Regression Analysis

Multiple Linear Regression

Nicoleta Serban, Ph.D.
Professor
School of Industrial and Systems Engineering

Regression Parameter Estimation

1

About This Lesson

2

Parameter Estimation $(\beta_0, \beta_1, \ldots, \beta_p), \sigma^2$

To estimate $(\beta_0, \beta_1, \ldots, \beta_p)$, find values $(\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p)$ that minimize squared error:

$$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} \left(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{i,1} + \ldots + \hat{\beta}_p x_{i,p})\right)^2 = \left(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\right)^{\mathrm{T}}\left(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\right)$$

By linear algebra (Orthogonal Decomposition Theorem) or differentiation:

$$\boldsymbol{X}^{\mathrm{T}}(\boldsymbol{y} - \widehat{\boldsymbol{y}}) = \boldsymbol{X}^{\mathrm{T}}\left(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\right) = 0$$

So

$$\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\widehat{\boldsymbol{\beta}} = \boldsymbol{X}^{\mathrm{T}}\boldsymbol{y}$$

If $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ is invertible,

$$\boxed{\widehat{\boldsymbol{\beta}} = \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y}}$$

Georgia Tech

3

Parameter Estimation $(\beta_0, \beta_1, \ldots, \beta_p), \sigma^2$

The fitted values are $\widehat{\boldsymbol{y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}$, and $\widehat{\boldsymbol{\beta}} = \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y}$, so

$$\widehat{\boldsymbol{y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}} = \boldsymbol{X}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y} = \mathbf{H}\boldsymbol{y}$$

where $\mathbf{H} \equiv \boldsymbol{X}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}$ is called the **hat matrix** because multiplying $\boldsymbol{y}$ by $\mathbf{H}$ gives $\widehat{\boldsymbol{y}}$.

The residuals are:

$$\hat{\boldsymbol{\varepsilon}} = \boldsymbol{y} - \widehat{\boldsymbol{y}} = \boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}} = \boldsymbol{y} - \mathbf{H}\boldsymbol{y} = (\mathbf{I} - \mathbf{H})\boldsymbol{y}$$

To estimate σ^2,

$$\widehat{\sigma}^2 = \hat{\boldsymbol{\varepsilon}}^{\mathrm{T}}\hat{\boldsymbol{\varepsilon}}/(n - p - 1)$$

Georgia Tech

4

Parameter Estimation $(\beta_0, \beta_1, \ldots, \beta_p), \sigma^2$

The fitted values are $\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$, and $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y}$, so

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

where $\mathbf{H} \equiv \boldsymbol{X}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}$ is called the **hat matrix** because multiplying $\boldsymbol{y}$ by $\mathbf{H}$ gives $\hat{\boldsymbol{y}}$.

The residuals are:

$$\hat{\boldsymbol{\varepsilon}} = \boldsymbol{y} - \hat{\boldsymbol{y}} = \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{y} - \mathbf{H}\boldsymbol{y} = (\mathbf{I} - \mathbf{H})\boldsymbol{y}$$

To estimate σ^2,

$$\hat{\sigma}^2 = \hat{\boldsymbol{\varepsilon}}^{\mathrm{T}}\hat{\boldsymbol{\varepsilon}}/(n - p - 1)$$

The estimator of σ^2 is MSE

Assuming $\varepsilon_1, \ldots, \varepsilon_n$ are normally distributed

- **MSE ~ χ^2 with *n-p-1* degrees of freedom**

Parameter Estimation

$$\hat{\sigma}^2 = \frac{\hat{\boldsymbol{\varepsilon}}^{\mathrm{T}}\hat{\boldsymbol{\varepsilon}}}{n-p-1} = \frac{\sum \hat{\boldsymbol{\varepsilon}}_i^2}{n-p-1} \sim \chi^2_{n-p-1}$$

(chi-squared distribution with *n-p*-1 degrees of freedom)

Assuming $\hat{\varepsilon}_i \sim \varepsilon_j \sim \mathrm{N}(0, \sigma^2)$

Estimating σ^2 ← Sample variance

This is the sample variance estimator, except we use *n-p*-1 degrees of freedom. **Why?**

Parameter Estimation

Recall that $\varepsilon_i = \left(y_i - (\beta_0 + \beta_1 x_{i,1} + \cdots + \beta_p x_{i,p}) \right)$

Replaced by $\hat{\varepsilon}_i = \left(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{i,1} + \cdots + \hat{\beta}_p x_{i,p}) \right)$

Use p+1 degrees of freedom because

$$\beta_0 \leftarrow \hat{\beta}_0$$
$$\beta_1 \leftarrow \hat{\beta}_1$$
$$\vdots$$
$$\beta_p \leftarrow \hat{\beta}_p$$

Thus, assuming that

$$\varepsilon_i \sim \mathrm{N}(0, \sigma^2)$$

$$\rightarrow \hat{\sigma}^2 = \mathrm{MSE} \sim \chi^2_{n-p-1}$$

(This is called the sampling distribution of $\hat{\sigma}^2$.)

Georgia Tech

7

Summary

Georgia Tech

8