Prob = round(logit_pred, 4),
+   Probit_Prob = round(probit_pred, 4),
+   Poisson_Count = round(poisson_pred, 2)
+ )
>
> print(comparison)
       Group OLS_Prob Logit_Prob Probit_Prob Poisson_Count
1  35F_75K_CG   0.3532     0.3580      0.3573          8.05
2 35F_150K_AD   0.2653     0.2488      0.2532          7.17
3  50M_75K_CG   0.2460     0.2293      0.2317          6.93
4 50M_150K_AD   0.1581     0.1501      0.1510          6.16
> # Calculate model comparison statistics
> cat("MODEL COMPARISON SUMMARY:\n")
MODEL COMPARISON SUMMARY:
> cat("==========================\n\n")
==========================

>
> # OLS metrics
> ols_r2 <- summary(ols_model_01)$r.squared
> ols_aic <- AIC(ols_model_01)
>
> # Logit metrics
> logit_aic <- AIC(logit_model)
> logit_null_dev <- logit_model$null.deviance
> logit_res_dev <- logit_model$deviance
> logit_pseudo_r2 <- 1 - (logit_res_dev / logit_null_dev)
>
> # Probit metrics
> probit_aic <- AIC(probit_model)
> probit_null_dev <- probit_model$null.deviance
> probit_res_dev <- probit_model$deviance
> probit_pseudo_r2 <- 1 - (probit_res_dev / probit_null_dev)
>
> # Poisson metrics
> poisson_aic <- AIC(poisson_model)
> poisson_null_dev <- poisson_model$null.deviance
> poisson_res_dev <- poisson_model$deviance
> poisson_pseudo_r2 <- 1 - (poisson_res_dev / poisson_null_dev)
>
> # Check Poisson overdispersion
> cat("Poisson Overdispersion Test:\n")
Poisson Overdispersion Test:
> poisson_dispersion <- sum(residuals(poisson_model, type = "pearson")^2) / poisson_model$df.residual
> cat("Dispersion parameter:", round(poisson_dispersion, 3), "\n")
Dispersion parameter: 1.37
> cat("(Values > 1 indicate overdispersion)\n\n")
(Values > 1 indicate overdispersion)

>
> # Create comparison table
> comparison_table <- data.frame(
+   Model = c("OLS", "Logit", "Probit", "Poisson"),
+   AIC = c(ols_aic, logit_aic, probit_aic, poisson_aic),
+   R2_PseudoR2 = c(ols_r2, logit_pseudo_r2, probit_pseudo_r2, poisson_pseudo_r2),
+   Interpretation = c("Linear Probability", "Log-odds", "Probit (z-scores)", "Count Data")
+ )
>
> print(comparison_table)
    Model       AIC R2_PseudoR2     Interpretation
1     OLS  346294.6  0.05046738 Linear Probability
2   Logit  336645.4  0.04756765           Log-odds
3  Probit  336695.8  0.04742500  Probit (z-scores)
4 Poisson 1612699.1  0.08443554         Count Data
> # Get Probit fitted probabilities
> probit_fitted <- predict(probit_model, type = "response")
> probit_predicted_class <- ifelse(probit_fitted > 0.5, 1, 0)
> probit_actual_class <- model.frame(probit_model)$MentalHealth_01
>
> # Create confusion matrix for probit
> probit_confusion_matrix <- table(Actual = probit_actual_class, Predicted = probit_predicted_class)
> print("Probit Confusion Matrix:")
[1] "Probit Confusion Matrix:"
> print(probit_confusion_matrix)
```

```
      Predicted
Actual      0      1
     0 249631    531
     1  75016    742
>
> # Calculate error rates
> if(nrow(probit_confusion_matrix) == 2 && ncol(probit_confusion_matrix) == 2) {
+    probit_type_I <- probit_confusion_matrix[1, 2] / sum(probit_confusion_matrix[1, ])
+    probit_type_II <- probit_confusion_matrix[2, 1] / sum(probit_confusion_matrix[2, ])
+
+    cat("\nProbit Type I Error Rate:", round(probit_type_I, 4))
+    cat("\nProbit Type II Error Rate:", round(probit_type_II, 4))
+ }

Probit Type I Error Rate: 0.0021
Probit Type II Error Rate: 0.9902
```

All Six Question/Answers Completed