Generalized Linear Models, Analysis of Variance,

Time Series, and Econometrics

Analysis of Variance (ANOVA)

ANOVA investigates special linear models, used for planning experiments or quality control. Here the matrix $\mathbf{X}$ of the deterministic predictors is a so-called *design-matrix* with 0-1 entries indicating that which predictors influence the response at all. For testing hypotheses, we will intensively use the following theorem and its corollaries.

Theorem 1 (Fisher–Cochran) *Let* $\mathbf{X} = (X_1, \dots, X_n)^T \sim \mathcal{N}_n(\mathbf{0}, \mathbf{I}_n)$ *be random vector and the quadratic forms* $Q = \mathbf{X}^T \mathbf{I}_n \mathbf{X} = \mathbf{X}^T \mathbf{X} = \sum_{i=1}^n X_i^2$ *and* $Q_j = \mathbf{X}^T \mathbf{A}_j \mathbf{X}$ $(j = 1, \dots, k)$ *be such that they satisfy*

$$Q = Q_1 + Q_2 + \cdots + Q_k,$$

where the $n \times n$ *symmetric matrix has rank* n_j $(j = 1, \dots, k \le n)$*. Then the random variables* $Q_1, Q_2, \dots, Q_k$ *are independent* $\chi^2(n_1)$-, $\chi^2(n_2)$- ,..., $\chi^2(n_k)$-*distributed if and only if*

$$\sum_{j=1}^k n_j = n.$$

Proof. One direction is trivial: if $Q_j \sim \chi^2(n_j)$, $j = 1, \dots, k$ and they are independent, then $\sum_{j=1}^k Q_j \sim \chi^2(\sum_{j=1}^k n_j)$. As $Q = \sum_{j=1}^k Q_j$ and $Q \sim \chi^2(n)$, it follows that $\sum_{j=1}^k n_j = n$.

In the other direction: assume that $\sum_{j=1}^k n_j = n$. Since $Q_j = \mathbf{X}^T \mathbf{A}_j \mathbf{X}$ and $\text{rank}(\mathbf{A}_j) = n_j$ $(j = 1, \dots, k)$, $\mathbf{A}_j$ has SD:

$$\mathbf{A}_j = \mathbf{U}_j \mathbf{\Lambda}_j \mathbf{U}_j^T = \mathbf{U}_j \tilde{\mathbf{\Lambda}}_j^{1/2} \Delta_j \tilde{\mathbf{\Lambda}}_j^{1/2} \mathbf{U}_j^T = \mathbf{B}_j \Delta_j \mathbf{B}_j^T,$$

where the diagonal matrix $\tilde{\mathbf{\Lambda}}_j$ contains the absolute values of the non-zero eigenvalues of $\mathbf{A}_j$ in the diagonal positions

$$N_j = \{\sum_{h=1}^{j-1} n_h + 1, \dots, \sum_{h=1}^{j} n_h\},$$

otherwise zeros. Because of the starting assumption, $\cup_{j=1}^k N_j = \{1, \dots, n\}$ is a disjoint union. The diagonal matrix Δ_j compensates for the signs, it contains ± 1s in the positions N_j. All the matrices are $n \times n$, and the matrix $\mathbf{B}_j = \mathbf{U}_j \tilde{\mathbf{\Lambda}}_j^{1/2}$ contains orthogonal eigenvectors, corresponding to the the non-zero eigenvalues of $\mathbf{A}_j$ in the columns of indices in N_j, otherwise zeros.

Therefore, with the notation $\mathbf{Z}_j = \mathbf{B}_j^T \mathbf{X}$,

$$Q_j = \mathbf{X}^T \mathbf{A}_j \mathbf{X} = \mathbf{X}^T \mathbf{B}_j \Delta_j \mathbf{B}_j^T \mathbf{X} = \mathbf{Z}_j^T \Delta_j \mathbf{Z}_j = \sum_{l \in N_j} \pm Z_j^2(l),$$

where $Z_j(l)$ denotes the lth coordinate of the vector $\mathbf{Z}_j$.

Because of the starting condition, the non-zero segments of the vectors $\mathbf{Z}_j$s can be concatenated into a vector $\mathbf{Z}$, and $\Delta := \sum_{j=1}^{k} \Delta_j$ be diagonal matrix (with ± 1s along its diagonal). Further, $\mathbf{B} := \sum_{j=1}^{k} \mathbf{B}_j$. With these,

$$\mathbf{X}^T \mathbf{X} = \sum_{j=1}^{k} \mathbf{X}^T \mathbf{A}_j \mathbf{X} = \sum_{i=1}^{n} \pm Z^2(i) = \mathbf{Z}^T \Delta \mathbf{Z} = \mathbf{X}^T \mathbf{B} \Delta \mathbf{B}^T \mathbf{X},$$

that for arbitrary $\mathbf{X}$ is only possible if

$$\mathbf{B} \Delta \mathbf{B}^T = \mathbf{I}_n,$$

i.e.,

$$\Delta = (\mathbf{B}\mathbf{B}^T)^{-1}.$$

But $\mathbf{B}\mathbf{B}^T$ is a Gram-matrix (positive semidefinite), and it is of full rank, so it is positive definite. Therefore, Δ should have all 1s along its diagonal. Consequently, $\Delta = \mathbf{I}_n$ and the matrix $\mathbf{B}$ is orthogonal.

Summarizing, $\mathbf{Z} = \mathbf{B}^T \mathbf{X} \sim \mathcal{N}_n(\mathbf{0}, \mathbf{I}_n)$ and

$$Q_j = \mathbf{X}^T \mathbf{A}_j \mathbf{X} = \sum_{l \in N_j} Z_j^2(l) \sim \chi^2(n_j), \quad j = 1, \dots, k.$$

Q_js are independent, because they are constituted with the sum of the squares of the i.i.d. $\mathcal{N}(0,1)$ coordinates of $\mathbf{Z}$ in the disjoint index sets $N_1, \dots, N_k$. This finishes the proof.

Note that the linear algebra content of this theorem is as follows. We provide the decomposition

$$\mathbf{I}_n = \mathbf{A}_1 + \cdots + \mathbf{A}_k,$$

where $\mathbf{A}_1, \dots, \mathbf{A}_k$ are symmetric matrices, and

$$\mathrm{rank}(\mathbf{A}_1) + \cdots + \mathrm{rank}(\mathbf{A}_k) = n.$$

Then the matrices are also idempotent, and correspond to orthogonal projections onto disjoint subspaces of $\mathbb{R}^n$, the direct sum of which is $\mathbb{R}^n$.

We also list some propositions which follow from the Fisher–Cochran theorem:

Proposition 1 *Consider $\mathbf{X} \in \mathcal{N}_n(\mathbf{0}, \mathbf{I}_n)$. Then*

- *(a) With some symmetric matrix $\mathbf{A}$, $\mathbf{X}^T \mathbf{A} \mathbf{X}$ is χ^2-distributed if $\mathbf{A}^2 = \mathbf{A}$. Then the degree of freedom of the χ^2-distribution*$=\mathrm{rank}(\mathbf{A}) = \mathrm{tr}(\mathbf{A})$.

- *(b) With some symmetric matrices $\mathbf{A}$, $\mathbf{B}$, the random variables $\mathbf{X}^T \mathbf{A} \mathbf{X} \sim \chi^2(a)$ and $\mathbf{X}^T \mathbf{B} \mathbf{X} \sim \chi^2(b)$ are independent if $\mathbf{A}\mathbf{B} = \mathbf{0}$ (then obviously, $a + b \le n$).*

- *(c) If $Q = Q_1 + Q_2$, $Q \sim \chi^2(a)$, $Q_1 \sim \chi^2(b)$, and Q_2 is positive semidefinite, then $Q_2 \sim \chi^2(a - b)$; further, Q_1 and Q_2 are independent.*

In the one-way case, we investigate whether different treatments (conditions) influence significantly some continuous measurements (response). For example, whether the GDP differs significantly in different countries. If the continuous measurement is normally distributed, this is the generalization of the t-test for more than two groups.

In the two-way case, two kinds of treatments are given on different levels, and we have a two-way classified table of continuous measurements. For example, whether the GDP differs significantly in different countries and under different economic regulations. Then we may investigate the effect of the countries, the effect of the regulations, and the interaction between them. We discuss these models with precise formulas.

ANOVA models:

- *One-way ANOVA*: Our sample is taken in k different groups: X_{ij}, $j = 1, \dots, n_i$, $i = 1, \dots, k$, and the sample size is $n = \sum_{i=1}^k n_i$. Assume that in group i our observations follow $\mathcal{N}(b_i, \sigma^2)$-distribution. It is important that the observations are independent with the same variance (homoscedasticity). We want to test the null-hypothesis

  $$H_0 : b_1 = b_2 = \dots = b_k,$$

  which is the generalization of the t-test for k groups. We use the decomposition $b_i = \mu + a_i$, where

  $$\mu = \frac{1}{n} \sum_{i=1}^k n_i b_i, \qquad a_i = b_i - \mu \quad (i = 1, \dots, k).$$

  Obviously,

  $$\sum_{i=1}^k n_i a_i = 0.$$

  With this notation, our model is

  $$X_{ij} = \mu + a_i + \varepsilon_{ij} \qquad (j = 1, \dots, n_i;\, i = 1, \dots, k) \tag{1}$$

  where $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$ are i.i.d. error terms. This model is, in fact, a linear model with the n-dimensional vectors

  $$\begin{aligned} \mathbf{Y} &:= (X_{11}, \dots, X_{1n_1}, X_{21}, \dots, X_{2n_2}, \dots, X_{k1}, \dots, X_{kn_k})^T \\ \underline{\varepsilon} &:= (\varepsilon_{11}, \dots, \varepsilon_{1n_1}, \varepsilon_{21}, \dots, \varepsilon_{2n_2}, \dots, \varepsilon_{k1}, \dots, \varepsilon_{kn_k})^T \end{aligned}$$

  and parameter-vector $\mathbf{a} = (a_1, \dots, a_k)^T$. With these, the above model is

  $$\mathbf{Y} = \mathbf{B}\mathbf{a} + \mu \mathbf{1} + \underline{\varepsilon}$$

  where the vector $\mathbf{1} \in \mathbb{R}^n$ has all 1 coordinates, and the $n \times k$ design matrix $\mathbf{B}$ is such that in its i-th column it contains all 0's, except the block i, where it contains 1's. This ensures that in group i only the parameter a_i appears in the model equation (1).

  The parameters are estimated by the method of least squares: we minimize

  $$\sum_{i=1}^k \sum_{j=1}^{n_i} \varepsilon_{ij}^2 = \sum_{i=1}^k \sum_{j=1}^{n_i} (X_{ij} - \mu - a_i)^2.$$

The least square estimates of the parameters are

$$\hat{\mu} = \bar{X}_{..} \quad \text{and} \quad \hat{a}_i = \bar{X}_{i.} - \bar{X}_{..} \quad (i = 1, \ldots, k),$$

where

$$\bar{X}_{i.} = \frac{1}{n_i} \sum_{j=1}^{n_i} X_{ij} \quad (i = 1, \ldots, k) \quad \text{and} \quad \bar{X}_{..} = \frac{1}{n} \sum_{i=1}^{k} \sum_{j=1}^{n_i} X_{ij}.$$

The minimum is

$$SSE = Q_e = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (X_{ij} - \hat{m} - \hat{a}_i)^2 = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_{i.})^2.$$

With

$$SSR = Q_a = \|\mathbf{B}\hat{\mathbf{a}}\|^2 = \sum_{i=1}^{n_i} n_i \hat{a}_i^2 = \sum_{i=1}^{k} n_i (\bar{X}_{i.} - \bar{X}_{..})^2,$$

we can decompose the total variation of the sample ($SST = Q$) into between-groups ($SSR = Q_a$) and within-groups ($SSE = Q_e$) variation as follows:

$$Q = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_{..})^2 = \sum_{i=1}^{k} \sum_{j=1}^{n_i} [(X_{ij} - \bar{X}_{i.}) + (\bar{X}_{i.} - \bar{X}_{..})]^2 =$$

$$= \sum_{i=1}^{k} \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_{i.})^2 + \sum_{i=1}^{k} \sum_{j=1}^{n_i} (\bar{X}_{i.} - \bar{X}_{..})^2 =$$

$$= \sum_{i=1}^{k} n_i (\bar{X}_{i.} - \bar{X}_{..})^2 + \sum_{i=1}^{k} \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_{i.})^2 = Q_a + Q_e.$$

This decomposition is summarized in the 1-way ANOVA table:

Cause of the dispersion	Sum of squares	Degrees of freedom	Empirical variance
Between groups	$Q_a = \sum_{i=1}^{k} n_i (\bar{X}_{i.} - \bar{X}_{..})^2$	$k - 1$	$s_a^2 = \frac{Q_a}{k-1}$
Within groups	$Q_e = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_{i.})^2$	$n - k$	$s_e^2 = \frac{Q_e}{n-k}$
Total	$Q = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_{..})^2$	$n - 1$	-

In the above model, we first investigate the null-hypothesis $\mu = 0$. If we reject it, we investigate the null-hypothesis, that there is no difference between the groups:

$$H_0 : a_1 = \cdots = a_k = 0, \quad \text{briefly,} \quad \mathbf{a} = \mathbf{0}.$$

In view of the Fisher–Cochran theorem and its consequences, $Q_e \sim \sigma^2\chi^2(n-k)$, irrespective whether H_0 holds or not. However, the expectation of the linear expressions in Q_a is

$$\mathbb{E}(\bar{X}_{i.} - \bar{X}_{..}) = \mathbb{E}(\bar{X}_{i.}) - \mathbb{E}(\bar{X}_{..}) = a_i - \frac{1}{n}\sum_{j=1}^{k} n_j a_j \qquad (i = 1, \dots, k)\,,$$

which can be zero for all i only if H_0 holds. In this case, with the Fisher–Cochran theorem and its consequences, $Q_a \sim \sigma^2\chi^2(k-1)$; further, Q_e and Q_a are independent of each other. Observe that the degrees of freedom in the decomposition

$$Q = Q_a + Q_e$$

are added together:

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

Therefore, with the notation

$$s_a^2 = \frac{Q_a}{k-1} \qquad \text{and} \qquad s_e^2 = \frac{Q_e}{n-k},$$

the test statistic

$$F = \frac{s_a^2}{s_e^2} = \frac{Q_a}{Q_e} \cdot \frac{n-k}{k-1} \sim \mathcal{F}(k-1, n-k)$$

follows Fisher F-distribution with the above degrees of freedom under H_0. Summarizing, if $F \geq F_\alpha(k-1, n-k)$, i.e., the between-group variances are significantly larger than the within-group-ones, then we reject H_0 with significance α.

- *Bartlett-test* for testing equality of the variances of the groups:

  $$H_0 \,:\, \sigma_1 = \cdots = \sigma_k.$$

  Based on the above grouped sample, the test statistic is

  $$B^2 = \frac{2.3026}{c}\left(f \lg S^{*2} - \sum_{i=1}^{k} f_i \lg S_i^{*2}\right),$$

  where $f_i = n_i - 1$ $(i = 1, \dots, k)$; $f = \sum_{i=1}^{k} f_i$; S_1^{*2}, …,S_k^{*2} are the corrected empirical variance within the groups, and

  $$S^{*2} = \frac{1}{f}\sum_{i=1}^{k} f_i S_i^{*2}, \quad c = 1 + \frac{1}{3(k-1)}\left(\sum_{i=1}^{k}\frac{1}{f_i} - \frac{1}{f}\right).$$

  Bartlett proved that for 'large' sample sizes, B^2 asymptotically follows $\chi^2(k-1)$-distribution. Therefore, if $B^2 \geq \chi^2_\alpha(k-1)$, then we reject H_0 with significance α.

- *Two-way ANOVA without interaction*: We have two-way classified data in $k \cdot p$ groups. Here one observation per group suffices. Let X_{ij} denote the continuous measurements ($i = 1, \ldots, k$; $j = 1, \ldots, p$), the sample size is $n = kp$. We assume that there is no interaction between the two treatments (based on the levels of which we form the groups).

  In our model, the independent, homoscedastic sample entries have $X_{ij} \sim \mathcal{N}(\mu + a_i + b_j, \sigma^2)$ distribution. Therefore, our linear model is the following:

  $$X_{ij} = \mu + a_i + b_j + \varepsilon_{ij}, \qquad (i = 1, \ldots, k;\ j = 1, \ldots, p) \tag{2}$$

  where $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$ are i.i.d. errors. The parameters a_i's and b_j's denote the non-interacting effects of the levels of the two treatments. We can assume (with the choice of μ) that

  $$\sum_{i=1}^{k} a_i = 0 \qquad \text{and} \qquad \sum_{j=1}^{p} b_j = 0.$$

  Here we do not specify the design matrices, but (2) also fits into the framework of linear models.

  Minimizing the objective function

  $$\sum_{i=1}^{k} \sum_{j=1}^{p} \varepsilon_{ij}^2 = \sum_{i=1}^{k} \sum_{j=1}^{p} (X_{ij} - \mu - a_i - b_j)^2$$

  by the method of least squares, we obtain the following estimates of the parameters:

  $$\begin{aligned} \hat{\mu} &= \bar{X}_{..}\,, \\ \hat{a}_i &= \bar{X}_{i.} - \bar{X}_{..} \qquad (i = 1, \ldots, k)\,,, \\ \hat{b}_j &= \bar{X}_{.j} - \bar{X}_{..} \qquad (j = 1, \ldots, p)\,, \end{aligned}$$

  where

  $$\begin{aligned} \bar{X}_{i.} &= \frac{1}{p} \sum_{j=1}^{p} X_{ij} \quad (i = 1, \ldots, k) \\ \bar{X}_{.j} &= \frac{1}{k} \sum_{i=1}^{k} X_{ij} \quad (j = 1, \ldots, p) \\ \bar{X}_{..} &= \frac{1}{n} \sum_{i=1}^{k} \sum_{j=1}^{p} X_{ij}. \end{aligned}$$

  With this, the minimum of our objective function is

  $$SSE = Q_e = \sum_{i=1}^{k} \sum_{j=1}^{p} (X_{ij} - \hat{\mu} - \hat{a}_i - \hat{b}_j)^2.$$

  Further, we have the decomposition

  $$Q = Q_a + Q_b + Q_e \tag{3}$$

of the total variance ($SST = Q$) into the variances caused by the a-effects, b-effects and the error (Q_a, Q_b, Q_e, where $SSR = Q_a + Q_b$).

This decomposition is summarized in the 2-way ANOVA table (without interaction):

Cause of the dispersion	Sum of squares	Degree of freedom	Empirical variance
a-effects	$Q_a = p\sum_{i=1}^{k}(\bar{X}_{i.} - \bar{X}_{..})^2$	$k-1$	$s_a^2 = \frac{Q_a}{k-1}$
b-effects	$Q_b = k\sum_{j=1}^{p}(\bar{X}_{.j} - \bar{X}_{..})^2$	$p-1$	$s_b^2 = \frac{Q_b}{p-1}$
Random error	$Q_e = \sum_{i=1}^{k}\sum_{j=1}^{p}(X_{ij} - \bar{X}_{i.} - \bar{X}_{.j} + \bar{X}_{..})^2$	$(k-1)(p-1)$	$s_e^2 = \frac{Q_e}{(k-1)(p-1)}$
Total	$Q = \sum_{i=1}^{k}\sum_{j=1}^{p}(X_{ij} - \bar{X}_{..})^2$	$kp-1$	-

If we have rejected the null-hypothesis $\mu = 0$, we compare the levels of both treatments, separately. To compare a-effects, we investigate

$$H_{0a}:\, a_1 = a_2 = \cdots = a_k = 0, \quad \text{briefly,} \quad \mathbf{a} = \mathbf{0}.$$

To compare b-effects, we investigate

$$H_{0b}:\, b_1 = b_2 = \cdots = b_p = 0, \quad \text{briefly,} \quad \mathbf{b} = \mathbf{0}.$$

In view of the Fisher–Cochran theorem and its consequences, since the degrees of freedoms of the terms in (3) are added together,

$$kp - 1 = (k-1) + (p-1) + (k-1)(p-1),$$

we have the following facts:

- $Q_e/\sigma^2 \sim \chi^2((k-1)(p-1))$, irrespective whether the above hypotheses hold or not.
- Under H_{0a}, $Q_a/\sigma^2 \sim \chi^2(k-1)$ and is independent of Q_e.
- Under H_{0b}, $Q_b/\sigma^2 \sim \chi^2(p-1)$ and is independent of Q_e.

Therefore, under H_{0a}, the test statistic

$$F_a = \frac{s_a^2}{s_e^2} \sim \mathcal{F}(k-1, (k-1)(p-1)),$$

and if $F_a \geq F_\alpha(k-1, (k-1)(p-1))$, then we reject H_{0a} with significance α.

Likewise, Therefore, under H_{0b}, the test statistic

$$F_b = \frac{s_b^2}{s_e^2} \sim \mathcal{F}(p-1, (k-1)(p-1)),$$

and if $F_b \geq F_\alpha(p-1, (k-1)(p-1))$, then we reject H_{0b} with significance α.

- *Two-way ANOVA with interaction*: Here we also have two-way classified data in $k \cdot p$ groups, but we have more than one (say, n) observations per cell, since there is interaction between the two treatments. The independent, homoscedastic sample entries are X_{ijl} $(i = 1, \dots, k;\ j = 1, \dots, p;\ l = 1, \dots, n)$. The sample size is kpn. Supposing that $X_{ijl} \sim \mathcal{N}(\mu + a_i + b_j + c_{ij}, \sigma^2)$, our linear model is the following:

$$X_{ijl} = \mu + a_i + b_j + c_{ij} + \varepsilon_{ijl}, \qquad (i = 1, \dots, k;\ j = 1, \dots, p),$$

where $\varepsilon_{ijl} \sim \mathcal{N}(0, \sigma^2)$ are i.i.d. errors, a_i's and b_j's denote the effects of the two treatments, whereas c_{ij}'s are the interactions. We assume that

$$\sum_{i=1}^{k} a_i = 0, \qquad \sum_{j=1}^{p} b_j = 0,$$
$$\sum_{i=1}^{k} c_{ij} = 0 \qquad (j = 1, \dots, p) \quad \text{és}$$
$$\sum_{j=1}^{p} c_{ij} = 0 \qquad (i = 1, \dots, k).$$

This model is also a linear one.

By the method of least squares, the minimum of

$$\sum_{i=1}^{k}\sum_{j=1}^{p}\sum_{l=1}^{n} \varepsilon_{ijl}^2 = \sum_{i=1}^{k}\sum_{j=1}^{p}\sum_{l=1}^{n} (X_{ijl} - \mu - a_i - b_j - c_{ij})^2$$

with respect to the parameters $\mu, a_1, \dots, a_k, b_1, \dots, b_p$, under the above constraints, is attained at

$$\hat{\mu} = \bar{X}_{...}\,,$$
$$\hat{a}_i = \bar{X}_{i..} - \bar{X}_{...} \qquad (i = 1, \dots, k)\,,$$
$$\hat{b}_j = \bar{X}_{.j.} - \bar{X}_{...} \qquad (j = 1, \dots, p)\,,$$
$$\hat{c}_{ij} = \bar{X}_{ij.} - \bar{X}_{i..} - \bar{X}_{.j.} + \bar{X}_{...} \qquad (i = 1, \dots, k; \quad j = 1, \dots, p)\,,$$

where

$$\bar{X}_{i..} = \frac{1}{pn}\sum_{j=1}^{p}\sum_{l=1}^{n} X_{ijl} \quad (i = 1, \dots, k)$$

$$\bar{X}_{.j.} = \frac{1}{kn}\sum_{i=1}^{k}\sum_{l=1}^{n} X_{ijl} \quad (j = 1, \dots, p)$$

$$\bar{X}_{ij.} = \frac{1}{n}\sum_{l=1}^{n} X_{ijl} \qquad (i = 1, \dots, k;\ j = 1, \dots, p)$$

$$\bar{X}_{...} = \frac{1}{kpn}\sum_{i=1}^{k}\sum_{j=1}^{p}\sum_{l=1}^{n} X_{ijl}.$$

The least square estimates of the parameters are

$$\begin{aligned}
\hat{m} &= \bar{X}_{...}\,, \\
\hat{a}_i &= \bar{X}_{i..} - \bar{X}_{...} \qquad (i = 1, \dots, k)\,, \\
\hat{b}_j &= \bar{X}_{.j.} - \bar{X}_{...} \qquad (j = 1, \dots, p)\,, \\
\hat{c}_{ij} &= \bar{X}_{ij.} - \bar{X}_{i..} - \bar{X}_{.j.} + \bar{X}_{...} \qquad (i = 1, \dots, k; \quad j = 1, \dots, p)\,,
\end{aligned}$$

and the minimum is

$$SSE = Q_e = \sum_{i=1}^{k}\sum_{j=1}^{p}\sum_{l=1}^{n}(X_{ijl} - \hat{m} - \hat{a}_i - \hat{b}_j - \hat{c}_{ij})^2.$$

For the decomposition

$$Q = Q_a + Q_b + Q_c + Q_e$$

we again have the ANOVA-table:

Cause of dispersion	Sum of squares	Degree of freedom	Empirical variance
a-effects	$Q_a = pn\sum_{i=1}^{k}(\bar{X}_{i..} - \bar{X}_{...})^2$	$k-1$	$s_a^2 = \frac{Q_a}{k-1}$
b-effects	$Q_b = kn\sum_{j=1}^{p}(\bar{X}_{.j.} - \bar{X}_{...})^2$	$p-1$	$s_b^2 = \frac{Q_b}{p-1}$
ab-interaction	$Q_c = n\sum_{i=1}^{k}\sum_{j=1}^{p}(\bar{X}_{ij.} - \bar{X}_{i..} - \bar{X}_{.j.} + \bar{X}_{...})^2$	$(k-1)(p-1)$	$s_c^2 = \frac{Q_c}{(k-1)(p-1)}$
Random error	$Q_e = \sum_{i=1}^{k}\sum_{j=1}^{p}\sum_{l=1}^{n}(X_{ijl} - \bar{X}_{ij.})^2$	$kp(n-1)$	$s_e^2 = \frac{Q_e}{kp(n-1)}$
Total	$Q = \sum_{i=1}^{k}\sum_{j=1}^{p}\sum_{l=1}^{n}(X_{ijl} - \bar{X}_{...})^2$	$kpn-1$	-

After rejecting the null-hypothesis $\mu = 0$, we investigate the interaction:

$$H_{0ab}: \; c_{ij} = 0, \quad (i = 1, \dots, k;\; j = 1, \dots, p).$$

If we accept it (no interaction), we investigate separately

$$H_{0a}: \; a_1 = a_2 = \dots = a_k = 0$$

and

$$H_{0b}: \; b_1 = b_2 = \dots = b_p = 0.$$

Using the Fisher–Cochran theorem and its consequences, further, the additivity of the degrees of freedoms,

$$kpn - 1 = (k-1) + (p-1) + (k-1)(p-1) + kp(n-1),$$

we have the following facts:

- $Q_e/\sigma^2 \sim \chi^2(kp(n-1))$, always.

– Under H_{0a}, $Q_a/\sigma^2 \sim \chi^2(k-1)$ and is independent of Q_e.
– Under H_{0b}, $Q_b/\sigma^2 \sim \chi^2(p-1)$ and is independent of Q_e.
– Under H_{0ab}, $Q_c/\sigma^2 \sim \chi^2((k-1)(p-1))$ and is independent of Q_e.

Therefore, we have the following test statistics: Under H_{0ab},

$$F_{ab} = \frac{s_c^2}{s_e^2} \sim \mathcal{F}((k-1)(p-1), kp(n-1)),$$

Under H_{0a},

$$F_a = \frac{s_a^2}{\tilde{s}_e^2} \sim \mathcal{F}(k-1, kpn-k-p+1).$$

Under H_{0b},

$$F_b = \frac{s_b^2}{\tilde{s}_e^2} \sim \mathcal{F}(k-1, kpn-k-p+1).$$

Then you can make the conclusions at significance α.

Note that the are so-called mixed ANOVA models, with different number of observations per cell, or with more than two factors. In these cases, we have to build up a design-matrix and use the Gauss normal equations to estimate the parameters.

- The ANOVA model can be extended to multivariate, grouped observations. In the 1-way Multivariate Analysis of Variance (MANOVA) setup, our p-variate measurements

$$\mathbf{Y}_{ij} \sim \mathcal{N}_p(\mathbf{m} + \mathbf{a}_i, \mathbf{C}) \qquad (j = 1, \dots, n_i;\ i = 1, \dots, k)$$

are assigned to k different groups, where $\sum_{i=1}^k \mathbf{a}_i = \mathbf{0}$ is assumed. Our inference is based on the decomposition

$$\mathbf{T} = \mathbf{B} + \mathbf{W}$$

of n times the $p \times p$ sample covariance matrix into *between- and within-group covariance matrices* in the following way:

$$\begin{aligned} \mathbf{T} &= \sum_{i=1}^{k}\sum_{j=1}^{n_i}(\mathbf{Y}_{ij} - \bar{\mathbf{Y}}_{..})(\mathbf{Y}_{ij} - \bar{\mathbf{Y}}_{..})^T \\ \mathbf{B} &= \sum_{i=1}^{k} n_i(\bar{\mathbf{Y}}_{i.} - \bar{\mathbf{Y}}_{..})(\bar{\mathbf{Y}}_{i.} - \bar{\mathbf{Y}}_{..})^T \\ \mathbf{W} &= \sum_{i=1}^{k}\sum_{j=1}^{n_i}(\mathbf{Y}_{ij} - \bar{\mathbf{Y}}_{i.})(\mathbf{Y}_{ij} - \bar{\mathbf{Y}}_{i.})^T \end{aligned} \qquad (4)$$

where $\bar{\mathbf{Y}}_{..} = \frac{1}{n}\sum_{i=1}^k \sum_{j=1}^{n_i} \mathbf{Y}_{ij}$ is the sample mean vector, while $\bar{\mathbf{Y}}_{i.} = \frac{1}{n_i}\sum_{j=1}^{n_i} \mathbf{Y}_{ij}$ is the mean vector of group i $(i = 1, \dots, k)$.

Time series

Definition 1 *The stochastic process* (X_t) *is strongly stationary if* $\forall n \in \mathbb{N}$ *and* $\forall h \in \mathbb{R}$*: the joint distribution of* $X_{t_1+h}, \dots, X_{t_n+h}$ *is the same as the joint distribution of* $X_{t_1}, \dots, X_{t_n}$ *for any time instances* $t_1 \le \dots \le t_n$*.*

Definition 2 *The stochastic process* (X_t) *is weakly stationary if* $\mathbb{E}(X_t) = \mu$ *exists and* $\forall t \in \mathbb{R}$ *and* $\forall h \in \mathbb{R}$*:*

$$\mathrm{Cov}(X_t, X_{t+h}) = \mathrm{Cov}(X_0, X_h) = c_h$$

does not depend on t*. Here* c_h *is called covariance-function.*

Note that in the Gaussian case the two notions are equivalent. Further, to any weakly stationary process there exists a strongly stationary Gaussian with the same covariance-function.

Usually, our time series is discrete (observed at equidistant times): $X_0, X_1, \dots, X_n, \dots$. If it is weakly stationary, then the covariance-function is

$$c_i = \mathrm{Cov}(X_i, X_0) = \mathrm{Cov}(X_{i+h}, X_h), \qquad i = 1, 2, \dots$$

The so-called *auto-correlation*s are:

$$r_i = \frac{\mathrm{Cov}(X_i, X_0)}{\sqrt{\mathrm{Var}(X_i)\mathrm{Var}(X_0)}} = \frac{c_i}{c_0}, \qquad i = 1, 2, \dots$$

The series of autocorrelations can be used for the characterization of the process, for example, to decide whether some of the following models fits to it, and also gives a hint for the number of parameters.

Definition 3 (ε_t) *is a white noise process if*

$$\mathbb{E}(\varepsilon_t) = 0, \quad \mathbb{E}(\varepsilon_t \varepsilon_s) = \delta_{ts}, \quad t, s = 1, 2, \dots$$

For example, when ε_t's are i.i.d. standard Gaussians.

- *AR*(p): *autoregressive process* of order p. It is defined by

  $$X_t = a_1 X_{t-1} + a_2 X_{t-2} + \dots + a_p X_{t-p} + \sigma\varepsilon_t, \quad t = 0, \pm 1, \pm 2, \dots$$

  where (ε_t) is a white noise process, and ε_t is uncorrelated with $X_{t-1}, X_{t-2}, \dots$. The coefficient vector $\mathbf{a} = (a_1, \dots, a_p)^T$ can be estimated as $\hat{\mathbf{a}} = \mathbf{C}^{-1}\mathbf{d}$, based on a sample for $t = 1, \dots, T$, where

  $$\hat{\mathbf{C}} = \begin{pmatrix} \hat{c}_0 & \hat{c}_1 & \dots & \hat{c}_{p-1} \\ \hat{c}_1 & \hat{c}_0 & \dots & \hat{c}_{p-2} \\ \vdots & \vdots & \vdots & \vdots \\ \hat{c}_{p-1} & \hat{c}_{p-2} & \dots & \hat{c}_0 \end{pmatrix}, \qquad \hat{\mathbf{d}} = \begin{pmatrix} \hat{c}_1 \\ \hat{c}_2 \\ \vdots \\ \hat{c}_p \end{pmatrix}.$$

- $ARMA(p, q)$: *autoregressive, moving average process* of order p, q. It is defined by

  $$X_t = a_1 X_{t-1} + a_2 X_{t-2} + \dots + a_p X_{t-p} + b_0\varepsilon_t + b_1\varepsilon_{t-1} + \dots + b_q\varepsilon_{t-q},$$

  where (ε_t) is a white noise process, and ε_t is uncorrelated with $X_{t-1}, \dots, X_{t-p}$. If $a_1 = \dots = a_p = 0$, it is a $MA(q)$ purely moving average process.

Note that every finite order AR process can equivalently written as an infinite MA.

We will not go into details, see e.g., the book of Karlin, S. and Taylor, H. M.: Stochastic Processes. We can also give conditions under which the above processes are weakly stationary.

Econometrics

This a collection of generalized regression, and mixed regression and ANOVA models, frequently using time series. Just as a sample, we enlist some situations when the classical linear model is not applicable, together with the method to solve them.

- *Heteroscedascity*: $\mathbf{Y} = \mathbf{X}\mathbf{a} + \underline{\varepsilon}$, $\mathrm{Var}(\underline{\varepsilon}) = \sigma^2\mathbf{B}$, where $\mathbf{B} > 0$ is a given positive definite matrix, while σ^2 (volatility) and $\mathbf{a}$ are to be estimated.

  The problem is solved by transforming the data. With the transformation

  $$\mathbf{Z} := \mathbf{B}^{-1/2}\mathbf{Y} \quad \text{and} \quad \mathbf{U} := \mathbf{B}^{-1/2}\mathbf{X},$$

  the model is
  $$\mathbf{Z} = \mathbf{U}\mathbf{a} + \underline{\varepsilon}', \quad \mathrm{Var}(\underline{\varepsilon}') = \sigma^2\mathbf{I}.$$

  Hence, $\hat{\mathbf{a}} = (\mathbf{U}^T\mathbf{U})^{-1}\mathbf{U}^T\mathbf{Z}$.

- *Dependence*: $Y = \mathbf{a}^T\mathbf{X} + \varepsilon$, where ε is correlated with $\mathbf{X}$. This may happen if both the predictors and the response are *endogenous*, and hence, not independent. In this case, the endogenous variables are to be substituted by *exogenous* ones ($\mathbf{Z}$), which are highly correlated with them, but uncorrelated with Y. For example:

  - Y: food-consumption/head (endogenous)
  - X: food prices (endogenous)
  - Z: farm prices (exogenous)

  The problem is solved by means of so-called *instrumental variables*: Z is highly correlated with X, but uncorrelated with the disturbance ε (and Y). After this, the *two-stage least squares method* is the following:

  - Regress $\mathbf{X}$ with $\mathbf{Z}$: $\hat{\mathbf{X}} = \mathbf{Z}(\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T\mathbf{X}$, which projects the column space of $\mathbf{X}$ onto the column space of $\mathbf{Z}$.
  - Estimate the parameter $\mathbf{a}$ of the usual linear model

    $$\mathbf{Y} = \hat{\mathbf{X}}^T\mathbf{a} + \underline{\varepsilon}'.$$

    The solution is

    $$\mathbf{a}^* = (\hat{\mathbf{X}}^T\hat{\mathbf{X}})^{-1}\hat{\mathbf{X}}^T\mathbf{Y} = (\hat{\mathbf{X}}^T\mathbf{X})^{-1}\hat{\mathbf{X}}^T\mathbf{Y}.$$

- *Analysis of Covariance*: this is an ANOVA model with continuous *concomitant variables*
  $$\mathbb{E}(\mathbf{Y}) = \mathbf{B}\mathbf{a} + \mathbf{D}\mathbf{c},$$

where $\mathbf{B}$ is $n \times k$ *design matrix* of 0-1 entries, $\mathbf{D}$ is $n \times \ell$ matrix of observations on the (continuous) concomitant variables; further, $\mathbf{a} \in \mathbb{R}^k$ and $\mathbf{c} \in \mathbb{R}^\ell$ are parameters to be estimated by minimizing

$$\sum_{i=1}^{n}(Y_i - b_{i1}a_1 - \cdots - b_{ik}a_k - d_{i1}c_1 - \cdots - d_{i\ell}c_\ell)^2.$$

The normal equations form the following system:

$$\mathbf{B}^T\mathbf{B}\mathbf{a} + \mathbf{B}^T\mathbf{D}\mathbf{c} = \mathbf{B}^T\mathbf{Y},$$
$$\mathbf{D}^T\mathbf{B}\mathbf{a} + \mathbf{D}^T\mathbf{D}\mathbf{c} = \mathbf{D}^T\mathbf{Y}.$$

If we can accept the zero hypothesis $\mathbf{c} = \mathbf{0}$, then we perform usual one-way ANOVA; otherwise, we eliminate the effect of the concomitant variables. See C. R. Rao: Linear statistical inference, page 291, where Y is the growth rate of pigs, the $k = 3$ factors are pen, sex, and type of food given, while the only concomitant variable is the initial weight.

Examples

The following examples are from the book
Gary Koop: Analysis of economic data, Wiley, 2005.

1. Example: Consider the following Multivariate Regression problem emerging in Electric Power Industry in the USA.

$$Y = \text{cost of production (million dollar/year)}$$
$$X_1 = \text{yield (kWh/year)}$$
$$X_2 = \text{cost of labour (dollar/year/worker)}$$
$$X_3 = \text{cost of capital (dollar/unit)}$$
$$X_4 = \text{cost of fuel (dollar/million BTU)}$$

Regression results:

	Coefficient	Standard error	t-value	p-value	Lower 95%	Upper 95%
Intercept	-70.49511	12.69501	-5.55298	1.76E-07	-95.6347	-45.3556
X_1	0.00474	0.00011	43.22597	3.41E-74	0.004514	0.004948
X_2	0.00363	0.00106	3.43660	0.000814	0.001537	0.005717
X_3	0.28008	0.12949	2.16301	0.032557	0.023663	0.536503
X_4	0.78346	0.16759	4.72566	6.391E-06	0.455154	1.11177

$R^2 = 0.94$, the p-value of the $H_0:\ R^2 = 0$ is $9.73E - 73$.

This shows that the regression is significant, and all the predictors are significant, except X_3. We can see that X_3 has relatively 'small' correlation with the other variables, where the correlation matrix of the predictors is

$$\begin{pmatrix} 1 & & & \\ 0.056399 & 1 & & \\ 0.021481 & -0.078686 & 1 & \\ 0.053507 & 0.318349 & 0.155224 & 1 \end{pmatrix}.$$

When we leave out X_3 from the regression, the results do not change significantly:

	Coefficient	Standard error	t-value	p-value	Lower 95%	Upper 95%
Intercept	-49.75804	8.44931	-5.88900	3.68E-08	-71.8765	-27.6396
X_1	0.00473	0.00011	42.6218	6.4E-74	0.004445	0.005027
X_2	0.00331	0.00006	3.12145	0.002259	0.000535	0.006091
X_4	0.851586	0.165266	5.15282	1.03E-06	0.418956	1.284216

$R^2 = 0.94$, the p-value of the $H_0 : R^2 = 0$ is $3.5E-73$.

If there were *collinearities* between the variables, we could not simply leave out one.

2. Example: the following Multivariate Regression problem is to predict the apartment prices in the USA.

$$\begin{aligned} Y &= \text{selling price} \\ X_1 &= \text{site area} \\ X_2 &= \text{number of bedrooms} \\ X_3 &= \text{number of bathrooms} \\ X_4 &= \text{number of levels} \end{aligned}$$

Regression results:

	Coefficient	Standard error	t-value	p-value	Lower 95%	Upper 95%
Intercept	-4009.5500	3603.109	-1.1128	0.266287	-11087.3	3068.248
X_1	5.42917	0.36925	14.70325	2.05E-41	4.703835	6.154513
X_2	2824.61379	1214.808	2.325153	0.020433	438.2961	5210.931
X_3	17105.1745	1734.434	9.862107	3.29E-21	13698.12	20512.22
X_4	7634.897	1007.974	7.574494	1.57E-13	5654.874	9614.92

$R^2 = 0.54$, the p-value of the $H_0 : R^2 = 0$ is $1.18E - 88$.

Therefore, the predicted price is

$$\hat{Y} = -4009.55 + 5.43X_1 + 2824.6X_2 + 17105.17X_3 + 7634.90X_4.$$

This means that in the apartments with the same number of bedrooms, bathrooms, and levels, the increase of the site with 1 $foot^2$ will result in the increase of the price with 5.43 dollar. Likewise, keeping the site are, number of bathrooms, and levels fixed, a 3-bedroom apartment costs with 2824.6 dollar more than a 2-bedroom one, on average, 'ceteris paribus'.

In fact, here 1-, 2-, or 3-way ANOVA can also be used with the discrete variables X_2, X_3, X_4, as well as an Analysis of Covariance with the site area as concomitant variable.