Recall the general mean-variance specification

\[ \begin{align*}
E(Y|x) &= f(x, \beta), \\
\text{var}(Y|x) &= \sigma^2 g(\beta, \theta, x)^2.
\end{align*} \]

The form of \( f(x, \beta) \) may be suggested by subject matter theory, or by exploration.

How can we decide on the form of \( g(\beta, \theta, x) \)?
If a Generalized Linear Model seems appropriate:
- the canonical link function is at least a starting point for $f(x, \beta)$;
- the associated variance function determines $g(\beta, \theta, x)$.

For most of the common instances of GLMs,

$$g(\beta, \theta, x) = f(x, \beta)^\theta$$

for an appropriate power $\theta$.

Otherwise, we study the residuals to gain knowledge about $g(\cdot)$. 
Residuals

Since

\[ g(\beta, \theta, x) = \frac{\text{var}(Y|x)}{\sigma^2} = \frac{\mathbb{E}\left\{ [Y - f(x, \beta)]^2 \right\}}{\sigma^2}, \]

a natural place to start is with the residuals \( r_j = Y_j - f\left(x_j, \hat{\beta}_{\text{OLS}}\right) \) from an initial unweighted (OLS) fit.

Plot the residuals against:

- Predicted values \( \hat{Y}_j = f\left(x_j, \hat{\beta}_{\text{OLS}}\right) \), possibly log-transformed;
- Covariates, elements of \( x_j \).

Non-constant variance often shows up as a wedge-shaped plot.
Example

Pharmacokinetics of indomethacin.

- Concentration $Y_j$ of indomethacin taken at $n = 11$ times points $x_j$ (hours) post-dose.
- The mean function

$$f(x_j, \beta) = e^{\beta_1} \exp(-e^{\beta_2} x_j) + e^{\beta_3} \exp(-e^{\beta_4} x_j)$$

- The variance tends to increase with the level of the response.
- A popular choice is the power-of-the-mean model $\sigma^2 f^{2\theta}(x_j, \beta)$, with, say, $\theta = 1$. 

Specifying the Variance Function
OLS and GLS fits

source("R-code/indo-GLS-nls.R")

The OLS fit (red line) and the GLS fit with $\theta = 1$ (blue line) are discernibly different.
OLS residuals

\[
\text{plot(fitted(nOLS), residuals(nOLS)^2, log = "x")}
\]

The squared OLS residuals vs the log of predicted values clearly shows a wedge shape, showing a rather severe increase in variance with level of the response.
Studentized Residuals

Consider the (matrix form of the) linear model

\[ E(Y) = X\beta, \]
\[ \text{var}(Y) = \sigma^2 W^{-1}. \]

OLS estimator is

\[ \hat{\beta}_{\text{OLS}} = (X^TX)^{-1}X^TY \]

with fitted values

\[ \hat{Y} = X\hat{\beta}_{\text{OLS}} = X(X^TX)^{-1}X^TY = HY, \]

where \( H = X(X^TX)^{-1}X^T \) is the hat matrix.
The residuals are

\[ r = Y - \hat{Y} = (I - H)Y. \]

If \( W = I \), i.e. the errors have the same variance, then

\[ \text{var}(r) = \sigma^2(I - H), \]

so the residuals do not in general have the same variance.

The studentized residuals are

\[ b_j = \frac{r_j}{\hat{\sigma}_{OLS} \sqrt{1 - h_{j,j}}}. \]

Here \( h_{j,j} \), the \( j^{th} \) diagonal entry in the hat matrix \( H \), measures the leverage of the \( j^{th} \) observation.
Extend the notions of leverage and studentization, approximately, to nonlinear mean models.

By linear approximation:

\[ r \approx \{ I_n - H(\beta) \} \{ Y - f(\beta) \} \]

where \( H(\beta) \) is the approximate hat matrix

\[ H(\beta) = X(\beta) \{ X(\beta)^T X(\beta) \}^{-1} X(\beta)^T \]

and \( X(\beta) \) is the gradient matrix

\[ X(\beta) = \begin{pmatrix}
  f_\beta(x_1, \beta) \\
  f_\beta(x_2, \beta) \\
  \vdots \\
  f_\beta(x_n, \beta)
\end{pmatrix} \]
Regard the diagonal elements of $H(\beta)$ as approximate “leverage values”.

We can define the studentized residuals in the same way as for the linear model, with approximate leverage values $h_{j,j}(\hat{\beta})$. 
Remarks

In a nonlinear model, the ramifications of “design” will be not only through the actual design points, but also through the behavior of the function $f$ at those points.

Unlike a linear model, a nonlinear model allows the changes in $f$ at different $x_j$ settings to be different, as the derivative of $f$ depends on both $x_j$ and $\beta$ in general.

Depending on how $f$ changes in different parts of the design space, different observations exert different amounts of influence on the fit.
The ordinary residual plot and the studentized residual plot

```r
source("R-code/nls2lm.R")
lOLS <- nls2lm(nOLS)
par(mfcol = c(2, 1))
plot(fitted(nOLS),
     residuals(lOLS) / summary(lOLS)$sigma, log = "x")
plot(fitted(nOLS), rstandard(lOLS), log = "x")
```

They look similar in this example.

**Note**

`rstandard()` produces Studentized residuals as defined earlier; `rstudent()` produces slightly different values ("Rstudent").
Transformations of studentized residuals

If \( g(\beta, \theta, x_j) = \exp\{\theta f(x_j, \beta)\} \), then

\[
\log[\{\text{var } Y_j\}^{1/2}] = \log \sigma + \theta f(x_j, \beta).
\]

So we plot \( \log |r_j| \) or \( \log |b_j| \) vs \( \hat{Y}_{OLS} = f(x_j, \hat{\beta}_{OLS}) \).

If \( g(\beta, \theta, x_j) = f^\theta(x_j, \beta) \), then

\[
\log[\{\text{var } Y_j\}^{1/2}] = \log \sigma + \theta \log f(x_j, \beta).
\]

So we plot \( \log |r_j| \) or \( \log |b_j| \) vs \( \log(\hat{Y}_{OLS}) = \log f(x_j, \hat{\beta}_{OLS}) \).
Log transformation and power transformation of the studentized residuals vs log transformation of the predicted values. It confirms that the power-of-the-mean model is reasonable.
Graphical verification

Check graphically for evidence that the assumed variance model accounts adequately for the form of the nonconstant variance.

Based on the original OLS fit, we decide on a tentative form for $g(\beta, \theta, x_j)^2$ and refit (e.g., by GLS-PL).
Define *weighted* residuals

\[ r_{\text{weighted}, j} = \frac{Y_j - f(x_j, \hat{\beta})}{g(\hat{\beta}, \hat{\theta}, x_j)}, \]

and *standardized* weighted residuals

\[ r_{\text{weighted}, j} / \hat{\sigma}. \]
To account for leverage, write as a vector,

\[ r_{\text{weighted}} = \mathbf{W}^{1/2} \left\{ \mathbf{Y} - f(\hat{\mathbf{\beta}}) \right\} \]

where

\[ \mathbf{W} = \text{diag} \left\{ \frac{1}{g(\hat{\mathbf{\beta}}, \hat{\mathbf{\theta}}, \mathbf{x}_1)^2}, \frac{1}{g(\hat{\mathbf{\beta}}, \hat{\mathbf{\theta}}, \mathbf{x}_2)^2}, \ldots \frac{1}{g(\hat{\mathbf{\beta}}, \hat{\mathbf{\theta}}, \mathbf{x}_n)^2} \right\}. \]

Write

\[ \mathbf{Y}^* = \mathbf{W}^{1/2} \mathbf{Y}, \]
\[ \mathbf{X}^*(\mathbf{\beta}) = \mathbf{W}^{1/2} \mathbf{X}(\mathbf{\beta}), \]
\[ f^*(\mathbf{\beta}) = \mathbf{W}^{1/2} f(\mathbf{\beta}). \]
Then another linear approximation gives

\[ r_{\text{weighted}} = Y^* - f^*(\hat{\beta}) \]

\[ \approx \{I_n - H^*(\beta)\} \{Y^* - f^*(\beta)\}. \]

Here \( H^*(\beta) \) is again an approximate hat matrix

\[ H^*(\beta) = X^*(\beta) \{X^*(\beta)^T X^*(\beta)\}^{-1} X^*(\beta)^T. \]

**Studentized weighted residuals** are then

\[ b_{\text{weighted},j} = \frac{r_{\text{weighted},j}}{\hat{\sigma} \sqrt{1 - h_{j,j}^*(\hat{\beta})}}. \]
The weighted residuals plots suggest that weighting according to the conjectured variance model has taken appropriate account of the nonconstant variance.
Final remarks about specifying variance function in this example:

- It is important to “get the variance right”.

Consider estimation of the terminal half-life, a parameter of great physical interest to pharmacologists.

The terminal half-life is the time that it takes the mean response in the “second phase” of the curve to decrease by half and is useful in determining appropriate dosing regimens. The terminal half-life is given by $\log 2/e^{\beta_4}$ hours here.
Substituting the estimate of $\beta_4$ yields an estimated half-life of 3.13 hours based on the OLS fit, and 3.96 hours based on the GLS-PL fit.

The difference in point estimates is nearly one hour, which in a clinical sense is quite a big difference.

Of course, we have not yet obtained standard errors for these point estimates, so whether or not this difference is of statistical importance is not clear.
REML Estimation of Variance

Motivation

Plots based on ordinary residuals may be misleading due to failure to account for “leverage”.

Will methods for estimation of variance parameters based on ordinary residuals also be subject to the same problem?

Recall PL estimating equations for $\sigma$ and $\theta$:

$$
\sum_{j=1}^{n} \left[ \frac{\{Y_j - f(x_j, \beta)\}^2}{\sigma^2 g(\beta, \theta, x_j)^2} - 1 \right] \times \left( \frac{1}{\nu_{\theta}(\beta, \theta, x_j)} \right) = 0.
$$
When plugging the true $\beta$,

$$
\sum_{j=1}^{n} \frac{(Y_j - f(x_j, \beta))^2}{\sigma^2 g(\beta, \theta, x_j)^2} \times \left( \begin{array}{c} 1 \\ \nu_\theta(\beta, \theta, x_j) \end{array} \right) = \sum_{j=1}^{n} \left( \begin{array}{c} 1 \\ \nu_\theta(\beta, \theta, x_j) \end{array} \right).
$$

because

$$
E \left[ \frac{(Y_j - f(x_j, \beta))^2}{\sigma^2 g(\beta, \theta, x_j)^2} \right] = 1.
$$
When plugging the estimated $\hat{\beta}$,

$$
\sum_{j=1}^{n} \frac{\left\{ Y_j - f(x_j, \hat{\beta}) \right\}^2}{\sigma^2 g(\hat{\beta}, \theta, x_j)^2} \times \begin{pmatrix} 1 \\ \nu_\theta(\hat{\beta}, \theta, x_j) \end{pmatrix} = \sum_{j=1}^{n} (1 - h_{j,j}^*(\hat{\beta})) \begin{pmatrix} 1 \\ \nu_\theta(\hat{\beta}, \theta, x_j) \end{pmatrix}.
$$

because

$$
E \left[ \frac{\left\{ Y_j - f(x_j, \hat{\beta}) \right\}^2}{\sigma^2 g(\hat{\beta}, \theta, x_j)^2} \right] \approx 1 - h_{j,j}^*(\hat{\beta}).
$$
Modified estimating equations:

\[
\sum_{j=1}^{n} \frac{\left\{ Y_j - f(x_j, \hat{\beta}) \right\}^2}{\sigma^2 g(\hat{\beta}, \theta, x_j)} \times \left( \begin{array}{c} 1 \\ \nu_\theta(\hat{\beta}, \theta, x_j) \end{array} \right) = \sum_{j=1}^{n} (1 - h_{j,j}^*) \left( \begin{array}{c} 1 \\ \nu_\theta(\hat{\beta}, \theta, x_j) \end{array} \right) = \left( \begin{array}{c} n - p \\ \sum_{j=1}^{n} (1 - h_{j,j}^*) \nu_\theta(\hat{\beta}, \theta, x_j) \end{array} \right),
\]

where \( \sum_{j=1}^{n} h_{j,j}^* = \text{trace}(H^*(\beta)) = \text{rank}(H^*(\beta)) = p \).
Solution for $\sigma^2$:

$$\hat{\sigma}^2 = \frac{1}{n - p} \sum_{j=1}^{n} \left\{ Y_j - f(x_j, \hat{\beta}) \right\}^2 \frac{1}{g(\hat{\beta}, \theta, x_j)^2}$$

That is, the modified PL equations yield the bias-adjusted estimator of $\sigma^2$. 
Recall that the original PL estimating equations arise from maximizing, for fixed $\beta$, the normal log-likelihood

$$PL(\hat{\beta}, \theta, \sigma) = -n \log \sigma - \sum_{j=1}^{n} \log g(\hat{\beta}, \theta, x_j) - \frac{1}{2} \sum_{j=1}^{n} \frac{(Y_j - f(x_j, \hat{\beta}))^2}{\sigma^2 g(\hat{\beta}, \theta, x_j)}.$$ 

The modified equations arise in the same way from maximizing

$$PL(\hat{\beta}, \theta, \sigma) + p \log \sigma - \frac{1}{2} \log \det \left\{ X(\hat{\beta})^T W(\hat{\beta}, \theta) X(\hat{\beta}) \right\}.$$ 

Specifying the Variance Function
Terminology

The extra terms can be viewed as “penalty terms” that impose a restriction on the solution, hence

\[ \text{REML} = \text{REstricted ML}. \]

Modified likelihoods of this form were first suggested based on the marginal distribution of residuals from a linear fit, hence

\[ \text{REML} = \text{REsidual ML}. \]

In some cases (linear mixed model, generalized linear model), they may be derived from the conditional distribution of the observations, conditioned on the sufficient statistic for \( \beta \). This is a partial, or restricted likelihood.