

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

$$y = \textit{herd} - \textit{year} - \textit{season} + BV + e$$

$$y = Xb + Zu + e$$

The “usual” Animal Model

$$y = Xb + Zu + e$$

$$E[u] = 0 \text{ and } E[e] = 0$$

$$\textit{therefore } E[y] = Xb$$

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

$$\textit{var}[y] = V = ZGZ' + R$$

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

1. Model Equation

2. Location Parameters

3. Dispersion Parameters

4. Distributional Assumptions

Fixed Effects – Linear Regression

$$y = Xb + e$$

$$E[u] = 0$$

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

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

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

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

If

$$y = Xb + e$$

then

$$K'y = K'Xb + K'e$$

for example, choosing $K' = X'$

$$X'y = X'Xb + X'e$$

and if $X'y = X'Xb$ then $X'e = 0$

so b is solution to $X'Xb = X'y$

Linear Regression

- Linear Regression

$$y = Xb + e$$

- Residual

$$e = y - Xb, \text{ with } E[e]=0, \text{ and } \text{var}[e]=\sigma_e^2$$

- 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

Estimation

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

which for full rank X is $\hat{b} = [X'X]^{-1}X'y$

$$E[\hat{b}] = E[[X'X]^{-1}X'y]$$

$$= [X'X]^{-1}X'E[y]$$

$$= [X'X]^{-1}X'Xb = b$$

$$\text{var}[\hat{b}] = \text{var}[[X'X]^{-1}X'y]$$

$$= [X'X]^{-1}X'\text{var}[y]X[X'X]^{-1}$$

$$= [X'X]^{-1}X'I\sigma_e^2X[X'X]^{-1}$$

$$= [X'X]^{-1}X'X[X'X]^{-1}\sigma_e^2$$

$$= [X'X]^{-1}\sigma_e^2$$

Linear functions of b

$k'b$ is estimated from $k'\hat{b}$

with $\text{var}[k'\hat{b}] = k'[X'X]^{-1}k\sigma_e^2$

X not full rank

$k'b$ is estimated from $k'\widehat{b}$

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

provided $k' = k'[X'X]^{-1}X'X$

rows of k' can be stacked in a matrix K

vector Kb is estimated from $K\widehat{b}$

with $\text{var} - \text{cov}[K\widehat{b}] = K[X'X]^{-1}K'\sigma_e^2$

provided $K = K[X'X]^{-1}X'X$

Residual Standard Error

$$\widehat{\sigma}_e^2 = MS_{ERROR} = SS_{ERROR} / df$$

$$= (y - X\widehat{b})'(y - X\widehat{b}) / (N - \text{rank}(X))$$

$$SS_{ERROR} = SS_{TOTAL} - SS_{MODEL}$$

$$= y'y - \widehat{b}'X'y$$

$$R^2 = SS_{MODEL/MEAN} / SS_{TOTAL/MEAN}$$

$$SS_{MODEL/MEAN} = SS_{MODEL} - SS_{MEAN}$$

$$SS_{MEAN} = N\bar{y}^2$$

$$SS_{TOTAL/MEAN} = SS_{TOTAL} - SS_{MEAN}$$

$$= y'y - N\bar{y}^2$$

Generalized Least Squares

$$\begin{aligned}y &= Xb + (Zu + e) \\ &= Xb + \varepsilon\end{aligned}$$

$$\text{var}[y] = V = ZGZ' + R$$

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

Weighted Least Squares

$$y = Xb + e$$

$$\text{var}[e] = R = D = \text{diag}(\sigma_{e_i}^2)$$

$$\hat{b} \text{ is solution to } X'D^{-1}Xb = X'D^{-1}y$$

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, C π etc)