The Actuarial Profession
making financial sense of the future

Health and Care Conference 2012

Chris Reynolds, PartnerRe
Niel Daniels, Daniels Actuarial Consulting

C4 : How Powerful are Your Rating Factors?

1 May 2012

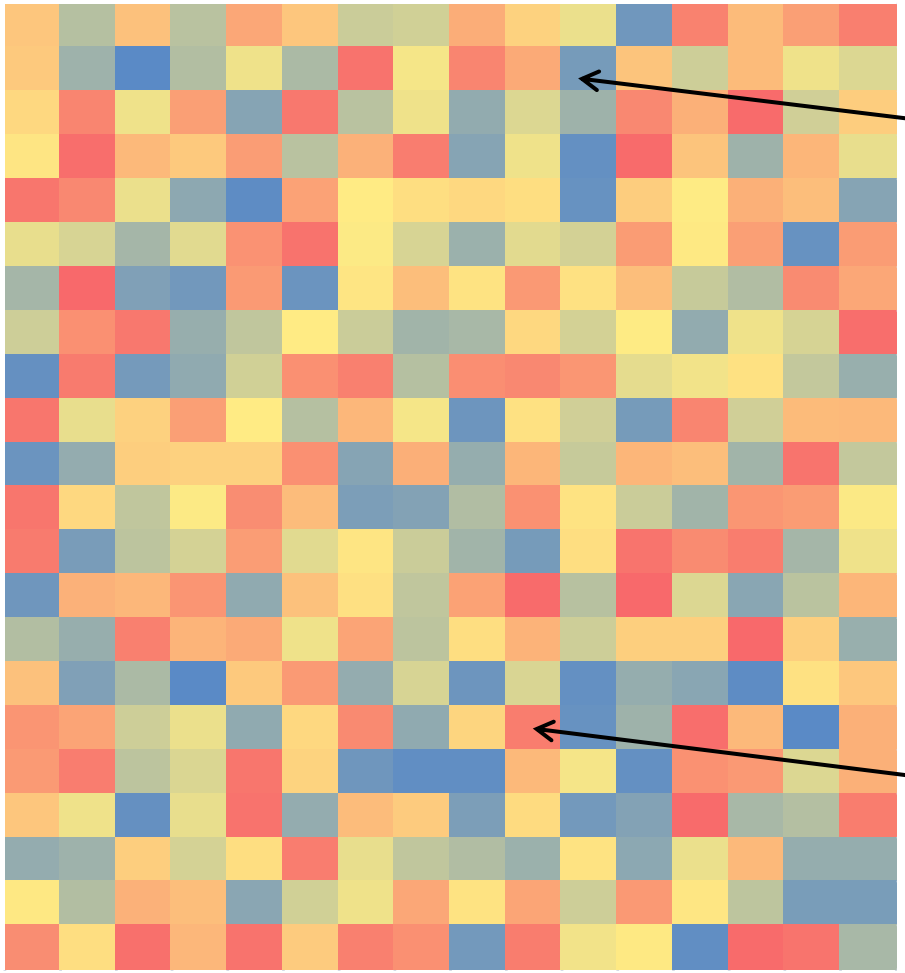
Disclaimer

- The views expressed today 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

Agenda

- Traditional 1-way Analysis
- GLM Theory
- Real Time Demo using R
- Creating a GLM model
- Results on a Critical Illness Dataset
- Removal of key factors
- Summary & Questions

Traditional 1 way analysis



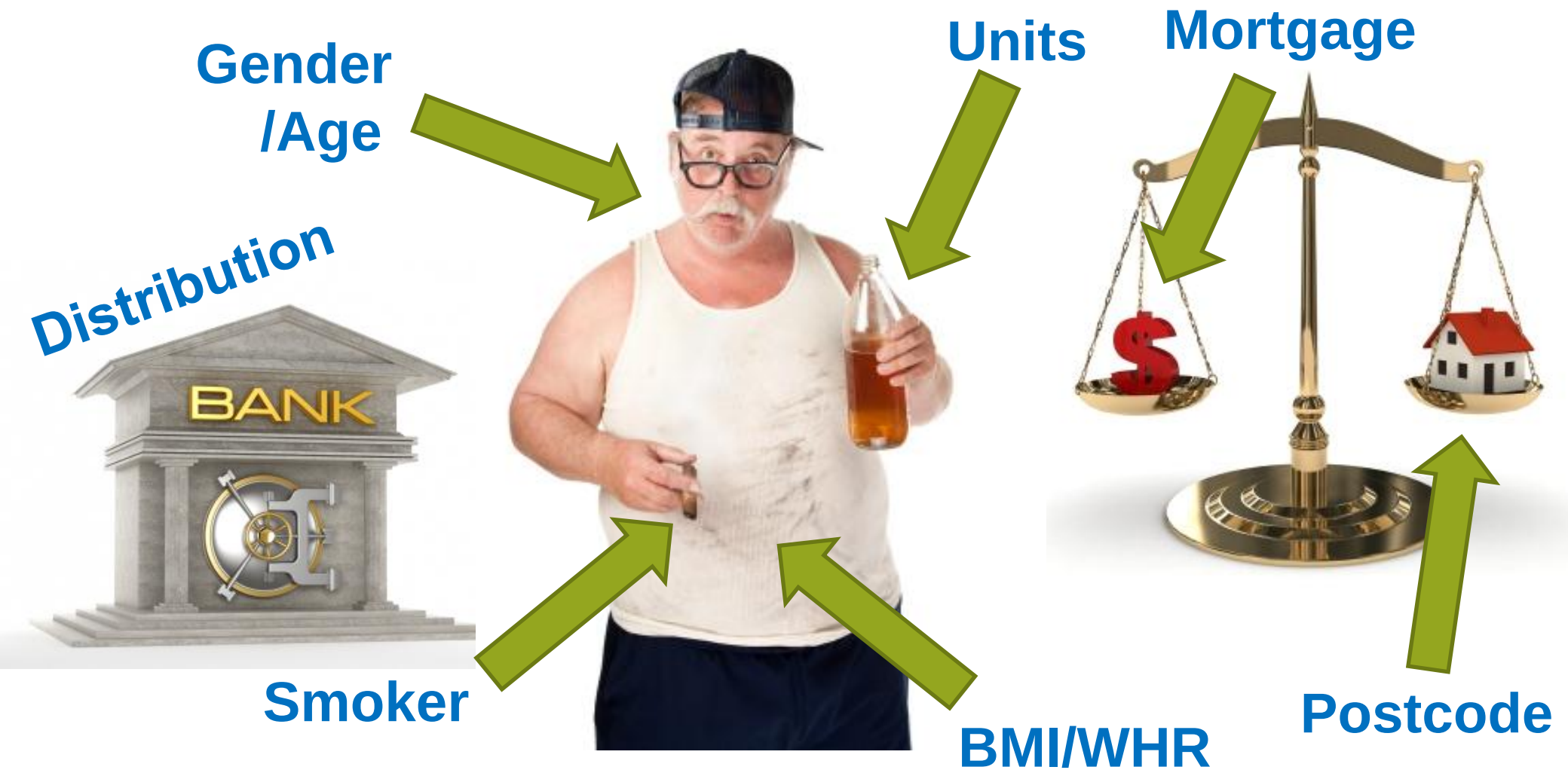
Male : Age 50 living in the countryside. Non smoker. SA - £250K. Buys from an IFA. Occupation class 1. Married with 3 children

and many more possible combinations

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

Source: PartnerRe

How do factors interact?



Source: istockphoto

Linear Regression

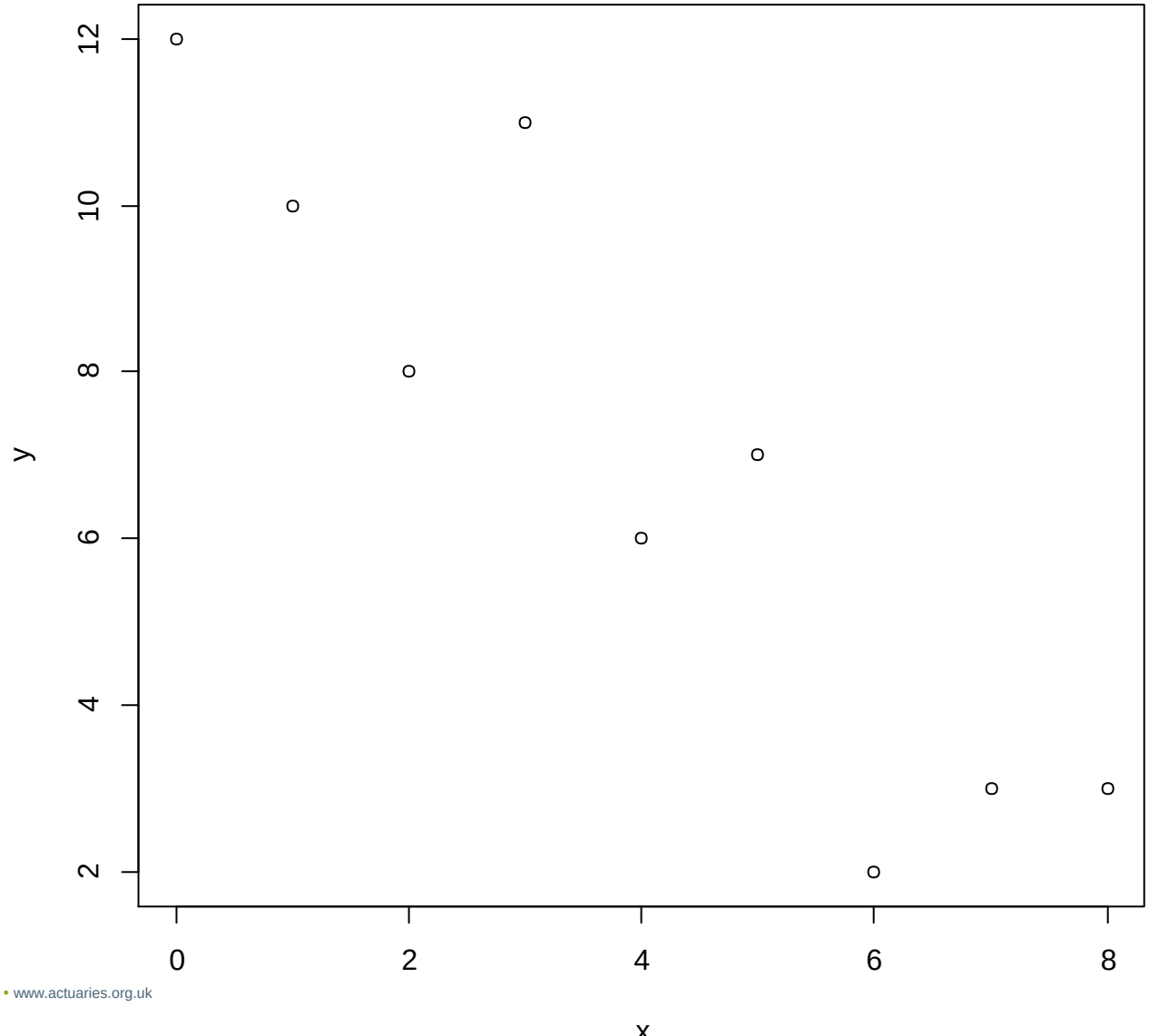
- **Random Structure**
  Responses vary even for constant values of the predictor

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

- **Systematic Structure**
  The simplest was to express the dependence of the response μ_i on the predictor x_i is to assume a linear function

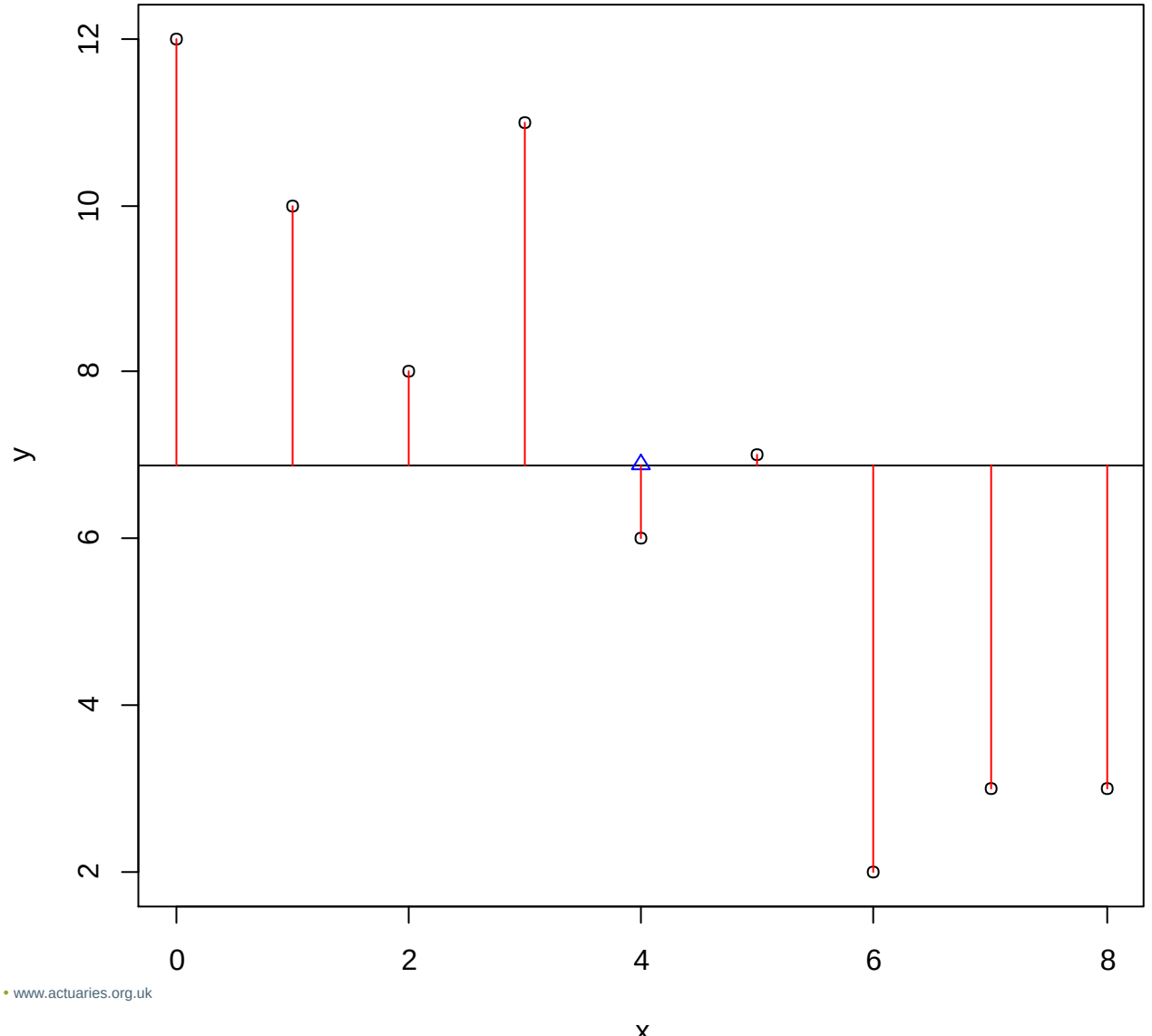
$$\mu_i = \alpha + \beta x_i$$

Linear Regression



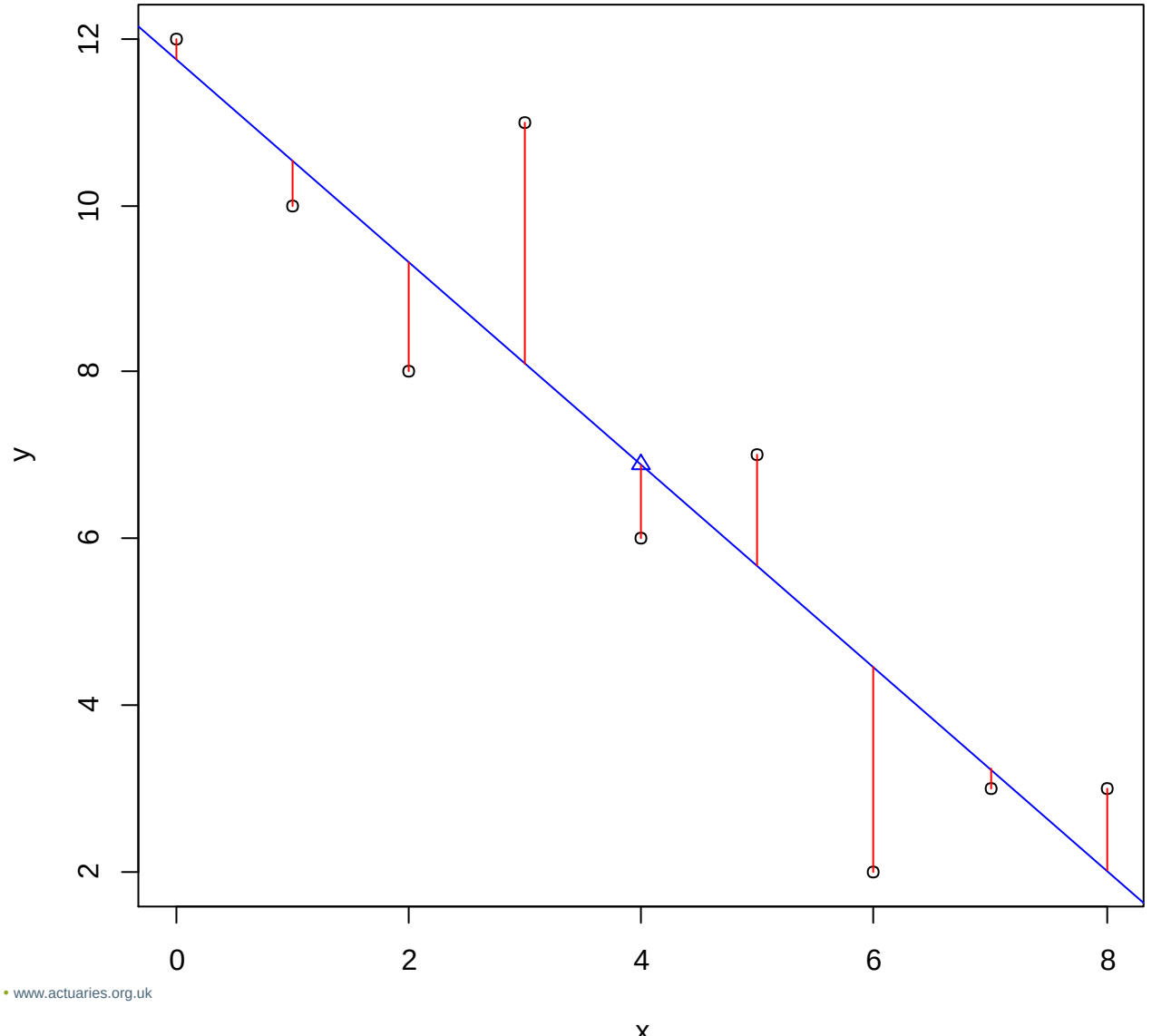
Source: PartnerRe

Linear Regression



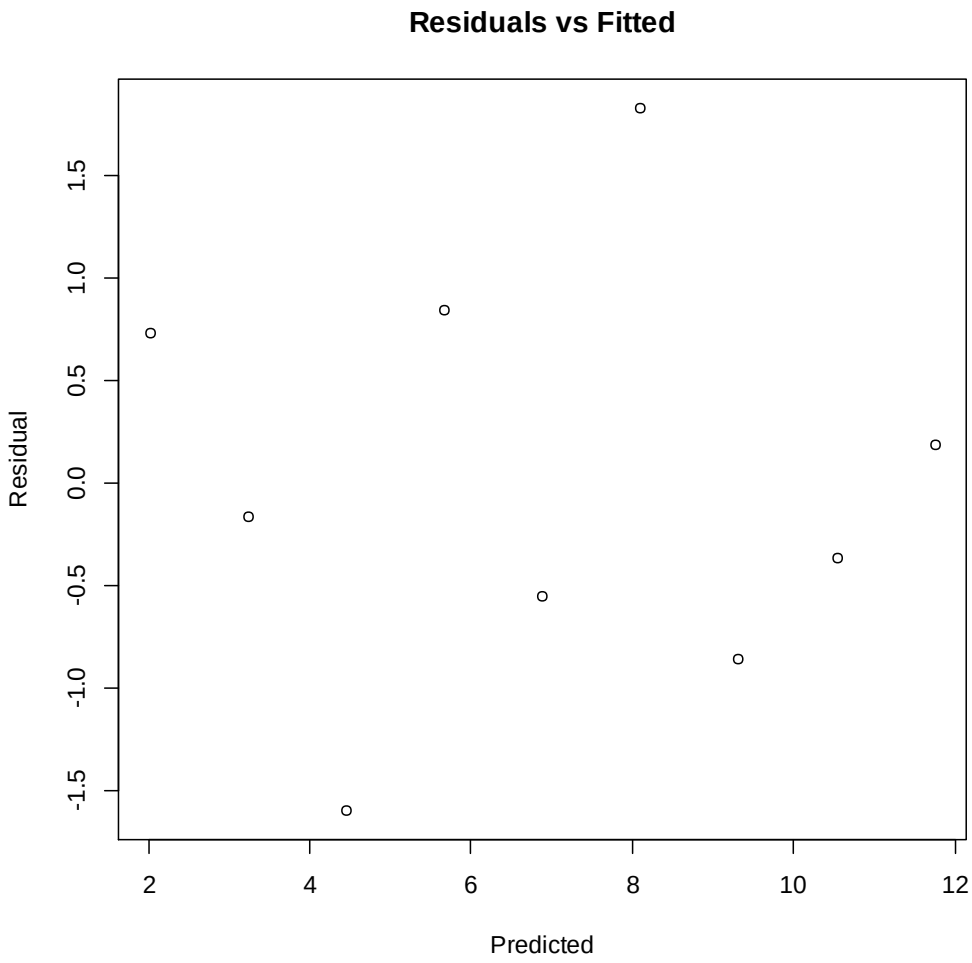
Source: PartnerRe

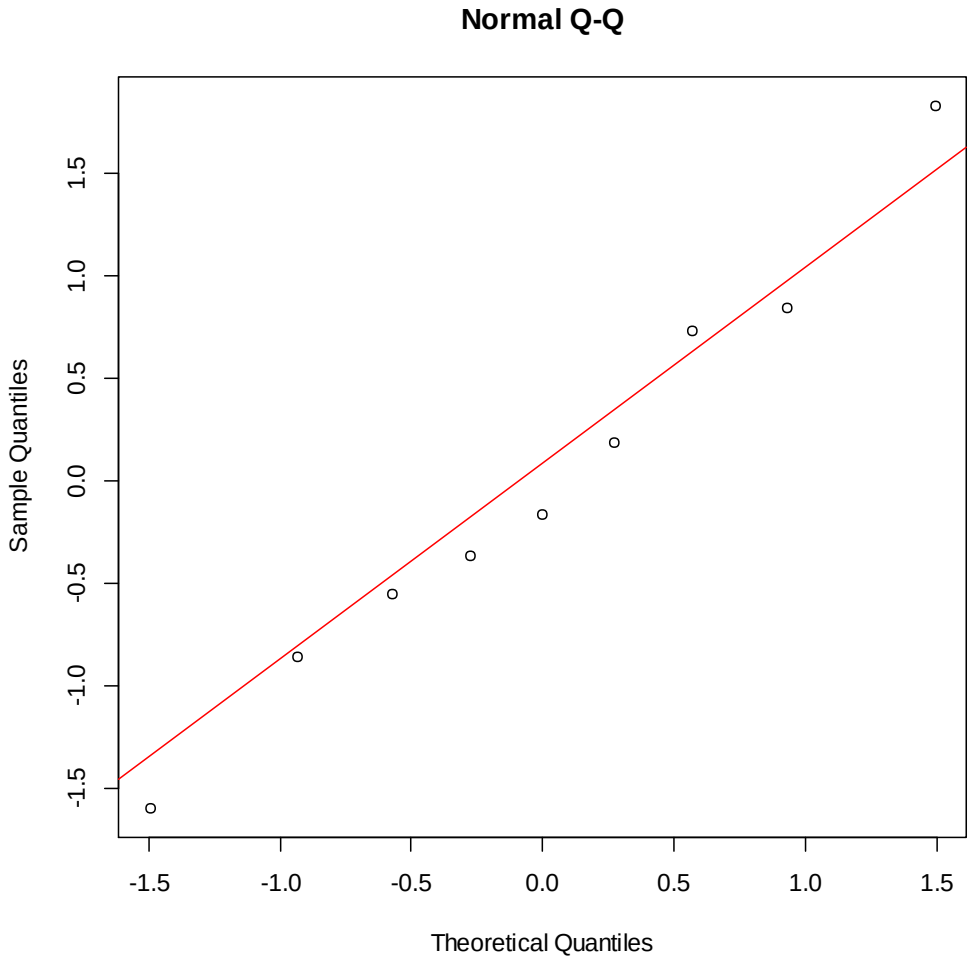
Linear Regression



Source: PartnerRe

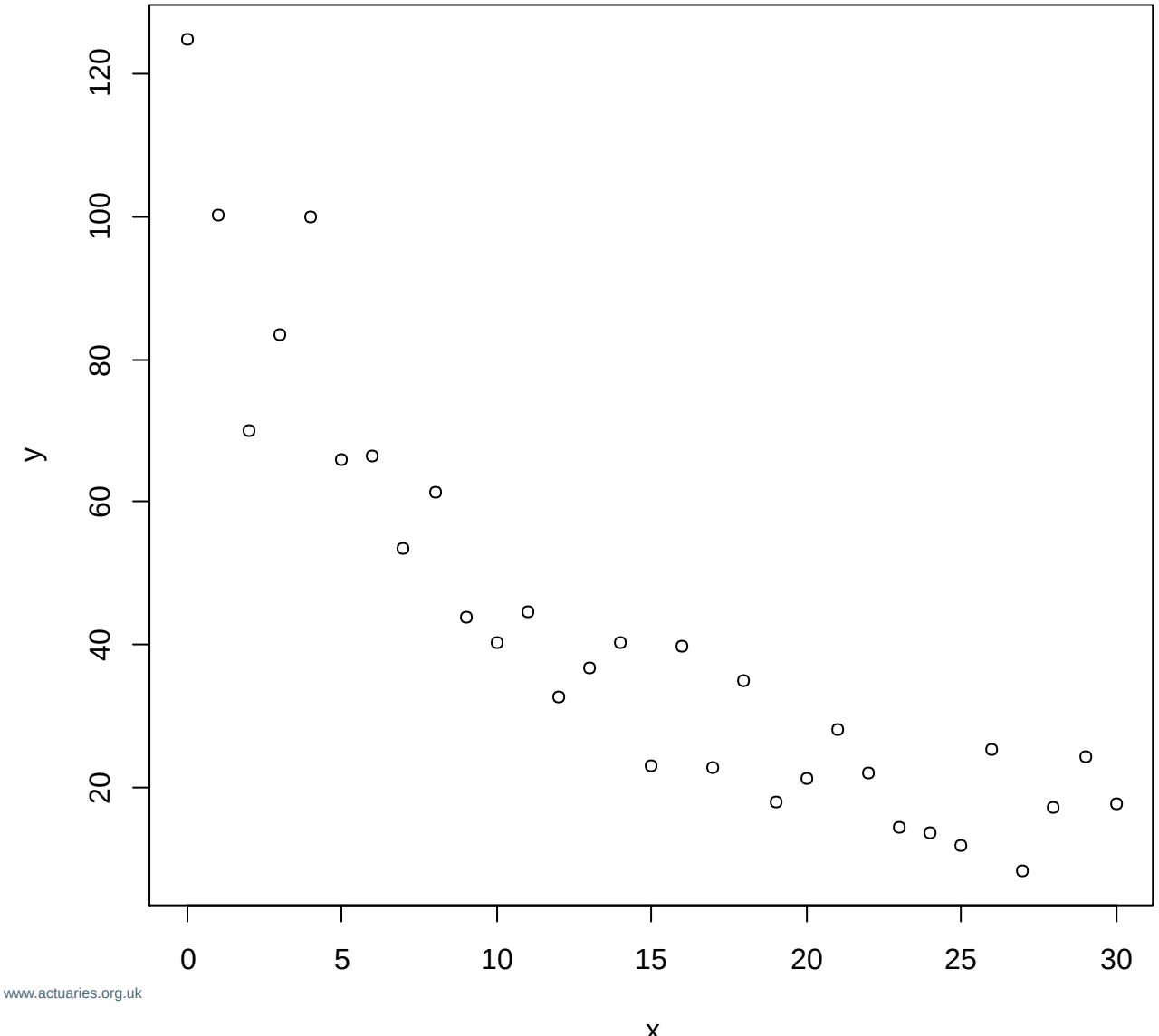
Linear Regression





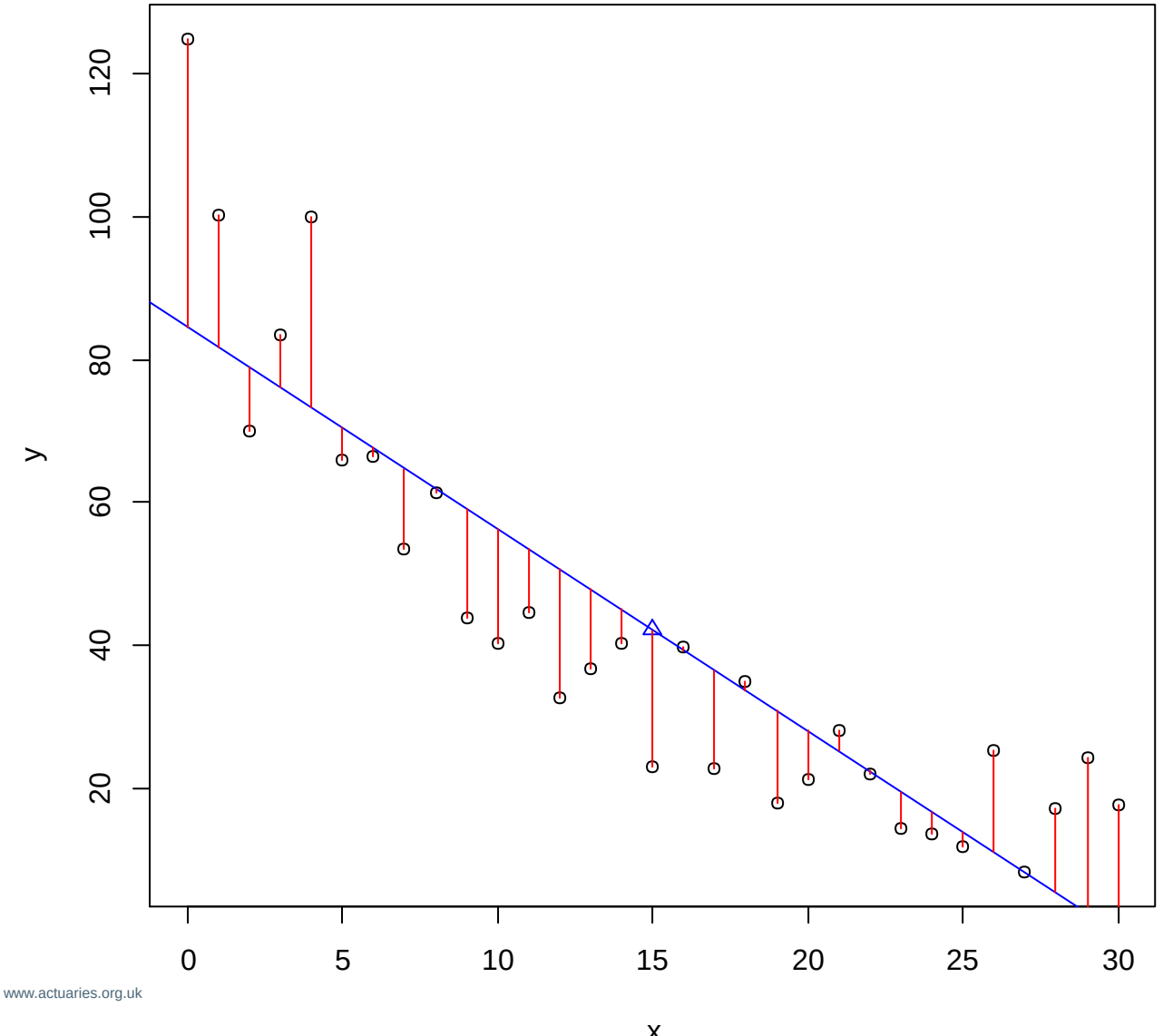
Source: PartnerRe

Linear Regression - Limitations



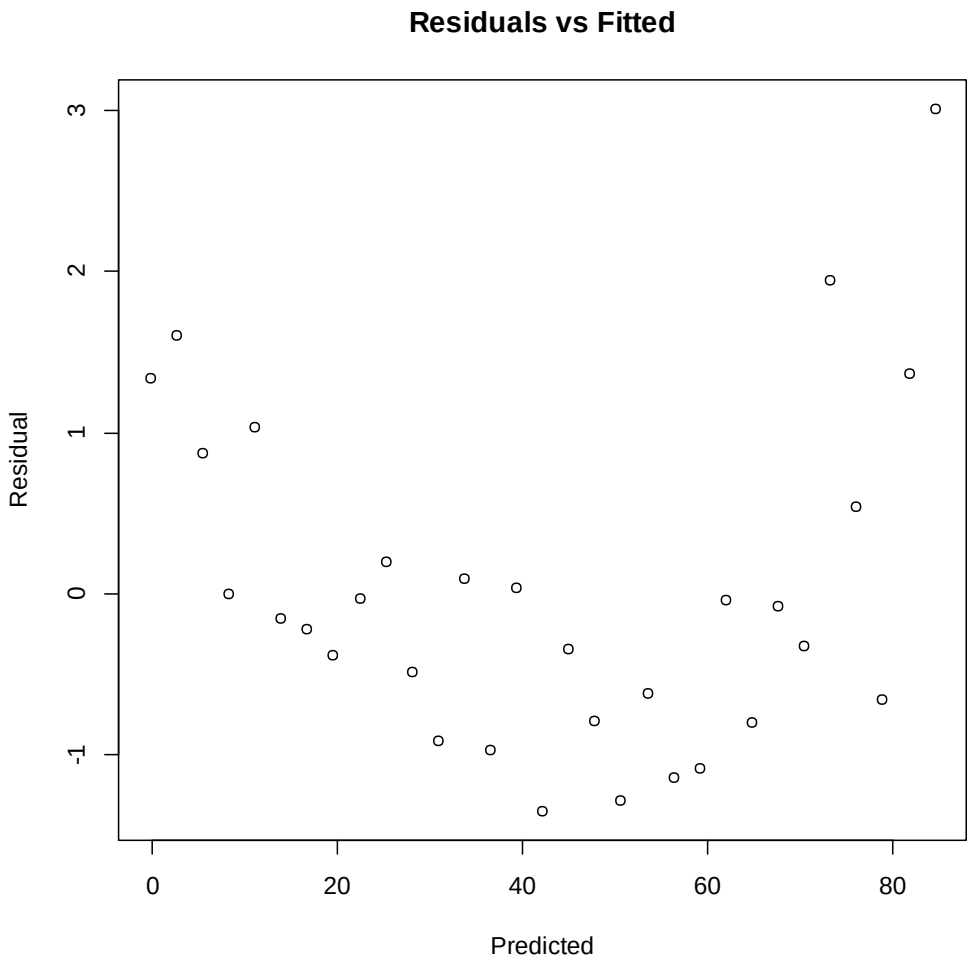
Source: PartnerRe

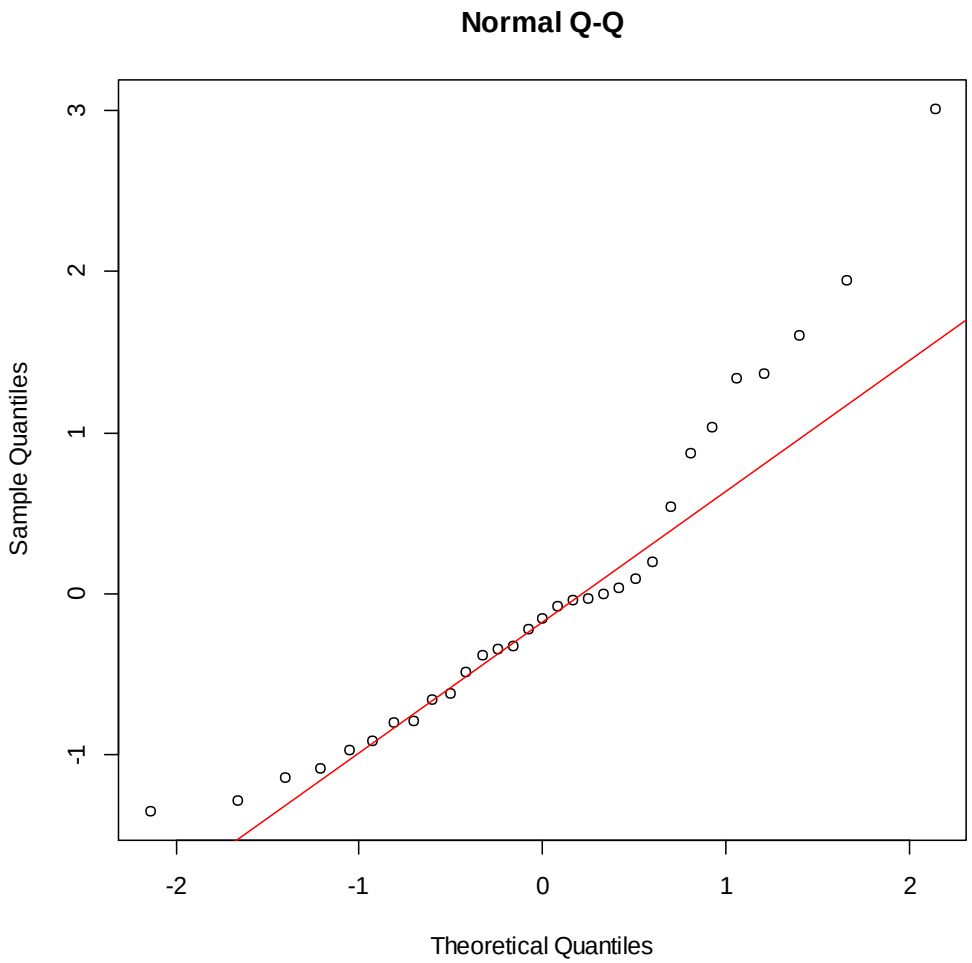
Linear Regression - Limitations



Source: PartnerRe

Linear Regression - Limitations





Source: PartnerRe

Linear Regression - Limitations

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

(2) A normal distribution for the response may be inappropriate;

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

What do we do???

We **generalise** the model framework.

The 3 part GLM Recipe

(1) Random Component

Identify the response variable Y and assume a probability distribution for it

(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 :

$$g\left(\mathrm{E}\left[Y_i\right]\right) = \alpha + \beta X_i$$

Our Model – Poisson Regression

$Y_i \sim Po(m_i)$ denotes the number of events from exposure n_i

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

The dependence of μ_i on the explanatory variables is usually modelled by

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

So that our generalised linear model is:

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

This is called the "offset"

Use the logarithmic log function gives:

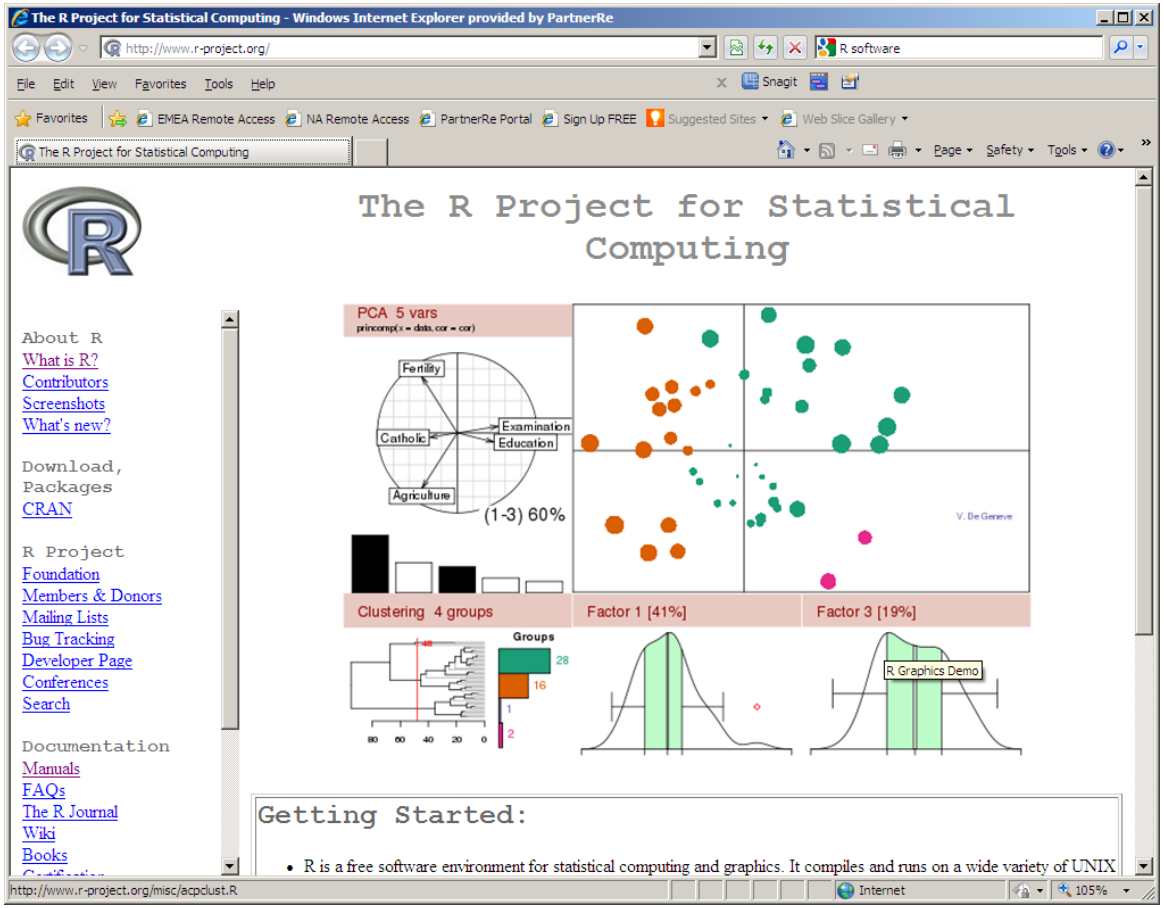
$$\log \mu_i = \log n_i + x_i \beta$$

R Software – What is it?

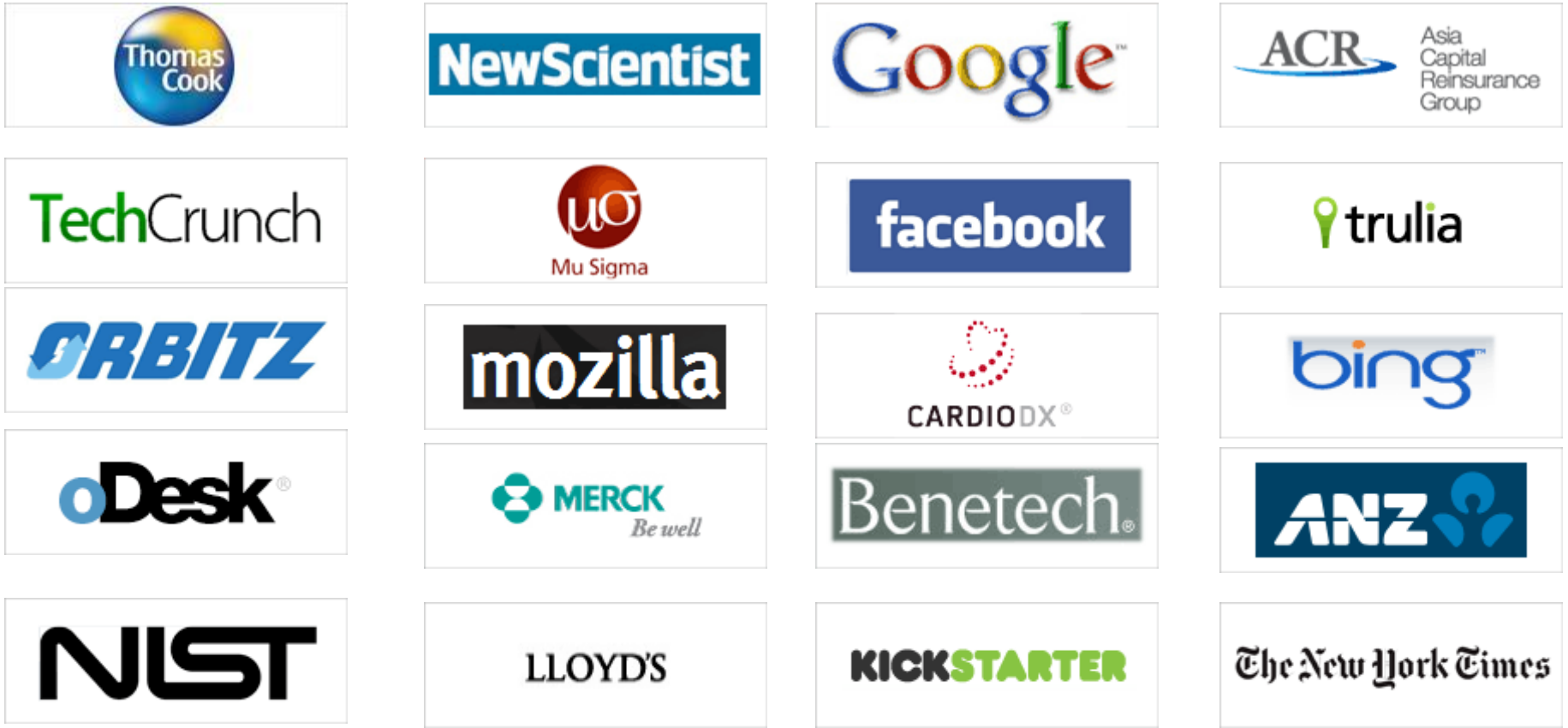
"R is an open-source, object-oriented statistical programming language. In the past decade, it has become the global lingua franca of statistics"

www.r-project.org

http://rcom.univie.ac.at/download.html



R – Should a company trust free software?



These companies do*

*Source : http://www.revolutionanalytics.com/what-is-open-source-r/companies-using-r.php

Real time demonstration

- Using HMD* mortality data
- Years 1960 onwards
- Ages 20 – 65
- UK, Ireland, Poland & Japan

* Human Mortality Database (www.mortality.org)

Real time demonstration

Creating a GLM model

- A number of questions need to be answered when deciding how to build a GLM model :
  - Trade off between complexity & ease of use
  - Base model on exposure or expected
  - Start simple & build vs start with saturated and strip down

Creating a GLM model
Complexity vs Ease of use

- Simpler models are always easier to understand
- But more complex models will usually fit better
- And having more rating factors will always fit better
- But is this spurious, do these extra factors add value
- There is no simple answer to where to compromise – practice makes perfect

- *There are tools that can fit the model – but you still need to think carefully about the appropriate model to use !*

Creating a GLM model
Do we model as Actual = fn (Exposure) or fn (Expected)

- Traditional actuarial experience analyses use an Actual / Expected approach
- This has the advantage of starting from our expected point of view
- So any rating factors that we derive can be directly applied to our underlying expected basis
- But, if we use exposure as our start point
  - We get a direct assessment of the risk rate
  - And an assessment of which rating factors are significant

Creating a GLM model
Start simple & build vs start with saturated and strip down

- The traditional approach might be to start simple and build up
- E.g. starting with age / sex / smoker and then adding sum assured / socio-economic / etc
- But it's often easier in the GLM world to start with all available rating factors and then strip out factors (or combinations of factors) that don't add much value

Critical Illness data
Our dataset

- Several companies
- Many thousands of claims
- All UK, Accelerated CI
- Chosen datasets with rich rating factors
  - eg socio-economic, rated status, joint v single, benefit shape, calendar year, etc
- We've taken a number of shortcuts in doing this analysis – eg no IBNR allowance was made – so we are focusing on the techniques and processes rather than on the real results

Model – Investigation 1

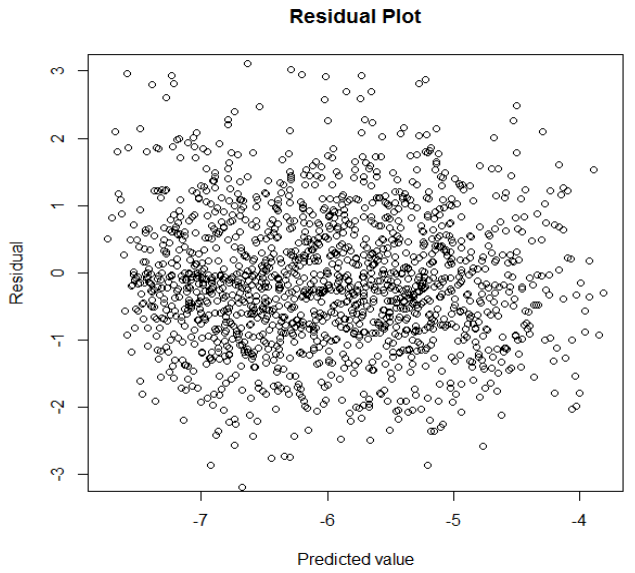
- Take a grouped dataset split by
  - Individual Age
  - Sex
  - Smoker
  - Duration
- Fit an AGE * SEX * SMK * DUR model
- Use Exposure as the “offset”
- Analyse the significance of each factor
- The principle of Parsimony (Occam’s Razor) – simplify the model to the “simplest acceptable model”

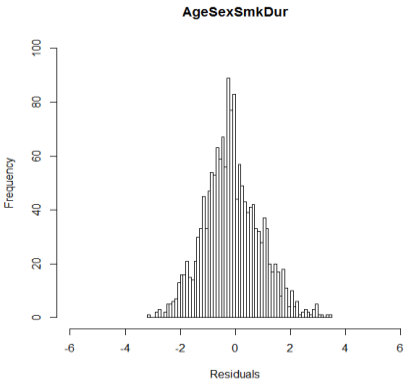
Model – Investigation 1 : Age * Sex * Smk * Dur

R Console (64-bit)

File Edit Misc Packages Windows Help

```
Call:
glm(formula = MtxG1$DTHS ~ offset(log(MtxG1$EXPO)) + MtxG1$AGE *
    MtxG1$SEX * MtxG1$SMOKER * MtxG1$DUR, family = poisson)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.1584  -0.7860  -0.1835   0.5170   3.4008

Coefficients:
                                          Estimate Std. Error z value Pr(>|z|)
(Intercept)                             -9.3654216  0.1570119 -59.648   <2e-16 ***
MtxG1$AGE                                0.0730491  0.0038214  19.116   <2e-16 ***
MtxG1$SEXM                              -0.2220890  0.2265305  -0.980   0.3269
MtxG1$SMOKERS                           -0.4327148  0.3937931  -1.099   0.2718
MtxG1$DUR                                0.0986692  0.0385242   2.561   0.0104 *
MtxG1$AGE:MtxG1$SEXM                     0.0008378  0.0054053   0.155   0.8768
MtxG1$AGE:MtxG1$SMOKERS                  0.0170185  0.0094561   1.800   0.0719 .
MtxG1$SEXM:MtxG1$SMOKERS                 0.3034664  0.5153018   0.589   0.5559
MtxG1$AGE:MtxG1$DUR                     -0.0019802  0.0009033  -2.192   0.0284 *
MtxG1$SEXM:MtxG1$DUR                    -0.0888574  0.0549402  -1.617   0.1058
MtxG1$SMOKERS:MtxG1$DUR                  0.0178025  0.0924713   0.193   0.8473
MtxG1$AGE:MtxG1$SEXM:MtxG1$SMOKERS       0.0013808  0.0123252   0.112   0.9108
MtxG1$AGE:MtxG1$SEXM:MtxG1$DUR           0.0026388  0.0012589   2.096   0.0361 *
MtxG1$AGE:MtxG1$SMOKERS:MtxG1$DUR       -0.0001962  0.0021508  -0.091   0.9273
MtxG1$SEXM:MtxG1$SMOKERS:MtxG1$DUR      -0.0808716  0.1218284  -0.664   0.5068
MtxG1$AGE:MtxG1$SEXM:MtxG1$SMOKERS:MtxG1$DUR  0.0010437  0.0028021   0.372   0.7095
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 4799.9  on 1569  degrees of freedom
Residual deviance: 1631.0  on 1554  degrees of freedom
AIC: 5776

Number of Fisher Scoring iterations: 5
```



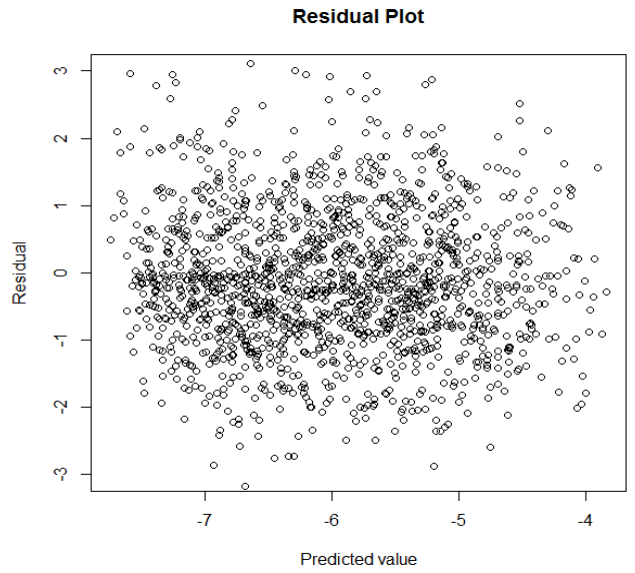


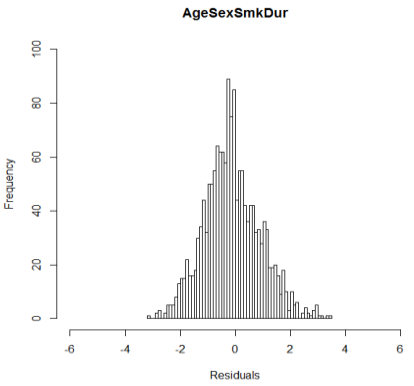
Model – Investigation 1 : Age * Sex * Smk * Dur

R Console (64-bit)

File Edit Misc Packages Windows Help

```
Call:
glm(formula = MtxG1$DTHS ~ MtxG1$AGE + MtxG1$SEX + MtxG1$SMOKER +
    MtxG1$DUR + MtxG1$AGE:MtxG1$SEX + MtxG1$AGE:MtxG1$SMOKER +
    MtxG1$SEX:MtxG1$SMOKER + MtxG1$AGE:MtxG1$DUR + MtxG1$SEX:MtxG1$DUR +
    MtxG1$SMOKER:MtxG1$DUR + MtxG1$AGE:MtxG1$SEX:MtxG1$SMOKER +
    MtxG1$AGE:MtxG1$SEX:MtxG1$DUR + MtxG1$AGE:MtxG1$SMOKER:MtxG1$DUR +
    MtxG1$SEX:MtxG1$SMOKER:MtxG1$DUR + offset(log(MtxG1$EXPO)),
    family = poisson)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.1475  -0.7850  -0.1855   0.5178   3.3938

Coefficients:
                                  Estimate Std. Error z value Pr(>|z|)
(Intercept)                     -9.3803722  0.1518533 -61.773  < 2e-16 ***
MtxG1$AGE                        0.0734147  0.0036933  19.878  < 2e-16 ***
MtxG1$SEXM                      -0.1920075  0.2116251  -0.907  0.36425
MtxG1$SMOKERS                   -0.3412458  0.3072574  -1.111  0.26673
MtxG1$DUR                        0.1032236  0.0365472   2.824  0.00474 **
MtxG1$AGE:MtxG1$SEXM             0.0001173  0.0050469   0.023  0.98146
MtxG1$AGE:MtxG1$SMOKERS          0.0148065  0.0073587   2.012  0.04421 *
MtxG1$SEXM:MtxG1$SMOKERS         0.1483638  0.3033009   0.489  0.62473
MtxG1$AGE:MtxG1$DUR             -0.0020887  0.0008554  -2.442  0.01461 *
MtxG1$SEXM:MtxG1$DUR            -0.0979055  0.0492775  -1.987  0.04694 *
MtxG1$SMOKERS:MtxG1$DUR         -0.0081670  0.0606525  -0.135  0.89289
MtxG1$AGE:MtxG1$SEXM:MtxG1$SMOKERS  0.0051063  0.0072042   0.709  0.47845
MtxG1$AGE:MtxG1$SEXM:MtxG1$DUR   0.0028496  0.0011245   2.534  0.01126 *
MtxG1$AGE:MtxG1$SMOKERS:MtxG1$DUR  0.0004180  0.0013786   0.303  0.76171
MtxG1$SEXM:MtxG1$SMOKERS:MtxG1$DUR -0.0362390  0.0219705  -1.649  0.09906 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 4799.9  on 1569  degrees of freedom
Residual deviance: 1631.1  on 1555  degrees of freedom
AIC: 5774.1
```



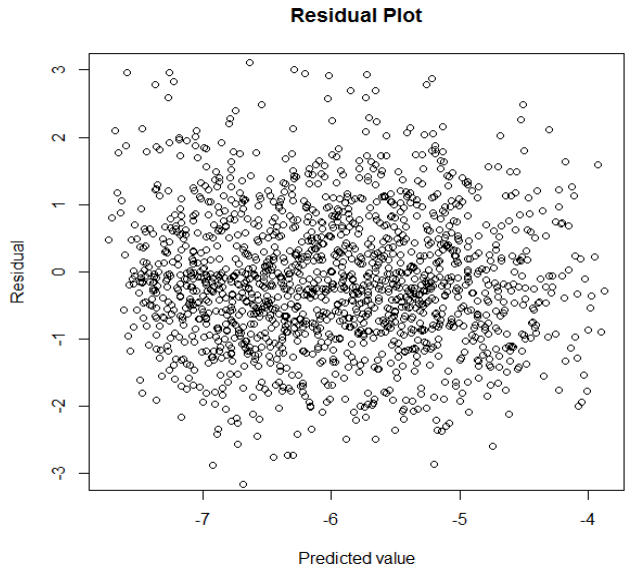


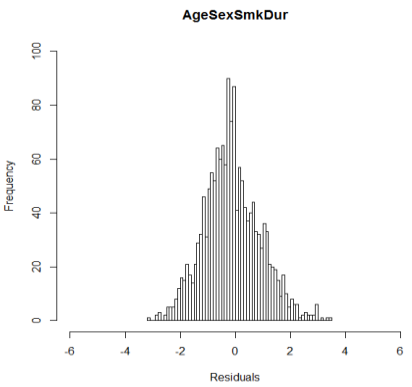
Model – Investigation 1 : Age * Sex * Smk * Dur

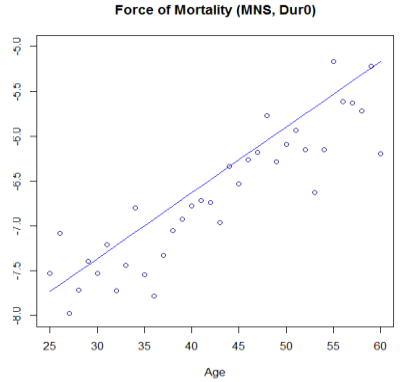
R Console (64-bit)

File Edit Misc Packages Windows Help

```
Call:
glm(formula = MtxG1$DTHS ~ MtxG1$AGE + MtxG1$SEX + MtxG1$SMOKER +
    MtxG1$DUR + MtxG1$AGE:MtxG1$SEX + MtxG1$AGE:MtxG1$SMOKER +
    MtxG1$SEX:MtxG1$SMOKER + MtxG1$AGE:MtxG1$DUR + MtxG1$SEX:MtxG1$DUR +
    MtxG1$SMOKER:MtxG1$DUR + MtxG1$AGE:MtxG1$SEX:MtxG1$SMOKER +
    MtxG1$AGE:MtxG1$SEX:MtxG1$DUR + MtxG1$SEX:MtxG1$SMOKER:MtxG1$DUR +
    offset(log(MtxG1$EXPO)), family = poisson)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.1369  -0.7853  -0.1862   0.5220   3.3985

Coefficients:
                                  Estimate Std. Error z value Pr(>|z|)
(Intercept)                     -9.370e+00  1.481e-01 -63.282  < 2e-16 ***
MtxG1$AGE                        7.317e-02  3.601e-03  20.317  < 2e-16 ***
MtxG1$SEXM                      -1.875e-01  2.111e-01  -0.888  0.37432
MtxG1$SMOKERS                   -4.035e-01  2.289e-01  -1.763  0.07795 .
MtxG1$DUR                        1.001e-01  3.508e-02   2.854  0.00432 **
MtxG1$AGE:MtxG1$SEXM             2.202e-05  5.038e-03   0.004  0.99651
MtxG1$AGE:MtxG1$SMOKERS          1.631e-02  5.429e-03   3.005  0.00266 **
MtxG1$SEXM:MtxG1$SMOKERS         1.487e-01  3.036e-01   0.490  0.62434
MtxG1$AGE:MtxG1$DUR             -2.015e-03  8.198e-04  -2.458  0.01398 *
MtxG1$SEXM:MtxG1$DUR            -9.916e-02  4.910e-02  -2.019  0.04344 *
MtxG1$SMOKERS:MtxG1$DUR         [illegible]
MtxG1$AGE:MtxG1$SEXM:MtxG1$SMOKERS  5.072e-03  7.203e-03   0.704  0.48128
MtxG1$AGE:MtxG1$SEXM:MtxG1$DUR  [illegible]
MtxG1$SEXM:MtxG1$SMOKERS:MtxG1$DUR -3.575e-02  2.190e-02  -1.633  0.10257
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 4799.9  on 1569  degrees of freedom
Residual deviance: 1631.2  on 1556  degrees of freedom
AIC: 5772.2

Number of Fisher Scoring iterations: 5
```





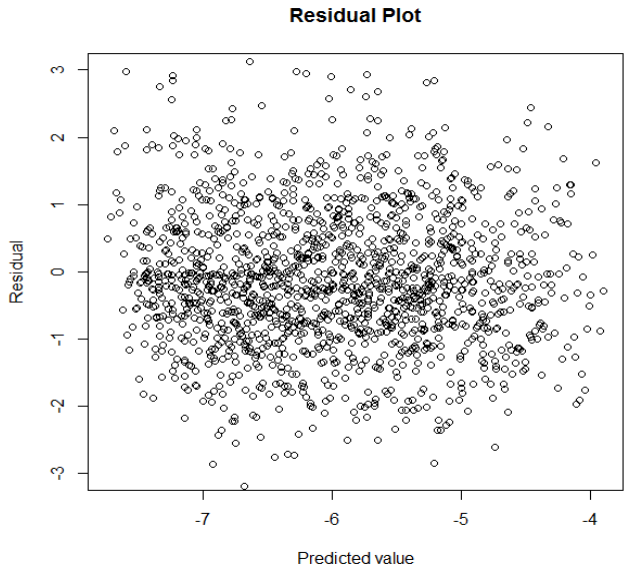


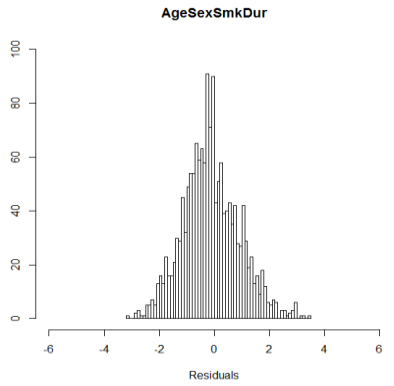
Model – Investigation 1 : Age * Sex * Smk * Dur

R Console (64-bit)

File Edit Misc Packages Windows Help

```
Call:
glm(formula = MtxG1$DTHS ~ MtxG1$AGE + MtxG1$SEX + MtxG1$SMOKER +
    MtxG1$DUR + MtxG1$AGE:MtxG1$SEX + MtxG1$AGE:MtxG1$SMOKER +
    MtxG1$SEX:MtxG1$SMOKER + MtxG1$AGE:MtxG1$DUR + MtxG1$SEX:MtxG1$DUR +
    MtxG1$SMOKER:MtxG1$DUR + MtxG1$AGE:MtxG1$SEX:MtxG1$DUR +
    MtxG1$SEX:MtxG1$SMOKER:MtxG1$DUR + offset(log(MtxG1$EXPO)),
    family = poisson)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.1569  -0.7894  -0.1904   0.5156   3.3818

Coefficients:
                                    Estimate Std. Error z value Pr(>|z|)
(Intercept)                       -9.3521881  0.1457922 -64.147  < 2e-16 ***
MtxG1$AGE                          0.0727051  0.0035426  20.523  < 2e-16 ***
MtxG1$SEXM                        -0.2262505  0.2038093  -1.110 0.266952
MtxG1$SMOKERS                     -0.5183070  0.1612034  -3.215 0.001303 **
MtxG1$DUR                          0.1012003  0.0350320   2.889 0.003867 **
MtxG1$AGE:MtxG1$SEXM               0.0009929  0.0048462   0.205 0.837666
MtxG1$AGE:MtxG1$SMOKERS            0.0191907  0.0035680   5.378 7.51e-08 ***
MtxG1$SEXM:MtxG1$SMOKERS           0.3512680  0.0977464   3.594 0.000326 ***
MtxG1$AGE:MtxG1$DUR               -0.0020334  0.0008195  -2.481 0.013091 *
MtxG1$SEXM:MtxG1$DUR              -0.1007753  0.0490574  -2.054 0.039953 *
MtxG1$SMOKERS:MtxG1$DUR            0.0079565  0.0166079   0.479 0.631884
MtxG1$AGE:MtxG1$SEXM:MtxG1$DUR     0.0028870  0.0011212   2.574 0.009771 **
MtxG1$SEXM:MtxG1$SMOKERS:MtxG1$DUR -0.0325515  0.0214269  -1.519 0.128714
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 4799.9  on 1569  degrees of freedom
Residual deviance: 1631.7  on 1557  degrees of freedom
AIC: 5770.7

Number of Fisher Scoring iterations: 5

> |
```



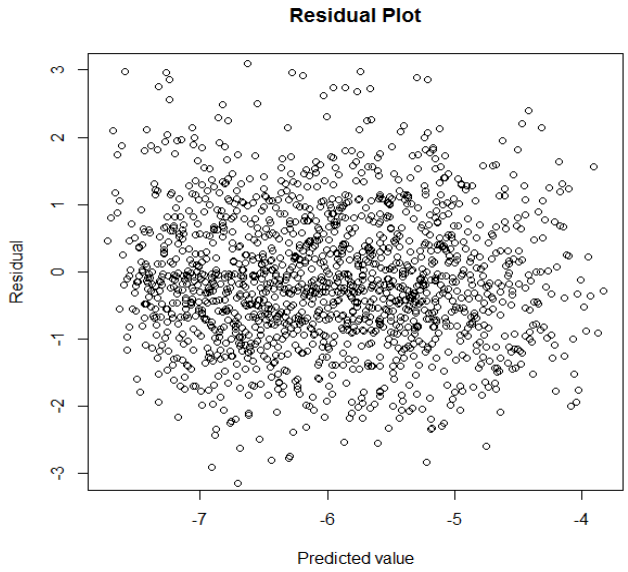


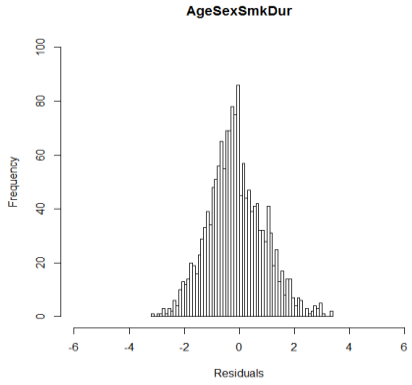
Model – Investigation 1 : Age * Sex * Smk * Dur

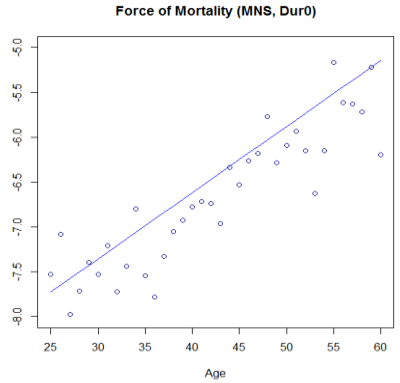
R Console (64-bit)

File Edit Misc Packages Windows Help

```
Call:
glm(formula = MtxG1$DTHS ~ MtxG1$AGE + MtxG1$SEX + MtxG1$SMOKER +
    MtxG1$DUR + MtxG1$AGE:MtxG1$SEX + MtxG1$AGE:MtxG1$SMOKER +
    MtxG1$SEX:MtxG1$SMOKER + MtxG1$AGE:MtxG1$DUR + MtxG1$SEX:MtxG1$DUR +
    MtxG1$SMOKER:MtxG1$DUR + MtxG1$AGE:MtxG1$SEX:MtxG1$DUR +
    offset(log(MtxG1$EXPO)), family = poisson)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.1160  -0.7795  -0.1856   0.5205   3.3630

Coefficients:
                                Estimate Std. Error z value Pr(>|z|)
(Intercept)                   -9.360434   0.145936 -64.140  < 2e-16 ***
MtxG1$AGE                      0.072614   0.003548  20.467  < 2e-16 ***
MtxG1$SEXM                    -0.202635   0.203031  -0.998  0.31826
MtxG1$SMOKERS                 -0.437615   0.151767  -2.883  0.00393 **
MtxG1$DUR                      0.103064   0.035004   2.944  0.00324 **
MtxG1$AGE:MtxG1$SEXM           0.001008   0.004842   0.208  0.83507
MtxG1$AGE:MtxG1$SMOKERS        0.018993   0.003565   5.327 9.97e-08 ***
MtxG1$SEXM:MtxG1$SMOKERS       0.228323   0.054352   4.201 2.66e-05 ***
MtxG1$AGE:MtxG1$DUR           -0.002000   0.000819  -2.442  0.01462 *
MtxG1$SEXM:MtxG1$DUR          -0.107251   0.048866  -2.195  0.02818 *
MtxG1$SMOKERS:MtxG1$DUR       -0.011237   0.010780  -1.042  0.29720
MtxG1$AGE:MtxG1$SEXM:MtxG1$DUR  0.002895   0.001121   2.582  0.00982 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 4799.9  on 1569  degrees of freedom
Residual deviance: 1634.0  on 1558  degrees of freedom
AIC: 5771

Number of Fisher Scoring iterations: 5

>
>
```





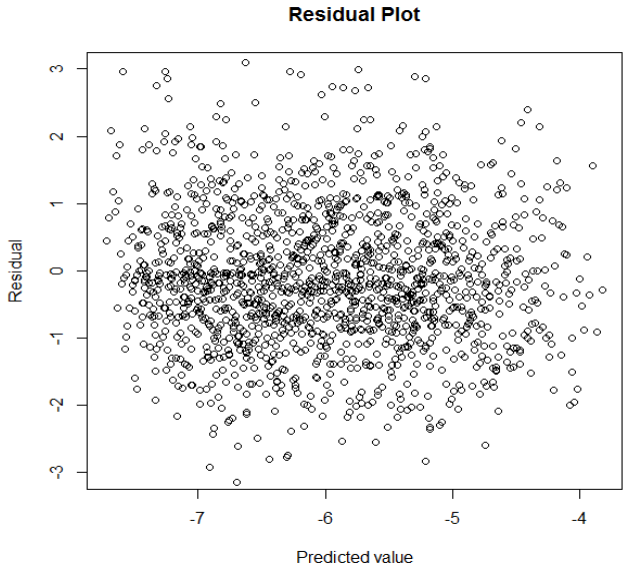


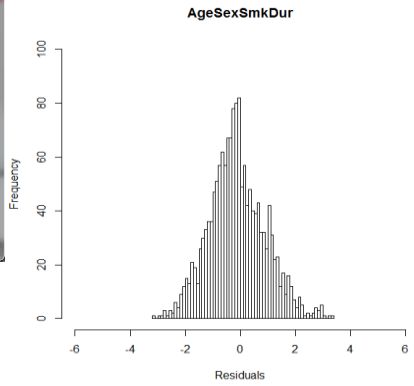
Model – Investigation 1 : Age * Sex * Smk * Dur

R Console (64-bit)

File Edit Misc Packages Windows Help

```
Call:
glm(formula = MtxG1$DTHS ~ MtxG1$AGE + MtxG1$SEX + MtxG1$SMOKER +
    MtxG1$DUR + MtxG1$AGE:MtxG1$SMOKER + MtxG1$SEX:MtxG1$SMOKER +
    MtxG1$AGE:MtxG1$DUR + MtxG1$SEX:MtxG1$DUR + MtxG1$SMOKER:MtxG1$DUR +
    MtxG1$AGE:MtxG1$SEX:MtxG1$DUR + offset(log(MtxG1$EXPO)),
    family = poisson)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.1206  -0.7790  -0.1857   0.5222   3.3622

Coefficients:
                                Estimate Std. Error z value Pr(>|z|)
(Intercept)                   -9.3814148  0.1056322 -88.812  < 2e-16 ***
MtxG1$AGE                      0.0731345  0.0025173  29.053  < 2e-16 ***
MtxG1$SEXM                    -0.1612144  0.0405479  -3.976 7.01e-05 ***
MtxG1$SMOKERS                 -0.4391762  0.1516075  -2.897 0.003770 **
MtxG1$DUR                      0.1070991  0.0291630   3.672 0.000240 ***
MtxG1$AGE:MtxG1$SMOKERS        0.0190225  0.0035625   5.340 9.31e-08 ***
MtxG1$SEXM:MtxG1$SMOKERS       0.2283091  0.0543478   4.201 2.66e-05 ***
MtxG1$AGE:MtxG1$DUR           -0.0020984  0.0006679  -3.142 0.001679 **
MtxG1$SEXM:MtxG1$DUR          -0.1150113  0.0316005  -3.640 0.000273 ***
MtxG1$SMOKERS:MtxG1$DUR       -0.0111854  0.0107767  -1.038 0.299307
MtxG1$AGE:MtxG1$SEXM:MtxG1$DUR  0.0030810  0.0006744   4.568 4.92e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 4799.9  on 1569  degrees of freedom
Residual deviance: 1634.1  on 1559  degrees of freedom
AIC: 5769

Number of Fisher Scoring iterations: 5
```



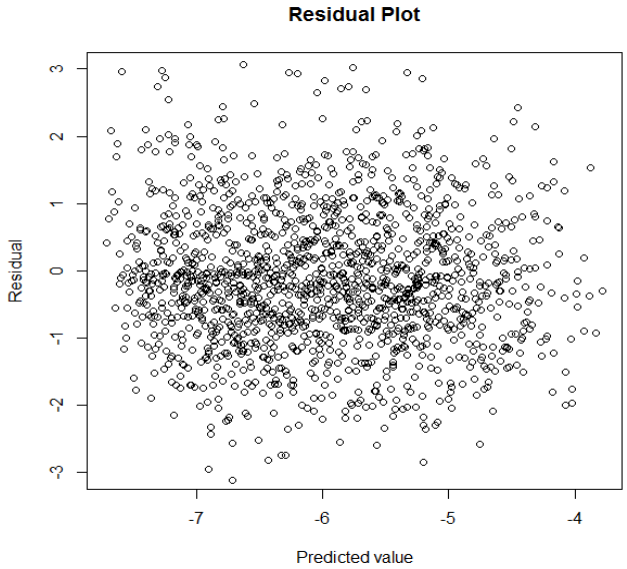


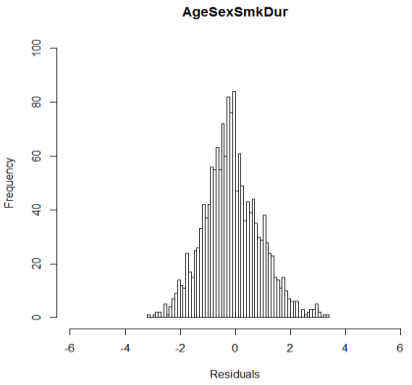
Model – Investigation 1 : Age * Sex * Smk * Dur

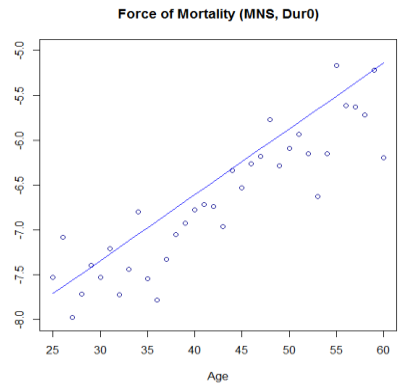
R Console (64-bit)

File Edit Misc Packages Windows Help

```
Call:
glm(formula = MtxG1$DTHS ~ MtxG1$AGE + MtxG1$SEX + MtxG1$SMOKER +
    MtxG1$DUR + MtxG1$AGE:MtxG1$SMOKER + MtxG1$SEX:MtxG1$SMOKER +
    MtxG1$AGE:MtxG1$DUR + MtxG1$SEX:MtxG1$DUR + MtxG1$AGE:MtxG1$SEX:MtxG1$DUR +
    offset(log(MtxG1$EXPO)), family = poisson)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.0928  -0.7835  -0.1818   0.5167   3.3781

Coefficients:
                              Estimate Std. Error z value Pr(>|z|)
(Intercept)                 -9.3803455  0.1053819 -89.013  < 2e-16 ***
MtxG1$AGE                    0.0732773  0.0025074  29.225  < 2e-16 ***
MtxG1$SEXM                  -0.1583553  0.0403594  -3.924 8.72e-05 ***
MtxG1$SMOKERS               -0.4465340  0.1515870  -2.946 0.003222 **
MtxG1$DUR                    0.1052758  0.0291085   3.617 0.000298 ***
MtxG1$AGE:MtxG1$SMOKERS      0.0181970  0.0034773   5.233 1.67e-07 ***
MtxG1$SEXM:MtxG1$SMOKERS     0.2281312  0.0543611   4.197 2.71e-05 ***
MtxG1$AGE:MtxG1$DUR         -0.0021005  0.0006678  -3.145 0.001658 **
MtxG1$SEXM:MtxG1$DUR        -0.1166803  0.0315617  -3.697 0.000218 ***
MtxG1$AGE:MtxG1$SEXM:MtxG1$DUR  0.0031008  0.0006742   4.599 4.24e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 4799.9  on 1569  degrees of freedom
Residual deviance: 1635.1  on 1560  degrees of freedom
AIC: 5768.1

Number of Fisher Scoring iterations: 5
```







Model – Investigation 1

- So the simplified model is:

$$\mu(x, y, \text{Sex}, \text{Smoker}) =$$

$$\exp\left(\begin{array}{l} -9.3803455 + 0.0732773x - 0.1583553\chi_{\text{MALE}} \\ -0.4465340\chi_{\text{SMOKER}} + 0.1052758y \\ +0.0181970x\chi_{\text{SMOKER}} + 0.2281312\chi_{\text{MALE}}\chi_{\text{SMOKER}} \\ -0.0021005xy - 0.1166803y\chi_{\text{MALE}} \\ +0.0031008xy\chi_{\text{MALE}} \end{array}\right)$$

- This can be easily programmed into Excel or R.

Model – Investigation 2

- Take a grouped dataset split by
  - Individual Age
  - Sex
  - Smoker
  - Duration

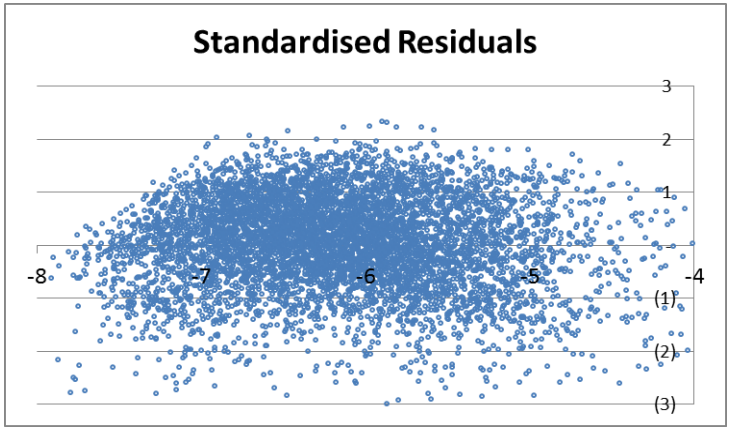
  and SA Band, Socioeconomic, Rated, Guaranteed/Reviewable, JL Status, Product Type.
- Fit the following model :
  AGE * SEX * (SMK + DUR) + SA + SE + R + GR + JS + Prod
- Use Exposure as the “offset”
- Analyse the significance of each factor and simplify

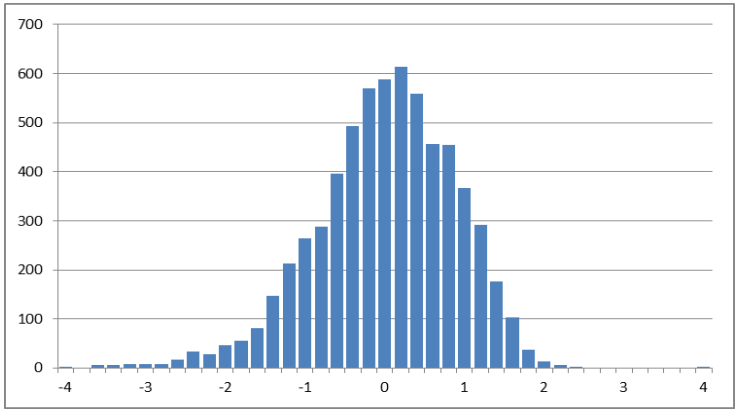
Model - Investigation 2 : AGE * SEX * (SMK + DUR) + SA + SE + R + GR + JS + Prod

R Console (64-bit)

File Edit Misc Packages Windows Help

```
    family = poisson)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.0382  -0.2892  -0.1459  -0.0588   7.8399

Coefficients:
                                  Estimate Std. Error z value Pr(>|z|)
(Intercept)                     -9.4858960  0.1695209 -55.957  < 2e-16 ***
MtxG1$AGE                        0.0723573  0.0040066  18.059  < 2e-16 ***
MtxG1$SEXM                      -0.2386803  0.2307640  -1.034  0.30099
MtxG1$SMOKERS                   -0.3742452  0.2271617  -1.647  0.09946 .
MtxG1$DUR                        0.1326609  0.0456577   2.906  0.00367 **
MtxG1$RR                         0.1211959  0.0300509   4.033 5.51e-05 ***
MtxG1$SJJ                       -0.0329484  0.0228954  -1.439  0.15013
MtxG1$SEU                        0.2264518  0.0371947   6.088 1.14e-09 ***
MtxG1$SEII                      -0.0062620  0.0336810  -0.186  0.85251
MtxG1$SEIII                      0.1514108  0.0313282   4.833 1.34e-06 ***
MtxG1$SEIV/V                     0.3947074  0.0344283  11.465  < 2e-16 ***
MtxG1$SASA_M                     0.0593600  0.0277533   2.139  0.03245 *
MtxG1$SASA_H                     0.0063091  0.0303842   0.208  0.83551
MtxG1$GRR                       -0.0531548  0.0291033  -1.826  0.06779 .
MtxG1$PRODFIB                   -0.3444491  0.2138521  -1.611  0.10725
MtxG1$PRODITA                   -0.3232213  0.1833071  -1.763  0.07785 .
MtxG1$AGE:MtxG1$SEXM             0.0005614  0.0055523   0.101  0.91946
MtxG1$AGE:MtxG1$SMOKERS          0.0162903  0.0053314   3.056  0.00225 **
MtxG1$SEXM:MtxG1$SMOKERS         0.0895085  0.3027016   0.296  0.76746
MtxG1$AGE:MtxG1$DUR             -0.0022762  0.0010850  -2.098  0.03592 *
MtxG1$SEXM:MtxG1$DUR            -0.1296263  0.0642486  -2.018  0.04364 *
MtxG1$AGE:MtxG1$SEXM:MtxG1$SMOKERS  0.0026340  0.0070385   0.374  0.70823
MtxG1$AGE:MtxG1$SEXM:MtxG1$DUR   0.0034718  0.0015040   2.308  0.02098 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 31316  on 119757  degrees of freedom
Residual deviance: 27945  on 119735  degrees of freedom
AIC: 41660
```



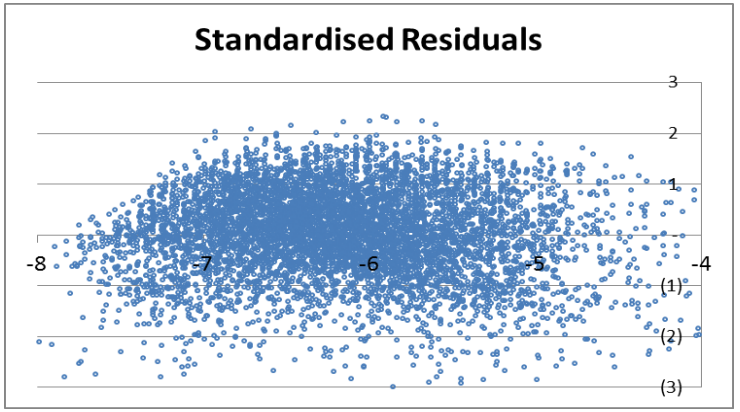


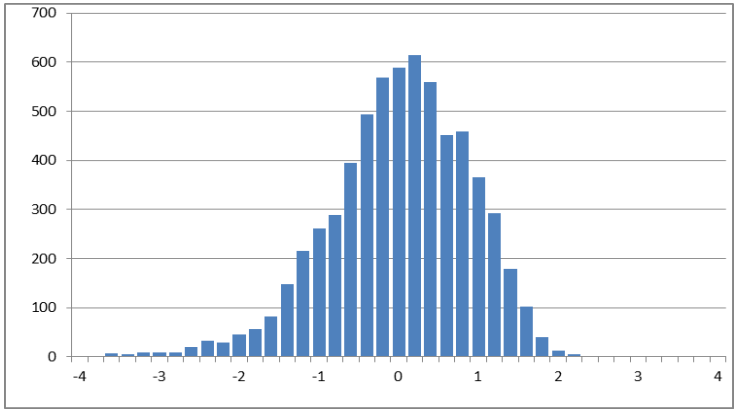
Model - Investigation 2 : AGE * SEX * (SMK + DUR) + SA + SE + R + GR + JS + Prod

R Console (64-bit)

File Edit Misc Packages Windows Help

```
Call:
glm(formula = MtxG1$DTHS ~ MtxG1$AGE + MtxG1$SEX + MtxG1$SMOKER +
    MtxG1$DUR + MtxG1$R + MtxG1$SE + MtxG1$GR + MtxG1$PROD +
    MtxG1$AGE:MtxG1$SMOKER + MtxG1$SEX:MtxG1$SMOKER + MtxG1$AGE:MtxG1$DUR +
    MtxG1$SEX:MtxG1$DUR + MtxG1$AGE:MtxG1$SEX:MtxG1$DUR + offset(log(MtxG1$EXPO)),
    family = poisson)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.0924  -0.2886  -0.1467  -0.0617   7.8488

Coefficients:
                                Estimate Std. Error z value Pr(>|z|)
(Intercept)                   -9.4938165  0.1170397 -81.116  < 2e-16 ***
MtxG1$AGE                      0.0727104  0.0027804  26.151  < 2e-16 ***
MtxG1$SEXM                    -0.2180423  0.0447562  -4.872 1.11e-06 ***
MtxG1$SMOKERS                 -0.4180789  0.1517190  -2.756 0.005858 **
MtxG1$DUR                      0.1386951  0.0371898   3.729 0.000192 ***
MtxG1$RR                       0.1171921  0.0298833   3.922 8.79e-05 ***
MtxG1$SEU                      0.2305645  0.0354618   6.502 7.94e-11 ***
MtxG1$SEIII                    0.1536476  0.0293628   5.233 1.67e-07 ***
MtxG1$SEIV/V                   0.3959840  0.0324140  12.216  < 2e-16 ***
MtxG1$GRR                     -0.0562391  0.0289695  -1.941 0.052220 .
MtxG1$PRODITA                 -0.3117066  0.1830479  -1.703 0.088593 .
MtxG1$AGE:MtxG1$SMOKERS        0.0173532  0.0034792   4.988 6.11e-07 ***
MtxG1$SEXM:MtxG1$SMOKERS       0.1997629  0.0544142   3.671 0.000241 ***
MtxG1$AGE:MtxG1$DUR           -0.0024391  0.0008646  -2.821 0.004784 **
MtxG1$SEXM:MtxG1$DUR          -0.1385034  0.0377147  -3.672 0.000240 ***
MtxG1$AGE:MtxG1$SEXM:MtxG1$DUR  0.0036985  0.0008072   4.582 4.60e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 31316  on 119757  degrees of freedom
Residual deviance: 27956  on 119742  degrees of freedom
AIC: 41658

Number of Fisher Scoring iterations: 8
```





Model – Investigation 2

- We can see there is a danger of overfitting the model
- The saturated model is:
  AGE * SEX * SMK * DUR * SA * SE * R * GR * JS * Prod
- This would involve:
  36 * 2 * 2 * 6 * 3 * 4 * 2 * 2 * 2 * 3
  = **248,832**
  data cells
- As a rule of thumb you need 3 times as many data points in your dataset, i.e. ~ 750,000
- So we need to simplify the model

Model – Investigation 3

- We can reduce the number of cells by grouping data
- Age seems like an obvious choice
- But does Dths ~ Exposure * q_{Age} make sense if age is grouped ?
- Easier to rationalise if we switch to using Dths ~ Expected

Model – Investigation 3

- Take a grouped dataset split by
  - 10 Year Age Band
  - Sex
  - Smoker
  - Duration

  and SA Band, Socio-economic, Rated, Guaranteed/Reviewable, Calendar Year, JL Status, Product Type.
- Fit the following model :
  AGE * SEX * (SMK + DUR) + SA + SE + R + GR + JS + Prod
- Use Expected as the “offset”
- Analyse the significance of each factor and simplify

Model – Investigation 3 : AGE * SEX * (SMK + DUR)

R Console (64-bit)

File Edit Misc Packages Windows Help

```
Call:
glm(formula = MtxG1$DTHS ~ offset(log(MtxG1$EXPE)) + MtxG1$AGE *
    MtxG1$SEX * MtxG1$SMOKER + MtxG1$AGE * MtxG1$SEX * MtxG1$DUR,
    family = quasipoisson)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.0031  -0.8308  -0.1527   0.4288   2.2771

Coefficients:
                                    Estimate Std. Error t value Pr(>|t|)
(Intercept)                        0.4731090  0.1413422   3.347 0.000969 ***
MtxG1$AGE                         -0.0096566  0.0033933  -2.846 0.004876 **
MtxG1$SEXM                         0.1012560  0.1949782   0.519 0.604093
MtxG1$SMOKERS                     -0.0831783  0.2147863  -0.387 0.698961
MtxG1$DUR                          0.0140704  0.0335575   0.419 0.675488
MtxG1$AGE:MtxG1$SEXM              -0.0044336  0.0046050  -0.963 0.336774
MtxG1$AGE:MtxG1$SMOKERS            0.0011954  0.0049629   0.241 0.809893
MtxG1$SEXM:MtxG1$SMOKERS           0.0212031  0.2829904   0.075 0.940346
MtxG1$AGE:MtxG1$DUR               -0.0004490  0.0007749  -0.579 0.562947
MtxG1$SEXM:MtxG1$DUR              -0.0068576  0.0459118  -0.149 0.881412
MtxG1$AGE:MtxG1$SEXM:MtxG1$SMOKERS -0.0005128  0.0064943  -0.079 0.937134
MtxG1$AGE:MtxG1$SEXM:MtxG1$DUR     0.0006037  0.0010415   0.580 0.562809
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for quasipoisson family taken to be 0.9713166)

    Null deviance: 309.33  on 218  degrees of freedom
Residual deviance: 214.25  on 207  degrees of freedom
AIC: NA
```

Now that we are targeting A/E, it is no surprise that most factors are no longer significant – showing that the expected basis (WP50) is a pretty good fit. Note the residual on age though ! Compare this to the output from the first model

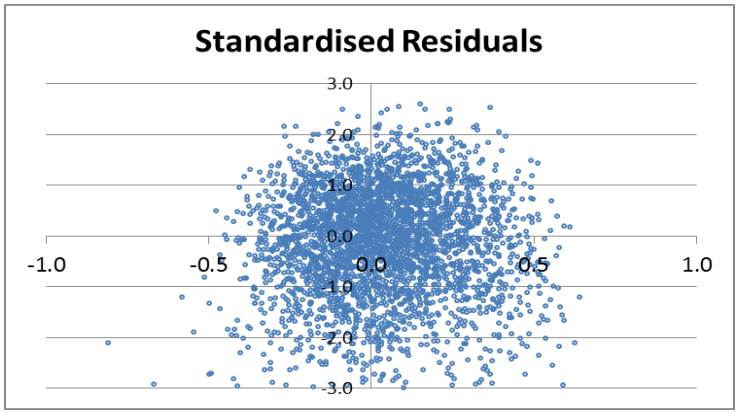
```
Coefficients:
                                    Estimate Std. Error t value Pr(>|t|)
(Intercept)                       -9.3705317  0.1534948 -61.048  < 2e-16 ***
MtxG1$AGE                          0.0730301  0.0037253  19.604  < 2e-16 ***
MtxG1$SEXM                        -0.1767165  0.2178201  -0.811  0.41732
MtxG1$SMOKERS                     -0.3900498  0.2353573  -1.657  0.09767 .
MtxG1$DUR                          0.1013510  0.0362272   2.798  0.00521 **
MtxG1$AGE:MtxG1$SEXM               0.0004614  0.0051929   0.089  0.92921
MtxG1$AGE:MtxG1$SMOKERS            0.0168522  0.0055251   3.050  0.00233 **
MtxG1$SEXM:MtxG1$SMOKERS           0.1312389  0.3133699   0.419  0.67542
MtxG1$AGE:MtxG1$DUR               -0.0020052  0.0008480  -2.365  0.01817 *
MtxG1$SEXM:MtxG1$DUR              -0.1093163  0.0505352  -2.163  0.03068 *
MtxG1$AGE:MtxG1$SEXM:MtxG1$SMOKERS  0.0022906  0.0072874   0.314  0.75332
MtxG1$AGE:MtxG1$SEXM:MtxG1$DUR      0.0029241  0.0011600   2.521  0.01181 *
```

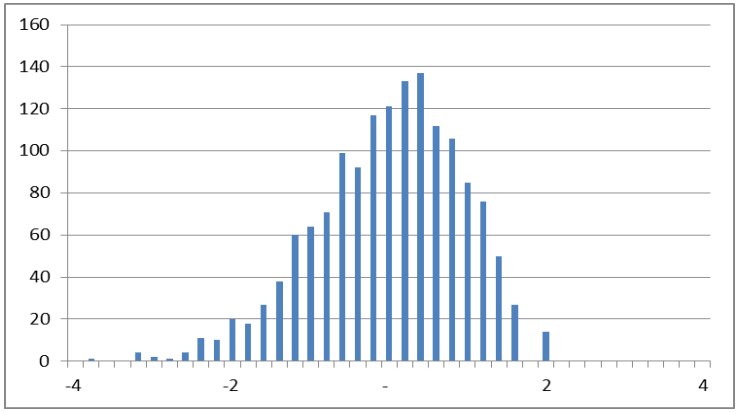
Model – Investigation 3 : AGE * SEX * (SMK + DUR) + SA + SE + R + GR + JS + Prod

R Console (64-bit)

File Edit Misc Packages Windows Help

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-4.2915  -0.4225  -0.1507  -0.0519   6.9821

Coefficients:
                                  Estimate Std. Error z value Pr(>|z|)
(Intercept)                      0.3256178  0.1639723   1.986 0.047055 *
MtxG1$AGE                       -0.0084515  0.0038335  -2.205 0.027480 *
MtxG1$SEXM                       0.0289411  0.2168156   0.133 0.893812
MtxG1$SMOKERS                   -0.0656133  0.2182195  -0.301 0.763662
MtxG1$DUR                        0.0266290  0.0445121   0.598 0.549678
MtxG1$RR                         0.0860224  0.0300376   2.864 0.004186 **
MtxG1$SJJ                       -0.0057765  0.0228357  -0.253 0.800299
MtxG1$SEU                        0.1724367  0.0370997   4.648 3.35e-06 ***
MtxG1$SEII                      -0.0332642  0.0336658  -0.988 0.323118
MtxG1$SEIII                      0.1247658  0.0312949   3.987 6.70e-05 ***
MtxG1$SEIV/V                     0.3694412  0.0343343  10.760  < 2e-16 ***
MtxG1$SASA_M                     0.0926463  0.0276362   3.352 0.000801 ***
MtxG1$SASA_H                     0.0341529  0.0301768   1.132 0.257735
MtxG1$GRR                       -0.1010563  0.0290063  -3.484 0.000494 ***
MtxG1$PRODFIB                   -0.3702445  0.2139398  -1.731 0.083523 .
MtxG1$PRODITA                   -0.3720952  0.1833525  -2.029 0.042418 *
MtxG1$AGE:MtxG1$SEXM            -0.0042319  0.0051668  -0.819 0.412757
MtxG1$AGE:MtxG1$SMOKERS          0.0008040  0.0050413   0.159 0.873287
MtxG1$SEXM:MtxG1$SMOKERS        -0.0172922  0.2875113  -0.060 0.952041
MtxG1$AGE:MtxG1$DUR             -0.0007419  0.0010444  -0.710 0.477494
MtxG1$SEXM:MtxG1$DUR            -0.0054936  0.0610872  -0.090 0.928343
MtxG1$AGE:MtxG1$SEXM:MtxG1$SMOKERS -0.0002817  0.0065977  -0.043 0.965949
MtxG1$AGE:MtxG1$SEXM:MtxG1$DUR   0.0007075  0.0014133   0.501 0.616640
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 8968.4  on 19486  degrees of freedom
Residual deviance: 8687.3  on 19464  degrees of freedom
AIC: 16907

Number of Fisher Scoring iterations: 7
```



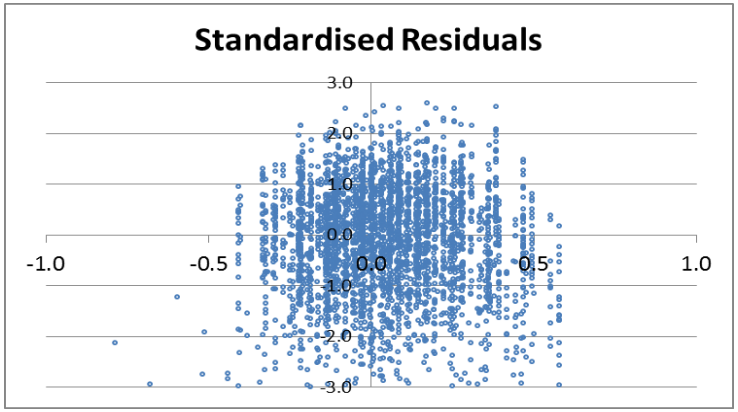


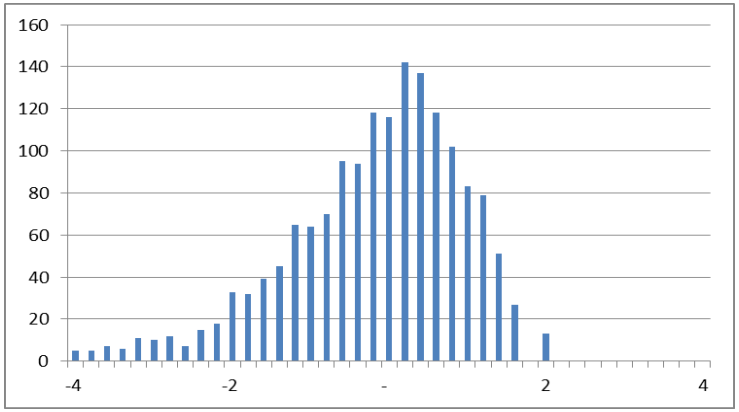
Model – Investigation 3 : AGE * SEX * (SMK + DUR) + SA + SE + R + GR + JS + Prod

R Console (64-bit)

File Edit Misc Packages Windows Help

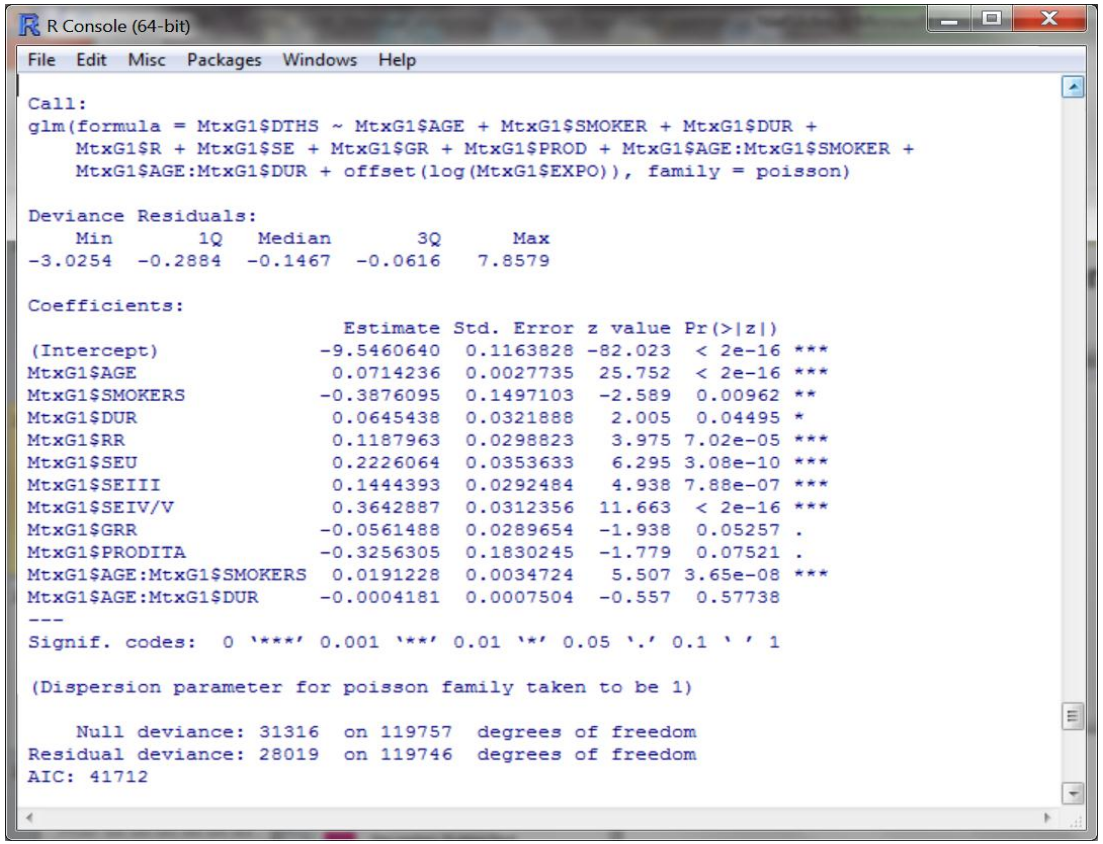
```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)    0.391916   0.066963   5.853 4.84e-09 ***
MtxG1$AGE     -0.010929   0.001353  -8.075 6.75e-16 ***
MtxG1$SEXM    -0.084191   0.022818  -3.690 0.000225 ***
MtxG1$RR       0.083724   0.029788   2.811 0.004944 **
MtxG1$SEU      0.169839   0.034307   4.951 7.40e-07 ***
MtxG1$SEIII    0.138432   0.029369   4.714 2.43e-06 ***
MtxG1$SEIV/V   0.383404   0.032363  11.847  < 2e-16 ***
MtxG1$SASA_M/H 0.076214   0.025269   3.016 0.002560 **
MtxG1$GRR     -0.115718   0.027393  -4.224 2.40e-05 ***
MtxG1$PRODFIB -0.378941   0.213800  -1.772 0.076327 .
MtxG1$PRODITA -0.381191   0.182912  -2.084 0.037159 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 8968.4  on 19486  degrees of freedom
Residual deviance: 8702.0  on 19476  degrees of freedom
AIC: 16898
```





What happens if we remove gender from Model 2?

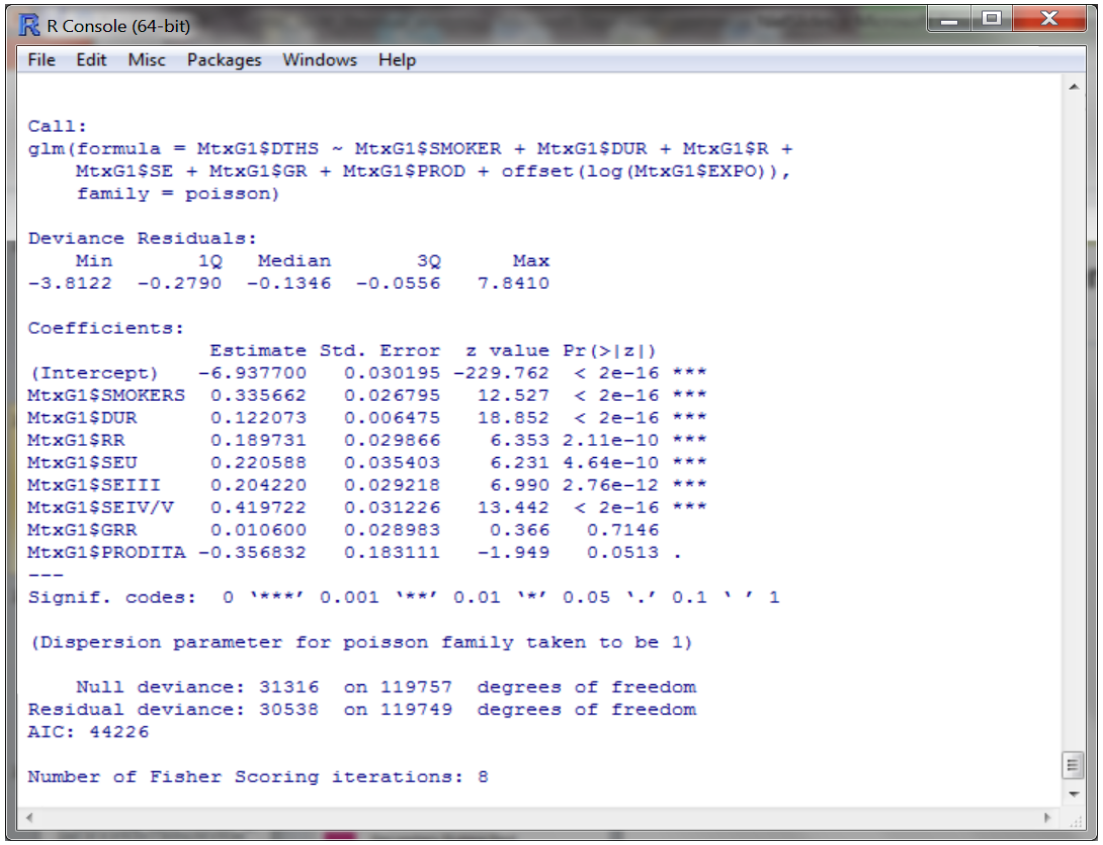
What happens if we remove gender from Model 2?



	AIC	Residual deviance
Gender/Age	41,658	27,956
Age	41,712	28,019

So losing gender as a rating factor only reduces the predictive power of the model slightly – and we can mitigate this to some degree through use of wider rating factors

What happens if we remove both gender and age from Model 2?



	AIC	Residual deviance
Gender/Age	41,658	27,956
Age	41,712	28,019
Neither	44,296	30,538

But loss of age as a rating factor significantly worsens our ability to predict the claim rate

Removing gender & age from Model 2

		30 MNS	40 MNS	50 MNS	30 FNS	40 FNS	50 FNS
Derived risk rate	Age & Sex	0.000717	0.001897	0.005021	0.000833	0.001885	0.004265
	Age only	0.000781	0.001932	0.004779	0.000781	0.001932	0.004779
	neither	0.002735	0.002735	0.002735	0.002735	0.002735	0.002735
Error cf age & sex rate	Age & Sex						
	Age only	9%	2%	-5%	-6%	2%	12%
	neither	282%	44%	-46%	228%	45%	-36%

So removing gender creates a small risk (which we can further mitigate by understanding how gender mix varies by key rating factors).

Removing age is a key risk !

Summary

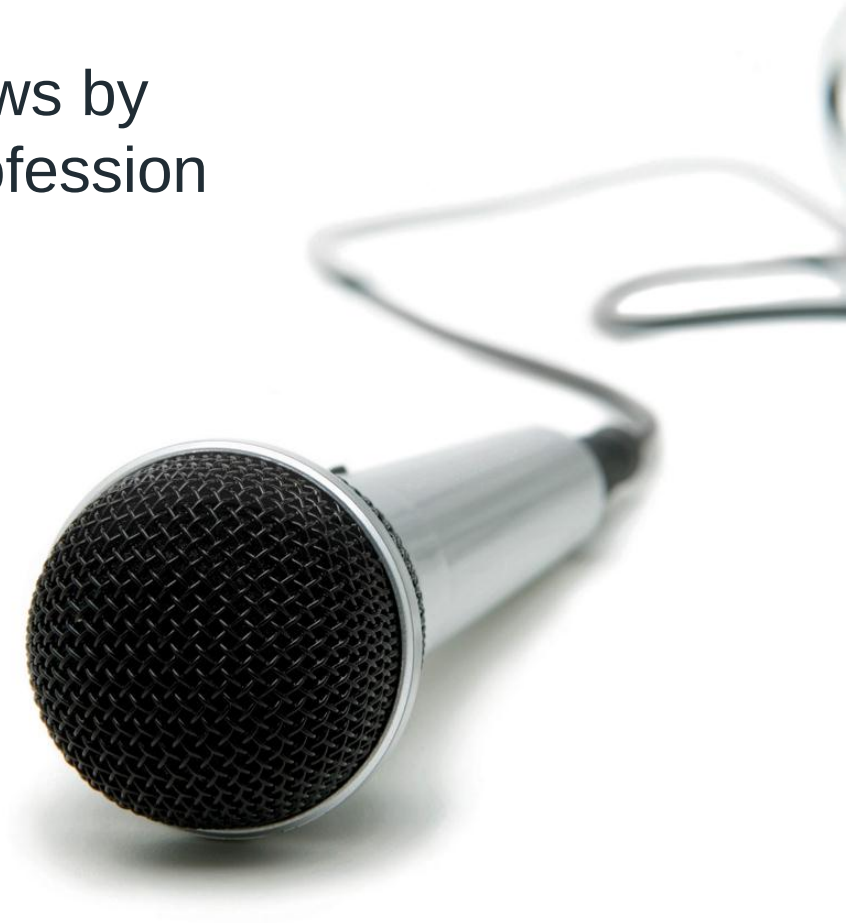
- Multiple regression is difficult, time consuming and always vulnerable to subjective decisions about what to include and what to leave out
- But decisions are **always** subjective to some extent. No system can do all the work in deciding what is the best model
- An overly complex model will lead to spurious results. Models can only be as complex as the data volume allows
- R allows us to get into GLMs quite quickly - so it is easy to get stuck in and start to learn more about your rating factors
- But, you'll need some statistics knowledge to interpret your results

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)

Questions or comments?

- Expressions of individual views by members of the Actuarial Profession and its staff are encouraged.
- The views expressed in this presentation are those of the presenter.



The Actuarial Profession
making financial sense of the future

Health and Care Conference 2012

Chris Reynolds, PartnerRe
Niel Daniels, Daniels Actuarial Consulting

C4 : How Powerful are Your Rating Factors?

1 May 2012