An Introduction to Linear Models

Models

- Concept of a Model Equation
- Other aspects of the model
  - Expected values, location parameters or first moments
  - Second moments or variance-covariance
  - Distributional assumptions

Simple Models

- Performance = Breeding + Feeding
- Phenotype = Genotype + Environment
- Animal Model – model equation

$$y = herd - year - season + BV + e$$

$$y = Xb + Zu + e$$

The "usual" Animal Model

$$y = Xb + Zu + e$$

1. Model Equation

$$E[u] = 0 \;\; and \;\; E[e] = 0$$

$$therefore \;\; E[y] = Xb$$

2. Location Parameters

$$var[u] = G = A\sigma_g^2 \;\; var[e] = R = I\sigma_e^2 \;\; \operatorname{cov}[u, e'] = 0$$

$$var[y] = V = ZGZ' + R$$

3. Dispersion Parameters

$$y \sim MVN[Xb, V]$$

4. Distributional Assumptions

Fixed Effects – Linear Regression

$$y = Xb + e$$

$$E[e] = 0$$

$$var[e] = R = I\sigma_e^2$$

$$Perhaps\ assume \quad e \sim N[0, I\sigma_e^2]$$

$$e_i \overset{iid}{\sim} N[0, \sigma_e^2]$$

Simple Linear Regression

$$y = Xb + e$$

$$b = \begin{bmatrix} \alpha \\ \beta \end{bmatrix} = \begin{bmatrix} \text{intercept} \\ \text{slope} \end{bmatrix}$$

$$X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}$$

Multiple Linear Regression

$$y = Xb + e$$

$$b = \begin{bmatrix} \alpha \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} = \begin{bmatrix} \text{intercept} \\ \text{slope}_1 \\ \vdots \\ \text{slope}_k \end{bmatrix}$$

$$X = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix}$$

Estimation

$$\begin{aligned} &\mathit{If} \\ &\quad y = Xb + e \\ &\mathit{then} \\ &\quad K'y = K'Xb + K'e \\ &\mathit{for\ example, choosing}\ K' = X' \\ &\quad X'y = X'Xb + X'e \\ &\mathit{and\ if}\ X'y = X'Xb\ \mathit{then}\ X'e = 0 \end{aligned}$$

$$\mathit{so}\ b\ \mathit{is\ solution\ to}\ X'Xb = X'y$$

Scheffe, The Analysis of Variance, 1959

Linear Regression

- Linear Regression
  $y = Xb + e$
- Residual
  $e = y - Xb$, with $E[e]=0$, and $\mathrm{var}[e]=I\sigma_e^2$
- Residual Sum of Squares
  $e'e = (y - Xb)'(y - Xb)$
  $= y'y - y'Xb - b'X'y + b'X'Xb$

Least Squares

- Residual Sum of Squares
  $e'e = y'y - y'Xb - b'X'y + b'X'Xb$
- Take derivatives with respect to vector b
  $de'e/db = - X'y - X'y + (X'X + (X'X)')b$
  set=0 and solve to find minima/maxima gives
  $X'Xb = X'y$
  known as the Least Squares Equations
  or the Normal Equations

Searle, Linear Models, 1971

Estimation

$$\widehat{b} \text{ is solution to } X'Xb = X'y$$
$$\text{which for full rank } X \text{ is } \widehat{b} = [X'X]^{-1}X'y$$
$$E[\widehat{b}] = E[[X'X]^{-1}X'y]$$
$$= [X'X]^{-1}X'E[y]$$
$$= [X'X]^{-1}X'Xb = b$$
$$var[\widehat{b}] = var[[X'X]^{-1}X'y]$$
$$= [X'X]^{-1}X'var[y]X[X'X]^{-1}$$
$$= [X'X]^{-1}X'I\sigma_e^2X[X'X]^{-1}$$
$$= [X'X]^{-1}X'X[X'X]^{-1}\sigma_e^2$$
$$= [X'X]^{-1}\sigma_e^2$$

Linear functions of b

$$k'b \text{ is estimated from } k'\widehat{b}$$
$$\text{with } var[k'\widehat{b}] = k'[X'X]^{-1}k\sigma_e^2$$

X not full rank

$$k'b \text{ is estimated from } k'\widehat{b}$$
$$\text{with } var[k'\widehat{b}] = k'[X'X]^{-}k\sigma_e^2$$
$$\text{provided } k' = k'[X'X]^{-}X'X$$

$$\text{rows of } k' \text{ can be stacked in a matrix } K$$
$$\text{vector } Kb \text{ is estimated from } K\widehat{b}$$
$$\text{with } var-\text{cov}[K\widehat{b}] = K[X'X]^{-}K'\sigma_e^2$$
$$\text{provided } K = K[X'X]^{-}X'X$$

Residual Standard Error

$$\widehat{\sigma_e^2} = MS_{ERROR} = SS_{ERROR}/df$$
$$= (y - X\widehat{b})'(y - X\widehat{b})/(N - rank(X))$$
$$SS_{ERROR} = SS_{TOTAL} - SS_{MODEL}$$
$$= y'y - \widehat{b}'X'y$$
$$R^2 = SS_{MODEL/MEAN}/SS_{TOTAL/MEAN}$$
$$SS_{MODEL/MEAN} = SS_{MODEL} - SS_{MEAN}$$
$$SS_{MEAN} = N\overline{y}^2$$
$$SS_{TOTAL/MEAN} = SS_{TOTAL} - SS_{MEAN}$$
$$= y'y - N\overline{y}^2$$

Generalized Least Squares

$$y = Xb + (Zu + e)$$
$$= Xb + \varepsilon$$

$$var[y] = V = ZGZ' + R$$

$$\widehat{b} \; is\ solution\ to\ X'V^{-1}Xb = X'V^{-1}y$$

Weighted Least Squares

$$y = Xb + e$$
$$var[e] = R = D = diag(\sigma^2_{e\,i})$$
$$\widehat{b} \; is\ solution\ to\ X'D^{-1}Xb = X'D^{-1}y$$

Hypothesis Testing

- To test hypotheses we need to know the distribution of the test statistic
  - Which is derived from the distribution of the residuals
    - Commonly assumed to be normally (iid) distributed

Linear Regression

1. Least Squares simple linear regression (unknown β_0 and β_1)
2. Gibbs Sampler with known σ_e^2
3. Bayesian Gibbs sampler with unknown σ_e^2
4. As above but with random not fixed β_1
5. Bayesian (multiple) linear regression (many random β's)
6. Various models (BLUP, BayesA, B, C, Cπ etc)