Multiple linear regression

Jean-Michel Marin

University of Montpellier
Faculty of Sciences

HAX912X - 2024/2025

1. Homoscedastic models
   - Hypotheses
   - Estimations of β and σ^2
   - Geometric properties

2. Heteroscedastic models

Homoscedastic models

Data

- n individuals : $i = 1, \dots, n$
- $y_i \in \mathbb{R}$ variable to explain (response)
- $\mathbf{x}_i = (x_{i1}, \dots, x_{ip}) \in \mathbb{R}^p$ values of p explanatory variables including the constant term constant $x_{i1} = 1, \forall i = 1, \dots, n$

Matrix notation

$$\mathbf{y} = (y_1, \dots, y_n) \quad ; \quad \mathbf{X} = \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \vdots \\ \mathbf{x}_n \end{bmatrix} = \begin{bmatrix} 1 & x_{12} & \dots & x_{1p} \\ 1 & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n2} & \dots & x_{np} \end{bmatrix}$$

Homoscedastic models

The multiple linear regression model is such that

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{u}$$

$\beta = (\beta_1, \ldots, \beta_p)$ is an unknown parameter
$\mathbf{u} = (u_1, \ldots, u_n)$ vector of residuals

We can write, $\forall i = 1, \ldots, n$

$$y_i = \beta_1 + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + u_i$$

y_i is a random variable for which we have a realization
$\mathbf{x}_i$ is a fixed vector

Homoscedastic models
Hypotheses

- the constant regressor is present in all models
- $n > p$ more observations than variables
- $\mathbf{X}^\mathsf{T}\mathbf{X}$ invertible, the rank of X is equal to p, the model is identifiable
- $\mathbb{E}[\mathbf{u}] = 0_n$ residuals are centered
- $\mathbb{E}[\mathbf{u}\mathbf{u}^\mathsf{T}] = \mathbb{V}(\mathbf{u}) = \sigma^2 I_n$ residuals are decorrelated and have the same variance σ^2

Homoscedastic models
Estimations of β and σ^2

Ordinary least squares (OLS) estimator of β

$$\hat{\beta} \in \arg\min_{\beta \in \mathbb{R}^p} \|\mathbf{y} - \mathbf{X}\beta\|^2 \left(= \arg\min_{\beta \in \mathbb{R}^p} \sum_{i=1}^{n} \left(y_i - \mathbf{x}_i^T \beta\right)^2 \right)$$

$$\boxed{\hat{\beta} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1} \mathbf{X}^T \mathbf{y}}$$

Unbiased estimator of σ^2

$$\boxed{\hat{\sigma}^2 = \frac{\|\mathbf{y} - \mathbf{X}\hat{\beta}\|^2}{n-p} = \frac{\mathrm{RSS}}{n-p}}$$

Estimator of the standard error of $\hat{\beta}_j$ $\boxed{\hat{\sigma}(\hat{\beta}_j) = \sqrt{\hat{\sigma}^2 \left(\mathbf{X}^T\mathbf{X}\right)^{-1}_{(j,j)}}}$

Homoscedastic models
Geometric properties

χ vector subspace of $\mathbb{R}^n$ generated by the columns of $\mathbf{X}$
$\mathrm{H} = \mathbf{X}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top$ orthogonal projector on χ
$\mathrm{K} = \mathrm{I}_n - \mathrm{H}$ projector on $\chi^\perp$

$$\boxed{\hat{\mathbf{y}} = \mathbf{X}\hat{\beta} = \mathbf{X}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{y} = \mathrm{H}\mathbf{y}}$$

$\hat{\mathbf{y}}$ orthogonal projection of $\mathbf{y}$ on χ

$$\boxed{\hat{\mathbf{u}} = \mathbf{y} - \hat{\mathbf{y}} = \mathbf{y} - \mathrm{H}\mathbf{y} = (\mathrm{I}_n - \mathrm{H})\mathbf{y} = \mathrm{K}\mathbf{y}}$$

$\hat{\mathbf{u}}$ orthogonal projection of $\mathbf{y}$ on $\chi^\perp$

$$\boxed{\hat{\mathbf{y}} \perp \hat{\mathbf{u}}}$$

Homoscedastic models
Geometric properties

According to the Pythagorean theorem

$$\|\mathbf{y} - \bar{\mathbf{y}}\mathbf{1}_n\|^2 = \|\mathbf{y} - \hat{\mathbf{y}}\|^2 + \|\hat{\mathbf{y}} - \bar{\mathbf{y}}\mathbf{1}_n\|^2$$

$$\boxed{\mathrm{TSS} = \mathrm{RSS} + \mathrm{ESS}}$$

Coefficient of determination

$$\boxed{R^2 = \frac{\mathrm{ESS}}{\mathrm{TSS}} = 1 - \frac{\mathrm{RSS}}{\mathrm{TSS}}}$$

Heteroscedastic models

$$\boxed{\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}}$$

$n > p$, $\mathbf{X}$ of full rank, identifiable model

$\mathbb{E}(\mathbf{u}) = 0_n$

$\mathbb{V}(\mathbf{u}) = \mathbb{E}\left[\mathbf{u}\mathbf{u}^T\right] = \sigma^2\Omega$ and $\Omega \neq I_n$

$\mathbb{C}(u_i, u_j) = \sigma^2\Omega_{ij}$ **heteroccedasticity and error correlation**

Heteroscedastic models

It is assumed that Ω is known

Let R be such that $\Omega^{-1} = R^T R$

Set $\mathbf{y}^* = R\mathbf{y}$, $\mathbf{X}^* = R\mathbf{X}$ and $\mathbf{u}^* = R\mathbf{u}$.

$$\mathbf{y}^* = R\mathbf{y} = R(\mathbf{X}\beta + \mathbf{u}) = R\mathbf{X}\beta + R\mathbf{u} = \mathbf{X}^*\beta + \mathbf{u}^*$$

$$\mathbb{E}(\mathbf{u}^*) = 0_n \quad ; \quad \mathbb{V}(\mathbf{u}^*) = R\mathbb{V}(\mathbf{u})R^T = \sigma^2 I_n$$

We have decorrelated the residuals and returned to the case of classical linear regression (homoscedastic)

Heteroscedastic models

We can calculate the Ordinary Least Squares (OLS) estimator for the homoscedastic model

$$\boxed{\mathbf{y}^* = \mathbf{X}^* \beta + \mathbf{u}^*}$$

$$\hat{\beta} \in \arg\min_{\beta \in \mathbb{R}^p} \|\mathbf{y}^* - \mathbf{X}^* \beta\|^2$$

$$\boxed{\hat{\beta} = \left((\mathbf{X}^*)^\mathsf{T} \mathbf{X}^*\right)^{-1} (\mathbf{X}^*)^\mathsf{T} \mathbf{y}^* = \left(\mathbf{X}^\mathsf{T} \Omega^{-1} \mathbf{X}\right)^{-1} \mathbf{X}^\mathsf{T} \Omega^{-1} \mathbf{y}}$$

This is the Generalized Least Squares (GLS) estimator of β

Heteroscedastic models

It's difficult to estimate Ω without additional constraints. Ω of particular parametric forms

General estimation method

If we have an estimator $\hat{\Omega}(\boldsymbol{\beta})$ of Ω, we can iterate until stabilization. we can iterate to stabilize

- estimate β with

$$\hat{\boldsymbol{\beta}}^{(t)} = \left(\mathbf{X}^T \left[\hat{\Omega}\left(\hat{\boldsymbol{\beta}}^{(t-1)}\right)\right]^{-1} \mathbf{X}\right)^{-1} \mathbf{X}^T \left[\hat{\Omega}\left(\hat{\boldsymbol{\beta}}^{(t-1)}\right)\right]^{-1} \mathbf{y}$$

- estimate Ω with

$$\hat{\Omega}\left(\hat{\boldsymbol{\beta}}^{(t)}\right)$$