A. INTRODUCTION

A general linear model says

\[ Y_i = X_i \alpha + \epsilon_i \]
\[ \epsilon_i \sim F(0, \Sigma_i) \]

which implies that \( Y_i \sim F(X_i \alpha, \Sigma_i) \).

What if it is NOT reasonable to assume that the expected outcome \( E[Y_i] \) is a linear function \( \alpha_0 + \alpha_1 X_1 + \ldots + \alpha_p X_p \) of the parameters of interest?

One possibility may be to transform the data, e.g., log, square - root, arc sine square - root.

- Often used for counts
- Sometimes used for proportions

Then we can use a linear model for the transformed outcome:

\[ f(Y_i) = X_i \alpha + \epsilon_i \]
\[ \epsilon_i \sim F(0, \Sigma_i) \]

and we can carry out all model fitting and testing procedures as before.

Note:
- Interpretation may be more difficult (e.g., arc sine)
- Variances/correlations are now among the transformed variables

The alternative is to change the model itself to a Generalized Linear Model (GLM) where we transform the mean of the outcome, instead of transforming the raw outcome data.
**EXAMPLE: BINARY DATA**

- Mean: if $Y_i$ is binary, then $E[Y_i]$ must take on values between 0 and 1 only. A function which satisfies this restriction is:
  
  $$E[Y_i] = f(X_i \alpha) = \frac{e^{X_i \alpha}}{1 + e^{X_i \alpha}}$$

  which when inverted yields:
  
  $$\log \left( \frac{E[Y_i]}{1 - E[Y_i]} \right) = g(E[Y_i]) = X_i \alpha$$

  $g(\cdot)$ is called the logit link. 

- Probability distribution: $Y_i \sim \text{Bernoulli}(\rho_i)$

- Variance: The Bernoulli distribution dictates that
  
  $$\text{Var}[Y_i] = v(E[Y_i]) = E[Y_i] (1 - E[Y_i])$$

  where $v(\cdot)$ is the variance function and is defined to be $v(E[Y_i]) = E[Y_i] (1 - E[Y_i])$ for a Bernoulli r.v.

**EXAMPLE: COUNT DATA**

- Mean: if $Y_i$ is a count, then $E[Y_i]$ can only take on non-negative values. A function which satisfies this restriction is:
  
  $$E[Y_i] = f(X_i \alpha) = e^{X_i \alpha}$$

  which when inverted yields:
  
  $$\ln(E[Y_i]) = g(E[Y_i]) = X_i \alpha$$

  $g(\cdot)$ is called the log link.

- A possible probability distribution: $Y_i \sim \text{Poisson}$

- Variance: A Poisson distribution dictates that
  
  $$\text{Var}[Y_i] = E[Y_i] = v(E[Y_i])$$

  so we now have variance function $v(E[Y_i]) = E[Y_i]$.

What’s the common thread?

- choose a distribution according to the form of the data
- choose a link which you think is reasonable according to the science and dictating how the covariates are related to the mean
- the form of the variance is then dictated by the distribution...
- ...which then dictates the form of the variance function $v(E[Y_i])$.

The links we’ve discussed here are called the **natural** or **canonical** links for these distributions. Others are possible; see the SAS PROC GENMOD documentation. Or if you are an R user, the `family` argument to the ‘glm()’ [or ‘gee()’] functions.

**NOTATION**

- $Y_i =$ outcome variable for subject $i$
- $X_i =$ vector of covariate values for subject $i$
- $Y_i \text{ iid} \sim \text{distribution e.g., Bernoulli (binomial), Poisson, Gamma, Normal}$
- $E[Y_i] = f(X_i \alpha)$ for a monotone function $f$
- $g(E[Y_i]) = g(f(X_i \alpha)) = X_i \alpha$ (since $g = f^{-1}$)
- $\text{Var}[Y_i] = \phi v(E[Y_i])$ for some function $v$

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

Where is the measurement error in these models?
\( v(E[Y_i]) \) is called the variance function and it dictates how the variance of \( Y_i \) is related to the mean of \( Y_i \).

\( \phi \) is called the dispersion parameter and it allows us to "inflate" the variance when needed. \( \phi \) is usually dictated by the distribution chosen for \( Y_i \).

By default, \( \phi = 1 \) for Bernoulli and Poisson models.

For lots of details on Generalized Linear Models, see McCullough & Nelder (1989) *Generalized Linear Models* Chapman & Hall.

---

**ESTIMATION**

Recall how GLMs with independent data are estimated: suppose we motivate via MLE of a normal distribution, which dictates the form of the likelihood:

\[
\ell = \prod_{i=1}^{n} \left( 2\pi \sigma^2 \right)^{-1/2} e^{-\frac{1}{2\sigma^2} (Y_i - X_i \alpha)^2} \\
= \left( 2\pi \sigma^2 \right)^{-\frac{n}{2}} e^{-\frac{1}{2\sigma^2} (Y - X\alpha)'(Y - X\alpha)}
\]

This is then maximized. How? With calculus:

- minimize \(-ln \ell\)
  
  \( \implies \) minimize \( \sum_{i=1}^{n} \frac{1}{\sigma^2} (Y_i - X_i \alpha)^2 \)

- take the derivative of \( \sum_{i=1}^{n} \frac{1}{\sigma^2} (Y_i - X_i \alpha)^2 \) and set it equal to 0:
  
  \[-2 \sum_{i=1}^{n} \frac{1}{\sigma^2} (Y_i - X_i \alpha) X_i = 0\]

- solve for \( \alpha \).

Can we use the same idea here? Yes, for distributions which belong to an exponential family, this equation from our normal model:

\[
\sum_{i=1}^{n} \frac{1}{\sigma^2} (Y_i - X_i \alpha) X_i = 0
\]

translates into:

\[
\sum_{i=1}^{n} \frac{1}{v(E[Y_i])} (Y_i - E[Y_i]) \frac{d}{d\alpha} E[Y_i] = 0
\]

or equivalently:

\[
\sum_{i=1}^{n} \frac{1}{v(f(X_i \alpha))} (Y_i - f(X_i \alpha)) f'(X_i \alpha) X_i = 0
\]

\( \uparrow \) inverse variance of \( Y_i \) from its mean
\( \uparrow \) deviation of \( Y_i \) from its mean
\( \uparrow \) derivative of the mean

for, e.g., Bernoulli or Poisson models. See SAS PROC GENMOD documentation for the formulas for the model likelihoods.
Things to note:
- Each deviation is weighted by the inverse of the variance, just like weighted least squares. (Which is also true for \( \sum_{i=1}^{n} \frac{1}{\sigma_i^2} (Y_i - X_i \alpha) X_i = 0 \), but since the variance is constant it cancels out of the equation.)
- The estimation of \( \alpha \) and of the variance function are now inseparable, because \( \nu (E[Y_i]) \) is a function of \( \alpha \) (whereas \( \sigma^2 \) is not).
- The equation to be solved

\[
\sum_{i=1}^{n} \frac{1}{\sigma_i^2} (Y_i - X_i \alpha) X_i = 0
\]

is not at all linear in \( \alpha \), whereas

\[
\sum_{i=1}^{n} \frac{1}{\sigma_i^2} (Y_i - X_i \alpha) X_i = 0
\]

is linear in \( \alpha \).

Because of all this, there is no longer an explicit equation for \( \hat{\alpha} \) like 
\( \hat{\alpha} = (X' \Sigma^{-1} X)^{-1} X' \Sigma^{-1} Y \), instead, \( \hat{\alpha} \) for must be found iteratively.

**ESTIMATION ALGORITHM:**

1. Guess at a value for \( \alpha \), call it \( \alpha^{(0)} \).
   
   (0) \begin{align*}
   \text{Using } \alpha^{(0)} \text{ calculate } & \frac{1}{\nu(f(X_i \alpha^{(0)}))} \\
   \text{Pretend the computation in (1) gives a known (rather than estimated) quantity and solve (*) for } \alpha^{(1)} \text{ using iteratively reweighted least squares or Newton-Raphson.} & \end{align*}

2. Solve (*) for \( \alpha^{(2)} \) etc.

   Thus we iterate between \( \nu(\cdot) \) and \( \alpha(\cdot) \) until \( \alpha^{(k)} \) is close to \( \alpha^{(k-1)} \) or \( \frac{\| \alpha^{(k)} - \alpha^{(k-1)} \|}{\| \alpha^{(k-1)} \|} < 1 \times 10^{-8} \).

3. **FITTED VALUES**

   There are two different quantities which could be predicted:
   - predicted mean \( E[Y_i] = f(X_i \hat{\alpha}) \)
     
     Example Bernoulli \( E[Y_i] = \frac{e^{X_i \hat{\alpha}}}{1 + e^{X_i \hat{\alpha}}} \equiv \hat{p}_i \)
   - predicted linear function \( g(E[Y_i]) = X_i \hat{\alpha} \)
     
     Example Bernoulli \( g(E[Y_i]) = X_i \hat{\alpha} \)

   There are multiple residual quantities as well, but they are all on the scale of the mean, not the linear function \( X_i \alpha \). (We’ll come back to these.)
GOODNESS OF FIT

The scaled deviance is defined as twice the difference between the maximum achievable log likelihood and the current fitted log likelihood at \( \hat{\alpha} \) assuming \( \phi \) is known:

\[
D^\ast (Y, E[Y]) = \frac{D(Y, E[Y])}{\phi} \quad \text{deviance}
\]

- Normal \( D = \sum_{i=1}^{n} (Y_i - E[Y_i]) \)
- Bernoulli \( D = \sum_{i=1}^{n} \left[ Y_i \log \left( \frac{Y_i}{E[Y_i]} \right) + (1 - Y_i) \log \left( \frac{1 - Y_i}{1 - E[Y_i]} \right) \right] \)
- Binomial \( D = 2 \sum_{i=1}^{n} \left[ Y_i \log \left( \frac{Y_i}{E[Y_i]} \right) - (Y_i - E[Y_i]) \right] \)
- Poisson \( D = 2 \sum_{i=1}^{n} \left( Y_i \log \left( \frac{Y_i}{E[Y_i]} \right) - (Y_i - E[Y_i]) \right) \)

The GENMOD documentation describes how the scaled deviance is computed when \( \phi \) is unknown.

GENMOD also computes the Pearson chi-square statistic:

\[
T(Y, E[Y]) = \sum_{j=1}^{m} \frac{Y_j - E(Y_j)}{\sqrt{v(f(X_j \hat{\alpha}))}}
\]

(average standardized residual) and the scaled Pearson chi-square statistic:

\[
T^\ast (Y, E[Y]) = \frac{T(Y, E[Y])}{\phi}
\]

A goodness-of-fit test can be carried out as either:

- if \( D^\ast > \chi^2_{n-p, 1-a} \) then reject \( H_0 \): model fit is adequate
- if \( T^\ast > \chi^2_{n-p, 1-a} \) then reject \( H_0 \): model fit is adequate

where \( p \) is the number of parameters in \( \alpha \).

Since the expected value of a \( \chi^2_{n-p} \) is equal to \( n - p \), an approximate guideline for ‘accepting’ \( H_0 \) is to see if \( D^\ast \) or \( T^\ast \) is approximately \( n - p \), or to see if \( \frac{D^\ast}{n-p} \) or \( \frac{T^\ast}{n-p} \) is approximately 1.

RESIDUALS

- Raw residual \( r_i = Y_i - E[Y_i] \)
- Pearson residual \( r_i^P = \frac{Y_i - E[Y_i]}{\sqrt{v(f(X_i \hat{\alpha}))}} \)
- standardized Pearson residual \( r_i^{SP} = \frac{r_i^P - \text{mean}(r_i^P)}{\text{var}(r_i^P)} \)

- deviance residual \( r_i^D = \text{sign}(r_i) \sqrt{d_i} \)
  where \( d_i = \text{contribution of the } j\text{th observation to the deviance} \)
- standardized deviance residual \( r_i^{SD} = \frac{r_i^D - \text{mean}(r_i^D)}{\text{var}(r_i^D)} \)
TESTING
Since there is no explicit formula for \( \hat{\alpha} \), there is no explicit distribution for \( \hat{\alpha} \). Approximately (for large \( n \)): \[
\hat{\alpha} \sim N\left(\alpha, \phi(\Delta'V^{-1}\Delta)^{-1}\right)
\]
where \( \Delta \) is the matrix of first derivatives of the vector \( f(X\alpha) \) with respect to \( \alpha \):
\[
\Delta_{jk} = \frac{\delta f(X_i\alpha)}{\delta \alpha_k} \quad \forall i \in \{1, \ldots, n\}
\]
\( V = \text{diag}\{v(f(X_i\alpha)), \quad i = 1, \ldots, n\} \)

A Wald chi-square test for \( \alpha \) is:
\[
H_0 : L\hat{\alpha} = 0
\]
\[
T^* = (L\hat{\alpha} - 0)' \left( L\Var(\hat{\alpha})L' \right)^{-1} (L\hat{\alpha} - 0)
\]
which is the same as the square of a Z-test when \( L \) has one row. Compare \( T^* \) to a \( \chi^2_{\text{rank}(L), 1-a} \).

Likelihood ratio tests can also be used to compare nested models:
\[
T^* = \left( -2 \text{ log likelihood of Reduced model} \right) - \left( -2 \text{ log likelihood of Full model} \right)
\]
The degrees of freedom are computed as
\[
r = (\# \text{ parameters in full model's } \alpha) - (\# \text{ parameters in reduced model's } \alpha)
\]
Reject \( H_0 \): reduced model is adequate if
\[
T^* > \chi^2_{r, 1-a} \quad \text{for } \ e.g., \ a = 0.05.
\]
There are no formal tests for the variance function, since the variance function is dictated by the distribution being used (e.g., Poisson) and the estimated \( \alpha \).
Note that we are again ignoring the variability from estimating \( V \).