STAT3035/8035
Tutorial 5

Marco Li

Contact: qingyue.li@anu.edu.au

Outline

Generalised linear models

- Response data: $X_1, \ldots, X_n$ with $E(X_i) = \mu_i$ - Historical claim amount data
- Predictor vectors (p-dimensional) : $u_1, \ldots, u_n$ with $u_i = (u_{i1}, \ldots, u_{ip})^T$ - Characteristics of policy holders (age, gender, etc.)
- 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 $\beta = (\beta_1, \ldots, \beta_p)$ - Key difference from linear models
- 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 - Assumption of claim amount distributions

Exponential families

- 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. $\phi = 1$ for exponential, Poisson or binomial distributions.
  - 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 $d(x,\phi) = \left(\phi^{-1}-1\right)\ln(x) - \phi^{-1}\ln(\phi) - \ln\left\{\Gamma\left(\phi^{-1}\right)\right\}$

Exponential families

- Some general properties of Exponential Families:

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

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

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

- Canonical link function
Log-likelihood (for estimating parameters in GLM) is:

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

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

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

Estimation of GLM parameters

- Using IRLS to calculate MLE of parameters (not examinable)
- SEs and CIs of parameters
  - $\text{Var}(\hat{\beta}) \approx \phi \left(U^T W U\right)^{-1} \approx \phi \left(U^T \hat{W} U\right)^{-1}$ where U is the matrix of the predictors and W is the diagonal matrix, see lecture slides for details
  - Predicted value at $u_0 = (u_{01}, \ldots, u_{0p})$ [Estimate for a new policy holder with characteristics u_0] is

$$\hat{\mu}\left(u_0\right) = \mathbb{E}(X|u_0) = h^{-1}\left(\hat{\eta}_0\right) = h^{-1}\left(u_0^T \hat{\beta}\right)$$

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

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

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

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

  - 95% CI for $\hat{\mu}\left(u_0\right)$ is $\left\{h^{-1}\left(c_l\right), h^{-1}\left(c_u\right)\right\}$ [Need to reverse endpoints if $h(\cdot)$ is decreasing function]

Model selection

- Residual deviance for a specific model with parameter β is:

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

- (Scaled) Deviance Statistic:

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

where, $\hat{X}_L$ is the fitted value from a "large" model and $\hat{X}_S$ is the fitted value from a smaller, "nested" model (i.e., the smaller model contains a subset of the predictors in the "large" model)

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

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

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 $D^*\left(\hat{X}_S, \hat{X}_L\right) \geq \chi^2_{(p-q)}(1-\alpha)$

Outline

Question

(Gamma GLM and Discrete Independent Variables) Suppose claims (in thousands of dollars) on home insurance policies are gamma distributed with mean, μ_i, and common shape parameter α. Further, suppose the mean claim size depends on the age of the house (in years), the size of the house (in square meters) and where the house is located according to a GLM with link function:

$$\ln(\mu_i) = \beta_1 \text{Age}_\text{i} + \beta_2 \text{Size}_\text{i} + \beta_3 I_{\text{Urban},i} + \beta_4 I_{\text{Suburban},i} + \beta_5 I_{\text{Rural},i}$$

where $I_{A,i}$ is an indicator of whether the i^{th} insured house is located in area A (e.g. a suburban home would have $I_{\text{Urban},i} = 0, I_{\text{Suburban},i} = 1$ and $I_{\text{Rural},i} = 0$). Based on a sample of data, a GLM produced a fitted model with a residual deviance of 15.971 and maximum likelihood estimates of the coefficients:

$$\widehat{\beta}_1 = -0.1114, \widehat{\beta}_2 = 0.0094, \widehat{\beta}_3 = 2.3270, \widehat{\beta}_4 = 2.4381, \widehat{\beta}_5 = 1.2421$$

Moreover, the dispersion parameter estimate for the GLM was $\widehat{\phi} = 0.6782$ and the unscaled variance-covariance matrix of the parameter estimate, $\left(U^T \widehat{W} U\right)^{-1}$, was:

	Age	Size	Urban	Suburban	Rural
Age	0.0027	0.000000	–0.0146	–0.0146	–0.0146
Size	0.0000	0.000022	–0.0033	–0.0033	–0.0033
Urban	–0.0146	–0.003300	0.6888	0.5777	0.5777
Suburban	–0.0146	–0.003300	0.5777	0.6888	0.5777
Rural	–0.0146	–0.003300	0.5777	0.5777	0.6888

Part (a)

A portfolio is constructed of 1000 policies with the following breakdown:

Urban			Suburban			Rural		
Size	Age	# of Policies	Size	Age	# of Policies	Size	Age	# of Policies
100	10	30	150	1	170	150	5	50
150	15	50	200	10	250	200	5	150
200	25	100	225	5	130	225	1	70

Use this table and the given results of the GLM to estimate the expected amount of the next claim made on the portfolio and its standard deviation. [NOTE: Assume that the claim rates for all policies are equal, so that each policy is equally likely to generate the next claim.]

Gamma Distribution

$$f_x(x;\alpha,\theta) = \frac{1}{\theta^\alpha \Gamma(\alpha)} x^{\alpha-1} \exp\left(-\frac{x}{\theta}\right), \quad x>0, \alpha>0, \theta>0$$

$$E_{\alpha,\theta}(X) = \alpha\theta \quad E_{\alpha,\theta}\left(X^2\right) = \alpha(\alpha+1)\theta^2$$

$$\mathrm{Var}_{\alpha,\theta}(X) = \alpha\theta^2 \quad m_X(t) = (1-t\theta)^{-\alpha}$$

where $\Gamma(\alpha) = \int_0^\infty x^{\alpha-1}e^{-x}dx$, and $\Gamma(\alpha+1) = \alpha\Gamma(\alpha)$, for $\alpha > 0$
If $\alpha = k$, a positive integer, then $\Gamma(k) = (k-1)!$

Solution (a)

Solution (a)

First, we need to construct a table of the predicted average claim size for each group of policies in the portfolio. We do this by using predicted values from the GLM. For example, the policies on 100m^2 urban homes which are 10 years old have a mean claim size of $\mu_{\{100\text{m}^2,\text{Urban},10\text{year}\}} = \exp\{-0.1114 \times 10 + 0.0094 \times 100 + 2.3270 \times 1 + 2.4381 \times 0 + 1.2421 \times 0\} = 8.611$. Similarly, for the rest of the policies, the mean claim sizes are calculated to be:

Urban			Suburban			Rural		
Size	Age	μ_i	Size	Age	μ_i	Size	Age	μ_i
100	10	8.611	150	1	41.959	150	5	8.126
150	15	7.893	200	10	24.633	200	5	13.002
200	25	4.145	225	5	54.386	225	1	25.680

So, the "portfolio-wide" mixture distribution in this case is just the weighted average of 9 different gamma distributions, each with shape parameter $\alpha = 1/\phi = 1/0.6782 = 1.4745$ and the various means (For gamma distribution, $\mu = \alpha\theta$, or $\theta = \mu/\alpha$). In other words, the density of the mixture distribution is (reparameterization needed):

$$f_X(x) = \sum_{i=1}^{9} \frac{n_i}{1000} \times \frac{1.4745^{1.4745}}{\mu_i^{1.4745}\Gamma(1.4745)} x^{0.4745} e^{-1.4745x/\mu_i}$$

where n_i and μ_i are the number of policies and mean claim size in each of the 9 policy groups of the portfolio. Note that we use the value $n_i/1000$ in this case, since we have assumed that each policy has the same claim rate, and thus is equally likely to produce the next claim.

Solution (a)

Using this distribution, or by recalling that the mean of a mixture distribution is just the expected value of the conditional means, we can find the overall expectation as:

$$\mathbb{E}X = \mathbb{E}[\mathbb{E}(X|\text{Policy Type})] = \frac{30}{1000} \times 8.611 + \cdots + \frac{70}{1000} \times 25.680 = 25.583$$

Finally, we recall that the second moment of a gamma is $\alpha(\alpha+1)\theta^2 = (\alpha+1)\mu^2/\alpha$. So,

$$\begin{aligned}\mathbb{E}X^2 &= \mathbb{E}\left[\mathbb{E}\left(X^2|\text{Policy Type}\right)\right] = \frac{30}{1000} \times \frac{2.4745}{1.4745} \times 8.611^2 + \cdots + \frac{70}{1000} \times \frac{2.4745}{1.4745} \times 25.680^2 \\ &= 1539.559\end{aligned}$$

Thus, the desired standard deviation is $\sqrt{1539.559 - 25.583^2} = 29.750$. Notice that the standard deviation is larger than the mean! This is an indication that the distribution in question is extremely skewed. Not surprising for home insurance, since most claims are about minor damage or theft, but some require the rebuilding of the whole house! Note that variances do not "mix" together like moments do! You can use the conditional variance formula, $\mathbb{V}X = \mathbb{E}[\mathbb{V}(X|\text{Policy Type})] + \mathbb{V}[\mathbb{E}(X|\text{Policy Type})]$ to get the above answer, however the final variance term is rather tedious to calculate.

Part (b)

Someone suggests that the average claim amounts for urban and suburban houses of the same size and age should be the same and has fit a model with a combined indicator for these two groups. The residual deviance for this smaller model was 16.025. Using this information, do you think that the suggestion that claims from urban and suburban houses of the same size and age are comparable in size is reasonable? Why or why not?

Chi-square Distribution Table

d.f.	.995	.99	.975	.95	.9	.1	.05	.025	.01
1	0.00	0.00	0.00	0.00	0.02	2.71	3.84	5.02	6.63
2	0.01	0.02	0.05	0.10	0.21	4.61	5.99	7.38	9.21
3	0.07	0.11	0.22	0.35	0.58	6.25	7.81	9.35	11.34

Solution (b)

Solution (b)

To test the adequacy of the smaller model, we use the scaled drop in deviance statistic:

$$D^* = \frac{16.025 - 15.971}{0.6782} = 0.0796$$

Comparing this to $\chi^2_{(1)}$ critical values at 95% level, 3.841459, clearly indicates that we should accept the hypothesis that the smaller model is adequate. Thus, it seems reasonable to assume that the claim sizes for urban and suburban houses with the same age and size have the same distribution.

Part (c)

(c) Suppose that you think the person from part (b) is not very reliable and you do not trust their residual deviance calculation. Can you test their claim using only the output from the model output given in the introduction to the exercise, which you do trust?

Part (c)

(c) Suppose that you think the person from part (b) is not very reliable and you do not trust their residual deviance calculation. Can you test their claim using only the output from the model output given in the introduction to the exercise, which you do trust?

Solution: An alternate method would be to test whether $\beta_3 - \beta_4 = 0$. The obvious estimate of this difference is $\hat{\beta}_3 - \hat{\beta}_4 = 2.3270 - 2.4381 = -0.1111$. To see whether this is significantly different from 0. we need to calculate the standard error of $\hat{\beta}_3 - \hat{\beta}_4$. To do so, notice that $\hat{\beta}_3 - \hat{\beta}_4 = c^T\hat{\beta}$, where $c^T = (0, 0, 1, -1, 0)$. We know that

$$\mathbb{V}\left(c^T\widehat{\beta}\right) = \widehat{\phi}c^T\left(U^T\hat{W}U\right)^{-1}c = 0.1507$$

Therefore, the appropriate standard error is $\sqrt{0.1507} = 0.3882$. Clearly, then, the value of -0.1111 is not significantly different from 0 (as it is only 0.1111/0.3882=0.2862 standard errors below 0). So, again, we see that we should accept the hypothesis that claims from urban and suburban houses of the same age and size have the same distribution.

Part (d)

(Advanced) The chosen GLM model structure has no intercept, can you explain why this must be the case? Moreover, suppose that we were to re-write the link structure of the model in the form:

$$\ln(\mu_i) = \beta_0 + \beta_1 \,\text{Age}_\text{i} + \beta_2 \,\text{Size}_\text{i} + \beta_3 I_{\text{Urban},i} + \beta_4 I_{\text{Suburban},i}$$

What would the new maximum likelihood estimates for the parameters be?

Part (d)

(Advanced) The chosen GLM model structure has no intercept, can you explain why this must be the case? Moreover, suppose that we were to re-write the link structure of the model in the form:

$$\ln(\mu_i) = \beta_0 + \beta_1 \text{ Age}_\text{i} + \beta_2 \text{ Size}_\text{i} + \beta_3 I_{\text{Urban},i} + \beta_4 I_{\text{Suburban},i}$$

What would the new maximum likelihood estimates for the parameters be?

Solution: If we included an intercept in the model with all three indicators, we would have an over-parameterized model, since their is an exact linear dependency between the three indicators and the intercept term (i.e., the sum of the three indicators equals the intercept). If we re-fit the model in the new form suggested, the intercept term now represents the rural homes (since the other indicators are zero in this case), so $\widehat{\beta}_0 = 1.2421$, while the β_3 and β_4 coefficients now represent the additional effect to mean claim size associated with being urban or suburban, respectively. In other words, $\widehat{\beta}_3 = 2.3270 - 1.2421 = 1.0849$ and $\widehat{\beta}_4 = 2.4381 - 1.2421 = 1.1960$