UNIVERSITEIT VAN AMSTERDAM

Amsterdam School of Economics

Computer class NLIST—Assignment 3A

Using GLM to analyze a portfolio and a BM-system

This assignment does GLM-analyses for the portfolio explored in Sec. 9.6, resembling the one studied in Appendix A.3, and also resembling the dataset that was the basis for the bonus-malus system developed in Ch. 6. See also the text on MART p. 252–261. In a previous assignment, we did some exploratory data analysis, and now we will do GLM-analyses.

Some things you will learn about:

- constructing fitted values from parameter estimates in a model with log-link
- estimating the dispersion parameter for a quasi-Poisson fit
- explaining the average claim size using a *gamma* model with log-link and weights
- overdispersion in case of mixed Poisson r.v.'s
- mean, variance and skewness of gamma and lognormal r.v.'s

1 Analyzing a bonus-malus system using GLM

To read and store the data as a dataframe, and to construct some other variables needed, do:

```
rm(list=ls(all=TRUE)) ## First remove traces of previous sessions
fn <- "http://www1.fee.uva.nl/ke/act/people/kaas/Cars.txt"
Cars <- read.table(fn, header=TRUE)
Bminus1 <- Cars$B - 1; Bis14 <- as.numeric(Cars$B==14)
Cars$A <- as.factor(Cars$A); Cars$R <- as.factor(Cars$R)
Cars$M <- as.factor(Cars$M); Cars$U <- as.factor(Cars$U)
Cars$B <- as.factor(Cars$B); Cars$WW <- as.factor(Cars$WW)
ActualWt <- c(650,750,825,875,925,975,1025,1075,1175,1375,1600)
W <- log(ActualWt/650)[Cars$WW]
```

See the assignment of last week for what `Cars` contains and what the other variables are. There we did some exploratory data analysis; here we use some GLMs to see how to incorporate bonus-malus class and car weight in the tariff. We also look at the claim severities, and combine a multiplicative gamma model for those with the model for the claim numbers. The resulting model for the risk premium is virtually identical to a simple direct model with a quasi-Poisson mean/variance relation, in which variances are proportional to the mean.

GLM analysis

We do some GLM estimations, explaining the average claim totals per policy by subsets of the risk factors. Since averages are taken over `Expo` numbers, the variances are to be divided by `Expo`, which is achieved by adding `wei=Expo`. Adding `offset=log(Expo)` and looking at totals rather than averages works just as well.

```
g1 <- glm(TotCl/Expo~R+A+U+W+Bminus1+Bis14, quasipoisson, wei=Expo, data=Cars)
g2 <- glm(TotCl/Expo~R+A+U+W+Bminus1+Bis14+M, quasipoisson, wei=Expo, data=Cars)
g3 <- glm(TotCl/Expo~R+A+U+W+B, quasipoisson, wei=Expo, data=Cars)
```

Calling `anova` helps to do the analysis-of-deviance. See Table 9.7.

```
> anova(g1,g2)
Analysis of Deviance Table

Model 1: TotCl/Expo ~ R + A + U + W + Bminus1 + Bis14
Model 2: TotCl/Expo ~ R + A + U + W + Bminus1 + Bis14 + M
  Resid. Df Resid. Dev Df Deviance
1      7515   38616941
2      7513   38614965  2   1975.8
> anova(g1,g3)
Analysis of Deviance Table

Model 1: TotCl/Expo ~ R + A + U + W + Bminus1 + Bis14
Model 2: TotCl/Expo ~ R + A + U + W + B
  Resid. Df Resid. Dev Df Deviance
1      7515   38616941
2      7504   38544506 11    72435
```

The scale factor ϕ can be estimated as the deviance in the fullest non-full model (`g3` in this case), divided by the number of degrees of freedom, that is, by $\widehat{\phi} \approx \frac{38544506}{7504} = 5137$. This means that the loss 72435 in deviance from `g3` to `g1`, by dropping 11 parameters, is in fact a loss of $\frac{72435}{5137} = 14.1$ in **scaled** deviance. The 95% critical value of a $\chi^2(11)$ db is 19.7.

So replacing the fixed decrease in the successive BM-class factors (except the last) by an arbitrary pattern does not significantly decrease the (scaled) deviance. The same holds for adding mileage `M` as an explanatory variable to `g1`.

The multiplicative coefficients estimated by the various models as they are listed in Table 9.8 are given by:

```
> options(digits=7)
> exp(coef(g1)); exp(coef(g2)); exp(coef(g3))
(Intercept)          R2          R3          A2          A3          U2
524.3016583   1.0842682   1.1916130   0.4147224   0.6184468   1.3841303
          W     Bminus1       Bis14
  2.3722083   0.8978647   1.1053665
(Intercept)          R2          R3          A2          A3          U2
522.6627527   1.0842767   1.1914111   0.4147232   0.6184538   1.3835062
```

```
          W      Bminus1        Bis14           M2           M3
  2.3721668    0.8978640    1.1053568    1.0073260    1.0014581
(Intercept)           R2           R3           A2           A3           U2
515.5320549    1.0843018    1.1916593    0.4143437    0.6178700    1.3841612
          W           B2           B3           B4           B5           B6
  2.3722369    0.9111279    0.8275175    0.7403718    0.6842609    0.6088526
         B7           B8           B9          B10          B11          B12
  0.5416103    0.4489065    0.4151901    0.3888576    0.3459030    0.3143452
        B13          B14
  0.2832722    0.2773037
```

Q_1 Are the values in Table 9.8, including those of the form 0.8979^k, correct?

Run

```
g1$y[4000]; g1$y["4000"]
g1$y[7000]; g1$y["7000"]
min(which(Cars$Expo==0))
```

`g1$y[4000]` produces the 4000th element of the vector of explained variables in `g1`. `g1$y["4000"]` produces the one with label `"4000"`, and these two coincide.

It is different for 7000. This is because from unit 5941 onwards, cells of youthful drivers in high BM-classes appear, and these cells are empty, so their weight `Expo==0`. They are excluded from the GLM data.

Now using the coefficients of `g1, g2, g3`, compute the fitted values for the cell 4000. Find the corresponding covariates using `model.matrix(g2)[4000,]`; you should find `B=7, WW=9, R=1, A=1, M=2, U=2`. Compare with the result of the observed value `g1$y[4000,]` and the fitted value `fitted(gX)[4000]` as computed by `R`.

Explain the result of `g2$family$linkinv(model.matrix(g2)[4000,]%*%coef(g2))`. □

Problems from MART 9.6

To reconstruct the portfolio needed and the other objects referred to, re-run the sccript on p. 1 and the first one on p. 2.

See the assignment of last week for ex. 9.6.1,2,3,7,11; ex. 9.6.8–10 are theory exercises.

In the following exercises, do an analysis of scaled deviance, estimating the scale factor by `Dev/Df` for a 'rich' model. Don't rely on the results of `test="Chisq"` in the `anova()` call.

Q_2 $=$ **9.6.4** In the model `g1` of Table 9.7, can `Bis14` be removed from the model (without getting a significantly worse fit? In model `g3`, can `B` or `W` be removed from the model? In `g1`, would it help to allow for separate coefficients for each weight class? □

Q_3 $\approx$ **9.6.5** The deviance of the same model as in Table 9.10, but without interactions, is 38 616 941 on 7515 df, the deviance with interactions is 38 408 588 on 7491 df. Do the interaction terms improve the fit significantly? □

Q_4 $\approx$ **9.6.6** Estimate a multiplicative model explaining `nCl/Expo` by `R, A, U, W, B` and `Bis14` using a quasi-Poisson error structure. Note: in the book there is also `M`.

Estimate a multiplicative model explaining `TotCl/nCl` by the same covariates, but now using `fam=Gamma(...)`. Use a log-link, and weights `nCl`.

Combine these two models for the risk premium into one.

Compare the coefficients of the combined models with those obtained by directly estimating `TotCl/Expo`. □