Life Conference and Exhibition 2012
Chris Reynolds, PartnerRe
Niel Daniels, Daniels Actuarial Consulting

F8: How Powerful are your Rating Factors?

(Pre-reading)

6th November 2012

Introduction

- This document serves as primer for the Life Conference 2012 presentation “How Powerful are Your Rating Factors?”
- Here we provide a brief introduction to the theory of Generalised Linear Modelling. It is not intended as a rigorous theoretical introduction to GLMs, but we do provide a number of sources that the interested reader may wish to consult.
- During the presentation Niel Daniels and Chris Reynolds will provide a practical demonstration of using GLMs with mortality data.
- We will give a live demonstration of using the software R to analyse a mortality dataset. This will include the use of GLMs and tree based methods.
- The presentation will also include the results of fitting a GLM to PartnerRe’s mortality data warehouse and the power of different rating factors.
- The talk will be a practical introduction to how you could use GLMs. It will not be an in-depth study of the statistical analysis underpinning GLMs and we will not be overly focused on numerical results.

Disclaimer

The views expressed within this document are those of the presenters and do not necessarily reflect those of their employers, and thus, their employers accept no liability as a result of any reliance you may have placed or action taken based upon the information outlined in this document / presentation

1

Traditional 1 way analysis

Traditionally, experience analyses have taken an expected table and then derived ratios of "Actual" over "Expected (A/E).

The expected tables are typically split by age, gender and smoker status.

Furthermore the A/Es are then summarised to understand the impact of key rating factors.

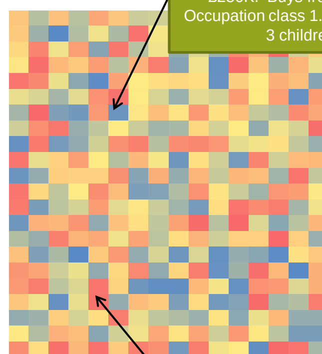
For example they may be summarised by Sum Assured, Socio-economic group, calendar year, etc.

However, it is difficult to deal with the interactions between these factors. For example, there will clearly be a significant interaction between sum assured and socio-economic group. Hence by looking at the factors individually we are at risk of double counting effects.

Life is a rich tapestry and there are there are many factors which impact a person's mortality.

The next page shows just some of the factors that may impact on an individual's expected mortality.

Male : Age 50 living in the countryside. Non smoker. SA - £250K. Buys from an IFA. Occupation class 1. Married with 3 children



Female : Age 30 living in the city centre. Smoker. SA - £30K. Buys from direct marketing. Occupation class 4. Single with 1 child.

… and many more possible combinations

Source: PartnerRe

Factors that impact expected mortality?

How can we allow for the potential interaction between all these factors?
We need to perform a multi-factor analysis.

Source: istockphoto

Linear Regression

- **Random Structure**
  There are 2 types of variables:
  1. Explanatory or predictor variables (e.g. age, sex, BMI, etc)
  2. Response variables (e.g. number of deaths)
- We start by recognizing that the response is random. We model this fact by treating the responses y_i as realizations of random variables $N(\mu, \sigma^2)$:

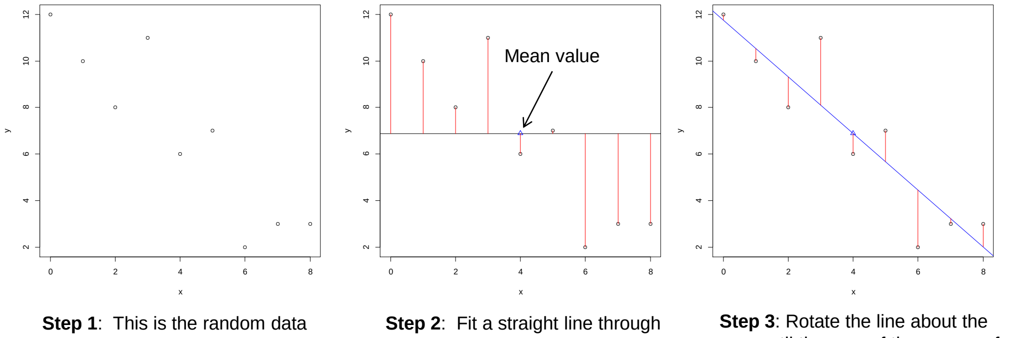
$$Y_i \sim N(\mu_i, \sigma^2)$$

- **Systematic Structure**
  In linear regression the focus is on the mean, namely $\mu_i = E[Y_i]$. One of the simplest relationships between the mean and the predictor is to use a straight line:

$$\mu_i = \alpha + \beta x_i$$

  In other words, the response, y_i, is a function of:
  - α - a fixed term
  - βx - a multiple of x
  - ε - an error term
- A simple example of this might be: Mortality = 80% + 10%*Socio-Economic Group + error

Graphical Illustration of Linear Regression



Step 1: This is the random data that we wish to perform a linear regression on.

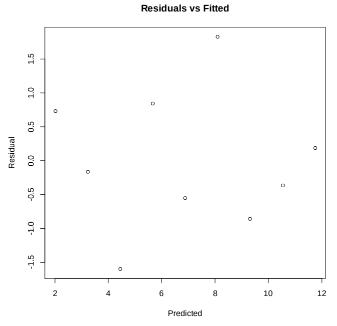
Step 2: Fit a straight line through the mean value. Calculate the errors (red lines) between the observed data and the fitted line.

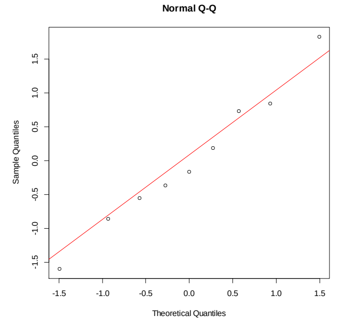
Step 3: Rotate the line about the mean until the sum of the square of the errors is minimised.

Source: PartnerRe

Linear Regression – Checking the Fit

- **Checking the fit.** After fitting the model, one should look at the (standardised) residuals. The residuals should be random in nature and have a normal distribution.





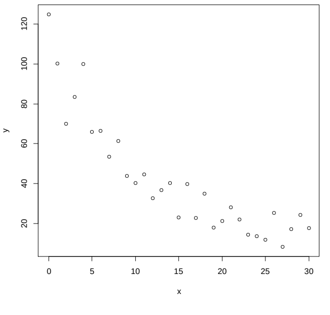
Residuals vs Fitted: This should show no discernible pattern or bias. Essentially you're looking for something that looks random, akin to looking at the sky on a clear night.

Q-Q Plot: The plot is checking for non-normality of errors. It is known as a q-q (quintile-quintile) plot. It plots the ranked samples from our distribution against a similar number of ranked quintiles taken from a normal distribution. If the errors are normally distributed then the points should sit on or close to the red line.

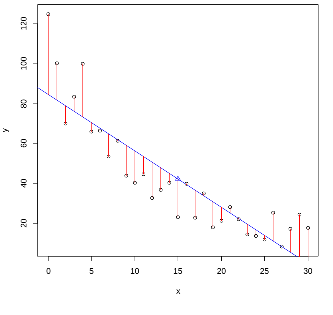
Source: PartnerRe

Linear Regression - Limitations

- **Limitations.** Unfortunately you soon come across datasets where fitting a straight line is inappropriate.



Now consider this dataset. Can we fit a linear model through it?



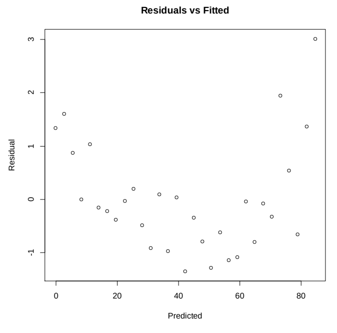
Yes we can, but how well does it fit?

The picture above suggest it's a reasonable fit, but let's analyse the residuals.

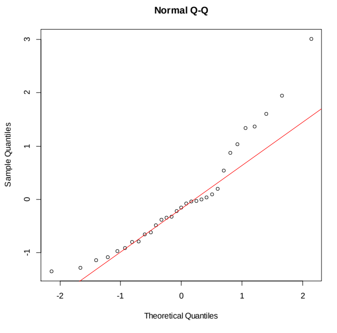
Source: PartnerRe

Linear Regression – Checking the Fit

- **Checking the fit.** After fitting the model, one should look at the residuals. The residuals should be random and nature and have a normal distribution.



Residuals vs Fitted: This should show no discernible pattern, but there appears to be a U shape here.



Q-Q Plot: There is a clear divergence between the sample and theoretical quintiles in the top right hand corner. The plot suggests non-normality of errors.

This suggests that the model is not a good fit for the data.

Source: PartnerRe

Linear Regression - Limitations

There are potentially 3 problems here:

(1) The relationship between the response and the predictor may not be linear.

(2) A normal distribution for the response is inappropriate;

(3) The variance will often increase linearly with the mean, so a constant variance assumption is inappropriate.

If these conditions don't hold, then the assumptions underlying linear regression don't hold. This was the case in the second example we considered.

What do we do in such a case?

We <u>generalise</u> the <u>linear model</u> framework.

The 3 part GLM Recipe

Setting up a Generalised Linear Model involves 3 key components:

(1) Random Component

Firstly you identify the response variable Y and assume a probability distribution for it. You are no longer restricted to a normal distribution.

(2) Systematic Component

Specify what the explanatory variables X are. This gives the linear component $\alpha + \beta X_i$

(3) Link

Specify the relationship g between the mean **E**[Y] and the systematic component X :

$$g(E[Y_i]) = \alpha + \beta X_i$$

Under linear regression g is just the identity function.

Using a link function is perfectly natural as often we don't see a simple linear relationship with things like claims frequencies. We know that risk rates follow more of an exponential shape, so we need to transform the data. Transformations can take a number of forms but a log transformation is often best for our work, since log(Mortality) is reasonably linear.

Poisson Regression

Consider Y, the number of events from exposure n. If we assume that Y has a Poisson distribution, $Y_i \sim Po(n_i)$, then the mean of Y will be

$$E(Y_i) = m_i = n_i \mu_i$$

where μ_i is the force of mortality.

The dependence of μ_i on the explanatory variables is modelled under the Gompertz mortality law by

$$\mu_i = e^{x_i \beta}$$

The implies that the mean will be:

$$E(Y_i) = m_i = n_i e^{x_i \beta}$$

Use the logarithmic function gives the generalised linear model:

$$\log \mu_i = \log n_i + x_i \beta$$

The term "log n_i" (i.e. log of the exposure) is a known constant which is called the "offset" in GLM parlance.

Further Reading

- CMI – Working Paper 58 (2011)
- An Introduction to Generalized Linear Models, Dobson & Barnett (2008)
- Statistics: An Introduction using R, Crawley (2005)
- Generalized Linear Models for Insurance Data, Jong & Heller (2008)
- Demystifying GLMs (Sessional Meeting - Australia), Henwood et al (1991)
- Risk classification in life insurance: methodology and case study, Gschlössl, Schoenmaekers and Denuit (2011)
- Actuarial Graduation Practice and Generalised Linear and Non-Linear Models, Renshaw (1991)