Group: Ege Yanik (I was alone)

Date: 11.14.2025

Rstudio Codes

```
# ============================================================

# WORK-FROM-HOME LOGIT & PROBIT MODELS —

# ============================================================


# install once if needed:

# install.packages(c("tidyverse","broom","margins","pROC"))


library(tidyverse)

library(broom)

library(margins)

library(pROC)


# ============================================================

# 1. Data prep

# ============================================================

data <- acs2021_couples %>%

  filter(!is.na(WFH)) %>%

  mutate(

    female  = ifelse(SEX == "Female", 1, 0),

    college = as.numeric(educ_college),

    advdeg  = as.numeric(educ_advdeg),

    married = as.numeric(Married),

    white   = as.numeric(white),

    asian   = as.numeric(Asian),

    age2    = AGE^2
```

```
  )

# ============================================================
# 2. Logit & Probit models
# ============================================================
logit_wfh <- glm(WFH ~ female + AGE + age2 + college + advdeg + married + white + asian,
          data = data, family = binomial(link = "logit"))
probit_wfh <- glm(WFH ~ female + AGE + age2 + college + advdeg + married + white + asian,
          data = data, family = binomial(link = "probit"))

summary(logit_wfh)
summary(probit_wfh)

# ============================================================
# 3. Marginal effects
# ============================================================
mfx_logit  <- margins(logit_wfh)
mfx_probit <- margins(probit_wfh)
summary(mfx_logit)
summary(mfx_probit)

# ============================================================
# 4. Coefficient (odds ratio) plot
# ============================================================
tidy(logit_wfh, conf.int = TRUE, exponentiate = TRUE) %>%
  filter(term != "(Intercept)") %>%
  mutate(term = fct_reorder(term, estimate)) %>%
  ggplot(aes(x = term, y = estimate, ymin = conf.low, ymax = conf.high)) +
  geom_pointrange() +
```

```
  coord_flip() +
  labs(title = "Logit Odds Ratios with 95% CIs (WFH ~ X)",
       x = "", y = "Odds ratio (exp(beta))")

# ============================================================
# 5. Average marginal effects plot
# ============================================================
ame_logit  <- margins(logit_wfh) %>% summary() %>% as_tibble()
ame_probit <- margins(probit_wfh) %>% summary() %>% as_tibble()
ame_logit$model  <- "Logit"
ame_probit$model <- "Probit"
bind_rows(ame_logit, ame_probit) %>%
  filter(factor != "(Intercept)") %>%
  mutate(factor = fct_reorder(factor, AME)) %>%
  ggplot(aes(factor, AME, fill = model)) +
  geom_col(position = position_dodge(width = 0.6)) +
  geom_hline(yintercept = 0, linetype = "dashed") +
  coord_flip() +
  labs(title = "Average Marginal Effects on P(WFH=1)",
       x = "", y = "AME (percentage-point change)")

# ============================================================
# 6. Predicted probability by age & education
# ============================================================
age_grid <- tibble(AGE = seq(min(data$AGE, na.rm=TRUE),
                             max(data$AGE, na.rm=TRUE), by=1),
                   age2 = AGE^2)

newdat <- expand_grid(
```

```
  age_grid,
  college = c(0,1),
  advdeg  = c(0,1)
) %>%
  filter(!(advdeg == 1 & college == 0)) %>%
  mutate(female=0, married=0, white=1, asian=0) %>%
  mutate(pred = predict(logit_wfh, newdata = ., type = "response"),
         edu_label = case_when(
           college==0 & advdeg==0 ~ "No College",
           college==1 & advdeg==0 ~ "College",
           advdeg==1              ~ "Advanced Degree"
         ))

ggplot(newdat, aes(x = AGE, y = pred, color = edu_label)) +
  geom_line(size = 1) +
  labs(title = "Predicted Probability of WFH by Age & Education (Logit)",
       x = "Age", y = "Predicted P(WFH=1)", color = "Education")

# ============================================================
# 7. Actual vs predicted by education
# ============================================================
data %>%
  mutate(pred_logit = predict(logit_wfh, type="response"),
         edu_grp = case_when(advdeg==1 ~ "Advanced",
                             college==1 ~ "College",
                             TRUE ~ "No College")) %>%
  group_by(edu_grp) %>%
  summarise(actual = mean(WFH, na.rm=TRUE),
            predicted = mean(pred_logit, na.rm=TRUE)) %>%
```

```
  pivot_longer(cols = c(actual, predicted), names_to="type", values_to="rate") %>%
  ggplot(aes(x = edu_grp, y = rate, fill = type)) +
  geom_col(position="dodge") +
  scale_y_continuous(labels=scales::percent_format(accuracy=1)) +
  labs(title="Actual vs Predicted WFH by Education", x="", y="Rate")

# ============================================================
# 8. Confusion matrices
# ============================================================
data <- data %>%
  mutate(p_logit=predict(logit_wfh,type="response"),
         p_probit=predict(probit_wfh,type="response"),
         yhat_logit=if_else(p_logit>0.5,1,0),
         yhat_probit=if_else(p_probit>0.5,1,0))

table_logit <- table(True=data$WFH, Pred=data$yhat_logit) %>% as.data.frame()
ggplot(table_logit, aes(x=Pred, y=True, fill=Freq)) +
  geom_tile() +
  geom_text(aes(label=Freq), color="white", fontface="bold") +
  scale_x_continuous(breaks=c(0,1)) +
  scale_y_continuous(breaks=c(0,1)) +
  labs(title="Confusion Matrix (Logit)", x="Predicted", y="Actual")

# ============================================================
# 9. ROC curves (Logit vs Probit)
# ============================================================
roc_logit  <- roc(response=data$WFH, predictor=data$p_logit)
roc_probit <- roc(response=data$WFH, predictor=data$p_probit)
plot(roc_logit, legacy.axes=TRUE, main="ROC Curves: Logit vs Probit")
```

```
plot(roc_probit, add=TRUE, lty=2)

legend("bottomright",

    legend=c(sprintf("Logit AUC = %.3f", auc(roc_logit)),

             sprintf("Probit AUC = %.3f", auc(roc_probit))),

    lty=c(1,2), bty="n")

# ============================================================
```

Outputs

```
Coefficients:
             Estimate Std. Error  z value Pr(>|z|)
(Intercept) -5.152e+00  4.398e-02 -117.123  < 2e-16 ***
female      -1.966e-01  7.653e-03  -25.688  < 2e-16 ***
AGE          1.386e-01  1.912e-03   72.475  < 2e-16 ***
age2        -1.680e-03  1.945e-05  -86.365  < 2e-16 ***
college      1.268e+00  8.868e-03  142.923  < 2e-16 ***
advdeg       1.442e+00  9.860e-03  146.232  < 2e-16 ***
married     -7.780e-02  1.100e-02   -7.074  1.5e-12 ***
white        2.283e-01  1.018e-02   22.424  < 2e-16 ***
asian        3.487e-01  1.520e-02   22.947  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 541306  on 776063  degrees of freedom
Residual deviance: 481788  on 776055  degrees of freedom
AIC: 481806

Number of Fisher Scoring iterations: 6
```

Interpretation

For this lab, I used the same ACS couples dataset but changed the topic to analyze which demographic and socioeconomic factors predict working from home. I estimated both logit and probit models with WFH as the dependent variable and variables such as gender, age, education, marital status, and race as predictors.

The results from both models were very consistent. Education turned out to be the strongest predictor college graduates and especially those with advanced degrees are significantly more likely to work from home. The relationship between age and remote work is nonlinear: the probability of working from home increases up to around age 40 and then gradually declines. Race also matters, with White and Asian individuals showing a higher likelihood of remote work. Gender and marital status have smaller

but statistically significant effects; women and married individuals are slightly less likely to work from home when holding other factors constant.

Both models fit the data well (AIC ≈ 482k), and the ROC curves show very similar predictive accuracy for logit and probit. The visualizations confirm that education has the largest marginal effect, while age exhibits an inverted-U pattern. Overall, this analysis shows how individual characteristics are correlated with remote-work probability, though education and marital status may not be fully exogenous since they are related to occupational selection.

The analysis shows that age is a major driver of whether the man is significantly older, while education affects both the baseline probability and the slope of this relationship. The equivalence between the interaction model and the subgroup models illustrates a key econometric concept.