Cause, Effect and Modeling

Dr. Ezra Hauer

CAUSE AND EFFECT IN OBSERVATIONAL STUDIES ON ROAD SAFETY; INTRODUCTION and CASE STUDY.

What distinguishes one from another?

Experiment vs. **Observational Study**

Before and After

Cross-Section

Two different questions:

Question A: **What is the safety effect of a certain treatment (cause known effect unknown)?**

Question B: What are the causes of certain events (cause unknown, effect known)?

Case Study – Rail Highway Grade Crossings

Start with Coleman & Stewart, 1976 (data for 32,000 crossings):

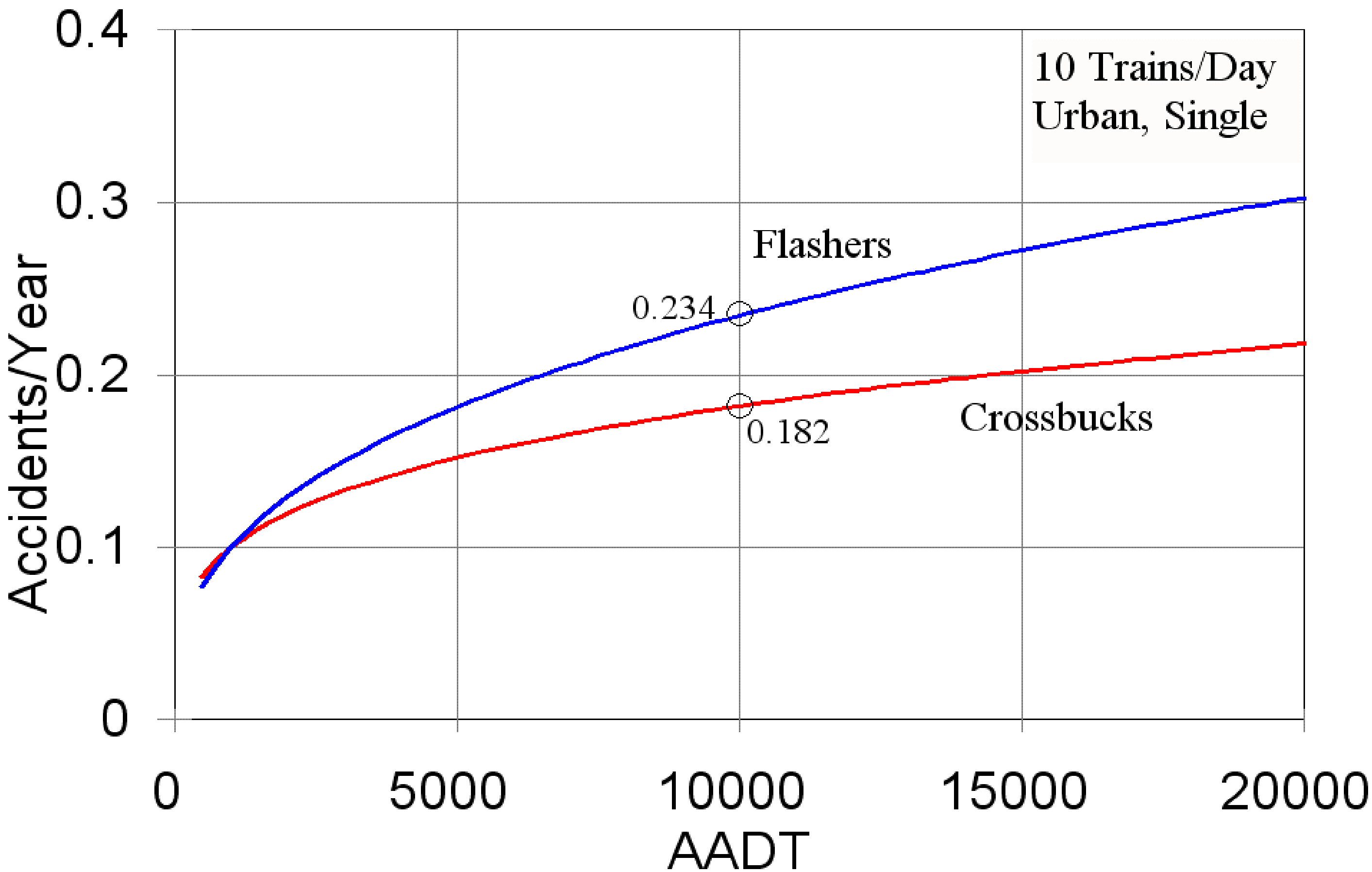
$$Accidents\ /Year = 10^{C_0}(Vehicles\ /Day)^{C_1}(Trains\ /Day)^{C_2}$$
$$(Trains\ /Day)^{C_3 \log_{10}(Trains\ /Day)}$$

Table of 12 sets of parameters was estimated

	Single Track		Multiple Track	
	Rural	Urban	Rural	Urban
Crossbucks	C_0, C_1, C_2,C_3	. . .	. . .	. . .
Flashers	. . .			
Gates	. . .			

Consider, e.g.,

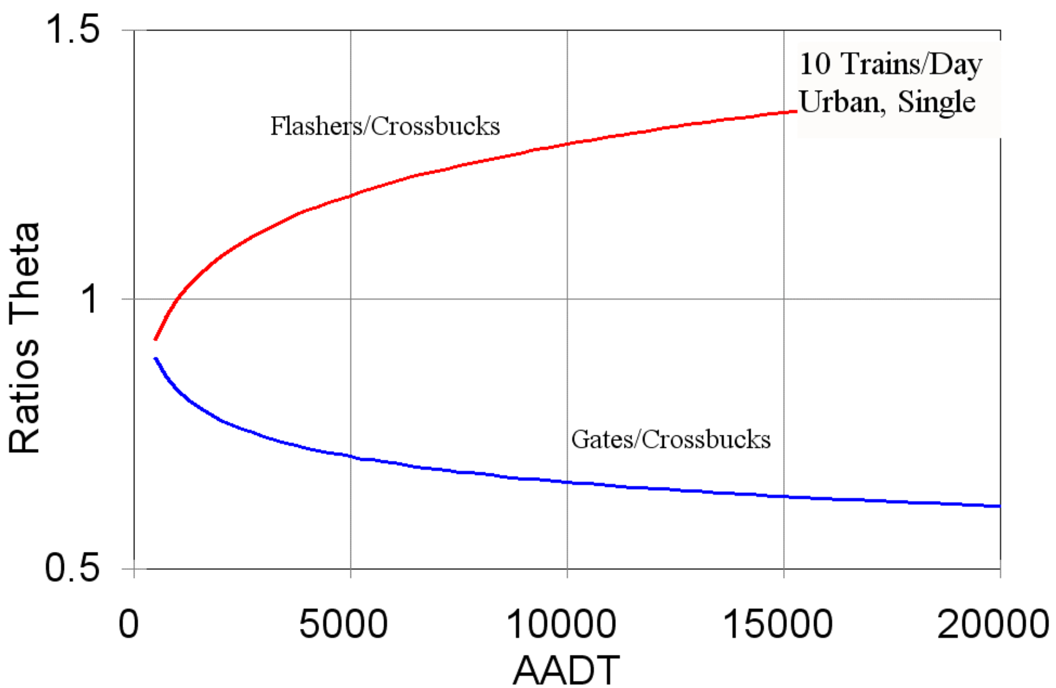
What is the safety effect of replacing 'Crossbucks' with 'Flashers' when AADT=10,000, Trains/day=10, Urban, Single Track?



δ=0.234-0.182 ?
Θ=0.234/0.182 ?
Unknown ?

} Customary

Cross-section study



Before/After Studies

	Calif. PUC	Morrisey	Coleman	Eck & Halkias	Farr & Hitz	Hauer& Persaud
Flashers/ Crossbucks	**0.36**	**0.35**	**0.29**	**0.31**	**0.30**	**0.49**
Gates/Cros -sbucks	**0.12**	**0.16**	**0.18**	**0.16**	**0.17**	**0.21**

Consider now Mengert (1980) who used data about more than 200,000 crossings

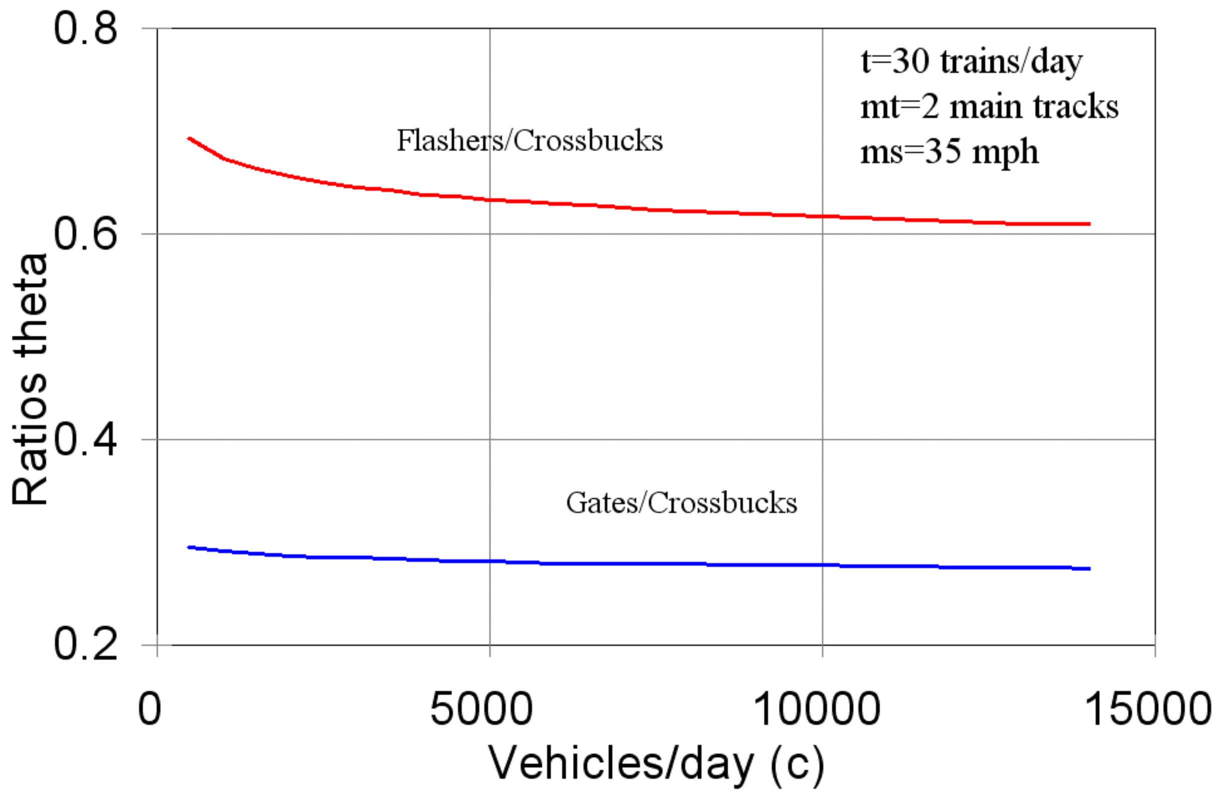
$$Accidents/year = 0.002268 \times \left(\frac{c \times t + 0.2}{0.2}\right)^{0.3334}$$

$$\times e^{0.209 \times (mt) \times \left(\frac{d+0.2}{0.2}\right)^{0.1336}} \times e^{-0.616(hp-1)} \times e^{0.0077 \times (ms)}$$

$$e^{-0.1 \times (ht-1)} \times e^{0.0 \times (hl-1)}$$

Vehicles/day (c); Trains/day (t); Main Tracks (mt); Thru Trains/day (d); Highway is paved (hp); Maximum timetable speed (ms); Highway type (ht); Highway Lanes (hl)

Cross-section study



Before/After Studies

	Calif. PUC	Morrisey	Coleman	Eck & Halkias	Farr & Hitz	Hauer& Persaud
Flashers/ Crossbucks	**0.36**	**0.35**	**0.29**	**0.31**	**0.30**	**0.49**
Gates/Cros -sbucks	**0.12**	**0.16**	**0.18**	**0.16**	**0.17**	**0.21**

Case Study Conclusions:

1. B/A studies in this case consistent (even if biased).

2. Disadvantage of B/A study is that it gives (usually) only the ‘average’ (typical) effect and how it depends on traits s not known. But for action one would like to know how effect depends on traits.

3. If safety effect depends on traits, it is difficult to say whether the results of two studies are consistent.

4. Two examine consistency one needs to know:
   a. What were the traits of the treated entities
   b. How safety effect depends on traits.

5. The two C-S studies examined produced inconsistent results.

Statistical Road Safety Modeling - SRSM

What is SRSM?

- Take data about past accidents and traits, of many sites
- Produce a best fitting equation: Expected Accidents=f(traits)

ourposes of SRSM?

te safety based on traits → Unproblematic

te effect of change in traits → Fraught with Difficulty

The Difficulty:

f(traits,10 ft.) = 2 accidents/mile-year

f(same traits, 11 ft.)= 1.8 accidents/mile-year

Do wider lanes cause a reduction in accidents?

We cannot say!

How do we know that?

Why can't we say?

Why can't we say

1. Missing variables

2. Imperfect 'modeling'

The aim of the paper: improve modeling - to improve our chances of making cause-effect predictions

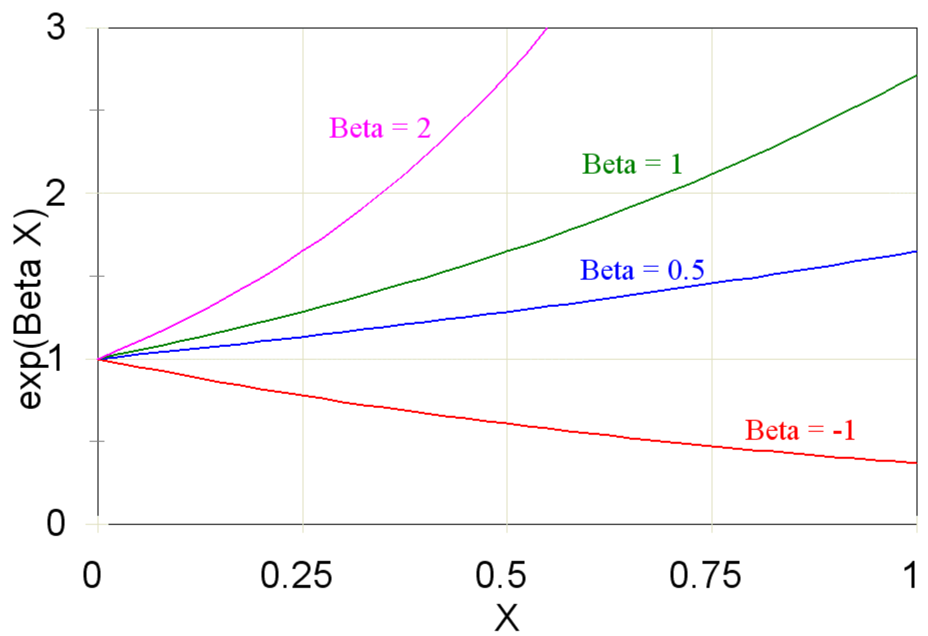
1. Model form

$y = Length \times (\beta_0 + \beta_1 X_1 + ...)$

$y = Length \times \beta_0 X_1^{\beta 1} ...$

$y = Length \times \beta_0 \exp(\beta_1 X_1 + ...)$

These model forms may not be able to represent reality well enough



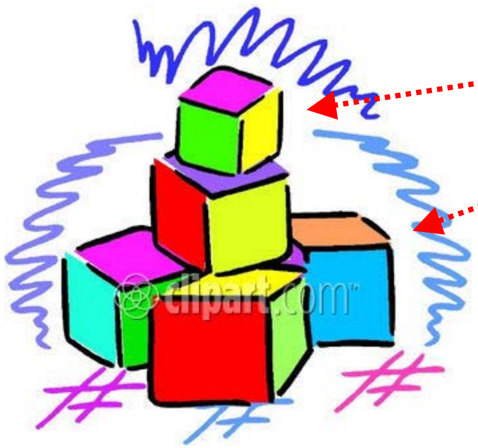
1. Model form (cont.)

A plausible general model form:

y= (Scale Parameter)×[Length×(Multiplicative Portion)+ (Additive Portion)]

Multiplicative Portion= f(ADT) × g(Shoulder Width) × h(Clearzone) × . . .
Additive Portion = p(ADT, No. of Driveways) + q(No. of Short Bridges) + ...

2. Which Variables and what should be the Functions f(), g(), ..., q()



Build model equation by adding blocks one-by-one.

Start with AADT

2. Which Variables and what should be the Functions (cont.)

Suppose that f(ADT) is in the model and introduction of g(Lane Width) is contemplated

Prepare Table

Lane Width	Predicted	Recorded	R=Rec./Pred.	σ
10’	160.9	163	1.01	0.08
11’	719.4	698	.97	0.04
12’	1251.2	1278	1.02	0.03
13’	308.7	307	0.99	0.06
…	…	…	…	…

Is there an orderly relationship between R and Lane Width?

No → Forget it (for now)

Yes → What function can represent it?

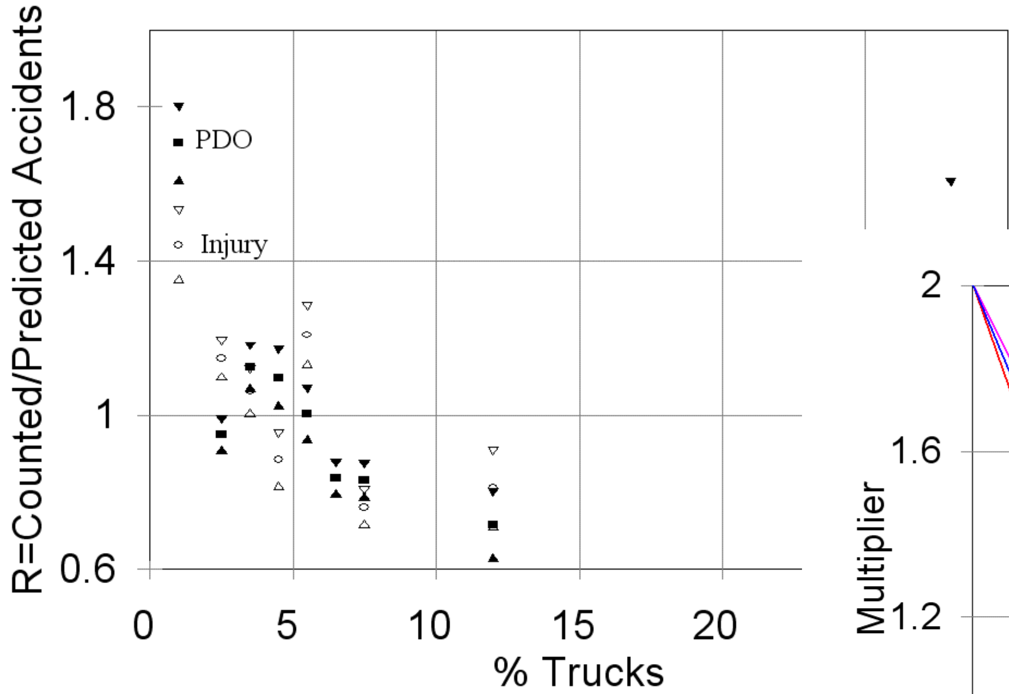
2. Which Variables and what should be the Functions (cont.)

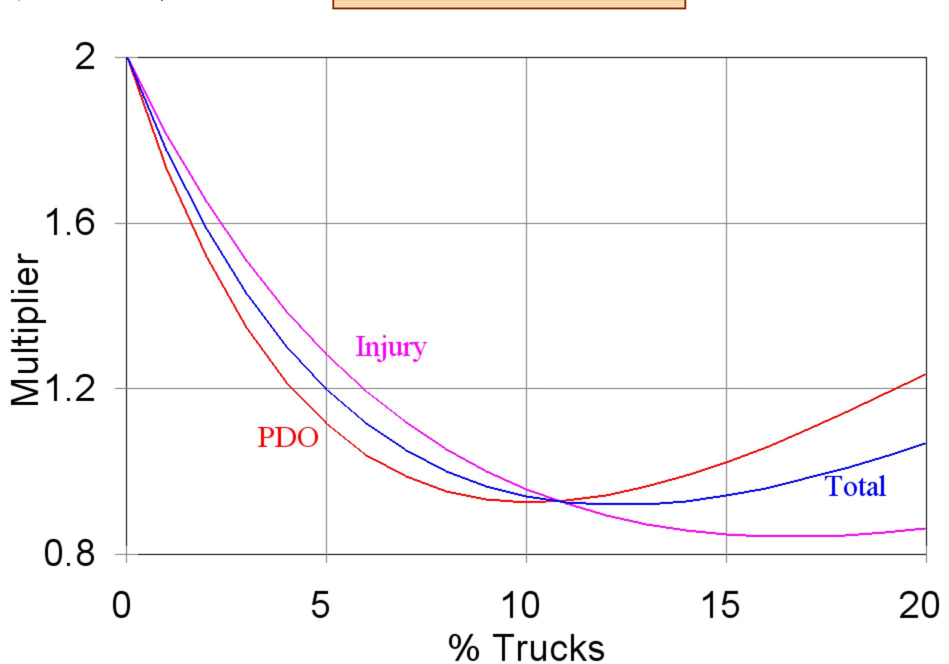
What function can represent the orderly relationship?

Tabulate R's

Plot and look

Seek Function





$$\beta_1 e^{\beta_2 (\%\, trucks)} + \beta_3 (\%\, trucks)$$

3. What (and how) to optimize?

Residuals
- Absolute
- Squared
- ...

Likelihood
- Poisson
- Neg. Bin.
- Neg. Multinomial.

Poisson

Assumption:
Sites with same Observed traits have same means.

Neg. Bin. / Neg. Multinomial.

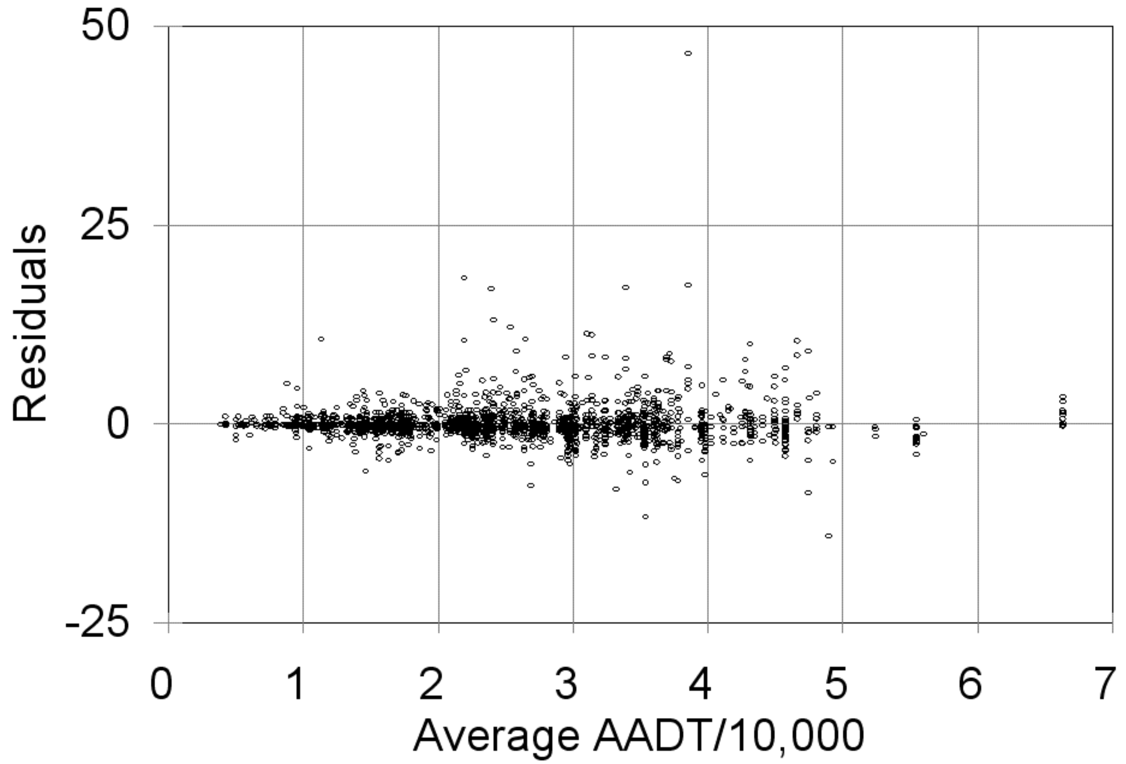
Assumption:
Sites with same observed traits do not have same means.

- Each year has its own random effect
- Random effect common to all years
  - Advantages

4.Residuals are the proof of the pudding.

Residual=Predicted accidents - Count of Accidents

Is this a good fit?

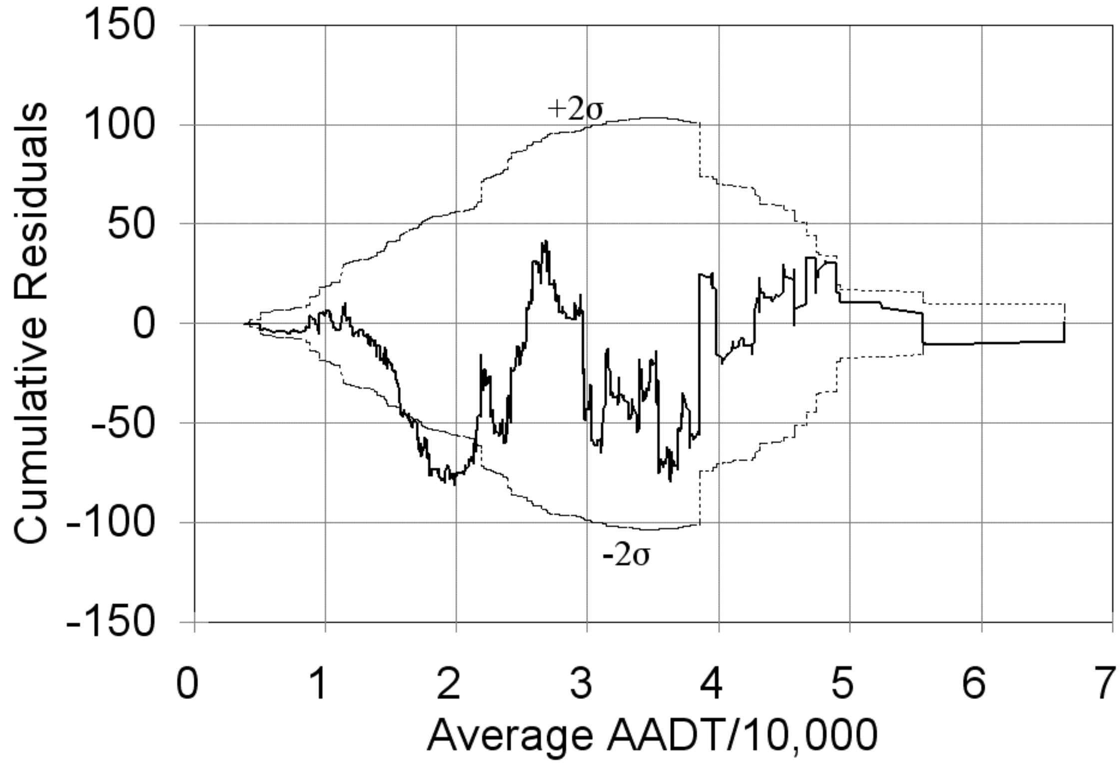


4.Residuals are the proof of the pudding (continued).

The CURE Plot - A Clearer Picture.

Advantages:

- Not only overall fit
- Outliers visible
- Contains clues to improved functional form
- Theory for bounds



Summary

Is it possible to get at cause & effect using haphazard data?

<u>To increase our chances</u>:

1. Use functional forms that satisfy basic logical considerations

2. Examine what kind of functional form is indicated by the data

3. Use a corresponding building block function

4. Use MN because it uses most data and is flexible

5. Examine goodness of fit by a device such as CURE

6. Do not follow routine; explore, backtrack, revise and experiment