Post-estimation Simulation

POSC 3410 – Quantitative Methods in Political Science

Steven V. Miller

Department of Political Science

Goal for Today

Provide intuitive quantities of interest from your regression.

Readable Regression Tables

Remember: your analysis should be as easily interpretable as possible.

- I should get a preliminary glimpse of effect size from a regression.
- Your y-intercept should be meaningful.

Standardizing variables helps.

- Creates meaningful zeroes (i.e. the mean).
- Coefficients communicate magnitude changes in x.
- Standardizing by two SDs allows for easy comparison with binary predictors.

Satisfy Your Audience

You need to relate your analysis to both me and your grandma.

- I will obviously know/care more about technical details.
- Grandma may not, but she may be a more important audience than me.

Her inquiries are likely be understandable. Examples:

- What's the expected tolerance of abortion for an 18-year-old black man?
- What's the increased probability of voting for a Republican for an increase of $20k in yearly income?

These are perfectly reasonable questions to ask of your analysis.

- If your presentation isn't prepared to answer her questions, you're not doing your job.

Statistical Presentations

Statistical presentations should:

1. Convey precise estimates of quantities of interest.
2. Include reasonable estimates of *uncertainty* around those estimates.
3. Require little specialized knowledge to understand Nos. 1 and 2.
4. Not bombard the audience with superfluous information.

We will do this with post-estimation simulation using draws from a multivariate normal distribution (King et al. 2000).

Estimating Uncertainty with Simulation

Any statistical model has a stochastic and systematic component.

- **Stochastic**: $Y_i \sim f(y_i \mid \theta_i, \alpha)$
- **Systematic**: $\theta_i = g(x_i, \beta)$

For a simple OLS model (i.e. a linear regression):

$$\begin{aligned} Y_i &= N(\mu_i, \sigma^2) \\ \mu_i &= X_i\beta \end{aligned}$$

Understanding our Uncertainty

We have two types of uncertainty.

1. **Estimation uncertainty**
    - Represents systematic components; can be reduced by increasing sample size.
2. **Fundamental uncertainty**
    - Represents stochastic component; exists no matter what (but can be modeled).

Getting our Parameter Vector

We want a **simulated parameter vector**, denoted as:

$$\hat{\gamma} \sim vec(\hat{\beta}, \hat{\alpha})$$

Central limit theorem says with a large enough sample and bounded variance:

$$\tilde{\gamma} \sim N(\hat{\gamma}, \hat{V}(\hat{\gamma}))$$

In other words: distribution of quantities of interest will follow a multivariate normal distribution with mean equal to $\hat{\gamma}$, the simulated parameter vector.

Getting our Quantities of Interest

This is a mouthful! Let's break the process down step-by-step.

1. Run your regression. Get your results.
2. Choose values of explanatory variable (as you see fit).
3. Obtain simulated parameter vector from estimating systematic component.
4. Simulate the outcome by taking random draw from the stochastic component.

Do this m times (typically $m = 1000$) to estimate full probability distribution of Y_c.

- Expected value $E(Y_c)$ = predicted value for linear models. It just averages over the fundamental uncertainty.

An Application with Zelig

Don't worry! We have software to make this easier.

- We'll be using the Zelig package in R.

I'll also be using sample data from my Github page.

- Question: What explains who voted for Romney or Obama in the 2012 presidential election?

Understanding our Sample Data

I took the 2014 wave of GSS data on American public opinion.

- **voteromney**: 1 if respondent (r) voted for Romney; 0 if r voted for someone else.
  - I also have an Obama version of this variable.
- **age**: age of respondent in years (18:89).
  - Note: Regrettably, GSS stops coding year after 89+.
- **race**: three category nominal variable (white, black, other).
- **collegeed**: 1 if r has at least a bachelors; 0 if otherwise.
- **partyid**: strength of affiliation with Republican party (0:6).
  - 0 = strong Democrat; 6 = strong Republican
- **female**: 1 if r is a woman; 0 if r is a man.
- **rincom06**: Income categories (1:25).
  - 1 = under \$1,000 earned in 2013; 25 = \$150k or over.

A Sample Regression

```
M1 <- glm(voteromney ~ age + female + collegeed + white + black + partyid + rin
          data=Data, family=binomial(link="logit"))
summary(M1)
```

```
##
## Call:
## glm(formula = voteromney ~ age + female + collegeed + white +
##     black + partyid + rincom06, family = binomial(link = "logit"),
##     data = Data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5752  -0.3637  -0.0510   0.3929   3.9658
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -4.872606   0.701502  -6.946 3.76e-12 ***
## age          0.011054   0.008298   1.332 0.182832
## female      -0.285121   0.228392  -1.248 0.211890
## collegeed    0.154290   0.233329   0.661 0.508447
## white        0.773461   0.452916   1.708 0.087685 .
## black       -3.125368   0.924890  -3.379 0.000727 ***
## partyid      1.221582   0.078809  15.501  < 2e-16 ***
## rincom06    -0.024922   0.020709  -1.203 0.228796
##
```

Standardizing the Variables

```
M2 <- glm(voteromney ~ z.age + female + collegeed + white + black +
            z.partyid + z.rincom06, data=Data, family=binomial(link="logit"))
summary(M2)
```

```
##
## Call:
## glm(formula = voteromney ~ z.age + female + collegeed + white +
##     black + z.partyid + z.rincom06, family = binomial(link = "logit"),
##     data = Data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5752  -0.3637  -0.0510   0.3929   3.9658
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -1.4611     0.4683  -3.120 0.001808 **
## z.age         0.3849     0.2890   1.332 0.182832
## female       -0.2851     0.2284  -1.248 0.211890
## collegeed     0.1543     0.2333   0.661 0.508447
## white         0.7735     0.4529   1.708 0.087685 .
## black        -3.1254     0.9249  -3.379 0.000727 ***
## z.partyid     4.7011     0.3033  15.501  < 2e-16 ***
## z.rincom06   -0.3019     0.2509  -1.203 0.228796
##
```

Something Nana Might Ask

What's the effect of Republican party ID on black voters?

- Let's see!

First, we re-estimate the model in Zelig.

- We also have to subset for complete cases.
- Zelig doesn't play nice with missing data.

The Model in Zelig

```
M3 <- zelig(voteromney ~ age + female + collegeed + white +
                black + partyid + rincom06,
model = "logit", data = Data2)
```

Notice Zelig's post-estimation simulation doesn't require intuitive zeroes.

- You should still do it for a regression table, though.
- Again, not having intuitive zeroes can still break a more complicated model.

Answering Nana's Question

What's the probability of a black vote for Romney among strong Republicans? How about strong Democrats?

```
M3.bsdem <- setx(M3, black = 1, white = 0, partyid = 0)
M3.bsrep <- setx(M3, black = 1, white = 0, partyid = 6)

M3.sim <- sim(M3, x = M3.bsdem, x1 = M3.bsrep)
summary(M3.sim)
```

Answering Nana's Question

```
summary(M3.sim)
```

```
##
## Model:  logit
## Number of simulations:  1000
##
## Values of X
##   (Intercept)      age    female collegeed white black partyid rincom06
## 2           1 46.95863 0.5429162   0.42606     0     1       0 16.07963
## attr(,"assign")
## [1] 0 1 2 3 4 5 6 7
##
## Values of X1
##   (Intercept)      age    female collegeed white black partyid rincom06
## 2           1 46.95863 0.5429162   0.42606     0     1       6 16.07963
## attr(,"assign")
## [1] 0 1 2 3 4 5 6 7
##
## Expected Values: E(Y|X)
##   mean    sd 50% 2.5% 97.5%
##  0.001 0.001   0    0 0.002
##
## Expected Values: E(Y|X1)
##   mean    sd   50%  2.5% 97.5%
##  0.366 0.165 0.359 0.093 0.724
##
## Predicted Values: Y|X
##  0 1
##  1 0
##
## Predicted Values: Y|X1
##      0     1
##  0.633 0.367
##
## First Differences: E(Y|X1) - E(Y|X)
##   mean    sd   50%  2.5% 97.5%
##  0.366 0.165 0.358 0.093 0.721
```

Another Question from Nana

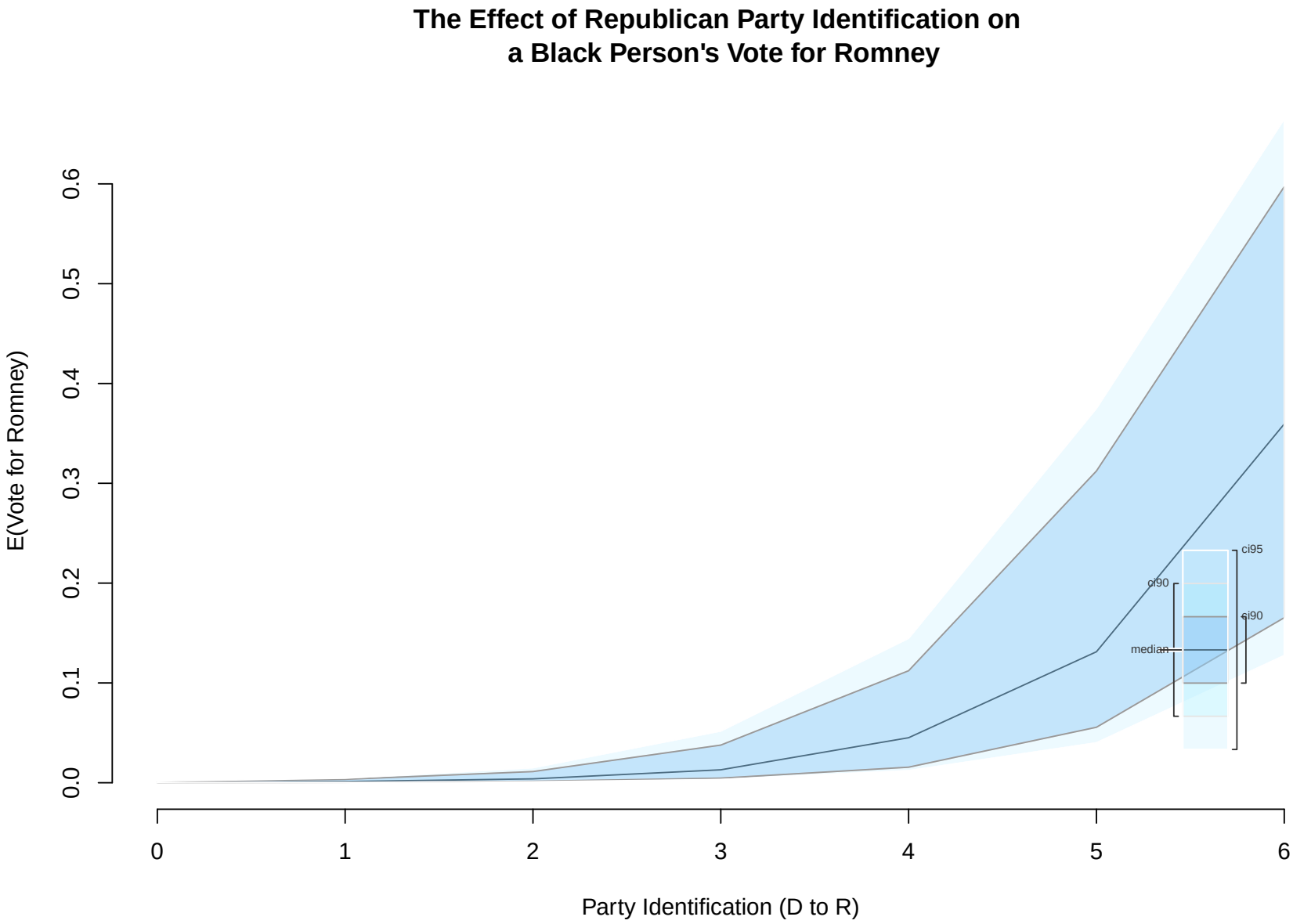
What's the effect of Republican party identification on Black voters?

```
M3.bsdemtosrep <- setx(M3, black = 1, white = 0, partyid = 0:6)

M3.sim2 <- sim(M3, x = M3.bsdemtosrep)

plot.ci(M3.sim2, xlab = "Party Identification (D to R)",
        ylab = "E(Vote for Romney)",
        main = "The Effect of Republican Party Identification
        on \n a Black Person's Vote for Romney",
        ci=c(90,95)
)
```

A Pretty Graph



What's Going on With the Black Vote among the Republicans?

Notice the huge confidence intervals for Black Republicans? What do you think is happening there?

There Aren't Many Black Republicans

```
table(Data$black, Data$partyid)
```

```
##
##         0   1   2   3   4   5   6
##   0 252 324 289 450 241 282 238
##   1 167  82  48  52   8  10   7
```

Note: Black respondents are the row; partyid is the column

- Also: 4 = "independent, near Republican"; 5 = "not a strong Republican"; 6 = "strong Republican"

There Weren't Many Black Voters for Romney Either

```
table(Data$black, Data$voteromney)
```

```
##
##         0   1
##   0 714 610
##   1 272   3
```

Note: Black respondents are the row; Romney votes are the column.

- If you're curious, two of those three black votes for Romney came from self-identified strong Republicans.
- One actually came from a self-identified strong Democrat!

The relative dearth of black Republicans and black Romney voters adds to the uncertainty in our predictions.

Conclusion

Regression provides all-else-equal effect sizes across the range of the data.

- You can extract meaningful quantities of interest from regression output itself.
- Typically, you'll need more to answer substantive questions and provide meaningful quantities of interest.

Post-estimation simulation from a multivariate normal distribution does this.

- When you start doing this yourselves, be prepared to provide quantities of interest for your audience.

Table of Contents