Lab 5: Generalised Linear Models

ACTL3142 and ACTL5110

Questions

Conceptual Questions

Components of a GLM

1. For the following members of the exponential dispersion family, give the density (including the domain), the mean and the variance:

   a. Normal(μ, σ^2)

   b. Poisson(μ)

   c. Binomial(m, p)

   d. Negbin(r, p)

   e. Gamma(α, β)

   f. Inverse Gaussian(α, β)

   Solution

2. $\star$ The density of the Binomial distribution is given by

$$f(y;p) = \frac{n!}{(n-y)!y!} p^y (1-p)^{n-y}$$

   Show that the Binomial distribution is a member of the exponential dispersion family with density

$$f(y;\theta,\psi) = \exp\left[\frac{y\theta - b(\theta)}{\psi} + c(y;\psi)\right].$$

   a. Give expressions for $b(\theta)$, $c(y;\psi)$ and ψ.

   b. List the three constituent parts of a generalized linear model.

c. Find the expression for the deviance of a binomial model.

Solution

Deviance and Scaled Deviance

3. $\star$ Verify that

$$\frac{D}{\psi} = -2\log\frac{L(\mathbf{y};\mu)}{L(\mathbf{y};\mathbf{y})} = \frac{2}{\psi}\sum_i \left(y_i \log\frac{y_i}{\hat{\mu_i}} - (y_i - \hat{\mu_i})\right)$$

is the scaled deviance for a Poisson distribution.

Solution

4. Verify that

$$\frac{D}{\psi} = -2\log\frac{L(\mathbf{y};\mu)}{L(\mathbf{y};\mathbf{y})} = \frac{2}{\psi}\sum_i \left(-\log\frac{y_i}{\hat{\mu_i}} + \frac{(y_i - \hat{\mu_i})}{\hat{\mu_i}}\right)$$

is the scaled deviance for a gamma distribution.

Solution

5. Show that the deviance for an Inverse Gaussian distribution has the following form:

$$D = \sum_{i=1}^{n}\frac{1}{\hat{\mu_i}^2}\frac{(y_i - \hat{\mu_i})^2}{\hat{\mu_i}}.$$

Solution

Fit a GLM and Evaluate the quality of a model

6. $\star$ (Question #9, ACTL3003/5106 Final Exam 2005)

A random variable Y is said to have an exponential dispersion model if its density can be expressed in the form

$$f_Y(y;\theta,\psi) = \exp\left(\frac{y\theta - b(\theta)}{\psi} + c(y;\psi)\right)$$

where θ and ψ are parameters and $b(\cdot)$ and $c(\cdot;\cdot)$ are both functions

Automobile insurance claims experience data of a French insurance company for a two-year period, beginning January 2001 and ending December 2002, is being modeled using a Generalized Linear Model (GLM) framework.

Assume that the number of claims for risk class i, Y_i, has a Poisson distribution with probability density function of the form

$$f_Y(y;\mu) = \frac{\mathrm{e}^{-\mu}\mu^y}{y!}, \text{ for } y = 0, 1, 2, \ldots$$

where its mean $\mu > 0$ is related to the variables `SEX`, `VEH_AGE`, `AGE`, and `LOYALTY` as

$$\log(\mu_i) = \log(\texttt{EXP}_i) + \beta_0 + \beta_1 \texttt{SEX}_i + \beta_2 \texttt{VEH\_AGE}_i(1) + \beta_3 \texttt{VEH\_AGE}_i(2) + \beta_4 \texttt{AGE}_i + \beta_5 \texttt{LOYALTY}_i$$

where detailed description of variables and their definitions are found below:

Variable	**Description**
SEX	**1** = male; **2** = female.
VEH_AGE	**1** = less than 1 year; **2** = between 1-2 years; **3** = 2 years and more.
AGE	**1** = 20 years and below; **2** = above 20 years.
LOYALTY	**1** = has been client for past 36 months; **0** = otherwise.
Y	total number of claims
EXP	total number of policies exposed to claims

Note that the variables `VEH_AGE`(1) and `VEH_AGE`(2) in the regression equation are the respective indicator variables for `VEH_AGE` of types 1 and 2.

SAS output for running `PROC GENMOD` on the data is provided below.

```
                              The GENMOD Procedure

                         Analysis Of Parameter Estimates

                                  Standard      Wald 95%          Chi-
    Parameter      DF  Estimate     Error   Confidence Limits   Square  Pr > ChiSq

    veh_age     1   1   -0.0136    0.0974   -0.2044    0.1772     0.02      0.8889
    veh_age     2   1   -0.0009    0.1394   -0.2741    0.2723     0.00      0.9948
    veh_age     3   0    0.0000    0.0000    0.0000    0.0000      .          .
    driver_age  1   1    0.8190    0.3059    0.2195    1.4185     7.17      0.0074
    driver_age  2   0    0.0000    0.0000    0.0000    0.0000      .          .
    loyalty     1   1   -0.1877    0.2306   -0.6396    0.2642     0.66      0.4155
    loyalty     2   0    0.0000    0.0000    0.0000    0.0000      .          .
    Scale           0    1.0000    0.0000    1.0000    1.0000

NOTE: The scale parameter was held fixed.

                        LR Statistics For Type 1 Analysis

                                                  Chi-
         Source          Deviance        DF     Square     Pr > ChiSq

         Intercept       199.9268
         sex             194.5739         -          -              -
         veh_age         194.5571         -          -              -
         driver_age      188.9746         -          -              -
         loyalty         188.3470         -          -              -
```

a. Show that the Poisson distribution can be written in exponential dispersion form. Identify the dispersion and canonical parameters, ψ and θ respectively, in terms of μ, to the extent possible.

b. Derive the expression for the deviance of the Poisson GLM model.

c. Based on the deviances provided in the SAS output, analyze the adequacy of the model

d. Explain the meaning of overdispersion in the context of a Poisson GLM model.

Solution

7. An insurance company has a set of n risks ($i = 1, 2, \ldots, n$) for which it has recorded the number of claims per month, Y_{ij}, for m months ($j = 1, 2, \ldots, m$). It is assumed that the

number of claims for each risk, for each month, are independent Poisson random variables with

$$\mathbb{E}(Y_{ij}) = \mu_{ij}.$$

These random variables are modelled using a Generalized Linear Model, with

$$\log \mu_{ij} = \beta_i, \text{ for } i = 1, 2, \ldots, n.$$

a. Derive the maximum likelihood estimator of β_i

b. Show that the deviance for this model is

$$2 \sum_{i=1}^{n} \sum_{j=1}^{m} \left(y_{ij} \log \frac{y_{ij}}{\bar{y}_i} - (y_{ij} - \bar{y}_i) \right)$$

where $\bar{y}_i = \frac{1}{m} \sum_{j=1}^{n} y_{ij}$.

c. A company has data for each month over a 2 year period. For one risk, the average risk of claims per month was 17.45. In the most recent month for this risk, there were 9 claims. Calculate the contribution that this observations makes to the deviance.

Solution

8. There are m male drivers in each of three age groups, and data on the number of claims made during the last year are available. Assume that the numbers of claims are independent Poisson random variables. If Y_{ij} is the number of claims for the jth male driver in group i ($i = 1, 2, 3; j = 1, 2, \ldots, m$), let $\mathbb{E}(Y_{ij}) = \mu_{ij}$ and suppose $\log(\mu_{ij}) = \alpha_i$.

a. Show that this is a Generalized Linear Model, identifying the link function and the linear predictor.

b. Determine the log-likelihood, and the maximum likelihood estimators of α_i for $i = 1, 2, 3$.

c. For a particular data set with 20 observations in each group, several models are fitted, with deviances as shown below:

```
Link function
```

Model 1 $\log(\mu_{ij}) = \alpha_i$ 60.40

Model 2 $\log(\mu_{ij}) = \begin{cases} \alpha, & \text{if } i = 1, 2 \\ \beta, & \text{if } i = 3 \end{cases}$ 61.64

Model 3 $\log(\mu_{ij}) = \alpha$ 72.53

i. Determine whether or not model 2 is a significant improvement over model 3, and whether or not model 1 is a significant improvement over model 2

ii. Interpret these three models

Solution

9. An insurance company tested for claim sizes under two factors, i.e. **CAR**, the insurance group into which the car was placed, and **AGE**, the age of the policyholder (i.e. two-way contingency table). It was assumed that the claim size y_i follows a gamma distribution i.e.

$$f(y_i) = \frac{1}{\Gamma(\nu_i)\, y_i} \left(\frac{y_i\, \nu_i}{\mu_i} \right)^{\nu_i} \exp\left(-\frac{y_i\, \nu_i}{\mu_i} \right) \quad \text{for} \;\; y_i \geq 0, \; \mu_i > 0, \; \nu_i = 1$$

with a log-link function. Analysis of a set of data for which $n = 8$ provided the following SAS output:

Observation	**Claim size**	**CAR type**	**Age group**	**Pred**	**Xbeta**	**Resdev**
1	27	1	1	25.53	3.24	0.30
2	16	1	2	24.78	3.21	-1.90
3	36	1	1		3.41	1.03
4	45	1	2	38.09	3.64	1.11
5	38	2	1	40.85	3.71	-0.46
6	27	2	2	36.97	3.61	-1.73
7	14	2	1		2.45	0.69
8	6	2	2	14.59	2.68	-2.55

Calculate the fitted claim sizes missing in the table.

Solution

Applied questions

1. $\star$ In this question, the vehicle insurance data set is used, `dataCar`. This data set is based on one-year vehicle insurance policies taken out in 2004 or 2005. There are 67856 policies of which 4624 had at least one claim.

   The data frame `dataCar` contains claim occurrence `clm`, which takes value 1 if there is a claim and 0 otherwise. The variable `veh_value` represents the vehicle value which takes value from $\$0 - \$350,000$. We will not be concerned about other variables at the moment.

In this question, we will build a logistic regression model to apply to the vehicle insurance data set. Previous study has shown that the relationship between the likelihood of occurrence of a claim and vehicle value are possibly quadratic or cubic.

a. Suppose the relationship between vehicle value and the probability of a claim is cubic, formulate the model and test significance of the coefficients.

b. Which variable(s) is/are significant at the 1% level?

c. Use AIC to determine which model is the best model: Linear, quadratic or cubic.

Solution

2. ⋆ Third party insurance is a compulsory insurance for vehicle owners in Australia. It insures vehicle owners against injury caused to other drivers, passengers or pedestrians, as a result of an accident.

   In this question, the third party claims data set `Third_party_claims.csv` is used. This data set records the number of third party claims in a twelve-month period between 1984-1986 in each of 176 geographical areas (local government areas) in New South Wales, Australia.

   a. Now consider a model for the number of claims (`claims`) in an area as a function of the number of accidents (`accidents`). Produce a scatter plot of `claims` against `accidents`. Do you think a simple linear regression model is appropriate?

   b. Fit a simple linear regression to the model and use the `plot` command to produce residual and diagnostic plots for the fitted model. What do the plots tell you?

   c. Now fit a Poisson regression model with `claims` as response and `log(accident)` as the predictor (include `offset=log(population)` in your code). Check if there is overdispersion in the model by computing the estimate of ψ.

   d. Now fit the regression model by specifying `family=quasipoisson`. Comment on the estimates of the parameters and their standard errors.

   Solution

Solutions

Conceptual Questions

Components of a GLM

1. See the lecture notes. The density and the mean are given in the table and the variance can be derived easily from the table with:

$$\sigma^2 = \psi \cdot V(\mu).$$

   Try to map some of the densities into the exponential family formulation.

2. You ought to be able to verify that the Binomial belongs to the family of Exponential Dispersion with

$$b(\theta) = n \log\left(1 + \mathrm{e}^{\theta}\right),\ c(y;\psi) = \log\left(\frac{n!}{(n-y)!y!}\right),\ \text{and } \psi = 1.$$

   You should also be able to show that

$$\theta = \log\frac{p}{1-p} = \log\frac{\mu}{n-\mu}.$$

   The three components of a generalized linear model are:

   a. *Stochastic Component*: The observations Y_i are independent and each follows an Exponential Dispersion distribution.
   b. *Systematic Component*: Every observation has a linear predictor $\eta_i = \sum_j x_{ij}\beta_j$ where x_{ij} denotes the jth explanatory variable, and
   c. *Link Function*: The expected value $\mathbb{E}[Y_i] = \mu_i$ is linked to the linear predictor η_i by the link function $\eta_i = g(\mu_i)$.

   Now to find the deviance of the binomial (deviance is also the scaled deviance since $\psi = 1$), we have

$$\begin{aligned}
\frac{D}{\psi} &= -2\log\left(\frac{\prod\limits_i \frac{n!}{(n-y_i)!y_i!}\left(\frac{\widehat{\mu}_i}{n}\right)^{y_i}(1-\frac{\widehat{\mu}_i}{n})^{(n-y_i)}}{\prod\limits_i \frac{n!}{(n-y_i)!y_i!}\left(\frac{y_i}{n}\right)^{y_i}(1-\frac{y_i}{n})^{(n-y_i)}}\right)\\
&= -2\log\left(\prod_i\left(\frac{\widehat{\mu}_i}{y_i}\right)^{y_i}\left(\frac{n-\widehat{\mu}_i}{n-y_i}\right)^{n-y_i}\right)\\
&= -2\sum_i\left[y_i\log\left(\frac{\widehat{\mu}_i}{y_i}\right) + (n-y_i)\log\left(\frac{n-\widehat{\mu}_i}{n-y_i}\right)\right]\\
&= 2\sum_{i=1}^{n}\left[y_i\log\left(\frac{y_i}{\widehat{\mu}_i}\right) + (n-y_i)\log\left(\frac{n-y_i}{n-\widehat{\mu}_i}\right)\right].
\end{aligned}$$

Deviance and Scaled Deviance

3. We know that if D denotes the deviance, the scaled deviance is

$$\frac{D}{\psi} = -2\log\left(\widehat{L}/\tilde{L}\right)$$

by definition, where $\widehat{L}$ is the likelihood computed using the MLE's $\widehat{\mu}$ under the current model replacing the μ, while $\widetilde{L}$ is the likelihood computed with the μ replaced by the estimates under the "full model", hence the actual observations y, in view of the remarks just below (8.22). To show that (8.23) results from this is basic algebra. To see this, note that

$$\begin{aligned}\frac{D}{\psi} &= -2\log\left(\frac{\prod_{i=1}^{n} \mathrm{e}^{-\widehat{\mu}_i}\widehat{\mu}_i^{y_i}/y_i!}{\prod_{i=1}^{n} \mathrm{e}^{-y_i}y_i^{y_i}/y_i!}\right)\\ &= -2\log\left(\prod_{i=1}^{n} \mathrm{e}^{-(\widehat{\mu}_i-y_i)}\left(\frac{\widehat{\mu}_i}{y_i}\right)^{y_i}\right)\\ &= -2\sum_{i=1}^{n}\left[-(\widehat{\mu}_i-y_i)+y_i\log\left(\frac{\widehat{\mu}_i}{y_i}\right)\right]\\ &= 2\sum_{i=1}^{n}\left[(\widehat{\mu}_i-y_i)-y_i\log\left(\frac{\widehat{\mu}_i}{y_i}\right)\right]\end{aligned}$$

4. To show that (8.26) results, following the discussion in the previous problem, we can verify that, for exponential dispersion models, the scaled deviance can be expressed as

$$\begin{aligned}\frac{D}{\psi} &= -2\log\left(\widehat{L}/\widetilde{L}\right)\\ &= 2\sum_{i=1}^{n}\left[\frac{y_i\left(\widetilde{\theta}_i-\widehat{\theta}_i\right)}{\psi} - \frac{b\left(\widetilde{\theta}_i\right)-b\left(\widehat{\theta}_i\right)}{\psi}\right].\end{aligned}$$

For Gamma, we have $\theta(\mu) = -1/\mu$ and $b(\theta) = -\log(-\theta) = \log\mu$, we than have

$$\begin{aligned}\frac{D}{\psi} &= 2\sum_{i=1}^{n}\left[\frac{y_i\left(1/\widehat{\mu}_i - 1/y_i\right)}{\psi} - \frac{\log y_i - \log\widehat{\mu}_i}{\psi}\right]\\ &= \frac{2}{\psi}\sum_{i=1}^{n}\left[\frac{y_i-\widehat{\mu}_i}{\widehat{\mu}_i} - \log\left(y_i/\widehat{\mu}_i\right)\right].\end{aligned}$$

Now, if the scale parameter were different for each observation according to some weight w_i, then it is easy to verify.

5. Recall that the scaled deviance for any member of the Exponential Dispersion family has the form

$$\begin{aligned}\frac{D}{\psi} &= 2\left[\ell\left(\widetilde{\theta};\mathbf{y}\right) - \ell\left(\widehat{\theta};\mathbf{y}\right)\right] \\ &= \frac{2}{\psi}\sum_{i=1}^{n}\left[\left(y_i\widetilde{\theta}_i - b\left(\widetilde{\theta}_i\right)\right) - \left(y_i\widehat{\theta}_i - b\left(\widehat{\theta}_i\right)\right)\right]\end{aligned}$$

where for the Inverse Gaussian, we have verified (in lecture) that

$$\psi = \beta/\alpha^2, \quad \theta = -\frac{1}{2}\left(\frac{\beta}{\alpha}\right)^2 = -\frac{1}{2\mu^2}, \quad \text{and} \quad b\left(\theta\right) = -\sqrt{-2\theta} = -1/\mu$$

Thus, the deviance can be expressed as

$$\begin{aligned}D &= 2\sum_{i=1}^{n}\left[\left(y_i\left(-\frac{1}{2y_i^2}\right) + \frac{1}{y_i}\right) - \left(y_i\left(-\frac{1}{2\widehat{\mu}_i^2}\right) - \frac{1}{\widehat{\mu}_i}\right)\right] \\ &= 2\sum_{i=1}^{n}\left[\frac{1}{2y_i}\left(1 + \frac{y_i^2}{\widehat{\mu}_i^2} - \frac{2y_i}{\widehat{\mu}_i}\right)\right] = \sum_{i=1}^{n}\left[\frac{1}{y_i}\left(1 - \frac{y_i}{\widehat{\mu}_i}\right)^2\right] \\ &= \sum_{i=1}^{n}\left[\frac{1}{y_i}\left(\frac{\widehat{\mu}_i - y_i}{\widehat{\mu}_i}\right)^2\right] = \sum_{i=1}^{n}\frac{1}{\widehat{\mu}_i^2}\frac{1}{y_i}\left(\widehat{\mu}_i - y_i\right)^2.\end{aligned}$$

This gives the desired result.

Fit a GLM and Evaluate the quality of a model

6. a. For a Poisson(μ) distribution, its density can be expressed as

$$\begin{aligned}f_Y(y) &= \frac{\mathrm{e}^{-\mu}\mu_y}{y!} \\ &= \frac{1}{y!}\exp\left(-\mu + y\ln\mu\right) \\ &= \exp\left(y\ln\mu - \mu - \ln y!\right)\end{aligned}$$

Thus, we see it belongs to the exponential dispersion family with

$$\theta = \ln\mu,\ \psi = 1,\ b(\theta) = \mu,\ c(y;\psi) = -\ln y!$$

b. Recall that the scaled deviance for any member of the Exponential Dispersion family has the form

$$\begin{aligned}\frac{D}{\psi} &= 2\left[\ell\left(\widetilde{\theta};\mathbf{y}\right) - \ell\left(\widehat{\theta};\mathbf{y}\right)\right] \\ &= \frac{2}{\psi}\sum_{i=1}^{n}\left[\left(y_i\widetilde{\theta}_i - b\left(\widetilde{\theta}_i\right)\right) - \left(y_i\widehat{\theta}_i - b\left(\widehat{\theta}_i\right)\right)\right]\end{aligned}$$

where for the Poisson, we have, from part 1 above that

$$\theta = \ln \mu, \ \psi = 1, \ b(\theta) = \mu, \ c(y;\psi) = -\ln y!$$

Note that $\widetilde{\boldsymbol{\theta}} = \boldsymbol{y}$ under the full model. Thus the deviance can be expressed as

$$\begin{aligned} D &= 2\sum_{i=1}^{n} \left[(y_i(\ln y_i) - y_i) - (y_i(\ln \widehat{\mu}_i)) - \widehat{\mu}_i\right] \\ &= 2\sum_{i=1}^{n} \left[y_i(\ln y_i - \ln \widehat{\mu}_i) - (y_i - \widehat{\mu}_i)\right] \end{aligned}$$

This gives us the desired form of the deviance of the Poisson.

c. First, summarizing the deviance statistics below:

Model	Deviance	Differences	df	ROT: Accept if D1-D2>2(p-q)	Significant?
intercept	199.9268	-		-	-
sex	194.5739	5.3529	1	Accept	Yes
veh_age	194.5571	0.0168	2	Reject	No
driver_age	188.9746	5.5825	1	Accept	Yes
loyalty	188.3470	0.6276	1	Reject	No

[ROT means Rule-of-Thumb, so there is no need to look up chi-square table.] Thus, the GLM model appears to be suitable/adequate with `SEX` and `DRIVER_AGE` as considered significant predictor variables for claims.

d. Overdispersion is a phenomenon that sometimes occurs in data that are modeled using the Poisson distribution. This is because the mean and the variance for a Poisson are both equal to the Poisson parameter, and if the observed variance from the data is much larger than the observed mean, then there is potential problem of over-dispersion. If the estimate of the dispersion after fitting the data, as measured by the either the deviance or Pearson's chi-square, divided by the degrees of freedom, is not near 1, then the data may be overdispersed if the dispersion estimate is greater than 1, or underdispersed if the dispersion estimate is less than 1. A convenient and simple way to model this situation is to allow the variance function of the Poisson distribution to have a multiplicative overdispersion factor. That is, we let the variance function be so that there is scale parameter ϕ, as in example in class, we have $V(\mu) = \phi\mu$.

7. a. The likelihood is $\prod_{i,j} \dfrac{\mu_{ij}^{y_{ij}} \mathrm{e}^{-\mu_{ij}}}{y_{ij}!}$ and the log-likelihood is therefore

$$\begin{aligned}\ell(\boldsymbol{\beta}) &= \sum_{i=1}^{n}\sum_{j=1}^{m}(y_{ij}\log\mu_{ij}-\mu_{ij}-\log y_{ij}!)\\ &= \sum_{i=1}^{n}\sum_{j=1}^{m}\left(y_{ij}\beta_i-\mathrm{e}^{\beta_i}-\log y_{ij}!\right)\\ &= \sum_{i=1}^{n}\beta_i\sum_{j=1}^{m}y_{ij}-\sum_{i=1}^{n}\mathrm{e}^{\beta_i}m-\sum_{i=1}^{n}\sum_{j=1}^{m}\log(y_{ij})!.\end{aligned}$$

Applying first order condition:

$$\frac{\partial\ell(\boldsymbol{\beta})}{\beta_i}=\sum_{j=1}^{m}y_{ij}-m\mathrm{e}^{\beta_i}=0$$

so that

$$\mathrm{e}^{\beta_i}=\frac{1}{m}\sum_{j=1}^{m}y_{ij}\triangleq\overline{y}_i$$

and the MLE is

$$\widehat{\beta}_i=\log\overline{y}_i.$$

b. The deviance is

$$\begin{aligned}2\left[\ell(y;y)-\ell(y;\mu)\right] &= 2\left[\begin{array}{c}\sum_{i=1}^{n}\sum_{j=1}^{m}(y_{ij}\log y_{ij}-y_{ij}-\log y_{ij}!)\\ -\sum_{i=1}^{n}\sum_{j=1}^{m}(y_{ij}\log\overline{y}_i-\overline{y}_i-\log y_{ij}!)\end{array}\right]\\ &= 2\sum_{i=1}^{n}\sum_{j=1}^{m}\left(y_{ij}\log\frac{y_{ij}}{\overline{y}_i}-(y_{ij}-\overline{y}_i)\right).\end{aligned}$$

c. The contribution to the deviance in this case is

$$\begin{aligned}D_{ij} &= 2\left(y_{ij}\log\frac{y_{ij}}{\overline{y}_i}-(y_{ij}-\overline{y}_i)\right)\\ &= 2\cdot\left(9\log\frac{9}{17.45}-(9-17.45)\right)=4.98.\end{aligned}$$

8. a. If Y has a Poisson distribution with mean parameter μ, then its density can be written as

$$f(y;\mu)=\mathrm{e}^{-\mu}\mu^y/y!=\exp\left(\frac{y\log\mu-\mu}{1}-\log y!\right)$$

which is of the exponential dispersion family form. The link function is the log so that $g(\mu)=\log\mu$ and the linear predictor is

$$\eta=\log\mu=\alpha_i.$$

So this is a Generalized Linear Model.

b. The likelihood is given by

$$\prod_{i=1}^{3}\prod_{j=1}^{m}\frac{\mu_{ij}^{y_{ij}}\mathrm{e}^{-\mu_{ij}}}{y_{ij}!}$$

so that the log-likelihood is

$$\sum_{i=1}^{3}\sum_{j=1}^{m}\left(y_{ij}\log\mu_{ij}-\mu_{ij}-\log y_{ij}!\right).$$

In terms of α_i, we re-write this as

$$\ell\left(\alpha_1,\alpha_2,\alpha_3\right)=-\sum_{i=1}^{3}m\mathrm{e}^{\alpha_i}+\sum_{i=1}^{3}y_{i+}\alpha_i+\text{constant}$$

where y_{i+} refers to the sum of the observations in the ith group. Differentiating, we get

$$\frac{\partial\ell\left(\alpha_i\right)}{\partial\alpha_i}=-m\mathrm{e}^{\alpha_i}+y_{i+}=0$$

so that the maximum likelihood estimator of α_i is

$$\widehat{\alpha}_i=\log\left(y_{i+}/m\right).$$

c. In comparing the models, notice the nesting: Model 3 is the smallest and is contained in Model 2 which is contained in Model 1. We may use our Rule of Thumb of significant improvement if the decrease in deviance is larger than twice the additional parameter. Here we summarize in table form:

Model	Deviance	Difference	d.f.	$D_1-D_2>2\left(p-q\right)$?	Significant improvement?
Model 3	72.53	-	-		
Model 2	61.64	10.89	1	Yes	Yes
Model 1	60.40	1.24	1	No	No

So Model 2 is a significant improvement from Model 3, but Model 1 is not a significant improvement from Model 1. Now, regarding interpretation of the models: Model 3 says that there is no difference in the average number of claims for the three age groups. Model 2 says that there is no difference in the average number of claims between age groups 1 and 2, but that the third age group may be different. Model 1 gives the possibility of different average number of claims for each age group.

9. We know that the linear predictor, for the ith observation, is

$$\eta_i = \log \mu_i = \sum_j x_{ij}\beta_j = x_i^T\beta \text{ (in vector form).}$$

Thus,

$$E\left(y_i\right) = \mu_i = \mathrm{e}^{x_i^T \cdot \beta}.$$

and therefore, the predicted values are

$$E\left(y_3\right) = \mathrm{e}^{3.41} = 30.27$$

and

$$E\left(y_7\right) = \mathrm{e}^{2.45} = 11.59.$$

Applied questions

1. a. Firstly, call `install.packages("insuranceData")` if you have not already done so. Then:

```
library(insuranceData)
data("dataCar")

# To get some impression on the structure of the data,
# i.e., what variables are there
names(dataCar)
```

```
 [1] "veh_value" "exposure"  "clm"       "numclaims" "claimcst0" "veh_body"
 [7] "veh_age"   "gender"    "area"      "agecat"    "X_OBSTAT_"
```

Note that the response (Y variable) takes only 0 and 1 values. This means (1) a binomial distribution with $n = 1$ and p is a reasonable model (2) the mean of the response ($\mathbb{E}[Y|X]$) is precisely p. Therefore, a logistic regression implies that we should choose the link function g such that

$$\log\left(\frac{p}{1-p}\right) = g^{-1}(p) = \eta = \alpha + \beta_1 X + \beta_2 X^2 + \beta_3 X^3,$$

where X is the vehicle value and we are considering the cubic formula.

We can now fit a glm using the following command.

```
car.glm1 <- glm(clm ~ veh_value + I(veh_value^2) + I(veh_value^3),
  family = binomial(link = "logit"), data = dataCar
)
```

b. `summary(car.glm1)`

```
Call:
glm(formula = clm ~ veh_value + I(veh_value^2) + I(veh_value^3),
    family = binomial(link = "logit"), data = dataCar)

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)    -2.9247606  0.0476282 -61.408  < 2e-16 ***
veh_value       0.2605947  0.0420331   6.200 5.66e-10 ***
I(veh_value^2) -0.0382409  0.0084167  -4.543 5.53e-06 ***
I(veh_value^3)  0.0008803  0.0002752   3.199  0.00138 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 33767  on 67855  degrees of freedom
Residual deviance: 33711  on 67852  degrees of freedom
AIC: 33719

Number of Fisher Scoring iterations: 6
```

The fit shows that all the coefficients are significant as the p-values are smaller than 0.01.

c.

```
# Linear
car.glmLinear <- glm(clm ~ veh_value,
  family = binomial(link = "logit"),
  data = dataCar
)
# Quadratic
car.glmQuadratic <- glm(clm ~ veh_value + I(veh_value^2),
  family = binomial(link = "logit"), data = dataCar
)
# Cubic
car.glmCubic <- glm(clm ~ veh_value + I(veh_value^2) + I(veh_value^3),
  family = binomial(link = "logit"), data = dataCar
)
## AIC
print(car.glmLinear$aic)
```

```
[1] 33749.12
```

```
print(car.glmQuadratic$aic)
```

```
[1] 33718.92
```

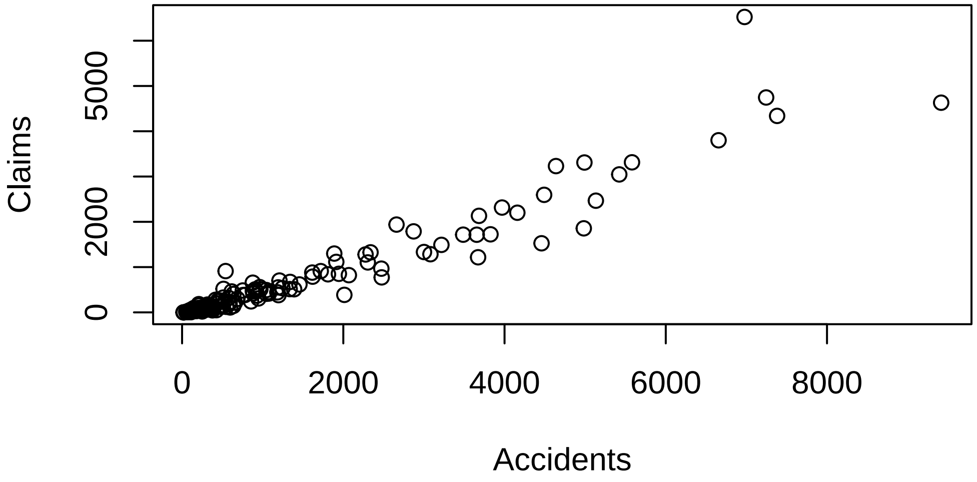
```
print(car.glmCubic$aic)
```

```
[1] 33718.72
```

The difference between the AIC of the cubic and quadratic models is less than one. This shows that if we include a cubic explanatory variable, the improvement of the fit quantified by AIC only decreases by 0.2. Therefore, when evaluating a model by the principal of parsimony, a quadratic model is preferred. Further, the AIC of the quadratic model is much less than that of the linear, suggesting that the linear model is inadequate.

2. a.

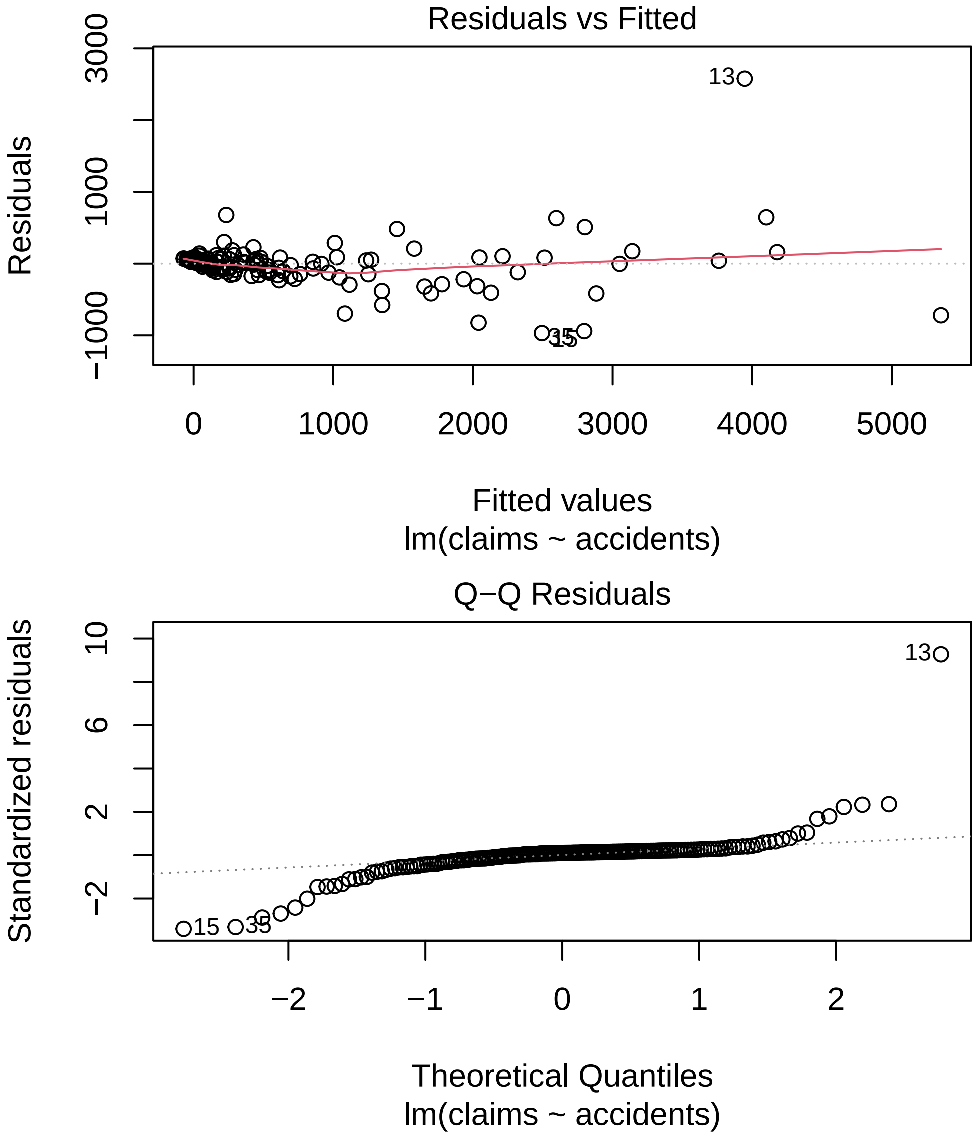
```
claimsdata <- read.csv("Third_party_claims.csv")
attach(claimsdata)
plot(accidents, claims, xlab = "Accidents", ylab = "Claims")
```

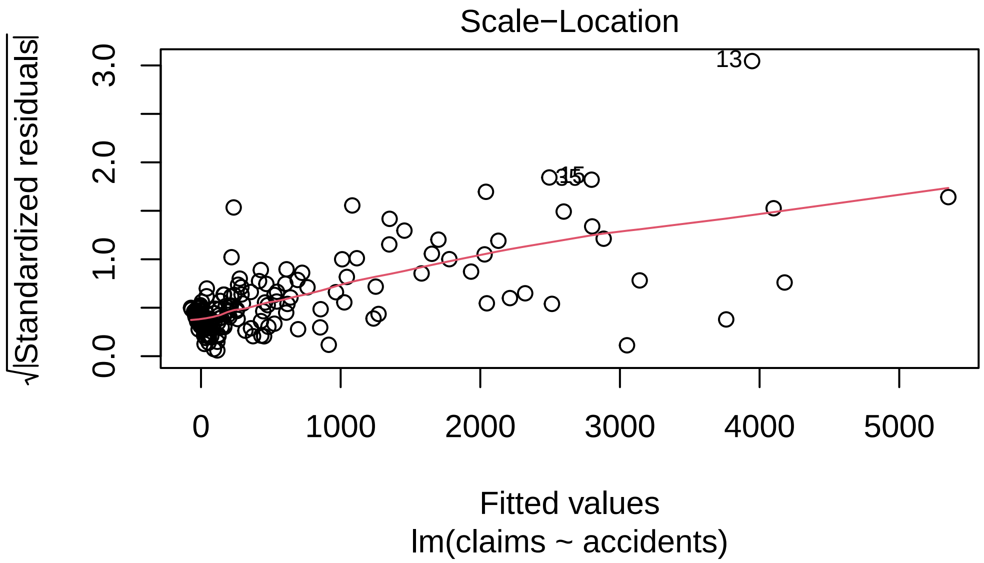


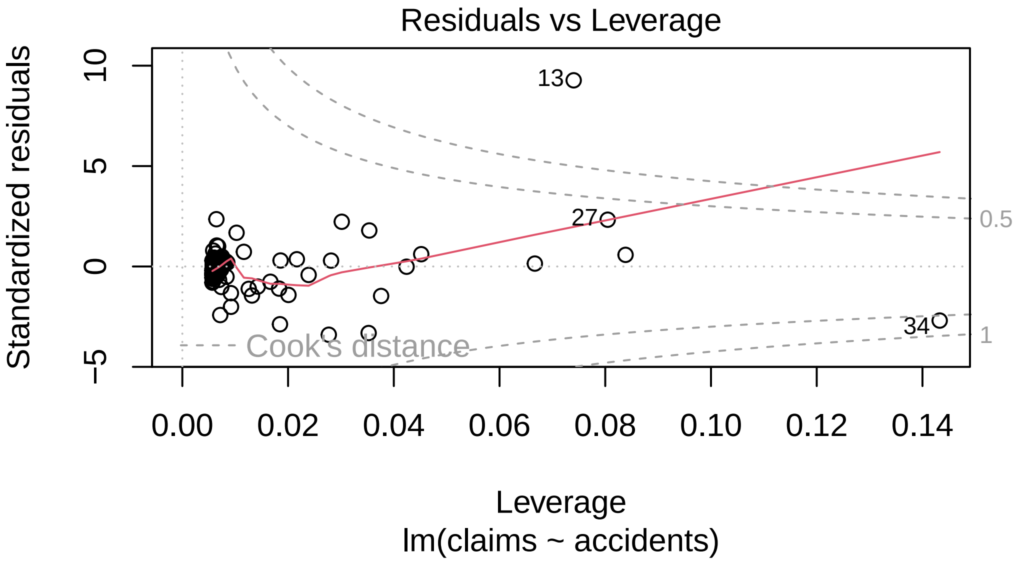
We can clearly see that there is a concentration of points around the origin make it difficult to discern the relationship between the predictor and response. The data is also strongly heteroskedastic, which means more variable for higher value of the predictor. This is a violation of the homoskedasticity assumption of linear model.

b.

```
third.lm <- lm(claims ~ accidents, offset = log(population))
plot(third.lm)
```


Residuals vs Fitted
Residuals
3000
1000
−1000
13
35
15
0
1000
2000
3000
4000
5000
Fitted values
lm(claims ~ accidents)
Q−Q Residuals
Standardized residuals
10
6
2
−2
13
15
35
−2
−1
0
1
2
Theoretical Quantiles
lm(claims ~ accidents)





The residuals vs fitted plot shows that the residual is clearly do not follow a standard normal distribution and the variance seems to inflate as the fitted value increases. Diagnostic checks indicate clear violation of the homoskedasticity assumption.

c.
```
third.poi <- glm(claims ~ log(accidents),
  family = poisson,
  offset = log(population)
)
summary(third.poi)
```

```

Call:
glm(formula = claims ~ log(accidents), family = poisson, offset = log(population))

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)    -7.093809   0.026992 -262.81   <2e-16 ***
log(accidents)  0.259103   0.003376   76.75   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 22393  on 175  degrees of freedom
Residual deviance: 15837  on 174  degrees of freedom
AIC: 17066

Number of Fisher Scoring iterations: 4
```

```
sum(resid(third.poi, type = "pearson")^2) / third.poi$df.residual
```

```
[1] 101.7168
```

The estimate of ψ takes a value of 101.7168. The inflated dispersion parameter suggests there is overdispersion in the data.

d.
```
third.qpoi <- glm(claims ~ log(accidents),
  family = quasipoisson,
  offset = log(population)
)
summary(third.qpoi)
```

```

Call:
glm(formula = claims ~ log(accidents), family = quasipoisson,
    offset = log(population))

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    -7.09381    0.27223 -26.058  < 2e-16 ***
log(accidents)  0.25910    0.03405   7.609 1.66e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for quasipoisson family taken to be 101.7172)
```

```
    Null deviance: 22393  on 175  degrees of freedom
Residual deviance: 15837  on 174  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 4
```

Model	Dispersion parameter	$\hat{\beta}_0$(se)	$\hat{\beta}_1$(se)
Poisson	ψ=1	-7.09381(0.02699)	0.25910(0.003376)
Quasi-Poisson	$\hat{\psi}$=101.7172	-7.09381(0.27223)	0.25910(0.03405)

The quasi-poisson estimates of β are identical to those of the Poisson model, but with standard errors larger by a factor of $\hat{\psi}^{1/2} = 10.085$.