Logistic and Ordered Logistic Regression

Note: The dataset used in this tutorial and the R Script are on Moodle:

Loading the 2016 CCES dataset

```
install.packages("foreign", dependencies=TRUE)
library(foreign)

dat <- read.dta(file.choose(), convert.factors=FALSE)
```

Recode Variables Quickly

1. Here, we recode a few variables that we will need later on.

```
table(dat$CC16_340a)
dat$ideology <- recode(dat$CC16_340a, "8=NA")
summary(dat$ideology)

table(dat$religpew)
dat$religion <- recode(dat$religpew, "1:2='Protestant-Catholic'; 3='Mormon';
4='Orthodox'; 5='Jewish'; 6:8='Other';  9='Atheist'; 10:11='Nothing'; else=NA")
table(dat$religion)

dat$religion <- factor(dat$religion, levels=c("Protestant-Catholic", "Mormon", "Orthodox",
"Jewish", "Other", "Nothing", "Atheist"))

table(dat$birthyr)
dat$age <- 2016 - dat$birthyr
table(dat$age)
summary(dat$age)

table(dat$faminc)
dat$income <- recode(dat$faminc, "31=NA; 97=NA")
table(dat$income)
summary(dat$income)

table(dat$gender)
dat$gender1 <- recode(dat$gender, "1=0; 2=1")
table(dat$gender)

table(dat$race)
dat$race1 <- recode(dat$race, "1='White'; 2='Black'; 3='Hispanic'; 4:8='Other'")
table(dat$race1)
```

```
dat$race1 <- factor(dat$race1, levels=c("White", "Black", "Hispanic", "Other"))

table(dat$CC16_410a)
dat$votechoice <- recode(dat$CC16_410a, "1='1'; 2='0'; else=NA")
table(dat$votechoice)
```

Multivariate Logistic Regression

1. Multivariate logistic regression allows us to estimate the effect that many independent variables have on one binary dependent variable. Example, what if we wanted to predict someone voting for Trump.

```
mod <- lm(votechoice ~ age + income + religion + gender1 + race1 + ideology, data=dat)

summary(mod)
```

Figure 1 presents the R output for the model that was estimated.

- The first aspect of the table to notice is either the t-statistics or the p-values. Remember, if the t statistic is less than -1.96 or greater than 1.96 the variable is significant. Further, if the p-value is smaller than 0.05 the variable is significant.
- If a variable is significant, you can move on to the directionality of the coefficient. If the variable is not significant, the investigation of the variable stops there.
- **Important**, we cannot directly interpret the coefficient in terms of the substantive impact of the number. For example, we can see that as political ideology increases (becomes more conservative) the likelihood of voting for Trump increases. BUT, we cannot say for certain the exact impact based on the number in the table.
- Let us also explore race (a nominal level variable), Here, we can see that all other races were less likely to vote for Trump over Clinton when compared to white respondents.
- Interpret the rest of the coefficients.

Figure 1: Regression Model Predicting Political Ideology

```
> summary(mod)

Call:
glm(formula = votechoice ~ age + gender1 + income + religion +
    race1 + ideology, family = binomial, data = dat)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.7737  -0.4759  -0.1335   0.4509   3.8339

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)      -4.567446   0.096832 -47.169  < 2e-16 ***
age               0.002886   0.001078   2.678  0.00741 **
gender1          -0.308208   0.032932  -9.359  < 2e-16 ***
income           -0.009944   0.005075  -1.959  0.05009 .
religionMormon    0.050205   0.148866   0.337  0.73593
religionOrthodox -0.033339   0.197224  -0.169  0.86576
religionJewish   -0.852580   0.097446  -8.749  < 2e-16 ***
religionOther    -1.045299   0.152233  -6.866 6.58e-12 ***
religionNothing  -0.294967   0.039176  -7.529 5.10e-14 ***
religionAtheist  -1.036426   0.085726 -12.090  < 2e-16 ***
race1Black       -3.423834   0.085327 -40.126  < 2e-16 ***
race1Hispanic    -1.166180   0.063849 -18.265  < 2e-16 ***
race1Other       -0.378779   0.065647  -5.770 7.93e-09 ***
ideology          1.173541   0.013431  87.377  < 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: 46546  on 33802  degrees of freedom
Residual deviance: 24805  on 33789  degrees of freedom
  (30797 observations deleted due to missingness)
AIC: 24833

Number of Fisher Scoring iterations: 6
```

In your research paper, the model should be presented in the manner below. You should not be presenting R output in your paper.

```
stargazer(mod)
```

Table 1: Logistic Regression Predicting Voting for Trump over Clinton

(Intercept)	-4.57*
	(0.10)
Age	0.00*
	(0.00)
Woman	-0.31*
	(0.03)
Family Income	-0.01
	(0.01)
Religion - Mormon	0.05
	(0.15)
Religion - Orthodox	-0.03
	(0.20)
Religion - Jewish	-0.85*
	(0.10)
Religion - Other	-1.05*
	(0.15)
Religion - Nothing	-0.29*
	(0.04)
Religion - Atheist	-1.04*
	(0.09)
Race - Black	-3.42*
	(0.09)
Race - Hispanic	-1.17*
	(0.06)
Race - Other	-0.38*
	(0.07)
Political Ideology	1.17*
	(0.01)
N	33803
AIC	24832.76
BIC	25304.75
$\log L$	-12360.38

Standard errors in parentheses

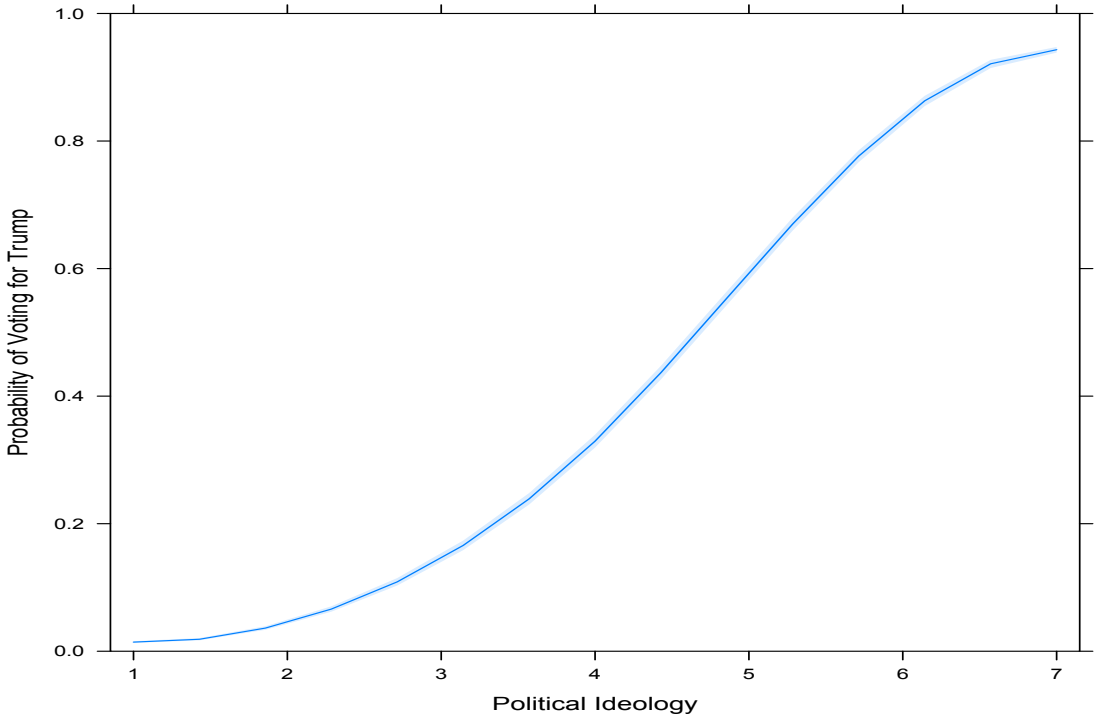
* indicates significance at $p < 0.05$

2. We could plot the predicted probabilities or substantive effect of the variables. For example, we could plot the predicted effect that political ideology had on voting for Trump.

```
eff <- effect("ideology", mod, default.levels=100, typical=median)
```

```
effect <- print(plot(eff, rescale.axis=F, rug=FALSE, xlab="Political Ideology",
ylab="Probability of Voting for Trump", main="", ylim=c(0,1)))
```

Figure 2: Effect of Ideology on Voting for Trump



Multivariate Ordered Logistic Regression

3. What about if we had a dependent variable that was ordered, but not continuous. Here, we would utilize ordered logistic regression. Say that we wanted to know how important immigration were to the respondents, but that the variable isn't exactly continuous, but instead ordered.

```
dat$immimp <- recode(dat$CC16_301d, "1='High'; 2='Somewhat High'; 3='Somewhat Low';
4='Very Low'; 5='None'")

dat$immimp <- factor(dat$immimp, levels=c("None", "Very Low", "Somewhat Low", "Somewhat
High", "High"))
```

Figure 3: Ordered Model Predicting Importance of Immigration

```
> summary(mod2)

Re-fitting to get Hessian

Call:
polr(formula = immimp ~ age + gender1 + income + religion + race1 +
    ideology, data = dat)

Coefficients:
                   Value Std. Error t value
age                0.01534   0.001320 11.6195
gender1            0.10027   0.038484  2.6056
income            -0.01917   0.005951 -3.2210
religionMormon    -0.09470   0.154457 -0.6131
religionOrthodox  -0.07537   0.255475 -0.2950
religionJewish    -0.03640   0.098690 -0.3688
religionOther      0.15063   0.146892  1.0254
religionNothing   -0.26669   0.047040 -5.6693
religionAtheist   -0.39999   0.073115 -5.4707
race1Black        -0.47805   0.073365 -6.5161
race1Hispanic      0.39742   0.105688  3.7603
race1Other        -0.03340   0.084175 -0.3968
ideology           0.36072   0.011678 30.8874

Intercepts:
                            Value    Std. Error t value
None|Very Low               -1.9739   0.1257   -15.7068
Very Low|Somewhat Low       -0.5493   0.1122    -4.8957
Somewhat Low|Somewhat High   0.9188   0.1100     8.3510
Somewhat High|High           2.4416   0.1124    21.7190

Residual Deviance: 24738.95
AIC: 24772.95
(54217 observations deleted due to missingness)
```

```
stargazer(mod2)
```

Table 2: Importance of Immigration

Age	0.015***
	(0.001)
Woman	0.100***
	(0.038)
Income	−0.019***
	(0.006)
Mormon	−0.095
	(0.154)
Orthodox	−0.075
	(0.255)
Jewish	−0.036
	(0.099)
Other	0.151
	(0.147)
Nothing	−0.267***
	(0.047)
Atheist	−0.400***
	(0.073)
Black	−0.478***
	(0.073)
Hispanic	0.397***
	(0.106)
Other	−0.033
	(0.084)
Political Ideology	0.361***
	(0.012)
None—Very Low	-1.974*** (0.126)
Very Low—Somewhat Low	-0.549*** (0.112)
Somewhat Low—Somewhat High	0.919*** (0.110)
Somewhat High—High	2.442*** (0.112)
Observations	10,383
AIC	24772.95
Note:	*p<0.1; **p<0.05; ***p<0.01

Lab Activity

In the 2020 Finland European Social Survey dataset, find a binary variable that you are interested in. Recode the variable and estimate a model with socio-demographic variables as the predictors. Next, find an ordinal level variable, recode it, and estimate a similar model. Interpret the results.