Linear Predictors

The covariates enter the GLM through the linear predictor η.

The predictor is linear in the coefficients to be estimated, not the covariates.

There are two types of covariates.

1. Variables, which are numbers.
2. Factors, which are categorical and have a finite number of categories.

We will write coefficients as $\beta_1, \beta_2, \ldots$

and the covariates as $x_1, x_2, \ldots$

Example

x_1 is age

x_2 is temperature

they are both variables

η could be

$$\beta_0 + \beta_1 x_1 + \beta_2 x_2$$

β_0 is ~~they~~ the y-intercept

η could be

$$\beta_0 + \beta_1 x_1$$

η could be

$$\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$$

called an interaction term

If $\beta_3 x_1 x_2$ is included, then so should $\beta_1 x_1$ and $\beta_2 x_2$.

$\eta = \beta_0 + \beta_3 x_1 x_2$ is not allowed.

Notation for η

Model	η	R Formula
1	β_0	Y~1
age	$\beta_0+\beta_1 X_1$	Y~X1
age+age^2 ~~age+ϕ~~	$\beta_0+\beta_1 X_1+\beta_2 X_1^2$	Y~X1+I(X1^2)
age+temperature	$\beta_0+\beta_1 X_1+\beta_2 X_2$	Y~X1+X2
age+temperature +age:temperature OR age*temperature	$\beta_0+\beta_1 X_1+\beta_2 X_2 + \beta_3 X_1 X_2$	Y~X1+X2+ X1:X2 OR Y~X1*X2

Model	μ	R Formula
age*temperature*weight	$\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_1 x_2 + \beta_5 x_1 x_3 + \beta_6 x_2 x_3 + \beta_7 x_1 x_2 x_3$	y~X1*X2*X3
age+temperature+ weight + age:temperature	$\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_1 x_2$	y~X1+X2+X3 + X_1:X_2

Factors

For a factor, we have a coefficient α_i or β_j where i or j can take any level (an element of the category). We can also ~~ha~~ have coefficients for interaction terms.

Model	η	R Formula
sex	α_i	y ~ sex
smoker/nonsmoker	β_j	y ~ s/n
sex + s/ns	$\alpha_i + \beta_j$	y ~ sex + s/ns
sex + s/ns + sex:s/ns OR sex * s/ns	$\alpha_i + \beta_j + \gamma_{ij}$	y ~ sex + s/ns + sex:s/ns OR y ~ sex * s/ns

We can combine variables and factors.

Model	η	R Formula
age	$\beta_0 + \beta_1 x_1$	Y~X1
sex	α_i	Y~sex
age+sex	$\alpha_i + \beta x_1$	Y~age+sex
age*sex	$\alpha_i + \beta_i x_1$	Y~age*sex

When we multiply two factors, we can't estimate all $n+m$ parameters, where i has n levels and j has ~~m~~ m levels. One parameter must be set to 0 (e.g. $\gamma_{00}=0$). So are really $n+m-1$ parameters.

The degrees of freedom of η is the number of (non-zero) coefficients in it.

Links

Let η denote the linear predictor.

Let $\mu = E(Y)$.

The link connects η and μ.

It is a function g for which

$$g(\mu) = \eta.$$

μ will be our prediction ~~to~~ for Y.

We would like to take inverses

$$\mu = g^{-1}(\eta).$$

In order to do this, g should be continuous and monotone increasing.

The usual link, called the canonical link

is $\theta(\mu)$

Unless, otherwise specified, the canonical link will be used.

Distribution	Canonical Link	R name
normal	$g(\mu)=\mu$	identity
Poisson	$g(\mu)=\log\mu$	log
Binomial	$g(\mu)=\ln\left(\frac{\mu}{1-\mu}\right)$	logit
Gamma	$g(\mu)=\frac{1}{\mu}$	inverse

Recall: $\mu=g^{-1}(\eta)$

Link	Inverse
identity	$g^{-1}(\eta)=\eta$
log	$g^{-1}(\eta)=e^{\eta}$ (this makes sense because for Poisson $\mu=\lambda>0$)
logit	set $g(\mu)=\ln\left(\frac{\mu}{1-\mu}\right)=\eta$ $\Rightarrow \frac{\mu}{1-\mu}=e^{\eta} \Rightarrow \mu=\frac{e^{\eta}}{1+e^{\eta}}$ (this makes sense because for Binomial $\mu=p\in[0,1]$)

~~inverse~~

Link

inverse

Inverse

inverse $g^{-1}(\mu) = \frac{1}{\mu}$

Using glm in R

To fit a GLM in R we use the command

model <- glm(y ~ ..., family = ...,

linear predictor

normal
poisson
binomial
gamma

link = ..., data = ...)

identity
log
logit
inverse

data.frame

if link is omitted the canonical link is used

The estimates of the parameters in η

are obtained by

summary(model)

The estimates are obtained by using maximum likelihood and numerical methods.

Call the estimates of parameters α_i, call them $\hat{\alpha}_i$.

Using these $\hat{\alpha}_i$ in η gives an estimator of η, called $\hat{\eta}$.

The prediction for Y, called fitted value,

is $$\hat{\mu} = g^{-1}(\hat{\eta}).$$

Significance of the Parameters

If a parameter is not significant, then it should not be in the model.

It is a fact that

$$\hat{\beta} \sim N(\beta, CRLB(\beta))$$

We use this fact to test

$$H_0 : \beta = 0 \quad vs \quad H_1 : \beta \neq 0$$

H_0 implies β is not significant.

Under H_0, $\tilde{\beta} \sim N(0, CRLB(\beta))$

Let $\widehat{CRLB(\beta)}$ be $CRLB(\beta)$ with β replaced by $\hat{\beta}$.

Under H_0 $$\frac{\hat{\beta} - 0}{\sqrt{\widehat{CRLB(\beta)}}} \sim N(0,1)$$

We reject H_0 at the 5% level if

$$\left| \frac{\hat{\beta}}{\sqrt{\widehat{CRLB}(\beta)}} \right| > 1.96$$

$$\Rightarrow \quad |\hat{\beta}| > 1.96 \sqrt{\widehat{CRLB}(\beta)}$$

or roughly $|\hat{\beta}| > 2\sqrt{\widehat{CRLB}(\beta)}$

In R the p-values for all coefficients are obtained by

summary (model)

We want measures of models saying how good they are.

The Saturated Model

The saturated model is a benchmark against which we can compare any other model. If a model has as many parameters as there are data points, it will fit the data perfectly.

I.e. $\hat{\mu}_i = y_i$ for all i,

where y_i ~~are~~ are the observed data.

Each model has a log-likelihood

$$\ell(\underline{y}; \theta, \phi) = \ln f_{\underline{Y}}(\underline{y}; \theta, \phi)$$

Example

Claim amounts per year are exponential and independent.

$$f_{Y_i}(y_i) = \frac{1}{\mu_i} \exp\left(-\frac{y_i}{\mu_i}\right)$$

$$\ln \prod_{i=1}^{n} f_{Y_i}(y_i) = -\sum_{i=1}^{n} \frac{y_i}{\mu_i} - \sum_{i=1}^{n} \ln(\mu_i)$$

For the saturated model $\hat{\mu}_i = y_i$

We obtain estimated log likelihood by substituting the estimated parameters $\hat{\alpha}_i$ for the parameters α_i

In this example, we obtain the likelihood

$$l_S = -\sum_{i=1}^{n} 1 - \sum_{i=1}^{n} \ln(y_i)$$

$$= -n - \sum_{i=1}^{n} \ln(y_i)$$