Selection Bias in a Study of the Effect of Friction on Expected Crash Counts on Connecticut Roads

John Ivan[1], Nalini Ravishanker[2], Brien Aronov[2], Sizhen Guo[1], Eric Jackson[1]

[1]Department of Civil and Environmental Engineering
University of Connecticut

[2]Department of Statistics
University of Connecticut

2009 Quality Productivity Research Conference

Outline

OVERVIEW OF PROBLEM

- Work funded by CT Cooperative Highway Research Program.
- Goal: To test if the effect of friction (measured as MeanSN40) has significant effect on (at least injurious) crashes on Connecticut roads, in the presense of typical control variables.
- We focus on friction as it is an actionable variable. That is, it is reasonably easy to change, unlike other physical characteristics and surroundings of the road.
- Roads are broken into two categories:
  - Divided Roads: Usually interstates, major highways. Roads that have a divider between directions
  - Undivided Roads: Roads where no such divider exists. Most town roads are in this category.
- The two categories of roads are analyzed seperately.

MEANSN40?

- The friction value of the road is measured by something called MeanSN40 (Mean Skid Number 40)
- It is the mean of a a series of tested friction values, standardized to a vehicle travelling at 40 MPH.
- The higher the number the better the friction, and (presumably) less accidents.
- Values below 30 are not thought possible, so readings under 30 are attributed to tester error, and thrown out.

FOUND DATA

- ConnDot has collected considerable data already (Found Data).
- However, this data may not be useful on its own.
- Found data is only on roads flagged with known issues.
- Could create a selection bias problem.

SELECTION BIAS PROBLEM

- Population frame of interest SHOULD be roads that have a non-trivial number of crashes.
- Selection Bias would come in if some extra factor, besides usual control variables is leading to different relationship between MeanSN40 and number of crashes
- Control Variables include: Speed Limit of road, Curvature of road, presence of intersections and/or driveways, volume of traffic on road, width of the shoulders on road, etc.

SELECTION BIAS SOLUTION(?)

- Overton (1993) has one approach to this problem:
- Idea is to match (as near as possible) the characteristics of the Found Data to that of a random sample we obtain later (Random Data).
- A well received idea as Found Data is not completely thrown away.

SAMPLE SIZE DETERMINATION

- We consider various usual count regressions techniques (Poisson, Overdispersed Poisson, Negative Binomial) in the GLIM framework with a log link function.
- In such models, the standard error of parameter estimates is also a function of independent data and other parameters!
- Using the basic ideas of Shieh (2001), we use Found Data to estimate parameters only for the purpose of the sample size calculation of the random sample.
- Shieh's results are based on a known distribution of the covariates.
- In our data, this is hard to find thus the actual calculation of sample size done via simulation.

COLLECTING RANDOM DATA

- The collection of data is a non-trivial process.
- CT Cooperative Highway Research Program, wants us to use as much of collected (both Found and Random) data as possible.
- Leads to some differences between characteristics of Found Data and Random Data, but overall not too bad.

HOW TO TELL IF THERE IS SELECTION BIAS?

- An obvious way to check for selection bias, is to compare the results from fitting the GLIMs the Random Data with results from just Found Data.
- Differences in interpretation would indicate that there may just be some selection bias there.
- We exhibit this with the following model, from a negative binomial regression with with the speed limit of the road and the total width of the roads shoulders and traffic volume as control variables.

DIVIDED ROADS PARAMETERS

- MeanSN40 parameter estimate for Random Data only:

variable	estimate	s.e. estimate	Z-value	p-value
MeanSN40	-0.01882	0.01775	-1.060	0.2894

- MeanSN40 parameter estimate Found Data only:

variable	estimate	s.e. estimate	Z-value	p-value
MeanSN40	-.07048	0.02359	-2.988	0.00281

UNDIVIDED ROADS PARAMETERS

- MeanSN40 parameter estimate for Random Data only:

variable	estimate	s.e. estimate	Z-value	p-value
MeanSN40	.022803	.006096	3.741	0.000184

- MeanSN40 parameter estimate Found Data only:

variable	estimate	s.e. estimate	Z-value	p-value
MeanSN40	-0.00576	0.00973	-0.592	0.553896

INTERPRETATION OF DIVIDED DATA

- In the divided roads, the Found data indicate a lack of evidence to support MeanSN40 as a significant factor, while the Random data does indicate statistical significance of MeanSN40
- The signs are estimates are what we expect them to be, even in the non-sginificant case of the Found data.
- This shows some possible selection bias, overall (combining Found and Random data) the coefficient for MeanSN40 is sufficient at a .1 level.

INTERPRETATION OF UNDIVIDED DATA

- In the undivided roads, the found data seem to indicate a significant (positive) effect of friction on crashes. This immediately sounds fishy, as we (more specifically, the subject matter experts, the engineers) believe that this effect should be negative
- The random data indicate a non-significant effect of friction on crashes.
- If data is put together, then effect turns out to be non-significant.
- The undivided data seems to indicate some highly influential selection bias.
- Had we simply used the Found data, we may have made an incorrect decision that friction is more important, and with opposite effect than expected, than it actually is.

SUMMARY

- Selection Bias seems to be present in this project. The model mentioned
- It can lead to an incorrect (and highly confusing) conclusion if not taken care of.
- Effects can be shown through this idea of matching attributes of the Found data with a random sample.
- Open Question: In combining the data, we weigh the Found and Random data the same. Is there some better way to weigh these data? Is it appropriate to use the Found data to generate priors for a Bayesian analysis of the Random data?

- THANK YOU!

References I

Overton, J. McC., Young, T.C. and Overton, W.S.
Using 'Found' Data To Augment A Probablity Sample: Procedure and Case Study.
Enviromental Monitoring and Assessment, 26:65–83, 1993.

Shieh, G.
Sample Size Calculations For Logistic and Poisson Regression Models.
Biometrika, 88:1193–1199, 2001.