Week 5 Notes Part 1 ANOVA

Code ▾

Author: Brendan Gongol

Last update: 07 January, 2023

1 Overview

In this Note one-way Analysis of Variance is introduced and relevant formalism and statistical distributions are discussed. We also discuss the meaning of ANOVA/F-test as applied to linear models. We conclude by looking at an example of ANOVA analysis in R using categorical linear model fit.

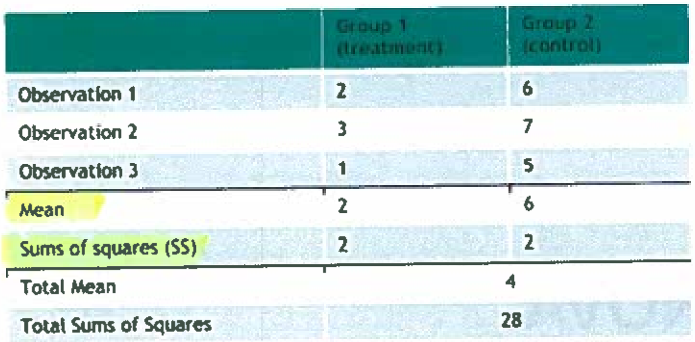
2 Analysis of Variance

When we were characterizing our linear model fits, we were using anova() function in order to assess the significance. However, we used it as a black box so far and were simply looking at the returned p-value. Analysis of variance (or ANOVA) is an important concept, and it is time to look at it in detail.

The "classical" set-up for what is known as "one-way ANOVA" is very similar to the t-test. Namely, we have n groups of measurements (i.e. samples), $n \geq 2$, with multiple measurements (e.g. blood pressures in a number of patients) in each group (for instance, healthy controls, disease stage I, disease stage II etc.). The null hypothesis we test is that the underlying means and variances of the respective distributions all the samples were drawn from are the same: $_1=...=_n $. Thus for $n = 2$, the null hypothesis is the same both for ANOVA and for the two-sample t-test with equal variance. As we will see in a second, ANOVA uses mathematical formalism that is very different from the t-test, but the reassuring news is that in the $n = 2$ case these two tests are equivalent (they will result in the same p-value – however always remember that we are considering the equal variance flavor of the t-test here!). With larger number of samples, $n > 2$, ANOVA is often a preferable way of testing. The reason is that regardless of the actual number of samples, n, ANOVA is still just one test applied to all samples at once. In contrast, if we were to use t-test, we could only run multiple pairwise comparisons between the samples when $n > 2$. As a result, while the latter procedure can be applied in practice in order to get better understanding of the data, the results of such tests are subject to multiple testing errors (when performing many tests, we are bound to eventually run into a case which is significant merely by chance, we have observed that in the homework simulations!). Moreover, all pair-wise comparisons within the group of the same n samples are not independent, so it can be hard to interpret the results in a rigorous way. On the flip-side, a single ANOVA test on all n samples may tell us that the null does not explain the data well (i.e. it is unlikely that all samples came from distributions with the same means and variances), but it cannot tell us which particular sample(s) violate the assumption (so you may need to investigate further).

Let us first consider a simple example. We study the number of the headache episodes per month in a group of patients treated with some anti-migraine drug (Group 1) and in an untreated group (Group 2). Let us further assume that there are just three people (i.e. three observations) in each group, which we label accordingly as Observation 1, ..., Observation 3 (note that these are just labels; in our example, observation 1 in group 1 and observation 1 in group 2 are different patients of course). Suppose the collected data are distributed as shown in the table below (the values in the table show the numbers of headache episodes during the given month in a particular patient):

@ $n = 2$ ANOVA + t-test result in same p-value
@ $n > 2$ - use ANOVA, if t-test is used results may have ↑ # testingerrors.



	Group 1 (treatment)	Group 2 (control)
Observation 1	2	6
Observation 2	3	7
Observation 3	1	5
Mean	2	6
Sums of squares (SS)	2	2
Total Mean	4	
Total Sums of Squares	28	

total # of people/observations = 6

In other words, we observed 2, 3, and 1 migraine episodes in the treated group, and 6, 7, and 5 episodes in the untreated group. One familiar way of assessing the difference between the two samples would be of course to run a two-sample t-test on the samples at hand, `c(2,3,1)` and `c(6,7,5)`. But let us do something different this time. Namely, in each group separately, we compute the mean μ_i and within-group sum of squares

sum $(observation_i - \mu_i)^2$

(1)

$$SS_i = \sum_{Groupi} (observation - \mu_i)^2$$

The results are shown in the second block in the table, right below the observations. We can also calculate separately the total mean μ_total and sum of squares of all six observations (across all groups and with respect to the total mean):

(2)

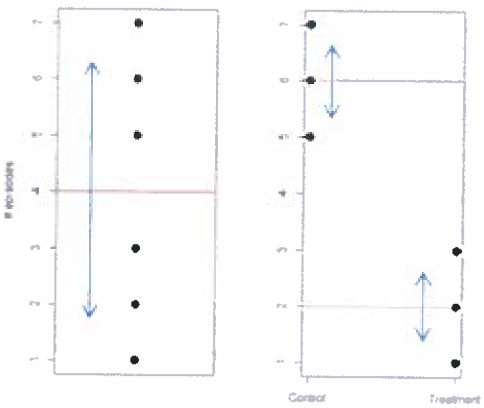
$$SS_total = \sum_{allgroups} (observation - \mu_total)^2$$

The results are shown in the last block of the table.

Let us now assume that the null hypothesis is correct, i.e. that the two samples have the same mean (and variance), in other words there is no difference between the treatment and the control groups. Let us defer the rigorous derivation until following sections and examine the problem at a more intuitive, semi-qualitative level here. Namely, we are going to forget for a second that our sample means are, of course, different from the underlying population mean(s). Let's assume for now that they are the same as the population means (or, in practice, close enough so that we do not care). In this case, under the null the (equal) means of the two samples are also equal to the total mean of all the observations, $\mu_1 = \mu_2 = \mu_{total}$. But if all the means are the same, then, as it is easy to see from the equations above, the total sum of squares across all observations (2) is equal to the sum over the individual groups, of the sums of squares (1) within each individual groups (samples), $SS_{total} = SS_1 + SS_2$. Indeed, when all the means are the same, it does not matter in what order we sum up the squared deviations of each observation from that (same) mean. Keep in mind, this is what should happen if the null hypothesis holds. In our example, however, $SS_1 + SS_2 = 4$, while $SS_{total} = 28$. This makes us think that null hypothesis might not be true (we will get to quantifying the significance of the discrepancy soon). The sum of squares within each group SS_1, SS_2 (which are just the variances, without the $1/N$ factors) can be interpreted as (remaining) noise: after we stratified patients into the groups, this is just the variability within study groups. In contrast, the difference between the groups $SS_{total} = -(SS_1 + SS_2)$ ($28 - 4 = 24$ in our example) is referred to as "between-group" variability of the "effect".

← if it were we'd expect $SS_{total} = 4$

The rationale is the following: consider all the data shown in the above table as a single sample, for all six subjects (see the figure below, left panel). We can characterize the sample with descriptive statistics such as mean (shown with red line) and variance (the latter is SS_{total}, up to a factor). Note that variance, when computed for all the observations against the total mean is quite large (shown by arrows). Then it occurs to us that there might be an association with treatment status (treated/untreated) hidden in the data. So we split the dataset into the groups based on this (categorical) variable. If there were no association with treatment status (equal means in the two groups), then $SS_1 + SS_2$ would be the same as SS_{total} we originally obtained. However, in the presence of an association (i.e. there is an effect, different groups have different means depending on the value of the treatment status), the sum $SS_1 + SS_2$ (remaining within-group spreads, shown by arrows in the right panel) would be smaller than the total variation in the whole dataset, SS_{total}, as we have seen in our example. In other words, the treatment status variable explains away large chunk of the total variability: between-the sample variability is a deterministic effect that we can ascribe to the treatment, within-the sample variability is the only case-to-case randomness that remains. Does this remind you "explained" variance vs remaining noise (residuals) we discussed in connection with linear models? It should, because it is the same thing. In fact, we do have a sort of linear model here, except that the explanatory variable X in this case is discrete (Treatment/Control). → the linear models we used from the last two weeks were continuous.



eg: j = c(walks, runs, no exercize)
p = 3
x_{ij} : individual data points within each group.
i : index of person/observation in that group
n_j : number of people in a group.

2.1 Between- and Within-sample Variance

This qualitative picture is most definitely something you should always keep in mind, but let's get more formal and follow up with a rigorous derivation.

Let us assume that we have our observations split into groups $j = 1...p$ (thus total of p groups), and j-th group contains nj observations (just like with the t-test, we can have different numbers of observations in each group, of course). We will denote each individual observation as x_{ij}, indicating its group membership j and with the index $\underline{i} = 1...n_j$ running within each group. Let us further define the total mean of all observations,

people/ samples, etc.

N = total # observations across all groups.
$\bar{x}$ = overall mean of all obs. in all groups.
x_{ij} the i-th obsv. in the j-th group.
p : total # groups
n_j : # observations in j-th group.

(3)

$$\frac{\text{Sum of all obsv}}{\text{total \# obsv.}}$$

sum of all individual obsv. across all groups.

sum # obsv in each group.

$$\bar{x} = (\sum_{j=1}^{p}\sum_{i=1}^{n_i} x_{ij})/N \ \ where \ \ N = \sum_{j=1}^{p} n_j$$

as well as the respective sample means within each group j:

(4)

$$\frac{\text{sum of obvs in one group}}{\text{\# obsev in same group.}}$$

sum of obvs in ~~j~~ group

$$\bar{x}_j = (\sum_{i=1}^{n_j} x_{ij})/n_j$$

← # obvs in group

For the total sum of squares we have (note that we are using a trick where we add and then subtract back the same value in each group):

(5)

$2 \times \boxed{\sum\sum(x_{ij} - \bar{x}_j)} \times \sum\sum(\bar{x}_j - \bar{x})$

+ − same value

error, missing ^2

$$SS_{total} = \sum_{j=1}^{p}\sum_{i=1}^{n_i}(x_{ij} - \bar{x})^2 = \sum_{j=1}^{p}\sum_{i=1}^{n_i}(x_{ij} - \bar{x}_j + \bar{x}_j - \bar{x})^2 = \sum_{j=1}^{p}\sum_{i=1}^{n_i}[(x_{ij} - \bar{x}_j)^2 + 2(x_{ij} - \bar{x}_j)(\bar{x}_j - \bar{x}) + (\bar{x}_j - \bar{x})^2]$$

expand

The second term in the last sum is equal to zero. Indeed, using (4) we obtain for that term:

$$\sum_{j=1}^{p}\sum_{i=1}^{n_i}(\overset{x}{X}_{ij} - \bar{x}_j) = \sum_j[(\bar{x}_j - \bar{x}) \cdot \boxed{\sum_i(x_{ij} - \bar{x}_j)}] = \sum_j[(\bar{x}_j - \bar{x})(\sum_i x_{ij} - \sum_i \bar{x}_j)] =$$

represents the sum of deviations from each obsv from its group mean.
↓
by definition, the sum of deviations from the mean within a group = 0.
↓
this is why we use SS, b/c sqr makes all values positive + cant cancel e/o out in a group.

$$\sum_j[(\bar{x}_j - \bar{x})(n_j\bar{x}_j - n_j\bar{x}_j)] = \sum_j[(\bar{x}_j - \bar{x}) \cdot 0] = 0$$

As the result, Eq. (5) can be equivalently rewritten as

(6)

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

where the within-group sum of squares (computed for each data point from the mean of its respective group) is given by

(7)

each obsv within a group

− group mean

$$SS_{within} = \sum_{j=1}^{p}\sum_{i=1}^{n_j}(x_{ij} - \bar{x}_j)^2 = \boxed{\sum_{j=1}^{p} SS_{within}^{(j)}}$$

(where we denote the sum of squared distances within j-th group from the mean of that particular group as SS_{within}^{(j)}), and the between-group sum of squares is given by

(8)

each obvs. within

overall mean of all obsv across all groups.

$$SS_{between} = \sum_{j=1}^{p}\sum_{i=1}^{n_j}(\underset{x_{ij}}{\cancel{\bar{x}_j}} - \bar{x})^2 = \boxed{\sum_{j=1}^{p} n_j(\bar{x}_j - \bar{x})^2}$$

← can also be written as $\sum_{j=1} n_j(\bar{x} - \bar{x}_j)^2$ swapped ← b/c $(x-y)^2 = (y-x)^2$

Note that if means of all the groups are equal (and thus equal to the total mean as well), then the between-the group variability term (8) is equal to 0, and $SS_{total} = SS_{within}$. This reproduces our qualitative picture from the last section. In a more realistic situation, even when underlying means are the same, the sample means will still fluctuate due to random sampling effects, but the differences between the (sample) group means and the total mean, and thus $SS_{between}$ are still expected to be small.

We are now in position to build a quantitative statistics employed by one-way ANOVA. This statistics is the ratio of mean sum of squares between the groups and mean sum of squares within the groups:

(9)

$$F = \frac{MSS_{between}}{MSS_{within}}$$

The only non-trivial part remaining is how to define the means most appropriately. Certainly it can be rigorously derived, but the full derivation is beyond the scope of our course. Instead, we will present some arguments (that actually go along the general line of the derivation) that are "mnemonic" enough to help you remember the correct answer.

Remember that the sums of squares in Eq. (1),(2) or (7),(8) are sample estimators and that the means used in those equations are estimates themselves, derived from the same sample. Let us invoke first an analogy with sample variance (standard deviation is simply a square root of variance, as you certainly remember). As we have discussed and observed in the homework, in a sample of size N, the unbiased estimator of the variance, i.e. the correct "average variance", is a sum of squared differences between observed values and the sample mean, divided by N-1. There are some hand-waving arguments in favor of this fact.

Note that the sum of square differences in the whole sample (as in S_{total}) always underestimates the sum of square differences from the true underlying mean. Indeed, the sample mean, by definition, is always located right "in the middle" of a sample, while the true underlying mean will be "off to one side" (or probably more accurate way of putting it is that it is the sample that is always a little bit "off to one side" from the true mean: the probability of the sample mean to be exactly equal to the true mean is zero for continuous distributions). Because the sample mean is "over-adjusted" to the given sample, the sum of squared differences of sample values from the sample mean will be smaller than if we used the true mean, so using smaller denominator will work in the right direction in order to combat that underestimation. Note also that the sample is already measured and its mean is fixed. for that sample. We are computing the variance around that mean, in other words given that mean. But if we make the mean of N variables fixed, only N-1 of them are going to be truly independent: the value of the Nth variable is fully determined by first N-1 values and by the given mean of the sample. This is why we use N-1 as the number of "degrees of freedom" (and this is why sample variance, properly scaled, follows chi-square distribution χ^2_{N-1}, not χ^2_N).

This explains why we need the correct number of degrees of freedom K: if we properly rescale sum of squares, it will follow standardized and well-studied chi-square distribution; this is the same kind of mathematical convenience as provided by properly rescaling sample mean so that the resulting T-statistics follows a standardized t-distribution. In other words: chi-square with N degrees of freedom is a probability distribution that describes sum of squares on N truly independent standardized normal variables (with mean=0 and sd=1). When we work with sample variance, all N members of the sum are

this means if sample: n=3, $\bar{x}=10$. If we freely choose 2 numbers, say 8 and 12 - it has to be a specific value to ensure $\bar{x}$ remains = to 10. ∴ the 3rd number isn't independent anymore.

← $\frac{8+12+x}{3}=10$ ← $x=10$. ← to keep the $\bar{x}=10$, the 3rd number must = 10,

That's why, when calculating sample variance we want only N-1 independent variables, b/c the last is determined by the mean and other numbers.

not completely independent: a non-trivial dependency is introduced by the fact that the mean itself is calculated from the same sample. In principle, this dependence could lead to a completely different distribution. Fortunately, it can be shown that sample variance still follows chi-square, but with reduced number of degrees of freedom.

Let us now return to the statistics (8) we are building for ANOVA. The sum of squares $SS_{between}$ (Eq. 8) includes p random variables $\bar{x}_j$, which are respective means in different, non-overlapping groups, so they are clearly unrelated and independent, and also normally distributed. Eq. (8) also includes the overall sample mean $\bar{x}$, and since it is computed from all observations and thus introduces a dependency between them, it takes away one degree of freedom, just like (and for the same reason as) in sample variance estimator. Hence, $SS_{between}$ should follow χ^2_{p-1} (with $p-1$ degrees of freedom). The expression for within-group variance (7) involves sample means, and since each of them is over-adjusted to the corresponding sample, we need to fix them (all p of them!), hence out of the N variables x_{ij} only $N-p$ remain truly independent (where N is the total number of measurements in all groups, Eq. (3)). The within-the group variance thus follows χ^2_{N-p}. The numbers of independent degrees of freedom are also the denominators we need to use to compute the mean sums of squares (just like we use $n-1$ in variance estimator). Hence, we arrive to:

t-statistic = $\frac{\text{Normal dist}}{\text{chi-square dist}}$.

F-statistic : $\frac{\text{chi-square dist}}{\text{chi-square dist}}$

(10)

$$MSS_{between} = \frac{SS_{between}}{p-1}$$

$$MSS_{within} = \frac{SS_{within}}{N-\cancel{1}p}$$

With these definitions, the ratio F formally defined in the equation (9) becomes a ratio of two variables each following, under the null hypothesis, a chi-square distribution (with $p-1$ and $N-p$ degrees of freedom, respectively). This ratio is known as F-statistics, and of course it is a random variable itself (as it depends on the randomly drawn sample). The ratio of two random variables, each characterized by a chi-square distribution, follows a known (although not too simple) distribution, referred to as F-distribution. This distribution (probability density) for degrees of freedom d_1, d_2 is given by (you do not need to remember it, just appreciate it):

$$f(x; d_1, d_2) = \frac{\sqrt{\frac{(d_1 x)^{d_1} d_2^{d_2}}{(d_1 x + d_2)^{d_1 + d_2}}}}{x B(\frac{d_1}{2}, \frac{d_2}{2})} = \frac{1}{B(\frac{d_1}{2}, \frac{d_2}{2})} (\frac{d_1}{d_2})^{(\frac{d_1}{2})} x^{\frac{d_1}{2}-1} (1 + \frac{d_1}{d_2} x)^{-\frac{d_1+d_2}{2}}$$

Now that we have the standardized F-distribution, the ANOVA test is straightforward: for p samples that we have at hand, we calculate between-the-group and within-the-group variances, following the prescriptions given above. From those, we calculate the value of the F-statistic, F_0. The F-distribution allows us to calculate the probability to observe value $F > F_0$ by random chance if the null hypothesis (equality of the means and variances of all the samples) is true. But the probability defined in this way is nothing but a p-value. If it is below our significance threshold (say, 0.05), then we reject the null hypothesis and conclude that the means or variances are in reality not the same according to the ANOVA test.

Note that according to Eq. (9) (and our intuition as well), the value of F-statistics increases (and corresponding p-value decreases) when between-sample variability increases (i.e. when variability "explained away" by the segregation of our measurements into the samples increases) or when the within-sample variability (remaining "unexplained" variability, or "noise") decreases. The F statistic compares variance b/w groups to variance w/in groups.

↑ F-value = ↓ p-value = group means are significantly different.

2.2 F-statistics of a Linear Model

Above we have considered a case of two or more samples and have seen how between- and within-sample variance can be used to define F statistics and calculate a p-value. Let us take a slightly different look at the same problem. Segregation of all the measurements into groups implies the presence of some "label", on which we sort our data points (treated/untreated, male/female, high-school education/associate degree/full college degree/advanced degree, etc). The value of the label is associated with each data point alongside the primary measurement we want to segregate on that label (i.e. blood pressure, age, annual income –whatever we want to study as a "function" of our label). While each "primary" measurement is a random variable (we do not know the value in the subject we are going to draw next), the "label" can be also viewed as such: we draw next subject randomly, and that subject is characterized by both primary variable (e.g. annual income) and the class label

label = categorical, primary = continuous.

variable (e.g. highest level of education attained). So the problem we were just solving (testing for a difference between the means in each group) can be reformulated as the problem of finding the dependence between, e.g. income and highest level of education random variables.

This is simply a change in interpretation, and this new point of view is completely equivalent to what we have been discussing so far, do not overthink it. However, it helps us better understand some subtleties. First, as we will see soon, linear models are perfectly capable of fitting not only Y vs continuous X (which is a customary way for us to think about "functions") but also Y vs categorical variable(s) – the interpretation of group label as realization of a random categorical variable works perfectly in this case; but what we want to concentrate on for now, is the F statistics for the "conventional" linear model fit we have been looking at so far.

If group label is treated as a random variable X (and each subject provides a realization of that variable), and ANOVA allows for arbitrary number of groups, then each point x_i in our linear model fit $y_i = ax_i + b + \varepsilon$ can be considered as such "group label". If we fit two samples X,Y, of size N , we will have N such groups defined by "categories" (measured values) $x_1, \dots, x_n$.

If we look at the sample Y alone, we can only quantify it using descriptive statistics, such as sample mean and variance, or equivalently total sum of squares:

$$SS_{total} = \sum_{i=1}^{n}(y_i - \bar{y})$$

where $\bar{y}$ is the total sample mean of all the observed values of Y.

The predictions (or fitted values) of our linear model $\bar{y} = ax_i + b$ play the roles of the separate "sample means" in each "category" xi. Note that we do expect the distributions of measured values yi to be centered at the predicted values and normally distributed around them. In other words, if we were to measure more values of y for precisely the same value of xi, we certainly believe they would be distributed around the predicted value and (hopefully) follow a gaussian. Hence it is fair to treat that predicted value as the mean even though in reality we usually do not sample multiple y at the same x location but instead have just one value yi at each distinct point ("class") xi. More formally, in the case of a linear model we can further apply the same algebraic transformations as in the previous section and arrive to the breakdown of the total variance of Y:

(11)

$$SS_{total} = \underset{\text{explained}}{SS_{between}} + \underset{\text{unexplained}}{SS_{within}} = \sum_i(\bar{y}_i - \bar{y})^2 + \sum_i(\overset{y_i}{\cancel{\bar{y}_i}} - \bar{y}_i)^2$$

The first term is the "explained", "between-group" variance: $\bar{y}_i$ is completely determined and driven in a deterministic way by the corresponding value of x. Hence whatever deviation $\bar{y}_i$ exhibits from the global total $\bar{y}$, it is the "effect" of x – just like the "effect" of treatment vs control we had when we considered discrete class variables. The second term is exactly the sum of squared residuals as you can easily see. This is the noise that remains unfitted in our linear model, the "unexplained" variance in the data. regular ~~of residual~~ (data point – mean of sample)^2 = normal SS.

The way F-statistics is defined in most meaningful way for linear models involves one extra step: we usually consider nested models, in a sense that we add new dependences, but the old model is contained within the new one. For instance, a model with intercept only $Y \sim 1$ is contained (or nested) within linear model $Y \sim X$ (because we can always set slope a to 0 in $y = ax + b$; thus we can always reproduce the fit with intercept only in a more general linear model); similarly, $Y \sim X$ is nested within $Y \sim X + X^2$. Since sum of squares of normal random variables is always distributed as chi-square (when properly scaled and with proper number of degrees of freedom), we have some freedom in defining the groups/models for comparison in a way that makes most sense, and we will still end up with some F-statistics. For linear model, in particular, we add coefficients one by one and define F, progressively, as the ratio of the change in the residual sum of squares with addition of each term to the model ("between"), divided by the sum of the squares of the residuals ("within"):

The F statistic is a ratio that compares how much variance is explained by adding a new factor (eg. education level) vs. how much variance is left unexplained.

$$F = \frac{\left(\frac{RSS_1 - RSS_2}{p2-p1}\right)}{\left(\frac{RSS2}{n-p2}\right)}$$

RSS → tells us how much variation in the data is not explained by the model.

Smaller RSS = better prediction

Here indexes 1 and 2 refer to the two models we are comparing (e.g. intercept only vs intercept and slope, or intercept and slope vs intercept, slope, and quadratic), p is the number of independent variables in the model, and RSS is the residual sum of squares. Note that in a simple case of a model with one independent variable we still have the "classical" definition of the anova test as was given before: RSS_2 is simply the unexplained noise in the model, while in the "no-model" 1, the "residual" sum of squares is actually the total variation SS_{total} (since "model" 1 has no dependence at all). But $SS_{total} - SS_{within} = SS_{between}$, so that we have the usual $SS_{between}$ in the numerator.

It can be shown that F defined in this way indeed follows F-distribution, and what we actually testing for is the "significance" in reduction of unexplained noise (residuals) attained by adding a new variable. This has an important consequence: in the F-test defined in such a way, the order in which the variables are added matters. Consider a simple illustration:

Hide

```
library(ALL); data(ALL)
y<-20+10*1:20+5*(1:20)*(1:20)+rnorm(20,0,200)
x<-1:20 # x,y as in homework 3
anova(lm(y~x+I(x^2)))
```

← 1st model linear (x) then quadratic (x^2)

```
## Analysis of Variance Table
##
## Response: y
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## x          1 9216054 9216054 362.147 6.732e-13 ***
## I(x^2)     1  643723  643723  25.295  0.000103 ***
## Residuals 17  432622   25448
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

← linear term improves model

← quadratic term improves more, ~~[illegible]~~

↑ noise model cant explain (RSS)

$\Sigma F = \underline{\underline{387.4}}$

slope (x) ~~[illegible]~~ against intercept I

Hide

```
anova(lm(y~I(x^2)+x))
```

↑ quadratic term

← 2nd model quadratic (x^2) then linear.

```
## Analysis of Variance Table
##
## Response: y
##           Df  Sum Sq Mean Sq  F value    Pr(>F)
## I(x^2)     1 9856430 9856430 387.3112 3.894e-13 ***
## x          1    3346    3346   0.1315    0.7214
## Residuals 17  432622   25448
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

← quadratic ~~linear~~ term first explains almost all the variation

Adding the linear term (x) after that does almost nothing.

Note how we fit two models with different order of X and X^2 variables. The coefficients of these two models are, of course, exactly the same (not shown, check for yourself!): indeed, all the coefficients are determined jointly in such a way that the fitted model minimizes the RSS (residual sum of squares). However as the example above shows, the anova results do differ! This is because anova adds variables one by one, in the order they are defined in the formula. In the first application of ANOVA above, we test linear term (slope) against intercept only (equivalently, against sample mean only, i.e. no predicted effect at all), and observe dramatic improvement in RSS and thus very significant p-value. Then we add another term (X^2), and test the new model against the previous one ($Y \sim X$, slope only); we still get quite significant reduction in RSS. However if we test the models the other way around (second application of ANOVA in the code fragment above), we observe that quadratic term alone brings tremendous improvement as compared to no dependence (sample mean, or intercept only), and then the improvement that a linear term can bring after that is not even significant.

This can be interpreted in a number of ways. First, it is clear that if there is a dependence in the data, even completely non-linear, adding a linear term will "explain" some variance in the data in most cases even if this term is completely inadequate (e.g. if we modeled our dataset without a linear term at all, with quadratic only): linear fit would be obviously still much better than no fit at all (sample mean only) and only diagnostics would tell us that something is wrong with the fit. In our case, there is indeed linear part in the dependence, but quadratic term is clearly more important and spans the larger range of y values. Thus adding quadratic term after linear one still results in significant improvement. If we add quadratic term first, we seem to capture almost all the important variance and adding linear term after that does not help much. We can confirm this by examining the "effect" sum of squares (first term in (11)) as well as residual sums of squares reported by the two anova runs.

It looks like the linear term is indeed much less important and "shines" only when compared to no fit at all. To further investigate the situation, we may want to run a quadratic only fit $Y \sim I(X^2)$ and check the residuals and diagnostics again. This investigation illustrates that fitting the model is a trivial task that reduces to some matrix calculations, but choosing the right model, in the presence of noise and without knowledge of the likely functional shape of the underlying process, can be very non-trivial.

3 ANOVA Examples

Let us now work through an example of ANOVA analysis. One of the annotations available in the `ALL` dataset is the "fusion protein". Biologically, fusion protein is an aberrant, chimera molecule that is literally constructed from parts of two different proteins fused together. There is a number of known highly oncogenic fusion proteins and the patients in the ALL study were tested for a few known isoforms of BCR/ABL1 fusion relevant for the disease.

Let us extract the fusion protein status in all the patients from the data and save it alongside gene expression levels of some gene:

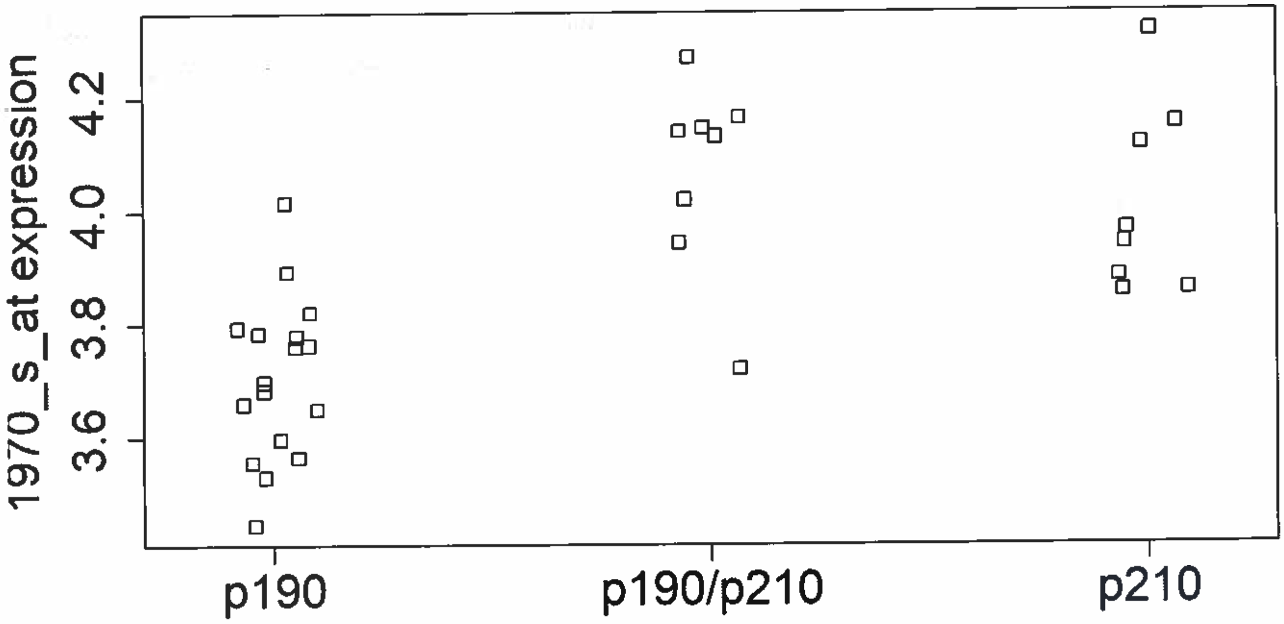
Hide

```
fp.df <- data.frame(fp=pData(ALL)[,"fusion protein"],
 g=exprs(ALL)["1970_s_at",])
table(fp.df$fp)
```

```
##
##      p190 p190/p210      p210
##        17         8         8
```

Hide

```
stripchart(fp.df$g~fp.df$fp,vertical=T,
ylab="1970_s_at expression",method='jitter',cex.lab=1.7,cex.axis=1.7)
```



Note that only a fraction of patients have fusion isoform reported (the rest either did not have the fusion detected, or were not tested): 17 patients have p190 isoform, and p210 and p190/p210 isoforms were found in 8 patients each; the rest of data are populated with NA. The plot of the expression levels of the selected gene, 1970_s_at, stratified by the fusion protein isoform is shown below (generated by the last command in the code fragment above).

It might have been guessed by now that our intention is to check if there is significant difference between expression levels of a gene in the three groups of patients, or as we discussed, the equivalent question is whether there is a significant association between expression level G and the fusion protein isoform categorical variable `Fp`.

After we formulated the problem as a search for the dependence between the two variables, it should not be at least too surprising that this task can be also performed using linear models:

dependent independent. → how does the categorical variable affect gene exprsn of gene 1970_s_at.

Hide

```
coef(lm(g~fp,fp.df))
```

```
## (Intercept) fpp190/p210       fpp210
##   3.7052303   0.3614408    0.3054996
```

Value A   Value B.

Hide

```
mean(fp.df$g[!is.na(fp.df$fp)])
```

– compute mean of column g, where fp is not N/A in the same row.

```
## [1] 3.866913
```

total mean   ∴ means of all fusions.

Hide

```
mean(fp.df$g[!is.na(fp.df$fp)&fp.df$fp=="p190"])
```

expression level mean of category p190.

```
## [1] 3.70523
```

= mean (190) = intercept.

Hide

```
mean(fp.df$g[!is.na(fp.df$fp)&fp.df$fp=="p190/p210"])
```

mean exprsn level of gene, ~~[illegible]~~ with label p190/p120.

```
## [1] 4.066671
```

= intercept + A × 1 = 3.705... + 0.361... = mean(p190/p210)

Hide

```
mean(fp.df$g[!is.na(fp.df$fp)&fp.df$fp=="p210"])
```

mean ~~ex~~ exprsn level of gene, with label p210.

```
## [1] 4.01073
```

= intercept + B × 1 = mean (p210).

total mean   mean (190)

Hide

```
17*(3.866913-3.7052303)^2+ # SSbetween
8*(3.866913-3.7052303-0.3614408)^2+ 8*(3.866913-3.7052303-0.3054996)^2
```

mean (190/210)   mean (210).

SSbetween.

```
## [1] 0.9290948
```

$$\sum_{j=1}^{p} n_j(\bar{x}-\bar{x}_j)^2$$

Hide

```
anova(lm(g~fp,fp.df))
```

```
## Analysis of Variance Table
##
## Response: g
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## fp         2 0.92909 0.46455  19.095 4.477e-06 ***
## Residuals 30 0.72984 0.02433
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

SSbetween groups = explained ~~ditto~~ variation [illegible]

low pvalue -> differences in mean for each group not by chance!!

The first command in the code segment above fits linear model of gene expression vs fusion protein status. Since `fp` is a categorical variable, we should be very careful in interpreting the coefficients of this “linear fit”. Remember that when we had a “conventional” linear regression fit, $y = ax + b$, the slope a was the “rate of change” of the variable y: the latter would change by a when x changes by 1. Moreover, when x is numerical, we can define “greater” and “less” relations, order the values, and can have changes in x from data point to data point equal to 1.7, 2.18 or 3.1415, as obvious as it sounds. In contrast, with categorical variable fp, we have three different values, but those values do not have any order, and we cannot define a “distance” between them (by how much does fp change from p190 to p210??), so we cannot really write a term like $a * fp$ in our model. Instead, when linear model is fitted on categorical data, each individual level of the categorical variable is encoded as 0 (not set) or 1 (set).

Next, you can remember that in “conventional” linear models we interpreted predicted value y_i as the “mean” of measurements we would expect to make when $x = x_i$ (even if in reality we had only one y_i measurement for each x_i, so we had to use other techniques to infer that mean). This interpretation remains exactly the same with linear models applied to categorical variables, except that in contrast to continuous variable x, when we have only a few levels of a categorical variable, it is likely that we have many measurements y_i for each level. Thus we can truly compute the mean in each category, and importantly this is all we can compute: since distance between the categorical labels is not defined, we cannot define a slope as well.

Let us now examine closely the coefficients calculated by our linear model in the code fragment shown above. We can see that the “Intercept” is the mean of the first category (p190), see the explicit calculation of the mean in that category few lines below.

Next value, let's denote it A, under `p190/p210`, should be interpreted as follows: when the category is p190/210 (i.e. the corresponding “flag” is 1, or “set”) then intercept+A gives the mean (in other words “predicted value”) of Y in that category. So in some sense, A is a slope, but only with respect to the indicator variable 0/1, `mean(p190/210)=intercept+A*1`. For p210 we need to use a different indicator variable, and different “slope” B, so that when the label is p210 (corresponding indicator is 1), then `mean(fpp210)=intercept+B*1`.

You can easily check that the coefficients reported by the linear model indeed give the means in the groups, when properly interpreted. This is left as a homework exercise.

Note that the means in the groups is the best and only “prediction” that we can make in this case, in the same way as a conventional linear model gives single predicted “mean” value for each “category” $x_i$in the data, and all the deviations of measured values of y from those predictions is unexplained noise.

nj

Let us now compute manually between-group sum of squares $SS_{between}$ explained noise following Eq. (8). In order to do that, we need the numbers of measurements in each group (17,8,8 as the `table()` command reported), we need to know the total mean (computed in the code fragment above) and also the group means (we will be using the coefficients returned by the linear model fit). Check the long numerical line in the code above and make sure you understand how in corresponds to Eq. 8. The result of our manual calculation of $SS_{between}$ is 0.92.... Finally, when we apply `anovaI()` to the linear model with categorical variable we just built, it does just the right thing: it calculates the multi-sample ANOVA statistics discussed in the first section of this Note. You can check, for instance, that the sum of squares (Sum. Sq. column) reported for variable `fp` (i.e. it is the sum of squares associated with that variable, in other words between-group, or “effect”) is equal to the value we produced manually (0.92...). The p-value corresponding to the resulting F-statistics (4.477e-06) indicates a very significant effect: the changes in (mean) expression level of the gene we selected in patients with different fusion protein isoforms seem to be very real.

The following code is almost the same except that we are looking now at the expression levels of a different gene, and we are also assessing the association between (mean) expression level and fusion protein status. There is no significant effect according to ANOVA in this case. Use this second calculation to practice more and to make sure you understand exactly all the numbers involved (what the coefficients of the categorical linear model are, how we compute $SS_{between} manually, etc.).

Hide

```
fp.df <- data.frame(fp=pData(ALL)[,"fusion protein"],
 g=exprs(ALL)["41682_s_at",])
table(fp.df$fp)
```

↑ different gene.

```
##
##      p190 p190/p210      p210
##        17         8         8
```

← same b/c of same patients

Hide

```
mean(fp.df$g[!is.na(fp.df$fp)])
```

total mean

```
## [1] 4.161295
```

Hide

dependent
independent

```
coef(lm(g~fp,fp.df))
```

```
## (Intercept) fpp190/p210      fpp210
##  4.14767131  0.04215518  0.01404192
```

A B

↑ mean(p190)

Hide

```
17*(4.161295-4.14767131)^2+ # SSbetween
 8*(4.161295-4.14767131-0.04215518)^2+ 8*(4.161295-4.14767131-0.01404192)^2
```

mean(p190/210) mean(p210)

```
## [1] 0.009669051
```

← SS between = explained variance.

Hide

```
anova(lm(g~fp,fp.df))
```

```
## Analysis of Variance Table
##
## Response: g
##           Df  Sum Sq  Mean Sq F value Pr(>F)
## fp         2 0.00967 0.004835  0.0515 0.9499
## Residuals 30 2.81759 0.093920
```

– high p-value → Accept null hyp. ~~these~~ The difference in means ~~are due to~~ is not due to an underlying relationship.

More ANOVA examples in WK5 pt2 topic notes.