Keyla Pereira
- Lab 8 + Addendum
Nancy Reyes Soto, Irania Matos, Kerra Sinanan, Heidy Collado.

Results:

LPM

	TRUE	
PRED	0	1
FALSE	203,909	9,306
TRUE	182,021	17,039

Logit

	TRUE	
PRED	0	1
FALSE	235,290	11,352
TRUE	150,640	14,993

Standardized LPM

	TRUE	
PRED	0	1
FALSE	180,978	8,244
TRUE	166,340	15,442

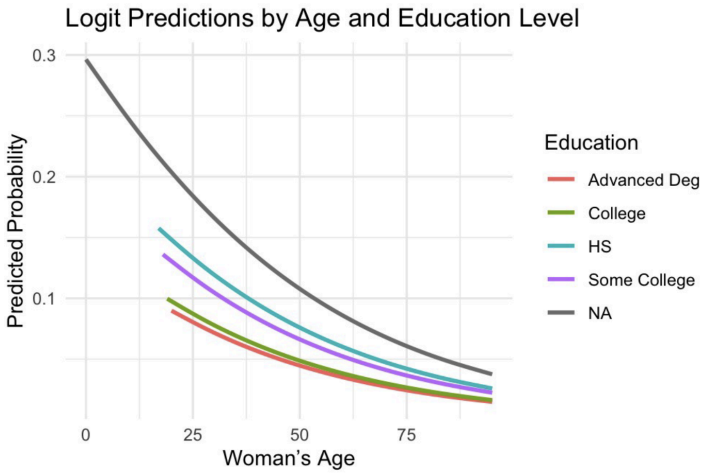
Standardized Logit

	TRUE	
PRED	0	1
FALSE	213,252	10,308
TRUE	134,066	13,378

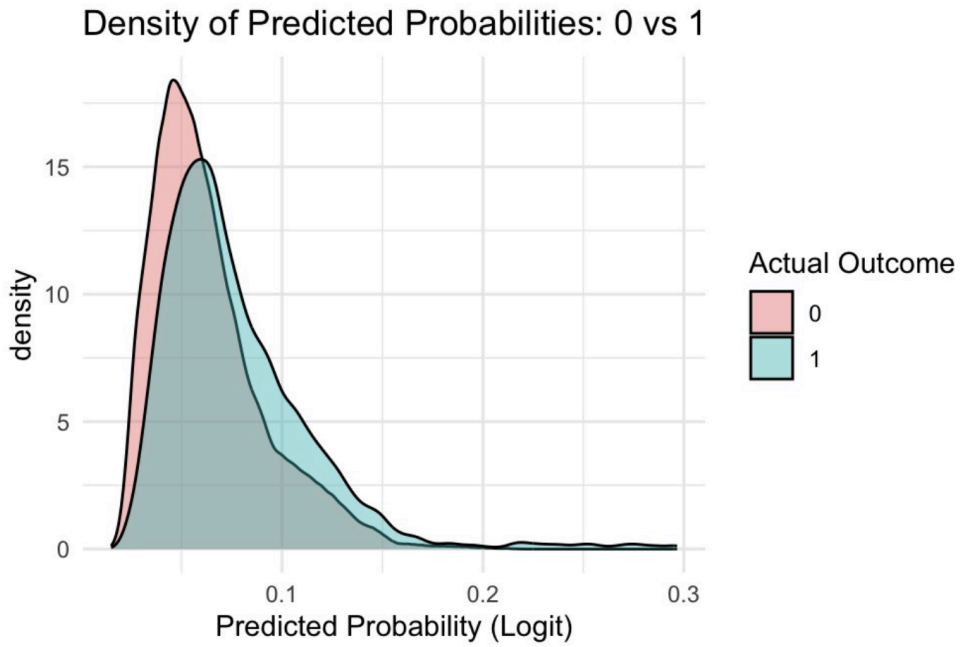
In the in-sample comparison, both the linear probability model and the logit model show the same broad prediction pattern: predicting a “0”is much easier than predicting a “1” man ≥10 years older.

The LPM predicts 17,039 true positives (cases where the man is ≥10 years older) and 9,306 false negatives. The logit model predicts slightly fewer true positives, 14,993, and more false negatives 11,352, but it compensates by producing fewer false positives.

This shows the trade-off between false positives and false negatives. The logit model is more conservative it predicts “1” less often. As a result, it lowers the false positive rate but increases false negatives. Both models agree that the outcome is rare, and both rely heavily on age and education as strong predictors.



There is a clear monotonic pattern: as education increases, the predicted probability of a large age gap declines. Age and education jointly shape the estimated probability, with younger and less educated women having the highest predicted values.



The densities overlap heavily, showing why prediction is difficult; the logit model does not sharply separate couples with large vs small age gaps.

```
library(ggplot2)
library(tidyverse)
library(haven)
library(standardize)


load("/Users/keylapereira/Downloads/ACS_2021_couples.RData")

ols_out1 <- lm(he_more_than_10yrs_than_her ~ educ_hs + educ_somecoll + educ_college + educ_advdeg + AGE,
data = trad_data)

pred_vals_ols1 <- predict(ols_out1, trad_data)
pred_model_ols1 <- (pred_vals_ols1 > mean(pred_vals_ols1))
table(pred = pred_model_ols1, true = trad_data$he_more_than_10yrs_than_her)

model_logit1 <- glm(he_more_than_10yrs_than_her ~ educ_hs + educ_somecoll + educ_college + educ_advdeg +
AGE, data = trad_data, family = binomial)
summary(model_logit1)

pred_vals <- predict(model_logit1, trad_data, type = "response")
pred_model_logit1 <- (pred_vals > mean(pred_vals))
table(pred = pred_model_logit1, true = trad_data$he_more_than_10yrs_than_her)

#Lab 8 Addendum

dat_use <- trad_data


d_y_varb <- data.frame(model.matrix(~ dat_use$he_more_than_10yrs_than_her))

d_educ_hs       <- data.frame(model.matrix(~ dat_use$educ_hs))
d_educ_somecoll <- data.frame(model.matrix(~ dat_use$educ_somecoll))
d_educ_college  <- data.frame(model.matrix(~ dat_use$educ_college))
d_educ_advdeg   <- data.frame(model.matrix(~ dat_use$educ_advdeg))
d_age           <- data.frame(model.matrix(~ dat_use$AGE))


dat_for_analysis_sub <- data.frame(
  d_y_varb[,2],
  d_educ_hs[,2],
  d_educ_somecoll[,2],
  d_educ_college[,2],
  d_educ_advdeg[,2],
  d_age[,2]
)


names(dat_for_analysis_sub) <- c(
  "he_more_than_10yrs_than_her",
  "HS", "SomeColl", "College", "AdvDeg",
  "Age"
)


set.seed(654321)
NN <- nrow(dat_for_analysis_sub)

restrict_1 <- (runif(NN) < 0.10)
summary(restrict_1)

dat_train <- subset(dat_for_analysis_sub, restrict_1)
dat_test  <- subset(dat_for_analysis_sub, !restrict_1)
```

```
sum(colSums(dat_train) == 0)



formula_sobj <- reformulate(
  names(dat_for_analysis_sub)[2:length(dat_for_analysis_sub)],
  response = "he_more_than_10yrs_than_her"
)

sobj <- standardize(formula_sobj, dat_train, family = binomial)
s_dat_test <- predict(sobj, dat_test)



model_lpm1 <- lm(sobj$formula, data = sobj$data)
summary(model_lpm1)

pred_vals_lpm <- predict(model_lpm1, s_dat_test)
pred_model_lpm1 <- (pred_vals_lpm > mean(pred_vals_lpm))

table(pred = pred_model_lpm1, true = dat_test$he_more_than_10yrs_than_her)

#Logit
model_logit1 <- glm(sobj$formula, family = binomial, data = sobj$data)
summary(model_logit1)

pred_vals_logit <- predict(model_logit1, s_dat_test, type = "response")
pred_model_logit1 <- (pred_vals_logit > mean(pred_vals_logit))

table(pred = pred_model_logit1, true = dat_test$he_more_than_10yrs_than_her)

#Graph 1

trad_data$educ_group <- case_when(
  trad_data$educ_hs == 1 ~ "HS",
  trad_data$educ_somecoll == 1 ~ "Some College",
  trad_data$educ_college == 1 ~ "College",
  trad_data$educ_advdeg == 1 ~ "Advanced Deg"
)

ggplot(trad_data, aes(x = AGE, y = pred_prob_logit, color = educ_group)) +
  geom_smooth(se = FALSE) +
  labs(
    x = "Woman's Age",
    y = "Predicted Probability",
    color = "Education",
    title = "Logit Predictions by Age and Education Level"
  ) +
  theme_minimal()


#Graph 2

ggplot(trad_data, aes(x = pred_prob_logit, fill = factor(he_more_than_10yrs_than_her))) +
  geom_density(alpha = 0.4) +
  labs(
    x = "Predicted Probability (Logit)",
    fill = "Actual Outcome",
    title = "Density of Predicted Probabilities: 0 vs 1"
  ) +
  theme_minimal()
```

Working Twice as Hard for Less Than Half as Much: A Sociolegal Critique of the Gendered Justifications Perpetuating Unequal Pay in Sports (by Shannon Morgan, 2021)

This article talks about why female athletes in different sports get paid much less than men, even when their performance is similar. Instead of just blaming the pay gap on "men's sports make more money," the author argues that deeper issues like gender bias and long-standing stereotypes also play a big role. The paper uses ideas from feminist theory and critical race theory to explain how women's sports have been undervalued for years. It looks at examples from basketball, soccer, and tennis to show that the gap isn't only about revenue. It's about how society views and supports women athletes.

Equal play, equal pay": moral grounds for equal pay in football's gender pay gap by A. Archer

The article argues that women's and men's national soccer teams should receive the same pay, not just because of performance or revenue, but because of fairness and history. The authors explain that even though some people justify the pay gap by saying men bring in more money, this misses the bigger picture. Women's soccer was held back for decades through lack of support and even official bans, so paying women less today continues that unfair treatment. The paper presents three arguments: equal pay as a basic labor right, equal pay as a message that women athletes are valued, and equal pay as a way to correct past discrimination. Overall, the article says national sports organizations have moral reasons—not just financial ones—to pay women and men equally.