Example: GLM estimation: Seizure data

A clinical trial was conducted in order to evaluate the impact of Progabide on the frequency of epileptic seizures. Patients were randomized to either receive or not receive Progabide. The data set contains information on:

- age at start of study (AGE; measured in years)
- baseline seizure count; defined as the number of seizures in the 8 weeks prior to the study's commencement (BASE)
- treatment indicator (Z; 1=treated, 0=placebo)
- seizure counts in each of 4 two-week periods (Y1, Y2, Y3, Y4)

The investigators define the outcome as total post treatment seizure count: $Y_i \equiv \sum_{j=1}^{4} Y_{ij}$.

(a) Fit the following model

$$E[Y_i] \quad = \quad \beta_0 + \beta_1 A_i + \beta_2 B_i + \beta_3 Z_i$$

using ordinary least squares and save the fitted values, $\widehat{Y}_i$, and studentized residuals, $\widehat{r}_i$.

- Systematic component:

$$\mu_i = \beta_0 + \beta_1 A_i + \beta_2 B_i + \beta_3 Z_i$$

- Random component:

$$Y_i \sim N(\mu_i, \sigma^2)$$

(b) Interpret β_3 coefficient. Test whether the treatment has an effect on the mean number of seizures.

- β_3: difference in mean seizure count between Progabide and Placebo treatment groups, adjusted for age and baseline seizure count.

- p-value= 0.5887, highly insignificant.

(c) Considering the data structure, is the above model appropriate?

- NO. Y is the seizure count. 1) Normality and 2) constant variance assumptions are not satisfied

(d) Plot the $\widehat{r_i}$ against $\widehat{Y_i}$. What evidence does this plot provide with respect to the model assumptions?

- Variance increases as the predicted Y increases. Equal variance assumption is not satisfied.

(e) Plot a histogram of the residuals, and do a q/q plot. Does the normality assumption appear reasonable?

- QQ plot does not follow the 45 degree line. Normality assumption is not satisfied.

(f) Are the $\widehat{Y_i}$ values all reasonable?

- Since Y_is are seizure counts, $\widehat{Y}_i$ should be > 0. In this data, 10% of $\widehat{Y}_i < 0$.

(g) Fit the transformed model,

$$E[T_i] \quad = \quad \beta_0 + \beta_1 A_i + \beta_2 B_i + \beta_3 Z_i.$$

where $T_i = \log(Y_i + 0.1)$. What is the sense in shifting Y_i in this case?

- to avoid $log(0)$.

(h) In transforming the response, what issues are we attempting to address?

- Non-normality and non-constant variance (heteroscedasticity)

(i) Based on log-transformed model, test $H_0 : \beta_3 = 0$. Compare your result to that from the original model.

- P-value=0.0657. Still insignificant at the level $\alpha = 0.05$.

(j) Assess whether the transformation was successful in remedying the lack of adherence to the model assumptions.

- Residual and QQ plots indicate that the transformation addressed the violation of normality and constant variance assumptions to some extent.

(k) Interpret β_3 based on the transformed model.

- β_3: Difference in mean log seizure count between Progabide and Placebo treatment groups, adjusted for age and baseline seizure count.
- Limitation: β_3 does not represent log mean seizure count difference.

(ℓ) Is $\exp\{\widehat{T}_i\}$ the mean seizure count estimate, i.e. estimate of $E(Y_i)$?

- Since

$$\exp\{\widehat{T}_i\} \approx \exp\{E(log(Y_i))\} \neq E\{\exp(log(Y_i))\} = E(Y_i)$$

So the answer is NO.

(m) Fit the following generalized linear model,

$$\log\{E[Y_i]\} \quad = \quad \beta_0 + \beta_1 A_i + \beta_2 B_i + \beta_3 Z_i.$$

using PROC GENMOD. Interpret β coefficients.

- Systematic component:

$$log(\lambda_i) = \beta_0 + \beta_1 A_i + \beta_2 B_i + \beta_3 Z_i$$

- Random component:

$$Y_i \sim Poisson(\lambda)$$

- β_3: Difference in log mean seizure count between Progabide and Placebo treatment groups, adjusted for age and baseline seizure count.

(n) Set up an iteratively re-weighted least squares algorithm to fit the previously specified GLM.

- Set β_0 and ζ

- In the j_{th} iteration, calculate

$$\begin{aligned}
\eta_{i,j} &= X_i^T \beta_j;\ \mu_{i,j} = exp(\eta_{ij});\ v(\mu_{i,j}) = \mu_{i,j} \\
\eta_j &= (\eta_{1,j}, ..., \eta_{n,j})';\quad \mu_j = (\mu_{1,j}, ..., \mu_{n,j})' \\
V_j &= diag\{v(\mu_{1,j}), ..., v(\mu_{n,j})\} \\
Z_j &= \eta_j + V_j^{-1}(Y - \mu_j) \\
\beta_{j+1} &= (X^T V_j X)^{-1} X^T V_j Z_j
\end{aligned}$$

- Stop when $||\beta_{j+1} - \beta_j|| < \zeta$

(o) Using PROC IML, Fit the GLM using iteratively re-weighted least squares.

See the SAS code.

(p) Recall that patients were followed for 8 weeks before randomization and 8 weeks after. An alternative to modeling number of seizures is to model whether or not the patient's number of seizures was reduced. This implies the binary response variate $I(Y_i < B_i)$.

- Write down an appropriate GLM for this response.

- Let $Y_i^* = I(Y_i < B_i)$

- Systematic component:

$$logit(\pi_i) = \beta_0 + \beta_1 A_i + \beta_2 B_i + \beta_3 Z_i$$

- Random component:

$$Y_i^* \sim Bernoulli(\pi_i)$$

◦ Set up an IRWLS algorithm.

- Use the same procedure in (n) with

$$\pi_{i,j} = \mu_{i,j} = \frac{exp(\eta_{i,j})}{1+exp(\eta_{i,j})}$$
$$v(\mu_{i,j}) = \pi_{i,j}(1 - \pi_{i,j})$$

◦ Code the IRWLS scheme using IML.

- See the SAS code.