

Assignment 3 – Solutions: Part 1 (ANES Voter Turnout)

Applied Quantitative Methods II, UC3M

1. Setup and data preparation

a) Load the dataset:

```
library(dplyr)
library(broom)
library(ggplot2)
library(modelsummary)
library(marginaleffects)

# Load pre-processed anes.csv from the course page
# (Here we process from raw ANES 2020 to replicate it)
raw = read.csv("https://raw.githubusercontent.com/franvillamil/AQM2/refs/heads/master/datasets/anes/anes_t

df = raw %>%
  transmute(
    voted = ifelse(V202109x < 0, NA, V202109x),
    age = ifelse(V201507x < 0, NA, V201507x),
    female = case_when(V201600 == 2 ~ 1, V201600 == 1 ~ 0, TRUE ~ NA_real_),
    education = case_when(
      V201511x == 1 ~ 10, V201511x == 2 ~ 12, V201511x == 3 ~ 14,
      V201511x == 4 ~ 16, V201511x == 5 ~ 20, TRUE ~ NA_real_),
    income = ifelse(V201617x < 0, NA, V201617x),
    party_id = ifelse(V201231x < 0, NA, V201231x)
  )
```

b) Drop observations with missing values:

```
df = na.omit(df)
nrow(df)

## [1] 6733
```

c) Overall turnout rate and summary statistics:

```
mean(df$voted)

## [1] 0.8609832

summary(df)

##      voted           age         female        education
##  Min.    :0.000   Min.    :18.00   Min.    :0.0000   Min.    :10.00
##  1st Qu.:0.000   1st Qu.:21.00   1st Qu.:0.0000   1st Qu.:12.00
##  Median :0.000   Median :24.00   Median :0.0000   Median :14.00
##  Mean    :0.861   Mean   :25.81   Mean   :0.0000   Mean   :14.81
##  3rd Qu.:1.000   3rd Qu.:27.00   3rd Qu.:1.0000   3rd Qu.:16.00
##  Max.   :1.000   Max.   :82.00   Max.   :1.0000   Max.   :20.00
```

```

## 1st Qu.:1.000 1st Qu.:37.00 1st Qu.:0.0000 1st Qu.:14.00
## Median :1.000 Median :52.00 Median :1.0000 Median :14.00
## Mean    :0.861 Mean    :51.38 Mean    :0.5394 Mean    :15.22
## 3rd Qu.:1.000 3rd Qu.:66.00 3rd Qu.:1.0000 3rd Qu.:16.00
## Max.   :1.000 Max.   :80.00 Max.   :1.0000 Max.   :20.00
##      income      party_id
## Min.   : 1.00  Min.   :1.000
## 1st Qu.: 6.00  1st Qu.:2.000
## Median :12.00  Median :4.000
## Mean   :11.81  Mean   :3.831
## 3rd Qu.:18.00  3rd Qu.:6.000
## Max.   :22.00  Max.   :7.000

```

2. Exploratory visualization

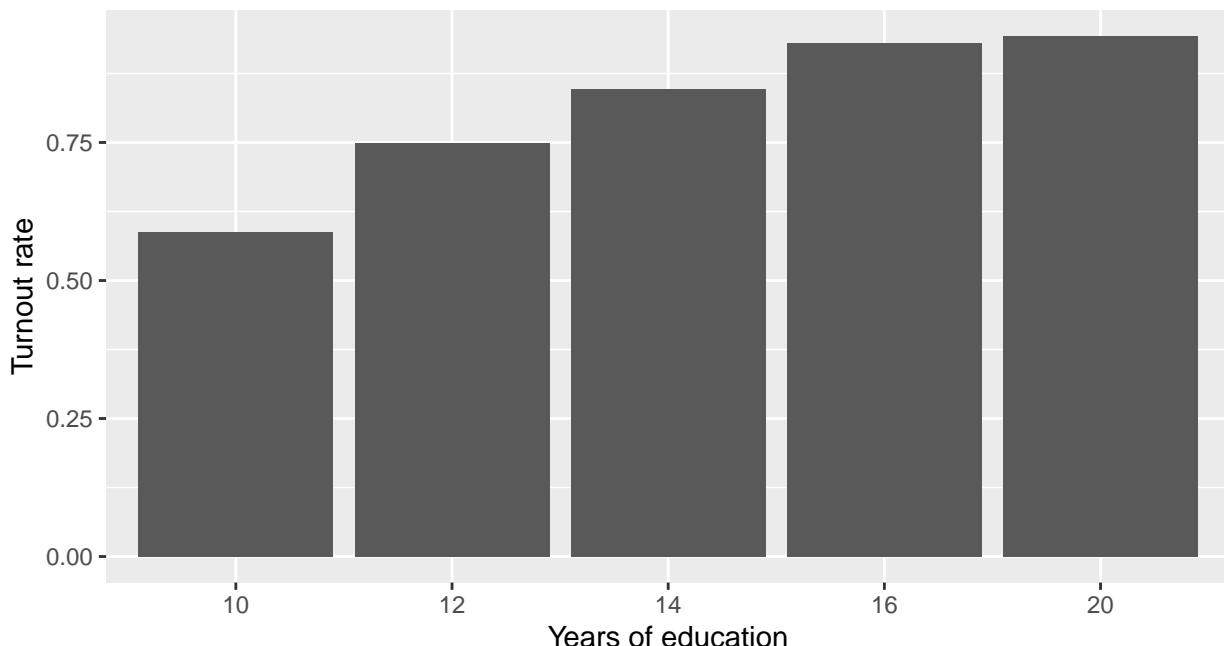
a) Bar chart of turnout by education level:

```

turnout_by_edu = df %>%
  group_by(education) %>%
  summarise(turnout = mean(voted))

ggplot(turnout_by_edu, aes(x = factor(education), y = turnout)) +
  geom_col() +
  labs(x = "Years of education", y = "Turnout rate")

```



b) Turnout increases with education: respondents with more years of education are more likely to report voting.
The pattern is monotonic.

3. Linear probability model

a–b) Estimate the LPM:

```
lpm = lm(voted ~ age + education + income + female, data = df)
tidy(lpm)
```

```
## # A tibble: 5 x 5
##   term      estimate std.error statistic p.value
##   <chr>     <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept) 0.240     0.0252     9.52 2.42e-21
## 2 age        0.00413    0.000233    17.7 1.17e-68
## 3 education  0.0193     0.00152    12.7 1.76e-36
## 4 income      0.00823    0.000641    12.8 2.53e-37
## 5 female      0.0344     0.00803    4.28 1.87e- 5
```

c) The coefficient on education represents the estimated change in the probability of voting for each additional year of education, holding the other variables constant.

d) Check predicted probabilities:

```
preds_lpm = predict(lpm)
sum(preds_lpm < 0)

## [1] 0

sum(preds_lpm > 1)

## [1] 802

range(preds_lpm)

## [1] 0.5150876 1.1708206
```

4. Logistic regression

a–b) Estimate the logit model:

```
logit = glm(voted ~ age + education + income + female,
            family = binomial, data = df)
tidy(logit)
```

```
## # A tibble: 5 x 5
##   term      estimate std.error statistic p.value
##   <chr>     <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept) -4.05     0.266    -15.2 1.99e-52
## 2 age         0.0367    0.00226    16.2 2.69e-59
## 3 education   0.222     0.0172    12.9 2.87e-38
## 4 income       0.0713    0.00620    11.5 1.23e-30
## 5 female       0.296     0.0764    3.87 1.08e- 4
```

c) Odds ratios:

```

exp(coef(logit))

## (Intercept)      age   education      income      female
## 0.01746474  1.03735990  1.24898963  1.07389559  1.34418610

```

The odds ratio for `education` indicates the multiplicative change in the odds of voting for each additional year of education. An odds ratio above 1 means more education is associated with higher odds of voting.

d) Verify all predicted probabilities are bounded:

```

preds_logit = predict(logit, type = "response")
range(preds_logit)

## [1] 0.2511085 0.9945010

```

All predicted probabilities are between 0 and 1.

5. Comparing LPM and logit

a) Average marginal effects:

```

avg_slopes(logit)

##
##          Term       Contrast Estimate Std. Error      z Pr(>|z|)      S 2.5 %
## age      mean(dY/dX)    0.00382  0.000226 16.90 <0.001 210.3 0.00338
## education mean(dY/dX)   0.02314  0.001760 13.15 <0.001 128.8 0.01969
## female    mean(1) - mean(0) 0.03101  0.008041  3.86 <0.001 13.1 0.01525
## income     mean(dY/dX)   0.00742  0.000633 11.72 <0.001 103.0 0.00618
## 97.5 %
## 0.00426
## 0.02659
## 0.04677
## 0.00866
##
## Columns: term, contrast, estimate, std.error, statistic, p.value, s.value, conf.low, conf.high, predict
## Type: response

```

b) The AMEs from the logit model are similar to the LPM coefficients, as expected when predicted probabilities are mostly in a moderate range. Both approaches tell a broadly similar story about the relationship between each predictor and voter turnout.

c) Side-by-side table:

```

modelsummary(list("LPM" = lpm, "Logit" = logit),
            vcov = list("robust", NULL), output = "markdown")

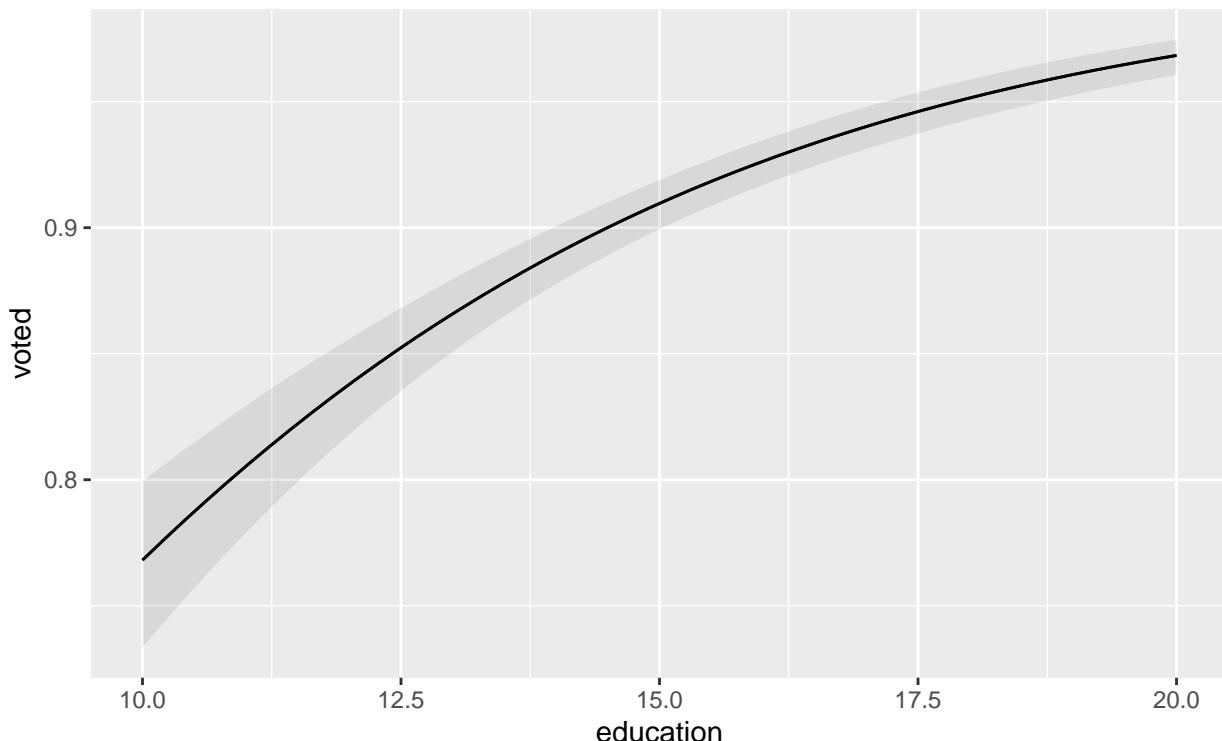
```

6. Predicted probabilities

a) Predicted probability across education:

	LPM	Logit
(Intercept)	0.240 (0.029)	-4.048 (0.266)
age	0.004 (0.000)	0.037 (0.002)
education	0.019 (0.001)	0.222 (0.017)
income	0.008 (0.001)	0.071 (0.006)
female	0.034 (0.008)	0.296 (0.076)
Num.Obs.	6733	6733
R2	0.110	
R2 Adj.	0.110	
AIC	4038.4	4646.7
BIC	4079.3	4680.8
Log.Lik.	-2013.218	-2318.343
F	165.848	157.632
RMSE	0.33	0.32
Std.Errors	HC3	

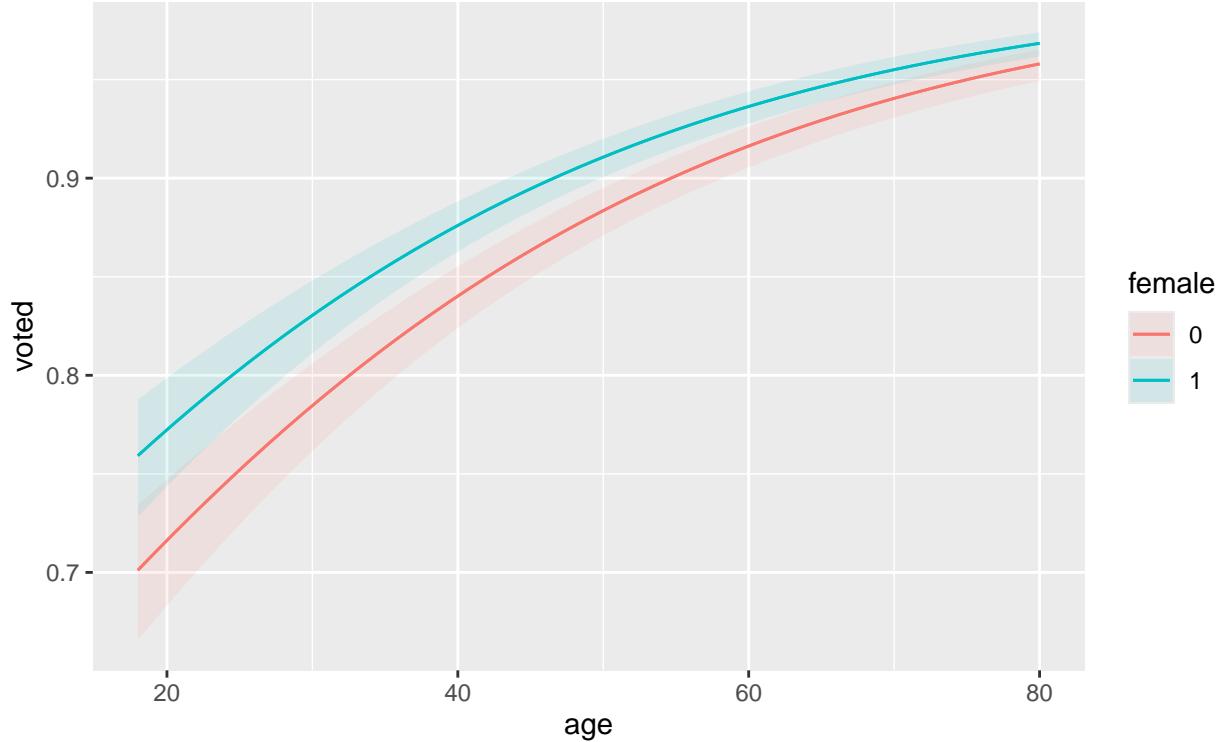
```
p1 = plot_predictions(logit, condition = "education")
p1
```



```
ggsave("pred_prob_education.png", p1, width = 6, height = 4)
```

b) Predicted probabilities by age and gender:

```
p2 = plot_predictions(logit, condition = c("age", "female"))
p2
```



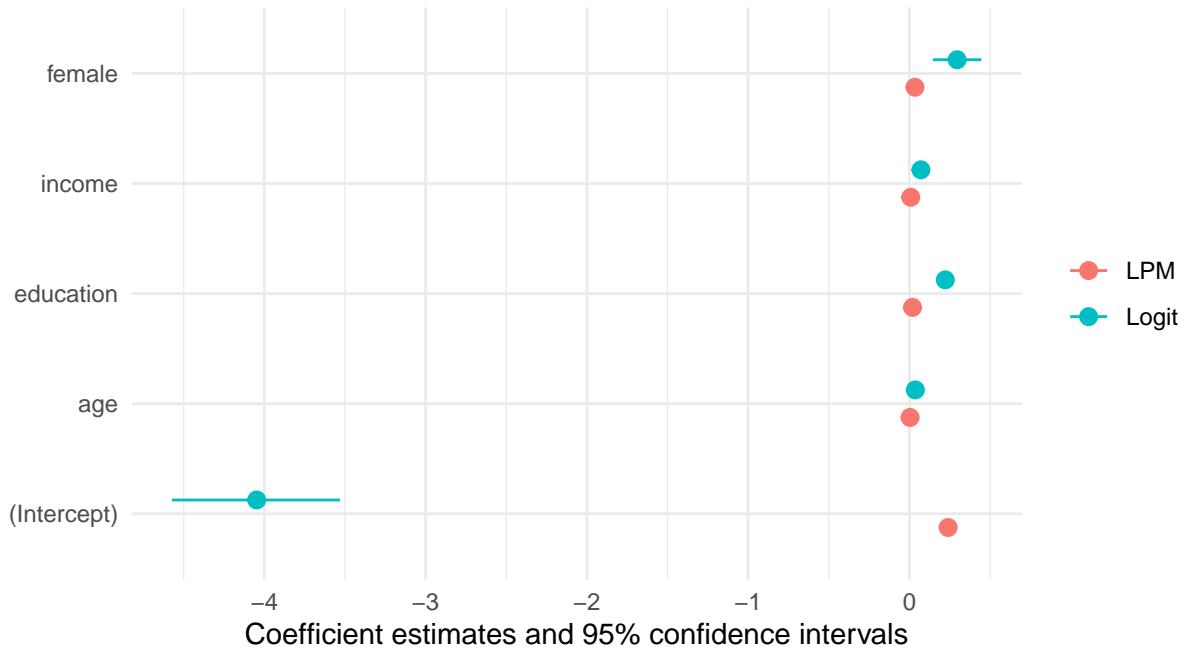
```
ggsave("pred_prob_age_gender.png", p2, width = 6, height = 4)
```

c) Education shows a clear positive relationship with turnout. Age also has a positive effect. The plot by gender shows that both men and women follow similar age-turnout patterns, with any gender gap being modest relative to the age effect.

7. Presenting results

a–b) Coefficient plot:

```
p3 = modelplot(list("LPM" = lpm, "Logit" = logit),
               vcov = list("robust", NULL))
p3
```



```
ggsave("coefplot_lpm_logit.png", p3, width = 6, height = 4)
```

- c) For this dataset, the LPM and logit lead to similar substantive conclusions: age, education, and income are all positively associated with turnout, and gender has a modest or negligible effect. The differences between LPM and logit matter more when predicted probabilities are close to the boundaries (0 or 1). In this sample, turnout is relatively common, so the linear approximation works reasonably well.