Institute of Actuaries of Australia

Does rainfall increase or decrease motor accidents?

Or, a reflection on the good, the bad and the ugly (in statistics)

Prepared by Gráinne McGuire

Presented to the Institute of Actuaries of Australia
16th General Insurance Seminar 9-12 November 2008
Coolum, Australia

This paper has been prepared for the Institute of Actuaries of Australia's (Institute) 4th Financial Services Forum 2008. The Institute Council wishes it to be understood that opinions put forward herein are not necessarily those of the Institute and the Council is not responsible for those opinions.

The Institute of Actuaries of Australia
Level 7 Challis House 4 Martin Place
Sydney NSW Australia 2000
Telephone: +61 2 9233 3466 Facsimile: +61 2 9233 3446
Email: actuaries@actuaries.asn.au Website: www.actuaries.asn.au

1 Abstract

There are lies, damn lies and statistics.

With apologies for the somewhat frivolous diversion into clichés, this paper looks at two issues:

- Firstly – the good, the bad and the ugly in statistics. How does the use of incorrect [bad] or somewhat inappropriate [ugly] methods lead to bad results [lies, damn lies, statistics]? What effect does the use of good methods have on the analysis?

- Secondly, does increased rainfall lead to more motor vehicle accidents? Or less? The analysis here, and other results in the literature suggest that, in fact, the answer is both. The effect is determined by the quantity of rain and when it actually fell.

Keywords

Claim frequency, generalised linear models, motor accidents, rainfall, regression, statistics.

2 Introduction

Does rainfall increase the incidence of motor vehicle accidents on any particular day? Common sense suggests it does – road conditions are more dangerous so surely it must. However, common sense also suggests that a cannon ball falls faster than a feather. While we might observe that to happen, we now know that it's not the case once we factor out the effects of air friction. So it is worthwhile to critically re-examine the rainfall question – does the daily accident rate increase or are other things going on of which we need to take into account.

Just like it was necessary to do the feather and cannon ball experiment properly (in a vacuum), so to is it necessary to get the statistics right when analysing rainfall and motor accidents. Some discussion of this point has been given in Davies et al (2004). The use of inappropriate methods can lead to misleading or incorrect results.

The main statistical analysis in this paper is similar to that in Eisenberg (2004). That paper looked at many of the points addressed here and demonstrated suitable statistical techniques for the analysis of rainfall data in relation to motor vehicle accidents.

This paper is structured as follows:

- A brief description of the data used is given in Section 3;
- An overview of the statistical methodology used is provided in Section 4;
- The results of the statistical analyses are given in Section 5;
- Some discussion of the appropriateness of the statistical methodology is given in Section 6;
- Concluding comments are made in Section 7.

So, let's head down to the OK Corral and have a shoot-out between the good, the bad and the ugly in statistical techniques to determine whether rainfall is a baddie or a goodie.

3 Data

3.1 Source data

All statistical analyses require data. Used in this paper are:

- Compulsory third party ("CTP") claims arising out of accidents in Perth from July 1993 to December 2005. For each of the claims the following was available:
  - Date of accident
  - Time of accident;
- Vehicle registrations in Western Australia;
- Daylight hours in Perth throughout the year;
- Rainfall data for Perth weather stations from July 1993 to December 2005.

3.2 Data manipulation

Some manipulation was necessary prior to analysing the data.

- Monthly Perth CTP claim frequency was calculated as the annualised monthly claims divided by the number of vehicles registered. For the later months, an estimate was made of the number of incurred but not reported claims;
- Rainfall statistics were based on the average rainfall measured at a number of Perth weather stations. Rainfall was calculated both monthly and daily. The numbers of days of rain per month were also calculated;
- Each claim was matched to a day of rain. "Rainfall days" run from 9am so accident days were also defined to run from 9am; i.e. accidents occurring at 8.59am and 9.01am, under this definition, would occur on different days.

4 Methodology

The methods used in this paper are restricted to regression techniques. Another possibility for exploratory work (as opposed to modelling) is matched sampling. See Davies et al (2004) for a description of this and its application to similar data from a different state.

4.1 Normal linear regression

The first technique considered is normal unweighted linear regression (using the "linest" function in excel). This:

- Assumes the responses are normally distributed
- Assumes that the relationship between the response and the covariates is linear, i.e. $\mathbf{Y} = \boldsymbol{\beta}\mathbf{X}$ where **Y** is the vector of responses, **β** is the parameter vector and **X** is a matrix of covariates.

Unweighted linear regression is used to do some analysis of monthly claim frequency data. More details and results are given in Section 5.1.

4.2 Generalised linear models

Generalised linear model ("GLM") methodology is also used in this paper. The particular type of GLM used here assumes that the claim frequency is distributed according to an Over-dispersed Poisson ("ODP"). A Poisson distribution is a natural choice for modelling numbers, hence its use here. However the standard Poisson assumes that the variance is equal to the mean. In many applications, the data exhibit more variability than this so it is common to use an ODP, where the variance is a scaled version of the mean for some scale value $\phi > 1$.

Where a Poisson distributed variable takes values 0,1,2, …, an ODP variable takes values $0, \phi, 2\phi, \ldots$ In practice, this is not a significant limitation. Parameter estimates from a Poisson GLM and an ODP GLM will be identical, but the parameter covariance matrix will be scaled by ϕ.

The ODP GLM was used to build a model of daily claim frequency, where each day is the rainfall day, starting at 9am. A description of the model and its results appears in Section 5.2.

5 Results – rainfall

Before beginning the analysis, a plot of monthly Perth annualised CTP claim frequency is given in Figure 1. The reduction in CTP claims over the years from July 1993 to December 2005 is evident from this plot.

Figure 1 – Monthly Perth CTP claim frequency

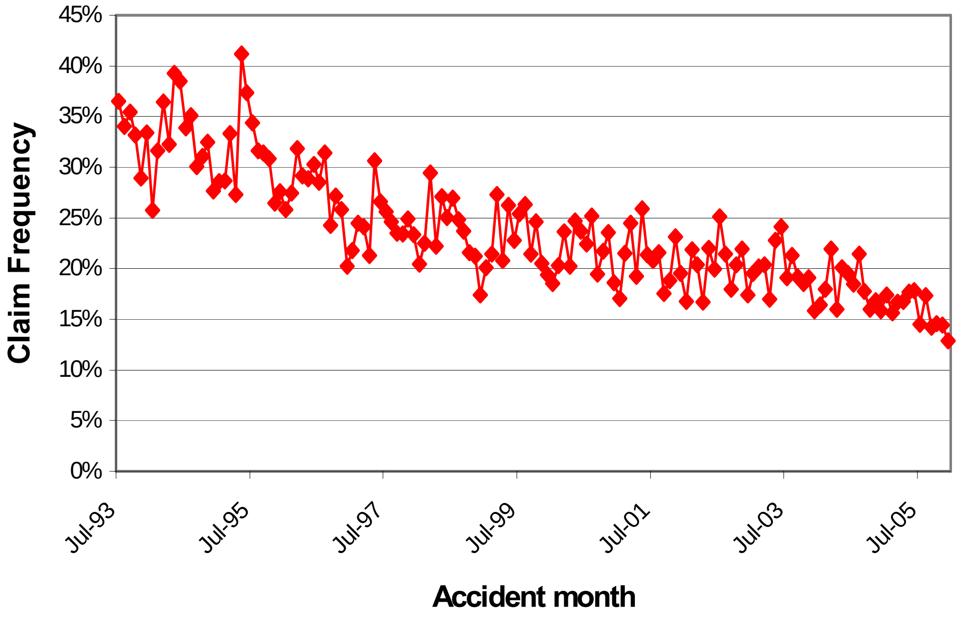
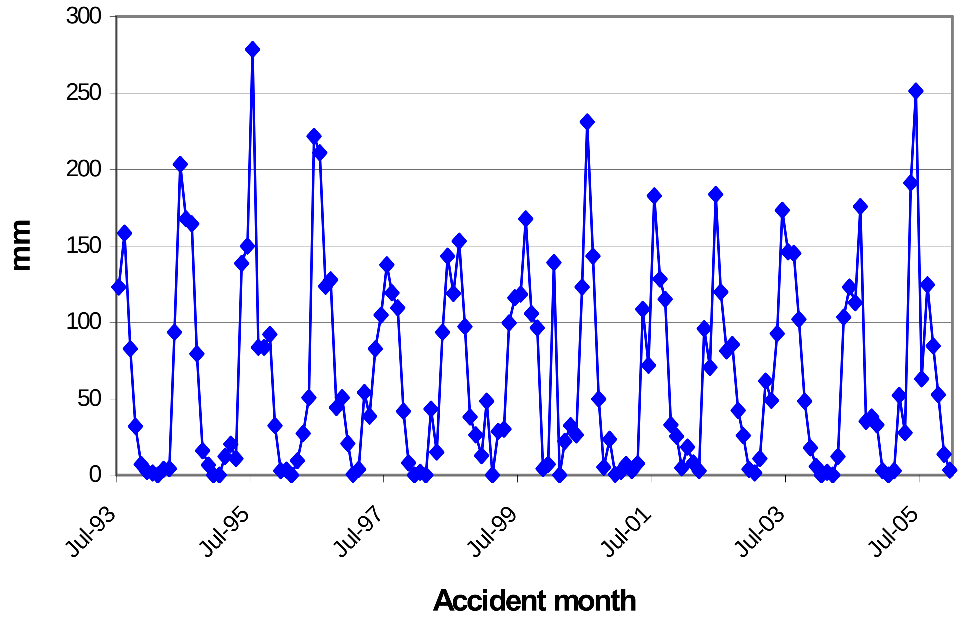


Figure 2 displays the monthly rainfall over this time period. The cyclic nature of rainfall is apparent here. Thus, rainfall looks unlikely to be a candidate for the overall fall in claim frequency, but may explain some of the fluctuations from month to month. Consequently, any modelling of rainfall should model the overall trend first, and then examine the residuals for any rainfall related effects.

Figure 2 – Monthly rainfall



5.1 Normal linear model results

5.1.1 Removing the overall downwards trend

The first stage is to factor out the overall trend in claim frequency. This is shown in Figure 3 below. Following this fit, the residuals may be extracted (shown in Figure 4) and can then be examined for dependencies on rainfall.

Figure 3 – Fitting the downward trend in frequency

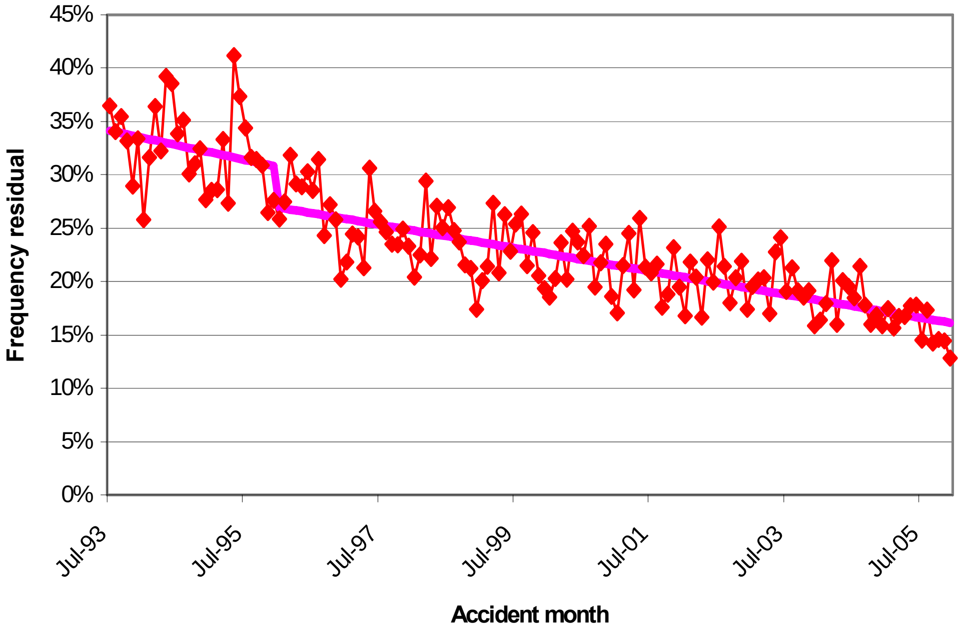
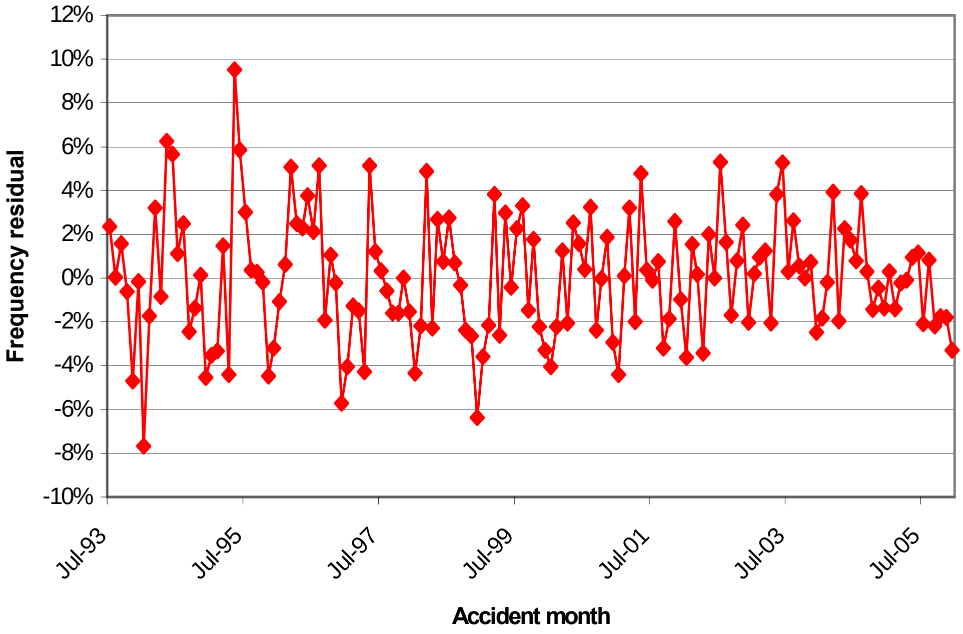


Figure 4 – Residuals after fitting of trend

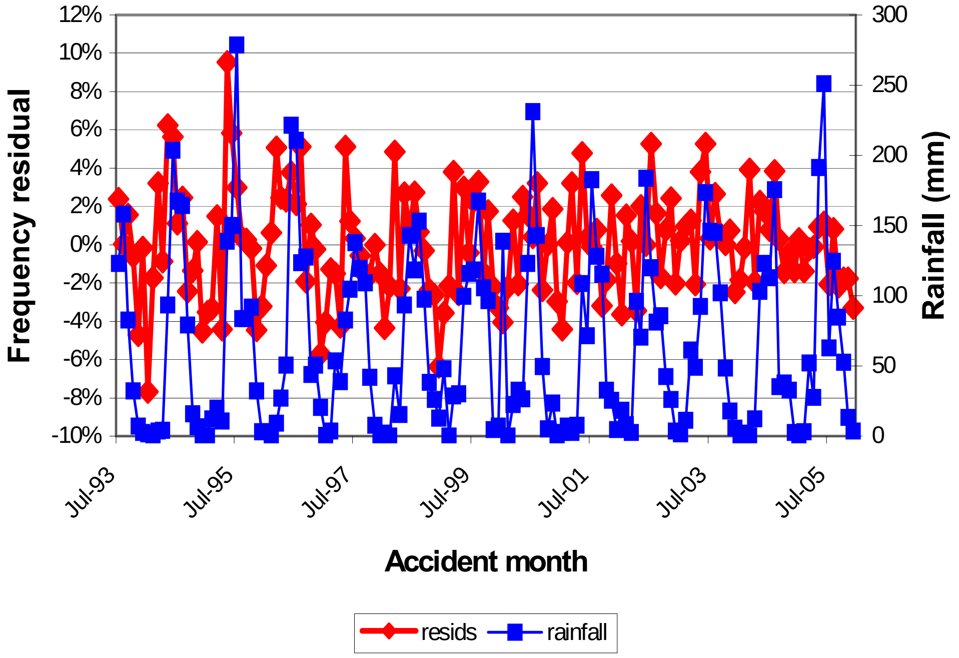


Of course it is possible (and preferable) to jointly model the overall trend and the rainfall. However, in the interests of ease of exploring the rainfall effect this has not been presented here. For this particular data set, there is little difference between the results of the two different analyses.

5.1.2 The effect of rainfall

The residuals (from Figure 4) are shown plotted on the same graph as monthly rainfall levels (Figure 5). To the naked eye, there do appear to be similarities between the two series in that both exhibit some degree of seasonality.

Figure 5 – Frequency residuals and rainfall

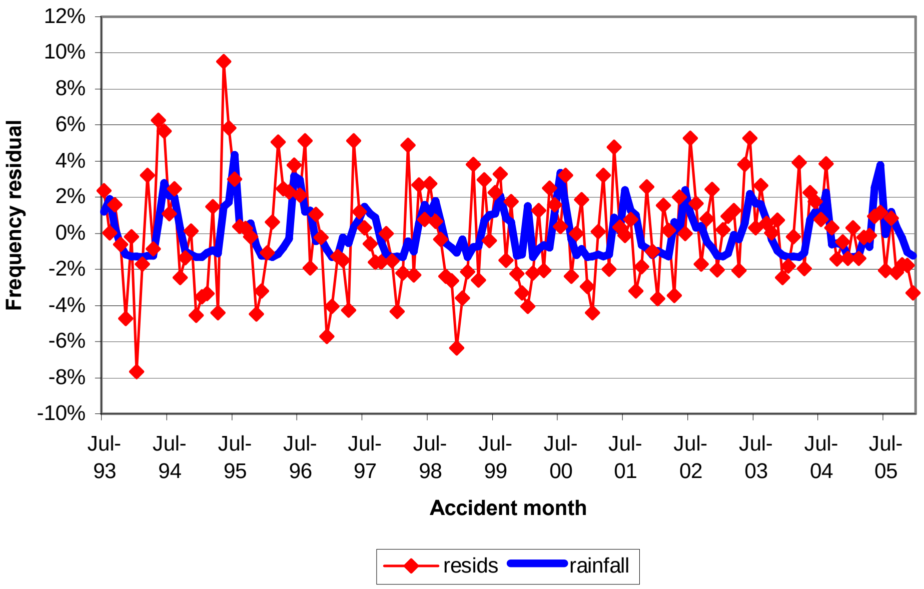


The frequency residuals have then been regressed against monthly rainfall using normal linear regression. The results are shown in Table 1 below. Looking at the t-test and F-test, it is seen that the rainfall parameter is very significant. The R^2 value suggests that there remains a high level of volatility in the series. The fitted values are shown in Figure 6.

Table 1 – Regression results for rainfall

Quantity	Estimate	Std error	t-value	Significant?
Intercept	-0.0132	0.0029	-4.5413	***
Rainfall	0.0002	0.0000	6.3744	***
R^2	22%			
df	148			
F-value	40.6330			***

Figure 6 – Fitted values from the regression on rainfall



However, the rainfall values and consequently the fitted values are strongly seasonal. Is it possible that the so-called rainfall effect is in fact a seasonal effect?

To test this, another seasonal covariate – average daylight hours per month – is added to the regression model. The results are shown in Table 2.

Table 2 – Results of regression on rainfall and daylight

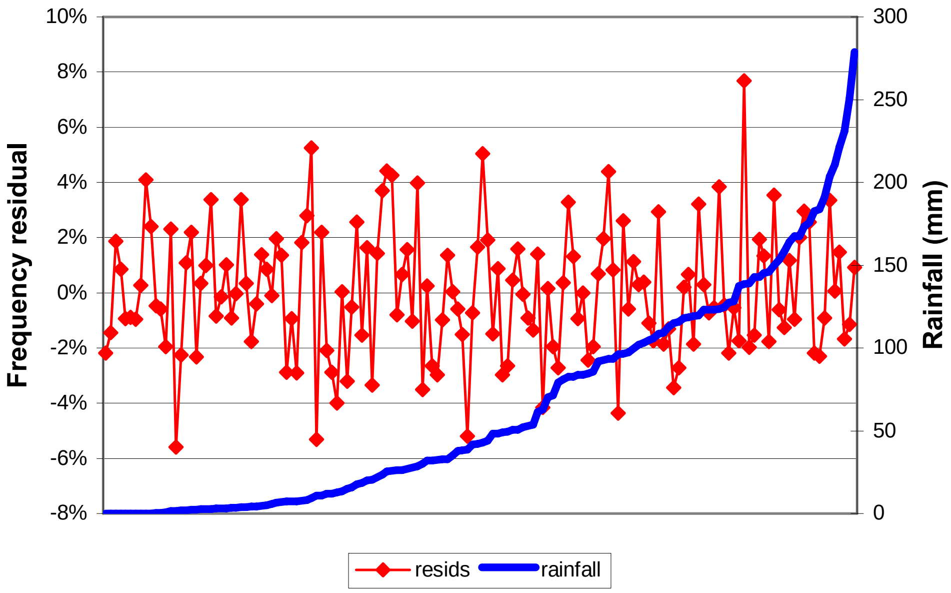
Quantity	Estimate	Std error	t-value	Significant?
Intercept	0.1213	0.0267	4.5345	***
Rainfall	3.E-05	5.E-05	0.6765	
Daylight	-0.0102	0.0020	-5.0554	***
R^2	33%			
df	147			
F-value	36.4664			***

From Table 2 it is seen that when daylight hours are added to the model, the parameter estimate for rainfall becomes insignificant. Further, the higher R^2 value indicates that this is a better model. Removing rainfall from the model does not lead to a deterioration in model fit. This suggests that the significance of rainfall was simply due to it acting as a proxy for a seasonal effect.

The relationship with daylight hours is an inverse one: fewer claims in the summer months when there is more daylight and conversely more claims in winter. It is possible that the extra claims are caused by more night-time driving in the darker months. It is also possible that people drive less in summer as more people take holidays then.

So is there any relationship between rainfall and claim frequency? Figure 7 is a plot of the frequency residuals after fitting daylight hours, ordered by rainfall. No relationship may be seen in this graph.

Figure 7 – Frequency residuals and rainfall, ordered by rainfall



Therefore, the end of the line has been reached and the case is closed – rainfall does not affect claim frequency in Perth. Or does it?

5.2 Poisson GLM Daily model

At this stage let's pause to reconsider the data. The models in Section 5.1 examine the effect of monthly rainfall on monthly claim numbers. No effect was found (other than, arguably, daylight hours serves as a general proxy for seasonality, and some times of the year are wetter than others which may impact claim numbers at that time of year).

However, it rarely rains for a month in Perth. Thus, any month is a mixture of wet days and dry days. Further, small amounts of rain may have little effect, particularly under the Australian sun where roads dry off extremely quickly. This, together with the noisy nature of claim numbers, means that any rainfall effect gets smeared out over the month and consequently becomes hard to detect.

This is exactly the problem tackled by Eisenberg (2004). His monthly analysis led to an even more extreme result than that above in that he found that higher levels of rainfall corresponded to fewer accidents.

In place of a monthly model, Eisenberg (2004) suggests modelling daily claim frequencies. This approach is superior to monthly modelling due to the shorter time scales used: rainfall and accidents on the same day are much more likely to be related than rainfall and accidents within the same month.

Therefore, an over-dispersed Poisson ("ODP") GLM was fitted to daily claim data. Covariates included accident month (to factor out the overall decreasing trend), month (to capture annual seasonality – for example claim frequencies tend to be lower in holiday months like December and January – so that this effect is not confused with rainfall), day of the week and rainfall.

Two forms of rainfall were considered:

- Rainfall on the day of the accident – to test the primary rain effect of the more rain, the more accidents;
- A lagged rainfall effect (represented here as rainfall two days prior to the accident) – to test a secondary rain effect – does recent but past rain lead to fewer accidents, possibly due to cleaner roads or a greater awareness amongst drivers of the need to drive with care.

Eisenberg (2004) demonstrated that the latter effect does occur so it is of interest to include it here. Rainfall two days prior rather than one day prior was used to better ensure that the rainfall did genuinely occur before the claim. While an accident time is available for the claims, rainfall can only be said to have occurred some time in the 24 hours up to 9am. Therefore, rainfall on a previous day may, in some cases, fall very close to the time of the accident, meaning that any effect is likely to be the primary rain effect (i.e. wet conditions lead to more accidents).

The ODP model found significant seasonality effects, both by month and by weekday. Rainfall effects, both primary and secondary, were found.

Figure 8 – Primary rainfall effect (rainfall on day of accident)

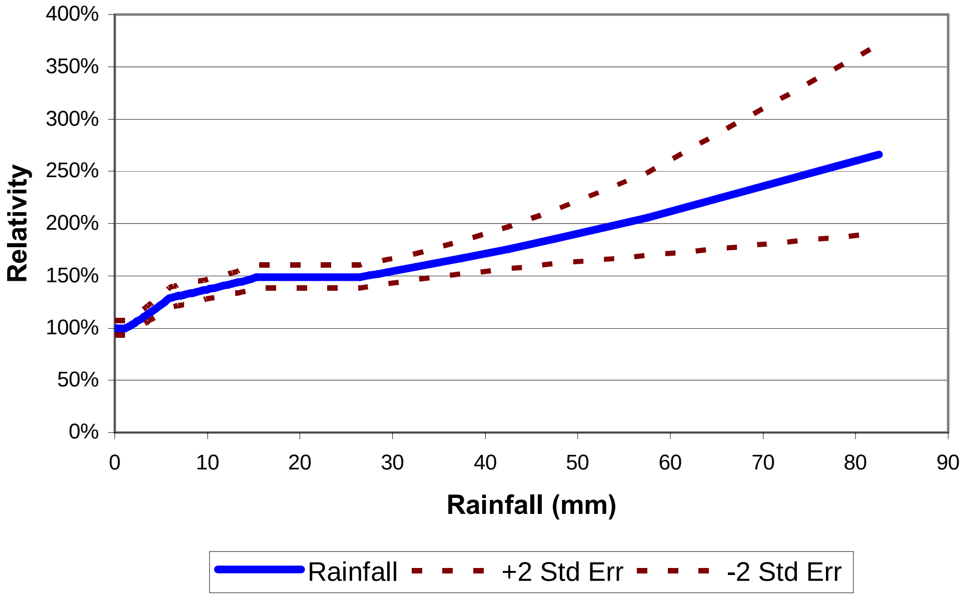
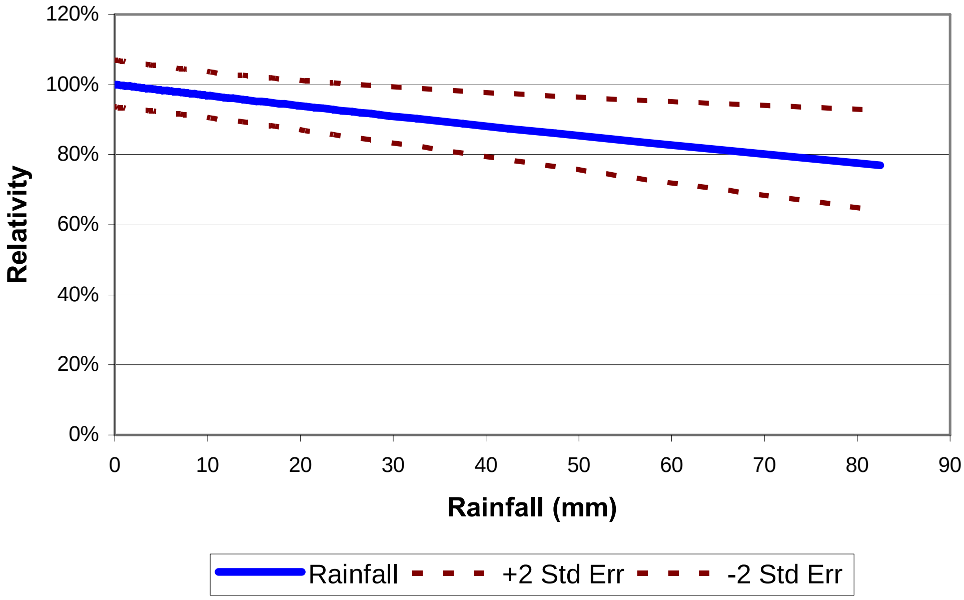


Figure 8 is a plot of the primary rainfall effect, i.e. rainfall on the day of accident, together with error bands showing plus and minus two standard errors. The effect is shown as a relativity where the relativity is 100% for zero rainfall. It shows a clear effect of rainfall on claim numbers with rainfall relativities consistently greater than 100% and monotonically increasing over the range analysed– the higher the rainfall, the more claims are estimated to occur.

The error bands widen as rainfall level increases due to fewer days of heavy rainfall. However, even in the extremes of the daily rainfall, the lower bound of the confidence interval for the rainfall effect exceeds 100%., indicating that, all else being equal, the presence of rain on a particular day is likely to lead to more accidents.

Figure 9 plots the estimated secondary effect – the impact of recent but past rain. This shows clearly that rain in the recent past on average leads to a reduction in claim numbers – though the effect is not as strong as the primary rain effect. Further the effect is less significant for lower levels of rainfall (the error bars straddle the 100% relativity [the level for zero rainfall]).

Figure 9 – Secondary rainfall effect (past rainfall)



Thus, the conclusion of this analysis is that rainfall is both good cop and bad cop:

- More CTP claims occurred when it rained. The higher the rainfall, the greater the effect tended to be;
- However, after periods of rain, particularly heavy rain, claim numbers reduced. This may be due to cleaner roads, as postulated by Eisenberg (2004), or due to the recent rainfall conditioning drivers to drive more cautiously.

6 Lies, damn lies and statistics

In this section, we return to the bad clichés and discuss the impacts of various types of statistical analysis on the results of the rainfall analysis. We start with the bad, then the ugly and conclude with the good.

6.1 The bad

Looking at Figure 1, it is tempting to propose an analysis to determine why the claim frequency is falling. After all, logic tells us that some of the following are candidates for the fall in frequency:

Vehicle factors

- Improvements in vehicle design and safety;
- Reductions in the average age of the Perth vehicle fleet;
- Greater multi-vehicle ownership where only one vehicle is driven at any one point in time.

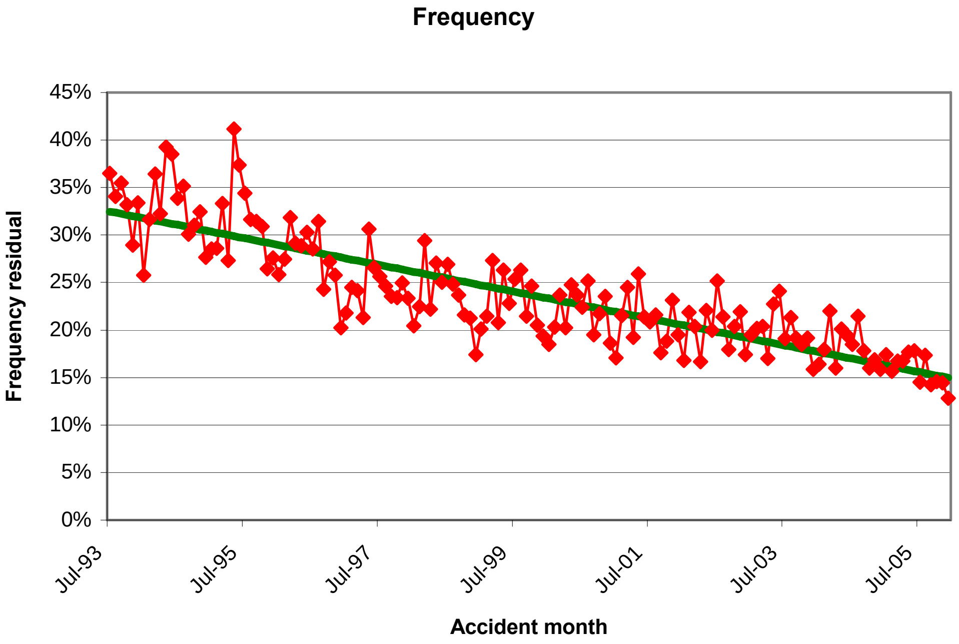
Environmental factors

- Road safety awareness campaigns and advertising;
- Cost of petrol;
- Unemployment rates;
- Lower rainfall levels;
- Improved roads and safety infrastructure (barriers, divided highways etc).

However an appropriate statistical model is difficult to define. Firstly, what exactly do we model for improvements in vehicle design and safety? Perhaps we could source the average age of the Perth vehicle fleet. Suppose that this does, indeed, show a reduction over the time period in question. We could use this as a covariate to model monthly claim frequency using a linear model or a GLM and would probably find a significant relationship. But does it mean anything?

The answer to this is “Not really”. Most monotonic series will be found to be a significant covariate in such an analysis. A facetious example is given in Figure 10 below where claim frequency is regressed on the author’s age, the result being the fitted green line shown in the graph. Normal linear regression was used and the effect of age was highly significant using a t-test. Obviously, however, while the two series may be correlated (-0.86 linear correlation for the series in Figure 10), there is not any causal relationship between them.

Figure 10 – Claim frequency regressed on author's age



While it is obviously unreasonable to expect a causal relationship between someone’s age and claim frequency, it is more reasonable to expect that things like the average age of the vehicle fleet, the unemployment rate or the rate of multi-vehicle ownership would impact the claim frequency. Proving this is another matter. Techniques such as time series analyses may be helpful here. Tests for co-integrated series may shed some light on this issue. However this is beyond the scope of this work and has not been considered further here.

The example in Figure 10 is an extreme example, but this problem of “correlation does not equal causation” is endemic to all regression analysis. Consider again the normal linear modelling of monthly claim frequency against rainfall and daylight hours. The results of this analysis suggested that rainfall was merely acting as a proxy for daylight hours. But is daylight hours acting as a proxy for something else? Something else that is really the driving force between claim frequency? For example, daylight hours are highest in December and January. This is also school holiday time – fewer cars on the roads – fewer accidents. Daylight hours are lowest in June and July – winter time – often more rain. So is the daylight hours series acting as a proxy for the combined effects of school holidays in December and January and more rain in winter time? Or other influences that we haven’t considered here. The analysis does not allow us to answer this question, so is **bad**.

6.2 The ugly

Even if the monthly data analysis did not have the problems associated with the correlation vs causation issue, the analysis still leaves a lot to be desired.

Before continuing the discussion it is important to note in practice the normal linear model would be fitted like the ODP GLM – i.e. in a multivariate way, with all covariates (including those that capture the trend) fitted at the same time. This type of modelling was not presented here for ease of exposition. In any case, for this particular example it makes little difference.

The response in such a model is claim frequency – a counting variable – yet it is modelled assuming a normal distribution – a continuous distribution. Thus there is a lack of congruence here. Further, claim frequency cannot be negative, but there is no condition in this model forcing the fitted values to yield a positive number.

This could be dealt with using a log normal model – where the data are first transformed by taking logs, before modelling using a normal linear model. However, such an approach means that a bias correction is required, something which can be difficult particularly when the true distribution is not exactly log normal. Further, there is still the discordance between integral counting quantities and the continuous strictly positive log normal distribution.

Another ugly aspect of the monthly modelling is that logically it does not make a lot of sense. Why would a claim on 1 January be affected by rainfall on 31 January? Why would a claim on 1 February not be affected by the rainfall on 31 January? Granted the monthly analysis does look at overall claim numbers and looks to see whether months of higher rainfall correspond to higher claims. However this suffers from two problems:

- Daily claim number data is noisy. Rainfall may only lead to one extra claim, so the signal may get lost in the noise;

- In any case the analysis of rainfall in Section 5.2 suggests that rainfall has two opposing effects. While the secondary effect is of lower magnitude than the primary effect, it may persist for longer and offset any extra claims due to the rainfall.

Therefore, any monthly analysis is **ugly** – it will suffer from loss of information.

6.3 The good

It can be argued that any rainfall analysis needs to be carried out at a much finer level than monthly. Davies et al (2004) tackled this problem by matched analysis – for example, they consider claims on a wet Tuesday of one month and compare them with claims from a dry Tuesday of the same month. In this way, all other things, apart from rainfall, are more or less equal. This allowed them to find general relationships between claim numbers and rainfall.

Here, a model of claim frequency is used to put numbers on the rainfall effect. Data were on a daily basis. An ODP GLM was used, with a log link. This means that the distribution is appropriate and that the resulting estimates will never be less than zero.

This analysis enabled the detection of both the primary and secondary rainfall effects, as discussed in Section 5.2. It is therefore **good**.

However, few things in this life are perfect and the analysis in Section 5.2 is no exception. Modelling daily claim numbers and daily rainfall is much better than a monthly analysis but the same problem of mismatching between the occurrence of the rain and the claim persists. For example a claim that happens at 9.01am is not affected by rain at 11am yet the rain would be ascribed to the same day as the accident in the data used here. At 8.59am the next day, particularly in warm sunny weather, the secondary rain effect may be in play (cleaner roads following rain) yet a claim at that time would be assumed to fall in the same day as the 11am rain and be subject to the primary rainfall effect.

The obvious solution to this problem is to model at an even more granular level – perhaps hourly. However, while claim data are available at this level, rainfall data are not.

The multivariate ODP GLM is still subject to the "correlation vs causation" problem. Therefore it is possible that there is some effect to which rainfall is correlated, but which is actually the cause of the changes in claim frequency ascribed to rainfall. It is impossible to completely eliminate this problem, but the risk of it occurring may be minimised.

For example, the initial monthly analysis (and common sense) suggests that rainfall is correlated with time of year. Therefore the ODP model contained accident month seasonality to capture things like school holidays etc. Daylight hours were also considered in the model, but were not found to be particularly significant. Capturing the accident month seasonality separately means that we can be more confident that the rainfall effect is really a rainfall effect.

7 Conclusion

This paper sets out to do two things: firstly to examine the effects of rainfall on motor vehicle bodily injury claim frequency, using data on CTP claims in the Perth metro area. Using GLM techniques, it was found that rainfall actually has a mixed effect – an immediate increase in claim frequency during or shortly after rainfall, but a subsequent decrease in claim frequency, perhaps due to cleaner roads following the rain, or more cautious driving. Both effects are greater for medium to high levels of rain.

A secondary topic is that it is important to get the statistical technique correct. Otherwise the analysis may return the wrong answer. To quote George Box (or whoever did actually say this – there is some controversy over this) – “All models are wrong. Some are useful”. The object of any statistical analysis should be to use a model that is not so badly wrong that the results are unusable.

Acknowledgements

I would like to thank the Insurance Commission of Western Australia for permission to use their data. Also, thanks are due to Bo Jing who assisted with the modelling work.

References

Davies, R., Winn, R. and Jiang, J. (2004). Determinants of claim frequency in CTP schemes. Accident Compensation Seminar, 2004, Institute of Actuaries of Australia.

Eisenberg, D. (2004). The mixed effects of precipitation on traffic crashes. Accident Analysis and Prevention, 36, 637-647.