

# Regression Diagnostics for GEE

Bahjat F. Qaqish  
(Bahjat\_Qaqish@unc.edu)

The University of North Carolina at Chapel Hill

# Outline

- Regression models for independent responses: Linear, logistic, loglinear
- Generalized Linear Models for independent responses
- Regression diagnostics for linear regression and GLMs for independent responses.
- Local versus global departures
- Approaches to model checking and diagnostics
- GEE
- Regression diagnostics for GEE
- Conclusion

## Warning

The term **cluster** here means a set of correlated observations.

It is completely different from the Kulldorf “cluster” which is a set of neighboring observations with unusually large values.

# Linear regression

- Response  $Y_i$ , covariates  $x_i$  on the  $i$ -th subject;  $i = 1, \dots, n$  independent subjects. Individual covariates  $x_{i1}, \dots, x_{ip}$ . Usually  $x_{i1} = 1$ , the intercept.
- Focus is on the expected value, the mean,  $E[Y_i] = \mu_i$ .

observed = expected + error

$$Y_i = \mu_i + (Y_i - \mu_i).$$

- Linear regression, the mean

$$E[Y_i] = \mu_i$$

equals the systematic component

$$\eta_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip}.$$

- The random error

$$\text{var}(Y_i - \mu_i) = \text{var}(Y_i) = \sigma^2$$

Constant variance. Normality assumed sometimes.

## Logistic regression (for 0/1 responses)

- Focus is on the expected value, the mean,  $E[Y_i] = P(Y_i = 1) = \mu_i$ .
- The logit of the mean

$$\log \frac{\mu_i}{1 - \mu_i} = \text{logit}(\mu_i)$$

equals the systematic component

$$\eta_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}.$$

- The random error

$$\text{var}(Y_i) = \mu_i(1 - \mu_i)$$

Variance changes with the mean (not constant).  
Bernoulli distribution.

## Loglinear models for counts

- Focus is on the expected value, the mean,  $E[Y_i] = \mu_i$ .
- The log of the mean

$$\log(\mu_i)$$

equals the systematic component

$$\eta_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}.$$

- The random random error

$$\text{var}(Y_i) = \mu_i$$

or more generally

$$\text{var}(Y_i) = \sigma^2 \mu_i$$

Variance proportional to the mean.

# Generalized Linear Models

- $Y_i$  is a random variable,

$$E[Y_i] = \mu_i,$$

$$\text{var}(Y_i) = \sigma^2 h(\mu_i).$$

Variance proportional to a known function of the mean.  $h()$  is the **variance function**.

- The systematic component:

$$\eta_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}.$$

- The **link function** links the random to the systematic components.

$$g(\mu_i) = \eta_i$$

# Fitting GLMs

Iteratively reweighted least squares (IRLS)

At the current estimate  $\hat{\beta}$ , compute

$$z_i = \hat{\eta}_i + (y_i - \mu_i) \frac{\partial \eta_i}{\partial \mu_i}$$

and

$$w_i^{-1} = V(\mu_i) \left( \frac{\partial \eta_i}{\partial \mu_i} \right)^2$$

Regress  $z$  on  $x$  with weight  $w$  to obtain a new  $\hat{\beta}$

Iterate to convergence

## Example: Correlated (Clustered) Responses

NC Early Cancer Detection Program

Health Maintenance Visit (HMV) in a given year,  
yes/no,

3900 patients (observations)

60 Medical practices (clusters)

20-200 patients per practice (cluster size), average  
65

Patient, physician and practice information

How does  $\Pr(\text{HMV})$  depend on patient, physician and practice characteristics? Logistic regression, GEE.

How does correlation within a practice depend on physician and practice characteristics?

Do large practices dominate the analysis?

## Notation for clustered responses

$K$  independent clusters

cluster: subject, family, pedigree, clinic

$i$  for cluster

$j$  and  $k$  for observations within a cluster

$n_i$  = cluster size.

Response vector  $Y_i = (Y_{i1}, \dots, Y_{in_i})^\top$ ,

$Y_{ij} \sim \text{Bernoulli}$ ,  $\text{pr}(Y_{ij} = 1) = E[Y_{ij}] = \mu_{ij}$

$\mu_{ijk} = E[Y_{ij}Y_{ik}] = \text{pr}(Y_{ij} = Y_{ik} = 1)$ .

Measures of dependence between  $Y_{ij}$  and  $Y_{ik}$ : Odds ratio, correlation, kappa, ...

I. Mean model:  $g_1(\mu_{ij}) = x_{ij}^\top \beta$ ,

II. Dependence model:  $g_2(\mu_{ij}, \mu_{ij}, \mu_{ijk}) = z_{ijk}^\top \alpha$ ,

## Analysis focus - mean vs correlation

- It is important at the outset to decide what the focus is.
- Focus on the mean structure: We would like to take correlation into account, although it is not the main focus of the analysis.
- Focus on the correlation structure: Careful modelling of correlation is required. Example: a study of the familial correlation of COPD. Risk of COPD (mean structure) is known to depend on smoking, age, etc. However, the main interest is modeling the sib-sib, sib-father and sib-mother correlations as a function of covariates.

## Three basic types of models

- Generalized Linear Models for independent responses have three basic extensions for dependent (correlated) responses:
  - Marginal (Population-Average) models
  - Conditional models
  - Random-effects (Subject-Specific, Mixed) models
- The names reflect the interpretation of the regression parameters  $\beta$ .
- Here we are concerned with populations, not samples.
- References:

Generalized Linear Models, P. McCullagh and J. A. Nelder. 2nd ed. Chapman & Hall, London, 1989.

Analysis of Longitudinal Data, Diggle, Heagerty, Liang & Zeger. 2nd ed. Oxford University Press, Oxford, 2002.

# Linear Regression

## Raw Materials of Model Checking

- Fitted values =  $\hat{\mu}$
- Residual variance  $s^2$
- Diagonals  $h$  of the projection matrix

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

- Standardized (const variance) residuals

$$\frac{y_i - \hat{\mu}_i}{\sqrt{1 - h_i}}$$

- Studentized standardized residuals

$$r'_i = \frac{y_i - \hat{\mu}_i}{s\sqrt{1 - h_i}}$$

$r_i'^2$  = the scaled reduction in residual SS after omitting observation  $i$

- Deletion residual

$$r_i^* = \frac{y_i - \hat{\mu}_i}{s_{(i)}\sqrt{1 - h_i}}$$

- The projection matrix for GLM's

$$H = W^{1/2} X (X^\top W X)^{-1} X^\top W^{1/2}$$

$W$  = the iterative weights

# 1. Systematic (Global) Departures

Two general approaches:

1. Embed the current model into a more general model, then use formal hypothesis tests.
2. Various residual plots from current model, smoothing and graphical methods.
  - Residuals: Standardized residuals against (VST transformed) fitted values. Look for curvature and change in spread with fitted value.
  - The variance function
  - The link function
  - Covariate scale (non-linearity), residual plots, generalized additive models

$$\eta = S_1(x_1) + S_2(x_2) + \cdots + S_p(x_p)$$

## 2. Isolated (Local) Departures

- Leverage:  $h_i$ , sum =  $p$ , average =  $p/n$ , 'large' leverage  $> 2p/n$  For GLMs, extreme  $x$  values may have low leverage.
- Consistency with the model: Look for large deletion residual. One-step approximations for GLMs.

- Influence:

$$\hat{\beta}_{(i)} - \hat{\beta}$$

- Influence: A weighted combination, Cook's distance

$$D_i = (\hat{\beta}_{(i)} - \hat{\beta})^\top (X^\top X) (\hat{\beta}_{(i)} - \hat{\beta}) / (ps^2)$$

$$D_i = r_i'^2 \frac{h_i}{p(1 - h_i)}$$

Influence  $\approx$  residual<sup>2</sup>  $\times$  leverage

For GLMs

$$D_i = (\hat{\beta}_{(i)} - \hat{\beta})^\top (X^\top W X) (\hat{\beta}_{(i)} - \hat{\beta}) / (p\hat{\phi})$$

and  $\hat{\beta}_{(i)}$  is the one-step approximation.

## Generalized Estimating Equations

- Model: Suppose  $Y_1, \dots, Y_K$  are independent vectors with means  $\mu_1, \dots, \mu_K$ , functions of  $\beta$ .

$$E[Y_{ij}] = \mu_{ij},$$

$$g(\mu_{ij}) = \eta_{ij} = x_{ij}^T \beta,$$

$$\text{Var}(Y_{ij}) = \phi h(\mu_{ij}),$$

$g$  is the link function,  $h$  the variance function,  $\eta_{ij}$  the linear predictor,  $\phi$  the scale parameter.

- Consider elementary estimating functions  $Y_i - \mu_i$ . Then the matrix  $\text{Var}(Y)$  is block-diagonal and the optimal linear combination is

$$\sum_{i=1}^K D_i^T (\text{Var}(Y_i))^{-1} (Y_i - \mu_i),$$

where  $D_i := \partial \mu_i / \partial \beta$ .

- The diagonals of  $\text{Var}(Y_i)$  are determined by

$$\text{Var}(Y_{ij}) = \phi h(\mu_{ij}).$$

The off-diagonals involve correlations that so far have not been defined. First write

$$\text{Var}(Y_i) = \Sigma_i = \phi A_i C_i A_i,$$

where  $A_i = \text{diag}(\sqrt{h(\mu_{ij})})$  and  $C_i = \text{Corr}(Y_i)$ .

Assumptions are made about the correlation matrix  $C_i$ . Specifically, it is parametrized by an  $s \times 1$  parameter vector  $\rho$ . An estimate of  $\rho$  is plugged-in and estimation proceeds. The assumed structure is called the **working correlation matrix**, denoted  $R_i$ , may not be identical to the true correlation, thus the different notation.

Define

$$V_i = A_i R_i A_i.$$

- The GEE is

$$\sum_{i=1}^K D_i^T V_i^{-1} (Y_i - \mu_i) = 0.$$

- Requirements:

A  $\sqrt{K}$ -consistent estimator of  $\phi$  at the true  $\beta$ .

A  $\sqrt{K}$ -consistent estimator of  $\rho$  at the true  $\beta$  and  $\phi$ .

Regularity conditions.

- The estimator  $\hat{\beta}$  is consistent and asymptotically Gaussian:

As  $K \rightarrow \infty$

$$\sqrt{K}(\hat{\beta}_K - \beta) \xrightarrow{d} N(0, S),$$

$$S := \lim_{K \rightarrow \infty} K H_1^{-1} H_2 H_1^{-1},$$

where

$$H_1 := \sum_{i=1}^K D_i^T V_i^{-1} D_i,$$

$$H_2 := \sum_{i=1}^K D_i^T V_i^{-1} \Sigma_i V_i^{-1} D_i.$$

- Variance estimation ( $S$ ):

The matrices  $H_1$  and  $H_2$  are evaluated at the estimates and  $\Sigma_i$  in  $H_2$  is replaced by

$$(Y_i - \mu_i)(Y_i - \mu_i)^T.$$

The estimator thus obtained is known as the **sandwich**, **robust**, or **empirical** variance estimator. This is because it is generally a consistent estimator of the true variance of  $\hat{\beta}$  even when the correlation is misspecified ( $R_i \neq C_i$ ).

- If the assumed correlation structure is correct, i.e. ( $R_i = C_i$ ), then  $H_1 = H_2$  and the asymptotic variance simplifies

$$S = \lim_{K \rightarrow \infty} K H_1^{-1}.$$

The estimator thus obtained is known as the **naive** or **model-based** variance estimator because it is generally an inconsistent estimator of the true variance of  $\hat{\beta}$  unless  $R_i = C_i$ .

- Some choices for  $R_i$ :

Exchangeable: For  $j \neq k$

$$\text{Corr}(Y_{ij}, Y_{ik}) = \rho.$$

Autoregressive: e.g. AR(1):

$$\text{Corr}(Y_{ij}, Y_{ik}) = \rho^{|j-k|}.$$

M-dependent: For  $|j - k| \leq m$

$$\text{Corr}(Y_{ij}, Y_{ik}) = \rho_{|j-k|}.$$

Unstructured:

$$\text{Corr}(Y_{ij}, Y_{ik}) = \rho_{jk}.$$

Independence: For  $j \neq k$

$$\text{Corr}(Y_{ij}, Y_{ik}) = 0.$$

(graphs in the paper)

# Summary

Fitting more general models vs residuals from current models

Global vs Local departures from the assumed model

Basic model-checking ingredients

One-step approximations for GLMs and GEEs

Correlation automatically adjusted for in GEE regression and diagnostics