Homework 8

Gergel Anastasia

6/27/2021

Exercise 1

$Y \sim$ Bernoulli (p), $\mathbb{E}(Y) = p$, $VAR(Y) = p(1-p)$, $\mathbb{E}[u_i|x] = 0$

- Since $Y \sim$ Bernoulli (p) (i.e. takes values of 0 or 1), the conditional expectation is

$$\mathbb{E}(Y|X) = \mathbb{E}(Y|x_i) = Pr(y_i = 1|x_i) = \beta_0 + \beta_1 x_i$$

- Show that $VAR(u_i|x_i) = (\beta_0 + \beta_1 x_i)[1 - (\beta_0 + \beta_1 x_i)]$. Recall that $\mathbb{E}[u_i^2|x] = 0$

$$\begin{aligned} VAR(u_i|x_i) &= \mathbb{E}((u_i - \mathbb{E}(u_i|x))^2|x_i) \\ &= \mathbb{E}((u_i - 0)^2|x_i) = \mathbb{E}(u_i^2|x_i) \\ &= \mathbb{E}[(Y - \mathbb{E}[Y|x_i])^2|x_i] \\ &= \mathbb{E}[Y^2|x_1] - \mathbb{E}[Y|x_i]^2 \end{aligned}$$

Note since Y can take only values 1 or o, $Y^2 = Y$. Hence,

$$\begin{aligned} \mathbb{E}[Y^2|x_1] - \mathbb{E}[Y|x_i]^2 &= \mathbb{E}[Y|x_1] - \mathbb{E}[Y|x_i]^2 \\ &= \beta_0 + \beta_1 x_i - (\beta_0 + \beta_1 x_i)^2 \\ &= (\beta_0 + \beta_1 x_i)[1 - (\beta_0 + \beta_1 x_i)] \end{aligned}$$

- Show why the u_i are heteroskedastic. The residuals are heteroskedastic because $VAR(u_i|x_i) = (\beta_0 + \beta_1 x_i)[1 - (\beta_0 + \beta_1 x_i)]$, i.e. random variables do not have the same finite variance, the value of the variance vary depending on values of x_i. For instance, when $x_i = 0$, $VAR(u_i|x_i) = \beta_0 - \beta_0^2$ but when $x_i = 1$, $VAR(u_i|x_i) = (\beta_0 + \beta_1)(1 - \beta_0 - \beta_1)$.

Exercise 2

- LPM:

$$\frac{\partial\ Pr(y = 1|\boldsymbol{x})}{\partial\ x_k} = \frac{\partial(\beta_0 + \beta_1 x_1 + ...\beta_k x_k + ...)}{\partial\ x_k} = \beta_k$$

- Probit, $\Phi(\cdot)$ – standard Normal distribution:

$$\frac{\partial\ Pr(y = 1|\boldsymbol{x})}{\partial\ x_k} = \beta_k \frac{\partial}{\partial x_k}\Phi(\boldsymbol{x'\beta}) = \beta_k \phi(\boldsymbol{x'\beta}),$$

where $\phi(\boldsymbol{x'\beta}) = \dfrac{\partial}{\partial x_k}\Phi(\boldsymbol{x'\beta})$ is the value of the standard normal PDF at $\boldsymbol{x'\beta}$.

Exercise 3

```
# (a)
y <- as.matrix(c(1, 1, 0, 0, 0))
pis <- c(.7, .8, .1, .08, .32)
mean(pis)
```

```
[1] 0.4
```

```
mean(y)
```

```
[1] 0.4
```

```
y.hat <- ifelse(pis >= .5, 1, 0)
y.hat
```

```
[1] 1 1 0 0 0
```

```
sum((y - pis)^2)/sum((y - mean(y))^2) # R^2 ps = 0.207
```

```
[1] 0.2073333
```

```
sum(y.hat == y)/length(y) # 100% accuracy
```

```
[1] 1
```

```
# (b)
pis2 <- c(.7, .7, .2, .08, .32)
y.hat2 <- ifelse(pis2 >= .05, 1, 0)
y.hat2
```

```
[1] 1 1 1 1 1
```

```
sum((y - pis2)^2)/sum((y - mean(y))^2) # R^2 ps = 0.274
```

```
[1] 0.274
```

```
sum(y.hat2 == y)/length(y) # 40% accuracy
```

```
[1] 0.4
```

(c) R^2_{Efron} is robust to the threshold we set for the predicted probability function in contrast to the accuracy measure. The accuracy measure is great for showing the percent of variation explained by the model on a given sample but its values depend too much on current observations and do not say anything about how our model would perform on a different sample. Relying on the accuracy measure only to adjust the model performance would lead to overfitting.

Exercise 4

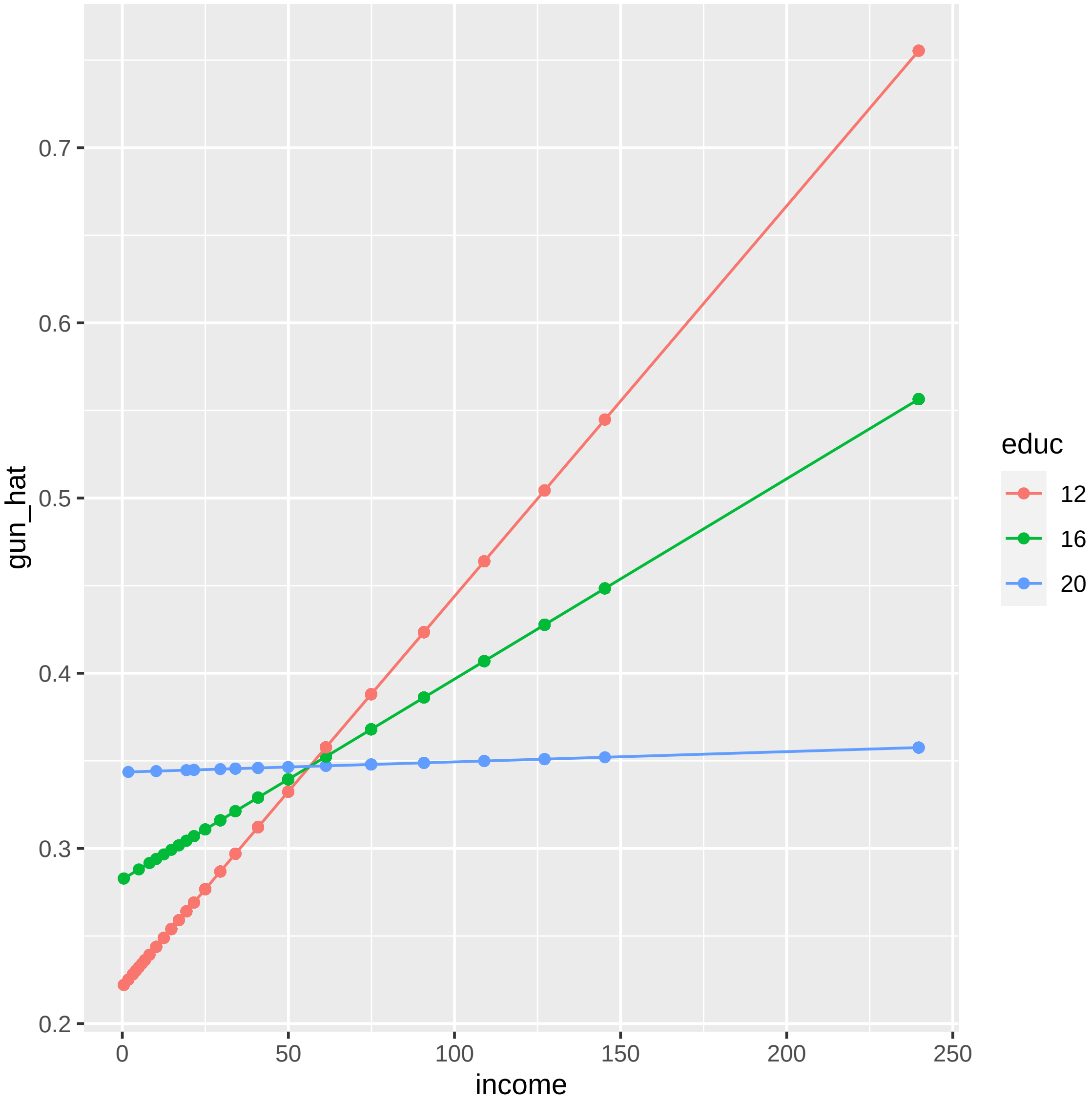
```
library(haven)

gss <- read_dta("/Users/herrhellana/Dropbox/_NYU studies/Quant I/home assignments/HW8/gss.dta")

# (a) LPM estimation
lmp_gss_1 <- lm(gun ~ educ + income + educ:income, data=gss)
gun_data <- cbind(lmp_gss_1$model, lmp_gss_1$fitted.values)
colnames(gun_data)[4] <- 'gun_hat'
```

```r
# (b) Construct a plot of the estimated relationship between income and Pr(gun = 1) at
# educ={12, 16, 20}, using different colors for each value of educ.
library(ggplot2)
gun_data2 <- gun_data[gun_data$educ %in% c(12, 16, 20), ]
gun_data2$educ <- as.factor(gun_data2$educ)

ggplot(data = gun_data2, aes(x = income, y = gun_hat, col = educ)) +
  geom_point() + geom_line()
```



(c) The relationship between income and gun ownership is more pronounced (is more strong) for respondents with 12 years of education. As years of education increase, its association with gun ownership lessens.

Exercise 5

```
cces <- read_dta("/Users/herrhellana/Dropbox/_NYU studies/Quant I/home assignments/HW8/cces.dta")
```

(a) Tabulate the unweighted distribution of the `educ` variable. Now tabulate it using the weight `commonweight`. Describe how these two tabulations are different and offer an explanation for the difference.

```
# unweighted
table(cces$educ) / sum(table(cces$educ))
```

```

         1          2          3          4          5          6 
0.03596667 0.27695000 0.21051667 0.10036667 0.23760000 0.13860000 
```

```
library(dplyr)
w <- count(cces, var = educ, wt = commonweight)
w$n/sum(w$n)
```

```
[1] 0.09424804 0.28544722 0.21850217 0.10281708 0.18986533 0.10912015
```

In the unweighted distribution of the `educ` variable, 9.4% of the sample has no degree, 19% has a college degree, and 11% has a post-grad degree, but in the weighted distribution of the `educ` variable, 3.6% of the sample has no degree, 24% has a college degree, and 14% has a post-grad degree. So weights increase the representation of those who have a degree to match the population distribution. The difference is probably due selection bias: people who tend to respond to surveys are usually educated. But those without education could be intimidated by answering survey questions or would not know the value of survey research.

(b) Do the same for the `CC18_317` variable.

```
# unweighted
table(cces[cces$race==1, 6]) / sum(table(cces[cces$race==1, 6]))
```

```

          1           2           3           4           5 
0.475077025 0.414365097 0.099581866 0.007784991 0.003191021 
```

```
# weighted
count(cces[cces$race==1,], var = CC18_317, wt = commonweight)$n /
  sum(count(cces[cces$race==1,], var = CC18_317, wt = commonweight)$n)
```

```
[1] 0.425625349 0.299196780 0.047518229 0.004141124 0.001983769 0.221534749
```

In the weigthed distribution, the percent of the white sample that voted for Clinton decreased from 41% to 30%, the percent of those who voted for Trump also decreased from 47.5% to 42.5%, as well as the percent of those who voted for someone else (from 10% to 4.75%). But the percent of those who did not recall or skipped increased: the weighted distribution even has an extra group for `skipped` the share of which was zero in the unweighted distribution. Probably, this difference is due selection bias and social desirability, i.e. respondents choose randomly a candidate instead of truthfully reporting that they did not vote or could not remember the candidate they voted for.

(c) Estimate LPM and probit models.

```
library(sandwich)
lmp_i <- lm(voted ~ income, data=cces)
heterosked_i <- vcovHC(lmp_i, type = 'HC1')

lmp_ii <- lm(voted ~ log(income), data=cces)
heterosked_ii <- vcovHC(lmp_ii, type = 'HC1')
```

```
library(clubSandwich)
probit_i <- glm(data=cces, voted ~ income, family = binomial(link = 'probit'))
vp_i <- vcovCR(probit_i, cluster = cces$inputstate, type = 'CR1')

probit_ii <- glm(data=cces, voted ~ log(income), family = binomial(link = 'probit'))
vp_ii <- vcovCR(probit_ii, cluster = cces$inputstate, type = 'CR1')


# (v)
library(stargazer)
stargazer(lmp_i, lmp_ii, probit_i, probit_ii, type = 'text', model.names = F,
          column.labels = c("LPM", "LPM", "Probit", "Probit"),
          se = list(sqrt(diag(heterosked_i)), sqrt(diag(heterosked_ii)),
                    sqrt(diag(vp_i)), sqrt(diag(vp_ii))))
```

```
================================================================================
                                               Dependent variable:
                                 -----------------------------------------------
                                                      voted
                                    LPM         LPM        Probit      Probit
                                    (1)         (2)          (3)         (4)
--------------------------------------------------------------------------------
income                            0.001***                0.004***
                                 (0.00002)                (0.0003)

log(income)                                   0.065***                0.308***
                                              (0.002)                  (0.010)

Constant                          0.835***    0.617***    0.894***     -0.025
                                  (0.003)     (0.009)      (0.029)     (0.048)

--------------------------------------------------------------------------------
Observations                       43,244      43,244      43,244      43,244
R2                                 0.013       0.028
Adjusted R2                        0.013       0.028
Log Likelihood                                           -15,710.750 -15,505.830
Akaike Inf. Crit.                                        31,425.490  31,015.670
Residual Std. Error (df = 43242)   0.326       0.323
F Statistic (df = 1; 43242)      589.478*** 1,251.944***
================================================================================
Note:                                               *p<0.1; **p<0.05; ***p<0.01
```

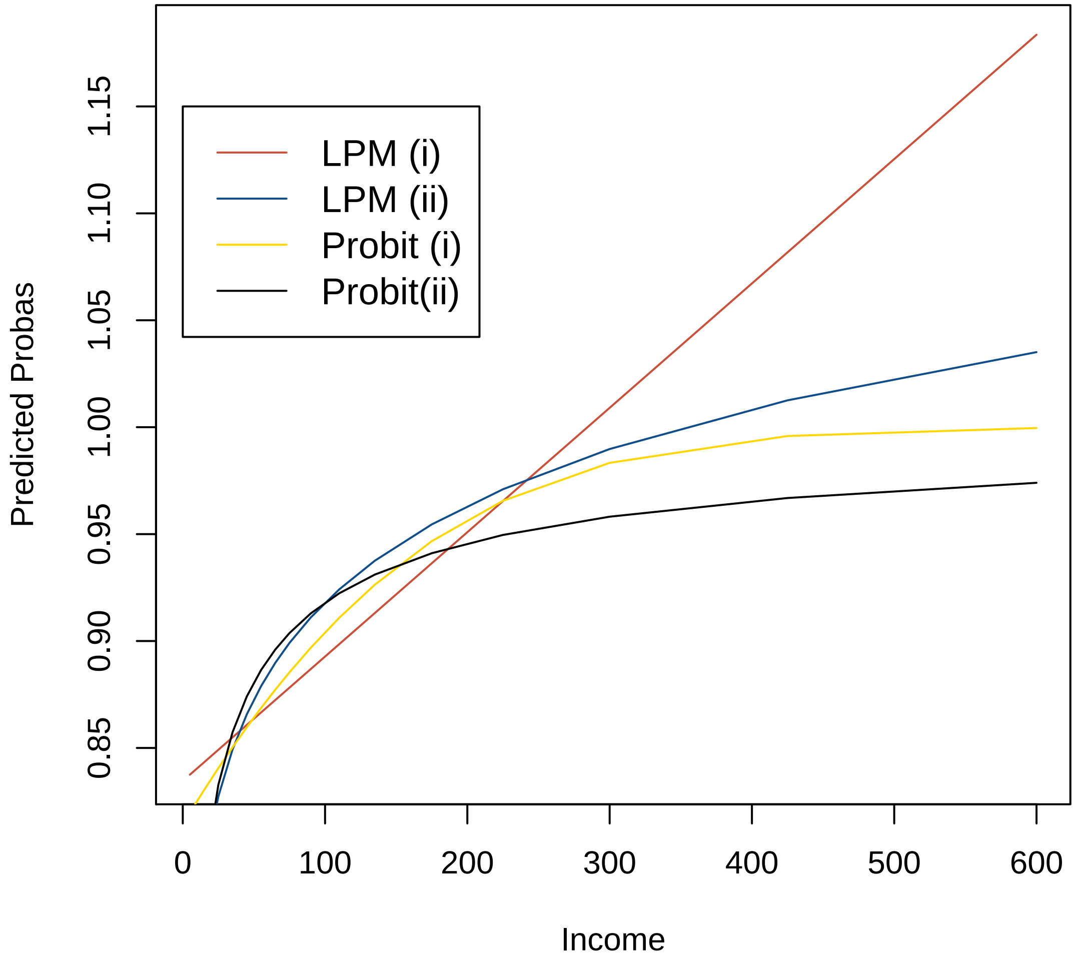
```
############################ plot ################################
# (vi)

ldv_data <- cbind(lmp_i$model, lmp_i$fitted.values,
                  lmp_ii$fitted.values, probit_i$fitted.values,
                  probit_ii$fitted.values)

colnames(ldv_data)[3:6] <- c("lpm_i", "lpm_ii", "probit_i", "probit_ii")

###
```

```r
{
plot(ldv_data$income, ldv_data$lpm_i, 'l', col = 'tomato3',
          xlab= "Income", ylab = "Predicted Probas")
lines(ldv_data$income, ldv_data$lpm_ii,  col = 'dodgerblue4')
lines(ldv_data$income, ldv_data$probit_i, 'l', col = 'gold')
lines(ldv_data$income, ldv_data$probit_ii, 'l')
legend(0, 1.15, legend=c("LPM (i)", "LPM (ii)", "Probit (i)", "Probit(ii)"),
       col=c("tomato3", "dodgerblue4", "gold", "black"), lty=1, cex=1.2)
}
```



(d) Replication

```r
cces <- cces %>% mutate(white = as.numeric(race==1))

probit_rep <- glm(data=cces, subset = educ %in% 1:6,
                  voted ~ income + gender
                  + white + as.factor(educ) + as.factor(inputstate),
                  family = binomial(link = 'probit'))
vp_rep <- vcovCR(probit_rep, cluster = cces$inputstate, type = 'CR1')


stargazer(probit_rep, type='text', omit = 'inputstate',
          title = 'DV: Reports Voting in the 2018 Election',
          omit.stat=c('ll', 'aic'), digits=2,
          covariate.labels = c("income ($1,000s)", "female",
                               "white", "high school graduate", "some college",
                               "2-year degree", "4-year degree", "Post-grad degree",
                               "intercept"),
          add.lines = c("state fixed effects", "yes"),
          se = list(sqrt(diag(vp_rep))))
```

```
DV: Reports Voting in the 2018 Election
================================================
                         Dependent variable:
                     ---------------------------
                                voted
------------------------------------------------
income (1,000s)               0.002***
                              (0.0002)

female                        -0.30***
                               (0.02)

white                          0.22***
                               (0.03)

high school graduate           0.28***
                               (0.04)

some college                   0.39***
                               (0.04)

2-year degree                  0.50***
                               (0.04)

4-year degree                  0.67***
                               (0.04)

Post-grad degree               0.83***
                               (0.04)

intercept                      0.68***
                               (0.05)

------------------------------------------------
state fixed effects
yes
Observations                   43,244
================================================
Note:                *p<0.1; **p<0.05; ***p<0.01
```