Generalised Linear Models

1. Introduction

- Model individual claim distributions for "similar" policies in a portfolio.

  · "Similarity" actually measured on the basis of various characteristics of the policyholders.

  · Models we use will be GLMs, since most claim distributions are not normal.

- Choose mixing distribution based on distribution of characteristics of policyholders in the portfolio.
- Create "portfolio-wide" distribution using mixture distribution procedure with "realistic" mixing distribution based on deterministic characteristics and a defensible model for the individual claim distributions.

2. Model Structure and Definition

- Response (claim amount) data: $X_1, \ldots, X_n$ with $E(X_i) = \mu_i$.
- Predictor vectors (p-dimensional): $u_1, \ldots, u_n$ with $u_i = (u_{i1}, \ldots, u_{ip})^T$.
- A *link function*, $h(\cdot)$, relating the mean response, μ_i, to the *linear predictor* $\eta_i = u_i^T \beta = u_{i1}\beta_1 + \ldots + u_{ip}\beta_p$, so that $\eta_i = h(\mu_i)$. This allows for non-linear relationships between the predictors and the response variable and requires estimation of the appropriate values for $\beta = (\beta_1, \ldots, \beta_p)$.
- A family of distributions for our response (e.g., Gamma, Poisson, etc.). Usually, we will try and choose an *exponential family* of distributions, which are specific types of families.

3. Exponential Families and Link Functions

- An exponential family (with dispersion) has *pdf*:

$$f_X(x;\mu,\phi) = \exp\left\{\frac{xb(\mu)-c(\mu)}{\phi} + d(x,\phi)\right\}$$

  for some functions $b(\mu)$, $c(\mu)$ and $d(x,\phi)$.

- Typically, $E(X)=\mu$. Also, ϕ is the *dispersion* parameter.
- Example - Gamma family in the (α,μ) parameterisation:

$$\frac{\alpha^\alpha}{\mu^\alpha\Gamma(\alpha)}x^{\alpha-1}e^{-\alpha x/\mu} = \exp\left[\frac{-x\mu^{-1}-\ln(\mu)}{\alpha^{-1}}+(\alpha-1)\ln(x)+\alpha\ln(\alpha)-\ln\{\Gamma(\alpha)\}\right],$$

  which has desired form when $\phi=\alpha^{-1}$, $b(\mu)=-\mu^{-1}$, $c(\mu)=\ln(\mu)$ and $d(x,\phi)=(\phi^{-1}-1)\ln(x)-\phi^{-1}\ln(\phi)-\ln\{\Gamma(\phi^{-1})\}$.

- Example - Normal family

$$\frac{1}{\sigma\sqrt{2\pi}}\exp\left\{\frac{-1}{2\sigma^2}(x-\mu)^2\right\} = \exp\left\{\frac{x\mu-(\mu^2/2)}{\sigma^2}-\frac{x^2}{2\sigma^2}-\log(\sigma\sqrt{2\pi})\right\},$$

  which if we let $\phi=\sigma^2$, $b(\mu)=\mu$, $c(\mu)=\mu^2/2$ and $d(x,\phi) = -(x^2/2\phi)-\log(\sqrt{2\pi\phi})$.

3. Exponential Families and Link Functions *(Continued)*

- Some general properties of Exponential Families:

$$E(X) = \mu = \frac{c'(\mu)}{b'(\mu)}$$

and $Var(X) = \phi V(\mu)$ with:

$$V(\mu) = \frac{c''(\mu) - \mu b''(\mu)}{b'(\mu)^2}.$$

- *Canonical* link function

· Log-likelihood is:

$$l(\beta, \phi) = \sum_{i=1}^{n} \left\{ \frac{X_i b(h^{-1}(u_i^T \beta)) - c(h^{-1}(u_i^T \beta))}{\phi} + d(X_i, \phi) \right\}.$$

so choosing $h(\cdot) = b(\cdot)$ gives nice mathematical reduction to:

$$l(\beta, \phi) = \sum_{i=1}^{n} \left\{ \frac{X_i \eta_i - c(h^{-1}(\eta_i))}{\phi} + d(X_i, \phi) \right\}.$$

4. Estimation of Parameters - MLEs and IRLS:

- The score equations for the MLE of β are:

$$\frac{\partial}{\partial \beta_j} l(\beta, \phi) = \frac{1}{\phi} \sum_{i=1}^{n} \frac{\partial \mu_i}{\partial \beta_j} b'(\mu_i)(X_i - \mu_i) = 0.$$

- Can be solved iteratively using a weighted least-squares (*IRLS*) algorithm:

  · Pick an initial value, $\hat{\beta}_{old}$;

  · Define $\hat{\eta}_i = u_i^T \hat{\beta}_{old}$, $\hat{\mu}_i = h^{-1}(\hat{\eta}_i)$ and

$$Z_i = \hat{\eta}_i + (X_i - \hat{\mu}_i) h'(\hat{\mu}_i)$$

  [NOTE: Don't need initial $\hat{\beta}_{old}$ if we use $\hat{\mu}_i = X_i$ and $\hat{\eta}_i = h(X_i)$.]

  · Fit a weighted regression $Z_i = u_{i1}\beta_1 + \ldots + u_{ip}\beta_{ip}$ using weights $\hat{w}_i = \left[\{h'(\hat{\mu}_i)\}^2 V(\hat{\mu}_i)\right]^{-1}$ to arrive at a new estimate

$$\hat{\beta}_{new} = (U^T \hat{W} U)^{-1} U^T \hat{W} Z$$

  where U is a matrix with the n rows $u_1, \ldots, u_n$, $\hat{W}$ is a diagonal matrix with diagonal elements $\hat{w}_1, \ldots, \hat{w}_n$ and $Z = (Z_1, \ldots, Z_n)^T$.

  · Set $\hat{\beta}_{old} = \hat{\beta}_{new}$ and repeat from Step 2.

4. Estimation of Parameters - MLEs and IRLS *(Continued)*:

- An Example - Data on Car Insurance Portfolio:

Vehicle Age (yrs)	Policyholder Age (yrs)	Claim Amount	Vehicle Age (yrs)	Policyholder Age (yrs)	Claim Amount
1	25	\$468.14	2	30	\$481.94
1	30	\$161.12	2	40	\$346.53
1	30	\$1750.33	2	55	\$244.26
1	35	\$1069.81	3	20	\$4644.47
1	40	\$1099.65	3	25	\$479.58
1	40	\$2313.55	3	25	\$3281.24
1	45	\$777.91	8	25	\$475.53
1	50	\$546.26	9	50	\$473.03
2	20	\$373.32	10	50	\$390.91
2	30	\$2021.32	10	55	\$561.98

· Assume claims are exponential (i.e., Gamma with $\alpha = 1$), so that $\phi = 1$ and $V(\mu) = \mu^2$.

· Assume $h(\mu) = 1/\mu$.

· Log-likelihood equation:

$$l(\beta) = \sum_{i=1}^{20} \left\{ \ln(\beta_0 + \beta_1 v_i + \beta_2 p_i) - X_i(\beta_0 + \beta_1 v_i + \beta_2 p_i) \right\}$$

· Score equations to be solved:

$$\sum_{i=1}^{20} \frac{1}{\beta_0 + \beta_1 v_i + \beta_2 p_i} - \sum_{i=1}^{20} X_i = 0$$
$$\sum_{i=1}^{20} \frac{v_i}{\beta_0 + \beta_1 v_i + \beta_2 p_i} - \sum_{i=1}^{20} v_i X_i = 0$$
$$\sum_{i=1}^{20} \frac{p_i}{\beta_0 + \beta_1 v_i + \beta_2 p_i} - \sum_{i=1}^{20} p_i X_i = 0$$

· Initial estimates: $\hat{\mu}_i = X_i$, $\hat{\eta}_i = 1/X_i$, $Z_i = 1/X_i$ and $\hat{w}_i = X_i^2$.

· First iteration of IRLS gives:

$$\begin{aligned}\hat{\beta}_{new} &= (U^T \hat{W} U)^{-1} U^T \hat{W} Z \\ &= (-7.8454 \times 10^{-4}, 9.626130 \times 10^{-5}, 3.676953 \times 10^{-5})^T\end{aligned}$$

· Converges to $\hat{\beta} = (-4.2614 \times 10^{-4}, 5.2056 \times 10^{-5}, 3.8283 \times 10^{-5})^T$.

5. Estimation of Parameters - SEs and CIs:

- Since $\hat{\beta}$ is MLE, use Fisher Information for SEs:

$$Var(\hat{\beta}) \approx \phi(U^T W U)^{-1} \approx \phi(U^T \hat{W} U)^{-1}$$

where U is the matrix of the predictors (including the intercept column) and W is the diagonal matrix, with diagonal elements $w_i = \left[\{h'(\mu_i)\}^2 V(\mu_i)\right]^{-1}$, and $\hat{W}$ is the estimate of W using the diagonal elements $\hat{w}_i$ as defined previously for the IRLS algorithm. [See Course Notes for algebra.]

- For preceding example:

$$\begin{aligned} Var(\hat{\beta}) &\approx (U^T \hat{W} U)^{-1} \\ &= \begin{pmatrix} 4.5489 \times 10^{-7} & -2.1248 \times 10^{-8} & -1.3349 \times 10^{-8} \\ -2.1248 \times 10^{-8} & 1.0426 \times 10^{-8} & -1.2391 \times 10^{-10} \\ -1.3349 \times 10^{-8} & -1.2391 \times 10^{-10} & 4.9209 \times 10^{-10} \end{pmatrix}. \end{aligned}$$

[Recall $\phi = 1$ since data assumed exponentially distributed.]

· So, approximate 95% CI for β_2 is:

$$3.8283 \times 10^{-5} \pm 1.96\sqrt{4.9209 \times 10^{-10}} = (-0.5196 \times 10^{-5},\ 8.1762 \times 10^{-5})$$

5. Estimation of Parameters - SEs and CIs *(Continued)*:

- Predicted value at $u_0 = (u_{01}, \ldots, u_{0p})$ is

$$\hat{\mu}(u_0) = h^{-1}(\hat{\eta}_0) = h^{-1}(u_0^T \hat{\beta})$$

- SE for $u_0^T \hat{\beta}$ calculated as:

$$\sqrt{Var(u_0^T \hat{\beta})} = \sqrt{u_0^T Var(\hat{\beta}) u_0} \approx \sqrt{\phi u_0^T (U^T \hat{W} U)^{-1} u_0}$$

- 95% CI for $\hat{\eta}_0 = u_0^T \hat{\beta}$ is:

$$(c_l,\, c_u) = \hat{\eta}_0 \pm 1.96 \sqrt{\phi u_0^T (U^T \hat{W} U)^{-1} u_0}$$

- 95% CI for $\hat{\mu}(u_0)$ is $\{h^{-1}(c_l), h^{-1}(c_u)\}$

  [NOTE: Need to reverse endpoints if $h(\cdot)$ is decreasing function.]

- Generally, need estimate of ϕ. More on this later, but sometimes, we know $\phi = 1$ as with exponential, Poisson or binomial distributed data.

- *Example (Continued):* For a 3 year-old car with a 40 year-old driver, the expected claim size is $\hat{\mu}\{(1, 3, 40)^T\} = h^{-1}(\hat{\eta}_0) = h^{-1}(\hat{\beta}_0 + 3\hat{\beta}_1 + 40\hat{\beta}_2) = 792.79$, and a 95% CI is:

$$h^{-1}\left\{ \hat{\eta}(u_0) \pm 1.96 \sqrt{\phi u_0^T (U^T \hat{W} U)^{-1} u_0} \right\} = (522.39,\, 1643.32)$$

  [NOTE: Endpoints reversed as $h(\cdot)$ is decreasing in this case.]

6. Model Selection and Analysis of Deviance:
   - Analogous to Sums-of-Squares and Analysis of Variance in Linear Regression.
   - Residual deviance for a specific model with parameter β is:

     $$D(\hat{X}, X) = 2\sum_{i=1}^{n} \big[X_i\{b(X_i) - b(\hat{X}_i)\} - \{c(X_i) - c(\hat{X}_i)\}\big],$$

     where $\hat{X}_i = h^{-1}(u_i^T \hat{\beta})$ are the fitted values and $\hat{\beta}$ is the MLE of β.

     [NOTE: See Course Notes for more detailed derivation.]
   - (Scaled) Deviance Statistic:

     $$D^*(\hat{X}_S, \hat{X}_L) = \frac{D(\hat{X}_S, X) - D(\hat{X}_L, X)}{\hat{\phi}_L},$$

     where, $\hat{X}_L = h^{-1}(u_{i,L}^T \hat{\beta}_L)$, $\hat{\beta}_L$ and $\hat{\phi}_L$ are the fitted values, MLE and dispersion estimate from a "large" model and $\hat{X}_S = h^{-1}(u_{i,S}^T \hat{\beta}_S)$ and $\hat{\beta}_S$ are the fitted values and MLE from a smaller, "nested" model (i.e., the smaller model contains a subset of the predictors in the "large" model). [NOTE: May want to use $\hat{\phi}_{full}$ for $\hat{\phi}_L$ in cases where both models under investigation are subsets of some larger, "full" model.]

6. Model Selection and Analysis of Deviance *(Continued)*:

- If small model has q of the p predictors used in the large model:

$$D^*(\hat{X}_S, \hat{X}_L) \sim \chi^2_{(p-q)}$$

  under the hypothesis that the small model is an adequate explanation. Thus, we reject the small model at significance level α (i.e., we accept that the larger model is a significantly better fit to the data) if $D^*(\hat{X}_S, \hat{X}_L) \geq \chi^2_{(p-q)}(1-\alpha)$.

- *Example (Continued)*: For our model of expected claims using both vehicle age and policyholder age:

$$\begin{aligned} D(\hat{X}_L, X) &= -2\sum_{i=1}^{n} \ln\{X_i(\hat{\beta}_0 + \hat{\beta}_1 p_i + \hat{\beta}_2 v_i)\} \\ &\quad + 2\sum_{i=1}^{n}\{X_i(\hat{\beta}_0 + \hat{\beta}_1 p_i + \hat{\beta}_2 v_i) - 1\} \\ &= 12.43122, \end{aligned}$$

  Similarly, for model with only policyholder age: $D(\hat{X}_S, X) = 12.72$. Therefore, deviance statistic for testing adequacy of smaller model is (remembering that $\phi = 1$ in this case) $D^\star(\hat{X}_S, \hat{X}_L) = 12.72 - 12.43 = 0.29$. This value is less than $\chi^2_{(1)}(0.95) = 3.84$. Thus, model without vehicle age is adequate. Indeed, model with no predictors (just intercept term) has residual deviance 16.50. Thus, deviance statistic is 16.50-12.43=4.07 which is less than $\chi^2_{(2)}(0.95) = 5.99$, indicating that we could get away without either predictor.

7. Model Diagnostics:

   - Residuals, $e_i = X_i - \hat{X}_i$ are heteroscedastic even if model fits since $Var(X_i) = \phi V(\mu_i)$ is not necessarily constant.
   - Pearson residuals:
$$r_i = \frac{e_i}{\sqrt{V(\hat{X}_i)}}.$$
   - Deviance residuals: Set $D_i = 2X_i\{b(X_i) - b(\hat{X}_i)\} - 2\{c(X_i) - c(\hat{X}_i)\}$, deviance residuals are:
$$d_i = \begin{cases} \sqrt{D_i} & \text{if } X_i > \hat{X}_i \\ -\sqrt{D_i} & \text{if } X_i < \hat{X}_i \end{cases}.$$
     Note $\sum_{i=1}^n d_i^2 = D(\hat{X}, X)$ (similar to sum of squared residuals equal to SSE). Also, $E\big(\sum_{i=1}^n d_i^2\big) \approx (n-p)\phi$, so can use deviance residuals to estimate ϕ as
$$\hat{\phi} = \frac{1}{n-p} E\left(\sum_{i=1}^n d_i^2\right)$$
   - Typical to plot residuals versus $\hat{\eta}_i = u_i^T \hat{\beta}$ rather than $\hat{X}_i$, since usually gives better visual assessment. Should appear homoscedastic. If not, data is *overdispersed*.

7. Model Diagnostics *(Continued)*:

- To fix *overdispersion* can:

  · Change link function. Check link using plot of $h(X_i)$ versus $\hat{\eta}_i = u_i^T \hat{\beta}$, should look linear if link is correct

  · Change $V(\mu)$ structure using new error distribution or varying dispersion parameters ϕ_i (amounts to weighting the data). Check $V(\mu)$ using plot of $|r_i|$ or $|d_i|$ versus $\hat{\eta}_i$, should show no trends if variance function is correct.

  · Change predictor structure (e.g., include/exclude predictors or change predictor scale, check latter using plot of $z_i = h(X_i) - u_i^T \hat{\beta} + u_{ij} \hat{\beta}_j$ versus u_{ij}, should look linear if predictor scale is correct).

8. Influence and Outliers:

- Leverages h_{ii} are the diagonal elements of:

$$H = \hat{W}^{1/2}U(U^T\hat{W}U)^{-1}U^T\hat{W}^{1/2}$$

- Can show $\sum_{i=1}^n h_{ii} = p$, so observation has high leverage when $h_{ii} > 2p/n$.
- Also, can show:

$$Var(r_i) \approx Var(d_i) \approx \hat{\phi}(1 - h_{ii}).$$

So, can create *Studentised* residuals. Easier to use to detect outliers. Also can construct *deleted* residuals (see Course Notes).

- Good overall measure of influence is *Cook's distance*:

$$C_i = \frac{(\hat{\beta}_{-i} - \hat{\beta})^T U^T \hat{W} U(\hat{\beta}_{-i} - \hat{\beta})}{p\hat{\phi}},$$

where $\hat{\beta}_{-i}$ is MLE from model fit without the i^{th} data point. Typically, use a barplot of C_i values and look for any which are excessively large.

Generalised Linear Models

9. Construction of Mixture Distributions:

- *Example (Continued)*: Suppose our portfolio has 1000 policies broken down by vehicle and policyholder age as follows:

Policyholder Ages (yrs)	Vehicle Ages (yrs) 1	2	5	10
25	135	105	40	20
35	115	125	20	40
45	70	90	30	10
55	80	80	10	30

Then $\hat{\mu}_i = \frac{1}{-4.2614\times10^{-4}+5.2056\times10^{-5}v_i+3.8283\times10^{-5}p_i}$ gives the fitted exponential mean for each category as:

Policyholder Ages (yrs)	Vehicle Ages (yrs) 1	2	5	10
25	\$1715.25	\$1574.65	\$1263.86	\$951.02
35	\$1035.37	\$982.42	\$851.74	\$697.18
45	\$741.47	\$713.91	\$642.30	\$550.30
55	\$577.53	\$560.67	\$515.54	\$454.54

Thus, since within each vehicle and policyholder age combination we have assumed claims are exponentially distributed, the "portfolio-wide" mixture distribution has *pdf*:

$$f_X(x) = \sum_{i=1}^{4}\sum_{j=1}^{4}\left(\frac{N_{ij}}{1000}\right)\frac{1}{\hat{\mu}_{ij}}e^{-x/\hat{\mu}_{ij}}$$

where N_{ij} is the number of policies with vehicle age category i and policyholder age category j (i.e., the entries in the first table above) and $\hat{\mu}_{ij}$ are the corresponding entries in the fitted mean table. [NOTE: This construction assumes that each policy is equally likely to make the next claim. If this is not the case, then we need additional weights to deal with the variable claim rates of the policies. This can be done my modelling the number of claims made on policies with various predictor values using, for example, a Poisson GLM.]