Generalized Linear Models (GLM)

- Regression is the study of the change of the distribution of one variable, y, according to the value of another variable, x.
- Both continous and discrete variables can be included in regression problems.
- Formally, a relationship between x and y usually means that the expected value of y is different for different values of x.
- At this stage, changes in σ^2 or other aspects of the distribution are not considered.
- When y is continous, changes in y are smooth (continuous and differentiable), and the model used is:

$$E(y|x) = g(x)$$

for some unknown smooth function $g(\cdot)$.

⋆ A variety of functions can be used to estimate $g(\cdot)$. These functions are called **scatterplot smoother**.

⋆ For a pair (x_i, y_i), $i = 1, \ldots, n$:

$$y_i = g(x_i) + \varepsilon_i$$

where $g(\cdot)$ is called the **regression function**.

- A chemical theory suggests that the temperature at which a certain reaction occurs is 180°.
- The results of 10 indipendent experiments, contained in `reaction.dat`, are:
  179.4, 177.5, 180.6, 178.1, 180.2, 179.3, 180.0, 178.3, 179.1, 181.1.
- In this example, $n = 10$ and

$$E(y_i) = \beta_0 \text{ for } i = 1, \ldots, n$$

- More generally, however, different observations are associated with different experimental conditions or different values of one or more **explanatory variables** as the following examples will illustrate.

- To compare two growth hormones, 19 chicken were tested, 11 being assigned to hormone A and 8 to hormone B.
- The following were the gains in weight (grams) over the period of experiment, contained in `hormones.dat`:

  Hormone A: 615, 645, 840, 615, 365, 450, 450, 530, 756, 785, 560.

  Hormone B: 870, 895, 1035, 700, 560, 756, 785, 980.

- $E(y_{ij}) = \beta_i$ for $j = 1, \ldots, n_i$; $i = 1, 2$.
- Thus, β_1 and β_2 are the expectations of y for hormones A and B, respectively.
- Typically, it is interesting in whether or not $\beta_1 = \beta_2$ or in the range of values for the difference $\beta_1 - \beta_2$.

- In an experiment to investigate the effects of a depressant drug, the reaction times of then male rats to a certain stimulus were measured after a specified dose of the drug have been administered to each rat.

- The results, contained in `drug.dat`, were as follows:

Rat number:	1	2	3	4	5	6	7	8	9	10
Dose (mg):	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
Reaction time (secs):	0.32	0.24	0.40	0.52	0.44	0.56	0.64	0.52	0.60	0.80

- A simple model for this data might be a straight line, i.e.

$$E(y_i|x_i) = \beta_0 + \beta_1 x_i \text{ for } i = 1, \ldots, n$$

- For simplicity, usually the above formula is written as $E(y_i) = \beta_0 + \beta_1 x_i$, the conditioning on x_i being understood.

- In the study of the polyesterification of fatty acids with glycols, the effect of temperature (°C) on the percentage conversion of the esterification process was investigated.
- The following data, contained in `polyesterification.dat`, are the results of an experiment using a catalyst of $4 \cdot 10^{-4}$ mole zinc chloride per 100 grams of fatty acid.

Temperature (°C):	175	200	225	250	275	300
Percentage conversion (%):	67	83	93	97	99	92

- This is similar to the previous example, but now a quadratic function of x seems sensible:

$$E(y_i) = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 \text{ for } i = 1, \ldots, n$$

- The data in the `oxidant.dat` file give levels of an air pollutant, "Oxidant", together with levels of four meteorological variables recorded on 30 days during one summer.

- Which, if any, of the four indipendent variables seem to be related to levels of oxidant?

- Again, a simple starting point would be:

$$E(y_i) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 \text{ for } i = 1, \ldots, n$$

Model formulae

Expression	Description
`y ~ x`	Simple regression
`y ~ 1 + x`	Explicit intercept
`y ~ -1 + x`	Through the origin
`y ~ x + x^2`	Quadratic regression
`y ~ x1 + x2 + x3`	Multiple regression
`y ~ G + x1 + x2`	Parallel regressions
`y ~ G / (x1 + x2)`	Separate regressions
`sqrt(y) ~ x + x^2`	Transformed
`y ~ G`	Single classification
`y ~ A + B`	Randomized block
`y ~ B + N * P`	Factorial in blocks
`y ~ x + B + N * P`	with covariate
`y ~ .  -x1`	All variables except `x1`
`y ~ .  + A:B`	Add interaction (`update`)
`Nitrogen ~ Times*(River/Site)`	More complex design

Common R functions for inference

Expression	Description
`coef(obj)`	regression coefficients
`resid(obj)`	residuals
`fitted(obj)`	fitted values
`summary(obj)`	analysis summary
`predict(obj, newdata = ndat)`	predict for newdata
`deviance(obj)`	residual sum of squares

⋆ All the examples discussed above can be expressed in the form:

$$E(y_i) = \beta_0 + \sum_{j=1}^{p} \beta_j x_{i,j}$$

⋆ The above formula can be rewritten as

$$\underline{\boldsymbol{\mu}} = E(\underline{\boldsymbol{y}}) = \boldsymbol{X}\underline{\boldsymbol{\beta}}$$

where $\boldsymbol{X}$ is a n x $(p+1)$ matrix, with all ones in the first columns, and $\underline{\boldsymbol{\beta}}$ is the vector of $\beta_0, \beta_1, \ldots, \beta_p$.

- In the example 1: $\boldsymbol{X} = (\underline{\mathbf{1}})$ and $\underline{\boldsymbol{\beta}} = (\beta_0)$

- In the example 2:

$$\boldsymbol{X} = \begin{pmatrix} 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 0 & 1 \\ \vdots & \vdots \\ 0 & 1 \end{pmatrix}_{n\text{x}2} \quad \underline{\boldsymbol{\beta}} = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}_{2\text{x}1}$$

⋆ In the example 3:

$$\boldsymbol{X} = \underset{n\text{x}2}{\begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}} \quad \underline{\boldsymbol{\beta}} = \underset{2\text{x}1}{\begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}}$$

⋆ In the example 4:

$$\boldsymbol{X} = \underset{n\text{x}3}{\begin{pmatrix} 1 & x_1 & x_1^2 \\ \vdots & \vdots & \vdots \\ 1 & x_n & x_n^2 \end{pmatrix}} \quad \underline{\boldsymbol{\beta}} = \underset{3\text{x}1}{\begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix}}$$

⋆ In the example 5:

$$\boldsymbol{X} = \begin{pmatrix} 1 & x_{1,1} & x_{1,2} & x_{1,3} & x_{1,4} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n,1} & x_{n,2} & x_{n,3} & x_{n,4} \end{pmatrix}_{n\text{x}5} \quad \underline{\boldsymbol{\beta}} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{pmatrix}_{3\text{x}1}$$

⋆ Models of the form $\mu = E(y_i) = \boldsymbol{X}\beta$ are called **linear models**, because they are linear in the **unknown parameters** comprising the elements of β.

- The file `hotdogs.dat` contains calories and sodium data for three types of hot dogs: Poultry, Beef, Meat.
- 20 Beef, 17 Meat and 17 Poultry hot dogs. These data may be thinked as samples from much larger populations.
- Looking at the density estimates it seems reasonable, at least as a first approximation, to model the distribution as Normal.
- The following model can be adopted (model 1):

$$B_1, \ldots, B_{20} \sim \text{i.i.d. } N(\mu_B, \sigma)$$

$$M_1, \ldots, M_{17} \sim \text{i.i.d. } N(\mu_M, \sigma)$$

$$P_1, \ldots, P_{17} \sim \text{i.i.d. } N(\mu_P, \sigma)$$

⋆ An equivalent formulation (model 2):

$$B_1, \ldots, B_{20} \sim \text{i.i.d. } N(\mu, \sigma)$$

$$M_1, \ldots, M_{17} \sim \text{i.i.d. } N(\mu + \delta_M, \sigma)$$

$$P_1, \ldots, P_{17} \sim \text{i.i.d. } N(\mu + \delta_P, \sigma)$$

⋆ There are three parameters for the population means and one for the standard deviation.

Model 1	Model 2	Interpretation	Approx value
μ_B	μ	MCC of Beef hot dogs	157
μ_M	$\mu + \delta_M$	MCC of Meat hot dogs	159
μ_P	$\mu + \delta_P$	MCC of Poultry hot dogs	119
$\mu_M - \mu_B$	δ_M		2
$\mu_P - \mu_B$	δ_P		-38
σ	σ	Standard deviation	29

MCC: Mean calorie content

⋆ Rewriting the data matrix in a slightly different form reveals some mathematical structure common to many models.

⋆ Let $(y_1, \cdot, y_{54})$ be the calorie contents. Two new variables are defined:

$$x_{1,i} = \begin{cases} 1 & \text{if the } i\text{-th hot dog is Meat} \\ 0 & \text{otherwise} \end{cases}$$

$$x_{2,i} = \begin{cases} 1 & \text{if the } i\text{-th hot dog is Poultry} \\ 0 & \text{otherwise} \end{cases}$$

⋆ $x_{1,i}$ and $x_{2,i}$ are indicator or dummy variables.

⋆ If there are k populations, then $k-1$ indicators suffice.

⋆ Model 2 can be stated as

$$y_i = \mu + \delta_M x_{1,i} + \delta_P x_{2,i} + \varepsilon_i$$

$$\varepsilon_1, \cdot, \varepsilon_{54} \sim \text{i.i.d. } N(0, \sigma)$$

⋆ Let

$$y = (y_1, \cdot, y_{54})^T$$

$$\underline{\boldsymbol{\beta}} = (\mu, \delta_M, \delta_P)^T$$

$$\underline{\boldsymbol{\varepsilon}} = (\varepsilon_1, \cdot, \varepsilon_{54})^T$$

$$\boldsymbol{X} = \begin{pmatrix} 1 & 0 & 0 \\ \vdots & \vdots & \vdots \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ \vdots & \vdots & \vdots \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ \vdots & \vdots & \vdots \\ 0 & 0 & 1 \end{pmatrix}_{54\text{x}3}$$

- The general form of the linear model:

$$\underline{\boldsymbol{y}} = \boldsymbol{X}\underline{\boldsymbol{\beta}} + \underline{\boldsymbol{\varepsilon}}$$

or

$$y_i = \beta_0 + \beta_1 \cdot x_{i1} + \beta_2 \cdot x_{i2} + \cdots + \beta_p \cdot x_{ip} + \varepsilon_i$$

or

$$y_i \sim N(\mu_i, \sigma)$$

with $\mu_i = \beta_0 + \sum_j \beta_j x_{j,i}$ and $\varepsilon_i \sim$ i.i.d. $N(0, \sigma)$.

- This model has $p + 2$ parameters: $\beta_0, \beta_1, \ldots, \beta_p, \sigma$.

- $\boldsymbol{X}\underline{\boldsymbol{\beta}} = \underline{\boldsymbol{\mu}}$: **systematic** part, **deterministic** part or **signal**.
- $\underline{\varepsilon}$: **random** or **noise** component.
- response = signal + noise
- Models in which the signal is a linear function of parameters (μ, δ_M, δ_P in the example) and response is a linear function of signal and noise, are linear models.

⋆ The likelihood function is

$$
\begin{aligned}
l(\beta_0, \ldots, \beta_p, \sigma) &= \prod_{i=1}^{n} p(y_i | \beta_0, \ldots, \beta_p, \sigma) = \\
&= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{1}{2} \left(\frac{y_i - \mu_i}{\sigma} \right)^2 \right) = \\
&= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{1}{2} \left(\frac{y_i - (\beta_0 + \sum_j \beta_j x_{j,i})}{\sigma} \right)^2 \right) = \\
&= (2\pi\sigma^2)^{-\frac{n}{2}} \exp -\tfrac{1}{2\sigma^2} \textstyle\sum_i (y_i - (\beta_0 + \sum_j \beta_j x_{j,i}))^2
\end{aligned}
$$

- The MLE (maximum likelihood estimate) is obtained by take the log of the likelihood expression, differentiate with respect to each parameter in turn, set the derivative equal to 0 and solve.

$$\log l(\beta_0, \ldots, \beta_p, \sigma) = c - n \log \sigma - \frac{1}{2\sigma^2} \sum_i (y_i - (\beta_0 + \sum_j \beta_j x_{j,i}))^2$$

$$\begin{cases} \frac{1}{\hat{\sigma}^2} \sum_i (y_i - (\hat{\beta}_0 + \sum_j \hat{\beta}_j x_{j,i})) = 0 \\ \frac{1}{\hat{\sigma}^2} \sum_i (y_i - (\hat{\beta}_0 + \sum_j \hat{\beta}_j x_{j,i})) x_{1,i} = 0 \\ \vdots \\ \frac{1}{\hat{\sigma}^2} \sum_i (y_i - (\hat{\beta}_0 + \sum_j \hat{\beta}_j x_{j,i})) x_{p,i} = 0 \\ -\frac{n}{\sigma} + \frac{1}{\hat{\sigma}^3} \sum_i (y_i - (\hat{\beta}_0 + \sum_j \hat{\beta}_j x_{j,i}))^2 = 0 \end{cases}$$

- The first $p+1$ equations can be multiplied by σ^2, yielding $p+1$ linear equations in the $p+1$ unknown βs.
- Because they are linear, equations can be solved by linear algebra:

$$\underline{\hat{\boldsymbol{\beta}}} = (\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\boldsymbol{T}}\underline{\boldsymbol{y}}$$

from which

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1,i} + \cdots + \hat{\beta}_p x_{p,i}$$

- $\hat{y}_i$ are the **fitted values**, while $r_i = y_i - \hat{y}_i$ are **residuals**, estimates of ε_i.
- The MLE for σ, $\hat{\sigma}$, is found from the last equation that can be rewritten as $-\frac{n}{\hat{\sigma}} + \frac{1}{\hat{\sigma}^3}\sum_i r_i^2 = 0$, from which $\hat{\sigma}^2 = \frac{1}{n}\sum_i r_i^2$.

- The main assumption of linear model is that the model is linear in the parameters.

- Other assumptions concern the error terms $\underline{\varepsilon}$:

1. $E(\varepsilon_i) = 0 \qquad (i = 1, \ldots, n)$
2. $\boldsymbol{X}$ is not stochastic
3. $\varepsilon_i \sim N(0, \sigma_\varepsilon^2) \qquad (i = 1, \ldots, n)$
4. $Corr(\varepsilon_i, \varepsilon_{i-j}) = 0 \qquad (i = 1, \ldots, n; 0 < j < i)$

- The residual normality can be checked through graphics and the normality tests, such as **Anderson-Darling test**.

- The Anderson-Darling test compute square mean difference between the empirical cumulative distribution (computed on residuals) and theoretical cumulative distribution of the normal distribution.

$$AD = \sum_{i=1}^{n} \frac{1-2i}{n} \{\ln(F_0[Z_{(i)}]) + \ln(1 - F_0[Z_{(n+1-i)}])\} - n$$

  when n is the sample size, $Z_{(i)}$ are observed ordered and standardized residuals and F_0 is the cumulative distribution function of the standard normal distribution.

- The value of AD increase when the difference between residuals distribution and gaussian cumulative distribution also increase.

- ⋆ If linear model assumptions are satisfied, $\underline{\hat{\boldsymbol{\beta}}}$ are random variables with a known distribution.

- ⋆ It is possible:

  1. to test hyphoteses of the form $H_0 : \beta_j = \beta_j^*$ that is to test that β_j (the true, and unknown value of the estimator) is equal to some (hypothesized) value β_j^*;
  2. to provide confidence intervals for the regression parameters β_j.

- To test β_j is equal to some value β_j^*, $j = 0, 1, \ldots, p$ the t-test statistic can be used:

$$t = \frac{\hat{\beta}_j - \beta_j^*}{\hat{se}(\beta_j)}$$

where $\hat{se}(\beta_j)$ denotes the estimated standard error of the estimator $\hat{\beta}_j$.

- t test statistics follow a **Student's t** distribution with $n - p - 1$ degrees of freedom, where $p + 1$ is the number of estimated parameters.

- The $1 - \alpha$ confidence interval for β_j is given by:

$$[\hat{\beta}_j - t_{n-p-1;\frac{\alpha}{2}} \hat{se}(\beta_j); \hat{\beta}_j + t_{n-p-1;\frac{\alpha}{2}} \hat{se}(\beta_j)]$$

- Linear regression can be used to fit a predictive model to an observed data set of $\underline{\boldsymbol{y}}$ and $\boldsymbol{X}$ values. After developing such a model, if an additional values of x, $\underline{\boldsymbol{x}}_{n+1}$, is then given without its accompanying value of y, the fitted model can be used to make a prediction of the value of y_{n+1}.

- It can be shown that the $1-\alpha$ confidence level for the regression for a given value of $\underline{\boldsymbol{x_*}}$ is limited by:

LCL: $f(\underline{\boldsymbol{x}}_*;\underline{\hat{\boldsymbol{\beta}}}) - t_{n-p-1;\,\frac{\alpha}{2}} \cdot \sqrt{\underline{\boldsymbol{x}}_*^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\underline{\boldsymbol{x}}_*} \cdot \hat{\sigma}_\varepsilon$

UCL: $f(\underline{\boldsymbol{x}}_*;\underline{\hat{\boldsymbol{\beta}}}) + t_{n-p-1;\,\frac{\alpha}{2}} \cdot \sqrt{\underline{\boldsymbol{x}}_*^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\underline{\boldsymbol{x}}_*} \cdot \hat{\sigma}_\varepsilon$

- ⋆ The confidence interval may be shown as the interval containing the "true" regression line with a given confidence level.
- ⋆ Instead, the prediction interval will contain, with a given confidence level, the true (and unknown) y_{n+1} value, given $\underline{\boldsymbol{x}}_{n+1}$ values.
- ⋆ The $1-\alpha$ confidence level for the prediction interval for a given value of $\underline{\boldsymbol{x}}_{n-1}$ is limited by:

  LPL: $f(\underline{\boldsymbol{x}}_*;\underline{\hat{\boldsymbol{\beta}}}) - t_{n-p-1;\,\frac{\alpha}{2}} \cdot \sqrt{\underline{\boldsymbol{x}}_*^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\underline{\boldsymbol{x}}_* + 1} \cdot \hat{\sigma}_\varepsilon$

  UPL: $f(\underline{\boldsymbol{x}}_*;\underline{\hat{\boldsymbol{\beta}}}) + t_{n-p-1;\,\frac{\alpha}{2}} \cdot \sqrt{\underline{\boldsymbol{x}}_*^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\underline{\boldsymbol{x}}_* + 1} \cdot \hat{\sigma}_\varepsilon$

- ⋆ Student's t test for independent samples is equivalent to the linear regression of the response variable on the grouping variable, where the grouping variable is recoded to have numerical values, if necessary.

- ⋆ The t test for indipendent samples p-value is equal to the slope regression parameter (β_1) p-value.

- The linear regression model says data are normally distributed about the regression line with constant standard deviation. The grouping variable takes on only two values. Therefore, there are only two locations along the regression line where there are data. "Homoschedastic normally distributed values about the regression line" is equivalent to "two normally distributed populations with equal variances".

- Thus, the hypothesis of equal means ($H_0 : \mu_A - \mu_B = 0$) is equivalent to the hypothesis that the regression coefficient of x is 0 ($H_0 : \beta_1 = 0$). That is, the population means are equal if and only if the regression line is horizontal.

- ⋆ More than two samples can be compared using ANOVA. In the ANOVA, the categorical variable is effect coded, which means that each category's mean is compared to the grand mean.
- ⋆ In the regression, the categorical variable is dummy coded, which means that each category's intercept is compared to the reference group's intercept.
- ⋆ So an ANOVA reports each mean and a p-value that says at least two are significantly different. A regression reports only one mean (as an intercept), and the differences between that one and all other means, but the p-values evaluate those specific comparisons.
- ⋆ Thus, the hypothesis of equal means ($H_0 : \mu_1 = \mu_2 = \cdots = \mu_p$) of the ANOVA is equivalent to the hypothesis that the regression dummy coefficients $\beta_1, \beta_2, \ldots, \beta_p$ are equal to zeros.

⋆ When distributional assumption about $\underline{e}$ is two restrictive and/or some of the indipendent variables are regarded as random variables:

$$\underline{y} = \boldsymbol{X}\underline{\beta} + \boldsymbol{Z}\underline{u} + \underline{e}$$

where:

- $\boldsymbol{Z}$: random-effects design matrix; can contains both continuous and dummy matrix, just like $\boldsymbol{X}$;
- $\underline{u}$: random-effects parameter.

- $\underline{u}$ and $\underline{e}$ are normally distributed with

$$E\begin{pmatrix} \underline{u} \\ \underline{e} \end{pmatrix} = \begin{pmatrix} \underline{0} \\ \underline{0} \end{pmatrix} \qquad Var\begin{pmatrix} \underline{u} \\ \underline{e} \end{pmatrix} = \begin{pmatrix} \boldsymbol{G} & \underline{0} \\ \underline{0} & \boldsymbol{R} \end{pmatrix}$$

- $Var(\underline{y}) = V = \boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}^T + \boldsymbol{R}$

- $\boldsymbol{V}$ can be modeled by defining $\boldsymbol{Z}$ and specifying covariance structures for $\boldsymbol{G}$ and $\boldsymbol{R}$.

- **Generalized least-squares** (GLS) requires knowledge of $\boldsymbol{V}$ and, therefore, of $\boldsymbol{R}$ and $\boldsymbol{G}$.
- Lacking of such information, some reasonable estimate of $\boldsymbol{V}$ can be used to minimize

$$(\underline{\boldsymbol{y}} - \boldsymbol{X}\underline{\boldsymbol{\beta}})^T \hat{\boldsymbol{V}}^{-1} (\underline{\boldsymbol{y}} - \boldsymbol{X}\underline{\boldsymbol{\beta}})$$

- A better approach entails using **likelihood-based** method, using the assumption that $\underline{\boldsymbol{u}}$ and $\underline{\boldsymbol{e}}$ are normally distributed.
- Two popular methods are ML and REML (Restricted/residual ML).

⋆ The log-likelihood functions are:

$$\text{ML:} \quad l(\boldsymbol{G}, \boldsymbol{R}) = -\tfrac{1}{2}\log|\boldsymbol{V}| - \tfrac{n}{2}\log \underline{\boldsymbol{r}}^T \boldsymbol{V}^{-1}\underline{\boldsymbol{r}} - \tfrac{n}{2}\left(1 + \log\left(\tfrac{2\pi}{n}\right)\right)$$

$$\text{REML:} \quad l_R(\boldsymbol{G}, \boldsymbol{R}) = -\tfrac{1}{2}\log|\boldsymbol{V}| - \tfrac{1}{2}\log|\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X}| + \\ -\tfrac{n-p}{2}\log(\underline{\boldsymbol{r}}^T\boldsymbol{V}^{-1}\underline{\boldsymbol{r}}) - \tfrac{n-p}{2}\left(1 + \log\left(\tfrac{2\pi}{n-p}\right)\right)$$

where

- $\underline{\boldsymbol{r}} = \underline{\boldsymbol{y}} - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X})^{-}\boldsymbol{X}^T\boldsymbol{V}^{-1}\underline{\boldsymbol{y}}$
- $p = \text{rank}(\boldsymbol{X})$

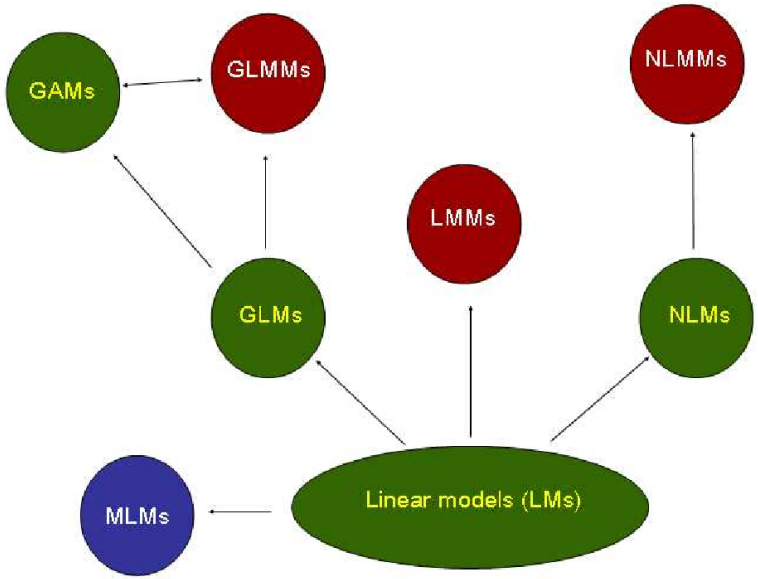
- Obtained $\hat{\boldsymbol{G}}$ and $\hat{\boldsymbol{R}}$, the mixed model equations can be solved:

$$E \begin{pmatrix} \boldsymbol{X}^T \hat{\boldsymbol{R}}^{-1} \boldsymbol{X} & \boldsymbol{X}^T \hat{\boldsymbol{R}}^{-1} \boldsymbol{Z} \\ \boldsymbol{Z}^T \hat{\boldsymbol{R}}^{-1} \boldsymbol{X} & \boldsymbol{Z}^T \hat{\boldsymbol{R}}^{-1} \boldsymbol{Z} + \hat{\boldsymbol{G}}^{-1} \end{pmatrix} \begin{pmatrix} \underline{\hat{\boldsymbol{\beta}}} \\ \underline{\hat{\boldsymbol{u}}} \end{pmatrix} = \begin{pmatrix} \boldsymbol{X}^T \hat{\boldsymbol{R}}^{-1} \underline{\boldsymbol{y}} \\ \boldsymbol{Z}^T \hat{\boldsymbol{R}}^{-1} \underline{\boldsymbol{y}} \end{pmatrix}$$

- The solutions are:

$$\underline{\hat{\boldsymbol{\beta}}} = (\boldsymbol{X}^T \hat{\boldsymbol{V}}^{-1} \boldsymbol{X})^- \boldsymbol{X}^T \hat{\boldsymbol{V}}^{-1} \underline{\boldsymbol{y}}$$

$$\underline{\hat{\boldsymbol{u}}} = \hat{\boldsymbol{G}} \boldsymbol{Z}^T \hat{\boldsymbol{V}}^{-1} (\underline{\boldsymbol{y}} - \boldsymbol{X} \underline{\hat{\boldsymbol{\beta}}})$$



Model	Acronym	R function
Linear Models	LM	`lm, aov`
Multivariate LMs	MLM	`manova`
Generalized LMs	GLM	`glm`
Linear Mixed Models	LMM	`lme, aov`
Non-linear Models	NLM	`nls`
Non-linear Mixed Models	NLMM	`nlme`
Generalized LMMs	GLMM	`glmmPQL`
Generalized Additive Ms	GAM	`gam`

- Collett (1991, p. 75) reports an experiment on the toxicity to the tobacco budworm Heliothis virescens of doses of the pyrethroid trans-cypermethrin to which the moths were beginning to show resistance.
- Batches of 20 moths of each sex were exposed for three days to the pyrethroid and the number in each batch that were dead or knocked down was recorded.
- The results were

	Dose (μg)					
Sex	1	2	4	8	16	32
Male	1	4	9	13	18	20
Female	0	2	6	10	12	16

- An experiment recorded the numbers of chromosomal abnormalities observed for various amounts and intensities of gamma radiation.
- The number of cells per measurement varied.
- The results contains, for each one of 27 measurements:
  - the number of cells per measurement;
  - the numbers of chromosomal abnormalities;
  - the amount of gamma radiation;
  - the intensity of gamma radiation.

- In the case of binomial data, response variable may be specified in three different ways:

1. If the response is a numeric vector it is assumed to hold the data in ratio form, $y_i = s_i/a_i$, in which case the a_is must be given as a vector of weights using the `weights` argument. If the a_i are all one, the default `weights` suffices.
2. If the response is a logical vector or a two-level factor it is treated as a 0/1 numeric vector and handled as previously.
3. If the response is a two-column matrix it is assumed that the first column holds the number of successes, s_i, and the second holds the number of failures, $a_i - s_i$, for each trial. In this case, no `weights` argument is required.

- Generalized Linear Models (**GLM**) are extensions of fixed-effects linear model to cases where standard linear model assumptions are violated.
- Standard linear model:

$$y_i = \beta_0 + \beta_1 \cdot x_{i1} + \beta_2 \cdot x_{i2} + \cdots + \beta_p \cdot x_{ip} + \varepsilon_i$$

The above formula can be written in a matrix format:

$$\underline{\boldsymbol{y}} = \boldsymbol{X}\underline{\boldsymbol{\beta}} + \underline{\boldsymbol{\varepsilon}}$$

or, in terms of expected values:

$$E(\underline{\boldsymbol{y}}) = \boldsymbol{X}\underline{\boldsymbol{\beta}}$$

Estimation is done by least squares, based on the assumption of normal errors.

⋆ GLM uses a likelihood-based procedure to fit $\boldsymbol{X}\underline{\boldsymbol{\beta}}$ to a function of $E(\underline{\boldsymbol{y}})$, suggested by the distribution of data.

⋆ $\underline{\boldsymbol{y}}$ is described as an additive function of **systematic** and **random** components. The first corresponds to $E(\underline{\boldsymbol{y}})$, the later to error.

$$\underline{\boldsymbol{y}} = \underline{\boldsymbol{\mu}} + \underline{\boldsymbol{\varepsilon}}$$

- $Var(\underline{\boldsymbol{y}}) = Var(\underline{\boldsymbol{\varepsilon}}) = R$.
- $\underline{\boldsymbol{\eta}} = \boldsymbol{X}\underline{\boldsymbol{\beta}}$, where $\underline{\boldsymbol{\eta}} = g(\underline{\boldsymbol{\mu}})$.
- $g(\underline{\boldsymbol{\mu}})$ is called the **link function**, because it links the linear model to the mean of $\underline{\boldsymbol{y}}$.

- Nelder and Wedderburn (1972) showed that the Maximum Likelihood Estimates (MLEs) for $\underline{\boldsymbol{\beta}}$ can be obtained iteratively solving:

$$\boldsymbol{X^T W X \beta = X^T W y^*}$$

with

- $\boldsymbol{W = D R^{-1} D}$
- $\underline{\boldsymbol{y^*}} = \underline{\hat{\boldsymbol{\eta}}} + (\underline{\boldsymbol{y}} - \underline{\tilde{\boldsymbol{\mu}}})\boldsymbol{D^{-1}}$
- $\underline{\boldsymbol{D}} = \partial \underline{\boldsymbol{\mu}} / \partial \underline{\boldsymbol{\eta}}$
- $\boldsymbol{R} = Var(\underline{\boldsymbol{\varepsilon}})$
- $\underline{\boldsymbol{\mu}} = E(\underline{\boldsymbol{y}})$

- Estimates of $\boldsymbol{D}$ and $\boldsymbol{R}$ are used in place of $\boldsymbol{D}$ and $\boldsymbol{R}$.

- In the case of standard linear model, $\underline{\boldsymbol{\eta}} = E(\underline{\boldsymbol{y}}) = \underline{\boldsymbol{\mu}}$ and $\boldsymbol{D} = \boldsymbol{I}$. Thus, $\boldsymbol{X^T R^{-1} X \beta = X^T R^{-1} y}$, the normal equation.

- ⋆ Elements for estimating $\underline{\boldsymbol{\beta}}$:
  - **link function**: determines $\underline{\boldsymbol{\nu}}$ and $\boldsymbol{D}$;
  - **probability function**: determines $\underline{\boldsymbol{\mu}}$ and $\boldsymbol{R}$.

- ⋆ The process for selecting a link function and the structure of the mean and variance can be understood by looking at the probability distribution, or, better said, to the likelihood function.

- ⋆ Three cases will be considered: the Binomial, the Poisson and the Normal.

- n trials, each trial with succes probability of π:

$$f(y_n) = \binom{n}{y_n} \pi^{y_n} (1-\pi)^{n-y_n}$$

- The log-likelihood function is:

$$l(\pi; y_n) = y_n \log\left(\frac{\pi}{1-\pi}\right) + n\log(1-\pi) + \log\binom{n}{y_n}$$

- A sample proportion, $y = y_n/n$, thus has a log-likelihood function of

$$l(\pi; y_n) = ny_n \log\left(\frac{\pi}{1-\pi}\right) + n\log(1-\pi) + \log\binom{n}{ny_n}$$

- Mean and variance of a sample proportion are:

$$E(y) = \pi \qquad\qquad Var(y) = \frac{\pi(1-\pi)}{n}$$

- The probability function of the Poisson distribution is:

$$f(y) = \frac{\lambda^y e^{-\lambda}}{y!}$$

- The log-likelihood function is:

$$l(\lambda; y) = y \log(\lambda) - \lambda - \log(y!)$$

- Mean and variance are:

$$E(y) = \lambda \qquad Var(y) = \lambda$$

⋆ The probability function of the Normal distribution is:

$$f(y) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(y-\mu)^2\right)$$

⋆ The log-likelihood function is:

$$l(\mu, \sigma^2; y) = \left(-\frac{1}{2\sigma^2}\right)(y-\mu)^2 - \frac{1}{2}\log(2\pi\sigma^2)$$

⋆ Mean and variance are:

$$E(y) = \mu \qquad\qquad Var(y) = \sigma^2$$

- The log-likelihood functions for these three distributions have a common form,

$$l(\theta, \phi; y) = \frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)$$

where

  - θ is the natural parameter,
  - ϕ is the scale parameter.

- θ is a function of the mean, $\theta(\mu)$.

- It is also possible express the variance as a function of the mean and $a(\phi)$:

$$Var(y) = V(\mu)a(\phi)$$

where $V(\mu)$ is the **variance function**.

	Sample proportion	Poisson	Normal
$E(y)$	π	λ	μ
$\theta(\mu)$	$\log(\pi/(1-\pi))$	$\log(\lambda)$	μ
$a(\phi)$	$1/n$	1	σ^2
$V(\mu)$	$\pi(1-\pi)$	λ	1
$Var(y)$	$\pi(1-\pi)/n$	λ	σ^2

- A distribution whose log-likelihood has the general form

$$l(\theta, \phi; y) = \frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)$$

  is a member of the **exponential family**.

- The GLM can be applied to data distributed according to the exponential family.

- If $y_1, \ldots, y_n$ is a random sample from such a family, the log-likelihood of y_i is

$$l(\theta_i, \phi_i; y_i) = \frac{y_i\theta_i - b(\theta_i)}{a(\phi_i)} + c(y_i, \phi_i)$$

- The joint log-likelihood of $y_1, \ldots, y_n$ is

$$l(\underline{\boldsymbol{\theta}}, \underline{\boldsymbol{\phi}}; y_1, \ldots, y_n) = \sum_i \frac{y_i\theta_i - b(\theta_i)}{a(\phi_i)} + c(y_i, \phi_i)$$

- ⋆ Observations are linear in the **natural parameter** θ.
  - For normally distributed data: $\theta = \mu$.
  - For Poisson distributed data: $\theta = \log(\lambda)$.
  - For Binomial distributed data: $\theta = \log(\pi/(1-\pi))$.

- ⋆ In these examples, $\theta(\mu)$ is used as a link function, they are called the **canonical link functions**.

⋆ The structure of the variance-covariance matrix of $\underline{\boldsymbol{y}}$, or errors, can be described in terms of the scale parameter and variance function. Specifically,

$$Var(\underline{\boldsymbol{\varepsilon}}) = \boldsymbol{R} = \boldsymbol{R_{\mu}^{1/2} A R_{\mu}^{1/2}}$$

where

- R_{μ} is the diagonal matrix whose i-th diagonal element is $V(\mu_i)$, the variance function for the i-th observation;
- $R_{\mu}^{1/2}$ is the diagonal matrix of square roots of the corresponding elements of R_{μ};
- A is the scale parameter matrix.

Distribution	$\boldsymbol{R_\mu}$	$\boldsymbol{A}$
Normal	$\boldsymbol{I}$	$\boldsymbol{I_{\sigma^2}}$
Poisson	$\text{diag}(\lambda_i)$	$\boldsymbol{I}$
Sample proportion	$\text{diag}(\pi_i/(1-\pi_i))$	$\text{diag}(1/n_i)$

- The **inverse link function** (sometimes referred to as the **mean function**) is defined as $h(\eta) = \underline{\boldsymbol{\mu}}$.
- It can be used to predict $\underline{\boldsymbol{\mu}}$ from $\underline{\hat{\boldsymbol{\beta}}}$.
- It is assumed that the relationship between $\underline{\boldsymbol{\eta}}$ and $\underline{\boldsymbol{\mu}}$ is one-to-one, thus $h(\eta) = g^{-1}(\eta)$. For some complex GLMs, this may be false.
- Because $\underline{\boldsymbol{\eta}}$ is estimated by $\boldsymbol{X}\hat{\boldsymbol{\beta}}$, $\hat{\mu} = h(\boldsymbol{X}\underline{\hat{\boldsymbol{\beta}}})$.
  - Normal:
    $$\eta = \mu \Rightarrow \mu = h(\boldsymbol{X}\underline{\hat{\boldsymbol{\beta}}}) = \boldsymbol{X}\underline{\hat{\boldsymbol{\beta}}}$$
  - Poisson:
    $$\eta = \log \lambda \Rightarrow \lambda = h(\boldsymbol{X}\underline{\hat{\boldsymbol{\beta}}}) = \exp \boldsymbol{X}\underline{\hat{\boldsymbol{\beta}}}$$
  - Sample proportion:
    $$\eta = \log\left(\pi/(1-\pi)\right) \Rightarrow \lambda = h(\boldsymbol{X}\underline{\hat{\boldsymbol{\beta}}}) = \exp \boldsymbol{X}\underline{\hat{\boldsymbol{\beta}}}/(1+\exp \boldsymbol{X}\underline{\hat{\boldsymbol{\beta}}})$$

- ⋆ Deviance is defined as:

$$Dev(\underline{\hat{\boldsymbol{\mu}}}; \underline{\boldsymbol{y}}) = 2\left(l(\underline{\boldsymbol{y}}; \underline{\boldsymbol{y}}) - l(\underline{\hat{\boldsymbol{\mu}}}; \underline{\boldsymbol{y}})\right)$$

  where
  - $l(y; y)$ is the value of the maximum log-likelihood achievable in a full model; i.e. it is the value of the log-likelihood for which θ is expressed as a function of y;
  - $l(\hat{\mu}; y)$ is the value of the log-likelihood over β, i.e. it is the vale of the log-likelihood for which θ is expressed as a function of $\hat{\mu}$.
- ⋆ The deviance is a generalization of the Sum of Squares for Error in the Analysis of Variance and the likelihood ratio χ^2 in contingency tables:
- ⋆ Deviance can be used to:
  - evaluate **goodness-of-fit**;
  - **test hypotheses**.

- Dev is a function of θ and ϕ.

- When ϕ is known then:
  - $Dev \sim \chi^2(n-p)$ where n is the number of observations and p is the number of parameters in $\underline{\boldsymbol{\beta}}$ (including intercept, if any);
  - therefore, Dev can be used as a χ^2 statistics to test the goodness-of-fit (GOF) of the model.

- When ϕ is unknown, it can be estimated and used to compute the **scaled deviance**:

$$Dev^*(\underline{\hat{\boldsymbol{\mu}}}; \underline{\boldsymbol{y}}) = Dev(\underline{\hat{\boldsymbol{\mu}}}; \underline{\boldsymbol{y}})/\hat{\phi}$$

- If $\underline{\boldsymbol{\beta}}$ is partitioned in $\underline{\boldsymbol{\beta_1}}$ and $\underline{\boldsymbol{\beta_2}}$ then

$$\boldsymbol{X}\underline{\boldsymbol{\beta}} = \boldsymbol{X_1}\underline{\boldsymbol{\beta_1}} + \boldsymbol{X_2}\underline{\boldsymbol{\beta_2}}$$

- The difference between the deviance of the full model and that of the model fitting $\boldsymbol{X_1}\underline{\boldsymbol{\beta_1}}$ ca be used as likelihood ratio (LR) to test

$$H_0 : \underline{\boldsymbol{\beta_2}} = \underline{\mathbf{0}}$$

$$LR \sim \chi^2 p - p_1$$

where p_1 is the number of parameters in $\underline{\boldsymbol{\beta_1}}$.

- When ϕ is unknown, the scaled difference can be used to test hypotheses.

- It can be shown that

$$Var(\underline{\boldsymbol{\beta}}) = \left(\boldsymbol{X^T W X}\right)^{-1}$$

then

$$Var\left(\boldsymbol{K^T}\underline{\boldsymbol{\beta}}\right) = \boldsymbol{K^T}\left(\boldsymbol{X^T W X}\right)\boldsymbol{K}$$

where $\boldsymbol{K^T}\underline{\boldsymbol{\beta}}$ is an estimable function.

- The **Wald statistic** for $H_0 = \boldsymbol{K^T}\underline{\boldsymbol{\beta}} = \boldsymbol{K^T}\underline{\boldsymbol{\beta_0}}$ is

$$\left(\boldsymbol{K^T}\underline{\boldsymbol{\beta}} - \boldsymbol{K^T}\underline{\boldsymbol{\beta_0}}\right)^T \left[\boldsymbol{K^T}\left(\boldsymbol{X^T W X}\right)^{-1}\boldsymbol{K}\right]\left(\boldsymbol{K^T}\underline{\boldsymbol{\beta}} - \boldsymbol{K^T}\underline{\boldsymbol{\beta_0}}\right)$$

- The scalar form of the Wald statistic is

$$\frac{\beta - \beta_0}{Var(\beta - \beta_0)}$$

- The Wald statistic is distributed as a χ^2_ν where $\nu = \text{rank}(\boldsymbol{K})$.

- Residuals are computed as in the linear model:

$$r_i = y_i - \hat{y}_i$$

- Residual diagnostic test the fulfillment of the assumptions of the generalized linear model:

1. testing the assumption of homogeneity of error variance;
2. testing for outliers.

- The assumption of homogeneity of error variance can be performed with the Levene's test.

- Levene's test is performed by computing:

$$Z_{ij} = |y_{ij} - E(y_{.j})|$$

  and then computing an F test on the j groups.

- A nonsignificant result indicates no heteroskedasticity.

- ⋆ Outliers can be detected:
  1. Look for standardized residuals greater than 3.5 or less than -3.5
  2. and look for high Cook's D, greater than $4p/(n-p-1)$. For instance, if $n = 100$, $p = 5$ then high Cook's D are higher than $4 \cdot 5/(100 - 5 - 1) = 20/94$.

- ⋆ Standardized residuals are computed by

$$\frac{r_i - \hat{\mu}_r}{\hat{\sigma}_r}$$

where $\hat{\mu}_r$ and $\hat{\sigma}_r$ are the mean and the standard deviation of r_i, respectively.

- Cook's D are computed by

$$\text{Cook's } D_i = \left(\frac{1}{p}\right)\left(\frac{h_{ii}}{1-h_{ii}}\right)\left(\frac{r_i^2}{\hat{\sigma}^2 \cdot (1-h_{ii})}\right)$$

where h_{ii} is the i-th diagonal entry of the hat matrix $\boldsymbol{H}$.

- Cook and Weisberg (1982) suggested that values of D_i that exceed 50% of the F distribution with p and $n-p$ degrees of freedom are large

- The hat matrix $\boldsymbol{H}$ transforms $\boldsymbol{Y}$ into the predicted scores. It is computed by

$$\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$$

- The diagonals for the hat matrix indicate which values could be outliers or not.
- The diagonals are therefore measures of leverage.
- Leverage is bounded by two limits: $1/n$ and 1. The closer the leverage is to unity, the more levarage the value has.
- The trace of the hat matrix is equal to the number of variables in the model.
- When the leverage is greater than $2p/n$ ther is high leverage according to Belsley et al. (1980), cited in Long J.F., Modern Methods in Data Analysis (page 262). For smaller samples, Vellman and Welsch (1981) suggested that $3p/n$ is the criterion.

- The leverage of outliers can be assessed:
  - constructing and analyzing studentized residuals;
  - constructing and analyzing the leverage of the high and low studentized residuals;
  - using Cook's D to help determine how problematic outliers are.

- Studentized residuals are computed by

$$\frac{r_i}{\hat{\sigma}\sqrt{1-h_{ii}}}$$

- ⋆ A Generalized Linear Mixed Model (GLMM) is an extension to the GLM in which the linear predictor contains random effects in addition to the usual fixed effects.
- ⋆ A GLMM use the **quasi-likelihood**, an extension of likelihood methods.
- ⋆ The quasi-likelihood function is defined as follows.
  - Given n observation y_i $(i = 1, \ldots, n)$ such that $E(y_i) = \mu_i$, $Var(y_i) \propto Var(\mu_i)$ (known function), $\mu = function(\beta_1, \ldots, \beta_p)$ the quasi-likelihood function is defined by

$$\frac{\partial Q(\mu_i, y_i)}{\partial \mu_i} = \frac{y_i - \mu_i}{Var(\mu_i)}$$

  - An example of $\mu = function(\beta_1, \ldots, \beta_p)$ is given by $\mu_i = h(\eta_i)$, $\eta_i = \sum_j X_{ij}\beta_j$, that is the inverse link function of a GLM
  - The log-likelihood function is a special case of $Q(\mu_i, y_i)$.

- ⋆ In the normal errors mixed model $\underline{y} = \boldsymbol{X}\underline{\beta} + \boldsymbol{Z}\underline{u} + \underline{e}$
  - the conditional mean of the observations given the random model effects is $E(\underline{y}|\underline{u}) = \mu = \boldsymbol{X}\underline{\beta} + \boldsymbol{Z}\underline{u}$,
  - the conditional variance is $Var(\underline{y}|\underline{u}) = \boldsymbol{R} = Var(\underline{e})$.

- ⋆ Observation can be described as $\underline{y} = \underline{\mu} + \underline{e}$ where $\underline{\mu}$ is the conditional mean, $E(\underline{y}|\underline{u})$, $u \sim MVN(\underline{0}, \boldsymbol{G})$.

- In the generalized linear mixed model, the **conditional distribution** of $\underline{\boldsymbol{y}}|\underline{\boldsymbol{u}}$ plays the same role as the distribution of $\underline{\boldsymbol{y}}$ in the fixed effects generalized linear model.

- The **conditional quasi-likelihood** of an observation, $\underline{\boldsymbol{y_i}}|\underline{\boldsymbol{y}}$ is

$$Q(u_i, y_i|u_i) = \left[\frac{y_i\theta_i - b(\theta_i)}{a(\phi_i)}\right]$$

⋆ The **joint quasi-likelihood** of an observation, $\underline{y_i}|\underline{y}$ of the observations, $\underline{y}$, is the sum of the quasi-likelihood of $\underline{y}|\underline{u}$ and $\underline{u}$. In matrix terms, the joint quasi-likelihood is

$$Q(\underline{\mu}, \underline{u}; \underline{y}) = \left[\underline{y}^{T} A^{-1} \underline{\theta} - (\underline{b_{\theta}^{1/2}})^{T} A^{-1} \underline{b_{\theta}^{1/2}})\right] + \frac{1}{2} \underline{u}^{T} G^{-1} \underline{u}$$

where:

- A is the matrix of $a(\phi_i)$,
- $\underline{\theta}$ is the vector of $\theta(\mu_i)$,
- $\underline{b_{\theta}}$ is the vector of $b(\theta_i)$.

⋆ The strategies for fitting a GLMM to $E(\underline{\boldsymbol{y}})$ are the same as those for a GLM:

- the form of the conditional quasi-likelihood determines the **variance structure** and contains the natural parameter,
- the conditional mean function $\theta(\mu)$ is used a **canonical link**.

$$\eta = \boldsymbol{X}\underline{\boldsymbol{\beta}} + \boldsymbol{Z}\underline{\boldsymbol{u}}$$

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

⋆ The strategies for fitting a GLMM to $E(\underline{\boldsymbol{y}})$ are the same as those for a GLM:

- the form of the conditional quasi-likelihood determines the **variance structure** and contains the natural parameter,
- the conditional mean function $\theta(\mu)$ is used a **canonical link**.

$$\eta = \boldsymbol{X}\underline{\boldsymbol{\beta}} + \boldsymbol{Z}\underline{\boldsymbol{u}}$$

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