Regression and ANOVA

Partial residual plot

A partial-residual plot characterizing the relationship between the dependent variable y and a specific independent variable x_i in a multiple-regression setting is constructed as follows:

1. A multiple regression is performed of y on all predictors other than x_i (i.e., $x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_k$), and the residuals are saved.
2. A multiple regression is performed of x_i on all other predictors (i.e., $x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_k$) and the residuals are saved.
3. The partial-residual plot is a scatter plot of the residuals in step 1 on the y-axis against the residual in step 2 on the x-axis.

The partial-residual plot reflects the relationship between y and x_i after each variable is adjusted for all other predictors in the multiple-regression analysis.

Partial correlation

1. We are interested in the association between two variables x and y but want to control for other covariates, $z_1, \ldots, z_k$.
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 $z_1, \ldots, z_k$ e_y = the residual from the linear regression of y on $z_1, \ldots, z_k$.

Multiple correlation

The association between y, dependent variable, and all the predictors $x_1, \ldots, x_k$. The multiple correlation is defined as the correlation between y and $b_1 x_1 + \ldots + b_k x_k$.

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 ith group. The jth observation in the ith 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 ith group and the overall mean

4. ϵ_{ij} represents random error about the mean for an individual observation from the ith 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)

5. Observe that

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

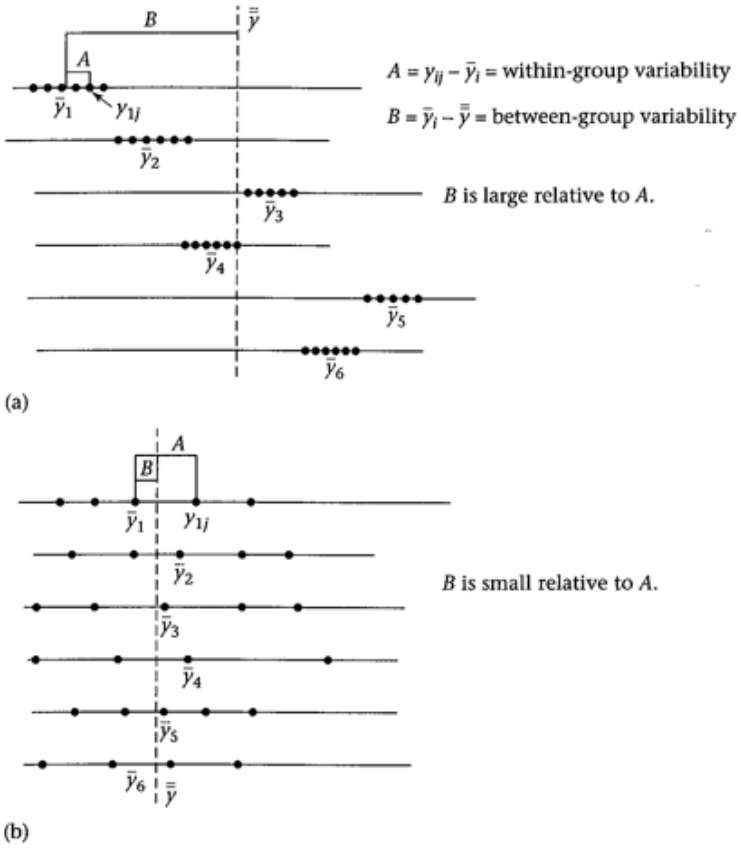
6. $\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{\bar{y}})^2$ is called the Total sum of squares (TSS)

7. $\sum_{i=1}^{k}\sum_{j=1}^{n_i}(\bar{y}_i - \bar{\bar{y}})^2$ is called The Beween 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: $\sum_{i=1}^{k} n_i \bar{y}_i^2 - y_{..}^2/n$

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



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 = Within $SS/(n-k)$

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 = Between MS/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				