Multiple Linear Regression

Lecture 5

Alexandra Chronopoulou

ILLINOIS

COLLEGE OF LIBERAL ARTS & SCIENCES

Department of Statistics
101 Illini Hall, MC-374
725 S. Wright St.
Champaign, IL 61820-5710

Learning objectives

In this lecture we will:

- Review of Random Vectors' Mean and Variance.
- Properties of LS Estimators
- The Gauss-Markov Theorem
- Maximum Likelihood in MLR

Least-Square Estimates

- In MLR the LS estimate $\hat{\beta}$ is given by

$$\hat{\beta} = \left(\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_p\right)^T == (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T y$$

- $\hat{\beta}$ is a random vector, since it is a function of $\mathbf{y}$ (which is random).

- For hypothesis testing, we need to understand the obtain distribution of $\hat{\beta}$.

Mean & Variance of Random Vectors: Review

Random Vectors: Mean

Let $\mathbf{Z}$ a random vector of size $m \times 1$, with components $Z_1, Z_2, \ldots, Z_m$. The mean of $\mathbf{Z}$ is equal to vector μ defined as:

$$\mu = \mathbb{E}(\mathbf{Z}) = \begin{pmatrix} \mathbb{E}(Z_1) \\ \mathbb{E}(Z_2) \\ \ldots \\ \mathbb{E}(Z_m) \end{pmatrix}$$

Variance of a Random Vector

The Variance of a random vector $\mathbf{Z}$ is a **matrix** – the Variance-Covariance matrix. This matrix is *symmetric* (why?) of size $m \times m$ with component (i, j) equal to the $\mathrm{Cov}(\mathrm{Z_i}, \mathrm{Z_j})$. Specifically,

$$
\begin{aligned}
\Sigma_{m\times m} = \mathrm{Cov}(\mathbf{Z}) &= \mathbb{E}\left((\mathbf{Z}-\mu)(\mathbf{Z}-\mu)^{\mathsf{T}}\right) \\
&= \begin{pmatrix} \mathrm{Var}(\mathrm{Z_1}) & \dots & \mathrm{Cov}(\mathrm{Z_1}, \mathrm{Z_m}) \\ \dots & \dots & \dots \\ \mathrm{Cov}(\mathrm{Z_m}, \mathrm{Z_1}) & \dots & \mathrm{Var}(\mathrm{Z_m}) \end{pmatrix}
\end{aligned}
$$

Affine Transformation of Random Vectors

Affine transformation Z

$$\mathbf{W} = \mathbf{a_{n\times 1}} + \mathbf{B_{n\times m}}\mathbf{Z_{m\times 1}}$$

– Mean & Covariance Matrix of **W**

$$\mathbb{E}(\mathbf{W}) = \mathbf{a} + \mathbf{B}\mu, \quad Cov(\mathbf{W}) = \mathbf{B}\mathbf{\Sigma}\mathbf{B}^T$$

Another transformation of Z

$$W = \mathbf{v}^{\mathbf{T}}\mathbf{Z} = v_1 Z_1 + v_2 Z_2 + \ldots + v_m Z_m$$

– Mean & Variance of W

$$\mathbb{E}(W) = \mathbf{v}^T\mu = \sum_{i=1}^{m} v_i \mu_i$$

$$\mathrm{Var(W)} = \mathbf{v}^T\mathbf{\Sigma}\mathbf{v} = \sum_{i=1}^{m} v_i^2 \mathrm{Var(Z_i)} + 2\sum_{\mathrm{i<j}} \mathrm{v_i v_j Cov(Z_i, Z_j)}$$

Linear Regression Model

$$\mathbf{y} = \mathbf{X}\beta + \varepsilon,$$

with $\mathbb{E}(\varepsilon) = \mathbf{0}$, and $\mathrm{Cov}(\varepsilon) = \sigma^2 \mathbf{I}_\mathrm{n}$.

– These assumptions imply that the response has mean and variance equal to:

$$\mathbb{E}(\mathbf{y}) = \mathbf{X}\beta, \quad Cov(\mathbf{y}) = \sigma^2 \mathbf{I}_n$$

LS Estimators

The LS estimators $\hat{\beta}$ are *unbiased*.

Indeed,

$$\begin{aligned}\mathbb{E}(\hat{\beta}) &= \mathbb{E}\left((\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{y}\right) = (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbb{E}(\mathbf{y}) \\ &= (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{X}\beta = \beta\end{aligned}$$

Mean & Covariance of the LS estimates (Cont.)

Variance-Covariance Matrix of $\hat{\beta}$

$$
\begin{aligned}
\mathrm{Cov}(\hat{\beta}) &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T \mathrm{Cov}(\mathbf{y}) \left((\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\right)^\mathsf{T} \\
&= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T \, \sigma^2 \, \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} \\
&= \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} \\
&= \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}
\end{aligned}
$$

Using the previous results we can also show the following properties for the fitted values $\hat{y}$ and the residuals r:

(a) $\mathbb{E}(\hat{\mathbf{y}}) = \mathbf{X}\beta$

(b) $\text{Cov}(\hat{\mathbf{y}}) = \sigma^2\mathbf{H}$

(c) $\mathbb{E}(\mathbf{r}) = \mathbf{0}$

(d) $Cov(\mathbf{r}) = \sigma^2(\mathbf{I_n} - \mathbf{H})$

(e) $\mathbb{E}(\hat{\sigma}^2) = \frac{1}{n-p}\,\mathbb{E}(\mathbf{r}^\mathbf{T}\mathbf{r}) = \frac{1}{n-p}\sigma^2(n-p) = \sigma^2$

Remark: It can be shown that $\frac{\mathbf{r}^\mathbf{T}\mathbf{r}}{\sigma^2} = \frac{RSS}{\sigma^2} \sim \chi^2_{n-p}$

Standard Error of $\hat{\beta}_1$

- $\hat{\beta}$ and $\hat{\sigma}^2$ are unbiased estimators of β and σ^2 respectively.
- We can plug-in the variance estimator $\hat{\sigma}^2$ to get an estimator for the covariance of $\hat{\beta}$.
- The standard errors of the $\hat{\beta}_i$ are the square roots of the elements of the diagonal of the covariance matrix $Cov(\hat{\beta}) = \hat{\sigma}^2(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}$.
  For example:
  $$se(\hat{\beta}_1) = \hat{\sigma}\sqrt{((\mathbf{X}^\mathsf{T}\mathbf{X})^{-1})_{11}}$$

If the errors are *uncorrelated*, have *equal variance* and *mean equal to zero*, the LS estimators have **the lowest variance within the class of linear estimators**.

- Suppose we are interested in estimating a linear combination of β of the form:

  $$\theta = \mathbf{c}^T \beta = \sum_{j=1}^{p} c_j \beta_j$$

  For example, estimating any element of β and estimating the mean response at a new value x^* are all special cases of this setup.

- Naturally, we can form an estimate of θ by plugging in the LS estimate β in the equation for θ:

$$\hat{\theta}_{LS} = \mathbf{c}^T \hat{\beta} = \mathbf{c}^\mathbf{T}(\mathbf{X}^\mathbf{T}\mathbf{X})^{-\mathbf{1}}\mathbf{X}^\mathbf{T}\mathbf{y}$$

This is a linear[1] and unbiased estimator of θ. Its mean square error can be calculated as:

$$MSE(\hat{\theta}_{LS}) = \mathbb{E}(\hat{\theta}_{LS} - \theta)^2 = Var(\hat{\theta}_{LS})$$

[1] It is a linear combination of the n data points $y_1, y_2, \ldots, y_n$

- Suppose there is another estimate of θ, which is also linear and unbiased. The following Theorem states that $\hat{\theta}_{LS}$ is always better in the sense that its MSE is always smaller (or at least, not bigger).

Gauss-Markov Theorem

The estimator $\hat{\theta}_{LS} = \mathbf{c}^T \hat{\beta}$ is the BLUE (best linear unbiased estimator) of the parameter $\mathbf{c}^T \beta$ for any vector $\mathbf{c} \in \mathbb{R}^p$.

Proof: Please see Supplemental Material.

- Recall the normality assumption for the regression model:

$$y_i = \mathbf{x}_i^T\beta + \varepsilon_i \; i = 1, \ldots, n, \text{ with } \varepsilon_i \sim N(0, \sigma^2)$$

- This implies that $\mathbf{y} \sim \mathbf{N_n}(\mathbf{X}\beta, \sigma^2\mathbf{I})$.
- We can show that the likelihood function can be written as:

$$L(\beta, \sigma^2|\mathbf{y}) \propto \frac{RSS}{n}^{-\frac{n}{2}}$$

- The value of β that maximizes the Likelihood function is the Maximum Likelihood Estimator (MLE) of β.
- This estimator is equal to the LS estimate of β.

Distribution of the Least-Squares estimates

- Recall the assumption for the linear regression model:

$$\mathbf{y} \sim \mathbf{N_n}(\mathbf{X}\beta, \sigma^2\mathbf{I})$$

- Any affine transformation of $\mathbf{y}$ will also have a Normal distribution[2].

- We can show that:

$$\hat{\beta} = (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{Y} \sim \mathbf{N_p}(\beta, \sigma^2(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1})$$
$$\hat{\mathbf{y}} = \mathbf{H}\mathbf{y} \sim \mathbf{N_n}(\mathbf{X}\beta, \sigma^2\mathbf{H})$$
$$\hat{\mathbf{r}} = (\mathbf{I_n} - \mathbf{H})\mathbf{y} \sim \mathbf{N_n}(\mathbf{0}, \sigma^2(\mathbf{I_n} - \mathbf{H}))$$

[2] They will also have a joint Normal distribution

Distribution of the Least-Squares estimates

Indeed, for the fitted values $\hat{\mathbf{y}}$ and the estimated residuals $\hat{\mathbf{e}} = \mathbf{r}$ we can calculate the mean and covariance matrices as follows:

$$\begin{aligned}
\mathbb{E}[\hat{\mathbf{y}}] &= \mathbf{H}\ \mathbb{E}[\mathbf{y}] = \mathbf{H}\mathbf{X}\beta = \mathbf{X}\beta \\
Cov(\hat{\mathbf{y}}) &= \mathbf{H}\sigma^2\mathbf{H}^T = \sigma^2\mathbf{H} \\
\mathbb{E}[\mathbf{r}] &= (\mathbf{I_n} - \mathbf{H})\mathbf{X}\beta = \mathbf{0} \\
Cov(\mathbf{r}) &= (\mathbf{I_n} - \mathbf{H})\sigma^2(\mathbf{I_n} - \mathbf{H})^T = \sigma^2(\mathbf{I_n} - \mathbf{H})
\end{aligned}$$

- Although $\mathbf{r}$ is a vector of dimension n, it always lies in a subspace of dimension $(n-p)$.
- $\mathbf{r}$ behaves like a random vector with a distribution $\mathbf{N}_{n-p}(\mathbf{0}, \sigma^2 \mathbf{I}_{n-p})$, so we have:

$$\hat{\sigma}^2 = \frac{||\mathbf{r}||^2}{n-p} \sim \sigma^2 \frac{\chi^2_{n-p}}{n-p}$$

- It can be show that $\hat{\mathbf{y}}$ and $\mathbf{r}$ are uncorrelated since they are in orthogonal spaces. Since they also have a joint normal distribution, they are independent.[3]

[3] Note that if two random variables are uncorrelated, they are not necessarily independent, unless they have a joint Normal distribution