STAT 500

Linear Models

Linear Models

Linear models provide a unified approach to many models

- One-way ANOVA (including two-independent samples)
- Block designs with fixed blocks (including matched pairs)
- Two-way ANOVA
- Simple Linear Regression
- Multiple Linear Regression

Linear Models

Any linear model can be written in the form

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

$$\begin{array}{cccccc}
\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} & = & \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1k} \\ x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix} & \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} & + & \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix} \\
\uparrow & & \uparrow & \uparrow & & \uparrow
\end{array}$$

response vector	the elements of design matrix X are known (non-random) values	unknown parameters	random errors (not observed)

Linear Models

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}$$ is a random vector

(1) $E(\mathbf{Y}) = X\boldsymbol{\beta}$ is a vector of expected responses for some known matrix X of constants and some unknown parameter vector $\boldsymbol{\beta}$

(2) $Var(\mathbf{Y}) = \Sigma$

(3) Complete the model by specifying a probability distribution for the possible values of $\mathbf{Y}$ or $\boldsymbol{\epsilon}$

Gauss-Markov Model

The linear model $\mathbf{Y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$ is called a Gauss-Markov model if

$$Var(\mathbf{Y}) = Var(\boldsymbol{\epsilon}) = \sigma^2 I$$

for some unknown constant σ^2.

For a Gauss-Markov Model

- The observations (and the random errors) are mutually uncorrelated
- Every observation (and every random error) has the same variance

Normal Theory Gauss-Markov Model

A normal theory Gauss-Markov model is a Gauss-Markov model where $\mathbf{Y}$ (and $\boldsymbol{\epsilon}$) has a multivariate normal distribution.

$$\mathbf{Y} \sim N\,(X\boldsymbol{\beta}\,,\ \sigma^2 I) \text{ implying } \boldsymbol{\epsilon} \sim N(\mathbf{0}\,,\ \sigma^2 I)$$

The additional assumption of a normal distribution is

(1) not needed for most estimation results

(2) used to create confidence intervals and perform tests of hypotheses

(3) used to obtain distributions for test statistics

Linear Model – Regression

Example 1: Yield of a chemical process

- Response Variable = Yield (Y)
- Explanatory Variable 1 = Temperature (x_1)
- Explanatory Variable 2 = Time (x_2)
- $n = 5$ observations

Linear Models: Regression Models

Example 1: Yield of a chemical process

Yield (%) Y	Temperature (^{o}F) X_1	Time (hr) X_2
77	160	1
82	165	3
84	165	2
89	170	1
94	175	2

Regression model

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \epsilon_i \quad \text{for } i = 1, 2, 3, 4, 5$$

Linear Models: Regression Models

$$
\begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \\ Y_5 \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} \\ 1 & x_{21} & x_{22} \\ 1 & x_{31} & x_{32} \\ 1 & x_{41} & x_{42} \\ 1 & x_{51} & x_{52} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \\ \epsilon_5 \end{bmatrix}
$$

ANOVA: Regression Models

Source of Variation	d.f.	Sums of Squares	Mean Square
Model	2	$\sum_{i=1}^{5}(\widehat{Y}_i - \bar{Y})^2 = \mathbf{Y}^T(P_X - P_1)\mathbf{Y}$	$\frac{1}{2}SS_{\text{model}}$
Error	2	$\sum_{i=1}^{5}(Y_i - \widehat{Y}_i)^2 = \mathbf{Y}^T(I - P_X)\mathbf{Y}$	$\frac{1}{2}SS_{\text{error}}$
C. total	4	$\sum_{i=1}^{5}(Y_i - \bar{Y})^2 = \mathbf{Y}^T(I - P_1)\mathbf{Y}$	

where $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$ and $\widehat{Y}_i = b_0 + b_1 x_{1i} + b_2 x_{2i}$

$P_X = X(X^TX)^{-1}X^T$ and $P_1 = \mathbf{1}(\mathbf{1}^T\mathbf{1})^{-1}\mathbf{1}^T$

ANOVA: Regression Models

The corrected total sum of squares is

$$SS_{\text{corrected total}} = \sum_{i=1}^{n} (Y_i - \bar{Y})^2 = (\mathbf{Y} - \bar{Y}\mathbf{1})^T(\mathbf{Y} - \bar{Y}\mathbf{1})$$

Note that

$$P_{\mathbf{1}}\mathbf{Y} = \mathbf{1}(\mathbf{1}^T\mathbf{1})^{-1}\mathbf{1}^T\mathbf{Y} = \mathbf{1}\frac{1}{n}\sum_{i=1}^{n} Y_i = \bar{Y}\mathbf{1}$$

and

$$(\mathbf{Y} - \bar{Y}\mathbf{1}) = \mathbf{Y} - P_{\mathbf{1}}\mathbf{Y} = (I - P_{\mathbf{1}})\mathbf{Y}$$

ANOVA: Regression Models

Then

$$\begin{aligned}
SS_{\text{corrected total}} = \sum_{i=1}^{n}(Y_i - \bar{Y})^2 &= (\mathbf{Y} - \bar{Y}\mathbf{1})^T(\mathbf{Y} - \bar{Y}\mathbf{1}) \\
&= ((I - P_{\mathbf{1}})\mathbf{Y})^T(I - P_{\mathbf{1}})\mathbf{Y} \\
&= \mathbf{Y}^T(I - P_{\mathbf{1}})^T(I - P_{\mathbf{1}})\mathbf{Y} \\
&= \mathbf{Y}^T(I - P_{\mathbf{1}})(I - P_{\mathbf{1}})\mathbf{Y} \\
&= \mathbf{Y}^T(I - P_{\mathbf{1}})\mathbf{Y}
\end{aligned}$$

because $(I - P_{\mathbf{1}})$ is a symmetric and idempotent matrix

ANOVA: Regression Models

$$SS_{Model} = \sum_{i=1}^{5} (\hat{Y}_i - \bar{Y})^2 = (\hat{\mathbf{Y}} - \bar{Y}\mathbf{1})^T(\hat{\mathbf{Y}} - \bar{Y}\mathbf{1})$$

Note that

$$\hat{\mathbf{Y}} = X\mathbf{b} = X(X^T X)^{-1} X^T \mathbf{Y} = P_X \mathbf{Y}$$

and

$$\hat{\mathbf{Y}} - \bar{Y}\mathbf{1} = P_X \mathbf{Y} - P_{\mathbf{1}} \mathbf{Y} = (P_X - P_{\mathbf{1}})\mathbf{Y}$$

ANOVA: Regression Models

Then

$$
\begin{aligned}
SS_{\text{model}} = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2 &= (\hat{\mathbf{Y}} - \bar{Y}\mathbf{1})^T (\hat{\mathbf{Y}} - \bar{Y}\mathbf{1}) \\
&= ((P_X - P_{\mathbf{1}})\mathbf{Y})^T (P_X - P_{\mathbf{1}})\mathbf{Y} \\
&= \mathbf{Y}^T (P_X - P_{\mathbf{1}})^T (P_X - P_{\mathbf{1}})\mathbf{Y} \\
&= \mathbf{Y}^T (P_X - P_{\mathbf{1}})(P_X - P_{\mathbf{1}})\mathbf{Y} \\
&= \mathbf{Y}^T (P_X - P_{\mathbf{1}})\mathbf{Y}
\end{aligned}
$$

because $(P_X - P_{\mathbf{1}})$ is a symmetric and idempotent matrix

ANOVA: Regression Models

Then

$$
\begin{aligned}
SS_{\text{residuals}} = \sum_{i=1}^{n} (Y_i - \widehat{Y}_i)^2 &= (\mathbf{Y} - \widehat{\mathbf{Y}})^T (\mathbf{Y} - \widehat{\mathbf{Y}}) \\
&= ((I - P_X)\mathbf{Y})^T (I - P_X)\mathbf{Y} \\
&= \mathbf{Y}^T (I - P_X)^T (I - P_X)\mathbf{Y} \\
&= \mathbf{Y}^T (I - P_X)(I - P_X)\mathbf{Y} \\
&= \mathbf{Y}^T (I - P_X)\mathbf{Y}
\end{aligned}
$$

because $I - P_X$ is a symmetric and idempotent matrix

ANOVA: Regression Models

Partition the corrected total sum of squares:

$$
\begin{aligned}
SS_{\text{corrected total}} &= \mathbf{Y}^T(I - P_{\mathbf{1}})\mathbf{Y} \\
&= \mathbf{Y}^T(I - P_X + P_X - P_{\mathbf{1}})\mathbf{Y} \\
&= \mathbf{Y}^T(I - P_X)\mathbf{Y} + \mathbf{Y}^T(P_X - P_{\mathbf{1}})\mathbf{Y} \\
&= SS_{\text{residuals}} + SS_{\text{regression model}}
\end{aligned}
$$

Linear Model – One-way ANOVA

Example 2: Blood coagulation times (in seconds) for blood samples from 12 different rats. Each rat was fed one of three diets, with 4 rats per diet.

- Response Variable = Blood coagulation times (Y)
- Explanatory Variable = Diet (A, B, or C)
- $n = 12$ observations

Linear Model – One-way ANOVA

Cell Means Model

$$\begin{bmatrix} Y_{11} \\ Y_{12} \\ Y_{13} \\ Y_{14} \\ Y_{21} \\ Y_{22} \\ Y_{23} \\ Y_{24} \\ Y_{31} \\ Y_{32} \\ Y_{33} \\ Y_{34} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{bmatrix} + \begin{bmatrix} \epsilon_{11} \\ \epsilon_{12} \\ \epsilon_{13} \\ \epsilon_{14} \\ \epsilon_{21} \\ \epsilon_{22} \\ \epsilon_{23} \\ \epsilon_{24} \\ \epsilon_{31} \\ \epsilon_{32} \\ \epsilon_{33} \\ \epsilon_{34} \end{bmatrix}$$

Linear Model – One-way ANOVA

Effects Model

$$\begin{bmatrix} Y_{11} \\ Y_{12} \\ Y_{13} \\ Y_{14} \\ Y_{21} \\ Y_{22} \\ Y_{23} \\ Y_{24} \\ Y_{31} \\ Y_{32} \\ Y_{33} \\ Y_{34} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{bmatrix} + \begin{bmatrix} \epsilon_{11} \\ \epsilon_{12} \\ \epsilon_{13} \\ \epsilon_{14} \\ \epsilon_{21} \\ \epsilon_{22} \\ \epsilon_{23} \\ \epsilon_{24} \\ \epsilon_{31} \\ \epsilon_{32} \\ \epsilon_{33} \\ \epsilon_{34} \end{bmatrix}$$

Linear Model – Two-Way ANOVA

Example 3: A full factorial experiment

- Experimental Units - 8 plots of trees - 5 trees per plot.
- Response Variable = Percentage of apples with spots (Y)
- Explanatory Variable 1 = Variety of Apple (A or B)
- Explanatory Variable 2 = Fungicide use (new or old)
- $n = 8$ observations

Linear Model – Two-Way ANOVA

Cell Means Model

$$\begin{bmatrix} Y_{111} \\ Y_{112} \\ Y_{121} \\ Y_{122} \\ Y_{211} \\ Y_{212} \\ Y_{221} \\ Y_{222} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \mu_{11} \\ \mu_{12} \\ \mu_{21} \\ \mu_{22} \end{bmatrix} + \begin{bmatrix} \epsilon_{111} \\ \epsilon_{112} \\ \epsilon_{121} \\ \epsilon_{122} \\ \epsilon_{211} \\ \epsilon_{212} \\ \epsilon_{221} \\ \epsilon_{222} \end{bmatrix}$$

Linear Model – Two-Way ANOVA

Effects Model

$$\begin{bmatrix} Y_{111} \\ Y_{112} \\ Y_{121} \\ Y_{122} \\ Y_{211} \\ Y_{212} \\ Y_{221} \\ Y_{222} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \tau_1 \\ \tau_2 \\ (\alpha\tau)_{11} \\ (\alpha\tau)_{12} \\ (\alpha\tau)_{21} \\ (\alpha\tau)_{22} \end{bmatrix} + \begin{bmatrix} \epsilon_{111} \\ \epsilon_{112} \\ \epsilon_{121} \\ \epsilon_{122} \\ \epsilon_{211} \\ \epsilon_{212} \\ \epsilon_{221} \\ \epsilon_{222} \end{bmatrix}$$