

Partial correlation

1. We are interested in the association between two variables x and y but want to control for other covariates, z_1, \dots, z_k .
2. Partial correlation is defined to be the Pearson correlation between two derived variables e_x and e_y where e_x = the residual from the linear regression of x on z_1, \dots, z_k and e_y = the residual from the linear regression of y on z_1, \dots, z_k .

Multiple correlation

The association between y , dependent variable, and all the predictors x_1, \dots, x_k . The multiple correlation is defined as the correlation between y and $b_1x_1 + \dots + b_kx_k$.

1 Multisample Inference

1. Comparing the means of more than two distributions/samples.
2. Whether or not passive smoking (exposure among nonsmokers to cigarette smokers in the atmosphere) has a measurable effect on pulmonary health.
3. Pulmonary function was measured in several ways in six groups
 - (a) Nonsmokers (NS): people who themselves did not smoke and were not exposed to cigarette smoke either at home or on the job.
 - (b) Passive smokers (PS): people who themselves did not smoke and were not exposed to cigarette smoke in the home but were employed for 20 or more years in an enclosed working area that routinely contained tobacco smoke.
 - (c) Noninhaling smokers (NI): people who smoked pipes, cigars, or cigarettes, but did not inhale.
 - (d) Light smokers (LS): people who smoked and inhaled 1-10 cigarettes per day for 20 or more years.
 - (e) Moderate smokers (MS): people who smoked and inhaled 11-39 cigarettes per day for 20 or more years.
 - (f) Heavy smokers (HS): people who smoked and inhaled 40 or more cigarettes per day for 20 or more years.

(g) Forced mid-expiratory flow (FEF) was used to assess pulmonary function

Table 12.1 FEF data for smoking and nonsmoking males

Group number, i	Group name	Mean FEF (L/s)	sd FEF (L/s)	n_i
1	NS	3.78	0.79	200
2	PS	3.30	0.77	200
3	NI	3.32	0.86	50
4	LS	3.23	0.78	200
5	MS	2.73	0.81	200
6	HS	2.59	0.82	200

Source: Reprinted by permission of the *New England Journal of Medicine*, 302(13), 720–723, 1980.

One-way ANOVA model

1. Suppose there are k groups with n_i observations in the i th group. The j th observation in the i th group will be denoted by y_{ij} . We can have the following model:

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

2. μ represents the underlying mean of all groups taken together.
3. α_i represents the difference between the mean of the i th group and the overall mean
4. ϵ_{ij} represents random error about the mean for an individual observation from the i th group

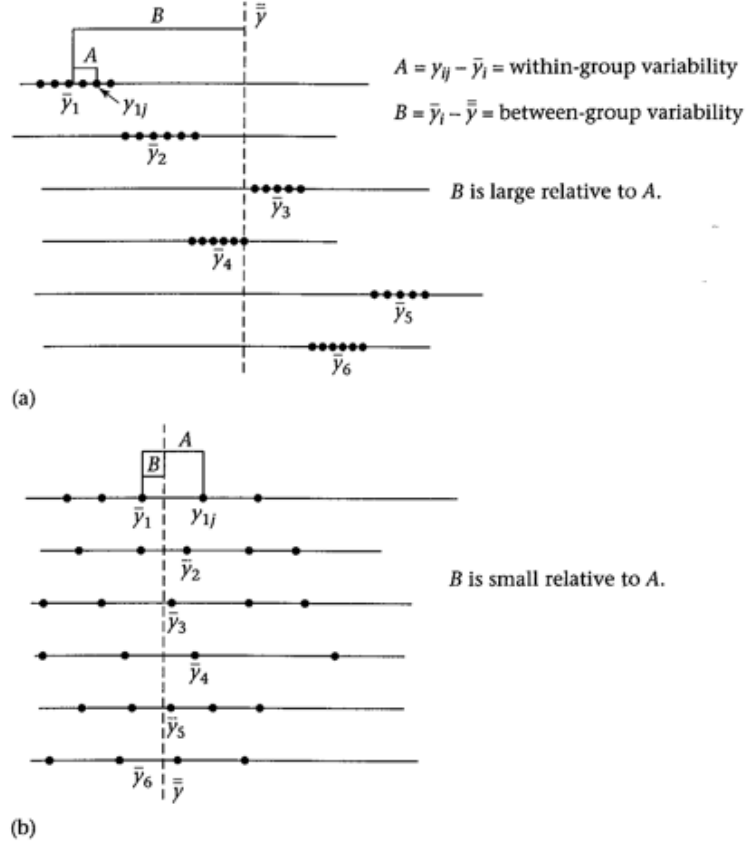
Overall comparison of group means

1. Hypothesis: H_0 : all $\alpha_i = 0$ versus H_1 : at least one $\alpha_i \neq 0$.
2. The mean FEF for the i th group is denoted by \bar{y}_i and the mean FEF of overall groups by $\bar{\bar{y}}$. The deviation of an individual observation from the overall mean can be represented by

$$y_{ij} - \bar{\bar{y}} = (y_{ij} - \bar{y}_i) + (\bar{y}_i - \bar{\bar{y}})$$

3. The first term on the right hand side: Deviation of an individual observation from the group mean for that observation (Within-group variability)
4. The second term on the right hand side: Deviation of a group mean from the overall mean (Between-group variability)

Figure 12.1 Comparison of between-group and within-group variability



5. Observe that

$$\sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{\bar{y}})^2 = \sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2 + \sum_{i=1}^k \sum_{j=1}^{n_i} (\bar{y}_i - \bar{\bar{y}})^2$$

6. $\sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{\bar{y}})^2$ is called the Total sum of squares (TSS)
7. $\sum_{i=1}^k \sum_{j=1}^{n_i} (\bar{y}_i - \bar{\bar{y}})^2$ is called The Between Sum of Squares (BSS)

8. $\sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2$ is called the Within Sum of Squares
9. Computational short form for BSS: $\sum_{i=1}^k n_i \bar{y}_i^2 - y_{..}^2/n$
10. Computational short form for WSS: $\sum_{i=1}^k (n_i - 1)s_i^2$ where $y_{..}$ = sum of observations across all the groups
11. Between mean Square = Between MS = $BSS/(k - 1)$
12. Within Mean Square = Within MS = $WSS/(n - k)$
13. The significance test will be based on the ratio of the Between MS to the Within MS.

Overall F Test for One-Way ANOVA To test the hypothesis $H_0: \alpha_i = 0$ for all i versus H_1 : at least one $\alpha_i \neq 0$, use the following procedure:

- (1) Compute the Between SS, Between MS, Within SS, and Within MS using Equation 12.5 and Definitions 12.5 and 12.6.
- (2) Compute the test statistic $F = \text{Between MS} / \text{Within MS}$, which follows an F distribution with $k - 1$ and $n - k$ df under H_0 .
- (3) If $F > F_{k-1, n-k, 1-\alpha}$ then reject H_0
If $F \leq F_{k-1, n-k, 1-\alpha}$ then accept H_0
- (4) The exact p -value is given by the area to the right of F under an $F_{k-1, n-k}$ distribution = $Pr(F_{k-1, n-k} > F)$.

Display of one-way ANOVA results

Source of variation	SS	df	MS	F statistic	p-value
Between	$\sum_{i=1}^k n_i \bar{y}_i^2 - \frac{y_{..}^2}{n} = A$	$k - 1$	$\frac{A}{k - 1}$	$\frac{A/(k - 1)}{B/(n - k)} = F$	$Pr(F_{k-1, n-k} > F)$
Within	$\sum_{i=1}^k (n_i - 1)s_i^2 = B$	$n - k$	$\frac{B}{n - k}$		
Total	Between SS + Within SS				

14. Test if the mean FEF scores differ significantly among the six groups in Table 12.1.
15. We can compute the between SS as 184.38 and within SS = 663.87. There are 1050 observations combined over all 6 groups. Between MS = $184.38/5 = 36.875$. Within MS = $663.87/(1050-6) = 663.87/1044 = 0.636$. $F = \text{Between MS}/\text{Within MS} = 36.875/0.636 = 58.0 \sim F_{5,1044}$ under null hypothesis. $F_{5,1044,.999} = qf(.999, 5, 1044) = 4.13 < 58.0 = F$. $p < 0.001$. At least two of the means are significantly different.

Comparison of Specific Groups in One-Way ANOVA

The previous test does not tell us which of the groups have means that differ from each other. The usual practice is to perform the F test first and if null hypothesis is rejected, then specific groups are compared. We want to test if groups 1 and 2 have means that are significantly different from each other.

\bar{Y}_1 is normally distributed with mean $\mu + \alpha_1$ and variance σ^2/n_1 and \bar{Y}_2 is normally distributed with mean $\mu + \alpha_2$ and variance σ^2/n_2 .

$$\begin{aligned}\bar{Y}_1 - \bar{Y}_2 &\sim N(\alpha_1 - \alpha_2, \sigma^2(1/n_1 + 1/n_2)) \\ z &= (\bar{Y}_1 - \bar{Y}_2) / \sqrt{\sigma^2(1/n_1 + 1/n_2)}\end{aligned}$$

Pooled Estimate of the Variance for One-Way ANOVA

$$s^2 = \sum_{i=1}^k (n_i - 1) s_i^2 / \sum_{i=1}^k (n_i - 1) = \left[\sum_{i=1}^k (n_i - 1) s_i^2 \right] / (n - k) = \text{Within MS}$$

t Test for the Comparison of Pairs of Groups in One-Way ANOVA (LSD Procedure)

Suppose we wish to compare two specific groups, arbitrarily labeled as group 1 and group 2, among k groups. To test the hypothesis $H_0: \alpha_1 = \alpha_2$ versus $H_1: \alpha_1 \neq \alpha_2$, use the following procedure:

- (1) Compute the pooled estimate of the variance $s^2 = \text{Within MS}$ from the one-way ANOVA.
- (2) Compute the test statistic

$$t = \frac{\bar{y}_1 - \bar{y}_2}{\sqrt{s^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}$$

which follows a t_{n-k} distribution under H_0 .

- (3) For a two-sided level α test,

$$\text{if } t > t_{n-k, 1-\alpha/2} \quad \text{or} \quad t < t_{n-k, \alpha/2}$$

then reject H_0

$$\text{if } t_{n-k, \alpha/2} \leq t \leq t_{n-k, 1-\alpha/2}$$

then accept H_0

Linear contrast

Compare the pulmonary function of the group of smokers who inhale cigarettes with the group of nonsmokers. The three groups of inhaling smokers could be combined to form one group of 600 inhaling smokers. However, these three groups were selected so as to be of the same size, whereas in the general population the proportions of light, moderate, and heavy smokers are not likely to be the same. Suppose large population surveys report that 70% of inhaling smokers are moderate smokers, 20% are heavy smokers, and 10% are light smokers. How can inhaling smokers as a group be compared with nonsmokers? A linear contrast (L) is any linear combination of the individual group means such that the linear coefficients add up to 0.

$$L = \sum_{i=1}^k c_i \bar{y}_i, \text{ where } \sum_{i=1}^k c_i = 0.$$

Linear contrast for the above example is $\bar{y}_1 - 0.1\bar{y}_4 - 0.7\bar{y}_5 - 0.2\bar{y}_6$.

Test the hypothesis that the underlying mean of the linear contrast defined in previous example is significantly different from 0. $s^2 = 0.636$. $L = \bar{y}_1 - 0.1\bar{y}_4 - 0.7\bar{y}_5 - 0.2\bar{y}_6 = 1.03$. $se(L) = \sqrt{s^2 \sum_{i=1}^k c_i^2 / n_i} = 0.070$. $t = L/se(L) = 1.03/0.070 = 14.69 \sim t_{1044}$. $qt(0.999, 1044) = 3.10$. $p < 0.001$. The linear contrast is very highly significant and the inhaling smokers as a group have much worse pulmonary function than the nonsmokers

$$L = \sum_{i=1}^k c_i \bar{y}_i \quad \text{where} \quad \sum_{i=1}^k c_i = 0 \quad \begin{array}{l} \text{Because } \text{Var}(\bar{y}_i) = s^2/n_i, \\ \text{Var}(L) = s^2 \sum_{i=1}^k c_i^2/n_i \end{array}$$

t Test for Linear Contrasts in One-Way ANOVA Suppose we want to test the hypothesis $H_0: \mu_L = 0$ versus $H_1: \mu_L \neq 0$, using a two-sided test with significance level $= \alpha$, where $y_{ij} \sim N(\mu + \alpha_i, \sigma^2)$, $\mu_L = \sum_{i=1}^k c_i \alpha_i$, and $\sum_{i=1}^k c_i = 0$.

- (1) Compute the pooled estimate of the variance $= s^2 = \text{Within MS}$ from the one-way ANOVA.
- (2) Compute the linear contrast

$$L = \sum_{i=1}^k c_i \bar{y}_i$$

- (3) Compute the test statistic

$$t = \frac{L}{\sqrt{s^2 \sum_{i=1}^k \frac{c_i^2}{n_i}}}$$

- (4) If $t > t_{n-k, 1-\alpha/2}$ or $t < t_{n-k, \alpha/2}$ then reject H_0 .
If $t_{n-k, \alpha/2} \leq t \leq t_{n-k, 1-\alpha/2}$ then accept H_0 .
- (5) The exact p -value is given by