THE AITKEN MODEL

- $\boldsymbol{y} = X\beta + \boldsymbol{\epsilon},\ \boldsymbol{\epsilon} \sim (\mathbf{0},\ \sigma^2 V)$
- Identical to the Gauss-Markov Linear Model except that Var $(\epsilon) = \sigma^2 V$ instead of $\sigma^2 I$.
- V is assumed to be a **known** nonsingular Variance matrix.
- The Normal Theory Aitken Model adds an assumption of normality: $\boldsymbol{\epsilon} \sim N(\mathbf{0},\ \sigma^2 V)$

- Observations are now correlated, or have unequal variances.
- but the correlations, or unequal variances, follow a known pattern

Examples - 1

- Analysis of averages

  The data to be analyzed are averages of unequal numbers of observations.
  Y_i is an average of n_i observations. Var $Y_i = \sigma^2/n_i$
  first 4 rows and columns of Var ϵ are:

$$\begin{bmatrix} \sigma^2/n_1 & 0 & 0 & 0 \\ 0 & \sigma^2/n_2 & 0 & 0 \\ 0 & 0 & \sigma^2/n_3 & 0 \\ 0 & 0 & 0 & \sigma^2/n_4 \end{bmatrix} = \sigma^2 \begin{bmatrix} 1/n_1 & 0 & 0 & 0 \\ 0 & 1/n_2 & 0 & 0 \\ 0 & 0 & 1/n_3 & 0 \\ 0 & 0 & 0 & 1/n_4 \end{bmatrix}$$

Examples - 2

- Analysis of data on a pedigree (genetic relationships among parents, children, grandchildren, ...)
- Genetic correlations between parents and children, among children, ..., all known.

$$\boldsymbol{Y} = \boldsymbol{X}\beta + \boldsymbol{\epsilon}$$

  First four rows of Var ϵ:

$$\sigma^2 \begin{bmatrix} 1 & \rho_{12} & \rho_{13} & \rho_{14} \\ \rho_{12} & 1 & \rho_{23} & \rho_{24} \\ \rho_{13} & \rho_{23} & 1 & \rho_{34} \\ \rho_{14} & \rho_{24} & \rho_{34} & 1 \end{bmatrix}$$

  where ρ_{ij} is the known genetic correlation among individuals i, j

Examples - 3

- Regression on data collected over time: $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$
  X_i = year (1990, 1991, ...)
- Assume errors follow an autoregressive process (more later), first 4 rows and columns of Var ϵ are:

$$\begin{bmatrix} \sigma^2 & \sigma^2\rho & \sigma^2\rho^2 & \sigma^2\rho^3 \\ \sigma^2\rho & \sigma^2 & \sigma^2\rho & \sigma^2\rho^2 \\ \sigma^2\rho^2 & \sigma^2\rho & \sigma^2 & \sigma^2\rho \\ \sigma^2\rho^3 & \sigma^2\rho^2 & \sigma^2\rho & \sigma^2 \end{bmatrix} = \sigma^2 \begin{bmatrix} 1 & \rho & \rho^2 & \rho^3 \\ \rho & 1 & \rho & \rho^2 \\ \rho^2 & \rho & 1 & \rho \\ \rho^3 & \rho^2 & \rho & 1 \end{bmatrix}$$

- Aitken model if ρ known.

Statistical analysis of Aitken model data

- Spectral Decomposition Theorem:
  - any positive definite symm. matrix $\boldsymbol{V}$ can be written as $\boldsymbol{V} = \boldsymbol{UDU}'$
  - $\boldsymbol{D}$ is a diagonal matrix of eigenvalues
  - $\boldsymbol{U}$ is a matrix of orthonormal eigenvectors with the property that $\boldsymbol{U}'\boldsymbol{U} = \boldsymbol{I}$.
- For any positive definite symmetric $\boldsymbol{V}$, there exists a nonsingular symmetric matrix $V^{1/2}$ such that $V^{1/2}\,V^{1/2} = V$.
- Given $\boldsymbol{U}$ and $\boldsymbol{D}$ for which $\boldsymbol{UDU}' = \boldsymbol{V}$, $V^{1/2} = \boldsymbol{U}\sqrt{\boldsymbol{D}}\boldsymbol{U}'$
- $V^{1/2}$ can be viewed as the "square root" of a matrix
- $V^{1/2}\,V^{1/2} = \boldsymbol{U}\sqrt{\boldsymbol{D}}\boldsymbol{U}'\boldsymbol{U}\sqrt{\boldsymbol{D}}\boldsymbol{U}' = \boldsymbol{U}\sqrt{\boldsymbol{D}}\boldsymbol{I}\sqrt{\boldsymbol{D}}\boldsymbol{U}' = \boldsymbol{UDU}' = \boldsymbol{V}$
- Define $V^{-1/2}$ as $(V^{-1})^{1/2}$
- Compute $V^{-1/2}$ by $\boldsymbol{U}(1/\sqrt{\boldsymbol{D}})\boldsymbol{U}'$,
  where $1/\sqrt{\boldsymbol{D}}$ is the diagonal matrix with elements $1/\sqrt{D_{ii}}$

Converting an Aitken model to GM model

- Our data model is $\boldsymbol{y} = \boldsymbol{X}\beta + \epsilon$, $\epsilon \sim (0, \sigma^2\boldsymbol{V})$
- Pre-multiple all terms by $\boldsymbol{V}^{-1/2}$
  $\boldsymbol{V}^{-1/2}\boldsymbol{y} = \boldsymbol{V}^{-1/2}\boldsymbol{X}\beta + \boldsymbol{V}^{-1/2}\epsilon$.
- This is a regression model, $\boldsymbol{Z} = \boldsymbol{W}\beta + \delta$ with
  $\boldsymbol{Z} = \boldsymbol{V}^{-1/2}\boldsymbol{y}, \qquad \boldsymbol{W} = \boldsymbol{V}^{-1/2}\boldsymbol{X}, \qquad \delta = \boldsymbol{V}^{-1/2}\epsilon.$
- Why do this? What is Var δ? $Var(\delta) = Var(\boldsymbol{V}^{-1/2}\epsilon)$
  $= \boldsymbol{V}^{-1/2}\sigma^2\boldsymbol{V}\boldsymbol{V}^{-1/2}$
  $= \sigma^2\boldsymbol{V}^{-1/2}\boldsymbol{V}^{1/2}\boldsymbol{V}^{1/2}\boldsymbol{V}^{-1/2} = \sigma^2\boldsymbol{I}$.
- After transformation, we have a Gauss-Markov Model!

Generalized Least Squares

- $\hat\beta_G = (\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^{-}\boldsymbol{V}^{-1}\boldsymbol{y}$ is a solution to the Aitken Equations:
  $$\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X}\boldsymbol{b} = \boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{y}$$
  which follow from the Normal Equations
  $$\begin{aligned} \boldsymbol{W}'\boldsymbol{W}\boldsymbol{b} &= \boldsymbol{W}'\boldsymbol{Z} \\ \Rightarrow \boldsymbol{X}'\boldsymbol{V}^{-1/2}\boldsymbol{V}^{-1/2}\boldsymbol{X}\boldsymbol{b} &= \boldsymbol{X}'\boldsymbol{V}^{-1/2}\boldsymbol{V}^{-1/2}\boldsymbol{y} \\ \Rightarrow \boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X}\boldsymbol{b} &= \boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{y}. \end{aligned}$$
- Solving the Normal Equations is equivalent to minimizing
  $$(\boldsymbol{Z} - \boldsymbol{W}\boldsymbol{b})'(Z - \boldsymbol{W}\boldsymbol{b}) \textit{ over } \boldsymbol{b} \in I\!R^p$$
- Now
  $(\boldsymbol{Z} - \boldsymbol{W}\boldsymbol{b})'(\boldsymbol{Z} - \boldsymbol{W}\boldsymbol{b}) = (\boldsymbol{V}^{-1/2}\boldsymbol{y} - \boldsymbol{V}^{-1/2}\boldsymbol{X}\boldsymbol{b})'(\boldsymbol{V}^{-1/2}\boldsymbol{y} - \boldsymbol{V}^{-1/2}\boldsymbol{X}\boldsymbol{b})$
  $= (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b})'\boldsymbol{V}^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b})$
- Thus, $\hat\beta_G = (\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^{-}\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{y}$ is a solution to this Generalized Least Squares problem.

Estimating Var $\hat\beta_G$ for full rank $\boldsymbol{X}$

- If $\boldsymbol{X}$ full rank, $\hat\beta_G = (\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{y}$

$$\begin{aligned} \text{Var } \hat\beta_G &= \text{E}\,(\hat\beta_G - \beta_G)(\hat\beta_G - \beta_G)' \\ &= \text{E}\,((\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{y} - (\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{V}^{-1}\,\text{E}\,\boldsymbol{y}) \times \\ &\quad ((\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{y} - (\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{V}^{-1}\,\text{E}\,\boldsymbol{y})' \\ &= ((\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{V}^{-1}(\boldsymbol{y} - \text{E}\,\boldsymbol{y})((\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{V}^{-1}(\boldsymbol{y} - \text{E}\,\boldsymbol{y}))' \\ &= ((\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{V}^{-1}(\boldsymbol{y} - \text{E}\,\boldsymbol{y})(\boldsymbol{y} - \text{E}\,\boldsymbol{y})'(\boldsymbol{V}^{-1}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^{-1} \\ &= \sigma^2((\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^{-1} \\ &= \sigma^2(\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^{-1} \end{aligned}$$

- Similar to OLS Var $\hat\beta$, except for extra $\boldsymbol{V}^{-1}$ in middle
- Var $\boldsymbol{C}\hat\beta_G = \boldsymbol{C}$ Var $\hat\beta\,\boldsymbol{C}'$
- E $\boldsymbol{y} = \boldsymbol{X}\hat\beta_G$
- Var $\hat{\boldsymbol{y}} = \sigma^2\boldsymbol{X}$ Var $\hat\beta_G\,\boldsymbol{X}'$

When $\boldsymbol{X}$ is not full rank:

- estimate E $\boldsymbol{y}$:
- Under the GM model, the best estimate of E $\boldsymbol{Z}$ is

$$\begin{aligned}\hat{\boldsymbol{Z}} &= P_W\boldsymbol{Z} = \boldsymbol{W}(\boldsymbol{W}'\boldsymbol{W})^-\boldsymbol{W}'\boldsymbol{Z} \\ &= \boldsymbol{V}^{-1/2}\boldsymbol{X}((\boldsymbol{V}^{-1/2}\boldsymbol{X})'\boldsymbol{V}^{-1/2}\boldsymbol{X})^-(\boldsymbol{V}^{-1/2}\boldsymbol{X})'\boldsymbol{V}^{1/2}\boldsymbol{y} \\ &= \boldsymbol{V}^{-1/2}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{V}^{-1/2}\boldsymbol{V}^{-1/2}\boldsymbol{X})^-\boldsymbol{X}'\boldsymbol{V}^{-1/2}\boldsymbol{V}^{-1/2}\boldsymbol{y} \\ &= \boldsymbol{V}^{-1/2}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^-\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{y}.\end{aligned}$$

- To estimate $E(\boldsymbol{y}) = \boldsymbol{V}^{1/2}E(\boldsymbol{Z})$, use

$$\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^-\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{y} = \boldsymbol{X}\hat{\beta}_G.$$

Estimating $\boldsymbol{C}\beta$ when $\boldsymbol{X}$ not full rank

- $\boldsymbol{C}\beta$ is estimable if $\boldsymbol{C}$ can written as a linear combination of rows of $\boldsymbol{W}$, i.e. $\boldsymbol{C} = \boldsymbol{A}\boldsymbol{W}$
- if $\boldsymbol{C}\beta$ estimable, the BLUE is the ordinary least squares (OLS) estimator using $\boldsymbol{Z}$ and $\boldsymbol{W}$

$$\begin{aligned}\boldsymbol{C}\hat{\beta} &= \boldsymbol{C}(\boldsymbol{W}'\boldsymbol{W})^-\boldsymbol{W}'\boldsymbol{Z} \\ &= \boldsymbol{C}(\boldsymbol{X}'\boldsymbol{V}^{-1/2}\boldsymbol{V}^{-1/2}\boldsymbol{X})^-\boldsymbol{X}'\boldsymbol{V}^{-1/2}\boldsymbol{V}^{-1/2}\boldsymbol{y} \\ &= \boldsymbol{C}(\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^-\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{y}.\end{aligned}$$

- $\boldsymbol{C}(\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^-\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{y} = \boldsymbol{C}\hat{\beta}_G$ is called a Generalized Least Squares (GLS) estimator.

Var $\boldsymbol{C}\hat{\beta}_G$

- Need to be careful because Var $\hat{\beta}_G$ doesn't exist

$$\begin{aligned}\text{Var } \boldsymbol{C}\hat{\beta}_G &= \text{Var } \boldsymbol{C}(\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^-\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{y} \\ &= \text{Var } \boldsymbol{C}(\boldsymbol{W}'\boldsymbol{W})^-\boldsymbol{W}'\boldsymbol{Z} \\ &= \text{Var } \boldsymbol{A}\boldsymbol{W}(\boldsymbol{W}'\boldsymbol{W})^-\boldsymbol{W}'\boldsymbol{Z} \\ &= \text{Var } \boldsymbol{A}\boldsymbol{P}_W\boldsymbol{Z} \\ &= \boldsymbol{A}\boldsymbol{P}_W\text{Var } \boldsymbol{Z}\boldsymbol{P}_W\boldsymbol{A}' \\ &= \boldsymbol{A}\boldsymbol{P}_W(\sigma^2\boldsymbol{I})\boldsymbol{P}_W\boldsymbol{A}' \\ &= \sigma^2\boldsymbol{A}\boldsymbol{P}_W\boldsymbol{P}_W\boldsymbol{A}' = \sigma^2\boldsymbol{A}\boldsymbol{P}_W\boldsymbol{A}' \\ &= \sigma^2\boldsymbol{A}\boldsymbol{W}(\boldsymbol{W}'\boldsymbol{W})^-\boldsymbol{W}'\boldsymbol{A}' \\ &= \sigma^2\boldsymbol{C}(\boldsymbol{W}'\boldsymbol{W})^-\boldsymbol{C}' \\ &= \sigma^2\boldsymbol{C}(\boldsymbol{X}'\boldsymbol{V}^{-1/2}\boldsymbol{V}^{-1/2}\boldsymbol{X})^-\boldsymbol{C}' \\ &= \sigma^2\boldsymbol{C}(\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^-\boldsymbol{C}'\end{aligned}$$

Weighted Least Squares

- When $\boldsymbol{V}$ is diagonal, the term "Weighted Least Squares" (WLS) is commonly used instead of GLS.
- Define $\boldsymbol{D}$ = diagonal matrix of inverse weights, $D_{ii} = 1/w_i$
- $(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b})'\boldsymbol{D}^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}) = \sum_{i=1}^{n} w_i\,(y_i - \boldsymbol{X}_i\beta)^2$
- When obs. have unequal variances, i.e. Var $\boldsymbol{y} = \text{diag}\left(\sigma_i^2\right)$, w_i is proportional to $1/\sigma_i^2$
- WLS assumes weights are known.
- If weights are estimated, e.g. using s_i^2, WLS analysis is in trouble (when small d.f. for each s_i^2) or approximate (when large d.f. for each s_i^2).

Summary of Aitken model results

- Aitken model: $\boldsymbol{y} = \boldsymbol{X}\beta + \epsilon$, $\epsilon \sim (0, \sigma^2 \boldsymbol{V})$
- equivalent to GM model: $\boldsymbol{Z} = \boldsymbol{W}\beta + \delta$, by premultiplying by $\boldsymbol{V}^{-1/2}$
- estimate $E(\boldsymbol{y}) = \boldsymbol{V}^{-1/2}E(\boldsymbol{Z})$ by $\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^{-}\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{y}$.
- estimate $\boldsymbol{C}\beta$ by $\boldsymbol{C}(\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^{-}\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{y}$
- estimate Var $\boldsymbol{C}\beta$ by $\hat{\sigma}^2\boldsymbol{C}(\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^{-}\boldsymbol{C}'$
- estimate Var δ by $\hat{\sigma}^2 = \frac{(\boldsymbol{y}-\boldsymbol{X}\hat{\beta}_G)'\boldsymbol{V}^{-1}(\boldsymbol{y}-\boldsymbol{X}\hat{\beta}_G)}{n-k}$
- which means estimating Var $\boldsymbol{y}$ by $\hat{\sigma}^2\boldsymbol{V}$
- If add normality, all inferential results from Stat 500 follow
- In particular, df. for $\hat{\sigma}^2 = N - \text{rank } \boldsymbol{X}$

What if $\boldsymbol{V}$ misspecified?

- Data follow Aitken model, $\boldsymbol{y} = \boldsymbol{X}\beta + \epsilon$, $\epsilon \sim (0, \sigma^2_G V)$, but analyzed using $\boldsymbol{y} = \boldsymbol{X}\beta + \epsilon$, $\epsilon \sim (0, \sigma^2 \boldsymbol{I})$?
- Results stated; derivations can be found in Kutner et al. (big white) or most Econometrics textbooks.
- E(y) unbiased, so $\boldsymbol{C}\hat{\beta}$ unbiased
- OLS estimates not as efficient as GLS, Var $\boldsymbol{C}\hat{\beta} >$ Var $\boldsymbol{C}\hat{\beta}_G$
- Var $\boldsymbol{C}\hat{\beta}$ **is not** $\sigma^2\boldsymbol{C}(\boldsymbol{X}'\boldsymbol{X})^{-}\boldsymbol{C}$. Bias can be considerable.
- Assume $\boldsymbol{X}$ full rank, so $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ exists

$$\begin{aligned}
\text{Var } \boldsymbol{C}\hat{\beta} &= \text{Var } \boldsymbol{C}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} = \boldsymbol{C}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}' \text{ (Var } \boldsymbol{y}) \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{C}' \\
&= \boldsymbol{C}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'(\sigma^2_G\boldsymbol{V})\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{C}' \\
&= \sigma^2_G\,\boldsymbol{C}(\boldsymbol{X}'\boldsymbol{X})^{-1}(\boldsymbol{X}'\boldsymbol{V}\boldsymbol{X})(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{C}' \qquad (1)
\end{aligned}$$

- Can't simplify!
- All inference is suspect, unless $(\boldsymbol{X}'\boldsymbol{V}\boldsymbol{X})(\boldsymbol{X}'\boldsymbol{X})^{-1}$ close to $\boldsymbol{I}$

Sandwich estimator

- However, equation (1) can be used to advantage!
- Can estimate Var $\boldsymbol{y}$ = Var from the empirical variance-covariance matrix of the residuals, Var $\epsilon = \epsilon\,\epsilon'$, or a modeled version of that.
- If you suspect a problem with homogeneous variances, independent errors, estimate Var $\boldsymbol{C}\hat{\beta}$ using (1).
- called "White's heteroscedastic consistent variance estimator" in econometrics
- Recent statistical literature has called this the "sandwich" estimator, Because of the meat $(\boldsymbol{X}'\boldsymbol{V}\boldsymbol{X})$ between the two slices of bread $(\boldsymbol{X}'\boldsymbol{X})$.

Example data analysis using Aitken model

- Study comparing metabolite concentrations in two genotypes of plants (A, B)
- Seedlings commonly grown in flats, here 9 seedlings per flat
- 4 flats per genotype. All 9 seedlings same genotype.
- Measure size of each seedling 2 weeks after germination.
- review of 500/401: what is the observational unit (ou)? what is the experimental unit?

Analysis - 2

- Seedlings too small to measure metabolite concentration on each.
- Combine together all plants in a flat. One measurement per flat.
- This is a physical average of the metabolite concentration in seedling.
- Goal: estimate mean difference between genotypes in size and in metabolite concentration.
- Problem: some seedlings died. Some responses are averages of 9 seedlings; some are averages of 5 seedlings.
- Standard analyses assume that death unrelated to size or metabolite conc.
- Many ways to model (and hence to analyze) data like this.

Analysis - 3

- For now, only consider metabolite data.
- **A** model (not the model) assumes that the variation among responses is due only to variation among seedlings.
- Y_{ij} is metabolite concentration measured in flat j of genotype i
- this is an average of n_{ij} seedlings.
- If only variation among seedlings, Var $Y_{ij} = \sigma^2/n_{ij}$
- Aitken model with $\boldsymbol{V} = \text{diag}(1/n_{ij})$

Analysis - 4

- The data:

Flat	1	2	3	4	5	6	7	8
Genotype	A	B	B	A	A	B	B	A
# seedlings	9	6	7	9	6	6	8	5
Metabolite	.0231	.0275	.0521	.0001	.0063	.0138	.1061	.0482

- The results:

	GLS		OLS	
Parameter	est.	s.e.	est.	s.e.
Group A mean	0.0168	0.0159	0.0194	0.0163
Difference	0.0373	0.0229	0.0305	0.0230
rMSE = $\hat{\sigma}$	0.085		0.032	

- Heteroscedasticity not very large (max Var / min Var) < 2
- Estimates similar, $\hat{\sigma}^2$ very different!

Analysis - 5

- Different parameters: Aitken model: Var $Y_{ij} = \sigma_G^2/n_i$, GM model: Var $Y_{ij} = \sigma^2$
- σ_G^2 is the variance among measurements of individual seedlings
- σ^2 from OLS is variance among flat means
- In fact, average $\hat{\sigma}_G^2/n_i = 0.085^2\ 0.1487 \approx 0.032^2$
- Very dependent on assumption of no variation other than seedling-seedling
- Assumes no additional variation among flats
- Common experience is that there is both flat-flat variation and seedling-seedling variation.
- Accounting for both requires a mixed model (coming up).
- Can estimate both variance components even though did not measure individual seedlings

R code for the Aitken model and weighted LS

```
# the metabolite example
# enter the data (genotype, # seedlings, and ave. conc.
genotype <-c('A','B','B','A','A','B','B','A')
gen <- as.factor(genotype)

nsdl <- c(9,6,7,9,6,6,8,5)
metab <- c(0.0231,0.0275,0.0521,0.0001,0.0063,0.0138,
  0.1061,0.0482)

# the ols analysis
ols.lm <- lm(metab~gen)
summary(ols.lm)
```



```
# an Aitken model analysis, by hand
# v is an 8 x 8 matrix with 1/nsdl on the diagonal
v <- diag(1/nsdl)
# diag(vector) creates a matrix with vector on diag
# diag(matrix) extracts the diagonal

# need to get the inverse square root matrix of v
#  use eigen function to do that

temp <- eigen(v)
# returns a list with two components: values, a vector,
#    and vectors, a matrix
```



```
u <- temp$vectors
d <- diag(temp$values)
# convert e-vals to a diagonal matrix

round(v - u %*% d %*% t(u), 5)  # check u d t' = v
all(v == u %*% d %*% t(u))
# another possible check, more sensitive to num. error

svi <- u %*% diag(1/sqrt(temp$values)) %*% t(u)
#  inverse square root matrix

# a check that svi does what we want it to do
v %*% svi %*% t(svi)
# svi svi' = v^{-1}, and v * v^{-1} = I
```



```
# use svi to transform y, and X
#  notation follows that in notes
z <- svi %*% metab
w <- svi %*% model.matrix(ols.lm)
  # could also use any other equivalent X matrix
  # w includes a column for intercept

gls.lm <- lm(z~-1+w) # need to suppress default intercept
summary(gls.lm)
```

```
# could also do this analysis using weighted least
#    squares, since V diagonal
#   weights are 1/variance= nsdl

wls.lm <- lm(metab~gen, weight=nsdl)
summary(wls.lm)

# some useful hints / tricks to construct different
#    sorts of V matrices

# AR(1) structure.  V has bands.
#   This sort of matrix is a toeplitz matrix

toeplitz(1:5)
```



```
# to generate a 8 x 8 ar(1) matrix with specified rho
rho <- 0.7
v <- toeplitz(rho^(0:7))
# sequence starts at 0 so diag = 1, ends at rho^7

# compound symmetry structure (obs within groups)
# we haven't yet talked about this model,
# it's here because it fits here, but you won't
#    need it for a while.
grp <- c(1,1,2,2,3,3,4,4,4)
# or use rep(1:4,c(2,2,2,3))
v <- outer(grp, grp, '==')
# true (i.e. 1) if grp[i] = grp[j]
k <- 2/1.5
# ratio of sigma^2_p / sigma^2_c
v <- v + diag(rep(k,length(grp)))
# add k to diag., also coerces T/F to 1/0 (good)
```


Maximum Likelihood

- Suppose $f(\boldsymbol{w}|\boldsymbol{\theta})$ is the probability density function (*pdf*) or probability mass function (*pmf*) of a random vector $\boldsymbol{w}$, where $\boldsymbol{\theta}$ is a $k \times 1$ vector of parameters.
- Given a value of the parameter vector $\boldsymbol{\theta}$, $f(\boldsymbol{w}|\boldsymbol{\theta})$ is a real-valued function of $\boldsymbol{w}$.
- The likelihood function $L(\boldsymbol{\theta}|\boldsymbol{w}) = f(\boldsymbol{w}|\boldsymbol{\theta})$ is a real-valued function of $\boldsymbol{\theta}$ for a given value of $\boldsymbol{w}$.
- $L(\boldsymbol{\theta}|\boldsymbol{w})$ is not a pdf. $\int_\theta L(\boldsymbol{\theta}|\boldsymbol{w}) d\theta \neq 1$.

- For any potential observed vector of values $\boldsymbol{w}$, define $\hat{\theta}(\boldsymbol{w})$ to be a parameter value at which $L(\boldsymbol{\theta}|\boldsymbol{w})$ attains its maximum value. $\hat{\theta}(\boldsymbol{w})$ is a <u>maximum likelihood estimator</u> (MLE) of $\boldsymbol{\theta}$.
- Invariance property of MLES: The MLE of a function of $\boldsymbol{\theta}$, say $\boldsymbol{g}(\boldsymbol{\theta})$, is the function evaluated at the MLE of $\boldsymbol{\theta}$: $\widehat{\boldsymbol{g}(\boldsymbol{\theta})} = \boldsymbol{g}(\hat{\boldsymbol{\theta}})$
- Often much more convenient to work with $l(\boldsymbol{\theta}|\boldsymbol{w}) = \log L(\boldsymbol{w}|\boldsymbol{\theta})$.
- If $l(\boldsymbol{\theta}|\boldsymbol{w})$ is differentiable, candidates for the MLE of $\boldsymbol{\theta}$ can be found by equating the score function

$$\frac{\partial l(\boldsymbol{\theta}|\boldsymbol{w})}{\partial \boldsymbol{\theta}} \equiv \begin{bmatrix} \frac{\partial l(\boldsymbol{\theta}|\boldsymbol{w})}{\partial \boldsymbol{\theta}} \\ \vdots \\ \frac{\partial l(\boldsymbol{\theta}|\boldsymbol{w})}{\partial \boldsymbol{\theta}_k} \end{bmatrix} \text{ to } \mathbf{0} \text{ and solving for } \boldsymbol{\theta}$$

- The score equations are $\frac{\partial l(\boldsymbol{\theta}|\boldsymbol{w})}{\partial \boldsymbol{\theta}} = \mathbf{0} \Rightarrow \frac{\partial l(\boldsymbol{\theta}|\boldsymbol{w})}{\partial \theta_j} = 0 \ \forall\, j = 1, ..., k$
- One strategy for obtaining an MLE is to find solution(s) of the score equations and verify that at least one such solution maximizes $l(\boldsymbol{\theta}|\boldsymbol{w})$.
- If the solution(s) to the score equations lie outside the appropriate parameter space, they are not MLE's

- Example: Normal Theory Gauss-Markov Linear Model

$$\underbrace{\boldsymbol{y}}_{n\times 1} = \underbrace{\boldsymbol{X}}_{n\times p}\underbrace{\beta}_{p\times 1} + \underbrace{\epsilon}_{n\times 1} \qquad \epsilon \sim N(\mathbf{0}, \sigma^2 I) \qquad \underbrace{\boldsymbol{\theta}}_{(p+1)\times 1} = \begin{bmatrix} \beta \\ \sigma^2 \end{bmatrix}$$

$$\begin{aligned} f(\boldsymbol{y}|\boldsymbol{\theta}) &= \frac{\exp\left\{\frac{-1}{2}(\boldsymbol{y}-x\beta)'(\sigma^2 I)^{-1}(\boldsymbol{y}-x\beta)\right\}}{(2\pi)^{n/2}|\sigma^2 I|^{1/2}} \\ &= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{\frac{-1}{2\sigma^2}(\boldsymbol{y}-x\beta)'(\boldsymbol{y}-x\beta)\right\} \end{aligned}$$

$$l(\boldsymbol{\theta}|\boldsymbol{y}) = -\frac{n}{2}\log 2\pi\sigma^2 - \frac{1}{2\sigma^2}(\boldsymbol{y}-x\beta)'(\boldsymbol{y}-x\beta)$$

- The score function is

$$\frac{\partial l(\boldsymbol{\theta}|\boldsymbol{y})}{\partial \boldsymbol{\theta}} = \begin{bmatrix} \frac{\partial l(\boldsymbol{\theta}|\boldsymbol{y})}{\partial \beta} \\ \frac{\partial l(\boldsymbol{\theta}|\boldsymbol{y})}{\partial \sigma^2} \end{bmatrix} = \begin{bmatrix} \frac{1}{\sigma^2}(x'\boldsymbol{y} - x'x\beta) \\ \frac{(\boldsymbol{y}-x\beta)'(\boldsymbol{y}-x\beta)}{2\sigma^4} - \frac{n}{2\sigma^2} \end{bmatrix}$$

- The score equations are

$\frac{\partial l(\boldsymbol{\theta}|\boldsymbol{y})}{\partial \boldsymbol{\theta}} = \mathbf{0} \Leftrightarrow \qquad x'x\beta = x'\boldsymbol{y} \qquad \sigma^2 = \frac{(\boldsymbol{y}-x\beta)'(\boldsymbol{y}-x\beta)}{n}$

- Here, any solution lies in the appropriate parameter space: $\beta \in I\!R^p$, $\sigma^2 \in I\!R^+$.
- A solution to the score equations is

$$\begin{bmatrix} \hat\beta \\ \frac{(\boldsymbol{y}-x\hat\beta)'(\boldsymbol{y}-x\hat\beta)}{n} \end{bmatrix},$$

where $\hat\beta$ is a solution to the normal equations
- Need to show this is a maximum of the likelihood function
- We already know that any solution to the normal equations, $\beta = \hat\beta$, minimizes $(\boldsymbol{y} - \boldsymbol{X}\beta)'(\boldsymbol{y} - \boldsymbol{X}\beta)$ for $\beta \in I\!R^p$.
- Thus,

$$\forall\, \sigma^2 > 0,\ l\left(\begin{bmatrix} \hat\beta \\ \sigma^2 \end{bmatrix} \mid \boldsymbol{y}\right) \geq l\left(\begin{bmatrix} \beta \\ \sigma^2 \end{bmatrix} \mid \boldsymbol{y}\right)\ \forall \beta \in I\!R^p$$

- The 2^{nd} derivative of $l(\boldsymbol{\theta}|\boldsymbol{y})$ with respect to σ^2 is < 0 so

$$\begin{bmatrix} \hat\beta \\ \frac{(\boldsymbol{y}-x\hat\beta)'(\boldsymbol{y}-x\hat\beta)}{n} \end{bmatrix} \text{ is an MLE of } \boldsymbol{\theta} = \begin{bmatrix} \beta \\ \sigma^2 \end{bmatrix}$$

- Thus, if $C\beta$ is estimable, the MLE of $C\beta$ is $C\hat\beta$ (by the Invariance Property of MLEs), which is the BLUE of $C\beta$
- Note that the MLE of σ^2 is not the unbiased estimator we have been using.

$E[\frac{(\boldsymbol{y}-x\hat{\boldsymbol{b}})'(\boldsymbol{y}-x\hat{\boldsymbol{b}})}{n}] = E(\frac{SSE}{n}) = \frac{n-p}{n}\sigma^2 < \sigma^2$

Thus, the MLE underestimates σ^2 on average.

- Now consider the general linear model

$$\boldsymbol{y} = X\beta + \boldsymbol{\epsilon},\ \boldsymbol{\epsilon}\ \sim\ N(\boldsymbol{0}, \boldsymbol{\Sigma}),$$

Where $\boldsymbol{\Sigma}$ is a positive definite covariance matrix whose entries depend on unknown parameters in some vector γ.
- For example,

$$\boldsymbol{\Sigma} = \sigma^2 \begin{bmatrix} 1 & \rho & \rho^2 \\ \rho & 1 & \rho \\ \rho^2 & \rho & 1 \end{bmatrix}, \gamma = \begin{bmatrix} \sigma^2 \\ \rho \end{bmatrix}$$

- Not Aitken model when ρ unknown.

- In general,

$$\boldsymbol{\theta} = \begin{bmatrix} \beta \\ \gamma \end{bmatrix} \text{ and } f(\boldsymbol{y}|\boldsymbol{\theta}) = \frac{\exp\left\{-\frac{1}{2}(\boldsymbol{y} - x\beta)'(\boldsymbol{\Sigma})^{-1}(\boldsymbol{y} - x\beta)\right\}}{(2\pi)^{n/2}|\boldsymbol{\Sigma}|^{1/2}}$$

$$l(\boldsymbol{\theta}|\boldsymbol{y}) = -\frac{1}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}(\boldsymbol{y} - x\beta)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{y} - x\beta) - \frac{n}{2}\log(2\pi)$$

- We know that for any positive definite covariance matrix $\boldsymbol{\Sigma}$, $(\boldsymbol{y} - x\beta)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{y} - x\beta)$ is minimized over $\beta \in I\!R^p$ by the GLS estimator $\hat{\beta}_g = (X'\boldsymbol{\Sigma}^{-1}X)^{-}X'\boldsymbol{\Sigma}^{-1}\boldsymbol{y}$.
- Thus, for any γ such that $\boldsymbol{\Sigma}$ is a positive definite covariance matrix,

$$l\left(\begin{bmatrix} \hat{\beta}_g \\ \gamma \end{bmatrix} \mid \boldsymbol{y}\right) \geq l\left(\begin{bmatrix} \beta \\ \gamma \end{bmatrix} \mid \boldsymbol{y}\right)\ \forall\, \beta \in I\!R^p$$

- We define the profile log likelihoood for γ to be

$$l^*(\gamma \mid \boldsymbol{y}) = l\left(\begin{bmatrix} \hat{\beta}_g \\ \gamma \end{bmatrix} \mid \boldsymbol{y}\right)$$

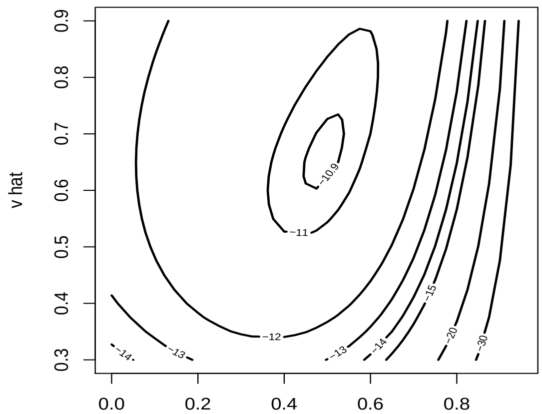
- The MLE of $\boldsymbol{\theta}$ is

$$\hat{\boldsymbol{\theta}} = \begin{bmatrix} \hat{\beta}_{\hat{g}} \\ \hat{\gamma} \end{bmatrix}$$

- Where $\hat{\gamma}$ is a maximizer of $l^*(\gamma|\boldsymbol{y})$ and $\hat{\boldsymbol{\Sigma}}$ is $\boldsymbol{\Sigma}$ with $\hat{\gamma}$ in place γ.
- In general, numerical methods are required to find $\hat{\gamma}$, a maximizer of $l^*(\gamma|\boldsymbol{y})$

- Numerical maximimization algorithms are iterative.
  - Require a starting value of γ
  - Attempt to find a better value, γ^*, in the sense that $l^*(\gamma^*|\boldsymbol{y}) > l^*(\gamma|\boldsymbol{y})$.
- Newton-Raphson algorithm:
  - estimate gradient vector and Hessian matrix at current γ.
  - This gives a quadratic approximation to log-likelihood.
  - Analytic maximum of quadratic gives γ^*.
  - replace γ by γ^*
  - Repeat until no improvement.
- Many details, will cover near end of semester.
- Will mention one now.

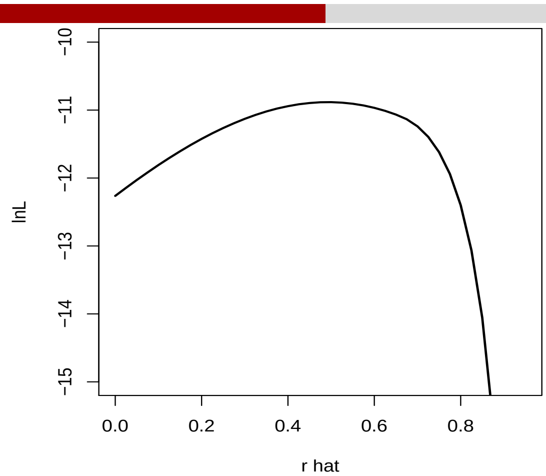
- Example: AR(1) model with unknown ρ



- Problem appears to be maximization over two parameters.
- Doesn't have to be. Remember the Aitken model:

$$\boldsymbol{y} = \boldsymbol{X}\beta + \epsilon, \;\; \epsilon \sim (\boldsymbol{0}, \sigma^2 \boldsymbol{V}),$$

  where $\boldsymbol{V}$ is a function of known constants
- e.g. for AR(1)
- $$\boldsymbol{\Sigma} = \sigma^2 \begin{bmatrix} 1 & \rho & \rho^2 & \rho^3 \\ \rho & 1 & \rho & \rho^2 \\ \rho^2 & \rho & 1 & \rho \\ \rho^3 & \rho^2 & \rho & 1 \end{bmatrix}$$
- $\hat{\beta}_g$ depends on ρ but not σ^2
- Use ρ as the parameter, maximize lnL over ρ
- Sometimes called "profiling out" the error variance



- The MLE of the variance component vector γ is often biased.
- For example, for the case of $\epsilon = \sigma^2\, I$, where $\gamma = \sigma^2$, the MLE of σ^2 is $\frac{(\boldsymbol{y}-X\hat{\beta})'(\boldsymbol{y}-X\hat{\beta})}{n}$ with expectation $\frac{n-p}{n}\sigma^2$.
- The MLE of σ^2 is often criticized for "failing to account for the loss of degrees of freedom needed to estimate β."

$$E[\frac{(\boldsymbol{y}-\boldsymbol{X}\hat{\beta})'(\boldsymbol{y}-\boldsymbol{X}\hat{\beta})}{n}] = \frac{n-p}{n}\sigma^2$$

$$< \sigma^2 = E[\frac{(\boldsymbol{y}-\boldsymbol{X}\beta)'(\boldsymbol{y}-\boldsymbol{X}\beta)}{n}]$$

- A variation, REML, is unbiased in simple models and less biased in many others.
- We'll see REML in detail when we talk about mixed models

Large N theory for MLE's

- Suppose θ is a $k \times 1$ parameter vector.
- Let $l(\theta)$ denote the log likelihood function.
- Under regularity conditions discussed, e.g., Casella and Berger, we have the following:
  - There is an estimator $\hat{\theta}$ that solves the likelihood equations $\frac{\delta l(\theta)}{\delta \theta} = \mathbf{0}$ and is a consistent estimator of θ, i.e., $\lim_{n\to\infty} Pr[\mid \hat{\theta} - \theta \mid > \epsilon] = 0$ for any $\epsilon > 0$.
  - For sufficiently large n, $\hat{\theta} \sim N(\theta, I^{-1}(\theta))$, where

-

$$\begin{aligned} \boldsymbol{I}(\boldsymbol{\theta}) &= E\left[\left(\frac{(\delta l(\boldsymbol{\theta}))}{\delta \boldsymbol{\theta}}\right)\left(\frac{\delta l(\boldsymbol{\theta})}{\delta \boldsymbol{\theta}}\right)'\right] \\ &= -E\left[\frac{\delta^2 l(\boldsymbol{\theta})}{\delta \boldsymbol{\theta} \delta \boldsymbol{\theta}'}\right] \end{aligned}$$

- Or, in scalar terms,

$$\boldsymbol{I}(\boldsymbol{\theta})_{ij} = -E\left[\frac{\delta^2 l(\boldsymbol{\theta})}{\delta \theta_i \delta \theta_j}\right],\ i = 1, .., k,\ j = 1, ..., k.$$

- $\boldsymbol{I}(\boldsymbol{\theta})$ is known as the Fisher Information matrix.
- $\boldsymbol{I}(\boldsymbol{\theta})$ can be approximated by $\frac{-\delta^2 l(\boldsymbol{\theta})}{\delta \boldsymbol{\theta} \delta \boldsymbol{\theta}'} \mid_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}$
- Often called "observed information"

Wald tests and confidence intervals

- Suppose for large n that $\hat{\theta} \sim N(\theta, \boldsymbol{V})$ and $\hat{\boldsymbol{V}}$ is a consistent estimator of $\boldsymbol{V}$.
- Then, in suitable large samples, $\hat{\boldsymbol{V}}^{-1/2}(\hat{\theta} - \theta) \sim N(\mathbf{0}, I)$ and $(\hat{\theta} - \theta)'\,\hat{\boldsymbol{V}}^{-1}(\hat{\theta} - \theta) \sim X_k^2$.
- An approximate $100(1-\alpha)$% confidence interval for θ_i is $\hat{\theta}_i \pm Z_{1-\alpha/2}\sqrt{\hat{\boldsymbol{V}}_{ii}(\theta)}$, where $Z_{1-\alpha/2}$ is the $1-\alpha/2$ quantile of N(0,1) and $\hat{\boldsymbol{V}}_{ii}(\theta)$ is element (i,i) of $\hat{\boldsymbol{V}}(\theta)$.
- An approximate p-value for testing $H_0 : \theta = \theta_0$ is $P[X_k^2 \geq (\hat{\theta} - \theta_0)'\,\hat{\boldsymbol{V}}^{-1}(\hat{\theta} - \theta_0)]$, where X_k^2 is a X^2 random variable with k degrees of freedom and k is the number of restrictions in the null hypothesis.
- The above confidence interval and test are based on the asymptotic normality of $\hat{\theta}$

Likelihood ratio based inference

- Suppose we wish to test the null hypothesis that a reduced model provides an adequate fit to a dataset relative to a more general full model that includes the reduced model as a special case.
- Under the regularity conditions mentioned previously, (REDUCED MODEL DEVIANCE) - (FULL MODEL DEVIANCE) $\overset{H_0}{\sim} \chi^2_{k_f - k_r}$, where k_f and k_r are the number of free parameters under the full and reduced models, respectively.
- This approximation can be reasonable if n is "sufficiently large".
- Note that the test statistic is equal to $-2log\Lambda$, where

$$\Lambda = \frac{\text{LIKELIHOOD MAXIMIZED UNDER REDUCED MODEL}}{\text{LIKELIHOOD MAXIMIZED UNDER THE FULL MODEL}}$$

- Λ is known as the <u>likelihood</u> <u>ratio</u>, and tests based on $-2log\Lambda$ are called <u>likelihood</u> <u>ratio</u> <u>tests</u>.

LRT's and Confidence Regions for a Subvector of θ:

- Suppose θ is a $k \times 1$ vector and is partitioned into vectors θ_1 $k_1 \times 1$ and θ_2 $k_2 \times 1$, where $k = k_1 + k_2$ and $\theta = \begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix}$
- Consider a test of $H_0 : \theta_1 = \theta_{10}$
- Suppose $\hat{\theta}$ is the MLE of θ and $\hat{\theta}_2(\theta_1)$ maximizes $l = \left(\begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix} \right)$ over θ_2 for any fixed value of θ_1.
- Then $2\left[l(\hat{\theta}) - l\left(\begin{bmatrix} \theta_{10} \\ \hat{\theta}_2(\theta_{10}) \end{bmatrix} \right) \right] \overset{H_0}{\sim} X^2_{k_1}$ by our previous result when n is "sufficiently large."

- Also, $Pr\left\{ 2[l\left(\hat{\theta}\right) - l\left(\begin{bmatrix} \theta_1 \\ \hat{\theta}_2(\theta_1) \end{bmatrix} \right)] \leq X^2_{k_1,\alpha} \right\} \approx 1 - \alpha \Rightarrow$ $Pr\left\{ l(\begin{bmatrix} \theta_1 \\ \hat{\theta}_2(\theta_1) \end{bmatrix}) \geq l(\hat{\theta}) - \frac{1}{2}X^2_{k_1,\alpha} \right\} \approx 1 - \alpha$
- Thus, the set of values of θ_1 that when maximizing over θ_2, yield a maximized likelihood within $\frac{1}{2}X^2_{k_1,\alpha}$ of the likelihood maximized over all θ, form a $100(1-\alpha)$% confidence region for θ_1
- The ultimate distributional dependence is the asymptotic N or χ^2 distribution.
  - Only holds for infitely large sample sizes
  - But, may be an appropriate approximate for practical but large sample sizes
  - may be inappropriate if sample sizes are too small
- Often, the above never worried about.

Why worry about a second ci method?

- When $\boldsymbol{Y} \sim N(\theta, \sigma^2 \boldsymbol{I})$, $\boldsymbol{C}\beta \sim N$ in any size sample. When inference uses $\hat{\sigma}^2$, the T and F distributions are exact for any size sample. No need for LR ci's or tests.
- Now, let $\boldsymbol{Y} indep \sim F(\beta)$, some arbitrary, non-normal distribution. Imagine that the sample size is sufficiently large that $\hat{\beta} approx. \sim N$. Now, you want inference on $\tau = \exp \beta_1$ or $\eta = \beta_2/\beta_1$.
- MLE of $\exp \boldsymbol{\beta}$ is $\exp \hat{\boldsymbol{\beta}}$, MLE of β_2/β_1 is $\hat{\boldsymbol{\beta}}_2/\hat{\boldsymbol{\beta}}_1$.
- se of $\exp \boldsymbol{\beta}$ or β_2/β_1 from Delta method.
- ci? If β *approx* $\sim N$, then $\exp \hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\beta}}_2/\hat{\boldsymbol{\beta}}_1$ are most definitely not normal.

- If sample size increased, eventually will converge to asymptotic normal distribution, but can't use Wald methods on current sample.
- log-likelihood invariant to reparameterization:
  - $l(\log \tau) = l(\beta_1)$.
  - so LR ci is $\tau s.t.$ $2\left[l(\hat{\beta}_1) - l(\log \tau)\right] < \chi^2_{1-\alpha,k}$
- for functions of multiple parameters, need an additional maximization:
  - so LR ci for η is $\eta s.t.$ $2\left[l(\hat{\beta}_1, \hat{\beta}_2) - \max_{\beta_1} l(\beta_1, \eta\beta_2)\right] < \chi^2_{1-\alpha,k}$
  - Just a computing exercise.
- In my experience, the χ^2 approximation for a LR statistic is usable at much smaller sample sizes than the N approximation for a $\hat{\beta}$, unless you're lucky in your choice of β.
- I know of no cases where the N approximation is usable at smaller sample sizes than the χ^2 approximation, unless $Y \sim N$.

Two/three notes for caution

- The regularity conditions do not hold if the true parameter or the value specified by the null hypothesis falls on the boundary of the parameter space.
- We will see examples of this soon, when we discuss mixed models
- One example, if a model has two random components (error and something else, call it $u \sim N(0, \sigma^2_u)$, testing $Ho : \sigma^2_u = 0$ is on the boundary of the parameter space
- None of the usual theory applies for this situation.
- because the regularity conditions are not met.

Generalized Linear Models

- Consider the normal theory Gauss-Markov linear model $\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon},\ \boldsymbol{\epsilon} \sim N(0, \sigma^2 I)$.
- Does not have to be written as function + error
- Could specify distribution and model(s) for its parameters
- i.e., $y_i \sim N(\mu_i, \sigma^2)$, where $\mu_i = \boldsymbol{X}_i'\boldsymbol{\beta}$ for all $i = 1, ..., n$ and $y_1, ..., y_n$ independent.
- This is one example of a generalized linear model.
- Here is another example of a GLM: $y_i \sim Bernoulli(\pi_i)$, where $\pi_i = \frac{exp(\boldsymbol{X}_i'\beta)}{1+exp(\boldsymbol{X}_i'\beta)}$ for all $i = 1, ..., n$ and $y_1, ..., y_n$ independent.
- In both examples, all responses are independent and each response is a draw from one type of distribution whose parameters may depend on explanatory variables through a known function of a linear predictor $\boldsymbol{X}_i'\boldsymbol{\beta}$.

- The normal and Bernoulli models (and many others) are special cases of a generalized linear model.
- These are models where:
  - The parameters are specified functions of $\boldsymbol{X}\beta$
  - The distribution is in the exponential scale family
  - i.e., y_i has a probability density function (or probability mass function, for discrete distribution)

$$\exp\left(\frac{\boldsymbol{\eta}(\boldsymbol{\theta})\boldsymbol{T}(y_i) - \boldsymbol{b}(\boldsymbol{\theta})}{a(\phi)} + c(y_i, \phi)\right) \quad (2)$$

  where $\boldsymbol{\eta}()$, $\boldsymbol{T}()$, $a()$, $b()$, and $c()$ are known functions and $\boldsymbol{\theta}$ is a vector of unknown parameters depending on $\boldsymbol{X}\beta$ and ϕ is either a known or unknown parameter.
- Exponential family / exponential class is (2) without the $a(\phi)$
- $a(\phi)$ includes "overdispersed" distributions in the family

- For example, the pdf for a normal distribution can be written as:

$$\exp\left(\frac{-1}{2\sigma^2}y_i^2 + \frac{\mu}{2\sigma^2}y_i - \frac{\mu^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2)\right)$$

- from which:
  - $\boldsymbol{\eta}(\boldsymbol{\theta}) = \left(\frac{\mu}{\sigma^2},\ \frac{-1}{2\sigma^2}\right)$
  - and $\boldsymbol{T}(y_i) = \left(y_i,\ y_i^2\right)$

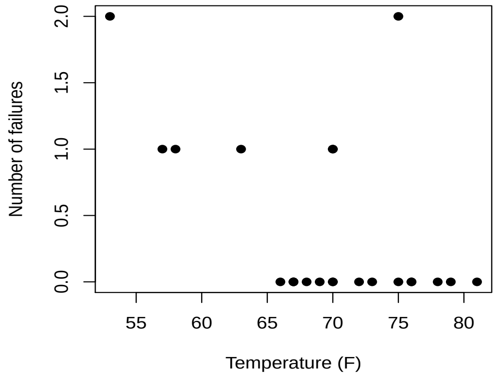
- This family includes many common distributions:

Distribution	$\boldsymbol{\eta}(\boldsymbol{\theta})'$	$\boldsymbol{T}(y_i)'$	$a(\phi)$
Normal	$\left(\frac{\mu}{\sigma^2},\ \frac{-1}{2\sigma^2}\right)$	$\left(y_i,\ y_i^2\right)$	1
Bernoulii	$\log\left(\frac{\pi}{1-\pi}\right)$	y_i	1
Poisson	$\log\lambda$	y_i	1
Overdisp. Poisson	$\log\lambda$	y_i	ϕ
Gamma	$\left(\frac{-1}{\theta},\ (k-1)\right)$	$\left(y_i,\ \log y_i\right)$	1

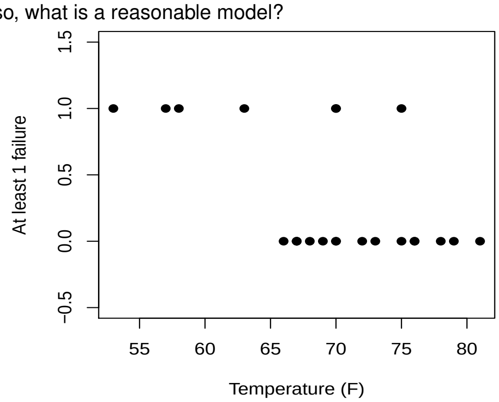
- A lot of stat theory results follow immediately from the exponential form
- e.g. $T(y_i)$ is the vector of sufficient statistics
- A lot of theory is much nicer when the distribution is parameterized in terms of $\boldsymbol{\eta}(\boldsymbol{\theta})$ instead of $\boldsymbol{\theta}$
- e.g. use $\log\left(\frac{\pi}{1-\pi}\right)$ as the parameter of a Bernoulli distribution instead of π.
- This is called the canonical or natural form

- A generalized linear model has:
  - A probability distribution in the exponential scale family
  - A linear predictor, $\boldsymbol{\eta} = \boldsymbol{X}\beta$
  - A link function that connects $E(\boldsymbol{y})$ and the linear predictor: $g(E(\boldsymbol{Y})) = g(\boldsymbol{\mu}) = \boldsymbol{\eta}$

- Example: Challenger O-ring data, with flight temperature



- Q: does P[one or more failures] depend on temperature? If so, what is a reasonable model?



- Need a reasonable model for independent observations taking values of 0 or 1.
- $y \sim Bernoulli(\pi)$ has probability mass function

$$f(n) = \begin{cases} \pi^y(1-\pi)^{1-y} & \text{for } y \in \{0, 1\} \\ 0 & \text{otherwise} \end{cases}$$

- Thus, $Pr(y = 0) = \pi^0(1-\pi)^{1-0} = 1 - \pi$ and
  $Pr(y = 1) = \pi^1(1-\pi)^{1-1} = \pi$.
  $E(y) = \sum_y yf(y) = 0(1-\pi) + 1 * \pi = \pi$
  $Var(y) = E(y) - \{E(y)\}^2 = \pi - \pi^2 = \pi(1-\pi)$
  Note that $Var(y)$ is a function of $E(y)$.

The Logistic Regression Model

- For $i = 1, ..., n;\ y_i \sim Bernoulli(\pi_i)$, where $\pi_i = \frac{exp(\boldsymbol{X}_i'\beta)}{1+exp(\boldsymbol{X}_i'\beta)}$ and $y_1, ..., y_n$ are independent.
- logit function: $g(\pi) = \log(\frac{\pi}{1-\pi})$.
  maps the interval (0,1) to the real line $(-\infty, \infty)$.
- Odds of event $A \equiv \frac{Pr(A)}{1-Pr(A)}$, so $\log(\frac{\pi}{1-\pi})$ is the log("odds").
- Note that:

$$\begin{aligned} g(\pi_i) &= \log(\frac{\pi}{1-\pi}) \\ &= \log[\frac{exp(\boldsymbol{X}_i'\beta)}{1+exp(\boldsymbol{X}_i'\beta)} / \frac{1}{1+exp(\boldsymbol{X}_i'\beta)}] \\ &= \log[exp(\boldsymbol{X}_i'\beta)] \\ &= \boldsymbol{X}_i'\beta \end{aligned}$$

- Thus, the logistic regression model says that $y_i \sim Bernoulli(\pi_i)$ where $\log(\frac{\pi_i}{1-\pi_i}) = \boldsymbol{X}_i'\beta$

- An initial model for the Challenger data is $\log(\frac{\pi_i}{1-\pi_i}) = \beta_0 + \beta_1 temp_i$
- The logistic regression model is one example of a Generalized Linear Model
- Reminder: a generalized linear model has:
    - A probability distribution in the exponential scale family: Bernoulli is.
    - A linear predictor, $\boldsymbol{\eta} = \boldsymbol{X}\beta$: $\boldsymbol{X}\beta = \beta_0 + \beta_1 temp_i$
    - A link function that connects $E(\boldsymbol{y})$ and the linear predictor: $g(E(\boldsymbol{Y})) = g(\boldsymbol{\mu}) = \boldsymbol{\eta}$
- In this model, the logit is the canonical link function.
- Interpretation of parameters especially easy because of the connection between change in log odds and β
- But, some other link function may fit the data better

Parameter estimation and inference

- Done using likelihood
- The likelihood function for logistic regression is

$$\begin{aligned} l(\beta|\boldsymbol{y}) &= \sum_{i=1}^{n} \log[\pi_i^{y_i}(1-\pi_i)^{1-y_i}] \\ &= \sum_{i=1}^{n} [y_i \log(\pi_i) + (1-y_i)\log(1-\pi_i)] \\ &= \sum_{i=1}^{n} [y_i \{\log(\pi_i) - \log(1-\pi_i)\} + \log(1-\pi_i)] \\ &= \sum_{i=1}^{n} [y_i \log(\frac{\pi}{1-\pi}) + \log(1-\pi_i)] \\ &= \sum_{i=1}^{n} [y_i \boldsymbol{X}_i'\beta - (1+\exp\{\boldsymbol{X_i\beta}\})] \end{aligned}$$

- The likelihood function has to be maximized numerically.
- Fisher's Scoring algorithm commonly used for likelihoods for Generalized Linear Models.
    - uses the expected value of the matrix of second derivatives (-Fisher Information matrix)
    - For Generalized Linear Models, Fisher' Scoring Method results in an iteratively reweighted least squares procedure.
    - The algorithm is presented for the general case in Section 2.5 of Generalized Linear Models $2^n d$ Edition (1989) by McCullagh and Nelder.
- Inference on β
    - For sufficiently large samples, $\hat{\beta}$ is approximately normal with mean β and a variance-covariance matrix that can be approximated by the estimated inverse of Fisher information matrix.
    - Use either Wald tests/intervals or likelihood ratio tests/profile likelihood intervals for inference

Interpretation of Logistic Regression Parameters:

- Let $\tilde{\boldsymbol{x}}' = (x_1, x_2, ..., x_{j-1}, x_j + 1, x_{j+1}, ..., x_p)'$
  In other words, $\tilde{\boldsymbol{X}}$ is the same as $\boldsymbol{x}$ except that the j^{th} explanatory variable has been increased by one unit.
  Let $\pi = \frac{exp(\boldsymbol{x}'\beta)}{1+exp(\boldsymbol{x}'\beta)}$ and $\tilde{\pi} = \frac{exp(\tilde{\boldsymbol{x}}'\beta)}{1+exp(\tilde{\boldsymbol{x}}'\beta)}$
- The odds ratio

$$\begin{aligned} \frac{\tilde{\pi}}{1-\tilde{\pi}} / \frac{\pi}{1-\pi} &= \exp\left\{\log(\frac{\tilde{\pi}}{1-\tilde{\pi}}) - \log(\frac{\pi}{1-\pi})\right\} \\ &= \exp\left\{\tilde{\boldsymbol{x}}'\boldsymbol{\beta} - \boldsymbol{x}'\boldsymbol{\beta}\right\} \\ &= \exp\{(x_j+1)\boldsymbol{\beta}_j - x_j\boldsymbol{\beta}_j\} \\ &= \exp\{\boldsymbol{\beta}_j\} \end{aligned}$$

- All other explanatory variables held constant, the odds of success at $x_j + 1$ are $\exp(\beta_j)$ times the odds of success at x_j.

- A 1 unit increase in the j^{th} explanatory variable (with all other explanatory variables held constant) is associated with a multiplicative change in the odds of success by the factor $\exp(\beta_j)$.
- This is true regardless of the initial value x_j.
- Effect on probability of event does depend on the initial odds

x_j	$P[Y_j = 1 \mid x_j]$	Odds$\mid x_j$	Odds$\mid x_j + 1$	$P[Y_j = 1 \mid x_j + 1]$
-5	0.0067	-5	-4	0.018
-1	0.268	-1	0	0.5
0	0.5	0	1	0.731
3	0.952	3	4	0.982

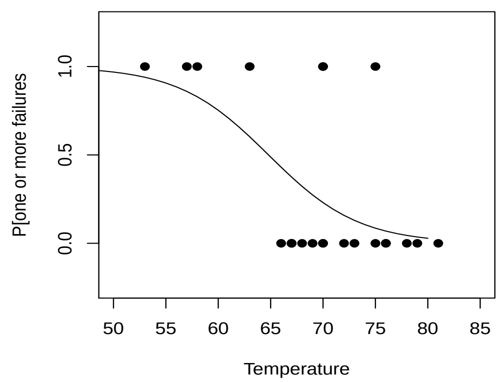
- If (L_j, U_j) is a $100(1-\alpha)\%$ confidence interval for β_j, then $(\exp\{L_j\}, \exp\{U_j\})$ is a $100(1-\alpha)\%$ confidence interval for $\exp\{\beta_j\}$.
- Also note that $\pi = \frac{\exp(\boldsymbol{x}'\beta)}{1+\exp(\boldsymbol{x}'\beta)} = \frac{1}{\frac{1}{\exp(\boldsymbol{x}'\beta)}+1} = \frac{1}{\exp(-\boldsymbol{x}'\beta)}$

- Thus, if (L_j, U_j) is a $100(1-\alpha)\%$ confidence interval for $\boldsymbol{x}'\beta$, then a $100(1-\alpha)\%$ confidence interval for π is $(\frac{1}{1+\exp(-L_j)}, \frac{1}{1+\exp(-U_j)})$
- Results for Challenger data

			95% ci	p-value	
Param.	estimate	se	(profile)	Wald	LRT
Int.	15.04	7.38	(3.33, 34.34)	0.042	
Temp.	-0.23	0.11	(-0.515, -0.061)	0.0320	0.0048

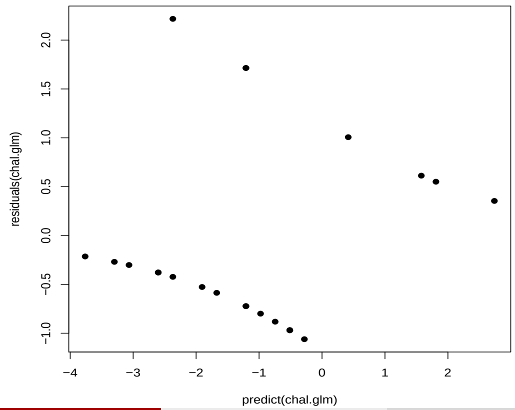
- Predictions:

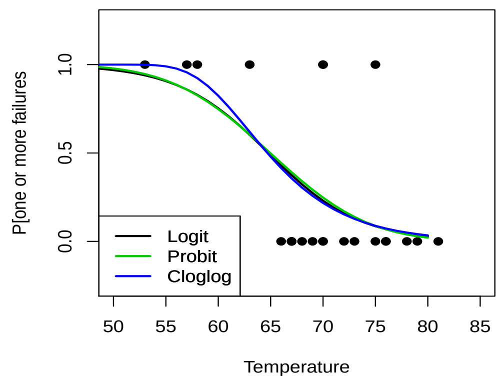
Temp	Xb	P[one or more failure]	se
50	3.43	0.968	0.061
60	1.11	0.753	0.191
70	-1.21	0.230	0.105
80	-3.53	0.028	0.039



Model assessment / diagnostics

- Harder with Bernoulli data because residuals not very informative
- If sufficient observations, define groups of obs with similar x_i compute $P[Y=1]$ for each group, compare to model predictions
- Consider more complicated models:
  logit $\pi_i = \beta_0 + \beta_1 x_i$: AIC = 24.38
  logit $\pi_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2$: AIC = 25.3887
- Consider different link functions:
  - Probit: $\Phi^{-1}(\pi_i) = \beta_0 + \beta_1 x_i$: AIC = 24.38
    where Φ^{-1} is inverse cdf of a normal distribution
  - Comp. log log: $\log(-\log \pi)) = \beta_0 + \beta_1 x_i$: AIC = 23.53
- If $Y \sim$ Poisson(λ), $P[Y=0] = \exp(-\lambda)$.
  $\log(-\log P[Y=0])) = \log \lambda$.





R code for 0/1 logistic regression

```
# fit a logistic regression to the Challenger data
#   coded as 1 = 1 or more failures

chal <- read.table('challenger2.txt', as.is=T, header=T)

plot(chal$temp,chal$fail)

# fit a bernoulli logistic regression using
#   default link (logit if binomial)

chal.glm <- glm(fail~temp, data=chal, family=binomial)
# family= specifies which exponential family dn to use
#   this sets the link function and the variance function
#   can override those defaults if needed, see ?glm
```



```
# all the usual helper functions:
#  print(), summary()
#  coef(),  vcov()
#  anova()
#  predict(), residual()

# most behave in the way you might expect
#  anova(), predict() and residual() are tricky

anova(chal.glm)   #  change in deviance only
#  to get a test, need to specify the appropriate dn
anova(chal.glm, test='Chi')   # Chi-square test
anova(chal.glm, test='F')     # if est. overdispersion
```

```
# predict() can return different types of predictions
#  ?predict.glm() gives the full story
predict(chal.glm)
# returns predictions on the Xb (linear predictor scale)
predict(chal.glm,type='response')
# returns predictions on the Y scale

# residuals() can return 5 different types of residuals.  see ?residuals.glm
residuals(chal.glm,type='pearson')
# pearson Chi-square residuals
residuals(chal.glm,type='deviance')
# deviance residuals

plot(predict(chal.glm),
  residuals(chal.glm,type='deviance'))
```



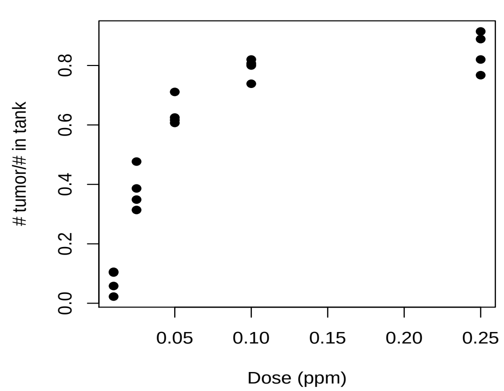
```
# How similar are the two types of resid?
plot(residuals(chal.glm, type='pearson'),
   residuals(chal.glm,type='deviance'))

# to overlay data and curve, need type='response'
plot(chal$temp,chal$fail,pch=19)
lines(30:85, predict(chal.glm,newdata=data.frame(temp=30:85),
  type='response')),

# change link functions
chal.glm2 <- glm(fail~temp,data=chal,
  family=binomial(link=probit))
chal.glm3 <- glm(fail~temp,data=chal,
  family=binomial(link=cloglog))
lines(30:85,
  predict(chal.glm2,newdata=data.frame(temp=30:85),
    type='response'), col=3)
lines(30:85,
  predict(chal.glm3,newdata=data.frame(temp=30:85),
    type='response'), col=4)
```


Logistic Regr. Model for Binomial Count Data

- Bernoulli model appropriate for 0/1 response on an individual
- What if data are # events out of # trials per subject?
- Example: Toxicology study of the carcenogenicity of aflatoxicol.
  - (from Ramsey and Schaefer, *The Statistical Sleuth*, p 641)
  - Tank of trout randomly assigned to dose of aflatoxicol
  - 5 doses. CRD. 4 replicate tanks per dose
  - 86-90 trout per tank
  - Response is # trout with liver tumor
- Could use Bernoulli model for each individual fish
- But, all fish in a tank have the same covariate values (dose)
- easier to analyze data in summarized form (# with tumor, # in tank)



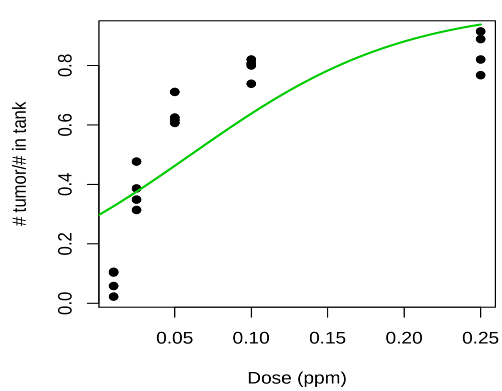
- Each response is # "events" out of # trials.
- $y_i \sim Binomial(m_i, \pi_i), \ \ i = 1, ..., n$, where m_i is a known number of trials for observation i.
  - $$\pi_i = \frac{\exp(\boldsymbol{x}_i'\boldsymbol{\beta})}{1 + \exp(\boldsymbol{x}_i'\boldsymbol{\beta})}$$
  - $y_1, ..., y_n$ are independent.
- Note: two levels of independence assumed in this model
  - each response, y_i, is independent
  - trials within each each response are independent $|\pi_i$
- $y_i \sim Binomial(m_i, \pi_i)$ when
  - $y_i = \sum_{j=1}^{m_i} Z_{ij}$, where $Z_{ij} \sim Bernoulli(\pi_i)$
  - **And** Z_{ij} independent
- May be an issue for both the Challenger and trout data sets
- Binomial model assumes no flight (tank) effects
- Data are same as $\approx$ 360 fish raised individually, or the same as one tank with $\approx$ 360 fish

- For now, assume Binomial model reasonable
- Facts about Binomial distributions: If $y_i \sim Binomial(m_i, \pi_i)$
  - $\mathrm{E(y_i) = m_i\pi_i}$
  - $\mathrm{Var(y_i) = m_i\pi_i(1 - \pi_i)}$
  - $f(y_i) = \left(\begin{array}{c} m_i \\ y_i \end{array}\right) \pi_i^{y_i}(1 - \pi_i)^{m_i - y_i}$ for $y_i \in \{0, ..., m_i\}$
  - $l(\beta|\boldsymbol{y}) = \sum_{i=1}^{n}[y_i \ \log(\frac{\pi_i}{1-\pi_i}) + m_i \log(1 - \pi_i)] + const =$ $\sum_{i=1}^{n}[y_i \ \boldsymbol{x}_i'\beta - m_i \log(1 + \exp\{-\boldsymbol{x}_i'\beta\})] + const.$
- $l(\beta|\boldsymbol{y}, \boldsymbol{m}, \boldsymbol{x})$ for $(y_1, m_1, x_1), (y_2, m_2, x_2), \ldots (y_n, m_n, x_n)$ same (apart from constant) as Bernoulli lnL: $l(\beta|\boldsymbol{z})$ for (z_{ij}, x_{ij}) =
  $(0, x_1), (0, x_1), \ldots (1, x_1), \ldots,$
  $(0, x_2), \ldots (1, x_2), \ldots,$
  $\cdots,$
  $(0, x_n), \ldots, (1, x_n), \ldots$
- MLE's $\hat{\beta}$ obtained by numerically maximizing $l(\beta|\boldsymbol{y})$ over $\beta \in I\!R^p$

- Results for trout data:

Coefficient	Estimate	se	z	p
Intercept	-0.867	0.076	-11.30	$<$0.0001
dose	14.33	0.937	15.30	$<$0.0001

- Looks impressive, but ...

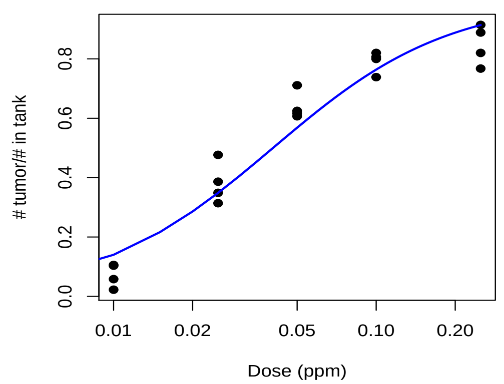


Testing lack of fit

- Remember ANOVA lack of fit test from 500
  - Compare fit of regression to fit of a means model.
  - Required replicated observations so you can estimate Var y
- Can use same ideas in logistic regression:

Model	df	Deviance	p-value
lin. reg.	18	277.047	
means	15	25.961	
diff.	3	251.09	$<$0.0001

- Better fit when X=log(dose), but still large LOF
- Problem is that P[tumor| dose] seems to asymptote at $\approx$0.8
- Standard logistic model asymptotes at 1.0



Residual deviance as a LOF test

- When obs. $\sim$ Binomial, there is another way to assess model fit
- Remember, don't need to estimate σ^2.
- Var $y_i = m_i\, E\, y_i(1 - E\, y_i)$
- Can fit (and use in sensible ways) a model with a separate parameter for each observation.
- Called a "saturated" model. Has one π_i parameter for each y_i observation.
- Logistic Regression Model — Saturated Model

Logistic Regression Model	Saturated Model
$y_i \sim Binomial(m_i, \pi_i)$	$y_i \sim Binomial(m_i, \pi_i)$
$y_i, ..., y_n$ *independent*	$y_i, ..., y_n$ *independent*
$\pi_i = \frac{\exp(\boldsymbol{x}_i'\beta)}{1+\exp(\boldsymbol{x}_i'\beta)}$ for some $\beta \in I\!R^p$	$\pi_i \in [0, 1]$ for $i = 1, ..., n$ with no other restrictions
p parameters	n parameters

- Can compare the fit of the two models using $D = -2\, lnL(\hat{\beta}|\boldsymbol{y}) - (-2\, lnL(\hat{\boldsymbol{\pi}}|\boldsymbol{y}))$
- This statistic is sometimes called the Deviance Statistic, the Residual Deviance, or just the Deviance.
- Under Ho: regression model fits the data, i.e. $\pi_i = \frac{\exp(\boldsymbol{x}_i'\beta)}{1+\exp(\boldsymbol{x}_i'\beta)}$, $D \sim \chi^2_{n-p}$
- To test lack of fit, compare D to χ^2_{n-p}
- This is an asymptotic result.
- The χ^2 approximation to the null distribution works reasonably well if $m_i \geq 5$ for most i.
- Do not have to fit two models to calculate the Deviance statistic
  - Let $\hat{\pi}_i = \frac{\exp(\boldsymbol{x}_i'\beta)}{1+\exp(\boldsymbol{x}_i'\beta)}$ for all $i = 1, ..., n$.
  - Then, the likelihood ratio statistic for testing the logistic regression model as the reduced model vs. the saturated model as the full model is $D = \sum_{i=1}^n 2[y_i log(\frac{y_i}{m_i\hat{\pi}_i}) + (m_i - y_i) log(\frac{m_i - y_i}{m_i - m_i\hat{\pi}_i})]$

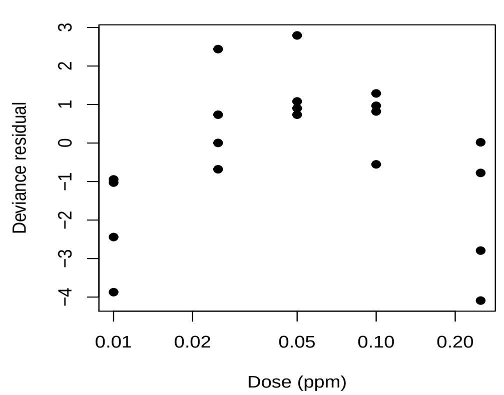
- For the trout data:

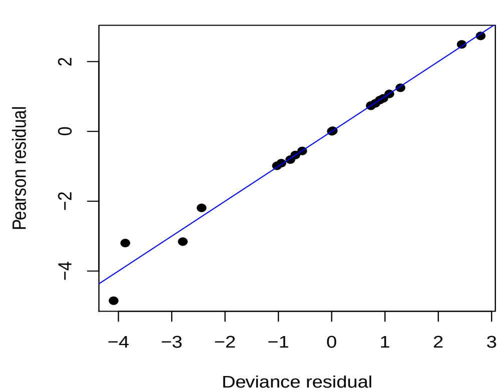
X	Residual Deviance	d.f.	p for LOF
dose	277.05	18	$<$0.0001
log dose	68.90	18	$<$0.0001

- Plotting data and curve works for one X variable. Harder for two X's and impossible for many X's
- Goal: a quantity like the residual, $y_i - \hat{y}_i$, in a linear regression that we can use to diagnose problems with the generalized linear model

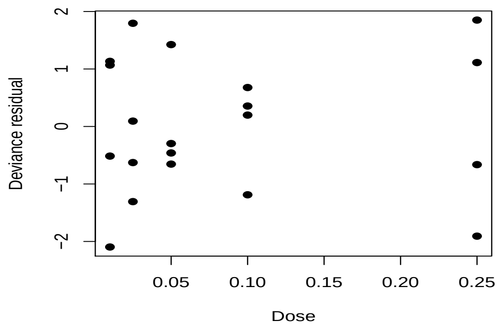
Defining a residual in a logistic regression

- in usual GM model, y_i have constant variance: $y_i \sim (\boldsymbol{X}\boldsymbol{\beta}, \sigma^2)$
- in a logistic regression Var y_i depends on π_i
- Two common definitions of residuals for logistic regression:
  1. Deviance residual:
     - $d_i = sign(y_i - m_i\hat{\pi}_i)\sqrt{2[y_i log(\frac{y_i}{m_i\hat{\pi}_c}) + (m_i - y_i)log(\frac{m_i - y_i}{m_i - m_i\hat{\pi}_c})]}$
     - The residual deviance statistic $D = \sum_{i=1}^n d_i^2$.
  2. Pearson χ^2 residual:
     - $r_i = \frac{y_i - m_i\hat{\pi}_i}{\sqrt{m_i\hat{\pi}_i(1-\hat{\pi}_i)}}$
     - Because $\sum_{i=1}^n r_i^2 = \sum_{i=1}^n (\frac{y_i - m_i\hat{\pi}_i}{\sqrt{m_i\hat{\pi}_i(1-\hat{\pi}_i)}})^2 = \sum_{i=1}^n (\frac{y_i - \hat{E}(y_i)}{\sqrt{\hat{Var}(y_i)}})^2$
- For large m_i's, both d_i and r_i should behave like standard normal random variables if the logistic regression model is correct.





- For fun, lets evaluate LOF of the means model
  5 parameters, one π_i for each dose



- But Residual deviance = 25.96 with 15 df. p = 0.038
- Test says that the means model doesn't fit the data!

Overdispersion:

- The LOF test is evaluating all aspects of the model.
- One is the fit of the model for π_i
- A second is the implied variance
  - For many GLM distributions, $Var(y_i)$ is a function of $E(y_i)$.
  - Logistic regression:
    $Var(y_i) = m_i\pi_i(1-\pi_i) = m_i\pi_i - \frac{(m_i\pi_i)^2}{m_i} = E(y_i) - \frac{[E(y_i)]^2}{m}$
  - So estimating π_i provides estimates of Var y_i
- If the variability of our response is greater than we should expect based on our estimates of the mean, we say that there is overdispersion.
- That is the problem with the means model for the trout data

Overdispersion

- Account for overdispersion by introducing an additional scale parameter, ϕ: $Var(y_i) = \phi\, m_i\pi_i(1-\pi_i)$
- Observations no longer have Binomial distributions, but have the variance pattern characteristic (apart from ϕ) of a Binomial distribution.
- Called a quasiBinomial distribution
- Can also use for underdispersion, but that rarely happens.
- The dispersion paramter ϕ can be estimated from
  - The residual deviance: $\frac{\sum_{i=1}^{n} d_i^2}{n-p}$
  - Or, the Pearson Chi-square statistic: $\frac{\sum_{i=1}^{n} r_i^2}{n-p}$
- Beware: can not distinguish between model lack of fit and overdispersion.
- Make sure model is reasonable before estimating ϕ

- For the trout data, using means model:
  - Residual deviance: 25.96 with 15 df.
  - $\hat{\phi} = 1.73$
- Because we have replicate tanks, can estimate E y_i/m_i and Var y_i/m_i for each dose. Compare Var y_i/m_i to Binomial implied variance. Approx. calculation because m_i not constant. Use $m_i = 88$, so implied Var $y_i/m_i \approx \hat{\pi}_i(1-\hat{\pi}_i)/88$
-

Dose	ave. $\boldsymbol{y}_i/N_i$	implied Var	sample Var	ratio
0.010	0.072	0.000764	0.00159	2.09
0.025	0.381	0.00268	0.00491	1.83
0.05	0.640	0.00262	0.00232	0.88
0.10	0.791	0.00187	0.00131	0.70
0.25	0.847	0.00147	0.00446	3.04
ave.				1.71

Adjusting infererence for overdispersion

- Estimation done by quasilikelihood, similar to a likelihood but only depends on mean and variance of the "distribution"
- Overdispersed Logistic regression: Var $y_i/m_i = \phi\, \pi_i(1-\pi_i)/m_i$
- Adjustments to infererence
  1. The estimated variance of $\hat{\beta}$ is multiplied by $\hat{\phi}$.
  2. For Wald type inferences (tests, ci's), the standard normal null distribution is replaced by t with $n-p$ degrees of freedom.
  3. A test statistic T that was assumed X^2_q under H_0 is replaced with $\frac{T}{q\hat{\phi}}$ and compared to an F distribution with q and $n-p$ degrees of freedom.
- Above are analogous to changes to inference for normal theory Gauss-Markov linear models if we switched from assuming $\sigma^2 = 1$ to assuming σ^2 was unknown and estimating it with MSE. (Here ϕ is like σ^2 and $\hat{\phi}$ is like MSE.)

Poisson regression

- What if y_i is a count?
- Often modelled by a Poisson distribution
- We begin with a review of the basics of the Poisson distribution.
- $y \sim Poisson(\mu) \Rightarrow$

$$\begin{aligned} f(y) &= \begin{cases} \frac{\mu^y e^{-\mu}}{y!} & \text{for } y \in \{0,1,2,...\} \\ 0 & \text{otherwise} \end{cases} \\ E(y) &= \mu \\ Var(y) &= \mu \end{aligned}$$

- μ must be ≥ 0

- The usual Poisson regression model: $y_i \sim Poisson(\mu_i)$ for $i = 1,..,n$
- $\mu_i \equiv \exp(\boldsymbol{x}_i'\boldsymbol{\beta})$
- $y_1, ..., y_n$ are independent.
- Note that $\mu_i \equiv \exp(\boldsymbol{x}_i'\boldsymbol{\beta}) \Leftrightarrow log(\mu_i) = \boldsymbol{x}_i'\boldsymbol{\beta}$
- Thus, log is the link function in this case.
- Both the Binomial and Poisson distributions are in the exponential family
- Both Logistic and Poisson regression are Generalized Linear Models.
- Estimation, inference, and model diagnosis analogous to those for logistic regression
- Differences: Var $y_i = \mu_i$ and log link instead of logit link

Interpretation of parameters in a Poisson model with a log link

- Consider two $\boldsymbol{X}_i$ vectors:
  $\boldsymbol{X}_1 = (\boldsymbol{X}_0, \boldsymbol{X}_1, \ldots \boldsymbol{X}_{j-1}, \quad \boldsymbol{X}_j, \quad \boldsymbol{X}_{j+1}, \ldots, \boldsymbol{X}_p)$, and
  $\boldsymbol{X}_2 = (\boldsymbol{X}_0, \boldsymbol{X}_1, \ldots \boldsymbol{X}_{j-1}, 1 + \boldsymbol{X}_j, \boldsymbol{X}_{j+1}, \ldots, \boldsymbol{X}_p)$
- $$\mu_2 = \exp(\boldsymbol{X}_2\boldsymbol{\beta}) = \exp(\boldsymbol{X}_1\boldsymbol{\beta} + \beta_j) = \mu_1 \exp(\beta_j)$$
- If X_j increases by one unit, the mean is multiplied by e^{β_j}
- Or, e^{β_j} is the multiplicative effect of a one unit change in X_j.

Using counts to model rates

- Sometimes the response is a count, but the real interest is a rate. e.g. Number of accidents per man-hours worked on a construction job
- Not Binomial because # man-hours may not be integer
  - define: y_i = # accidents, o_i = # man-hours,
  - Interested in $r_i = y_i/o_i$ as a function of covariates
- Model: $y_i \sim Poisson(o_i\, r_i) = Poisson(o_i \exp[f(\boldsymbol{X}_i, \beta)]$
- $\log \mu_i = \log o_i + \boldsymbol{X}_i\beta$
- $\log o_i$ is like another X variable, but associated β is exactly 1
- $\log o_i$ is called an offset
- NB: need $\log o_i$, not o_i, as the offset

R code for Binomial and Poisson models

```
# analysis of the trout aflatoxicol data

trout <- read.table('trout.txt',as.is=T,header=T)

# for binomial data, the response is a 2 col. matrix
#    (# event, # not event)
#   Often n is the number tested!

plot(trout$dose, trout$y/trout$n)

trout$f <- cbind(trout$y,trout$n-trout$y)

trout.glm <- glm(f~dose,data=trout,
   family=binomial)
```



```
# lack of fit test relative to means model
trout$dose.f <- factor(trout$dose)
trout.glm2 <- glm(f~dose.f,data=trout,
   family=binomial)

anova(trout.glm,trout.glm2,test='Chi')

# or could get in one fit from sequential SS
trout.glm2b <- glm(f~dose+dose.f,data=trout,
   family=binomial)
anova(trout.glm2b,test='Chi')
```



```
trout.glm3 <- glm(f~log(dose),data=trout,
   family=binomial)
plot(trout$dose,residuals(trout.glm3,log='x',
   type='deviance'))

# estimate and account for overdispersion
trout.glm4 <- glm(f~log(dose),data=trout,
   family=quasibinomial)
summary(trout.glm4)
# glm() uses Pearson Chi2/residual df as
#   estimate of overdisp.

anova(trout.glm4,test='F')
# use F instead of Chi2 when estimating over disp.
```

```
#  Here's how to fit a Poisson model - use trout
#    data, but assume count ~Poisson

trout.pglm <- glm(y~log(dose),data=trout,
   family=poisson)
summary(trout.pglm)
trout.pglm2 <- glm(y~log(dose),data=trout,
   family=quasipoisson)

# include offset as a vector, remember needs
#   to be a log scale value, # so to consider
#   # tested (trout$n) as the basis for a rate,
trout.pglm3 <- glm(y~log(dose),data=trout,
   family=quasipoisson, offset=log(n))
```