

Irania Matos

## Exam 2

### Question 1:

For the White group (D=0):

$$\begin{aligned}\beta_0 &= \gamma_0 \\ \beta_1 &= \gamma_1\end{aligned}$$

Non-White group (D=1):

$$\begin{aligned}\alpha_0 &= \gamma_0 + \gamma_2 \\ \alpha_1 &= \gamma_1 + \gamma_3\end{aligned}$$

- $\gamma_2$  measures how the intercept differs between NW and W.  
If  $\gamma_2 > 0$ , NW households start at a higher predicted probability.
- $\gamma_3$  measures how the slope with respect to Age differs between NW and W.  
If  $\gamma_3 > 0$ , the effect of Age on the age gap is stronger for NW households.
- The combined model exactly reproduces the two separate regression lines.

### Questions 2 :

#### **Education**

F = 213.51, p-value < 2.2e-16df = 6 restrictions

Because the p-value is effectively zero, we reject the null hypothesis that all education categories have no effect. Education is a highly significant predictor of mental health (K4SUM), even after controlling for income, age, gender, and race.

## Joint Hypothesis Test: Income

Output:

$F = 4771.3$   $p\text{-value} < 2.2e-16$   $df = 7$  restrictions

This p-value is also effectively zero, so we strongly reject the null that income has no effect on the K4SUM mental-health score. Income is an extremely strong predictor of mental health.

## Compare Education vs Income

Education:

- $F = 213.5$
- Very significant

Income:

- $F = 4771$
- Even more significant

Both education and income are jointly statistically significant predictors of K4SUM mental health scores. The income coefficients have a much larger F-statistic, indicating that income explains more variation in mental health than education in this model. Income appears to be the stronger predictor of mental health in the Household Pulse Survey data.

## Question 3:

For my analysis I picked prime working-age adults, defined as individuals ages 25 to 54. Restricting to this group reduces heterogeneity from very young adults (students) and older individuals (retirees), making education and income more comparable across respondents.

On average, adults in this age group score 7.49 on the Kessler-4 mental distress scale, which ranges from 4 (no symptoms) to 16 (symptoms nearly every day). This suggests moderate and widespread mental-health strain across the working-age population.

Among adults ages 25 to 54, mental-health distress (K4SUM) averages about 7.5, indicating moderate symptom levels. The group is highly educated, with more than half of respondents having at least some college education. Income is widely dispersed, with a median of \$82,500..

Question 4:

- a) I include four main predictors:

Age is included as a continuous predictor because mental-health symptoms often vary across the life cycle. Age is plausibly exogenous, since short-run mental-health symptoms cannot change a person's age.

Education is included as a categorical variable because it reflects long-term socioeconomic status and coping resources. Education is mostly determined earlier in life, so it is largely exogenous to current mental-health episodes (though long-term illness may affect schooling, creating mild endogeneity).

Income is included because financial stress is strongly correlated with mental health. Income is not fully exogenous, since mental-health problems can affect labor-market outcomes (reverse causality). However, income is a central socioeconomic variable and must be included.

Gender differences in mental-health reporting are well-documented, so gender is included as an important demographic predictor. Gender is exogenous to the mental-health outcome being measured.

- b) The estimates from the OLS model appear highly plausible. The signs and magnitudes of coefficients match well-known relationships in mental-health research: distress declines with age, declines with higher education and income, and is higher among women and gender minorities.
- c) The F-test results were:  $F = 1869$   $p\text{-value} < 2.2e-16$  Degrees of freedom = 7 restrictions. The null hypothesis states that income has no effect on the probability of high psychological distress. The test strongly rejects the null hypothesis. Income is jointly statistically significant. At least one (and likely many) income categories differ from zero. Income is a very important predictor of mental-health distress, even after controlling for education, age, and gender.

- d) Low-income, high-school graduate female (age 35)

Predicted probability: 0.451

This means she has about a 45% chance of reporting high mental-health distress. Low income and lower education both increase the predicted likelihood of distress.

2. High-income, college-graduate male (age 40).

Predicted probability: 0.203

This means he has about a 20% chance of high distress.

Higher education and income substantially reduce the predicted probability.

e) FALSE 271185 10906 TRUE 10448 11909

Type I Error (False Positive)-The model predicted MentalHealth\_01 = 1 (high distress) but the person actually had 0.

Type II Error (False Negative)- The model predicted 0 (no high distress), but the person actually had 1.

### Question 5 : Logit Model

- a. I picked age because mental-health symptoms often vary across the life cycle. Age is exogenous because psychological distress cannot influence a person's chronological age. I also picked Race because racial groups differ in stress exposure, discrimination, and mental-health access. Race is also an exogenous demographic characteristic.
- b. The logit estimates are plausible and statistically significant. Age has a negative sign and very large z-statistic, showing that younger adults are more likely to experience psychological distress. All race coefficients are significant as well: Black and "other" race categories show higher predicted distress than White respondents, while Asians show lower distress. These patterns match well-known differences in mental-health reporting and demographics, confirming the model's plausibility.
- c. The null hypothesis is that race has no effect on the probability of high psychological distress. The result of the test was **Chi-square = 1749.9** with a **p-value < 2.2e-16**. Because the p-value is effectively zero, I strongly reject the null hypothesis. Race is jointly statistically significant in the logit model. At least one race category has a meaningful effect on mental-health distress after controlling for age.
- d. The Black 30-year-old has a slightly higher predicted probability of high mental-health distress(0.3596705) compared to an otherwise similar White 30-year-old(0.3486628). This difference reflects the positive coefficient on RaceBlack in the logit model.
- e. true  
pred FALSE TRUE

FALSE 309615 131282

The logit model makes 0 Type I errors and 131,282 Type II errors. This means the model never incorrectly predicts high distress when the person is actually low distress, but it fails to identify every single true case of high distress.

- f. The logit model produces similar coefficient signs to OLS, showing that younger adults and certain racial groups are more likely to experience distress. However, the prediction performance differs: the logit model predicts almost no TRUE cases when using the 0.5 cutoff, leading to many Type II errors. OLS produces more balanced predictions and identifies more TRUE cases, although at the cost of more Type I errors. Overall, both models agree on the direction of effects, but OLS performs better in catching true cases of high mental distress.

#### Question 6:

For my additional model, I estimated a random forest classifier using Age and Race as predictors. Random forests are non-parametric models that automatically capture nonlinearities and interactions without needing to specify them manually.

Compared with the logit model, the random forest showed different strengths. It predicted more TRUE cases because random forests can adjust to complex patterns, whereas the logit model using Age and Race rarely produced fitted probabilities above 0.5.

A key weakness of the random forest is that it is less interpretable than logit—there are no simple coefficients. However, its predictive performance is typically stronger, especially when relationships are nonlinear.

Overall, the random forest provides better classification performance than the simple logit model, but the logit model remains more interpretable and easier to explain.

true

pred FALSE TRUE

FALSE 309615 131282

TRUE 0 0

