STA 610L: Module 2.3

One way ANOVA (distribution of estimates and linear combinations)

Dr. Olanrewaju Michael Akande

STA 610L

Linear model estimates

Consider a very simple one-sample linear model given by $y_i = \mu + \varepsilon_i$, $\varepsilon_i \sim N(0, \sigma^2)$.

In matrix notation, this model can be written as

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} \left(\mu \right) + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

with the vector $\varepsilon \sim N(0_{n\times 1}, \sigma^2 I_{n\times n})$.

STA 610L

MLEs

Recalling that the normal distribution for one observation is given by

$$\frac{1}{\sqrt{2\pi\sigma}} \exp -\frac{1}{2}(y_i - \mu)^2.$$

We can obtain the likelihood by taking the product over all n independent observations:

$$\begin{aligned} L(y, \mu, \sigma) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma}} \exp\left\{-\frac{1}{2}\frac{(y_i - \mu)^2}{\sigma^2}\right\} \\ &= \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n}{2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu)^2\right\}. \end{aligned}$$

To find the MLE solve for the parameter values that make the first derivative equal to 0 (often we work with the log-likelihood as it is more convenient).

STA 610L

MLEs

The log-likelihood is given by

$$\ell(y, \mu, \sigma^2) = \frac{n}{2}\log\frac{1}{2\pi\sigma^2} - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu)^2$$

$$= -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu)^2$$

STA 610L

MLEs

To find the MLE of μ, take the derivative

$$\begin{aligned}
\frac{\partial \ell(\mu, \sigma^2)}{\partial \mu} &= 0 - \frac{1}{2\sigma^2} \sum_{i=1}^{n} 2(y_i - \mu)(-1) \\
&= \frac{1}{\sigma^2} \left(\sum_{i=1}^{n} y_i - n\mu \right)
\end{aligned}$$

Setting this equal to zero, we obtain the MLE

$$\begin{aligned}
n\widehat{\mu} &= \sum_{i=1}^{n} y_i \\
\widehat{\mu} &= \frac{\sum_{i=1}^{n} y_i}{n} = \overline{y}
\end{aligned}$$

STA 610L

MLEs

To find the MLE of σ^2 take the derivative

$$
\begin{aligned}
\frac{\partial \ell(\mu, \sigma^2)}{\partial \sigma^2} &= 0 - \frac{n}{2}\frac{1}{\sigma^2} - \frac{1}{2(-1)(\sigma^2)^2}\sum_{i=1}^{n}(y_i - \mu)^2 \\
&= -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}\sum_{i=1}^{n}(y_i - \mu)^2
\end{aligned}
$$

Setting to 0 and solving for the MLE, using the MLE of μ we just found, we obtain

$$
\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2.
$$

Note this MLE of σ^2 is **not** the usual (unbiased) sample variance s^2. We will return to this point later in the course.

STA 610L

Properties of MLEs

For any MLE $\hat{\theta}$,

- $\hat{\theta} \to \theta$ as $n \to \infty$ (if the model is correct)
- $\hat{\theta} \sim N\left(\theta, \frac{1}{n}\mathcal{I}^{-1}\right)$, where $\mathcal{I}$ is the Fisher information.
- Alternatively, $\hat{\theta} \sim N\left(\theta, \mathrm{Var}(\hat{\theta})\right)$, where $\mathrm{Var}(\hat{\theta}) \approx \left[\frac{d^2 l(\theta|y)}{d\theta^2}\right]^{-1}$ in large samples

For the hierarchical model, this gives us a method for getting approximate 95% confidence intervals for mean parameters (and functions of them).

However, since the variance itself actually includes the unknown parameter, we would have to rely on an estimated version.

STA 610L

Information

The observed information matrix is the matrix of second derivatives of the negative log-likelihood function at the MLE (Hessian matrix):

$$J(\hat{\theta}) = \left\{ -\frac{\partial^2 \ell(\theta \mid y)}{\partial\theta_j \partial\theta_k} \right\} \Big|_{\theta=\hat{\theta}}$$

The inverse of the information matrix gives us an estimate of the variance/covariance of MLE's:

$$\widehat{\mathrm{Var}}(\hat{\theta}) \approx J^{-1}(\hat{\theta})$$

The square roots of the diagonal elements of this matrix give approximate SE's for the coefficients, and the MLE $\pm$ 2 SE gives a rough 95% confidence interval for the parameters.

Motivating example: cycling safety

In the cycling safety study, after we found evidence that the rider's distance from the curb was related to passing distance (the overall F test), we wanted to learn what kind of relationship existed (pairwise comparisons).

Each pairwise comparison is defined by a linear combination of the parameters in our model.

Consider the treatment means model $y_{ij} = \mu_j + \varepsilon_{ij}$.

We are interested in which $\mu_j \neq \mu_j'$.

STA 610L

Distribution of least squares estimates

Recall in the linear model, the least squares estimate $\widehat{\beta} = (X'X)^{-1}X'y$.

Its covariance is given by $\mathrm{Cov}(\widehat{\beta}) = \sigma^2(X'X)^{-1}$.

In large samples, or when our errors are exactly normal, $\widehat{\beta} \sim N\left(\beta, \sigma^2(X'X)^{-1}\right)$.

Linear combinations

In order to test whether the means in group 1 and 2 are the same, we need to test a hypothesis about a *linear combination* of parameters.

The quantity $\sum_j a_j\mu_j$ is a *linear combination*. It is called a contrast if $\sum_j a_j = 0$.

Using matrix notation, we often express a hypothesis regarding a linear combination of regression coefficients as

$$\begin{aligned} H_0: &\quad \theta = C\beta = \theta_0 \\ H_a: &\quad \theta = C\beta \neq \theta_0, \end{aligned}$$

where often $\theta_0 = 0$.

STA 610L

LINEAR COMBINATIONS

For example, suppose we have three groups in the model $y_{ij} = \mu_j + \varepsilon_{ij}$ and want to test differences in all pairwise comparisons. We can set

- $\beta = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{pmatrix}$,
- $C = \begin{pmatrix} 1 & -1 & 0 \\ 1 & 0 & -1 \\ 0 & 1 & -1 \end{pmatrix}$, and
- $\theta_0 = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$,

so that our hypothesis is that $\begin{pmatrix} \mu_1 - \mu_2 \\ \mu_1 - \mu_3 \\ \mu_2 - \mu_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$.

Distributional results for linear combinations

Using basic properties of the multivariate normal distribution, we have

$$C\widehat{\beta} \sim N\left(C\beta, \sigma^2 C(X'X)^{-1}C'\right).$$

Using this result, you can derive the standard error for any linear combination of parameter estimates, which can be used in constructing confidence intervals.

You could also fit a reduced model subject to the constraint you wish to test (e.g., same mean for groups 1 and 2), and then use either a partial F test or a likelihood-ratio test (F is special case of LRT) to evaluate the hypothesis that the reduced model is adequate.

We will implement this later in R.

What's next?

Move on to the readings for the next module!