

Chapter 5

Hypothesis testing in Linear Models

Julius van der Werf

Analysis of Variance

Analysis of variance or ANOVA are the common way to test whether certain effect are significant in linear models. The method, introduced by R.A. Fischer in the early 20th century, is based on calculating sums of squares which provides a measure of variation. Sums of squares are standardized to 'mean squares' by accounting for the number levels (strictly by accounting for the degrees of freedom). An effect is defined based on a grouping according to the various levels of that effect. Whether such an allocation into grouping is meaningful can be assessed by comparing the group differences with the random differences that occur within groups.

There are 4 basic components of an ANOVA table, based on

- Total sum of squares
- Sum of squares due to the model
- Sum of squares due to the mean
- Residual sums of squares.

As an example, we can take a simple model with one main sire effect (a_i).

$$y_{ij} = \mu + a_i + e_{ij}$$

Assume N observations, with s sires, with $N/s=n$ is the number daughters per sire. Then, the ANOVA table is as follows,

Source	df	Sum of Squares	Mean Squares	EMS
Mean	1	SSM	SSM	
Sires	s-1	SSA	SSA/(s-1)	$n\sigma_s^2 + \sigma_e^2$
Error	N-s	SSE	SSE/(N-s)	σ_e^2
Total	N	SST		

Where

$$SST = y'V^{-1}y = \sum_{i=1}^s \sum_{j=1}^n y_{ij}^2 \quad (\text{the latter holds if } V=I)$$

y_{ij} is an observation on the j^{th} daughter of the i^{th} sire. The total sum of squares (SST) is therefore the sum of each of the observations squared.

$$SSM = y'V^{-1}1(1'V^{-1}1)^{-1}1'V^{-1}y = (\sum y)^2 / N = N * \bar{y}_{..}^2 \quad (\text{assuming } V=I)$$

The mean sum of squares is therefore N times the means squared.

$$SSA = y'V^{-1}X(X'V^{-1}X)^{-1}X'V^{-1}y = n \sum_{i=1}^s (\bar{y}_i - \bar{y}_{..})^2$$

The sum of squares due to a particular effect (e.g. the sire effect) is therefore the sum over all observations of the estimated (sire) effect in each observation squared (in balanced data this is the difference between the progeny group mean of a sire and the overall mean).

The sum of squares due to the total model is (incl. mean)

$$SSR = y'V^{-1}X(X'V^{-1}X)^{-1}X'V^{-1}y = \hat{b}'X'V^{-1}y$$

is simply the solution vector times the right hand side.

$$SSE = \sum_{i=1}^s \sum_{j=1}^n (y_{ij} - \bar{y}_i)^2$$

The sum of squares due to the residual (error) is the sum over all observations of the residual effect in each observation squared (this is the difference between the observation and its group mean).

The coefficient of determination of the model is $R^2 = \frac{SSR - SSM}{SST - SSM}$

This is also the squared correlation between predicted and observed data, or the proportion of total variation (besides the mean) that is explained by the model.

To get a feel for variance-between and variance- within groups, and how the first relates to resemblance (covariance) between observed values within a group, we look at an example. Consider 3 repeated measures on each of five animals. Hence, observations are grouped by animal, and we look at resemblance between repeated measurements.

Five animals are shown, each with 3 measurements.

Example data set 1:

Sheep No.:	1	2	3	4	5
Day 1	21	24	27	20	27
Day 2	22	26	30	19	24
Day 3	20	25	30	18	27
Means:	21	25	29	19	26

Example data set 2:

Sheep No.:	1	2	3	4	5
Day 1	17	21	25	22	24
Day 2	20	29	28	16	22
Day 3	23	28	34	16	32

Means:	20	26	29	18	26
---------------	----	----	----	----	----

By simply looking at the data we can already observe that

- the variation of observed values on the same animal is larger in data set 2.
- the observation on the same animal are more 'alike' in data set 1.
- the variation of the means is slightly larger in data set 2.

This 'gut-feel' about the data can be formally quantified with an analysis of variance.

We will perform this ANOVA based the example data sets. We expect to find

- more random error in data set 2.
- a lower repeatability in data set 2.

More detail about analysis of variance

What do the true effects look like?

To help understand an analysis of variance, the following Tables demonstrate 'knowledge of the underlying parameters' In reality we do not have this knowledge, but the example shows that larger effects means more variance. It also shows how variance components reflect 'similarity' (repeatability) of repeated performances.

Observed phenotypes P for each measure are the sums of permanent (P_p) and temporary (P_t) effects: $P = P_p + P_t$

We can call the temporary effects 'measurement error'. We look again at the data, but now with the underlying effects. The actual measurements (P) are shown in **bold**, the other numbers are the underlying effects, and the means.

Consider again example data set 1, where the measurement error (P_t) is low:

ANIMAL:	1	2	3	4	5
	$P_p + P_t = P$	$P_p + P_t = P$	$P_p + P_t = P$	$P_p + P_t = P$	$P_p + P_t = P$
3 Measurements each	22 -1 21	24 0 24	27 0 27	19 1 20	25 2 27
	22 0 22	24 2 26	27 3 30	19 0 19	25 -1 24
	22 -2 20	24 1 25	27 3 30	19 -1 18	25 2 27
MEANS:	22 -1 21	24 1 25	27 2 29	19 0 19	25 1 26

Now consider example data set 2 where the measurement error (P_t) is high:

ANIMAL:	1	2	3	4	5
	$P_p + P_t = P$	$P_p + P_t = P$	$P_p + P_t = P$	$P_p + P_t = P$	$P_p + P_t = P$
3 Measurements each	22 -5 17	24 -3 21	27 -2 25	19 3 22	25 -1 24
	22 -2 20	24 5 29	27 1 28	19 -3 16	25 -3 22
	22 1 23	24 4 28	27 +7 34	19 -3 16	25 7 32
MEANS:	22 -2 20	24 2 26	27 2 29	19 -1 18	25 1 26

We want to quantify differences between P_p :

variance **Between** groups σ_B^2

and differences between P_t :

variances **Within** groups σ_w^2

Analysis of Variance

We want to get an *estimate* of the variance between groups, as we cannot measure the P_p effects directly (we cannot see P_p values - just phenotypic values P).

The variance of observed group means (i.e. means per animal) (21, 25, 29, 19 and 26 in example data 1) is made up of the variance of mean permanent effects (σ_B^2 : 22, 24, 27, 19, 25) plus one n^{th} of the variance of mean temporary effects (σ_W^2/n , -1, 1, 2, 0, 1). This variance of observed group means is determined by taking the squared differences of the group means (as a deviation from the overall mean), leading to the *sums of squares due to group effects*. We divide these SS groups by the degrees of freedom for groups (equal to the number of comparisons we can make between groups). We expect the *means squares for groups* to contain 3 times (because of 3 values per group) the variance due to groups means, i.e. $3 * (\sigma_B^2 + \frac{\sigma_W^2}{n}) = \sigma_w^2 + 3\sigma_B^2$.

We can estimate the contribution of the variance of temporary effects within groups by taking all deviations within groups (we estimate the group mean and take the deviation of each record from each group mean. These deviations are called *residual effects* and if we square all these within group deviations, we obtain the *residual sums of squares*. If the residual sums of squares are divided by the number of residuals that we can compare (this is the *degrees of freedom* for the residual) than we obtain an estimate of the residual variance: σ_W^2 . For this example:

Analysis of variance Example data set 1.

Effect	Degr. of Free.	Sums of Squares	Mean Squares	Expected Mean Squares
Mean	1	8640		
Group effect (Between groups)	4	192	48	$\sigma_w^2 + 3\sigma_B^2$
Residual (Within groups)	10	18	1.8	σ_w^2
Total	15	8850		

Here is how these figures are calculated ...

- 1) sums of squares due to means: $15 * 24^2 = 8640$
- 2) sums of squares due to group differences:
 $3 * (21^2 + 25^2 + 29^2 + 19^2 + 26^2) = 8832$
 corrected for mean:
 $3 * ((21-24)^2 + (25-24)^2 + (29-24)^2 + (19-24)^2 + (26-24)^2) = 192$
 or directly: $8832 - 8640 = 192$
- 3) total sums of squares
 $21^2 + 22^2 + \dots$ <all individual weightings squared> $\dots + 26^2 = 8850$
- 4) residual sums of squares
 total SS - SS groups = $8850 - 8832 = 18$
 notice that also: $(-1)^2 + 0 + (-2)^2 + \dots + (2)^2 = 18$

The estimated variance components for example data set 1:

Between groups	$\sigma_B^2 = 15.4$
Within groups	$\sigma_w^2 = 1.8$
Total variance is	$\sigma_B^2 + \sigma_w^2 = 17.2$
Repeatability = intra-class correlation =	$15.4/17.2 = 0.895$

5: Hypothesis testing

Variance of the group means	$s_B^2 + \frac{s_W^2}{n} = 15.4 + 1.8/3 = 16$
-----------------------------	---

Analysis of Variance example data set 2:

Effect	Degr.of Free.	Sums of Squares	Mean Squares	Expected Mean Squares
Mean	1	8640		
Group effect (Between groups)	4	254.4	63.6	$\sigma_w^2 + 3\sigma_B^2$
Residual (Within groups)	10	178	17.8	σ_w^2
Total	15	8929		

<u>The estimated variance components for example data set 2:</u>	
Between groups	$\sigma_B^2 = 15.3$
Within groups	$\sigma_w^2 = 17.8$
Total variance is	$\sigma_B^2 + \sigma_w^2 = 33.1$
Repeatability = intra-class correlation =	15.3/33.1 = 0.46
Variance of the group means	$s_B^2 + \frac{s_W^2}{n} = 15.3 + 17.8/3 = 21.2$

Comparing the two data sets:

The values of Pt are much larger in data set 2: we have larger measurement errors → The repeatability is lower

The group means are nearly the same (essentially, we have the same animals), they are only changed due to more variation in measurement error. The variance of the group means is a bit higher in data set 2.

Summary of the example

- | |
|--|
| <p>It is not critical to be able to do all these sums, they serve more as an illustration. What is important is to get the concept:</p> <ul style="list-style-type: none"> ▪ The extent of differences between the groups and differences within the groups can be quantified. We call these <i>variance components</i>. ▪ The variance components provide information about “how much alike” different observations within a group are. If differences between groups are large in relation to the differences we observe within groups, than observations within the same groups are very much ‘alike’. If the variance between groups is large, the observations within the group have more <i>covariance</i>. ▪ Since the covariance among related animals is due to genetic components, the between group (full-sibs or half sibs) variance component can be used to determine genetic variance. |
|--|

Hypothesis testing

Requirements: Assume that y has a multivariate distribution. Hypothesis testing requires knowing the distributions of sums of squares. A sum of squares, say $y'Qy$, will have a chi-squared distribution if QV is idempotent (i.e. this matrix times itself is equal to itself) and if y is MVN.

The most relevant sums of squares are:

$$SS_{\text{Total}} = y'Q_T y \quad \text{where } Q_T = V^{-1}$$

$$\text{and } SS_{\text{model}} = y'Q_R y \quad \text{where } Q_R = V^{-1}X(X'V^{-1}X)^{-1}X'V^{-1}$$

$$\text{and } SS_{\text{Residual}} = y'Q_E \quad \text{where } Q_E = V^{-1} - V^{-1}X(X'V^{-1}X)^{-1}X'V^{-1}$$

It can be proven that SSR and SSE are independent chi-square variables ($Q_R V Q_E = 0$).

Testing the model:

The ratio of two independent central chi-square variables has an F-distribution. The adequacy of the whole model is tested as:

$$F_M = \frac{SSR / r(X)}{SSE / (N - r(X))}$$

where $r(X)$ is the rank of X and N is the total number of observations.

The whole model is usually significant as it contains the mean (which is usually significantly different from zero). It is more useful to test subsets of the parameter vector b . Various functions of b can be tested. An hypothesis test consists of

1. The null hypothesis
2. the alternative hypothesis
3. a test statistic
4. a probability level or rejection region

The null hypothesis can be written as

$$H'b = c \quad \text{or:} \quad H'b - c = 0$$

Where:

H must be of full column rank

$H'b$ must be an estimable function

If these conditions are met, $H'b$ is testable. The test statistic is

$$F = \frac{s / r(H')}{SSE / (N - r(X))}$$

where $s = (H'\beta - c)'(H'CH)^{-1}(H'\beta - c)$

and $C = (X'V^{-1}X)^{-1}$

Example data set 5 (repeat)

Calf ID	Age of Dam (yr)	Breed	Growth Rate (kg/day)
1	2	AN	2.10
2	3	AN	2.15
3	4	AN	2.20
4	5+	HE	2.35
5	5+	HE	2.33
6	2	HE	2.22
7	3	HE	2.25
8	3	HE	2.27
9	4	SM	2.50
10	5+	SM	2.60
11	2	SM	2.40
12	2	SM	2.45

The model

$$y_{ijk} = b_0 + b_i + b_j + e_{ijk}$$

where

y_{ijk} is an observation on the growth rate of calves

b_0 is the overall mean

b_i is an effect due to the age of dam of the calf ($i = 1, \dots, 4$)

b_j is an effect due to the breed of the calf ($j=1, \dots, 3$)

e_{ijk} is the residual for each observation

The model written in matrix notation is

$$y = Xb + e$$

LS equations: $(X'X) = X'y$; with a possible solution vector is $\hat{b} = (X'X)^{-1}X'y$;

$$\begin{bmatrix} 12 & 4 & 3 & 2 & 3 & 3 & 5 & 4 \\ 4 & 4 & 0 & 0 & 0 & 1 & 1 & 2 \\ 3 & 0 & 3 & 0 & 0 & 1 & 2 & 0 \\ 2 & 0 & 0 & 2 & 0 & 1 & 0 & 1 \\ 3 & 0 & 0 & 0 & 3 & 0 & 2 & 1 \\ 3 & 1 & 1 & 1 & 0 & 3 & 0 & 0 \\ 5 & 1 & 2 & 0 & 2 & 0 & 5 & 0 \\ 4 & 2 & 0 & 1 & 1 & 0 & 0 & 4 \end{bmatrix} \begin{bmatrix} b_0 \\ b_{11} \\ b_{12} \\ b_{13} \\ b_{14} \\ b_{21} \\ b_{22} \\ b_{23} \end{bmatrix} = \begin{bmatrix} 27.82 \\ 9.17 \\ 6.67 \\ 4.70 \\ 7.28 \\ 6.45 \\ 11.42 \\ 9.95 \end{bmatrix} \rightarrow \begin{bmatrix} \hat{b}_0 \\ \hat{b}_{11} \\ \hat{b}_{12} \\ \hat{b}_{13} \\ \hat{b}_{14} \\ \hat{b}_{21} \\ \hat{b}_{22} \\ \hat{b}_{23} \end{bmatrix} = \begin{bmatrix} 0.000 \\ 0.000 \\ 0.052 \\ 0.082 \\ 0.147 \\ 2.105 \\ 2.204 \\ 2.430 \end{bmatrix}$$

Properties of Solutions

While there are an infinite number of different solution vectors to the GLS equations the sum of squares due to the model is unique.

$$SSR = \mathbf{b}'\mathbf{X}'\mathbf{y} = \mathbf{y}'\mathbf{X}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = 64.737$$

Therefore when it comes to testing the model, the values for mean squares and F-tests are likewise independent of the solution vector.

Analysis of Variance

For the example data, the basic analysis of variance table is

Source	d.f	Sums of Squares	Means Square	F - stat
Total (SST)	N-1 = 12	$\mathbf{y}'\mathbf{y} = 64.740$	5.3950	10295.8
Mean (SSM)	1	$N\bar{y}^2 = 64.496$	64.4960	123084.0**
Model (SSR)	r = 6	$\mathbf{b}'\mathbf{X}'\mathbf{y} = 64.7371$	10.7895	20590.6**
Residual (SSE)	N - r = 6	$\mathbf{y}'\mathbf{y} - \mathbf{b}'\mathbf{X}'\mathbf{y} = 0.0031$	0.0005	

The estimate of the error variance is given by the residual means square (MSE). The test for the adequacy of the model is given by the calculation of an F-statistic for the model:

$$F_M = MSR / MSE$$

which at $P < 0.05$, is highly significant for this example. A significant F_M indicates that the solution vector is not a null vector and that the model explains some of the major sources of variation. This is generally significant because the solution vector includes the mean of the observations, which is usually different from zero. Alternatively the multiple correlation coefficient, R^2 can be used to determine the amount of variation accounted for by the model where

$$\begin{aligned} R^2 &= (SSR - SSM) / (SST - SSM) \\ &= 0.987 \end{aligned}$$

The higher the value for R^2 the better the model, which in this case is a very adequate description of the variation in calf growth rates.

Hypothesis Tests

As mentioned above the test of a models appropriateness is usually significant because the solution vector for the model includes the mean of observations. Therefore when testing the importance of a model, it is of greater worth to test the significance of the elements of b other than the mean.

The General Linear Hypothesis Procedure

Using the generalised linear hypothesis procedure, the test of SSR is directly partitioned into sub-hypotheses which test the various estimable functions of b . For the current example the two tests of interest would be

- i) age of dam effects
- ii) breed of calf effects

For the general linear hypothesis, the null hypothesis for testing age of dam effects would be $H_1'b = 0$ where

$$H_1' = \begin{bmatrix} 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & -1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & -1 & 0 & 0 & 0 \end{bmatrix}$$

and

$$H_1'b = \begin{bmatrix} b_{11} - b_{12} \\ b_{11} - b_{13} \\ b_{11} - b_{14} \end{bmatrix}$$

For $H_1'b$ to be testable

- i) H_1' must have full row rank
- ii) $H_1'b$ must be an estimable function (determined using methods described previously)

and since these conditions hold, $H_1'b$ is testable.

The sum of squares for $H_1'b$ is

$$s_1 = (H_1'b)'(H_1'(X'X)^{-1}H_1)^{-1}H_1'b$$

and given

$$H_1'b = \begin{bmatrix} -0.052 \\ -0.082 \\ -0.147 \end{bmatrix}$$

then

$$s_1 = 0.0357$$

with 3 degrees of freedom, ie. $n_1 - 1$. The F-test is

$$F_1 = \frac{s_1 / r(H_1')}{\sigma_e^2} = 22.7$$

which at the 0.05 level means that the differences among age of dam groups is significantly different from zero and accounts for some of the variation explained by the model.

Similarly to test breed of calf differences

$$H_2' = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & -1 \end{bmatrix},$$

$$H_2'b = \begin{bmatrix} -0.099 \\ -0.325 \end{bmatrix}$$

and the sums of squares for breed effects = 0.174 with 2 degrees of freedom. In summary of the partitions of SSR for calf growth data

Source	d.f.	Sum of Squares	Means Square	F-test
Model	6	64.737	10.7895	20590.6
Mean	1	64.496	64.4960	123084.0
Age of Dam	3	0.0357	0.0119	22.7**
Breed of Calf	2	0.174	0.0870	166.0**

which shows that both elements of the model are significantly different from zero and both explain some of the variation in calf growth rates.

Reduction Notation

Another means of testing the appropriateness of a model and its elements is to examine the significance of the reduction in sums of squares after regressing y on each element of the model separately. In the current example, the full model could be written in component form as

$$y = \mathbf{m}l + X_1 b_i + X_2 b_j + e$$

with the different sub-models being

- Model 1: $y = \mathbf{m}l + X_2 b_j + e$
- Model 2: $y = \mathbf{m}l + X_1 b_i + e$
- Model 3: $y = \mathbf{m}l + e$

The notation for the reductions of these models are:

- $R(\mathbf{m}, b_i, b_j)$ = the sums of squares due to fitting the full model
- $R(\mathbf{m}, b_j)$ = the reduction sums of squares due to fitting model 1
- $R(\mathbf{m}, b_i)$ = the reduction sums of squares due to fitting model 2
- $R(\mathbf{m})$ = the reduction sums of squares due to fitting the mean

and for each of these models reductions in sums of squares for each of these models would be obtained by constructing OLS equations and then solving to give

$$\begin{aligned} R(\mathbf{m}, b_i, b_j) &= 64.737 \\ R(\mathbf{m}, b_j) &= 64.701 \\ R(\mathbf{m}, b_i) &= 64.563 \end{aligned}$$

$$R(\mathbf{m}) = 64.496$$

To test the null hypothesis $\mathbf{b} = 0$, ie. that the differences in age of dam effects are not significantly different from zero

$$\begin{aligned} s_1 &= R(\mathbf{m}, b_i, b_j) - R(\mathbf{m}, b_j) \\ &= R(b_i | \mathbf{m}, b_j) \\ &= 0.0357 \end{aligned}$$

with 3 degrees of freedom. Similarly for the differences in breed of calf

$$\begin{aligned} s_2 &= R(\mathbf{m}, b_i, b_j) - R(\mathbf{m}, b_i) \\ &= R(b_j | \mathbf{m}, b_i) \\ &= 0.174 \end{aligned}$$

with 2 degrees of freedom. Finally to test the significance of the model after correcting for the mean gives

$$\begin{aligned} s_3 &= R(\mathbf{m}, b_i, b_j) - R(\mathbf{m}) \\ &= R(b_i, b_j | \mathbf{m}) \\ &= 0.241 \end{aligned}$$

While these results are the same as those obtained from the general linear hypothesis procedure, the method requires that individual analyses of variance be performed for each effect. In this way, the former technique is a much simpler and easier method in that the sums of squares for fixed effects and their significance in the model can be determined with greater efficiency.

Exercises for linear model testing

1. A simple ANOVA

To get some feel for why it is useful to calculate sums of square in the construction and testing of statistical models for prediction, consider the following example.

Suppose we have 4 observations and a one-way classification with 2 levels (A and B). Calculate the sum of squares for the total, the mean, the model and the residual. Residual 1 refers to a model where only the mean is fitted and residual 2 to a model where also the class effect is fitted.

Calculate sums of squares 'by hand' based on the numbers the column.

	class	Observation	Mean	Residual 1	Predicted Y	Residual 2
	A	8				
	A	9				
	B	11				
	B	12				
Sum of Squares						

2. One way classification

Given are data on three pig breeds. Estimate the breed effect

	Yorkshire	Landrace	Pietrain
	800	600	600
	700	700	
	600		

Test whether the breed effect is significant

Test whether the Yorkshire breed is significantly different from Landrace.

3. Two way classification

Given are data of daily growth of cattle in an experiment, where 3 feeding levels were tested. Given are observations on some bulls.

		Feeding level		
		Pasture Queensland	Feedlot	Pasture Victoria
Breed	Angus	200, 165	450, 460	300, 350
	Hereford	220	426, 390, 430	310, 320, 330
	Brahman	260, 240, 235	380, 450	280
	St. Gertrude	245, 220, 250, 240	420, 440,	300

In a one-way analysis:

Analyze the effect of breed. Test the general effect

5: Hypothesis testing

Analyze the effect of feeding regime. Test general effect.

Repeat the previous analysis in a 2-way classification.