STAT 3690 Lecture Note

Part VI: Linear model

Zhiyang Zhou (zhiyang.zhou@umanitoba.ca, zhiyanggeezhou.github.io)

2023/Mar/08 14:37:31

Multivariate linear regression

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)

- Model (population version):

$$Y \mid X_1, \ldots, X_q \sim \left(\sum_{j=1}^{q} X_j \beta_j, \sigma^2\right)$$

  - Equiv. $Y = \sum_{j=1}^{q} X_j \beta_j + \varepsilon$ with $\varepsilon \perp\!\!\!\perp [X_1, \ldots, X_q]^\top$ and $\varepsilon \sim (0, \sigma^2)$
  - Univariate linear regression: $q = 2$ with $X_1 = 1$
  - Multiple linear regression: $q > 2$ with $X_1 = 1$
- Model (sample version):

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

  - $\boldsymbol{Y} = [Y_1, \ldots, Y_n]^\top$
  - Design matrix

$$\boldsymbol{X} = \begin{bmatrix} X_{11} & \cdots & X_{1q} \\ \vdots & \ddots & \vdots \\ X_{n1} & \cdots & X_{nq} \end{bmatrix}_{n \times q}$$

    * $\mathrm{rk}(\boldsymbol{X}) = q$
  - $\boldsymbol{\beta} = [\beta_1, \ldots, \beta_q]^\top$
  - $\boldsymbol{\varepsilon} = [\varepsilon_1, \ldots, \varepsilon_n]^\top \sim (\mathbf{0}_n, \sigma^2 \mathbf{I}_n)$, independent of $\boldsymbol{X}$

- Least squares (LS) estimation (no need of normality)
  - $\hat{\boldsymbol{\beta}}_{\mathrm{LS}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{Y}$
    * $\mathrm{E}(\hat{\boldsymbol{\beta}}_{\mathrm{LS}} \mid \boldsymbol{X}) = \boldsymbol{\beta}$
  - $\hat{\sigma}^2_{\mathrm{LS}} = (n-q)^{-1} (\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_{\mathrm{LS}})^\top (\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_{\mathrm{LS}}) = (n-q)^{-1} \boldsymbol{Y}^\top (\mathbf{I} - \mathbf{H}) \boldsymbol{Y}$
    * $n \times n$ hat matrix $\mathbf{H} = \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top$
    * $\mathrm{E}(\hat{\sigma}^2_{\mathrm{LS}} \mid \boldsymbol{X}) = \sigma^2$

- ML estimation (under normality)

  - $\hat{\boldsymbol{\beta}}_{\text{ML}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{Y} = \hat{\boldsymbol{\beta}}_{\text{LS}}$
    * $\hat{\boldsymbol{\beta}}_{\text{ML}} \mid \boldsymbol{X} \sim \text{MVN}_q(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}^\top \boldsymbol{X})^{-1})$
  - $\hat{\sigma}^2_{\text{ML}} = n^{-1}\boldsymbol{Y}(\mathbf{I} - \mathbf{H})\boldsymbol{Y} = n^{-1}(n-q)\hat{\sigma}^2_{\text{LS}}$
    * Given $\boldsymbol{X}$, $n\hat{\sigma}^2_{\text{ML}}/\sigma^2 = (n-q)\hat{\sigma}^2_{\text{LS}}/\sigma^2 \sim \chi^2(n-q)$

- Inference (under normality)
  - To infer $\boldsymbol{a}^\top\boldsymbol{\beta}$, given $\boldsymbol{a} \in \mathbb{R}^q$ (e.g., to compare β_1 and β_2 by checking $\boldsymbol{a}^\top\boldsymbol{\beta} = \beta_1 - \beta_2$ with $\boldsymbol{a} = [1, -1, 0, \ldots, 0]^\top$)
    * Estimator: $\boldsymbol{a}^\top\hat{\boldsymbol{\beta}}_{\text{ML}}$
    * $100 \times (1-\alpha)\%$ confidence interval for $\boldsymbol{a}^\top\boldsymbol{\beta}$:

$$\boldsymbol{a}^\top\hat{\boldsymbol{\beta}}_{\text{ML}} \pm \hat{\sigma}_{\text{LS}} \cdot t_{1-\alpha/2,n-q}\sqrt{\boldsymbol{a}^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{a}}$$

  - To predict $Y_0 = \boldsymbol{X}_0^\top\boldsymbol{\beta} + \varepsilon_0$ with $\boldsymbol{X}_0$ different from each row of $\boldsymbol{X}$
    * Prediction: $\hat{Y}_0 = \boldsymbol{X}_0^\top\hat{\boldsymbol{\beta}}_{\text{ML}}$
    * $100 \times (1-\alpha)\%$ prediction interval for Y_0

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

Multivariate linear regression

- Model (population version):

$$Y_1, \ldots, Y_p \mid X_1, \ldots, X_q \sim ([X_1, \ldots, X_q]\mathbf{B}, \boldsymbol{\Sigma})$$

  - Equiv. $[Y_1, \ldots, Y_p] = [X_1, \ldots, X_q]\mathbf{B} + \boldsymbol{\varepsilon}^\top$ with p-vector $\boldsymbol{\varepsilon} \perp\!\!\!\perp [X_1, \ldots, X_q]$ and $\boldsymbol{\varepsilon} \sim (\mathbf{0}_p, \boldsymbol{\Sigma})$
    * Unknown coefficients

$$\mathbf{B} = \begin{bmatrix} b_{11} & \cdots & b_{1p} \\ \vdots & \ddots & \vdots \\ b_{q1} & \cdots & b_{qp} \end{bmatrix}_{q\times p} = \begin{bmatrix} \boldsymbol{b}_{1\cdot}^\top \\ \vdots \\ \boldsymbol{b}_{q\cdot}^\top \end{bmatrix} = \begin{bmatrix} \boldsymbol{b}_{\cdot 1} & \cdots & \boldsymbol{b}_{\cdot p} \end{bmatrix}$$

      · $\boldsymbol{b}_{i\cdot}^\top$: the ith row of $\mathbf{B}$
      · $\boldsymbol{b}_{\cdot j}$: the jth column of $\mathbf{B}$
- Model (sample version):

$$\underset{n\times p}{\boldsymbol{Y}} = \underset{n\times q}{\boldsymbol{X}}\,\underset{q\times p}{\mathbf{B}} + \underset{n\times p}{\boldsymbol{E}}$$

  - Response

$$\boldsymbol{Y} = \begin{bmatrix} Y_{11} & \cdots & Y_{1p} \\ \vdots & \ddots & \vdots \\ Y_{n1} & \cdots & Y_{np} \end{bmatrix}_{n\times p}$$

  - Design matrix

$$\boldsymbol{X} = \begin{bmatrix} X_{11} & \cdots & X_{1q} \\ \vdots & \ddots & \vdots \\ X_{n1} & \cdots & X_{nq} \end{bmatrix}_{n\times q}$$

    * $\text{rk}(\boldsymbol{X}) = q < p + q \le n$
  - Error

$$\boldsymbol{E} = \begin{bmatrix} e_{11} & \cdots & e_{1q} \\ \vdots & \ddots & \vdots \\ e_{n1} & \cdots & e_{nq} \end{bmatrix}_{n\times q} = \begin{bmatrix} \boldsymbol{e}_{1\cdot}^\top \\ \vdots \\ \boldsymbol{e}_{n\cdot}^\top \end{bmatrix}$$

* $\boldsymbol{e}_{i\cdot} \perp\!\!\!\perp [X_{i1}, \ldots, X_{iq}]$
* $\boldsymbol{e}_{i\cdot} \overset{\text{iid}}{\sim} (\mathbf{0}_p, \boldsymbol{\Sigma})$

- Relationship with MANOVA
  - MANOVA models can be expressed as multivariate linear regression with a carefully selected $\boldsymbol{X}$.
- Exercise 6.1: rephrase the following one-way MANOVA model

$$\boldsymbol{Y}_{ij} = \boldsymbol{\mu} + \boldsymbol{\tau}_i + \boldsymbol{E}_{ij}, \quad j = 1, \ldots, n_i, \quad i = 1, \ldots, m$$

into a multivariate linear regression model, where $\boldsymbol{E}_{ij} \overset{\text{iid}}{\sim} \text{MVN}_p(\mathbf{0}, \boldsymbol{\Sigma})$ and $\sum_i \boldsymbol{\tau}_i = 0$.

- LS estimation (no need of normality)
  - $\hat{\mathbf{B}}_{\text{LS}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{Y}$
    * $\text{E}(\hat{\mathbf{B}}_{\text{LS}} \mid \boldsymbol{X}) = \mathbf{B}$
  - $\hat{\boldsymbol{\Sigma}}_{\text{LS}} = (n-q)^{-1}(\boldsymbol{Y} - \boldsymbol{X}\hat{\mathbf{B}}_{\text{LS}})^\top(\boldsymbol{Y} - \boldsymbol{X}\hat{\mathbf{B}}_{\text{LS}}) = (n-q)^{-1}\boldsymbol{Y}^\top(\mathbf{I} - \mathbf{H})\boldsymbol{Y}$
    * $\text{E}(\hat{\boldsymbol{\Sigma}}_{\text{LS}} \mid \boldsymbol{X}) = \boldsymbol{\Sigma}$
- ML estimation (under normality)
  - $\hat{\mathbf{B}}_{\text{ML}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{Y} = \hat{\mathbf{B}}_{\text{LS}}$
  - $\hat{\boldsymbol{\Sigma}}_{\text{ML}} = n^{-1}\boldsymbol{Y}^\top(\mathbf{I} - \mathbf{H})\boldsymbol{Y} = n^{-1}(n-q)\hat{\boldsymbol{\Sigma}}_{\text{LS}}$
    * Given $\boldsymbol{X}$, $n\hat{\boldsymbol{\Sigma}}_{\text{ML}} \sim W_p(\boldsymbol{\Sigma}, n-q)$

- Inference (under normality)
  - To infer $\mathbf{B}^\top \boldsymbol{a}$, given $\boldsymbol{a} \in \mathbb{R}^q$ (e.g., to compare the 1st and 2nd rows of $\mathbf{B}$, i.e., $\boldsymbol{b}_{1\cdot}$ and $\boldsymbol{b}_{2\cdot}$, by checking $\mathbf{B}^\top \boldsymbol{a} = \boldsymbol{b}_{1\cdot} - \boldsymbol{b}_{2\cdot}$ with $\boldsymbol{a} = [1, -1, 0, \ldots, 0]^\top$)
    * Estimator: $\hat{\mathbf{B}}_{\text{ML}}^\top \boldsymbol{a}$
    * $100 \times (1-\alpha)\%$ confidence region for $\mathbf{B}^\top \boldsymbol{a}$

$$\left\{ \boldsymbol{u} \in \mathbb{R}^p : (\boldsymbol{u} - \hat{\mathbf{B}}_{\text{ML}}^\top \boldsymbol{a})^\top \hat{\boldsymbol{\Sigma}}_{\text{LS}}^{-1} (\boldsymbol{u} - \hat{\mathbf{B}}_{\text{ML}}^\top \boldsymbol{a}) \leq \boldsymbol{a}^\top (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{a} \cdot \frac{p(n-q)}{n-p-q+1} F_{1-\alpha, p, n-p-q+1} \right\}$$

  - To predict $\boldsymbol{Y}_0 = \mathbf{B}^\top \boldsymbol{X}_0 + \boldsymbol{E}_0$ with newly observed $\boldsymbol{X}_0 \in \mathbb{R}^q$
    * Prediction: $\hat{\boldsymbol{Y}}_0 = \mathbf{B}_{\text{ML}}^\top \boldsymbol{X}_0$
    * $100 \times (1-\alpha)\%$ prediction region for $\boldsymbol{Y}_0$

$$\left\{ \boldsymbol{u} \in \mathbb{R}^p : (\boldsymbol{u} - \hat{\boldsymbol{Y}}_0)^\top \hat{\boldsymbol{\Sigma}}_{\text{LS}}^{-1} (\boldsymbol{u} - \hat{\boldsymbol{Y}}_0) \leq \{1 + \boldsymbol{X}_0^\top (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}_0\} \cdot \frac{p(n-q)}{n-p-q+1} F_{1-\alpha, p, n-p-q+1} \right\}$$

  - To infer $\boldsymbol{a}^\top \boldsymbol{Y}_0 = \boldsymbol{a}^\top(\mathbf{B}^\top \boldsymbol{X}_0 + \boldsymbol{E}_0)$, given $\boldsymbol{a} \in \mathbb{R}^p$ and newly observed $\boldsymbol{X}_0 \in \mathbb{R}^q$
    * Prediction: $\boldsymbol{a}^\top \hat{\boldsymbol{Y}}_0 = \boldsymbol{a}^\top \mathbf{B}_{\text{ML}}^\top \boldsymbol{X}_0$
    * $100 \times (1-\alpha)\%$ prediction interval for $\boldsymbol{a}^\top \boldsymbol{Y}_0$

$$\boldsymbol{a}^\top \hat{\boldsymbol{Y}}_0 \pm \sqrt{\boldsymbol{a}^\top \hat{\boldsymbol{\Sigma}}_{\text{LS}} \boldsymbol{a} \cdot \{1 + \boldsymbol{X}_0^\top (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}_0\}} \cdot t_{1-\alpha/2, n-q}$$

  - $100 \times (1-\alpha)\%$ simultaneous prediction intervals for $\boldsymbol{a}_k^\top \boldsymbol{Y}_0$, $k = 1, \ldots, m$, given $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_m \in \mathbb{R}^p$ and newly observed $\boldsymbol{X}_0 \in \mathbb{R}^q$
    * (Bonferroni)

$$\boldsymbol{a}_k^\top \hat{\boldsymbol{Y}}_0 \pm \sqrt{\boldsymbol{a}_k^\top \hat{\boldsymbol{\Sigma}}_{\text{LS}} \boldsymbol{a}_k \cdot \{1 + \boldsymbol{X}_0^\top (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}_0\}} \cdot t_{1-\alpha/(2m), n-q}$$

    * (Scheffé's)

$$\boldsymbol{a}_k^\top \hat{\boldsymbol{Y}}_0 \pm \sqrt{\boldsymbol{a}_k^\top \hat{\boldsymbol{\Sigma}}_{\text{LS}} \boldsymbol{a}_k \cdot \{1 + \boldsymbol{X}_0^\top (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}_0\} \cdot \frac{p(n-q)}{n-p-q+1} F_{1-\alpha, p, n-p-q+1}}$$

Testing for nested models

- $H_0 : \mathrm{E}(\boldsymbol{Y} \mid \boldsymbol{X}) = \boldsymbol{X}_{(0)}\mathbf{B}_{(0)}$ (nested model) vs. $H_1 : \mathrm{E}(\boldsymbol{Y} \mid \boldsymbol{X}) = \boldsymbol{X}_{(0)}\mathbf{B}_{(0)} + \boldsymbol{X}_{(1)}\mathbf{B}_{(1)}$ (full model)
  - When $\boldsymbol{X}_{(0)}$ has only the column of ones, we are testing the empty model (i.e., only the intercept) against the full model.
  - When $\boldsymbol{X}_{(1)}$ only contains one column, we are testing for the significance of that variable.
- Likelihood ratio

$$\lambda = \left(\frac{\det \hat{\boldsymbol{\Sigma}}_{\mathrm{ML},H_0}}{\det \hat{\boldsymbol{\Sigma}}_{\mathrm{ML}}}\right)^{-n/2} = \left[\det\left\{(\hat{\boldsymbol{\Sigma}}_{\mathrm{ML},H_0} - \hat{\boldsymbol{\Sigma}}_{\mathrm{ML}})\hat{\boldsymbol{\Sigma}}_{\mathrm{ML}}^{-1} + \mathbf{I}\right\}\right]^{-n/2}$$

- Alternatives to the likelihood ratio
  - Suppose $\eta_1 \geq \cdots \geq \eta_p$ are eigenvalues of $(\hat{\boldsymbol{\Sigma}}_{\mathrm{ML},H_0} - \hat{\boldsymbol{\Sigma}}_{\mathrm{ML}})\hat{\boldsymbol{\Sigma}}_{\mathrm{ML}}^{-1}$
  - Wilks' lambda: $\prod_i (1+\eta_i)^{-1}$
  - Pillai's trace: $\sum_i \{\eta_i(1+\eta_i)^{-1}\}$
  - Hotelling-Lawley trace: $\sum_i \eta_i$
  - Roy's largest root: $\eta_1(1+\eta_1)^{-1}$
  - When $\boldsymbol{X}_{(1)}$ has only one column, all four tests are equivalent; as n increases, all four tests give similar results.

Information criteria

- Akaike's information criterion (AIC)

$$-\ln \mathit{Likelihood} + 2 \times \text{number of parameters to estimate}$$

  - Number of parameters to estimate in $\mathbf{B}$ and $\boldsymbol{\Sigma}$: $pq + p(p+1)/2$
  - The smaller, the better.
- Bayesian information criterion (BIC)

$$-\ln \mathit{Likelihood} + \ln n \times \text{number of parameters to estimate}$$

- Model selection using information criteria proceeds as follows
  - Select models of interest $M_1, \ldots, M_K$. They do not need to be nested.
    * Candidate models should be selected using domain-specific expertise, if possible. Or, you can go through all possible models.
  - Compute the specific information criterion for each model.
  - Select the model with the smallest value of the information criterion.

Multivariate influence measures

- Hat matrix $\mathbf{H} = [h_{ij}]_{n\times n} = \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top$
- Leverage: the influence of $\boldsymbol{Y}_{i\cdot}^\top$ (the ith row of $\boldsymbol{Y}$) on $\hat{\boldsymbol{Y}}_{i\cdot}$ ($= h_{ii}\boldsymbol{Y}_{i\cdot} + \sum_{j\neq i} h_{ij}\boldsymbol{Y}_{j\cdot}$); specifically, $\boldsymbol{Y}_{i\cdot}$ is said to have a high leverage if h_{ii} is large compared to the other diagonal entries of hat matrix $\mathbf{H}$
- (Externally) Studentized residuals

$$T_i^2 = \frac{\hat{\boldsymbol{e}}_{i\cdot}^\top \hat{\boldsymbol{\Sigma}}_{\mathrm{LS},(-i)}^{-1} \hat{\boldsymbol{e}}_{i\cdot}}{1 - h_{ii}}$$

- $\hat{\boldsymbol{e}}_{i\cdot}^\top$: the ith row of residual matrix $\hat{\boldsymbol{E}} = (\mathbf{I} - \mathbf{H})\boldsymbol{Y}$
- $\hat{\boldsymbol{E}}_{(-i)\cdot}^\top$: the remaining part of $\hat{\boldsymbol{E}}$ with Row i removed
- $\hat{\boldsymbol{\Sigma}}_{\mathrm{LS},(-i)} = (n-q-1)^{-1}\hat{\boldsymbol{E}}_{(-i)\cdot}^\top\hat{\boldsymbol{E}}_{(-i)\cdot}$: LS estimator of $\boldsymbol{\Sigma}$ where we have removed Row i from the residual matrix
- The ith observation may be considered as a potential outlier if

$$T_i^2 > \frac{p(n-q-1)}{n-p-q}F_{1-\alpha,p,n-q-1}$$

  * $F_{1-\alpha,p,n-q-1}$: the $1-\alpha$ quantile of $F(p, n-q-1)$

- (Multivariate) Cook's distance

$$D_i = \frac{h_{ii}}{q(1-h_{ii})^2}\hat{\boldsymbol{e}}_{i\cdot}^\top\hat{\boldsymbol{\Sigma}}_{\mathrm{LS}}^{-1}\hat{\boldsymbol{e}}_{i\cdot}$$

  - The Cut-off is far from unique even for multiple linear regression (i.e., the case with $p = 1$)
  - Pay attention to a small set of observations that has substantially higher values than the remaining observations

---

Normality of residuals

- Check the normality of residuals following Lecture Note Part 3
- Apply Box-Cox transformation to colums of $\boldsymbol{Y}$
