ECON 3412 HW 6

David Chen, dc3451

November 25, 2020

Problem 1

a,b,c,d

We have the following table, which has values generated by the following `R`:

```
mroz <- read_dta("./mroz.dta")

mroz_lm <- lm(inlf ~ nwifeinc + educ + exper + expersq +
                age + kidslt6 + kidsge6, data = mroz)
coeftest(mroz_lm, vcovCL)

mroz_probit <- probit(inlf ~ nwifeinc + educ + exper + expersq +
                        age + kidslt6 + kidsge6, data = mroz)
coeftest(mroz_probit, vcovCL)

mroz_logit <- glm(inlf ~ nwifeinc + educ + exper + expersq +
                    age + kidslt6 + kidsge6, data = mroz,
                  family = binomial(link = "logit"))
coeftest(mroz_logit, vcovCL)

nrow(mroz[(predict(mroz.lm) >= 0.5) == mroz$inlf,])
nrow(mroz[(pnorm(predict(mroz.logit)) >= 0.5) == mroz$inlf,])
nrow(mroz[(((1 + exp(predict(mroz.logit)) ^ -1)^-1) >= 0.5) == mroz$inlf,])
```

e

From the table, we have that $t = \frac{-0.2618}{0.03178} = -8.2374 < -1.96$, so it is significant at the 5% level.

Independent Variables	LPM (OLS)	Probit (MLE)	Logit (MLE)
nwifeinc	-0.00340517 (0.00152493)	-0.01202374 (0.00531057)	-0.0213452 (0.0090782)
educ	0.03799530 (0.00726604)	0.13090473 (0.02581922)	0.2211704 (0.0444509)
exper	0.03949239 (0.00581002)	0.12334759 (0.01885370)	0.2058695 (0.0322914)
expersq	-0.00059631 (0.00019000)	-0.00188708 (0.00060072)	-0.0031541 (0.0010124)
age	-0.01609081 (0.00239901)	-0.05285267 (0.00835318)	-0.0880244 (0.0144392)
kidslt6	-0.26181047 (0.03178320)	-0.86832850 (0.11620366)	-1.4433541 (0.2031606)
kidsge6	0.01301223 (0.01353293)	0.03600496 (0.04529575)	0.0601122 (0.0798824)
constant	0.58551922 (0.15225987)	0.27007677 (0.50517502)	0.4254524 (0.8597302)
% Correctly Predicted	73.43%	73.57%	73.57%
Log-likelihood value	-	-401.30	-401.76
Pseudo R^2	0.2642	0.2206	0.2197

f

This is a test that *exper* and *expersq* are jointly insignificant, and we get an F-statistic of 67.17 from `linearHypothesis(mroz.lm, c("exper = 0", "expersq = 0"), vcov = vcovCL)`, so they are jointly significant at the 5% level ($p < 2.2e-16$).

Holding the other factors constant, we expect to see that the probability of a woman being in the labor force changes nonlinearly with the amount of years they've spent in the labor market (since $t = -3.13$ for *expersq*). Note that $0.3949exper - 0.000596exper^2$ reaches a maximum at $0.3949/(2 \cdot 0.00059) \approx 33$ years and vanishes at 0 and ≈ 66 years as a result. This tells us that holding all else in the regression constant first starting out, the first few years of experience increase (on average) the likelihood of being in the work force the most, and that after ≈ 33 of working, that women actually see that their likelihood foe being in the labor force start to decline with more years of experience (though still better than no experience until 66 years of experience!), probably since soon after that point, people are

starting to retire. The likelihood of being in the workforce is maximized at about 33 years of experience.

g

From the table, we have that $t = \frac{-0.86833}{0.1162} = -7.4725 < -1.96$, so it is significant at the 5% level.

h

This is a test that *exper* and *expersq* are jointly insignificant, and we get an χ^2-statistic of 97.21 from `linearHypothesis(mroz.probit, c("exper = 0", "expersq = 0"), vcov = vcovCL)`, so they are jointly significant at the 5% level ($p < 2.2e-16$).

This is roughly the same interpretation as section f, since we have that Φ is a monotonically increasing function, and the signs of $exper, expersq$ have not changed. Note that $0.12334exper - 0.00188exper^2$ reaches a maximum at $0.12334/(2 \cdot 0.00188) \approx 32.5$ years and vanishes at 0 and ≈ 65 years as a result. Then, the likelihood of being in the workforce holding all else in the regression constant increases (on average) with experience until ≈ 32.5 years of being in the workforce, after which the likelihood starts to decrease again with each additional year, and the total effect of experience after 65 is worse than 0 years of employment. Hard to discern comparative magnitudes of effect since Φ is heavily nonlinear, but the effect on the z-score is direct: highest impact when first starting out and at around 65 years of experience, minimal effect at around 32.5, likelihood of being in the workforce maximized at about 32.5 years of experience.

Problem 2

a

```
> vote <- read_dta("./vote1.dta")
> vote$winA <- vote$voteA > 50
> vote$dSpend <- vote$expendA - vote$expendB
> vote.lm <- lm(winA ~ expendB + dSpend + prtystrA + democA, data = vote)
> coeftest(vote.lm, vcovCL)

t test of coefficients:

               Estimate  Std. Error t value  Pr(>|t|)
(Intercept)  0.07197900  0.13425537  0.5361 0.5925744
expendB     -0.00015686  0.00012344 -1.2707 0.2056052
```

```
dSpend       0.00098666  0.00011353  8.6905 3.117e-15 ***
prtystrA     0.00696709  0.00243076  2.8662 0.0046858 **
democA       0.18657501  0.05210329  3.5809 0.0004481 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

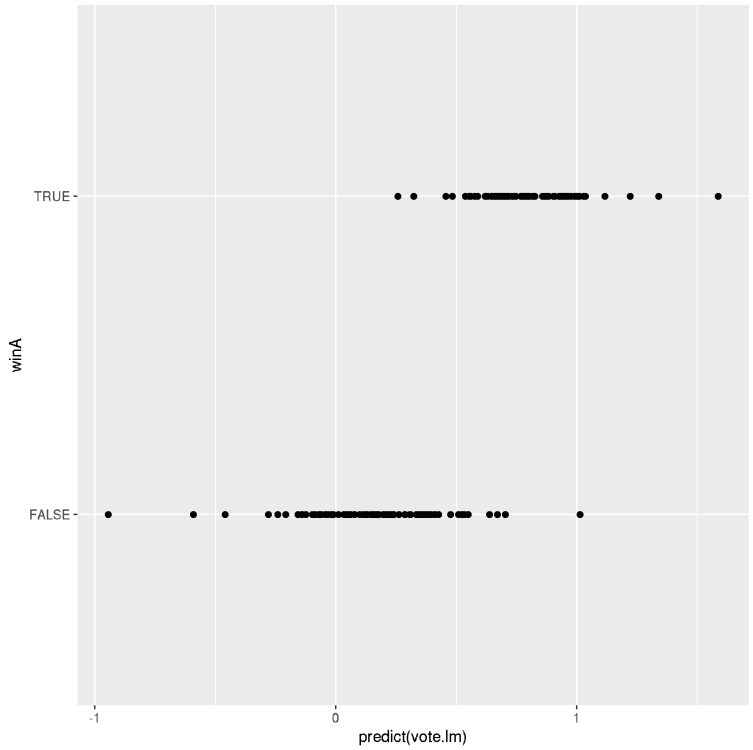
i

The coefficient on *dSpend* is 0.000986, so for each thousand dollars that party A outspent party B, holding all else constant, their likelihood of winning is increased on average by 0.0986% (so a million dollars would increase the likelihood by an expected amount of 9.8%). Note that holding the others constant means holding *expendB* constant, so this interpretation is exactly the same as if party A had spend an extra thousand dollars (if we regress on *expendA*, we get the same coefficient of 0.000986 on *expendA*, so this checks out). The important part is the sign, which tells us that more spending increases the chance of a victory.

When *dSpend* goes from -250 to -249, the likelihood of winning increases (holding the other factors constant) by $(-249(0.000986) - -250(0.000986) = 0.000986)$ which is the same as when *dSpend* goes from 0 to 1. This makes sense; after all, it is a *linear* probability model (and we have only modeled a linear fit), so the effect of a unit change is the same everywhere. However, this might not be the best fit: for example, increasing the amount you outspend the other party seems more important when the spending is close. If you're already outspending the other party by a few million, another thousand probably won't matter, but another thousand might matter more when the spending is close.

ii

```
ggplot(data = vote) + geom_point(aes(x = predict(vote.lm), y = winA))
```



iii

`nrow(vote[(predict(vote.lm) > 0.5) == vote$winA,])` returns 160, so there are 160 correctly predicted wins, out of a total of 173 total districts, for a % of $160/173 = 92.5\%$.

b

```
> vote.probit <- probit(winA ~ expendB + dSpend + prtystrA + democA, data = vote)
> coeftest(vote.probit, vcovCL)

t test of coefficients:

              Estimate  Std. Error t value  Pr(>|t|)
(Intercept) -1.68679531  1.08237805 -1.5584   0.12102
expendB     -0.00189816  0.00074718 -2.5404   0.01198 *
dSpend       0.00617806  0.00112223  5.5052 1.358e-07 ***
prtystrA     0.02948302  0.01855579  1.5889   0.11397
democA       0.84355220  0.44218524  1.9077   0.05814 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

i

The big problem is that the linear probability model outputs some probabilities that are outside of $[0, 1]$. In particular, there is one county where the predicted value of $winA$ is -0.944! Using probit will allow for the probabilities to actually be on the unit interval. Further, it might account for the potential nonlinearity discussed part in a.i.

ii

The interpretation of the coefficient of $dSpend$ is that holding all else constant, the effect of outspending party B by a thousand dollars (equivalent to increasing party A spending by 1000) increases the $z-$value of party A winning by 0.006178. In effect, this means that the more party A spends, the more likely this model predicts them to win.

The mean marginal effects (given as the average of the derivatives w.r.t. $dSpend$) are 0.000666 at $dSpend = -250$, 0.0019 at $dSpend = 0$, and 0.00116 at $dSpend = 250$. This is reasonable, since we have that the CDF of the normal distribution is nonlinear, so we expect the marginal effect (here technically the derivative w.r.t $dSpend$). In a applied sense, this is also reasonable: the marginal effect is largest when the spending race is close ($dSpend = 0$) compared to when it isn't, like we said we might expect in part a.i.

We do this by computing the derivative by hand:

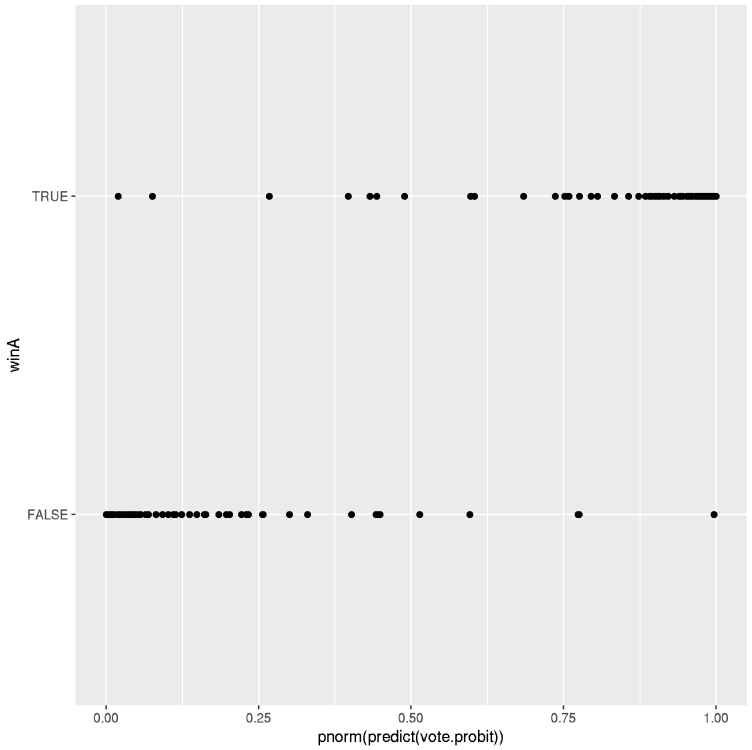
$$\frac{\partial \Phi}{\partial X} = \frac{\beta_X}{\sqrt{2\pi}} e^{-(\beta_0 + \sum \beta_{X_i} X_i)^2/2}$$

```
> mean(0.006178 * 1 / sqrt(2 * pi) * exp(-(-1.686795 + -0.001898 * vote$expendB +
                                           0.029483 * vote$prtystrA +
                                           0.843552 * vote$democA +
                                           0.006178 * -250)^2 / 2))
> mean(0.006178 * 1 / sqrt(2 * pi) * exp(-(-1.686795 + -0.001898 * vote$expendB +
                                           0.029483 * vote$prtystrA +
                                           0.843552 * vote$democA +
                                           0.006178 * 0)^2 / 2))
> mean(0.006178 * 1 / sqrt(2 * pi) * exp(-(-1.686795 + -0.001898 * vote$expendB +
                                           0.029483 * vote$prtystrA +
                                           0.843552 * vote$democA +
                                           0.006178 * 250)^2 / 2))
```

which comes out to the values above.

iii

```
ggplot(data = vote) + geom_point(aes(x = pnorm(predict(vote.probit)), y = winA))
```



iv

`nrow(vote[pnorm((predict(vote.probit)) > 0.5) == vote$winA,])` returns 161, so there are 161 correctly predicted wins, out of a total of 173 total districts, for a % of 161/173 = 93.1%.

v

The interpretation of the linear model is very clear: a unit change in any of the factors has a linear effect on the predicted likelihood of winning the district, irregardless of the values of the other regressors. They also predict (almost) the same amount of districts, so perhaps the linear model is more clear and equally effective here.