family of model

Generalised Linear Models

ACTL3142 & ACTL5110 Statistical Machine Learning for Risk Applications

UNSW SYDNEY

Lecture Outline

- **Introduction to GLMs**
- The components of a GLM
- Fit a GLM

Generalised linear models

- The linear, logistic and Poisson regression model have common properties and can be summarised in a unified framework
- Framework consists of a systematic and distribution part:
  - $X\beta$ — 1. **Systematic component**: describes the mean structure
  - 2. **Stochastic component**: describes the individual variation of the response around the mean — distribution of Y
  - 3. Link function (Link between stochastic and systematic component).
- This class of models is called Generalised Linear Models (GLM)
- The class of GLMs has played a key role in the development of statistical modelling and of associated software
- The class of GLMs has numerous application in Actuarial Science

Generalised linear models

	2-3 **Linear Regression**	4 **Logistic Regression**	4 **Poisson Regression**	**Generalised Linear Models**
Type of Data	Continuous	Binary (Categorical)	Count	Flexible
Use	Prediction of continuous variables	Classification	Prediction of the number of events	Flexible
Distribution of Y	Normal	Bernoulli (Binomial for multiple trials)	Poisson	Exponential Family
$\mathbb{E}[Y\|X]$	$X\beta$	$\frac{e^{X\beta}}{1+e^{X\beta}}$	$e^{X\beta}$	$g^{-1}(X\beta)$
Link Function Name	Identity	Logit	Log	Depends on the choice of distribution
Link Function Expression	$\eta(\mu) = \mu$	$\eta(\mu) = \log\left(\frac{\mu}{1-\mu}\right)$	$\eta(\mu) = \log(\mu)$	Depends on the choice of distribution

Insurance Applications

- Application are numerous
  - Mortality Modelling
  - Rate making (Modelling Claims Frequency and severity)
  - Loss reserving
- Models used are often *multiplicative*, hence linear on the log-scale.
- Claim numbers are generally Poisson, or Poisson with over-dispersion. These distributions are not symmetric and their variance is proportional to mean.
- Claim amounts are skewed to the right densities, shaped like for example Gamma.

When to use a GLM?

Use GLMs when or constant (linear regression)

- variance not constant and / or
- when errors not normal. Or normal (linear regression)

Cases when we might use GLMs include: when response variable is

- count data expressed as proportions (e.g. logistic regression)
- count data that are not proportions (e.g. log-linear models of counts)
- binary response variable (e.g. dead or alive)
- data on time to death where the variance increases faster than linearly with the mean (e.g. time data with gamma errors).

Many basic statistical methods (regression, *t*-test) assume constant variance—but often untenable. Hence value of GLMs.

Error structure

Many kinds of data have non-normal errors

- errors that are strongly skewed
- errors that are kurtotic – heavy tailed
- errors that are strictly bounded (as in proportions)
- errors that cannot lead to negative fitted values (as in counts)

Error structure

GLM allows specification of a variety of different error distributions:

- Poisson errors, useful with count data
- binomial errors, useful with data on proportions
- gamma errors, useful with data showing a constant coefficient of variation
- exponential errors, useful with data on time to death (survival analysis)

Lecture Outline

- Introduction to GLMs
- **The components of a GLM**
- Fit a GLM

The components of a GLM

- A Generalised Linear Model (GLM) has three components:
  1. A *systematic component* allowing for inclusion of covariates or explanatory variables (captures location). – $X\beta$.
  2. A *stochastic component* specifying the error distributions (captures spread) – dist. of Y.
  3. A parametric *link function* linking the stochastic and systematic components by associating a function of the mean to the covariates.

$$Y \longleftrightarrow X\beta.$$

UNSW SYDNEY

The Systematic Component

- The systematic component is a linear predictor η, that is, a function (with linear coefficients) of the covariates, sometimes called explanatory variables.
- Consider the following linear (in its coefficients) model:

$$\eta_i = \mathrm{x}_i\beta = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip},$$

where x_i is the i'th row of X and there are p predictor variables (or covariates) affecting the response.

- The mean μ_i (location) of the response will depend on η_i in that

$$\begin{aligned}\eta_i &= g(\mu_i)\\ \mu_i &= g^{-1}(\mathrm{x}_i\beta) = g^{-1}(\eta_i),\end{aligned}$$

where g is called the *link* function (unsurprisingly!).

$\mathbb{E}[Y_i|X] = \mu_i$

$\mu_i \overset{g}{\longleftrightarrow} X\beta \; (\eta_i)$

UNSW SYDNEY

The Stochastic Component

$X \sim Exp(\lambda)$

$f_X(x) = \lambda e^{-\lambda x}$

- We are now interested in building spread around our linear model for the mean
- We will use a the **exponential dispersion family**.
- We say Y comes from an exponential dispersion family if its density has the form

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

Here θ and ψ are location and scale parameters, respectively. Note in the book they use different representation but they are equivalent.

- θ known as canonical or natural parameter of the distribution.
- $b(\theta)$ and $c(y;\psi)$ are known functions and specify the distribution

$b(\theta) = \frac{1}{2}\mu^2$ for normal $\psi = \sigma^2$ $\theta = \mu$.

Examples of Exponential Dispersion Families

- Normal $N\left(\mu, \sigma^2\right)$ with $\theta = \mu$ and $\psi = \sigma^2$.
- Gamma(α, β) with $\theta = -\beta/\alpha = -\frac{1}{\mu}$ and $\psi = 1/\alpha$.
- Inverse Gaussian(α, β) with $\theta = -\frac{1}{2}\beta^2/\alpha^2 = -\frac{1}{2\mu^2}$ and $\psi = \beta/\alpha^2$.
- Poisson(μ) with $\theta = \log \mu$ and $\psi = 1$.
- Binomial(m, p) with $\theta = \log\left[p/\left(1-p\right)\right] = \log\left[\mu/(m-\mu)\right]$ and $\psi = 1$.
- Negative Binomial(r, p) with $\theta = \log(1-p) = \log(\mu p/r)$ and $\psi = 1$.

Example - Gamma

The gamma(α, β) distributions belong to the exponential dispersion families. Its density is

$$
\begin{aligned}
f(y) &= \frac{\beta^\alpha y^{\alpha-1} e^{-\beta y}}{\Gamma(\alpha)} \\
&= \exp\left(-\log\Gamma(\alpha) + \alpha\log\beta + (\alpha-1)\log y - \beta y\right) \\
&= \exp\left(\frac{-\frac{\beta}{\alpha}y - \left(-\log\left(\frac{\beta}{\alpha}\right)\right)}{\frac{1}{\alpha}} + \frac{\log\frac{1}{1/\alpha}}{\frac{1}{\alpha}} - \log\Gamma\left(\frac{1}{1/\alpha}\right) + \left(\frac{1}{1/\alpha} - 1\right)\log y\right) \\
&= \exp\left(\frac{y\theta - b(\theta)}{\psi} + c(y;\psi)\right)
\end{aligned}
$$

where $\theta = -\frac{\beta}{\alpha}$, $\psi = \frac{1}{\alpha}$, $b(\theta) = -\log(-\theta)$,

$c(y;\psi) = \frac{\log\frac{1}{\psi}}{\psi} + (\frac{1}{\psi} - 1)\log y - \log\Gamma(\frac{1}{\psi})$

$X \sim \text{Gamma}(\alpha, \beta)$

$E[X] = \frac{\alpha}{\beta} = \mu$

$\theta = -\frac{1}{\mu}$

Some Properties of the exponential family

- The moment generating function can be expressed as

$$M_Y(t) = \mathbb{E}(e^{Yt}) = \exp\left[\frac{b(\theta + t\psi) - b(\theta)}{\psi}\right].$$

- The cumulant generating function immediately follows:

$$\kappa_Y(t) = \log M_Y(t) = \frac{b(\theta + t\psi) - b(\theta)}{\psi}.$$

which can be used to determine

- mean: $\kappa_1 = \left.\frac{\partial \kappa_Y(t)}{\partial t}\right|_{t=0} = b'(\theta) = \mathbb{E}(Y) = \mu$
- variance: $\kappa_2 = \left.\frac{\partial^2 \kappa_Y(t)}{\partial t^2}\right|_{t=0} = \psi b''(\theta) = \mathrm{Var}(Y)$

μ depends on θ.

$\updownarrow$

θ depends on μ.

Mean

$\theta = (b')^{-1}(\mu)$

Notice $\mu = b'(\theta)$ and so mean μ depends on location parameter θ or θ *depends on* μ. So we sometimes write $\theta = \theta(\mu)$.

Examples

- Normal $N(\mu, \sigma^2)$: $\theta = \mu$ and hence $\theta(\mu) = \mu$.
- Gamma(α, β): $\theta = -\beta/\alpha = -\frac{1}{\mu}$ and hence $\theta(\mu) = -\frac{1}{\mu}$.
- Inverse Gaussian(α, β): $\theta = -\frac{1}{2}\beta^2/\alpha^2 = -\frac{1}{2\mu^2}$ and hence $\theta(\mu) = -\frac{1}{2\mu^2}$.
- Poisson(μ) with $\theta = \log\mu$ and hence $\theta(\mu) = \log\mu$.
- Binomial(m, p) with $\theta = \log[p/(1-p)] = \log[\mu/(m-\mu)]$ and hence $\theta(\mu) = \log[\mu/(m-\mu)]$.
- Negative Binomial(r, p) with $\theta = \log(1-p) = \log(\mu p/r)$ and hence $\theta(\mu) = \log(1-p) = \log(\mu p/r)$.

If $X \sim Bin(1, p)$

$E[X] = p = \mu$

Example

For the gamma(α, β) random variable Y, find $\theta(\mu)$.

Answer: The density can be written as

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

where $\theta = -\frac{\beta}{\alpha}$, $\psi = \frac{1}{\alpha}$, $b(\theta) = -\log(-\theta)$, $c(y;\psi) = \frac{\log\frac{1}{\psi}}{\psi} + (\frac{1}{\psi} - 1)\log y - \log\Gamma(\frac{1}{\psi})$. Then

$$\mu = \mathbb{E}[Y] = b'(\theta) = -\frac{1}{\theta}.$$

Therefore,

$$\theta(\mu) = \theta = -\frac{1}{\mu}$$

Variance Function

- The variance is sometimes expressed as $\text{Var}(Y) = \psi V(\mu)$ where clearly

$$V(\mu) = b''(\theta(\mu))$$

and is called the variance function.

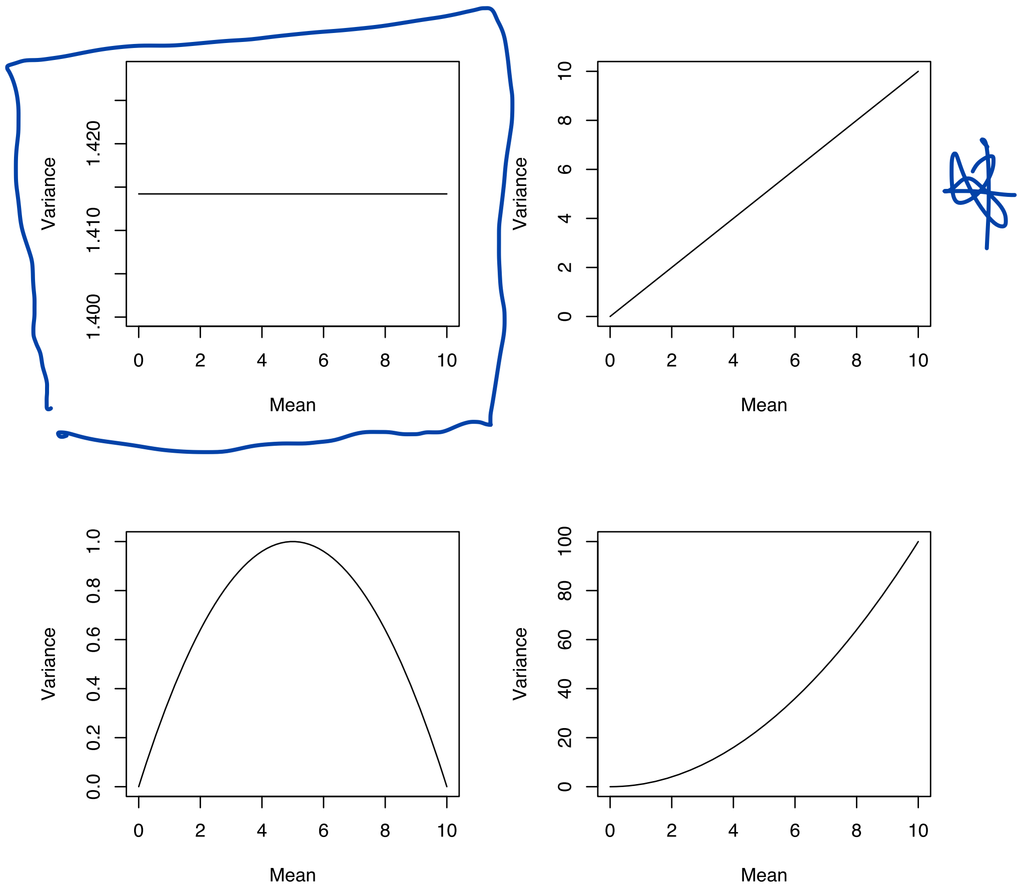
Examples

- Normal $N(\mu, \sigma^2)$ with $V(\mu) = 1$ Variance of Y does not depend on μ.
- Gamma(α, β) with $V(\mu) = \mu^2$
- Inverse Gaussian(α, β) with $V(\mu) = \mu^3$
- Poisson(μ) with $V(\mu) = \mu$ Variance of Y is linear in μ.
- Binomial(m, p) with $V(\mu) = \mu(1 - \mu/m)$
- Negative Binomial(r, p) with $V(\mu) = \mu(1 + \mu/r)$

How does the variance of Y depend on μ.

Mean-variance relationship

- With linear regression, central assumption is constant variance (top left-hand graph)
- Count data: response is integer, lots of zeros—variance may increase linearly with mean (top right)
- Proportion data: count of number of failures of events or successes, variances will be inverted U-shape (bottom left)
- If response follows gamma distribution (e.g. time-to-death data), variance increases faster than linearly with mean (bottom right).



Example

Show that the variance function of Gamma(α, β) GLM is $V(\mu) = \mu^2$.

Answer:

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

where $\theta = -\frac{\beta}{\alpha}, \psi = \frac{1}{\alpha}, b(\theta) = -\log(-\theta)$.

Note that $\mu = \mathbb{E}[Y] = b'(\theta) = -\frac{1}{\theta}$, and therefore

$$\theta = -\frac{1}{\mu}.$$

So $V(\mu) = b''(\theta) = \frac{1}{\theta^2} = \mu^2$.

$V(\mu) = \mu^2$

$= \frac{\alpha^2}{\beta^2}$

$\psi = \frac{1}{\alpha}$

If $X \sim \text{Gamma}(\alpha, \beta)$

$\text{Var}(X) = \frac{\alpha}{\beta^2}$

$\psi \cdot V(\mu)$

$= \frac{1}{\alpha} \cdot \frac{\alpha^2}{\beta^2}$

The link between both components

$\theta = (b')^{-1}(\mu)$

- The mean μ_i is connected to the linear predictor $\mathbf{x}_i\beta$ through

$$\mu_i = g^{-1}\big(\mathbf{x}_i\beta\big) = g^{-1}(\eta_i) \quad \text{or} \quad \eta_i = g(\mu_i)$$

  where g is called the *link* function.

- If $g(\cdot) \equiv \theta(\cdot)$, that is, if

$$\theta_i = \eta_i,$$

$(b')^{-1}(\mu) = X\beta$.

  we say we have a *canonical link*, or natural link function.

$\mu = b'(X\beta)$

UNSW SYDNEY

Canonical Link Functions

$b(\mu) = \frac{1}{2}\mu^2$

- Some canonical links are:

Distribution	**Canonical Link** $g(\mu)$	**Called**
Normal	$g(\mu) = \theta(\mu) = \mu$	Identity
Poisson	$g(\mu) = \theta(\mu) = \log \mu$	Log link
Binomial	$g(\mu) = \theta(\mu) = \log\left(\frac{\mu}{m-\mu}\right)$	Logit
Gamma	$g(\mu) = \theta(\mu) = -1/\mu$	Reciprocal

$\log(\lambda_i) = X_i\beta$

$\log\left(\frac{p}{1-p}\right) = X\beta.$

Summary of GLM components

A GLM models an n-vector of independent response variables, $\mathbf{Y}$ using

1. **Random component:** For each observation y_i we use an exponential dispersion model

$$f(y_i;\theta_i) = \exp\left[\frac{y_i\theta_i - b(\theta_i)}{\psi} + c(y_i;\psi)\right]$$

   where θ_i is the canonical parameter, ψ is a dispersion parameter and function $b(\cdot)$ and $c(\cdot,\cdot)$ are known.

2. **Sytematic component:** $\eta_i = \mathrm{x}_i\beta = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$, the linear predictors with $\beta = (\beta_0, \beta_1, \ldots, \beta_p)$ regression parameters.
3. Parametric **link function**: the link function $g(\mu_i) = \eta_i = \mathrm{x}_i\beta$ combines the linear predictor with the mean $\mu_i = \mathbb{E}[Y_i]$. The link is called canonical if $\theta_i = \eta_i$

Lecture Outline

- Introduction to GLMs
- The components of a GLM
- **Fit a GLM**

Procedure

$\theta = (b')^{-1}(\mu)$

Constructing a GLM consists of the following steps:

- Choose a response distribution $f(y_i)$ and hence choose $b(\theta)$.
- Choose a link $g(\mu)$. — When in doubt canonical link
- Choose explanatory variables x in terms of which $g(\mu)$ is to be modeled.
- Collect observations $y_1, \cdots, y_n$ on the response y and corresponding values $x_1, \cdots, x_n$ on the explanatory variables x.
- Fit the model by estimating β and, if unknown, ψ.
- Given the estimate of β, generate predictions (or fitted values) of y for different settings of X and examine how well the model fits. Also the estimated value of β will be used to see whether or not given explanatory variables are important in determining μ.

Assignment

1. Binomial
2. Canonical (logit)
3. Data done
4. Done
5. Software
6. Interpretation

Case Study: Motor claims illustration

- Consider the data used in McCullagh and Nelder (1989), page 299, related to motor claims. So the responses are claims.
- There are three factors used:
  - policyholder age (PA), with 8 levels, 17-20, 21-24, 25-29, etc.
  - car group (CG), with 4 levels, A, B, C, and D.
  - vehicle age (VA), with 4 levels, 0-3, 4-7, 8-9, 10+
- There are a total of 123 different cells of data.

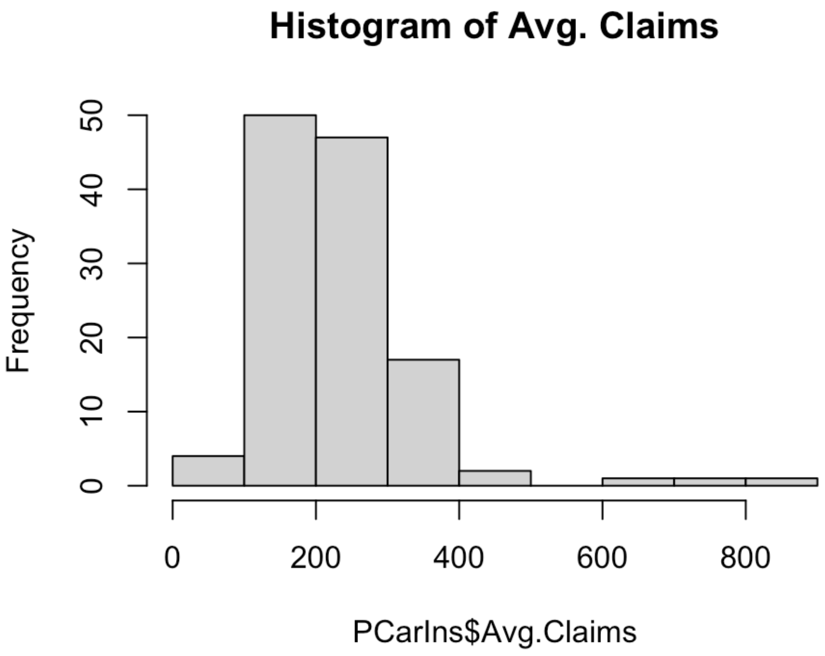
Case Study: Motor claims illustration

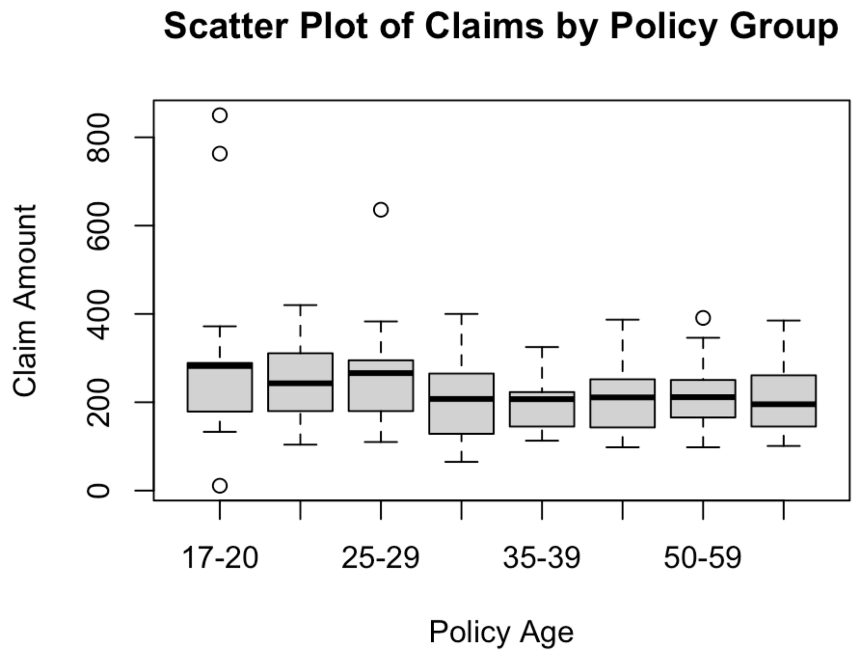
```r
library(tidyverse)

PCarIns <- read_csv("PrivateCarIns1975-Data.csv")
PCarIns <- PCarIns %>% filter(Numb.Claims>0) # remove the 3 categories with no claims

str(PCarIns)
```

```
spc_tbl_ [123 × 7] (S3: spc_tbl_df/tbl_df/tbl/data.frame)
 $ Pol.Age    : num [1:123] 1 1 1 1 1 1 1 1 1 1 ...
 $ Cpol.Age   : chr [1:123] "17-20" "17-20" "17-20" "17-20" ...
 $ Car.Group  : chr [1:123] "A" "A" "A" "A" ...
 $ Veh.Age    : num [1:123] 1 2 3 4 1 2 3 4 1 2 ...
 $ Cveh.Age   : chr [1:123] "0-3" "4-7" "8-9" "10+" ...
 $ Avg.Claims : num [1:123] 289 282 133 160 372 249 288 11 189 288 ...
 $ Numb.Claims: num [1:123] 8 8 4 1 10 28 1 1 9 13 ...
 - attr(*, "spec")=
  .. cols(
  ..   Pol.Age = col_double(),
  ..   Cpol.Age = col_character(),
  ..   Car.Group = col_character(),
  ..   Veh.Age = col_double(),
  ..   Cveh.Age = col_character(),
  ..   Avg.Claims = col_double(),
  ..   Numb.Claims = col_double()
  .. )
 - attr(*, "problems")=<externalptr>
```

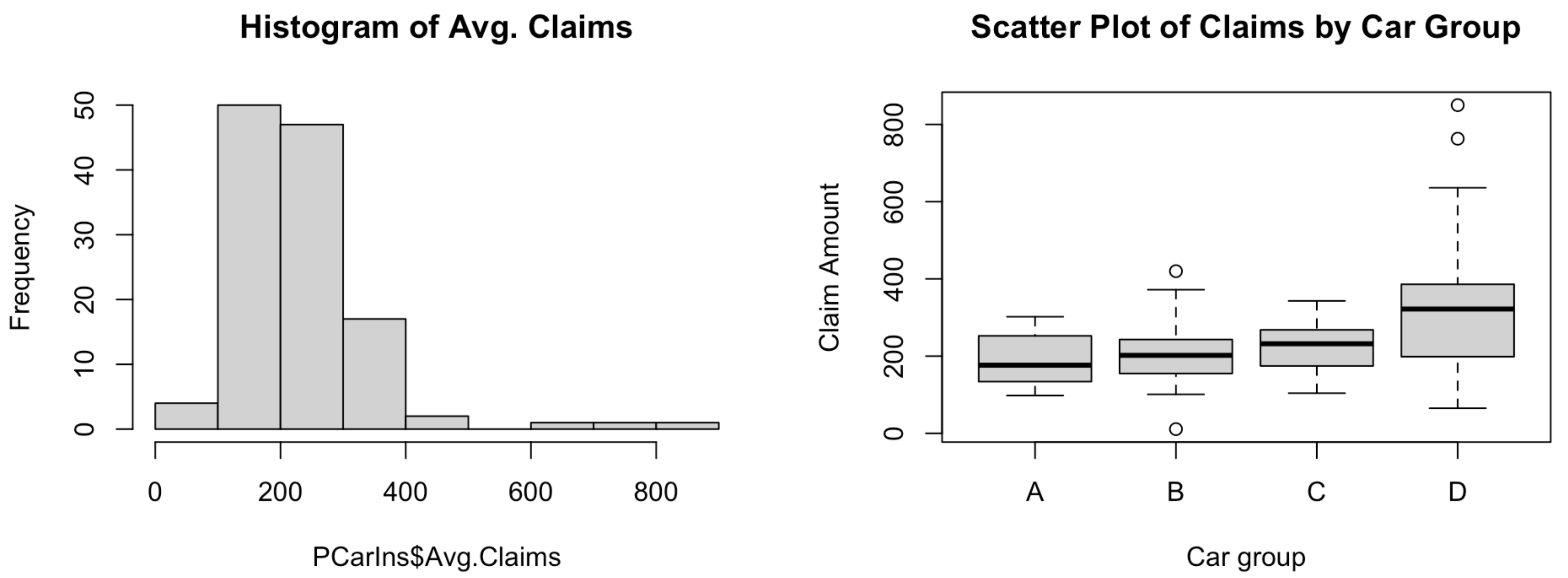
```r
# convert to categorical
PCarIns <-  PCarIns %>%
  mutate(Cpol.Age = factor(Cpol.Age),
         Car.Group = factor(Car.Group),
         Cveh.Age = factor(Cveh.Age))
```

Case Study: Motor claims illustration

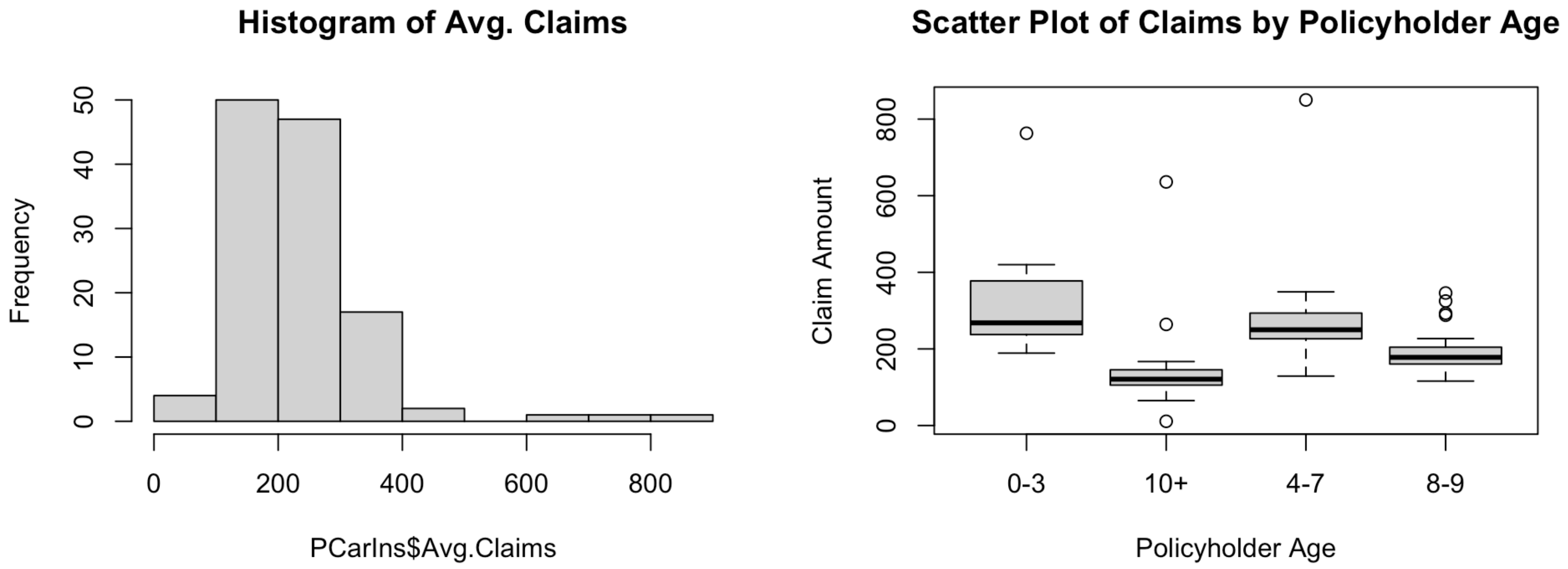




Case Study: Motor claims illustration



Case Study: Motor claims illustration



How could we model the relationship between claim Amounts and the covariates?

Maximum Likelihood Estimation

- The parameters in a GLM are estimated using maximum likelihood.
- For each observation y_i the contribution to the likelihood is

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

- Given vector $\mathbf{y}$, an observation of $\mathbf{Y}$, MLE of β is possible. Since the y_i are mutually independent, the likelihood of β is

$$L(\beta) = \prod_{i=1}^{n} f(y_i;\theta_i).$$

UNSW SYDNEY

Maximum Likelihood Estimation

So for n independent observations $y_1, y_2, \ldots, y_n$, we have

$$L(\mathbf{y}; \mu) = \prod_{i=1}^{n} \exp\left[\frac{y_i\theta_i - b(\theta_i)}{\psi} + c\left(y_i; \psi\right)\right].$$

Take log to obtain the log-likelihood as

$$\ell(\mathbf{y}; \mu) = \sum_{i=1}^{n} \left[\frac{y_i\theta_i - b(\theta_i)}{\psi} + c(y_i; \psi)\right].$$

Example

Consider a GLM model with the canonical link and gamma distribution. The density of response variable is

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

with $b(\theta) = -\log(-\theta)$.

Moreover, with canonical link, we have $\theta_i = \theta(\mu_i) = g(\mu_i) = \mathbf{x}_i\beta$.

The log-likelihood is

$$\ell(\mathbf{y};\mu) = \sum_{i=1}^{n}\left(\frac{y_i\theta_i - b(\theta_i)}{\psi} + c(y_i;\psi)\right)$$

Example (continued)

$$\frac{\partial}{\partial \beta_j}\ell(\mathbf{y};\mu) = \sum_{i=1}^{n} \frac{y_i - b'(\theta_i)}{\psi}\frac{\partial \theta_i}{\partial \beta_j}$$

$$= \sum_{i=1}^{n} \frac{y_i - (-\frac{1}{\theta_i})}{\psi} x_{ij}$$

$$= \cancel{\sum_{i=1}^{n} \frac{y_i + \frac{1}{\mathrm{x}_i\beta}}{\psi} x_{ij}}$$

Solving $\sum_{i=1}^{n} \cancel{\frac{y_i + \frac{1}{\mathrm{x}_i\beta}}{\psi}} x_{ij} = 0$ for all j gives the MLE of β.

$$\frac{(y_i + X_i\beta)}{\psi(X_i\beta)} \cdot x_{ij}$$

Normal dist.

$\theta(\mu) = \mu$ $\quad b(\theta) = \frac{1}{2}\mu^2$

$\psi = \sigma^2$

$$\frac{\partial \theta_i}{\partial \beta_j} = x_i$$

$\mu = X\beta$ Canonical link

$$\sum \frac{y_i - \mu}{\sigma^2} \cdot x_{ij}$$

$$= \sum \frac{(y_i - x_i\beta)}{\sigma^2} x_{ij}$$

Equivalent of OLS

UNSW SYDNEY

Case Study: Motor claims illustration - Gamma GLM

```r
pcarins.glm <- glm(Avg.Claims~ Cpol.Age + Car.Group + Cveh.Age,  weights=Numb.Claims,
                   family=Gamma,  data = PCarIns)
summary(pcarins.glm)
```

link function is $\frac{1}{\mu}$.

$\mu = \frac{1}{X\beta}$.

$X\beta > 0$

```
Call:
glm(formula = Avg.Claims ~ Cpol.Age + Car.Group + Cveh.Age, family = Gamma,
    data = PCarIns, weights = Numb.Claims)

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    3.411e-03  4.179e-04   8.161 6.31e-13 ***
Cpol.Age21-24  1.014e-04  4.363e-04   0.232 0.816664
Cpol.Age25-29  3.500e-04  4.124e-04   0.849 0.397942
Cpol.Age30-34  4.623e-04  4.106e-04   1.126 0.262652
Cpol.Age35-39  1.370e-03  4.192e-04   3.268 0.001447 **
Cpol.Age40-49  9.695e-04  4.046e-04   2.396 0.018284 *
Cpol.Age50-59  9.164e-04  4.080e-04   2.246 0.026691 *
Cpol.Age60+    9.201e-04  4.157e-04   2.213 0.028958 *
Car.GroupB     3.765e-05  1.687e-04   0.223 0.823776
Car.GroupC    -6.139e-04  1.700e-04  -3.611 0.000463 ***
Car.GroupD    -1.421e-03  1.806e-04  -7.867 2.84e-12 ***
Cveh.Age10+    4.154e-03  4.423e-04   9.390 1.05e-15 ***
Cveh.Age4-7    3.663e-04  1.009e-04   3.632 0.000430 ***
Cveh.Age8-9    1.651e-03  2.268e-04   7.281 5.45e-11 ***
---
```

The "Null Model" and "Full Model"

- With a GLM we estimate Y_i by $\hat{\mu}_i$
- For n data points we can estimate up to n parameters
- **Null model**: the systematic component is a constant term only.

$$\hat{\mu}_i = \bar{y}, \quad \text{for all} \quad i = 1, 2, \ldots, n$$

  - Only one parameter $\rightarrow$ too simple
- **Full or saturated model**: Each observation has its own parameter.

$$\hat{\mu}_i = y_i, \quad \text{for all} \quad i = 1, 2, \ldots, n$$

  - All variations can be explained by the covariates $\rightarrow$ no explanation of data possible

Deviance and Scaled Deviance

The log-likelihood in the full model gives

$$\ell(\mathbf{y};\mathbf{y}) = \sum_{i=1}^{n} \left[\frac{y_i\tilde{\theta}_i - b(\tilde{\theta}_i)}{\psi} + c(y_i;\psi) \right]$$

where $\tilde{\theta}_i$ are the canonical parameter values corresponding to $\mu_i = y_i$ for all $i = 1, 2, \ldots, n$.

UNSW SYDNEY

Deviance and Scaled Deviance

- Let $\widehat{\mu}$ denote the M.L.E. of chosen model.
- One way of assessing the fit of a given model is to compare it to the model with the "closest" possible fit: the full model
- The **likelihood ratio criterion** compares a model with its associated full model.

$$-2\log\left[\frac{L(\mathbf{y};\widehat{\mu})}{L(\mathbf{y};\mathbf{y})}\right] = 2[\ell(\mathbf{y};\mathbf{y}) - \ell(\mathbf{y};\widehat{\mu})] = 2\sum_{i=1}^{n}\left[\frac{y_i(\tilde{\theta}_i - \hat{\theta}_i)}{\psi} - \frac{b(\tilde{\theta}_i) - b(\hat{\theta}_i)}{\psi}\right]$$
$$= \frac{D(y,\widehat{\mu})}{\psi}$$

$\frac{\sum d_i}{\psi}$

- $D(y,\widehat{\mu})$ is called the **deviance** and $D(y,\widehat{\mu})/\psi$ the **scaled deviance**.
- Deviance plays much the same role for GLMs that RSS plays for ordinary linear models. (For ordinary linear models, deviance *is* RSS.)

Example

The scaled deviance of a gamma(α, β) is

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

Exponential Dispersions and their Deviances

We drop the subscript $i = 1, 2, \ldots, n$

Deviances are:

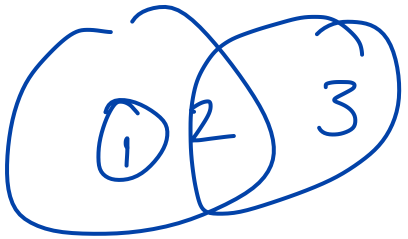
Distribution	**Deviance** $D(y, \hat{\mu})$
Normal	$\sum (y - \hat{\mu})^2$
Poisson	$2 \sum [y \log(y/\hat{\mu}) - (y - \hat{\mu})]$
Binomial	$2 \sum [y \log(y/\hat{\mu}) + (m - y) \log((m - y)/(m - \hat{\mu}))]$
Gamma	$2 \sum [- \log(y/\hat{\mu}) + (y - \hat{\mu})/\hat{\mu}]$
Inverse Gaussian	$\sum (y - \hat{\mu})^2/(\hat{\mu}^2 y)$

Scaled Deviance as a Measure of Model Fit

- The scaled deviance is actually a measure of the fit of the model. It has approximately (asymptotically true) a chi-squared distribution with degrees of freedom equal to the number of observations minus the number of estimated parameters.

$$\frac{D(y,\hat{\mu})}{\psi} \to \chi^2_{n-(p+1)} \quad \text{when} \quad n \to \infty$$

- Thus, we can use the scaled deviance usually for comparing models that are nested (one model is a subset of the other) by looking at the difference in the deviance and comparing it with the chi-squared table.
- Reminder: a significant value (at the 5% level) for a χ^2 distribution with ν degrees of freedom is approximately 2ν.



Model selection

- **Nested models:** Wald test, score test, likelihood ratio test (drop-in deviance test)
- **Non-Nested models:** Use $AIC = -2\ell(\mathbf{y};\widehat{\mu}) + 2d$ (the smaller the better)

- Cross validation

Model selection (Nested models)

- Model 1: $\eta = \beta_0 + \beta_1 x_1 + \ldots \beta_q x_q$,
- Model 2: $\eta = \beta_0 + \beta_1 x_1 + \ldots \beta_q x_q + \underbrace{\beta_{q+1} x_{q+1} + \ldots \beta_p x_p}_{\text{additional}}$

Is Model 2 an improvement over Model 1?

$$H_0 : \beta_{q+1} = \cdots = \beta_p = 0$$

$$H_a : \text{at least one } \beta_j \text{ is non-zero}$$

Model selection (Nested models)

- Consider two models,
  - Model 1: q parameters, with scaled deviance D_1; $\sim \chi^2_{n-(q+1)}$
  - Model 2: p parameters ($p > q$), with scaled deviance D_2. $\sim \chi^2_{n-(p+1)}$
- Model 2 is a significant improvement over Model 1 (a more parsimonious model), if $D_1 - D_2 >$ the critical value obtained from a $\chi^2(p-q)$ distribution.
- Since

$$\mathbb{P}\left[\chi^2(\nu) > 2\nu\right] \approx 5\%,$$

the following rule of thumb can be used as an approximation:

$$\text{model 2 is preferred if } D_1 - D_2 > 2(p-q).$$

Rule of thumb

Case Study: Motor claims illustration - Null Model

```r
#Null model
pcarins.glm.NULL <- glm(Avg.Claims~ 1,  weights=Numb.Claims, family=Gamma,
                              data = PCarIns)
pcarins.glm.NULL
```

```

Call:  glm(formula = Avg.Claims ~ 1, family = Gamma, data = PCarIns,
    weights = Numb.Claims)

Coefficients:
(Intercept)
   0.004141

Degrees of Freedom: 122 Total (i.e. Null);  122 Residual
Null Deviance:      649.9
Residual Deviance: 649.9    AIC: 99520
```

Case Study: Motor claims illustration - Deviance analysis

```
#analysis of the deviance table
print(anova(pcarins.glm, test="Chi"))
```

```
Analysis of Deviance Table

Model: Gamma, link: inverse

Response: Avg.Claims

Terms added sequentially (first to last)

          Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                        122     649.87
Cpol.Age   7   82.178       115     567.69 3.801e-12 ***
Car.Group  3  228.309       112     339.38 < 2.2e-16 ***
Cveh.Age   3  214.602       109     124.78 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Is Cpol.Age worth adding?

82.178 > 2x7 ?

Yes, so worth adding at 5%

Order of ANOVA matters!

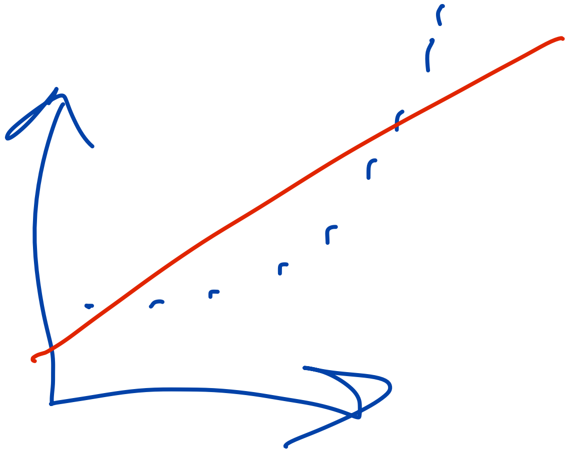
Continued

The scaled deviance statistics are provided below:

Model	Deviance	First Diff.	d.f.	Mean Deviance
1	649.9			
PA	567.7	82.2	7	11.7 >2 ✓
PA+CG	339.4	228.3	3	76.1 >2 ✓
PA+CG+VA	124.8	214.7	3	71.6 >2 ✓
+PA·CG	90.7	34.0	21	1.62 <2 ✗
+PA·VA	71.0	19.7	21	0.94
+CG·VA	65.6	5.4	9	0.60 ✓
Complete	0.0	65.6	58	1.13

Residuals in GLMs

- Residuals are a primary tool for assessing how well a **model fits** the data.
- They can also help to detect the form of the variance function and to diagnose problem observations.
- We consider three different kinds of residuals:
  - deviance residuals: $r_i^D = \text{sign}(y_i - \widehat{\mu}_i)\sqrt{d_i}$ where d_i is contribution of ith observation to the scaled deviance (drawing on idea that deviance is akin to RSS).
  - Pearson residuals: $r_i^P = \dfrac{y_i - \widehat{\mu}_i}{\sqrt{V(\widehat{\mu}_i)}}$.
  - response residuals: they are simply $y_i - \widehat{\mu}_i$.



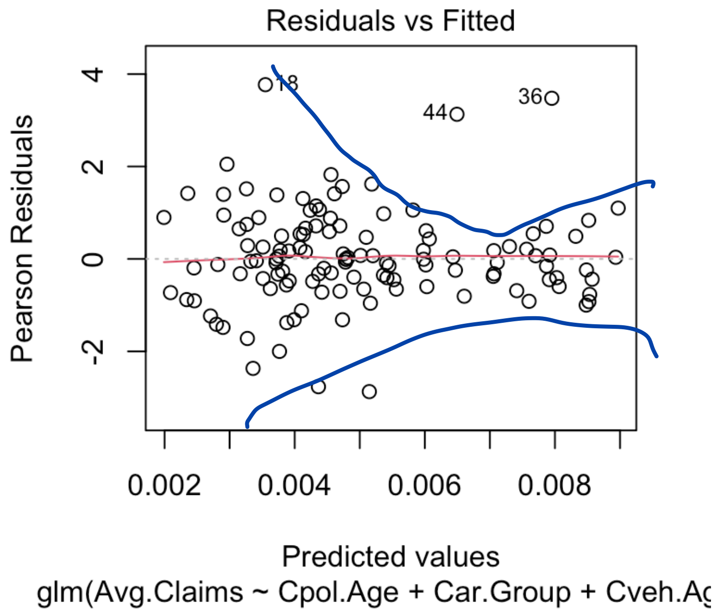
Residuals in GLMs (continued)

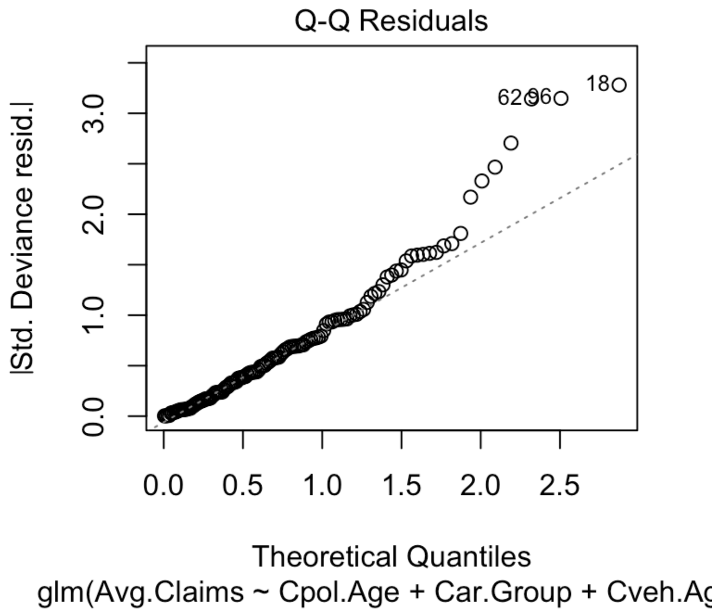
- If the model is correct and the sample size n is large, then the (scaled) deviance is approximately $\chi^2_{n-(p+1)}$.
- The expected value of the deviance is thus $n - (p+1)$, and one expects each case to contribute approximately $(n-(p+1))/n \approx 1$ to the deviance. If $|d_i|$ is much greater than 1, then case i is contributing too much to the deviance (contributing to lack of fit), indicating a departure from the model assumptions for that case.
- Typically deviance residuals are examined by plotting them against fitted values or explanatory variables.

Case Study: Motor claims illustration - Residual

```
plot(pcarins.glm, which = 1:2)
```

Indicates lack of fit





Pattern happening, suggesting lack of fit.

Case Study: Motor claims illustration - Residual

```r
plot(pcarins.glm, which = c(3,5))
```