PART 2:
LINEAR REGRESSION

TMA4267 L7
07.02.2017

Model definition [F 3.1.0]

Y = variable of primary interest (response, dependent variable)

$x_1, x_2, \dots, x_k$ = regressors, explanatory variables, independent variables, covariates

Assumptions:

$$Y = \underbrace{f(x_1, x_2, \dots, x_k)}_{\text{systematic component}} + \underset{\substack{\uparrow \\ \text{error term}}}{\varepsilon}$$

1) Systematic component is a linear combination of the covariates.

$$f(x_1, x_2, \dots, x_k) = \overbrace{\underbrace{\beta_0 + \beta_1 x_1}_{\text{simple}} + \beta_2 x_2 + \dots + \beta_k x_k}^{\text{multiple}}$$

$$\underset{\substack{p \times 1 \\ p = k+1}}{x} = \begin{bmatrix} 1 \\ x_1 \\ x_2 \\ \vdots \\ x_k \end{bmatrix} \qquad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} \qquad f(x) = x^T \beta$$

2) Additive errors $Y = x^T\beta + \varepsilon$

Restrictive? Maybe → transformations?

Data and design matrix

We collect independent data (Y_i, x_i) for $i = 1, \dots, n$

pairs

$x_i = (1, x_{i1}, x_{i2}, \dots, x_{ik})^T$

response

$$\underset{n\times 1}{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, \quad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

$$\underset{n\times p}{\mathbb{X}} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ \vdots & & & & \\ 1 & x_{n1} & x_{n2} & & x_{nk} \end{bmatrix} \leftarrow \text{individual/observational unit}$$

covariate

design matrix. We will assume that $n >> p$.

and $\text{rank}(\mathbb{X}) = p$

n = number of observations

p = number of covariates + 1 (intercept)

Q: What can make $\text{rank}(\mathbb{X}) < p$?

Ex: Munich rent index: $n = 3082$

Y = rent or rent pr sq.m $\rightarrow$ $Y = \begin{bmatrix} 5.26 \\ 0.41 \\ \vdots \\ \end{bmatrix}$

$$\mathbb{X} = \begin{bmatrix} 1 & 20 & 1970 & 1 & 0 & 0 & 0 & 916 \\ 1 & 41 & 1941 & 1 & 0 & 0 & 1 & 206 \\ \vdots & \vdots & \vdots & 2 & 1 & 1 & 1 & \vdots \\ & & & \vdots & 1 & \vdots & 1 & \vdots \\ 1 & & \vdots & 3 & \vdots & & \vdots & \vdots \\ & & & & & 1 & 1 & \end{bmatrix}$$

area — year of c — location — bath — kitch — c. heate — district

$\mathbb{X}$ is chosen so that we have a good model.

The classical linear model

$Y_i = x_i^T \beta + \varepsilon_i$ ← one person

$$\underset{n\times 1}{Y} = \underset{n\times p}{\mathbb{X}} \underset{p\times 1}{\beta} + \underset{n\times 1}{\varepsilon}$$

ε: unobserved random vector

1) $E(\varepsilon) = 0$

2) $Cov(\varepsilon) = E(\varepsilon\varepsilon^T) - \underset{0}{E(\varepsilon)}\,\underset{0}{E(\varepsilon)^T} = \sigma^2 I$

   $Var(\varepsilon_i) = \sigma^2$ for all i ← homoscedastic errors

   $Cov(\varepsilon_i, \varepsilon_j) = 0$ for $i \neq j$ ← uncorrelated errors

3) The design matrix has rank $rank(\mathbb{X}) = k+1 = p$

4) If we in addition assume that $\varepsilon \sim N_n(0, \sigma^2 I)$ then we have a normal linear regression model.

   What does the imply for the distribution of Y?

$Y \sim N_n$ since $Y = \underbrace{X\beta}_{\text{constant}} + \underset{\substack{\uparrow \\ N_n}}{\varepsilon}$, and

$$E(Y) = E(X\beta + \varepsilon) = X\beta + \underbrace{E(\varepsilon)}_{=0} = X\beta$$

$$Cov(Y) = Cov(X\beta + \varepsilon) = 0 + Cov(\varepsilon) = \sigma^2 I$$

$$Y \sim N_n(X\beta, \sigma^2 I)$$

The covariates X may be regarded as random variables, and then the assumptions (1)+(2) are made conditional on $X = x$, so $E(\varepsilon \mid X = x) = 0$ and $Cov(\varepsilon \mid X = x) = \sigma^2 I$

If we isteed assume that

$$\begin{bmatrix} Y \\ X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix} \sim N_{n+1} \Rightarrow \begin{array}{l} E(Y \mid X = x) = \text{linear in } x \\ Var(Y \mid X = x) = \text{not dependent on } x \end{array}$$

Model parameters, estimates and residuals

[F 3.1.1]

$$Y = \mathbb{X}\beta + \varepsilon \quad , E(\varepsilon) = 0, \ \mathrm{Cov}(\varepsilon) = \sigma^2 I$$

The $\underbrace{\text{model parameters}}_{\text{the unknown}}$ are $\underset{p\times 1}{\beta}$, σ^2

We will develop estimators:

$$\hat{\beta} = (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T Y$$ by least squares (LS) and maximum likelihood. (ML)

$$\hat{\sigma}^2 = \frac{1}{n-p}(Y - \mathbb{X}\hat{\beta})^T(Y - \mathbb{X}\hat{\beta})$$ by restricted maximum likelihood (REML)

Further: Y is a random vector with mean $\mathbb{X}\beta$, and estimator for $E(Y) = \mathbb{X}\beta$ is $\hat{Y} = \mathbb{X}\hat{\beta}$.

The error ε is a random vector with $E(\varepsilon) = 0$ and $\mathrm{Cov}(\varepsilon) = \sigma^2 I$, but ε is not <u>observed</u>.

unknown

$$Y = \mathbb{X}\beta + \varepsilon$$ unobserved

observed   observed

Our best guess for the error is the residual vector ($\hat{\varepsilon}$, e)

$$\hat{\varepsilon} = Y - \hat{Y} = Y - \mathbb{X}\hat{\beta}$$

So, the residuals can be calculated, and we may think of the residuals as <u>predictions of the errors</u>

Be aware: don't mix errors ε (unobserved) with residuals $\hat{\varepsilon}$ ("observed").

The residuals will be used to assess model assumptions as proxies for the errors.

Model assumption [F 3.1.2]

1) Linearity of covariates $Y = X\beta + \varepsilon$

$$y_i = \beta_0 + \beta_1 x_{i1} + \varepsilon_i \qquad \text{ok}$$
$$\beta_0 + \beta_1 z_{i1}^2 + \varepsilon_i \qquad \text{ok}$$
$$\beta_0 + \beta_1 \log(z_{i1}) + \varepsilon_i \qquad \text{ok}$$
$$\beta_0 + \beta_1 \sin(\beta_2 z_{i1}) + \varepsilon_i \qquad \text{not ok}$$

If the relationship between Y and x_1 is nonlinear $\Rightarrow$ ok to use polynomial (or similar) in x_1. More advanced: nonparametric regression

2) Homoscedastic error variance: $Cov(\varepsilon) = \sigma^2 I$

Need to check that $Var(\varepsilon)$ does not vary systematic across observations.

Look at covariate vs residuals $\rightarrow$ trend? fan out, fan in.

Solution if problem: If we knew $Cov(\varepsilon) = \Sigma$

we may use a socalled general linear model, with weighted least squares (LecEx3.P4), but Σ is in general unknown.

Remaining: 3) uncorrelated errors
4) additive error.