LECTURE 10.3:

Multiple and partial correlations

4/27/2023

BRIEF REVIEW

Full multiple linear model

A full multiple liner model relates a single *outcome* variable $y$to multiple $(k > 1)$*predictor* variables $x_1, \ldots, x_k$ and an unobserved error e:

$$y \; = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + e,$$

where $e \sim N(0, \sigma^2)$ and $(\beta_0,\ \beta_1, \ldots,\ \beta_k,\ \sigma^2)$ are unknown parameters.

For a given data set $\{(y_i, x_i = (x_{i1}, \dots, x_{ik})): i = 1, \dots, n\}$, we fit the ***genuine data generating mechanism*** by aforesaid ***multiple linear model***. Thus far, we have learnt to:

1. estimate the unknown parameters by the method of Least Squares;

2. use $R^2 = \frac{\text{Reg SS}}{\text{Total SS}}$ to measure the strength of linear relationship;

3. perform F test for joint hypothesis H_0: $\beta_1 = \cdots = \beta_k = 0$ versus H_a: least one $\beta_j \neq 0$.

4. perform partial t test or partial F test for partial hypothesis H_{0j}: $\beta_j = 0$ versus H_{aj}: $\beta_j \neq 0$.

Note 1: In multiple linear regression, we often **assume** normality ($e \sim N(0, \sigma^2)$) , *linearity* ($\mathrm{E}(y|x_1, \dots, x_k) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$), homoscedasticity ($\sigma^2$ does not depends on x values) and *independence* (i.i.d. sample). In other words, real model is assumed to be identical to postulated model, up to unknown parameters.

Note 2: For given data points, we can compute the estimates even if all the assumptions are not true. In general, postulated statistical models are not identical to the data generating mechanism. The discrepancy between the real model and the linear model determines the performance of the estimators.

- **Note 3:** As illustrated, $t^2\left(x_j \middle| \text{other } x's\right) = F(x_j|\text{other } x's)$ for arbitrary specific data points $(n > k + 1)$, *no matter all the assumptions* are true.

- **Note 4:** If all the assumptions are true, then under the null, $t^2\left(x_j \middle| \text{other } x's\right) = F(x_j|\text{other } x's) \sim F_{1,n-k-1}$. Even without normality, the statistics may approximately follows $F_{1,n-k-1}$ for large sample size.

1. Multiple Correlation

Sometimes, we are interested in ***existence*** and ***strength*** *of the association* (linear relationship) *between one variable y and a set of variables* $(x_1, \dots, x_k)$ when considered as a group (**Fig. 1**) other than the effects themselves. The *multiple correlation* is a useful measure for such association.

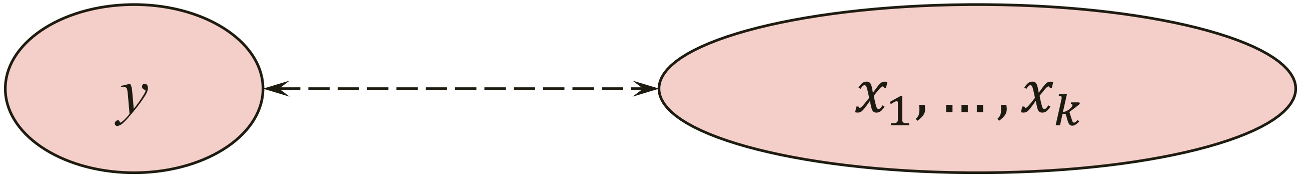

Figure 1: Existence and strength of the linear relationship between variable y and a set of variables $x_1, \dots, x_k$.

1.1. Definition

Definition 1.1: The ***coefficient of the multiple correlation*** (ϱ) between variable y and a set of variables $\{x_1, \dots, x_k\}$is defined as the maximum possible correlation coefficient between $y$and a linear combination of $\{x_1, \dots, x_k\}$.

Mathematically, we have

$$\varrho = \sqrt{c'\Omega^{-1}c},$$

where $c = (\rho_{y,x_1}, \dots, \rho_{y,x_k})'$ is the $k \times 1$ vector of *cross-correlations* and $\Omega = (\rho_{x_i,x_j})$ is the ***correlation matrix*** of ***inter-correlations*** between variable $x_1, \dots, x_k$.

Note 1: This population level definition allows for arbitrary variable y and an arbitrary set of variables $\{x_1, \dots, x_k\}$. It does not require normality of any of the variables, linearity relationship or even the dependence assumption.

A large ϱ value indicates a strong association (linear relationship) between y and the set of predictors. If $\varrho = 0$, then there is no linear relationship between y and $\{x_1, \dots, x_k\}$.

Note 2: For given data points $\{(y_i, x_i = (x_{i1}, \dots, x_{ik})): i = 1, \dots, n\}$, the multiple correlation coefficient *is estimated* by the Pearson correlation between $(y_1, \dots, y_n)$ and $(\hat{y}_1, \dots, \hat{y}_n)$, where $\hat{y}_i = \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_k x_{ik}$, and $\hat{\beta}_1, \dots, \hat{\beta}_k$ are the least-squares estimates of $\beta_1, \dots, \beta_k$. The ***sample multiple correlation coefficient*** equals to

$$\hat{\varrho} = \sqrt{R^2} = \sqrt{(\mathrm{RegSS})/(\mathrm{TotSS})}.$$

1.2. F test for multiple correlation

- The hypothesis is $H_0\colon \varrho = 0$ versus $H_a\colon \varrho \neq 0$.

- F statistic: $F = \frac{R^2}{1-R^2}\frac{n-k-1}{k}$.

- Null distribution:

  (i) If (y_i, x_i)'s are i.i.d. and (y_i, x_i) follows a joint normal distribution, then under the null, $F \sim F_{k,n-k-1}$, the centralized F distribution with $(k, n-k-1)$ degrees of freedom.

  (ii) If (y_i, x_i)'s are i.i.d. and the sample size is large, then under the null F approximately $F_{k,n-k-1}$.

1.2. Cont'd

- The exact p-value of the observed F value is given by $p = \Pr\left(F_{k,n-k-1} > F\right)$.

- Decision rule: Reject the H_0 at nominal level $\alpha \in (0,1)$ if $F > F_{k,n-k-1,1-\alpha}$ (equivalently, $p < \alpha$), where the critical value $F_{k,n-k-1,1-\alpha,}$ is the $1-\alpha$ percentile of $F_{k,n-k-1}$.

Note 3: The above F statistic is mathematically the F statistic for the joint hypothesis $H_0\colon \beta_1 = \cdots = \beta_k = 0$ versus $H_a\colon$ least one $\beta_j \neq 0$.

Prove it by the definitions of RegSS, TotSS, R^2 and F statistics.

Example 1 (Hypertension, Pediatrics): Estimate the multiple correlation between SBP and the predictors (weight, age) based on the pediatric blood pressure data in **Table 1**.

Table 1: Sample data for infant birth weight, age and blood pressure for 16 infants.

i	Weight (oz) (x_1)	Age (days) (x_2)	SBP (mm HG) (y)
1	135	3	89
2	120	4	90
3	100	3	83
4	105	2	77
5	130	4	92
6	125	5	98
7	125	2	82
8	105	3	85
9	120	5	96
10	90	4	95
11	120	2	80
12	95	3	79
13	120	3	86
14	150	4	97
15	160	3	92
16	125	3	88

Solution: Refer to **Example 1**. The R^2 for the regression model is

$$R^2 = \frac{\text{Reg SS}}{\text{Res SS}} = \frac{591.03564}{670.93750} = 0.8809.$$

Hence, the multiple correlation is estimated as $\sqrt{0.8809} = 0.94$.

This indicates a strong association between y and the set of predictors {weight, age}.

Since $n = 16$, $k = 2$ and $R^2 = 0.8809$, the realized F statistic is

$$F = \frac{R^2}{1-R^2}\frac{n-k-1}{k} = \frac{0.8809}{1-0.8809}\frac{16-2-1}{2} = 48.07599,$$

which is identical to the F value in **Example 1** for the hypothesis on the regression coefficients.

Since the realized $F = 48.08 > F_{2,13,0.95}$ (the p-value of the F value $< 0.0001 < 0.05$), we have firm evidence to reject $H_0: \varrho = 0$. This analysis suggests that the association is significant.

2. Partial Correlation

It is important to assess the degree of association between two variables (y, x_j) after controlling for other variables $\{x_1, \dots x_{j-1}, x_{j+1}, x_k\}$.

In context, I rewrite x_j as x, and rewrite variables $\{x_1, \dots x_{j-1}, x_{j+1}, x_k\}$ as $\boldsymbol{z} = (z_1, \dots, z_{k-1})$. The partial correlation (**Figure 8.2**) accomplish this goal.

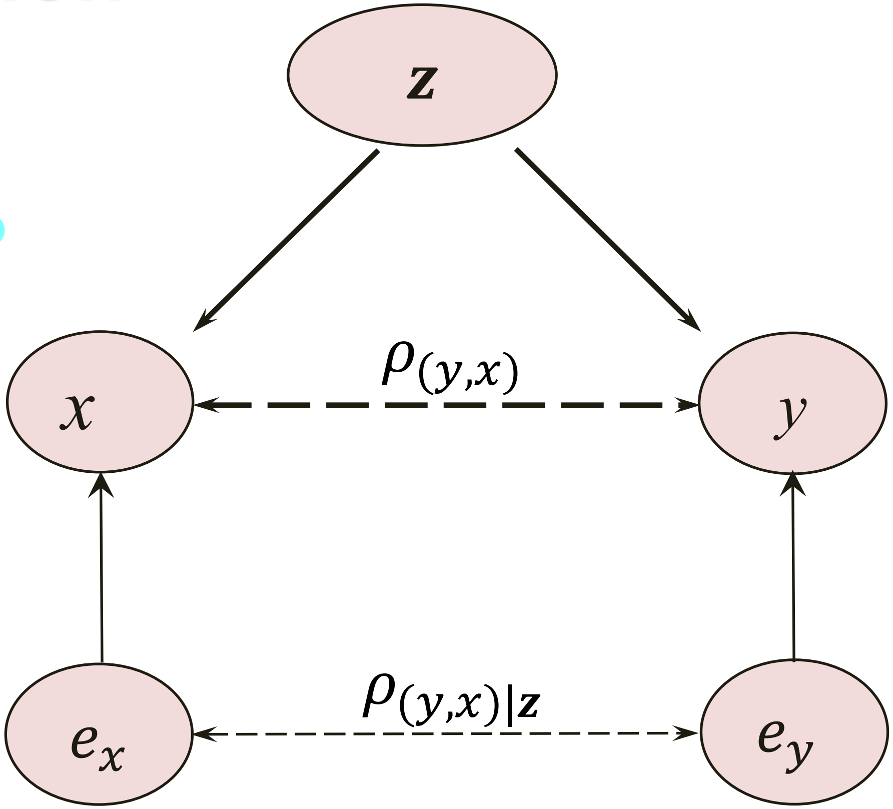

Figure 2: Correlation and partial correlation

2.1. Definition

Definition 2: The partial correlation coefficient two variables (y, x) after controlling for $\boldsymbol{z}$ is defined as $\rho_{(y,x)|\boldsymbol{z}}$, the Pearson correlation between two residual variables e_x and e_y in the multiple linear regressions $x = \gamma_0 + \gamma_1 z_1 + \cdots + \gamma_k z_{k-1} + e_x$ and $y = \gamma_0 + \gamma_1 z_1 + \cdots + \gamma_k z_{k-1} + e_y$.

Note 4: Partial Correlation Coefficient is a measure of the strength of the association between two variables (y, x) after controlling for the effects of other variables $\boldsymbol{z}$.

Note 6: In particular, for three random variables (y, x_1, x_2), the Partial Correlation Coefficient between y and x_1 is define as

$$\rho_{(y,x_1)|x_2} = \frac{\rho_{y,x_1} - \rho_{y,x_2}\rho_{x_1,x_2}}{\sqrt{\left(1 - \rho_{y,x_2}^2\right)\left(1 - \rho_{x_1,x_2}^2\right)}},$$

where the notation $\rho_{\cdot,\cdot}$ denotes the Pearson Moment Product Correlation Coefficient between two random variables.

The Sample Partial Correlation Coefficient between y and x_1 is define as

$$r_{(y,x_1)|x_2} = \frac{r_{y,x_1} - r_{y,x_2} r_{x_1,x_2}}{\sqrt{\left(1 - r_{y,x_2}^2\right)\left(1 - r_{x_1,x_2}^2\right)}},$$

where the notation $r_{\cdot,\cdot}$ denotes the Sample Pearson Moment Product Correlation Coefficient between two random variables.

Definition 3: The ***partial coefficient of determination*** of x after controlling for $\boldsymbol{z}$ is defined as the proportion that the variability in y not explained by $\boldsymbol{z}$ is now explained by x:

$$R^2_{(y,x)|\boldsymbol{z}} = \frac{\mathrm{RegSS}(x|\boldsymbol{z})}{\mathrm{ResSS}(\boldsymbol{z})}.$$

Note 5: For a given data set, $r^2_{(y,x)|\boldsymbol{z}} = R^2_{(y,x)|\boldsymbol{z}}$ (See Example 2).

2.2. HYPOTHESIS TESTING FOR PARTIAL CORRELATION

We want to know whether an individual variable x is significantly associated with y after controlling for the other variables $\boldsymbol{z}$.

The hypothesis is H_0: $\rho_{(y,x)|\boldsymbol{z}} = 0$ versus H_a: $\rho_{(y,x)|\boldsymbol{z}} \neq 0$.

Example 2 (Hypertension, Pediatrics): Estimate the partial correlation between SBP and the and weight based on the pediatric blood pressure data in **Table 1**.

Solution:y =SBP, x_1 = WEIGHT, x_2 = AGE. The cross- and inter-correlations are

$(r_{y,x_1} = 0.44109, r_{y,x_2} = 0.87084)$ and $r_{x_1,x_2} = 0.10683$.

Hence, $r^2_{y,x_2} = 0.758362, r^2_{x_1,x_2} = 0.011413$. The Partial Correlation of SBP and WEIGHT (controlling for AGE) is given by

$$r_{(y,x_1)|x_2} = \frac{r_{y,x_1} - r_{y,x_2} r_{x_1,x_2}}{\sqrt{\left(1 - r_{y,x_2}^2\right)\left(1 - r_{x_1,x_2}^2\right)}}$$

$$= \frac{.44109 - (.87084)(.10683)}{\sqrt{(1 - .758362)(1 - .011413)}}$$

$$= 0.71213.$$

From R outputs:

$\text{RegSS}(x_1, x_2)$ =591.03564
$\text{RegSS}(x_2) = 508.8166$
$\text{ResSS}(x_2) = 162.12093$
$\text{RegSS}(x_1|x_2) = \text{Reg SS}(x_1, x_2) - \text{Reg SS}(x_2)$
$= 591.03564 - 508.8166 = 82.2191.$

$$R^2_{(y,x_1)|x_2} = \frac{\text{RegSS}(x_1|x_2)}{\text{ResSS}(x_2)} = 0.507147.$$

That is, 50.71% of the variability in SBP not explained by AGE is now explained by WEIGHT. I obtain that $r^2_{(y,x_1)|x_2} = 0.5071291 = R^2_{(y,x_1)|x_2}$ up to round error.

Practice

1. Download R code and play with the SBP data to estimate and test for multiple and partial correlations

2. Analyze the FEV data to estimate and test for multiple and partial correlations (ignoring gender and smoke data at this moment).