PART 2: LINEAR REGRESSION

TMA4267 L8

10.02.2017

$$\underset{\text{response}}{Y} = \underset{\text{design matrix}}{\mathbb{X}} \underset{\text{paremeter}}{\beta} + \underset{\text{RV}}{\varepsilon}$$

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} \qquad \begin{bmatrix} 1 & x_{11} \cdots & x_{1k} \\ \vdots & & \\ 1 & x_{n1} \cdots & x_{nk} \end{bmatrix} \qquad \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

response

design matrix

$n \times p$, $p = k+1$

errors unobservable

$$Y_i = \beta_0 + \beta_1 \cdot x_{i1} + \beta_2 \cdot x_{i2} + \dots + \beta_k \cdot x_{ik} + \varepsilon_i$$

and assume:

$E(\varepsilon) = 0$ and $Cov(\varepsilon) = \sigma^2 I$

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

residuals

Model assumptions [F3.1.2 cont.]

1) Linearity of covariates:

plot each x vs y

each x vs residuals

2) Homoscedastic error variance

Plot x vs $\hat{\varepsilon}$

$\hat{y}$ vs $\hat{\varepsilon}$

3) Uncorrelated errors: $\mathrm{Cov}(\varepsilon_i, \varepsilon_j) = 0$

If data comes from measurements in time or space errors may be autocorrelated.

$$\varepsilon_i = \rho \cdot \varepsilon_{i-1} + u_i$$

popular time serie model

u_i ~ $N(0,1)$

But, autocorrelation may also be due to misspecification of the model, e.g. a missing (unobserved) covariate, or by modelling a linear instead of a nonlinear relationship.

4) Additivity of errors: $Y = X\beta + \varepsilon$

$$Y = \exp\{\beta_0 + \beta_1 x_1 + \dots + \beta_k x_k + \varepsilon\}$$

$$= \exp(\beta_0) \cdot \exp(\beta_1 x_1) \cdot \dots \cdot \exp(\beta_k x_k) \cdot \exp(\varepsilon)$$

has multiplicative errors.

Transforming the model (Y) using $\ln$ gives additive errors.

Coveriates: how to include in the linear regression [F3.1.3]

1) Continuous x_1

* Y vs x_1 linear
* transform x_1 $\left(\ln x_1, \frac{1}{x_1}, \sqrt{x_1}\right)$
* use polynomial in x_1.

2) Categorical ← nominal (green, red, blue) ← district
← ordinal (good, average, top) ← location

Ex: Y = rentsqm
X_1 = location — 1: average, 2: good, 3: top

$\hat{\beta}_1 = 0.39$ → the effect of good location is twice the effect of average location

a) Linear coding

b) Dummy variable coding

Ex: location →

aloc	1 0	←	1794
gloc	1 0	←	1210
tloc	1	←	78

$\underbrace{}_{3082}$

see slide

$$\underset{3082\times 4}{\mathbb{X}} = \begin{array}{cccc} \beta_0 & aloc & gloc & tloc \\ \end{array}$$
$$\begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{bmatrix} \quad \begin{array}{l} \text{average} \cdot 1794 \\ \text{good} \cdot 1210 \\ \text{top} \cdot 78 \end{array}$$

Problem: $\text{rank}(\mathbb{X}) = 3$, not unique solution.

How to solve:

a) not include intercept

b) omit one of the dummy variables, the omitted category is called the reference category.

Ex: let average be omitted: contr.treatment

$$rentsqm_i = \beta_0 + \beta_2\, gloc_i + \beta_3 \cdot tloc_i + \varepsilon_i$$

average loc$_i$: $\widehat{rentsqm} = \hat{\beta}_0$

good : $\widehat{rentsqm} = \hat{\beta}_0 + \hat{\beta}_2$

top : $\widehat{rentsqm} \quad \hat{\beta}_0 + \hat{\beta}_3$

c) add a restriction: sum-to-zero

$$\sum_{j=1}^{3} \beta_j = 0 \qquad \text{R. contr.sum}$$

Effect coding (important in Part 3+4)

We have $X_1 = \begin{cases} 1 & \text{average} \\ 2 & \text{good} \\ 3 & \text{top} \end{cases}$

$$Z_1 = \begin{cases} 1 & \text{if } x_1 = 1 \\ -1 & \text{if } x_1 = 3 \\ 0 & \text{else } (x_1 = 2) \end{cases} \qquad Z_2 = \begin{cases} 1 & \text{if } x_1 = 2 \\ -1 & \text{if } x_1 = 3 \\ 0 & \text{else } (x_1 = 1) \end{cases}$$

$$y_i = \alpha_0 + \alpha_1 Z_1 + \alpha_2 \cdot Z_2 + \varepsilon$$

$$\alpha_3 = -\alpha_1 - \alpha_2.$$

3) Interactions

Is the effect (on Y) of a change in x_1 dependent on the value of another covariate x_2?

Lego: Y = birth weight child
x_1: glucose level of mother
x_2 = BMI of mother ($\frac{kg}{m^2}$)

figure 1: $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$

figure 2: High glucose will have a different effect on birth weight when BMI is low compared to when BMI is high. $\Rightarrow$ we have an interaction between x_1 and x_2.

a) continuous x_1 and x_2

simplest solution: $Y_i = \beta_0 + \underbrace{\beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1} \cdot x_{i2}}_{f(x_1, x_2)} + \varepsilon_i$

many complex solutions possible

b) categorical: may do the same as for continuous. Easiest solution: define new variable with all combinations of x_1 and x_2

premium kitchen or not — 2 levels

premium bath or not — 2 levels

x_3 with 4 levels

only use this.

dummy variable coding.

Paremeter estimation [F3.2]

Estimator for β [F3.2.1]

1) Maximum likelihood $\qquad Y = X\beta + \varepsilon$

If $\varepsilon \sim N_n(0, \sigma^2 I)$ then $Y \sim N_n(X\beta, \sigma^2 I)$

Alt 1: $Y_1, Y_2, \dots, Y_n$ independent

$E(Y_i) = x_i^T\beta$, $Var(Y_i) = \sigma^2$

$$\rightarrow f(y_i; \overset{x_i^T\beta}{\overset{\downarrow}{\mu}}, \sigma) = \frac{1}{\sqrt{2\pi}} \frac{1}{\sigma} \cdot e^{-\frac{1}{2\sigma^2}(y_i - \underset{\underset{x_i^T\beta}{\parallel}}{\mu})^2}$$

$$\underset{\underset{f(y)}{\parallel}}{L(\beta, \sigma^2)} = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \frac{1}{\sigma} \exp\left\{-\frac{1}{2\sigma^2}(y_i - x_i^T\beta)^2\right\}$$

joint density of $Y_1, \dots, Y_n$

$$= \left(\frac{1}{2\pi}\right)^{\frac{n}{2}} \frac{1}{\sigma^n} \exp\left\{-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - x_i^T\beta)^2\right\}$$

homework

$$(*) = \left(\frac{1}{2\pi}\right)^{\frac{n}{2}} \frac{1}{\sigma^n} \exp\left\{-\frac{1}{2\sigma^2} \underbrace{(y - X\beta)^T(y - X\beta)}_{LS(\beta)}\right\}$$

maximizing L wrt β is the same a minimizing $LS(\beta)$ wrt β

with respect to

Alt 2: $Y \sim N_n(\mu, \Sigma)$

$$f(y; \mu, \Sigma) = \left(\frac{1}{2\pi}\right)^{\frac{n}{2}} [\det(\Sigma)]^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(y - \mu)^T \Sigma^{-1} (y - \mu)\right\}$$

Homework: $\mu = X\beta$, $\Sigma = \sigma^2 I \Rightarrow$ get the same $L(\beta, \sigma^2)$ as $(*)$

2) Least squares: minimize $LS(\beta)$ wrt β

$$LS(\beta) = (y - \mathbb{X}\beta)^T (y - \mathbb{X}\beta)$$

i) $LS(\beta) = y^T y - y^T \mathbb{X}\beta - \beta^T \mathbb{X}^T y + \beta^T \mathbb{X}^T \mathbb{X}\beta$

observe: $\underbrace{y^T \mathbb{X}\beta}_{\text{scalar}} = \beta^T \mathbb{X}^T y$

$$LS(\beta) = y^T y - 2y^T \mathbb{X}\beta + \beta^T \mathbb{X}^T \mathbb{X}\beta$$

ii) To minimize $LS(\beta)$ wrt β we may solve

$$\frac{\partial LS(\beta)}{\partial \beta} = 0$$

$$\begin{pmatrix} \frac{\partial LS(\beta)}{\partial \beta[1]} \\ \frac{\partial LS(\beta)}{\partial \beta[2]} \\ \vdots \\ \frac{\partial LS(\beta)}{\partial \beta[p]} \end{pmatrix}$$

"Need" two rules for derivatives wrt vector:

Rule 1: $\frac{\partial}{\partial \beta} (\underset{1\times p}{d^T} \underset{p\times 1}{\beta}) = \frac{\partial}{\partial \beta} \left(\sum_{i=1}^{p} d_i \beta_i \right)$

$$= \underline{\underline{d}}$$

Rule 2: $\frac{\partial}{\partial \beta} (\beta^T D \beta) = \frac{\partial}{\partial \beta} \left(\sum_{j=1}^{p} \sum_{k=1}^{p} \beta_j d_{jk} \cdot \beta_k \right)$

$\{d_{ik}\}$

$= (D + D^T)\beta$ and $2D\beta$ when $D = D^T$.

7

Homework: $\frac{\partial}{\partial \beta} LS(\beta)$ with these two rules!