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**: $Y_i \sim f(y_i \,|\, \theta_i, \alpha)$
- **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.

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 a pro-choice attitude in an American citizen?

Understanding our Sample Data

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

- **prochoice**: normally distributed "latent pro-choice sentiment"

Estimated from Y/N questions about whether abortion should be allowed. . .

1. For any reason (**abany**)
2. Possible defect in the child (**abdefect**)
3. Woman is married, but wants no more children (**abnomore**)
4. Health risk to the mother (**abhlth**)
5. Woman is poor, can't afford children (**abpoor**)
6. Woman is pregnant as result of rape (**abrape**)
7. Woman is not married; doesn't want to marry the man (**absingle**)

Understanding our Sample Data

Here are our predictor variables:

- Age (in years)
- Female
- Black
- Other race, not white
- Hispanic ethnicity
- College education
- Respondent is an atheist
- Religious activity [never:once a day, 0:9]
- Party ID [strong Dem:strong GOP, 0:6]
- Income [under $1k:over $150k, 0:24]

A Sample Regression

```
M1 <- lm(prochoice ~ age +  female  + black + otherrace + hispanic +
           collegeed + atheist +  relactiv + partyid +rincom06,
         data=Data)

M2 <- glm(abany ~ age +  female  + black + otherrace + hispanic +
            collegeed + atheist + relactiv + partyid + rincom06,
          data=Data, family=binomial(link="logit"))

stargazer(M1, M2, header=FALSE, font.size="footnotesize", style="ajps",
          title="Explaining Pro-Choice Attitudes (GSS, 2014)",
          omit.stat=c("ll","aic", "rsq", "adj.rsq", "f", "ser"),
          model.names= FALSE, dep.var.labels.include = FALSE,
          covariate.labels=c("Age", "Female", "Race=Black", "Race=Other",
                            "Hispanic", "College Educated",
                            "Atheist", "Religious Activity",
                            "Party ID (D to R)", "Income", "Constant"),
          single.row=TRUE
          )
```

A Sample Regression

Table 1: Explaining Pro-Choice Attitudes (GSS, 2014)

	Model 1	**Model 2**
Age	0.002 (0.002)	−0.001 (0.006)
Female	0.103* (0.057)	0.252 (0.161)
Race=Black	−0.037 (0.085)	−0.074 (0.234)
Race=Other	0.067 (0.097)	0.101 (0.277)
Hispanic	−0.365*** (0.089)	−0.800*** (0.249)
College Educated	0.304*** (0.063)	0.878*** (0.181)
Atheist	0.192*** (0.068)	0.497** (0.195)
Religious Activity	−0.103*** (0.013)	−0.214*** (0.040)
Party ID (D to R)	−0.126*** (0.015)	−0.308*** (0.045)
Income	0.020*** (0.005)	0.047*** (0.015)
Constant	0.024 (0.135)	0.136 (0.380)
N	863	863

***$p < .01$; **$p < .05$; *$p < .1$

Standardizing the Variables

```
M1 <- lm(prochoice ~ z.age +  female  + black + otherrace + hispanic +
           collegeed + atheist +  z.relactiv + z.partyid + z.rincom06,
         data=Data)

M2 <- glm(abany ~ z.age +  female  + black + otherrace + hispanic +
            collegeed + atheist + z.relactiv + z.partyid + z.rincom06,
          data=Data, family=binomial(link="logit"))

stargazer(M1, M2, header=FALSE, font.size="footnotesize", style="ajps",
          title="Explaining Pro-Choice Attitudes (GSS, 2014)",
          omit.stat=c("ll","aic", "rsq", "adj.rsq", "f", "ser"),
          model.names= FALSE, dep.var.labels.include = FALSE,
          covariate.labels=c("Age (Standardized)", "Female", "Race=Black",
                             "Race=Other", "Hispanic",
                             "College Educated", "Atheist",
                             "Religious Activity (Standardized)",
                             "Party ID (D to R, Standardized)",
                             "Income (Standardized)", "Constant"),
          single.row=TRUE
          )
```

Comparing the Results

Table 2: Explaining Pro-Choice Attitudes (GSS, 2014)

	Model 1	**Model 2**
Age (Standardized)	0.086 (0.071)	−0.019 (0.203)
Female	0.103* (0.057)	0.252 (0.161)
Race=Black	−0.037 (0.085)	−0.074 (0.234)
Race=Other	0.067 (0.097)	0.101 (0.277)
Hispanic	−0.365*** (0.089)	−0.800*** (0.249)
College Educated	0.304*** (0.063)	0.878*** (0.181)
Atheist	0.192*** (0.068)	0.497** (0.195)
Religious Activity (Standardized)	−0.482*** (0.063)	−1.003*** (0.188)
Party ID (D to R, Standardized)	−0.483*** (0.060)	−1.183*** (0.173)
Income (Standardized)	0.237*** (0.063)	0.564*** (0.179)
Constant	−0.095* (0.053)	−0.423*** (0.150)
N	863	863

***$p < .01$; **$p < .05$; *$p < .1$

Something Nana Might Ask

What's the effect of Republican party ID on an atheist's pro-choice attitudes?

- Let's see!

First, we re-estimate the model in Zelig.

The Model in Zelig

```r
M3 <- zelig(abany ~ age + female + black + otherrace +
              hispanic + collegeed + atheist +
              partyid + relactiv + rincom06,
            model="logit", data= Data)
```

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 an atheist's pro-choice attitude among strong Republicans? How about strong Democrats?

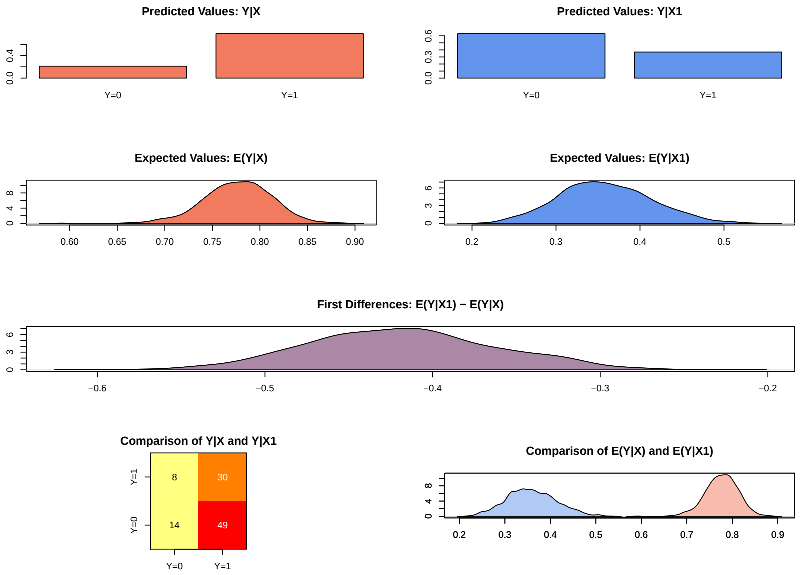
```
M3.athsdem <- setx(M3, atheist = 1, partyid = 0)
M3.athsrep <- setx(M3, atheist = 1, partyid = 6)

M3.sim <- sim(M3, x = M3.athsdem, x1 = M3.athsrep)
```

Answering Nana's Question

	E(Y)	**SD**	**95% Interval**
Atheist, Strong Dem	0.777	0.035	(0.709, 0.845)
Atheist, Strong GOP	0.361	0.052	(0.259, 0.463)
First Difference	-0.416	0.056	(-0.525, -0.307)

Answering Nana's Question



Another Question from Nana

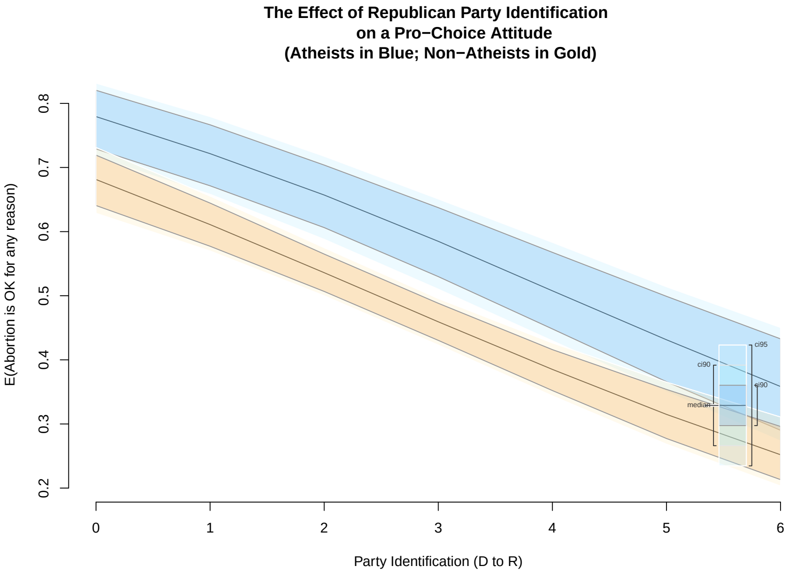
What's the effect of Republican party identification on atheists, overall?

```
M3.athsdemtosrep <- setx(M3, atheist=1, partyid = 0:6)
M3.nathsdemtosrep <- setx(M3, atheist=0, partyid = 0:6)

M3.sim2 <- sim(M3, x = M3.athsdemtosrep, x1=M3.nathsdemtosrep)

ci.plot(M3.sim2, xlab = "Party Identification (D to R)",
        ylab = "E(Abortion is OK for any reason)",
        main = "The Effect of Republican Party Identification
        on a Pro-Choice Attitude \n (Atheists in Blue; Non-Athei
        ci=c(90,95)
)
```

A Pretty Graph



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