Linear and Generalized Linear Models
Lectures Notes (STAT 244, Fall 2014)

Won I. Lee

1 Introduction

1.1 GLM Components

Three components of a GLM The 3 components are:

1. Random component: distribution of y_i, i.i.d.
   - Response variable y has exponential dispersion family
   - $\sum_i y_i$ is sufficient statistic
2. Linear predictor: $\eta = \mathbf{X}\beta$ with $n \times p$ model matrix $\mathbf{X}$ and parameters β
   - x_{ij} is value of explanatory variable x_j for observation i
   - $\mathbf{x_i} = (x_{i1}, \dots, x_{ip})$
   - $\eta_i = \sum_j \beta_j x_{ij}$
   - $\mathbf{X} = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix}$
3. Link function: g linking mean to linear predictor; $g[E(\mathbf{y})] = \eta = \mathbf{X}\beta$
   - $g(\mu_i) = \sum_j \beta_j x_{ij}$
   - Canonical link: g s.t. transform μ_i to natural parameter θ_i; then we have concave log-likelihood, simple likelihood equations, Fisher scoring = Newton-Raphson, etc.
   - Binary response: logit ($\theta_i = \text{logit}(\mu_i) = \text{logit}(\pi_i)$)
   - Count response: log ($\theta_i = \log(\mu_i)$)
   - Continuous response: identity ($\theta_i = \mu_i$)

Why GLMs? We can transform data instead. But this requires a transformation that yields simultaneously: 1) approximate normality; 2) homoscedasticity. This often conflicts with each other.

For GLMs, two separate choices/degrees of freedom: 1) choice of link function; 2) choice of random component. Gives freedom to model and fit data well without having to worry about normality or homoscedasticity.

Finally, GLM models $g[E(y_i)]$, so we can say that $E(y_i) = g^{-1}(\mathbf{x_i}\beta)$, i.e. we have direct interpretability of parameters.

1.2 Quantitative vs. Qualitative Variables

Types of Explanatory Variables In linear predictors, they can be:

- Quantitative: simple linear regression; single term $\beta_j x_j$ and single column in $\mathbf{X}$
- Qualitative: ANOVA, odds ratios (binary); if c categories, require $c-1$ terms (indicators) in linear predictor and $c-1$ columns in $\mathbf{X}$ (i.e. one is baseline)
- Mized: i.e. interaction of quantitative $\times$ qualitative; ANCOVA (analysis of covariance due to interaction term)
- Ordinal: ordered categorical variables can be treated as either quantitative or qualitative

1.3 Model Matrices and Vector Spaces

Matrices Induce Vector Spaces Consider all possible $\eta = \mathbf{X}\beta$ for all possible β. This is:

$$\eta = \beta_1 \mathbf{X}_1 + \cdots + \beta_p \mathbf{X}_p$$

i.e. a linear combination of the *columns* of $\mathbf{X}$. Thus, η lives in the **column space** of $\mathbf{X}$:

$$C(\mathbf{X}) = \{\eta : \eta = \mathbf{X}\beta\} = \{\mathbf{X}\beta : \beta \in \mathbb{R}^p\}$$

This is called the *model space* of the GLM. Properties:

- Models with matrices $\mathbf{X}_a, \mathbf{X}_b$ are equivalent if $C(\mathbf{X}_a) = C(\mathbf{X}_b)$
- If model a is nested in model b, then $C(\mathbf{X}_a) \subset C(\mathbf{X}_b)$

Dimension of $C(\mathbf{X})$ Rank of the model matrix $\mathbf{X}$ is equal to number of linearly independent columns, so:

$$\dim(C(\mathbf{X})) = \text{rank}(\mathbf{X}) \leq p$$

If equal p, then $\mathbf{X}$ has full rank. If not full rank, then $\dim(N(\mathbf{X})) > 0$; i.e. model matrix has redundancies, or aliasing.

- Extrinsic: When variable (usually quantitative) just happens to be linear combination of the others (collinearity)
- Intrinsic: Inherent redundancy in matrix, i.e. when one-way ANOVA has both intercept term (all 1) and all indicators (no baseline)

One-Way ANOVA Used for comparing means across different groups/categories, each group labeled by an indicator I_i. Suppose c groups, $i = 1, \ldots, c$, and $j = 1, \ldots, n_i$ observations in each group.

$$g[E(y_{ij})] = \beta_0 + \beta_i = \beta_0 + \beta_1 I_{i1} + \cdots + \beta_c I_{ic}$$

Significance test of null hypothesis, $H_0 : \mu_1 = \cdots = \mu_c$. Combining terms:

$$\mathbf{y} = (y_{11}, \ldots, y_{1n_1}, \ldots, y_{c1}, \ldots, y_{cn_c})$$
$$\beta = (\beta_0, \beta_1, \ldots, \beta_c)$$

This results in the non-identifiable, intrinsically aliased model matrix:

$$\mathbf{X} = \begin{pmatrix} \mathbf{1}_{n_1} & \mathbf{1}_{n_1} & \cdots & \mathbf{0}_{n_1} \\ \mathbf{1}_{n_2} & \mathbf{0}_{n_2} & \cdots & \mathbf{0}_{n_2} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{1}_{n_c} & \mathbf{0}_{n_c} & \cdots & \mathbf{1}_{n_c} \end{pmatrix}$$

1.4 Identifiability and Estimability

Identifiability Parameters β are identifiable if whenever $\beta^* \neq \beta \Rightarrow \mathbf{X}\beta^* \neq \mathbf{X}\beta$.

Another characterization is $\mathbf{X}\beta^* = \mathbf{X}\beta \Rightarrow \beta^* = \beta$. This is equivalent to $\mathbf{X}$ being invertible; columns of $\mathbf{X}$ being linearly independent; and $\mathbf{X}$ having full rank.

Example: One-Way ANOVA. The model matrix above is not identifiable because: $\beta = (\beta_0, \beta_1, \ldots, \beta_c)$ and $\beta^* = (\beta_0 + d, \beta_1 - d, \ldots, \beta_c - 3)$ both yield the same linear predictor, namely $\beta_0 + \beta_i$. Thus, we drop the baseline category 1, and get:

$$\mathbf{X} = \begin{pmatrix} \mathbf{1}_{n_1} & \mathbf{0}_{n_1} & \cdots & \mathbf{0}_{n_1} \\ \mathbf{1}_{n_2} & \mathbf{1}_{n_2} & \cdots & \mathbf{0}_{n_2} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{1}_{n_c} & \mathbf{0}_{n_c} & \cdots & \mathbf{1}_{n_c} \end{pmatrix}$$

Thus, our new parameters are $\beta = (\beta_0, \beta_2, \ldots, \beta_c)$ and $\beta_0 = \mu_1$ and $\beta_i = \mu_i - \mu_1$.

Ways to achieve identifiability:

- Drop a parameter: first-category ($\beta_1 = 0$) or last-category baseline ($\beta_c = 0$)
- Add a constraint: $\sum_i n_i \beta_i = 0$ or $\sum_i \beta_i = 0$

General Identifiability $\mathbf{a}^T\beta$ is identifiable if $\mathbf{a}^T\beta^* \neq \mathbf{l}^T\beta \Rightarrow \mathbf{X}\beta^* \neq \mathbf{X}\beta$ (allows for linear combinations and selecting out subsets of parameters)

Estimability $\mathbf{a}^T\beta$ is estimable if $\exists$ coefficients $\mathbf{c}$ such that $E(\mathbf{c}^T\mathbf{y}) = \mathbf{a}^T\beta$.

Note that the definition implies that **all** estimable quantities are *linear combinations of the means.* If β is identifiable, all quantitatives $\mathbf{a}^T\beta$ are estimable.

2 Linear Models: Least Squares Theory

Notation: $\mathbf{y} = (y_1, \ldots, y_n)$ and $\mu_i = E(y_i)$; $\mu = (\mu_1, \ldots, \mu_n)$. The covariance matrix is: $\mathbf{V} = \text{var}(\mathbf{y}) = E[(\mathbf{y} - \mu)(\mathbf{y} - \mu)^T]$

Linear Model: $\mu = \mathbf{X}\beta$ and $\mathbf{V} = \sigma^2\mathbf{I}$ (i.e. identity link with i.i.d. homoscedastic errors)

$$\mathbf{y} = \mathbf{X}\beta + \epsilon, \epsilon \sim \mathbf{0}, \sigma^2\mathbf{I}$$

(This additive structure makes no sense for most GLMs, such as logistic, log-linear, etc., but does for normal linear model and latent variable formulations.)

2.1 Least Squares Fitting

Least Squares How do we get best estimates of parameters $\hat{\beta}$ and fitted values $\hat{\mu} = \mathbf{X}\hat{\beta}$? Use least squares:

$$\min \|\mathbf{y} - \hat{\mu}\|^2 = \min \sum_i \left(y_i - \sum_j \beta_j x_{ij} \right)^2$$

Least squares corresponds to maximum likelihood when $y_i \sim \mathcal{N}(\mu_i, \sigma^2)$.

Normal Equations Minimize squared error by differentiating $L(\beta) = \sum_i (y_i - \mu_i)^2 = \sum_i (y_i - \sum_j \beta_j x_{ij})^2$:

$$\frac{\partial L}{\partial \beta_j} = \sum_i (y_i - \hat{\mu}_i) x_{ij} = 0$$

$$\Rightarrow \boxed{\sum_i y_i x_{ij} = \sum_i \hat{\mu}_i x_{ij}}$$

These are **normal equations**; solving yields estimates $\hat{\beta} = \mathbf{X}^{-1}\hat{\mu}$. Uing matrix algebra:

$$L(\beta) = \|\mathbf{y} - \mathbf{X}\beta\|^2$$

Use matrix derivatives:

$$\frac{\partial(\mathbf{a}^T\beta)}{\partial\beta} = \mathbf{a}$$

$$\frac{\partial(\beta^T\mathbf{A}\beta}{\partial\beta} = (\mathbf{A} + \mathbf{A}^T)\beta$$

This yields the matrix **normal equations**:

$$\mathbf{X}^T\mathbf{y} = \mathbf{X}^T\mathbf{X}\hat{\beta} \Rightarrow \boxed{\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}}$$

Hat Matrix Note that:

$$\hat{\mu} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} = \mathbf{H}\mathbf{y}$$

where we define the **hat matrix**: $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ and is $n \times n$. $\mathbf{H}$ projects y onto $C(\mathbf{X})$, the model space; $\hat{\mu} \in C(\mathbf{X})$. Recall that, using $\beta = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$:

$$E(\hat{\beta}) = \beta, \text{var}(\hat{\beta}) = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$$

Bivariate Regression Let $E(y_i) = \beta_0 + \beta_1 x_i$, with x_i being a quantitative variable. Then the normal equations yield:

$$\sum_i y_i = n\beta_0 + \beta_1 \sum_i x_i, \sum_i x_i y_i = \beta_0 \sum_i x_i + \beta_1 \sum_i x_i^2$$

$$\Rightarrow \hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}, \hat{\beta}_1 = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2}$$

But we see that the Pearson product-moment correlation is:

$$r = \text{corr}(x, y) = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2}\sqrt{\sum_i (y_i - \bar{y})^2}} = \hat{\beta}_1 \frac{s_x}{s_y}$$

So we see that: $\hat{\beta}_1 s_x = r s_y$, that is a change in s_x in x only yields a change in r in $\hat{\mu}$, so we have regression towards the mean.

Orthogonal Subspaces, Residuals Key results from linear algebra:

- $\mathbf{u}, \mathbf{v}$ are orthogonal if $\mathbf{u}^T\mathbf{v} = 0$
- Orthogonal complement if $\mathbf{W}$, vector subspace of $\mathbb{R}^n$, is the set of all $\mathbf{v}$ orthogonal to every $\mathbf{u} \in \mathbf{W}$.
- $\dim(\mathbf{W}) + \dim(\mathbf{W}^\perp) = n$
- Every $\mathbf{y} \in \mathbb{R}^n$ has a unique orthogonal decomposition into $\mathbf{y} = \mathbf{y}_W + \mathbf{y}_{W^\perp}$

$C(\mathbf{X})^\perp$ is the set of all vectors that are orthogonal to all vectors in $C(\mathbf{X})$; since the columns are in $C(\mathbf{X})$, we must have $\mathbf{X}_i^T\mathbf{v} = 0$, where $\mathbf{X}_i$ is a column of $\mathbf{X}$. Thus, $\mathbf{X}^T\mathbf{v} = \mathbf{0}$, so:

$$C(\mathbf{X})^\perp = N(\mathbf{X}^T)$$

Now we define the **residual**: $\mathbf{e} = \mathbf{y} - \mathbf{X}\hat{\beta}$.

From the normal equations, $\mathbf{X}^T(\mathbf{y} - \mathbf{X}\hat{\beta}) = \mathbf{X}^T\mathbf{e} = 0$ so we must have $\mathbf{e} \in N(\mathbf{X}^T) = C(\mathbf{X})^\perp$

2.2 Projections Onto Model Spaces

Projection Matrices A square matrix $\mathbf{P}$ is a projection matrix onto vector subspace $\mathbf{W}$ iff:

1. $\mathbf{y} \in \mathbf{W} \Rightarrow \mathbf{P}\mathbf{y} = \mathbf{y}$
2. $\mathbf{y} \in \mathbf{W}^\perp \Rightarrow \mathbf{P}\mathbf{y} = 0$

Equivalently, $\mathbf{P}$ is project iff:

1. $\mathbf{P}$ is symmetric
2. $\mathbf{P}^2 = \mathbf{P}$, i.e. $\mathbf{P}$ is idempotent

Properties of projection matrices include:

- $\mathbf{P}$ projects onto the space spanned by the columns of $\mathbf{P}$, that is $C(\mathbf{P})$
- $\mathbf{y} = \mathbf{y}_P + \mathbf{y}_{P^\perp}$ uniquely decomposes, so that $\mathbf{P}\mathbf{y} = \mathbf{y}_P$ is unique
- Projection matrix onto any subspace $\mathbf{W}$ is unique
- If $\mathbf{P}$ projects onto $\mathbf{W}$, then $\mathbf{I} - \mathbf{P}$ projects onto $\mathbf{W}^\perp$, so that $\mathbf{y} = \mathbf{P}\mathbf{y} + (\mathbf{I} - \mathbf{P})\mathbf{y}$
- Eigenvalues of $\mathbf{P}$ are all 0 or 1
- $\text{rank}(\mathbf{P}) = \text{trace}(\mathbf{P})$, since the rank of a symmetric matrix is number of nonzero eigenvalues
- If $\{\mathbf{P}_i\}$ are symmetric matrices such that $\sum_i \mathbf{P}_i = \mathbf{I}$, then the following are equivalent: 1) P_i are idempotent; 2) $\mathbf{P}_i\mathbf{P}_j = 0$ for all i, j; 3) $\sum_i \text{rank}(\mathbf{P}_i) = n$

Projection Matrices for Linear Model Spaces Let $\mathbf{P}_X$ be the projection matrix onto $C(\mathbf{X})$. We have the following properties:

- If $\mathbf{X}$ is full rank, then $\mathbf{P}_X = \mathbf{H}$
- If $\mathbf{X}, \mathbf{W}$ are equivalent models, that is $C(\mathbf{X}) = C(\mathbf{W})$, then $\mathbf{P}_X = \mathbf{P}_W$
- When model a is nested in b, i.e. $C(\mathbf{X}_a) \subset C(\mathbf{X}_b)$, then $\mathbf{P}_a\mathbf{P}_b = \mathbf{P}_b\mathbf{P}_a = \mathbf{P}_a$ and $\mathbf{P}_b - \mathbf{P}_a$ are projection matrices

Orthogonal Parameters If $\mathbf{X}_1$ is orthogonal with $\mathbf{X}_2$, then the effects of the reduced model $\mu = \beta_1\mathbf{X}_1$ is the same as the effects of the full model $\mu = \beta_{1\cdot2}\mathbf{X}_1 + \beta_{2\cdot1}X_2$. Suppose that $\mathbf{X} = (\mathbf{X}_1 : \mathbf{X}_2)$. Then:

$$\mathbf{X}^T\mathbf{X} = \begin{pmatrix} \mathbf{X}_1^T\mathbf{X}_1 & 0 \\ 0 & \mathbf{X}_2^T\mathbf{X}_2 \end{pmatrix}, \mathbf{X}^T\mathbf{y} = \begin{pmatrix} \mathbf{X}_1^T\mathbf{y} \\ \mathbf{X}_2^T\mathbf{y} \end{pmatrix}$$

$$\Rightarrow \beta = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \Rightarrow \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} = \begin{pmatrix} (\mathbf{X}_1^T\mathbf{X}_1)^{-1}\mathbf{X}_1^T\mathbf{y} \\ (\mathbf{X}_2^T\mathbf{X}_2)^{-1}\mathbf{X}_2^T\mathbf{y} \end{pmatrix}$$

so the parameters are exactly the same as when fitted separately.

Pythagoras' Theorem for Linear Models Because of orthogonality properties of the projection onto the model space, we can apply Pythagoras' theorem:

- Unique least squares fit: $\|\mathbf{y} - \mathbf{P}_X\mathbf{y}\| \leq \|\mathbf{y} - \mathbf{z}\|$ for all $\mathbf{z} \in C(\mathbf{X})$
- True and sample residuals: $\|\mathbf{y} - \mu\|^2 = \|\mathbf{y} - \hat{\mu}\|^2 + \|\hat{\mu} - \mu\|^2$ (assuming that the model is correct, i.e. $\mu \in C(\mathbf{X})$)
- Data = fit + residuals (sum of squares): $\|\mathbf{y}\|^2 = \|\hat{\mu}\|^2 + \|\mathbf{y} - \hat{\mu}\|^2$

2.3 Linear Model Examples

Null Model $E(y_i) = \beta$ (no explanatory variables) Then, the model matrix and projection matrix are:

$$\mathbf{X} = \mathbf{1}_n, \mathbf{P}_X = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T = \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T$$

This yields the fitted values: $\hat{\mu} = \mathbf{P}_X\mathbf{y} = \bar{y}\mathbf{1}_n$

The corresponding sum of squares is: $\mathbf{y}^T\mathbf{y} = \mathbf{y}^T\mathbf{P}_X\mathbf{y} + \mathbf{y}^T(\mathbf{I} - \mathbf{P}_X)\mathbf{y} \Rightarrow \sum_i y_i^2 = n\bar{y}^2 + \sum_i(y_i - \bar{y})^2$

One-Way Layout The non-identifiable model matrix and generalized inverses are:

$$\mathbf{X} = \begin{pmatrix} \mathbf{1}_{n_1} & \mathbf{1}_{n_1} & \cdots & \mathbf{0}_{n_1} \\ \mathbf{1}_{n_2} & \mathbf{0}_{n_2} & \cdots & \mathbf{0}_{n_2} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{1}_{n_c} & \mathbf{0}_{n_c} & \cdots & \mathbf{1}_{n_c} \end{pmatrix}, (\mathbf{X}^T\mathbf{X})^- \begin{pmatrix} 0 & 0 & \cdots & 0 \\ 0 & 1/n_1 & \cdot & 0 \\ \vdots & \vdots\ddots & \vdots & \\ 0 & 0 & \cdots & 1/n_c \end{pmatrix}$$

Alternatively, we can use the first-category baseline constraint:

$$\mathbf{X} = \begin{pmatrix} \mathbf{1}_{n_1} & \mathbf{0}_{n_1} & \cdots & \mathbf{0}_{n_1} \\ \mathbf{1}_{n_2} & \mathbf{1}_{n_2} & \cdots & \mathbf{0}_{n_2} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{1}_{n_c} & \mathbf{0}_{n_c} & \cdots & \mathbf{1}_{n_c} \end{pmatrix}$$

Either way, we get the projection matrix:

$$\mathbf{P}_X = \begin{pmatrix} \frac{1}{n_1}1_{n_1}1_{n_1}^T & 0 & \cdots & 0 \\ 0 & \frac{1}{n_2}1_{n_2}1_{n_2}^T & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{1}{n_c}1_{n_c}1_{n_c}^T \end{pmatrix}$$

which yields: $\hat{\mu} = \mathbf{P}_X\mathbf{y} = (\bar{y}_1, \ldots, \bar{y}_1, \ldots, \bar{y}_c, \ldots, \bar{y}_c)$

The relevant sum of squares decomposition for one-way ANOVA is:

$$y_{ij} = \bar{y} + (\bar{y}_i - \bar{y}) + (y_{ij} - \bar{y}_i)$$

i.e. obs = overall mean + between-groups + within-groups. This corresponds to using the $\mathbf{P}_0$ and $\mathbf{P}_X$ projection matrices for the null model and the one-way layout model, respectively, yielding:

$$\mathbf{y}^T\mathbf{y} = \mathbf{y}^T[\mathbf{P}_0 + (\mathbf{P}_X - \mathbf{P}_0) + (\mathbf{I} - \mathbf{P}_X)]\mathbf{y}$$

$$\Rightarrow \sum_{i=1}^{c}\sum_{j=1}^{n_i} y_{ij}^2 = n\bar{y}^2 + \sum_{i=1}^{c}(\bar{y}_i - \bar{y})^2 + \sum_{i=1}^{c}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2$$

which yields the ANOVA table:

Source	Projection matrix	df	SS
Mean	$\mathbf{P}_0$	1	$n\bar{y}^2$
Groups	$\mathbf{P}_X$	$c-1$	$\sum_{i=1}^{c}(\bar{y}_i - \bar{y})^2$
Error	$\mathbf{I} - \mathbf{P}_X$	$n-c$	$\sum_{i=1}^{c}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2$
Total	$\mathbf{I}$	n	$\sum_{i=1}^{c}\sum_{j=1}^{n_i} y_{ij}^2$

Two-Way Layout Suppose we have two facts rather than one (i.e. rows are treatments, columns are experimental blocks). Let there be $i = 1, \dots, r$ rows and $j = 1, \dots, c$ columns. The model is:

$$E(y_{ij}) = \beta_0 + \beta_i + \gamma_j$$

with $\beta_1 = \gamma_1 = 0$ for identifiability. Letting $\mathbf{y} = (y_{11}, \dots, y_{1c}, \dots, y_{r1}, \dots, y_{rc})$, the relevant projections are:

$$\mathbf{P}_r = \begin{pmatrix} 1/c & \cdots & 1/c & \cdots & 0 & \cdots & 0 \\ \dots & \ddots & \vdots & & \vdots & \ddots & \vdots \\ 1/c & \cdots & 1/c & \cdots & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & \cdots & 0 & \cdots & 1/c & \cdots & 1/c \\ \vdots & \ddots & \vdots & & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & \cdots & 1/c & \cdots & 1/c \end{pmatrix}, \mathbf{P}_c = \frac{1}{r}\begin{pmatrix} \mathbf{I}_{r\times r} & \cdots & \mathbf{I}_{r\times r} \\ \vdots & \ddots & \vdots \\ \mathbf{I}_{r\times r} & \cdots & \mathbf{I}_{r\times r} \end{pmatrix}$$

which project onto separate one-way layouts for the row factor and the column factor separately. That is:

$$\mathbf{P}_r\mathbf{y} = (\bar{y}_{1\cdot}, \dots, \bar{y}_{1\cdot}, \dots, \bar{y}_{c\cdot}, \dots, \bar{y}_{c\cdot})$$

$$\mathbf{P}_c\mathbf{y} = (\bar{y}_{\cdot 1}, \dots, \bar{y}_{\cdot r}, \dots, \bar{y}_{\cdot 1}, \dots, \bar{y}_{\cdot r})$$

This yields the ANOVA table:

Source	Projection matrix	df	SS
Mean	$\mathbf{P}_0$	1	$rc\bar{y}^2$
Rows	$\mathbf{P}_r - \mathbf{P}_0$	$r-1$	$c\sum_{i=1}^{r}(\bar{y}_{i\cdot} - \bar{y})^2$
Columns	$\mathbf{P}_c - \mathbf{P}_0$	$c-1$	$r\sum_{j=1}^{c}(\bar{y}_{\cdot j} - \bar{y})^2$
Error	$\mathbf{I} - \mathbf{P}_r - \mathbf{P}_c + \mathbf{P}_0$	$(r-1)(c-1)$	$\sum_{i=1}^{r}\sum_{j=1}^{c}(y_{ij} - \bar{y}_{i\cdot} - \bar{y}_{\cdot j} + \bar{y})^2$
Total	$\mathbf{I}$	$n = rc$	$\sum_{i=1}^{r}\sum_{j=1}^{c} y_{ij}^2$

2.4 Summarizing Variability in Linear Models

We can use the fact that the residual is in the error space to glean information about the error term ϵ.

Estimating Error Variance We assume that the error term has $\text{var}(\epsilon) = \sigma^2\mathbf{I}$, so we want to estimate σ^2. We use the fact that:

$$E(\mathbf{y}^T\mathbf{A}\mathbf{y}) = \text{trace}(\mathbf{AV}) + \mu^T\mathbf{A}\mu$$

where $\mathbf{V}$ is the variance of the error term, that is $\mathbf{V} = \sigma^2\mathbf{I}$. Using $\mathbf{A} = \mathbf{I} - \mathbf{P}_X$, we have:

$$E[\mathbf{y}^T(\mathbf{I} - \mathbf{P}_x)\mathbf{y}] = \text{trace}[(\mathbf{I} - \mathbf{P}_X)\sigma^2\mathbf{I}] + \mu^T(\mathbf{I} - \mathbf{P}_X)\mu = \sigma^2\text{trace}(\mathbf{I} - \mathbf{P}_X) = \sigma^2(n-p)$$

$$\Rightarrow \boxed{E\left[\frac{\mathbf{y}^T(\mathbf{I} - \mathbf{P}_X)\mathbf{y}}{n-p}\right] = \sigma^2}$$

So that $s^2 = \frac{\mathbf{y}^T(\mathbf{I}-\mathbf{P}_X)\mathbf{y}}{n-p} = \frac{\sum_i(y_i - \hat{\mu}_i)^2}{n-p}$ is an unbiased estimator for σ^2; that is, the average error taken with respect to the dimension of the error space, $n-p$. s^2 is called the error mean square.

SSE and SSR We split up the sums of squares in ANOVA fashion, to get:

$$\sum_i (y_i - \bar{y})^2 = \sum_i (\hat{\mu}_i - \bar{y})^2 + \sum_i (\mathbf{y}_i - \hat{\mu}_i)^2$$

- Total sum of squares (TSS): $\sum_i (y_i - \bar{y})^2$, that is the variability in y_i after correcting for the overall mean (i.e. from null model)
- Regression sum of squares (SSR): $\sum_i (\hat{\mu}_i - \bar{y})^2$, that is the variability in y_i explained by the model
- Error sum of squares (SSE): $\sum_i (y_i - \hat{\mu}_i)^2$, that is the variability in y_i unexplained by the full model

For the one-way layout, $SSR = \sum_i n_i(\bar{y}_i - \bar{y})^2$ = Between-groups SS, whereas $SSE = \sum_i \sum_j (y_{ij} - \bar{y}_i)^2$ = Within-groups SS.

Adding Variables on SSE/SSR When we add more explanatory variables, SSE decreases monotonically while SSR increases monotonically (since we can set new $\beta_p = 0$).

Sequential Sums of Squares Consider p explanatory variables $x_1, \ldots, x_p$, entered into model 1 at a time. We get incremental SSR:

$$SSR(x_1), SSR(x_2|x_1), \ldots, SSR(x_p|x_1, \ldots, x_{p-1})$$

where, say, $SSR(x_2|x_1) = \sum_i (\hat{\mu}_{i12} - \hat{\mu}_{i1})^2$ from fitting with both x_1, x_2 vs. fitting with only x_1 (from orthogonal decomposition). Note:

$$SSR(x_1, \ldots, x_p) = SSR(x_1) + SSR(x_2|x_1) + \cdots + SSR(x_p|x_1, \ldots, x_{p-1})$$

Partial Sums of Squares We can consider full conditional SSR of x_i given all other x_{-i}:

$$SSR(x_1|x_2, \ldots, x_p), SSR(x_2|x_1, x_3, \ldots, x_p), \ldots, SSR(x_p|x_1, \ldots, x_{p-1})$$

that is, additional variability explained by x_i given all other variables are already in the model.

R^2

$$R^2 = \frac{SSR}{TSS} = \frac{TSS - SSE}{TSS} = \frac{\sum_i (y_i - \bar{y})^2 - \sum_i (y_i - \hat{\mu}_i)^2}{\sum_i (y_i - \bar{y})^2}$$

so R^2 measures the proportional reduction in error from null model to full model; $R^2 \in [0, 1]$.

Multiple Correlation Another way to measure predictive power: sample correlation between y_i and $\hat{\mu}_i$. (Note: $\bar{\hat{\mu}} = \bar{y}$ due to normal equations with intercept term.)

$$\text{corr}(\mathbf{y}, \hat{\mu}) = \frac{\sum_i (y_i - \bar{y})(\hat{\mu}_i - \bar{\hat{\mu}})}{\sqrt{\sum_i (y_i - \bar{y})^2}\sqrt{\sum_i (\hat{\mu}_i - \bar{\hat{\mu}})^2}} = \frac{\sum_i (\hat{\mu}_i - \bar{y})^2}{\sqrt{\sum_i (y_i - \bar{y})^2}\sqrt{\sum_i (\hat{\mu}_i - \bar{y})^2}}$$

$$\Rightarrow \boxed{\text{corr}(\mathbf{y}, \hat{\mu}) = +\sqrt{R^2} = R}$$

Adjusted R^2 When: 1) n is small; 2) p is large, R^2 is overoptimistic. Thus, we can use the *adjusted* R^2:

$$\text{adj. } R^2 = 1 - \frac{SSE/(n-p)}{TSS/(n-1)} = 1 - \frac{n-1}{n-p}(1 - R^2)$$

2.5 Residuals, Leverage, and Influence

Residuals are in error space ⇒ orthogonal to model space ⇒ contain information in data not explained by model ⇒ used to investigate model lack of fit.

Plots of Residuals for Model Fit $\text{corr}(\mathbf{e}, \hat{\mu}) = 0$ due to orthogonality, so we can plot $\mathbf{e}$ vs. $\hat{\mu}$ to check lack of fit (should have slope 0). Possible problems:

1. Heteroscedasticity: "fan-shaped" plot of $\mathbf{e}$ vs. $\hat{\mu}$, i.e. non-constant variance
2. Nonlinearity: "U-shaped" plot; signals higher-order terms neded

Other diagnostic: histogram of residuals should be approximately Normal.

Standardized/Studentized Residuals Recall that:

$$\text{var}(\hat{\mu}) = \sigma^2\mathbf{H}, \text{var}(\mathbf{e}) = \sigma^2(\mathbf{I} - \mathbf{H})$$

so the residuals are correlated and don't have variance 1. We want all residuals to have variance 1, so we standardized:

$$r_i = \frac{y_i - \hat{\mu}_i}{s\sqrt{1 - h_{ii}}}$$

so that $\text{var}(r_i) = \frac{1}{s^2(1-h_{ii})}\sigma^2(1 - h_{ii}) \approx 1$. The studentized residual is obtained by estimating s with all observations besides i. Standardized residual describes how many estimated standard deviations e_i falls from 0.

Leverage $h_{ii} = [\mathbf{H}]_{ii}$ is leverage of observation i. If $h_{ii} \approx 1$, then y_i has a large influence on $\hat{\mu}_i$. Properties:

- $\hat{\mu}_i = \sum_j h_{ij}y_j \Rightarrow \frac{\partial \hat{\mu}_i}{\partial y_i} = h_i i$
- Since we assume y_i are uncorrelated:

$$\text{Cov}(y_i\hat{\mu}_i) = \text{Cov}\left(\mathbf{y}_i, \sum_j h_{ij}y_j\right) = \sum_j h_{ij}\text{Cov}(y_i, y_j) = h_{ii}\text{Cov}(y_i, y_i) = h_{ii}\sigma^2$$

  and since $\text{var}\hat{\mu}_i = \sigma^2 h_{ii}$, we have:

$$\text{corr}(y_i, \hat{\mu})i) = \frac{\sigma^2 h_{ii}}{\sqrt{\sigma^2 \cdot \sigma^2 h_{ii}}} = \sqrt{h_{ii}}$$

- With p explanatory variables, leverages have mean $\frac{p}{n}$
- Larger deviation of x_i from $\bar{x}$ yields higher leverage

Cook's Distance To be influential, observation must have: 1) large leverage; 2) large standardized residual. We can combine measures to get Cook's distance:

$$D_i = r_i^2\left[\frac{h_{ii}}{p(1 - h_{ii})}\right] = \frac{(y_i - \hat{\mu}_i)^2}{ps^2}\frac{h_{ii}}{(1 - h_{ii})^2}$$

"Adjusting for Other Variables" The effect of x_i in a model of $x_1, \ldots, x_p$ is the same as: 1) regressing y on x_{-i}; 2) regressing x_i on x_{-i}; 3) effect of regressing residuals from (1) on residuals from (2).

Example. Consider $E(y_i) = \beta_{1\cdot 2}x_{i1} + \beta_{2\cdot 1}x_{i2}$. 1) Regress $E(y_i) = \beta_2 x_{i2}$; 2) Regress $E(x_{i1}) = \beta_{12}x_{i2}$. The normal equations are: 1) $\sum_i x_{i2}(y_i - \hat{\beta}_2 x_{i2}) = 0$; 2) $\sum_i x_{i2}(x_{i1} - \hat{\beta}_{12}x_{i2}) = 0$. Similar equations for multiple regression. Plugging in and solving yields:

$$\hat{\beta}_{1\cdot 2} = \frac{\sum_i(y_i - \hat{\beta_2 x}_{i2})(x_{i1} - \hat{\beta}_{12}x_{i2})}{\sum_i(x_{i1} - \hat{\beta}_{12}x_{i2})^2}$$

But this is exactly the effect of regressing residuals from (1), $y_i - \hat{\beta}_2 x_{i2}$ on the residuals from (2), $x_{i1} - \hat{\beta}_{12}x_{i2}$. From this we also see that plugging into the regression of residuals equation,

$$\hat{\beta}_{2\cdot 1} = \hat{\beta}_2 - \hat{\beta}_{1\cdot 2}\hat{\beta}_{12}$$

i.e. the subtracted term represents omitted variable bias from trying to estimate the effect of x_1 without including x_2.

2.6 Gauss-Markov Theorem

Why least squares? We've noted a number of good properties, such as:

- The least squares estimate $\hat{\mu}$ is maximally correlated with $\mathbf{y}$
- It yields nice interpretability in terms of orthogonal subspaces, and orthogonal decomposition in terms of fitted values and residuals
- It corresponds to maximum likelihood estimation under normality assumption

We add another *optimality condition* about least squares:

Gauss-Markov Theorem. If $E(\mathbf{y}) = \mathbf{X}\beta$ holds and $\mathbf{X}$ has full rank with $\text{var}(\mathbf{y}) = \sigma^2\mathbf{I}$, then the least squares estimator $\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$ is the *best linear unbiased estimator* (BLUE) of β. That is, for any quantity $\mathbf{a}^T\beta$, the estimator $\mathbf{a}^T\hat{\beta}$ has the minimum variance among all estimators that are: 1) linear in $\mathbf{y}$; 2) unbiased.

If we add normality to $\mathbf{y}$, then the least squares estimator becomes *minimum variance unbiased estimator* (MVUE); i.e., the restriction of linearity in $\mathbf{y}$ is removed.

2.7 Generalized Least Squares

If $\mathbf{y}$ not i.i.d, that is $\text{var}(\mathbf{y}) = \sigma^2\mathbf{V}$ with $\mathbf{V} \neq \mathbf{I}$, use GLS. Use spectral decomposition to write $\mathbf{V} = \mathbf{Q}\Lambda\mathbf{Q}^T$ and $\mathbf{V}^{1/2} = \mathbf{Q}\Lambda^{1/2}\mathbf{Q}^T$ for orthogonal $\mathbf{Q}$. Let $\mathbf{y}^* = \mathbf{V}^{-1/2}\mathbf{y}$ and $\mathbf{X}^* = \mathbf{V}^{-1/2}\mathbf{X}$; then $E(\mathbf{y}^*) = \mathbf{V}^{-1/2}\mathbf{X}\beta = \mathbf{X}^*\beta$ and $\text{var}(\mathbf{y}^*) = \sigma^2\mathbf{V}^{-1/2}\mathbf{V}(\mathbf{V}^{-1/2})^T = \sigma^2\mathbf{I}$ so $\mathbf{y}*$ satisfies OLS.

Minimize squared error: $(\mathbf{y}^* - \mathbf{X}^*\beta)^T(\mathbf{y}^* - \mathbf{X}^*\beta) = (\mathbf{y} - \mathbf{X}\beta)^T\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\beta)$ so the normal equations are: $[(\mathbf{X}^*)^T\mathbf{X}^*]\beta = (\mathbf{X}^*)^T\mathbf{y}^* \Rightarrow (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})\beta = \mathbf{X}^T\mathbf{V}^{-1}\mathbf{y}$ and therefore:

$$\boxed{\hat{\beta}_{GLS} = (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{y}}$$

- Unbiased: $E(\hat{\beta}_{GLS}) = (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}E(\mathbf{y}) = \beta$
- Covariance: $\text{var}(\hat{\beta}_{GLS}) = \sigma^2(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}$
- BLUE estimator for β; MVUE and ML under normality
- Hat matrix: $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}$ not necessarily projection because need not be symmetric ($\hat{\mu} = \mathbf{X}\hat{\beta}_{GLS} = \mathbf{X}(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{y}$)
- Generalized projection: if $\mathbf{u} \in C(\mathbf{X})$, then $\mathbf{H}\mathbf{u} = \mathbf{u}$; and if $\mathbf{v} \in C(\mathbf{X})^\perp = \mathcal{N}(\mathbf{X}^T)$, then $\mathbf{H}\mathbf{v} = 0$ (since $(\mathbf{u}, \mathbf{v}) = 0$)
- Estimated variance: If $\text{rank}(\mathbf{X}) = r$, $s^2 = \frac{(\mathbf{y}^* - \mathbf{X}^*\hat{\beta})^T(\mathbf{y}^* - \mathbf{X}^*\hat{\beta})}{n-r} = \frac{(\mathbf{y} - \hat{\mu})^T\mathbf{V}^{-1}(\mathbf{y} - \hat{\mu})}{n-r}$

3 Normal Linear Models

Normal Linear Model: In addition to $\mu = \mathbf{X}\beta$ and $\mathbf{V} = \text{var}(\mathbf{y}) = \sigma^2\mathbf{I}$, assume that y_i follow Normal distribution, that is: $\mathbf{y} \sim \mathcal{N}(\mathbf{X}\beta, \sigma^2\mathbf{I})$, or $\mathbf{y} = \mathbf{X}\beta + \epsilon$ where $\epsilon \sim \mathcal{N}(0, \sigma^2\mathbf{I})$.

3.1 Normal and Related Distributions

Multivariate Normal Denoted $\mathbf{y} \sim \mathcal{N}(\mu, \mathbf{V})$; properties include:

- PDF: $f(\mathbf{y}) = (2\pi)^{-n/2}|\mathbf{V}|^{-1/2} \exp\left[-\frac{1}{2}(\mathbf{y}-\mu)^T\mathbf{V}^{-1}(\mathbf{y}-\mu)\right]$
- $\mathbf{x} = \mathbf{A}\mathbf{y} + \mathbf{b} \Rightarrow \mathbf{x} \sim \mathcal{N}(\mathbf{A}\mu + \mathbf{b}, \mathbf{A}\mathbf{V}\mathbf{A}^T)$
- If $y = \begin{pmatrix}\mathbf{y}_1 \\ \mathbf{y}_2\end{pmatrix}$, i.e. partitions, with $\mathbf{V} = \begin{pmatrix}\mathbf{V}_{11} & \mathbf{V}_{12} \\ \mathbf{V}_{21} & \mathbf{V}_{22}\end{pmatrix}$, then:
  $\mathbf{y}_1 \perp \mathbf{y}_2$ iff $\mathbf{V}_{12} = 0$ (i.e. independence iff uncorrelated)
- As corollary, if $\mathbf{V} = \sigma^2\mathbf{I}$, then $\mathbf{y}_i \sim \mathcal{N}(\mu_i, \sigma^2)$ and $y_i \perp y_j$ for all i, j

Chi-Squared Denoted χ^2_p for p degrees of freedom:

- If $y_i \sim \mathcal{N}(0,1)$ i.i.d, then $\sum_{i=1}^p y_i^2 \sim \chi^2_p$
- Generally: if $\mathbf{y} \sim \mathcal{N}(\mu, \mathbf{V})$ is p-dimensional, then:
$$(\mathbf{y}-\mu)^T\mathbf{V}^{-1}(\mathbf{y}-\mu) \sim \chi^2_p$$
- Moments: $E[\chi^2_p] = p$ and $\text{var}(\chi^2_p) = 2p$

t Distribution Denoted t_p for p degrees of freedom:

- If $z \sim \mathcal{N}(0,1)$ and $x \sim \chi^2_p$, $x \perp z$, then:
$$\frac{z}{\sqrt{x/p}} \sim t_p$$
- Symmetric about 0: $E(t_p) = 0$ and $\text{var}(t_p) = \frac{p}{p-2}$ $(p > 2)$
- Converges to $\mathcal{N}(0,1)$ as $p \to \infty$
- Suppose $y_1, \dots, y_n \sim \mathcal{N}(\mu, \sigma^2)$, sample mean $\bar{y}$ and sample variance s^2. Under null hypothesis $H_0 : \mu = \mu_0$:
$$z = \frac{\bar{y}-\mu_0}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1) \text{ and } x = \frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{n-1}$$
$$\Rightarrow \frac{z}{\sqrt{x/(n-1)}} = \frac{\bar{y}-\mu_0}{s/\sqrt{n}} \sim t_{n-1}$$
  and larger values of $|t|$ mean stronger evidence against H_0

F Distribution Denoted $F_{p,q}$ for degrees of freedom p, q:

- If $x \sim \chi^2_p$, $y \sim \chi^2_q$, $x \perp y$, then:
$$\frac{x/p}{y/q} \sim F_{p,q}$$
- Mean: $E(F_{p,q}) = \frac{q}{q-2}$ (for $q > 2$)
- $(t_p)^2 = F_{1,p}$

Noncentral Distributions Used to analyze test statistics when null hypothesis does not hold.

- **Chi-Squared:** If $\mathbf{y}_i \sim \mathcal{N}(\mu_i, 1)$, then noncentrality parameter $\lambda = \sum_{i=1}^{p} \mu_i$ and $\sum_{i=1}^{p} y_i \sim \chi^2_{p,\lambda}$
  Moments are: $E(\chi^2_{p,\lambda}) = p + \lambda$; $\text{var}(\chi^2_{p,\lambda}) = 2(p + 2\lambda)$
  More generally, if p-dimensional $\mathbf{y} \sim \mathcal{N}(\mu, \mathbf{V})$, then: $\mathbf{y}^T\mathbf{V}^{-1}\mathbf{y} \sim \chi^2_{p,\lambda}$ with $\lambda = \mu^T\mathbf{V}^{-1}\mu$
- **t Distribution**: If $z \sim \mathcal{N}(\mu, 1)$, $x \sim \chi^2_p$, $x \perp z$, then:
$$\frac{z}{\sqrt{x/p}} \sim t_{p,\mu}$$
  with degrees of freedom p and noncentrality μ (from z)
  Skewed in direction of sign of μ; $t_{p,\mu} \to \mathcal{N}(\mu, 1)$ as $p \to \infty$
- **F Distribution**: If $x \sim \chi^2_{p,\lambda}$, y$\sim \chi^2_q$, $x \perp y$, then:
$$\frac{x/p}{y/q} \sim F_{p,q,\lambda}$$
  with mean $1 + \frac{\lambda}{p}$ for large q.

Cochran's Theorem and Normal Quadratic Forms Some preliminary results:

- If $\mathbf{y} \sim \mathcal{N}(\mu, \mathbf{V})$ and $\mathbf{A}$ is symmetric, then:
$$\mathbf{y}^T\mathbf{A}\mathbf{y} \sim \chi^2_{r,\mu^T\mathbf{A}\mu} \Leftrightarrow \mathbf{AV} \text{ is idempotent of rank } r$$
- Letting $\mathbf{A} = \mathbf{P}$ for $\mathbf{y} \sim \mathcal{N}(\mu, \sigma^2\mathbf{I})$, and since $\mathbf{y}/\sigma \sim \mathcal{N}(\mu/\sigma, \mathbf{I})$:
$$\mathbf{y}^T\mathbf{P}\mathbf{y}/\sigma^2 \sim \chi^2_{r,\mu^T\mathbf{P}\mu/\sigma^2}$$
- Using standardized $(\mathbf{y} - \mu)/\sigma$, we have the important result:
$$\frac{1}{\sigma^2}(\mathbf{y} - \mu)^T\mathbf{P}(\mathbf{y} - \mu) \sim \chi^2_r \Leftrightarrow \mathbf{P} \text{ is projection matrix of rank } r$$
  which tells us: degrees of freedom = rank of $\mathbf{P}$ = dimension of vector space projected to by $\mathbf{P}$

Cochran's Theorem. Suppose n observations $\mathbf{y} \sim \mathcal{N}(\mu, \sigma^2\mathbf{I})$ and $\mathbf{P}_1, \dots, \mathbf{P}_k$ are projection matrices s.t. $\sum_i \mathbf{P}_i = \mathbf{I}$. Then:

1. $\{\mathbf{y}^T\mathbf{P}_i\mathbf{y}\}$ are independent
2. $\frac{1}{\sigma^2}\mathbf{y}^T\mathbf{P}_i\mathbf{y} \sim \chi^2_{r_i,\lambda_i}$, with $r_i = \text{rank}(\mathbf{P}_i)$ and $\lambda_i = \frac{1}{\sigma^2}\mu^T\mathbf{P}_i\mu$

3.2 Significance Tests for Normal Linear Model

Cochran's Theorem is useful because it can be applied to prove more or less any significant test result for normal linear models.

Introduction: One-Way ANOVA $E(y_{ij}) = \beta_0 + \beta_i$, with baseline constraint. Consider $H_0 : \mu_1 = \cdots = \mu_c$, or equivalently $H_0 : \beta_1 = \cdots = \beta_c$. Under H_0, we have $E(y_{ij}) = \beta_0$, or the null model. We use decomposition:
$$\mathbf{I} = \mathbf{P}_0 + (\mathbf{P}_X - \mathbf{P}_0) + (\mathbf{I} - \mathbf{P}_X)$$
with $\mathbf{P}_X$ having blocks $\frac{1}{n_i}\mathbf{1}_{n_i}\mathbf{1}_{n_i}^T$ and $\mathbf{P}_0 = \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T$. Applying Cochran's Theorem, $\mathbf{P}_X - \mathbf{P}_0$ and $\mathbf{I} - \mathbf{P}_X$ are both projection matrices and are perpendicular, so:
$$\frac{1}{\sigma^2}\mathbf{y}^T(\mathbf{P}_X - \mathbf{P}_0)\mathbf{y} = \frac{1}{\sigma^2}\sum_{i=1}^{c} n_i(\bar{y}_i - \bar{y})^2 \sim \chi^2_{c-1,\lambda}$$
$$\frac{1}{\sigma^2}\mathbf{y}^T(\mathbf{I} - \mathbf{P}_X)\mathbf{y} = \frac{1}{\sigma^2}\sum_{i=1}^{c}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2 \sim \chi_{n-c}$$

where $\lambda = \frac{1}{\sigma^2}\mu^T(\mathbf{P}_X - \mathbf{P}_0)\mu = \frac{1}{\sigma^2}\sum_i n_i(\mu_i - \bar{\mu})^2$ and the quadratic forms are independent. Thus, we can create an F test:

$$F = \frac{\sum_i n_i(\bar{y}_i - \bar{y})^2/(c-1)}{\sum_i \sum_j (y_{ij} - \bar{y}_i)^2/(n-c)} \sim F_{c-1,n-c,\lambda}$$

Under H_0, we have $\lambda = 0, df_1 = c-1, df_2 = n-c$, so expected value $\frac{n-c}{n-c-2}$, and larger F values are stronger evidence against H_0.

$$\text{p-value} = P(F_{c-1,n-c} > F_{obs})$$

Source	df	SS	F_{obs}
Mean	1	$n\bar{y}^2$	
Groups	$c-1$	$\sum_{i=1}^c(\bar{y}_i - \bar{y})^2$	$\frac{\sum_i n_i(\bar{y}_i - \bar{y})^2/(c-1)}{\sum_i \sum_j (y_{ij} - \bar{y}_i)^2/(n-c)} \sim F_{c-1,n-c,\lambda}$
Error	$n-c$	$\sum_{i=1}^c \sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2$	
Total	n	$\sum_{i=1}^c \sum_{j=1}^{n_i} y_{ij}^2$	

Comparing Nested Models Let simpler model be M_0 with p_0 parameters, projection $\mathbf{P}_0$, and complicated model be M_1 with p_1 parameters, projection $\mathbf{P}_1$. Decomposition yields $\mathbf{I} = \mathbf{P}_0 + (\mathbf{P}_1 - \mathbf{P}_0) + (\mathbf{I} - \mathbf{P}_1)$ with the sum of squares decomposition:

$$\mathbf{y}^T\mathbf{y} = \mathbf{y}^T\mathbf{P}_0\mathbf{y} + \mathbf{y}^T(\mathbf{P}_1 - \mathbf{P}_0)\mathbf{y} + \mathbf{y}^T(\mathbf{I} - \mathbf{P}_1)\mathbf{y}$$

$\mathbf{y}^T(\mathbf{P}_1 - \mathbf{P}_0)\mathbf{y} = \mathbf{y}^T(\mathbf{I} - \mathbf{P}_0)\mathbf{y} - \mathbf{y}^T(\mathbf{I} - \mathbf{P}_1)\mathbf{y} = \sum_i(y_i - \hat{\mu}_{i0})^2 - \sum_i(y_i - \hat{\mu}_{i1})^2 = SSE_0 - SSE_1 = \sum_i(\hat{\mu}_{i1} - \hat{\mu}_{i0})^2 = SSR(M_1|M_0)$. Similarly, $\mathbf{y}^T(\mathbf{I} - \mathbf{P}_1)\mathbf{y} = \sum_i(y_i - \hat{\mu}_{i1})^2 = SSE_1$. $\mathbf{I} - \mathbf{P}_1$ has df $n - p_1$ while $\mathbf{P}_1 - \mathbf{P}_0$ has df $p_1 - p_0$. Thus, we have:

$$\frac{1}{\sigma^2}\mathbf{y}^T(\mathbf{P}_1 - \mathbf{P}_0)\mathbf{y} = \frac{SSE_0 - SSE_1}{\sigma^2} \sim \chi^2_{p_1 - p_0,\lambda}$$

$$\frac{1}{\sigma^2}\mathbf{y}^T(\mathbf{I} - \mathbf{P}_1)\mathbf{y} = \frac{SSE_1}{\sigma^2} \sim \chi^2_{n-p_1}$$

with $\lambda = \frac{1}{\sigma^2}\mu^T(\mathbf{P}_1 - \mathbf{P}_0)\mu = \frac{\|\mu_1 - \mu_0\|^2}{\sigma^2}$ which is 0 under H_0. Thus, under H_0:

$$F = \frac{(SSE_0 - SSE_1)/(p_1 - p_0)}{SSE_1/(n-p_1)} = \frac{SSR(M_1|M_0)/(p_1 - p_0)}{s^2} \sim F_{p_1 - p_0, n-p_1,\lambda}$$

where s^2 is the σ^2 estimator under M_1.

Example: All Effects Equal 0. Let $M_1 : E(y_i) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_{p-1}x_{i,p-1}$ and $M_0 : E(y_i) = \beta_0$ be the null model. Consider $H_0 : \beta_1 = \cdots = \beta_{p-1} = 0$. For M_0, we have $\mathbf{P}_0 = \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T$ and the SS decomposition is:

$$\mathbf{y}^T\mathbf{y} = \mathbf{y}^T\mathbf{P}_0\mathbf{y}^T + \mathbf{y}^T(\mathbf{P}_1 - \mathbf{P}_0)\mathbf{y} + \mathbf{y}^T(\mathbf{I} - \mathbf{P}_1)\mathbf{y}$$

with the same ANOVA table as in the one-way layout.

Non-null Behavior of F Statistic. How large can we expect $SSE_0 - SSE_1 = \|\hat{\mu}_1 - \hat{\mu}_0\|^2$ to be under non-null? Let μ_1 be true mean under M_1, and μ_0 be projection of μ_1 onto M_0. Then the numerator has expectation:

$$E\|\hat{\mu}_1 - \hat{\mu}_0\|^2 = E[\mathbf{y}^T(\mathbf{P}_1 - \mathbf{P}_0)\mathbf{y}] = \text{trace}[(\mathbf{P}_1 - \mathbf{P}_0)\sigma^2\mathbf{I}] + \mu_1^T(\mathbf{P}_1 - \mathbf{P}_0)\mu_1 = \sigma^2(p_1 - p_0) + \|\mu_1 - \mu_0\|^2$$

$$E\left[\frac{\|\hat{\mu}_1 - \hat{\mu}_0\|^2}{p_1 - p_0}\right] = \sigma^2 + \frac{\|\mu_1 - \mu_0\|^2}{p_1 - p_0}$$

while the denominator has expectation:

$$E\|\mathbf{y} - \hat{\mu}_1\|^2 = E[\mathbf{y}^T(\mathbf{I} - \mathbf{P}_1)\mathbf{y}] = \text{trace}[(\mathbf{I} - \mathbf{P}_1)\sigma^2\mathbf{I}] + \mu_1^T(\mathbf{I} - \mathbf{P}_1)\mu_1 = (n - p_1)\sigma^2$$

$$E\left[\frac{\|\mathbf{y} - \hat{\mu}\|^2}{n - p_1}\right] = \sigma^2$$

regardless of whether H_0 is true.

Power. The *power* of the F test is defined as:

$$\text{Power} = P(F_{p_1-p_0,n-p_1,\lambda} > F_{p_1-p_0,n-p_1}(0.95))$$

i.e. the probability that the nocentral F rv exceeds the F statistic under the null H_0.

Testing General Linear Hypothesis $H_0 : \Lambda\beta = 0$ for $l \times p$ matrix Λ; l independent constraints on β. Properties include:

- Estimator $\Lambda\hat{\beta}$ is BLUE (Gauss-Markov)
- $\Lambda\hat{\beta} \sim \mathcal{N}[\Lambda\beta, \sigma^2\Lambda(\mathbf{X}^T\mathbf{X})^{-1}\Lambda^T]$
- $(\Lambda\hat{\beta} - 0)^T[\sigma^2\Lambda(\mathbf{X}^T\mathbf{X})^{-1}\Lambda^T]^{-1}(\Lambda\hat{\beta} - 0) \sim \chi^2_l$
- $F = \frac{(\Lambda\hat{\beta})^T[\Lambda(\mathbf{X}^T\mathbf{X})^{-1}\Lambda^T]^{-1}(\Lambda\hat{\beta})/l}{SSE/(n-p)} \sim F_{l,n-p}$ since $SSE/\sigma^2 \sim \chi^2_{n-p}$
- $\Lambda\beta = 0$ is special case M_0 of full model; let $\mathbf{W}$ be matrix s.t. $C(\mathbf{W}) \perp C(\Lambda)$; then $\beta = \mathbf{W}\gamma$, so $E(\mathbf{y}) = \mathbf{X}\beta = \mathbf{X}\mathbf{W}\gamma = \mathbf{X}_0\gamma$ for simpler $\mathbf{X}_0 = \mathbf{X}\mathbf{W}$.

Example: Single Parameter Equals 0. For testing $H_0 : \beta_j = 0$, let $\Lambda = \lambda = (0, 0, \ldots, 0, 1, 0, \ldots, 0)$ in j^{th} slot. This yields:

$$F = \frac{(SSE_0 - SSE_1)/1}{SSE_1/(n-p)} = \frac{\hat{\beta}_j^2}{(SE_j)^2} \sim F_{1,n-p}$$

3.3 Confidence Intervals for Normal Linear Models

Confidence intervals yield more information than significance tests because they provide the entire range of plausible values. We obtain confidence intervals by *inverting significance tests*.

For Parameter Invert test of $H_0 : \beta_j = \beta_{j0}$, yielding test statistic:

$$t = \frac{\hat{\beta}_j - \beta_{j0}}{SE_j} \sim t_{n-p}$$

where $SE_j = \sqrt{[s^2(\mathbf{X}^T\mathbf{X})^{-1}]_{jj}}$ of estimated covariance matrix of $\hat{\beta}$. Residuals uncorrelated with $\hat{\beta}$ since error space/model space, and s^2 function of residuals, so $\hat{\beta} \perp s^2$ and numerator/denominator are independent.

$100(1-\alpha)\%$ CI has p-value $> \alpha$, or $|t| < t_{\alpha/2,n-p}$, so that:

$$\beta_{j0} \in \hat{\beta}_j \pm t_{\alpha/2,n-p}(SE_j)$$

For True Mean To get CI for fitted value (i.e. true mean), note if $\hat{\mu} = \mathbf{x}_0\hat{\beta}$, then $\text{var}(\hat{\mu}) = \text{var}(\mathbf{x}_0\hat{\beta}) = \sigma^2\mathbf{x}_0(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_0^T$ so that when we standardize,

$$z = \frac{\mathbf{x}_0\hat{\beta} - \mathbf{x}_0\beta}{\sigma\sqrt{\mathbf{x}_0(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_0^T}} \sim \mathcal{N}(0,1)$$

$$\Rightarrow t = \frac{\mathbf{x}_0\hat{\beta} - \mathbf{x}_0\beta}{s\sqrt{\mathbf{x}_0(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_0^T}} \sim t_{n-p}$$

since $(n-p)s^2/\sigma^2 \sim \chi^2_{n-p}$ by Cochran. The resulting CI for μ is:

$$\mu \in \mathbf{x}_0\hat{\beta} \pm t_{\alpha/2,n-p}s\sqrt{\mathbf{x}_0(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_0^T}$$

Note if $\mathbf{x}_0 = \mathbf{x}_i$ for some obs i, then the square root term is just h_{ii}.

For Future Prediction At given $\mathbf{x}_0$, suppose predict future y; $y = \mathbf{x}_0\beta + \epsilon$, $\epsilon \sim \mathcal{N}(0, \sigma^2)$. From fitting, $y = \mathbf{x}_0\hat{\beta} + e$ where $e = y - \hat{\mu}$, so that:

$$\text{var}(e) = \text{var}(y - \hat{\mu}) = \text{var}(y) + \text{var}(\hat{\mu}) = \sigma^2(1 + \mathbf{x}_0(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_0^T)$$

since $y \perp y_1, \ldots, y_n$ used for $\hat{\mu}$. Thus:

$$\frac{y - \hat{\mu}}{s\sqrt{1 + \mathbf{x}_0(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_0^T}} \sim t_{n-p}$$

so the $100(1-\alpha)\%$ *prediction interval* is:

$$y \in \hat{\mu} \pm t_{\alpha/2,n-p}s\sqrt{1 + \mathbf{x}_0(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_0^T}$$

4 Generalized Linear Models: Fitting and Inference

Generalized Linear Model: 1) Non-normal $\mathbf{y}$; 2) Non-identity g.

4.1 Exponential Dispersion Family

Properties For y_i from EDF:

- PDF: $f(y_i;\theta_i,\phi) = \exp\left[\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} - c(y_i,\phi)\right]$
- θ_i is natural parameter; ϕ is dispersion parameter
- Generally, $a(\phi)=1$ (natural exponential family); $a(\phi)\phi/w_i$ for weight w_i known (i.e. binomial)
- $\mu_i = E(y_i) = b'(\theta_i)$ and $\text{var}(y_i) = b''(\theta_i)a(\phi)$ (using exp. score = 0 and second partials of l results)

Poisson, Binomial, Normal, Gamma All in EDF:

- Poisson: $f(y_i;\mu_i) = \frac{\mu_i^{y_i}e^{-\mu_i}}{y_i!} = \exp[y_i\log\mu_i - \mu_i - \log(y_i!)]$ so we have:

$$\theta_i = \log(\mu_i), b(\theta_i) = \exp(\theta_i), a(\phi) = 1$$

- Binomial: Let $n_iy_i \sim \text{Bin}(n_i,\pi_i)$ so y_i is sample proportion.

$$f(y_i;n_i,\pi_i) = \binom{n_i}{n_iy_i}\pi_i^{n_iy_i}(1-\pi_i)^{n_i-n_iy_i} = \exp\left[\frac{y_i\theta_i - \log(1-\exp(\theta_i))}{1/n_i} + \log\binom{n_i}{n_iy_i}\right]$$

where $\theta_i = \log[\pi_i/(1-\pi_i)] = \text{logit}(\pi_i$ and $b(\theta_i) = \log[1+\exp(\theta_i)]$, $a(\phi) = 1/n_i$

- Normal: $f(y_i;\mu_i,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{(y_i-\mu_i)^2}{2\sigma^2}\right] = \exp\left[\frac{y_i\mu_i - \mu_i^2/2}{\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2) - \frac{y_i^2}{2\sigma^2}\right]$:

$$\theta_i = \mu_i, b(\theta_i) = \frac{1}{2}\theta_i^2, a(\phi) = \sigma^2$$

- Gamma: $f(y;\mu,k) = \frac{(k/\mu)^k}{\Gamma(k)}y^{k-1}e^{-ky/\mu}$ with $E(y)=\mu$ and $\text{var}(y) = \mu^2/k$

$$\theta = -\frac{1}{\mu}, b(\theta) = -\log(-\theta), \phi = \frac{1}{k}$$

Canonical Link $g:\mu_i \mapsto \theta_i$ results in direct relationship $\theta_i = \eta_i = \sum_j \beta_j x_{ij}$ (good things: Newton-Raphson = Fisher scoring, always concave, sufficient statistics = expected values)

4.2 Likelihood Equations and Asymptotics

Sufficient Statistics $l(\beta) = \sum_i l_i = \sum_i \frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + \sum_i c(y_i,\phi)$. When g is canonical link, $\theta_i = \sum_j \beta_j x_{ij}$, so when $a(\phi)$ is constant, the kernel is:

$$\sum_i y_i(\sum_j \beta_j x_{ij}) = \sum_j \beta_j(\sum_i y_i x_{ij})$$

so the sufficient statistics are $\sum_i y_i x_{ij}$ for all $j = 1,\ldots,p$

Likelihood Equations For ML, want $\frac{\partial l(\beta)}{\partial\beta_j} = 0$ for all j; using chain rule:

$$\frac{\partial l_i}{\partial\beta_j} = \frac{\partial l_i}{\partial\theta_i}\frac{\partial\theta_i}{\partial\mu_i}\frac{\partial\mu_i}{\partial\eta_i}\frac{\partial\eta_i}{\partial\beta_j}$$

$$\frac{\partial l_i}{\partial\theta_i} = \frac{y_i-\mu_i}{a(\phi)}, \frac{\partial\mu_i}{\partial\theta_i} = b''(\theta_i) = \frac{\text{var}(y_i)}{a(\phi)}, \frac{\partial\eta_i}{\partial\beta_j} = x_{ij}$$

$$\Rightarrow \frac{\partial l(\beta)}{\partial\beta_j} = \sum_i \frac{\partial l_i}{\partial\beta_j} = \boxed{\sum_i \frac{(y_i-\mu_i)x_{ij}}{\text{var}(y_i)}\frac{\partial\mu_i}{\partial\eta_i} = 0}$$

Let $\mathbf{D} = \text{diag}\left(\frac{\partial\mu_i}{\partial\eta_i}\right)$, and $\mathbf{V}$ be covariance matrix. Then:

$$\mathbf{X}^T\mathbf{D}\mathbf{V}^{-1}(\mathbf{y}-\mu) = 0$$

Mean-Variance Relation If y_i in EDF, then relation between mean and variance $\text{var}(y_i) = v(\mu_i)$ completely determines distribution.

- Poisson: $v(\mu_i) = \mu_i$
- Binomial: $v(\mu_i) = \frac{\mu_i(1-\mu_i)}{n_i}$
- Normal: $v(\mu_i) = \sigma^2$ (constant)
- Gamma: $v(\mu_i) = \frac{\mu_i^2}{k}$

Asymptotics of Parameter Estimators By ML properties, for large n $\hat{\beta}$ is: 1) efficient; 2) approximately Normal. Moreover, covariance matrix of $\hat{\beta}$ is $\text{var}(\hat{\beta}) = \mathcal{J}^{-1}$, the Fisher information matrix:

$$\mathcal{J} = \left(-E\left[\frac{\partial^2 l(\beta)}{\partial\beta_i \partial\beta_j}\right]\right)$$

Using the ML second derivative result,

$$-E\left(\frac{\partial^2 l_i}{\partial\beta_j\partial\beta_k}\right) = E\left[\left(\frac{\partial l_i}{\partial\beta_j}\right)\left(\frac{\partial l_i}{\partial\beta_k}\right)\right] = \frac{x_{ij}x_{ik}}{\text{var}(y_i)}\left(\frac{\partial\mu_i}{\partial\eta_i}\right)^2$$

$$\Rightarrow -E\left[\frac{\partial^2 l(\beta)}{\partial\beta_i\partial\beta_j}\right] = \sum_i \frac{x_{ij}x_{ik}}{\text{var}(y_i)}\left(\frac{\partial\mu_i}{\partial\eta_i}\right)^2$$

so let $\mathbf{W} = \text{diag}\left(\frac{(\partial\mu_i/\partial\eta_i)^2}{\text{var}(y_i)}\right)$, then we have: $\mathcal{J} = \mathbf{X}^T\mathbf{W}\mathbf{X}$

$$\boxed{\hat{\beta} \sim \mathcal{N}[\beta, (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}]}$$

Asymptotics of Fitted Values Note that $\hat{\eta} = \mathbf{X}\hat{\beta} \Rightarrow \text{var}(\hat{\eta}) = \mathbf{X}\text{var}(\hat{\beta})\mathbf{X}^T \approx \mathbf{X}(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T$. We want $\text{var}(\hat{\mu}$, and we can use delta method:

$$h(y) - h(\mu) \approx h'(\mu)(y-\mu) \Rightarrow \text{var}[h(y)] \approx [h'(\mu)]^2\text{var}(y)$$

In the vector vase, $\text{var}[\mathbf{h}(\mathbf{y})] \approx \left(\frac{\partial\mathbf{h}}{\partial\mu}\right)\mathbf{V}\left(\frac{\partial\mathbf{h}}{\partial\mu}\right)^T$ for the Jacobian $\left(\frac{\partial\mathbf{h}}{\partial\mu}\right)$. So using $\mathbf{D} = \text{diag}(\partial\mu_i/\partial\eta_i)$:

$$\text{var}(\hat{\mu}) \approx \mathbf{D}\mathbf{X}(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{D}$$

Model Misspecification Even if we specified wrong distribution for $\mathbf{y}$, as long as we used EDF: $\hat{\beta} \xrightarrow{p} \beta$ as long as linear predictor and link are correct.

4.3 GLM Parameter Inference: LRT, Wald, Score

In order to: 1) say if a parameter estimate is significantly non-zero; 2) establish confidence intervals for the true parameters, we need tests of significance. There are three standard methods:

Likelihood-Ratio Test Let $H_0 : \beta_j = 0$. Then define $l_0 = \max_\beta l(\beta)|_{\beta_j=0}$ and $l_1 = \max_\beta l(\beta)$. Then as $n \to \infty$:

$$\boxed{-2(l_0 - l_1) \sim \chi_1^2}$$

This can be extended to multiple parameters $\beta = (\beta_0, \beta_1)$ and $H_0 : \beta_0 = 0$ leads to $\chi^2_{|\beta_0|}$ and general linear hypothesis $H_0 : \Lambda\beta = 0$ leads to χ^2_l where Λ adds l constraints.

Wald Test Recall: $SE_{\hat{\beta}} \approx \sqrt{(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}}$ so estimating that using: $\hat{SE}_{\hat{\beta}} = \sqrt{(\mathbf{X}^T\hat{\mathbf{W}}\mathbf{X})^{-1}}$ where $\hat{\mathbf{W}}$ is $\mathbf{W} = \frac{(\partial\mu_i)/\partial\eta_i)^2}{\text{var}(y_i)}$ evaluated at $\hat{\eta}_i = \sum_j \hat{\beta}_j x_{ij}$. To test $H_0 : \beta_j = \beta_{j0}$, using $\hat{SE}_j = (\hat{SE}_{\hat{\beta}})_{jj}$:

$$\boxed{z = \frac{\hat{\beta}_j - \beta_{j0}}{\hat{SE}_j} \sim \mathcal{N}(0,1)}$$

$$z^2 \sim \chi_1^2$$

For multiple parameters $\beta = (\beta_0, \beta_1)$, testing $H_0 : \beta_0 = 0$:

$$z^2 = \hat{\beta}_0^T [\hat{\text{var}}(\hat{\beta})]_{\beta_0}^{-1} \hat{\beta}_0 \sim \chi^2_{|\beta_0|}$$

where $[\hat{\text{var}}(\hat{\beta})]_{\beta_0} = (\mathbf{X}^T \hat{\mathbf{W}} \mathbf{X})$ using only the rows/columns corresponding to β_0.

Problems: 1) Useless at boundary; 2) Depends on scale

Score Test Testing $H_0 : \beta = \beta_0$:

$$\boxed{z^2 = \frac{[\partial l(\beta)/\partial \beta_0]^2}{-E[\partial^2 l(\beta)/\partial \beta_0^2]} \sim \chi_1^2}$$

where the derivatives are evaluated at $\beta = \beta_0$.

Confidence Intervals We again get CI by inverting the test.

- Likelihood-Ratio Test: For $H_0 : \beta = \beta_0$: $\beta_0 \in \{\beta : -2[l(\beta) - l(\hat{\beta})] > \chi_1^2(\alpha)\}$
- Wald Test: $\frac{|\hat{\beta} - \beta_0|}{SE} < z_{\alpha/2} \Rightarrow \beta_0 \in \hat{\beta} \pm z_{\alpha/2}(SE)$
- Score Test: Depends on likelihood; generally close to Wald interval

When n small or $\hat{\beta}$ very non-normal (i.e. Wald and LRT CI differ greatly) then Wald fails, so *use LRT*.

Profile Likelihood For multiparameter models, i.e. $\beta = (\beta_0, \psi)$, best CI is obtained by maximizing $l(\beta)$ at each possible value of β_0. That is: 1) plug in β_0 into $l(\beta)$; 2) maximize $l(\beta)$ over all other ψ, yielding maximum nuisance parameters $\hat{\psi}(\beta_0)$; 3) use the *profile log-likelihood function* $l(\beta_0, \hat{\psi}(\beta_0))$. The *profile likelihood CI* for true β_0 is:

$$-2[l(\beta_0, \hat{\psi}(\beta_0)) - l(\hat{\beta}_0, \hat{\psi})] < \chi_1^2(\alpha)$$

4.4 Deviance and Model Checking/Comparison

For normal linear models, we used Cochran's Theorem and F statistics to tell whether model fit well (nested models). Can't do that for GLMs, so we use deviance (LRT).

Deviance Compare log-likelihood of model with saturated model; let $l(\mu; \mathbf{y})$ be log-likelihood in terms of $\mu = g^{-1}(\theta)$, then $l(\hat{\mu}; \mathbf{y})$ is maximum of log-likelihood under model, $l(\mathbf{y}; \mathbf{y})$ is log-likelihood under saturated model (separate parameter for each obs $\tilde{\mu} = \mathbf{y}$).

Likelihood-ratio statistic: $-2[l(\hat{\mu}; \mathbf{y}) - l(\mathbf{y}; \mathbf{y})] = 2 \sum_i \frac{y_i(\tilde{\theta} - \hat{\theta}) - b(\tilde{\theta}) + b(\hat{\theta})}{a(\phi)}$

Generally, $a(\phi) = \phi / w_i$, so then:

Deviance $\boxed{D(\mathbf{y}; \hat{\mu}) = 2 \sum_i w_i [y_i(\tilde{\theta} - \hat{\theta}) - b(\tilde{\theta}) + b(\hat{\theta})]}$

and: $-2[l(\hat{\mu}; \mathbf{y}) - l(\mathbf{y}; \mathbf{y})] = \frac{D(\mathbf{y}; \hat{\mu})}{\phi}$ (so LRT statistic = scaled deviance)

- Poisson GLM: Using canonical link, $\hat{\theta}_i = \log(\hat{\mu}_i)$ and $b(\theta_i) = \exp(\theta_i)$, with $w_i = 1$ so:

  $$D(\mathbf{y}; \hat{\mu}) = 2 \sum_i [y_i \log(y_i / \hat{\mu}_i) - y_i + \hat{\mu}_i]$$

  If there is intercept term, likelihood equations yield $\sum_i y_i = \sum_i \hat{\mu}_i$:

  $$D(\mathbf{y}; \hat{\mu}) = 2 \sum_i y_i \log(y_i / \hat{\mu}_i)$$

- Normal GLM: $D(\mathbf{y}; \hat{\mu}) = 2 \sum_i \left[y_i(y_i - \hat{\mu}_i) - \frac{\mathbf{y}_i^2}{2} + \frac{\hat{\mu}_i^2}{2} \right] = \sum_i (y_i - \hat{\mu}_i)^2 = SSE$

$$\boxed{\text{Maximize likelihood} \Leftrightarrow \text{Minimize deviance}}$$

Model Comparison In normal linear models, we used SSE comparisons to compare models. Generalize to GLMS:

1. **Likelihood-Ratio Test**: Suppose M_0 nested in M_1, so $l(\hat{\mu}_1; \mathbf{y}) \geq l(\hat{\mu}_0; \mathbf{y})$. Consider likelihood-ratio test of $H_0 : M_0$ holds:

$$-2[l(\hat{\mu}_0; \mathbf{y}) - l(\hat{\mu}_1; \mathbf{y})] = -2[l(\hat{\mu}_0; \mathbf{y}) - l(\mathbf{y}; \mathbf{y})] + 2[l(\hat{\mu}_1; \mathbf{y}) - l(\mathbf{y}; \mathbf{y})] = D(\mathbf{y}; \hat{\mu}_0) - D(\mathbf{y}; \hat{\mu}_1)$$

if $\phi = 1$, as in Poisson/Binomial, which has deviance form, so:

$$G^2(M_0|M_1) = D(\mathbf{y}; \hat{\mu}_0) - D(\mathbf{y}; \hat{\mu}_1) = 2\sum_i w_i[y_i(\hat{\theta}_{1i} - \hat{\theta}_{0i}) - b(\hat{\theta}_{1i} + b(\hat{\theta}_{0i})]$$

$$\boxed{G^2(M_0|M_1) = D(\mathbf{y}; \hat{\mu}_0) - D(\mathbf{y}; \hat{\mu}_1) \sim \chi_{p_1 - p_0}}$$

under the null hypothesis (M_0 holds)
Using the fact that deviance $\approx$ LRT statistic so $D(\mathbf{y}; \hat{\mu}_1) \sim \chi^2_{n-p_1}$, we have:

$$\frac{[D(M_0) - D(M_1)]/(p_1 - p_0)}{D(M_1)/(n - p_1)} \sim F_{p_1 - p_0, n - p_1}$$

2. **Score/Pearson Statistics**: For GLM with $\text{var}(y_i) = v(\mu_i)$ and $\phi = 1$:

$$X^2 = \sum_i \frac{(y_i - \hat{\mu}_i)^2}{v(\hat{\mu})}$$

This is the *generalized Pearson chi-squared statistic*; original was $X^2 = \sum_i (\text{obs}-\text{fitted})^2/\text{fitted}$ which holds when GLM is Poisson ($v(\hat{\mu}) = \hat{\mu}$). For testing nested M_0 in M_1:

$$\boxed{X^2(M_0|M_1) = \sum_i \frac{(\hat{\mu}_{1i} - \hat{\mu}_{0i})^2}{v(\hat{\mu}_{0i})} \sim \chi^2_{p_1 - p_0}}$$

which is quadratic approximation to $G(M_0|M_1)$, the deviance statistic. Often has better behavior asymptotically.

Asymptotics of Residuals Unlike in LM case where $\mathbf{y} = \hat{\mu} + (\mathbf{y} - \hat{\mu})$ yielded orthogonal decomposition, in GLM Case, $\mu = g^{-1}(\eta)$ need not constitute vector space, so projections/orthogonality don't hold. We suppose that $\hat{\mu}$ and residuals are asymptotically uncorrelated. Using $\mathbf{W}$ and $\mathbf{D}$ as before, we have: $\mathbf{V} = \text{var}(\mathbf{y}) = \mathbf{D}\mathbf{W}^{-1}\mathbf{D}$, and $\text{var}(\mathbf{y}) \approx \text{var}(\hat{\mu}) + \text{var}(\mathbf{y} - \hat{\mu})$ under asymptotic uncorrelatedness. Thus,

$$\text{var}(\mathbf{y} - \hat{\mu}) \approx \mathbf{V} - \text{var}(\hat{\mu}) \approx \mathbf{D}\mathbf{W}^{-1}\mathbf{D} - \mathbf{D}\mathbf{X}(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{D}$$

$$\Rightarrow \text{var}(\mathbf{y} - \hat{\mu}) \approx \mathbf{D}\mathbf{W}^{-1/2}[\mathbf{I} - \mathbf{W}^{1/2}\mathbf{X}(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}^{1/2}]\mathbf{W}^{-1/2}\mathbf{D} = \mathbf{V}^{1/2}[\mathbf{I} - \mathbf{H}_W]\mathbf{V}^{1/2}$$

where $\mathbf{H}_W = \mathbf{W}^{1/2}\mathbf{X}(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}^{1/2}$ is projection matrix (hat matrix) for $V^{-1/2}(\mathbf{y} - \mu)$.

Pearson, Deviance, Standardized Residuals Three kinds of residuals for GLMS:

1. **Pearson residual** $\boxed{e_i = \frac{y_i - \hat{\mu}_i}{\sqrt{v(\hat{\mu}_i)}}}$

   Note that: $X^2 = \sum_i e_i^2 \sim \chi^2_1$ for Poisson and Binomial; for Poisson, $e_i = (y_i - \hat{\mu}_i)/\sqrt{\hat{\mu}_i}$, whereas for Binomial, $e_i = (y_i - \hat{\pi}_i)/\sqrt{\hat{\pi}_i(1 - \hat{\pi}_i)/n_i}$.

2. **Deviance residual** $d_i = 2w_i[y_i(\tilde{\theta}_i - \hat{\theta}_i) - b(\tilde{\theta}_i) + b(\hat{\theta}_i)]$ so that $D(\mathbf{y}; \hat{\mu}) = \sum_i d_i$. Then:

   Deviance residual: $\boxed{\sqrt{d_i} \times \text{sign}(y_i - \hat{\mu}_i)}$

3. **Standardized residual**: Pearson/deviance residuals have variance < 1 because compare y_i to $\hat{\mu}_i$ rather than μ_i. Using generalized hat matrix $\mathbf{H}_W = \mathbf{W}^{1/2}\mathbf{X}(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}^{1/2}$ and $\hat{h}_{ii} = (\hat{H}_W)_{ii}$, we have:

   Standardized residual: $\boxed{r_i = \frac{e_i}{\sqrt{1 - \hat{h}_{ii}}}}$

4.5 GLM Fitting

Unlike normal equations, likelihood equations are nonlinear in $\hat{\beta}$, so need iterative schemes.

Newton-Raphson Use quadratic approximations to iterate solution to maximum:

$$\mathbf{u} = \left(\frac{\partial l(\beta)}{\partial \beta_1}, \ldots, \frac{\partial l(\beta)}{\partial \beta_p}\right)$$

$$\mathbf{H} = \left(\frac{\partial^2 l(\beta)}{\partial \beta_i \partial \beta_j}\right)$$

where $\mathbf{H}$ is the Hessian matrix, or observed information. Let $\mathbf{u}^{(t)}, \mathbf{H}^{(t)}$ be score/Hessian evaluated at $\beta^{(t)}$. Using Taylor:

$$l(\beta) \approx l(\beta^{(t)}) + (\mathbf{u}^{(t)})^T(\beta - \beta^{(t)}) + \frac{1}{2}(\beta - \beta^{(t)})^T \mathbf{H}^{(t)}(\beta - \beta^{(t)}) \Rightarrow \frac{\partial l(\beta)}{\partial \beta} \approx \mathbf{u}^{(t)} + \mathbf{H}^{(t)}(\beta - \beta^{(t)}) = 0$$

$$\Rightarrow \boxed{\beta^{(t+1)} = \beta^{(t)} - (\mathbf{H}^{(t)})^{-1}\mathbf{u}^{(t)}}$$

Fisher Scoring Uses expected information, not observed information. Recall:

$$\mathcal{J} = -E\left[\frac{\partial^2 l(\beta)}{\partial \beta_i \partial \beta_j}\right] = \mathbf{X}^T\mathbf{W}\mathbf{X}$$

so let $\mathcal{J}^{(t)}$ be $\mathcal{J}$ evaluated at $\beta^{(t)}$; $\mathcal{J}^{(t)} = \mathbf{X}^T\mathbf{W}^{(t)}\mathbf{W}$. Equivalently to Newton-Raphson:

$$\boxed{\beta^{(t+1)} = \beta^{(t)} + (\mathcal{J}^{(t)})^{-1}\mathbf{u}^{(t)}}$$

Example: Binomial Parameter. Consider single set of binomial observation, $ny \sim \text{Bin}(n, \pi)$ and consider estimating the maximum parameter $\hat{\pi}$ (rather than β, as usual). Then $l(\pi) = ny \log \pi + (n - ny)\log(1 - \pi) + \log\binom{n}{ny}$. Thus, the derivatives are: $u = \frac{\partial l(\pi)}{\partial \pi} = \frac{ny - n\pi}{\pi(1-\pi)}$ and $H = -\left[\frac{ny}{\pi^2} + \frac{n-ny}{(1-\pi)^2}\right] \Rightarrow E[H] = \frac{n}{\pi(1-\pi)}$ So we can use:

1. Newton-Raphson: $\pi^{(t+1)} = \pi^{(t)} - (H^{(t)})^{-1}u^{(t)}$, which does do the right thing
2. Fisher Scoring: $\pi^{(t+1)} = \pi^{(t)} + \left[\frac{n}{\pi^{(t)}(1-\pi^{(t)})}\right]^{-1} \frac{ny - n\pi^{(t)}}{\pi^{(t)}(1-\pi^{(t)})} = \pi^{(t)} + (y - \pi^{(t)}) = y$ so achieved in one step.

Fisher Scoring = IRLS Fisher scoring is equivalent to iteratively reweighted least squares on the adjusted response, $z_i = \sum)jx_{ij}\beta_j^{(t)} + (y_i - \mu_i^{(t)})\frac{\partial \eta_i^{(t)}}{\partial \mu_i^{(t)}} = \eta_i^{(t)} + (y_i - \mu_i^{(t)})\frac{\partial \eta_i^{(t)}}{\partial \mu_i^{(t)}}$. For the linear model $\mathbf{z} = \mathbf{X}\beta + \epsilon$, with ϵ covariance $\mathbf{V}$, the generalized LS estmator is: $\hat{\beta} = (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{z}$. The score vector is $\mathbf{u} = \mathbf{X}^T\mathbf{D}\mathbf{V}^{-1}(\mathbf{y} - \mu)$, and we see that $\mathbf{D}\mathbf{V}^{-1} = \mathbf{W}\mathbf{D}^{-1}$ for diagonal $\mathbf{V}$. Thus, $\mathbf{u} = \mathbf{X}^T\mathbf{W}\mathbf{D}^{-1}(\mathbf{y} - \mu)$, and the Fisher scoring equations are: $\mathcal{J}^{(t)}\beta^{(t+1)} = \mathcal{J}^{(t)}\beta^{(t)} + \mathbf{u}^{(t)}$. Thus,

$$\jmath^{(t)}\beta^{(t)} = \mathbf{X}^T\mathbf{W}^{(t)}\mathbf{X}\beta^{(t)} + \mathbf{x}^T\mathbf{W}^{(t)}(\mathbf{D}^{(t)})^{-1}(\mathbf{y} - \mu^{(t)}) = \mathbf{X}^T\mathbf{W}^{(t)}[\mathbf{X}\beta^{(t)} + (\mathbf{D}^{(t)})^{-1}(\mathbf{y} - \mu^{(t)})] = \mathbf{X}^T\mathbf{W}^{(t)}\mathbf{z}^{(t)}$$

and $J^{(t)}\beta^{(t+1)} = \mathbf{X}^T\mathbf{W}^{(t)}\mathbf{W}\beta^{(t+1)}$ so that:

$$\beta^{(t+1)} = (\mathbf{X}^T\mathbf{W}^{(t)}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}^{(t)}\mathbf{z}^{(t)}$$

Equivalence for Canonical Link For canonical link $\theta_i = \eta_i$, we have: $\partial \mu_i / \partial \eta_i = b''(\theta_i)$, so $\frac{\partial l_i}{\partial \beta_j} = \frac{(y_i - \mu_i)x_{ij}}{a(\phi)} \Rightarrow \frac{\partial^2 l_i}{\partial \beta_j \partial \beta_k} = -\frac{x_{ij}}{a(\phi)}\left(\frac{\partial \mu_i}{\partial \beta_k}\right)$ which is independent of y_i, so:

$$\boxed{\mathbf{H} = -\mathcal{J}}$$

and so Newton-Raphson = Fisher scoring for GLMs with canonical link.

4.6 Model/Variable Selection

Stepwise Procedures Forward selection vs. backward elimination

Bias-Variance Tradeoff MSE = variance + $(\text{bias})^2$ so simpler model has higher bias, but may have lower variance $\Rightarrow$ lower overall MSE.

AIC Kullback-Leibler divergence: $KL[p, p_M(\hat{\beta}_M)] = E\left[\log\left(\frac{p(\mathbf{y}^*)}{p_M(\mathbf{y}^*;\hat{\beta}_M)}\right)\right]$ measures distance between true distribution $p(\cdot)$ and model fitted distribution $p_M(\cdot;\hat{\beta}_M)$

AIC: minimize $E[KL(p, p_M(\hat{\beta}_M))] \Leftrightarrow \min E[-E\log(p_M(\mathbf{y};\hat{\beta}_M))]$ where outer with respect to set of models, inner with respect to p. $l(\hat{\beta}_M)$ is biased estimator for $E[E\log(p_M(\mathbf{y};\hat{\beta}_M))]$ but can be reduced using number of parameters in M. Thus:

$$\boxed{\text{AIC} = -2[l(\hat{\beta}) + |M|]}$$

where $|M|$ is the number of parameters in model M.

Predictive Power Two measures of summarizing predictive power (i.e. R^2 in linear models):

1. $\text{corr}(\mathbf{y}, \hat{\mu})$: analog of multiple correlation (but not necessarily non-decreasing with more parameters)
2. Likelihood Ratio: let l_M be maximized log-likelihood for model M; l_S for saturated; l_0 for null model, then:
$$\frac{l_M - l_0}{l_S - l_0} \in [0, 1]$$

Collinearity Relations among explanatory variables may reduce validity and effects:

$$\text{var}(\hat{\beta}_j) = \frac{1}{1 - R_j^2}\left[\frac{\sigma^2}{\sum_i (x_{ij} - \bar{x})^{@}}\right]$$

where R_j^2 is R^2 in predicting x_j using x_{-j} and $VIF_j = \frac{1}{1-R_j^2}$ is variance inflation factor. (So as variables are collinear, R_j^2 goes up and $\text{var}(\hat{\beta}_j) \to \infty$.)

5 Binary Models

For binary response, assume $n_i y_i \sim \text{Bin}(n_i, \pi_i)$. Two sample sizes: 1) n_i is number of Bern trials in single binomial obs; 2) N is number of binomial obs. Let $\mathbf{n} = (n_1, \ldots, n_N)$ be samples sizes, $n = \sum_i n_i$ overall Bern obs.

Two data types: 1) *ungrouped data* has $\mathbf{n} = (1, \ldots, 1)$ and large-sample asymptotics $= N \to \infty$; 2) *grouped data* has $n_i > 1$ with (usually) categorical variables, same values in a group, and small-dispersion asymptotics $= n_i \to \infty$ with N constant.

Same estimates $\hat{\beta}$ and SE for grouped/ungrouped, but deviance changes (different saturated model).

5.1 Link Functions

Latent Variable Model Threshold model with ungrouped data: 1) $\exists$ unobserved continuous y_i^* s.t. $y_i^* = \sum_j \beta_j x_{ij} + \epsilon_i$; 2) ϵ_i has mean 0, CDF F; 3) threshold τ s.t. $y_i = 0$ if $y_i^* \leq \tau$ and $y_i = 1$ if $y_i^* > \tau$. Then:

$$\begin{aligned} P(y_i = 1) = P(y_i^* > \tau) &= P\left(\sum_j \beta_j x_{ij} + \epsilon_i > \tau\right) \\ &= 1 - P\left(\epsilon_i \leq \tau - \sum_j \beta_j x_{ij}\right) \\ &= 1 - F\left(\tau - \sum_j \beta_j x_{ij}\right) \end{aligned}$$

since data doesn't indicate what τ is, can take $\tau = 0$ WLOG, and can use standard F (since multiply all parameters by constant). Generally F is symmetric about 0, so $F(z) = 1 - F(-z)$ and:

$$P(y_i = 1) = F\left(\sum_b \beta_j x_{ij}\right) \Rightarrow \boxed{F^{-1}[P(y_i = 1)] = \sum_{j=1}^{p} \beta_j x_{ij}}$$

so the link function corresponds to inverse CDF for some latent distribution.

Link Functions/Models Possible link functions are:

1. Probit: $F = \Phi$ so $\Phi^{-1}[P(y_i = 1)] = \sum_j \beta_j x_{ij}$
2. Logit: $F(z) = \frac{e^z}{1+e^z}$ is logistic distribution, so $F^{-1} = \text{logit}$ and $\text{logit}[P(y_i = 1)] = \sum_j \beta_j x_{ij}$
3. Log-Log: $F(z) = \exp[-\exp(-(x-a)/b)]$ (Type I extreme-value distribution) so that: $-\log[-\log P(y_i = 1)] = \sum_j \beta_j x_{ij}$

5.2 Logistic Regression: Interpretation

$$\boxed{\pi_i = \frac{\exp(\sum_j \beta_j x_{ij})}{1 + \exp(\sum_j \beta_j x_{ij})}}$$

$$\boxed{\text{logit}(\pi_i) = \sum_j \beta_j x_{ij}}$$

Interpreting β Interpretations depending on quantitative/qualitative:

- Quantitative x: $\frac{\partial \pi_i}{\partial x_{ij}} = \beta_j \frac{\exp(\sum_j \beta_j x_{ij})}{1+\exp(\sum_j \beta_j x_{ij})} = \beta_j \pi_j (1 - \pi_j)$ so that at steepest, $\pi_i = 1/2$:

$$\frac{\partial \pi_i}{\partial x_{ij}} = \frac{\beta_j}{4}$$

- Qualitative x: Let x be binary indicator, $\text{logit}(\pi_i) = \beta_0 + \beta_1 x$ (2×2 contingency table). Then $\text{logit}[P(y=1|x=1)] - \text{logit}[P(y=1|x=0)] = \beta_1$ so that e^{β_1} is odds ratio:

$$e^{\beta_1} = \frac{P(y=1|x=1)/[1-P(y=1|x=1)]}{P(y=1|x=0)/[1-P(y=1|x=0)]}$$

  If there are multiple variables, odds of $P(y=1)$ multiply by e^{β_j} for unit increase in x_j:

$$e^{\beta_j} = \frac{P(y=1|x_j=u+1)/[1-P(y=1|x_j=u+1)]}{P(y=1|x_j=u)/[1-P(y=1|x_j=u)]}$$

Case-Control Studies Retrospective studies fine for logistic regression since:

$$e^{\beta} = \frac{P(y=1|x=1)/P(y=0|x=1)}{P(y=1|x=0)/P(y=0|x=0)} = \frac{P(x=1|y=1)/P(x=0|y=1)}{P(x=1|y=0)/P(x=0|x=0)}$$

i.e. we can reverse response/explanatory and still get odds ratio interpretation.

Predictive Power Two main ways to summarize predictive power:

1. Classification table: cross-classify y with prediction $\hat{y}$; i.e. use $\hat{y}_i = 1$ if $\hat{\pi}_i > \pi_0$ and $\hat{y}_i = 0$ otherwise (i.e. $pi_0 = 0.5$, $\pi_0 = \bar{y}$). Then:

$$\text{sensitivity} = P(\hat{y}=1|y=1) \text{ and specificity} = P(\hat{y}=0|y=0)$$

   but depends strongly on cutoff π_0.
2. ROC curve: Let tpr = sensitivity and fpr = $1 -$ specificity.
   ROC curve = plot tpr (y) as function of fpr (x); generally concave
   If $pi_0 \approx 1$ then tpr = fpr = 0; If $\pi_0 \approx 0$ then tpr = fpr = 1.
   Concordance index = area under ROC curve = proportion of all pairs (i,j) such that $y_i = 1, y_j = 0$ and $\hat{\pi}_i > \hat{\pi}_j$.
3. Correlation measure: $\text{corr}(\mathbf{y}, \hat{\mu})$ is useless because $\mathbf{y}$ is 0 or 1. Better measure is $\text{corr}(\mathbf{y}^*, \hat{\mu})$, i.e. $\mathbf{y}^* = \mu + \epsilon$ and $\hat{\mu} = \sum_j \beta_j x_{ij}$.

5.3 Logistic Regression: Inference

Use likelihood equations and Newton-Raphson/Fisher Scoring, like other GLMs:

$$\sum_{i=1}^{N} \frac{(y_i - \hat{\mu}_i)x_{ij}}{\text{var}(y_i)} \frac{\partial \mu_i}{\partial \eta_i} = \sum_{i=1}^{N} \frac{n_i(y_i - \pi_i)x_{ij}}{\pi_i(1-\pi_i)} f(\eta_i) = 0$$

since $\mu_i = F(\eta_i)$ for CDF F resulting in PDF f. In terms of β:

$$\sum_{i=1}^{N} \frac{n_i(y_i - F(\sum_j \beta_j x_{ij}))x_{ij} f(\sum_j \beta_j x_{ij})}{F(\sum_j \beta_j x_{ij})[1 - F(\sum_j \beta_j x_{ij})]} = 0$$

Likelihood Equations For logistic regression: $F(z) = \frac{e^z}{1+e^z}$, $f(z) = \frac{e^z}{(1+e^z)^2} = F(z)[1-F(z)]$ so:

$$\boxed{\sum_{i=1}^{N} n_i(y_i - \pi_i)x_{ij} = 0}$$

and if $\mathbf{X}$ is the $N \times p$ model matrix, with totals $s_i = n_i y_i$, then:

$$\mathbf{X}^T \mathbf{s} = \mathbf{X}^T E(\mathbf{s})$$

i.e. as with all canonical link: sufficient statistic = expected value.

Asymptotic Covariance Matrix of Estimators $\mathcal{J} = \mathbf{X}^T \mathbf{W} \mathbf{X}$, and $w_i = \frac{(\partial \mu_i / \partial \eta_i)^2}{\text{var}(y_i)} = n_i \pi_i (1-\pi_i)$ so the estimated covariance matrix for large samples is:

$$\boxed{\hat{\text{var}}(\hat{\beta}) = (\mathbf{X}^T \hat{\mathbf{W}} \mathbf{X})^{-1} = (\mathbf{X}^T \text{diag}[n_i \hat{\pi}_i (1-\hat{\pi}_i)] \mathbf{X})^{-1}}$$

Wald is Suboptimal 1) Scale-dependent; 2) Aberrant behavior when effect is large.

For null model $\text{logit}(\pi) = \beta_0$, and $H_0 : \beta_0 = 0$, then on totals scale, $z^2 = \text{logit}(y)^2[ny(1-y)]$ while on proportion scale, $z^2 = \frac{(y-0.5)^2}{y(1-y)/n}$ which are different.

Fisher Exact Test Used when n is small relative to p; eliminate nuisance parameters by conditioning on their sufficient statistics. Consider logistic regression with single binary x and small N, ungrouped: $\text{logit}[P(y_i = 1)] = \beta_0 + \beta_1 x_i$. Interested in β_1; β_0 is nuisance.

Kernel of log-likelihood is: $\sum_i y_i\theta_i = \sum_i y_i(\beta_0+\beta_1 x_i) = \beta_0 \sum_i y_i + \beta_1 \sum_i x_i y_i$ so $\sum_i y_i$ is sufficient for β_0, and $\sum_i x_i y_i$ for β_1. To eliminate β_0, consider $\sum_i x_i y_i = s_1$ while conditioning on $\sum_i y_i = s_0 + s_1$ where s_0 is binomial success totals when $x = 0$ (n_0)and s_1 is total for $x = 1$ (n_1).

Consider $H_0 : \beta_1 = 0 \Leftrightarrow \pi_0 = \pi_1$. Let $\pi = \frac{e^{\beta_0}}{1+e^{\beta_0}}$ under H_0 and consider:

$$P(s_1 = t, s_0 = u) = \binom{n_0}{t}\pi^t(1-\pi)^{n_0-t}\binom{n_1}{u}\pi^u(1-\pi)^{n_1-u}$$

$$P(s_0 + s_1 = v) = \binom{n_0+n_1}{v}\pi^v(1-\pi)^{n_0+n_1-v}$$

$$\Rightarrow P(s_1 = t | s_0 + s_1 = v) = \frac{\binom{n_1}{t}\binom{n_0}{v-t}}{\binom{n_0+n_1}{v}}$$

which is independent of β_0. To test $H_0 : \beta_1 = 0$ vs. $H_a : \beta_1 > 0$, we use: $P(s_1 \geq t | s_1 + s_0)$ where t is observed s_1 value.

Limited: we need sufficient statistics for nuisance parameters; only exist for canonical link GLMs.

5.4 Logistic Regression: Fitting

Iterative Fitting Since logit is canonical, Newton-Raphson = Fisher scoring. We can express derivatives as:

$$u_j^{(t)} = \sum_i (s_i - n_i\pi_i^{(t)})x_{ij} \Rightarrow \mathbf{u}^{(t)} = \mathbf{X}^T(\mathbf{s} - \mu^{(t)})$$

$$(\mathbf{H})_{jk}^{(t)} = -\sum_i x_{ij}x_{ik}n_i\pi_i^{(t)}(1-\pi_i^{(t)}) \Rightarrow \mathbf{H}^{(t)} = -\mathbf{X}^T\text{diag}[n_i\pi_i^{(t)}(1-\pi_i^{(t)})]\mathbf{X}$$

where $\pi_i^{(t)} = \frac{\exp(\sum_j \beta_j^{(t)} x_{ij})}{1+\exp(\sum_j \beta_j^{(t)} x_{ij})}$, $\mu_i^{(t)} = n_i\pi_i^{(t)}$ so that the update is:

$$\boxed{\beta^{(t+1)} = \beta^{(t)} + \left(\mathbf{X}^T\text{diag}[n_i\pi_i^{(t)}(1-\pi_i^{(t)})]\mathbf{X}\right)^{-1}\mathbf{X}^T(\mathbf{s} - \pi^{(t)})}$$

Infinite Estimates Fitting runs into problems when *complete separation* or *quasi-complete separation* occurs. Quick example: $y = 1$ at $x = 1, 2, 3$ and $y = 0$ and $x = 4, 5, 6$; then $\hat{\beta}_0 = -3.5\hat{\beta}_1$ and $\hat{\beta}_1 = \infty$.

Signs: 1) very large standard errors (since log-likelihood is near-flat); 2) perfect prediction ($\hat{\pi}_i = 1$ if $y_i = 1$ and vice versa); 3) maximized log-likelihood is basically 0.

Quasi-complete separation when cases exist with both outcomes on hyperplane; still infinite estimate, but log-likelihood < 0. (Often happens when $y_i = 1$ or 0 for every obs with certain value of categorical variable)

We can still do: 1) LRT of $\beta_1 = 0$ vs. $\hat{\beta}_1 = \infty$ comparing log-likelihoods at these values; 2) invert test to get confidence interval, i.e. (L, ∞) where $H_0 : \beta_1 = L$ has p-value α.

5.5 Deviance and Model Comparison/Checking

1) LRT to check more complex model is better (if not, current model is probably fine); 2) Global goodness-of-fit tests (Pearson chi-squared or deviance)

Deviance For grouped data, saturated model has $\tilde{\pi}_i = y_i$ (sample proportion), so LRT statistic comparing model to saturated is:

$$-2\left[\sum_i (n_i y_i \log(\hat{\pi}_i) + (n_i - n_i y_i)\log(1-\hat{\pi}_i)) - \sum_i (n_i y_i \log(y_i) + (n_i - n_i y_i)\log(1-y_i))\right]$$

$$G^2 = D(\mathbf{y};\hat{\mu}) = 2\sum_i n_i y_i \log\frac{n_i y_i}{n_i\hat{\pi}_i} + 2\sum_i (n_i - n_i y_i)\log\frac{n_i - n_i y_i}{n_i - n_i\hat{\pi}_i} = 2\sum_i \text{obs}\times\log\left(\frac{\text{obs}}{\text{fitted}}\right) \sim \chi^2_{N-p}$$

Pearson Statistic $X^2 = \sum_{2N\text{cells}} \frac{(\text{obs}-\text{fitted})^2}{\text{fitted}} = \sum_i \frac{(n_i y_i - n_i\hat{\pi}_i)^2}{n_i\hat{\pi}_i} + \sum_i \frac{[(n_i - n_i y_i)-(n_i - n_i\hat{\pi}_i)]^2}{n_i - n_i\hat{\pi}_i}$

$$\Rightarrow \boxed{X^2 = \sum_{i=1}^{N}\frac{(y_i-\hat{\pi}_i)^2}{\hat{\pi}_i(1-\hat{\pi}_i)/n_i} \sim \chi^2_{N-p}}$$

Again, X^2 is a quadratic approximation of G^2, and $|X^2 - G^2| \xrightarrow{p} 0$ under H_0. But X^2 converges to χ^2_{N-p} faster than G^2, so provides more reliable estimates when small success/failures.

Also, chi-squared under H_0 ***only*** for grouped data!! Even for grouped data, if N is big with n_i small, then not really chi-squared.

However, even if ungrouped, we can still use $G^2(M_0|M_1) = D(M_0) - D(M_1) \sim \chi^2_{p_1-p_0}$ under $H_0 : M_0$ holds.

Residuals Use Deviance/Pearson statistic (global goodness-of-fit) or LRT/deviance comparison (model comparison) to select a model; then use residuals to determine microscopic fits.

1. Pearson residual: $e_i = \frac{y_i - \hat{\pi}_i}{\sqrt{\hat{\pi}_i(1-\hat{\pi}_i)/n_i}}$
   so that $X^2 = \sum_i e_i^2$
2. Deviance residual: $d_i = \sqrt{2\left[n_i y_i \log\left(\frac{n_i y_i}{n_i\hat{\pi}_i}\right) + (n_i - n_i y_i)\log\left(\frac{n_i - n_i y_i}{n_i - n_i\hat{\pi}_i}\right)\right]} \times \text{sign}(e_i)$
   so that $D(\mathbf{y};\hat{\mu}) = \sum_i d_i^2$
3. Standardized residual: $r_i = \frac{y_i - \hat{\pi}_i}{\sqrt{\hat{\pi}_i(1-\hat{\pi}_i)(1-\hat{h}_{ii})/n_i}} \sim \mathcal{N}(0,1)$ if model holds
   where $\hat{h}_{ii} = (\hat{\mathbf{H}}_W)_{ii}$ for $\hat{\mathbf{H}}_W = \hat{\mathbf{W}}^{1/2}\mathbf{X}(\mathbf{X}^T\hat{\mathbf{W}}\mathbf{X})^{-1}\mathbf{X}^T\hat{\mathbf{W}}^{1/2}$ and $\hat{\mathbf{W}} = n_i\hat{\pi}_i(1-\hat{\pi}_i)$

5.6 Probit and Log-Log Models

Probit Models $\Phi^{-1}(\pi_i) = \sum_j \beta_j x_{ij}$ and $\pi_i = \Phi\left(\sum_j \beta_j x_{ij}\right)$

- Interpreting parameters: $\frac{\partial\pi_i}{\partial x_{ij}} = \beta_j\phi(\sum_j \beta_j x_{ij})$ so at max, 0, rate of increase is $0.4\cdot\beta_j$ (compare to $0.25\cdot\beta_j$ for logistic)
- Logistic comparison: ML parameter estimates in logistic are 1.8 times estimates in probit (because standard deviation of logistic is $pi/\sqrt{3}$ times probit)
- Predictive power: Use ROC curve and $\text{corr}(\mathbf{y}^*, \hat{\mu})$ as in logistic
- Fitting: Use likelihood equations with Φ, ϕ and iterative (Newton-Raphson $\neq$ Fisher scoring)
- Asymptotics: $\hat{\text{var}}(\hat{\beta}) = (\mathbf{X}^T\hat{\mathbf{W}}\mathbf{X})^{-1}$ where $\hat{w}_i = \frac{n_i\phi(\eta_i)^2}{\Phi(\eta_i)[1-\Phi(\eta_i)]}$

Log-Log/Complementary Log-Log Models Both probit and logistic are symmetric response distributions ($\text{logit}(\pi_i) = -\text{logit}(1-\pi_i)$). Log-log/complementary log-log useful when response for π_i is not symmetric.

1. **Log-Log Model** $\pi_i = \exp[-\exp(\sum_j \beta_j x_{ij})]$ or $-\log[-\log(\pi_i)] = \sum_j \beta_j x_{ij}$
   Approaches 0 sharply; approaches 1 slowly
2. **Complementary Log-Log Model**
   $\pi_i = 1 - \exp[-\exp(\sum_j \beta_j x_{ij})]$ or $\log[-\log(1-\pi_i)] = \sum_j \beta_j x_{ij}$
   Approaches 0 slowly; approaches 1 sharply

6 Multinomial Models

Binomial = two categories. Multinomial = c categories. Can be either nominal (no natural category ordering) or ordinal (categories ordered).

$\pi_{ij} = P(y_i = j) = P(y_{ij} = 1)$ s.t. $\sum_{j=1}^{c} \pi_{ij} = 1$; $\mathbf{y}_i = (y_{i1}, \ldots, y_{ic})$ s.t. $\sum_j y_{ij} = 1$. Finally,

$$p(y_{i1}, \ldots, y_{ic}) = \pi_{i1}^{y_{i1}} \cdots \pi_{ic}^{y_{ic}}$$

6.1 Nominal Response: Baseline-Category Logit

Baseline-Category Logits Need to consider all categories exchangeably, so: 1) pick a baseline category, i.e. c; 2) form logits of every other category w.r.t c (i.e. conditional probability of being in category j given in category j or c). Basically treat each j, c pair as binary model.

Baseline logits: $\log \frac{\pi_{i1}}{\pi_{ic}}, \ldots, \log \frac{\pi_{i,c-1}}{\pi_{ic}}$ where the j^{th} category logit is:

$$\log \frac{\pi_{ij}}{\pi_{ic}} = \log \left[\frac{P(y_{ij} = 1 | y_{ij} = 1 \text{ or } y_{ic} = 1)}{1 - P(y_{ij} = 1 | y_{ij} = 1 \text{ or } y_{ic} = 1)} \right] = \text{logit} \left[P(y_{ij} = 1 | y_{ij} = 1 \text{ or } y_{ic} = 1) \right]$$

letting $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})$ be explanatory variable values for subject i and $\beta_j = (\beta_{j1}, \ldots, \beta_{jp})$ be parameters for j^{th} baseline logit (i.e. exp. var. by subject, parameters by logit equation):

$$\boxed{\log \frac{\pi_{ij}}{\pi_{ic}} = \mathbf{x}_i \beta_j = \sum_{k=1}^{p} \beta_{jk} x_{ik}}$$

simultaneously describes effects of $\mathbf{x}_i$ on all $c-1$ baseline logits; effects vary according to j category. Also, determines effects on all other logits, since:

$$\log \frac{\pi_j}{\pi_k} = \log \frac{\pi_j}{\pi_c} - \log \frac{\pi_k}{\pi_c} = \mathbf{x}_i (\beta_j - \beta_k)$$

Nominal: if all outcome category labels are permuted, and parameters permuted according, then model still holds!

Multivariate GLM Generalizing GLM to multivariate response: $\mathbf{g}(\mu_i) = \mathbf{X}_i \beta$ where $\mathbf{g}$ is multivariate; $\mathbf{X}_i$ is model matrix (generally $\mathbf{x}_i$ repeated $|\mathbf{g}|$ times, but can differ for each g_i). $\mathbf{y}_i$ is from multivariate EDF:

$$f(\mathbf{y}_i; \theta_i, \phi) = \exp \left[\frac{\mathbf{y}_i^T \theta_i - b(\theta_i)}{a(\phi)} + c(\mathbf{y}_i, \phi) \right]$$

Multinomial $\in$ Multivariate EDF: $y_i = (y_{i1}, \ldots, y_{i,c-1})$ since $y_{ic} = 1 - (y_{i1} + \cdots + y_{i,c-1})$ so redundant; $\mu_i = (\mu_{i1}, \ldots, \mu_{i,c-1})$ and we can express baseline logit model as:

$$g_j(\mu_i) = \log \left[\frac{\mu_{ij}}{1 - (\mu_{i1} + \cdots + \mu_{i,c-1})} \right], \mathbf{X}_i \beta = \begin{pmatrix} \mathbf{x}_i & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{x}_i & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{x}_i \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_{c-1} \end{pmatrix}$$

where each $\beta_j = (\beta_{j1}, \ldots, \beta_{jp})$

Multinomial likelihood is: $\sum_{j=1}^{c-1} y_{ij} \log \pi_{ij} + \left(1 - \sum_{j=1}^{c-1} y_{ij}\right) \log \pi_{ic} = \sum_{j=1}^{c-1} \log \frac{\pi_{ij}}{\pi_{ic}} + \log \pi_{ic}$

so $\theta_j = \log \frac{\pi_{ij}}{\pi_{ic}}$: baseline logit is the natural parameter and canonical link!

Fitting Important formulas:

$$\boxed{\pi_{ij} = \frac{\exp(\mathbf{x}_i \beta_j)}{1 + \sum_{k=1}^{c-1} \exp(\mathbf{x}_i \beta_k)}}$$

$$\boxed{\pi_{ic} = \frac{1}{1 + \sum_{k=1}^{c-1} \exp(\mathbf{x}_i \beta_k)}}$$

with $\beta_c = \mathbf{0}$ for identifiability (also $\exp(0) = 1$, as needed).

The likelihood equations are:

$$l(\beta;\mathbf{y}) = \log\left[\prod_{i=1}^{N}\left(\prod_{j=1}^{c}\pi_{ij}^{y_{ij}}\right)\right] = \sum_{i=1}^{N}\left[\sum_{j=1}^{c-1} y_{ij}(\mathbf{x}_i\beta_j) - \log\left(1+\sum_{j=1}^{c-1}\exp(\mathbf{x}_i\beta_j)\right)\right]$$
$$= \sum_{j=1}^{c-1}\left[\sum_{k=1}^{p}\beta_{jk}\left(\sum_{i=1}^{N}x_{ik}y_{ij}\right)\right] - \sum_{i=1}^{N}\log\left[1+\sum_{j=1}^{c-1}\exp(\mathbf{x}_i\beta_j)\right]$$

so sufficient statistics are $\sum_i x_{ik}y_{ij}$. Taking derivatives:

$$\frac{\partial l(\beta;\mathbf{y})}{\partial\beta_{jk}} = \sum_{i=1}^{N}x_{ik}y_{ij} - \sum_{i=1}^{N}\left[\frac{x_{ik}\exp(\mathbf{x}_i\beta_j)}{1+\sum_{l=1}^{c-1}exp(\mathbf{x}_i\beta_l)}\right] = \sum_{i=1}^{N}x_{ik}(y_{ij}-\pi_{ij}) = 0$$

$$\Rightarrow \boxed{\sum_{i=1}^{N}x_{ik}y_{ij} = \sum_{i=1}^{N}x_{ik}\pi_{ij}}$$

so sufficient statistic = expected value, as in all canonical link.

Differentiating the log-likelihood again, we have:

$$\frac{\partial^2 l(\beta;\mathbf{y})}{\partial\beta_{jk}\partial\beta_{jk'}} = -\sum_{i=1}^{N}x_{ik}x_{ik'}\pi_{ij}(1-\pi_{ij}),\ \frac{\partial^2 l(\beta;\mathbf{y})}{\partial\beta_{jk}\partial\beta j'k'} = \sum_{i=1}^{N}x_{ik}x_{ik'}\pi_{ij}\pi_{ij'}$$

$$\Rightarrow (\mathcal{J})_{j,j'} = -\frac{\partial^2 l(\beta;\mathbf{y})}{\partial\beta_j\partial\beta_j'} = \sum_{i=1}^{N}pi_{ij}[I(j=j')-\pi_{ij'}]\mathbf{x}_i^T\mathbf{x}_i$$

where each are blocks of size $p\times p$, and there are $(c-1)^2$ of them. We also have: $\hat{\beta}\sim\mathcal{N}(\beta,\mathcal{J}^{-1})$

Deviance and Inference After fitting, need to do: 1) significance tests for parameters; 2) confidence intervals; 3) model comparisons. We can use LRT, Wald, or score for significance tests: i.e. $H_0 = \beta_{1k} = \beta_{2k} = \cdots = \beta_{c-1,k} = 0$ can be done using LRT with maximized likelihood with/without variable x_k; has χ^2_{c-1} distribution.

Deviance/Pearson Statistic: For grouped data, let y_{ij} = proportion of observations in setting i in category j, then multinomial likelihood is: $\prod_i\prod_j \pi_{ij}^{n_iy_{ij}}$ and deviance compares log-likelihood at model fit $\hat{\pi}_{ij}$ and at saturated $\tilde{\pi}_{ij} = y_{ij}$ resulting in:

$$\boxed{G^2 = 2\sum_{i=1}^{N}\sum_{j=1}^{c}n_iy_{ij}\log\frac{n_iy_{ij}}{n_i\hat{\pi}_{ij}} = 2\sum \text{obs}\times\log\frac{\text{obs}}{\text{fitted}} \sim \chi^2_{(N-p)(c-1)}}$$

$$\boxed{X^2 = \sum_{i=1}^{N}\sum_{j=1}^{c}\frac{(n_iy_{ij}-n_i\hat{\pi}_{ij})^2}{n_i\hat{\pi}_{ij}} = \sum\frac{(\text{obs}-\text{fitted})^2}{\text{fitted}} \sim \chi^2_{(N-p)(c-1)}}$$

where $df = N(c-1) - p(c-1) = (N-p)(c-1)$ because that's number of multinomial probabilities modeled minus number of parameters ($\beta_c = 0$). (i.e. N = number of combinations of explanatory variable values.)

6.2 Ordinal Response: Cumulative Logit

If categories are ordered, use cumulative logits; generally fewer parameters, so model parsimony!

Cumulative Logit Models Now let $y_i = j$ represent subject i falling into category j; equivalent to $y_{ij} = 1$. Consider cumulative probabilities $P(y_i \le j) = \pi_{i1} + \cdots + \pi_{ij}$.

Cumulative logits: $\text{logit}[P(y_i\le j)] = \log\frac{\pi_{i1}+\cdots+\pi_{ij}}{\pi_{i,j+1}+\cdots+\pi_{ic}}$

Cumulative logit model: Consider being in categories $1,\ldots,j$ as "success", categories $j+1,\ldots,c$ as "failure". Then:

$$\boxed{\text{logit}[P(y_i\le j)] = \alpha_j + \mathbf{x}_i\beta}$$

where each cumulative logit has different intercept but same slope; α_j increasing in j (i.e. same shape logit curves, do not cross). Ordinal because if arbitrary permutation of labels, then model need not hold!

Proportional odds structure: Note that:

$$\log \frac{P(y_i \leq j|\mathbf{x}_i = \mathbf{u})/P(y_i > j|\mathbf{x}_i = \mathbf{u})}{P(y_i \leq j|\mathbf{x}_i = \mathbf{v})/P(y_i > j|\mathbf{x}_i = \mathbf{v})} = \text{logit}[P(y_i \leq j|\mathbf{x}_i = \mathbf{u})] - \text{logit}[P(y_i \leq j|\mathbf{x}_i = \mathbf{v})] = (\mathbf{u}-\mathbf{v})\beta$$

so *cumulative odds ratio* (odds ratio of cumulative probabilities at different values of $\mathbf{x}_i$) is proportional to $e^{(\mathbf{u}-\mathbf{v})\beta}$. Every unit increase in x_{ik} results in odds of $y_i \leq j$ multiplying by e^{β_k}.

Latent Variable Motivation Motivate common effect β: suppose linear y_i^* s.t. $y_i^* = \mathbf{x}_i\beta + \epsilon_i$ and $\epsilon_i \sim G(\cdot)$, i.e. $\mu_i = \mathbf{x}_i\beta$ and $y_i^* \sim G(y_i^* - \mu_i)$. Cutpoints $-\infty = \alpha_0 < \alpha_1 < \cdots < \alpha_c = \infty$ so that $y_i = j$ iff $\alpha_{j-1} < y_i^* \leq \alpha_j$. Then: $P(y_i \leq j) = P(y_i^* \leq \alpha_j) = G(\alpha_j - \mathbf{x}_i\beta)$, so the link function is G^{-1} and $G^{-1}[P(y_i \leq j)] = \alpha_j - \mathbf{x}_i\beta$. (Note: $-$ instead of $+$ here: if $\beta_k > 0$ and as x_{ik} increases, each $P(y_i \leq j)$ decreases, so less probability of being at low end of scale, so y_i tends to be larger at higher values of x_{ik}.) *Same effects* β regardless of selection of cutpoints!

Cumulative Link Models $G^{-1}[P(y_i \leq j)] = \alpha_j + \mathbf{x}_i\beta$. Effects are same for each cumulative probability; G is CDF of error term.

Cumulative probit if $G = \Phi$ for standard normal; again effects $\pi/\sqrt{3}$ times bigger in logit model. 1-unit increase in x_{ik} corresponds to β_k increase in $E(y_i^*)$.

Predictive Power Use $\text{corr}(\mathbf{y}^*, \hat{\mathbf{y}}^*)$, that is:

$$R^2 \approx \text{corr}(\mathbf{y}^*, \hat{\mathbf{y}}^*)^2 = \frac{\text{var}(\hat{y}^*)}{\hat{\text{var}}(\mathbf{y}^*)} = \frac{\text{var}(\hat{\mathbf{y}}^*)}{\text{var}(\hat{y}^*) + \text{var}(\epsilon)}$$

where $\text{var}(\epsilon) = 1$ for probit, $\pi/\sqrt{3}$ for logit.

Fitting Consider again multicategory indicator $\mathbf{y}_i = (y_{i1}, \ldots, y_{ic})$ and cumulative link model $G^{-1}[P(y_i \leq j)] = \alpha_j + \mathbf{x}_i\beta$. The likelihood is:

$$\prod_{i=1}^{N}\prod_{j=1}^{c} \pi_{ij}^{y_{ij}} = \prod_{i=1}^{N}\prod_{j=1}^{c}[P(y_i \leq j) - P(y_i \leq j-1)]^{y_{ij}}$$

$$\Rightarrow \boxed{l(\alpha, \beta) = \sum_{i=1}^{N}\sum_{j=1}^{c} y_{ij} \log[G(\alpha_j + \mathbf{x}_i\beta) - G(\alpha_{j-1} + \mathbf{x}_i\beta)]}$$

Then the likelihood equations are (with g being PDF of G):

$$\frac{\partial l}{\partial \beta_k} = \sum_{i=1}^{N}\sum_{j=1}^{c} y_{ij}x_{ik}\frac{g(\alpha_j + \mathbf{x}_i\beta) - g(\alpha_{j-1} + \mathbf{x}_i\beta)}{G(\alpha_j + \mathbf{x}_i\beta) - G(\alpha_{j-1} + x_i\beta)} = 0$$

$$\frac{\partial l}{\partial \alpha_k} = \sum_{i=1}^{N}\sum j = 1^c y_{ij}\frac{\delta_{jk}g(\alpha_j + \mathbf{x}_i\beta) - \delta_{j-1,k}g(\alpha_{j-1} + \mathbf{x}_i\beta)}{G(\alpha_j + \mathbf{x}_i\beta) - G(\alpha_{j-1} + x_i\beta)} = 0$$

Model Checking Cumulative logit/proportional odds assumes: 1) location varies (i.e. α_j differs by j); 2) constant variability (β constant). This results in *stochastic ordering*: $P(y_i \leq j|\mathbf{x}_i = \mathbf{u}) \leq P(y_i \leq j|\mathbf{x}_i = \mathbf{v})$ or $P(y_i \leq j|\mathbf{x}_i = \mathbf{u}) \geq P(y_i \leq j|\mathbf{x}_i = \mathbf{v})$ for **all** j! (If this is violated, cumulative logits might not fit well.)

Score test: Can check if separate effects β_j fit better than common β by using score test $H_0 : \beta_1 = \cdots = \beta_c = \beta$ (since score test only uses log-likelihood at H_0, i.e. common effects, so no problems with fitting with β_j.)

Using OLS for Ordinal Problems: 1) No clear-cut choice for category to numerical score; 2) Ordinal outcome is consistent with $[\alpha_{j-1}, \alpha_j]$ interval of response; OLS doesn't consider this error; 3) OLS does not yield estimated prob. for each category given x_i; 4) Non-constant variability due to floor/ceiling effects violates OLS; 5) Floor/ceiling effects can yield spurious interactions effects.

7 Count Models

7.1 Poisson Loglinear Model

Poisson Distribution Properties include:

- PMF: $p(y;\mu) = \frac{\mu^y e^{-\mu}}{y!}$
- Moments: $E(y_i) = \mu$, $\text{var}(y_i) = \mu$, and $\text{skew}(y_i) = 1/\sqrt{\mu}$, with $\text{mode}(y_i) = \lfloor \mu \rfloor$

We have two ways of fitting count data assuming $y_i \sim \text{Pois}(\mu_i)$.

1. **Variance Stabilization + OLS**: Since Poisson has non-constant variance, we can transform y_i so transformed values have constant variance. By delta method, $\text{var}[g(y)] \approx [g'(\mu)]^2 \text{var}(y)$ so using $g(y) = \sqrt{y}$: $\text{var}(\sqrt{y}) \approx \left(\frac{1}{2\sqrt{\mu}}\right)^2 \mu = \frac{1}{4}$!
   So fit $E(\sqrt{\mathbf{y}}) = \mathbf{X}\beta$ using OLS. But: 1) effects hard to interpret; 2) other transforms might fit linear predictor better (i.e. $\log(y_i)$ or y_i itself).
2. **Poisson Loglinear GLM**: Using $\log \mu_i = \sum_j \beta_j x_{ij}$, model is:

$$\boxed{\log \mu_i = \sum_{j=1}^{p} \beta_j x_{ij} \text{ or } \log \mu = \mathbf{X}\beta}$$

   The likelihood equations become: $\sum_i x_{ij}(y_i - \mu_i) = 0$
   Exponential relation: $\boxed{\mu_i = (e^{\beta_1})^{x_{i1}} \cdots (e^{\beta_p})^{x_{ip}}}$, i.e. 1-unit increase in x_{ij} multiples μ_i by e^{β_j}

Model Fitting As usual, Newton-Raphson = Fisher Scoring for canonical log link; and asymptotically/estimated covariance of $\hat{\beta}$ is: $\hat{\text{var}}(\hat{\beta}) = (\mathbf{X}^T \hat{\mathbf{W}} \mathbf{X})^{-1}$ with $w_i = \mu_i$.

Model Checking/Comparison Again, we use global goodness-of-fits: Deviance or Pearson

Deviance: $D(\mathbf{y};\hat{\mu}) = 2\sum_i \left[y_i \log\left(\frac{y_i}{\hat{\mu}}\right) - y_i + \hat{\mu}_i\right]$ but if there is intercept term, then by likelihood equations, $\sum_i y_i = \sum_i \hat{\mu}_i$, so:

$$\boxed{G^2 = D(\mathbf{y};\hat{\mu}) = 2\sum_{i=1}^{n} \left[y_i \log\left(\frac{y_i}{\hat{\mu}_i}\right)\right]}$$

Pearson Statistic: $\boxed{X^2 = \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i}}$

Both statistics are χ^2_{n-p} when n is fixed and μ_i grows unboundedly (i.e. contingency tables with fixed cells and sample size within each cell growing).

But neither reveals **how** the model fails. Better to compare (i.e. LRT/Deviance comparison) with more complex model, i.e. Poisson $\subset$ Negative binomial.

Residuals For Poisson GLM:

- **Pearson residual**: $e_i = \frac{y_i - \hat{\mu}_i}{\sqrt{\hat{\mu}_i}}$
- **Deviance residual**: components of deviance d_i as usual
- **Standardized residual**: $r_i = \frac{y_i - \hat{\mu}_i}{\sqrt{\hat{\mu}_i(1-\hat{h}_{ii})}}$

Also: compare observed counts to fitted counts; generally too low for 0 and high outcomes

Example: One-Way Layout Suppose y_{ij} is count variable in one-way layout of obs j in group i, $i = 1, \ldots, c$ and $j = 1, \ldots, n_i$, $n = \sum_i n_i$. Let $y_{ij} \sim \text{Pois}(\mu_{ij})$; model common means in groups, $\log(\mu_{ij}) = \beta_i$ ($\beta_0 = 0$ for identifiability). Then $\log \mu = \mathbf{X}\beta$ with:

$$\mu = \begin{pmatrix} \mu_1 \mathbf{1}_{n_1} \\ \mu_2 \mathbf{1}_{n_2} \\ \vdots \\ \mu_c \mathbf{1}_{n_c} \end{pmatrix}, \mathbf{X}\beta = \begin{pmatrix} \mathbf{1}_{n_1} & \mathbf{0}_{n_1} & \cdots & \mathbf{0}_{n_1} \\ \mathbf{0}_{n_2} & \mathbf{1}_{n_2} & \cdots & \mathbf{0}_{n_2} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0}_{n_c} & \mathbf{0}_{n_c} & \cdots & \mathbf{1}_{n_c} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_c \end{pmatrix}$$

Likelihood equations for β_i are: $\sum_{j=1}^{n_i} (y_{ij} - \hat{\mu}_i) = 0$ so that $\hat{\mu}_i = \bar{y}_i \Rightarrow \hat{\beta}_i = \log \bar{y}_i$.

Since $\hat{w}_{ii} = \hat{\mu}_i = \bar{y}_i$, we have: $\hat{\text{var}}(\hat{\beta}) = (\mathbf{X}^T \hat{\mathbf{W}} \mathbf{X})^{-1} = \text{diag}\left(\frac{1}{n_i \bar{y}_i}\right)$ so $\hat{\beta}_i$ are uncorrelated and since $\frac{\mu_h}{\mu_i} = \exp(\beta_h - \beta_i)$, $\text{var}(\beta_h - \beta_i) = \text{var}(\beta_h) + \text{var}(\beta_i)$ and the $100(1-\alpha)$% CI for the ratio of means:

$$\frac{\mu_h}{\mu_i} \in \exp\left[(\hat{\beta}_h - \hat{\beta}_i) \pm z_{\alpha/2} \sqrt{\frac{1}{n_h \bar{y}_h} + \frac{1}{n_i \bar{y}_i}}\right]$$

$H_0 : \mu_1 = \cdots = \mu_c$ by using Deviance comparison/LRT, which equals: $2 \sum_{i=1}^{c} n_i \bar{y}_i \log\left(\frac{\bar{y}_i}{\bar{y}}\right) \approx \chi^2_{c-1}$

Global GOF tests: $G^2 = 2 \sum_{i=1}^{c} \sum_{j=1}^{n_i} y_{ij} \log\left(\frac{y_{ij}}{\bar{y}_i}\right)$ and $X^2 = \sum_{i=1}^{c} \sum_{j=1}^{n_i} \frac{(y_{ij} - \bar{y}_i)^2}{\bar{y}_i} \sim \chi^2_{\sum_i (n_1 - 1)}$

7.2 Contingency Tables: Poisson = Multinomial

Independent Poisson counts in cells = multinomial models once conditioned on total sample size. Explore independence/association/interaction structure by specifying models with interaction terms (vs. not).

Poisson = Multinomial Independent Poisson $(y_1, \ldots, y_c)$, means $(\mu_1, \ldots, \mu_c)$; total $n = \sum_j y_j \sim \text{Pois}(\sum_j \mu_j)$. Then conditional probability of $(y_1, \ldots, y_c)$ given n is:

$$P\left[y_1 = n_1, \ldots, y_c = n_c | \sum_{j=1}^{c} y_j = n\right] = \frac{P(y_1 = n_1, \ldots, y_c = n_c)}{P(\sum_j y_j = n)} = \left(\frac{n!}{n_1! \cdots n_c!}\right) \prod_{j=1}^{c} \pi_j^{n_j}$$

where $\pi_j = \frac{\mu_j}{\sum_i \mu_i}$; i.e. multinomial with n, pi_j.

Example: Two-Way Contingency Table Two categorical variables, A and B, $r \times c$ table; y_{ij} with $A = i$, $B = j$. Model: $\mu_{ij} = \mu \phi_i \psi_j$ s.t. $\sum_i \phi_i = \sum_j \psi_j = 1$. Then, log model is additive: $\log \mu_{ij} = \beta_0 + \beta_i^A + \beta_j^B$ (main effects, no interaction; identifiability requires first-category baseline)

Multinomial: Conditional on $\sum_i \sum_j y_{ij} = n$, we have $\sum_i \sum_j \mu_{ij} = \mu$, so $\pi_{ij} = \mu_{ij}/\mu = \phi_i \psi_j$, and since $\sum_i \phi_i = 1, \sum_j \psi_j = 1$, we must have $\phi_i = \pi_{i+}$ and $\psi_j = \pi_{+j}$. Thus: $\boxed{\{\pi_{ij} = \pi_{i+}\pi_{+j}\}}$ and so category responses in A vs. B are **independent**! (i.e. $P(A = i, B = j) = P(A = i)P(B = j)$)

Poisson: Consider 2×2 table, $\beta_1^A = \beta_1^B = 0$ for identifiability, then:

$$\log \mu = \begin{pmatrix} \log \mu_{11} \\ \log \mu_{12} \\ \log \mu_{21} \\ \log \mu_{22} \end{pmatrix} = \mathbf{X}\beta = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_2^A \\ \beta_2^B \end{pmatrix}$$

Deriving the likelihood equations, with $\log \mu_{ij} = \beta_0 + \beta_i^A + \beta_j^B$, we have log-likelihood kernel:

$$l(\mu) = \sum_{i=1}^{r} \sum_{j=1}^{c} y_{ij} \log(\mu_{ij}) - \sum_{i=1}^{r} \sum_{j=1}^{c} \mu_{ij} = n\beta_0 + \sum_{i=1}^{r} y_{i+} \beta_i^A + \sum_{j=1}^{c} y_{+j} \beta_j^B - \sum_{i=1}^{r} \sum_{j=1}^{c} \exp(\beta_0 + \beta_i^A + \beta_j^B)$$

$$\frac{\partial l}{\partial \beta_i^A} = y_{i+} - \sum_{j=1}^{c} \exp(\beta_0 + \beta_i^A + \beta_j^B) = y_{i+} - \mu_{i+} \; , \; \frac{\partial l}{\partial \beta_j^B} = y_{+j} - \mu_{+j}$$

So ML fitted values are: $\boxed{\left\{\hat{\mu}_{ij} = \frac{y_{i+} y_{+j}}{n}\right\}}$ (equivalent to multinomial: $\hat{\pi}_{i+} = y_{i+}/n, \hat{\pi}_{+j} = y_{+j}/n$)

Parameters: Multinomial has $(r-1) + (c-1)$, while Poisson has $1 + (r-1) + (c-1)$.

Pearson Statistic: $X^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(y_{ij} - \hat{\mu}_{ij})^2}{\hat{\mu}_{ij}} \sim \chi^2_{(r-1)(c-1)}$ (since $(rc - 1) - (r-1) - (c-1)$)

Example: Adding Interaction Term Suppose $\log \mu_{ij} = \beta_0 + \beta_i^A + \beta_j^B + \gamma_{ij}^{AB}$, interaction term γ_{ij}^{AB}; model matrix has cross-products of $r-1$ row indicators and $c-1$ column indicators. (i.e. $\gamma_{1j}^{AB} = \gamma_{i1}^{AB} = 0$, so for first column/row, we just have $\beta_0 + \beta_i^A$ or $\beta_0 + \beta_j^B$; yields $1 + (r-1) + (c-1) + (r-1)(c-1) = rc$, so model is now saturated)

Interpretation: odds ratios. For $r = c = 2$, the log odds ratio is:

$$\log \frac{\pi_{11}/\pi_{21}}{\pi_{12}/\pi_{22}} = \log \frac{\mu_{11}\mu_{22}}{\mu_{12}\mu_{21}} = \gamma_{11}^{AB} + \gamma_{22}^{AB} - \gamma_{12}^{AB} - \gamma_{21}^{AB} = \gamma_{22}^{AB}$$

so $e^{\gamma_{22}^{AB}}$ is odds ratio between being in $A = 1$ vs$A = 2$ given in $B = 1$ over $B = 2$.

General Interactions for Multiway Tables Consider three-way table, A, B, C, with $r \times c \times l$ cells; independent cell counts $\{y_{ijk}\}$ or multinomial cell prob. $\{\pi_{ijk}\}$ with $\sum_i \sum_j \sum_k \pi_{ijk} = 1$.

1. **Mutual independence**: $\boxed{P(A = i, B = j, C = k) = P(A = i)P(B = j)P(C = k)}$, that is $\pi_{ijk} = \pi_{i++}\pi_{+j+}\pi_{++k}$ or $\log \mu_{ijk} = \beta_0 + \beta_i^A + \beta_j^B + \beta_k^C$ (independence = additive)
2. **Joint independence**: $\boxed{P(A = i, B = j, C = k) = P(A = i)P(B = j, C = k)}$: A is jointly independent of B, C. That is, $\pi_{ijk} = \pi_{i++}\pi_{+jk}$ or $\log \mu_{ijk} = \beta_0 + \beta_i^A + \beta_j^B + \beta_k^C + \gamma_{jk}^{BC}$
3. **Conditional independence**: $\boxed{P(A = i, B = j|C = k) = P(A = i|C = k)P(B = j|C = k)}$ then A, B are conditionally independent given C (i.e. consider separate two-way tables between A, B for each value of C; then in each two-way table, A, B are independent.)
   Then $\pi_{ijk} = \frac{\pi_{i+k}\pi_{+jk}}{\pi_{++k}}$ and $\log \mu_{ijk} = \beta_0 + \beta_i^A + \beta_j^B + \beta_k^C + \gamma_{ik}^{AC} + \gamma_{jk}^{BC}$
4. **Homogenous association**: All pairs can be conditionally dependent:

   $$\log \mu_{ijk} = \beta_0 + \beta_i^A + \beta_j^B + \beta_k^C + \gamma_{ij}^{AB} + \gamma_{ik}^{AC} + \gamma_{jk}^{BC}$$

   Similar interpretation as interaction term in two-way model: consider fixed $C = k$, then **conditional association** between A, B is specified by odds ratios: $\theta_{ij(k)} = \frac{\mu_{ijk}\mu_{rck}}{\mu_{ick}\mu_{rjk}}$ i.e. to baseline categories r, c. Then the log odds for $r = c = 2$ are: $\log \theta_{11(k)} = \log \frac{\mu_{11k}\mu_{22k}}{\mu_{12k}\mu_{21k}} = \gamma_{11}^{AB} + \gamma_{22}^{AB} - \gamma_{12}^{AB} - \gamma_{21}^{AB} = \gamma_{22}^{AB}$ so that $\theta_{ij(1)} = \cdots = \theta_{ij(l)}$ for every i, j (without three-factor term) $\Rightarrow$ **homogeneous association**.

Fitting in Contingency Tables Generally likelihood equations equate observed counts = fitted values for the highest-order terms, i.e.:

1) Mutual independence: $y_{i++} = \hat{\mu}_{i++}, y_{+j+} = \hat{\mu}_{+j+}, y_{++k} = \hat{\mu}_{++k}$

2) Homogenous association: $y_{ij+} = \hat{\mu}_{ij+}, y_{i+k} = \hat{\mu}_{i+k}, y_{+jk} = \hat{\mu}_{+jk}$

Loglinear $\leftrightarrow$ Logistic Models Loglinear = symmetric category classifications, model joint distribution of categorical variables; Logistic = distinguish response vs. explanatory classifications.

Consider homogeneous association model, with A as response, B, C as explanatory; i.e. condition on n_{+jk} for each combination of B, C values, so $c \times l$ logits. Let $r = 2$, then:

$$\log \frac{P(A = 1|B = j, C = k)}{P(A = 2|B = j, C = k)} = \log \frac{\mu_{1jk}}{\mu_{2jk}} = \log \mu_{1jk} - \log \mu_{2jk} = (\beta_1^A - \beta_2^A) + (\gamma_{1j}^{AB} - \gamma_{2j}^{AB}) + (\gamma_{1k}^{AC} - \gamma_{2k}^{AC})$$

$$\Rightarrow \text{logit}[P(A = 1|B = j, C = k)] = \lambda + \delta_j^B + \delta_k^C$$

Same thing can be done if $r > 2$ using baseline-logits for A in terms of $B, C, \ldots$ So note that the log-odds ratio at, say, different values of B are:

$$\log \frac{P(A = 1|B = u, C = k)/P(A = 2|B = u, C = k)}{P(A = 1|B = v, C = k)/P(A = 2|B = v, C - k)} = \delta_u^B - \delta_v^B$$

so the interaction terms are exactly the log-odds ratios, as in loglinear case.

7.3 Negative Binomial GLMs

Overdispersion: Poisson has variance = mean; but count data often has variance > mean, often due to heterogeneity (mixture of Poisson; not all explanatory variables in model)

Negative Binomial = Gamma Mixture of Poisson

$$y|\lambda \sim \text{Pois}(\lambda)$$

$$\lambda \sim \text{Gamma}(\mu, k)$$

Then $E(\lambda) = \mu$, $\text{var}(\lambda) = \frac{\mu^2}{k}$, so that $E(y) = E[E(y|\lambda)] = \mu$ and $\text{var}(y) = E[\text{var}(y|\lambda)] + \text{var}[E(y|\lambda)] = E(\lambda) + \text{var}(\lambda) = \mu + \frac{\mu^2}{k} > \mu$.

Marginal y over Gamma mixture yields **Negative Binomial**:

- PDF: $p(y; \mu, k) = \frac{\Gamma(y+k)}{\Gamma(k)\Gamma(y+1)} \left(\frac{\mu}{\mu+k}\right)^y \left(\frac{k}{\mu+k}\right)^k$
- Natural parameter: $\theta_i = \log \frac{\mu_i}{\mu_i+k}$ for fixed k
- Dispersion parameter: $\gamma = 1/k$ (NBin $\to$ Pois as $\gamma \to 0$)
- Moments: $E(y) = \mu$, $\text{var}(y) = \mu + \gamma\mu^2$

Negative Binomial GLMs Use log link rather than canonical (natural parameter above); treat γ as constant for all i but unknown.

- Link: $\log \mu_i$
- Log-likelihood:

$$l(\beta, \gamma; \mathbf{y}) = \sum_{i=1}^{n} [\log\Gamma(y_i + 1/\gamma) - \log\Gamma(1/\gamma) - \log\Gamma(y_i + 1)] + \sum_{i=1}^{n} \left[y_i \log\left(\frac{\gamma\mu_i}{1+\gamma\mu_i}\right) - \left(\frac{1}{\gamma}\right)\log(1+\gamma\mu_i)\right]$$

- Likelihood equations: $\sum_{i=1}^{n} \frac{(y_i-\mu_i)x_{ij}}{\mu_i+\gamma\mu_i^2}\left(\frac{\partial\mu_i}{\partial\eta_i}\right) = 0$
- Hessian: $\frac{\partial^2 l}{\partial\beta_j\partial\gamma} = -\sum_i \frac{(y_i-\mu_i)x_{ij}}{(1+\gamma\mu_i)^2}\left(\frac{\partial\mu_i}{\partial\eta_i}\right)$
  so $E\left[\frac{\partial^2 l}{\partial\beta_j\partial\gamma}\right] = 0$ and β, γ are orthogonal, and $\hat{\beta}, \hat{\gamma}$ are asymptotically independent.
- Fitting: $\hat{w}_i = \frac{\hat{\mu}_i}{1+\gamma\hat{\mu}_i}$ and $\hat{\text{var}}(\hat{\beta}) = (\mathbf{X}^T\hat{\mathbf{W}}\mathbf{X})^{-1}$ with log link.
- Deviance: $D(\mathbf{y}; \hat{\mu}) = 2\sum_i \left[y_i \log\left(\frac{y_i}{\hat{\mu}_i}\right) - \left(y_i + \frac{1}{\hat{\gamma}}\right)\log\left(\frac{1+\hat{\gamma}y_i}{1+\hat{\gamma}\hat{\mu}_i}\right)\right]$

Model Comparison: Poisson vs. NBin Use LRT with $H_0: \gamma = 0$ (or informally AIC values). But since $\gamma = 0$ is on boundary, the LRT statistic is 1/2 point mass at 0 and 1/2 chi-squared, df = 1, so the p-value is 1/2 what we obtain by treating LRT statistic as χ^2_1.

7.4 Zero-Inflated GLMs

Often counts of 0 are much larger than expected for Poisson; i.e. random vs. structural zero $\Rightarrow$ zero-inflation. Less problematic for negative binomial, but still can be problem if two modes (i.e. mode at 0, mode > 0).

Zero-Inflated Poisson (ZIP) Mixture model of: 1) point mass at 0; 2) count distribution (Poisson):

$$y_i \sim \begin{cases} 0 & \text{with probability} \quad 1-\phi_i \\ \text{Pois}(\lambda_i) & \text{with probability} \quad \phi_i \end{cases}$$

- Unconditional PMF:

$$P(y_i = 0) = (1-\phi_i) + \phi_i e^{-\lambda_i}, P(y_i = j) = \phi_i \frac{\lambda_i^j e^{-\lambda_i}}{j!}$$

- Model: $\text{logit}(\phi_i) = \mathbf{x}_{1i}\beta_1$ and $\log(\lambda_i) = \mathbf{x}_{21}\beta_2$

- Latent variable: $z_i = 0 \Rightarrow y_i = 0$, $z_i = 1 \Rightarrow y_i \sim \text{Pois}(\lambda_i)$; $P(z_i = 0) = 1 - \phi_i, P(z_i = 1) = \phi_i$
- Moments: $E(y_i) = E[E(y_i|z_i)] = (1 - \phi_i) \cdot 0 + \phi_i\lambda_i = \phi_i\lambda_i$
  $\text{var}(y_i) = E[\text{var}(y_i|z_i)]+\text{var}[E(y_i|z_i)] = [(1-\phi_i)\cdot 0+\phi_i\lambda_i]+[(1-\phi_i)(0-\phi_i\lambda_i)^2+\phi_i(\lambda_i-\phi_i\lambda_i)^2] = \phi_i\lambda_i[1 + (1 - \phi_i)\lambda_i] > E(y_i)$ (overdispersion)
- Log-likelihood:

$$l(\beta_1, \beta_2) = \sum_{y_i=0} \log[1+e^{\mathbf{x}_{1i}\beta_1} e^{-exp(\mathbf{x}_{2i}\beta_2)}] - \sum_{i=1}^{n} \log(1+e^{\mathbf{x}_{1i}\beta_1}) + \sum_{y_i>0} [\mathbf{x}_{1i}\beta_1 + y_i\mathbf{x}_{2i}\beta_2 - e^{\mathbf{x}_{2i}\beta_2} - \log(y_i!)]$$

- Simpler parametrization: ZIP model has many parameters β_1, β_2 compared to Poisson. Instead, consider: $\mathbf{x}_{1i} = \mathbf{x}_{2i}$ and $\beta_2 = \tau\beta_1$
  Interpretability also ruined because parameters do not directly effect $E(y_i) = \phi_i\lambda_i$; one solution is to do null model for ϕ_i (so $E(y_i)$ proportional to λ_i)

Zero-Inflated Negative Binomial (ZINB) Same as Poisson, except negative binomial on count part; useful when still **overdispersion** after applying ZIP model

Hurdle Model "Hurdle" crossing 0; $P(y_i > 0) = \pi_i$, $P(y_i = 0) = 1 - \pi_i$; truncated model for $y_i|y_i > 0$

- PMF: $P(y_i = 0) = 1 - \pi_i, P(y_i = j) = \pi_i \frac{f(j;\mu_i)}{1-f(0;\mu_i)}$
- Model: $\text{logit}(\pi_i) = \mathbf{x}_{1i}\beta_1$ and $\log(\mu_i) = \mathbf{x}_{2i}\beta_2$
- Log-likelihood: $l(\beta_1, \beta_2) = l_1(\beta_1) + l_2(\beta_2)$ with:

$$l_1(\beta_1) = \sum_{y_i=0} \log(1 - \pi_i) + \sum_{y_i>0} \log(\pi_i) = \sum_{y_i>0} \mathbf{x}_{1i}\beta_1 - \sum_{i=1}^{n} \log(1 + e^{\mathbf{x}_{1i}\beta})$$

$$l_2(\beta_2) = \sum_{y_i>0} \left[\log f\left(y_i; e^{\mathbf{x}_{2i}\beta_2}\right) - \log[1 - f\left(0; e^{\mathbf{x}_{2i}\beta_2}\right)]\right]$$

8 Quasi-Likelihood

QL is motivated by two points:

1. Overdispersion: i.e. for Poisson, restriction of variance = mean made the fit very poor for many data sets.

2. Mean-variance relation: Likelihood equations **only** depend on distribution of y_i through μ_i and $v(\mu_i)$.

So instead of specifying distribution for y_i, just pick mean-variance relation $v(\mu_i)$, which seems appropriate for given data; along with: 1) link function; 2) linear predictor.

8.1 Variance Inflation for Poisson/Binomial GLMs

To motivate QL methods, we use QL to deal with variance inflation in Poisson/Binomial models.

QL Approach to Variance Inflation Suppose standard model (i.e. Poisson/Binomial) assumes $v^*(\mu_i)$, but actual variance may be different, i.e.:

$$\boxed{\text{var}(y_i) = v(\mu_i) = \phi v^*(\mu_i)}$$

for constant ϕ ($\phi > 1$ is overdispersion case.)

- Substitute $v(\mu_i)$ into likelihood equations; ϕ drops since equal to zero: $\sum_i \frac{(y_i-\mu_i)x_{ij}}{v(\mu_i)} \left(\frac{\partial \mu_i}{\partial \eta_i}\right) = 0 \Rightarrow \sum_i \frac{(y_i-\mu_i)x_{ij}}{v^*(\mu_i)} \left(\frac{\partial \mu_i}{\partial \eta_i}\right) = 0$ so identical to likelihood equations for GLM with variance $v^*(\mu_i)$.
- Fits/estimates identical; $w_i = \frac{(\partial\mu_i/\partial\eta_i)^2}{\text{var}(y_i)} = \frac{(\partial\mu_i/\partial\eta_i)^2}{\phi v^*(\mu_i)}$ so asymptotic var$(\hat{\beta}) = (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1} = \phi(\mathbf{X}^T\mathbf{W}^*\mathbf{W})^{-1}$ for the QL-adjusted model. (i.e. $\boxed{SE_{QL} = \sqrt{\phi} \times SE_{standard}}$)
- Pearson statistic: $X^2 = \sum_i \frac{(y_i-\hat{\mu}_i)^2}{v^*(\hat{\mu}_i)}$ for standard model.
  If variance inflation, then X^2 doesn't fit well; for QL model, want $X^2/\phi \approx \chi^2_{n-p}$ so $E(X^2/\phi) \approx n-p \Rightarrow E[X^2/(n-p)] \approx \phi$ and:

$$\boxed{\hat{\phi} = \frac{X^2}{n-p} = \frac{1}{n-p}\sum_{i=1}^{n} \frac{(y_i-\hat{\mu}_i)^2}{\hat{\mu}_i}}$$

So steps to fitting QL approach are:

1. Fit standard GLM with variance $v^*(\mu_i)$, and use p ML estimates $\hat{\beta}$
2. Multiply standard SE estimates by $\sqrt{\hat{\phi}} = \sqrt{X^2/(n-p)}$

Overdispersed Poisson $v(\mu_i) = \phi\mu_i$, with identical parameter estimates, and Pearson statistic: $X^2 = \sum_i \frac{(y_i-\hat{\mu}_i)^2}{\hat{\mu}_i}$ so $\hat{\phi} = X^2/(n-p)$ for variance-inflation estimate

Overdispersed Binomial Let $n_i y_i \sim \text{Bin}(n_i, \pi_i)$; overdispersion due to: 1) heterogeneity due to unobserved variables; 2) positive correlation between Bern trials (alternative: use Beta-Binomial)

Variance function: $v(\mu_i) = \phi\pi_i(1-\pi_i)/n_i$

Pearson statistic/estimate: $\hat{\phi} = \frac{X^2}{n-p} = \frac{1}{n-p}\sum_i \frac{(y_i-\hat{\pi}_i)^2}{\hat{\pi}_i(1-\hat{\pi}_i)/n_i}$

Note: Does **not** work for ungrouped data, because necessarily var$(y_i) = \pi_i(1-\pi_i)$ structurally

8.2 Beta-Binomial Models

Handling Binomial overdispersion (without structural problems as in variance-inflation) due to: 1) correlated trials; 2) unobserved heterogeneity

1) Correlated Bernoulli Trials Let $y_{i1}, \ldots, y_{in_i}$ be n_i Bernoulli trials for $y_i = \sum_{j=1}^{n_i} \frac{y_{ij}}{n_i}$. If trials not independent, i.e. $\text{corr}(y_{ij}, y_{ik}) = \rho$: $\text{var}(y_{ij}) = \pi_i(1-\pi_i)$, $\text{Cov}(y_{ij}, y_{ik}) = \rho\pi_i(1-\pi_i)$, so:

$$\text{var}(y_i) = \frac{1}{n_i^2}\text{var}(\sum_{j=1}^{n_i} y_{ij}) = \frac{1}{n_i^2}\left[\sum_{j=1}^{n_i}\text{var}(y_{ij}) + 2\sum_{j<k}\text{Cov}(y_{ij}, y_{ik})\right] = \frac{1}{n_i^2}[n_i\pi_i(1-\pi_i) + n_i(n_i-1)\rho\pi_i(1-\pi_i)]$$

$$\Rightarrow \boxed{\text{var}(y_i) = [1+\rho(n_i-1)]\frac{\pi_i(1-\pi_i)}{n_i}}$$

so overdispersion when $\rho > 0$ (also works when $n_i = 1$ since just binomial variance)

Using QL with $v(\pi_i) = [1+\rho(n_i-1)]\frac{\pi_i(1-\pi_i)}{n_i}$, the estimates differ from ML estimates (since $1+\rho(n_i-1)$ term doesn't drop out of likelihood equations). Iterative method:

1. Solve quasi-likelihood equations for $\hat{\beta}$ given $\hat{\rho}$: $\sum_i \frac{(y_i - \hat{\pi}_i)x_{ij}}{[1+\hat{\rho}(n_i-1)]\hat{\pi}_i(1-\hat{\pi}_i)/n_i} = 0$
2. Use updated $\hat{\beta}$ to solve: $X^2 = \sum_i \frac{(y_i-\hat{\pi}_i)^2}{[1+\hat{\rho}(n_i-1)]\hat{\pi}_i(1-\hat{\pi}_i)/n_i} = n - p$ (Pearson to expected value)

2) Heterogeneity: Mixture Model (Beta-Binomial) Mixture model over π for $s = ny$:

$$s|\pi \sim \text{Bin}(n, \pi)$$

$$\pi \sim \text{Beta}(\alpha_1, \alpha_2)$$

Properties of the Beta distribution:

- PDF: $f(\pi; \alpha_1, \alpha_2) = \frac{\Gamma(\alpha_1+\alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)}\pi^{\alpha_1-1}(1-\pi)^{\alpha_2-1}$ for $\alpha_1, \alpha_2 > 0$
- Shapes: uniform ($\alpha_1 = \alpha_2 = 1$); unimodal symmetric ($\alpha_1 = \alpha_2 > 1$); unimodal skewed left ($\alpha_1 > \alpha_2 > 1$) or right ($\alpha_2 > \alpha_1 > 1$); U-shaped ($\alpha_1, \alpha_2 < 1$)
- Re-parametrization: $\mu = \frac{\alpha_1}{\alpha_1+\alpha_2}$ and $\theta = \frac{1}{\alpha_1+\alpha_2}$
- Moments: $E(\pi) = \mu$ and $\text{var}(\pi) = \mu(1-\mu)\frac{\theta}{1+\theta}$
- **Beta-Binomial**: Marginal of $s = ny$:

$$p(s; n, \mu, \theta) = \binom{n}{s}\frac{\left[\prod_{k=0}^{s-1}(\mu + k\theta)\right]\left[\prod_{k=0}^{n-s-1}(1-\mu+k\theta)\right]}{\prod_{k=0}^{n-1}(1+k\theta)}$$

- Marginal moments: $E(y) = \mu$ and $\text{var}(y) = \left[1+(n-1)\frac{\theta}{1+\theta}\right]\frac{\mu(1-\mu)}{n}$
- Correlation: $\rho = \frac{\theta}{1+\theta}$ is **exactly** the correlation between Bernoulli trials
- Model: assume θ identical for all observations; say $n_i y_i \sim \text{Beta-Bin}(n_i, \mu_i, \theta)$ then use **logit link**: $\text{logit}(\mu_i) = \mathbf{x}_i\beta$ (can use Newton-Raphson, but Beta-Bin **not** in EDF!)
- If not actually Beta-Binomial, estimates $\hat{\beta}$ are **not robust** or consistent.

8.3 Model Misspecification and Robust Estimation

Unlike Beta-Binomial mixture model, QL methods **are robust** to model misspecification!

Estimating Equations The **quasi-score / estimating equations** are:

$$\boxed{\mathbf{u}(\beta) = \sum_{i=1}^{n}\left(\frac{\partial\mu_i}{\partial\beta}\right)^T\frac{y_i-\mu_i}{v(\mu_i)} = \mathbf{0}}$$

i.e. using the fact that $\frac{\partial\mu_i}{\partial\beta_j} = \frac{\partial\mu_i}{\partial\eta_i}x_{ij}$.

Quasi-score function $u_j(\beta)$ is an **unbiased estimating function** because $E[u_j(\beta)] = 0$. For unbiased estimating function, the estimating equations yield estimator $\hat{\beta}$.

Quasi-Likelihood Properties QL treats quasi-score $\mathbf{u}(\beta)$ as derivative of quasi-log-likelihood function, which yields nice properties like ML:

- If μ_i, $v(\mu_i)$ are correct, then QL estimators $\hat{\beta}$ are asymptotically efficient for estimators locally linear in y_i
- $\hat{\beta}$ are asymptotically normal with $\mathbf{V} \approx \left[\sum_{i=1}^n \left(\frac{\partial \mu_i}{\partial \beta}\right)^T [v(\mu_i)]^{-1} \left(\frac{\partial \mu_i}{\partial \beta}\right)\right]^{-1}$
- **Key result**: $\hat{\beta}$ are **consistent** for β even if $v(\mu_i)$ is misspecified! (as long as link function + linear predictor are correct)

Robust Covariance Estimation: Sandwich Matrix Generally, $\text{var}(y_i) \neq \mathbf{v}(\mu_i)$; then the asymptotic $\mathbf{V}$ is incorrect. To find $\text{var}(\beta)$, use Taylor expansion of $\mathbf{u}(\beta)$: $\mathbf{u}(\hat{\beta}) \approx \mathbf{u}(\beta) + \frac{\partial \mathbf{u}(\beta)}{\partial \beta}(\hat{\beta} - \beta)$ and since $\mathbf{u}(\hat{\beta}) = \mathbf{0}$ by definition, $(\hat{\beta} - \beta) \approx -\left(\frac{\partial \mathbf{u}(\beta)}{\partial \beta}\right)^{-1} \mathbf{u}(\beta)$ so that $\text{var}(\hat{\beta}) \approx \left(\frac{\partial \mathbf{u}(\beta)}{\partial \beta}\right)^{-1} \text{var}[\mathbf{u}(\beta)] \left(\frac{\partial \mathbf{u}(\beta)}{\partial \beta}\right)^{-1}$.

But $\left(\frac{\partial \mathbf{u}(\beta)}{\partial \beta}\right)$ is Hessian of quasi-log-likelihood, so symmetric and $-\left(\frac{\partial \mathbf{u}(\beta)}{\partial \beta}\right)^{-1} = \mathbf{V}$ is inverse information matrix for specified model; and

$\text{var}[\mathbf{u}(\beta)] = \text{var}\left[\sum_{i=1}^n \left(\frac{\partial \mu_i(\beta)}{\partial \beta}\right)^T \frac{y_i - \mu_i}{v(\mu_i)}\right] = \sum_{i=1}^n \left(\frac{\partial \mu_i(\beta)}{\partial \beta}\right)^T \frac{\text{var}(y_i)}{[v(\mu_i)]^2} \left(\frac{\partial \mu_i(\beta)}{\partial \beta}\right)$ and so:

$$\text{var}(\hat{\beta}) \approx \mathbf{V}\left[\sum_{i=1}^n \left(\frac{\partial \mu_i(\beta)}{\partial \beta}\right)^T \frac{\text{var}(y_i)}{[v(\mu_i)]^2} \left(\frac{\partial \mu_i(\beta)}{\partial \beta}\right)\right]\mathbf{V}$$

which simplifies to $\mathbf{V}$ if $\text{var}(y_i) = v(\mu_i)$. But generally we don't know $\text{var}(y_i)$, so we estimate: $\mu_i \to \hat{\mu}_i$ and $\text{var}(y_i) \to (y_i - \hat{\mu}_i)^2$ and obtain the **sandwich estimator**:

$$\boxed{\text{var}(\hat{\beta}) \approx \hat{\mathbf{V}}\left[\sum_{i=1}^n \left(\frac{\partial \hat{\mu}_i(\beta)}{\partial \beta}\right)^T \frac{(y_i - \hat{\mu}_i)^2}{[v(\hat{\mu}_i)]^2} \left(\frac{\partial \hat{\mu}_i(\beta)}{\partial \beta}\right)\right]\hat{\mathbf{V}}}$$

Sandwich estimator is robust: whether or not $v(\mu_i)$ is correct, n times estimator converges in probability to asymptotic covariance matrix of $\sqrt{n}(\hat{\beta} - \beta)$!

Example: Poisson Misspecification: Suppose model $y_i \sim \text{Pois}(\mu_i)$, but actually $\text{var}(y_i) = \mu_i^2$; consider null model $\mu_i = \beta \Rightarrow \frac{\partial \mu_i}{\partial \beta} = 1$, so: $u(\beta) = \sum_{i=1}^n \left(\frac{\partial \mu_i}{\partial \beta}\right)[v(\mu_i)]^{-1}(y_i - \mu_i) = \sum_{i=1}^n \frac{y_i - \mu_i}{\mu_i} = \sum_{i=1}^n \frac{y_i - \beta}{\beta} = 0$ so $\hat{\beta} = \bar{y}$ and model-based variance is: $V = \left[\sum_{i=1}^n \left(\frac{\partial \mu_i}{\partial \beta}\right)[v(\mu_i)]^{-1}\left(\frac{\partial \mu_i}{\partial \beta}\right)\right]^{-1} = \frac{\beta}{n}$ so that $\hat{V} = \frac{\bar{y}}{n}$.

The true variance of $\hat{\beta}$ using $\text{var}(y_i) = \mu_i^2$ is: $\frac{\beta^2}{n} = \frac{\bar{y}^2}{n}$ which is different when $\bar{y} > 1$. The robust sandwich estimator (since we don't know $\text{var}(y_i)$) is, using $\mu_i = \beta = \bar{y}$, $\sum_i \frac{(y_i - \bar{y})^2}{n^2}$

9 Correlated Data

Possible cases: 1) Survey asks for opinions on related questions/topics, so answers will be correlated; 2) Clinical trial observes same subjects over time, and measurements from each time point are correlated.

Notation: $\mathbf{y}_i = (y_{i1}, \ldots, y_{id})$, i.e. each subject i has cluster of d obs (i.e. one subject observed over d time points); $\mathbf{x}_{ij}$ is row vector of p explanatory variables for y_{ij}; $\mu_{ij} = E(y_{ij})$.

Two types of models: 1) **marginal model** (model each marginal y_{ij} and use correlation structure for SE); 2) **generalized linear mixed model** (model entire cluster, using random effect for each cluster)

Two types of effects: 1) **between-subject** (between-cluster); 2) **within-subject** (within-cluster).

Example: 2×2 **Design.** Suppose treatments A, B given at times $1, 2$ ($d = 2$); treatment = between-subjects, time = within-subjects. (y_{i1}^A, y_{i2}^A) and (y_{i1}^B, y_{i2}^B) are for subject i in A or B. Let $\text{corr}(y_{i1}^X, y_{i2}^X) = \rho$ and $\text{corr}(y_{it}^A, y_{ju}^B) = 0$, $\text{var}(y_{it}^A) = \text{var}(y_{it}^B) = \sigma^2$. Let $\bar{y}_t^A = \frac{1}{n}\sum_{i=1}^n y_{it}^A$ and $\bar{y}_t^B = \frac{1}{n}\sum_{i=1}^n y_{it}^B$. Then between-subjects effect is $b = \frac{\bar{y}_1^A + \bar{y}_2^A}{2} - \frac{\bar{y}_1^B + \bar{y}_2^B}{2}$ and within-subjects effect is $w = \frac{\bar{y}_1^A + \bar{y}_1^B}{2} - \frac{\bar{y}_2^A + \bar{y}_2^B}{2}$. Then we have $\text{var}(b) = \frac{\sigma^2(1+\rho)}{n}$ and $\text{var}(w) = \frac{\sigma^2(1-\rho)}{n}$, but if we assume independence than they are both $\frac{\sigma^2}{n}$, so standard errors are too small for $\text{var}(b)$ and too large for $\text{var}(w)$.

9.1 Marginal Models and GLMMs

Marginal Model $\boxed{g(\mu_{ij}) = \mathbf{x}_{ij}\beta}$ for all $i = 1, \ldots, n$ and $j = 1, \ldots, d$ (for ***between-cluster effects***)

i.e. models marginal distribution of each y_{ij}, so GLM structure for each y_{ij}.

Example: y_{ij} is score on test j for student i, with GPA x_i, so then $\beta = (\beta_{01}, \beta_{11}, \ldots, \beta_{0d}, \beta_{1d})$ and $\mathbf{x}_{ij} = (0, 0, \ldots, 1, x_i, \ldots, 0, 0)$

GLMM $\boxed{g[E(y_{ij}|\mathbf{u}_i)] = \mathbf{x}_{ij}\beta + \mathbf{z}_{ij}\mathbf{u}_i}$ for $i = 1, \ldots, n$ and $j = 1, \ldots, d$ (for ***within-cluster effects***)

β are **fixed effects** (constant) and $\mathbf{u}_i$ are **random effects** (has probability distribution)

Generally $\mathbf{u}_i \sim \mathcal{N}(\mathbf{0}, \Sigma_{\mathbf{u}})$ i.i.d.; common $\mathbf{u}_i$ for all j, which leads to correlation; given conditional of $(y_{i1}, \ldots, y_{id})|\mathbf{u}_i$, distribution is specified for $\mathbf{y}$.

Intuition: β must apply to **all** subjects identically if they have the same values of the explanatory variables $\mathbf{x}$; but random effects apply to each individual differently while preserving model parsimony (if we wanted to include $\mathbf{u}_i$ as fixed effect, we'd have to have a separate parameter for each person, so $p \propto n$, while now we only have $\Sigma_{\mathbf{u}}$ added); $\mathbf{u}_i$ variability reflects that different subjects with identical $\mathbf{x}_i$ may be heterogeneous due to unobserved variables.

Example: Random-Intercepts Model. Let $\mathbf{z}_{ij}\mathbf{u}_i = u_i$, i.e. add a random intercept. If y_{ij} is score on exam j and x_i = GPA, then: $E(y_{ij}|u_i) = \beta_{0j} + \beta_{1j}x_i + u_i = (\beta_{0j} + u_i) + \beta_{1j}x_i$ which adds separate intercept $\beta_{0j} + u_i$ for each subject!

Example: Matched-Pairs, Binary-Normal Model. Let (y_{i1}, y_{i2}) be matched pair of observations for subject i, with success = 1. Compare $P(y_{i1} = 1)$ and $P(y_{i2} = 1)$.

- **Marginal model**: $\text{logit}[P(y_{ij} = 1)] = \beta_0 + \beta_1 x_j$ for $x_1 = 0, x_2 = 1$; ***average*** over all observations and use Binomial; i.e. consider success/failure totals n_{11} (success/success), n_{12} (success/failure), n_{21} (failure/success), n_{22} (failure/failure). β_1 is the log odds ratio comparing success in observation 2 vs. observation 1 (over entire population) so **population-averaged** effect
- **GLMM**: $\text{logit}[P(y_{ij} = 1|u_i)] = \beta_0 + \beta_1 x_j + u_i$; uses ***individual*** contingency table; β_1 is log odds ratio at the individual level so **subject-specific** effect ($\mathbf{u}_i$ basically centers regression at mean of each subject, so β_1 can be steeper to take care of each individual effect)

The population-averaged = subject-specific effect if **identity link**, but not for any other links. For example above, $\hat{\beta}_1^{marginal} = \log \frac{n_{+1}/n_{+2}}{n_{1+}/n_{2+}}$ while $\hat{\beta}_1^{GLMM} = \log \frac{n_{21}}{n_{12}}$

GLMM → Marginal To find the between-cluster effects for GLMM (for which it's not natural), we have to integrate out $\mathbf{u}_i$ using LIE; i.e. $E(y_i) = E[E(y_i|u_i)] = E[g^{-1}(\mathbf{x}_{ij}\beta + \mathbf{z}_{ij}\mathbf{u}_i)]$; leads to exact same marginal model if identity link; different form otherwise

9.2 Normal Linear Mixed Model

Start with simplest, normal linear mixed model: $E(y_{ij}|\mathbf{u}_i) = \mathbf{x}_{ij}\beta + \mathbf{z}_{ij}\mathbf{u}_i$ i.e. $y_{ij} = \mathbf{x}_{ij}\beta + \mathbf{z}_{ij}\mathbf{u}_i + \epsilon_{ij}$ where β is $p \times 1$ vector of fixed effects, $\mathbf{u}_i \sim \mathcal{N}(0, \Sigma_\mathbf{u})$ is $q \times 1$ vector of random effects, $\epsilon_{ij} \sim \mathcal{N}(0, \sigma_e^2)$. Generally, $\mathbf{y}_i = \mathbf{X}_i\beta + \mathbf{Z}_i\mathbf{u}_i + \epsilon_i$ ($\mathbf{X}_i$ is $d \times p$ model matrix, $\mathbf{Z}$ is $d \times q$ model matrix for random effects, $\epsilon_i \sim \mathcal{N}(\mathbf{0}, \sigma_e^2\mathbf{I})$). $E(\mathbf{y}_i|\mathbf{u}_i) = \mathbf{X}_i\beta + \mathbf{Z}_i\mathbf{u}_i$ and $\text{var}(\mathbf{y}_i) = \mathbf{Z}_i\Sigma_\mathbf{u}\mathbf{Z}_i^T + \sigma_e^2\mathbf{I}$.

Random-Intercepts Model: $\mathbf{u}_i = u_i$, $\mathbf{Z}_i = \mathbf{1}$ and $\text{var}(u_i) = \sigma_u^2$. Then $\text{var}(\mathbf{y}_i) = \sigma_u^2\mathbf{1}\mathbf{1}^T + \sigma_e^2\mathbf{I}$ so that $\text{corr}(y_{ij}, y_{ik}) = \frac{\sigma_u^2}{\sigma_u^2+\sigma_e^2}$ for $j \neq k$ (exchangeable/compound symmetry)

9.3 GLMM Fitting and Inference

No closed-form likelihood, so model fitting is difficult.

Marginal Likelihood/Maximum Likelihood GLMM is two-stage: 1) conditional on $\mathbf{u}_i$, fit a GLM with known effect $\mathbf{z}_{ij}\mathbf{u}_i$; 2) $\mathbf{u}_i \sim \mathcal{N}(\mathbf{0}, \Sigma_\mathbf{u})$ so fit parameters.

Marginal likelihood is: To fit likelihood for $\beta, \Sigma_\mathbf{u}$, integrate out random effects:

$$\boxed{Ł(\beta, \Sigma_\mathbf{u}; \mathbf{y}) = f(\mathbf{y}; \beta, \Sigma_\mathbf{u}) = \int f(\mathbf{y}|\mathbf{u}; \beta) f(\mathbf{u}; \Sigma_\mathbf{u}) d\mathbf{u}}$$

Example: Logistic-Normal Random-Intercepts Model.

$$L(\beta, \sigma_u^2; \mathbf{y}) = \prod_{i=1}^{n} \left[\int_{-\infty}^{\infty} \prod_{j=1}^{d} \left(\frac{\exp(\mathbf{x}_{ij}\beta + u_i)}{1 + \exp(\mathbf{x}_{ij}\beta + u_i)} \right)^{y_{ij}} \left(\frac{1}{1 + \exp(\mathbf{x}_{ij}\beta + u_i)} \right)^{1-y_{ij}} f(u_i; \sigma_u^2) du_i \right]$$

Need to approximate this numerically and then maximize: 1) Gauss-Hermite quadrature; 2) Monte-Carlo; 3) Laplace approximation; 4) EM algorithm

GLMM Inference Inference for fixed effects is standard (i.e. LRT for nested models); but for random effects is more complex (because if variance = 0, then on boundary, so likelihood-based inference doesn't work); i.e. $H_0 : \sigma_u^2 = 0$ vs. $H_a : \sigma_u^2 > 0$ has the mixed distribution of $\frac{1}{2}\delta_0 + \frac{1}{2}\chi_1^2$ so the p-value is $\frac{1}{2}P(\chi_1^2 > t_{obs})$

9.4 Marginal Model Fitting and Inference

ML fitting generally only possible for multivariate normal response; if not, we need to use multivariate QL, i.e. GEE.

Multivariate Normal Regression $\mathbf{y}_i = (y_{i1}, \ldots, y_{id})$ and $y_{ij} = \mathbf{x}_{ij}\beta + \epsilon_{ij}$ with $\epsilon_i \sim \mathcal{N}(\mathbf{0}, \mathbf{V}_i)$ so that $\mathbf{y} \sim \mathcal{N}(\mathbf{X}\beta, \mathbf{V})$ where $\mathbf{X}$ is stacked $\mathbf{X}_i$ of dimension $dn \times p$ then we have GLS estimator $\hat{\beta} = (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{y}$

Generalized Estimating Equations (GEE) Lack of discrete distributions that can show correlation structures; use QL-like method, where we specify: 1) $\mu_{ij} = E(y_{ij})$; 2) $v(\mu_{ij})$; 3) **working correlation structure** $\text{corr}(y_{ij}), y_{ik})$. Simple correlation structures:

- Exchangeable: $\text{corr}(y_{ij}, y_{ik}) = \alpha$
- Autoregressive: $\text{corr}(y_{ij}, y_{ik}) = \alpha^{|j-k|}$
- Independent: $\text{corr}(y_{ij}, y_{ik}) = 0$
- Unstructured: $\text{corr}(y_{ij}, y_{ik}) = \alpha_{jk}$

When link function + linear predictor are correct, GEE estimator $\hat{\beta}$ are still consistent for β even if correlation is incorrect. But standard errors are wrong, so we need to use robust sandwich estimator.

Marginal model: $g(\mu_{ij}) = \mathbf{x}_{ij}\beta$; $\mathbf{V}_i$ is working covariance matrix for $\mathbf{y}_i$ based on working correlation matrix $\mathbf{R}(\alpha)$; if $\mathbf{R}(\alpha)$ is true correlation, then $\mathbf{V}_i = \text{var}(\mathbf{y}_i)$. Let $\mathbf{D}_i = \frac{\partial \mu_i}{\partial \beta}$ be $d \times p$ matrix of jk elements $\frac{\partial \mu_{ij}}{\partial \beta_k}$. Recall: univariate QL estimating equations were: $\sum_i \left(\frac{\partial \mu_i}{\partial \beta}\right)^T [v(\mu_i)]^{-1}(y_i - \mu_i) = \mathbf{0}$, so the multivariate analog is **generalized estimating equations**:

$$\boxed{\sum_{i=1}^{n} \mathbf{D}_i^T \mathbf{V}_i^{-1}(\mathbf{y}_i - \mu_i) = \mathbf{0}}$$

GEE estimator $\hat{\beta}$ is solution to GEE equations. Iterated method: 1) estimate β given current estimate of α; 2) estimate α given current estimate of β using moment estimation (pairwise empirical correlation). Then: $(\hat{\beta} - \beta) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathbf{V}_G/n)$ where:

$$\text{var}(\hat{\beta}) \approx \frac{\mathbf{V}_G}{n} \approx \left[\sum_{i=1}^{n} \mathbf{D}_i^T \mathbf{V}_i^{-1} \mathbf{D}_i\right]^{-1} \left[\sum_{i=1}^{n} \mathbf{D}_i^T \mathbf{V}_i^{-1} [\text{var}(\mathbf{y}_i)] \mathbf{V}_i^{-1} \mathbf{D}_i\right] \left[\sum_{i=1}^{n} \mathbf{D}_i^T \mathbf{V}_i^{-1} \mathbf{D}_i\right]^{-1}$$

Estimated sandwich matrix $\hat{\mathbf{V}}_G/n$ for $\hat{\beta}$ replaces $\beta \to \hat{\beta}$, $\phi \to \hat{\phi}$, $\alpha \to \hat{\alpha}$, and $\text{var}(\mathbf{y}_i) \to (\mathbf{y}_i - \hat{\mu}_i)(\mathbf{y}_i - \hat{\mu}_i)^T$

Disadvantages of GEE approach:

1. No likelihood: can't do likelihood methods (i.e. LRT, deviance) for fit, model comparison, inference
2. Categorical data: "correlation" not really natural for discrete data
3. Stronger missing data assumption: compared to ML, strong missing data; GEE must have MCAR, but ML only requires MAR

Important Formulae

$$E[\mathbf{y}^T\mathbf{A}\mathbf{y}] = \text{trace}(\mathbf{AV}) + \mu^T\mathbf{A}\mu$$

$$\frac{\partial(\mathbf{a}^T\beta)}{\partial\beta} = \mathbf{a}$$

$$\frac{\partial(\beta^T\mathbf{A}\beta)}{\partial\beta} = (\mathbf{A} + \mathbf{A}^T)\beta$$

Likelihood results: for log-likelihood l:

$$E\left(\frac{\partial l}{\partial\theta}\right) = 0$$

$$-E\left(\frac{\partial^2 l}{\partial\theta^2}\right) = E\left(\frac{\partial l}{\partial\theta}\right)^2$$

$$-E\left(\frac{\partial^2 l_i}{\partial\beta_j\partial\beta_k}\right) = E\left[\left(\frac{\partial l_i}{\partial\beta_j}\right)\left(\frac{\partial l_i}{\partial\beta_k}\right)\right]$$