A practical introduction to extreme value modelling

Harry southworth

Data Clarity Consulting Ltd

harry@dataclarityconsulting.co.uk

2016-02-04

Statistical models

- Most statistical models are concerned with estimating (or predicting) *expected* values
  - Linear regression, analysis of variance
  - Generalized linear models, survival analysis
- Sometimes the expected value is of no interest and it is the *unexpected*, extreme values that matter
  - What is the oldest age a human can achieve?
  - What is the highest windspeed a tall building must be designed to withstand?
  - How many concurrent calls must a telecommunications network be designed to handle?
  - How far can the value of assets fall in a year?
  - Extreme values of certain chemicals in the blood indicate liver damage. Are they associated with a medicine being used? Are they related to the dose of that medicine?

Drug safety: ALT

- According to published guidance (CTC):

	Grade	**Severity**
$ULN \leq ALT < 2.5 \times ULN$	1	Mild
$2.5 \times ULN \leq ALT < 5 \times ULN$	2	Moderate
$5 \times ULN \leq ALT < 20 \times ULN$	3	Severe
$ALT > 20 \times ULN$	4	Life threatening

- ULN can be thought of as the units of measurement for ALT
- In general, it appears to make more sense to worry about changes from baseline than absolute values, and a 5-fold increase is considered actionable

ALT is short for *alanine aminotransferase*

Example: troglitazone

Troglitazone (for treatment of diabetes). FDA review states

> *Mean [ALT] levels fell in patients receiving troglitazone in phase 3 trials... It was also stated that 2.2% of patients in phase 3 trials had an [ALT] level exceeding* $3 \times$ *ULN... What was not appreciated by [FDA] was that many of the patients classified as ALT* $> 3 \times$ *ULN actually had ALT values that were VERY much higher than* $3 \times$ *ULN... 23 patients had treatment-emergent ALT values over* $3 \times$ *ULN... In 14 of these 23 patients, the ALT value exceeded* $8 \times$ *ULN... and in 5/23 patients the ALT value exceeded* $30 \times$ *ULN.*

The drug was withdrawn from the market after reports of liver failure and death

And another application...

Extreme rainfall events and floods

Photograph: Christopher Thomond for the Guardian

(http://www.theguardian.com/uk-news/2015/dec/28/uk-floods-2015-york-suffers-phone-internet-outages-cash-machines)

Photograph: Fliker alh1 (https://www.flickr.com/photos/allan_harris/23675911959)

The A591 north of Grasmere, 2016-01-02 (photo by me).

The A591 north of Grasmere, 2016-01-02 (photo by me).

Some quotes

Calderdale council leader Tim Swift said: "It's just obvious that the scale of flooding events over the last 10 years has been dramatically greater than anything we've had before..."

http://www.theguardian.com/environment/2015/dec/27/floods-army-called-continue-devastate-northern-england

The Guardian, 2015-12-27

Some quotes

The Met Office say that Honister Pass has set a UK rain record for any 24-hour period... Provisional data has a new 48 hour rain record at Thirlmere...

In conclusion, the latest devastating floods in Cumbria are consistent with what we would expect from a warming world – we are seeing impacts now, and the risks of more frequent and severe rainfall are increasing every year...

https://www.foe.co.uk/sites/default/files/downloads/floods-climate-foe-briefing-december-2015-94324.pdf

Friends of the Earth, 2015-12-08

Questions

Is it *obvious* that the scale of flooding events over the last 10 years has been *dramatically* greater than anything we've seen before?

Is it the case that the risks of more frequent and severe rainfall are increasing every year?

- How unusual was the rainfall of December 2015?
- How high can monthly rainfall get?
- Are large rainfall events getting worse?
- Are large rainfall events getting more common?

Data

Weather station data available from the UK Met Office:
http://www.metoffice.gov.uk/public/weather/climate-historic

The right-hand panel there tells us we have

- Mean maximum and minimum temperature (tmax, tmin)
- Days of air frost (af)
- Total rainfall (rain)
- Total sunshine duration (sun)

Blue appears to indicate a station that has closed.

Click a few to find out when they opened.

Click "Historical station data" in the pop-ups to see the data

What now?

Having motivated the use of extreme value modelling, the rest of the session proceeds as follows:

- Brief and superficial history of statistical extremes
- Short introduction to the generalized extreme value distribution
- Short introduction to the generalized Pareto distribution
- Some aspects of model fitting
- Model checking
- Approaches to inference
- Worked examples
- Closing comments for the first session

... after which, the practical session.

For a much more detailed treatment,

http://www.cces.ethz.ch/projects/hazri/EXTREMES/talks/colesDavisonDavosJan08.pdf

A brief and superficial history of statistical extremes

Extreme value modelling is a mature branch of statistics:

- First full-length textbook by Gumbel (1958)
  - "A first draft of the manuscript was written in 1949..."
- Gumbel traces recognition of the problem back to Nicolaus Bernoulli, 1709
- Key publication by R. A. Fisher and L. H. C. Tippett, 1928, resulting in a flurry of research that continues to this day
- Gumbel's applications were largely hydrological

The Generalized Extreme Value distribution

Key to the *theory* of extreme values is the *Generalized Extreme Value distribution* (GEV)

Put $M_n = max(Y_1, Y_2, \cdots, Y_n)$ for IID observations $Y_1, \cdots, Y_n$

Then, $M_n \rightarrow$ upper endpoint of the distribution of Y

Normalizing, $\frac{M_n - a_n}{b_n} \rightarrow GEV(\xi)$ as $n \rightarrow \infty$ for suitable a_n, b_n

Compare with the well-known Central Limit Theorem:
$\frac{\bar{Y} - \mu}{\sigma} \rightarrow N(0,1)$ as $n \rightarrow \infty$

As $N()$ is useful for modelling *means*, $GEV()$ is useful for modelling *maxima*

Comments on GEV

For an outline proof of the asymptotic argument that maxima are *GEV*() see Coles (2001), Section 3.1.4

GEV gives us a starting point, but:

- Modelling maxima implies we take lots of samples, and throw away *all* observations other than the maxima
  - With the weather station data we could treat each station as providing a sample, take the maximum observed rainfall per station and throw away the rest: for Newton Rigg, throw away 658 values and keep just 1
  - ... or we could use the maximum value observed each year: throw away 11 $12^{th}s$ of the data and assume asymptotic behaviour has already kicked in at a sample size of 12

That's a hugely inefficient use of data. However,

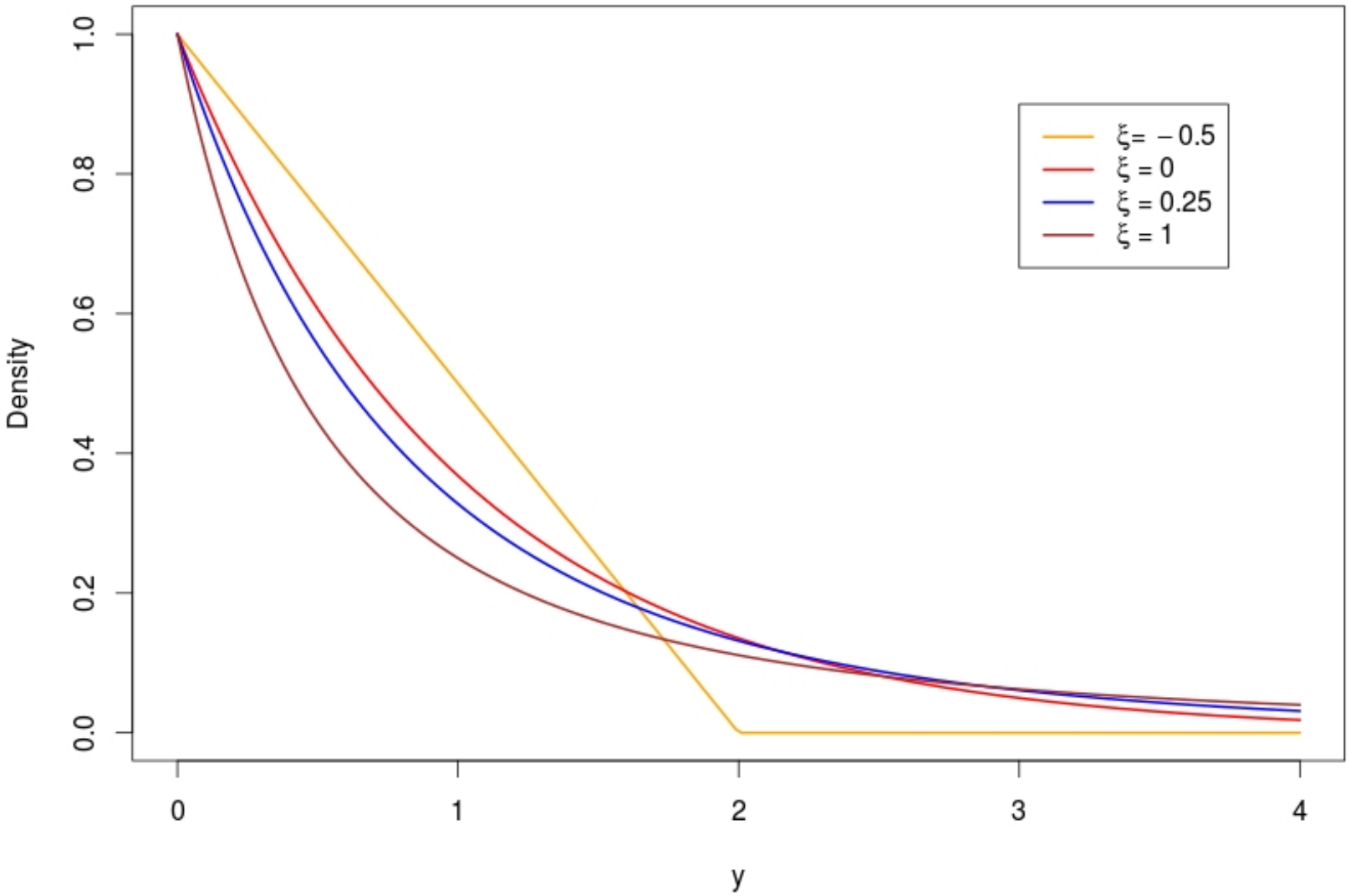
- The existence of GEV implies the existence of the *Generalized Pareto distribution* (GP) which makes more efficient use of data and should be preferred in practice

The Generalized Pareto distribution

If the assumptions underpinning the existence of GEV are true, then there exists a threshold, u, exceedances of which follow a GP distribution:

$$\begin{aligned} Y|Y > u &\sim GP(\sigma, \xi) \\ P(Y \leq y|Y > u) &= 1 - \left(1 + \frac{(Y-u)\xi}{\sigma}\right)^{-1/\xi} \end{aligned}$$

For an outline of the proof, see Coles (2001) Section 4.2.2


Density
1.0
0.8
0.6
0.4
0.2
0.0
0
1
2
3
4
y
ξ= −0.5
ξ = 0
ξ = 0.25
ξ = 1

Comments on the GP distribution

Again, the mathematical details justifying use of the distribution are left thin and focus is on practical aspects

Threshold u needs somehow to be selected on a data-driven basis

There are two parameters, σ and ξ: the *scale* and *shape* parameters respectively. One or both of these can depend on covariates: e.g. $\sigma = f(X\beta)$ for covariates X and β to be estimated from the data

We know the form of the cumulative distribution function, so we can derive a log-likelihood: $\ell(\sigma, \xi) = -n \log \sigma - (1 + 1/\xi)\Sigma_{i=1}^{n} \log(1 + (y - u)\xi/\sigma)$

Armed with a likelihood function, we can apply the usual array of inferential techniques: maximum likelihood estimation (MLE), profile likelihood, Bayesian posterior simulation, penalized likelihood

The shape parameter

Note that

- $\xi = 0 \implies$ equivalent to the Exponential distribution
- $\xi < 0 \implies$ shorter-tailed distributions with upper limit $u - \sigma/\xi$
  - $\xi = -1 \implies$ equivalent to Uniform distribution
  - $\xi < -1/2 \implies$ regularity conditions fail
  - $\hat{\xi}$ becomes *superefficient* but experience suggests it is more likely that the model is a poor fit (in the applications I'm used to)
- $\xi > 0$ we get heavier-tailed distributions
  - $\xi \geq 1/2 \implies$ infinite variance
  - $\xi \geq 1 \implies$ infinite expectation

Comments on inference for the GP distribution

Asymptotic MLE results generally not reliable

- $\hat{\sigma}, \hat{\xi} \to N()$ slowly as $n \to \infty$
- $\implies$ constructing interval estimates with standard errors will result in poor coverage properties

Coles (2001) uses profile likelihood methods to construct appropriate asymmetric interval estimates. However

- Profile likelihood intervals tend to be a little too narrow
- The problem gets worse as the number of parameters increases
- The computational burden increases as the number of parameters increases
- Bayesian computational approaches don't suffer these problems

Practical inference

In practice, σ and ξ have no straightforward physical interpretation. Interpretation is usually performed by estimating high quantiles of the fitted distribution. In the field, these are referred to as *return levels*

The M-observation return level is the level that we expect to be exceeded only once every M observations

$$y_M = u + \frac{\sigma}{\xi}[(M\theta_u)^{\xi} - 1]$$

where θ_u is the proportion of observations exceeding threshold u

As such, we can manipulate the simulated posterior densities of σ and ξ to obtain the *posterior predictive distribution* of the return level, and thus interval estimates

A general approach to modelling for the GP distribution

My currently preferred approach to inference is:

1. Select a threshold using standard techniques (we're coming onto that)
2. Fit models by MLE and using your favourite method for model selection (I use Akaike Information Criterion – AIC)
3. Examine diagnostic plots and go back to step 1 if necessary
4. Once happy with the model, use Markov chain Monte Carlo (MCMC) to simulate from the posterior distributions of σ and ξ
5. Use the posterior distributions of σ and ξ to make inferences about high quantiles of y

Threshold selection

Two standard tools to aid threshold selection

Mean residual life (MRL) plot

- Compute the mean of y above increasing thresholds u. Plot it
- Above a suitable threshold, the MRL plot should be linear (accounting for increasing uncertainty as the available sample size decreases)

Parameter stability plots

- Above a suitable threshold, $\hat{\xi}$ and an adjusted estimate of $\hat{\sigma}$, $\hat{\sigma}^*$, should be constant
- Plot them, with approximate confidence intervals, over increasing thresholds u

See Coles (2001), Sections 4.3.1 and 4.3.4 for further details

Example: clinical trial safety data

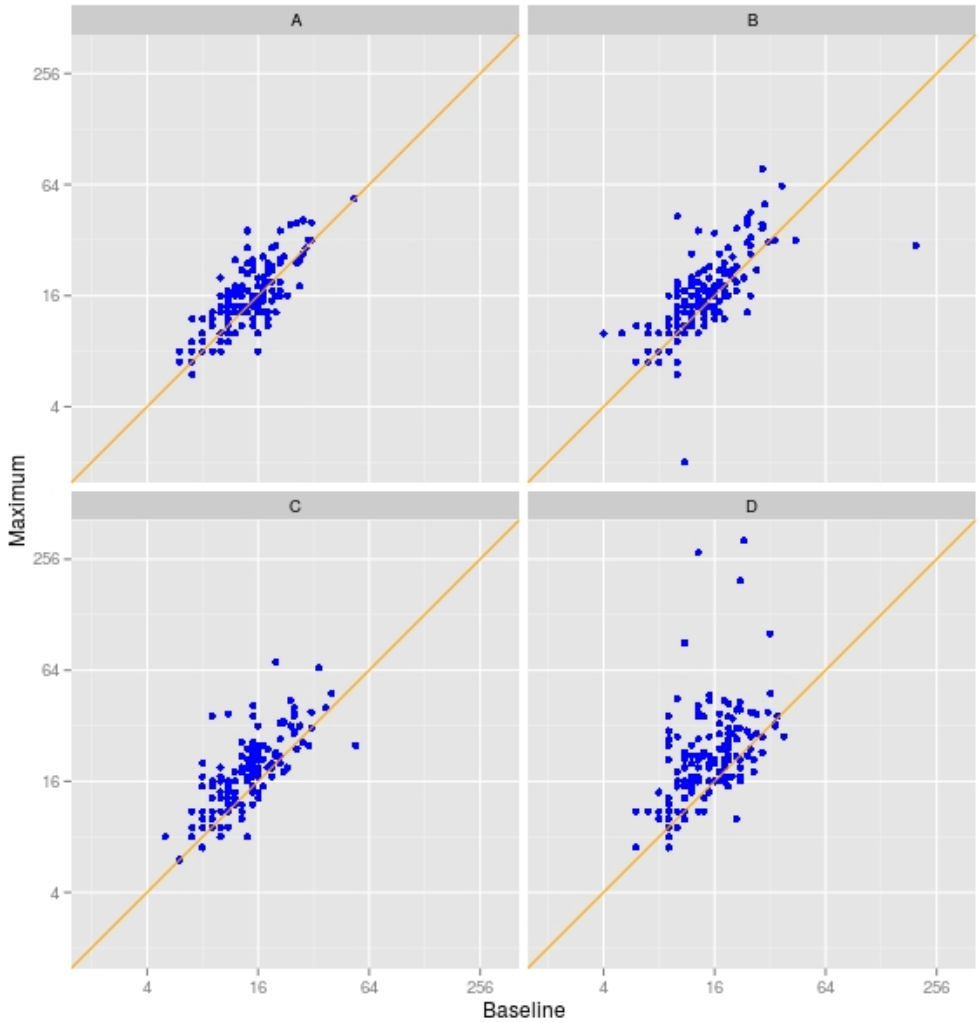
Background:

- ALT is a chemical in the blood that, at extreme levels, indicates potential liver injury
- A clinical trial was performed with 4 doses of an experimental drug, approximately 160 patients per group
- ALT was recorded at baseline (immediately before the patients started to take the drug) and after 6 weeks of daily treatment
- Doses are labelled A (lowest) to D (highest)

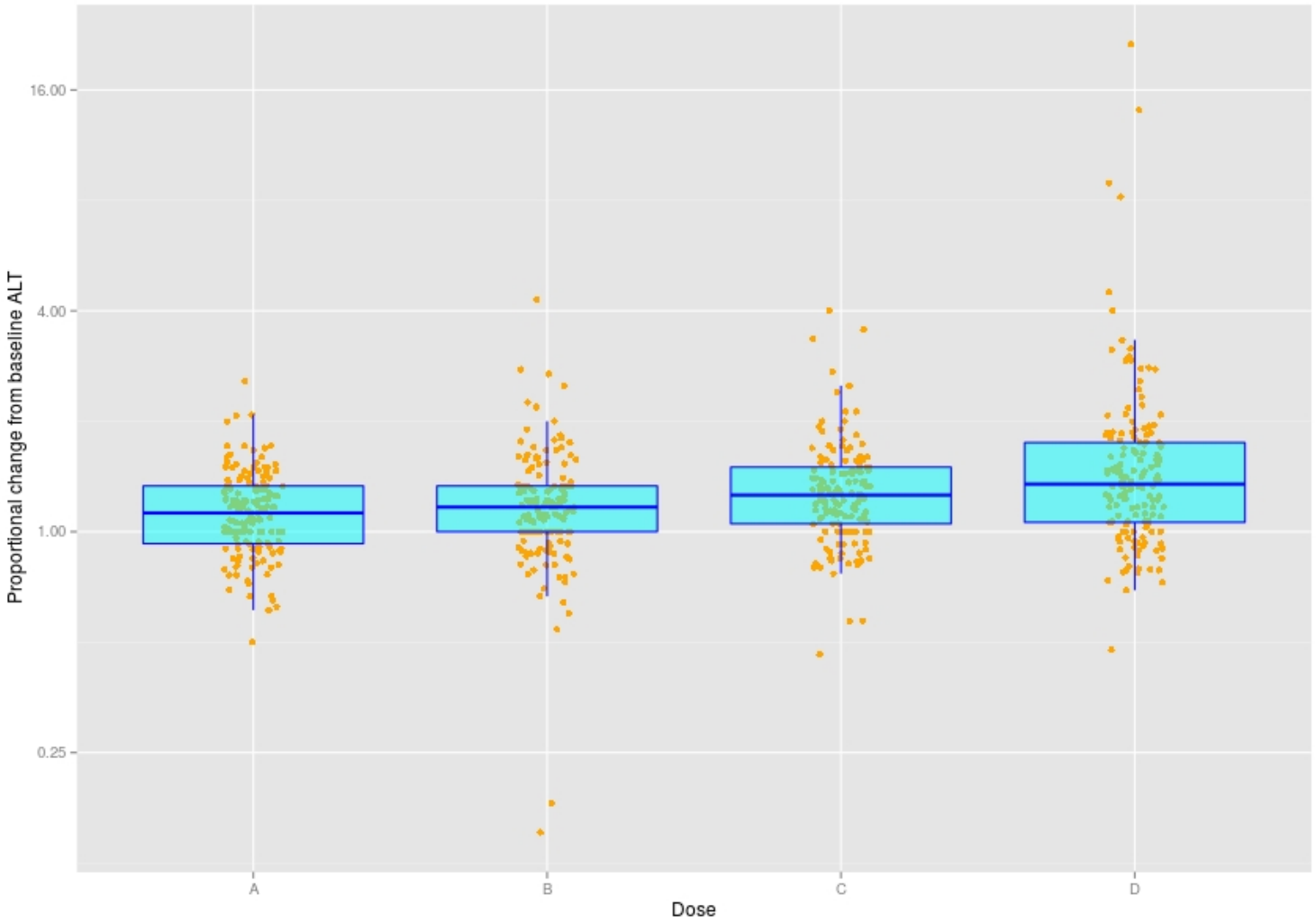
Because baseline values vary by patient, we work with change from baseline

- Specifically, proportional change from baseline
- A 5-fold increase from baseline would be an actionable value
- We take logs because the data are easier to view that way and the models appear to fit better

Shiftplots of the liver data



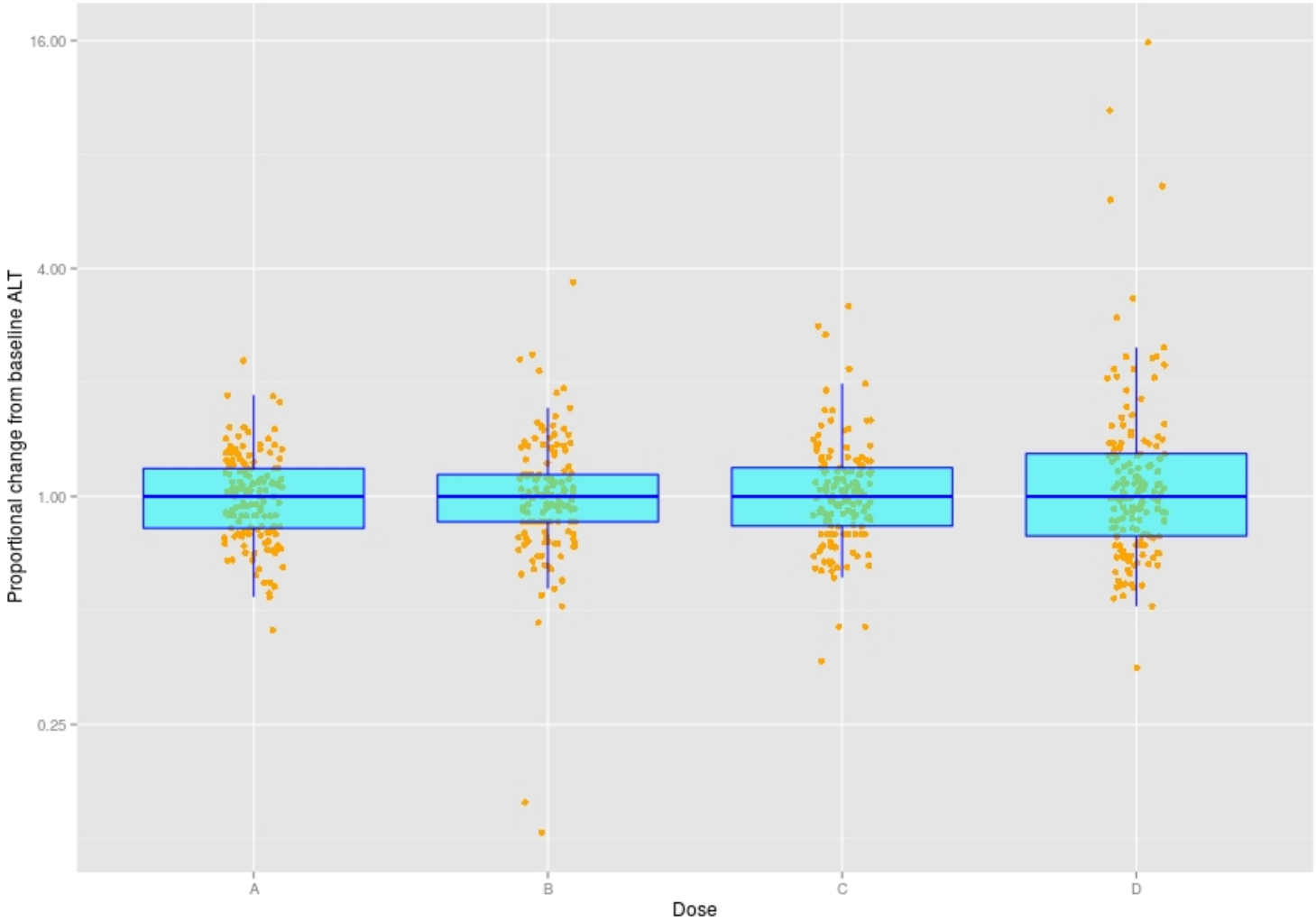
Boxplots of the liver data



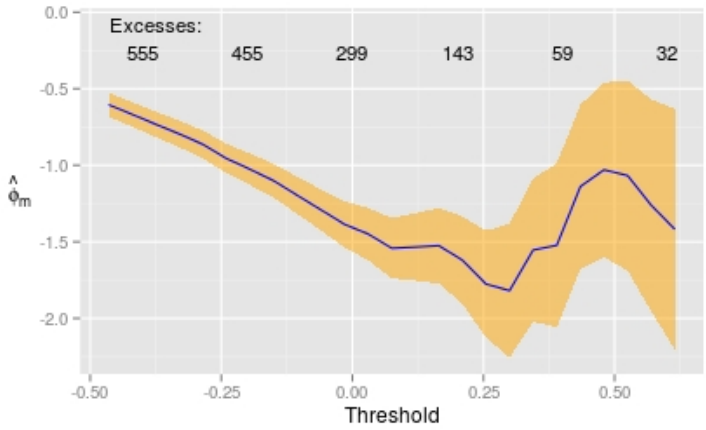
The boxplots of proportional change from baseline indicate a trend in dose

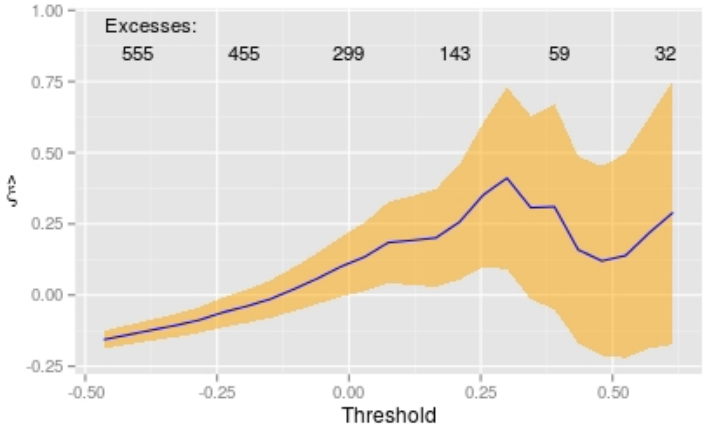
- Suggests the threshold in the GP distribution should increase by dose
- Divide by the median in each dose group
- We will have to reapply the medians later in the analysis to estimate return levels

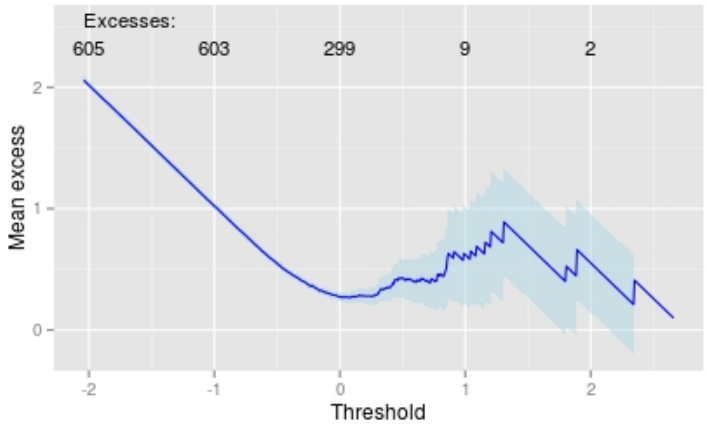
Boxplots of the liver data, median adjusted



Threshold selection plots for the liver data







Comments on threshold selection

A threshold of 0 might just about be ok

Note that the theory implies that a threshold exists above which a GP distribution ought to be a good approximation, *not* that the threshold be *high*

Whatever threshold is chosen, it is essential to examine diagnostic plots from the fitted GP distribution. If there is lack of fit, a higher threshold should then be attempted. We use

- PP plot
- QQ plot
- 'Return level' plot (no covariates)
- Fitted distribution with histogram of observed data (no covariates)
- Plot of covariates against residuals

Experience suggests the QQ-plot is the most useful

Model selection for the liver data

Models fit with log-linear terms for dose in $\phi = \log \sigma$, ξ and both

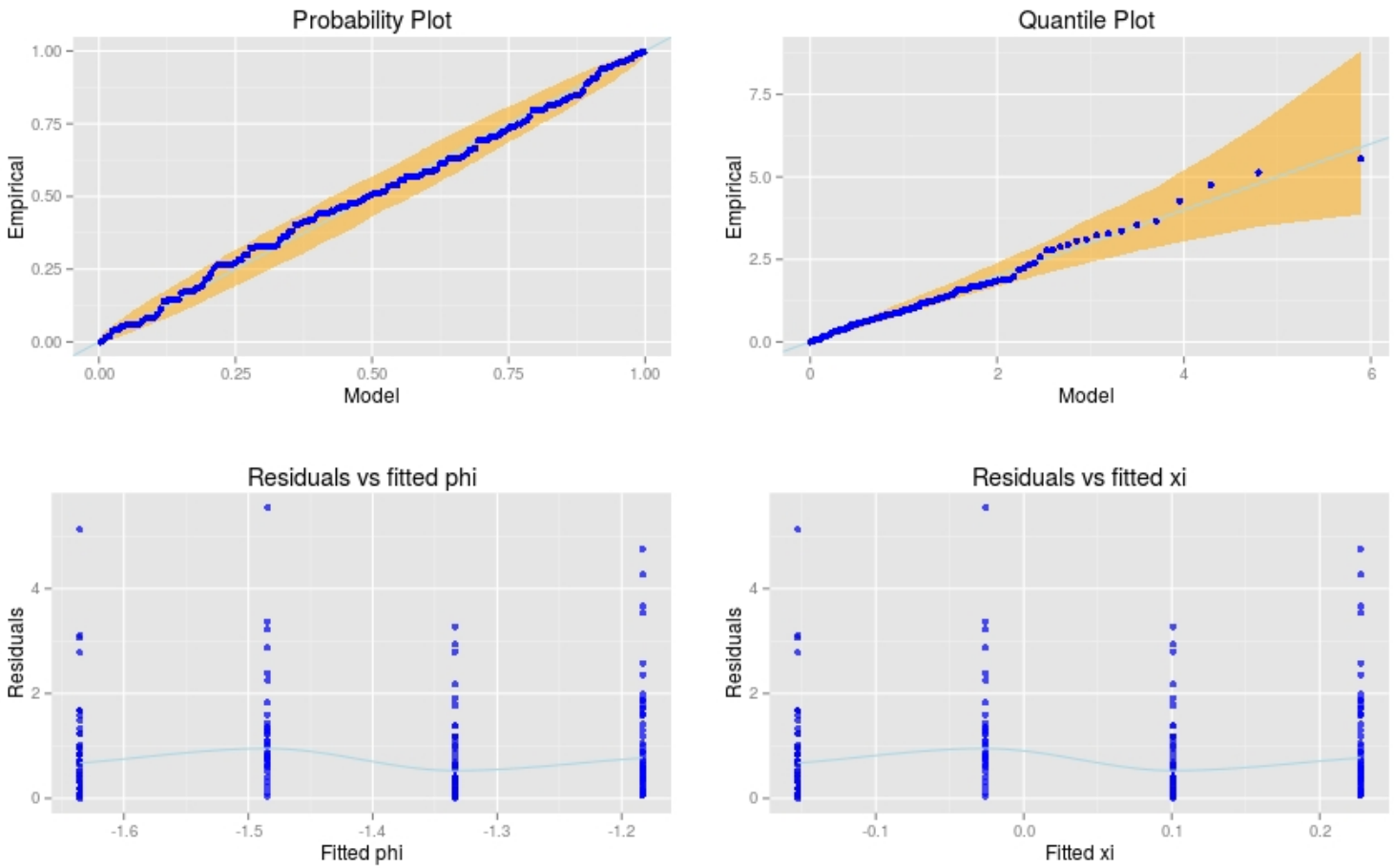
Turned out a threshold a little above 0 was ok

Model with term for dose in both was preferred by AIC, though there's very little in it between the model with a term for dose in just ξ and in practice we might fit a few models and see how the predicted return levels vary between them

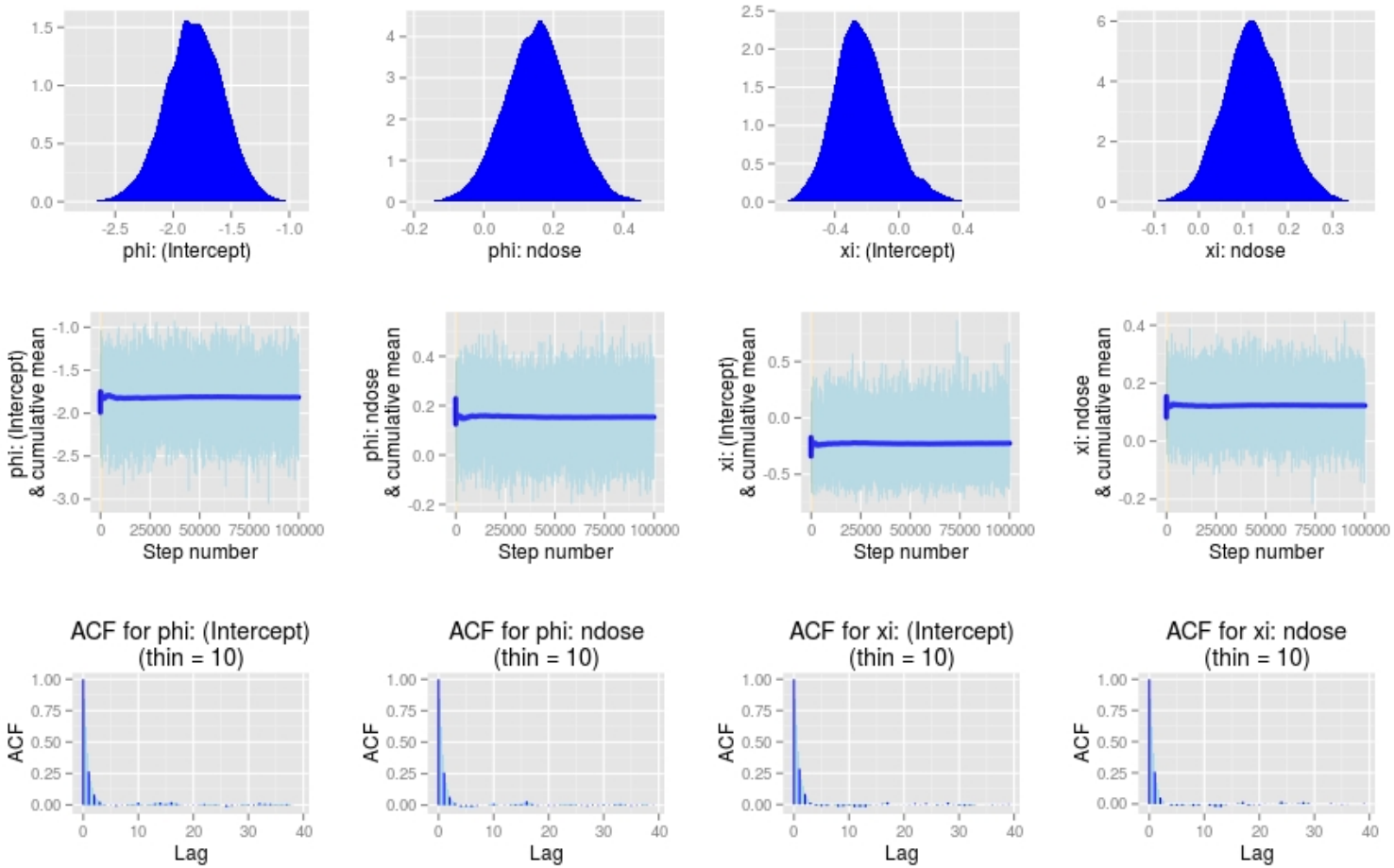
$\log \sigma = f(\log dose), \xi = f(\log dose)$

The null model (i.e. with no dose-response) was, though, clearly rejected

Diagnostic plots for the preferred liver model

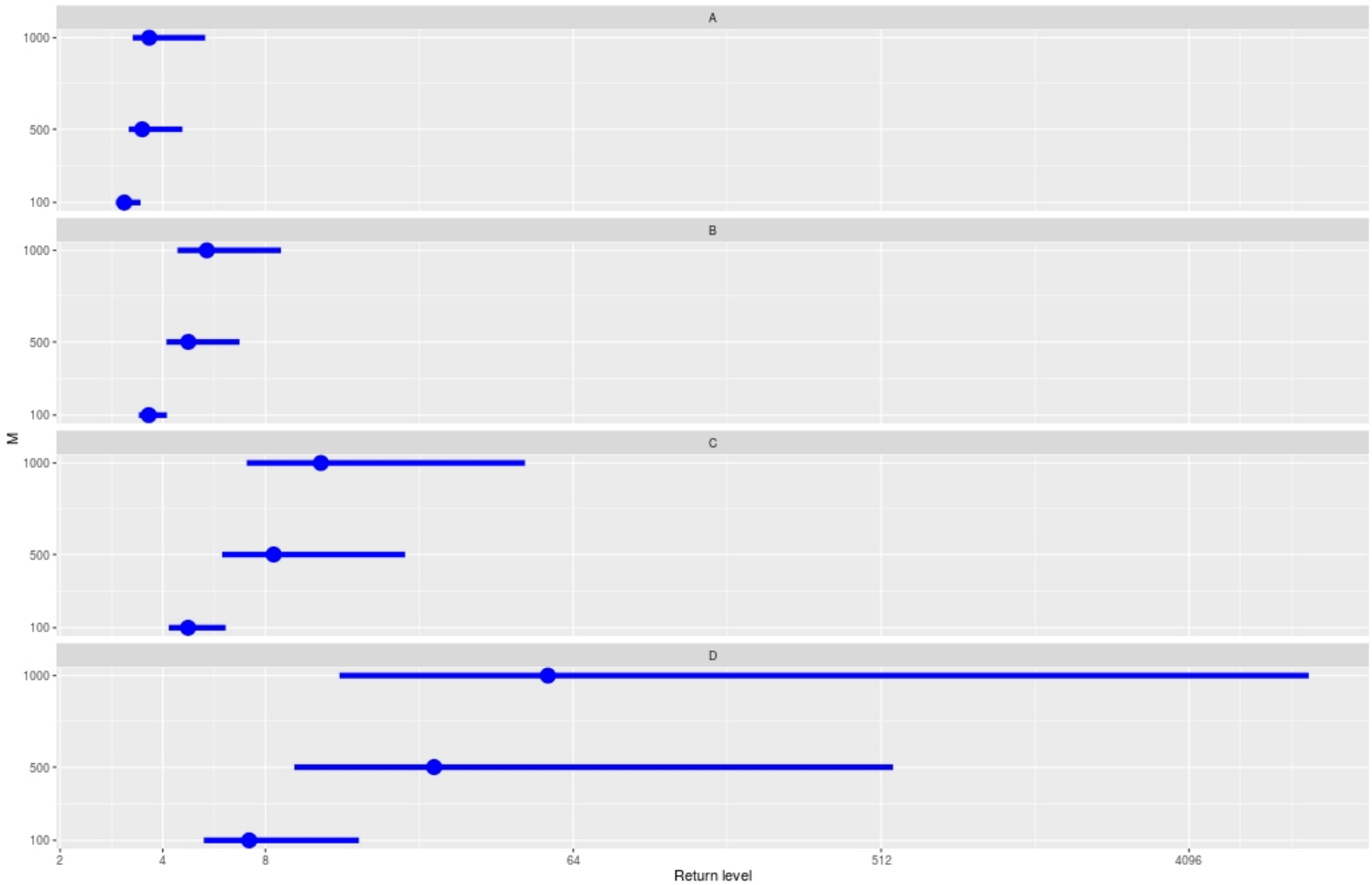


Diagnostic MCMC plots for the preferred liver model



A fat hairy caterpillar

Predicted return levels for the liver data



Comments on the liver data

The original analysis (Southworth & Heffernan, 2012) predicted values on the absolute scale (because that's how medics like to see the data) and was considerably more complicated

Various sources of uncertainty not accounted for here (e.g. uncertainty about the true medians). It would be dominated by the uncertainty due to extrapolation

Some of the predicted return levels might be physically impossible to observe because the patient's liver would have completely dissolved before ALT could get that high

Newton Rigg monthly rainfall data

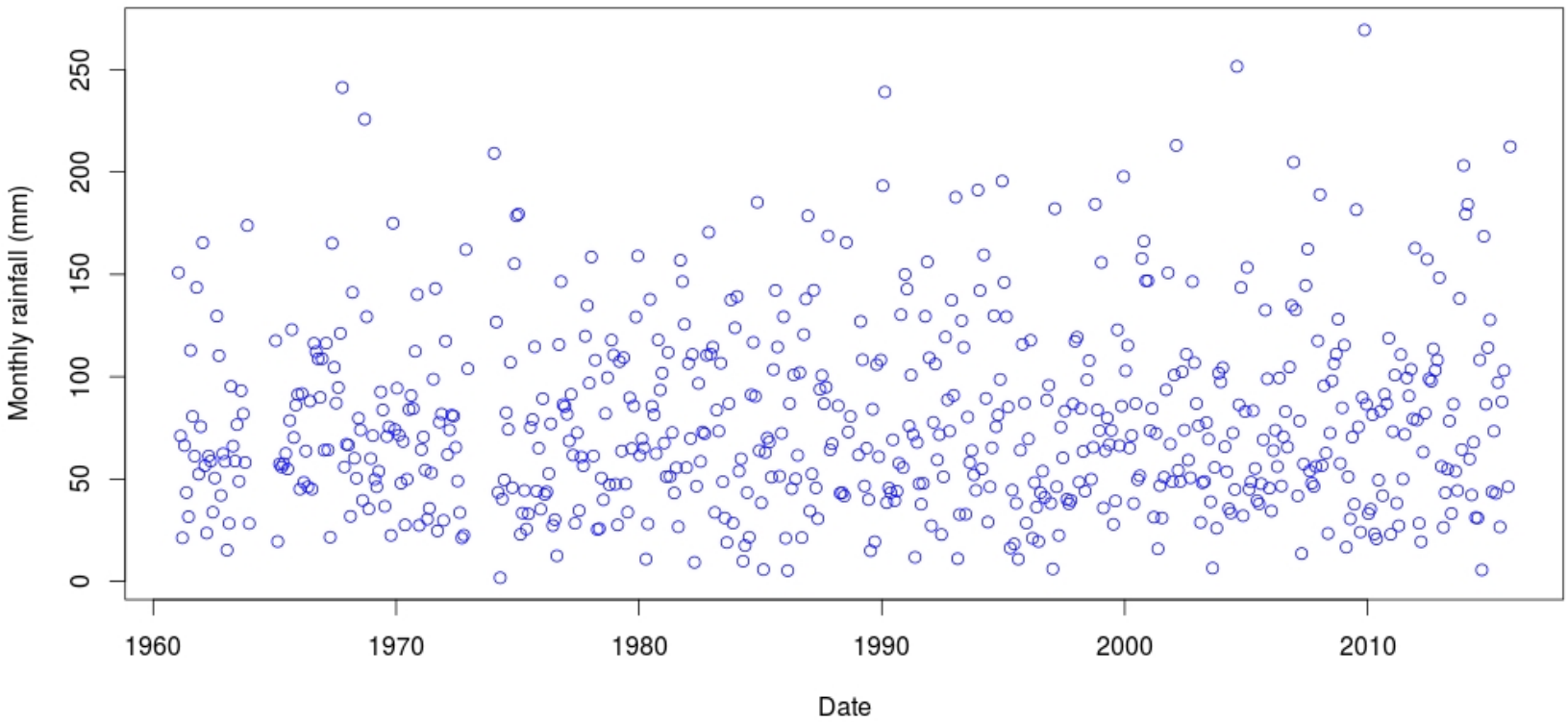
Total monthly rainfall data from the Newton Rigg weather station

Fit GP models and see how weird the recent rainfall events have been

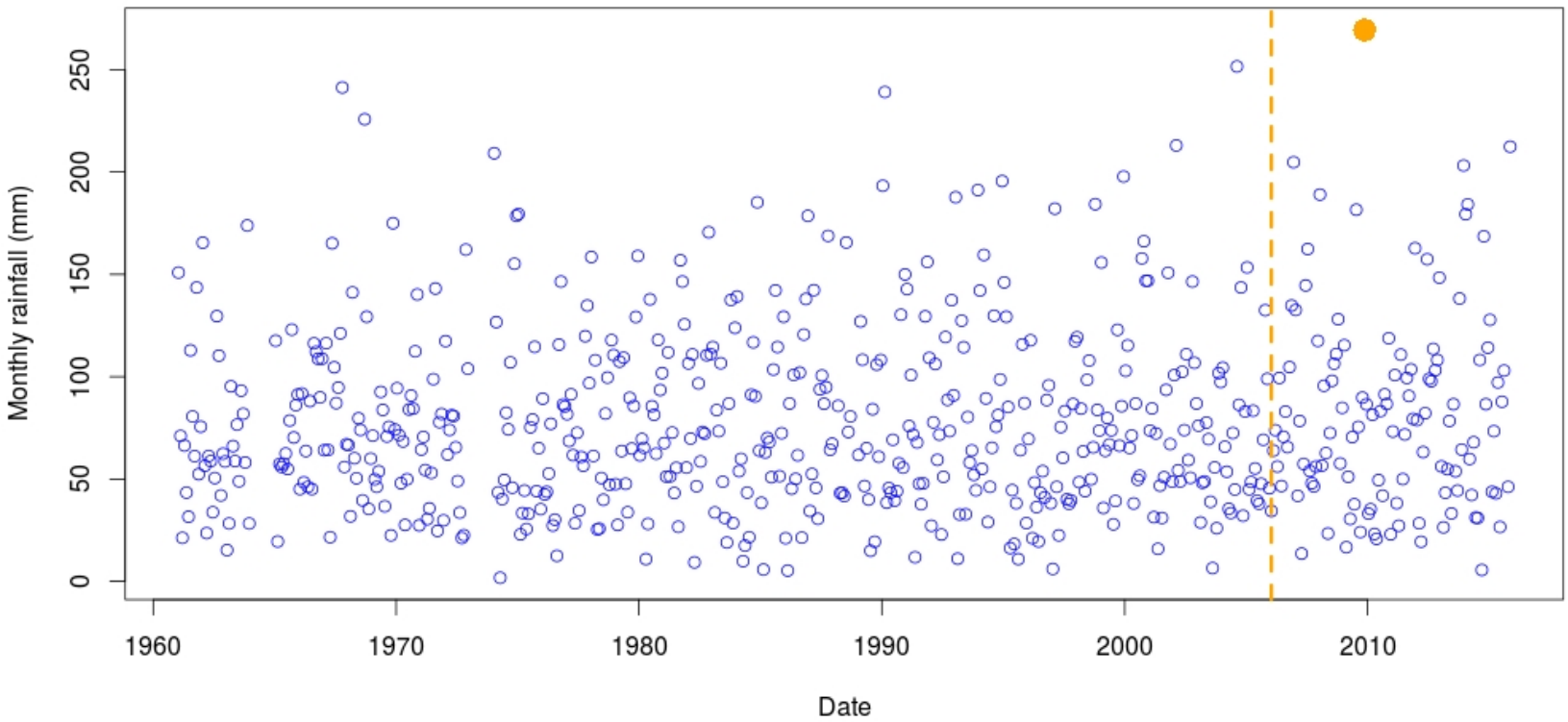
- obvious, dramatic, more frequent and severe...

First, though, take a look at the data...

Newton Rigg rainfall data



Newton Rigg rainfall data



The 2009 rainfall event

Offer starts in Weekend Terms and conditions apply

THE TIMES

SATURDAY

Max 15C, min 3C | November 21 2009 | timesonline.co.uk | Newspaper of the Year | No 69799 | £1.50

Once every 1,000 years rain falls like this

Hundreds of homes evacuated after floods

Steve Bird, Lindsay McIntosh

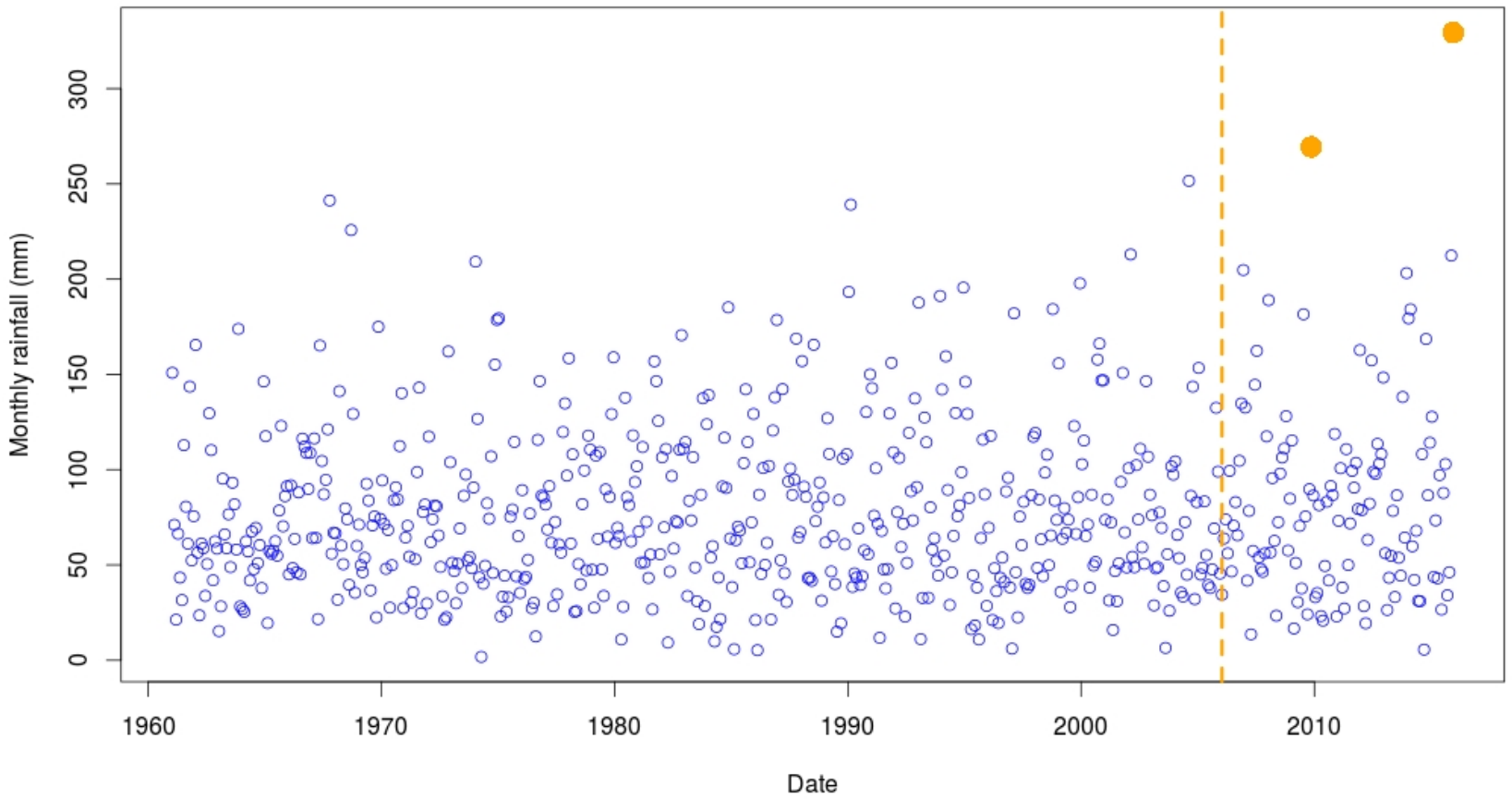
The full and devastating impact of England's worst recorded day of rain was still emerging last night as tributes were paid to a policeman swept away by floodwaters while trying to save others.

PC Bill Barker was helping motorists stranded on a bridge over the Derwent in the Cumbrian town of Workington when it collapsed. His body was discovered hours later on a nearby beach.

homes and businesses were evacuated, many of them ruined by floodwater and mud.

Jerry Graham, Cumbria's Assistant Chief Constable, said that PC Barker and a colleague had gone on to the bridge to help drivers who were trying to cross it. He said: "It was obvious they were going to put themselves in danger so PC Barker went to try and protect them. The bridge gave way just due to the volume of water and PC Barker went into the water and was swept away."

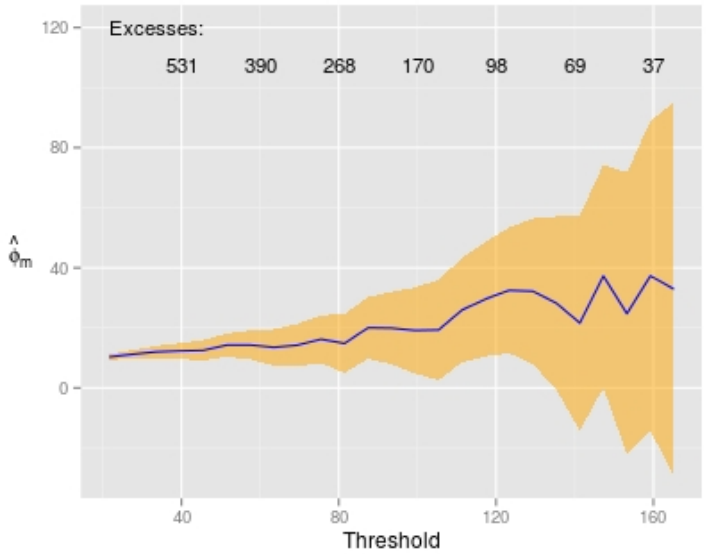
Newton Rigg rainfall data including December 2015

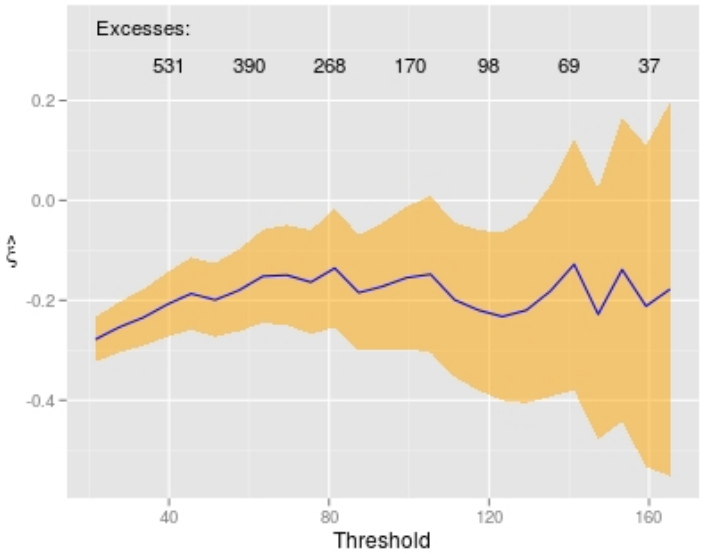


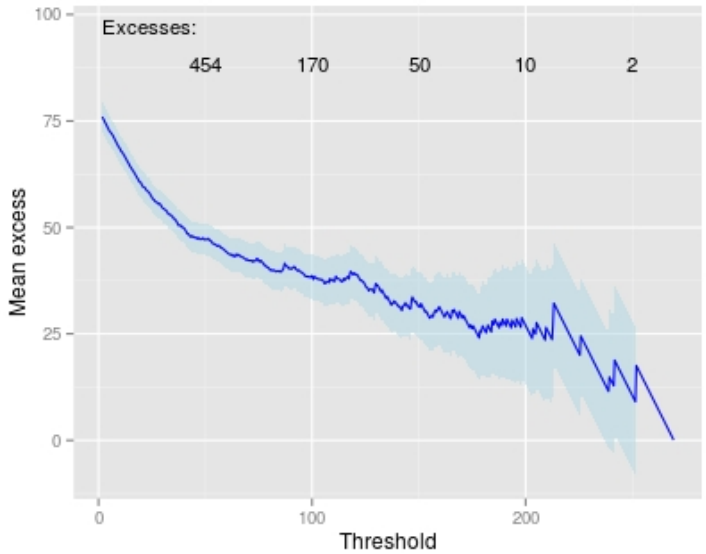
Extreme value modelling for Newton Rigg rainfall

Proceed by omitting the December 2015 observation and using the remaining data to see how unusual that observation is

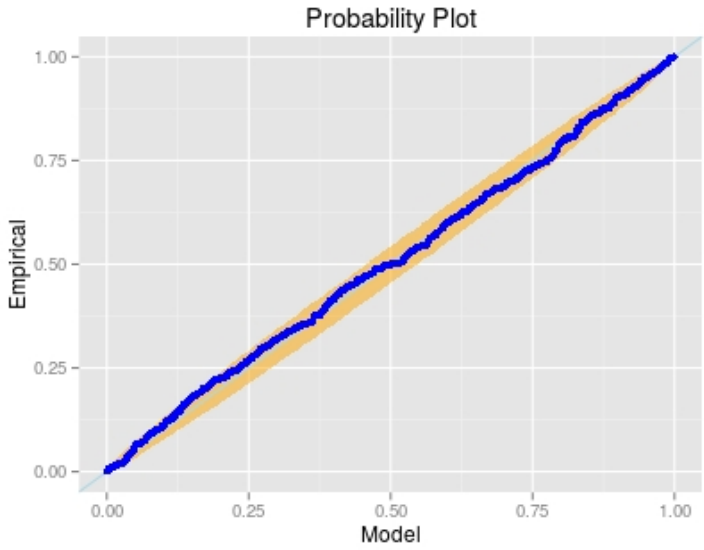
Threshold selection plots for the Newton Rigg rainfall data

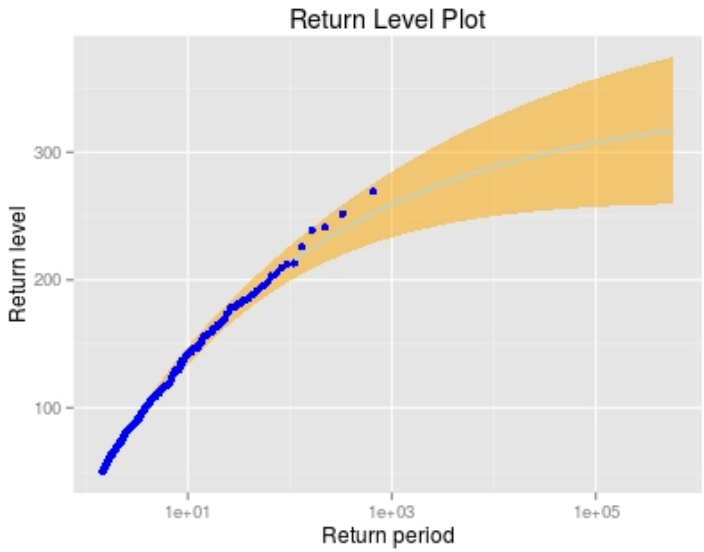


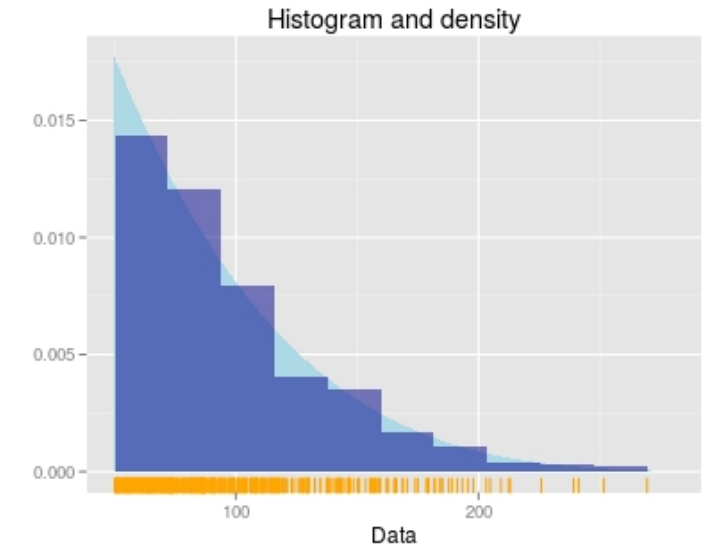




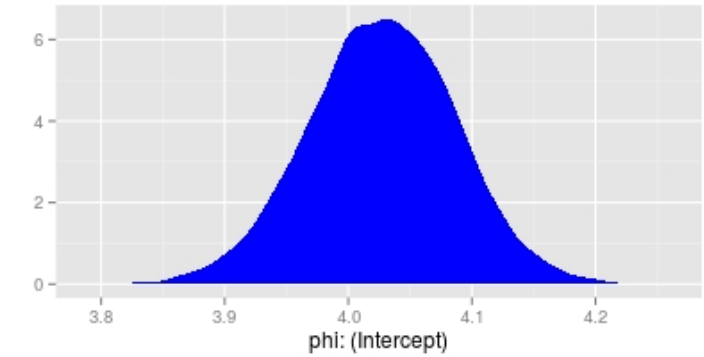
Diagnostic plots for the Newton Rigg rainfall data

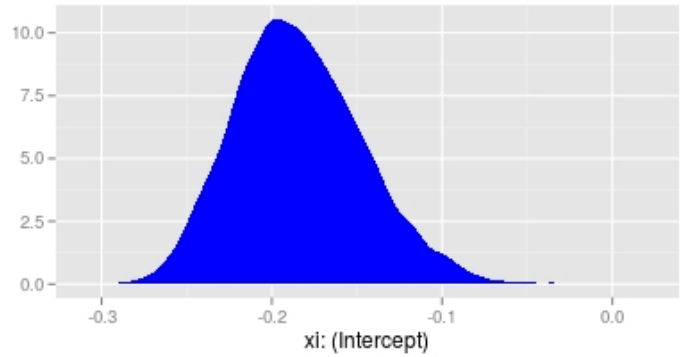


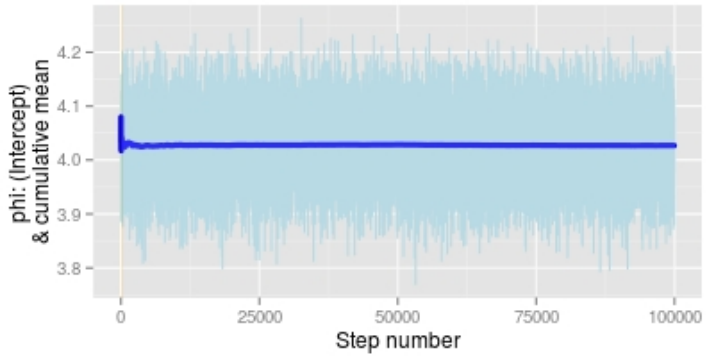


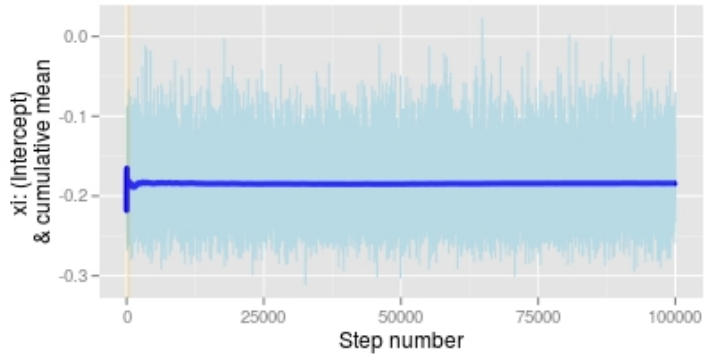


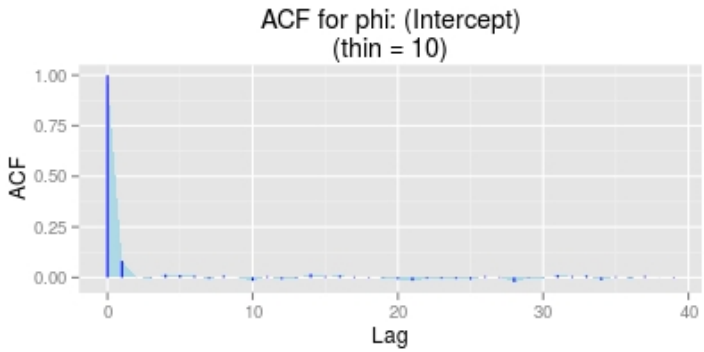
Diagnostic MCMC plots for the Newton Rigg rainfall data

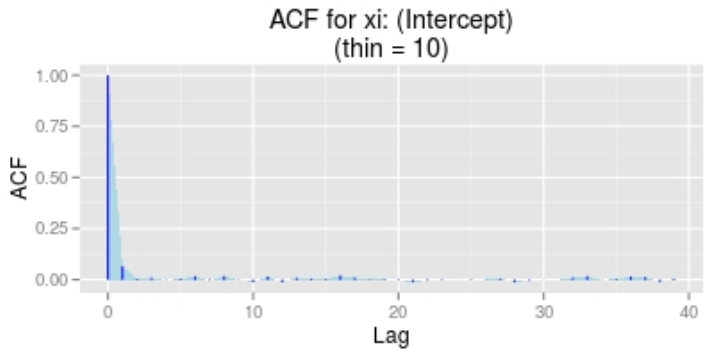




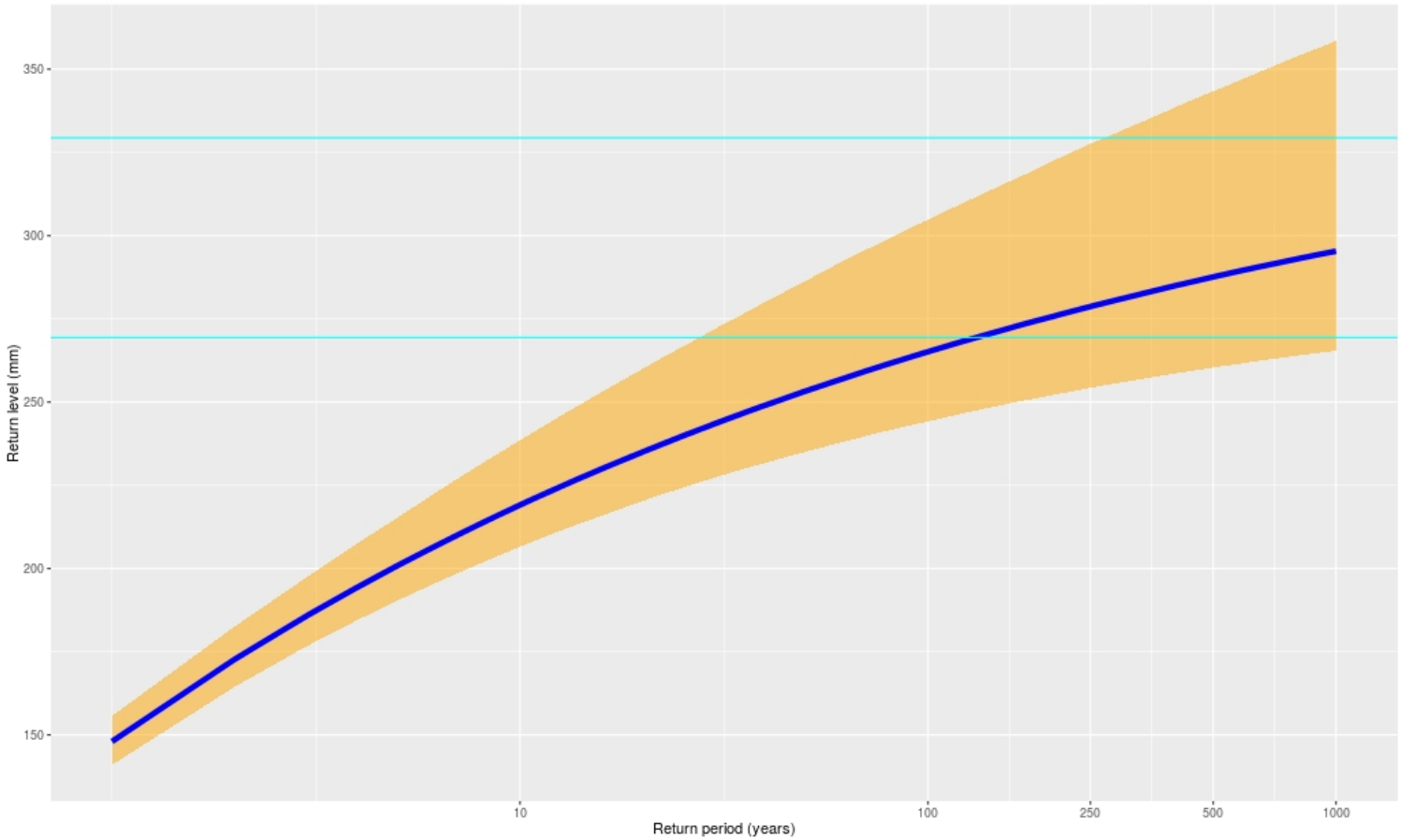








Return level plot for the Newton Rigg rainfall data



Once every 1000 years?

"Once every 1000 years" is possibly overstating the severity of the 2009 event

Best estimate is that the November 2009 event was a 1 in 150 year event

But

- If we exclude data from November 2009 onwards, we get a different estimate
- Accounting for uncertainty, the November 2009 event falls inside the 95% interval for it being a 1 in 30(ish) year event
- The analysis is very simple and assumes stationarity, ignoring seasonal effects
- Newton Rigg station is not the location where the most rain fell

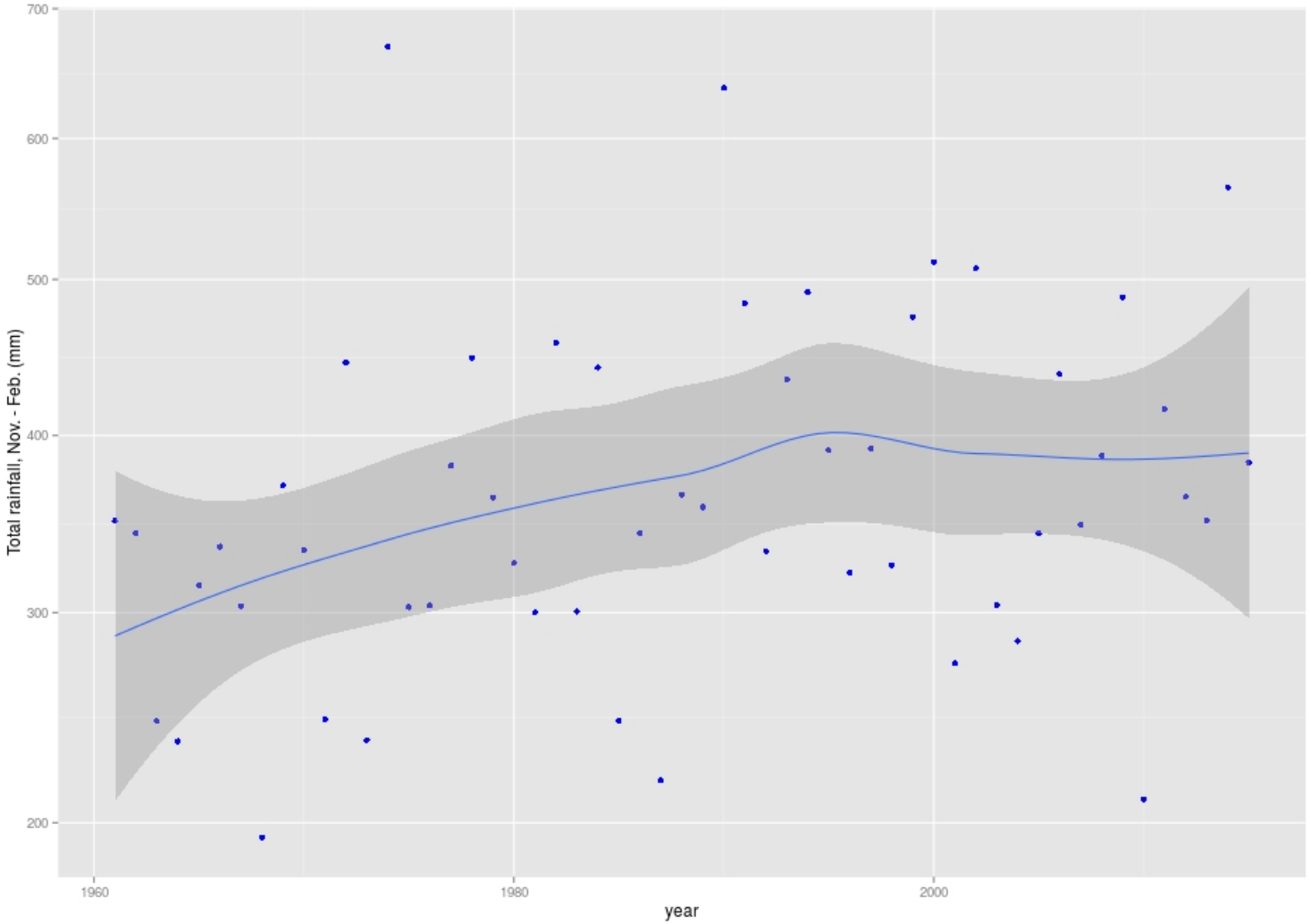
Comments on Newton Rigg rainfall data

Eyeballing the data reveals the December 2015 event to be 'dramatic', but it isn't clear that there is a trend or that there was a shift 10 years ago

But:

- We have included all data. Do the summer months tell us anything about rainfall in the autumn and winter?
- Looking at total rainfall from November – February suggests an increase over time that stabilized about 15 – 20 years ago
- (Choice of November – February is rather data-driven and...
- slightly more formal inference reveals the strength of evidence is not at all compelling)
- This is *data dredging* and we need to stop

Total winter rainfall



Further comments on the floods

> *As expected the undercroft in my home here at Bishopthorpe Palace is flooded again - we are fortunate however that back in the 13th century they built with flooding in mind, such that when the water subsides it soon washes through the original flood drains made for the purpose*
>
> *(http://www.archbishopofyork.org/articles.php/3385/statement-from-the-archbishop-of-york)*

Statement from the Archbishop of York, 2015-12-27

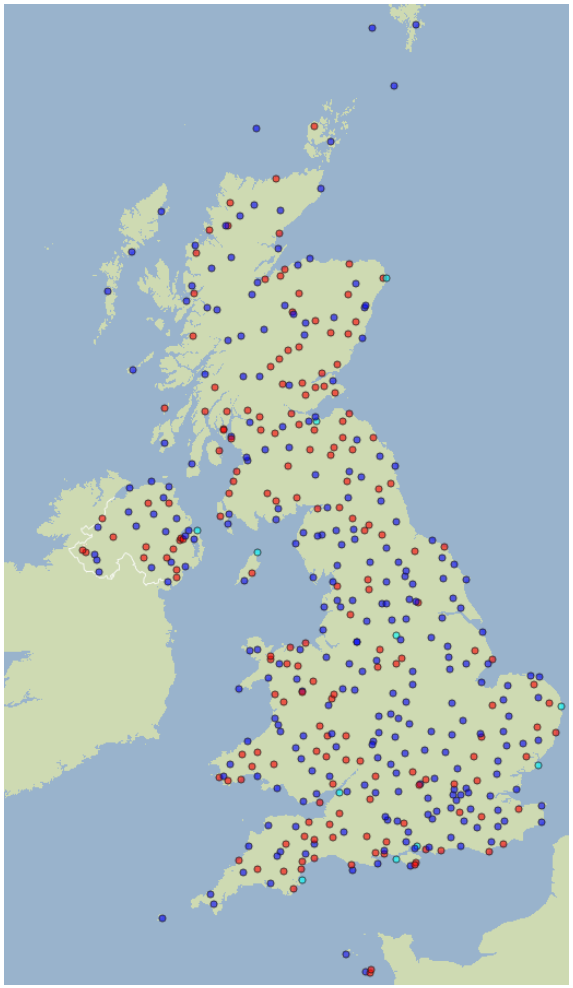
So maybe flooding isn't becoming more common and the problem is that humans have evolved to detect patterns in randomness.

More caveats

- Flooding and monthly rainfall are not the same thing
  - Perhaps the way land is managed causes less water retention on the hillsides (suggested in the Friends of the Earth document)
  - Perhaps the way rivers are managed causes them to flood more often
- The system being studied could change
  - Analysis of ancient silt deposits suggests flooding was more common in the past
    (http://www.cam.ac.uk/research/news/unprecedented-storms-and-floods-are-more-common-than-we-think)
  - Climate change has always occurred and is known to be occurring now

What about the broken records?

Remember the 24-hour rainfall record was broken at Honister Pass, and the 48-hour record was broken at Thirlmere



There are 455 weather stations in the UK

Final comments

Data is available from many other weather stations, so we might get a better picture if we study more of them (the next session)

Floods have always occurred once in a while, they continue to do so, and probably will in the future

References

- Common Terminology Criteria for Adverse Events, U. S. Department of Health and Human Services, v4.03, 2010,
- E. J. Gumbel, Statistics of Extremes, Columbia University Press, 1958
- S. Coles, An Introduction to Statistical Modeling of Extreme Values, Springer, 2001
- H. Southworth and J. E. Heffernan, Extreme value modelling of laboratory safety data from clinical studies, Pharmaceutical Statistics, 2012