MS-C1620 Statistical inference

9 Linear regression II

Jukka Kohonen

Department of Mathematics and Systems Analysis
School of Science
Aalto University

Academic year 2020–2021
Period III–IV

Contents

1. Multiple linear regression
2. Parameter estimation
3. Assessing model fit
4. Inference for model parameters
5. Extensions

Multiple linear regression

The simple linear regression can be extended to multiple linear regression incorporating several explanatory variables.

Multiple linear regression, assumptions

- Consider n observations $(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_n, y_n)$ and assume that the p-dimensional $\mathbf{x}_i$ are non-random.
- Assume, that $p < n$.
- Assume, that the values of the variable y depend linearly on the values of the variable x,

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i, \quad i = 1, \ldots, n,$$

where the regression coefficients $\beta_0, \beta_1, \ldots, \beta_p$ are unknown constants.

Multiple linear regression

The multiple linear regression model is usually coupled with the following additional assumptions.

Multiple linear regression, assumptions, continued

- The expected value of the errors is $\mathrm{E}[\varepsilon_i] = 0$ for all $i = 1, \dots, n$.
- The errors have the same variance $\mathrm{Var}[\varepsilon_i] = \sigma^2$.
- The errors are uncorrelated i.e. $\rho(\varepsilon_i, \varepsilon_j) = 0, \quad i \neq j$.
- The errors are i.i.d. *(a stronger version of the previous two assumptions)*.

Multiple linear regression

Under the previous assumptions, the random variables y_i have the following properties:

- Expected value: $\mathrm{E}[y_i] = \beta_0 + \boldsymbol{\beta}^\top \mathbf{x}_i, \quad i = 1, \ldots, n,$
- Variance: $\mathrm{Var}(y_i) = \mathrm{Var}(\varepsilon_i) = \sigma^2.$
- Correlation: $\rho(y_i, y_j) = 0, \quad i \neq j.$
- If we chose to assume that the errors are i.i.d., then y_i are independent of each other.

Multiple linear regression, parameters

The multiple linear model

$$y_i = \beta_0 + \boldsymbol{\beta}^\top \mathbf{x}_i + \varepsilon_i, \quad i = 1, \ldots, n,$$

has the following parameters: regression coefficients β_0, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$ and the error variance $\mathrm{Var}(\varepsilon_i) = \sigma^2$.

These parameters are usually unknown and have to be *estimated* from the observations.

Under the assumption that $\mathrm{E}[\varepsilon_i] = 0$, for all $i = 1, \ldots, n$, the simple linear model can be given as

$$y_i = \mathrm{E}[y_i] + \varepsilon_i, \quad i = 1, \ldots, n,$$

where $\mathrm{E}[y_i] = \beta_0 + \boldsymbol{\beta}^\top \mathbf{x}_i$ is the systematic part and ε_i is the random part of the model.

Multiple linear regression, parameter interpretation

The systematic part

$$\mathrm{E}[y_i] = \beta_0 + \boldsymbol{\beta}^\top \mathbf{x}_i$$

of the linear model defines the regression plane

$$"y = \beta_0 + \boldsymbol{\beta}^\top \mathbf{x}."$$

- The intercept β_0 tells the expected value of the response when the explanatory variable vector **x** is the zero vector.
- The regression coefficient β_j tells how much the expected value of the response variable y changes when the value of the explanatory variable x_j grows by one unit *and the other variables stay constant*.
- The error variance $\mathrm{Var}(\varepsilon_i) = \sigma^2$ describes the magnitude of the random deviations of the observed values from the regression plane.

Contents

1. Multiple linear regression
2. Parameter estimation
3. Assessing model fit
4. Inference for model parameters
5. Extensions

Multiple linear regression, objective

The estimates $\hat{\beta}_0, \hat{\boldsymbol{\beta}}$ should be chosen such that the fitted values/predictions,

$$\hat{y}_i = \hat{\beta}_0 + \hat{\boldsymbol{\beta}}^\top \mathbf{x}_i,$$

best match the observations in some suitable sense.

Again, the most popular solution method is the method of least squares.

Let $\mathbf{X}$ be a $n \times (p+1)$ matrix whose first column in full of ones and the remaining columns correspond to the observed values of the p explanatory variables. Let the n-vector $\mathbf{y}$ contain the observed response values.

The method of least squares

- In the method of least squares we choose the estimates by minimizing the sum of squared differences between the observations y_i and the fitted values $\hat{y}_i$,

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}\left(y_i - (\hat{\beta}_0 + \hat{\boldsymbol{\beta}}^\top \mathbf{x}_i)\right)^2.$$

- Denoting $\hat{\boldsymbol{\beta}}^* = (\hat{\beta}_0, \hat{\boldsymbol{\beta}}^\top)^\top$, the explicit solution is

$$\hat{\boldsymbol{\beta}}^* = (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \mathbf{y}.$$

Unstability of the solution

The least squares solution requires inverting the matrix $\mathbf{X}^\top\mathbf{X}$. If its rank is smaller than p, then some of the explanatory variables are fully linearly dependent. In that case, some of the variables can be excluded from the analysis without losing any information.

If $\mathbf{X}^\top\mathbf{X}$ is full rank, it could still be that it is very close to being singular. This corresponds to multicollinearity, the case where the explanatory variables are not necessarily fully linearly dependent but still exhibit large correlations.

Multicollinearity can make the regression coefficients unstable and its presence can be investigate using the variance inflation factors of the explanatory variables.

Variance inflation factor

Variance inflation factor (VIF) for an explanatory variable x_{ik} is defined as:

$$VIF_k = \frac{1}{1 - R_k^2},$$

where R_k^2 is the coefficient of determination for a model where x_{ik} is the dependent variable and the remaining predictors are used as explanatory variables.

VIF is calculated separately for each explanatory variable x_{ik}. If the variable x_{ik} is uncorrelated with the other explanatory variables, then the VIF $= 1$.

If an explanatory variable has VIF ≥ 10, multicollinearity is likely present, and some of the variables should be dropped from the model.

Roughly, the aim is to select variables such that the coefficient of determination (of the $(\mathbf{x}_i, y_i)$-model) is as high as possible and the explanatory variables are as uncorrelated as possible.

Contents

1. Multiple linear regression
2. Parameter estimation
3. Assessing model fit
4. Inference for model parameters
5. Extensions

Fitted values and residuals

- Recall that the fitted value of the variable y_i, i.e., the value given to the response variable by the estimated regression plane at the point $\mathbf{x}_i$, is

$$\hat{y}_i = \hat{\beta}_0 + \hat{\boldsymbol{\beta}}^\top \mathbf{x}_i, \quad i = 1, \ldots, n.$$

- The residual $\hat{\varepsilon}_i$ of the estimated model is the difference

$$\hat{\varepsilon}_i = y_i - \hat{y}_i, \quad i = 1, \ldots, n$$

  between the observed value y_i (of the variable y) and fitted value $\hat{y}_i$.
- The smaller the residuals of the estimated model are, the better the regression model explains the observed values of the response variable.

Residual mean square estimation

Under the regression assumptions, an unbiased estimate for the error variance $\mathrm{Var}(\varepsilon_i) = \sigma^2$ is given by

$$\hat{\sigma}^2 = \frac{1}{n-p-1}\sum_{i=1}^{n}(\hat{\varepsilon}_i - \bar{\hat{\varepsilon}})^2 = \frac{1}{n-p-1}\sum_{i=1}^{n}\hat{\varepsilon}_i^2 = \frac{1}{n-p-1}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2.$$

Note that the divisor equals the sample size n minus the number of estimated regression parameters $p+1$.

Coefficient of determination

- Coefficient of determination (also known as "R-squared") gives a single number with which to assess the accuracy of the model fit.
- Coefficient of determination is defined as

$$R^2 = 1 - \frac{SSE}{SST},$$

  where

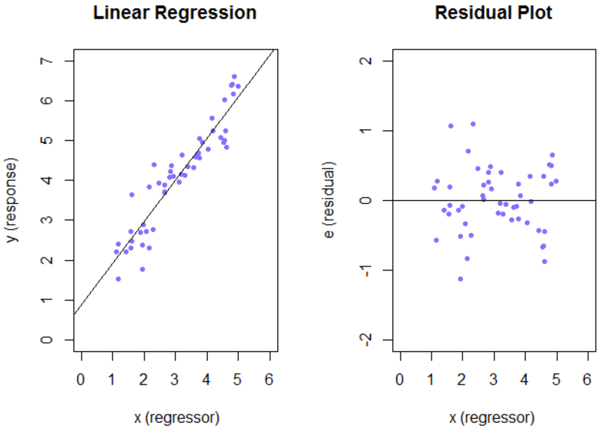
$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2 \quad \text{and} \quad SSE = \sum_{i=1}^{n} \hat{\varepsilon}_i^2$$

  measure the variation of the data "before" and "after" fitting the model.
- If SSE is small compared to SST, the model has managed to *explain* a large proportion of the variance in the data.
- We always have $0 \leq R^2 \leq 1$.

Diagnostics

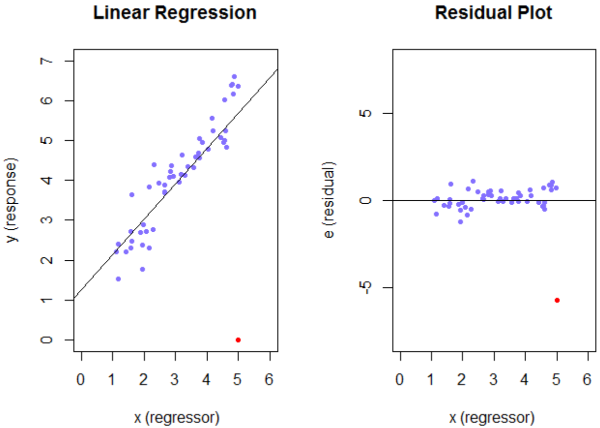
- The verification of the assumptions of a regression model is called diagnostics.
- The diagnostics are usually performed by using the plots of the residuals versus the fitted values (or the explanatory variable x_i if there is only a single one).
- If the model assumptions hold, the residuals,
    1. are approximately evenly distributed on both sides of zero,
    2. have constant variance regardless of the value of x (no heteroscedasticity),
    3. exhibit no unusual (non-linear) patterns in general.
- Additionally, if the sample size is small and we cannot rely on the central limit theorem, the normality of the residuals should be tested.

Example, linear regression



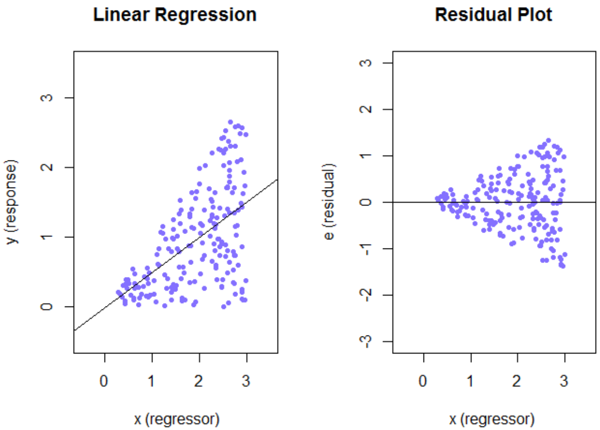
Kuva: Example of mostly satisfactorily looking residuals.

Example, outlier



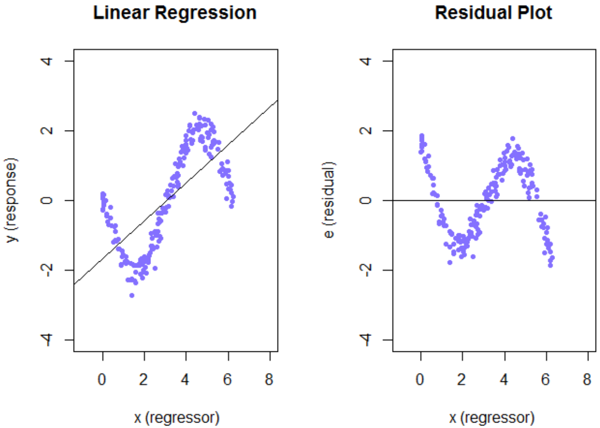
Kuva: Diagnostics can also be used to spot outliers.

Example, heteroskedasticity



Kuva: The variance of the residuals depends clearly on the value of x (heteroskedasticity).

Example, non-linear dependence



Kuva: The residuals exhibit clear non-linear dependency on x.

Contents

1 Multiple linear regression

2 Parameter estimation

3 Assessing model fit

4 Inference for model parameters

5 Extensions

Inference for model parameters

Analogous results as for the simple linear regression (confidence intervals, hypothesis tests) also exist for multiple linear regression under the assumptions that

Multiple linear regression, assumptions, continued

- The errors ε_i are i.i.d.
- The errors ε_i are normally distributed.

The assumption of normality can be replaced by a *large enough* sample size.

However, we will not go through the theory behind them here.

Bootstrap in linear models

In addition to the standard normality/CLT-based inference output by the software, bootstrap can be used to obtain confidence intervals for the model parameters. A bootstrap sample can be created in two ways.

- **Observation resampling:** we simply draw a bootstrap sample of size n from amongst the observations $(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n)$ and fit a linear model to it (and repeat the same B times).
- **Residual resampling:** a bootstrap resample is obtained as $(\mathbf{x}_1, y_1^*), \dots, (\mathbf{x}_n, y_n^*)$, where $y_i^* = \hat{y}_i + \hat{\varepsilon}_i^*$ and $(\hat{y}_1, \dots, \hat{y}_n)$ are the fitted values of the original sample and $\hat{\varepsilon}_1^*, \dots, \hat{\varepsilon}_n^*$ is a bootstrap sample of the residuals of the model for the original sample.

The residual resampling is suited for situations where we want to preserve the explanatory variable structure of the original data also in the bootstrap resamples (e.g. two treatment groups of certain sizes).

Contents

1. Multiple linear regression

2. Parameter estimation

3. Assessing model fit

4. Inference for model parameters

5. Extensions

Extensions of the multiple regression model

The following slides list some of the most common cases where extensions to the standard linear regression methodology are required.

Problem: *My data exhibits non-linear dependencies.*
Solution:

- Transform your predictors. That is, $y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \varepsilon_i$ is still a multiple regression model.
- Add interaction terms, $\beta_{12} x_{i1} x_{i2}$ (note: interaction terms make interpreting the coefficients difficult).
- A more automated solution is given by non-linear regression, such as *kernel regression*.

Problem: *I have several response variables.*
Solution: Model them simultaneously using *multivariate regression:*

$$\mathbf{y}_i = \boldsymbol{\beta}_0 + \mathbf{B}^\top \mathbf{x}_i + \boldsymbol{\varepsilon}_i, \quad i = 1, \dots, n.$$

Extensions of the multiple regression model

Problem: *My data contains a large number of outliers and I would rather not discard them all.*
Solution: Use a fitting method which is less sensitive to outlying values such as "the method of least absolute values", or ℓ_1-regression.

Problem: *My data points exhibit significant correlations.*
Solution: Depending on the nature of the correlations, use either *mixed models* (correlation within groups of subjects), *time series* methods (temporal correlation) or something else.

Extensions of the multiple regression model

Problem: *My response variable takes values in some specific subset of $\mathbb{R}$. That is, the ranges of the two sides of the model equation $\mathrm{E}[y_i] = \beta_0 + \boldsymbol{\beta}^\top \mathbf{x}_i$ do not match.*
Solution: Use a suitable *link function* to transform your model equation, $g(\mathrm{E}[y_i]) = \beta_0 + \boldsymbol{\beta}^\top \mathbf{x}_i$. This leads, e.g., to *logistic regression* and *log-linear models*.

Problem: *Too many variables, $p \geq n$.*
Solution: This makes the matrix $\mathbf{X}^\top \mathbf{X}$ singular, meaning that no least squares solution exist. This problem can be solved by regularized regression estimates discussed briefly next time.