Logit and Probit Notes

```
library(haven)
setwd("/Users/jessicakreese/Desktop/r studio/722_Class_Notes")
d <- read_dta("cps08for729a.dta")
```

Logit Model Summary

```
# Run logit on vote.

logit.model <- glm(vote ~ close + as.numeric(age) + as.numeric(edu7cat), family=binomial

summary(logit.model)
```

```

Call:
glm(formula = vote ~ close + as.numeric(age) + as.numeric(edu7cat),
    family = binomial(link = "logit"), data = d)

Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)       -3.930361   0.384551 -12.914   <2e-16 ***
close             -0.007521   0.005021  -1.498    0.134
as.numeric(age)    0.040661   0.003107  13.086   <2e-16 ***
as.numeric(edu7cat)  0.640868   0.046286  14.001   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2797.2  on 2187  degrees of freedom
Residual deviance: 2424.4  on 2184  degrees of freedom
  (258 observations deleted due to missingness)
AIC: 2432.4

Number of Fisher Scoring iterations: 4
```

Interpreting Age Coefficient

- **On average and holding all else constant, as age increases, the probability of voting increases.**

Interpreting Education Coefficient

- **On average and holding all else constant, as education increases, the probability of voting increases**

Interpreting the Closing Date Coefficient

- **As the closing date for registration gets further from election day, the probability of voting goes down on average and holding all else constant.**

```
# Regression sample

sample <- d[complete.cases(divote, dsclose, dsage, dsedu7cat)==T,]
```

The Mechanics Of Obtaining The Predicted Probabilities

```
# Get the predicted probabilities for the valid observations.

preds <- predict(logit.model, type = "response")
options(digits=4)
summary(preds)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.0998  0.5258  0.6870  0.6627  0.8132  0.9851
```

```
# List the coefficients.

coefs <- coef(logit.model)
coefs
```

```
       (Intercept)              close     as.numeric(age) as.numeric(edu7cat)
         -3.930361         -0.007521          0.040661          0.640868
```

```
# Calculate the predicted probability by hand, first the way I learned before
# I was smart enough to look at the Stata manual.
# Note that we are calculating the predictions just for those cases that are in the samp
# exp = E in the regular equation

plogit <- 1/(1+exp(-1*((-0.00752053*sample$close) + (0.04066068*as.numeric(sample$age))
                       + (0.6408677*as.numeric(sample$edu7cat)) + (-3.930361 ))))
summary(plogit)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.0998  0.5258  0.6870  0.6627  0.8132  0.9851
```

```
# What a pain it is to get all of the coefficients and paste them in & it is harder to c
# So here is the better approach that uses the names of the coefficients as R stores the
# the object 'coefs_logit' we created to make the code easier to type and read.

plogit2 <- 1/(1+exp(-1*(coefs['(intercept)'] + coefs['close']*sample$close + coefs['age'
                        coefs['edu7cat']*as.numeric(sample$edu7cat))))
```

```
## Or equivalently:
plogit2 <- 1/(1+exp(-1*(coefs[1] + coefs[2]*sample$close + coefs[3]*as.numeric(sample$ag
                        coefs[4]*as.numeric(sample$edu7cat))))
## I will use the numeric one from now on, but the other option is equally fine
summary(plogit2)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.0998  0.5258  0.6870  0.6627  0.8132  0.9851
```

```
# Here is the same thing using the invlogit feature.
library(arm)
```

```
Loading required package: MASS

Loading required package: Matrix

Loading required package: lme4


arm (Version 1.14-4, built: 2024-4-1)

Working directory is /Users/jessicakreese/Desktop/r studio/722_Class_Notes
```

```
plogit3 <- invlogit(coefs[2]*sample$close + coefs[3]*as.numeric(sample$age) +
                        coefs[4]*as.numeric(sample$edu7cat) + coefs[1])
summary(plogit3)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.0998  0.5258  0.6870  0.6627  0.8132  0.9851
```

```
# Create a variable so that the R and hand calculations can be compared.
testl <- preds - plogit
test2 <- preds - plogit2
testl3 <- preds - plogit3
summary(cbind(testl,testl2,testl3))
```

```
     testl              testl2              testl3
Min.   :-2.50e-08   Min.   :-4.44e-16   Min.   :-3.33e-16
1st Qu.:-7.74e-09   1st Qu.:-5.55e-17   1st Qu.:-1.11e-16
Median : 3.20e-10   Median : 0.00e+00   Median : 0.00e+00
Mean   : 4.30e-10   Mean   :-1.12e-17   Mean   :-1.27e-17
3rd Qu.: 7.17e-09   3rd Qu.: 0.00e+00   3rd Qu.: 0.00e+00
Max.   : 4.68e-08   Max.   : 2.22e-16   Max.   : 3.33e-16
```

```
# So, the difference is all zeros (even with the less precise method to generate the plo

# Now probit.

probit.model <- glm(vote ~ close + as.numeric(age) + as.numeric(edu7cat), family=binomia
summary(probit.model)
```

```

Call:
glm(formula = vote ~ close + as.numeric(age) + as.numeric(edu7cat),
    family = binomial(link = "probit"), data = d)

Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)       -2.26432    0.17491  -12.95   <2e-16 ***
close             -0.00471    0.00299   -1.58    0.12
as.numeric(age)    0.02361    0.00179   13.16   <2e-16 ***
as.numeric(edu7cat)  0.37646   0.02629   14.32   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2797.2  on 2187  degrees of freedom
Residual deviance: 2429.7  on 2184  degrees of freedom
  (258 observations deleted due to missingness)
AIC: 2438

Number of Fisher Scoring iterations: 4
```

```
# Get the predicted probabilities for the valid observations.

preds.probit <- predict(probit.model, type = "response")
summary(preds.probit)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.102   0.527   0.680   0.662   0.807   0.993
```

```
# List the coefficients.
coefs_probit <- coef(probit.model)
coefs_probit
```

```
       (Intercept)              close     as.numeric(age) as.numeric(edu7cat)
         -2.264323         -0.004708          0.023608          0.376455
```

```
# calculate the predicted probability by hand, using the "better approach".

pprobit <- pnorm(coefs_probit[2]*sample$close + coefs_probit[3]*as.numeric(sample$age) +
                     coefs_probit[4]*as.numeric(sample$edu7cat) + coefs_probit[1])
summary(pprobit)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.102   0.527   0.680   0.662   0.807   0.993
```

```
# create a variable so that the R and hand calculations can be compared.

testp <- preds.probit - pprobit
summary(testp)
```

```
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
-2.22e-16  0.00e+00  0.00e+00 -7.50e-19  0.00e+00  2.22e-16
```

Counter-factual Analysis

```
# Continuing with probit, what is the effect of changing closing date so that all have c
#
# compute predicted prob setting close = 0 for all.

ppclose0 <- pnorm(coefs_probit[2]*0 + coefs_probit[3]*as.numeric(sample$age) +
                      coefs_probit[4]*as.numeric(sample$edu7cat) + coefs_probit[1])
```

For each observation, we now have the baseline operation. So, what if America said everyone can register on election day.

```
# Now compute the effect, which we might define as the difference between the predicted

effect <- ppclose0 - pprobit
summary(effect)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.0000  0.0182  0.0357  0.0327  0.0493  0.0563
```

Above, is the "status quo" treatment effect. We're imagining everyone now lives in a state where the closing date is zero and we're comparing it to the probability of voting in the state where they actually live in. That's how we're going to get our effect.

- *Mike uses the mean number to observe the effect.*

The effect means that: on average, holding all else constant, if America said everyone can register on election day, the probability of voting goes up by 3%.

```
# Now examine the effect by education (R&W (1980) theorize that the effect should decrea
library(plyr)
sample$effect <- effect
options(digits = 3)
ddply(sample, .edu7cat, summarise, N = length(edu7cat), mean = mean(effect), sd = sd(eff
                                   min = min(effect), max = max(effect))
```

```
  edu7cat   N   mean      sd    min    max
1       1  11 0.0393 0.01225 0.0178 0.0560
2       2  81 0.0389 0.01824 0.0000 0.0563
3       3 103 0.0382 0.01714 0.0000 0.0563
4       4 724 0.0379 0.01810 0.0000 0.0563
5       5 623 0.0346 0.01733 0.0000 0.0563
6       6 380 0.0247 0.01414 0.0000 0.0510
7       7 186 0.0144 0.00901 0.0000 0.0358
```

The policies seem to be effecting people with lower levels of education rather than higher levels of education.

```
# Another way to think about this is to compare a scenario in which none had close = 0 a
# We already have a prediction for close = 0, so get a prediction for close = 30.

ppclose30 <- pnorm(coefs_probit[2]*30 + coefs_probit[3]*as.numeric(sample$age) +
                       coefs_probit[4]*as.numeric(sample$edu7cat) + coefs_probit[1])
summary(ppclose30)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.101   0.515   0.671   0.651   0.795   0.991
```

The is for when the closing date is zero and getting the effect for each individual person.

```
# Now compute the effect, which we might define as the difference between the predicted
# rate with close = 0 for all and close = 30 for all.

effect2 <- ppclose0 - ppclose30
summary(effect2)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.0038  0.0178  0.0406  0.0441  0.0545  0.0563
```

The effect goes from 6.5% to 4.4%.