NTNU
Norwegian University of
Science and Technology

TMA4267 Linear Statistical Models V2014 (13)

Multiple linear regression, properties of estimators [3.2-3.4]

Mette Langaas

To be lectured: February 17, 2014
wiki.math.ntnu.no/emner/tma4267/2014v/start/

Acid rain in Norwegian lakes

Measured pH in Norwegian lakes explained by content of

- x1: SO_4: sulfate (the salt of sulfuric acid),
- x2: $N0_3$: nitrate (the conjugate base of nitric acid),
- x3: *Ca*: calsium,
- x4: latent *Al*: aluminium,
- x5: organic substance,
- x6: area of lake,
- x7: position of lake (Telemark or Trøndelag),

TMA4267: Lecture 13, Ch 3.2-3.3-3.4

Multiple linear regression (MLR)

$$\underbrace{Y}_{\substack{n\times 1 \\ \text{response}}} = \underbrace{X}_{\substack{n\times p \\ \text{design} \\ \text{matrix}}} \underbrace{\beta}_{\substack{p\times 1 \\ \text{parameter} \\ \text{vector}}} + \underbrace{\varepsilon}_{\substack{n\times 1 \\ \text{error}}}$$

i) where ε_i are i.i.d with $E(\varepsilon_i) = 0$ and $Var(\varepsilon_i) = \sigma^2$.

ii) And add normality $\underset{n\times 1}{\varepsilon} \sim N_n(0, \underbrace{\sigma^2 I}_{n\times n})$.

Normal equations: $(X^T X)\hat{\beta} = X^T Y$

Estimator: $\hat{\beta} = (X^T X)^{-1} X^T Y$

Multiple linear regression model

$$Y_i \quad = \quad \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i, \text{ for } i = 1, ..., n$$

Matrix formulation:

$$\underset{(n\times 1)}{\boldsymbol{Y}} \quad = \quad \underset{(n\times p)}{\boldsymbol{X}} \underset{(p\times 1)}{\boldsymbol{\beta}} + \underset{(n\times 1)}{\boldsymbol{\varepsilon}}$$

$$E(\boldsymbol{\varepsilon}) = \underset{(n\times 1)}{\mathbf{0}} \quad \text{and} \quad Cov(\boldsymbol{\varepsilon}) = \underset{(n\times 1)}{\sigma^2 \boldsymbol{I}}$$

where

— $\boldsymbol{\beta}$ and σ^2 are unknown parameters and

— the design matrix $\boldsymbol{X}$ has ith row $[x_{i1}, x_{i2}, ..., x_{ip}]$.

Performing MLR in practice

Numerical note:

The normal equations can in practice be solved by using the QR decomposition of X. See book p 68-69 for details.

$$X = QR$$

$$(X^T X) b = X^T y \iff \underbrace{R}_{\text{upper triangular}} b = \underbrace{Q^T}_{\text{ortogonal}} y$$

solved by backsubstitution

R: try out the L13.r script to analyse acid rain

Main function to use is "lm".

Finding: $\hat{\beta} = \begin{pmatrix} 5.68 \\ -0.32 \\ \vdots \\ 0.09 \end{pmatrix}$

Least squares estimation

- $\boldsymbol{Y} = \boldsymbol{X}\beta + \varepsilon$ with $E(\varepsilon) = \mathbf{0}$ and $\mathrm{Cov}(\varepsilon) = \sigma^2 \boldsymbol{I}$.
- Let $\boldsymbol{X}$ has full rank $p \leq n$.
- The Least Squares Estimate (LSE) of β is given by

$$\hat{\beta} = (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{Y}$$

Properties of $\hat{\beta}$ [3.3]

First, not assume normality, only assume

$$E(\varepsilon) = 0 \quad \text{and} \quad Cov(\varepsilon) = \sigma^2 I.$$

And, $Y = X\beta + \varepsilon$,

$$E(Y) = X\beta \qquad Cov(Y) = \sigma^2 I$$

The LS estimator $\hat{\beta} = \underbrace{(X^TX)^{-1}X^T}_{G}Y = GY$

1) Observe that $\hat{\beta}$ is linear in the data Y.

2) Mean: $E(\hat{\beta}) = E(GY) = G\underbrace{E(Y)}_{X\beta} = \underbrace{(X^TX)^{-1}X^TX}_{I}\beta = \beta$

$\hat{\beta}$ is an unbiased estimator for β.

3) Covariance:

$$Cov(\hat{\beta}) = Cov(GY) = G\underbrace{Cov(Y)}_{\sigma^2 I}G^T$$

$$= (X^TX)^{-1}X^T\sigma^2 I\left[(X^TX)^{-1}X^T\right]^T$$

$$= \sigma^2\underbrace{(X^TX)^{-1}X^TX}_{I}(X^TX)^{-1} = \sigma^2(X^TX)^{-1}$$

$$= \sigma^2 \cdot C^{-1} \quad \text{(book notation)}$$

Properties of LS-estimates

$\hat{\beta} = (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{y}$ has

$$E(\hat{\beta}) = \boldsymbol{\beta} \text{ and } \mathrm{Cov}(\hat{\beta}) = \sigma^2 (\boldsymbol{X}^T \boldsymbol{X})^{-1}$$

The matrix $C = X^T X$ is called the information metrix, and is SPD if X has full rank.
(symmetrix, positive definite)

Note that in a designed experiment we may ourselves define X.
In the topic of Design of Experiments (DOE) X is designed to be optimal in some sense.
For example X may be chosen to minimize a function of $(X^T X)^{-1}$, e.g make the variences or covariances small.
We will look at a socalled 2^k experiment, where X is chosen with orthogonal columns so that $(X^T X)$ is a diagonal matrix, and thus all $Cov(\hat{\beta}_j, \hat{\beta}_k) = 0$.

Best linear unbiased estimator (BLUE)

general matrix

Among all unbiased linear estimator $\hat{\beta} = BY$

$\hat{\beta} = (X^T X)^{-1} X^T Y$ has the minimum varience in each component

$\Rightarrow \hat{\beta}$ is BLUE.

Proof: page 73 of BF.

Gauss' LS theorem

— $\boldsymbol{Y} = \boldsymbol{X}\beta + \varepsilon$ with $E(\varepsilon) = \mathbf{0}$ and $\mathrm{Cov}(\varepsilon) = \sigma^2\boldsymbol{I}$.
— And $\boldsymbol{X}$ has full rank $p \leq n$.
— $\hat{\beta} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$.

Then, for any vector $\boldsymbol{c}$, the estimator

$$\boldsymbol{c}^T\hat{\beta}$$

of $\boldsymbol{c}^T\beta$ has the *smallest possible variance* among all linear estimators that are *unbiased* for $\boldsymbol{c}^T\beta$.

Maximum likelihood estimator of σ

Assume Y_i indendent $N(X_i\beta, \sigma^2)$

row vector of X matrix

Log-likelihood

Lecture 12

$$\ell(\beta, \sigma^2) = \ln L(\beta, \sigma)$$

$$= \frac{n}{2}\ln(2\pi) - n\log\sigma - \frac{1}{2\sigma^2}(y - X\beta)^T(y - X\beta)$$

$$\frac{\partial \ell}{\partial \sigma} = 0 - \frac{n}{\sigma} + \frac{1}{\sigma^3}(y - X\beta)^T(y - X\beta) = 0$$

At the likelihood maximum $\beta = \hat{\beta}$, so we insert $\hat{\beta}$ and solve for σ^2 to get $\hat{\sigma}^2$:

$$\hat{\sigma}^2 = \frac{1}{n}\underbrace{(Y - X\hat{\beta})^T(Y - X\hat{\beta})}_{SSE}$$

SSE $\leftarrow$ a RV and a statistic

error

New notation

$$Y = X\beta + \varepsilon$$ error

$$\underset{n\times 1}{\hat{Y}} = X\hat{\beta}$$ is called fitted values

$$\underset{n\times 1}{e} = Y - \hat{Y} = Y - X\hat{\beta}$$ ← called residuals

$$\begin{bmatrix} e_1 \\ \vdots \\ e_n \end{bmatrix} \quad SSE = (Y - X\hat{\beta})^T (Y - X\hat{\beta}) = e^T e$$

sums of squared residuals

ML estimation of σ

The MLE for σ^2 is

$$\hat{\sigma}^2 = \frac{1}{n}(\boldsymbol{Y} - \boldsymbol{X}\hat{\beta})^T(\boldsymbol{Y} - \boldsymbol{X}\hat{\beta}) = \frac{1}{n}\sum_{i=1}^{n}(y_i - X_i\hat{\beta})^2$$

This may also be written in a slightly different way:

— First, fitted values of $\boldsymbol{Y}$:

$$\hat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\beta} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$$

— The residuals:

$$\boldsymbol{e} = \boldsymbol{Y} - \hat{\boldsymbol{Y}} = (\boldsymbol{Y} - \boldsymbol{X}\hat{\beta})$$

— and then the Sums-of-squares-of-error:

$$\text{SSE} = \boldsymbol{e}^T\boldsymbol{e} = (\boldsymbol{Y} - \boldsymbol{X}\hat{\beta})^T(\boldsymbol{Y} - \boldsymbol{X}\hat{\beta})$$

— with leads to

$$\hat{\sigma}^2 = \frac{1}{n}\text{SSE}$$

Projection metrices

putting a hat on y (H)
(in book use P — overloaded)

$$\hat{Y} = X\hat{\beta} = \underbrace{X(X^TX)^{-1}X^T}_{H}Y = HY$$

$$e = Y - \hat{Y} = Y - HY = (I-H)Y$$

Observe:

1) $\underset{n\times n}{HH} = X\underbrace{(X^TX)^{-1}X^TX}_{I}(X^TX)^{-1}X^T = X(X^TX)^{-1}X^T = H$

$A^2 = A$
idempotent

2) Homework: $(I-H)(I-H) = (I-H)$

3) $H\cdot(I-H) = H - HH = H - H = 0$

In general: A linear transformation A is a projection (into a vector space V) if $A^2 = AA = A$ (idempotent).

Then $V = \underbrace{\text{Im}(A)}_{} \oplus \underbrace{\text{Ker}(A)}_{}$

Image of A
= subspace spanned by the colums of A

direct sum

any vector in V can be uniquly written as a sum of a vector in Im(A) and a vector in Ker(A)

kernel of A
= subspace spanned by the column vectors x such that
$Ax = 0$. Also called null space

Also: $\text{Im}(A) = \text{Ker}(I - A)$ and $\text{Ker}(A) = \text{Im}(I - A)$

In our case:

$$H = X(X^TX)^{-1}X^T$$

$$H^T = H$$

and $(I-H)^T = (I-H)$

If a projection matrix A is symmetric it is also orthogonal. That is, A is an orthogonal projection,

This means that Im(A) and ker(A) are orthogonal vector spaces. And, thus are Im(A) and Im(I-A) orthogonal vector spaces.

Back to H, $(I-H)$, MLR:

$\hat{Y} = HY$ is the projection of Y onto the space spanned by the columns of H. And, $HX = X(X^TX)^{-1}X^TX = X$ so the column space of H and X are the same.

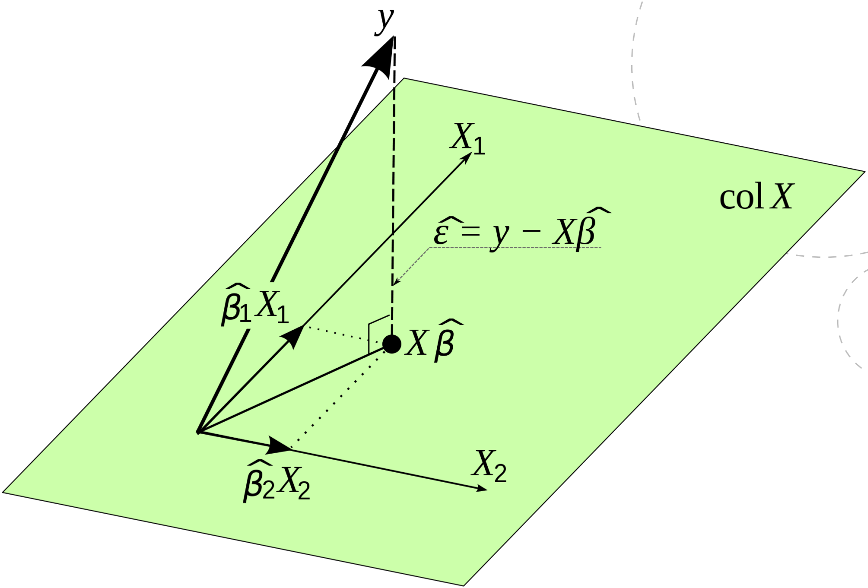
Further: $e = (I-H)Y$ is the projection of Y onto the space orthogonal to the space spanned by the columns of H.

Geometry of Least Squares

- Mean response vector: $E(\boldsymbol{Y}) = \boldsymbol{X}\beta$
- As β varies, $\boldsymbol{X}\beta$ spans the model plane of all linear combinations. I.e. the space spanned by the columns of $\boldsymbol{X}$: the column-space of $\boldsymbol{X}$.
- Due to random error (and unobserved covariates), $\boldsymbol{y}$ is not exactly a linear combination of the columns of $\boldsymbol{X}$.
- LS-estimation chooses $\hat{\beta}$ such that $\boldsymbol{X}\beta$ is the point in the column-space of $\boldsymbol{X}$ that is closes to $\boldsymbol{y}$.

Geometry of Least Squares (cont.)

- The residual vector $\boldsymbol{e} = \boldsymbol{Y} - \hat{\boldsymbol{Y}}$ is perpendicular to the column-space of $\boldsymbol{X}$.
- Multiplication by $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$ projects a vector onto the column-space of $\boldsymbol{X}$.
- Multiplication by $\boldsymbol{I} - \boldsymbol{H} = \boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$ projects a vector onto the space perpendicular to the column-space of $\boldsymbol{X}$.



http://en.wikipedia.org/wiki/File:OLS_geometric_interpretation.svg

Properties of symmetric projection matrices

A projection $\boldsymbol{A}$ matrice is idempotent, $\boldsymbol{A}^2 = \boldsymbol{A}$. A symmetric projection matrix is orthogonal.

1. The eigenvalues of a projection matrix are 0 and 1.
2. The rank of a symmetric matrix (actually: a diagonalizable quadratic matrix) equals the number of nonero eigenvaluse of the matrix. Should be known from previous courses.
3. (Combining 1+2). If a $(n \times n)$ symmetric projection matrix $\boldsymbol{A}$ has rank r then r eigenvalues are 1 and $n - r$ are 0.
4. The trace and rank of a symmetric projection matrix are equal: $tr(\boldsymbol{A}) = \text{rank}(\boldsymbol{A})$.