DESCRIPTIVE STATISTICS

Dr Alina Gleska

Institute of Mathematics, PUT

May 27, 2018

The power model

$$y = \alpha x^{\beta} e$$

can be transformed into the linear one by taking logarithms

$$ln(y) = ln(\alpha) + \beta ln(x) + lne,$$

so it can be written as

$$y^* = \alpha^* + \beta x^* + e^*.$$

We have obtained a linear regression model for the logarithms of the variable Y with respect to the logarithms of the variable X.

After using the least squares method we obtain the estimates of the parameters α^* and β:

$$\alpha = \exp(\alpha^*), \quad \beta = \beta^*.$$

The parameter α defines the theoretical level of the variable Y for the unit value of X, and the parameter β defines the percentage change of the variable Y when the variable X changes by 1%.

The exponential model

$$y = \alpha\beta^x e$$

can be transformed into the linear one by taking logarithms

$$ln(y) = ln(\alpha) + xln(\beta) + lne,$$

so it can be written as

$$y^* = \alpha^* + \beta^* x + e^*.$$

We have obtained a linear regression model for the logarithms of the variable Y with respect to the value of the variable X.

After using the least squares method we obtain the estimates of the parameters α^* i β^*:

$$\alpha = \exp(\alpha^*), \quad \beta = \exp(\beta^*).$$

The parameter α defines the theoretical level of the variable Y in the period just before the first tested period (e.g. for $X = 0$), and β can be used for the definition of the medium-period tempo of change of the variable Y: $r_g = (\beta - 1) \cdot 100\%$. The quantity r_g is the constant rate of change and defines the relative increase of the variable Y according to constant time unit.

EXAMPLE

The Pearson's correlation indices are used to measure the strength of the quantity variables, regardless of the shape of the relationship between them.
How to find the Pearson's correlation indices:

- group the values of the variables X and Y in the table with dimensions $r \times k$ (r - the number of rows, k - the number of columns),
- calculate the total variance of the variable X and/or the total variance of the variable Y using formulas:

$$S^2(X) = \frac{1}{n}\sum_{i=1}^{r}(x_i^0)^2 n_{i\bullet} - \overline{x}^2, \quad S^2(Y) = \frac{1}{n}\sum_{j=1}^{k}(y_j^0)^2 n_{\bullet j} - \overline{y}^2,$$

- calculate the intergroup variance for the variable X and/or for the variable Y by formulas:

$$S_m^2(X) = \frac{1}{n}\sum_{j=1}^{k}(\overline{x_j} - \overline{x})^2 n_{\bullet j}, \quad S_m^2(Y) = \frac{1}{n}\sum_{i=1}^{r}(\overline{y_i} - \overline{y})^2 n_{i\bullet},$$

where $\overline{x}, \overline{y}$ are the arithmetic means calculated by formulas:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{r} x_i^0 n_{i\bullet}, \quad \overline{y} = \frac{1}{n}\sum_{j=1}^{k} y_j^0 n_{\bullet j},$$

and $\overline{x}_j$ is the j-th arithmetic mean for the variable X provided $Y = y_j$:

$$\overline{x}_j = \frac{1}{n_{\bullet j}} \sum_{i=1}^{r} x_i^0 n_{ij} \quad \text{for} \quad j = 1, 2, \ldots, k,$$

and $\overline{y}_i$ is the i-th arithmetic mean for the variable Y provided $X = x_i$:

$$\overline{y}_i = \frac{1}{n_{i\bullet}} \sum_{j=1}^{k} y_j^0 n_{ij} \quad \text{for} \quad i = 1, 2, \ldots, r.$$

The Pearson's correlation indices are given by formulas:

- for the variable X with respect to the variable Y:

$$r_{P(X,Y)} = \sqrt{\frac{S_m^2(X)}{S^2(X)}},$$

- for the variable Y with respect to the variable X:

$$r_{P(Y,X)} = \sqrt{\frac{S_m^2(Y)}{S^2(Y)}}.$$

The Pearson's correlation indices $r_{P(X,Y)}$ and $r_{P(Y,X)}$ belong to the interval $[0,1]$ and say nothing about the direction of the correlation. They are equal to 0 if there is no correlation between variables; they are equal to 1, if there is a functional relationship between variables.
They are mostly asymmetric, e.g. $r_{P(X,Y)} \neq r_{P(Y,X)}$. Moreover we have the inequalities:

$$r^2_{P(X,Y)} > r^2 \quad \text{and} \quad r^2_{P(Y,X)} > r^2,$$

where r is the Pearson's linear correlation coefficient.

The measures of nonlinearity are used for checking the level of the nonlinearity of a relationship:

- the variable X with respect to the variable Y: $M_{K(X,Y)} = r^2_{P(X,Y)} - r^2$,
- the variable Y with respect to the variable X: $M_{K(Y,X)} = r^2_{P(Y,X)} - r^2$.

The measures of nonlinearity also belong to the interval $[0, 1]$; if their values are less or equal to 0.2 we have a linear relationship, and if they are bigger than 0.2 we have a nonlinear relationship.

We use the coefficients of the partial correlation and the multiple correlation when we want to measure the strength of the relationship between many variables X_j ($j = 1, 2, \ldots, m$). We have two typical situations:

- we want to estimate the strength of the correlation between any two variables X_{j_1} and X_{j_2} ($1 \leq j_1 \neq j_2 \leq m$), without any influence of other variables (the partial correlation),
- we want to estimate the strength of the correlation between the concrete variable X_{j_1} and the rest of variables (the multiple correlation).

We calculate the coefficients of the partial correlation using the formula:

$$r_{jl \bullet \Omega} = \frac{-\tilde{R}_{jl}}{\sqrt{\tilde{R}_{jj}\tilde{R}_{ll}}},$$

where the two first letters jl at the index $jl \bullet \Omega$ mean, that we calculate the coefficients of the partial correlation for the variables X_j and X_l without the influence of the variables with indices from the set
$\Omega = \{1, 2, \ldots, j-1, j+1, \ldots, l-1, l+1, \ldots, m\}$, $\tilde{R}_{jl}$, $\tilde{R}_{jj}$, $\tilde{R}_{ll}$ are cofactors of the correlation matrix R with dimensions $m \times m$.
The correlation matrix R we create using the Pearson's linear correlation coefficients for each pair of variables (separately).

The correlation matrix R is symmetric, with $r_{jj} = 1$ lying on its principal diagonal. So:

$$R = \begin{bmatrix} 1 & r_{12} & r_{13} & \dots & r_{1m} \\ r_{21} & 1 & r_{23} & \dots & r_{2m} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ r_{m1} & r_{m2} & r_{m3} & \dots & 1 \end{bmatrix}$$

The minor $det(R_{jl})$ of the correlation matrix R is the determinant of a square matrix obtained from R by removing the j-th row and the l-th column. The cofactor $\tilde{R}_{jl}$ of the correlation matrix R we define as follows:

$$\tilde{R}_{jl} = (-1)^{j+l}\det(R_{jl}) =$$

$$= (-1)^{j+l}\det\begin{bmatrix} 1 & \cdots & r_{1,l-1} & r_{1,l+1} & \cdots & r_{1m} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ r_{j-1,1} & \cdots & r_{j-1,l-1} & r_{j-1,l+1} & \cdots & r_{j-1,m} \\ r_{j+1,1} & \cdots & r_{j+1,l-1} & r_{j+1,l+1} & \cdots & r_{j+1,m} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ r_{m1} & \cdots & r_{m,l-1} & r_{m,l+1} & \cdots & 1 \end{bmatrix}$$

The cofactor $\tilde{R}_{jj}$ is defined by the minor $det(R_{jj})$ obtained from the correlation matrix R by removing the j-th row and j-th column, and the cofactor $\tilde{R}_{ll}$ – by removing the l-th row and the l-th column.

The coefficients of the partial correlation $r_{jl \bullet \Omega}$ belong to the interval $[-1, 1]$, and their values can be greater or smaller than the value of the total correlation coefficient. Moreover, they can have different signs. The interpretation of the coefficients of the partial correlation is the same as for the Pearson's linear correlation coefficient (provided we exclude the influence of the other variables).

The coefficient of the multiple correlation is defined by:

$$R_{j\bullet\Omega} = \sqrt{1 - \frac{\det(R)}{\det(R_{jj})}},$$

where j at the index $j \bullet \Omega$ means, that we calculate the coefficient of the multiple correlation between the variable X_j and the set of variables indexed by the elements of the set $\Omega = \{1, 2, \ldots, j-1, j+1, \ldots, m\}$, and R_{jj} is the cofactor of the correlation matrix R.

The coefficient of the multiple correlation $R_{j\bullet\Omega}$ measures the strength of the correlation and belongs to the interval $[0,1]$. The closer $R_{j\bullet\Omega}$ to 1, the stronger relationship between the variable X_j and the set of the rest variables. If this value is close to 0, there is no relationship between them .
EXAMPLE