MULTIPLE LM: ESTIMATION

parameters to estimate: $\left.\begin{matrix}\beta_1,\dots,\beta_p \\ \sigma^2\end{matrix}\right\}$ $\Rightarrow$ denote with $\theta = (\beta_1,\dots,\beta_p,\sigma^2)$ $\Rightarrow$ parameter space $\Theta = \mathbb{R}^p \times (0,+\infty)$

data: random sample $(y_1,\dots,y_n)$; covariates $(x_{i1},\dots,x_{ip})$ for $i=1,\dots,n$

MODEL: $Y_i \sim N(\mu_i, \sigma^2)$ independent for $i=1,\dots,n$

with $\mu_i = \beta_1 x_{i1} + \dots + \beta_p x_{ip}$

density $f(y_1,\dots,y_n) = \prod_{i=1}^n f(y_i) = \prod_{i=1}^n \phi(y_i;\mu_i,\sigma^2)$ with $\mu_i = \underline{\tilde{x}}_i^T \underline{\beta}$

likelihood

$$L(\theta) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(y_i-\mu_i)^2\right\}$$

$$= (2\pi)^{-\frac{n}{2}} (\sigma^2)^{-\frac{n}{2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^n (y_i-\mu_i)^2\right\} \qquad \mu_i = \underline{\tilde{x}}_i^T \underline{\beta}$$

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

log-likelihood

$$\ell(\theta) = \cancel{-\frac{n}{2}\log 2\pi} - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^n (y_i-\underline{\tilde{x}}_i^T\underline{\beta})^2$$

$$= -\frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^n (y_i-\underline{\tilde{x}}_i^T\underline{\beta})^2$$

$$\sum_{i=1}^n (y_i-\underline{\tilde{x}}_i^T\underline{\beta})^2 = (\underline{y}-X\underline{\beta})^T(\underline{y}-X\underline{\beta}) = S(\underline{\beta})$$ sum of squared residuals

for fixed σ^2, maximizing the likelihood is equivalent to minimizing $S(\underline{\beta})$, independently of the value of σ^2

$$\Rightarrow \hat{\underline{\beta}} = \underset{\underline{\beta}\in\mathbb{R}^p}{\operatorname{argmin}}\, S(\underline{\beta})$$

Similar to the simple linear model, the ML estimators are the same that we obtain minimizing the sum of squared residuals (OLS estimation).

To find $\hat{\underline{\beta}}$ we need to solve $\frac{\partial}{\partial\underline{\beta}} S(\underline{\beta}) = 0$

where $S(\underline{\beta}) = (\underline{y}-X\underline{\beta})^T(\underline{y}-X\underline{\beta}) =$

$$= \underline{y}^T\underline{y} - \underbrace{\underline{y}^T X\underline{\beta}}_{1\times 1} - \underbrace{\underline{\beta}^T X^T \underline{y}}_{1\times 1} + \underline{\beta}^T X^T X \underline{\beta}$$

$$= \underline{y}^T\underline{y} - 2\underline{y}^T X\underline{\beta} + \underline{\beta}^T X^T X\underline{\beta}$$

useful properties of derivatives:

consider: $\underline{a}$ $(p\times 1)$ vector of constants

A $(p\times p)$ matrix of constants

- $\frac{\partial}{\partial\underline{\beta}} \underline{a}^T\underline{\beta} = \underline{a}$ $(p\times 1)$
- $\frac{\partial}{\partial\underline{\beta}} \underline{\beta}^T A \underline{\beta} = 2A\underline{\beta}$ $(p\times 1)$

Hence,

$$\frac{\partial}{\partial\underline{\beta}} S(\underline{\beta}) = \frac{\partial}{\partial\underline{\beta}}\left(\underline{y}^T\underline{y} - 2\underbrace{\underline{y}^T X}_{1\times p}\underline{\beta} + \underline{\beta}^T \underbrace{X^T X}_{p\times p}\underline{\beta}\right)$$

$$= 0 - 2X^T\underline{y} + 2X^TX\underline{\beta}$$

$\Rightarrow \frac{\partial}{\partial\underline{\beta}} S(\underline{\beta}) = 0$

$\Rightarrow X^TX\underline{\beta} = X^T\underline{y}$

$\Rightarrow \hat{\underline{\beta}} = (X^TX)^{-1}X^T\underline{y}$

notice that

$$-2X^T(\underline{y}-X\underline{\beta}) = 0 \Rightarrow \begin{cases}\underline{x}_1^T(\underline{y}-X\underline{\beta}) = 0 \\ \vdots \\ \underline{x}_p^T(\underline{y}-X\underline{\beta}) = 0\end{cases}$$

"normal equations"

Remark

To solve the equation, X^TX has to be nonsingular (invertible).

This is ensured by assumption ③ of ABSENCE of MULTICOLLINEARITY (i.e. $\operatorname{rank}(X) = p$).

We have found a critical point. Is it a minimum?

Hessian: $\frac{\partial^2}{\partial\underline{\beta}\partial\underline{\beta}^T} S(\underline{\beta}) = \frac{\partial^2}{\partial\underline{\beta}\partial\underline{\beta}^T}\left(-2X^T\underline{y} + 2X^TX\underline{\beta}\right) = 2X^TX \Big|_{\underline{\beta}=\hat{\underline{\beta}}} = 2X^TX$ has to be positive definite

Recall: Σ is positive definite if $\forall \underline{a}\neq\underline{0}$, $\underline{a}^T\Sigma\underline{a} > 0$

does it hold for X^TX?

$\underline{a}^TX^TX\underline{a} = (X\underline{a})^T(X\underline{a}) \geq 0$ and it is $=0 \iff X\underline{a} = 0$

since we required X to have full rank $\Rightarrow X\underline{a} = 0 \iff \underline{a} = \underline{0}$

$\Rightarrow \underline{a}^TX^TX\underline{a} > 0 \Rightarrow 2X^TX$ is positive definite

$\Rightarrow \hat{\underline{\beta}} = (X^TX)^{-1}X^T\underline{y}$ is the minimum of $S(\underline{\beta})$

and the MAXIMUM LIKELIHOOD ESTIMATE

The maximum likelihood estimator is $\hat{\underline{B}} = (X^TX)^{-1}X^T\underline{Y}$

- ESTIMATE of σ^2

$$\ell(\theta) = \ell(\underline{\beta},\sigma^2) = -\frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}(\underline{y}-X\underline{\beta})^T(\underline{y}-X\underline{\beta})$$

$$\ell(\hat{\underline{\beta}},\sigma^2) = -\frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}(\underline{y}-X\hat{\underline{\beta}})^T(\underline{y}-X\hat{\underline{\beta}})$$

$$\frac{\partial}{\partial\sigma^2}\ell(\hat{\underline{\beta}},\sigma^2) = -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}(\underline{y}-X\hat{\underline{\beta}})^T(\underline{y}-X\hat{\underline{\beta}})$$

$$\frac{\partial}{\partial\sigma^2}\ell(\hat{\underline{\beta}},\sigma^2) = 0$$

$$\Rightarrow -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}(\underline{y}-X\hat{\underline{\beta}})^T(\underline{y}-X\hat{\underline{\beta}}) = 0$$

$$\Rightarrow -\frac{1}{2(\sigma^2)^2}\left[n\sigma^2 - (\underline{y}-X\hat{\underline{\beta}})^T(\underline{y}-X\hat{\underline{\beta}})\right] = 0 \quad\Rightarrow\quad \hat{\sigma}^2 = \frac{(\underline{y}-X\hat{\underline{\beta}})^T(\underline{y}-X\hat{\underline{\beta}})}{n} = \frac{\underline{e}^T\underline{e}}{n}$$

$$\frac{\partial^2}{\partial(\sigma^2)^2} = \frac{n}{2(\sigma^2)^2} - \frac{\cancel{2}}{\cancel{2}(\sigma^2)^3}(\underline{y}-X\hat{\underline{\beta}})^T(\underline{y}-X\hat{\underline{\beta}})$$

$$= \frac{n}{2(\sigma^2)^2} - \frac{1}{(\sigma^2)^3}\cdot n\hat{\sigma}^2$$

$$\frac{\partial^2}{\partial(\sigma^2)^2}\ell(\sigma^2;\hat{\underline{\beta}})\Big|_{\sigma^2=\hat{\sigma}^2} \Rightarrow \frac{n}{2(\hat{\sigma}^2)^2} - \frac{n\cancel{\hat{\sigma}^2}}{(\hat{\sigma}^2)^{\cancel{3}\,2}} = \frac{n}{2(\hat{\sigma}^2)^2} - \frac{n}{(\hat{\sigma}^2)^2} = -\frac{n}{2(\hat{\sigma}^2)^2} < 0 \quad \text{it's a max}$$

The maximum likelihood estimator is $\hat{\Sigma}^2 = \frac{(\underline{Y}-X\hat{\underline{B}})^T(\underline{Y}-X\hat{\underline{B}})}{n} = \frac{\underline{E}^T\underline{E}}{n}$

Similarly to the case of the simple linear model, one can show that $\hat{\Sigma}^2$ is biased:

$$\mathbb{E}[\hat{\Sigma}^2] = \frac{n-p}{n}\sigma^2$$

We can define an UNBIASED estimator of the variance: $S^2 = \frac{(\underline{Y}-X\hat{\underline{B}})^T(\underline{Y}-X\hat{\underline{B}})}{n-p} = \frac{\underline{E}^T\underline{E}}{n-p} = \frac{n}{n-p}\hat{\Sigma}^2$

the denominator is
$n - \#$ columns of X

Remarks

- the normal equations imply $\begin{cases}(\underline{y}-X\hat{\underline{\beta}})^T\underline{x}_1 = 0 & \rightarrow \underline{e}^T\underline{x}_1 = 0 \\ \vdots & \\ (\underline{y}-X\hat{\underline{\beta}})^T\underline{x}_p = 0 & \rightarrow \underline{e}^T\underline{x}_p = 0\end{cases} \Rightarrow \underline{e}^TX = 0$

  $\Rightarrow$ ORTHOGONALITY between the residuals and the columns of X

- if we include the intercept $\underline{x}_1 = \underline{1}$

  $\underline{e}^T\underline{x}_1 = 0 \Rightarrow \underline{e}^T\underline{1} = 0 \Rightarrow \sum_{i=1}^n e_i = 0 \Rightarrow \bar{e} = 0 \Rightarrow$ the residuals have mean $= 0$

- the vector of the predicted values is

$$\hat{\underline{y}} = \hat{\underline{\mu}} = X\hat{\underline{\beta}}$$

$$= \underbrace{X\cdot(X^TX)^{-1}X^T}_{P}\underline{y} = \underset{n\times n}{P}\cdot\underset{n\times 1}{\underline{y}}$$

The matrix P is:

- symmetric: $P^T = (X(X^TX)^{-1}X^T)^T = X(X^TX)^{-1}X^T = P$
- idempotent: $P\cdot P = X\cancel{(X^TX)^{-1}X^TX}(X^TX)^{-1}X^T = X(X^TX)^{-1}X^T = P$

P is called the PROJECTION MATRIX (details in the next class...)

- the vector of the residuals is

$$\underline{e} = \underline{y} - \hat{\underline{y}} = \underline{y} - X\hat{\underline{\beta}} = \underline{y} - P\underline{y} = (I_n - P)\underline{y}$$

The matrix $(I_n - P)$ is:

- symmetric: $(I_n-P)^T = I_n^T - P^T = I_n - (X(X^TX)^{-1}X^T)^T = I_n - X(X^TX)^{-1}X^T = I_n - P$
- idempotent: $(I_n-P)(I_n-P) = I_n - P - P + P^2 = I_n - 2P + \underset{=P}{P^2} = I_n - 2P + P = I_n - P$