Topics for today

- *When to use non-normal methods*
- *Interpreting effects in loglinear and logistic models*
- *Augmenting GzLMs to account for correlated responses*

Related reading: Sections 1 through 5 in the Non-normal notes, but with very little on Section 3 (Introduction to generalized linear models), which was covered in more detail by Gary Grunwald. In addition, Section 5 of 'Interpreting effects for loglinear and logistic models' is contained here.

1 Introduction

- Topic: models for outcome variables that are not normally distributed.
- Methods for modeling non-normal correlated data are also discussed in BIOS7712.
- We have learned how to use linear mixed models to fit clustered data with continuous and approximately normally distributed outcome variables. The models are versatile in handling random effects as well as repeated measures over time.
- For other types of outcome variables that involve clustered data, such as counts or binary outcomes, we can use generalized estimating equations (GEE) or employ generalized linear mixed models (GzLMM) as discussed in this chapter.

- Sometimes we can still employ normal theory models even when the outcome variable is non-continuous or non-normal. For example, a count outcome with a wide range of observed values not too close to zero may be well approximated by the normal curve.

- In other cases we might be able to apply a transformation to a non-normal outcome so that it is approximately normal.

  - In particular, outcome variables with right-skewed distributions are very common (e.g., cell counts), especially when there is a lower bound (typically zero).

  - A natural log transformation might make the distribution approximately normal.

  - But how then are effects from the model interpreted? (We shall soon see.)

2 *Determining when and when not to use normal theory methods*

- For a given outcome variable, a normal-theory model (e.g., GLM, LMM) may work adequately even if the observed data are not perfectly normal. Typically, the model fit will be fairly robust to violations of the normal assumption as long as the distribution is not too skewed or does not have a high percentage of data on one or more individual values, or when sample sizes are large enough that the central limit theorem comes into play.

- Consider the following examples of response variables and how you might model them.

    1. Y = FEV_1: slightly right skewed; true lower bound of 0 although P(Y=0)=0 or negligible for a non-error blow.

    2. Y = forced exhaled nitric oxide (FeNO): moderately right-skewed; lower bound of 0 but P(Y=0)=0 or negligible.

    3. Y = expenditures for health clinics; can be considered continuous; right skewed but with P(Y=0)>0 and possible that P(Y=0)>>0 (e.g., 20% or more).

    4. Y = whether child had an asthma exacerbation in a given week (y/n).

    5. Y = percentage of patients that adhere to doctor's directions, based on large n.

    6. Y = number of times albuterol was used in a day by a child to treat asthma. Counts of use typically range from 0 to 6, but most commonly are 0, 1 or 2.

3 Generalized linear models (GzLM)

- Covered by Gary

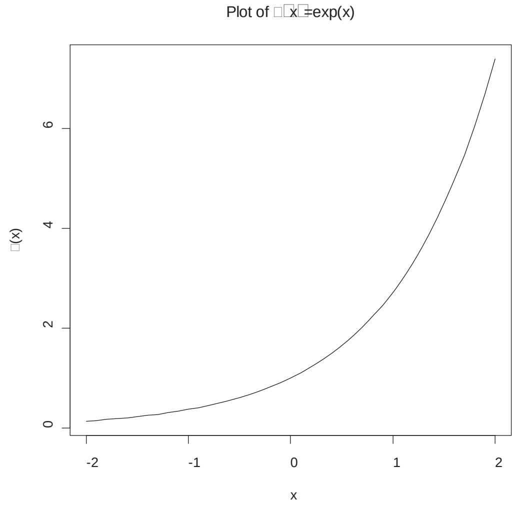
- Recall that for the Poisson, the *canonical link* or *natural link* is in fact the natural log link.

- Count outcomes can often be modeled using a Poisson distribution, although it is often necessary to add a dispersion parameter into the model (also discussed later).

- Regression of a Poisson variable on one or more predictors is often referred to as *Poisson regression*. For the Poisson, using the natural log link leads to

$$\mu_i = g^{-1}(\mathbf{x}_i^r \boldsymbol{\beta}) = e^{\mathbf{x}_i^r \boldsymbol{\beta}} .$$

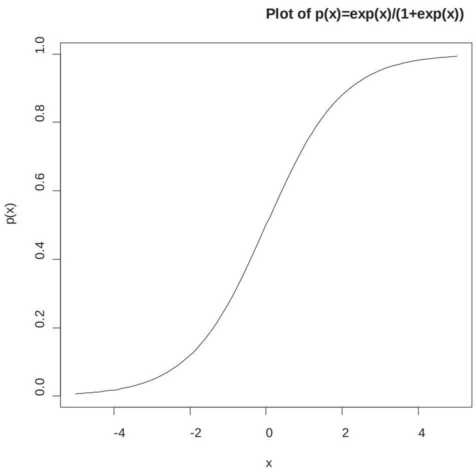
Here is a plot of this function with one covariate and β_0=0 and β_1=1.



- For the binomial, consider $\hat{p}_i = Y_i / n_i$, where $Y_i \sim Bin(n_i, p_i)$. Of course *p* must be bound between 0 and 1 and intuitively would be a continuous function of $\mathbf{x}_i^r$. In order to maintain these characteristics, one possibility is to set

$$p_i = \frac{e^{\mathbf{x}_i^r \boldsymbol{\beta}}}{1 + e^{\mathbf{x}_i^r \boldsymbol{\beta}}} . \qquad (3)$$

This corresponds to the logit link. (Can you show?) Function (3) is plotted to the right, for one continuous covariate, for β_0 = 0 and β_1=1; this demonstrates the signature 'slanted S' shape. However, for certain applications, this may not be apparent since the range of *x* values may only cover a portion of the 'S'.



From Section 5 of 'Interpreting effects for loglinear and logistic models'

- There are two common situations where loglinear models are used: (i) when the outcome variable is log transformed to be approximately normal, and (ii) when a log link is used for a count variable (e.g., Poisson regression). For such models, we can easily derive multiplicative effects.

- Let's first consider case (i); for simplicity, consider the simple model

$$ln(Y_{ij}) = \beta_0 + \beta_1 x_{ij} + \varepsilon_{ij} ,$$

for subject *i* and time *j*.

- Then

$$Y_{ij} = e^{\beta_0 + \beta_1 x_{ij} + \varepsilon_{ij}} ,$$

which implies that

$$\begin{aligned} E(Y_{ij} | X_{ij} = x_{ij}) &= E(e^{\beta_0 + \beta_1 x_{ij} + \varepsilon_{ij}}) \\ &= E(e^{\beta_0 + \varepsilon_{ij}} \mathrm{e}^{\beta_1 x_{ij}}) \\ &= e^{\beta_1 x_{ij}} E(e^{\beta_0 + \varepsilon_{ij}}) \\ &= e^{\beta_1 x_{ij}} c \end{aligned}$$

and

$$E(Y_{ij} \mid X_{ij} = x_{ij} + 1) = e^{\beta_1 (x_{ij}+1)} c ,$$

where *c* is a constant. From these, $E(Y \mid x+1) / E(Y \mid x) = \mathrm{e}^{\beta_1}$. In words, the multiplicative increase in the mean of *Y* for a 1-unit increase in x is e^{β_1}. Or the relative increase in the mean of *Y* for a 1-unit increase in *x* is $100(e^{\beta_1} - 1)\%$.

- Case (ii) (log link in generalized linear models) differs from (i) in that the natural log is taken on $E(Y)$ and not Y itself. Consider (1), with one simple linear predictor:

  $$ln(\mu_i) = \beta_0 + \beta_1 x_i$$

  Exponentiating both sides yields

  $$\mu_i = e^{\beta_0 + \beta_1 x_i}$$

  Thus, β_1 also has a multiplicative effect interpretation, as before. The log is the natural link for Poisson outcomes and consequently beta parameters associated with predictors in Poisson regression have relative increase interpretations when the natural link is used.

- Some things differ between cases (i) (logged outcome for a normal theory model) and (ii) (log link for a count outcome) above.
  - In (i) we model the mean of the logged Y values, so inverting back to original units will yield the geometric mean (usually closer to the median than the arithmetic mean for right skewed distributions). In (ii), we model the log of the mean (where ‘mean’ is the arithmetic mean).
  - For (i), if we assume that normal theory model assumptions are met after log transformation (including constant variance), then equal variances on that scale will imply that variances on the original scale will increase as the mean of Y increases. This is why the log transformation is sometimes used to ‘stabilize’ variance. In (ii), we assume constant variance on the regular scale [i.e., for Y, not log(Y)]

- For logit link models, odds ratios are derived readily. To illustrate this using the simple linear logit model:

  $$log\left[\left(\pi\left(x\right)/(1-\pi\left(x\right)\right)\right]=\beta_0+\beta_1 x, \qquad \text{where } \pi\left(x\right)=P(Y=1|X=x).$$

  Exponentiating both sides yields

  $$\pi\left(x\right)/\left(1-\pi(x)\right)=e^{\beta_0+\beta_1 x}. \qquad (4)$$

  But note that

  $$\pi\left(x+1\right)/\left(1-\pi(x+1)\right)=e^{\beta_0+\beta_1(x+1)}. \qquad (5)$$

  Thus, if we divide (5) by (4), we yield an odds ratio (a ratio of ratios):

  $$\left[\pi\left(x+1\right)/\left(1-\pi(x+1)\right)\right]/\left[\pi\left(x\right)/\left(1-\pi(x)\right)\right]=e^{\beta_1}.$$

  In words, for a 1-unit increase in *x*, the odds of an event (associated with *Y*=1) increases e^{β_1} times.

4 Augmenting GzLMs to account for correlated responses

4.1 Generalized estimating equations (GEE)

- One option in modeling correlated non-normal outcome data is to use generalized estimating equations (GEE) that can be applied to GzLMs for longitudinal data. To begin, initial estimates are obtained assuming data are independent, using the usual GzLM methodology. GEE are then applied iteratively to obtain estimates of interest accounting for the correlated data.

- GEE does not work with a true likelihood and thus does not actually fit a true covariance matrix. Rather, it uses what is called a working covariance matrix. But the forms of the working structures that can be used are ones we are familiar with (e.g., AR(1), exchangeable – i.e., CS).

- The specific steps for GEE are as follows. This is just a sketch; for more detail see the SAS Help Documentation or other references listed at the end of this subsection.

  1. Use standard GzLM theory to obtain initial estimates of $\boldsymbol{\beta}$.
  2. Compute working correlations based on standardized residuals, the current estimate of $\boldsymbol{\beta}$ and the assumed covariance structure.
  3. Compute an estimate of $\mathbf{V}_i = \text{Var}(\mathbf{Y}_i)$.
  4. Update the estimate of $\boldsymbol{\beta}$ using the new estimate of $\mathbf{V}_i$.
  5. Repeat steps 2-4 until convergence.

- GEE is considered a general type of quasi-likelihood estimation (QLE) since it is an estimation method that is not built on maximum likelihood principles and only requires the form of the mean, the variance as a function of the mean, correlation parameters (via the ‘working’ correlation matrix), and scale parameter (see Liang and Zeger, 1986).

- After the GEE process is complete, model-based and empirical forms of $\text{Var}(\hat{\boldsymbol{\beta}})$ can be obtained in order to conduct tests involving $\boldsymbol{\beta}$. The forms of these variances (e.g., see Hedeker, p. 137-38) are analogous to the model-based and empirical forms of variances of beta estimates in mixed models.

- The default in SAS is to use the empirical estimates (sometimes also called robust, or sandwich estimators). These estimators have the advantage that they are robust to miss-specifications of the (working) covariance structure.

- However, for smaller sample sizes the use of residuals often leads to an underestimated standard error; the smaller the sample size, the worse the underestimation. (A recent student, Yu Zhang, examined this issue for count outcomes and successfully defended his Master’s Thesis on this topic.)

- In order to obtain the model-based variance estimators, include MODELSE as an option in the REPEATED statement. If this is done, there are various ways to adjust the variance estimates by scale parameter estimates, some of which are listed below.

  - Adjust the variance estimates by a scale parameter by including ϕ as a scalar in Var($\mathbf{Y}_{\mathrm{i}}$) within the GEE estimation process. This can be achieved by not including a SCALE or NOSCALE option in the MODEL statement. A standardized Pearson statistic is used to estimate ϕ.

  - Fix the scale parameter at 1 in the GEE estimation process but then adjust variance estimates by a factor of $\sqrt{\phi}$, using a Pearson or deviance statistic to estimate ϕ. Include the PSCALE option to use the Pearson statistic or the DSCALE option to use the deviance statistic in the MODEL statement.

  - Do not adjust variance estimates by a scale parameter. This can be achieved by including the NOSCALE option in the MODEL statement.

- Note that for GzLM fits, the inclusion of the scale parameter in the GEE process [a scalar in the equation of $Var(Y_i)$] will affect $Var(\hat{\boldsymbol{\beta}})$, but not $\hat{\boldsymbol{\beta}}$ itself, as is the case for standard GzLMs.

- For GzLMs, the theory is developed for independent responses from subjects (i.e., cross-sectional data). GEE is then an extension for clustered data (e.g., longitudinal data). Thus, the notation of GzLMs can be modified to account for this. Specifically, we can use

$$\eta_{ij} = \mathbf{x}_{ij}^{r}\boldsymbol{\beta}$$

  to denote linear predictor for subject i at time j; $\mathbf{x}_{ij}^{r}$ is a row vector with elements x_{vij}, where v denotes the covariable (formerly denoted by j; j is now used to index time). The response for subject i at time j is then denoted as Y_{ij}. Other quantities can be generalized similarly (e.g., see Hedeker, 2006).

References:

Liang K-YL, Zeger S. (1986) Longitudinal Data Analysis Using Generalized Linear Models, Biometrika 73(1): 13-22. [Original article on GEE.]

Hedeker D., Gibbons RD. (2006) Longitudinal Data Analysis, Wiley, NJ, Chapter 8: Generalized Estimating Equations (GEE) Models.

SAS Help Documentation: SAS/STAT, The GENMOD Procedure, Details, Generalized Estimating Equations, v. 9.1 and 9.2, Cary, NC.

4.2 Application of GEE with a count outcome

- Count outcome variables can often be fit with a Poisson distribution, perhaps with the addition of a scale parameter, if necessary, if there is over- or under-dispersion. The following count outcome example is from my work, illustrating a significant association between daily doser medication use and air pollution. The data are fit using GzLM/GEE. These data tend to be underdispersed relative to the Poisson distribution (i.e., the variance tends to be less than the mean).

Code:

GEE is invoked when the REPEATED statement is included in the PROC GENMOD code.

```
PROC GENMOD DATA = y4dir.y4data;
  CLASS id friday;
  MODEL dnew_spuff = temp pressure humidity friday date mmaxpm25
     / noscale DIST = poisson;
  REPEATED SUBJecT = id / TYPE = AR(1) modelse; RUN;
```

Condensed output:

The GENMOD Procedure Model Information Data Set Y4DIR.Y4DATA Distribution Poisson Link Function Log Dependent Variable dnew_spuff	Number of Observations Used 5928 Class Level Information Class Levels id 57	Algorithm converged. GEE Fit Criteria QIC 10917.8584 QICu 10869.9274

Output using empirical SE's (default with GEE in SAS). Empirical SE's will be more robust to model misspecifications.

```
                     Analysis Of GEE Parameter Estimates
                      Empirical Standard Error Estimates

                         Standard   95% Confidence
Parameter   Estimate     Error        Limits            Z Pr > |Z|
Intercept     -6.9736    3.4398 -13.7156   -0.2317   -2.03    0.0426
temp          -0.0060    0.0014  -0.0086   -0.0033   -4.44    <.0001
pressure       0.0010    0.0019  -0.0028    0.0048    0.50    0.6154
humidity      -0.0036    0.0006  -0.0049   -0.0024   -5.85    <.0001
friday    0    1.2139    0.0810   1.0551    1.3726   14.99    <.0001
friday    1    0.0000    0.0000   0.0000    0.0000     .        .
date           0.0004    0.0002  -0.0000    0.0008    1.80    0.0720
mmaxpm25       0.0019    0.0007   0.0005    0.0033    2.63    0.0086
```

Output using model-based SE's. Note that there is no scale parameter in the model (i.e., set to 1), so underdispersion is not accounted for properly.

```
                     Analysis Of GEE Parameter Estimates
                     Model-Based Standard Error Estimates

                         Standard   95% Confidence
Parameter   Estimate     Error        Limits            Z Pr > |Z|
Intercept     -6.9736    5.3021 -17.3655    3.4183   -1.32    0.1884
temp          -0.0060    0.0018  -0.0095   -0.0025   -3.40    0.0007
pressure       0.0010    0.0031  -0.0050    0.0070    0.32    0.7489
humidity      -0.0036    0.0010  -0.0056   -0.0017   -3.68    0.0002
friday    0    1.2139    0.0440   1.1276    1.3001   27.59    <.0001
friday    1    0.0000    0.0000   0.0000    0.0000     .        .
date           0.0004    0.0003  -0.0002    0.0009    1.27    0.2045
mmaxpm25       0.0019    0.0012  -0.0005    0.0043    1.57    0.1176
Scale          1.0000     .         .         .        .        .

NOTE: The scale parameter was held fixed.
```

Can you interpret the parameter estimate for mmaxpm25? (Remember, by default the GzLM for the Poisson distribution uses the log link.)

If you include the PSCALE or DSCALE options in the MODEL statement, the model-based SE's will be a bit closer to the SE's using empirical methods:

Abbreviated output when including PSCALE to the right of '/' in the MODEL statement:

Output using model-based SE's for model with scale parameter (i.e., quasi-likelihood). Note that the beta estimates are still the same, but the SE's are smaller (compared with model-based SE's with the scale parameter) since they account for underdispersion.

```
                Analysis Of GEE Parameter Estimates
                Model-Based Standard Error Estimates

                       Standard   95% Confidence
Parameter    Estimate     Error        Limits            Z Pr > |Z|
Intercept     -6.9736    4.5340 -15.8601    1.9129   -1.54   0.1240
temp          -0.0060    0.0015  -0.0090   -0.0030   -3.98   <.0001
pressure       0.0010    0.0026  -0.0041    0.0061    0.37   0.7082
humidity      -0.0036    0.0008  -0.0053   -0.0020   -4.31   <.0001
friday    0    1.2139    0.0376   1.1401    1.2876   32.26   <.0001
friday    1    0.0000    0.0000   0.0000    0.0000     .       .
date           0.0004    0.0002  -0.0001    0.0009    1.48   0.1379
mmaxpm25       0.0019    0.0010  -0.0001    0.0039    1.83   0.0672
Scale          0.8551     .         .         .        .       .

NOTE: The scale parameter was held fixed.
```

Scale in this table is $\sqrt{\hat{\varphi}}$ using previous notation. E.g., SE's here are 0.8551 times the SE's from the previous table that used no scale parameter. Thus, the approach here adjusts for underdispersion.

The 'Scale' estimate is $\sqrt{\hat{\varphi}}$, and the SE's here are equivalent to those from the model-based SE's when NOSCALE is used, times $\sqrt{\hat{\varphi}}$. In this case the SE's are still larger than when using the empirical approach, despite the scale adjustment. When including DSCALE, the square root of phi is 0.8956, so that the adjustment to the original model-based SE's are even less.

The following is the partial output obtained when not including any SCALE or NOSCALE option in the MODEL statement. Here, the phi parameter is involved in the GEE estimation, but note that the results are not too different than the previous one in which the variance estimates were only adjusted after the GEE estimation.

```
PROC GENMOD DATA = y4dir.y4data;
  CLASS id friday;
  MODEL dnew_spuff = temp pressure humidity friday date mmaxpm25
     / DIST=poisson;
  REPEATED SUBJecT = id / TYPE = AR(1) modelse; RUN;
```

```
                Analysis Of GEE Parameter Estimates
                Model-Based Standard Error Estimates

                       Standard   95% Confidence
Parameter    Estimate     Error        Limits            Z Pr > |Z|
Intercept     -6.9736    4.5468 -15.8853    1.9380   -1.53   0.1251
temp          -0.0060    0.0015  -0.0090   -0.0030   -3.96   <.0001
pressure       0.0010    0.0026  -0.0042    0.0061    0.37   0.7090
humidity      -0.0036    0.0008  -0.0053   -0.0020   -4.29   <.0001
friday    0    1.2139    0.0377   1.1399    1.2878   32.17   <.0001
friday    1    0.0000    0.0000   0.0000    0.0000     .       .
date           0.0004    0.0003  -0.0001    0.0009    1.48   0.1390
mmaxpm25       0.0019    0.0010  -0.0001    0.0039    1.82   0.0680
Scale          0.8576     .         .         .        .       .

NOTE: The scale parameter for GEE estimation was computed as the square root of the normalized
      Pearson's chi-square.
```