STAT 3690 Lecture 16

zhiyanggeezhou.github.io

Zhiyang Zhou (zhiyang.zhou@umanitoba.ca)

Mar 07, 2022

What is a linear model?

- Responses are linear functions with respect to unknown parameters.

Univariate/multiple linear regression (J&W Sec. 7.2–7.5)

- Interested in the relationship between random scalar Y and random q-vector $[X_1, \dots, X_q]^\top$
- Model
  - Population version: $Y \mid X_1, \dots, X_q \sim ([1, X_1, \dots, X_q]\boldsymbol{\beta}, \sigma^2)$, where $\boldsymbol{\beta} = [\beta_0, \dots, \beta_q]^\top$, i.e.,
    - $\mathrm{E}(Y \mid X_1, \dots, X_q) = [1, X_1, \dots, X_q]\boldsymbol{\beta} = \beta_0 + \sum_{j=1}^q X_j \beta_j$
    - $\mathrm{var}(Y \mid X_1, \dots, X_q) = \sigma^2$
  - Sample version $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$
    - $\mathbf{Y} = [Y_1, \dots, Y_n]^\top$ and design matrix

$$\mathbf{X} = \begin{bmatrix} 1 & X_{11} & \cdots & X_{q1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & \cdots & X_{nq} \end{bmatrix}_{n \times (q+1)}$$

      - Independent realizations $[Y_i, X_{i1}, \dots, X_{iq}]^\top \sim [Y, X_1, \dots, X_q]^\top$, $i = 1, \dots, n$
      - $\mathrm{rk}(\mathbf{X}) = q + 1 < p + q + 1 \le n$
    - $\boldsymbol{\varepsilon} = [\varepsilon_1, \dots, \varepsilon_n]^\top \sim (\mathbf{0}_n, \sigma^2 \mathbf{I}_n)$

- Least squares (LS) estimation (no need of normality)
  - $\hat{\boldsymbol{\beta}}_{\mathrm{LS}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Y}$
  - $\hat{\sigma}^2_{\mathrm{LS}} = (n - q - 1)^{-1}(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}_{\mathrm{LS}})^\top(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}_{\mathrm{LS}}) = (n - q - 1)^{-1}\mathbf{Y}^\top(\mathbf{I} - \mathbf{H})\mathbf{Y}$
    - Hat matrix $\mathbf{H} = [h_{ij}]_{n \times n} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$
      - Symmetric
      - Idempotent: $\mathbf{H}^2 = \mathbf{H}\mathbf{H} = \mathbf{H}$
      - $\mathrm{rk}(\mathbf{H}) = \mathrm{rk}(\mathbf{X})$
      - Each eigenvalue of $\mathbf{H}$ is either zero or one
    - $\mathrm{E}(\hat{\sigma}^2_{\mathrm{LS}}) = \sigma^2$

a. Prove that $\hat{\beta}_{LS} = \arg\min_{\beta} Q(\beta) = (Y - X\beta)^T (Y - X\beta) = \arg\min_{\beta} Y^TY - 2\beta^T X^T Y + \beta^T X^T X\beta$

Known: $\frac{\partial Q(\beta)}{\partial \beta} = -2X^TY + 2X^TX\beta$ (Matrix calculus)

Let $\partial Q(\beta)/\partial\beta = 0$, then $X^TY = X^TX\beta$.

So, $\hat{\beta}_{LS} = (X^TX)^{-1}X^TY$ is a stationary point, i.e., a candidate for the minimizer

Actually, for any β,

$$Q(\beta) = (Y - X\hat{\beta}_{LS} + X\hat{\beta}_{LS} - X\beta)^T (Y - X\hat{\beta}_{LS} + X\hat{\beta}_{LS} - X\beta)$$
$$= \{Y - X\hat{\beta}_{LS} + X(\hat{\beta}_{LS} - \beta)\}^T \{Y - X\hat{\beta}_{LS} + X(\hat{\beta}_{LS} - \beta)\}$$
$$= (Y - X\hat{\beta}_{LS})^T(Y - X\hat{\beta}_{LS}) + \underbrace{(\hat{\beta}_{LS} - \beta)^T X^T X(\hat{\beta}_{LS} - \beta)}_{\geq 0} + 2\underbrace{(Y - X\hat{\beta}_{LS})^T X(\hat{\beta}_{LS} - \beta)}_{(1)}$$

$$(1) = (Y - HY)^T H X(\hat{\beta}_{LS} - \beta) = Y^T \underbrace{(I - H)H}_{0} X(\hat{\beta}_{LS} - \beta) = 0$$

That is, $Q(\beta) \geq (Y - X\hat{\beta}_{LS})^T(Y - X\hat{\beta}_{LS}) = Q(\hat{\beta}_{LS})$ for all β.

b. Prove that $E(\hat{\sigma}^2_{LS}) = \sigma^2$

$$Y^T(I-H)Y = (Y - X\beta + X\beta)^T (I - H)(Y - X\beta + X\beta)$$
$$= (Y - X\beta)^T(I - H)(Y - X\beta) + 2\beta^T \underbrace{X^T(I - H)}_{0}(Y - X\beta) + \beta^T \underbrace{X^T(I - H)X}_{0}\beta$$
$$= tr\{(I-H)(Y - X\beta)(Y - X\beta)^T\}$$

$$\therefore E\{Y^T(I-H)Y \mid X\} = E[tr\{(I-H)(Y - X\beta)(Y - X\beta)^T\} \mid X]$$
$$= tr[(I-H)\, E\{(Y - X\beta)(Y - X\beta)^T \mid X\}]$$
$$= tr\{(I-H)\, Cov(Y - X\beta \mid X)\}$$
$$= tr\{(I-H)\, Cov(Y \mid X)\}$$
$$= tr\{\sigma^2 (I-H)\} = \sigma^2(n - q - 1)$$

$\therefore E(\hat{\sigma}^2_{LS} \mid X) = \sigma^2$ for all X

$\therefore E(\hat{\sigma}^2_{LS}) = \sigma^2$

- Maximum likelihood (ML) estimation (in need of (conditional) normality)
  - $\hat{\boldsymbol{\beta}}_{\mathrm{ML}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Y} = \hat{\boldsymbol{\beta}}_{\mathrm{LS}}$
    * Given $\mathbf{X}$, $\hat{\boldsymbol{\beta}}_{\mathrm{ML}} \sim MVN_{q+1}(\boldsymbol{\beta}, \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1})$
  - $\hat{\sigma}^2_{\mathrm{ML}} = n^{-1}\mathbf{Y}(\mathbf{I} - \mathbf{H})\mathbf{Y} = n^{-1}(n - q - 1)\hat{\sigma}^2_{\mathrm{LS}}$
    * Given $\mathbf{X}$, $n\hat{\sigma}^2_{\mathrm{ML}}/\sigma^2 = (n - q - 1)\hat{\sigma}^2_{\mathrm{LS}}/\sigma^2 \sim \chi^2(n - q - 1)$

$\because \ln L(X\beta, \sigma) = const - n\ln\sigma - \frac{1}{2\sigma^2}(Y - X\beta)^T(Y - X\beta)$

$\therefore$ for each σ, maximize $\ln L(X\beta, \sigma) \Leftrightarrow$ minimize $(Y - X\beta)^T(Y - X\beta)$

$\therefore \hat{\beta}_{ML} = \hat{\beta}_{LS}$

$\therefore \hat{\sigma}_{ML} = \arg\max_{\sigma} \ln L(X\hat{\beta}_{LS}, \sigma)$

Let $\partial \ln L(X\hat{\beta}_{LS}, \sigma)/\partial\sigma = -n/\sigma + \sigma^{-3}(Y - X\hat{\beta}_{LS})^T(Y - X\hat{\beta}_{LS}) = 0$

Then $\hat{\sigma} = \{(Y - X\hat{\beta}_{LS})^T(Y - X\hat{\beta}_{LS})/n\}^{1/2}$ is a stationary point.

$$\because \partial^2 \ln L(X\hat{\beta}_{LS}, \sigma)/\partial\sigma^2 \Big|_{\sigma=\hat{\sigma}} = n\hat{\sigma}^{-2} - 3\hat{\sigma}^{-4}(Y - X\hat{\beta}_{LS})^T(Y - X\hat{\beta}_{LS})$$
$$= \frac{n^2 - 3n^2}{(Y - X\hat{\beta}_{LS})^T(Y - X\hat{\beta}_{LS})} < 0$$

$\therefore \hat{\sigma} = \arg\max_{\sigma} \ln L(X\hat{\beta}_{LS}, \sigma) = \hat{\sigma}_{ML}$

- Inference (in need of (conditional) normality)
  - Inference on $\boldsymbol{a}^\top \boldsymbol{\beta}$, given $\boldsymbol{a} \in \mathbb{R}^{q+1}$
    - Estimator $\boldsymbol{a}^\top \hat{\boldsymbol{\beta}}_{\mathrm{ML}}$
    - $100(1-\alpha)\%$ confidence interval for $\boldsymbol{a}^\top \boldsymbol{\beta}$:

$$\boldsymbol{a}^\top \hat{\boldsymbol{\beta}}_{\mathrm{ML}} \pm t_{1-\alpha/2, n-q-1} \hat{\sigma}_{\mathrm{LS}} [\boldsymbol{a}^\top (\mathbf{X}^\top \mathbf{X})^{-1} \boldsymbol{a}]^{1/2}$$

  - Inference on $Y_0 = \mathbf{X}_0^\top \boldsymbol{\beta} + \varepsilon_0$ with a new observation vector given $\mathbf{X}_0 = [1, X_{01}, \ldots, X_{0q}]^\top \in \mathbb{R}^{q+1}$
    - Prediction $\hat{Y}_0 = \mathbf{X}_0^\top \hat{\boldsymbol{\beta}}_{\mathrm{ML}}$
    - $100(1-\alpha)\%$ prediction interval for Y_0

$$\mathbf{X}_0^\top \hat{\boldsymbol{\beta}}_{\mathrm{ML}} \pm t_{1-\alpha/2, n-q-1} \hat{\sigma}_{\mathrm{LS}} [1 + \mathbf{X}_0^\top (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}_0]^{1/2}$$

Given X and $a \in \mathbb{R}^{q+1}$,

$$E(a^T\beta - a^T\hat{\beta}_{ML}) = a^T(\beta - E\hat{\beta}_{ML}) = 0_{q+1}$$

$$\begin{aligned} var(a^T\beta - a^T\hat{\beta}_{ML}) &= a^T cov(\beta - \hat{\beta}_{ML}) a \\ &= a^T cov(\hat{\beta}_{ML}) a \\ &= a^T (X^TX)^{-1} X^T (\sigma^2 I) X (X^TX)^{-1} a \\ &= \sigma^2 a^T (X^TX)^{-1} a \end{aligned}$$

$\because$ $a^T\hat{\beta}_{ML}$ is normally distributed

$\therefore$ $a^T\hat{\beta}_{ML} \sim N(a^T\beta, \sigma^2 a^T(X^TX)^{-1}a)$

$\therefore$ $\dfrac{a^T(\hat{\beta}_{ML} - \beta)}{\sigma\{a^T(X^TX)^{-1}a\}^{1/2}} \sim N(0,1)$

$\therefore$ $\dfrac{a^T(\hat{\beta}_{ML} - \beta)}{\cancel{\sigma}\{a^T(X^TX)^{-1}a\}^{1/2}} \Big/ \dfrac{\sqrt{n}\hat{\sigma}_{ML}}{\cancel{\sigma}\sqrt{n-q-1}} = \dfrac{a^T(\hat{\beta}_{ML} - \beta)}{\hat{\sigma}_{LS}\{a^T(X^TX)^{-1}a\}^{1/2}} \sim t(n-q-1)$

$\therefore$ $Pr\left(a^T\hat{\beta}_{ML} - t_{1-\frac{\alpha}{2}, n-q-1} \hat{\sigma}_{LS}\{a^T(X^TX)^{-1}a\}^{1/2} \le a^T\beta \le a^T\hat{\beta}_{ML} + t_{1-\frac{\alpha}{2}, n-q-1} \hat{\sigma}_{LS}\{a^T(X^TX)^{-1}a\}^{1/2}\right) = 1-\alpha$

Similarly, given X and X_0,

$$E(\hat{Y}_0 - Y_0) = E(X_0^T\hat{\beta}_{ML} - X_0^T\beta - \varepsilon_0) = 0$$

$$var(\hat{Y}_0 - Y_0) = var(X_0^T\hat{\beta}_{ML}) + var(\varepsilon_0) = X_0^T cov(\hat{\beta}_{ML}) + \sigma^2 = \sigma^2 X_0^T(X^TX)^{-1}X_0 + \sigma^2$$

$\because$ $\hat{Y}_0 - Y_0$ is normal

$\therefore$ $\hat{Y}_0 - Y_0 \sim N(0, \sigma^2\{1 + X_0^T(X^TX)^{-1}X_0\})$

$\therefore$ $\dfrac{\hat{Y}_0 - Y_0}{\cancel{\sigma}\{1 + X_0^T(X^TX)^{-1}X_0\}^{1/2}} \Big/ \dfrac{\sigma_{LS}\cancel{\sqrt{n-q-1}}}{\cancel{\sigma}\cancel{\sqrt{n-q-1}}} \sim t(n-q-1)$

$\therefore$ $Pr\left(\hat{Y}_0 - t_{1-\frac{\alpha}{2}, n-q-1}\sigma_{LS}\{1 + X_0^T(X^TX)^{-1}X_0\}^{1/2} \le Y_0 \le \hat{Y}_0 + t_{1-\frac{\alpha}{2}, n-q-1}\sigma_{LS}\{1 + X_0^T(X^TX)^{-1}X_0\}^{1/2}\right) = 1-\alpha$