Bamba Cisse

Homework8_B2000

```
library(dplyr)
library(ggplot2)
library(broom)
library(forcats)
library(scales)
trad_data$pred_prob <- predict(model_logit1, type = "response")
```

1

```
# A tidy helper for weighted/binomial CIs by group
rate_ci <- function(df, group_var) {
  df %>%
    group_by({{ group_var }}) %>%
    summarise(
      n = n(),
      rate = mean(he_more_than_3yrs_than_her, na.rm = TRUE),
      se = sqrt(rate * (1 - rate) / n),
      ci_low  = pmax(0, rate - 1.96 * se),
      ci_high = pmin(1, rate + 1.96 * se),
      .groups = "drop")
}



coef_plot <- tidy(model_logit1, conf.int = TRUE, exponentiate = TRUE) %>%
  filter(term != "(Intercept)") %>%                    # drop intercept
  mutate(term = fct_reorder(term, estimate))           # order by OR
```

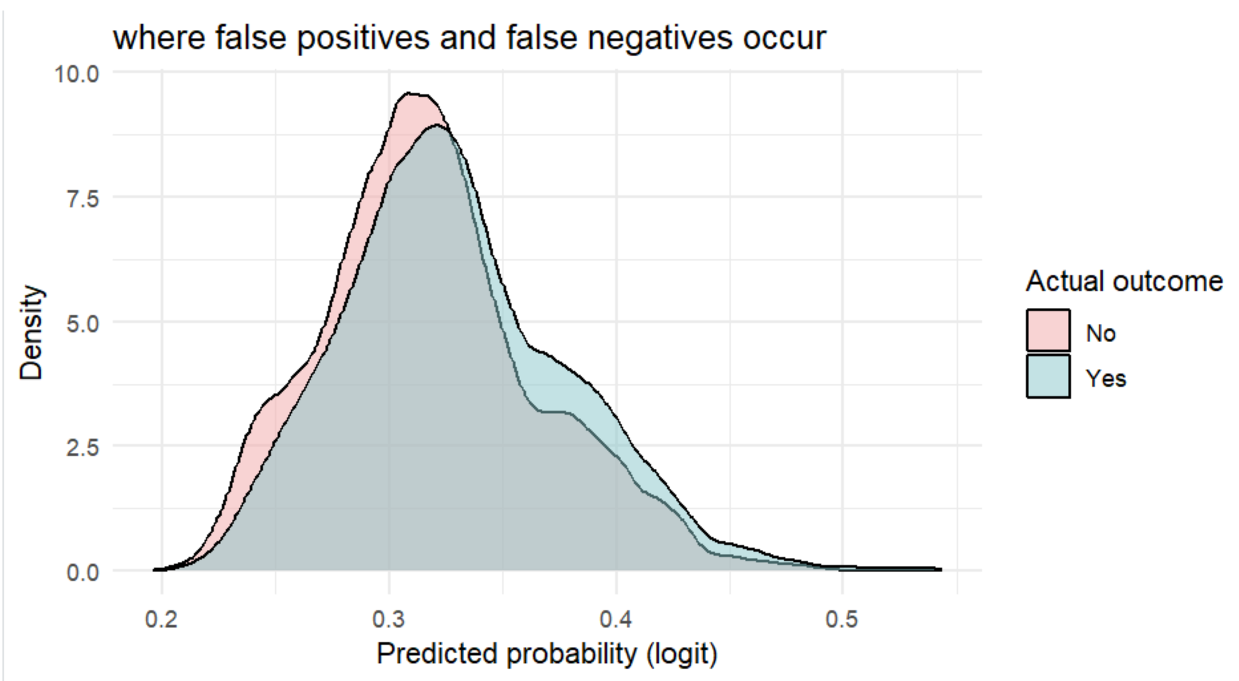
```r
ggplot(coef_plot, aes(x = term, y = estimate)) +
  geom_hline(yintercept = 1, linetype = "dashed") +
  geom_point(size = 2) +
  geom_errorbar(aes(ymin = conf.low, ymax = conf.high), width = 0.15) +
  coord_flip() +
  scale_y_log10(labels = label_number(accuracy = 0.01)) +
  labs(title = "Odds ratios from logit: Pr(He ≥3 years older)",
       x = NULL, y = "Odds ratio (log scale)") +
  theme_minimal()
```

Graph1: Odds ratios from the logistic model

This coefficient plot (odds ratios on a log scale) shows which predictors most strongly relate to the event "he is ≥3 years older." Points above 1 increase the odds; below 1 decrease them; error bars are 95% CIs. The chart makes it easy to see which education indicators and age are statistically significant and the direction/magnitude of their associations after adjustment.

2- Distribution of Predicted Probabilities

```
ggplot(trad_data, aes(x = pred_prob, fill = factor(he_more_than_3yrs_than_her))) +
 geom_density(alpha = 0.5, color = "black") +
 scale_fill_manual(
  name = "Actual outcome",
  values = c("0" = "#F2A7A7", "1" = "#87C6C9"),
  labels = c("0" = "No", "1" = "Yes")
 ) +
 labs(title = "where false positives and false negatives occur",
     x = "Predicted probability (logit)", y = "Density") +
 theme_minimal()
```



Graph2: Distribution of Predicted Probabilities

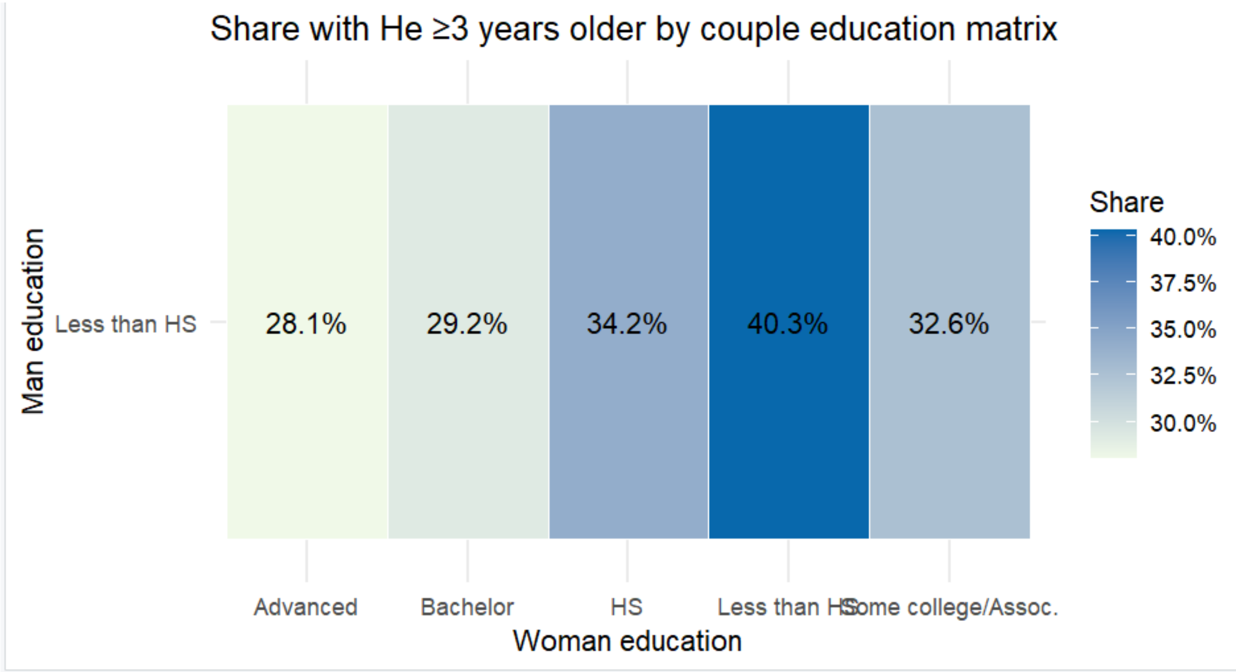
This density plot shows the distribution of predicted probabilities from the logistic model, divided by the actual outcomes ("Yes" for the man being ≥3 years older, "No" otherwise). The overlap between the pink (No) and teal (Yes) areas indicates where the model struggles to separate the two groups, representing potential false positives and false negatives. The clearer the separation between the two distributions, the stronger the model's predictive power.

3

```
# Use your numeric encodings to bin into labels
lab_from_num <- function(x){
  dplyr::case_when(
    x >= 18 ~ "Advanced", x >= 16 ~ "Bachelor",
    x >= 14 ~ "Some college/Assoc.", x >= 12 ~ "HS",
    TRUE ~ "Less than HS"  )
}

heat_df <- trad_data %>%
  mutate(w_edu = lab_from_num(educ_numeric),
         h_edu = lab_from_num(h_educ_numeric)) %>%
  filter(!is.na(w_edu), !is.na(h_edu)) %>%
  group_by(h_edu, w_edu) %>%
  summarise(rate = mean(he_more_than_3yrs_than_her, na.rm = TRUE),
            n = n(), .groups = "drop")

ggplot(heat_df, aes(x = w_edu, y = h_edu, fill = rate)) +
  geom_tile(color = "white") +
  geom_text(aes(label = scales::percent(rate, accuracy = 0.1))) +
  scale_fill_gradient(low = "#f0f9e8", high = "#0868ac", labels = percent) +
  labs(title = "Share with He ≥3 years older by couple education matrix",
       x = "Woman education", y = "Man education", fill = "Share") +
  theme_minimal()
```



Graph3: Education Matching Heatmap

This heatmap visualizes how the likelihood of the man being at least three years older varies across combinations of the couple's education levels. Darker blue shades represent higher probabilities of a ≥3-year age gap. The diagonal cells (where both partners have similar education levels) reveal assortative matching, while off-diagonal areas show how mismatched education pairs relate to larger or smaller age differences. It provides an intuitive picture of how educational pairing interacts with partner age dynamics.