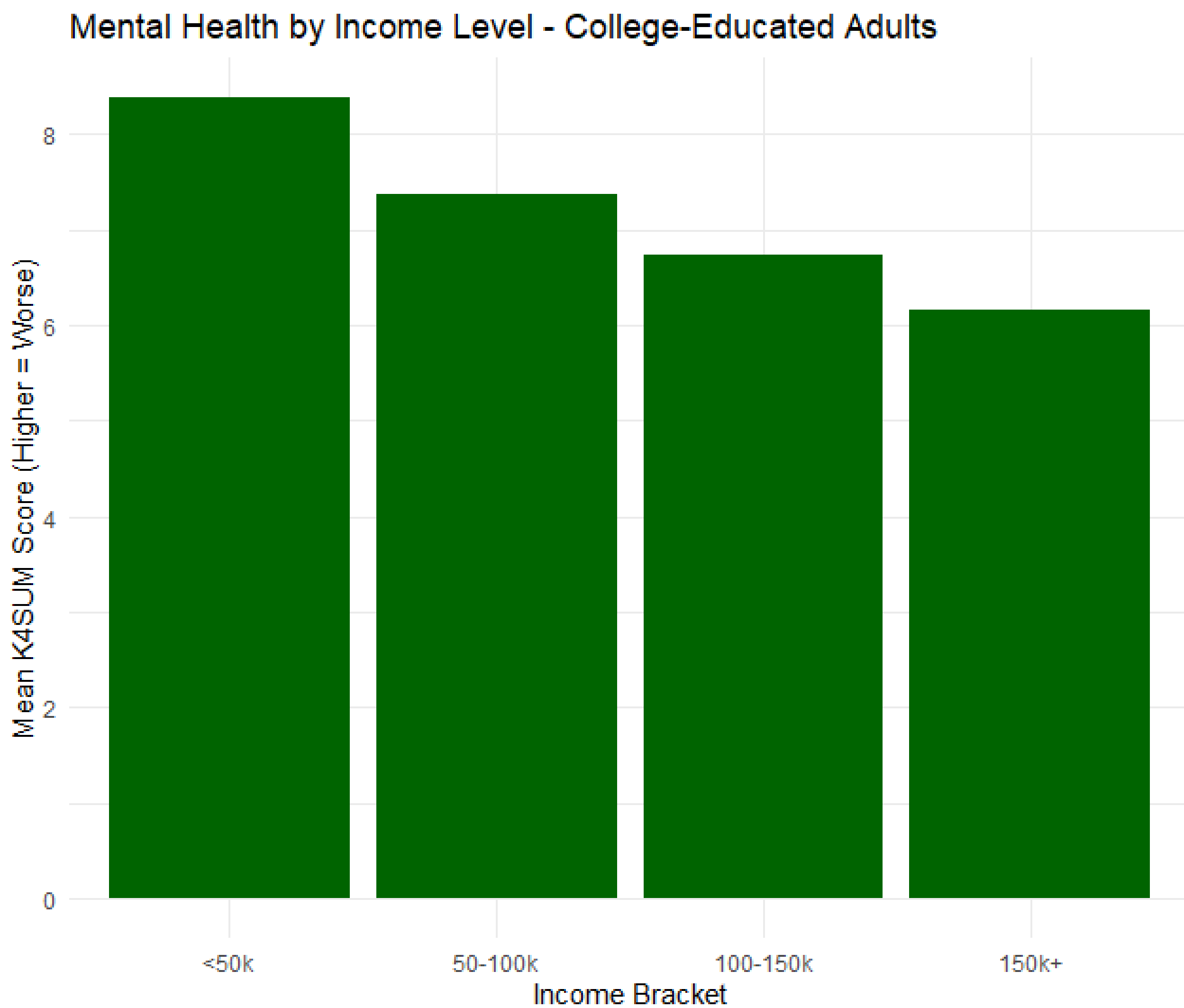
Michael Stewart
ECON B2000
11/20/2025

EXAM 2

```
library(tidyverse)
library(modelsummary)
library(ggplot2)
library(stargazer)
library(class)
library(car)

# Clear environment
rm(list = ls())

# Load
load("C:/Users/Michael/Desktop/ECON/ECON EXAM/d_HHP2020_24/d_HHP2020_24.RData")

# Check data
head(d_HHP2020_24)
names(d_HHP2020_24)

## Question 1

# When D = 0 (White households):
# Y = $\gamma_0 + \gamma_1$Age
# So: $\beta_0 = \gamma_0$ and $\beta_1 = \gamma_1$

# When D = 1 (Non-White households):
# Y = $\gamma_0 + \gamma_1$Age + $\gamma_2(1) + \gamma_3(1) \cdot$Age
# Y = $(\gamma_0 + \gamma_2) + (\gamma_1 + \gamma_3)$Age
# So: $\alpha_0 = \gamma_0 + \gamma_2$ and $\alpha_1 = \gamma_1 + \gamma_3$

# Summary:
# $\gamma_2$ = difference in intercepts between NW and W households
# $\gamma_3$ = difference in age slopes between NW and W households
# The interaction model allows us to test whether the relationship between age and the outcome differs by race.
# This is helpful since running separate regressions gives us the same information,
# but doesn't allow for a formal test of whether those differences are statistically significant.

## Question 2

# Check if there are bracket versions already
names(d_HHP2020_24)

# Create age brackets
d_HHP2020_24 <- d_HHP2020_24 %>%
  mutate(AgeBracket = cut(Age, breaks = c(24, 34, 44, 54, 64),
                          labels = c("25-34", "35-44", "45-54", "55-64")))

# OLS Regression
ols_model1 <- lm(K4SUM ~ AgeBracket + Education + income_midpoint, data = d_HHP2020_24)
```

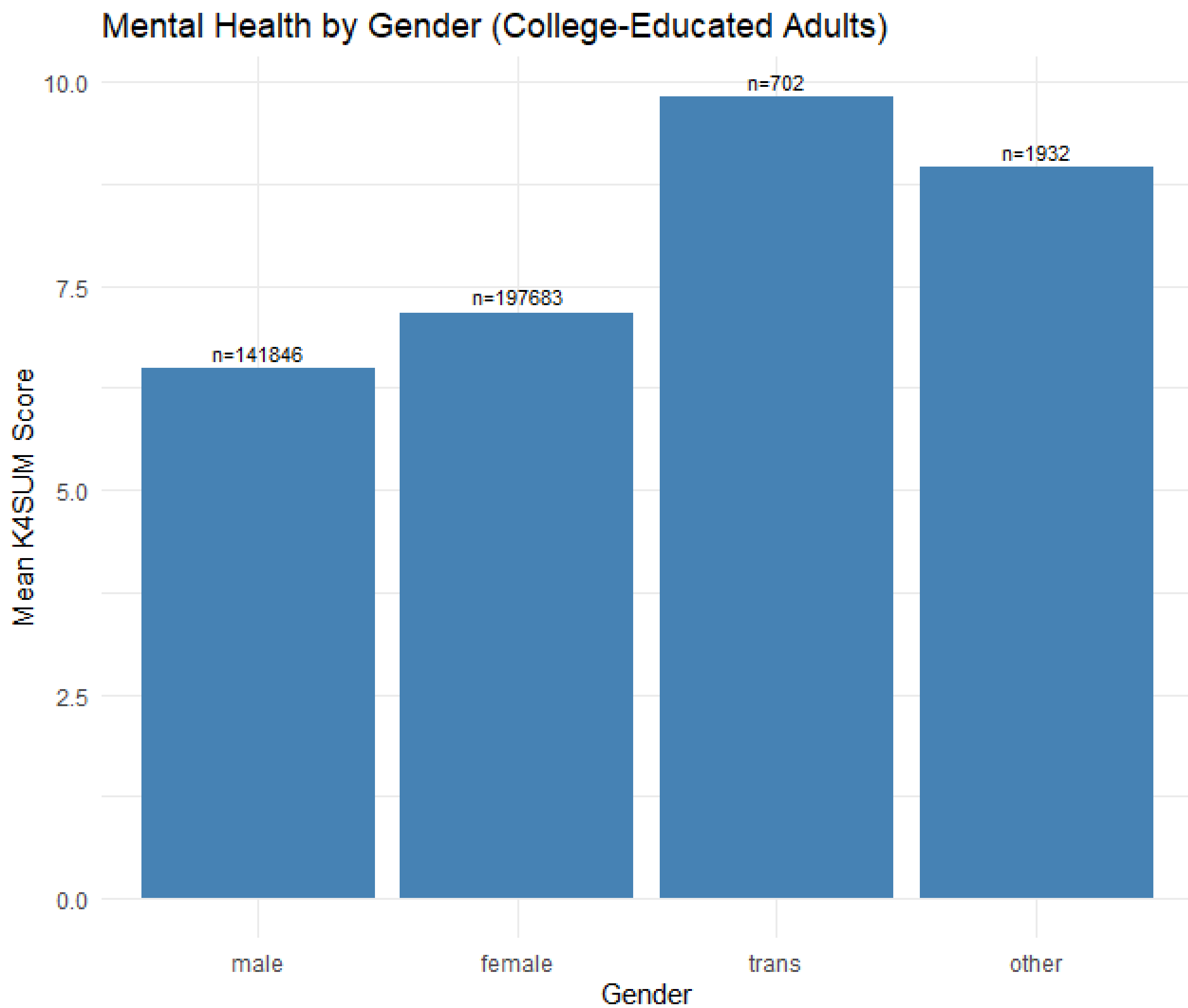
```r
summary(ols_model1)

# Joint test for ALL education coefficients
linearHypothesis(ols_model1, c("Educationsome hs = 0",
                               "Educationhigh school = 0",
                               "Educationsome college = 0",
                               "Educationassoc deg = 0",
                               "Educationcollege grad = 0",
                               "Educationadv degree = 0"))

# Test 2: Income coefficient = 0 (t-test from summary)

## Question 2 Summary:
# Both education and income are significant predictors of mental health.
# Joint F-test for all education coefficients: F(6, 558545) = 185.38, p < 2.2e-16.
# Education is highly significant overall - people with more education report better mental health.
# For example: advanced degrees are associated with 0.241 points lower K4SUM compared to
baseline.
# Income test p-value: < 2e-16 (also highly significant).
# Higher income is associated with better mental health (coefficient = -0.0000122).

## Question 3

# Focus on working-age adults (25-64) with at least a college degree.
# This allows analyzing mental health among those actively engaged in the labor market with higher
education.

# Create subsample: working-age, college-educated
subsample <- d_HHP2020_24 %>%
  filter(Age >= 25 & Age <= 64,
         Education %in% c("college grad", "adv degree"))

# Summary statistics
summary(subsample$K4SUM)
summary(subsample$Age)
summary(subsample$income_midpoint)
table(subsample$Gender)
table(subsample$Mar_Stat)
table(subsample$workloss)

# Subsample contains N=378,749 working-age (25-64) adults with college degrees.
# Of these, 342,163 have complete K4SUM data (36,586 missing).

# Summary statistics of interesting variables:
# Mean K4SUM: 6.9 (SD: moderate mental health issues on average)
# Mean age: 45 years (range: 25-64)
# Mean income: $125,627
# 58% female, 42% male
# 63% married
# 16% experienced recent household job loss

# VIZ: Mean K4SUM by income bracket
subsample %>%
  filter(!is.na(K4SUM), !is.na(income_midpoint)) %>%
```

```
  mutate(income_bracket = cut(income_midpoint,
                              breaks = c(0, 50000, 100000, 150000, 250000),
                              labels = c("<50k", "50-100k", "100-150k", "150k+"))) %>%
  group_by(income_bracket) %>%
  summarise(mean_K4SUM = mean(K4SUM)) %>%
  ggplot(aes(x = income_bracket, y = mean_K4SUM)) +
  geom_bar(stat = "identity", fill = "darkgreen") +
  labs(title = "Mental Health by Income Level - College-Educated Adults",
       y = "Mean K4SUM Score (Higher = Worse)",
       x = "Income Bracket") +
  theme_minimal()
```



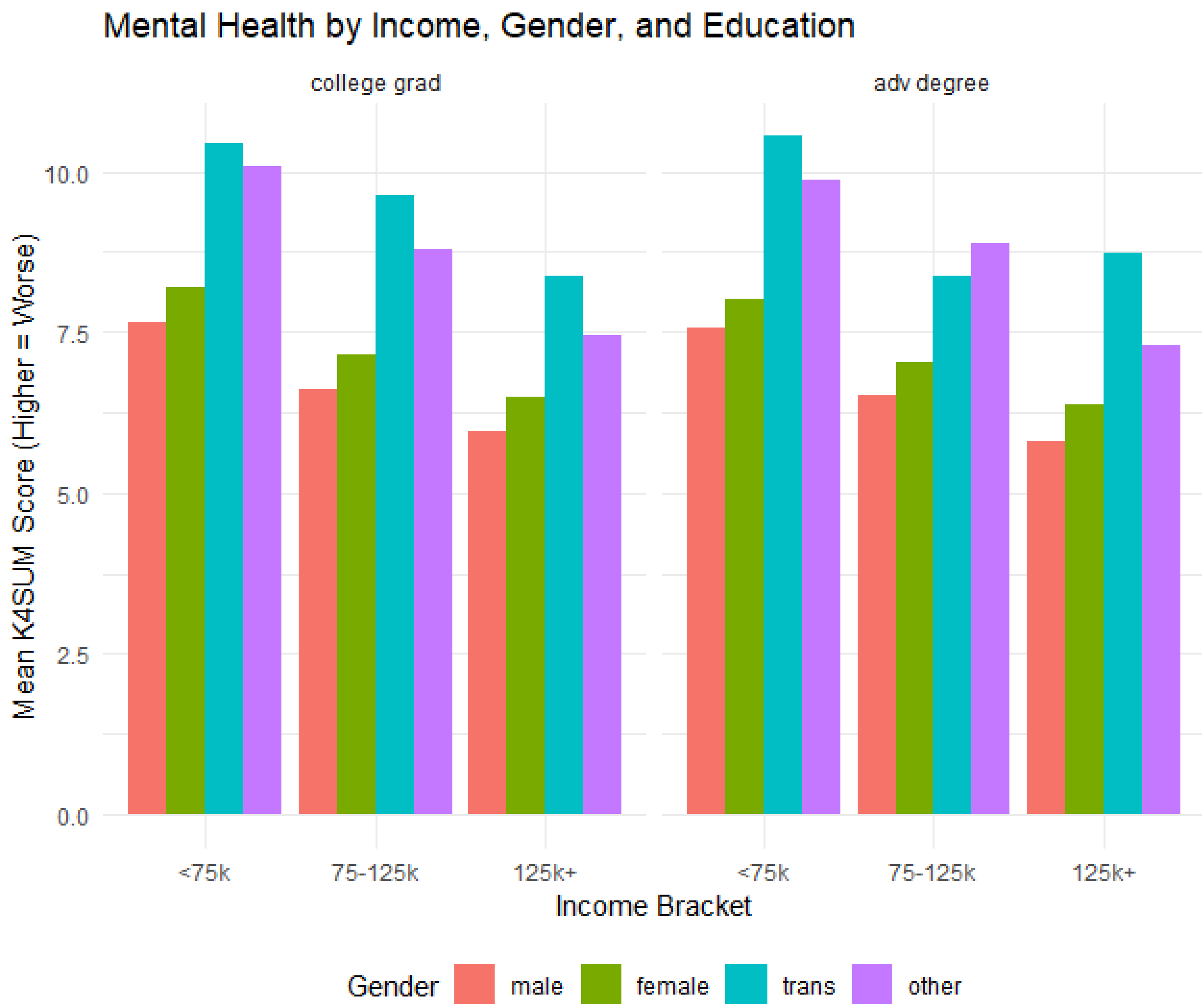

```
# VIZ: Mean K4SUM by Gender
subsample %>%
  filter(!is.na(K4SUM)) %>%
  group_by(Gender) %>%
  summarise(mean_K4SUM = mean(K4SUM),
            n = n()) %>%
  ggplot(aes(x = Gender, y = mean_K4SUM)) +
  geom_bar(stat = "identity", fill = "steelblue") +
```

```
  geom_text(aes(label = paste0("n=", n)), vjust = -0.5, size = 3) +
  labs(title = "Mental Health by Gender (College-Educated Adults)",
       y = "Mean K4SUM Score", x = "Gender") +
  theme_minimal()
```




```
# Notable: trans and other gender individuals show elevated mental health concerns,
# though they represent <1% of the sample (n=2,925).

# VIZ: Mental health by income and gender (all genders), faceted by education
subsample %>%
  filter(!is.na(K4SUM), !is.na(income_midpoint)) %>%
  mutate(income_bracket = cut(income_midpoint,
                              breaks = c(0, 75000, 125000, 250000),
                              labels = c("<75k", "75-125k", "125k+"))) %>%
  group_by(Education, Gender, income_bracket) %>%
  summarise(mean_K4SUM = mean(K4SUM), n = n(), .groups = "drop") %>%
  ggplot(aes(x = income_bracket, y = mean_K4SUM, fill = Gender)) +
  geom_bar(stat = "identity", position = "dodge") +
  facet_wrap(~Education) +
  labs(title = "Mental Health by Income, Gender, and Education",
       y = "Mean K4SUM Score (Higher = Worse)",
```

```
      x = "Income Bracket") +
  theme_minimal() +
  theme(legend.position = "bottom")
```




```
# VIZ: Mental health by income and gender (male/female only), faceted by education
subsample %>%
  filter(!is.na(K4SUM), !is.na(income_midpoint), Gender %in% c("male", "female")) %>%
  mutate(income_bracket = cut(income_midpoint,
                              breaks = c(0, 75000, 125000, 250000),
                              labels = c("<75k", "75-125k", "125k+"))) %>%
  group_by(Education, Gender, income_bracket) %>%
  summarise(mean_K4SUM = mean(K4SUM), n = n(), .groups = "drop") %>%
  ggplot(aes(x = income_bracket, y = mean_K4SUM, fill = Gender)) +
  geom_bar(stat = "identity", position = "dodge") +
  facet_wrap(~Education) +
  labs(title = "Mental Health by Income and Gender - Male and Female Only",
      subtitle = "Faceted by Education Level",
      y = "Mean K4SUM Score (Higher = Worse)",
      x = "Income Bracket") +
  theme_minimal() +
```

```
  theme(legend.position = "bottom")
```


```
## Question 4

# Create binary mental health variable
subsample <- subsample %>%
  mutate(MentalHealth_01 = ifelse(K4SUM > 8, 1, 0))

# Check it
table(subsample$MentalHealth_01, useNA = "always")

# OLS with binary outcome and interaction
ols_binary <- lm(MentalHealth_01 ~ Age + Gender + Education + income_midpoint +
                 Gender:Education, data = subsample)
summary(ols_binary)

# Q4a:
# I include Age, Gender, Education, and income as predictors, with a Gender×Education interaction.
# Exogeneity is questionable - income could be affected by mental health (reverse causality),
# and unobserved factors like family background likely influence both education and mental health.
```

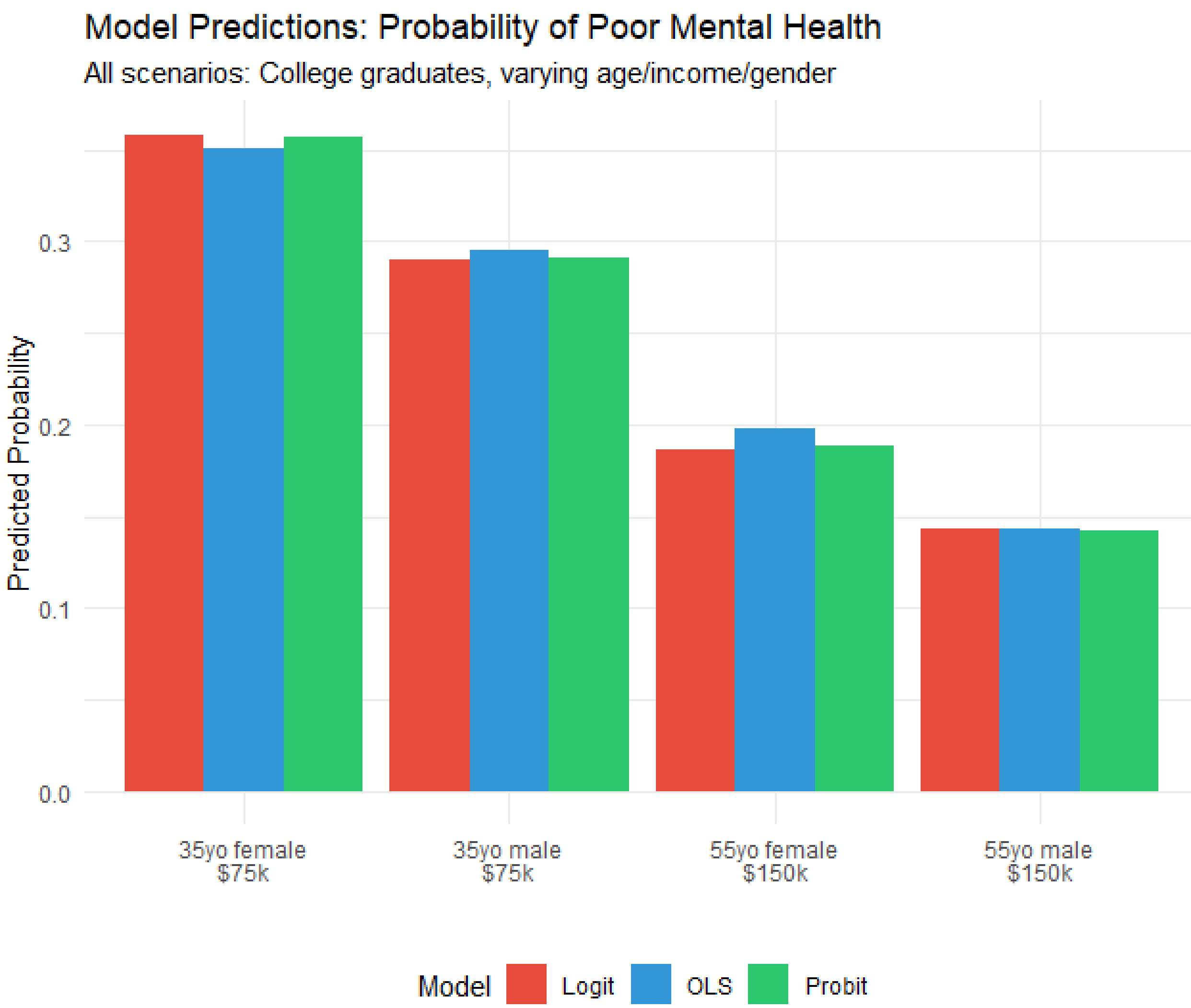
```
# The interaction tests whether the education effect differs by gender.

# Q4b:
# Results are plausible: older age reduces poor mental health probability (-0.0036 per year).
# Females have higher risk (+0.055), higher income protective (-0.0000011 per dollar).
# Most coefficients are highly significant (p<0.001).
# The female×advanced degree interaction (-0.011, p=0.000178) suggests education benefits women's
mental health more than men's.

# Joint test for ALL education-related terms (main effect + interactions)
linearHypothesis(ols_binary, c("Educationadv degree = 0",
                               "Genderfemale:Educationadv degree = 0",
                               "Gendertrans:Educationadv degree = 0",
                               "Genderother:Educationadv degree = 0"))

# Joint test for all education-related terms (main effect + all interactions):
# F(4, 317256) = 8.82, p = 4.047e-07.
# Education effects ARE statistically significant when considering both the main effect
# and how education interacts with gender. The significant female×education interaction
# (p=0.000178) drives this result.

# Q4d:
# Predictions
pred_data <- data.frame(
  Age = c(35, 35, 55, 55),
  Gender = c("male", "female", "male", "female"),
  Education = c("college grad", "college grad", "adv degree", "adv degree"),
  income_midpoint = c(75000, 75000, 150000, 150000)
)
pred_data$predicted_prob <- predict(ols_binary, newdata = pred_data)
pred_data

# Predicted probabilities:
# 35-year-old male college grad at $75k: 29.5%;
# same female: 35.1%;
# 55-year-old male with advanced degree at $150k:
# 14.3%; same female: 18.7%.
# Older age and higher income reduce risk;
# females have higher risk.

# Q4e:
subsample <- subsample %>%
  mutate(predicted_class = ifelse(predict(ols_binary, newdata = subsample) > 0.5, 1, 0))

# Confusion matrix
table(Actual = subsample$MentalHealth_01, Predicted = subsample$predicted_class)

# Calculate error rates
confusion <- table(Actual = subsample$MentalHealth_01, Predicted = subsample$predicted_class)
confusion

# Type I error: predict poor mental health when actually good
type1 <- confusion[1, 2] / sum(confusion[1, ])
```

```r
# Type II error: predict good mental health when actually poor
type2 <- confusion[2, 1] / sum(confusion[2, ])

cat("Type I error rate:", type1, "\n")
cat("Type II error rate:", type2, "\n")

# Type I error rate: 0.21% (509 false positives out of 242,760 actually healthy).
# Type II error rate: 99.0% (73,782 false negatives out of 74,506 actually poor mental health).
# The model is very conservative - it rarely predicts poor mental health, leading to many missed cases
but few false alarms.

## Question 5

# Logit model
logit_model <- glm(MentalHealth_01 ~ Age + Gender + Education + income_midpoint +
                     Gender:Education,
                   data = subsample,
                   family = binomial(link = "logit"))
summary(logit_model)

# Q5a:
# I use the same predictors as Q4 (Age, Gender, Education, income, Gender×Education interaction).
# Exogeneity concerns remain - income and mental health may have reverse causality, and unobserved
factors affect both education and mental health.
# However, logit is more appropriate than OLS for binary outcomes because it constrains predicted
probabilities to [0,1] and models the log-odds rather than assuming a linear probability.

# Q5b:
# Results are plausible: older age reduces poor mental health risk (coef=-0.020).
# Females have higher risk (+0.314), higher income protective (-0.0000065).
# Most coefficients are highly significant (p<0.001).
# Unlike OLS, the female × education interaction is not significant in logit (p=0.364).

# Q5c:
# Joint test for ALL education-related terms in logit
linearHypothesis(logit_model, c("Educationadv degree = 0",
                                "Genderfemale:Educationadv degree = 0",
                                "Gendertrans:Educationadv degree = 0",
                                "Genderother:Educationadv degree = 0"))

# Joint test for all education-related terms (main effect + all interactions):
# χ²(4) = 20.04, p = 0.00049.
# Education effects ARE statistically significant in the logit model when considering
# both the main effect and interactions with gender. This differs from testing the
# main effect alone (p=0.057), showing the importance of accounting for how education's
# effect varies by gender.

# Q5d:
pred_data_logit <- data.frame(
  Age = c(35, 35, 55, 55),
  Gender = c("male", "female", "male", "female"),
  Education = c("college grad", "college grad", "adv degree", "adv degree"),
  income_midpoint = c(75000, 75000, 150000, 150000)
)
```

```r
pred_data_logit$predicted_prob <- predict(logit_model, newdata = pred_data_logit, type = "response")
pred_data_logit

# Predicted probabilities (logit):
# 35-year-old male college grad at $75k: 29.0%;
# same female: 35.8%;
# 55-year-old male with advanced degree at $150k: 14.0%;
# same female: 17.9%.
# Very similar to OLS predictions, showing older age and higher income reduce risk.

# Q5e:
logit_preds <- predict(logit_model, newdata = subsample, type = "response")

subsample <- subsample %>%
  mutate(predicted_class_logit = ifelse(!is.na(MentalHealth_01),
                                        ifelse(logit_preds > 0.5, 1, 0),
                                        NA))

confusion_logit <- table(Actual = subsample$MentalHealth_01, Predicted =
subsample$predicted_class_logit)
confusion_logit

type1_logit <- confusion_logit[1, 2] / sum(confusion_logit[1, ])
type2_logit <- confusion_logit[2, 1] / sum(confusion_logit[2, ])

cat("Logit Type I error:", type1_logit, "\n")
cat("Logit Type II error:", type2_logit, "\n")

# Logit Type I error: 0.29% (697 false positives).
# Type II error: 98.8% (73,600 false negatives).
# Very similar to OLS errors - both models are extremely conservative,
# rarely predicting poor mental health, leading to many missed cases.

# Q5f:
# Logit and OLS produce very similar predictions and error rates.
# Logit Type I: 0.29% vs OLS: 0.21%; Type II: 98.8% vs OLS: 99.0%.
# Both models are extremely conservative, rarely predicting poor mental health.
# Logit predicted probabilities are similar to OLS (e.g., 35yo female: 35.8% vs 35.1%).
# Logit is theoretically superior for binary outcomes as it constrains probabilities to [0,1].
# AIC for logit: 329,664 (lower AIC indicates better fit when comparing models).

## Question 6 - Probit Model

probit_model <- glm(MentalHealth_01 ~ Age + Gender + Education + income_midpoint +
                      Gender:Education,
                    data = subsample,
                    family = binomial(link = "probit"))
summary(probit_model)

# Predictions with probit
pred_data_probit <- data.frame(
  Age = c(35, 35, 55, 55),
  Gender = c("male", "female", "male", "female"),
  Education = c("college grad", "college grad", "adv degree", "adv degree"),
```

```
  income_midpoint = c(75000, 75000, 150000, 150000)
)
pred_data_probit$predicted_prob <- predict(probit_model, newdata = pred_data_probit, type =
"response")
pred_data_probit

# Probit confusion matrix
probit_preds <- predict(probit_model, newdata = subsample, type = "response")

subsample <- subsample %>%
  mutate(predicted_class_probit = ifelse(!is.na(MentalHealth_01),
                                         ifelse(probit_preds > 0.5, 1, 0),
                                         NA))

confusion_probit <- table(Actual = subsample$MentalHealth_01, Predicted =
subsample$predicted_class_probit)
confusion_probit

type1_probit <- confusion_probit[1, 2] / sum(confusion_probit[1, ])
type2_probit <- confusion_probit[2, 1] / sum(confusion_probit[2, ])

cat("Probit Type I error:", type1_probit, "\n")
cat("Probit Type II error:", type2_probit, "\n")

# Compare AICs
cat("\nModel Comparison (AIC):\n")
cat("OLS: N/A (not comparable)\n")
cat("Logit AIC:", AIC(logit_model), "\n")
cat("Probit AIC:", AIC(probit_model), "\n")

# VIZ: Compare predicted probabilities across models
pred_comparison <- data.frame(
  Age = c(35, 35, 55, 55),
  Gender = c("male", "female", "male", "female"),
  Education = rep("college grad", 4),
  income_midpoint = c(75000, 75000, 150000, 150000)
)

# Get predictions from all three models
pred_comparison$OLS <- predict(ols_binary, newdata = pred_comparison)
pred_comparison$Logit <- predict(logit_model, newdata = pred_comparison, type = "response")
pred_comparison$Probit <- predict(probit_model, newdata = pred_comparison, type = "response")

# Reshape for plotting
comparison_long <- pred_comparison %>%
  mutate(Scenario = paste0(Age, "yo ", Gender, "\n$", income_midpoint/1000, "k")) %>%
  select(Scenario, OLS, Logit, Probit) %>%
  pivot_longer(cols = c(OLS, Logit, Probit),
               names_to = "Model", values_to = "Probability")

# Plot
ggplot(comparison_long, aes(x = Scenario, y = Probability, fill = Model)) +
  geom_bar(stat = "identity", position = "dodge") +
  scale_fill_manual(values = c("OLS" = "#3498db",      # blue
```

```
                    "Logit" = "#e74c3c",    # red
                    "Probit" = "#2ecc71")) + # green
  labs(title = "Model Predictions: Probability of Poor Mental Health",
       subtitle = "All scenarios: College graduates, varying age/income/gender",
       y = "Predicted Probability",
       x = "") +
  theme_minimal() +
  theme(legend.position = "bottom")
```




```
## Question 6 Summary:
# estimated a probit model as an alternative to logit for binary outcomes.
# Probit uses a normal CDF link function instead of logistic.
#
# Results are similar across all three models:
# - Predicted probabilities nearly identical (e.g., 35yo female: OLS 35.1%, Logit 35.8%, Probit 35.7%)
# - Error rates comparable: Probit Type I: 0.22%, Type II: 99.0%
# - All three models are extremely conservative, rarely predicting poor mental health
#
# Model comparison:
# - Logit AIC: 329,664 (slightly better)
```

```
# - Probit AIC: 329,718
# - Lower AIC indicates logit fits marginally better
#
# Strengths: Logit/Probit constrain probabilities to [0,1], theoretically appropriate for binary outcomes
# Weaknesses: All models have very high Type II error rates (miss 99% of poor mental health cases),
# suggesting we need better predictors or different threshold than 0.5 for classification.
```

```
library(tidyverse)
> library(modelsummary)
> library(ggplot2)
> library(stargazer)
> library(class)
> library(car)
>
> # Clear environment
> rm(list = ls())
>
> # Load
> load("C:/Users/Michael/Desktop/ECON/ECON EXAM/d_HHP2020_24/d_HHP2020_24.RData")
>
> # Check data
> head(d_HHP2020_24)
  Age Gender    Education Mar_Stat income_midpoint  Race     Hispanic Number_people_HH
1  34 female college grad  Married           62500 white not Hispanic                4
2  65   male some college divorced           30000 white not Hispanic                1
3  44 female college grad  Married          225000 other not Hispanic                2
4  56   male some college divorced           12500 white not Hispanic                2
5  57 female   adv degree    never           62500 white not Hispanic                1
6  44 female   adv degree  Married          125000 white not Hispanic                2
  Number_kids_HH Number_adults_HH private_health_ins public_health_ins
1              2                2                  0                 0
2              0                1                  0                 0
3              0                2                  0                 0
4              0                2                  0                 0
5              0                1                  0                 0
6              0                2                  0                 0
                         work_kind                          workloss DOWN ANXIOUS WORRY INTERES
1          employed by private co                                no    1       4     3
2                             <NA>                               no    4       3     4
3 employed by nonprofit or charity                               no    1       1     1
4                             <NA> yes recent household loss of work    4       4     4
5 employed by nonprofit or charity                               no    2       2     1
6          employed by private co                                no    2       3     2
  YEAR Begin_Date K4SUM income_midpoint_factor
1   20 2020-04-23     9                  62500
2   20 2020-04-23    15                  30000
3   20 2020-04-23     4                 225000
4   20 2020-04-23    16                  12500
5   20 2020-04-23     7                  62500
6   20 2020-04-23     9                 125000
> names(d_HHP2020_24)
 [1] "Age"                "Gender"             "Education"
 [4] "Mar_Stat"           "income_midpoint"    "Race"
 [7] "Hispanic"           "Number_people_HH"   "Number_kids_HH"
[10] "Number_adults_HH"   "private_health_ins" "public_health_ins"
[13] "work_kind"          "workloss"           "DOWN"
[16] "ANXIOUS"            "WORRY"              "INTEREST"
[19] "YEAR"               "Begin_Date"         "K4SUM"
```

```
[22] "income_midpoint_factor"
>
> ## Question 1
>
> # When D = 0 (White households):
> # Y = γ₀ + γ₁Age
> # So: β₀ = γ₀ and β₁ = γ₁
>
> # When D = 1 (Non-White households):
> # Y = γ₀ + γ₁Age + γ₂(1) + γ₃(1)·Age
> # Y = (γ₀ + γ₂) + (γ₁ + γ₃)Age
> # So: α₀ = γ₀ + γ₂ and α₁ = γ₁ + γ₃
>
> # Summary:
> # γ₂ = difference in intercepts between NW and W households
> # γ₃ = difference in age slopes between NW and W households
> # The interaction model allows us to test whether the relationship between age and the outcor
differs by race.
> # This is helpful since running separate regressions gives us the same information,
> # but doesn't allow for a formal test of whether those differences are statistically signific
>
> ## Question 2
>
> # Check if there are bracket versions already
> names(d_HHP2020_24)
 [1] "Age"                    "Gender"                 "Education"
 [4] "Mar_Stat"               "income_midpoint"        "Race"
 [7] "Hispanic"               "Number_people_HH"       "Number_kids_HH"
[10] "Number_adults_HH"       "private_health_ins"     "public_health_ins"
[13] "work_kind"              "workloss"               "DOWN"
[16] "ANXIOUS"                "WORRY"                  "INTEREST"
[19] "YEAR"                   "Begin_Date"             "K4SUM"
[22] "income_midpoint_factor"
>
> # Create age brackets
> d_HHP2020_24 <- d_HHP2020_24 %>%
+   mutate(AgeBracket = cut(Age, breaks = c(24, 34, 44, 54, 64),
+                          labels = c("25-34", "35-44", "45-54", "55-64")))
>
> # OLS Regression
> ols_model1 <- lm(K4SUM ~ AgeBracket + Education + income_midpoint, data = d_HHP2020_24)
> summary(ols_model1)

Call:
lm(formula = K4SUM ~ AgeBracket + Education + income_midpoint,
    data = d_HHP2020_24)

Residuals:
    Min      1Q  Median      3Q     Max
-5.2403 -2.5524 -0.8587  1.7432 10.9463
```

```
Coefficients:
                        Estimate Std. Error  t value Pr(>|t|)    
(Intercept)            9.227e+00  5.895e-02  156.516  < 2e-16 ***
AgeBracket35-44       -3.062e-01  1.362e-02  -22.479  < 2e-16 ***
AgeBracket45-54       -5.520e-01  1.389e-02  -39.734  < 2e-16 ***
AgeBracket55-64       -1.197e+00  1.363e-02  -87.820  < 2e-16 ***
Educationsome hs      -1.238e-01  6.983e-02   -1.773  0.07619 .  
Educationhigh school  -1.695e-01  5.963e-02   -2.842  0.00448 ** 
Educationsome college  1.650e-01  5.898e-02    2.798  0.00515 ** 
Educationassoc deg    -6.136e-02  5.973e-02   -1.027  0.30427    
Educationcollege grad -2.354e-01  5.890e-02   -3.996 6.44e-05 ***
Educationadv degree   -2.413e-01  5.917e-02   -4.078 4.54e-05 ***
income_midpoint       -1.216e-05  7.473e-08 -162.657  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.352 on 558545 degrees of freedom
  (426234 observations deleted due to missingness)
Multiple R-squared:  0.07853,	Adjusted R-squared:  0.07852 
F-statistic:  4760 on 10 and 558545 DF,  p-value: < 2.2e-16

> 
> # Joint test for ALL education coefficients
> 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 ~ AgeBracket + Education + income_midpoint

  Res.Df     RSS Df Sum of Sq      F    Pr(>F)    
1 558551 6288269                                  
2 558545 6275772  6     12497 185.38 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> 
> # Test 2: Income coefficient = 0 (t-test from summary)
> 
> ## Question 2 Summary:
> # Both education and income are significant predictors of mental health.
```

```
> # Joint F-test for all education coefficients: F(6, 558545) = 185.38, p < 2.2e-16.
> # Education is highly significant overall - people with more education report better mental
health.
> # For example: advanced degrees are associated with 0.241 points lower K4SUM compared to base
> # Income test p-value: < 2e-16 (also highly significant).
> # Higher income is associated with better mental health (coefficient = -0.0000122).
>
> ## Question 3
>
> # Focus on working-age adults (25-64) with at least a college degree.
> # This allows analyzing mental health among those actively engaged in the labor market with h
education.
>
> # Create subsample: working-age, college-educated
> subsample <- d_HHP2020_24 %>%
+  filter(Age >= 25 & Age <= 64,
+         Education %in% c("college grad", "adv degree"))
>
> # Summary statistics
> summary(subsample$K4SUM)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
  4.000   4.000   6.000   6.909   8.000  16.000   36586
> summary(subsample$Age)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  25.00   36.00   45.00   45.14   54.00   64.00
> summary(subsample$income_midpoint)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
  12500   62500  125000  125627  175000  225000   60104
> table(subsample$Gender)

  male female  trans  other
157018 218806    760   2165
> table(subsample$Mar_Stat)

  Married   widowed  divorced separated     never
   239504      5238     43344      4855     83157
> table(subsample$workloss)

yes recent household loss of work                                no
                            60384                            312618
>
> # Subsample contains N=378,749 working-age (25-64) adults with college degrees.
> # Of these, 342,163 have complete K4SUM data (36,586 missing).
>
> # Summary statistics of interesting variables:
> # Mean K4SUM: 6.9 (SD: moderate mental health issues on average)
> # Mean age: 45 years (range: 25-64)
> # Mean income: $125,627
> # 58% female, 42% male
> # 63% married
> # 16% experienced recent household job loss
```

```
>
> # VIZ: Mean K4SUM by income bracket
> subsample %>%
+  filter(!is.na(K4SUM), !is.na(income_midpoint)) %>%
+  mutate(income_bracket = cut(income_midpoint,
+                              breaks = c(0, 50000, 100000, 150000, 250000),
+                              labels = c("<50k", "50-100k", "100-150k", "150k+"))) %>%
+  group_by(income_bracket) %>%
+  summarise(mean_K4SUM = mean(K4SUM)) %>%
+  ggplot(aes(x = income_bracket, y = mean_K4SUM)) +
+  geom_bar(stat = "identity", fill = "darkgreen") +
+  labs(title = "Mental Health by Income Level - College-Educated Adults",
+       y = "Mean K4SUM Score (Higher = Worse)",
+       x = "Income Bracket") +
+  theme_minimal()
>
> # VIZ: Mean K4SUM by Gender
> subsample %>%
+  filter(!is.na(K4SUM)) %>%
+  group_by(Gender) %>%
+  summarise(mean_K4SUM = mean(K4SUM),
+            n = n()) %>%
+  ggplot(aes(x = Gender, y = mean_K4SUM)) +
+  geom_bar(stat = "identity", fill = "steelblue") +
+  geom_text(aes(label = paste0("n=", n)), vjust = -0.5, size = 3) +
+  labs(title = "Mental Health by Gender (College-Educated Adults)",
+       y = "Mean K4SUM Score", x = "Gender") +
+  theme_minimal()
>
> # Notable: trans and other gender individuals show elevated mental health concerns,
> # though they represent <1% of the sample (n=2,925).
>
> # VIZ: Mental health by income and gender (all genders), faceted by education
> subsample %>%
+  filter(!is.na(K4SUM), !is.na(income_midpoint)) %>%
+  mutate(income_bracket = cut(income_midpoint,
+                              breaks = c(0, 75000, 125000, 250000),
+                              labels = c("<75k", "75-125k", "125k+"))) %>%
+  group_by(Education, Gender, income_bracket) %>%
+  summarise(mean_K4SUM = mean(K4SUM), n = n(), .groups = "drop") %>%
+  ggplot(aes(x = income_bracket, y = mean_K4SUM, fill = Gender)) +
+  geom_bar(stat = "identity", position = "dodge") +
+  facet_wrap(~Education) +
+  labs(title = "Mental Health by Income, Gender, and Education",
+       y = "Mean K4SUM Score (Higher = Worse)",
+       x = "Income Bracket") +
+  theme_minimal() +
+  theme(legend.position = "bottom")
>
> # VIZ: Mental health by income and gender (male/female only), faceted by education
> subsample %>%
```

```
+  filter(!is.na(K4SUM), !is.na(income_midpoint), Gender %in% c("male", "female")) %>%
+  mutate(income_bracket = cut(income_midpoint,
+                              breaks = c(0, 75000, 125000, 250000),
+                              labels = c("<75k", "75-125k", "125k+"))) %>%
+  group_by(Education, Gender, income_bracket) %>%
+  summarise(mean_K4SUM = mean(K4SUM), n = n(), .groups = "drop") %>%
+  ggplot(aes(x = income_bracket, y = mean_K4SUM, fill = Gender)) +
+  geom_bar(stat = "identity", position = "dodge") +
+  facet_wrap(~Education) +
+  labs(title = "Mental Health by Income and Gender - Male and Female Only",
+       subtitle = "Faceted by Education Level",
+       y = "Mean K4SUM Score (Higher = Worse)",
+       x = "Income Bracket") +
+  theme_minimal() +
+  theme(legend.position = "bottom")
>
> ## Question 4
>
> # Create binary mental health variable
> subsample <- subsample %>%
+  mutate(MentalHealth_01 = ifelse(K4SUM > 8, 1, 0))
>
> # Check it
> table(subsample$MentalHealth_01, useNA = "always")

     0      1   <NA>
262556  79607  36586
>
> # OLS with binary outcome and interaction
> ols_binary <- lm(MentalHealth_01 ~ Age + Gender + Education + income_midpoint +
+                    Gender:Education, data = subsample)
> summary(ols_binary)

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

Residuals:
     Min       1Q   Median       3Q      Max
-0.66629 -0.27088 -0.18121 -0.04716  0.97066

Coefficients:
                                   Estimate Std. Error t value Pr(>|t|)
(Intercept)                       5.015e-01  3.637e-03 137.874  < 2e-16 ***
Age                              -3.563e-03  6.834e-05 -52.134  < 2e-16 ***
Genderfemale                      5.541e-02  2.041e-03  27.154  < 2e-16 ***
Gendertrans                       2.674e-01  2.019e-02  13.244  < 2e-16 ***
Genderother                       2.275e-01  1.315e-02  17.302  < 2e-16 ***
Educationadv degree              -1.735e-06  2.312e-03  -0.001 0.999401
income_midpoint                  -1.085e-06  1.122e-08 -96.733  < 2e-16 ***
Genderfemale:Educationadv degree -1.124e-02  2.999e-03  -3.748 0.000178 ***
```

```
Gendertrans:Educationadv degree  -2.070e-02  3.344e-02  -0.619 0.535873
Genderother:Educationadv degree  -2.539e-02  1.989e-02  -1.277 0.201775
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4132 on 317256 degrees of freedom
  (61483 observations deleted due to missingness)
Multiple R-squared:  0.04985,	Adjusted R-squared:  0.04982
F-statistic:  1849 on 9 and 317256 DF,  p-value: < 2.2e-16

>
> # Q4a:
> # I include Age, Gender, Education, and income as predictors, with a Gender×Education interac
> # Exogeneity is questionable - income could be affected by mental health (reverse causality),
> # and unobserved factors like family background likely influence both education and mental he
> # The interaction tests whether the education effect differs by gender.
>
> # Q4b:
> # Results are plausible: older age reduces poor mental health probability (-0.0036 per year).
> # Females have higher risk (+0.055), higher income protective (-0.0000011 per dollar).
> # Most coefficients are highly significant (p<0.001).
> # The female×advanced degree interaction (-0.011, p=0.000178) suggests education benefits wor
mental health more than men's.
>
> # Joint test for ALL education-related terms (main effect + interactions)
> linearHypothesis(ols_binary, c("Educationadv degree = 0",
+                               "Genderfemale:Educationadv degree = 0",
+                               "Gendertrans:Educationadv degree = 0",
+                               "Genderother:Educationadv degree = 0"))

Linear hypothesis test:
Educationadv degree = 0
Genderfemale:Educationadv degree = 0
Gendertrans:Educationadv degree = 0
Genderother:Educationadv degree = 0

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

  Res.Df   RSS Df Sum of Sq      F    Pr(>F)    
1 317260 54173                                  
2 317256 54167  4    6.0259 8.8234 4.047e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
> # Joint test for all education-related terms (main effect + all interactions):
> # F(4, 317256) = 8.82, p = 4.047e-07.
> # Education effects ARE statistically significant when considering both the main effect
> # and how education interacts with gender. The significant female×education interaction
> # (p=0.000178) drives this result.
```

```
>
> # Q4d:
> # Predictions
> pred_data <- data.frame(
+  Age = c(35, 35, 55, 55),
+  Gender = c("male", "female", "male", "female"),
+  Education = c("college grad", "college grad", "adv degree", "adv degree"),
+  income_midpoint = c(75000, 75000, 150000, 150000)
+ )
> pred_data$predicted_prob <- predict(ols_binary, newdata = pred_data)
> pred_data
  Age Gender    Education income_midpoint predicted_prob
1  35   male college grad           75000      0.2954141
2  35 female college grad           75000      0.3508264
3  55   male   adv degree          150000      0.1427828
4  55 female   adv degree          150000      0.1869548
>
> # Predicted probabilities:
> # 35-year-old male college grad at $75k: 29.5%;
> # same female: 35.1%;
> # 55-year-old male with advanced degree at $150k:
> # 14.3%; same female: 18.7%.
> # Older age and higher income reduce risk;
> # females have higher risk.
>
> # Q4e:
> subsample <- subsample %>%
+  mutate(predicted_class = ifelse(predict(ols_binary, newdata = subsample) > 0.5, 1, 0))
>
> # Confusion matrix
> table(Actual = subsample$MentalHealth_01, Predicted = subsample$predicted_class)
      Predicted
Actual      0      1
     0 242251    509
     1  73782    724
>
> # Calculate error rates
> confusion <- table(Actual = subsample$MentalHealth_01, Predicted = subsample$predicted_class)
> confusion
      Predicted
Actual      0      1
     0 242251    509
     1  73782    724
>
> # Type I error: predict poor mental health when actually good
> type1 <- confusion[1, 2] / sum(confusion[1, ])
>
> # Type II error: predict good mental health when actually poor
> type2 <- confusion[2, 1] / sum(confusion[2, ])
>
> cat("Type I error rate:", type1, "\n")
```

```
Type I error rate: 0.002096721
> cat("Type II error rate:", type2, "\n")
Type II error rate: 0.9902827
>
> # Type I error rate: 0.21% (509 false positives out of 242,760 actually healthy).
> # Type II error rate: 99.0% (73,782 false negatives out of 74,506 actually poor mental health
> # The model is very conservative - it rarely predicts poor mental health, leading to many mis
cases but few false alarms.
>
> ## Question 5
>
> # Logit model
> logit_model <- glm(MentalHealth_01 ~ Age + Gender + Education + income_midpoint +
+                      Gender:Education,
+                    data = subsample,
+                    family = binomial(link = "logit"))
> summary(logit_model)

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

Coefficients:
                                   Estimate Std. Error z value Pr(>|z|)
(Intercept)                       3.016e-01  2.076e-02  14.528   <2e-16 ***
Age                              -2.032e-02  3.974e-04 -51.132   <2e-16 ***
Genderfemale                      3.139e-01  1.193e-02  26.323   <2e-16 ***
Gendertrans                       1.149e+00  1.009e-01  11.389   <2e-16 ***
Genderother                       1.038e+00  6.571e-02  15.791   <2e-16 ***
Educationadv degree              -2.791e-02  1.464e-02  -1.906   0.0566 .
income_midpoint                  -6.491e-06  6.911e-08 -93.918   <2e-16 ***
Genderfemale:Educationadv degree -1.651e-02  1.818e-02  -0.908   0.3637
Gendertrans:Educationadv degree  -1.465e-03  1.668e-01  -0.009   0.9930
Genderother:Educationadv degree  -2.369e-02  1.001e-01  -0.237   0.8129
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 345855  on 317265  degrees of freedom
Residual deviance: 329644  on 317256  degrees of freedom
  (61483 observations deleted due to missingness)
AIC: 329664

Number of Fisher Scoring iterations: 4

>
> # Q5a:
> # I use the same predictors as Q4 (Age, Gender, Education, income, Gender×Education interacti
> # Exogeneity concerns remain - income and mental health may have reverse causality, and unobs
factors affect both education and mental health.
```

```
> # However, logit is more appropriate than OLS for binary outcomes because it constrains predi
probabilities to [0,1] and models the log-odds rather than assuming a linear probability.
>
> # Q5b:
> # Results are plausible: older age reduces poor mental health risk (coef=-0.020).
> # Females have higher risk (+0.314), higher income protective (-0.0000065).
> # Most coefficients are highly significant (p<0.001).
> # Unlike OLS, the female × education interaction is not significant in logit (p=0.364).
>
> # Q5c:
> # Joint test for ALL education-related terms in logit
> linearHypothesis(logit_model, c("Educationadv degree = 0",
+                                 "Genderfemale:Educationadv degree = 0",
+                                 "Gendertrans:Educationadv degree = 0",
+                                 "Genderother:Educationadv degree = 0"))

Linear hypothesis test:
Educationadv degree = 0
Genderfemale:Educationadv degree = 0
Gendertrans:Educationadv degree = 0
Genderother:Educationadv degree = 0

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

  Res.Df Df  Chisq Pr(>Chisq)
1 317260
2 317256  4 20.041  0.0004902 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
> # Joint test for all education-related terms (main effect + all interactions):
> # χ²(4) = 20.04, p = 0.00049.
> # Education effects ARE statistically significant in the logit model when considering
> # both the main effect and interactions with gender. This differs from testing the
> # main effect alone (p=0.057), showing the importance of accounting for how education's
> # effect varies by gender.
>
> # Q5d:
> pred_data_logit <- data.frame(
+  Age = c(35, 35, 55, 55),
+  Gender = c("male", "female", "male", "female"),
+  Education = c("college grad", "college grad", "adv degree", "adv degree"),
+  income_midpoint = c(75000, 75000, 150000, 150000)
+ )
> pred_data_logit$predicted_prob <- predict(logit_model, newdata = pred_data_logit, type =
"response")
> pred_data_logit
  Age Gender    Education income_midpoint predicted_prob
1  35   male college grad           75000      0.2897854
```

```
2  35 female college grad           75000      0.3583573
3  55   male   adv degree          150000      0.1397243
4  55 female   adv degree          150000      0.1794371
>
> # Predicted probabilities (logit):
> # 35-year-old male college grad at $75k: 29.0%;
> # same female: 35.8%;
> # 55-year-old male with advanced degree at $150k: 14.0%;
> # same female: 17.9%.
> # Very similar to OLS predictions, showing older age and higher income reduce risk.
>
> # Q5e:
> logit_preds <- predict(logit_model, newdata = subsample, type = "response")
>
> subsample <- subsample %>%
+  mutate(predicted_class_logit = ifelse(!is.na(MentalHealth_01),
+                                        ifelse(logit_preds > 0.5, 1, 0),
+                                        NA))
>
> confusion_logit <- table(Actual = subsample$MentalHealth_01, Predicted =
subsample$predicted_class_logit)
> confusion_logit
      Predicted
Actual      0      1
     0 242063    697
     1  73600    906
>
> type1_logit <- confusion_logit[1, 2] / sum(confusion_logit[1, ])
> type2_logit <- confusion_logit[2, 1] / sum(confusion_logit[2, ])
>
> cat("Logit Type I error:", type1_logit, "\n")
Logit Type I error: 0.002871148
> cat("Logit Type II error:", type2_logit, "\n")
Logit Type II error: 0.9878399
>
> # Logit Type I error: 0.29% (697 false positives).
> # Type II error: 98.8% (73,600 false negatives).
> # Very similar to OLS errors - both models are extremely conservative,
> # rarely predicting poor mental health, leading to many missed cases.
>
> # Q5f:
> # Logit and OLS produce very similar predictions and error rates.
> # Logit Type I: 0.29% vs OLS: 0.21%; Type II: 98.8% vs OLS: 99.0%.
> # Both models are extremely conservative, rarely predicting poor mental health.
> # Logit predicted probabilities are similar to OLS (e.g., 35yo female: 35.8% vs 35.1%).
> # Logit is theoretically superior for binary outcomes as it constrains probabilities to [0,1]
> # AIC for logit: 329,664 (lower AIC indicates better fit when comparing models).
>
> ## Question 6 - Probit Model
>
> probit_model <- glm(MentalHealth_01 ~ Age + Gender + Education + income_midpoint +
```

```
+                       Gender:Education,
+                     data = subsample,
+                     family = binomial(link = "probit"))
> summary(probit_model)

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

Coefficients:
                                   Estimate Std. Error z value Pr(>|z|)
(Intercept)                       1.484e-01  1.221e-02  12.151   <2e-16 ***
Age                              -1.201e-02  2.321e-04 -51.736   <2e-16 ***
Genderfemale                      1.851e-01  6.944e-03  26.650   <2e-16 ***
Gendertrans                       7.077e-01  6.228e-02  11.363   <2e-16 ***
Genderother                       6.327e-01  4.055e-02  15.601   <2e-16 ***
Educationadv degree              -1.278e-02  8.283e-03  -1.543    0.123
income_midpoint                  -3.718e-06  3.936e-08 -94.463   <2e-16 ***
Genderfemale:Educationadv degree -1.524e-02  1.044e-02  -1.460    0.144
Gendertrans:Educationadv degree  -1.059e-02  1.030e-01  -0.103    0.918
Genderother:Educationadv degree  -2.705e-02  6.160e-02  -0.439    0.661
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 345855  on 317265  degrees of freedom
Residual deviance: 329698  on 317256  degrees of freedom
  (61483 observations deleted due to missingness)
AIC: 329718

Number of Fisher Scoring iterations: 4

>
> # Predictions with probit
> pred_data_probit <- data.frame(
+  Age = c(35, 35, 55, 55),
+  Gender = c("male", "female", "male", "female"),
+  Education = c("college grad", "college grad", "adv degree", "adv degree"),
+  income_midpoint = c(75000, 75000, 150000, 150000)
+ )
> pred_data_probit$predicted_prob <- predict(probit_model, newdata = pred_data_probit, type =
"response")
> pred_data_probit
  Age Gender    Education income_midpoint predicted_prob
1  35   male college grad           75000      0.2909204
2  35 female college grad           75000      0.3573144
3  55   male   adv degree          150000      0.1395212
4  55 female   adv degree          150000      0.1807091
>
> # Probit confusion matrix
```

```
> probit_preds <- predict(probit_model, newdata = subsample, type = "response")
>
> subsample <- subsample %>%
+  mutate(predicted_class_probit = ifelse(!is.na(MentalHealth_01),
+                                        ifelse(probit_preds > 0.5, 1, 0),
+                                        NA))
>
> confusion_probit <- table(Actual = subsample$MentalHealth_01, Predicted =
subsample$predicted_class_probit)
> confusion_probit
      Predicted
Actual      0      1
     0 242234    526
     1  73756    750
>
> type1_probit <- confusion_probit[1, 2] / sum(confusion_probit[1, ])
> type2_probit <- confusion_probit[2, 1] / sum(confusion_probit[2, ])
>
> cat("Probit Type I error:", type1_probit, "\n")
Probit Type I error: 0.002166749
> cat("Probit Type II error:", type2_probit, "\n")
Probit Type II error: 0.9899337
>
> # Compare AICs
> cat("\nModel Comparison (AIC):\n")

Model Comparison (AIC):
> cat("OLS: N/A (not comparable)\n")
OLS: N/A (not comparable)
> cat("Logit AIC:", AIC(logit_model), "\n")
Logit AIC: 329664.5
> cat("Probit AIC:", AIC(probit_model), "\n")
Probit AIC: 329717.9
>
> # VIZ: Compare predicted probabilities across models
> pred_comparison <- data.frame(
+  Age = c(35, 35, 55, 55),
+  Gender = c("male", "female", "male", "female"),
+  Education = rep("college grad", 4),
+  income_midpoint = c(75000, 75000, 150000, 150000)
+ )
>
> # Get predictions from all three models
> pred_comparison$OLS <- predict(ols_binary, newdata = pred_comparison)
> pred_comparison$Logit <- predict(logit_model, newdata = pred_comparison, type = "response")
> pred_comparison$Probit <- predict(probit_model, newdata = pred_comparison, type = "response")
>
> # Reshape for plotting
> comparison_long <- pred_comparison %>%
+  mutate(Scenario = paste0(Age, "yo ", Gender, "\n$", income_midpoint/1000, "k")) %>%
+  select(Scenario, OLS, Logit, Probit) %>%
```

```
+  pivot_longer(cols = c(OLS, Logit, Probit),
+               names_to = "Model", values_to = "Probability")
>
> # Plot
> ggplot(comparison_long, aes(x = Scenario, y = Probability, fill = Model)) +
+  geom_bar(stat = "identity", position = "dodge") +
+  scale_fill_manual(values = c("OLS" = "#3498db",      # blue
+                                "Logit" = "#e74c3c",    # red
+                                "Probit" = "#2ecc71")) + # green
+  labs(title = "Model Predictions: Probability of Poor Mental Health",
+       subtitle = "All scenarios: College graduates, varying age/income/gender",
+       y = "Predicted Probability",
+       x = "") +
+  theme_minimal() +
+  theme(legend.position = "bottom")
>
> ## Question 6 Summary:
> # I estimated a probit model as an alternative to logit for binary outcomes.
> # Probit uses a normal CDF link function instead of logistic.
> #
> # Results are very similar across all three models:
> # - Predicted probabilities nearly identical (e.g., 35yo female: OLS 35.1%, Logit 35.8%, Prob
35.7%)
> # - Error rates comparable: Probit Type I: 0.22%, Type II: 99.0%
> # - All three models are extremely conservative, rarely predicting poor mental health
> #
> # Model comparison:
> # - Logit AIC: 329,664 (slightly better)
> # - Probit AIC: 329,718
> # - Lower AIC indicates logit fits marginally better
> #
> # Strengths: Logit/Probit constrain probabilities to [0,1], theoretically appropriate for bin
outcomes
> # Weaknesses: All models have very high Type II error rates (miss 99% of poor mental health
cases),
> # suggesting we need better predictors or different threshold than 0.5 for classification.




 >
```