UNIVERSITY OF MANNHEIM
School of Social Sciences

Quantitative Methods in Political Science: Wrap Up

Thomas Gschwend | Oliver Rittmann | Viktoriia Semenova | David M. Grundmanns

Week 13 - 1 December 2021

Overview: Week 13

What you have learned

A Guidebook for Empirical Research

Data Essay

Where to go from here

Evaluation

What you have learned

Statistical Control

- Consider the following table on graduate school admissions at the UC Berkeley in 1973:[1]

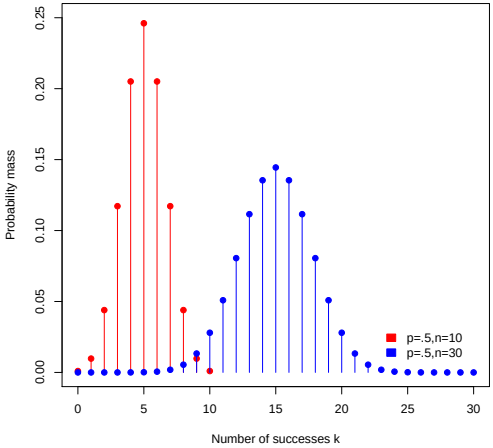
	Male	Female
Admit	1198	557
Reject	1493	1278

- Why is the message of this 2x2 contingency table hard to decipher?

[1] Source: Bickel et al. 1975.

- The binomial probability mass function is:

$$f(k; n, p) = P(K = k) = \binom{n}{k} p^k (1-p)^{n-k}$$



Statistical Inference

- We survey 50 voters how popular a recent reform package is, yielding $y =$

  15 16 12 17 14 13 15 16 12 14 17 15 12 15 14 16 16 14 13 12 13 15 16 14 15

  11 13 13 16 15 17 14 12 15 14 13 16 17 14 15 16 14 13 14 13 15 17 11 14 15

  with an average value of $\hat{\theta} = 14.36$ and a standard deviation of $\hat{\sigma} = 1.61$.
- The standard error of the estimate $\hat{\theta}$ of the population mean is

$$\text{SE}(\hat{\theta}) = \frac{\hat{\sigma}}{\sqrt{N}} = \frac{1.61}{\sqrt{50}} = 0.228$$

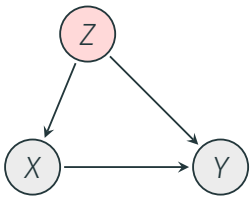
- The 95% confidence interval ranges from

$$14.36 \pm 1.96 \times 0.228 = [13.91, 14.81]$$

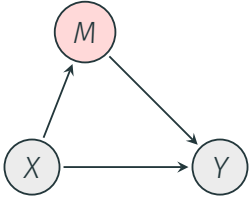
.

Causal Graphs and Statistical Control

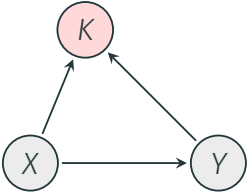
- Causal paths can be open or blocked. Statistical control can open or block paths, depending on whether a variable is:
  1. A confounder Z induces a non-causal association between X and Y. Controlling for a confounder blocks the non-causal path $X \leftarrow Z \rightarrow Y$.
  2. A mediator M captures a specific mechanism that translates a causal effect of X on Y. Controlling for a mediator blocks the *indirect* causal effect $X \rightarrow M \rightarrow Y$.
  3. A collider K is variable that is a descendant of both X and Y. Controlling for a collider opens the non-causal path $X \rightarrow K \leftarrow Y$.



(a) Confounder



(b) Mediator



(c) Collider

Ordinary Least Square Estimation

- Given a zero conditional mean, constant error variance, independent error distribution, and for the multiple regression model no perfect collinearity, the OLS gives us the best linear unbiased estimator (BLUE).
- This means that OLS estimators are unbiased and have minimum variance (Gauss-Markov Theorem).
- The linear model can be summarized as:

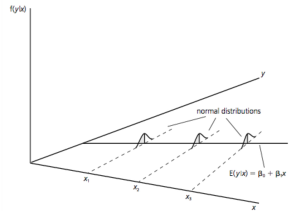
$$Y_i = X_i\beta + \varepsilon_i$$

$$\epsilon_i \sim N(0, \sigma^2)$$

Or, alternatively:

$$Y_i \sim N(\mu_i, \sigma^2) \quad \text{Stochastic component}$$

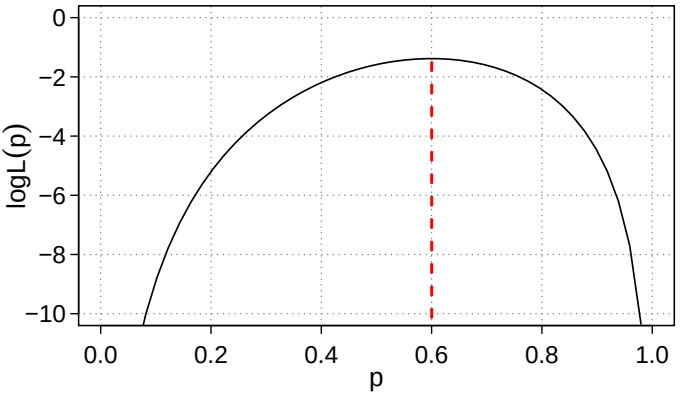
$$\mu_i = X_i\beta \quad \text{Systematic component}$$



Maximum Likelihood Estimation

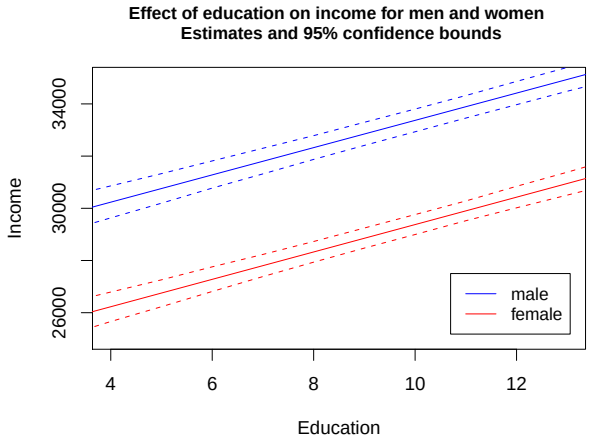
- Example: plot of the binomial log likelihood function
- We find the maximum via

$$\frac{dlogL(p)}{dp} = 0$$



Simulation

- Calculate expected values and confidence bound for each value of a continuous covariate.



A Generalized Linear Model (GLM) consist three components

1. *Stochastic Component*, specifying the conditional distribution of the dependent variable Y_i
2. *Systematic Component*, a linear function of predictors, e.g.,

$$\eta_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_k X_{ik} \tag{1}$$

3. *Link Function* $h(\cdot)$ which transforms the expectation of the dependent variable, $\mu_i = E(Y_i)$, to the linear predictor

$$h(\mu_i) = \eta_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_k X_{ik} \tag{2}$$

Logit-Probit Models

- Remember that for a continuous cdf F we have:

$$P(y_i = 1) = F(\mathbf{x}_i'\boldsymbol{\beta})$$

- If F is a logistic distribution response function, we get

$$P(y_i = 1) = g(\mathbf{x}_i'\boldsymbol{\beta}) = \frac{exp(\mathbf{x}_i'\boldsymbol{\beta})}{1 + exp(\mathbf{x}_i'\boldsymbol{\beta})},$$

  the logit model.

- If F is a standard normal distribution response function, we get

$$P(y_i = 1) = \Phi(\mathbf{x}_i'\boldsymbol{\beta}),$$

  the probit model.

- Deriving the likelihood function:

$$L(\lambda|Y) = \prod_{i=1}^{n} \frac{e^{-\lambda_i} \lambda_i^{y_i}}{y_i!}$$

$$\ln L(\lambda|Y) = \sum_{i=1}^{n} (y_i ln(\lambda_i) - \lambda_i - ln(y_i!))$$

$$\ln L(\beta|Y) = \sum_{i=1}^{n} (y_i (X_i \beta) - e^{X_i \beta})$$

- As before, we maximize the log-likelihood to get $\hat{\beta}$.

```
## 3. Generate Y ~ N(mu,sigma^2)
mu <- cbind(1,X) %*% b  # systematic component

# Choose sigma.est
sigma.est <- 1

# Generate Y
Y <- rnorm(100000,mu,sigma.est)   # (stochastic component)

# Population data
pop <- data.frame(cbind(Y,X))
head(pop)

# Let's put all that into a function

gen.pop <- function(mus,varcov,b,sigma.est){
  X <- mvrnorm(n=100000,mu=mus,Sigma=varcov) # Generate IV
  mu.y <- cbind(1,X) %*% b
  Y <- rnorm(100000,mu.y,sigma.est)
  population <- data.frame(cbind(Y,X))
  return(population)
}

# No multicollinearity
cor.mat1 <- matrix(c(1,0,0,1),nrow=2,ncol=2)
varcov1 <- sds.mat %*% cor.mat1 %*% sds.mat
pop1 <- gen.pop(mus=mus,varcov=varcov1,b=b,sigma.est=sigma.est)
```

Regression Tables

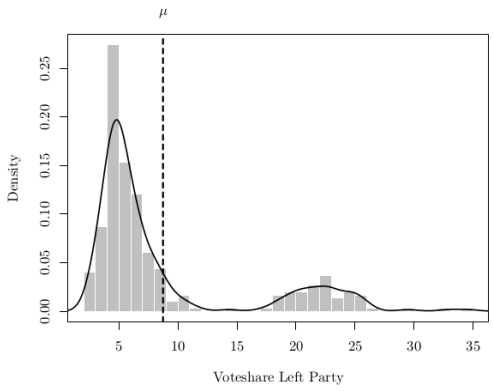
- Let's consider presidential popularity in addition to economic growth.
- Popularity for presidents prior to election ranges between 31% and 74%.
- Our model then becomes:

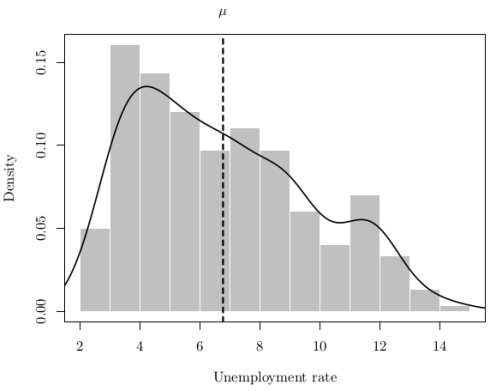
$$VoteShare = \beta_0 + \beta_1 Growth + \beta_2 Approval + e$$

- Results of the regression of vote share on growth and approval rating:

Variable	Model	SE
Constant	34.83	2.77
Growth	0.81	0.27
Approval	0.32	0.06
R^2 = 0.81		
Obs. = 15		

Plots


μ
Density
0.00
0.05
0.10
0.15
0.20
0.25
5
10
15
20
25
30
35
Voteshare Left Party


μ
Density
0.00
0.05
0.10
0.15
2
4
6
8
10
12
14
Unemployment rate

One last time

Roadmap

- Understand and model stochastic processes ✓
- Understand statistical inference ✓
- Implement it mathematically and learn how to estimate it
  - OLS ✓
  - Maximum Likelihood ✓
- Implement it using software
  - R ✓
  - Basic programming skills ✓

Think about where you started and what you've learned.

Be proud of yourselves!

A Guidebook for Empirical Research

5 Steps Towards Better Empirical Research

1. Generate testable hypotheses.
2. Operationalize your theoretical constructs.
3. Specify an appropriate statistical model.
4. Check the robustness of your results.
5. Present your results in a convincing way.

Step 1: Generate Testable Hypotheses

- No matter if your theory comes from the literature, a formal model, or “falls from heaven”, make sure to derive testable hypotheses.
- These hypotheses are best framed as concise “the higher *X*, the higher (lower) *Y*” statements.
  - Hypothesis: *The higher citizens’ levels of income, the more important environmental issues become for them.*

Step 1: Generate Testable Hypotheses

Key aspects to remember

- Make always sure that you understand what the theoretical claim is.
- The more specific your hypothesis, the better you can test it.
- Check if your hypothesis is conditional or not.

Step 2: Operationalize your Theoretical Constructs

- Rarely, your theory is about GDP, but about economic wealth.
- This requires that we operationalize the theoretical constructs from our theories as empirical variables that we can measure.
- Consider the following examples:
  - 'GDP' is a way to operationalize the theoretical construct 'economic wealth'.
  - 'Executive constraints' is a way to operationalize the theoretical construct 'democracy'.
  - 'Trade volume' is a way to operationalize the theoretical construct 'level of globalization'.
  - 'Corruption' is a way to operationalize the theoretical construct 'good governance'.
- At least for your key independent variables, make a case why your operationalization is preferable to other measures.
  - Is it robust to alternative operationalizations (if theory or conventional wisdom cannot give clear guidance)?
- Key independent variables are the ones for which you test your hypotheses.

Step 2: Operationalize your Theoretical Constructs

Key aspects to remember

- Mostly, it is advisable to use standard ways to operationalize your constructs.
- Always defend your choice.
- Ideally, show that your results continue to hold even if you choose an alternative operationalization.

Step 3: Specify an appropriate Statistical Model

- What is the data generating process that leads to your dependent variable? The dependent variable (conditional on X) determines the type of statistical model that you need to choose (linear; log-linear; logit or probit model; count model).
- Check the distribution and summary statistics of your key independent variables, preferably for all the variables you use. It has become good empirical practice to provide this information in an online appendix.
- Be aware of missing values.
- For your main results table (use English labels rather than abbreviations!), present a set of models which become gradually more inclusive. Start with the most basic model, then add e.g., year-fixed effects, region-fixed effects, additional political control variables, and economic control variables.

Step 3: Specify an appropriate Statistical Model

Key aspects to remember

- Always plot distributions of your variables.
- Check summary statistics, correlations, and carefully inspect your data first before you run econometric models.
- Think about the causal order of your IVs. Don't control for consequences of your key causal variable!
- There is no 'correct specification'; it is your job to defend your particular model selection.

Step 4: Check the Robustness of your Results

- Robustness checks are a key exercise for good data analysis.
- Robustness checks (sensitivity analysis) show if your results hold even though the model specification is "slightly" changed.
- Conceptually, the main reasons for robustness checks is omitted variable bias and unobserved heterogeneity, i.e. there exists unmeasured (unobserved) differences between units that are associated with the (observed) variables of interest.
- The following checks can be useful:
  - Include additional control variables (often variables that others in the literature have already used).
  - Operationalize key variables differently.
  - Subset your data (Do the results also hold in a random subset?)
  - Exclude or control for specific countries and years.
  - Exclude outliers (use residual plots or distributions of key variables).
  - Assess consequences of violations of model assumptions (e.g., through *Monte Carlo Simulations*)

Step 4: Check the Robustness of your Results

Key aspects to remember

- Check how your conclusions change when your assumptions change.
- If most of your results hold within some boundaries, this is sufficient.
- It is better to admit the limits of your findings than to not check robustness at all.

Step 5: Present your Results in a Convincing Way

- Make sure that you present your results in the most convincing way possible.
- The presentation of your results is the most important part of your data analysis. If people do not understand what you are saying, ...

  ...they will not give you a good grade,
  ...they will not accept your papers,
  ...they will not cite your work and
  ...they will not be convinced by it.
- It is all about interpreting substantive effects. We do not care if $\hat{\beta} = .5$ or $\hat{\beta} = 0.25$, but we care about what this means.

Step 5: Present your Results in a Convincing Way

- Consider the following example:
  "The estimated effect of corruption on private electricity generation in India is sizable (-2.162) and highly statistically significant at level of $p < 0.01$."
- Alternatively, we could say the following:
  "Similarly, a one unit increase in corruption, which corresponds to an increase by 14 corruption cases per 1,000 inhabitants or a shift from mean corruption levels in Kerala (about 13 cases in 1,000 inhabitants) to Himachal Pradesh (about 28 cases in 1,000 inhabitants), is associated with a decrease in expected private electricity generation by 3%."

Step 5: Present your Results in a Convincing Way

- No matter whether you use OLS models or non-linear models, there is no excuse for not presenting substantive effects. Readers do have a right to know what your results mean.
- For OLS models, the example on the previous slide is a good illustration of how results can be substantiated.
- For non-linear models, you need to calculate quantities of interest and simulate uncertainty with the techniques we taught you.
- Always define the scenario(s) your are simulating clearly in the text.

Step 5: Present your Results in a Convincing Way

Key aspects to remember

- Presenting your results in the best way possible (often as graphs) is key. Spend a lot of time on presenting your research. It is worth the effort!
- Readers have a right to learn what your results mean in substantive terms.
- Thinking about substantive effects also helps you to better understand what the implications of your results are.

Data Essay

Data Essay - Formal Requirements (updated)

- We will provide you with data and a theory that you are supposed to test empirically.
- Your data essay should have 2000 words ($\pm$ 10%, without bibliography). Please indicate the word count on your paper!
- The data essay should emphasize the substantive, statistical, and causal significance of your analysis and the write-up should read very much like the results section of a published article.
- You will have to upload both your PDF and your R code to Ilias by the deadline, i.e. December 15th, 2021, 10h00. The files should be personalized and called `lastnameDataessay.pdf` and `lastnameRcode.R` (or `lastnameRcode.Rmd`, respectively).
- Collaboration is not permitted.
- In case of any questions you can contact us via e-mail, book a slot in our office hours and ask questions on Slack so that the answers will be available to everyone.

Marking Scheme

	Max.	Points	Comment
Descriptive statistics	10		
Description of variables Presentation/Description			
Model Selection/Model Fit	25		
Justification - Chosen Model(s) - Selection of covariates Description - statistical significance Presentation - Table coef, se, loglik, N Model Fit			
Quantities of interest	25		
Some quantities of interest Sensible scenarios Substantial magnitude of findings Description Presentation			
Robustness	10		
Discussion of Robustness Robustness			
Conclusion	5		
Argument Limitations, further research			
R-Code	10		
Well-documented Runs Smoothly			
Overall Impression	15		
Presentation, Language, Coherence, Creativity			
Total			

Where to go from here

Where to go from here...

More advanced methods currently used by political scientists...

- Discrete choice models: Extensions of ordered and multinomial models allowing for nested choices, correlated choices, individual heterogeneity ...
- TSCS models for repeated cross-section observations, e.g.. datasets of OECD countries observed at different time points
- Dynamic models for individual or household panel data
- “Multilevel” models for nested data structures, e.g. parliamentarians within parties
- Measurement error models: modeling & correcting for measurement error,
- Latent variable models: estimating individual positions on unobserved dimensions, e.g. decision maker’s ideal point on some policy dimension
- Matching / “synthetic control” methods for causal inference from longitudinal data or natural experiments
- Bayesian Inference
- Actual experiments ....

Also.... go read the Journals

Read the journals you wanna get published in!

- American Political Science Review (APSR)
- American Journal of Political Science (AJPS)
- Journal of Politics (JoP)
- Political Analysis (PA)

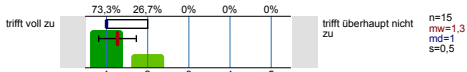
- British Journal of Political Science (BJPolS)
- European Journal of Political Research (EJPR)
- Political Science Research and Methods (PSRM)

- International Organizations (IO)
- Electoral Studies
- Legislative Studies Quarterly (LSQ)

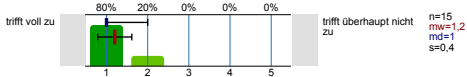
Evaluation - What can we improve?

Your Evaluation of this Course

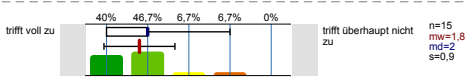
Ich habe in dieser Veranstaltung etwas Sinnvolles und Wichtiges gelernt.



Mein Verständnis für das Studienfach hat sich durch die Veranstaltung weiterentwickelt.

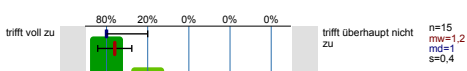


Mein Interesse an den behandelten Inhalten ist durch diese Veranstaltung gestiegen.



Kontext

Das Lehrmaterial (Skript, Folien, Literatur, etc.) war gut zugänglich und wurde rechtzeitig zur Verfügung gestellt.



Der Umfang des behandelten Stoffes war angemessen.