Chapter 11
Asymptotic Evaluations
Analysis of Variance

Outline

Basic Set-up

For linear models, we begin with a general structure

$$y = X\beta + \epsilon.$$

- y is a matrix whose rows form a series of multivariate measurements, the response variables,
- X is a matrix of explanatory variables,
- β is a matrix of parameters, and
- ϵ is a matrix containing residuals (i.e., errors or noise).

If the residuals have a multivariate normal distribution, then least squares estimation is a maximum likelihood estimation procedure for the β.

Multiple Linear Regression

For multiple linear regression:

- $y = (y_1, y_2, \ldots, y_n)^T$ is a column vector of responses,
- X is a matrix of predictors,

$$X = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1k} \\ 1 & x_{21} & \cdots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{nk} \end{pmatrix}.$$

- $\beta = (\beta_0, \beta_1, \ldots, \beta_k)^T$ is a column vector of parameters, and
- $\epsilon = (\epsilon_1, \epsilon_2, \ldots, \epsilon_n)^T$ is a column vector of “errors”.

$$y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{1k} + \epsilon_i.$$

Analysis of Variance

Example. The data on 30 forest plots in Borneo are the number of trees per plot.

	never logged	logged 1 year ago	logged 8 years ago
n_j	12	12	9
$\bar{y}_j$	23.750	14.083	15.778
s_j	5.065	4.981	5.761

We compute these statistics from the data $y_{11}, \ldots y_{n_1 1}$, $y_{12}, \ldots y_{n_2 2}$ and $y_{13}, \ldots y_{n_3 3}$,

$$\bar{y}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} y_{ij} \quad \text{and} \quad s_j^2 = \frac{1}{n_j - 1} \sum_{i=1}^{n_j} (y_{ij} - \bar{y}_j)^2.$$

Overview

- The basic question is: Are these means the same (the null hypothesis) or not (the alternative hypothesis)?
- The basic idea of the test is to examine the ratio of $s^2_{between}$, the variance between groups (indicated by the variation in the center lines of the boxes) and $s^2_{residual}$, a statistic that measures the variances within groups.
- If the resulting ratio test statistic is sufficiently large, then we say, based on the data, that the means of these groups are distinct and we reject H_0.

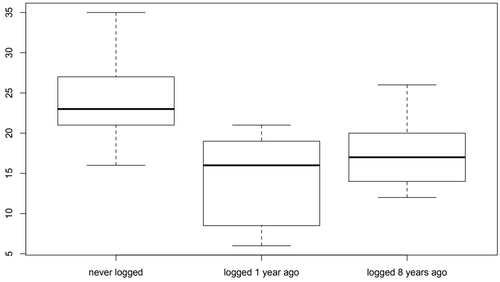


Figure: Side-by-side boxplots of the number of trees per plot.

One Way Analysis of Variance

The hypothesis for one way analysis of variance is

$$H_0 : \mu_j = \mu_k \text{ for all } j, k \quad \text{and} \quad H_1 : \mu_j \neq \mu_k \text{ for some } j, k.$$

The data $\{y_{ij}, 1 \leq i \leq n_j, 1 \leq j \leq q\}$ represents that we have n_j observation for the j-th group and that we have q groups. The total number of observations is denoted by $n = n_1 + \cdots + n_q$. The model is

$$y_{ij} = \mu_j + \epsilon_{ij}$$

where ϵ_{ij} are independent $N(0, \sigma^2)$ random variables with σ^2 unknown.

For X, here called the design matrix, x_{ij} is 1 if the i-th observation belongs to group j and 0 otherwise.

Linear Models

Assume that $\beta \in \mathbb{R}^m$ and that X is a $n \times m$ matrix of rank $m < n$. Let $Y_1, \ldots, Y_n$ are independent normally distributed random variables with mean vector $\mu = X\beta$. Then, the likelihood ratio test of the hypothesis

$$H_0 : A\beta = 0 \quad \text{versus} \quad H_1 : A\beta \neq 0.$$

where A is a $r \times m$ matrix has critical region

$$C = \{\mathbf{y}; F(\mathbf{y}) \geq F_0\}.$$

F is given by

$$F(\mathbf{y}) = \frac{\sum_{k=1}^n (y_k - \widehat{\widehat{\mu}}_k)^2 - \sum_{k=1}^n (y_k - \widehat{\mu}_k)^2}{\sum_{k=1}^n (y_k - \widehat{\mu}_k)^2} \frac{n-m}{r}.$$

Linear Models

For the expression

$$\sum_{k=1}^{n}(y_k - \mu_k)^2,$$

- the vector $\widehat{\mu}$ is the minimum value under the restriction $\mu = X\beta$, and
- The vector $\widehat{\widehat{\mu}}$ is the minimum value under the pair of restrictions $\mu = X\beta$ and $A\beta = 0$.

Proof. The likelihood function

$$L(\beta, \sigma^2|\mathbf{x}, \mathbf{y}) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp -\frac{1}{2\sigma^2}\sum_{k=1}^{n}(y_k - \mu_k)^2 = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp -\frac{1}{2\sigma^2}(\mathbf{y} - \mu)^T(\mathbf{y} - \mu)$$

Linear Models

The likelihood ratio

$$\Lambda(\mathbf{x},\mathbf{y}) = \frac{\sup\{L(\beta,\sigma^2|\mathbf{x},\mathbf{y}); \mathbf{y} = X\beta, A\beta = 0\}}{\sup\{L(\beta,\sigma^2|\mathbf{x},\mathbf{y}); \mathbf{y} = X\beta\}}$$

For the numerator, let $\hat{\beta}$ be the maximum likelihood estimator for the parameter β and let $\widehat{\mu} = X\hat{\beta}$. Then, the the maximum likelihood estimator for σ^2 is

$$\widehat{\sigma^2} = \frac{1}{n}\sum_{k=1}^{n}(y_k - \widehat{\mu}_k)^2 = \frac{1}{n}(\mathbf{y} - \widehat{\mu})^T(\mathbf{y} - \widehat{\mu})$$

Therefore

$$L(\hat{\beta},\widehat{\sigma^2}|\mathbf{x},\mathbf{y}) = \frac{\exp -\frac{n}{2}}{(2\pi\widehat{\sigma^2})^{n/2}}.$$

Linear Models

Similarly, for the denominator, let $\widehat{\widehat{\mu}} = X\hat{\hat{\beta}}$ and $\widehat{\widehat{\sigma^2}}$ be the corresponding maximum likelihood estimates when the null hypothesis is true. Then,

$$L(\hat{\hat{\beta}}, \widehat{\widehat{\sigma^2}}|\mathbf{x}, \mathbf{y}) = \frac{\exp -\frac{n}{2}}{(2\pi\widehat{\widehat{\sigma^2}})^{n/2}}.$$

Consequently, the likelihood ratio test,

$$\lambda_0 \geq \Lambda(\mathbf{x}, \mathbf{y}) = \left(\frac{\widehat{\sigma^2}}{\widehat{\widehat{\sigma^2}}}\right)^n \qquad \lambda_0^{-1/n} - 1 \leq \frac{\widehat{\widehat{\sigma^2}}}{\widehat{\sigma^2}} - 1$$

$$\begin{aligned} F(\mathbf{y}) &= (\lambda_0^{-1/n} - 1)\frac{n-m}{r} \leq \left(\frac{\widehat{\widehat{\sigma^2}}}{\widehat{\sigma^2}} - 1\right)\frac{n-m}{r} \\ &= \frac{\sum_{k=1}^n (y_i k - \widehat{\widehat{\mu}}_k)^2 - \sum_{k=1}^n (y_k - \widehat{\mu}_k)^2}{\sum_{k=1}^n (y_k - \widehat{\mu}_k)^2}\frac{n-m}{r}. \end{aligned}$$

Sample Means

For the F statistic, we introduce two types of sample means:

- For Θ, differentiate with respect to μ_j. The maximum valueis the within group means, the sample mean inside each of the groups,

$$\hat{\mu}_j = \bar{y}_j = \frac{1}{n_j}\sum_{i=1}^{n_j} y_{ij}, \quad j = 1. \ldots, q.$$

- For Θ_0, The For μ_j are all equal to some value For μ. So, differentiate with respect to μ to see that the maximum value is the mean of the data taken as a whole, known as the grand mean,

$$\hat{\hat{\mu}} = \bar{\bar{y}} = \frac{1}{n}\sum_{j=1}^{q}\sum_{i=1}^{n_j} y_{ij} = \frac{1}{n}\sum_{j=1}^{q} n_j \bar{y}_j,$$

the weighted average of the $\bar{y}_j$ with weights n_j, the sample size in each group.

Exercise. For the Borneo rain forest example, show that the grand mean is 18.06055.

Sums of Squares

For the numerator of $F(\mathbf{y})$, we have total sums of squares

$$SS_{total} = \sum_{k=1}^{n}(y_k - \widehat{\mu}_k) = \sum_{j=1}^{q}\sum_{i=1}^{n_j}(y_{ij} - \overline{\overline{y}})^2,$$

the total square variation of individual observations from their grand mean. The test statistic is determined by decomposing SS_{total}. We first rewrite the interior sum as

$$\sum_{i=1}^{n_j}(y_{ij} - \overline{\overline{y}})^2 = \sum_{i=1}^{n_j}(y_{ij} - \overline{y}_j)^2 + n_j(\overline{y}_j - \overline{\overline{y}})^2 = (n_j - 1)s_j^2 + n_j(\overline{y}_j - \overline{\overline{y}})^2.$$

Here, s_j^2 is the unbiased sample variance based on the observations in the j-th group.

Exercise. Show the first equality above. (Hint: Begin with the difference in the two sums.)

Sums of Squares

Here $m = q$, the number of groups. Also,

$$A = \begin{pmatrix} 1 & 0 & \cdots & 0 & -1 \\ 0 & 1 & \cdots & 0 & -1 \\ \vdots & \vdots & \ddots & 0 & -1 \\ 0 & 0 & \cdots & 1 & -1 \end{pmatrix}$$

has $q-1$ rows showing $\mu_j = \mu_q,\ j = 1, \ldots, q-1$. Thus, $\text{rank}(A) = q - 1$.

$$\begin{aligned} F(\mathbf{y}) &= \frac{\sum_{k=1}^n (y_k - \widehat{\widehat{\mu}}_k)^2 - \sum_{k=1}^n (y_k - \widehat{\mu}_k)^2}{\sum_{k=1}^n (y_k - \widehat{\mu}_k)^2} \frac{n-m}{r}. \\ &= \frac{\sum_{j=1}^q n_j(\bar{y}_j - \bar{\bar{y}})^2/(q-1)}{\sum_{j=1}^q (n_j - 1)s_j^2/(n-q)} = \frac{SS_{between}/(q-1)}{SS_{residual}/(n-q)} \end{aligned}$$

Sums of Squares

This analysis yields a decomposition of the variation

$$SS_{total} = SS_{residual} + SS_{between}$$

with

$$SS_{residual} = \sum_{j=1}^{q} \sum_{i=1}^{n_j} (y_{ij} - \bar{y}_j)^2 = \sum_{j=1}^{q} (n_j - 1)s_j^2 \quad \text{and} \quad SS_{between} = \sum_{j=1}^{q} n_j(\bar{y}_j - \bar{\bar{y}})^2.$$

For the rain forest example, we find that

$$SS_{residual} = (12-1)\cdot 5.065^2 + (12-1)\cdot 4.981^2 + (9-1)\cdot 5.761^2 = 820.6234$$

and

$$SS_{between} = 12\cdot(23.750-\bar{\bar{y}})^2 + 12\cdot(14.083-\bar{\bar{y}})^2 + 9\cdot(15.778-\bar{\bar{y}})^2) = 625.1793$$

Sums of Squares

source of variation	degrees of freedom	sums of squares	mean square
between groups	$q-1$	$SS_{between}$	$s^2_{between} = SS_{between}/(q-1)$
residuals	$n-q$	$SS_{residual}$	$s^2_{residual} = SS_{residual}/(n-q)$
total	$n-1$	SS_{total}	

- The $q-1$ degrees of freedom between groups is derived from the q groups minus 1 degree of freedom used to compute $\overline{\overline{y}}$.
- The $n-q$ degrees of freedom within the groups is derived from the $n_j - 1$ degree of freedom used to compute the variances s_j^2.

Sums of Squares

The analysis of variance information for the Borneo rain forest data is summarized in the table below.

source of variation	degrees of freedom	sums of squares	mean square
between groups	2	625.2	312.6
residuals	30	820.6	27.4
total	32	1445.8	

The F Statistic

The test statistic is

$$F = \frac{s^2_{between}}{s^2_{residual}} = \frac{SS_{between}/(q-1)}{SS_{residual}/(n-q)}.$$

- Under the null hypothesis, F is a constant multiple of the ratio of two independent χ^2 random variables, namely $SS_{between}$ and $SS_{residual}$.
- This ratio is called an F random variable with $q-1$ numerator degrees of freedom and $n-q$ denominator degrees of freedom and written $F_{q-1,n-q}$

F Statistic

.

For the rain forest data,

$$F = \frac{s^2_{between}}{s^2_{residual}} = \frac{312.6}{27.4} = 11.43.$$

The critical value for an 0.01 level test is 5.390. So, we reject H_0 stating mean number of trees does not depend on logging history.

```
> 1-pf(11.43,2,30)
[1] 0.0002041322
> qf(0.99,2,30)
[1] 5.390346
```

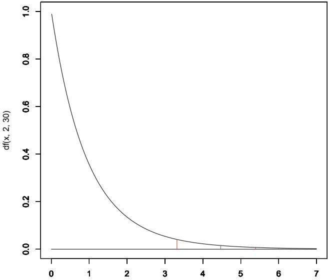


Figure: Upper tail critical values. The density for an $F_{2,30}$ random variable. The indicated values 3.316, 4.470, and 5.390 are critical values for significance levels $\alpha = 0.05$, 0.02, and 0.01, respectively.

Exercise. Use R to determine these critical values.

Confidence Intervals

Confidence intervals are determined using the data from all of the groups as an unbiased estimate $s^2_{residuals} = SS_{residuals}/(n-q)$ for the variance, σ^2. This allows us to increase the degrees of freedom in the t distribution and reduce the margin of error. Thus, the γ-level confidence interval for μ_j is

$$\bar{y}_j \pm t_{(1-\gamma)/2,n-q} s_{residual}/\sqrt{n_j}.$$

The interval for the difference in $\mu_j - \mu_k$ is similar to that for a pooled two-sample t confidence interval,

$$\bar{y}_j - \bar{y}_k \pm t_{(1-\gamma)/2,n-q} s_{residual}\sqrt{\frac{1}{n_j}+\frac{1}{n_k}}.$$

The 95% confidence interval for mean number of trees on a lot logged 1 year ago

$$14.083 \pm 2.042\frac{\sqrt{27.4}}{\sqrt{12}} = 14.083 \pm 4.714 = (9.369, 18.979).$$

Exercise. Give the 95% confidence interval for the difference in trees between plots never logged plots versus logged 8 years ago.

Honey Bee Queen Development Time

- The development time for a European queen in a honey bee hive is suspected to depend on the temperature of the hive.
- To examine this, queens are reared in a low (31.1° C), a medium (32.8° C) and a high temperature hive (34.4° C).
- The hypothesis is that higher temperatures increase metabolism rate and thus reduce the time needed from the time the egg is laid until an adult queen honey bee emerges from the cell.

Figure: Emerging adult honey bee queen

Honey Bee Queen Development Time

The hypothesis is

$$H_0 : \mu_{low} = \mu_{med} = \mu_{high} \quad \text{versus} \quad H_1 : \mu_{low}, \mu_{med}, \mu_{high} \text{ differ}$$

where μ_{low}, μ_{med}, and μ_{high} are, respectively, the mean development time in days for queen eggs reared in a low, a medium, and a high temperature hive.

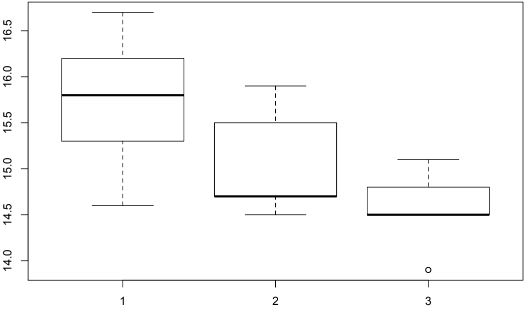
Here are the data and a boxplot:

```
> ehblow<-c(16.2,14.6,15.8,15.8,15.8,15.8,16.2,16.7,15.8,16.7,15.3,14.6,
  15.3,15.8)
> ehbmed<-c(14.5,14.7,15.9,15.5,14.7,14.7,14.7,15.5,14.7,15.2,15.2,15.9,
  14.7,14.7)
> ehbhigh<-c(13.9,15.1,14.8,15.1,14.5,14.5,14.5,14.5,13.9,14.5,14.8,14.8,
  13.9,14.8,14.5,14.5,14.8,14.5,14.8)
> boxplot(ehblow,ehbmed,ehbhigh)
```

Honey Bee Queen Development Time

```
> ehb<-c(ehblow,ehbmed,ehbhigh)
> temp<-c(rep(1,length(ehblow)),
  rep(2,length(ehbmed)),
  rep(3,length(ehbhigh)))
> ftemp<-factor(temp,c(1:3))
> anova(lm(ehb~ftemp))
Analysis of Variance Table

Response: ehb
          Df Sum Sq Mean Sq F value    Pr(>F)
ftemp      2 11.222  5.6111  23.307 1.252e-07 ***
Residuals 44 10.593  0.2407
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```



Contrasts

After completing a one way analysis of variance, resulting, as above, in rejecting the null hypotheses, a typical follow-up procedure is the use of contrasts. Contrasts use as a null hypothesis that some linear combination of the means equals to zero.

To see if the mean queen development time for medium hive temperature is midway between the time for the high and low temperature hives, we have the contrast,

$$H_0 : \frac{1}{2}(\mu_{low} + \mu_{high}) = \mu_{med} \quad \text{versus} \quad H_1 : \frac{1}{2}(\mu_{low} + \mu_{high}) \neq \mu_{med}$$

or

$$H_0 : \frac{1}{2}\mu_{low} - \mu_{med} + \frac{1}{2}\mu_{high} = 0 \quad \text{versus} \quad H_1 : \frac{1}{2}\mu_{low} - \mu_{med} + \frac{1}{2}\mu_{high} \neq 0.$$

Contrasts

Notice that, under the null hypothesis, the mean

$$E\left[\frac{1}{2}\bar{Y}_{low} - \bar{Y}_{med} + \frac{1}{2}\bar{Y}_{high}\right] = \frac{1}{2}\mu_{low} - \mu_{med} + \frac{1}{2}\mu_{high} = 0$$

and the variance

$$\mathrm{Var}\left(\frac{1}{2}\bar{Y}_{low} - \bar{Y}_{med} + \frac{1}{2}\bar{Y}_{high}\right) = \frac{1}{4}\frac{\sigma^2}{n_{low}} + \frac{\sigma^2}{n_{med}} + \frac{1}{4}\frac{\sigma^2}{n_{high}}.$$

This leads to the test statistic

$$t = \frac{\frac{1}{2}\bar{y}_{low} - \bar{y}_{med} + \frac{1}{2}\bar{y}_{high}}{s_{residual}\sqrt{\frac{1}{4n_{low}} + \frac{1}{n_{med}} + \frac{1}{4n_{high}}}} = \frac{\frac{1}{2}15.743 - 15.043 + \frac{1}{2}14.563}{0.4906\sqrt{\frac{1}{4\cdot 14} + \frac{1}{14} + \frac{1}{4\cdot 19}}} = 0.7005.$$

The p-value,

```
> 2*(1-pt(0.7005,44))
[1] 0.487303
```

again, is considerably too high to reject the null hypothesis.

Contrasts

If we want to see if the rain forest has seen a change in logged areas over the past 8 years in the mean number of trees. This can be written as

$$H_0 : \mu_2 = \mu_3 \quad \text{versus} \quad H_1 : \mu_2 \neq \mu_3$$

or

$$H_0 : \mu_2 - \mu_3 = 0 \quad \text{versus} \quad H_1 : \mu_2 - \mu_3 \neq 0$$

Under the null hypothesis, the test statistic has a t-distribution with $n - q = 33 - 3 = 30$ degrees of freedom. Here

$$t = \frac{\bar{y}_2 - \bar{y}_3}{s_{residual}\sqrt{\frac{1}{n_2} + \frac{1}{n_3}}} = \frac{14.083 - 15.778}{5.234\sqrt{\frac{1}{12} + \frac{1}{9}}} = -0.7344,$$

Exercise. Compute the p-value for this two-sided test and comment on the strength of the evidence against the null hypothesis.

Contrasts

Exercise. Under the null hypothesis appropriate for one way analysis of variance, with n_j observations in group $j = 1, \ldots, q$ and $\bar{Y}_j = \sum_{i=1}^{n_j} Y_{ij}/n_j$,

$$E[c_1\bar{Y}_1 + \cdots + c_q\bar{Y}_q] = c_1\mu_1 + \cdots + c_q\mu_q, \quad \mathrm{Var}(c_1\bar{Y}_1 + \cdots + c_q\bar{Y}_q) = \frac{c_1^2\sigma^2}{n_1} + \cdots + \frac{c_q^2\sigma^2}{n_q}.$$

In general, a contrast begins with a linear combination of the means

$$\psi = c_1\mu_1 + \cdots + c_q\mu_q.$$

The hypothesis is

$$H_0 : \psi = 0 \quad \text{versus} \quad H_1 : \psi \neq 0.$$

For sample means, $\bar{y}_1, \ldots, \bar{y}_q$, the test statistic is

$$t = \frac{c_1\bar{y}_1 + \cdots + c_q\bar{y}_q}{s_{residual}\sqrt{\frac{c_1^2}{n_1} + \cdots + \frac{c_q^2}{n_q}}}.$$

which, under the null hypothesis, has a t distribution with $n - q$ degrees of freedom.