STAT 410 - Linear Regression

Lecture 13

Meng Li

Department of Statistics

Oct. 31, 2017

Multicollinearity

- The LS estimator of $\boldsymbol{\beta}$ is $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$.
- Let $\mathbf{X} = [\mathbf{X}_1, \ldots, \mathbf{X}_p]$ where $\mathbf{X}_j$ is the vector to contain the jth regressor.
  - The vectors $\mathbf{X}_1, \ldots, \mathbf{X}_p$ is called *linearly dependent* if there is a set of constants $t_1, \ldots, t_p$ that are not all zero such that

$$\sum_{j=1}^{p} t_j \mathbf{X}_j = 0.$$

- Multicollinearity means the matrix $\mathbf{X}'\mathbf{X}$ is close to a singular matrix where $(\mathbf{X}'\mathbf{X})^{-1}$ does not exist.
- Possible sources:
  - Regressors are just highly correlated in a lot of experiments
  - Over-complete model, e.g., one-way ANOVA without constraint
  - $p > n$, e.g., big data
  - etc.

Consequences

- Numerically, it is not stable.
  - Let the eigenvalues of $\mathbf{X}'\mathbf{X}$ be $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$.
  - We conclude that $\lambda_j \geq 0$ for $j = 1, \ldots, p$ because $\mathbf{X}'\mathbf{X}$ is positive semi-definite.
  - However, some of λ_j could be very small and close to 0.
  - The **condition number** of $\mathbf{X}'\mathbf{X}$ is defined as

$$\kappa = \frac{\lambda \max}{\lambda_{\min}} = \frac{\lambda_1}{\lambda_p}.$$

  - A large condition number indicates the numerical instability when calculating the inverse of $\mathbf{X}'\mathbf{X}$.

- Inflated variance of $\hat{\beta}_j$.
  - Recall: the covariance matrix of $\hat{\boldsymbol{\beta}}$ is $\mathrm{Var}(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$.
  - Let C_{jj} be the (j,j)th entry of $(\mathbf{X}'\mathbf{X})^{-1}$, then $\mathrm{Var}\hat{\beta}_j = \sigma^2 C_{jj}$.
  - Multicollinearity leads to a large value of C_{jj} thus inflate the variance. Consequently, β_j tend to be not significant if we conduct a t test.
  - The **variance inflation factor**, or VIF_j is defined by

$$\text{VIF}_j = \frac{1}{1 - R_j^2},$$

    where R_j is the determination coefficient from the regression of x_j on the other regressors.
  - VIFs help identify which regressors are involved in the multicollinearity.
  - In practice, $\text{VIF} > 5$ is considered an important level of multicollinearity. If the maximum VIF of a given model exceeds 10 it is an indication that multicollinearity may be adversely influencing the least squares estimates.

- Large absolute values of $\hat{\beta}_j$.
  - The expected squared distance of $\hat{\boldsymbol{\beta}}$ from $\boldsymbol{\beta}$ is

$$\mathrm{E}\{(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})'(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})\} = \mathrm{tr}(\mathrm{E}\{(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})'\}) \quad (1)$$

$$= \mathrm{tr}(\mathrm{Var}(\hat{\boldsymbol{\beta}})) = \sigma^2 \mathrm{tr}(\mathbf{X}'\mathbf{X})^{-1}. \quad (2)$$

  - We further simply it to

$$\mathrm{E}\{(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})'(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})\} = \sigma^2 \sum_{j=1}^{p} \frac{1}{\lambda_j},$$

    as the trace of a matrix is equal to the sum of its eigenvalues.
- Multicollinearity may bring a switched sign of $\hat{\beta}_j$ thus affect the model interpretation.

- Let’s go to the R markdown file for various examples.
- What’s next:
  - We shall learn model building techniques to select the best set of variables.
  - We shall use regularization in model fitting, therefore we include all variables in the model and avoid undesirable performances of parameter estimation.