Preliminary Analysis of ANES Data

Patrick

2025-03-19

Quarto

Quarto enables you to weave together content and executable code into a finished document. To learn more about Quarto see https://quarto.org. Hello this is nice. We can write some code. Add some comments.

$$\bar{x} = 1/N \sum x_i$$

Running Code

When you click the **Render** button a document will be generated that includes both content and the output of embedded code. You can embed code like this:

```
# this is a usual comment

1 + 1
```

```
[1] 2
```

```
2 + 4
```

```
[1] 6
```

You can add options to executable code like this

```
[1] 4
```

The `echo: false` option disables the printing of code (only output is displayed).

Load Packages and Data

First we load our packages. We'll add more later.

```
library(here)
library(broom)
library(tidyverse)
library(modelsummary)
library(marginaleffects)
library(lmtest)
library(sandwich)
```

Now let's load our data. We will use the script that we have already written.

```
source(here("code/01-load_anes.R"))

anes
```

```
# A tibble: 8,280 x 13
       id mode    age female   edu college social misinfo_russia
    <dbl> <fct> <dbl>  <dbl> <dbl>   <dbl>  <dbl>          <dbl>
 1 200015 Web      46      0     6       1  0.875              1
 2 200022 Web      37      1     3       0  0.375             -1
 3 200039 Web      40      1     2       0  0.5               -1
 4 200046 Web      41      0     4       0  0.5               -1
 5 200053 Web      72      0     8       1  0.125             -1
 6 200060 Web      71      1     3       0  0                 -1
 7 200084 Web      37      1     4       0  0.375              1
 8 200091 Web      45      1     2       0  0.25              -1
 9 200107 Web      70      1     2       0  0.5               -1
10 200114 Web      43      0     4       0  0.875              1
# i 8,270 more rows
# i 5 more variables: confident_russia <dbl>, misconf_russia <dbl>,
#   misinfo_warm <dbl>, confident_warm <dbl>, misconf_warm <dbl>
```

Presenting Regression Results

First we have to run a few regressions.

```r
m1a <- lm(misconf_russia ~ social + college +
            age + female + mode, data = anes)
summary(m1a)
```

```
Call:
lm(formula = misconf_russia ~ social + college + age + female +
    mode, data = anes)

Residuals:
    Min      1Q  Median      3Q     Max
-0.9979 -0.4819 -0.1756  0.3961  1.6045

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.0678304  0.0354383   1.914   0.0557 .
social      -0.1700124  0.0364281  -4.667 3.11e-06 ***
college     -0.2116326  0.0145857 -14.510  < 2e-16 ***
age         -0.0044224  0.0004877  -9.069  < 2e-16 ***
female       0.0140873  0.0144537   0.975   0.3298
modePhone    0.0490508  0.0582347   0.842   0.3997
modeVideo   -0.0784512  0.0373307  -2.102   0.0356 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5963 on 6885 degrees of freedom
  (1388 observations deleted due to missingness)
Multiple R-squared:  0.04497,   Adjusted R-squared:  0.04414
F-statistic: 54.03 on 6 and 6885 DF,  p-value: < 2.2e-16
```

```r
m1b <- lm(misconf_russia ~ social + I(social^2) + college +
            age + female + mode, data = anes)
summary(m1b)
```

```
Call:
lm(formula = misconf_russia ~ social + I(social^2) + college +
    age + female + mode, data = anes)

Residuals:
    Min      1Q  Median      3Q     Max
```

```
-0.9723 -0.4790 -0.1787  0.3940  1.6012

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.0446459  0.0371885   1.201   0.2300
social       0.0316291  0.1048123   0.302   0.7628
I(social^2) -0.2498675  0.1217871  -2.052   0.0402 *
college     -0.2126240  0.0145903 -14.573   <2e-16 ***
age         -0.0044698  0.0004881  -9.158   <2e-16 ***
female       0.0125818  0.0144690   0.870   0.3846
modePhone    0.0528310  0.0582503   0.907   0.3645
modeVideo   -0.0805521  0.0373360  -2.157   0.0310 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5962 on 6884 degrees of freedom
  (1388 observations deleted due to missingness)
Multiple R-squared:  0.04555,   Adjusted R-squared:  0.04458
F-statistic: 46.94 on 7 and 6884 DF,  p-value: < 2.2e-16
```

```r
m1c <- lm(misconf_russia ~ social * college + I(social^2) * college +
            age + female + mode, data = anes)
summary(m1c)
```

```
Call:
lm(formula = misconf_russia ~ social * college + I(social^2) *
    college + age + female + mode, data = anes)

Residuals:
    Min      1Q  Median      3Q     Max
-0.9702 -0.4753 -0.1705  0.3958  1.6010

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)         0.0050335  0.0396223   0.127  0.89891
social              0.2731156  0.1378411   1.981  0.04759 *
college            -0.1077620  0.0390652  -2.759  0.00582 **
I(social^2)        -0.4985179  0.1658486  -3.006  0.00266 **
age                -0.0044948  0.0004881  -9.209  < 2e-16 ***
female              0.0137248  0.0144742   0.948  0.34305
modePhone           0.0563612  0.0582381   0.968  0.33319
```

```
modeVideo           -0.0799365  0.0373191  -2.142  0.03223 *
social:college      -0.5860933  0.2100605  -2.790  0.00528 **
college:I(social^2)  0.5910766  0.2450554   2.412  0.01589 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5959 on 6882 degrees of freedom
  (1388 observations deleted due to missingness)
Multiple R-squared:  0.04673,   Adjusted R-squared:  0.04548
F-statistic: 37.48 on 9 and 6882 DF,  p-value: < 2.2e-16
```

Regression Tables

```
modelsummary(list(m1a, m1b, m1c),
             coef_map = c("social" = "Social Media Exposure",
                          "college" = "College Education",
                          "social:college" = "Social Media X College",
                          "I(social^2)" = "Social Media Squared",
                          "college:I(social^2)" = "Social Media Squared X College",
                          "age" = "Age",
                          "female" = "Gender (Female)",
                          "modePhone" = "Survey Mode (Phone)",
                          "modeVideo" = "Survey Mode (Video)",
                          "(Intercept)" = "Constant"),
             gof_map = c("nobs", "r.squared", "rmse"),
             vcov = "HC3")
```

Working with multiple models / loops over models

```
models <- anes %>%
  group_by(female, mode) %>%
  do(Model = lm(misconf_russia ~ social * college + I(social^2) * college + age,
                data = .))

models
```

```
# A tibble: 7 x 3
# Rowwise:
```

	(1)	(2)	(3)
Social Media Exposure	−0.170	0.032	0.273
	(0.037)	(0.107)	(0.145)
College Education	−0.212	−0.213	−0.108
	(0.015)	(0.015)	(0.040)
Social Media X College			−0.586
			(0.213)
Social Media Squared		−0.250	−0.499
		(0.122)	(0.173)
Social Media Squared X College			0.591
			(0.245)
Age	−0.004	−0.004	−0.004
	(0.000)	(0.000)	(0.000)
Gender (Female)	0.014	0.013	0.014
	(0.015)	(0.015)	(0.015)
Survey Mode (Phone)	0.049	0.053	0.056
	(0.066)	(0.066)	(0.066)
Survey Mode (Video)	−0.078	−0.081	−0.080
	(0.035)	(0.035)	(0.035)
Constant	0.068	0.045	0.005
	(0.036)	(0.038)	(0.041)
Num.Obs.	6892	6892	6892
R2	0.045	0.046	0.047
RMSE	0.60	0.60	0.60

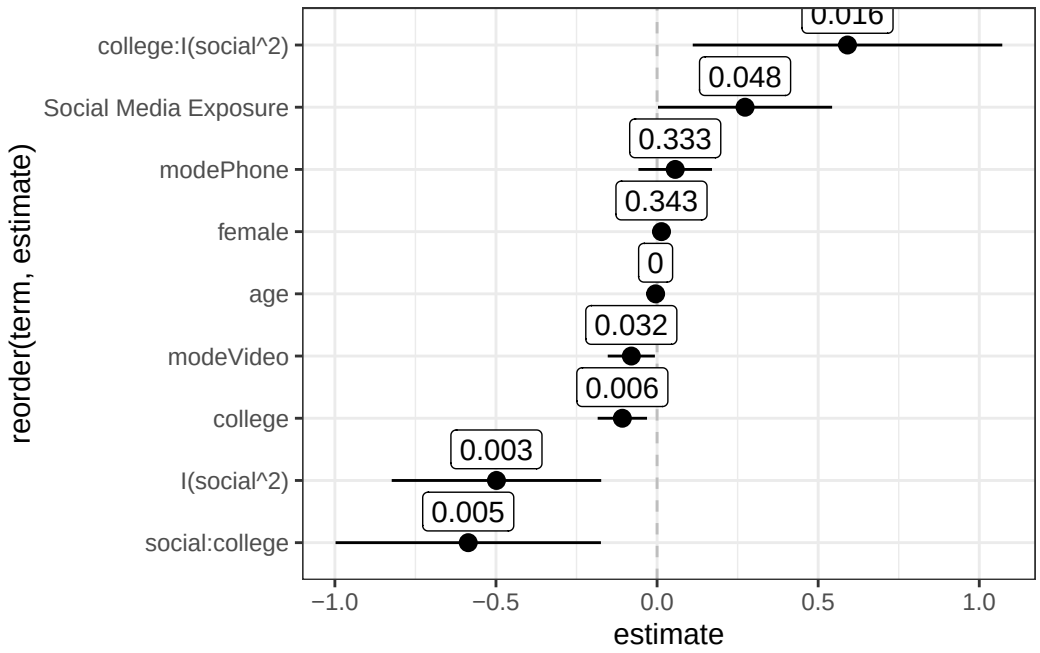
```
  female mode  Model
   <dbl> <fct> <list>
1      0 Web   <lm>
2      0 Phone <lm>
3      0 Video <lm>
4      1 Web   <lm>
5      1 Phone <lm>
6      1 Video <lm>
7     NA Web   <lm>
```

```
summary(models$Model[[1]])
```

```

Call:
lm(formula = misconf_russia ~ social * college + I(social^2) *
    college + age, data = .)

Residuals:
    Min      1Q  Median      3Q     Max
-0.9265 -0.5060 -0.2131  0.4462  1.5643

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)         -0.0406221  0.0599515  -0.678   0.4981
social               0.2527750  0.2098384   1.205   0.2284
college             -0.1181033  0.0560866  -2.106   0.0353 *
I(social^2)         -0.4990996  0.2488951  -2.005   0.0450 *
age                 -0.0031954  0.0007635  -4.185 2.93e-05 ***
social:college      -0.6982970  0.3139748  -2.224   0.0262 *
college:I(social^2)  0.7136796  0.3718227   1.919   0.0550 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6152 on 2997 degrees of freedom
  (543 observations deleted due to missingness)
Multiple R-squared:  0.04678,   Adjusted R-squared:  0.04487
F-statistic: 24.51 on 6 and 2997 DF,  p-value: < 2.2e-16
```

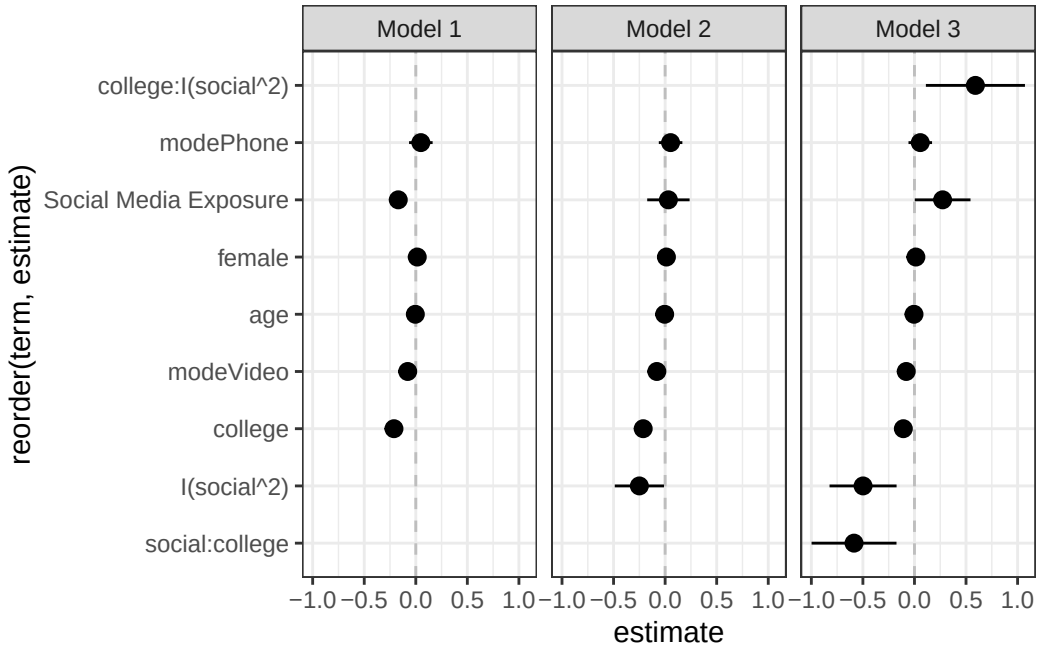
Coefficient Plots

```r
m1c %>%
  tidy(conf.int = T) %>%
  filter(term != "(Intercept)") %>%
  mutate(term = recode(term,
                       `social` = "Social Media Exposure")) %>%
  ggplot(aes(x = estimate, xmin = conf.low, xmax = conf.high,
             y = reorder(term, estimate))) +
  geom_vline(xintercept = 0, lty = "dashed", col = "grey") +
  geom_pointrange() +
  geom_label(aes(label = round(p.value, 3)), nudge_y = .5) +
  theme_bw()
```

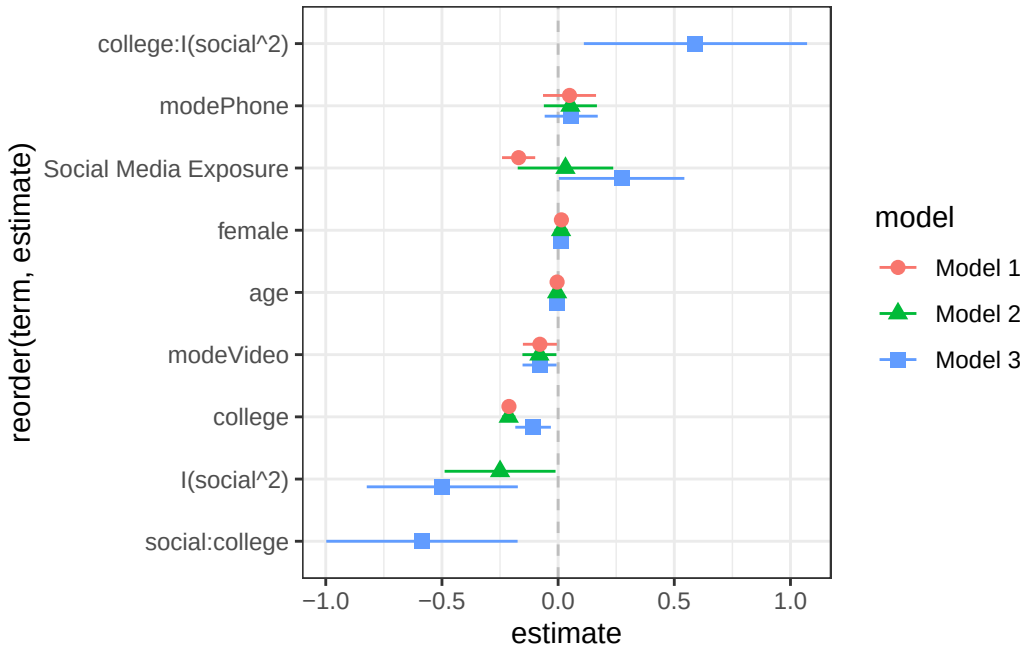



```r
list(
  "Model 1" = m1a,
  "Model 2" = m1b,
  "Model 3" = m1c
) %>%
  map_dfr(tidy, conf.int = T, .id = "model") %>%
  filter(term != "(Intercept)") %>%
  mutate(term = recode(term,
                       `social` = "Social Media Exposure")) %>%
  ggplot(aes(x = estimate, xmin = conf.low, xmax = conf.high,
             y = reorder(term, estimate))) +
```

```
  geom_vline(xintercept = 0, lty = "dashed", col = "grey") +
  geom_pointrange() +
  theme_bw() +
  facet_wrap(~model)
```

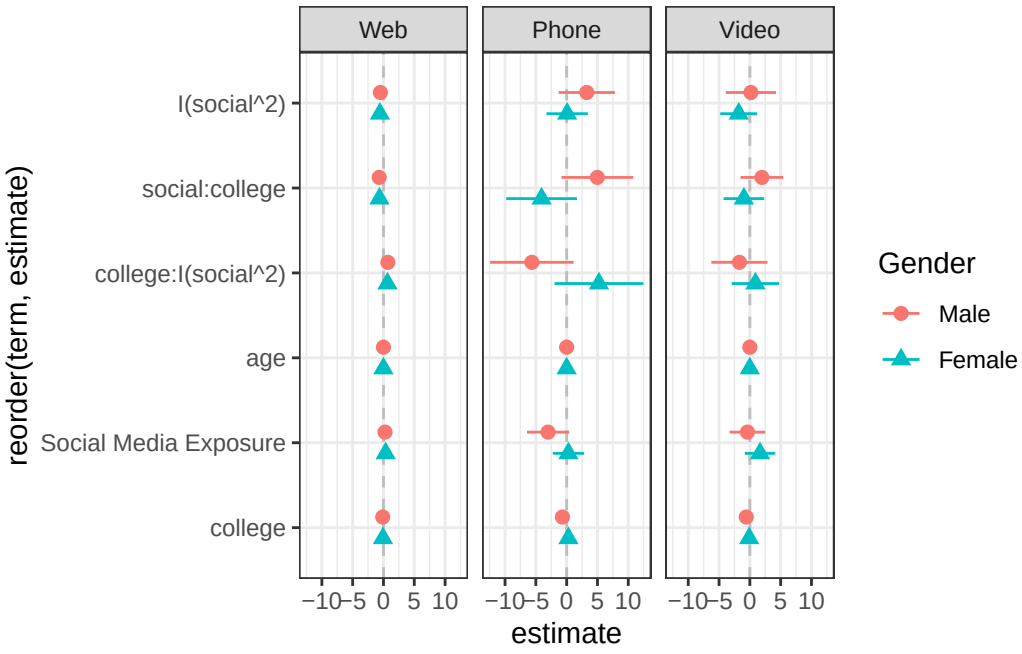



```
list(
  "Model 1" = m1a,
  "Model 2" = m1b,
  "Model 3" = m1c
) %>%
  map_dfr(tidy, conf.int = T, .id = "model") %>%
  filter(term != "(Intercept)") %>%
  mutate(term = recode(term,
                       `social` = "Social Media Exposure")) %>%
  ggplot(aes(x = estimate, xmin = conf.low, xmax = conf.high,
             y = reorder(term, estimate),
             shape = model,
             col = model)) +
  geom_vline(xintercept = 0, lty = "dashed", col = "grey") +
  geom_pointrange(position = position_dodge2(width = .5, reverse = T)) +
  theme_bw()
```




```
models %>%
  mutate(Model = list(tidy(Model, conf.int = T))) %>%
  unnest() %>%
  na.omit() %>%
  filter(term != "(Intercept)") %>%
  mutate(term = recode(term,
                       `social` = "Social Media Exposure"),
         Gender = recode_factor(female,
                                `0` = "Male",
                                `1` = "Female")) %>%
  ggplot(aes(x = estimate, xmin = conf.low, xmax = conf.high,
             y = reorder(term, estimate),
             shape = Gender,
             col = Gender)) +
  geom_vline(xintercept = 0, lty = "dashed", col = "grey") +
  geom_pointrange(position = position_dodge2(width = .5, reverse = T)) +
  theme_bw() +
  facet_wrap(~mode, ncol = 3)
```

```
Warning: `cols` is now required when using `unnest()`.
i Please use `cols = c(Model)`.
```



Robust Standard Errors and such

```
bptest(m1a)
```

```
    studentized Breusch-Pagan test

data:  m1a
BP = 66.525, df = 6, p-value = 2.105e-12
```

```
coeftest(m1a, vcov = vcovHC(m1a, type = "HC2")) %>%
  tidy()
```

```
# A tibble: 7 x 5
  term        estimate std.error statistic  p.value
  <chr>          <dbl>     <dbl>     <dbl>    <dbl>
1 (Intercept)  0.0678   0.0355       1.91  5.62e- 2
2 social      -0.170    0.0367      -4.63  3.79e- 6
3 college     -0.212    0.0146     -14.5   3.90e-47
4 age         -0.00442  0.000485    -9.13  9.04e-20
5 female       0.0141   0.0145       0.971 3.32e- 1
```

```
6 modePhone    0.0491   0.0656       0.747 4.55e- 1
7 modeVideo   -0.0785   0.0350      -2.24  2.51e- 2
```

```
# help(package = "sandwich")
```

Testing multiple restrictions and other hypotheses

```
m1a
```

```

Call:
lm(formula = misconf_russia ~ social + college + age + female +
    mode, data = anes)

Coefficients:
(Intercept)       social      college          age       female    modePhone
   0.067830    -0.170012    -0.211633    -0.004422     0.014087     0.049051
  modeVideo
  -0.078451
```

```
m1c
```

```

Call:
lm(formula = misconf_russia ~ social * college + I(social^2) *
    college + age + female + mode, data = anes)

Coefficients:
        (Intercept)              social             college
           0.005034            0.273116           -0.107762
        I(social^2)                 age              female
          -0.498518           -0.004495            0.013725
          modePhone           modeVideo      social:college
           0.056361           -0.079936           -0.586093
college:I(social^2)
           0.591077
```

```r
anova(m1a, m1c)
```

```
Analysis of Variance Table

Model 1: misconf_russia ~ social + college + age + female + mode
Model 2: misconf_russia ~ social * college + I(social^2) * college + age +
    female + mode
  Res.Df    RSS Df Sum of Sq      F   Pr(>F)
1   6885 2448.2
2   6882 2443.7  3    4.5041 4.2282 0.005398 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
hypotheses(m1a, "social = college")
```

```
     Hypothesis Estimate Std. Error    z Pr(>|z|)   S   2.5 % 97.5 %
 social=college   0.0416     0.0409 1.02    0.309 1.7 -0.0386  0.122
```

```r
hypotheses(m1a, "social = -female")
```

```
     Hypothesis Estimate Std. Error     z Pr(>|z|)    S  2.5 %  97.5 %
 social=-female   -0.156     0.0383 -4.07   <0.001 14.4 -0.231 -0.0808
```

```r
hypotheses(m1a, "exp(college + female) = 0.1", vcov = "HC2")
```

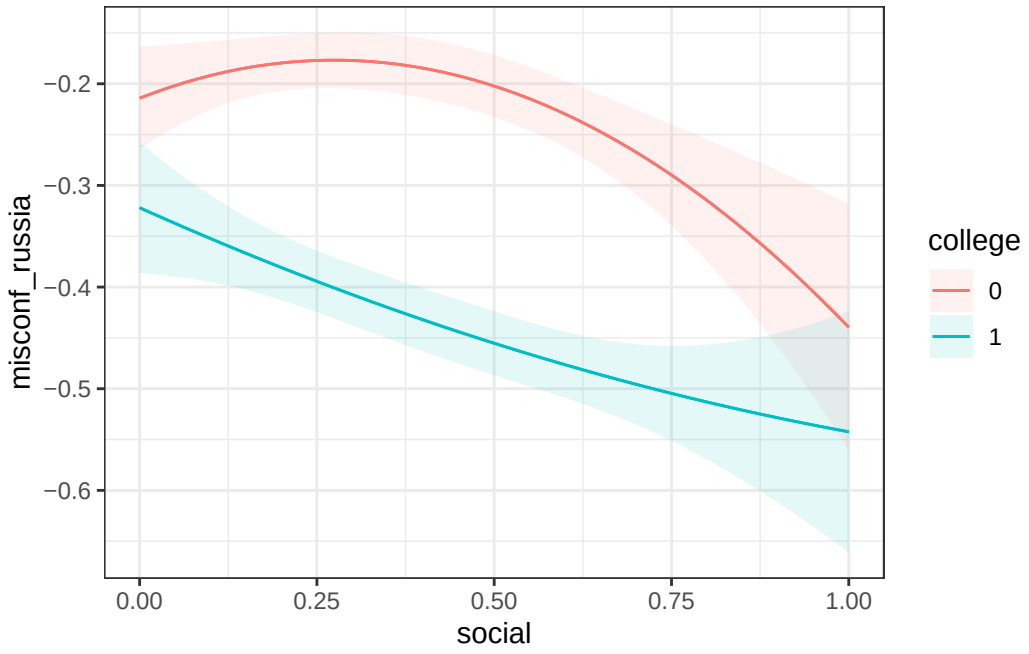
```
             Hypothesis Estimate Std. Error    z Pr(>|z|)   S 2.5 % 97.5 %
 exp(college+female)=0.1    0.721     0.0174 41.5   <0.001 Inf 0.687  0.755
```

Computing expected values and marginal effects

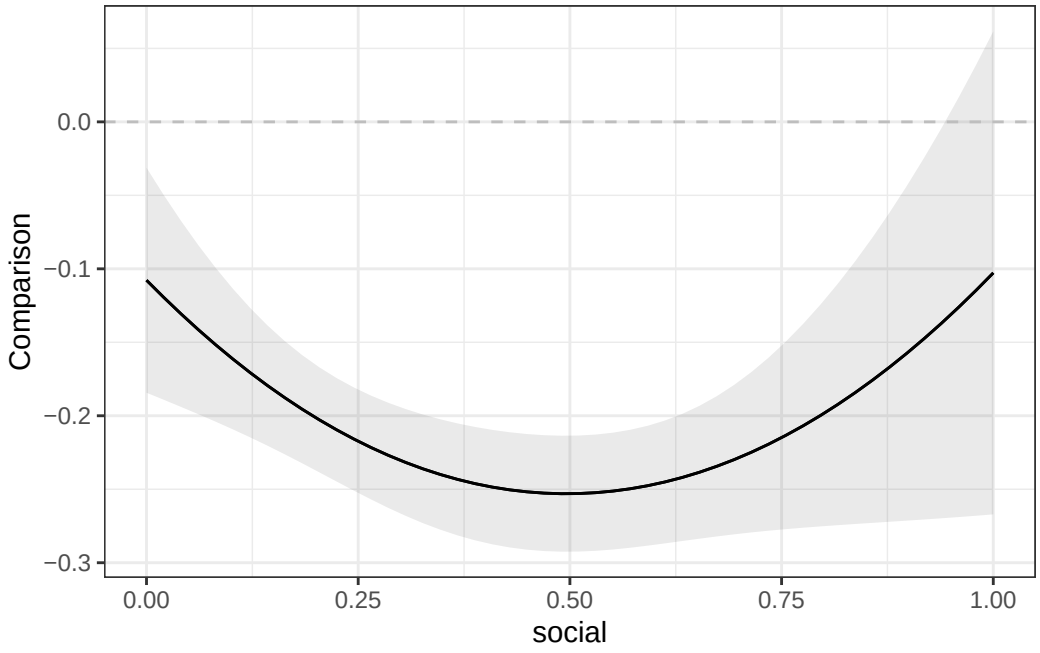
```r
summary(m1c)
```

```
Call:
lm(formula = misconf_russia ~ social * college + I(social^2) *
    college + age + female + mode, data = anes)

Residuals:
    Min      1Q  Median      3Q     Max
-0.9702 -0.4753 -0.1705  0.3958  1.6010

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)          0.0050335  0.0396223   0.127  0.89891
social               0.2731156  0.1378411   1.981  0.04759 *
college             -0.1077620  0.0390652  -2.759  0.00582 **
I(social^2)         -0.4985179  0.1658486  -3.006  0.00266 **
age                 -0.0044948  0.0004881  -9.209  < 2e-16 ***
female               0.0137248  0.0144742   0.948  0.34305
modePhone            0.0563612  0.0582381   0.968  0.33319
modeVideo           -0.0799365  0.0373191  -2.142  0.03223 *
social:college      -0.5860933  0.2100605  -2.790  0.00528 **
college:I(social^2)  0.5910766  0.2450554   2.412  0.01589 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5959 on 6882 degrees of freedom
  (1388 observations deleted due to missingness)
Multiple R-squared:  0.04673,   Adjusted R-squared:  0.04548
F-statistic: 37.48 on 9 and 6882 DF,  p-value: < 2.2e-16
```

```r
plot_predictions(m1c, condition = c("social", "college")) +
  theme_bw()
```




```
plot_comparisons(m1c, variables = "college",
                 condition = "social") +
  geom_hline(yintercept = 0, lty = "dashed", col = "grey") +
  theme_bw()
```



Going beyond linear models

Generalized Linear Models

```
anes$russia_dummy <- anes$misconf_russia > 0
m2a <- glm(russia_dummy ~ social + college + age + female,
           family = binomial("logit"), data = anes)
summary(m2a)
```

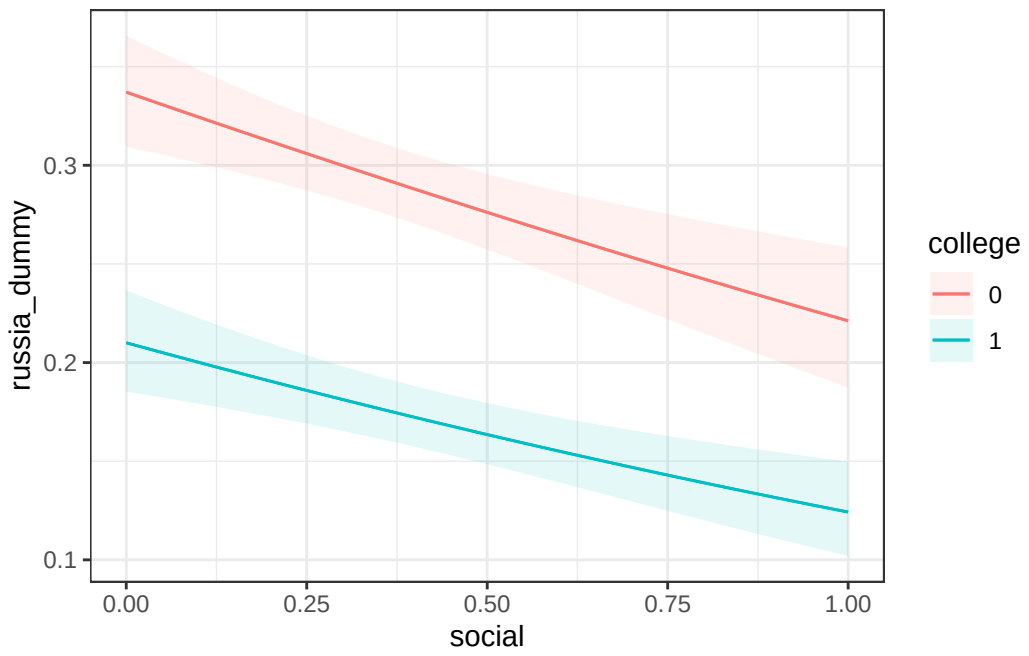
```

Call:
glm(formula = russia_dummy ~ social + college + age + female,
    family = binomial("logit"), data = anes)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.183720   0.136312  -1.348    0.178
social      -0.591546   0.143557  -4.121 3.78e-05 ***
college     -0.660760   0.059204 -11.161  < 2e-16 ***
age         -0.008313   0.001905  -4.364 1.27e-05 ***
female      -0.057032   0.057029  -1.000    0.317
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 7666.6  on 6891  degrees of freedom
Residual deviance: 7499.7  on 6887  degrees of freedom
  (1388 observations deleted due to missingness)
AIC: 7509.7

Number of Fisher Scoring iterations: 4
```

```
m2b <- glm(russia_dummy ~ social + college + age + female,
           family = binomial("probit"), data = anes)
summary(m2b)
```

```

Call:
glm(formula = russia_dummy ~ social + college + age + female,
    family = binomial("probit"), data = anes)
```

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.136278   0.080896  -1.685   0.0921 .
social      -0.347733   0.084330  -4.124 3.73e-05 ***
college     -0.385830   0.034213 -11.277  < 2e-16 ***
age         -0.004835   0.001123  -4.306 1.66e-05 ***
female      -0.033636   0.033489  -1.004   0.3152
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 7666.6  on 6891  degrees of freedom
Residual deviance: 7499.8  on 6887  degrees of freedom
  (1388 observations deleted due to missingness)
AIC: 7509.8

Number of Fisher Scoring iterations: 4
```

```r
plot_predictions(m2b, condition = c("social", "college")) +
  theme_bw()
```



Other useful packages for modeling

```r
## IV regression, robust regression and more
library(estimatr)
# help(package = "estimatr")

## Panel regression
library(plm)
```

```
Attaching package: 'plm'

The following objects are masked from 'package:dplyr':

    between, lag, lead
```

```r
# help(package = "plm")

## Mixed effects models
library(lme4)
```

```
Loading required package: Matrix


Attaching package: 'Matrix'

The following objects are masked from 'package:tidyr':

    expand, pack, unpack
```

Other resources

Multiple imputation with mice: https://amices.org/mice/

Drawing maps with ggmaps: https://socviz.co/maps.html

Time Series: https://cran.r-project.org/web/views/TimeSeries.html

Machine Learning: https://cran.r-project.org/web/views/MachineLearning.html