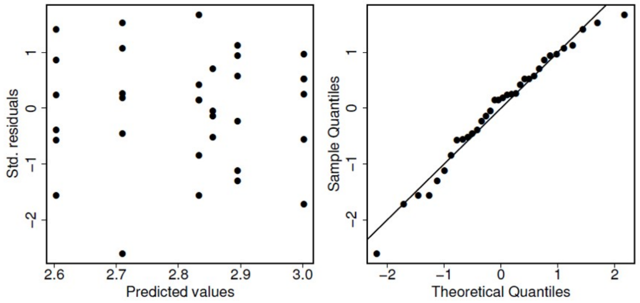
One-way ANOVA model. In the one-way ANOVA setup with k groups, the group means $\alpha_1, \dots, \alpha_k$ are parameters, and we write the one-way ANOVA model

$$y_i = \alpha_{g(i)} + e_i, \quad i = 1, \dots, n,$$

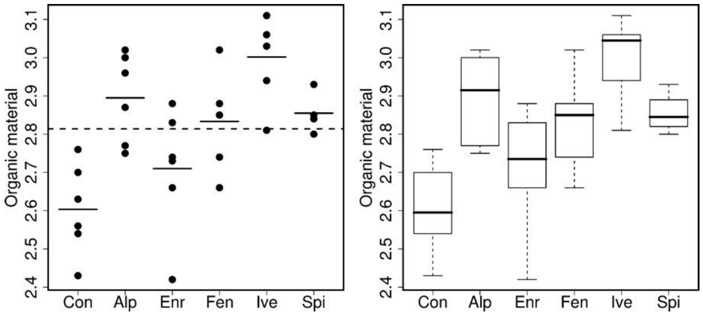
where $g(i) = x_i$ is a “categorical” explanatory variable that corresponds to a “numerical” variable y_i. The remainder terms $e_1, \dots, e_n$ are independent and $N(0, \sigma^2)$-distributed. In other words, it is assumed that there is a normal distribution for each group, with **means that are different from group to group** and given by the α’s but with the **same standard deviation in all groups** (namely, σ) representing the **within-group variation**. The parameters of the model are $\alpha_1, \dots, \alpha_k$ and σ, where α_j is the expected value (or the population average) in the j-th group.

Example: Dung decomposition. An experiment with dung from heifers was carried out in order to explore the influence of antibiotics on the decomposition of dung organic material. As part of the experiment, $n = 34$ heifers were divided into six groups. No antibiotics were added for heifers in the first group (the control group with $j = 1$), and antibiotics of different types (alpha-Cypermethrin, Enrofloxacin, Fenbendazole, Ivermectin, Spiramycin) were added to the feed for heifers in the remaining five groups labeled $j = 2, \dots, 6$. For each heifer, a bag of dung was dug into the soil, and after eight weeks the amount of organic material was measured for each bag.

The residual analysis is illustrated below: Residual plot (left) and QQ-plot (right) of the standardized residuals. The straight line has intercept zero and slope one. There are only six possible predicted values (one for each group), and the variation of the standardized residuals seems to be roughly the same for all groups.



Strip chart of the antibiotics data is obtained in the left below: Data points with group sample means (solid line segments) and the total mean of all observations (dashed line) in the left. Parallel boxplots are presented in the right below.



The observations together with group means (solid lines) and the total mean (dashed line) are shown in the strip chart. The amount of organic material appears to be lower for the control group compared to any of the five types of antibiotics, suggesting that decomposition is generally inhibited by antibiotics. However, there is variation from group to group (between-group variation) as well as a relatively large variation within each group (within-group variation). The within-group variation seems to be roughly the same for all types, except perhaps for spiramycin, but that is hard to evaluate because there are fewer observations in that group.

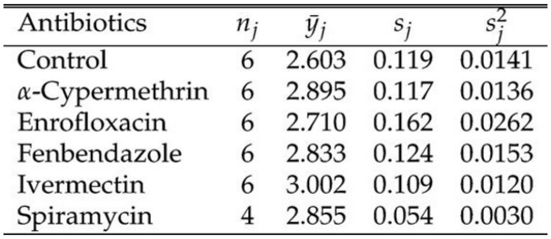
Group means and SD's. Consider the situation with n observations split into k groups. Label the groups $j = 1$ through $j = k$. Let $g(i)$ denote the group for observation i. Then $g(i)$ has one of the values $1, \ldots, k$. The sample mean $\bar{y}_j$ and sample standard deviation s_j in j-th group are given by

$$\bar{y}_j = \frac{\sum_{i:g(i)=j} y_i}{n_j}$$

$$s_j = \sqrt{\frac{\sum_{i:g(i)=j} (y_i - \bar{y}_j)^2}{n_j - 1}}$$

where n_j is the size of j-th group and the summation is over all observations i's in j-th group. The sample mean $\bar{y}_j$ becomes the estimate $\hat{\alpha}_j$.

Example: Dung decomposition. The sample means and the sample standard deviations are computed for each group separately. We find the same indications as we did in the boxplots. On average the amount of organic material is lower for the control group than for the antibiotics groups, and except for the spiramycin group the standard deviations are roughly the same in all groups.



Antibiotics	n_j	$\bar{y}_j$	s_j	s_j^2
Control	6	2.603	0.119	0.0141
α-Cypermethrin	6	2.895	0.117	0.0136
Enrofloxacin	6	2.710	0.162	0.0262
Fenbendazole	6	2.833	0.124	0.0153
Ivermectin	6	3.002	0.109	0.0120
Spiramycin	4	2.855	0.054	0.0030

Standard error in one-way ANOVA. In the ANOVA setup the residual variance s^2 is given by

$$s^2 = \frac{\sum_{i=1}^{n} (y_i - \bar{y}_{g(i)})^2}{n - k}$$

which we call the **pooled variance** estimate. In the one-way ANOVA case we are very often interested in the **differences** or **contrasts** between group levels rather than the levels themselves. Hence, we are interested in quantities $\alpha_j - \alpha_l$ for two groups j and l. Then the estimate is simply the difference between the two estimates, and the corresponding standard error is given by

$$\mathrm{SE}(\hat{\alpha}_j - \hat{\alpha}_l) = s\sqrt{\frac{1}{n_j} + \frac{1}{n_l}}$$

The formulas above are particularly useful for two samples ($k = 2$).

Consider the one-way ANOVA model

$$y_i = \alpha_{g(i)} + e_i, \quad i = 1, \dots, n,$$

where $g(i)$ denotes the group corresponding to the i-th observation and $e_1, \dots, e_n$ are independent and $N(0, \sigma^2)$-distributed. Then the estimates $\hat{\alpha}_1, \dots, \hat{\alpha}_k$ for the group mean are simply the group averages $\bar{y}_1, \dots, \bar{y}_k$, and the corresponding standard errors are given by

$$\mathrm{SE}(\hat{\alpha}_j) = s\sqrt{\frac{1}{n_j}}$$

It suggests that mean parameters for groups with many observations (large n_j) are estimated with greater precision than mean parameters with few observations.

Between-group variation. **Between-group variation** refers to differences between the groups, and it is calculated by

$$\mathrm{SS}_{grp} = \sum_{j=1}^{k} n_j (\bar{y}_j - \bar{y})^2; \quad \mathrm{MS}_{grp} = \frac{\mathrm{SS}_{grp}}{k-1}$$

As illustrated in the strip chart for the antibiotics example, variation between the different treatments is represented as deviation between the group means $\bar{y}_j$ (horizontal line segments) and the overall mean (dashed line):

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$$

When we examine the between-group variation, the k group means essentially act as our observations. The mean squared difference MS_{grp} per group becomes the between-group variation.

Hypothesis test. Consider the comparison of groups

$$y_i = \alpha_{g(i)} + e_i, \quad i = 1, \dots, n,$$

where $g(i)$ is the group that observation i belongs to and $e_1, \dots, e_n$ are residuals. As usual, k denotes the number of groups. In a typical linear model, it tests the null hypothesis that $\alpha_j = 0$. However, in this study we are interested in whether there is no difference between the groups. Thus, the null hypothesis is given by

$$H_0 : \alpha_1 = \cdots = \alpha_k$$

and the alternative hypothesis is the opposite; namely, that at least two α's are different.

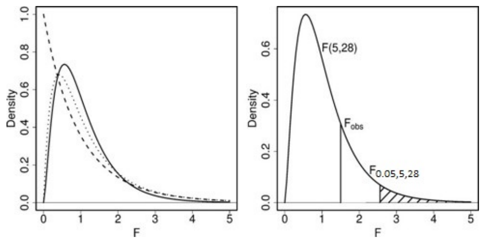
$$H_A : \alpha_j \neq \alpha_l \text{ for some pair } \{j, l\}.$$

F-test. If there is no difference between any of the groups (H_0 is true), then the group averages will be of similar size and be similar to the total mean . Hence, MS_{grp} will be small. On the other hand, if groups 1 and 2, say, are different (H_0 is false), then the group means will be somewhat different and cannot be similar; thus, MS_{grp} will be large. "Small" and "large" should be measured relative to the within-group variation. We use the test statistic

$$F_{obs} = \frac{\text{MS}_{grp}}{\text{MS}_e}$$

where MS_{grp} is standardized with the residual variance $\text{MS}_e = s^2$. Note that large values of F_{obs} are critical; that is, not in agreement with the null hypothesis.

If the null hypothesis is true, then F_{obs} comes from a so-called F-distribution with $(k-1, n-k)$ degrees of freedom. Notice that there is a pair of degrees of freedom (not just a single value) and that the relevant degrees of freedom are the same as those used for computation of MS_{grp} and MS_e. The density for the F distribution is shown for three different pairs of degrees of freedom in the left panel below.



Mechanism of rejection. This disagreement is equivalent to F_{obs} being larger, and the corresponding p-value are often inserted in an analysis of variance table. The p-value of being smaller than 0.05 indicates significance evidence toward the disagreement between groups. Since only large values of F_{obs} are critical, we calculate the p-value by

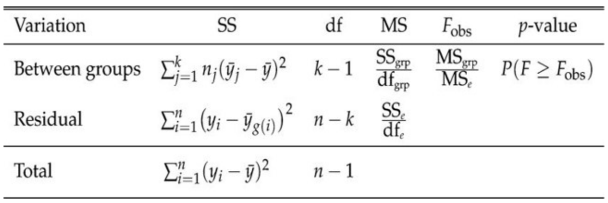
$$P(F \geq F_{obs})$$

where F follows the F-distribution with $(k-1, n-k)$ degrees of freedom. The hypothesis is rejected if the p-value is 0.05 or smaller (if $\alpha = 0.05$ is the significance level). In particular, H_0 is rejected on the 5% significance level if $F_{obs} \geq F_{0.05,k-1,n-k}$.

Analysis of variance. Within-group variation refers to the variation in each of the groups:

$$\text{SS}_e = \sum_{i=1}^{n} (y_i - \bar{y}_{g(i)})^2; \quad \text{MS}_e = \frac{\text{SS}_e}{n-k} = s^2$$

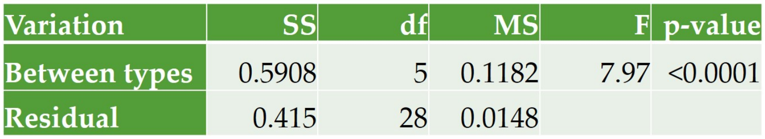
which we call the **mean square error**. A large between-group variation is an indication of differences between the groups, but if the within-group variation is also large, then the differences may be due to random variation. It is the distinction between different sources of variation that has given analysis of variance (ANOVA) its name.



Variation	SS	df	MS	F_{obs}	p-value
Between groups	$\sum_{j=1}^{k} n_j(\bar{y}_j - \bar{y})^2$	$k-1$	$\frac{\text{SS}_{\text{grp}}}{\text{df}_{\text{grp}}}$	$\frac{\text{MS}_{\text{grp}}}{\text{MS}_e}$	$P(F \geq F_{\text{obs}})$
Residual	$\sum_{i=1}^{n} \left(y_i - \bar{y}_{g(i)}\right)^2$	$n-k$	$\frac{\text{SS}_e}{\text{df}_e}$		
Total	$\sum_{i=1}^{n} (y_i - \bar{y})^2$	$n-1$			

ANOVA is associated with the estimation of "variation" within and between groups in order to analyze the differences among means. Ronald Fisher introduced the term variance and proposed its formal analysis in his article in 1918. It provides a procedure of F-test to find whether two or more population means are equal (null hypothesis), and therefore generalizes the t-test beyond two means. Under the null hypothesis the test statistic F has an F-distribution which was named after Fisher.

Example: Dung decomposition. The values are listed in an ANOVA table as follows:



Variation	SS	df	MS	F	p-value
Between types	0.5908	5	0.1182	7.97	<0.0001
Residual	0.415	28	0.0148		

The F value of 7.97 is very extreme, corresponding to the very small p-value less than 0.0001. Thus, we reject the hypothesis and conclude that there is strong evidence of group differences. Subsequently, we need to quantify the conclusion further. Which groups are different and how large are the differences?

Pairwise comparisons. Sometimes interest is in particular groups from the experiment, and we want to compare j-th group and l-th group, say. Still, the analysis is carried out using all data since this makes the estimate of the standard deviation more precise. In a sense we borrow information from all observations when we estimate the standard deviation, even though we use only the data from the two groups in question to estimate the mean difference. The null hypothesis is

$$H_0 : \alpha_j = \alpha_l \text{ (or, equivalently } H_0 : \alpha_j - \alpha_l = 0)$$

and we consider the two-sided alternative hypothesis

$$H_A : \alpha_j \neq \alpha_l \text{ (or, equivalently } H_A : \alpha_j - \alpha_l \neq 0)$$

In the ANOVA setup the residual variance s^2 is given by

$$s^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y}_{g(i)})^2}{n-k}$$

which we call the **pooled variance** estimate. In the one-way ANOVA case we are very often interested in the **differences** or **contrasts** between group levels. Hence, we are interested in quantities $\alpha_j - \alpha_l$ for two groups j and l. Then the estimate is simply the difference between the two estimates, and the corresponding standard error is given by

$$\text{SE}(\hat{\alpha}_j - \hat{\alpha}_l) = s\sqrt{\frac{1}{n_j} + \frac{1}{n_l}}$$

The formulas above are particularly useful for two samples ($k = 2$).

The difference is significant on the significance level α if and only if

$$\frac{|\hat{\alpha}_j - \hat{\alpha}_l|}{\text{SE}(\hat{\alpha}_j - \hat{\alpha}_l)} \geq t_{\alpha/2,n-k}$$

Here the critical value is obtained from to the t-distribution with $n-k$ degree of freedom. The value

$$(t_{\alpha/2,n-k})\mathrm{SE}(\hat{\alpha}_j - \hat{\alpha}_l) \tag{9.1}$$

is called the **margin of error** for $(1-\alpha)\%$ confidence interval of difference between j-th group and l-th group. Hence, we can see significant difference if the magnitude $|\hat{\alpha}_j - \hat{\alpha}_l|$ of difference is larger than the margin (9.1) of error. This should not be done uncritically, though, due to the multiple testing problem.

Example: Dung decomposition. Recall that the residual standard deviation is obtained by

$$s = \sqrt{0.01482} = 0.1217$$

The margin of error for 95% confidence interval of difference between Control and Spiramycin is given by

$$(2.048)(0.1217)\sqrt{\frac{1}{6}+\frac{1}{4}} = 0.161$$

whereas the margin of error for all other comparisons becomes

$$(2.048)(0.1217)\sqrt{\frac{1}{6}+\frac{1}{6}} = 0.144$$

For the Spiramycin group, we find that

$$\hat{\alpha}_{spiramycin} - \hat{\alpha}_{control} = 0.252 > 0.161,$$

so the group is significantly different from the control group. On the other hand, there is no significant difference between the enrofloxacin group and the control group since

$$\hat{\alpha}_{enroflox} - \hat{\alpha}_{control} = 0.107 < 0.144.$$

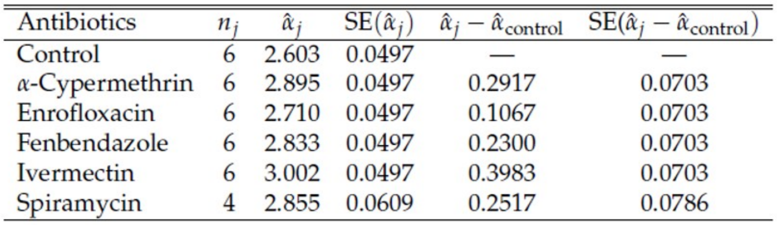
Using the same arguments for the remaining three antibiotic types, we conclude that the amount of organic material is significantly lower for the control groups than for all other groups, except the enrofloxacin group.

R coding: Dung decomposition. In the one-way ANOVA analysis, we want to test the hypothesis of an overall effect. The test is reported by the summary from the `aov()` function:

```
Model <- aov(org ~ type)
summary(Model)
```

The output contains one line per source of variation, and for each source it lists the degrees of freedom, the SS-value, and the MS-value. Moreover, the F-test for the effect of type is carried out: the value of F-test and the associated p-value are reported. Notice how the degrees of freedom for the test, here $(5, 28)$, also appear in the output.

We use the antibiotics data for illustration of multiple comparison, and reproduce the following table.

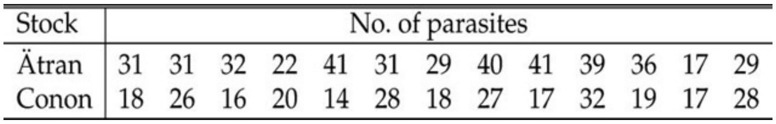


Antibiotics	n_j	$\hat{\alpha}_j$	SE$(\hat{\alpha}_j)$	$\hat{\alpha}_j - \hat{\alpha}_{\text{control}}$	SE$(\hat{\alpha}_j - \hat{\alpha}_{\text{control}})$
Control	6	2.603	0.0497	—	—
α-Cypermethrin	6	2.895	0.0497	0.2917	0.0703
Enrofloxacin	6	2.710	0.0497	0.1067	0.0703
Fenbendazole	6	2.833	0.0497	0.2300	0.0703
Ivermectin	6	3.002	0.0497	0.3983	0.0703
Spiramycin	4	2.855	0.0609	0.2517	0.0786

Since the variable type contains text values, R automatically uses it as a factor. The `lm()` and `summary()` calls are used as follows:

```
summary(lm(org ~ type - 1))
summary(lm(org ~ type))
```

Example: Parasite counts for salmons. An experiment with two difference salmon stocks, from River Conon in Scotland and from River Atran in Sweden, was carried out as follows. Thirteen fish from each stock were infected and after four weeks the number of a certain type of parasites was counted.

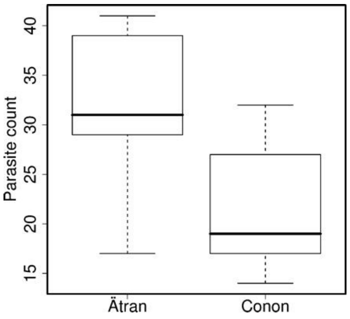


Stock	No. of parasites												
Ätran	31	31	32	22	41	31	29	40	41	39	36	17	29
Conon	18	26	16	20	14	28	18	27	17	32	19	17	28

The statistical model for the salmon data is given by

$$y_i = \alpha_{g(i)} + e_i, \quad i = 1, \ldots, 26$$

where $g(i)$ is either 1="Atran" or 2="Conon" and $e_1, \ldots, e_{26}$ are from $N(0, \sigma^2)$. In other words, Atran observations are from $N(\alpha_1, \sigma^2)$, and Conon observations are from $N(\alpha_2, \sigma^2)$.



Parallel boxplots for the two samples are shown above, and the sample mean and sample standard deviations are obtained for each group separately.

$$\bar{y}_1 = 32.23, \qquad s_1 = 7.28$$
$$\bar{y}_2 = 21.54, \qquad s_2 = 5.81.$$

We can compute the pooled variance estimate by

$$s^2 = \frac{(12)s_1^2 + (12)s_2^2}{24} = 43.40$$

The difference in parasite counts is estimated by

$$\hat{\alpha}_1 - \hat{\alpha}_2 = 32.23 - 21.54 = 10.69$$

with a standard error of

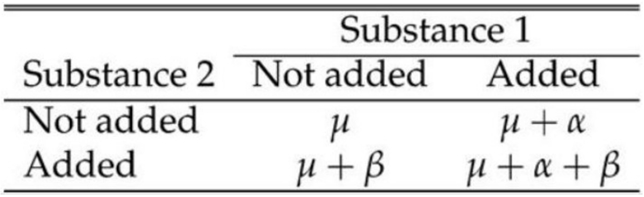
$$\mathrm{SE}(\hat{\alpha}_1 - \hat{\alpha}_2) = s\sqrt{\frac{1}{13} + \frac{1}{13}} = 2.58$$

The 95% confidence interval for the difference is given by

$$10.69 \pm (t_{0.025,24})(2.58) = (5.36, 16.02)$$

We see that the data is not in accordance with a difference of zero between the stock means. Thus, the data suggests that Atran salmons are more susceptible than Conon salmons to parasites.

Additive two-way analysis of variance. We can think of a feeding experiment to examine the average weight of some animal and we have a reference feeding strategy and two substances we can add to the food. We denote the average weight for the reference group by μ and the average increase in weight for substances 1 and 2 by α and β, respectively. Then we can summarize the average values for the groups by different feeding strategies.



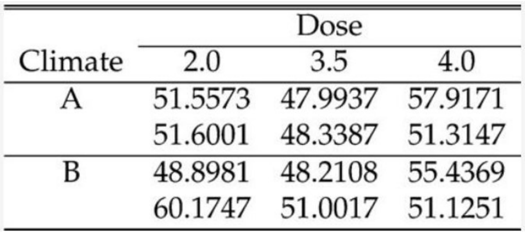
	Substance 1	
Substance 2	Not added	Added
Not added	μ	$\mu + \alpha$
Added	$\mu + \beta$	$\mu + \alpha + \beta$

The two-way analysis of variance (also called the additive two-way analysis of variance model) uses two categorical explanatory variables. Let $g(i)$ and $h(i)$ denote the functions that define the groups of the two categorical variables for i-th observation, and consider the two-way additive model

$$y_i = \alpha_{g(i)} + \beta_{h(i)} + e_i, \quad i = 1, \ldots, n$$

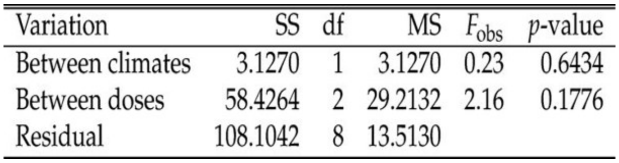
The model extends the one-way analysis of variance in the same way that the simple linear regression model extends to multiple linear regression.

Example: Cucumber disease. This study examines how the spread of a disease in cucumbers depends on climate and amount of fertilizer. Two different climates were used: (A) change to day temperature 3 hours before sunrise and (B) normal change to day temperature. Fertilizer was applied in 3 different doses: 2.0, 3.5, and 4.0 units. The amount of infection on standardized plants was recorded after a number of days, and two plants were examined for each combination of climate and dose.



	Dose		
Climate	2.0	3.5	4.0
A	51.5573	47.9937	57.9171
	51.6001	48.3387	51.3147
B	48.8981	48.2108	55.4369
	60.1747	51.0017	51.1251

Here we have 2 categorical variables: climate (with 2 possible categories) and dose (with 3 possible categories). One way to think about the design of a two-way analysis of variance model is that we can place each of our observations in exactly one cell of the two-way table. Hypothesis tests are analogous to their one-way analysis of variance counterparts. Null hypotheses state that there is no difference among the levels of the first and second explanatory variables, respectively.



Variation	SS	df	MS	F_{obs}	p-value
Between climates	3.1270	1	3.1270	0.23	0.6434
Between doses	58.4264	2	29.2132	2.16	0.1776
Residual	108.1042	8	13.5130		

The two hypotheses of interest for the cucumber data are:

$$H_0 : \alpha_A = \alpha_B$$

and

$$H_0 : \beta_{2.0} = \beta_{3.5} = \beta_{4.0}$$

The alternative hypotheses are that at least two α's are unequal or that at least two of the β's are different, respectively. A consequence of the two-way additive model is that the contrast between any two levels for one of the explanatory variables is the same for every category.

Example: Pork color over time. The investigators seek to examine if there is a systematic change in the brightness from a tristimulus color measurement. The color was measured from a pork chop from each of ten pigs at days 1, 4, and 6 after storage. We write the model as

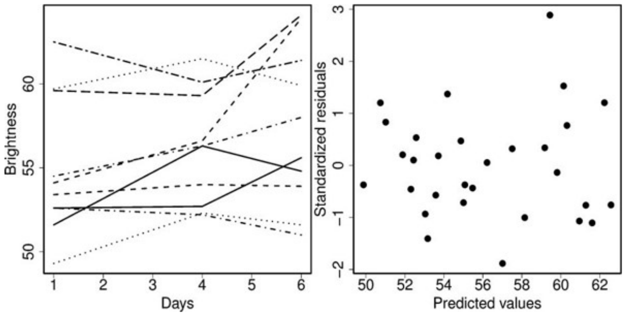
$$y_i = \alpha_{g(i)} + \beta_{h(i)} + e_i, \quad i = 1, \ldots, n$$

The function $g(i)$ defines which pig (levels $1, \ldots, 10$) each of the 30 observations corresponds to, while $h(i)$ is the similar function for days. We include pig as an explanatory variable because we suspect that meat brightness might depend on which specific pig the pork was cut from. There could be an effect of pig, and we seek to account for that by including pig in the model even though we are not particularly interested in being able to compare any pair of pigs like, say, pig 2 and pig 7.

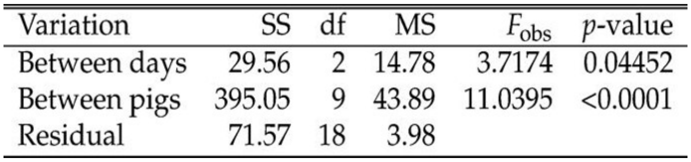
The primary hypothesis of interest is

$$H_0 : \beta_1 = \beta_4 = \beta_6$$

which corresponds to no change in brightness over time. Excluding the explanatory variable pig from the analysis might blur the effects of day since the variation among pigs may be much larger than the variation between days. This is called a block experiment with pigs as “blocks.” Sometimes observational units are grouped in such blocks and the observational units within a block are expected to be more similar than observations from different blocks. We expect observations taken on the same pig to be potentially more similar than observations taken on different pigs.

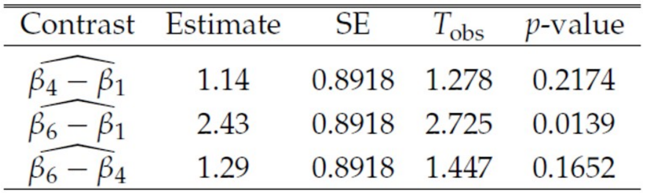


Left panel shows interaction plot of the change in meat brightness for 10 pigs measured at days 1, 4, and 6 after storage. Right panel shows the residual plot for the two-way analysis of variance of the pork data.



Variation	SS	df	MS	F_{obs}	p-value
Between days	29.56	2	14.78	3.7174	0.04452
Between pigs	395.05	9	43.89	11.0395	<0.0001
Residual	71.57	18	3.98		

We can conclude that there is a borderline significant effect of days. The test for pigs can also be seen in the analysis of variance table, and while this may be of little interest for the manufacturers producing pork, since that is nothing they can control, we can still see that it is highly significant.



Contrast	Estimate	SE	T_{obs}	p-value
$\widehat{\beta_4 - \beta_1}$	1.14	0.8918	1.278	0.2174
$\widehat{\beta_6 - \beta_1}$	2.43	0.8918	2.725	0.0139
$\widehat{\beta_6 - \beta_4}$	1.29	0.8918	1.447	0.1652

The p-value from ANOVA table only tells us that not all days have the same level. The contrasts give us more information about the different days. From the contrasts we see that the difference in days primarily stems from a difference between days 1 and 6. We can also see that the brightness scores decrease as time increases since the contrasts are all positive. The average difference in brightness between days 1 and 6 is 2.43.