7. Analysis of Variance (ANOVA)

7.1

An overview of ANOVA

What’s ANOVA?

- ANOVA refers to statistical models and associated procedures, in which the observed variance is partitioned into components due to different explanatory variables.
- ANOVA was first developed by R. A. Fisher in the 1920s and 1930s. Thus, it is also known as Fisher's analysis of variance, or Fisher's ANOVA.

What does ANOVA do?

- It provides a statistical test concerning if the means of several groups are all equal.
- In its simplest form, ANOVA is equivalent to Student's *t*-test when only two groups are involved.

Types of ANOVA

- **One-way ANOVA** --- involves only a single factor in the experiment.
- **two-way/multiple-way ANOVA** --- two or more factors are relevant.
- **Factorial ANOVA** --- there is replication at each combination of levels in a two way/multi-way ANOVA**.**
- **Mixed-design ANOVA** --- a factorial mixed-design, in which one factor is a between-subjects variable and the other is within-subjects variable.
- **Multivariate analysis of variance** (**MANOVA**) --- more than one dependent variable involved in the analysis.

Basic Assumptions

- **Independence** — cases are independent.
- **Normality** — data are normally distributed in each of the groups.
- **Homogeneity of variances** — variance of data are the same in all the groups (**Homoscedasticity**).
- *The above form the common assumption that the errors are independently, identically, and normally distributed for fixed-effect models.*

LOGIC OF ANOVA (1)

□ The fundamental technique of ANOVA is to partition the total **sum of squares** into components related to the effects involved in the model.

SSY = SSA + SSE

dfY = dfA + dfE

MSA = SSA/dfA; MSE = SSE/dfE

LOGIC OF ANOVA (2)

□ *MSE* is the pooled variance obtained by combining the individual group variance, and thus it provides an estimate of the population variance.

□ *MSA* is also an estimate of in the absence of true group effects, but it includes a term related to differences between group means when there are group effects.

□ Thus, a test for significant difference between the group means can be performed by comparing the two variance estimates, that is, F = MSA/MSE

LOGIC OF ANOVA (3)

- Under the null hypothesis of identical means, the value of the *F* statistic is ideally 1, but it is expected to have some variation around that value.
- Statistically, it is an *F* distribution with (*k*-1, *n*-*k*) degrees of freedom, assuming that all group means are equal.

FOLLOW UP TESTS

- If a statistically significant effect is found in ANOVA, one or more tests of appropriate kinds will follow up, in order to assess which groups are different from which other groups or to test various other focused hypotheses.
- For example, Tukey's test most commonly compare every group mean with every other group mean and typically incorporate some methods to control Type I errors.

11

7.2

One-way ANOVA

The data model

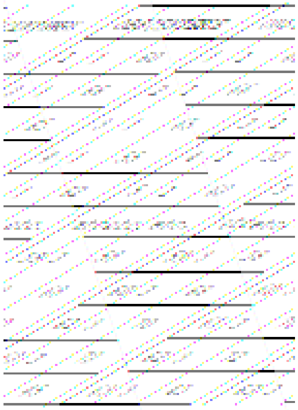
12

$$y_{ij} = \bar{y}_{\cdot\cdot} + (\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot}) + (y_{ij} - \bar{y}_{i\cdot})$$

$$\downarrow$$

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

where $\varepsilon_{ij} \sim N\left(0, \sigma^2\right)$



Decomposition of the total sum of squares

$$\sum_i \sum_j \left(y_{ij} - \bar{y}_{\bullet\bullet}\right)^2 = \sum_i n_i \left(\bar{y}_{i\bullet} - \bar{y}_{\bullet\bullet}\right)^2 + \sum_i \sum_j \left(y_{ij} - \bar{y}_{i\bullet}\right)^2$$

SSY = SSA + SSE

Degrees of freedom

$$n - 1 = (k - 1) + (n - k)$$

$$dfY = dfA + dfE$$

Mean squares and F statistic

$$MSA = \frac{SSA}{dfA} = \frac{\sum_i n_i \left(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot}\right)^2}{k-1}$$

$$MSE = \frac{SSE}{dfE} = \frac{\sum_i \sum_j \left(y_{ij} - \bar{y}_{i\cdot}\right)^2}{n-k}$$

$$F = \frac{MSA}{MSE}$$

Example

- □ The “red cell folate” data, described by Altman (1991, p208)
- □ 22 observations, a numeric variable *folate* and a factor *ventilation*.
- □ Three level of ventilation: “N2Q+O2,24h”, “N2O+O2,op”, and “O2,24h”.
- □ > attach(red.cell.folate)
- □ > str(red.cell.folate)
- □ 'data.frame': 22 obs. of 2 variables:
- □ $ folate : num 243 251 275 291 347 354 380 392 206 210 ...
- □ $ ventilation: Factor w/ 3 levels "N2O+O2,24h","N2O+O2,op",..: 1 1 1 1 1 1 1 1 2 2 ...

ANOVA using *anova* and *lm*

```
> anova(lm(folate~ventilation))
Analysis of Variance Table

Response: folate
            Df Sum Sq Mean Sq F value  Pr(>F)
ventilation  2  15516    7758  3.7113 0.04359 *
Residuals   19  39716    2090
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Interpretation of regression coefficients

- The regression coefficients for a factor variable do not have the usual meaning as the slope of a regression analysis with a numeric explanatory variable.

```
> summary(lm(folate~ventilation))
……
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)         316.62      16.16  19.588 4.65e-14 ***
ventilationN2O+O2,op -60.18      22.22  -2.709   0.0139 *
ventilationO2,24h   -38.62      26.06  -1.482   0.1548
---
```

Multiple test problem

- Consider *k* independent tests, T1, T2, …, Tk, each with a significance probability, say, Pr(Ti) = α .
- The probability that at least one of them comes out significant is Pr(T1+T2+…+Tk) $\le$ Pr(T1) + Pr(T2) + … + Pr(Tk) = $n\alpha$.
- Suppose α =0.05, then the chance of having at least one positive result in 10 test is up to 50%.
- Thus, **the *p*-values tend to be exaggerated.**

Bonferroni correction

- The Bonferroni correction is a method used to address the problem of multiple comparisons by dividing the significance level by the number of tests , or, equivalently, by multiplying the *p*-values by the number of test

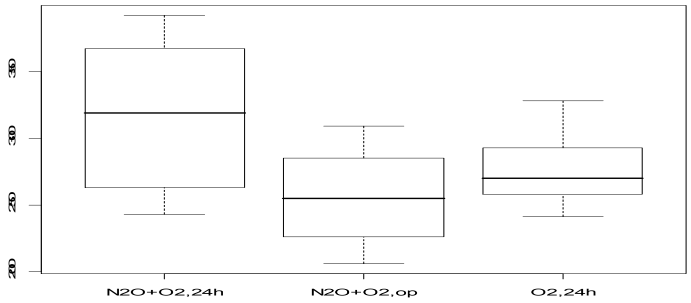
- Let Pr(T1+T2+…+Tk) = α, where α is the significance level for the entire series of tests.
- Let Pr(T1) = Pr(T2) = … = Pr(Tk) = β.
- Then, $\alpha \le k\beta$, or $\beta \le \alpha / k$.

Multiple comparison

- The function *pairwise.t.test* is available to carry out all possible two-group comparisons, and meanwhile making adjustments for multiple comparisons, e.g., via **Bonferroni correction**

```
> pairwise.t.test(folate,ventilation, p.adj="bonferroni")
    Pairwise comparisons using t tests with pooled SD
data:  folate and ventilation
                N2O+O2,24h    N2O+O2,op
N2O+O2,op     0.042              -
O2,24h          0.464              1.000
P value adjustment method: bonferroni
```

Interpretation of results by plots



Testing of homogeneity of variance (1)

```
> bartlett.test(folate~ventilation)

        Bartlett test of homogeneity of variances

data:  folate by ventilation
Bartlett's K-squared = 2.0951, df = 2, p-value = 0.3508

> fligner.test(folate~ventilation)

        Fligner-Killeen test of homogeneity of variances

data:  folate by ventilation
Fligner-Killeen:med chi-squared = 5.5244, df = 2, p-value = 0.06315
```

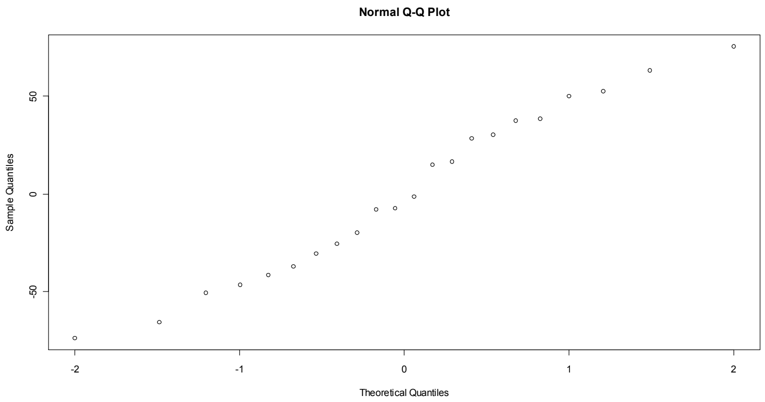
The Levene's test (1)

- Insensitive to non-normality; more appropriate for testing of homogeneity of variance.
  - Compute the absolute values of the residuals from the original linear regression analysis;
  - Fit a linear model by regressing these absolute residuals on the same set of explanatory variables;
  - Significant group effects are indicative of violation of the homoscedasticity assumption.

The Levene's test (2)

```
> g<-lm(folate~ventilation)
> summary(lm(abs(g$res)~ventilation))
……
  Coefficients:
                         Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)              51.625     6.673     7.737   2.74e-07 ***
ventilationN2O+O2,op  -21.353     9.171     -2.328   0.0311 *
ventilationO2,24h      -25.625    10.759    -2.382  0.0278 *
……
```

Diagnostics of normality



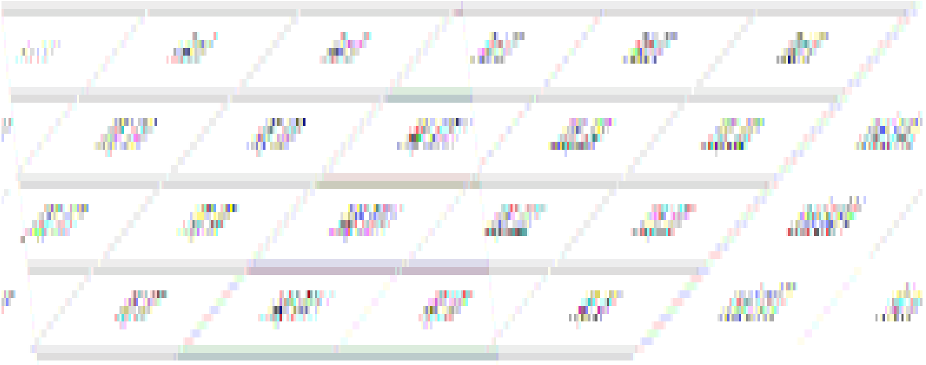
7.3

Two-way ANOVA

The data model

$$y_{ij} = \bar{y}_{\cdot\cdot} + \left(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot}\right) + \left(\bar{y}_{\cdot j} - \bar{y}_{\cdot\cdot}\right) + \left(y_{ij} - \bar{y}_{i\cdot} - \bar{y}_{\cdot j} + \bar{y}_{\cdot\cdot}\right)$$

$$y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}$$



Decomposition of total sum of squares

$$SSY = \sum_i \sum_j \left(y_{ij} - \bar{y}_{\cdot\cdot} \right)^2$$

$$= n\sum_i \left(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot} \right)^2 + m\sum_j \left(\bar{y}_{\cdot j} - \bar{y}_{\cdot\cdot} \right)^2 + \sum_i \sum_j \left(y_{ij} - \bar{y}_{i\cdot} - \bar{y}_{\cdot j} + \bar{y}_{\cdot\cdot} \right)^2$$

$$= SSA + SSB + SSE$$

$$MSA = \frac{SSA}{dfA} = \frac{n\sum_i \left(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot} \right)^2}{m-1} \qquad MSB = \frac{SSB}{dfB} = \frac{m\sum_j \left(\bar{y}_{\cdot j} - \bar{y}_{\cdot\cdot} \right)^2}{n-1}$$

Mean squares & F statistic

$$MSA = \frac{SSA}{dfA} = \frac{n\sum_i \left(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot} \right)^2}{m-1}$$

F = MSA/MSE

$$MSB = \frac{SSB}{dfB} = \frac{m\sum_j \left(\bar{y}_{\cdot j} - \bar{y}_{\cdot\cdot} \right)^2}{n-1}$$

F = MSB/MSE

$$MSE = \frac{SSE}{dfE} = \frac{\sum_i \sum_j \left(y_{ij} - \bar{y}_{i\cdot} - \bar{y}_{\cdot j} + \bar{y}_{\cdot\cdot} \right)^2}{(m-1)(n-1)}$$

Example --- data

```
> heart.rate <- data.frame(
+   hr = c(96,110,89,95,128,100,72,79,100,
+          92,106,86,78,124,98,68,75,106,
+          86,108,85,78,118,100,67,74,104,
+          92,114,83,83,118,94,71,74,102),
+   subj=gl(9,1,36),
+   time=gl(4,9,36,labels=c(0,30,60,120)))

> str(heart.rate)
'data.frame':   36 obs. of  3 variables:
 $ hr  : num  96 110 89 95 128 100 72 79 100 92 ...
 $ subj: Factor w/ 9 levels "1","2","3","4",..: 1 2 3 4 5 6 7 8 9 1 ...
 $ time: Factor w/ 4 levels "0","30","60",..: 1 1 1 1 1 1 1 1 1 2 ...
```

Two-way ANOVA

```
> anova(lm(hr~subj + time))
Analysis of Variance Table

Response: hr
          Df  Sum Sq  Mean Sq   F value    Pr(>F)
subj       8  8966.6   1120.8   90.6391    4.863e-16 ***
time       3  151.0     50.3     4.0696    0.01802 *
Residuals 24  296.8     12.4
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

7.4

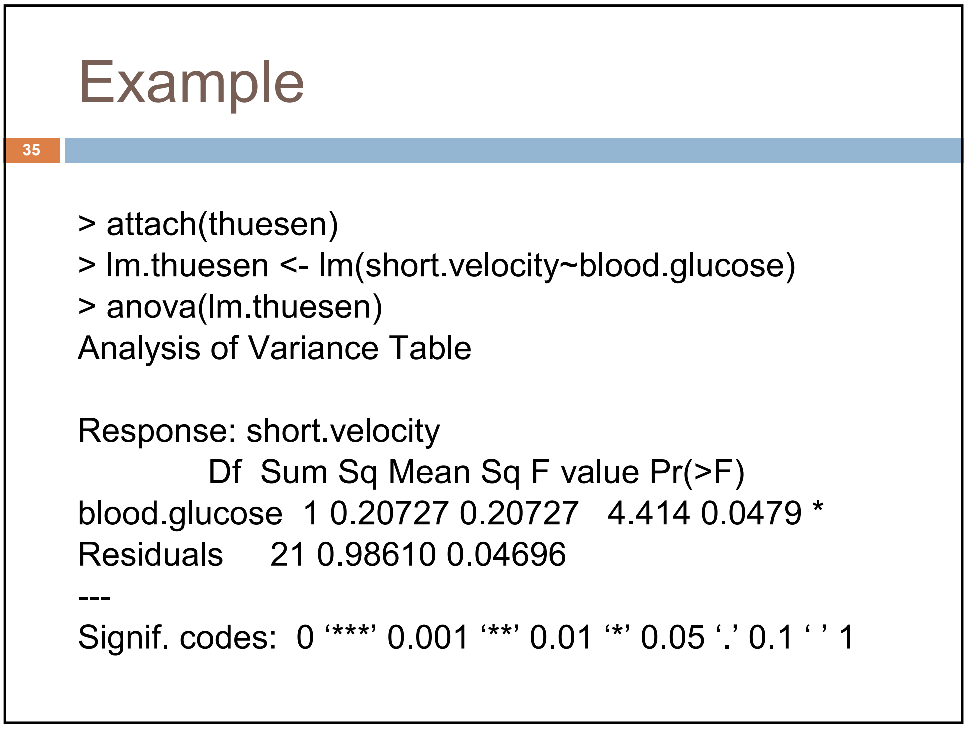
ANOVA in regression analysis

Sum of squares

$$SSY = \sum_i \left(y_i - \overline{y}. \right)^2$$

$$SSM = \sum_i \left(\hat{y}_i - \overline{y}. \right)^2$$

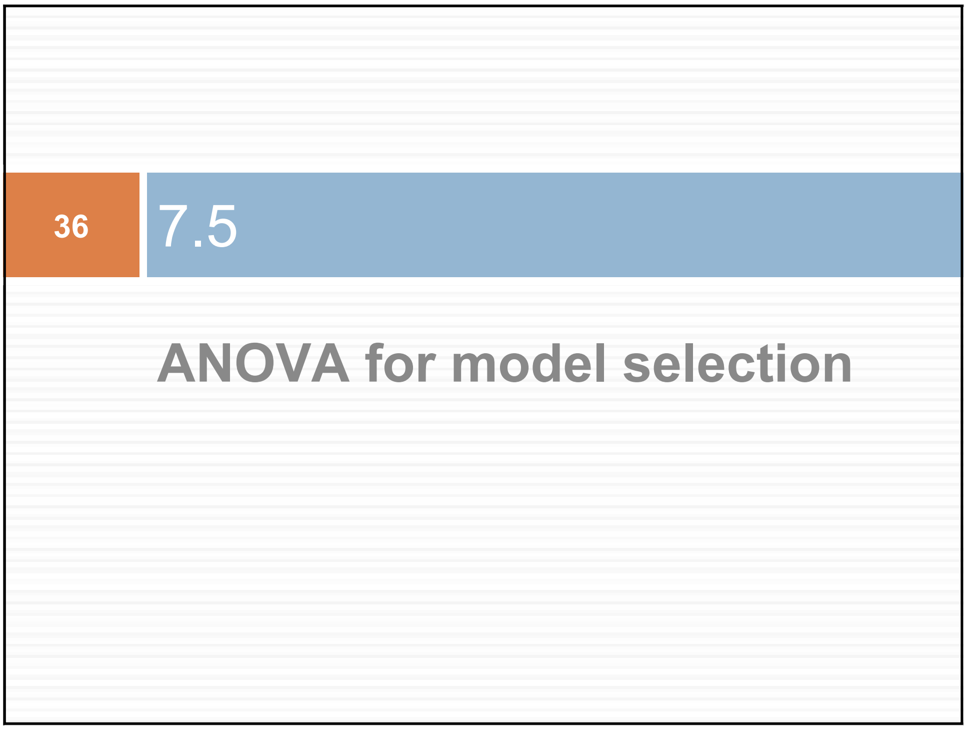
$$SSR = \sum_i \left(y_i - \hat{y}_i \right)^2$$



Example

```
> attach(thuesen)
> lm.thuesen <- lm(short.velocity~blood.glucose)
> anova(lm.thuesen)
Analysis of Variance Table

Response: short.velocity
              Df  Sum Sq Mean Sq F value Pr(>F)
blood.glucose  1 0.20727 0.20727   4.414 0.0479 *
Residuals     21 0.98610 0.04696
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```



7.5

ANOVA for model selection

Models & null hypothesis

- □ Full model:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

- □ Reduced model:

$$\mathbf{y} = \mathbf{1}\mu + \boldsymbol{\varepsilon}$$

- □ Null hypothesis:

$$H_0 : \beta_1 = \ldots = \beta_{k-1} = 0$$

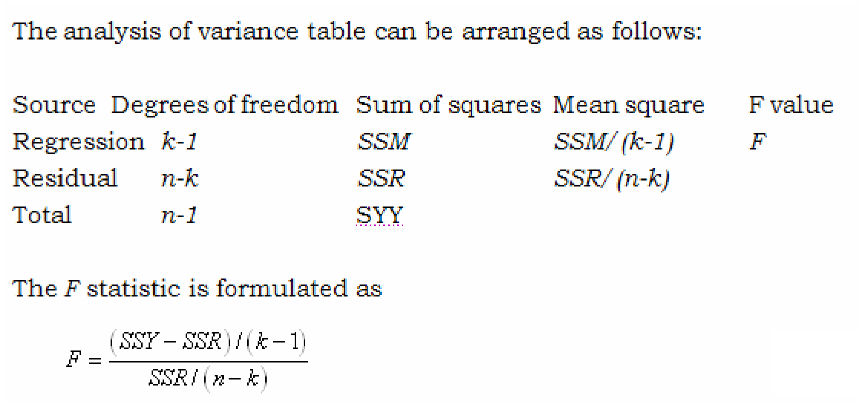
Sum of squares

$$SSY = \left(\mathbf{y} - \overline{\mathbf{y}}\right)'\left(\mathbf{y} - \overline{\mathbf{y}}\right)$$

$$SSR = \hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}} = \left(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right)'\left(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right)$$

SSM = SSY - SSR

ANOVA table



The analysis of variance table can be arranged as follows:

Source	Degrees of freedom	Sum of squares	Mean square	F value
Regression	*k-1*	*SSM*	*SSM/ (k-1)*	*F*
Residual	*n-k*	*SSR*	*SSR/ (n-k)*	
Total	*n-1*	SYY		

The *F* statistic is formulated as

$$F = \frac{(SSY - SSR)/(k-1)}{SSR/(n-k)}$$

Full model vs. reduced model

```
> gfit4<-lm(Species~Elevation+Nearest+Scruz+Adjacent,data=gala)
> y<-as.vector(gala$Species)
> SYY<-sum((y-mean(y))^2)
> SYY
[1] 381081.4
> RSS<-sum(gfit4$res^2)
> RSS
[1] 93469.08
> F<-((SYY-RSS)/4)/(RSS/25)
> F
[1] 19.23178
> 1-pf(F,4,25)
[1] 2.44953e-07
```

Comparing two models

```
> gfit2<-lm(Species~Elevation+Nearest,data=gala)
> anova(gfit4,gfit2)
Analysis of Variance Table

Model 1: Species ~ Elevation + Nearest + Scruz + Adjacent
Model 2: Species ~ Elevation + Nearest
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1     25  93469
2     27 173241 -2    -79771 10.668 0.0004469 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```