APPENDIX B MODELS

B.1 Introduction to Generalised Linear Models

B.1.1 Overview

A generalised linear model ("GLM") is a generalisation of ordinary least squares regression that is able to deal with non-normally distributed response variables. Given a response variable *y* and a set of independent variables or predictors $x_1, x_2, \ldots, x_n$, a GLM models the dependency as:

$$y = h^{-1}\left(\sum_{i=1}^{n} \beta_i x_i\right) + \varepsilon_i \qquad \text{(B.1)}$$

And

$$E(y) = \mu = h^{-1}\left(\sum_{i=1}^{n} \beta_i x_i\right) \qquad \text{(B.2)}$$

Where

$h^{-1}()$ is the **link function**

β_i (i=1, 2, …, n) is the **parameter** corresponding to the dependent variable x_i

ε_i is an **error** term.

Note that

$$\eta = \sum_{i=1}^{n} \beta_i x_i \qquad \text{(B.3)}$$

is referred to as the **linear predictor** and that the GLM may be written as:

$$y = h^{-1}(\eta) + \varepsilon_i \qquad \text{(B.4)}$$

Thus, a GLM consists of three components:

1. A probability distribution;
2. A link function; and
3. A linear predictor.

B.1.2 Further detail

Probability distribution

In the equations (B.1) and (B.4) above, the error term ε_i is determined by the probability distribution of the response variable. Common distributions that may be used include:

- Normal;
- Poisson;
- Gamma;
- Inverse Gaussian; and
- Binomial.

The choice of distribution is informed by the response variable. For example, counts are naturally modelled by a Poisson distribution while strictly positive continuous quantities may be appropriately handled by a Gamma or Inverse Gaussian distribution depending on the distribution of the response values. Probabilities may be modelled using a Binomial distribution.

Link function

The link function $h^{-1}()$ gives the relationship between the mean of the distribution and the linear predictor. There are many possibilities for the link function including (but not limited to):

- Identity link: $h^{-1}(\eta) = \eta$;
- Log link: $h^{-1}(\eta) = \exp(\eta)$; and
- Logit link: $h^{-1}(\eta) = \exp(\eta)/(1 + \exp(\eta))$.

It is usually convenient to choose a link function which matches the domain of the link function to the range of the response variable's mean. In other words, if a response must be positive (for example, an average benefit payment), then a log link will ensure that the fitted value μ (B.2) is positive. If the modelled quantity is a probability (for example, the probability of transitioning off benefit in the next quarter), then the logit link ensures that the fitted value lies between 0 and 1, as probabilities must.

Linear predictor

The linear predictor (equation B.3) is the quantity which incorporates the information about the independent variables into the model and is typically denoted by η. η is expressed as linear combinations of unknown parameters β_i and independent variables x_i (i=1, 2, …). The x_i are known variables.

In all cases, once the probability distribution and the link function have been selected, the linear predictor (B.3) needs to be constructed. The steps to doing this include:

- Identify the list of independent variables or predictors (x_i) to be considered;

- Using data exploration, modelling techniques, statistical tests and prior knowledge, identify those predictors that are useful predictors of the response variable. Note

that this may include functions of the predictors, rather than the raw predictors themselves; and then

- Estimate the parameters β_i using GLM software.

The list of variables considered for the key benefits is given in Section 4.3.3. For the minor benefits, the variables are listed in Sections 4.5.2 and 4.5.3.

Functions of the predictors

The predictors or independent variables may be used as follows.

- **In their raw forms**: For example, had_dpb with two levels 0 and 1;
- **As categorical groupings of the original variable:** For example, age may be banded into a number of groups (<18, 18-29, 30-39 etc);
- **As indicator functions depending on the value of the original variable where one condition is assigned the value 1 and the complementary position 0:** For example, letting I(age ≥ 30) be 1 for age ≥ 30 and 0 otherwise would fit a step term at age 30;
- **As a spline for underlying raw predictors which are numeric or ordinal (e.g. age, benefit quarter, duration on benefit):** The dependency of a linear predictor on duration could be modelled (if appropriate) by a combination of several line segments. For instance, if the linear predictor varied in a linear fashion with duration with one slope from duration 1 to 4, a different slope from 4 to 12 and a third slope from 12 onwards, then using three line pieces(1-4, 4-12 and 12+) would capture this dependency. The points 4 and 12 where the resulting fitted spline bends are referred to as knot points.
- **As Interaction terms:** All of the above may be used as interaction terms. For example a duration effect may be well fitted by one spline for those aged under 30 and another for those aged 30 and above. This could be accommodated by interacting the spline with the I(age ≥ 30) term.

B.1.3 Model fitting approach

Our typical approach to fitting a model includes the following:

- First fit a saturated model including most, if not all, raw predictors as well as any known interactions. For continuous predictors like age, or categorical ordered predictors like duration, we would usually fit the predictor as a grouped version (e.g. for age which is in quarter years, we might fit it as integer years).
- Simplify the model by:
  - Removing insignificant parameters;
  - Grouping together related parameters with similar estimated values; and
  - Using splines where this is warranted.

- Using diagnostics check to see if there is evidence of poor fitting which may suggest the need for some interactions. Add additional terms as required until a satisfactory fit is obtained.

B.1.4 References

The following books give a complete introduction to GLMs:

- McCullagh P. and Nelder J. (1989). *Generalized linear models, second edition*. Chapman and Hall, London UK.

- Dobson A. J. (2002). *An introduction to generalized linear models, second edition*. Chapman & Hall/CRC, Florida USA.

For a discussion on the application of GLMs in contexts similar to the modelling of the MSD benefit liabilities (e.g. claim size and claim numbers modelling in insurance), the following papers provide some starting points.

- England, P. D. and Verrall, R. J. (2002). Stochastic claims reserving in general insurance. British Actuarial Journal, **8** 443-544.

- Haberman, S. and Renshaw, A. E. (1996). Generalized linear models and actuarial science. *The Statistician*, **45** 407-436.

- Mulquiney, P. and Taylor, G. (2007). Modeling Mortgage Insurance as a multi-state process. *Variance* **1**, 81-102.

- Taylor, G. and McGuire, G (2004). Loss reserving with GLMs: a case study. *Casualty Actuarial Society Discussion Paper Program 2004*. Available at http://www.casact.org/pubs/dpp/dpp04/04dpp327.pdf

B.2 Application of GLMs to the MSD valuation

B.2.1 Types of GLMs used in the MSD valuation

To fit a GLM to any part of the MSD experience we need to set out the form of each of the three components of the GLM:

1. A probability distribution;
2. A link function; and
3. A linear predictor.

This section discussed the choice of the first two while the next section considers the linear predictor.

As discussed in Section B.1.1, the choice of probability distribution and link function is informed by the response variable. For MSD, we have several categories of models:

- Probability models;

- Multivariate models;
- Average benefit payment models; and
- Numbers of newcomers to the benefits system.

Probability models

Examples of these include the TRA and NOB models (Section 4.3.3) which model, respectively, the probabilities of staying on the same benefit and the probabilities of moving off benefit given that the person leaves the current benefit. For these, a binomial distribution is a suitable probability function while the logit link ensures that the linear predictor (any real number) maps to a probability between 0 and 1.

Multivariate models

For the key benefits, the probabilities of moving to a different benefit type given that the person changes benefit and does not move off benefit are modelled using a multinomial probability distribution since there are four possible outcomes (or five in the case where the person starts on the NOB state). Strictly speaking, rather than probabilities, this model estimates numbers in each of the four states and thus a log link is appropriate (i.e. numbers are therefore positive). The probabilities may be obtained by scaling the estimated values so that they sum to one.

Average Benefit Payment models

ABP models exist for all benefit types. For most of these, we have used a gamma distribution and a log link so that claim sizes are strictly positive, but continuous. For some models, there are a significant number of APBs of zero; for these we have used a Poisson distribution with a log link, which is capable of handling zero ABPs.

Number of newcomer models

There are 6 models for the number of newcomers entering the system: one for each key benefit type, including OTH but not NOB, and one for minor benefits. The DPB and minor benefit models have been constructed using Poisson models with a log link. The remainder have been constructed using averages of past experience.

B.2.2 Model validation for the MSD valuation

It is very important to ensure that model fits are adequate. There are a number of tests that have been used for the MSD modelling, listed below.

Residual diagnostics

For a well-specified model, the standardised deviance residuals should be approximately normally distributed. Residual scatterplots and histograms may indicate regions of poor matches to the assumptions. Heteroscedastic residuals (for example, indicated by a fanning out or fanning in of residuals plotted against the linear value) suggest that the probability distribution is unsuitable or that weights for the existing probability distribution

may be required to give a better match. Another useful diagnostic is a Q-Q plot against the selected probability distribution.

Parameter diagnostics

A series of diagnostics related to individual parameters is also available. The most widely cited is the *p*-value, which a value describing the statistical significance of that parameter. It is the probability of seeing the observed effect in the data if there was actually no genuine parameter effect. A very low p-value (below 0.01 say), means the chance of that parameter being a spurious effect is very small (less than 1%).

Actual versus Expected Analysis

This analysis compares actual and expected values for different subsets of the data and is useful for identifying missing parameters. A simple plot of the actuals and expected in a number of groups (e.g. 20-30), ordered by expected values, can often indicate that the model is missing some important interactions. Looking at actual and expected values for specific data subsets may help identify what is missing.

Model comparison tools and statistics

The Akaike Information Criterion (AIC) is sometimes useful in comparing one model to another. Gains charts are a useful visual aid in distinguishing between models.

Backtesting

Although not specific to GLMs, backtesting is a particularly useful model testing method for a valuation such as the one described here where a number of probability models and ABP models are combined to project forward future cohorts on benefits and future cash flows. As its name suggests, backtesting involves the projection of a historical cohort (e.g. all those who received a benefit in the last 12 months at June 2000, or all those receiving DPB at December 2003 etc) to the end of the experience. Actual experience may then be compared with the projected experience to identify any deficiencies exposed by the chaining of models.

B.3 Description of the models spreadsheet appendix

B.3.1 Introduction

As part of the appendices to the valuation report, the spreadsheet "02_models.xlsm" has been provided, giving detailed descriptions of all the models used in constructing the valuation. Most of the sheets describe GLMs and have identical layouts, while the remaining sheets describe non-GLM models. For each sheet that describes a GLM, we have included the following components:

- **Parameter tables:** These tables list each effect in the model, its corresponding parameter estimate, its statistical significance plus the formula for constructing the

effect from the raw underlying variables. Some further specific comments on the parameter table:

- The intercept is the one effect in the linear predictor that does not have an attached formula – it is a constant effect applied to all observations;
- The "ProbChiSq" column is the p-value arising from the Chi-squared test of parameter significance. In loose terms, it is the probability of the parameter not being a real effect;
- The Scale parameter, listed at the end of each model, relates to how the mean estimate is related to an observation's variance. For more details, consult the texts suggested above.

- **Linear predictor plots:** These show the fitted effects for calendar quarter, duration (in quarters) and client age *assuming all other effects are held constant at their base levels*. These charts are given on a linear predictor scale, rather than the actual (probability or payment amount) scale. DPB models also include plots for number of children and youngest child age, where appropriate.

- **Actual versus expected plots:** These show the average actual and predicted values of the response by calendar quarter, duration (in quarters) and client age. Note these are averaged over the entire modelling dataset, and it is possible that their shape differs from the linear predictor plots due to correlations between predictor variables. These charts are on an actual response scale. DPB models also include plots for number of children and youngest child age, where appropriate.

- **Model notes:** These tables contain some brief comments on trends observed in the model, with an emphasis on calendar quarter effects.

The effects included in the models are a function of raw underlying variables. The following table lists those variables and gives brief descriptions:

Table B.1 **Description of raw variables used in GLMs**

Variable	Description
Age, age_band	Age of client. Generally age is in quarter years (e.g. 23.25), while age_band will be truncated to age in years
Dur_current_ben	Duration of current spell on benefit, in quarters
Ben_qtr	Calendar quarter of benefit payment, a date set to the end of that quarter
UErate	The unemployment rate for the corresponding calendar quarter
Qtr, dur_season	Takes values 1/2/3/4, indicating the quarter (Mar = 1, Jun = 2 etc)
Dur	Duration since first entering the system, measured in quarters
Start_qtr	Quarter client first entered the welfare system, measured in quarters
Active_dur	Number of quarters between start_qtr and the valuation

	date
Dur_diff, dur_lag	The difference between the valuation date and the (later) benefit quarter, measured in quarters
had_dpb, had_inv, had_nob, had_oth, had_sic, had_ueb	Binary flag (0-1) indicating whether client has previously received DPB, IB, NOB, OTH, SB and UB respectively

B.3.2 Worksheet descriptions

The following table gives a description of each of the sheets included in the models spreadsheet. ABP is the "average benefit paid", estimated on a quarterly basis.

Table B.2 **Description of sheets of models appendix**

Sheet name	GLM?	Distribution, link	Description
d_chi	Yes	Binomial, logit	Probability model for the chance that a DPB client has a new youngest child in that quarter.
d_chi2	No	N/A	Distributions used for child models for DPB clients: • Number of children when joining DPB • Age of youngest child when joining DPB • Age of youngest child if positive response in D_chi model • Distribution for changing number of children for continuing DPB clients
d_tra	Yes	Binomial, logit	Probability that DPB client remains in DPB in the next quarter.
d_nob	Yes	Binomial, logit	Probability that a client who leaves DPB moves to NOB state (no benefit)
d_mul	Yes	Multivariate logistic model	Multivariate probability model for eventual state of a client who leaves DPB and does not moves to NOB
d_abp	Yes	Gamma, log	ABP for DPB benefits paid to clients receiving DPB
d_as	Yes	Gamma, log	ABP for AS benefits paid to clients receiving DPB
d_da	Yes	Gamma, log	ABP for DA benefits paid to clients receiving DPB
i_tra	Yes	Binomial, logit	Probability that IB client remains in IB in the next quarter.
i_nob	Yes	Binomial, logit	Probability that a client who leaves IB moves to NOB state (no benefit)
i_mul	Yes	Multivariate logistic model	Multivariate probability model for eventual state of a client who leaves IB and does not move to NOB
i_abp	Yes	Gamma, log	ABP for IB benefits paid to clients receiving IB
i_as	Yes	Gamma, log	ABP for AS benefits paid to clients receiving IB
i_da	Yes	Gamma, log	ABP for DA benefits paid to clients receiving IB
s_tra	Yes	Binomial, logit	Probability that SB client remains in SB in the next quarter.
s_nob	Yes	Binomial, logit	Probability that a client who leaves SB moves to NOB state (no benefit)
s_mul	Yes	Multivariate logistic model	Multivariate probability model for eventual state of a client who leaves SB and does not move to NOB

Sheet name	GLM?	Distribution, link	Description
s_abp	Yes	Gamma, log	ABP for SB benefits paid to clients receiving SB
s_as	Yes	Gamma, log	ABP for AS benefits paid to clients receiving SB
s_da	Yes	Gamma, log	ABP for DA benefits paid to clients receiving SB
u_tra	Yes	Binomial, logit	Probability that UB client remains in UB in the next quarter.
u_nob	Yes	Binomial, logit	Probability that a client who leaves UB moves to NOB state (no benefit)
u_mul	Yes	Multivariate logistic model	Multivariate probability model for eventual state of a client who leaves UB and does not move to NOB
u_abp	Yes	Gamma, log	ABP for UB benefits paid to clients receiving UB
u_as	Yes	Gamma, log	ABP for AS benefits paid to clients receiving UB
u_da	Yes	Gamma, log	ABP for DA benefits paid to clients receiving UB
o_tra	Yes	Binomial, logit	Probability that OTH client remains in OTH in the next quarter.
o_nob	Yes	Binomial, logit	Probability that a client who leaves OTH moves to NOB state (no benefit)
o_mul	Yes	Multivariate logistic model	Multivariate probability model for eventual state of a client who leaves OTH and does not move to NOB
n_tra	Yes	Binomial, logit	Probability that NOB client remains in NOB in the next quarter.
n_mul	Yes	Multivariate logistic model	Multivariate probability model for eventual state of a client who leaves NOB
dpb-csi_prob	Yes	Binomial, logit	Probability that a client active in the system is receiving a DPB-CSI benefit
dpb-csi_abp	Yes	Poisson, log	ABP for DPB-CSI for those receiving that benefit
WA-WB_prob	Yes	Binomial, logit	Probability that a client active in the system is receiving a WA-WB benefit
WA-WB_abp	Yes	Poisson, log	ABP for WA-WB for those receiving that benefit
eb_prob	Yes	Binomial, logit	Probability that a client active in the system is receiving a EB benefit
eb_abp	Yes	Poisson, log	ABP for EB for those receiving that benefit
orp_prob	Yes	Binomial, logit	Probability that a client active in the system is receiving a ORP benefit
orp_abp	Yes	Poisson, log	ABP for ORP for those receiving that benefit
as_prob	Yes	Binomial, logit	Probability that a client active in the system is receiving an AS benefit but not a key benefit
as_abp	Yes	Poisson, log	ABP for AS for those receiving that benefit but not in receipt of a key benefit
da_prob	Yes	Binomial, logit	Probability that a client active in the system is receiving a DA benefit but not a key benefit
da_abp	Yes	Poisson, log	ABP for DA for those receiving that benefit but not in receipt of a key benefit
cda_prob	Yes	Binomial, logit	Probability that a client active in the system is receiving a CDA benefit
cda_abp	Yes	Poisson, log	ABP for CDA for those receiving that benefit
ccs_prob	Yes	Binomial, logit	Probability that a client active in the system is receiving a CCS benefit
ccs_abp	Yes	Poisson, log	ABP for CCS for those receiving that benefit

Sheet name	GLM?	Distribution, link	Description
ccs2_prob	Yes	Binomial, logit	Probability that a client active outside the system is receiving a CCS benefit
ccs2_abp	Yes	Poisson, log	ABP for CCS for those receiving that benefit and outside the system
ei_prob	Yes	Binomial, logit	Probability that a client active in the system is receiving a EI benefit
ei_abp	Yes	Poisson, log	ABP for EI for those receiving that benefit
ei2_prob	Yes	Binomial, logit	Probability that a client active outside the system is receiving an EI benefit
ei2_abp	Yes	Poisson, log	ABP for EI for those receiving that benefit and outside the system
hs_prob	Yes	Binomial, logit	Probability that a client active in the system is receiving an HS benefit
hs_abp	Yes	Poisson, log	ABP for HS for those receiving that benefit
hs2_prob	Yes	Binomial, logit	Probability that a client active outside the system is receiving an HS benefit
hs2_abp	Yes	Poisson, log	ABP for HS for those receiving that benefit and outside the system
Loa_prob	Yes	Binomial, logit	Probability that a client active outside the system is receiving a Recoverable Assistance payment
Loa_abp	Yes	Poisson, log	ABP for those receiving a Recoverable Assistance payment
react_prob	Yes	Binomial, logit	Probability that a person who is not in the current client liability re-enters the system (i.e. starts receiving a benefit again)
minor_new	Yes	Poisson, log	Number of genuinely new people entering the welfare system
key_new	No	N/A	Time series models for people joining key liability benefits, whether new or having been off benefits for at least 40 quarters.

Interested readers will note that the parameter tables for most of the "prob" models contain a series of terms of the form *bq_last*active_eq_dur*start_qtr3 ="XXX"*. These are additional terms introduced towards the end of the modelling process to ensure the probability level projected matches the levels seen in recent quarters.