The Curious Complications of Confounding Covariates

Derek de Montrichard

CIBC

Autumn, 2011

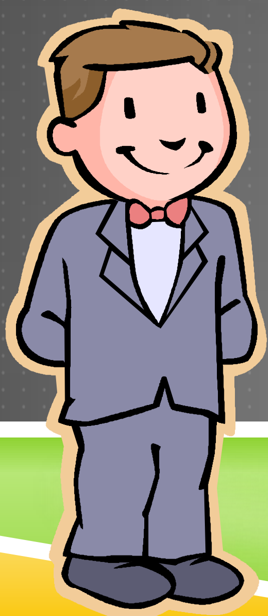
DISCLAIMER

- ALL NUMBERS AND EXAMPLES IN THE FOLLOWING PRESENTATION ARE FOR DEMONSTRATIONAL PURPOSES ONLY. THE REPORT DOES NOT REFLECT ANY ACTUAL DATA FROM HISTORAL STUDIES, BUT INSTEAD SHOWS HOW THESE EFFECTS COULD ENTER TRUE EXPERIMENTS

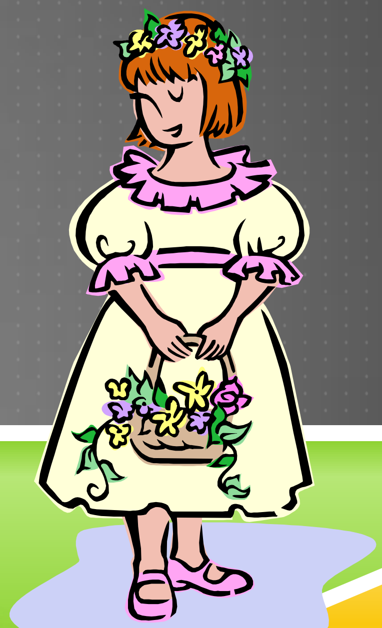
THE BIG QUESTION

- Who will live longer?

Boys



Girls



THE BIGGER QUESTION

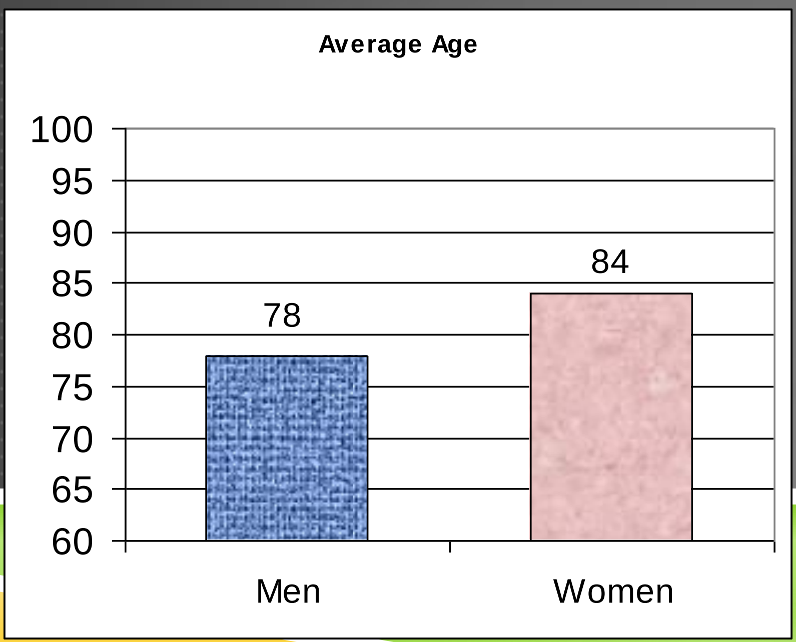
- Whom should you ask?

Statistician

Game Theorist

Answers May Vary

Statistician Answer



Game Theorist Answer

Average Age		Factory Worker	
		Yes	No
Gender	Men	72	92
	Women	70	90

Univariate Effects do not Match Combined Effects

Single Effect

		Average Age	Effect
Gender	Men	78	
	Women	84	6
Factory Worker	Yes	71	
	No	91	20

Quite a difference!!!

Joint Effect

Average Age		Factory Worker		Effect
		Yes	No	
Gender	Men	72	92	
	Women	70	90	-2
Effect			20	

Disjoint Caused by Correlation of Independent Variables

% of population		Factory Worker	
		Yes	No
Gender	Men	35%	15%
	Women	15%	35%

- Because the factors are not independent, the end results can be strange and difficult to interpret
- This event in data is known as Simpson's Paradox

Example #2 : Smoking is Good For You!

- Parsing the data creatively can lead you to believe smokers outlive non smokers

Average Age			
		Yes	No
Smoker?	Yes	55	85
	No	50	84

- BUT
- True average age for smokers is much less than non-smokers

Example #2 : Why smoking is misleading (and probably not too good for you)

Average Age		Had Lung Cancer	
		Yes	No
Smoker?	Yes	55	85
	No	50	84

Conditional Distribution		Had Lung Cancer	
		Yes	No
Smoker?	Yes	50%	50%
	No	5%	95%

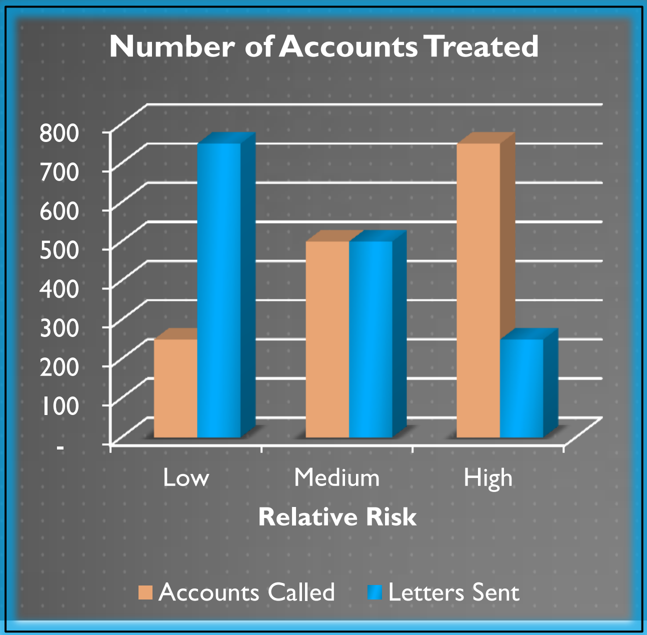
- By splitting on a causal relationship, we've made the initial condition look better than it truly is

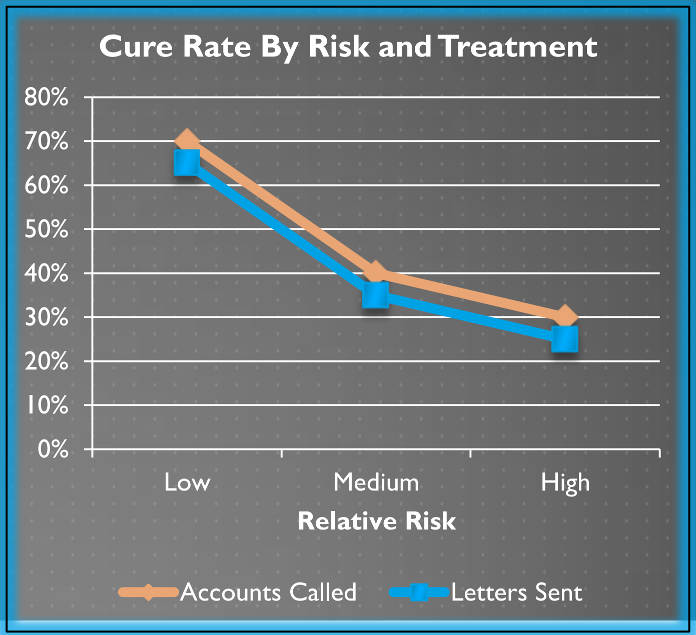
Example 3 : Loans and Collections Actions

- Typically, lending institutions have two treatments for handling overdue accounts:
  - Send a letter
  - Make a phone call
- Cure rates for each treatment are as follows:
  - Letters : 48%
  - Phone Calls : 40%
- Does this mean that sending letters increases the cure rate?

Example 3 : Risk Defines Treatment

- The higher the risk of the account being bad, the more likely we are to call rather than send a generic letter





- When controlling for the covariate (risk), making a phone call will increase the cure rate by 5 percentage points

Example 4 : The Boys of Summer

- Our scouts are tracking two players… who should we consider to be the better hitter?

	Batting Average	
	Year 1	Year 2
Gary Weinrib	0.250	0.320
Alex Živojinović	0.275	0.333

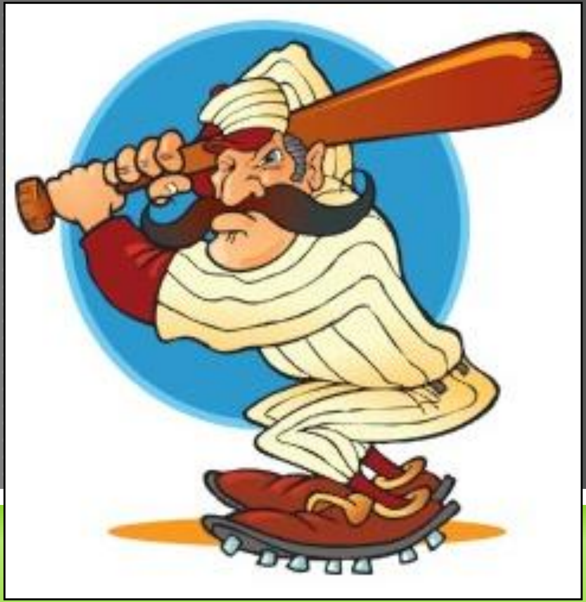


Image courtesy of http://www.zazzle.com/cartoon_baseball_batter_card-137871616801255430

Example 4 : Extreme conditions leads to simple solution

- At bats for each player by year are ***completely*** different due to injuries / playing time

	Batting Average		At Bats		Hits		
	Year 1	Year 2	Year 1	Year 2	Year 1	Year 2	Total BA
Gary Weinrib	0.250	0.320	12	400	3	128	0.318
Alex Živojinović	0.275	0.333	300	12	83	4	0.277

- The more fair comparison would be the grouped effect (.318 >> 0.277)

Example 4 : Nuanced conditions lead to complex solutions (and lots of arguments)

- At bats for each player by year are *slightly* different and unbalanced

	Batting Average		At Bats		Hits		
	Year 1	Year 2	Year 1	Year 2	Year 1	Year 2	Total BA
Gary Weinrib	0.250	0.320	200	475	50	152	0.299
Alex Živojinović	0.275	0.333	355	175	98	58	0.294

- What is the fair comparison? Do we include the covariate?
  - Do we include or exclude covariate if years are 2009 and 2010?
  - Do we include or exclude covariate if years are 2001 and 2010?
  - Which effects can be replicated for 2012?

Keys for Covariates

- If we want to measure the effect of treatment A in combination with covariate B, covariate analysis does increase accuracy of the overall model
- We have to be careful with cause / effect relationship in interpreting parameter estimates
    - If A causes B, then having both A and B in the model can dilute the true effect of A
    - If B causes A, then it is necessary to have both A and B in the model
    - If B is independent of A, both variables can be in the model
- These keys are especially important if A is something we want to change in the overall population

Let's Go Back to Girls vs. Boys…

- Who does live longer?
- Who will live longer?

Average Age		Factory Worker	
		Yes	No
Gender	Men	72	92
	Women	70	90

- ANSWER: Currently, women live longer than men. However, if all factors could be made to be equal, then men could outlive women by two years. As it stands now, men work in harder conditions which leads to lower life expectancies. In the future, this gender inequality may no longer be true, as more women enter manufacturing industries; or, that the manufacturing sector collapses and no jobs remain regardless of gender. If these factors can indeed be made to be balanced across gender, or if these factors are indirectly caused and responsible by gender still remains open for interpretation
- OR: 84 > 78

WHEW!

Conclusions

- Covariate analysis is essential and can lead to more accurate final predictions on your dependent variable
- If the covariates are correlated with the key dependent variables, interpreting the betas can be confusing at best, and misleading at worst
- The modeler / statistician must understand cause and effect within independent variables (or at least document possible causal relationships)
- In presenting results for treatment effects, show both univariate effects (actual and predicted values) and overall modeled effects
- Examples shown only contain 2 dimensions. It is more complex to analyze over n-dimensional space
- Expect difficult questions from Vetting and Peer Review

Questions?

email at derek.montrichard@cibc.com

or visit http://sascanada.ning.com/profile/DerekdeMontrichard