Multiple 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"))
```

Multivariate Linear Regression

1. Multivariate linear regression allows us to estimate the effect that many independent variables have on one continuous dependent variables. Example, what if we wanted to know all of the variables that impact political ideology.

```
mod <- lm(ideology ~ age + income + religion + gender1 + race1, 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 estimate or coefficient. If the variable is not significant, the investigation of the variable stops there.

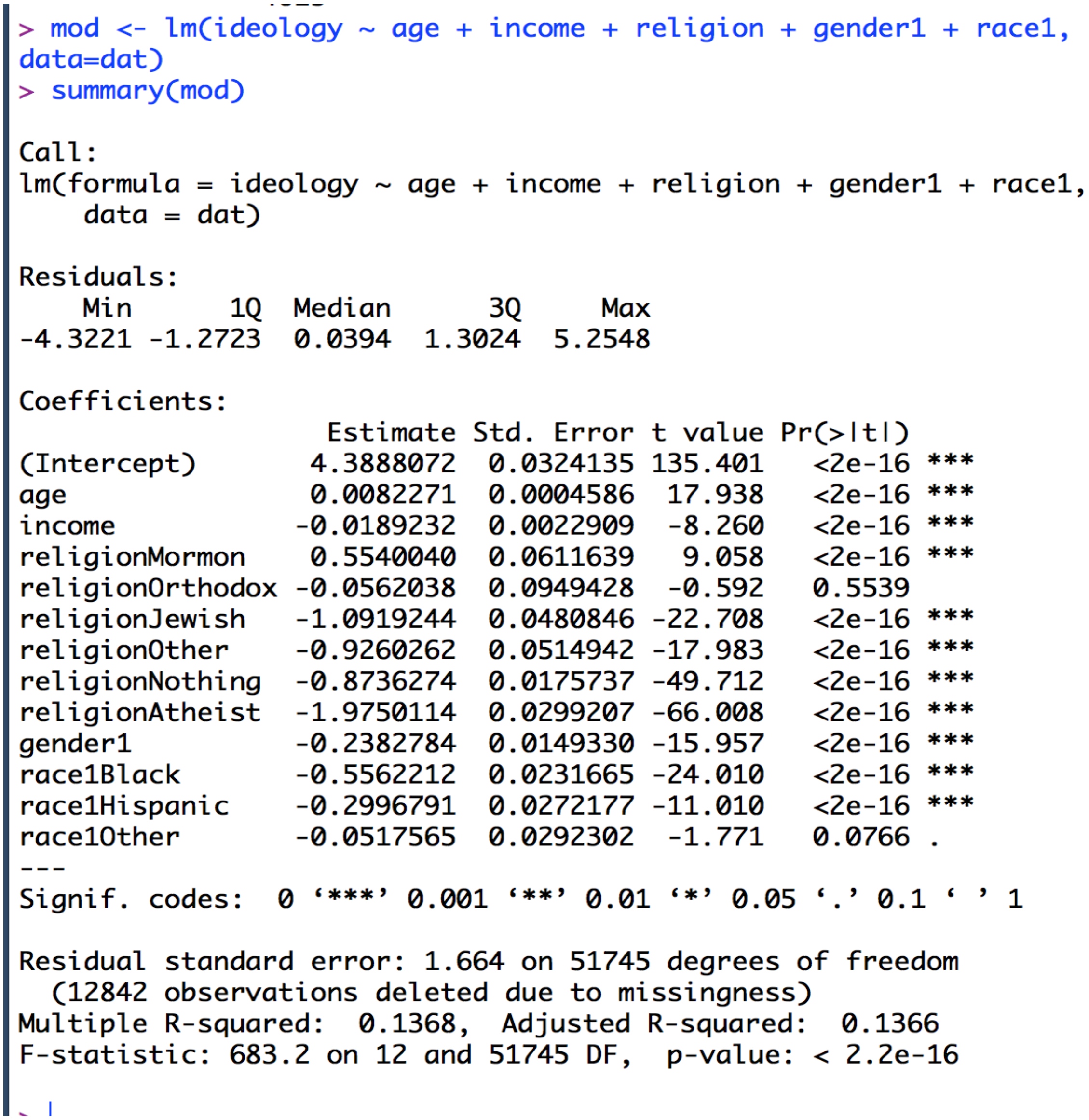
- First, if the intercept is statistically significant, we interpret the estimate. The intercept tells us that when all other independent variables are at zero, a respondent's ideology will be at 4.38. What does this mean substantively?

- Then, we can start looking at specific coefficients. Let us explore income (a ratio-interval level variable), remember the income variable is measured on a scale from 1-15 and the ideology variables is measured from 1-7. The income coefficient indicates to us that when a respondent increases in one category of income they decrease in ideology by .02.

- Let us also explore religion (a nominal level variable), Here, each religious category/coefficient is in comparison to the Protestant/Catholic category we created. Therefore, we are only testing whether each religious category is statistically different on ideology in comparison to Protestants/Catholics. Here, we see that Mormons are statistically more conservative, Orthodox Americans are statistically the same in terms of ideology, and Jewish, Other, Nothing, and Atheists are statistically more liberal.

- Interpret the rest of the coefficients.

Figure 1: Regression Model Predicting Political Ideology



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

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

Residuals:
    Min      1Q  Median      3Q     Max
-4.3221 -1.2723  0.0394  1.3024  5.2548

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       4.3888072  0.0324135 135.401   <2e-16 ***
age               0.0082271  0.0004586  17.938   <2e-16 ***
income           -0.0189232  0.0022909  -8.260   <2e-16 ***
religionMormon    0.5540040  0.0611639   9.058   <2e-16 ***
religionOrthodox -0.0562038  0.0949428  -0.592   0.5539
religionJewish   -1.0919244  0.0480846 -22.708   <2e-16 ***
religionOther    -0.9260262  0.0514942 -17.983   <2e-16 ***
religionNothing  -0.8736274  0.0175737 -49.712   <2e-16 ***
religionAtheist  -1.9750114  0.0299207 -66.008   <2e-16 ***
gender1          -0.2382784  0.0149330 -15.957   <2e-16 ***
race1Black       -0.5562212  0.0231665 -24.010   <2e-16 ***
race1Hispanic    -0.2996791  0.0272177 -11.010   <2e-16 ***
race1Other       -0.0517565  0.0292302  -1.771   0.0766 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.664 on 51745 degrees of freedom
  (12842 observations deleted due to missingness)
Multiple R-squared:  0.1368,	Adjusted R-squared:  0.1366
F-statistic: 683.2 on 12 and 51745 DF,  p-value: < 2.2e-16
```

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

```
install.packages("stargazer")
library(stargazer)
stargazer(mod)
```

Table 1: Multivariate Regression Predicting Political Ideology

(Intercept)	4.39*
	(0.03)
Age	0.01*
	(0.00)
Woman	-0.24*
	(0.01)
Income	-0.02*
	(0.00)
Religion - Mormon	0.55*
	(0.06)
Religion - Orthodox	-0.06
	(0.09)
Religion - Jewish	-1.09*
	(0.05)
Religion - Other	-0.93*
	(0.05)
Religion - Nothing	-0.87*
	(0.02)
Religion - Atheist	-1.98*
	(0.03)
Race - Black	-0.56*
	(0.02)
Race - Hispanic	-0.30*
	(0.03)
Race - Other	-0.05
	(0.03)
N	51758
R^2	0.14
adj. R^2	0.14
Resid. sd	1.66

Standard errors in parentheses

* indicates significance at $p < 0.05$

Plotting Effects

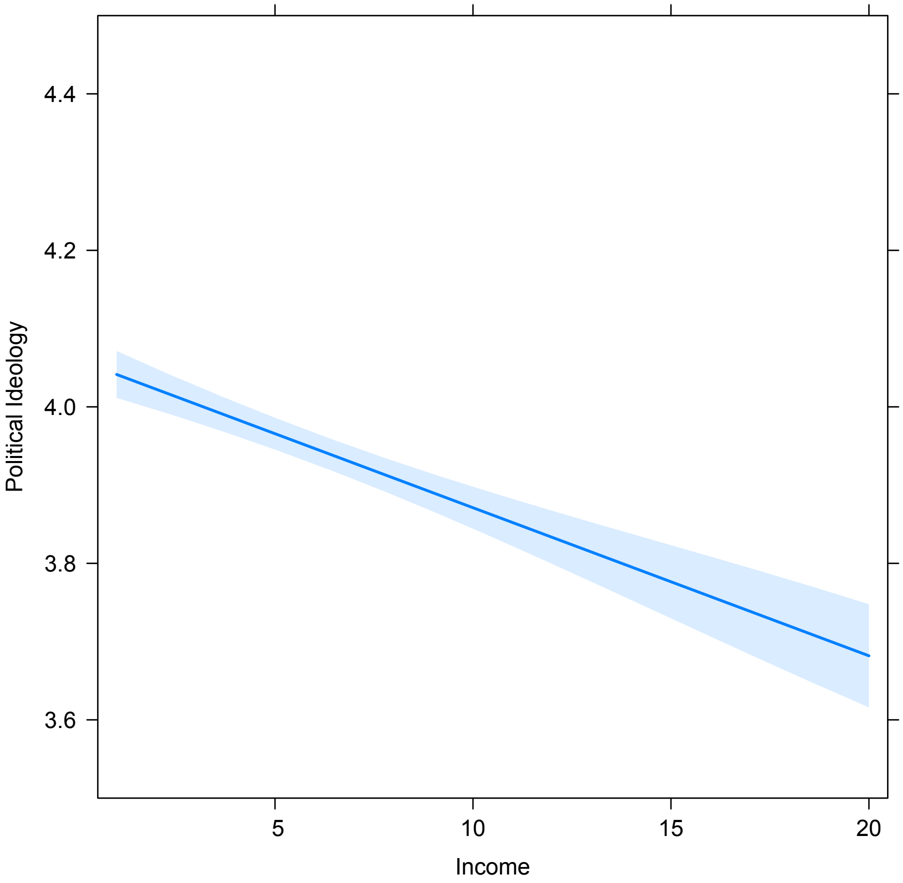
1. It is possible for a variable to be statistically significant, but substantively unimportant. This result is especially common when there is a lot of data. Thus, it is important to visually show the substantive effect of an independent variable on a dependent variable.

```
install.packages("effects")
library(effects)

eff1 <- effect("income", mod, default.levels=100, typical=median)

effect1 <- print(plot(eff1, rescale.axis=F, rug=FALSE, xlab="Income", ylab=
"Political Ideology", main="", ylim=c(3.5,4.5)))
```

Figure 2: The Effect of Income on Political Ideology



Lab Activity

In the 2020 Finland European Social Survey dataset, find any continuous variable that you find interesting and might want to investigate. Using all of the socio-demographic variables we have recoded in the past, and any independent variables you find relevant, estimate a multiple regression model. Plot the effect of the main independent variable of interest on the dependent variable.