Advanced Quantitative Research Methodology, Lecture Notes: Model Dependence in Counterfactual Inference[1]

Gary King

GaryKing.org

April 6, 2014

References

- King, Gary and Langche Zeng. "The Dangers of Extreme Counterfactuals," *Political Analysis*, 14, 2, (2007): 131-159.
- King, Gary and Langche Zeng. "When Can History be Our Guide? The Pitfalls of Counterfactual Inference," *International Studies Quarterly*, 2006, 51 (March, 2007): 183–210.

- Related Software: WhatIf, MatchIt, Zelig, CEM

`http://GKing.Harvard.edu/projects/cause.shtml`

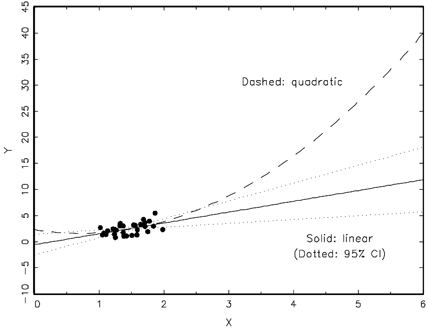
Counterfactuals

- Three types:
  1. Forecasts Will the U.S. be in Afghanistan in 2016?
  2. Whatif Questions What would have happened if the U.S. had not invaded Iraq?
  3. Causal Effects What is the causal effect of the Iraq war on U.S. Supreme Court decision making? (a factual minus a counterfactual)
- Counterfactuals are some part of most research

Model Dependence in Practice

- How do you conduct empirical analyses?
  - collect the data over many months or years.
  - finish recording and merging.
  - sit in front of your computer with nobody to bother you.
  - run one regression.
  - run another regression with different control variables.
  - run another regression with different functional forms.
  - run another regression with different measures.
  - run yet another regression with a subset of the data.
  - end up with 100 or 1000 *different* estimates.
  - put 1 or maybe 5 regression results in the paper.
- What's the problem?
  - Some specification is designated as the "correct" one, only after looking at the estimates.
  - Is this a true test of an ex ante hypothesis or merely a demonstration that it is *possible* to find results consistent with your favorite hypothesis?

Which model would you choose? (Both fit the data well.)



- Compare prediction at $x = 1.5$ to prediction at $x = 5$
- How do you choose a model? R^2? Some "test"? "Theory"?
- The bottom line: answers to some questions don't exist in the data.
- Same for what if questions, predictions, and causal inferences

Model Dependence Proof

Model Free Inference

To estimate $E(Y|X = x)$ at x, average many observed Y with value x

Assumptions (Model-Based Inference)

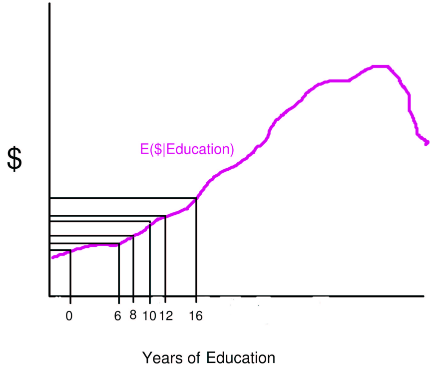
1. Definition: model dependence at x is the difference between predicted outcomes for any two models that fit about equally well.
2. The functional form follows strong continuity (think smoothness, although it is less restrictive)

Result

The maximum degree of model dependence: solely a function of the distance from the counterfactual to the data

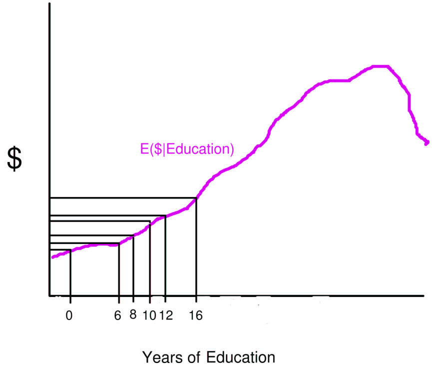
- Randomly select a large number of infants
- Randomly assign them to **0,6,8,10,12,16** years of education
- Assume 100% compliance, and no measurement error, omitted variables, or missing data
- Regress cumulative salary in year 17 on education
- We find a coefficient of $\hat{\beta} = \$1,000$, big t-statistics, narrow confidence intervals, and pass every test for auto-correlation, fit, normality, linearity, homoskedasticity, etc.

What Inferences Would You Be Willing to Make?



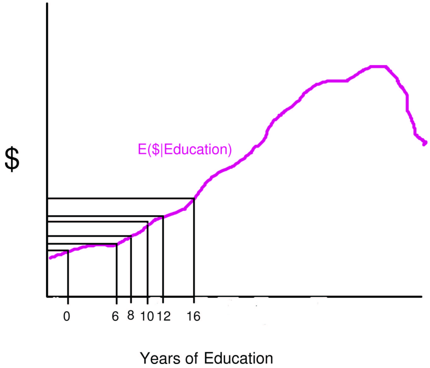
- A Factual Question: How much salary would someone receive with 12 years of education (a high school degree)?
- The model-free estimate: mean(Y) among those with $X = 12$.
- The model-based linear estimate: $\hat{Y} = X\hat{\beta} = 12 \times \$1,000 = \$12,000$

Counterfactual Inferences with Interpolation



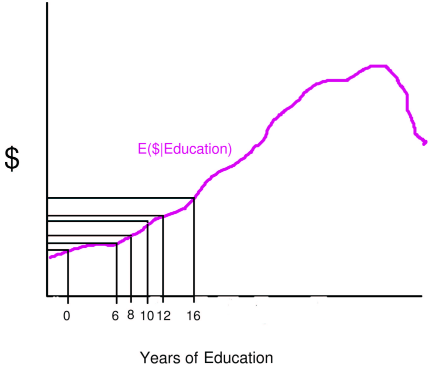
- How much salary would someone receive with 14 years of education (an Associates Degree)?
- Model free estimates impossible.
- $\hat{Y} = X\hat{\beta} = 14 \times \$1,000 = \$14,000$

Counterfactual Inference with Extrapolation



- How much salary would someone receive with 24 years of education (a Ph.D.)?
- $\hat{Y} = X\hat{\beta} = 24 \times \$1,000 = \$24,000$

Another Counterfactual Inference with Extrapolation



- How much salary would someone receive with 53 years of education?
- $\hat{Y} = X\hat{\beta} = 53 \times \$1,000 = \$53,000$
- Recall: the regression passed every test and met every assumption; identical calculations worked for the other questions.
- What's changed? How would we recognize it when the example is less extreme or multidimensional?

Model Dependence with One Explanatory Variable

- Suppose Y is starting salary; X is education in 10 categories.
- To estimate $E(Y|X)$: we need 10 parameters, $E(Y|X = x_j)$, $j = 1, \ldots, 10$.
- Model-free method: average 50 observations on Y for each value of X
- Model-based method: regress Y on X, summarizing 10 parameters with 2 (intercept and slope).
- The difference between the 10 we need and the 2 we estimate with regression is pure assumption.
- (If X were continuous, we would be reducing ∞ to 2, also by assumption)

Model Dependence with Two Explanatory Variables

Variables: X (education) and Z, parent's income, both with 10 categories

- How many parameters do we now need to estimate? 20? Nope. Its $10 \times 10 = 100$. This is the curse of dimensionality: the number of parameters goes up geometrically, not additively.
- If we run a regression, we are summarizing 100 parameters with 3 (an intercept and two slopes).
- But what about including an interaction? Right, so now we're summarizing 100 parameters with 4.
- The difference: an enormous assumption based on convenience, not evidence or theory.

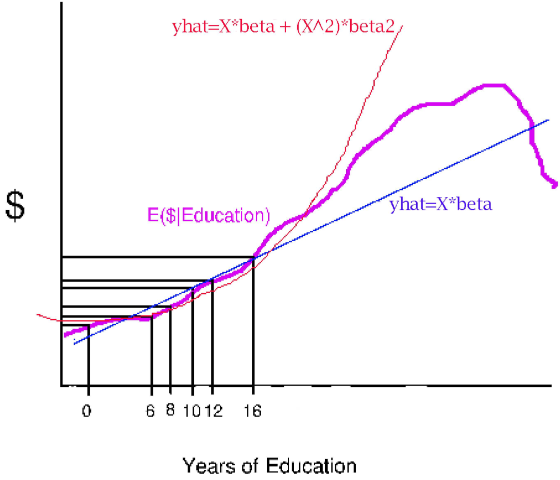
Model Dependence with Many Explanatory Variables

- Suppose: 15 explanatory variables, with 10 categories each.
  - need to estimate 10^{15} (a quadrillion) parameters with how many observations?
  - Regression reduces this to 16 parameters, by assumption.
- Suppose: 80 explanatory variables.
  - 10^{80} is more than the number of atoms in the universe.
  - Yet, with a few simple assumptions, we can still run a regression and estimate only 81 parameters.
- The curse of dimensionality introduces huge assumptions, often recognized.

We Ask: How Factual is your Counterfactual?

- Readers have the right to know: is your counterfactual close enough to data so that statistical methods provide *empirical* answers?
- If not, the same calculations will be based on indefensible model assumptions. With the curse of dimensionality, its too easy to fall into this trap.
- A good existing approach: *Sensitivity testing*, but this requires the user to specify a class of models and then to estimate them all and check how much inferences change
- Our alternative "Convex Hull" approach:
  - Specify your explanatory variables, X.
  - Assume $E(Y|X)$ is (minimally) smooth in X
  - No need to specify models (or a class of models), estimators, or dependent variables.
  - Results of one run apply to the class of all models, all estimators, and all dependent variables.

Interpolation vs Extrapolation in one Dimension



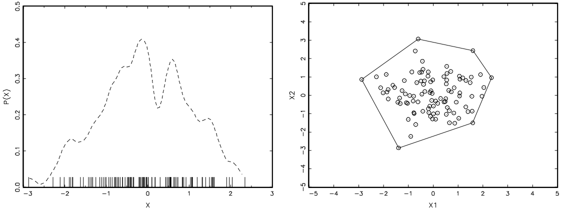


Figure: The Convex Hull

- Interpolation: Inside the convex hull
- Extrapolation: Outside the convex hull
- Works mathematically for any number of X variables
- Software to determine whether a point is in the hull (which is all we need) without calculating the hull (which would take forever), so its fast; see `http://GKing.harvard.edu/whatif`

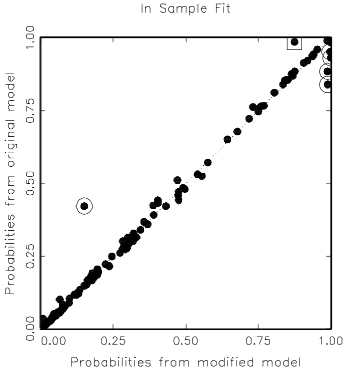
Replication: Doyle and Sambanis, APSR 2000

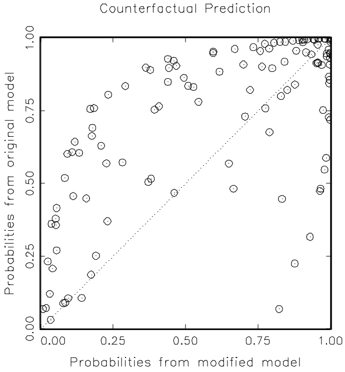
- Data: 124 Post-World War II civil wars
- Dependent variable: peacebuilding success
- Treatment variable: multilateral UN peacekeeping intervention (0/1)
- Control variables: war type, severity, and duration; development status; etc...
- Counterfactuals: UN intervention switched (0/1 to 1/0) for each observation
- Percent of counterfactuals in the convex hull: 0%
- Thus, without estimating any models, we know inferences will be model dependent; for illustration, let's find an example....

Doyle and Sambanis, Logit Model

	Original Model			Modified Model		
Variables	Coeff	SE	P-val	Coeff	SE	P-val
Wartype	−1.742	.609	.004	−1.666	.606	.006
Logdead	−.445	.126	.000	−.437	.125	.000
Wardur	.006	.006	.258	.006	.006	.342
Factnum	−1.259	.703	.073	−1.045	.899	.245
Factnum2	.062	.065	.346	.032	.104	.756
Trnsfcap	.004	.002	.010	.004	.002	.017
Develop	.001	.000	.065	.001	.000	.068
Exp	−6.016	3.071	.050	−6.215	3.065	.043
Decade	−.299	.169	.077	−0.284	.169	.093
Treaty	2.124	.821	.010	2.126	.802	.008
UNOP4	3.135	1.091	.004	.262	1.392	.851
Wardur*UNOP4	—	—	—	.037	.011	.001
Constant	8.609	2.157	0.000	7.978	2.350	.000
N		122			122	
Log-likelihood		-45.649			-44.902	
Pseudo R^2		.423			.433	

Doyle and Sambanis: Model Dependence





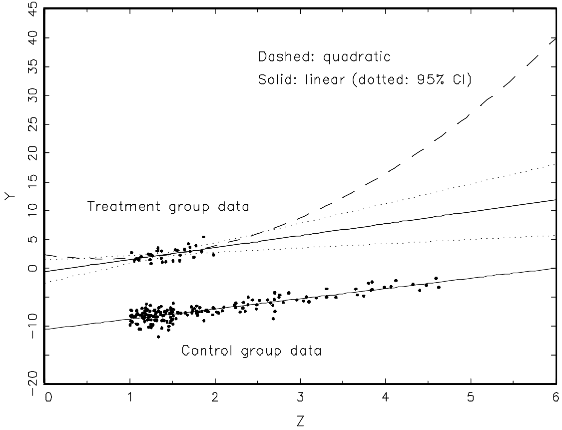
Biases in Causal Inference: A New Decomposition

$$d = \text{mean}(Y|D=1) - \text{mean}(Y|D=0)$$

$$\text{bias} \equiv E(d) - \theta = \Delta_o + \Delta_p + \Delta_i + \Delta_e$$

- Δ_o Omitted variable bias (ignorability) (you know this!)
- Δ_p Post-treatment bias (check this with theory!)
- Δ_i Interpolation bias (Usually not so bad; use models or matching)
- Δ_e Extrapolation bias (check this with data!)

Interpolation vs Extrapolation Bias



Causal Effect of Multidimensional UN Peacekeeping Operations