GOV 2001/ 1002/ Stat E-200 Section 12 Research Designs and Multiple Equation Models

Solé Prillaman

Harvard University

April 22, 2015

Logistics

Reading Assignment- UPM Ch. 8, King et al 2001, Honaker and King 2010

Last Section- Wednesday April 29, only at 6pm. Open Q& A.

Last Office Hours-	Soleé	Friday 2-4pm
	Stephen	Monday 11am-1pm

Replication Paper

- Due April 29 on Canvas. (Optional extension until May 4 at 5:00pm)
- Upload your paper and an anonymized version of your paper.
- Upload all of your files to the class dataverse (paper, code, data)
- You will then receive **four** other papers to read, rank, and comment on
- Your comments will be due on May 11

Final Exam

- Take-home exam through Canvas
- Available starting on April 29
- Due May 6 on Canvas
- ONLY for extension school students

Quiz Break!

Why can we say that $L(\theta) \propto \prod_{i=1}^{n} p(y_i|\theta)$? I.e. How do we link the likelihood of our parameters to the probability distribution of our data?

Outline

Causal Effects

Causal Effect:

$$Y_i(1) - Y_i(0)$$

Average treatment effect (ATE):

$$E[Y(1) - Y(0)] = E[Y(1)] - E[Y(0)]$$

Average treatment effect on the treated (ATT):

$$E[Y(1|T = 1) - Y(0|T = 1)] = E[Y_t(1) - Y_t(0)]$$

We can't estimate the ATE and ATT because of unobserved potential outcomes.

Fundamental Problem of Causal Inference

In an ideal world, we would see this:

$Unit_i$	X_i^1	X_i^2	X_i^3	T_i	$Y_i(0)$	$Y_i(1)$	$Y_i(1) - Y_i(0)$
1	2	1	50	0	69	75	6
2	3	1	98	0	111	108	-3
3	2	2	80	1	92	102	10
4	3	1	98	1	112	111	-1

Fundamental Problem of Causal Inference

But in the real world, we see this:

$Unit_i$	X_i^1	X_i^2	X_i^3	T_i	$Y_i(0)$	$Y_i(1)$	$Y_i(1) - Y_i(0)$
1	2	1	50	0	69	?	?
2	3	1	98	0	111	?	?
3	2	2	80	1	?	102	?
4	3	1	98	1	?	111	?

Fundamental Problem of Causal Inference

- For control units, $Y_i(1)$ is the **counterfactual**.
- For treatment units, $Y_i(0)$ is the **counterfactual**.

Our goal: Find a way to estimate this counterfactual.

Decomposition of Estimation Error

- Difference in means: $\hat{\text{ATE}} = E[Y|D=1] - E[Y|D=0]$
- Estimation Error:

$$\begin{aligned}
\Delta &\equiv \text{PATE} - \hat{\text{ATE}} \\
&= \text{PATE} - \text{SATE} + \text{SATE} - \hat{\text{ATE}} \\
&= \Delta_S + \Delta_T \\
&= (\Delta_{S_X} + \Delta_{S_U}) + (\Delta_{T_X} + \Delta_{T_U})
\end{aligned}$$

- Error due to Δ_S (sample selection), Δ_T (treatment imbalance), and each due to observed (X_i) and unobserved (U_i) covariates

Effects of Design Components on Estimation Error

Design Choice	Δ_{S_X}	Δ_{S_U}	Δ_{T_X}	Δ_{T_U}
Random sampling	$\overset{\text{avg}}{=} 0$	$\overset{\text{avg}}{=} 0$		
Stratified Random Sampling	$= 0$	$\overset{\text{avg}}{=} 0$		
Focus on SATE rather than PATE	$= 0$	$= 0$		
Weighting for nonrandom sampling	$= 0$	$=?$		
Large sample size	$\to?$	$\to?$	$\to?$	$\to?$
Random treatment assignment			$\overset{\text{avg}}{=} 0$	$\overset{\text{avg}}{=} 0$
Complete blocking			$= 0$	$=?$
Exact matching			$= 0$	$=?$
Assumption				
No selection bias	$\overset{\text{avg}}{=} 0$	$\overset{\text{avg}}{=} 0$		
Ignorability				$\overset{\text{avg}}{=} 0$
No omitted variables				$= 0$

The Benefits of Major Research Designs

	Δ_{S_X}	Δ_{S_U}	Δ_{T_X}	Δ_{T_U}
Randomized clinicial trials	$\neq 0$	$\neq 0$	$\overset{\text{avg}}{=} 0$	$\overset{\text{avg}}{=} 0$
Social Science Field Experiment	$\neq 0$	$\neq 0$	$\to 0$	$\to 0$
Survey Experiment	$\to 0$	$\to 0$	$\to 0$	$\to 0$
Observational Study (Representative data set, Well-matched)	≈ 0	≈ 0	≈ 0	$\neq 0$
Observational Study (Unrepresentative but partially, correctable data, well-matched)	≈ 0	$\neq 0$	≈ 0	$\neq 0$
Observational Study (Unrepresentative data set, Well-matched)	$\neq 0$	$\neq 0$	≈ 0	$\neq 0$

Research Designs and Estimation Strategies

So here are some possible solutions

1. Randomized or Experimental Studies
2. Observational studies
   - Matching
   - Instrumental Variables
   - Regression Discontinuity Design
   - Difference-in-Differences
   - Synthetic Controls

Research Designs and Estimation Strategies

Always think about:

- What is your assignment mechanism? How was treatment assigned?
- How can we estimate a counterfactual?

Randomized Experiments

The gold standard of scientific research.

- The simplest way (conceptually) to compare treated and control units.
- Units are randomly assigned to receive treatment and control.
- We still can't estimate individual-level causal effects, but we can estimate the population **average treatment effect**, $E[Y_i(1) - Y_i(0)]$.
- Principles of Experimentation: Replication, Randomization, Blocking

Randomized Experiments - Assumptions

1. **Ignorability:** There are no factors out there that affect both the probability of treatment and the outcome.
   - If the treatment was assigned in a truly random fashion, you are generally ok.
   - But if it wasn't, then you are in trouble!
2. **SUTVA:** We have assumed that assigning treatment to one unit doesn't affect the outcome for another unit.
   - Not always reasonable if there are potential peer effects!

Randomized Experiments - Dealing with Non-compliance

What about non-compliance?

- Did the units actually "take" the treatment as assigned?
  - **Never-taker:** Unit never takes treatment
  - **Always-taker:** Unit always takes treatment
  - **Complier:** Unit takes the treatment only when assigned
  - **Defier:** Unit takes treatment only when not assigned

Randomized Experiments - Dealing with Non-compliance

With non-compliance, what did you randomize? Assignment to treatment, **not** treatment

Solutions:

- Focus on "intent to treat" rather than on actual treatment.
- Use an Instrumental Variables approach with the intent to treat as an instrument.
  - Only estimates the effect of treatment on **compliers!**

Observational Data

We only observe our data **after** the experiment occurred and we have no control over treatment assignment.

1. Gather dataset.
2. Estimate ATE or ATT with a *model*.

Observational Data - Assumptions

1. **Ignorability:** Include covariates to get conditional ignorability. Goal: Treatment assignment is independent of the outcomes (Y) given covariates X (must know assignment mechanism for this).

$$(Y(1), Y(0)) \perp T | X$$

2. **SUTVA:** simply assumed (a problematic assumption most of the time)

Observational Data - Concerns

- Ignorability does not hold because of omitted variable bias.
  - Don't include all the variables that makes treatment assignment independent of Y.
- Ignorability does not hold because of unobserved variables.
  - Don't include all the variables that makes treatment assignment independent of Y because unobserved.
- SUTVA assumption
- Model Dependence
  - We try to alleviate the curse of dimensionality and problem of continuous covariates by specifying a model.
  - Estimates of ATE or ATT may differ depending on the model you specify.

Matching to Ameliorate Model Dependence

- If we had pairs of observations with:
  - perfect **balance**
  - differed only on treatment assignment
  - $\Rightarrow$ perfect conditional ignorability
- Same treatment effect regardless of the model.
- Reduces model dependence

- Assume that $\mathbf{X}$ perfectly characterizes the assignment mechanism – no unmeasured confounders (omitted variables) and no unobserved confounders!
- And remember: Don't match on post-treatment variables!

Instrumental Variables to Find Random Variation

Find an instrumental variable Z that:

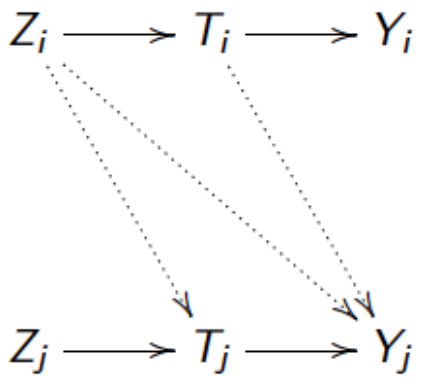
- is randomly assigned (or assignment is ignorable)
- affects Y only through T
- $\Rightarrow$ induce ignorability for a subpopulation

Example: Y = post-Vietnam War civilian mortality; T = serving in the military during Vietnam War; Z = draft lottery

Instrumental Variables - Assumptions

1. **Ignorability** of the Instrument Z
   Example: Assignment of draft status was random.
2. **SUTVA**: Z_i does not affect T_j and Y_j and T_i does not affect Y_j for all $i \neq j$ (non-interference).

Figure: SUTVA Assumption implies absence of dotted arrows.

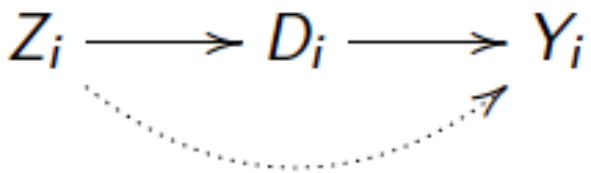


Example: The veteran status and civilian mortality of any man was not affected by the draft status of others.

Instrumental Variables - Assumptions

3. **Exclusion Restriction:** Any effect of Z on Y must be via an effect of Z on T.

Figure: Exclusion restriction implies absence of dotted arrow.



Example: Civilian mortality risk was not affected by draft status once veteran status is taken into account.

4. **First Stage:** Nonzero Average Causal Effect of Z on T.
   Example: Having a low lottery number increases the average probability of service.

5. **Monotonicity:** No Defiers
   Example: There is no one who would have served if given a high lottery number, but not if given a low lottery number.

Instrumental Variables - Estimation

If all the assumptions hold, then the **Local Average Treatment Effect (LATE)** of T on Y is

$$\text{LATE} = \frac{\text{Effect of Z on Y}}{\text{Effect of Z on T}}$$

It is only a local average treatment effect because it's the effect of T on Y for the subpopulation of **compliers**, and not the whole population.

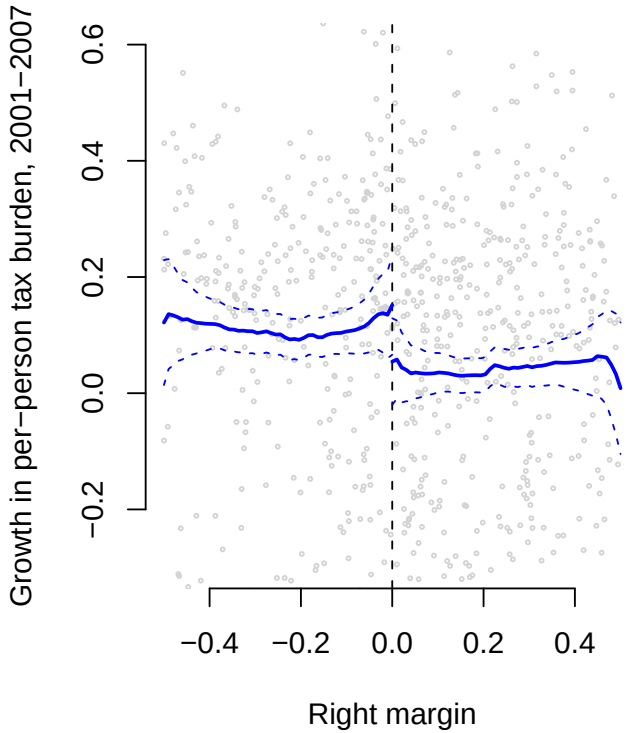
Regression Discontinuity to Exploit Randomness

If you can find a rule or "natural experiment" such that:

- There is some pre-treatment variable x with a cutoff value – below the cut-off value the unit receives treatment and above the cutoff it receives control.
- The cutoff should be arbitrary and not related to the other covariates.
- There should be no real reason to think that the units below the cut-off are substantively that different from the units above the cut-off.
- $\Rightarrow$ ignorability

Regression Discontinuity - Example

In an election a right-wing party wins if they get just over 50% of the vote and loses if they get just under 50%. What is the effect of right-wing governments on tax policies? (Eggers 2013)



Regression Discontinuity - Issues

- Gives us an estimate of the **LATE** - estimate around the cutoff value of x.
- Always pay attention to the bandwidth size.
- Is the rule truly random?

Difference-in-Differences for Unobserved Confounders

We can use two units observed at two time points to get complete **ignorability**.

- Imagine two sets of units (treated and control, $D = 1, 0$) each observed at two time points ($T = 1, 2$)
- There is some intervention on the treated units in $T = 2$ but not the controls
- Take First Differences and Compare

$$(E[Y|D = 1, T = 2] - E[Y|D = 1, T = 1]) - (E[Y|D = 0, T = 2] - E[Y|D = 0, T = 1])$$

Difference-in-Differences - Example

How do inflows of immigrants affect unemployment for native workers? One natural experiment is the Mariel Boatlift - in 1980 thousands of cubans emigrated to the Miami. (Card 1990)

Differences-in-differences estimates of the effect of immigration on unemployment[a]

	Group	Year		
		1979 (1)	1981 (2)	1981–1979 (3)
	Whites			
(1)	Miami	5.1 (1.1)	3.9 (0.9)	− 1.2 (1.4)
(2)	Comparison cities	4.4 (0.3)	4.3 (0.3)	− 0.1 (0.4)
(3)	Difference Miami-comparison	0.7 (1.1)	− 0.4 (0.95)	− 1.1 (1.5)
	Blacks			
(4)	Miami	8.3 (1.7)	9.6 (1.8)	1.3 (2.5)
(5)	Comparison cities	10.3 (0.8)	12.6 (0.9)	2.3 (1.2)
(6)	Difference Miami-comparison	− 2.0 (1.9)	− 3.0 (2.0)	− 1.0 (2.8)

[a] Notes: Adapted from Card (1990, Tables 3 and 6). Standard errors are shown in parentheses.

Difference-in-Differences - Assumptions

- **Parallel Trends:** In the absence of treatment, the treated and control units would have had the same trend over time.
- **Sample Composition:** The composition of the sample should be the same over the pre-treatment and post-treatment periods.

Synthetic Controls to find Counterfactuals at Aggregate Levels

If we have a treated unit and several untreated units, we can find an appropriate and ignorable comparison by taking a combination of the untreated units.

- The **synthetic control** unit is a weighted average of all potential comparison units.
- Weights are chosen by comparing pre-treatment covariates of the untreated units with the treated unit.

Synthetic Controls - Example

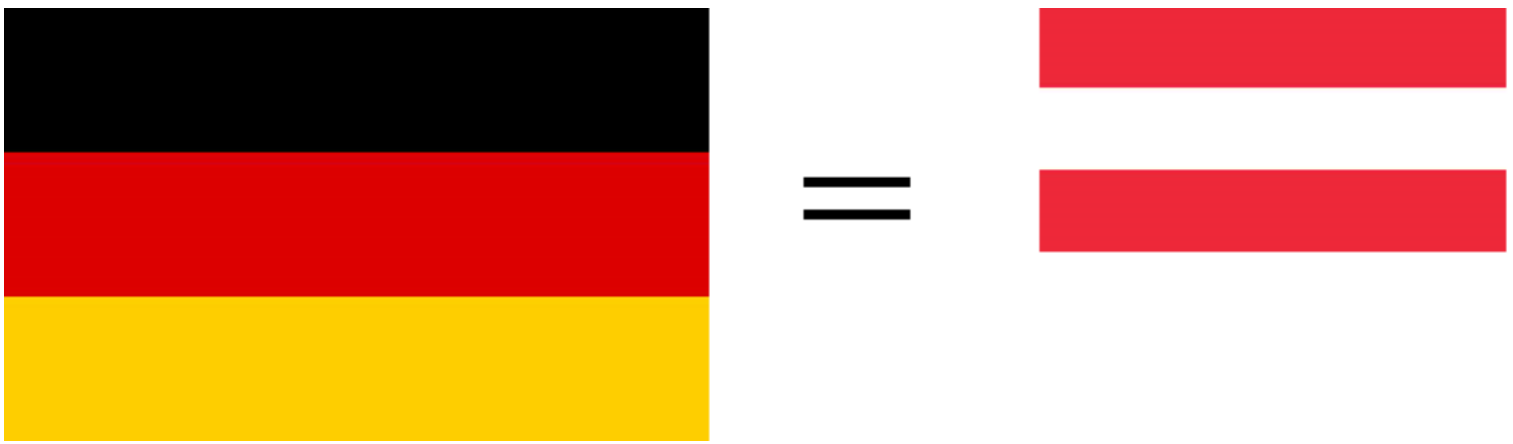
Did reunification affect GDP per capita in West Germany? (Abadie, Diamond, and Hainmueller 2012)

Country	Weight	Country	Weight
Australia	0	Netherlands	0.10
Austria	0.42	New Zealand	0
Belgium	0	Norway	0
Denmark	0	Portugal	0
France	0	Spain	0
Greece	0	Switzerland	0.11
Italy	0	United Kingdom	0
Japan	0.16	United States	0.22

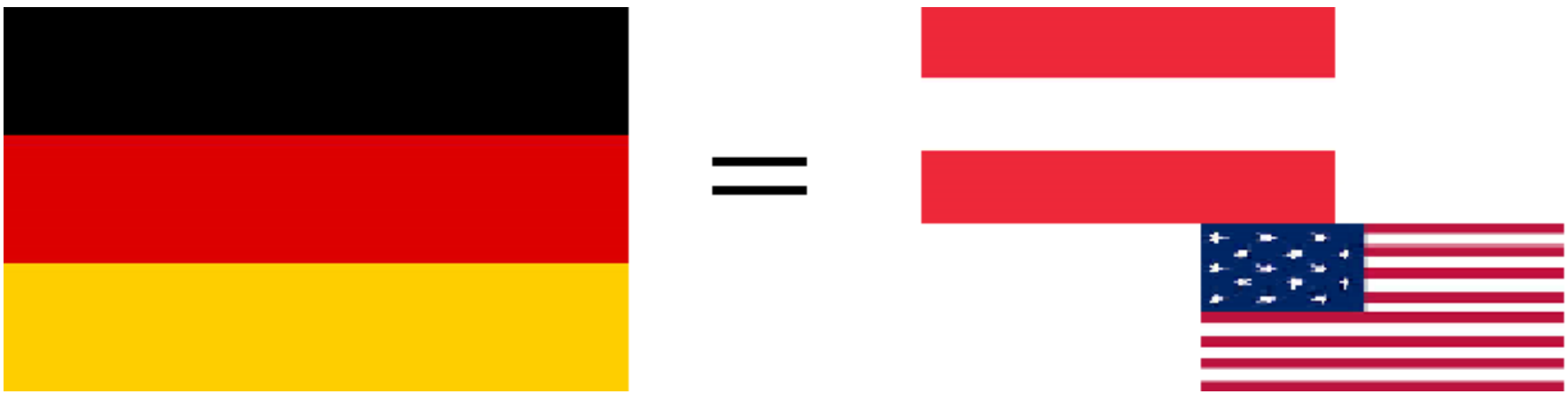
Synthetic Controls - Example

=

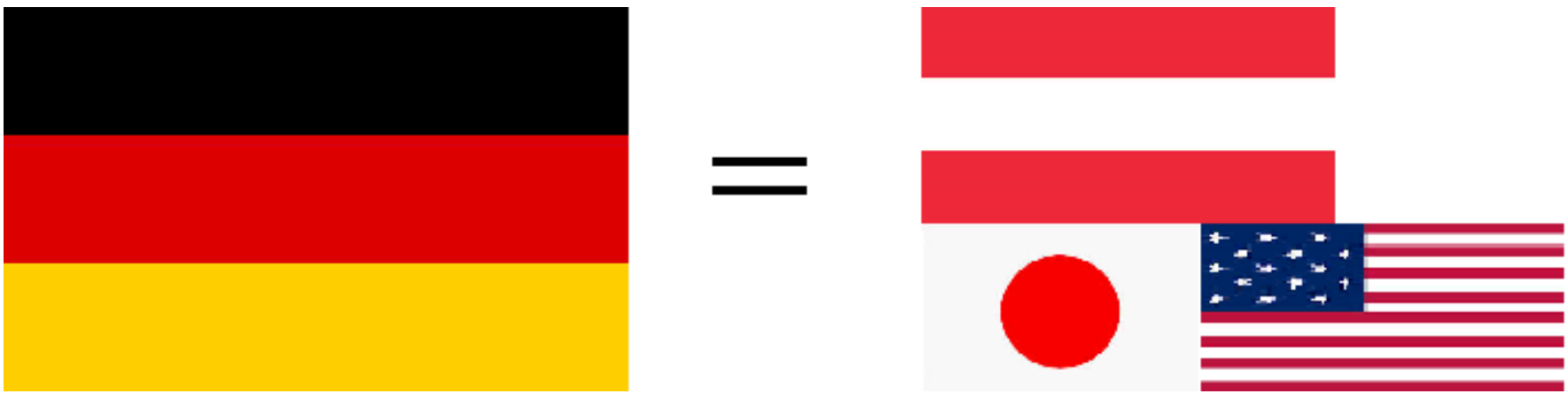
Synthetic Controls - Example



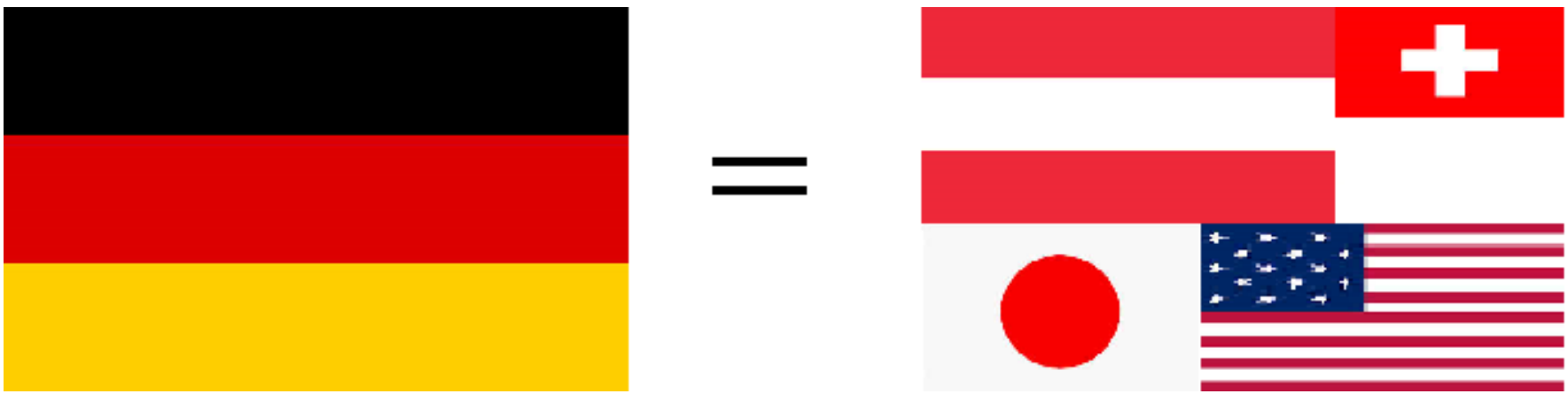
Synthetic Controls - Example



Synthetic Controls - Example

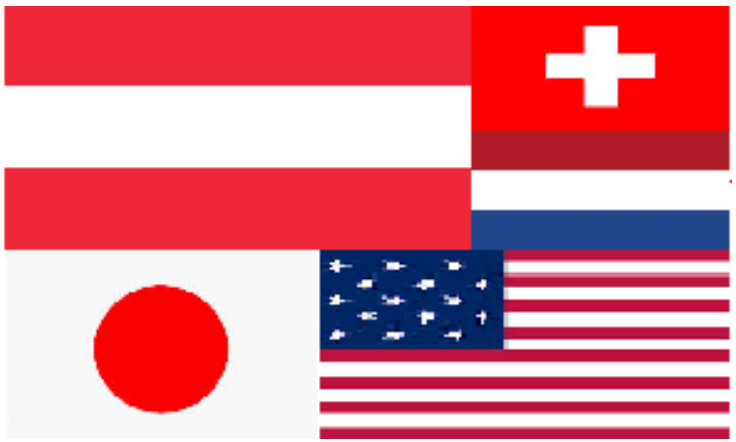


Synthetic Controls - Example



Synthetic Controls - Example

=



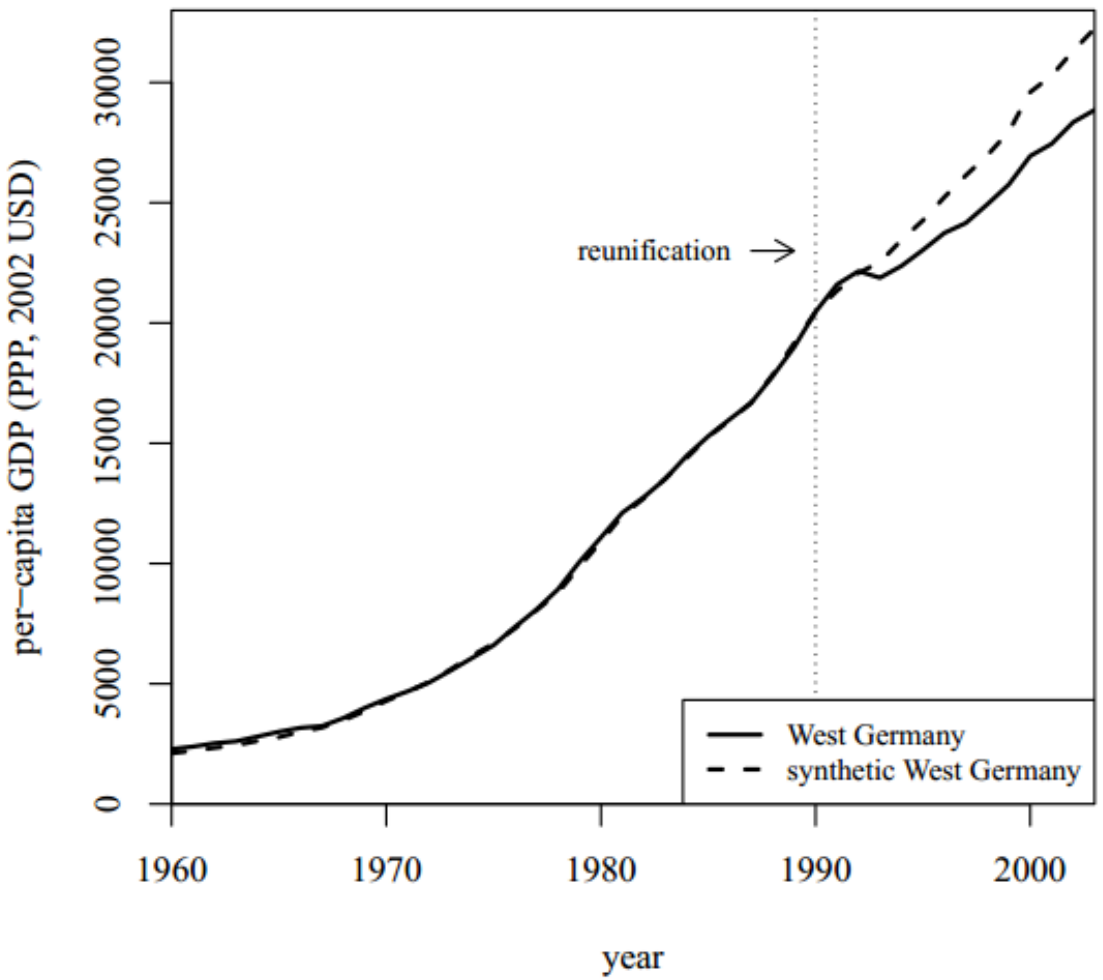
Synthetic Controls - Example

Did reunification affect GDP per capita in West Germany? (Abadie, Diamond, and Hainmueller 2012)

	West Germany	Synthetic West Germany	OECD Sample
GDP per capita	15808.9	15800.9	8021.1
Trade Openness	56.8	56.9	31.9
Inflation Rate	2.6	3.5	7.4
Industry Share	34.5	34.4	34.2
Schooling	55.5	55.2	44.1
Investment Rate	27.0	27.0	25.9

Synthetic Controls - Example

Did reunification affect GDP per capita in West Germany? (Abadie, Diamond, and Hainmueller 2012)



Synthetic Controls - Assumptions

- In the absence of treatment, the treated unit and the synthetic control would behave the same.
- Treated unit must not be an extreme case to ensure the existence of a comparison unit.

References

Angrist, Joshua D., Guido W. Imbens and Donald B. Rubin. 1996. "Identification of Causal Effects Using Instrumental Variables." *Journal of the American Statistical Association* 91(434):444-455.

Angrist, Joshua David and Jörn Steffen Pischke. *Mostly Harmless Econometrics.* Princeton University Press.

Holland, Paul W. 1986. "Statistics and Causal Inference." *Journal of the American Statistical Association* 81(396): 945-960.

Pearl, Judea. *Causality: models, reasoning, and inference*.

Rubin, Donald B. *Matched Sampling for Causal Effects*

Outline

Multiple Equation Models

Occasionally, the system we are studying produces multiple outputs:

- The number of presidential vetoes and the number of Congressional overrides in any given year
- The number of hostile acts directed toward Israel and the number of hostile acts directed toward Palestinians
- The monthly unemployment rate in the United States and the monthly inflation rate.

Multiple Equation Models

You could build a separate model for each outcome variable or you could model them using multiple equation modeling.

When is it better to use multiple equation models?

- When the system outcomes (the Y_i's) are either:
  - (1) **stochastically dependent**: e.g., dependence in Y_{1i} and Y_{2i}
  - (2) **parametrically dependent**: e.g., when $\theta_{j=1}$ and $\theta_{j=2}$ are deterministically related

Stochastic Dependence

Example model:

1. $\vec{Y}_i \sim f_{BVN}(\vec{y}_i | \vec{\mu}_i, \Sigma)$
2. $\vec{\mu}_i = X_i \beta$

$$
\begin{aligned}
\vec{y}_i &= [y_{1i}, y_{2i}] \\
\vec{\mu}_i &= [\mu_{1i}, \mu_{2i}] \\
\Sigma &= \begin{pmatrix} \sigma_1^2 & \sigma_{1,2} \\ \sigma_{2,1} & \sigma_2^2 \end{pmatrix}
\end{aligned}
$$

Stochastic Dependence

- With stochastic independence, we can factor the likelihood into two components

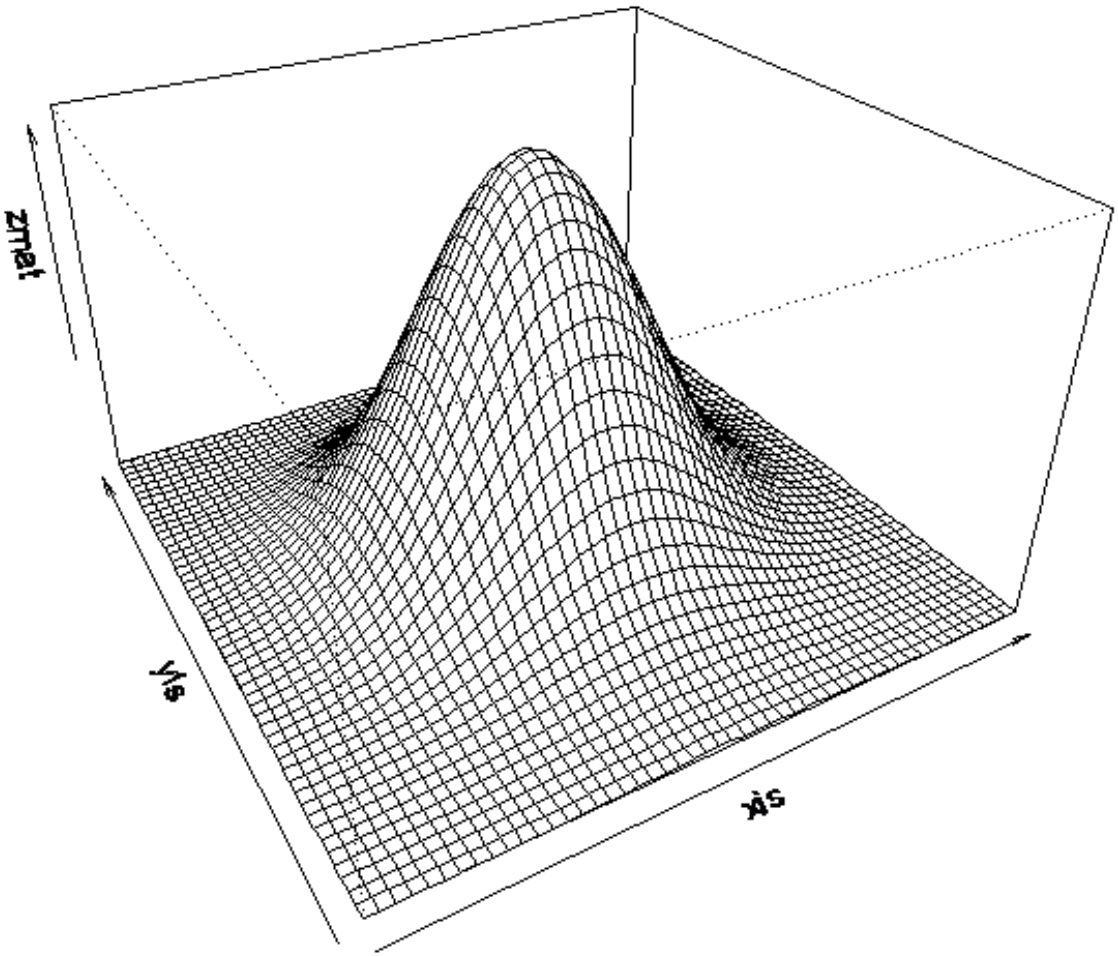
$$\begin{aligned} L(\mu, \Sigma|y) &= \prod_{i=1}^{n} f(y_{1i}, y_{2i}|\vec{\mu}_i, \Sigma) \\ &= \prod_{i=1}^{n} f(y_{1i}|\mu_{1i}, \sigma_1^2) f(y_{2i}|\mu_{2i}, \Sigma_2) \\ \ell(\mu, \Sigma|y) &= \sum_{i=1}^{n} \ln f(y_{1i}|\mu_{1i}, \sigma_1^2) + \sum_{i=1}^{n} \ln f(y_{2i}|\mu_{2i}, \sigma_2^2) \end{aligned}$$

- With stochastic dependence we can't do this. The vector of y's are jointly determined by the probability density function.

$$L(\mu, \Sigma|y) \propto \prod_{i=1}^{n} BVN(\vec{y}_i|\vec{\mu}_i, \Sigma)$$

Stochastic Dependence

What does the distribution of $\vec{y_i}$ look like?



Parametric Dependence

- Parametric independence means we can treat one of the θ's as a constant.

$$\begin{aligned} \max_{\mu_1,\Sigma} \ell(\mu, \Sigma|y) &= \sum_{i=1}^{n} \ln f(y_{1i}|\mu_{1i}, \sigma_1^2) + \sum_{i=1}^{n} \ln f(y_{2i}|\mu_{2i}, \sigma_2^2) \\ \Rightarrow \max_{\mu_1,\Sigma} \ell(\mu, \Sigma|y) &= \sum_{i=1}^{n} \ln f(y_{1i}|\mu_{1i}, \sigma_1^2) \end{aligned}$$

- With parametric dependence we can't do this. Say $\mu_{1i} = \mu_{2i}$, then we can't drop any terms because:

$$\begin{aligned} \ell(\mu, \Sigma|y) &= \sum_{i=1}^{n} \ln f(y_{1i}|\mu_{1i}, \sigma_1^2) + \sum_{i=1}^{n} \ln f(y_{2i}|\mu_{2i}, \sigma_2^2) \\ &= \sum_{i=1}^{n} \ln f(y_{1i}|\mu_{1i}, \sigma_1^2) + \sum_{i=1}^{n} \ln f(y_{2i}|\mu_{1i}, \sigma_2^2) \end{aligned}$$

Multiple Equation Example

Let's look at the `grunfeld` data in `Zelig`.

- Observations from 1935 to 1954 of 7 variables for two firms: General Electric and Westinghouse. The variables are
  - Ige and Iw = Gross investment for GE and W, respectively;
  - Fge and Fw=Market value of Firm at beginning of the year;
  - Cge and Cw= Capital stock measure at beginning of the year.
- We are interested in modeling investment as a function of market value and capital stock.

Multiple Equation Example

- Y: $J \times n$ matrix with $j = 1, ..., J$ and $i = 1, ..., n$
  - j indexes dependent variables (2)
  - i indexes observations (20 years)

For the `grunfeld` data, $Y =$

```
  Ige     Iw
33.1 12.93
45.0 25.90
77.2 35.05
44.6 22.89
48.1 18.84
74.4 28.57
```

Multiple Equation Example

- X: We're going to separate X into two parts:
  - X_{i1}: the covariates specific to the mean for dependent variable 1, in this case GE.
  - X_{i2} the covariates specific to the mean for dependent variable 2, in this case Westinghouse.

For the `grunfeld` data, $X =$

X_{i1}

Fge	Cge
1170.6	97.8
2015.8	104.4
2803.3	118.0
2039.7	156.2
2256.2	172.6
2132.2	186.6

X_{i2}

Fw	Cw
191.5	1.8
516.0	0.8
729.0	7.4
560.4	18.1
519.9	23.5
628.5	26.5

SURM

Seemingly Unrelated Regression Model:

- Stochastic Component

$$\vec{Y}_i \sim MVN(\vec{y_i}|\vec{\mu_i}, \Sigma)$$

where

- $\vec{Y}_i$ and $\vec{\mu_i}$ are $J \times 1$
- Σ is $J \times J$
- Σ is symmetric and full; Why? Stochastic Dependence

- Systematic Component

$$\mu_{i1} = X_{i1}\vec{\beta}$$
$$\mu_{i2} = X_{i2}\vec{\gamma}$$

where

- β: Estimates that predict GE mean
- γ: Estimates that predict Westinghouse mean

SURM

- The Likelihood

$$\begin{aligned} L(\mu, \Sigma) &= \prod_{i=1}^{n} MVN(\vec{y_i}|\vec{\mu_i}, \Sigma) \\ &= \prod_{i=1}^{n} (2\pi)^{-\frac{J}{2}} |\Sigma|^{-\frac{1}{2}} exp\left[-\frac{1}{2}(\vec{y_i} - \vec{\mu_i})'\Sigma^{-1}(\vec{y_i} - \vec{\mu_i})\right] \end{aligned}$$

- What if Σ had off-diagonal elements that were zero? Stochastic Independence

Multiple Equation Example

We can operationalize this model in R

```
formula <- list(mu1 = Ige ~ Fge + Cge,
                mu2 = Iw ~ Fw + Cw)
sur.out <- systemfit(formula = formula,
                method = "SUR", data = grunfeld)
sur.out

# To get the variance - covariance matrix
vcov.sur<-vcov(sur.out)
vcov.sur
# We can see that there are dependencies in our parameters across mo
```

Multiple Equation Example

We can operationalize this model in R

```
sur.out

Coefficients:
mu1_(Intercept)          mu1_Fge          mu1_Cge mu2_(Intercept)
    -27.7193171        0.0383102        0.1390363       -1.2519882
```

Are the μ's correlated?

```
summary(sur.out)

The correlations of the residuals
          mu1      mu2
mu1 1.000000 0.765043
mu2 0.765043 1.000000
```

Multinomial Choice Models

Suppose you have a choice among different, non-ordered things:

- Democrat, Republican, Independent
- Republican candidates for president
- at a restaurant between beef, chicken, or vegetarian
- to wear a blue, yellow, or green sweater

The choice sets are not ordered. But we can use generalizations of ordered logit and ordered probit to analyze.

Multinomial Choice Example

We'll use a running example of vote choice between three candidates:

- Suppose for each individual i, we observe her vote with

$$V_{ij} = (V_{i1}, V_{i2}, V_{i3})$$

  where $V_{ij} = 1$ if the person votes for candidate j and is zero otherwise.
- We could use the ordered logit/probit from the beginning of the semester. But this assumes candidates are ordered along a single dimension.
- Multinomial choice models relax this assumption; we assume there is no order to the choice under consideration, just categories.

Multinomial Logit

- **Stochastic Component:**

$$\vec{V}_i \sim Multinomial(\vec{v}_i|\vec{\pi}_i)$$

where

$$Pr(V_{ij} = 1|X_i, \beta_j) = \pi_{ij}$$

- **Systematic Component:**

$$\pi_{ij} = \frac{exp(X_i\beta_j)}{\sum_{k=1}^{3} exp(X_i\beta_k)}$$

Multinomial Logit

- Suppose we have n voters and 3 candidates.
- The likelihood model for the data, having observed an $n \times K$ matrix of covariates X and an $n \times 3$ matrix of votes V is:

$$L(\beta|X, V) \propto \prod_{i=1}^{n} \prod_{j=1}^{3} \pi_{ij}^{v_{ij}}$$

We can operationalize this in `R` using `Zelig` or the `mlogit` package.

Mutlinomial Logit

Let's look at the 1988 presidential vote in Mexico.

Our hypothesis is that vote for the ruling PRI party varies with age.

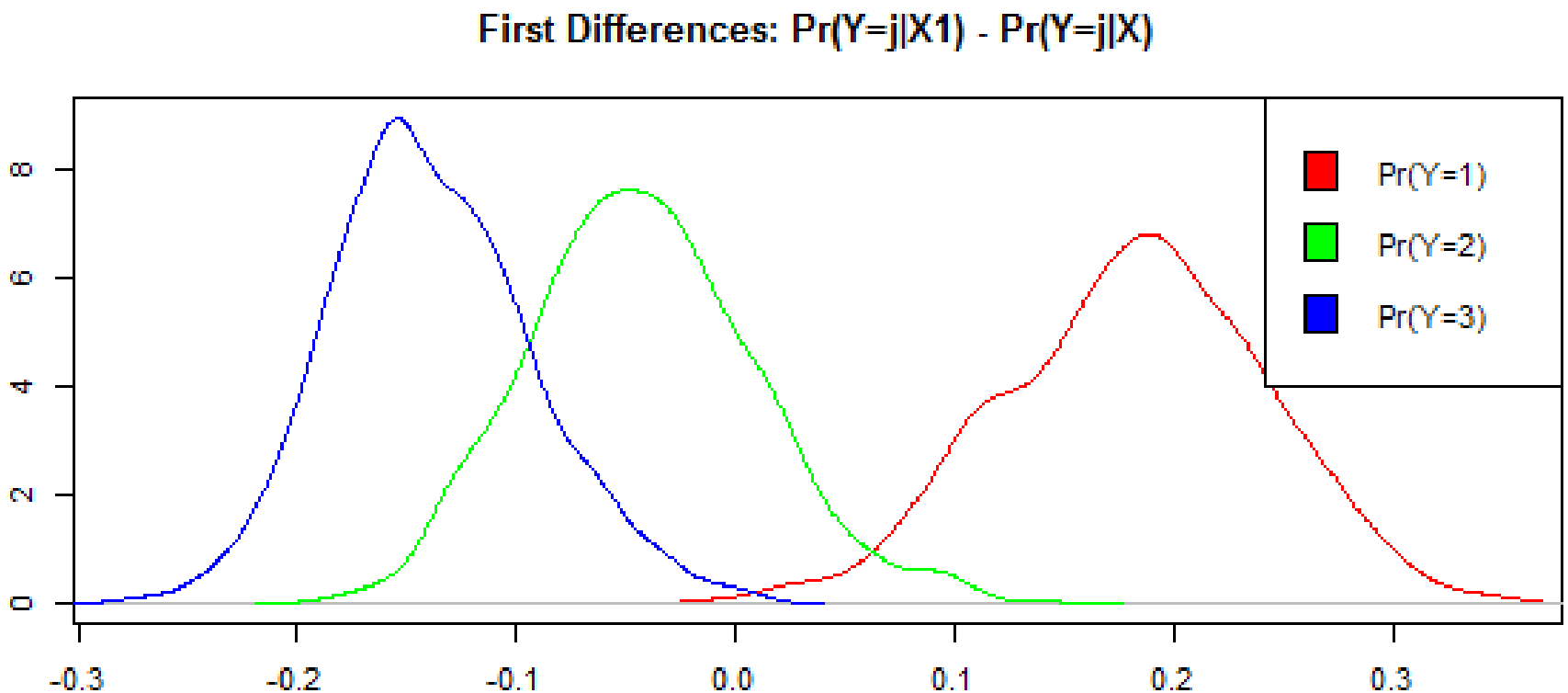
```
data(mexico)

ml.out <- zelig(as.factor(vote88) ~ age + female,
                model = "mlogit", data = mexico)
ml.out

Coefficients:
(Intercept):1 (Intercept):2         age:1         age:2      female:
  0.104832885  -0.500933247   0.016842559   0.008101289   0.26598923
```

Multinomial Logit

Let's look at the difference in expected vote choice for the youngest and the oldest.

```
x.young <- setx(ml.out, age = min(mexico$age))
x.old<- setx(ml.out, age = max(mexico$age))
ml.sim <- sim(ml.out, x1 = x.old, x = x.young)

summary(ml.sim)
First Differences: Pr(Y=k|X1) - Pr(Y=k|X)
              mean          sd       2.5%        97.5%
Pr(Y=1)  0.1831760 0.06264429  0.0618293  0.30277899
Pr(Y=2) -0.0406195 0.05047512 -0.1350517  0.06689080
Pr(Y=3) -0.1425565 0.04747605 -0.2322232 -0.04139632
```

Multinomial Logit



Multinomial Logit

- The multinomial logit assumes independence of irrelevant alternatives (IIA).
- My choice between Shake Shack and Russell House does not depend on the option of Grafton Street. I.e. if Grafton Street closed, the customers would equally go between Shake Shack and Russell House.
- Why do you need IIA? Because, under multinomial logit, your decision between options 1 and 2 never depends on option 3:

$$\frac{\pi_i(1)}{\pi_i(2)} = \frac{\frac{exp(X_i\beta_1)}{\sum exp(X_i\beta_j)}}{\frac{exp(X_i\beta_2)}{\sum exp(X_i\beta_j)}} = \frac{exp(X_i\beta_1)}{exp(X_i\beta_2)}$$

Multinomial Probit

- Relaxes the IIA assumption.

Suppose that voter i receives utility from voting for candidate j:

$$U_{ij}^* \sim N(u_{ij}^* | \mu_{ij}, \Sigma)$$

where Σ is the 3×3 var-cov matrix.

The voter will vote for whoever gives her the highest utility, so we have the following observation mechanism

$$V_{ij} = \begin{cases} 1 & \text{if } U_{ij}^* > U_{ij'}^* \forall j \neq j' \\ 0 & \text{otherwise.} \end{cases}$$

Multinomial Probit

Stochastic Component:

$$V_i \sim Multinomial(v_i|\pi_i)$$

Systematic Component:

$$Pr(V_{ij} = 1) = \pi_{ij}$$

such that

$$\sum_{j=1}^{3} \pi_{ik} = 1$$

Multinomial Probit

To compute π_{ij}, we need to calculate the following:

$$\pi_{ij} = Pr(U^*_{ij} > U^*_{ij'}) \forall j' \neq j$$

which is a messy series of integrals. For example, for $j = 3$:

$$Pr(U^*_{i3} > U^*_{ij} \; \forall j \neq 3) = \int_{-\infty}^{\infty} \int_{-\infty}^{U_{i3}} \int_{-\infty}^{U_{i3}} f(U_1, U_2, U_3) dU_1 dU_2 dU_3.$$

But it's less messy than it looks!

Multinomial Probit

To compute π_{ij}, we need to calculate the following:

$$\pi_{ij} = Pr(U^*_{ij} > U^*_{ij'}) \forall j' \neq j$$

which is a messy series of integrals. For example, for $j = 3$:

$$Pr(U^*_{i3} > U^*_{ij} \; \forall j \neq 3) = \int_{-\infty}^{\infty} \int_{-\infty}^{U_{i3}} \int_{-\infty}^{U_{i3}} f(U_1, U_2, U_3) dU_1 dU_2 dU_3.$$

- We no longer need to assume IIA because we have included other options directly in this integral.
- However, this is hard to identify because of potential correlations between unobserved latent dimensions.
- Note: Imai and Van Dyck's `MNP` package will calculate multinomial probit using Bayesian methods.

Outline

Missing Data: A Preview

This could be our data matrix:

Observation	*Age*	*Education*	*Income*
1	13	4	
2	56	32	$100,000
3	24		$ 30,000
4		12	$ 20,000
⋮	⋮	⋮	⋮

Multiple Equation Models and Missing Data

Wait, it's a multiple equation model!

Observation	*Age*	*Education*	*Income*
$Y_{1j} =$	(13	4	)
$Y_{2j} =$	(56	32	\$100,000)
$Y_{3j} =$	(24		\$ 30,000)
$Y_{4j} =$	(	12	\$ 20,000)
$\vdots$	$\vdots$	$\vdots$	$\vdots$

Understanding Missingness in Our Data

How was this missingness generated? We can characterize the missingness mechanism as:

- **Missing completely at random (MCAR)**: missingness purely random; unrelated to variables in data or any unobserved variables
- **Missing at random (MAR)**: missingness related to *observed* data
- **Not missing at random (NMAR)**: missingness related to *unobserved* data

When can we use multiple imputation?

Multiple Equation Models and Missing Data

We can model this the same way we've modeled multiple equation models:

$$L(\mu, \Sigma|D) \propto \prod_{i=1}^{n} MVN(D_i|\mu, \Sigma)$$

- As usual, we can make μ a function of some covariates

Multiple Equation Models and Missing Data

But aren't some of the data missing?

Observation	*Age*	*Education*	*Income*
$Y_{1j} =$	(13	4	?)
$Y_{2j} =$	(56	32	$100,000)
$Y_{3j} =$	(24	?	$ 30,000)
$Y_{4j} =$	(?	12	$ 20,000)
$\vdots$	$\vdots$	$\vdots$	$\vdots$

Multiple Imputation

We'll integrate out the missing components:

$$\begin{aligned} L(\mu, \Sigma | D_{obs}) &\propto \prod_{i=1}^{n} \int MVN(D_i | \mu, \Sigma) dD_{mis} \\ &= \prod_{i=1}^{n} MVN(D_{i,obs} | \mu_{obs}, \Sigma_{obs}) \end{aligned}$$

- Use the SURM model for the observed data to get estimates of μ and Σ
- Use these estimates to simulate to estimate the missing observations
- Include in your model all X's that you think affect missingness! (And all X's in your model)

Multiple Imputation

Steps to multiple imputation:

1. Impute m values for each missing element
2. Create m completed data sets
3. Run your statistical model on each imputed data set
4. Calculate the point estimate *across* imputed data sets:

$$\bar{q} = \frac{1}{m} \sum_{j=1}^{m} q_j$$

5. Calculate the standard error for this estimate *across* imputed data sets:

$$S[\bar{q}] = \sqrt{mean(SE_k^2) + Var[q_j] \times (1 + \frac{1}{m})}$$

AMELIA

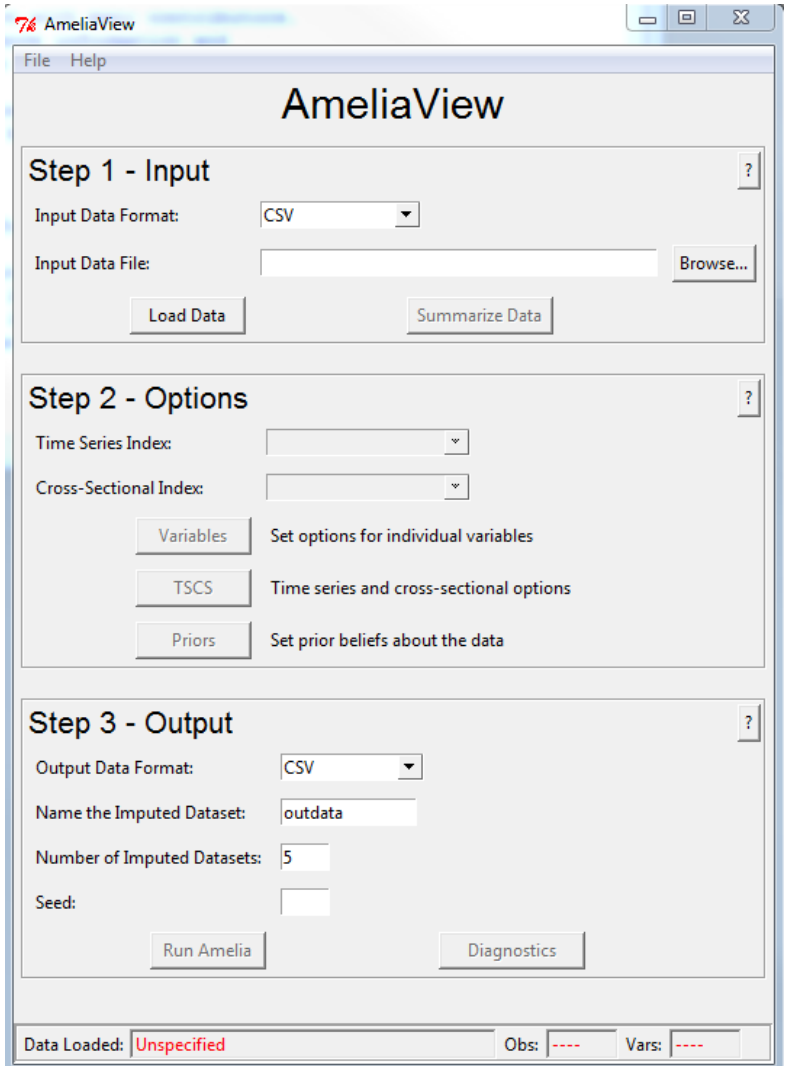
```
library(Amelia)
data(africa)

africa[1,]
  year       country gdp_pc  infl trade civlib population
1 1972 Burkina Faso    377 -2.92 29.69    0.5    5848380

set.seed(1234)
a.out <- amelia(x = africa, cs = "country", ts = "year",
  logs = "gdp_pc", m = 5)
```

`a.out` is a *list* of 5 imputed datasets, each of which can be accessed using `a.out$imputations[[i]]`.

GUI

You can use the GUI by typing:

```
AmeliaView()
```



Estimation with Imputed Data

Now, we just want to estimate a basic regression model with our imputed data, but we have 5 datasets!

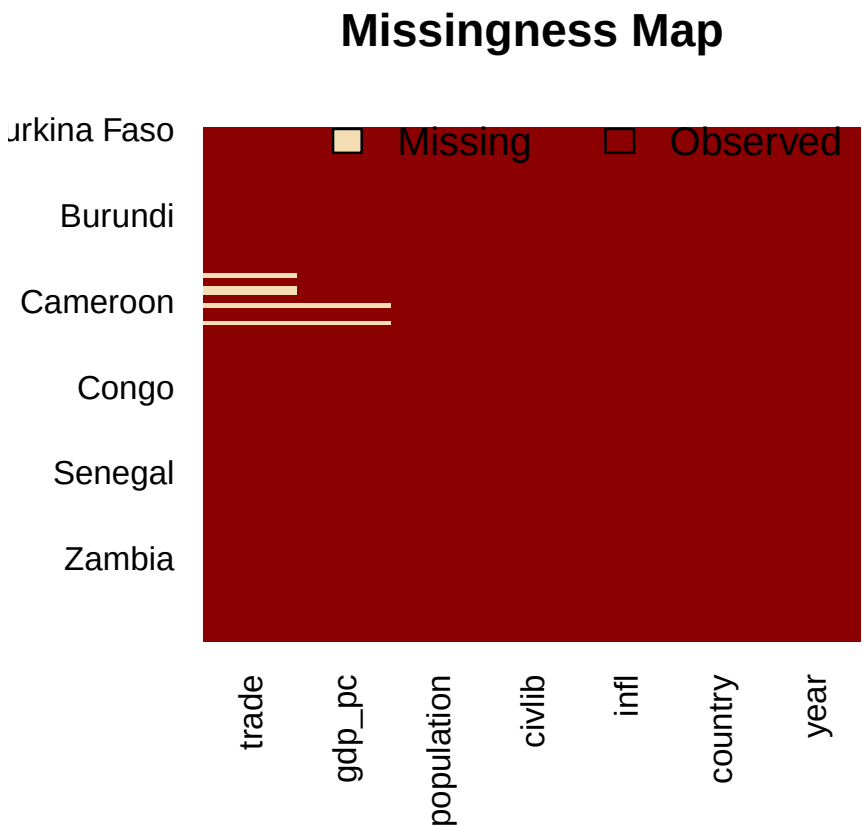
```
z.out.imp <- zelig(trade ~ log(population) + log(gdp_pc) + infl
  + civlib, data = a.out$imputations, model = "ls")
summary(z.out.imp)

Coefficients:
                   Value Std. Error t-stat   p-value
(Intercept)     112.4097   45.47779  2.472 1.345e-02
log(population) -17.8513    2.36646 -7.543 4.595e-14
log(gdp_pc)      31.3875    2.31108 13.581 1.834e-41
infl              0.2605    0.05836  4.463 8.075e-06
civlib           27.2104    6.67210  4.078 4.59$e-05
```

`Zelig` will automatically combine the results of the different models, but if a model you are using isn't programmed in `Zelig`, it isn't hard to combine your estimates.

DIAGNOSTICS

The missingness map gives an overall sense of the shape and extent of the missingness.

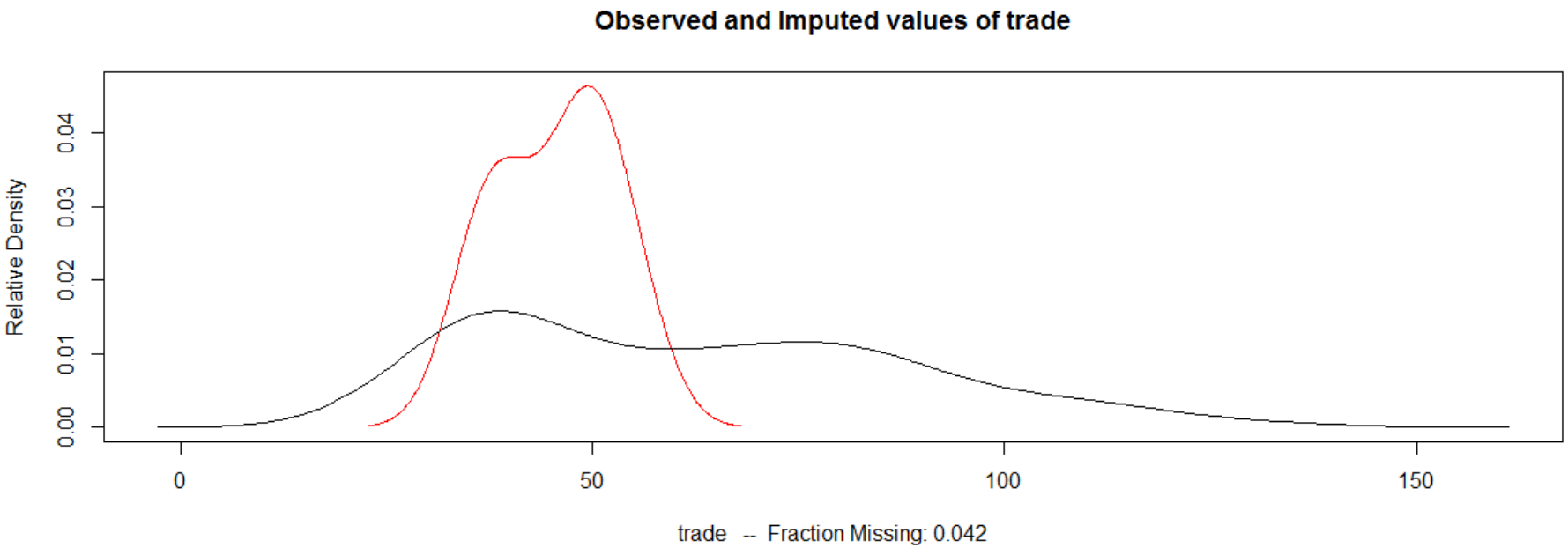
```
missmap(a.out)
```



Diagnostics

Plotting the Amelia object contrasts empirical and imputed densities.

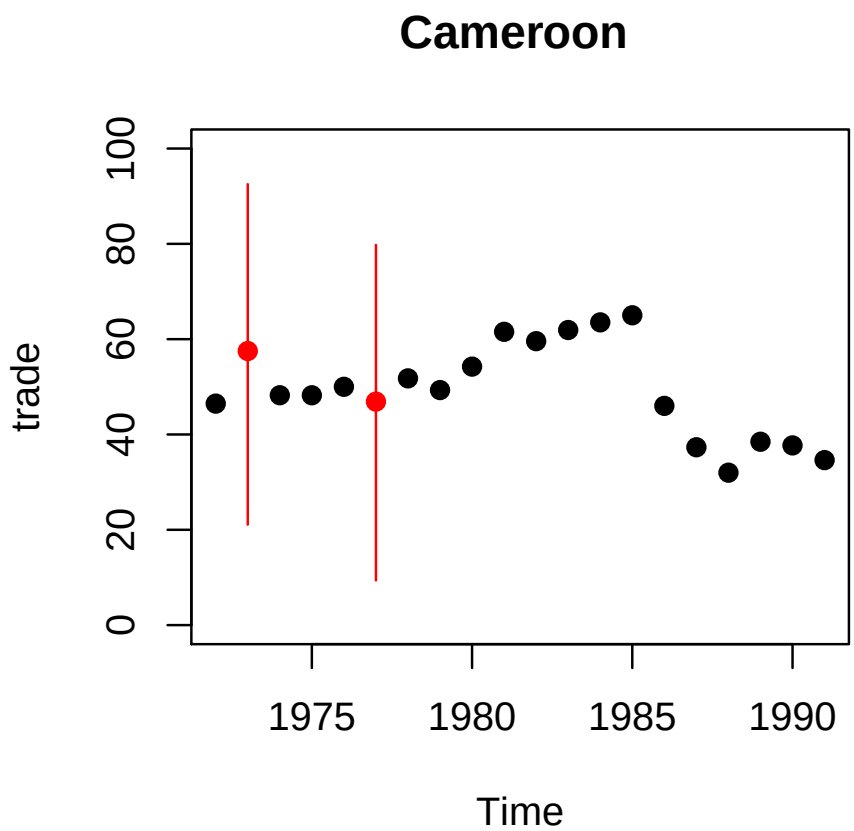
```
plot(a.out)
```



DIAGNOSTICS

Plotting Time-series cross-sectional plots

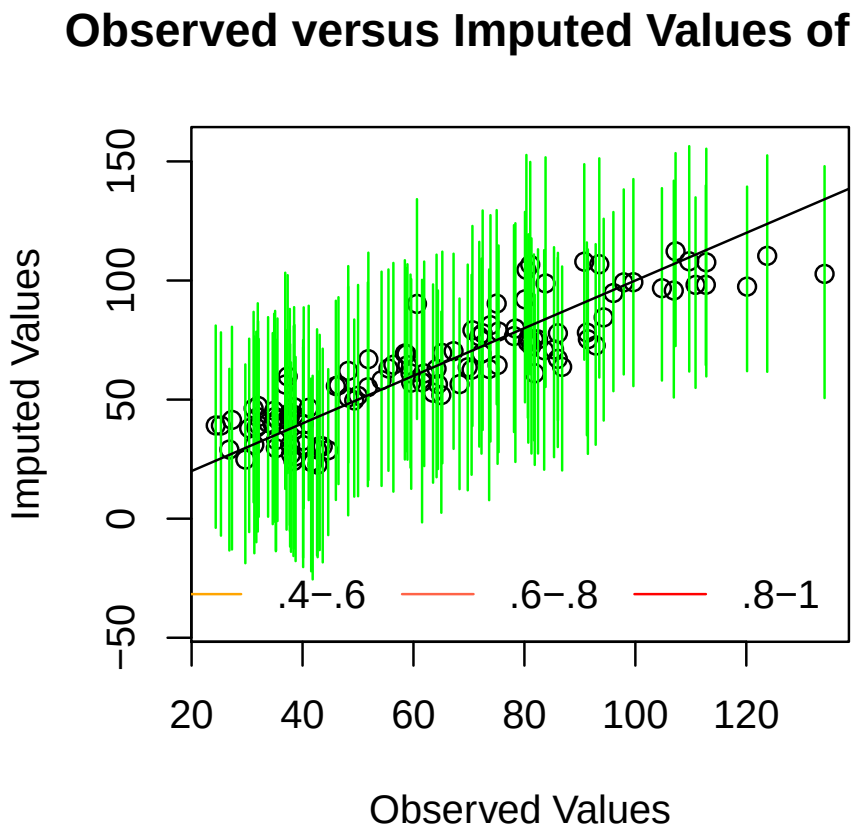
```
tscsPlot(a.out, var = "trade", cs = "Cameroon")
```



Diagnostics

Overimputation for a specific variable tests the imputation model by imagining that each observation is missing and generating some imputations to check performance.

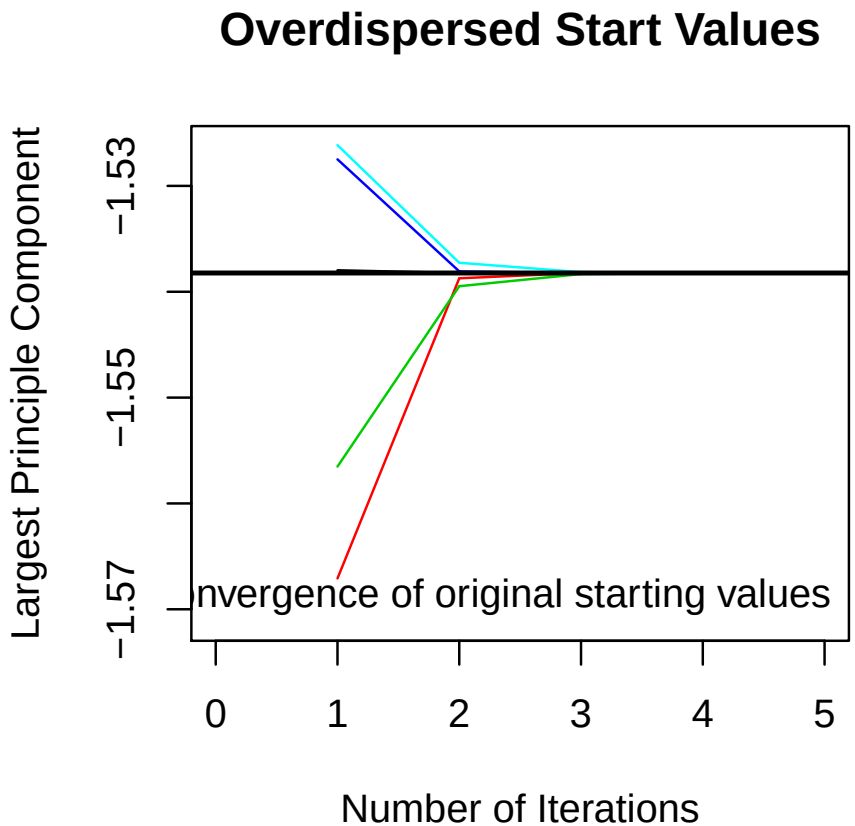
```
overimpute(a.out, var = "trade")
```



DIAGNOSTICS

The disperse function starts the algorithm at some unlikely values to check that amelia hasn't found a local rather a global maximum for the likelihood of the complete data.

```
disperse(a.out, dims = 1, m = 5)
```



Considerations: Transformations

Recall that our model assumes multivariate normal data. This suggests some issues:

1. Ordinal variables: imputation is faster and more informative if ordinal variables are permitted to be continuous, but if undesirable use `ords` argument to constrain.
2. Nominal (unordered) variables: amelia automatically converts into factors and imputes accordingly if a variable is passed to the `noms` argument.
3. Various transformations (logarithmic, root) are pre-programmed to make skewed distributions more normal.

CONSIDERATIONS: TIME SERIES/PANEL DATA

1. Use the `ts` and `cs` arguments to tell amelia the panel structure. It uses this information for constructing time series correctly.
2. `polytime` can be used to add in polynomial time trends (up to a cubic) and `intercs = TRUE` creates a separate trend for each panel.
3. Add in lags and leads if you think it will help predict the missing data. Use `lags` and `leads` with some variable.

Things to remember

1. Set the seed!
2. Include any variable in the analysis model in your imputation model.
3. Don't impute things that don't make sense.
4. Check diagnostics (and think carefully about applicability).
5. Remember transformations, polynomials and data structure.