Assignment 3 – Solutions: Part 2 (STAR — High School Graduation)

Applied Quantitative Methods II, UC3M

1. Data preparation

a) Load and create factor variables:

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

star = read.csv("https://raw.githubusercontent.com/franvillamil/AQM2/refs/heads/master/datasets/star/star.

star$classtype = factor(star$classtype,
  levels = 1:3,
  labels = c("Small", "Regular", "Regular+Aide"))

star$race = factor(star$race,
  levels = 1:6,
  labels = c("White", "Black", "Asian", "Hispanic",
             "Native American", "Other"))
```

b) Create binary `small` indicator:

```
star$small = ifelse(star$classtype == "Small", 1, 0)
```

c) Drop observations with missing `hsgrad`:

```
df = star %>% filter(!is.na(hsgrad))
nrow(df)
```

```
## [1] 3047
```

d) Graduation rate overall and by class type:

```
mean(df$hsgrad)
```

```
## [1] 0.8332786
```

```
df %>%
  group_by(classtype) %>%
  summarise(grad_rate = mean(hsgrad), n = n())
```

```
## # A tibble: 3 x 3
##   classtype    grad_rate     n
##   <fct>            <dbl> <int>
## 1 Small            0.836   902
## 2 Regular          0.825  1081
## 3 Regular+Aide     0.839  1064
```

Students in small classes have a slightly higher graduation rate than those in regular or regular+aide classes.

2. LPM and logit

a) LPM:

```
lpm1 = lm(hsgrad ~ small, data = df)
tidy(lpm1)
```

```
## # A tibble: 2 x 5
##   term        estimate std.error statistic p.value
##   <chr>          <dbl>     <dbl>     <dbl>   <dbl>
## 1 (Intercept)  0.832     0.00805   103.      0
## 2 small        0.00375   0.0148      0.254   0.800
```

b) Logit:

```
logit1 = glm(hsgrad ~ small, family = binomial, data = df)
tidy(logit1)
```

```
## # A tibble: 2 x 5
##   term        estimate std.error statistic   p.value
##   <chr>          <dbl>     <dbl>     <dbl>     <dbl>
## 1 (Intercept)   1.60      0.0578    27.7   4.92e-169
## 2 small         0.0271    0.107      0.254 8.00e-  1
```

c) The LPM coefficient on `small` is the estimated difference in graduation probability between small and non-small classes. Because this is experimental data, it has a causal interpretation.

d) AME from logit:

```
avg_slopes(logit1)
```

```
##
##   Term Contrast Estimate Std. Error     z Pr(>|z|)   S   2.5 % 97.5 %
##  small    1 - 0  0.00375     0.0147 0.255    0.799 0.3 -0.0251 0.0326
##
## Columns: term, contrast, estimate, std.error, statistic, p.value, s.value, conf.low, conf.high
## Type:  response
```

The AME is very close to the LPM coefficient, as expected when the outcome probability is moderate (around 0.8).

3. Adding controls

a) Controlled models:

```
lpm2 = lm(hsgrad ~ small + race + yearssmall, data = df)
logit2 = glm(hsgrad ~ small + race + yearssmall,
             family = binomial, data = df)
```

b) Compare bivariate and controlled coefficients on `small`:

```
tidy(lpm1) %>% filter(term == "small") %>% select(term, estimate)
```

```
## # A tibble: 1 x 2
##   term  estimate
##   <chr>    <dbl>
## 1 small  0.00375
```

```
tidy(lpm2) %>% filter(term == "small") %>% select(term, estimate)
```

```
## # A tibble: 1 x 2
##   term  estimate
##   <chr>    <dbl>
## 1 small  -0.0756
```

The coefficient on `small` changes little when controls are added. This is consistent with successful randomization: because treatment was randomly assigned, covariates are balanced across groups and do not confound the treatment effect.

c) Interpret `yearssmall` from the logit:

```
tidy(logit2)
```

```
## # A tibble: 7 x 5
##   term                estimate std.error statistic   p.value
##   <chr>                  <dbl>     <dbl>     <dbl>     <dbl>
## 1 (Intercept)            1.83     0.0712   25.6    4.97e-145
## 2 small                 -0.562    0.189    -2.98   2.91e-  3
## 3 raceBlack             -0.834    0.102    -8.20   2.32e- 16
## 4 raceAsian             -0.192    1.10     -0.175  8.61e-  1
## 5 raceNative American   11.7    535.        0.0219 9.83e-  1
## 6 raceOther             11.9    377.        0.0316 9.75e-  1
## 7 yearssmall             0.210    0.0572    3.66   2.48e-  4
```

```
avg_slopes(logit2, variables = "yearssmall")
```

```
##
##        Term Estimate Std. Error    z Pr(>|z|)    S  2.5 % 97.5 %
##  yearssmall   0.0283     0.0077 3.67   <0.001 12.0 0.0132 0.0434
##
## Columns: term, estimate, std.error, statistic, p.value, s.value, conf.low, conf.high
## Type:  response
```

The coefficient on `yearssmall` captures the cumulative effect of each additional year in a small class on the probability of graduating. The AME converts the log-odds coefficient to a probability scale.

4. Predicted probabilities

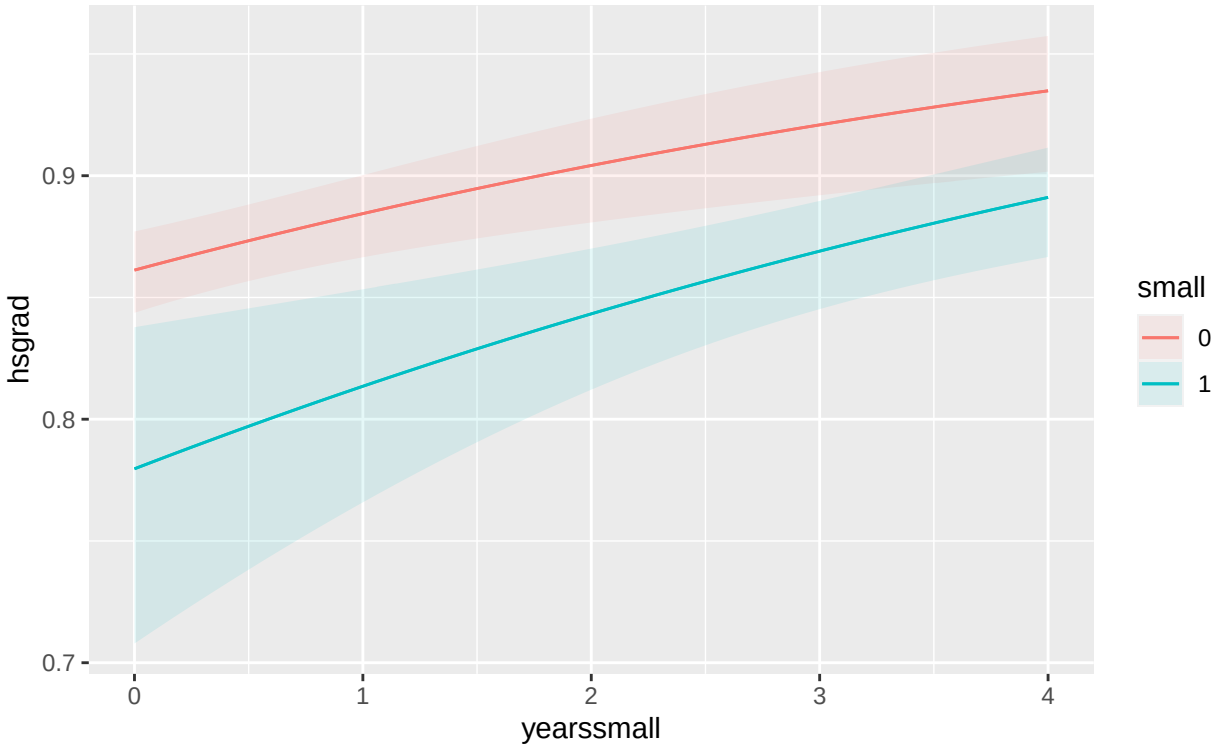
a) Predicted probabilities for specific profiles:

```
predictions(logit2,
  newdata = datagrid(
    small = c(1, 0),
    race = c("White", "Black"),
    yearssmall = c(3, 0)))
```

```
##
##  small  race yearssmall Estimate Pr(>|z|)     S 2.5 % 97.5 %
##      0 White          0    0.861   <0.001 479.4 0.844  0.877
##      0 White          3    0.921   <0.001 145.8 0.892  0.943
##      0 Black          0    0.729   <0.001  98.6 0.695  0.762
##      0 Black          3    0.835   <0.001  60.7 0.780  0.878
##      1 White          0    0.780   <0.001  34.0 0.708  0.838
##      1 White          3    0.869   <0.001 266.0 0.845  0.890
##      1 Black          0    0.606   0.0312   5.0 0.510  0.694
##      1 Black          3    0.742   <0.001  68.6 0.698  0.782
##
## Columns: rowid, estimate, p.value, s.value, conf.low, conf.high, hsgrad, small, race, yearssmall
## Type:  invlink(link)
```

b) Plot across `yearssmall` by `small`:

```
p1 = plot_predictions(logit2, condition = c("yearssmall", "small"))
p1
```



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

5. Interactions

a) Interaction model:

```
logit3 = glm(hsgrad ~ small * race + yearssmall,
             family = binomial, data = df)
```

b) Marginal effect of `small` by race:

```
avg_slopes(logit3, variables = "small", by = "race")
```

```
##
##   Term          Contrast            race     Estimate Std. Error         z
##  small mean(1) - mean(0) White           -0.076464381   0.029265 -2.612798
##  small mean(1) - mean(0) Black           -0.102992190   0.047833 -2.153167
##  small mean(1) - mean(0) Asian            0.222953892   0.197711  1.127673
##  small mean(1) - mean(0) Native American -0.000001064   0.000570 -0.001868
##  small mean(1) - mean(0) Other           -0.000000273   0.000966 -0.000283
##  Pr(>|z|)   S    2.5 %   97.5 %
##   0.00898 6.8 -0.13382 -0.01911
##   0.03131 5.0 -0.19674 -0.00924
##   0.25946 1.9 -0.16455  0.61046
##   0.99851 0.0 -0.00112  0.00112
##   0.99977 0.0 -0.00189  0.00189
##
## Columns: term, contrast, race, estimate, std.error, statistic, p.value, s.value, conf.low, conf.high, p
## Type:  response
```

c) The small class effect differs somewhat across racial groups. Some groups show a larger benefit from small classes, though confidence intervals are wide for groups with fewer observations (Asian, Hispanic, Native American). The interaction results suggest that the benefits of small classes may not be uniform across all student populations.

6. Presenting results and discussion

a) Comparison table:

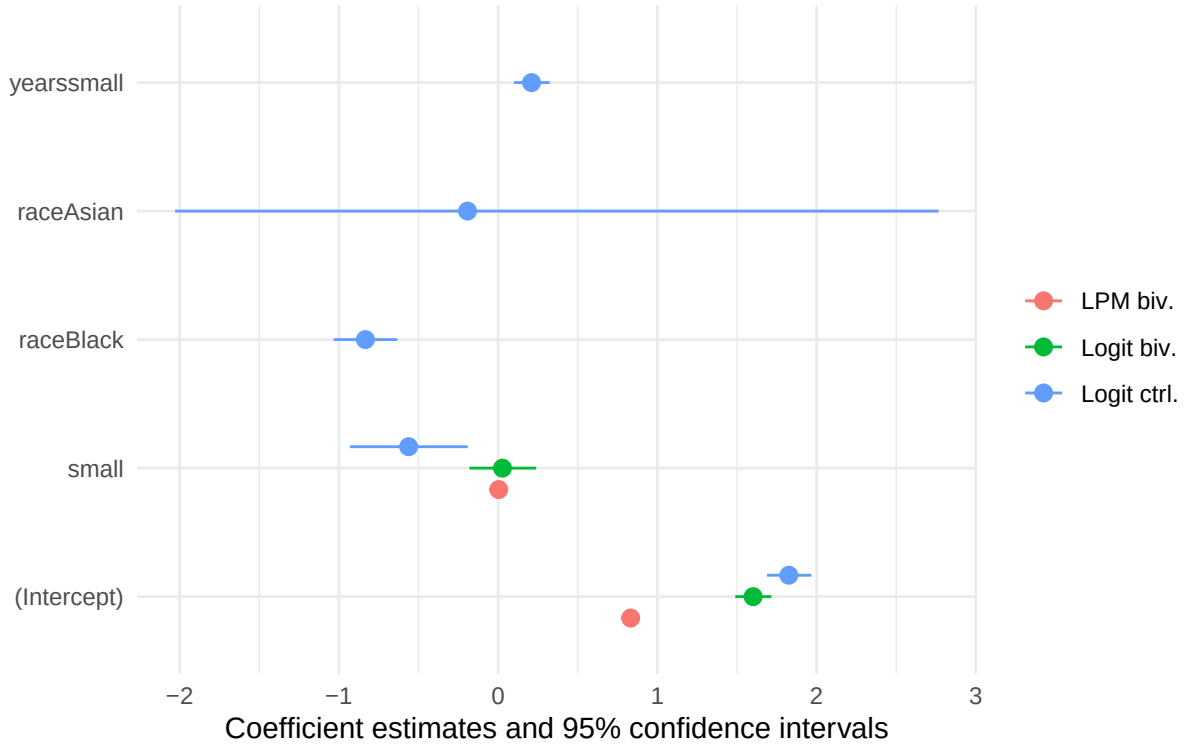
```
modelsummary(
  list("LPM biv." = lpm1, "LPM ctrl." = lpm2,
       "Logit biv." = logit1, "Logit ctrl." = logit2),
  vcov = list("robust", "robust", NULL, NULL),
  output = "markdown")
```

	LPM biv.	LPM ctrl.	Logit biv.	Logit ctrl.
(Intercept)	0.832	0.860	1.601	1.826
	(0.008)		(0.058)	(0.071)
small	0.004	-0.076	0.027	-0.562

	LPM biv.	LPM ctrl.	Logit biv.	Logit ctrl.
	(0.015)		(0.107)	(0.189)
raceBlack		-0.127		-0.834
				(0.102)
raceAsian		-0.024		-0.192
				(1.100)
raceNative American		0.140		11.741
				(535.411)
raceOther		0.165		11.926
				(377.410)
yearssmall		0.027		0.210
				(0.057)
Num.Obs.	3047	3047	3047	3047
R2	0.000	0.028		
R2 Adj.	0.000	0.026		
AIC	2638.7	2563.0	2750.2	2679.4
BIC	2656.8	2611.2	2762.2	2721.5
Log.Lik.	-1316.361	-1273.489	-1373.096	-1332.681
F	0.065		0.064	13.460
RMSE	0.37	0.37	0.37	0.37
Std.Errors	Robust	Robust		

b) Coefficient plot:

```
p2 = modelplot(
  list("LPM biv." = lpm1, "LPM ctrl." = lpm2,
       "Logit biv." = logit1, "Logit ctrl." = logit2),
  vcov = list("robust", "robust", NULL, NULL))
p2
```




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

c) The STAR data provide experimental evidence that small class sizes have a positive effect on high school graduation. The estimated effect is modest but consistent across specifications. The LPM and logit results are very similar, both in magnitude and significance, which is expected when the outcome probability is not extreme. This similarity confirms that functional form assumptions make little practical difference here. The experimental design gives this evidence more credibility than observational studies, where class size correlates with school resources, neighborhood characteristics, and student composition. Random assignment eliminates these confounders by design, allowing a causal interpretation of the treatment effect. The stability of the small-class coefficient when controls are added further supports successful randomization.