Multiple Linear Regression

What you are going to learn

- Concept of Multiple Linear Regression
- Special Cases
- Properties of the Residual
- Multiple Linear Regression in Matrix Notation
- Identification of the Coefficients
- The Coefficient of Determination
- Multiple Linear Quasi-Regression
- Statistical Models of Multiple Linear Regressions
- The General Linear Model

Multiple Linear Regression: Definition

Let Y and $X_1, ..., X_m$ be numerical random variables on a common probability space with finite expected values, positive, finite variances and a non-singular covariance matrix S_{xx}. The regression $E(Y| X_1, ..., X_m)$ is called *linear in* $(X_1, ..., X_m)$, if

$$E(Y| X_1, ..., X_m) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_m X_m.$$

Multiple Linear Regression: Special Case I

Note that in the definition presented on the last transparency it is not assumed that the regressors $X_1, ..., X_m$ are defined independently of each other. This is only necessary if we consider the conditional regressions $E_{X_2=x_2, ..., X_m=x_m}(Y|\, X_1)$.

We have already seen in Chap. 9, that a simple quadratic regression $E(Y|\, X)$ is linear in (X, X^2), if

$$E(Y \mid X) = \beta_0 + \beta_1 X + \beta_2 X^2.$$

Multiple Linear Regression: Special Case II

A second special case of the multiple linear regression is the following:

$$E(Y \mid X_1, X_2) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 \cdot X_2.$$

Note that the regression $E(Y|\, X_1, X_2)$ is linear in $(X_1, X_2, X_1 \cdot X_2)$ and that

$$E(Y \mid X_1, X_2) = E(Y \mid X_1, X_2, X_1 \cdot X_2).$$

Multiple Linear Regression: Special Case III

If the regressor X has n values $x_1, \ldots, x_n$, the cell means model

$$E(Y \mid X) = \beta_0 + \beta_1 \cdot I_1 + \beta_2 \cdot I_2 + \ldots + \beta_n \cdot I_n, \quad \text{with } \beta_0 = 0,$$

is also a special case of a multiple linear regression. The regression $E(Y \mid X)$ is always linear in $(I_1, \ldots, I_n)$, even if the regression $E(Y \mid X)$ is not linear in (X).

The Properties of the Residual

The residual is defined by

$$e := Y - E(Y \mid X_1, \ldots, X_m)$$

It has the following properties

$$E(e) = 0,$$

$$E(e \mid X_1, \ldots, X_m) = 0,$$

$$Cov[e, f(X_1, \ldots, X_m)] = 0,$$

$$Cov(e, X_i) = 0 \quad \text{for } i = 1, \ldots, m,$$

where $f(X_1, \ldots, X_m)$ can be any numerical function of the regressors.

Matrix Notation I

If you gather all regressors in one row vector $\boldsymbol{x}' = (X_1 \ldots X_m)$ and the regression coefficients $\beta_1, \ldots, \beta_m$ in an m-dimensional column vector $\mathrm{b} = (\beta_1 \ldots \beta_m)'$, then you can write the multiple linear regression in matrix notation:

$$E(\boldsymbol{y} \mid \boldsymbol{x}) = \beta_0 + \boldsymbol{x}' \mathrm{b} = \beta_0 + (X_1 \ldots X_m) \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_m \end{pmatrix}.$$

The regressand $\boldsymbol{y} = (Y)$ is a vector with only one component, Y. Therefore β_0 is a real number.

Matrix Notation II

If we define the row vector $\boldsymbol{z}' := (1 \ X_1 \ldots X_m)$ and the column vector $\mathrm{g} := (\beta_0 \ \beta_1 \ldots \beta_m)'$, the last equation simplifies to:

$$E(\boldsymbol{y} \mid \boldsymbol{x}) = E(\boldsymbol{y} \mid \boldsymbol{z}) = \boldsymbol{z}' \mathrm{g} = (1 \ X_1 \ldots X_m) \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_m \end{pmatrix}.$$

Identification of the Coefficients I

In order to identify β_0 and the components of $b = (\beta_1 \ldots \beta_m)'$ we need the expected values of the regressand and the regressors as well as the covariance matrices S_{xx} and S_{xy}. The identification equation of the constant of the regression β_0 is:

$$\beta_0 = E(\boldsymbol{y}) - E(\boldsymbol{x})' b = E(Y) - [E(X_1) \ \ldots E(X_m)] \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_m \end{pmatrix}.$$

Identification of the Coefficients II

Pre-multiplying both sides of the following equation

$$S_{\boldsymbol{xy}} = Cov(\boldsymbol{x}, \boldsymbol{y}) = Cov(\boldsymbol{x}, \beta_0 + \boldsymbol{x}' b + e) = Cov(\boldsymbol{x}, \beta_0 + b'\boldsymbol{x} + e)$$

$$= Cov(\boldsymbol{x}, \boldsymbol{x})\, b = S_{\boldsymbol{xx}}\, b$$

by the inverse $S_{\boldsymbol{xx}}^{-1}$ yields

$$S_{\boldsymbol{xx}}^{-1} S_{\boldsymbol{xx}}\, b = S_{\boldsymbol{xx}}^{-1} S_{\boldsymbol{xy}}.$$

Because $S_{\boldsymbol{xx}}^{-1} S_{\boldsymbol{xx}} = \mathbf{I}$ is the identity matrix, this simplifies to

$$b = S_{\boldsymbol{xx}}^{-1} S_{\boldsymbol{xy}},$$

which is the general formula for the identification of the coefficients of the regression.

Identification of the Coefficients: Two Regressors I

Let's say you have two regressors X_1 and X_2 and apply these formulas. This leads to the same identification equations given in Chapter 9:

$$\beta_0 = E(Y) - E(\boldsymbol{x})' \mathrm{b} = E(Y) - [\beta_1 E(X_1) + \beta_2 E(X_2)] = E(Y) - \beta_1 E(X_1) - \beta_2 E(X_2).$$

The covariance matrix and the variance covariance matrix are

$$\mathrm{S}_{\boldsymbol{xx}} = \begin{pmatrix} Var(X_1) & Cov(X_1, X_2) \\ Cov(X_1, X_2) & Var(X_2) \end{pmatrix} \text{ and } \mathrm{S}_{\boldsymbol{xy}} = \begin{pmatrix} Cov(X_1, Y) \\ Cov(X_2, Y) \end{pmatrix}.$$

Cramer's Rule as given in the last chapter yields

$$\mathrm{S}_{\boldsymbol{xx}}^{-1} = \frac{1}{Var(X_1)Var(X_2) - Cov(X_1, X_2)^2} \begin{pmatrix} Var(X_2) & -Cov(X_1, X_2) \\ -Cov(X_1, X_2) & Var(X_1) \end{pmatrix}.$$

Identification of the Coefficients: Two Regressors II

Multiplying $\mathrm{S}_{\boldsymbol{xx}}^{-1}$ with $\mathrm{S}_{\boldsymbol{xy}}$ yields

$$\beta_1 = \frac{Var(X_2)\ Cov(X_1, Y) - Cov(X_2, Y)\ Cov(X_1, X_2)}{Var(X_1)\ Var(X_2) - Cov(X_1, X_2)^2},$$

$$\beta_2 = \frac{Var(X_1)\ Cov(X_2, Y) - Cov(X_1, Y)\ Cov(X_1, X_2)}{Var(X_1)\ Var(X_2) - Cov(X_1, X_2)^2}.$$

Coefficient of Determination

The variance $Var[E(\boldsymbol{y}|\boldsymbol{x})]$ of the regression is:

$$Var[E(\boldsymbol{y}|\boldsymbol{x})] = Var(\beta_0 + \boldsymbol{x}'\mathrm{b}) = Var(\boldsymbol{x}'\mathrm{b}) = Var(\mathrm{b}'\boldsymbol{x}) = = \mathrm{b}'\,Var(\boldsymbol{x})\,\mathrm{b} =$$

$$= \mathrm{b}'\,\mathrm{S}_{\boldsymbol{xx}}\,\mathrm{b},$$

where $Var(\boldsymbol{x}) = \mathrm{S}_{\boldsymbol{xx}}$ is the $(m \times m)$ variance-covariance matrix of the regressors $X_1, ..., X_m$ and b is the m-dimensional column vector of the regression coefficients $\beta_1, ..., \beta_m$. Hence, the multiple coefficient of determination is:

$$R^2_{Y|X_1,...,X_m} = [\mathrm{b}'\,\mathrm{S}_{\boldsymbol{xx}}\,\mathrm{b}] \,/\, Var(Y).$$

Coefficient of Determination: Special Cases

If you have only two regressors X_1 and X_2, the last equation yields

$$R^2_{Y|X_1,X_2} = [ß_1^2 Var(X_1) + ß_2^2 Var(X_2) + 2\,ß_1\,ß_2 Cov(X_1, X_2)] \;\; / \; Var(Y),$$

which is identical to the equation of Chapter 9.

If all regressors $X_1, ..., X_m$ are *pairwise uncorrelated*, then the equation simplifies to

$$R^2_{Y|X_1,...,X_m} = \left(\sum_{i=1}^{m} ß_i^2\, Var(X_i)\right)\frac{1}{Var(Y)}.$$

Multiple Linear Quasi-Regression: Definition

Definition 14.2. Given the assumptions of Definition 14.1, we define the *multiple lineare Quasi-Regression*, denoted $Q(Y \mid X_1, ..., X_m)$ or $Q(\boldsymbol{y} \mid \boldsymbol{x})$, as that linear combination $\beta_0 + \beta_1 X_1 + ... + \beta_m X_m = \beta_0 + \mathrm{b}' \boldsymbol{x}$ of the components of $\boldsymbol{x} = (X_1 \ldots X_m)'$, for which

$$Y = \beta_0 + \beta_1 X_1 + ... + \beta_m X_m + \boldsymbol{n}$$

holds with

$$E(\boldsymbol{n}) = 0,$$

and

$$Cov(\boldsymbol{n}, X_1) = ... = Cov(\boldsymbol{n}, X_m) = 0.$$

Multiple Linear Quasi-Regression: Alternative Definition

Definition 14.3. Given the same assumptions as before, we can also define $Q(\boldsymbol{y} \mid \boldsymbol{x})$, as the linear combination $\beta_0 + \mathrm{b}' \boldsymbol{x}$, that minimizes the following function of b_0 und $\boldsymbol{b}$, the *least squares criterion*

$$LS(b_0, \boldsymbol{b}) = E[[Y - (b_0 + \boldsymbol{x}' \boldsymbol{b})]^2].$$

The number b_0 and the vector $\boldsymbol{b}$, for which the function $LS(b_0, \boldsymbol{b})$ is minimal, are denoted with β_0 and b, respectively. The multiple linear quasi-regression is then defined by:

$$Q(\boldsymbol{y} \mid \boldsymbol{x}) = \beta_0 + \boldsymbol{x}' \mathrm{b}.$$

The Coefficient of Determination of the Multiple Linear Quasi-Regression

The identification formulas for the coefficients of the multiple linear quasi-regression are identical to the ones of the (real) multiple linear regression.

The identification for the coefficient of determination of the multiple linear quasi-regression is also identical to the one of the multiple linear regression. It is

$$Q^2_{Y|X_1,\ldots,X_m} := Var[Q(\boldsymbol{y}\,|\,\boldsymbol{x})] \,/\, Var(Y) = [\mathrm{b}^{\complement}\, S_{\boldsymbol{xx}}\, \mathrm{b}] \,/\, Var(Y).$$

Statistical Models for the Multiple Linear Regression

So far we have always considered a single unit trial: Drawing an observational unit u out of a population and registering the values of the regressand and the regressors. Statistical models, however, refer to N random experiments, in which information about the parameters to be estimated is collected.

Models With Stochastic Regressors

Models with stochastic regressors consider the case, in which the single unit trial is repeated N-times. This yields N vectors $(Y_i\ X_{i1}\ \ldots\ X_{im})$, $i = 1, \ldots, N$, each representing the outcome of the random experiment i. Different assumptions about the distribution of these vectors can be made, e.g. that the vectors $(Y_i\ X_{i1}\ \ldots\ X_{im})$ are independent and each one is $(m + 1)$-*variate normally distributed*.

Models With Fixed Regressors

Other models estimate, *within* each of the combinations $x_1, ..., x_m$ of the regressors $X_1, ..., X_m$, the expected values $E(Y \mid X_1 = x_1, ..., X_m = x_m)$ of Y. Hence, within each of these combinations the regressand Y is observed many times. The values $x_1, ..., x_m$ are not chosen at random but are considered to be fixed quantities, that characterize the design of the experiment. This is why these models are called models with *fixed* or *non-stochastic* regressors. The most important of these models with fixed regressors is the General Linear Model (GLM).

The General Linear Model I

The GLM is defined by the following assumptions:

$$\boldsymbol{y} = \mathbf{X}\,\mathbf{ß} + \boldsymbol{e}$$

$$\boldsymbol{e} \sim \mathcal{N}(\mathbf{0}, \mathrm{s}^2\,\mathbf{I}),$$

where $\boldsymbol{y} = (Y_1 \ldots Y_i \ldots Y_N)'$ denotes the column vector of the "dependent" variables of a sample with size N. The *design matrix* $\mathbf{X}$ consists of $N \times (m + 1)$ fixed numbers. Every row of $\mathbf{X}$ consists of the vectors $\mathbf{x}_i' := (1\ x_{i1} \ldots x_{im})$, which are the combinations of values of the regressors, within which the Y_i is observed. The constant 1 in the first position of $\mathbf{x}_i'$ implies that the regression constant $ß_0$ is the first component of $\mathbf{ß} = (ß_0\ ß_1 \ldots ß_m)'$. The vector $\boldsymbol{e} = (e_1 \ldots e_i \ldots e_N)'$ consists of the residuals $e_i := Y_i - (\mathbf{x}_i'\,\mathbf{ß})$.

The General Linear Model II

The second assumption $\boldsymbol{e} \sim \mathcal{N}(\mathbf{0}, \mathrm{s}^2\,\mathbf{I})$ implies that $\boldsymbol{e}$ has a multivariate normal distribution with the vector of expected values $E(\boldsymbol{e}) = \mathbf{0}$ and the $N \times N$ covariance matrix $\mathbf{S}_{\boldsymbol{ee}} = \mathrm{s}^2\,\mathbf{I}$. That is, the residuals e_i are uncorrelated and have all the same variances. The latter is called the assumption of homoscedasticity.

Implications of these assumptions are

$$E(\boldsymbol{y}) = \mathbf{X}\,\mathbf{ß},$$

and

$$\mathbf{S}_{\boldsymbol{yy}} = \mathrm{s}^2\,\mathbf{I}.$$

The General Linear Model: Identification

If $\mathbf{X}'\mathbf{X}$ is non-singular then the vector of the regression coefficients can be estimated by:

$$\hat{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{y}.$$

This formula is derived by minimizing the least-square criterion

$$LS(\boldsymbol{b}) = (\boldsymbol{y} - \mathbf{X}\boldsymbol{b})'(\boldsymbol{y} - \mathbf{X}\boldsymbol{b}).$$

Of interest is also the covariance matrix

$$S_{\hat{b}\hat{b}} = s^2(\mathbf{X}'\mathbf{X})^{-1}$$

of these estimates. The square roots of the components on the main diagonal of this matrix are the standard errors of the regression coefficients.

Estimating the Coefficient of Determination

The formula to estimate the coefficient of determination is

$$\hat{R}^2 = \frac{\boldsymbol{y}'\mathbf{X}\hat{b} - \boldsymbol{N}\cdot\bar{Y}^2}{\boldsymbol{y}'\boldsymbol{y} - \boldsymbol{N}\cdot\bar{Y}^2} = \frac{\textit{sum of squares of the regression}}{\textit{total sum of squares}},$$

where $\bar{Y} = (1/N)\cdot\sum_{i=1}^{N} Y_i$.

Specifying the Null Hypothesis I

There are two strategies of specifying hypotheses about dependencies:

First strategy: Comparing

$$Q(Y \mid X_1, X_2, \ldots, X_{m-p}) = \gamma_0 + \gamma_1 X_1 + \ldots + \gamma_{m-p} X_{m-p}$$

with

$$E(Y \mid X_1, \ldots, X_m) = \beta_0 + \beta_1 X_1 + \ldots + \beta_m X_m.$$

The null hypothesis tested can be written in three different versions:

H_0: $\beta_{m-p+1} = \beta_{m-p+2} = \ldots = \beta_m = 0$ *version 1*

H_0: $Q(Y \mid X_1, \ldots, X_{m-p}) = E(Y \mid X_1, X_2, \ldots, X_m)$. *version 2*

H_0: $R^2_{Y|X_1,\ldots,X_m} - Q^2_{Y|X_1,\ldots,X_{m-p}} = 0$ *version 3*

Specifying Hypothesis II

Second strategy: A more general way is testing the *General Linear Hypothesis* (GLH)

$$H_0: \mathbf{A}\, b - d = \mathbf{0},$$

where the matrix **A** and the vector d are specified according to the hypothesis to be tested. The matrix **A** has to contain $p \leq m$ linear independent rows.

Test of Significance within the GLM I

Testing a null hypothesis in version 1

$$H_0: \beta_{m-p+1} = \beta_{m-p+2} = ... = \beta_m = 0,$$

that some of the coefficients of the multiple linear regression are zero, is done by estimating $\hat{R}_E^2$ and $\hat{R}_Q^2$ as seen before. Again, please note that this is done by different design matrices and different regression coefficients, i.e. with m regressors for the multiple linear regression and $m-p$ regressors for the multiple linear quasi-regression. Given the assumptions of the GLM the following statistic

$$F = \frac{(\hat{R}_E^2 - \hat{R}_Q^2) / p}{(1 - \hat{R}_E^2) / (N - m - 1)},$$

is F-distributed with $df_1 = p$ and $df_2 = N - m - 1$, where N is the sample size.

Test of Significance within the GLM II

For the second strategy – the general linear hypothesis (GLH)

$$H_0: \mathbf{A}\,\mathrm{b} - \mathrm{d} = \mathbf{0},$$

the following statistic is computed:

$$F = \frac{\hat{Q}_h / p}{\hat{Q}_e / (N - m - 1)},$$

where p is the number of (linear independent) rows of the matrix **A** of the GLH (the number of simultaneously tested single hypotheses), and further

$\hat{Q}_h = (\mathbf{A}\hat{b} - \mathrm{d})' [\mathbf{A}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}']^{-1} (\mathbf{A}\hat{b} - \mathrm{d})$ *SSQ of the hypothesis*

$\hat{Q}_e = \boldsymbol{y}'\boldsymbol{y} - \boldsymbol{y}'\,\mathbf{X}\hat{b}$. *SSQ of the error term*

This statistic is also F-distributed with $df_1 = p$ the degrees of freedom for the numerator and $df_2 = N - m - 1$ the degrees of freedom for the denominator.