Lecture 3: Normal Linear Models — Types and Matrix Form

MATH3823 Generalised Linear Models

Richard P Mann

MATH3823 Generalised Linear Models

Reading

Course notes: Chapter 2, Sections 2.3–2.4

`www.richardpmann.com/MATH3823`

Types of Normal Linear Models

p	Explanatory Variables	Model Name
1	Quantitative	Simple linear regression
> 1	Quantitative	Multiple linear regression
1	Dichotomous (2 levels)	Two-sample t-test
1	Polytomous (k levels)	One-way ANOVA
> 1	Qualitative	Multi-way ANOVA
> 1	Mixed (quant. + qual.)	ANCOVA

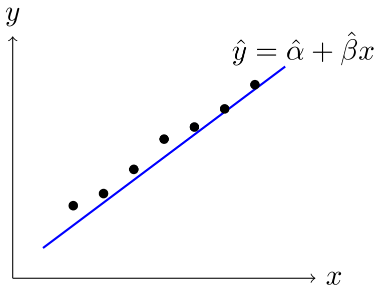
Key insight: All of these are *special cases* of the general linear model.

Simple Linear Regression

Model:

$$y_i = \alpha + \beta x_i + \epsilon_i, \quad \epsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$$

Example: Height vs. weight, dose vs. response



Multiple Linear Regression

Model:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \epsilon_i$$

Example: House price depending on size, age, location, etc.

Interpretation:

- $\beta_j =$ change in $\mathbb{E}[Y]$ per unit increase in x_j, *holding other variables constant*
- This is a **partial** or **adjusted** effect

Warning: In observational studies, "holding constant" is conceptual, not causal.

Two-Sample t-Test as a Linear Model

Setting: Compare means of two groups

Traditional formulation:

$$Y_{1j} \sim \mathcal{N}(\mu_1, \sigma^2), \quad j = 1, \ldots, n_1$$
$$Y_{2j} \sim \mathcal{N}(\mu_2, \sigma^2), \quad j = 1, \ldots, n_2$$

As a linear model:

$$y_i = \alpha + \gamma \cdot \text{Group}_i + \epsilon_i$$

where $\text{Group}_i = 0$ for group 1, $\text{Group}_i = 1$ for group 2.

Parameters:

- $\alpha = \mu_1$ (mean of reference group)
- $\gamma = \mu_2 - \mu_1$ (difference in means)

One-Way ANOVA as a Linear Model

Setting: Compare means across k groups

Model:

$$y_{ij} = \mu + \alpha_i + \epsilon_{ij}, \quad i = 1, \ldots, k, \quad j = 1, \ldots, n_i$$

Problem: The model is **overparameterized**.

- We have $k + 1$ parameters $(\mu, \alpha_1, \ldots, \alpha_k)$
- But only k group means to estimate

Solution: Add a constraint (identifiability condition)

Parameter Constraints

Option 1: Corner constraint (R default)

$$\alpha_1 = 0$$

- Group 1 is the reference category
- μ = mean of group 1
- α_i = difference between group i and group 1

Option 2: Sum-to-zero constraint

$$\sum_{i=1}^{k} \alpha_i = 0$$

- μ = grand mean (average of group means)
- α_i = deviation of group i from grand mean

Matrix Formulation: The General Linear Model

For n observations:

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

where:

- $\mathbf{Y} = (Y_1, \ldots, Y_n)'$ is the $n \times 1$ response vector
- $\mathbf{X}$ is the $n \times p$ **design matrix**
- $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ is the $p \times 1$ parameter vector
- $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)'$ is the $n \times 1$ error vector

Assumptions:

$$\boldsymbol{\epsilon} \sim \mathcal{N}_n(\mathbf{0}, \sigma^2 \mathbf{I}_n)$$

Design Matrix: Simple Linear Regression

Model: $y_i = \alpha + \beta x_i + \epsilon_i$

$$\underbrace{\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}}_{\mathbf{Y}} = \underbrace{\begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}}_{\mathbf{X}} \underbrace{\begin{pmatrix} \alpha \\ \beta \end{pmatrix}}_{\boldsymbol{\beta}} + \underbrace{\begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}}_{\boldsymbol{\epsilon}}$$

- First column of **X** is all 1s (for intercept)
- Second column contains the x values

Design Matrix: One-Way ANOVA

Model: $y_{ij} = \mu + \alpha_i + \epsilon_{ij}$ with $k = 3$ groups

Full (overparameterized) design matrix:

$$\mathbf{X}_{\text{full}} = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \end{pmatrix} \quad \text{(not full rank)}$$

With corner constraint ($\alpha_1 = 0$):

$$\mathbf{X} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{pmatrix} \quad \boldsymbol{\beta} = \begin{pmatrix} \mu \\ \alpha_2 \\ \alpha_3 \end{pmatrix}$$

Constructing the Design Matrix

Recipe:

1. Start with a column of 1s (intercept)
2. For each **quantitative** variable: add one column of values
3. For each **qualitative** variable with k levels:
    - Create k dummy (indicator) columns
    - Remove one column to avoid singularity
4. For interactions: multiply corresponding columns element-wise

Result: A full-rank $n \times p$ matrix where $p =$ number of free parameters.

Least Squares Solution in Matrix Form

Residual sum of squares:

$$\mathrm{RSS}(\boldsymbol{\beta}) = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})$$

Normal equations:

$$\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{Y}$$

Least squares estimator:

$$\boxed{\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}}$$

Properties:

- $\mathbb{E}[\hat{\boldsymbol{\beta}}] = \boldsymbol{\beta}$ (unbiased)
- $\mathrm{Var}[\hat{\boldsymbol{\beta}}] = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$

Fitted Values and Residuals

Fitted values:

$$\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = \mathbf{H}\mathbf{Y}$$

where $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ is the **hat matrix**.

Residuals:

$$\mathbf{r} = \mathbf{Y} - \hat{\mathbf{Y}} = (\mathbf{I} - \mathbf{H})\mathbf{Y}$$

Error variance estimate:

$$\hat{\sigma}^2 = \frac{\mathbf{r}'\mathbf{r}}{n-p} = \frac{\mathrm{RSS}}{n-p}$$

Viewing the Design Matrix in R

```r
# Create a factor variable
group <- factor(c("A", "A", "B", "B", "C", "C"))

# See the design matrix
model.matrix(~ group)
```

Output:

```
  (Intercept) groupB groupC
1           1      0      0
2           1      0      0
3           1      1      0
4           1      1      0
5           1      0      1
6           1      0      1
```

Note: Group A is the reference (no column for it).

Summary

Key points:

- Many statistical tests are special cases of linear models
- The matrix formulation $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ unifies all cases
- Design matrix $\mathbf{X}$ encodes the model structure
- Qualitative variables require dummy coding
- Constraints are needed to avoid overparameterization
- Least squares solution: $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$

Next lecture: Model notation, R formula syntax, and fitting in practice.