A NEW MATHEMATICAL MODEL OF AUSTRALIAN DISABILITY EXPERIENCE

David Service and David Pitt

ABSTRACT

In Service & Pitt (2002) the authors presented a Generalised Linear Model of disability income claim termination rates based on the data from 1980 to 1998. The model was simplified for the purposes of the paper and was intended to demonstrate the possibilities of using such an approach in the preparation of “standard” tables.

In this paper, the authors present a full scale GLM of incidence rates using the 1995 to 1998 data. All available characteristics present in the data are included. A comparison of the GLM with IAD89-93 shows the difference in experience between the two data periods. The significance of the individual characteristics shows the variation in the model according to the characteristics which IAD89-93 does not include.

Overall the model shows a goodness of fit at a level with significance greater than 95%.

Keywords: Disability Insurance, disability experience, generalised linear models

David Service FIA, FIAA
Centre for Actuarial Research
Australian National University
david.service@anu.edu.au

David Pitt BEc, BSc, FIAA
Centre for Actuarial Research
Australian National University
david.pitt@anu.edu.au

1. INTRODUCTION

In Service & Pitt (2002) we presented a simplified Generalised Linear Model ("GLM") of disability income claim termination rates based on experience data from 1980 to 1998. That model was intended to demonstrate the possibilities of using such an approach in the preparation of "standard" tables.

In this paper, we present a full scale GLM of incidence rates using the 1995 to 1998 data.

The paper is set out in the following sections

1. Introduction
    This introduction
2. Why A GLM?
    A discussion of the reasons why a GLM is ideal for modelling disability experience
3. Data
    Description of the data used in developing the GLM
4. The Incidence Model
    The GLM for incidence rates
5. Comparison with IAD89-93
    A comparison of the rates from the model and IAD89-93
6. Goodness of Fit Analysis
    The results of the tests for goodness of fit
7. Conclusions
    The conclusions drawn from the analyses of the paper

Bibliography

Appendix A
    S Plus Output for the GLM

Appendix B
    The encoding scheme for each characteristic

Acknowledgements

The data on which this paper is based was provided to the authors by the Disability Committee of the Institute of Actuaries of Australia. The extensive assistance provided by that Committee is gratefully acknowledged.

Michael Martin and Steven Stern, from the School of Finance and Applied Statistics at the Australian National University, provided valuable statistical advice during the preparation of this paper.

2. WHY A GLM?

Models of disability experience have, for many years, been mathematical in form. One of the earliest was Miller & Courant (1974). Since then most published standard tables have been expressed using mathematical forms, at least in part, rather than just raw data which has been smoothed without the use of explicit mathematical formulae.

IAD89-93 ("IAD") is no exception. It is partially in mathematical form. It provides incidence rates differentiated by gender, occupation, deferment period and smoker status and claim termination rates differentiated by gender, occupation, deferment period and duration.

However, these characteristics are only a subset of the characteristics which are expected to impact on claims incidence rates. In Service & Pitt (2002) the large differences between claim termination experience according to various characteristics were clearly shown. Indeed the differences for some characteristics not included in IAD were greater than those which are included.

The similarly large differences for incidence rates are apparent from examining the form of the final GLM presented in Section 4 of this paper. The characteristics in IAD are not sufficient to capture all the expected variation in experience differentiated by appropriate characteristics. This omission of potentially important characteristics from the calculation of expected claims can lead to incorrect conclusions being drawn as to the reasons for differences between actual and expected. It may not be that the experience has worsened (or improved??) but that the business mix has changed when measured by the characteristics not included in IAD (or most other "standard" tables). A more comprehensive measure of expected may capture these changes in business mix. An inadequate classification of the risk factors that describe disability income insurance can also lead to adverse selection for the office.

For many years the Disability Committee reports have shown material differences between the experience of individual companies. It may be that these differences have more to do with different business mixes than to different underwriting and claims business practices.

Using a GLM for the standard table enables all the crucial characteristics to be included and the difficulties just noted will be of much lower impact. It is true, of course, that the effect of a particular characteristic may change over time. However, by refitting the GLM such changes will be readily apparent.

The prime advantage of using a GLM as the model for expected claims is its ability to capture the combined influence of all the characteristics which impact on the claims behaviour of the business. We are left with much less guessing about the reasons for deviations in actual experience from expected. We have a more reliable model of expected incidence rates based on the particular characteristics of the block of business under consideration.

In addition, a GLM ensures that the predicted incidence rates will be "smooth" since they are the output of a mathematical function and, finally, that the resulting standard table is a single model rather than a big range of different tables.

This last point makes the implementation of a GLM in the modelling software potentially simpler and easier to update. Since the expected claims calculation is a single mathematical formula it can be incorporated as one subroutine with all potential characteristics as inputs even if all the characteristics are not necessarily used. This facilitates updating as only the code internal to the single subroutine needs changing.

3. DATA

We were provided with the individual data records on which the Disability Committee's 2000 Report was based. While each data record corresponds to a single policy any data which could identify individual companies or policies had been stripped before the dataset was given to us. As a result the summary of the data is identical to that provided in the Committee's report.

The characteristics contained in the individual records and the characteristics which were included in the GLM are set out in the following table. A complete description of the way in which each characteristic was encoded for the GLM is contained in Appendix B.

Characteristic	Included in GLM
Gender	YES
Occupation	YES
Deferment	YES
Smoker Status	YES
Disability Definition	YES
Age	YES
Date of Entry	NO
Policy Expiry Age	NO
Benefit Period	YES
Benefit Type	YES
Medical Evidence	YES
Contract Type	YES
No Claim Bonus	YES
AIDS Exclusion	YES
Benefit Insured	YES
Duration	YES

The base data selection criteria were the same as adopted by the Disability Committee i.e. only Australian business, only non-cancellable policies and only lives insured under individual policies.

Of the available characteristics, two were excluded in the derivation of the GLM. While it is clearly desirable to include as many explanatory variables as possible, there is also a need to reduce the number of possible cells in order that sufficient data remains in each cell to allow the fitting of the GLM to proceed adequately. For this reason characteristics were excluded where it was reasonable to assume that the information they represented was already covered by another characteristic or where there was a reasonable expectation that the characteristic was unlikely to have a material impact on claims experience.

Date of Entry was, given that only four years of data were being aggregated, regarded as adequately covered by Duration. Policy Expiry Age was regarded both as unlikely to have a material impact on claims and also, were there to be an impact, to be covered by the inclusion of Benefit Period.

In some cases the range of values recorded in the data was reduced by combining some individual values of the characteristic. These cases are detailed in Appendix B.

Because the processing of the individual data records used the Disability Committee's new software system – IDEAS - the calculation of the exposure and the definition of new claims have changed marginally from those used to prepare the Committee's 2000 Report. The total exposure used for this paper was 434,423,997 days or 1,189,388 years compared to

1,153,816 years shown in Table B of the Committee's report and the number of new claims, before any adjustment for partial benefits, was 30,784 compared to 30,437 in Table A of the Committee's report. These differences are not regarded as material in terms of fitting the GLM.

4. THE INCIDENCE MODEL

A GLM with Poisson error and logarithmic link function was fitted to the claim incidence rate data. See Service and Pitt (2002) for a description use that actuaries have made of GLMs over the past twenty years. The final model chosen is of the form

$$E\left[\log\left(\mathrm{Y}\right)\right] = \log(\text{exposure}) + \beta_0 + \sum_{j=1}^{35} \beta_j x_j + \sum_{\text{selected } i,j;k=1}^{13} \beta_k x_i x_j$$

where Y denotes actual number of claims multiplied by benefit percent, the x_j are covariates and the β_j are regression coefficients derived using maximum likelihood estimation (McCullagh and Nelder (1989)). The exposure is the amount of time that disability income insurance holders are exposed to the risk of becoming disabled and is recorded in days. The $x_i x_j$ terms are interaction terms. These will be discussed further in this section.

This model is fit using an offset for log(exposure). The offset ensures that the model is fit with a regression coefficient for log(exposure) equal to one. This means that the model can be rewritten as

$$E\left(\log\left(\text{IncRate}\right)\right) = \beta_0 + \sum_{j=1}^{35} \beta_j x_j + \sum_{\text{selected } i,j;k=1}^{13} \beta_k x_i x_j$$

where IncRate denotes the incidence rate of claims measured on a per day basis.

The coefficients of the fitted model for each variable and for the statistically significant interaction terms are shown in Table 4.1. The output generated by the S-Plus statistical package for the model is shown in Appendix A. The *t*-ratios used to test for statistical significance of each coefficient estimate is also included in the table. A description of each covariate is given in Appendix B. The fitted intercept for the model is –5.281746.

Table 4.1 GLM Coefficients and Significance

Covariate	Coefficient Estimate	Significance	Covariate	Coefficient Estimate	Significance
Gender	0.274745	2.93	Benamount6	0.211641	6.10
Age	0.246015	15.81	Benamount7	0.254199	5.52
$\sqrt{\text{Age}}$	-2.310697	-18.63	Smoker	0.156685	10.61
OccupationB	0.921944	7.49	Aids	0.222879	15.04
OccupationC	1.805061	22.77	Duration1	0.207573	12.56
OccupationD	2.000999	23.57	Duration2	0.083477	5.89
Definition2	-0.058871	-4.03	Ncb	0.067367	4.69
Definition3	-0.043319	-0.89	Contract1	-0.321532	-4.76
Definition4	-0.219713	-4.58	Contract2	-0.078040	-3.99
Definition5	-0.356431	-8.86	Contract3	0.149045	3.06
Definition6	-1.158268	-1.46	Medical	-0.205617	-6.11
Deferment2	0.906221	1.20	Age*OccupationB	-0.009219	-3.38
Deferment3	-0.048222	-0.06	Age*OccupationC	-0.019752	-11.39
Deferment4	-2.540355	-3.09	Age*OccupationD	-0.021182	-11.42
Deferment5	-2.612055	-3.20	Gender*Age	0.004870	2.40
Deferment6	-4.371066	-3.14	Gender*OccupationB	-0.175858	-2.95
Deferment7	-3.533438	-1.52	Gender*OccupationC	-0.474074	-9.73
Deferment8	-3.333802	-6.72	Gender*OccupationD	-0.563529	-7.08
Benperiod1	-1.078854	-13.62	Age*Deferment2	-0.025920	-1.96
Benperiod2	0.0196288	1.39	Age*Deferment3	-0.023097	-1.75
Benamount2	0.044054	2.20	Age*Deferment4	0.011881	0.80
Benamount3	0.174378	9.06	Age*Deferment5	0.002877	0.20
Benamount4	0.200430	9.88	Age*Deferment6	0.027720	1.08
Benamount5	0.235900	8.73	Age*Deferment7	-0.011860	-0.25

The above table shows that thirteen interaction terms are included in the model for claim incidence rates. Interaction terms add to the flexibility of the model and enable a more realistic description of the underlying data. To consider an example we see in the above table that the gender*age interaction is statistically significant. This means that the impact of age on incidence rates is dependent on the value taken by gender in the model. The gender by age interaction term has a positive coefficient meaning that for a unit increase in age the resulting predicted incidence rate increases more when gender is 1 (female) than when gender is 0 (male).

The Poisson error structure employed in this model means that the variance increases in line with the size of the fitted value. A check was placed on the model to ensure that this proportional increase in variance with the fitted values was appropriate for this data. The test indicated that there was no overdispersion.

Another important characteristic of this analysis was the treatment of the smoker variable. In the original data, approximately 3.5% of the exposure was not classified as either smoker or non-smoker. One alternative in the model fitting procedure was to ignore this 3.5% of the data. It is not possible to fit the GLM using a subset of the required covariates for certain data points. An alternative method is to attempt to predict the value of the missing smoker value by employing logistic regression. This analysis fits a regression model where the response variable is the probability that the policyholder was a smoker and the explanatory variables are the other covariates including claim information available in the data. The logistic regression methodology ensures that the fitted values for the probability of being a smoker are on the range 0 to 1. These fitted probabilities are then used to fill the gaps where the data on smoker status was not available.

5. COMPARISON WITH IAD89-93

An analysis of the main differences between the incidence rates from the GLM and the existing rates from IAD89-93 is given in this section.

For the GLM values all other rating variables are included by weighting by exposure. For IAD the smoker / non-smoker variable is weighted by exposure.

Table 5.1 compares the smoothed annual incidence rates using the GLM from Section 4 and the rates from IAD89-93 for 2-week deferment period policies.

Table 5.1 IAD vs GLM incidence rates - two week deferment, males

	Occupation A	Occupation B	Occupation C	Occupation D
Age 22 (IAD)	1.8889%	4.0643%	5.4911%	5.7715%
Age 22 (GLM)	1.6661%	3.3402%	6.4850%	7.5635%
Age 27 (IAD)	1.7401%	3.4906%	4.9069%	5.3417%
Age 27 (GLM)	1.5625%	2.8803%	5.4560%	6.2806%
Age 32 (IAD)	1.8441%	3.2985%	4.8465%	5.4489%
Age 32 (GLM)	1.6318%	2.8408%	5.1082%	5.8771%
Age 37 (IAD)	2.1474%	3.4744%	5.1656%	5.9374%
Age 37 (GLM)	1.8429%	3.0734%	5.2040%	5.9217%
Age 42 (IAD)	2.6645%	4.0039%	5.8266%	6.7658%
Age 42 (GLM)	2.1955%	3.5029%	5.5956%	6.3472%
Age 47 (IAD)	3.4226%	4.8797%	6.7981%	7.8987%
Age 47 (GLM)	2.7603%	4.1802%	6.2835%	7.0909%
Age 52 (IAD)	4.4471%	6.1014%	8.0550%	9.3069%
Age 52 (GLM)	3.6175%	5.2749%	7.4288%	8.3044%
Age 57 (IAD)	5.7930%	7.6759%	9.5784%	10.9671%
Age 57 (GLM)	4.9656%	6.6699%	8.9480%	10.1073%
Age 62 (IAD)	7.5205%	9.6170%	11.3560%	12.8620%
Age 62 (GLM)	7.0301%	8.8306%	11.2966%	13.0054%
Age 67 (IAD)	9.4457%	11.7093%	13.2333%	14.8477%
Age 67 (GLM)	10.2979%	12.4523%	14.8078%	17.2813%

Graph 5.1 IAD vs GLM incidence rates - two week deferment, males, A

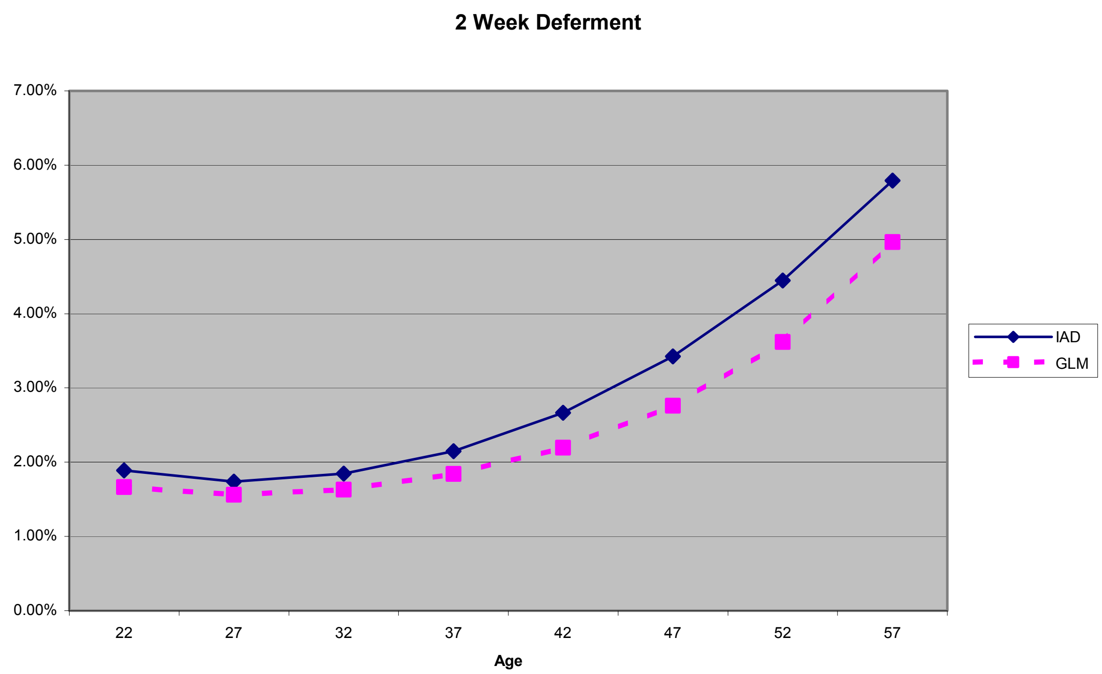
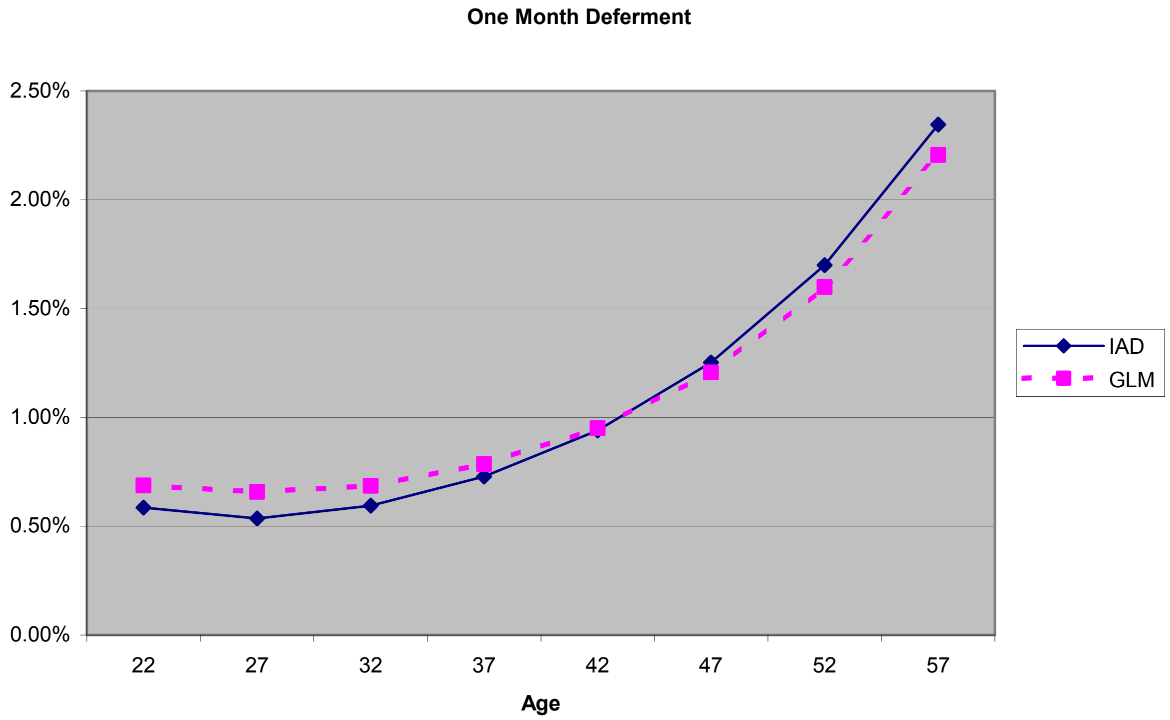


Table 5.2 compares the incidence rates using the GLM from Section 4 and the rates from IAD89-93 for one-month deferment period policies.

Table 5.2 IAD vs GLM incidence rates - one month deferment, males

	Occupation A	Occupation B	Occupation C	Occupation D
Age 22 (IAD)	0.5856%	1.1248%	2.2204%	2.2662%
Age 22 (GLM)	0.6874%	1.3810%	2.6977%	3.2659%
Age 27 (IAD)	0.5358%	1.0037%	1.9808%	2.1374%
Age 27 (GLM)	0.6578%	1.2108%	2.2889%	2.7232%
Age 32 (IAD)	0.5947%	1.0333%	1.9486%	2.1758%
Age 32 (GLM)	0.6861%	1.1913%	2.1743%	2.5688%
Age 37 (IAD)	0.7290%	1.1581%	2.1002%	2.3628%
Age 37 (GLM)	0.7853%	1.2686%	2.2314%	2.6056%
Age 42 (IAD)	0.9400%	1.3428%	2.4340%	2.7153%
Age 42 (GLM)	0.9504%	1.4557%	2.4194%	2.8215%
Age 47 (IAD)	1.2517%	1.6247%	3.0144%	3.3107%
Age 47 (GLM)	1.2068%	1.7347%	2.7519%	3.1868%
Age 52 (IAD)	1.6995%	2.1345%	3.9507%	4.2741%
Age 52 (GLM)	1.5995%	2.1383%	3.2830%	3.7740%
Age 57 (IAD)	2.3454%	3.0658%	5.4235%	5.7822%
Age 57 (GLM)	2.2051%	2.7362%	4.0876%	4.6466%
Age 62 (IAD)	3.2603%	4.6743%	7.6233%	8.0478%
Age 62 (GLM)	3.1133%	3.4621%	5.1661%	5.9687%
Age 67 (IAD)	4.3157%	6.5442%	9.9974%	10.6651%
Age 67 (GLM)	4.4394%	4.6979%	6.9509%	7.4860%

Graph 5.2 IAD vs GLM incidence rates - one month deferment, males, A



The key message from the above comparison is that the differences between the incidence rates predicted by the GLM and the incidence rates in IAD89-93 are, overall, not great although in some cells there are significant differences.

The difference between the IAD and the GLM rates is greatest at age 67 in the above table but of course the amount of exposure at this age group is very small.

It should also be noted that a global comparison between IAD and the GLM rates cannot be made from the above table since the comparison takes no account of the different levels of exposure.

6. GOODNESS OF FIT ANALYSIS

In this section the fit of the GLM described in Section 4 is analysed. A simple check of the goodness of fit is achieved by comparing the crude incidence rates with the rates predicted from the GLM. A chi-squared goodness of fit test on a 3 way table data is given. The table used includes age, occupation and is for males. The data in the table is aggregated across all other rating variables employed in the model weighted by exposure.

Table 6.1 shows the value of the fitted rates minus the crude incidence rates for the GLM described in Section 4.

Table 6.1 Goodness of Fit Analysis (absolute values)

	Occupation A	Occupation B	Occupation C	Occupation D
Age 22	-0.73%	-1.52%	-1.47%	-1.80%
Age 27	-0.06%	-0.20%	-0.41%	0.10%
Age 32	-0.15%	-0.19%	0.16%	0.46%
Age 37	-0.18%	-0.14%	0.15%	0.01%
Age 42	-0.15%	-0.27%	0.12%	-0.42%
Age 47	-0.24%	0.03%	0.14%	0.51%
Age 52	-0.13%	0.17%	-0.04%	-0.58%
Age 57	-0.13%	0.12%	-0.07%	-0.24%
Age 62	-1.75%	-2.43%	-2.83%	2.03%
Age 67	2.75%	-2.06%	3.47%	2.45%

It is clear from the above table that the fitted rates at the higher ages are not as close to the crude rates. This is partly due to the lack of data at the higher ages and therefore the greater volatility in the reported rates at these ages. In the construction of IAD89-93 it was noted that the higher ages were often ignored in the fitting process or constraints on the fitted values were employed to ensure that the more volatile rates at these ages did not effect the fitted rates too much.

A global test of the fit of the model using a chi squared test indicates that the fitted rates generated by the GLM adhere closely to the crude rates at greater than 95% significance.

7. CONCLUSIONS

The GLM whose full specification is set out in Appendix A is a very good fit to the claim incidence data for the years 1995 to 1998 with significance greater than 95%.

In the authors' view the use of a GLM incorporating all material characteristics of the data provides a more robust way of modelling expected claims and would enable a better understanding of the reasons for the usual deviation between actual and expected incidence experience.

To use a GLM for industry "standard" tables would greatly simplify the graduation work normally involved in producing such a table. Indeed, it would be possible to use the GLM described in this paper as the latest incarnation for a "standard" table without further work provided the demonstrated goodness of fit is regarded as acceptable.

The extension of these GLM techniques to the issue of claim termination rates is the author's next project and work has already started. It is expected to be finalised in the new future,

BIBLIOGRAPHY

Brockman, M. J., and Wright, T. S. (1992). Statistical motor rating: making effective use of your data. *Journal of the Institute of Actuaries* **119**, 457.

Disability Committee of the Institute of Actuaries of Australia (2002). Report of the Disability Committee, Institute of Actuaries of Australia, 2002.

Haberman, S. and Pitacco, E. (1999). Actuarial Models for Disability Insurance. *Chapman and Hall/CRC.*

Haberman, S., and Renshaw, A. E. (1996). Generalized linear models and actuarial science. *The Statistician* **45**, 41-65.

McCullagh P. and Nelder J. A. (1989) Generalized Linear Models. Second Edition. *Chapman and Hall/CRC.*

Miller, J.H. and Courant, S. (1974). A Mathematical Model of the Incidence of Disability. *Transactions of the Society of Actuaries Vol 25*

Nelder, J. A. and W. Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society A* ***135**, 370-384.*

Renshaw, A. E. (1991). Actuarial Graduation and Generalised Linear and Non-Linear Models. *Journal of the Institute of Actuaries* ***118**, II, 295-312*

Renshaw, A. E. and Haberman, S. (1995). On the graduation associated with a multiple state model for permanent health insurance. *Insurance, Mathematics and Economics,* ***17,*** 1-17.

Service D.A. & Pitt D. (2002) *Disability Claims – Does Anyone Recover?* Australian Actuarial Journal Issue 4, *615-659*

Appendix A

S Plus Output for the GLM

```
Call:
glm(formula = actualbenpct ~ offset(log(exposure)) + gender1 + agecon + sqrt(agecon) + occupationB
+ occupationC + occupationD + definition2 +
          definition3 + definition4 + definition5 + definition6 + deferment2 + deferment3 + deferment4 +
deferment5 + deferment6 + deferment7 +
          deferment8 + benperiod1 + benperiod2 + benamount2 + benamount3 + benamount4 +
benamount5 + benamount6 + benamount7 + smokercon + aids1
+
          duration1 + duration2 + agecon * occupationB + agecon * occupationC + agecon *
occupationD + gender1 * agecon + gender1 * occupationB +
          gender1 * occupationC + gender1 * occupationD + agecon * deferment2 + agecon *
deferment3 + agecon * deferment4 + agecon * deferment5 +
          agecon * deferment6 + agecon * deferment7 + ncb2 + contract1 + contract2 + contract3 +
medical1, family = poisson)

Coefficients:
 (Intercept)   gender1    agecon sqrt(agecon) occupationB occupationC occupationD definition2
definition3 definition4 definition5 definition6
   -5.281746 0.2747452 0.2460146    -2.310697    0.921944    1.805061    2.000999 -0.05887085 -
0.04331885  -0.2197125  -0.3564313   -1.158268

 deferment2  deferment3 deferment4 deferment5 deferment6 deferment7 deferment8 benperiod1
benperiod2 benamount2 benamount3 benamount4
  0.9062213 -0.04822189  -2.540355  -2.612055  -4.371066  -3.533438
-3.333802  -1.078854 0.01962875 0.04405446  0.1743777  0.2004298

 benamount5 benamount6 benamount7 smokercon     aids1 duration1  duration2      ncb2  contract1
contract2 contract3   medical1
  0.2358995  0.2116414  0.2541993 0.1566849 0.2228785 0.2075731 0.08347734 0.06736696 -
0.3215316 -0.07804049 0.1490454 -0.2056168

 agecon:occupationB agecon:occupationC agecon:occupationD gender1:agecon
gender1:occupationB gender1:occupationC gender1:occupationD
       -0.00921935        -0.01975208        -0.02118229     0.00486997         -0.1758584         -
0.4740741          -0.5635285

 agecon:deferment2 agecon:deferment3 agecon:deferment4 agecon:deferment5 agecon:deferment6
agecon:deferment7
      -0.02592019       -0.02309725        0.01188085        0.002876845        0.02772008       -
0.01185985

Degrees of Freedom: 202516 Total; 202467 Residual
Residual Deviance: 68951.67
```

Appendix B

The Encoding Scheme for each Characteristic

The variables in the table below, apart from agecon which is a continuous predictor, take the value 1 if the description is true and 0 otherwise.

When all indicator variables for a particular characteristic are zero, the remaining possible value is true.

Variable Name	Description
definition2	own occupation for 2 years then any suitable
definition3	any suitable
definition4	own occupation for 5 years then any suitable
definition5	own occupation for 3 years then any suitable
definition6	own occupation for 1 years then any suitable
gender1	Female
occupationB	Occupation Class is B
occupationC	Occupation Class is C
occupationD	Occupation Class is D
deferment2	14 day deferment period
deferment3	1 month deferment period
deferment4	2 month deferment period
deferment5	3 month deferment period
deferment6	6 month deferment period
deferment7	1 year deferment period
deferment8	>1 year deferment period
agecon	age in quinquennial groups from 17 to 72
benperiod1	benefit period <expiry age
benperiod2	benefit period until expiry or lifetime
benamount2	1500-2000 per month
benamount3	2000-2500 per month
benamount4	2500-3500 per month
benamount5	3500-5000 per month
benamount6	5000-8000 per month
benamount7	>8000 per month
bentype2	increasing benefit
bentype3	level, out of working hours only
bentype4	increasing, out of working hours only
medical1	medical evidence required
contract1	level guaranteed premiums
contract2	level non-guaranteed premiums
contract3	stepped guaranteed premiums
ncb2	no claim bonus
smoker1	smoker
aids1	AIDS covered
duration1	less than 1 year since policy purchase
duration2	between 1 and 2 years since policy purchase