Road Surface characteristics and traffic accident rates on New Zealand's state highway network

Robert Davies

Statistics Research Associates

http://www.statsresearch.co.nz

Joint work with Marian Loader, Peter Cenek & Opus International Consultants.

Funded by Transfund

Copy of my report

- There is a copy of my report and some related reports on

  http://robertnz.com

- Look in the section “statistical analysis”

Why am I giving this talk

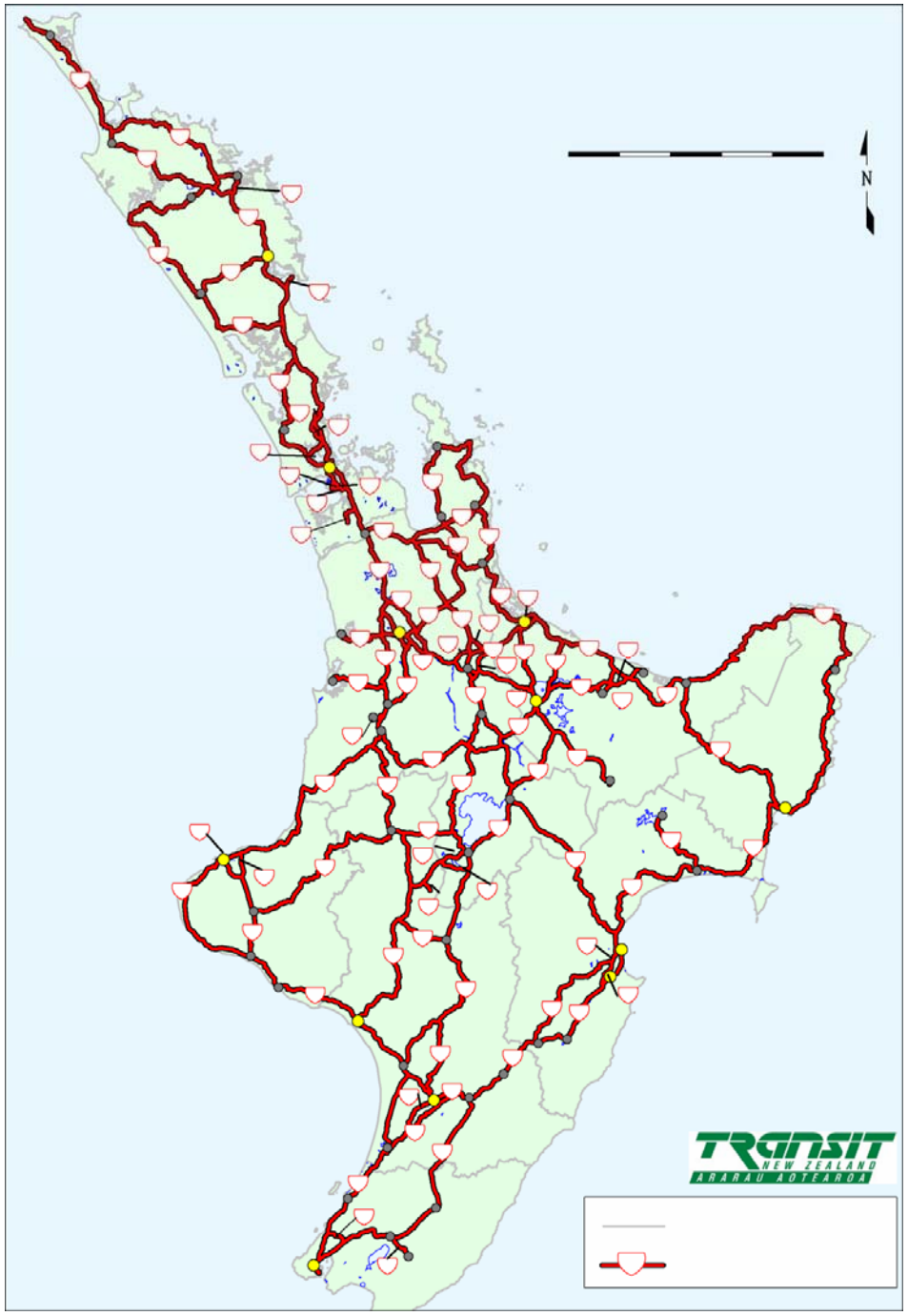
- It represents a slightly unusual analysis
- The data may be typical of the kind of data we might expect from automatic measuring devices
- It gets some interesting and possibly important results
- There are some open questions
- I may have time to talk about my approach to the statistical computing

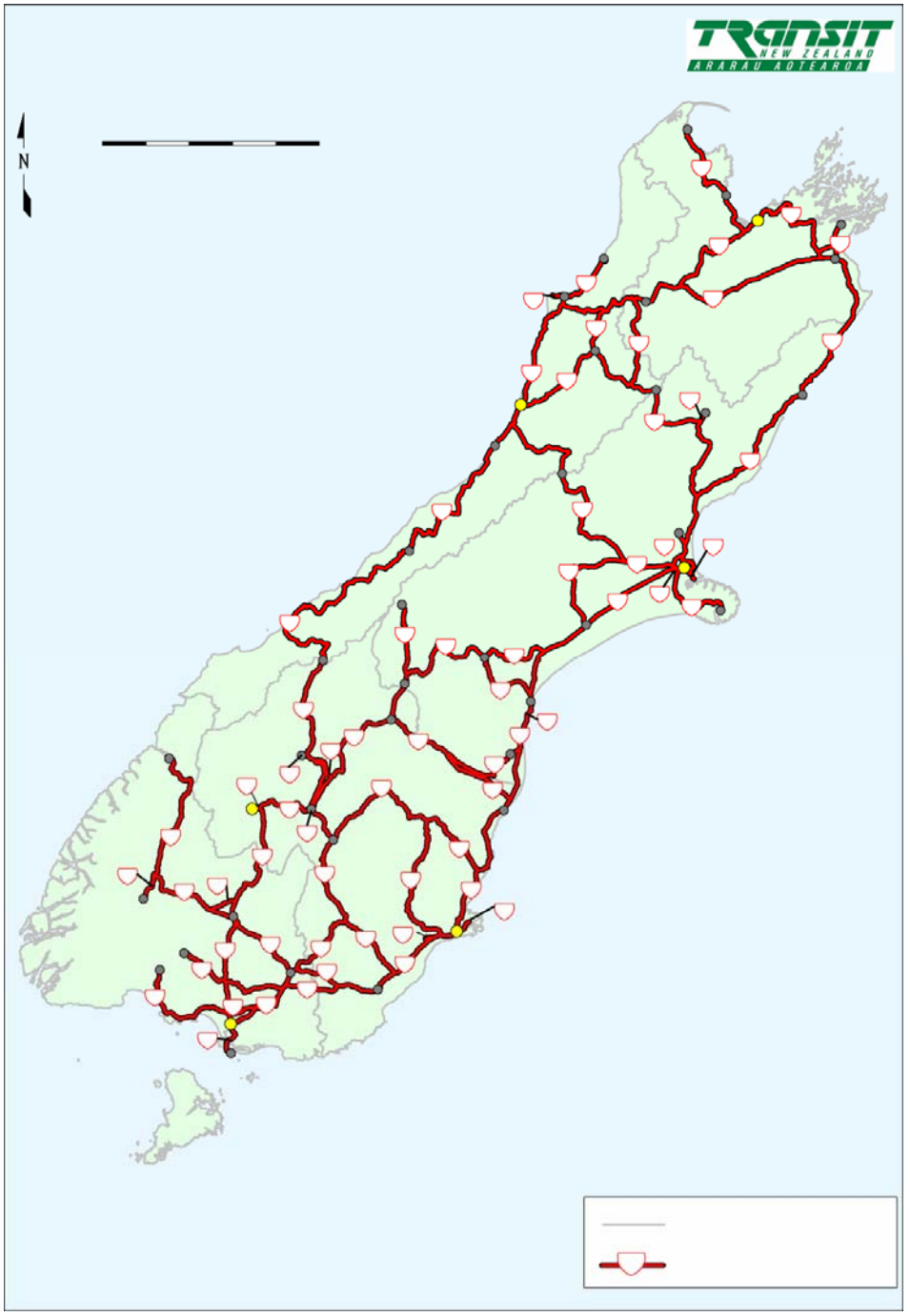
Overview

- I want to relate road crash (traffic accident) rates to road characteristics:
  - Curvature
  - Skid resistance
  - Gradient
  - etc

State highway network

- Network of major roads maintained by the central government
- 10,000 kilometres
- We will be considering 2 lane roads (most of the network)
  - exclude divided roads
  - exclude multi lane roads
  - exclude freeways
- Map on next page


N
TRANSIT
NEW ZEALAND
ARARAU AOTEAROA


N
TRANSIT
NEW ZEALAND
ARARAU AOTEAROA

Road crashes (accidents)

- Consider reported crashes where there is an injury or death
- Around 3200 per year on the State Highway network
- Reporting rate
  - 100% for fatal
  - ~50% for serious injury
  - low for minor injury

Three sets of data

- *SCRIM-plus* data for *state highway network*
- *Land Transport* road crash data
- *TransitNZ* traffic volume data

Six years of data – 1997 to 2002

First data set – *SCRIM-plus* data

curvature	(10 metre intervals)
gradient	(10 metre intervals)
crossfall	(10 metre intervals)
skid resistance	(10 metre intervals)
roughness	(20 metre intervals)
texture	(10 metre intervals)
rut depth	(20 metre intervals)

skid-site – priority for high skid resistance

One million data-points on each side of the road for each year.

First data set - descriptions

- Crossfall – slope across road
- Texture – shows how well water runs off surface
- Rut depth – road surface tends to be lower where car wheels normally go – *rut depth* measures the depth of this
- Skid site – see next page

skid site	Description	Notes	scrim site investigatory level
4	Normal roads	All normal roads. (Undivided carriageways only)	0.4
3	Approaches to road junctions	Approaches to road junctions. Down gradients 5-10%	0.45
2	Curve <250m rad. Gradient>10%	Curve <250m radius. Gradient > 10%.	0.5
1	Highest priority	Railway level crossing, approaches to roundabouts, traffic lights, pedestrian crossings and similar hazards.	0.55
5	Divided carriageway	Divided carriageways	0.35

Second data set – *LTSA* crash data

Reported injury and fatal crashes

- location
- movement classification (e.g. overtaking)
- road condition (e.g. wet)

14,000 crashes over 6 years

(number is lower than given previously because we couldn’t locate all the crashes)

Categories for crashes possibly involving skidding

A	overtaking and lane change
B	head on
C	lost control or off road (straight roads)
D	cornering
E	collision with obstruction
F	rear end
G	turning versus same direction
H	crossing (no turns)
J	crossing (vehicle turning)
K	merging
L	right turn against
M	manoeuvring
N	pedestrians crossing road
P	pedestrians other
Q	miscellaneous

Third data set – *TransitNZ* data

- Average daily traffic (ADT)
- Urban or rural
- Road width
- Number of lanes

Preliminary approach to the analysis: two way tables

Crashrisk

crashes per 100 million km of vehicle travel

Skid site category	Skid resistance range					
	< 0.3	0.3 to 0.4	0.4 to 0.5	0.5 to 0.6	0.6 to 0.7	> 0.7
4 (normal)	17	16	13	13	14	12
3 (junctions)	44	29	27	26	23	32
2 (curves, hills)	62	39	33	31	31	33
1 (highest priority)	0	44	52	47	47	40

Crashrisk – notes about table

- Yellow shows where we have sufficient data to make inferences
- Note increase if we go down or left in the table
- More pronounced for *wet* crashes
- Wet rates are lower because we can't allow for the % of time roads are wet

Number of crashes

Skid site category	Skid resistance range					
	< 0.3	0.3 to 0.4	0.4 to 0.5	0.5 to 0.6	0.6 to 0.7	> 0.7
4 (normal)	17	322	2650	3094	811	36
3 (junctions)	10	163	1249	1259	220	13
2 (curves, hills)	12	200	942	832	211	12
1 (highest priority)	0	35	216	191	36	1

Crashrisk
Wet crash rate

Skid site category	Skid resistance range					
	< 0.3	0.3 to 0.4	0.4 to 0.5	0.5 to 0.6	0.6 to 0.7	> 0.7
4 (normal)	2	5	3	2	2	1
3 (junctions)	4	7	6	5	2	3
2 (curves, hills)	36	15	10	8	8	8
1 (highest priority)	0	13	9	9	11	0

Main method of analysis: modified Poisson regression model

Poisson regression model (with offset)

Start with a Poisson regression model: each 10 metre section can generate crashes at a rate

$$\boldsymbol{a} \exp(\boldsymbol{L})$$

where $\boldsymbol{a}$ is the average daily traffic volume

$\boldsymbol{L}$ is a linear combination of predictor variables ($\boldsymbol{a}$ also appears in the $\boldsymbol{L}$ term)

The actual crash risk is given by

$$\exp(\boldsymbol{L})$$

times some suitable factor to get the units right.

Two problems:

1: We don’t know the location of the crashes very accurately.
Therefore average the crash-rates from the model over 100 metres from each side of the site that we want the observed crash-rate for.

2: We don’t believe the records for the direction of the vehicle.
Therefore sum over the two sides of the road (we aren’t doing dual carriageway so there are always two sides).

Model is no longer standard Poisson regression.

But will we can still fit by maximum likelihood

I tried four different sets of crash data:

- **the complete data**;
- the crashes most likely to have involved skidding;
- the crashes where the road was wet;
- the crashes most likely to have involved skidding and where the road was wet.

Two analyses on each

- **involving all of the predictors and spline or polynomial functions of the variables**;
- reduced model using simplified functions (no splines) and a smaller number of predictors.

Predictor	DF	1% pt	chi sq.
year	5	15.1	102.1
region	6	16.8	122.2
urban rural	1	6.6	203.2
skid site category	2	9.2	2015.8
spline(log curvature)	5	15.1	2036.8
poly(log daily traffic)	2	9.2	281.4
spline(gradient)	5	15.1	258.7
poly(skid resistance)	2	9.2	125.1
spline(log roughness)	3	11.3	56.1

Predictor	DF	1% pt	chi sq.
spline(log roughness)	3	11.3	56.1
skid site * skid res.	2	9.2	28.9
spline(sqrt rut_depth)	4	13.3	27.4
cway_width	1	6.6	23.4
texture	1	6.6	4.3
lanes_category	1	6.6	3.5
irr_width	1	6.6	0.3
crossfall	1	6.6	0.0
abs(crossfall)	1	6.6	0.4

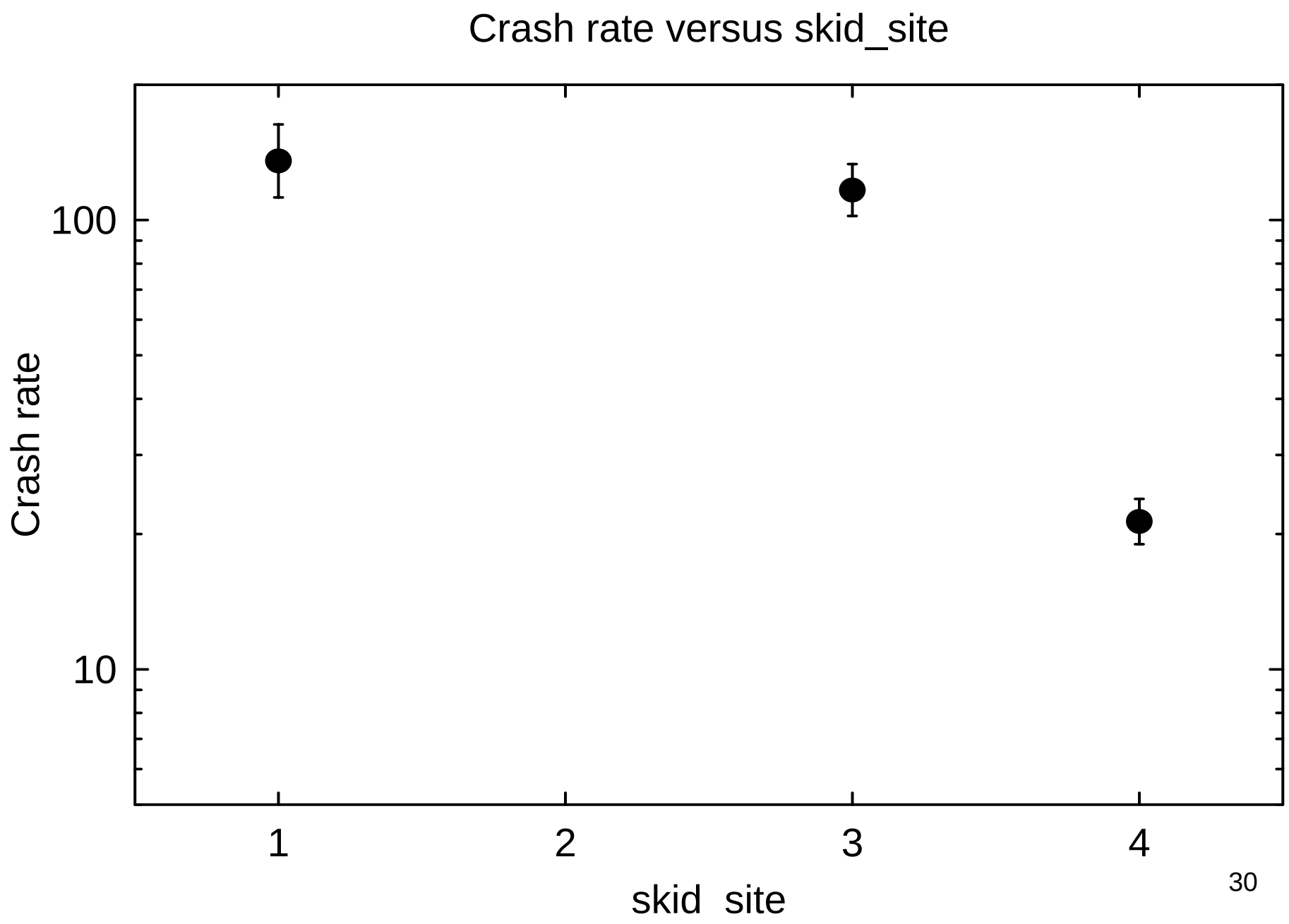
Significances

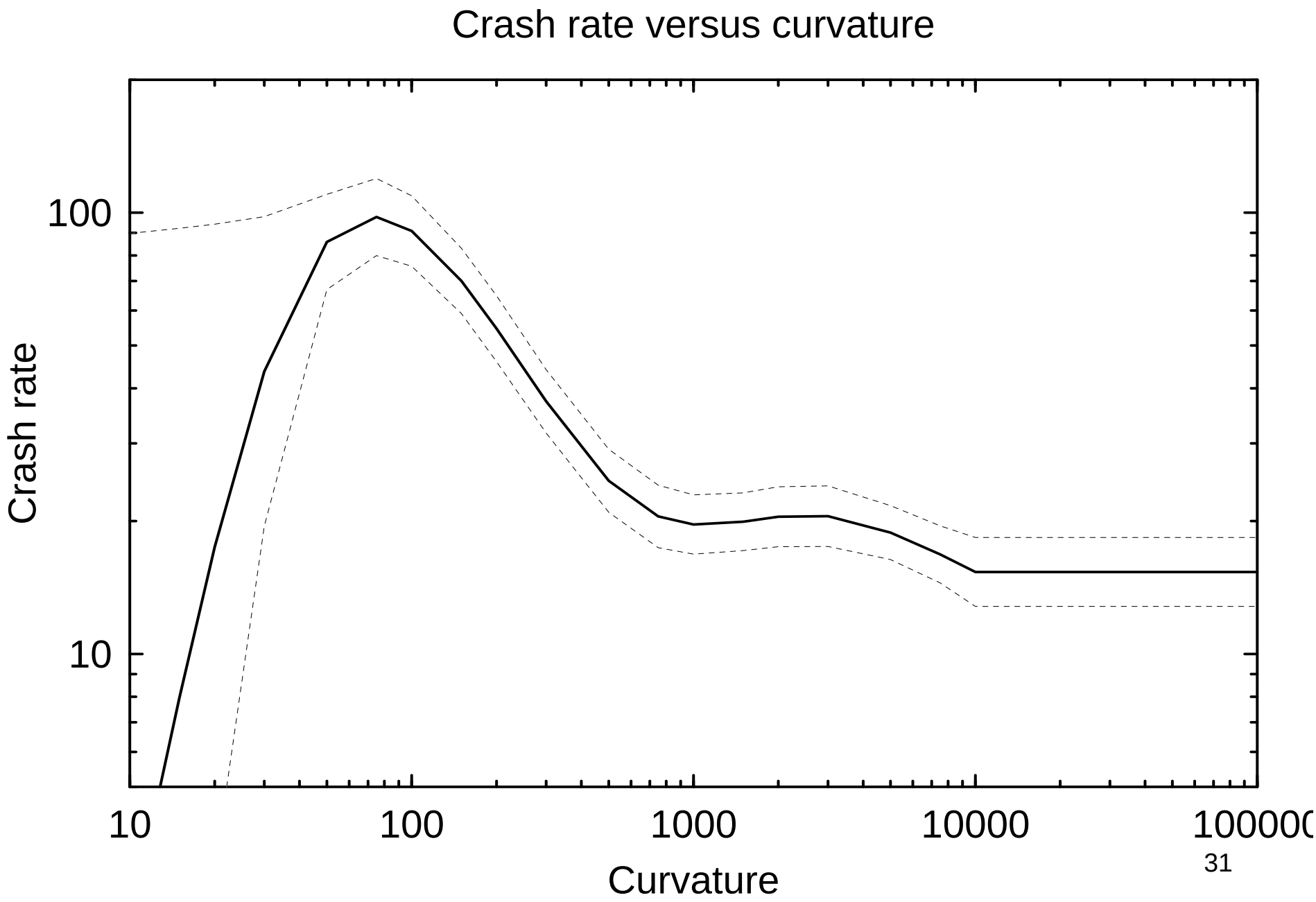
- Don’t believe significance levels
- Only include effects down to *spline(log roughness)*
- Next sheet shows significance levels for the 4 analyses
- “Significant” effects are shown in yellow
- iri = roughness – doubtful whether this should be included

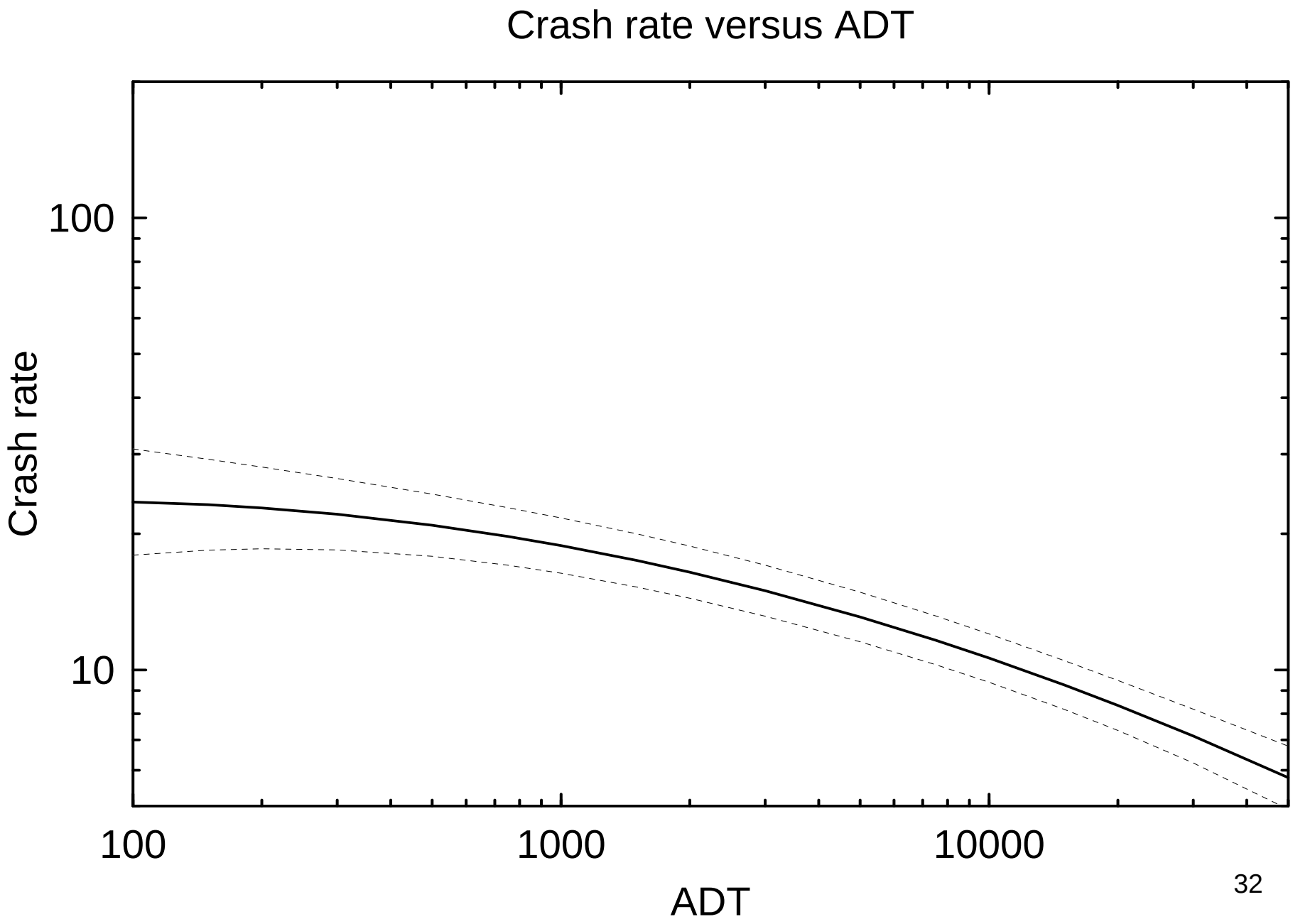
Predictor variable	df	1% pt.	Chi-squared values			
			All	Selected	Wet	Wet selected
year	5	15.09	102.10	56.56	46.76	39.46
region	6	16.81	122.23	93.91	81.25	73.10
urban_rural	1	6.63	203.20	148.60	3.89	64.69
skid_site	2	9.21	2015.70	206.72	315.64	45.64
spline6(log10_curvature)	5	15.09	2036.80	2874.40	1365.80	1620.80
poly2_log10_ADT	2	9.21	281.44	300.17	55.99	48.30
spline6(gradient)	5	15.09	258.68	15.26	24.10	12.42
poly2_scrim-0.5000	2	9.21	125.05	148.28	147.93	172.11
spline4(log10_iri)	3	11.34	56.12	51.36	18.46	15.23
skid_site*(scrim-0.5000)	2	9.21	28.89	12.59	13.86	2.61
spline5(sqrt_rut_depth)	4	13.28	27.39	3.06	2.70	0.80
cway_width	1	6.63	23.42	0.04	13.29	0.42
texture	1	6.63	4.27	0.64	0.82	0.16
lanes_category	1	6.63	3.48	0.03	5.50	0.07
irr_width	1	6.63	0.31	1.89	4.24	1.24
crossfall	1	6.63	0.02	0.00	0.43	0.46
abs_crossfall	1	6.63	0.35	0.01	1.11	0.35

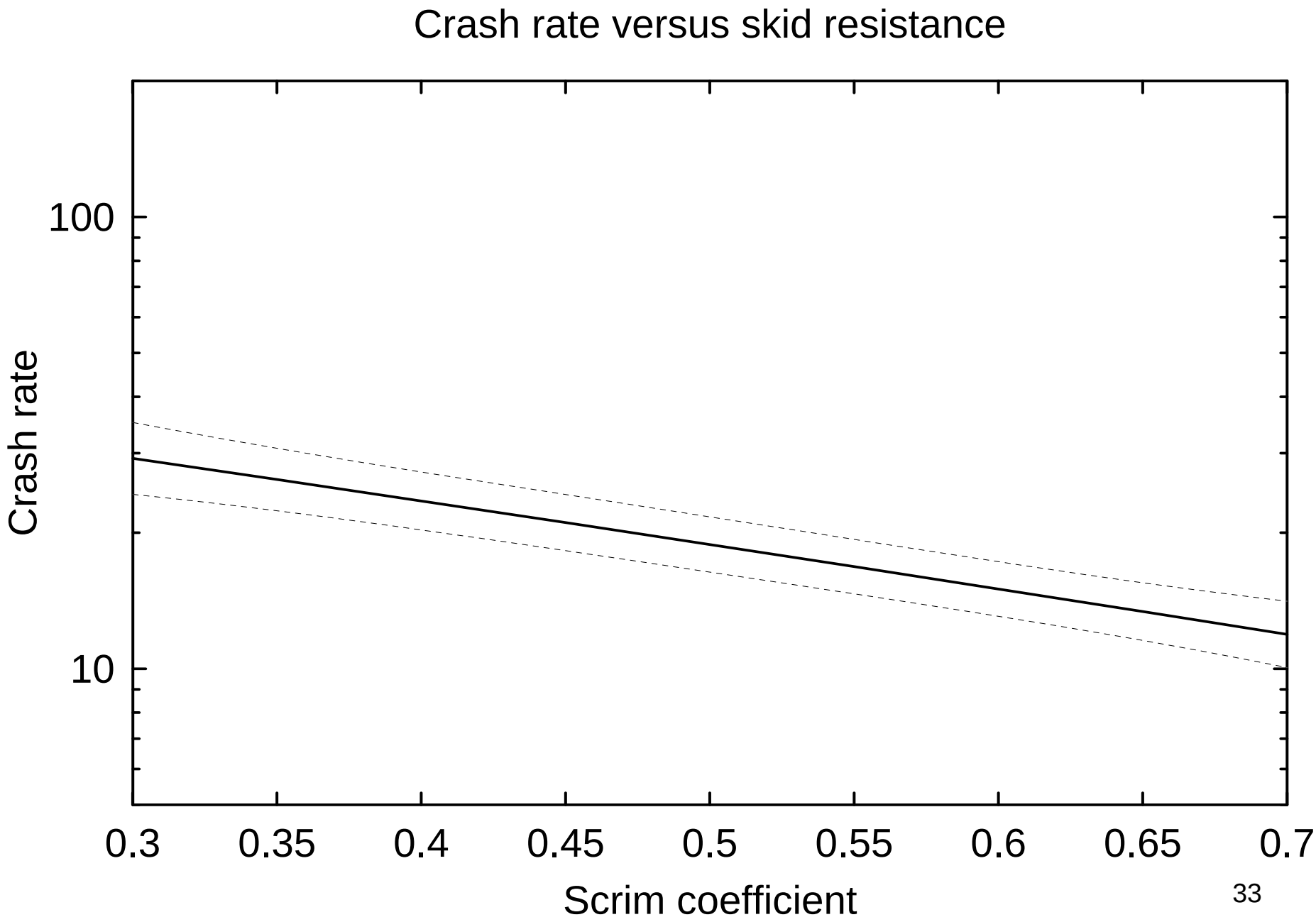
Graphs

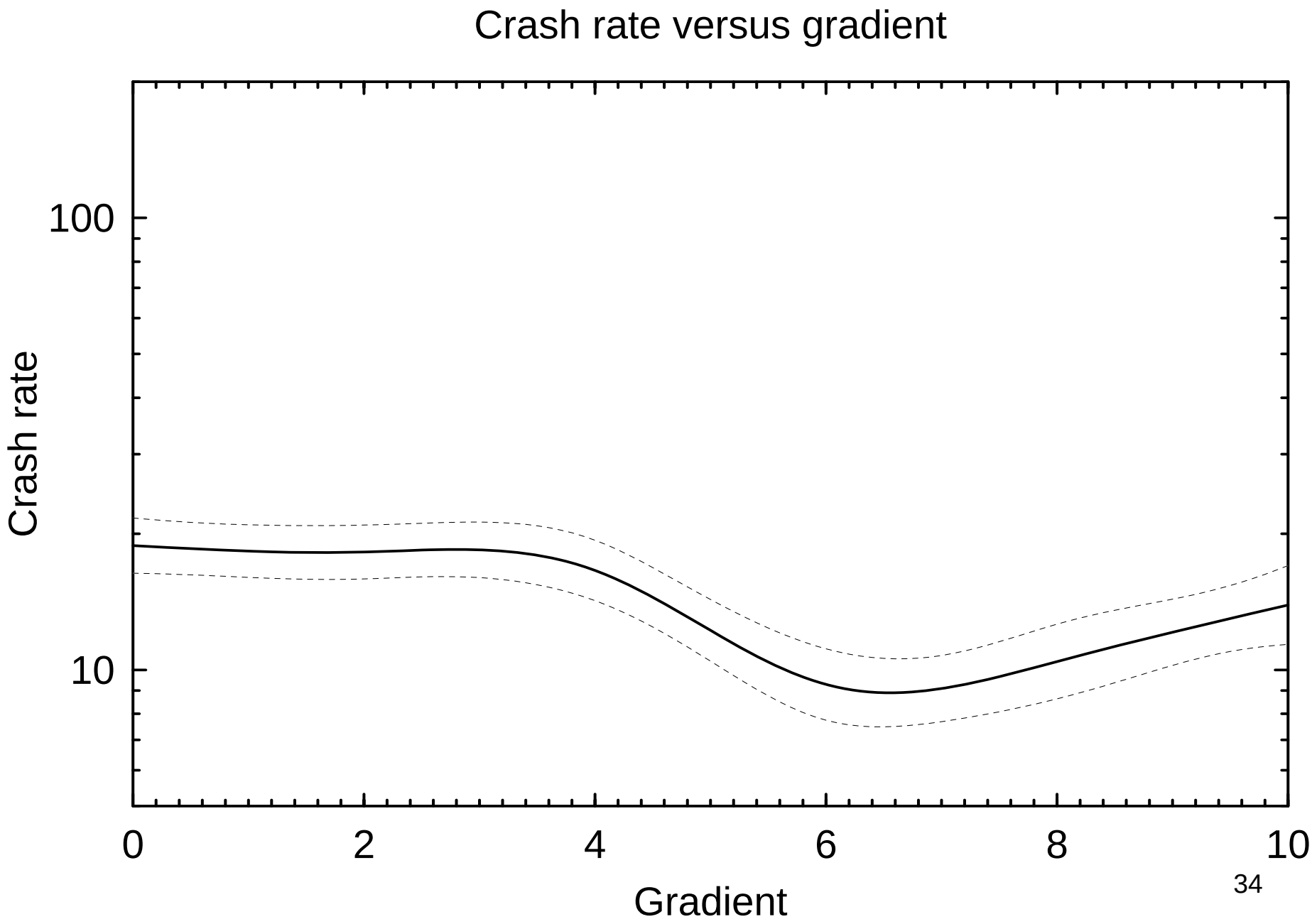
- Graphs of estimates of effects
- Select the most commonly occurring set of X variables
- Then vary each one in turn and graph the predicted values
- Don’t really believe the confidence bounds
- *Gradient effect* probably wrong

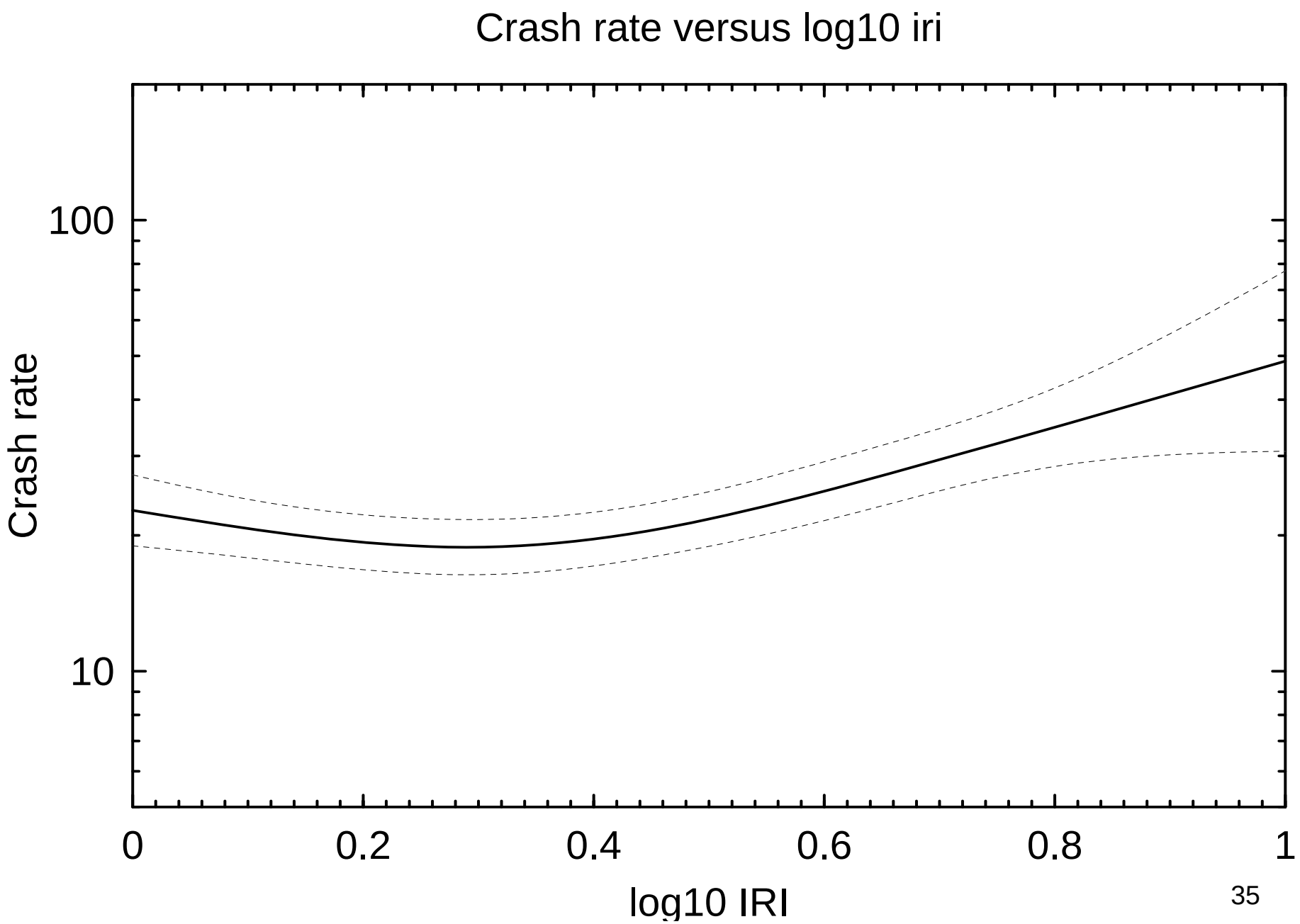

Crash rate versus skid_site
100
10
Crash rate
1
2
3
4
skid site


Crash rate versus curvature
100
10
Crash rate
10
100
1000
10000
100000
Curvature

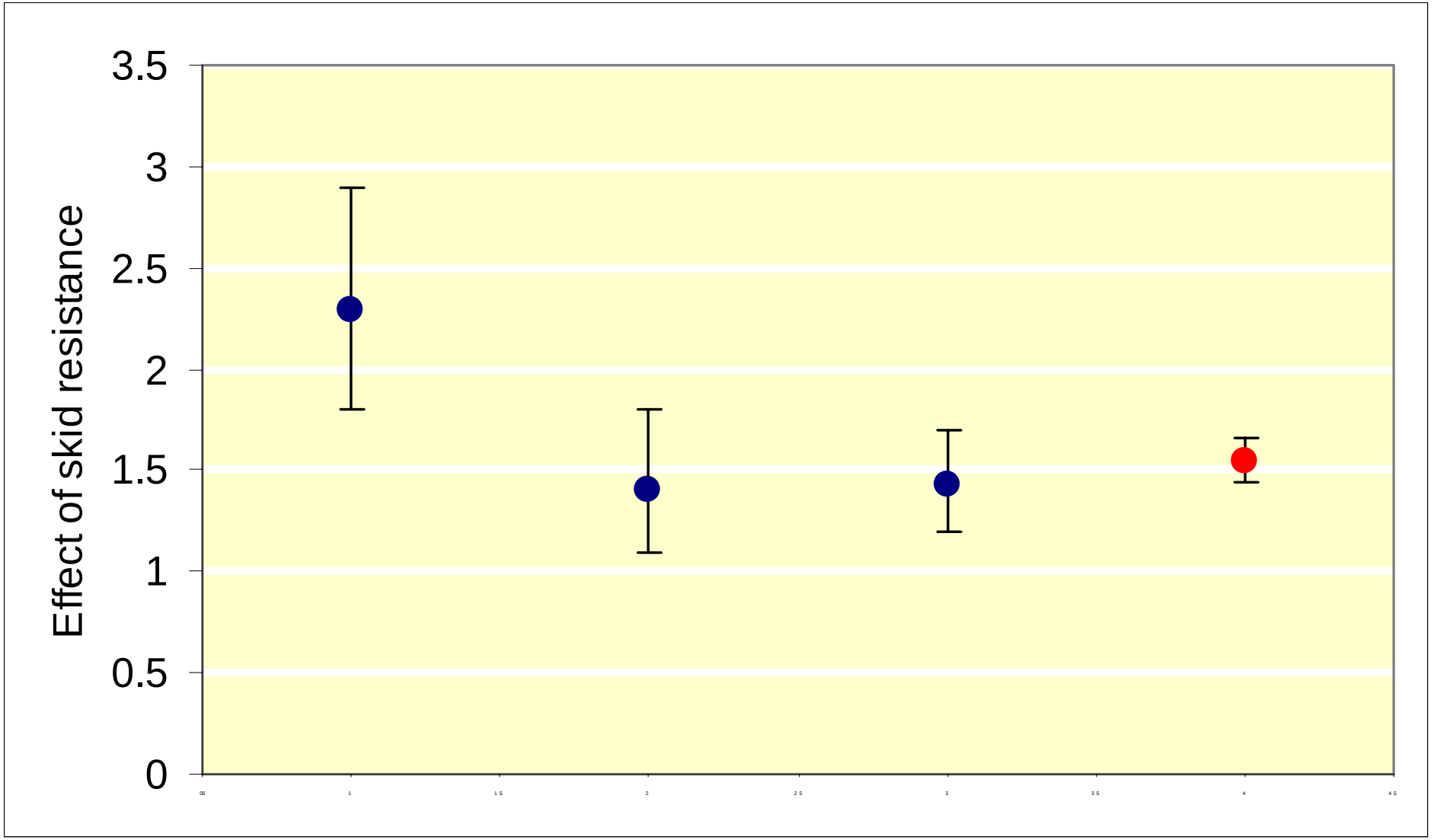

Crash rate versus ADT
100
10
Crash rate
100
1000
10000
ADT


Crash rate versus skid resistance
100
10
Crash rate
0.3
0.35
0.4
0.45
0.5
0.55
0.6
0.65
0.7
Scrim coefficient


Crash rate versus gradient
100
10
Crash rate
0
2
4
6
8
10
Gradient

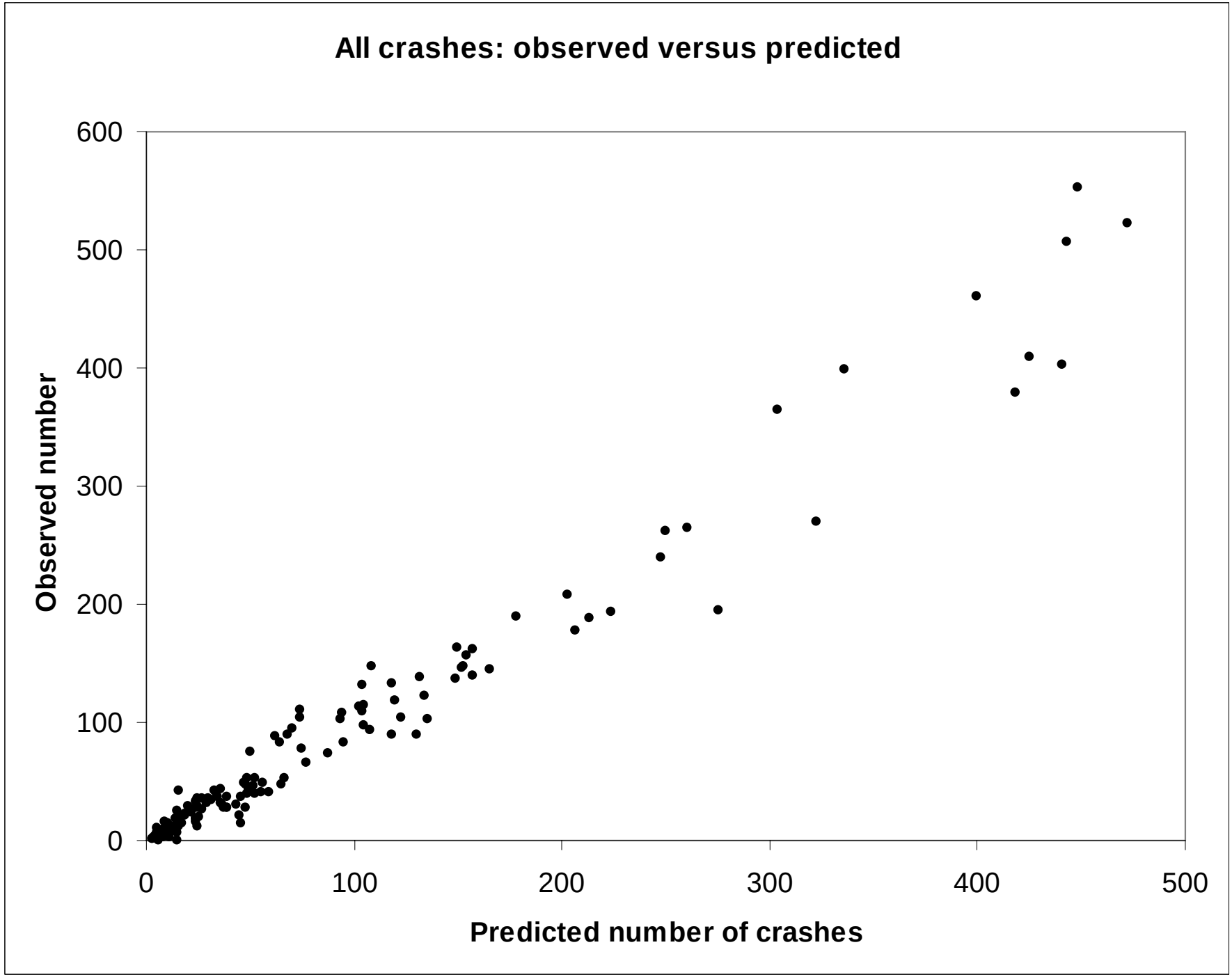

Crash rate versus log10 iri
100
10
Crash rate
0
0.2
0.4
0.6
0.8
1
log10 IRI

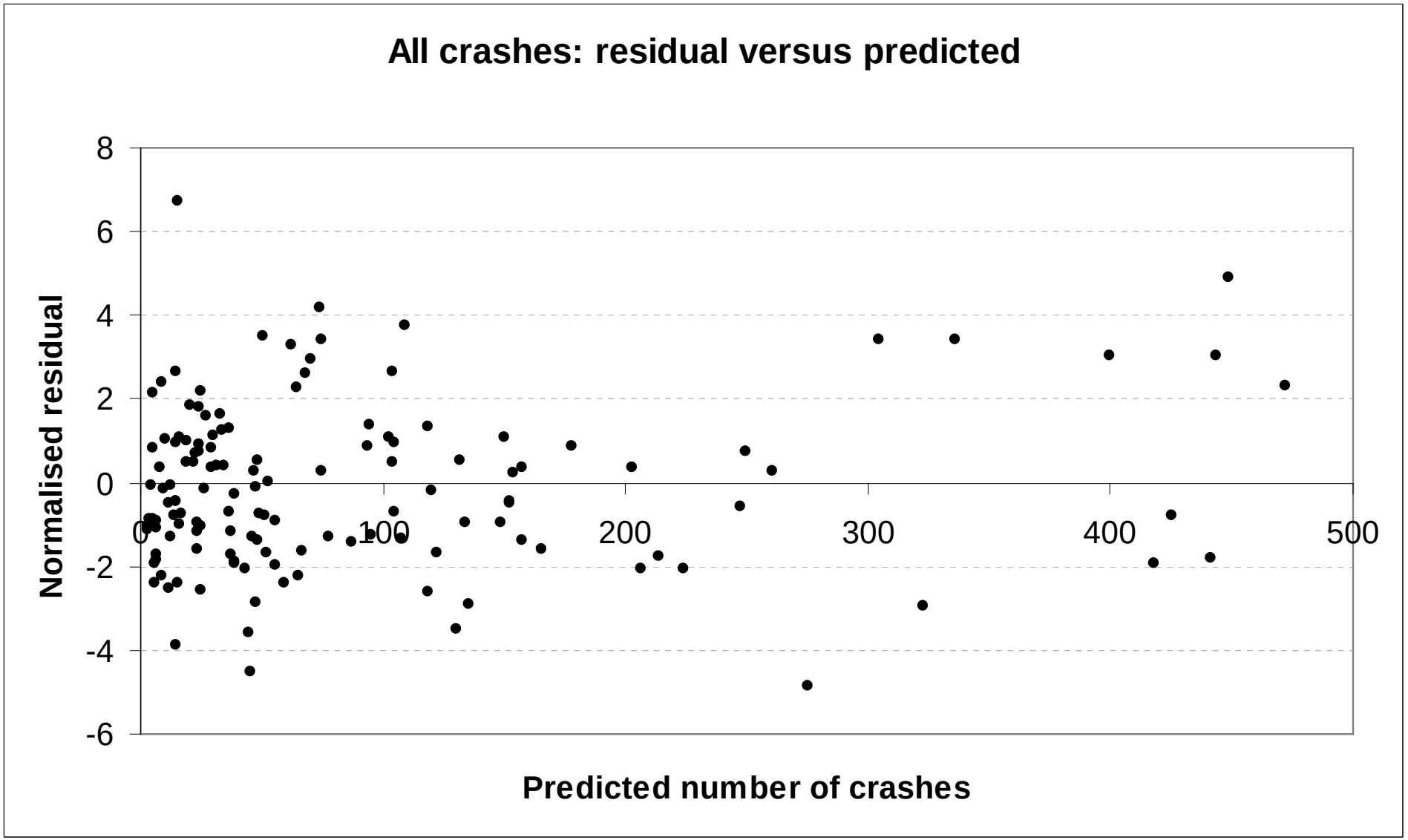
How does this study compare with other studies?

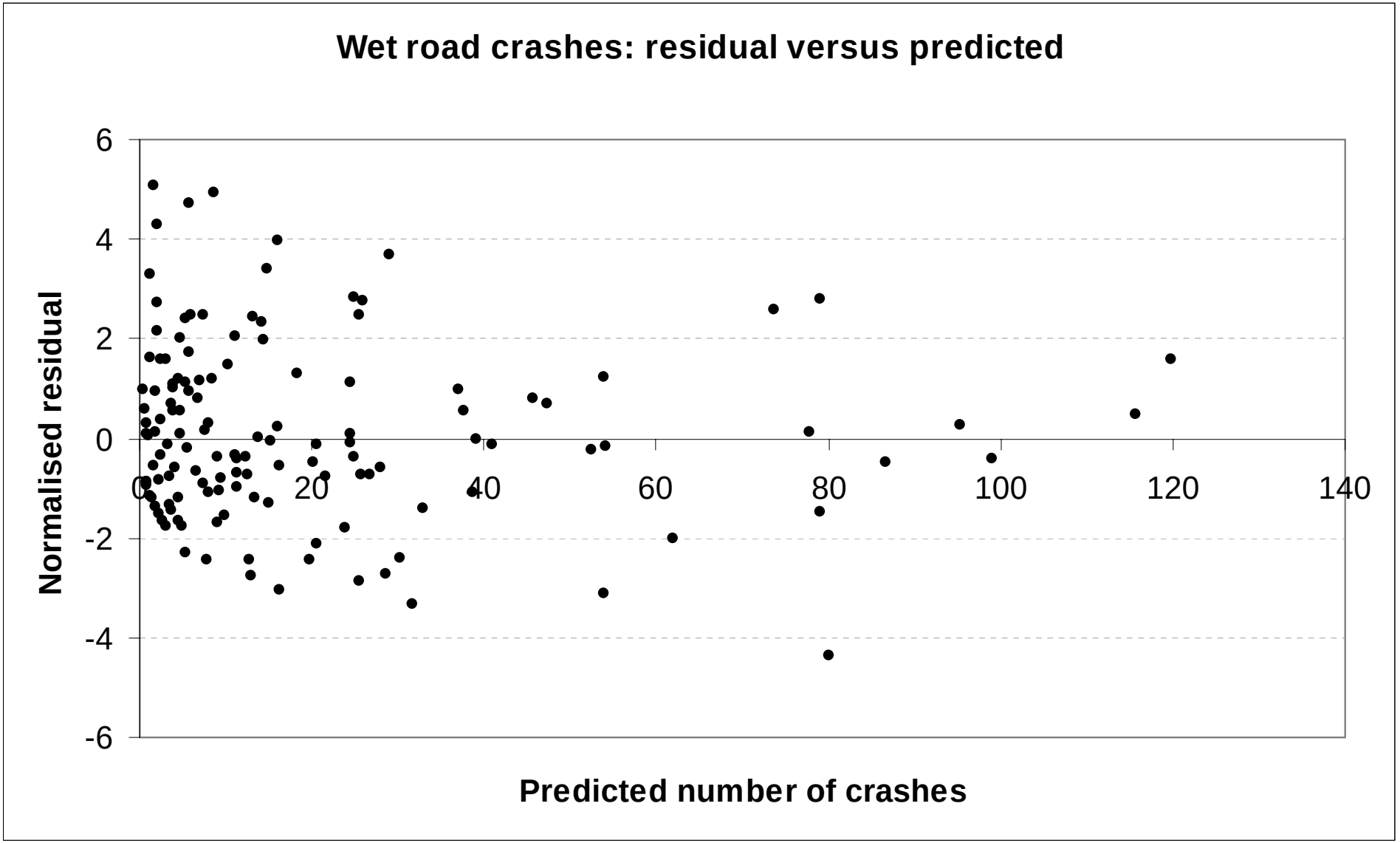


Residuals

- Doesn’t make sense to look at residuals in usual sense
- Divide up roads by highway number intersected by region (8 regions in all)
- Look at predicted and observed number of crashes
- Residuals seem to be too large by a factor of 2


All crashes: observed versus predicted
Observed number
Predicted number of crashes
0
100
200
300
400
500
600


All crashes: residual versus predicted
Normalised residual
8
6
4
2
0
-2
-4
-6
0
100
200
300
400
500
Predicted number of crashes


Wet road crashes: residual versus predicted
Normalised residual
6
4
2
0
-2
-4
-6
0
20
40
60
80
100
120
140
Predicted number of crashes

Does averaging matter?

- We have been averaging over 210 metres
- What happens if we vary this?
- Calculate log likelihood and chi-squared values when we vary averaging length
- Yellow in following slides shows best values

Does averaging matter – all crashes

averaging length	log likelihood	skid-site chi-sq	curvature chi-sq	skid res. chi-sq
30	-83,944	1769	1115	131
90	-83,839	2899	1117	133
110	-83,829	2148	1191	134
130	-83,839	2153	1217	135
210	-83,988	1832	1265	144

Does averaging matter – wet crashes

averaging length	log likelihood	skid-site chi-sq	curvature chi-sq	skid res. chi-sq
30	-17,584	33	677	162
210	-17,466	31	989	184
310	-17,459	28	1019	194
410	-17,478	33	1000	186

What-if study

- Upgrade the skid-resistance on all skid-site 2 (curvature < 250m radius or gradient > 10%)
- How many crashes would we save in 2001?
- How much road would we have to upgrade?

What-if study: skid-site 2 locations, 2001 data

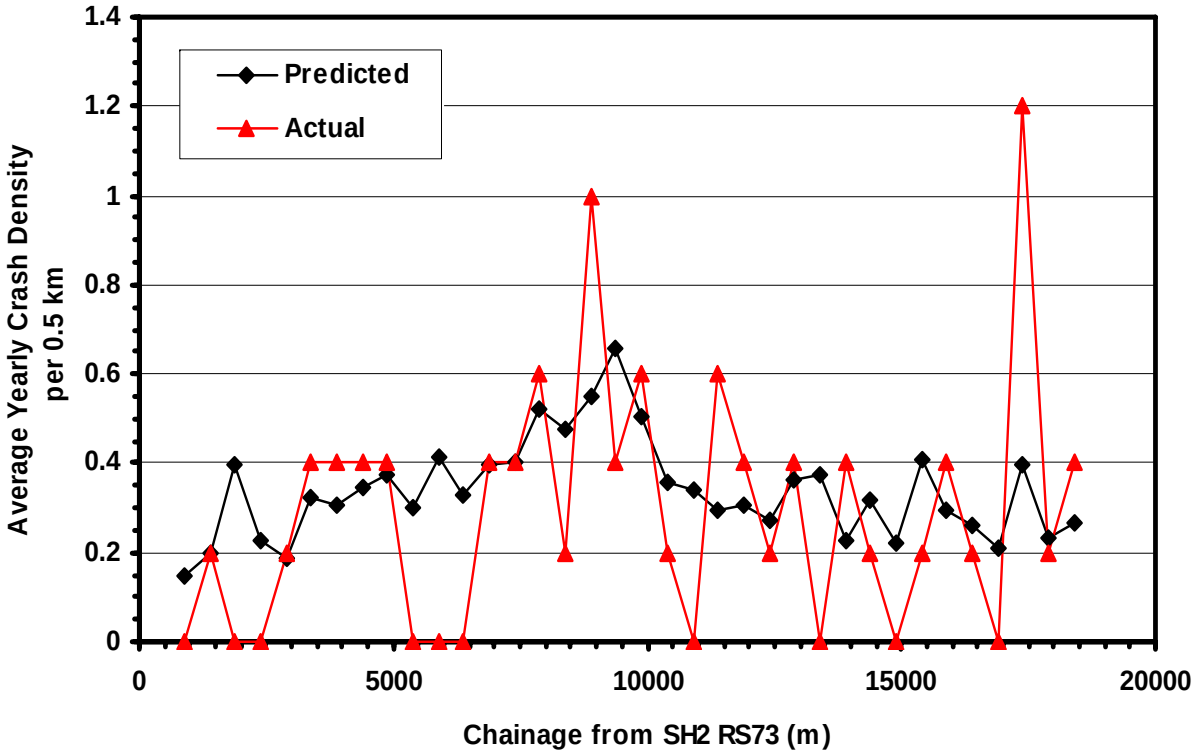
min. skid resistance	fix for traffic ≥	fix length	predicted crashes	saved crashes
0	0	0	370	0
0.4	0	120	369	2
0.5	0	719	357	13
0.6	0	1574	330	41
0.4	1000	93	369	2
0.5	1000	545	358	12
0.6	1000	1055	333	37
0.4	5000	18	370	1
0.5	5000	98	365	5
0.6	5000	169	356	14

Case study

- Karangahake Gorge
- How do observed and predicted values agree
- Details on website

Karangahake Gorge

Predicted and actual



Discussion

- How credible are the results?
- How to handle additional error structure
- Use of 10 metre sections rather than combining into (e.g.) 100 metre sections
- Danger in taking data beyond its *design accuracy*
- How to present confidence intervals on graphs of effects of variables
- Use C++ as statistical programming language

How credible are the results (i)?

- Retrospective study
- Predictor variables subject to error
- Road properties not in model
- Different skid-site 1 (and 3) characteristics
- Don’t know fraction of time road is wet
- Additional error structure?
- Choice of averaging length
- Non-linearity and interactions

How credible are the results (ii)?

- The results make sense
- They are stable – in that small changes to the analysis don't make much difference to the results
- The skid resistance results are similar to those from earlier studies
- Analysis on 1997-1999 & on 2000-2002 seem to agree
- But – we need international comparisons

Additional error structure

- Can’t use residual deviance as scaling factor
- Divide network into *blocks* – what length – does it matter?
- Use hidden Markov point process model?
- Could traffic flow data be the problem?

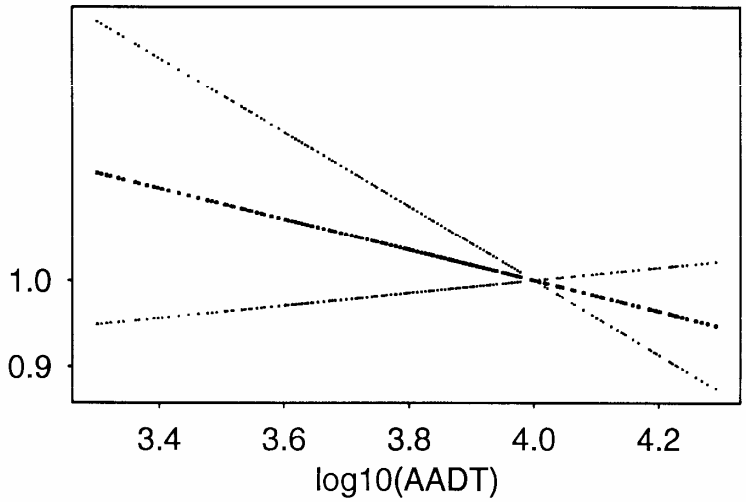
More discussion

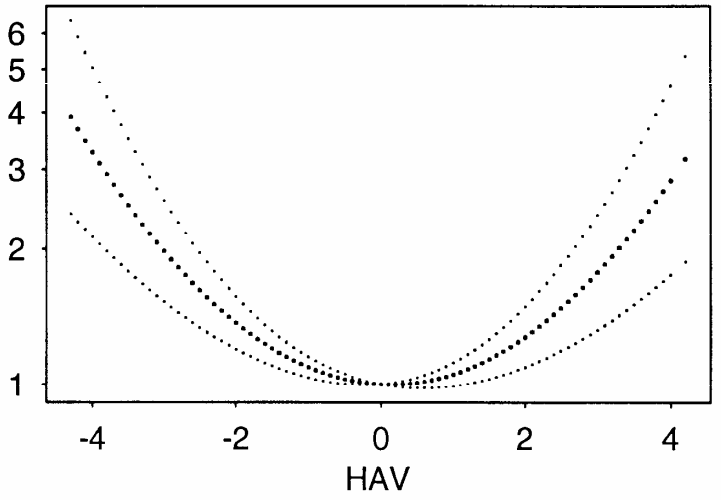
- How credible are the results?
- How to handle additional error structure
- Use of 10 metre sections rather than combining into (e.g.) 100 metre sections
- Danger in taking data beyond its *design accuracy*
- How to present confidence intervals on graphs of effects of variables
- Use C++ as statistical programming language

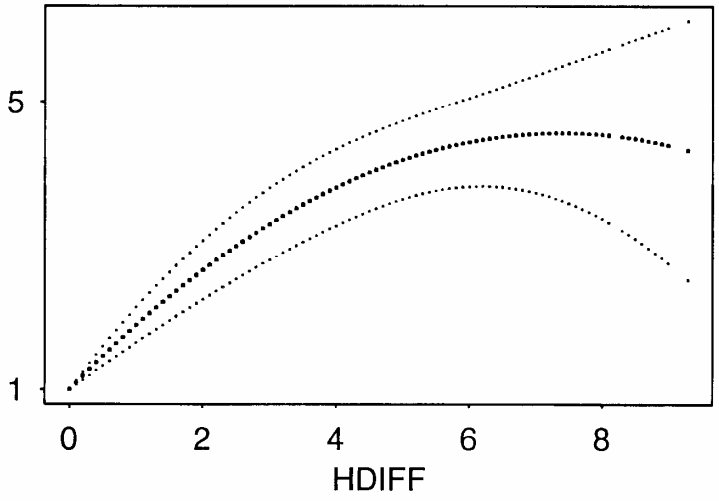
Alternative way of showing confidence intervals

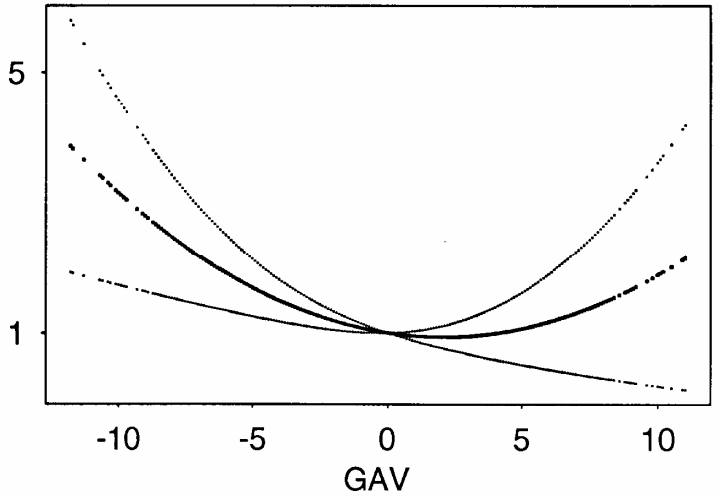
- This from a previous study
- Select a particular value of the predictor
- Then show the confidence intervals for effect of changing to a different value of the predictor

Relative risk functions









Programming details

- Preliminary processing using SAS at Opus and SQL server on my computer
- Main model fitting using C++ programs using my matrix package, automatic differentiation package and a new array and model formula package
- Plots by *Gnuplot*
- Fit runs take about 1 hour
- See http://robertnz.com

Array and model formula package

- C++ package
- Named array (Array, Name, Missing value indicator)
- Factor
- Model Formula
- Try to get expressiveness of R and Splus
- Compiled code speed
- Flexibility of C++ code
- Not ready for release – not on website

That’s all