STAT 410 - Linear Regression

Lecture 6

Meng Li

Department of Statistics

Sep.21, 2017

Inferences in MLR - Multivariate normal

- A random vector $\boldsymbol{Y} = [Y_1, Y_2, \ldots, Y_p]'$ is said to have a multivariate normal (or Gaussian) distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, if its probability density function is given by

$$f(\boldsymbol{Y}) = \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\boldsymbol{Y}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{Y}-\boldsymbol{\mu})\right\}.$$

- Notation: $\boldsymbol{Y} \sim \text{MVN}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where

$$\boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \cdots \\ \mu_p \end{bmatrix}, \quad \boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_p^2 \end{bmatrix}.$$

- Property: for any (non-singular) matrix $\boldsymbol{A}$ and a vector $\boldsymbol{b}$, the random vector $\boldsymbol{AY}+\boldsymbol{b} \sim \text{MVN}(\boldsymbol{A\mu}+\boldsymbol{b}, \boldsymbol{A\Sigma A'})$. (Why?)
- This property implies that any linear combination of $\boldsymbol{Y}$ is normally distributed, including any of its marginal distribution.

Inference in MLR

- In order to allow inference in multiple linear regression, we again assume that $\varepsilon_i \overset{i.i.d.}{\sim} N(0, \sigma^2)$ as in SLR.
- In other words, we assume $\boldsymbol{\varepsilon} \sim \text{MVN}(0, \sigma^2 \boldsymbol{I})$.
- This implies that the response vector $\mathbf{y} \sim \text{MVN}(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{I})$.
- Thus, the LS estimators $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ is a multivariate normal:

$$\hat{\boldsymbol{\beta}} \sim \text{MVN}(\boldsymbol{\beta}, \sigma^2 (\mathbf{X}'\mathbf{X})^{-1}). \quad (1)$$

- Let C_{ij} be the ij-th entry of $(\mathbf{X}'\mathbf{X})^{-1}$, then

$$\text{Var}(\hat{\beta}_i) = \sigma^2 C_{ii}, \quad \text{Cov}(\hat{\beta}_i, \hat{\beta}_j) = \sigma^2 C_{ij}.$$

(Marginal) Hypothesis testing and Confidence interval

- Hypothesis test on any single regression coefficient:

$$H_0 : \beta_j = 0, \quad \text{v.s.} \quad H_1 : \beta_j \neq 0.$$

- Test statistics:

$$t = \frac{\hat{\beta}_j}{se(\hat{\beta}_j)} = \frac{\hat{\beta}_j}{\sqrt{\hat{\sigma}^2 C_{jj}}}.$$

- Reject H_0 if $|t| \geq t_{\alpha/2,n-k-1}$. or use p-value.
- Note: this is a marginal test!
- A $100(1-\alpha)$ percent C.I. for the regression coefficient β_j is

$$\hat{\beta}_j - t_{\alpha/2,n-k-1} se(\hat{\beta}_j) \leq \beta_j \leq \hat{\beta}_j + t_{\alpha/2,n-k-1} se(\hat{\beta}_j).$$

- Note: this is a marginal confidence interval!

Confidence interval - continued

- Let $\mathbf{x}_0 = [x_{00}, x_{01}, \ldots, x_{0k}]'$ be any point at which we will estimate the mean response $\mu_0 = \mathrm{E}(y|\mathbf{x}_0)$, where $x_{00} = 1$ is the intercept term.
- Point estimate: $\hat{\mu}_0 = \mathbf{x}_0' \hat{\boldsymbol{\beta}}$
- A $100(1-\alpha)$ percent CI on the mean response:

$$[\hat{\mu}_0 - t_{\alpha,n-k-1} se(\hat{\mu}_0), \hat{\mu}_0 + t_{\alpha,n-k-1} se(\hat{\mu}_0)],$$

  where $se(\hat{\mu}_0) = \sqrt{\hat{\sigma}^2 \mathbf{x}_0' (\mathbf{X}'\mathbf{X})^{-1} \mathbf{x}_0}$. (Why?)
- A $100(1-\alpha)$ percent prediction interval for a future observation is

$$\hat{y}_0 - t_{\alpha,n-k-1} \sqrt{\hat{\sigma}^2 (1 + \mathbf{x}_0' (\mathbf{X}'\mathbf{X})^{-1} \mathbf{x}_0)} \leq y_0 \leq \hat{y}_0 + t_{\alpha,n-k-1} \sqrt{\hat{\sigma}^2 (1 + \mathbf{x}_0' (\mathbf{X}'\mathbf{X})^{-1} \mathbf{x}_0)},$$

  where $\hat{y}_0 = \hat{\mu}_0 = \mathbf{x}_0' \hat{\boldsymbol{\beta}}$.