21 Lecture 21: March 15

Last time

- Diagnosing nonlinearity (JF chapter 12)
- Data transformation (JF chapter 4)

Today

- Collinearity (JF chapter 13, RD 8.3.2)
- Principal component analysis (JF 13.1.1, RD 8.3.4)

Additional reference

"A First Course in Linear Model Theory" by Nalini Ravishanker and Kipak K. Dey.

Collinearity

In linear model

$$\begin{aligned}\mathbf{Y} &= \mathbf{X}\beta + \epsilon\\ \epsilon &\sim \mathcal{N}(0, \sigma^2\mathbf{I}_n)\end{aligned}$$

Collinearity (or multicollinearity) exists when there is "near-dependency" between the columns of the design matrix $\mathbf{X}$.

- Two or more columns.
- In other words, high correlation between explanatory variables.
- the data/model pair is ill-conditioned when $\mathbf{X}^T\mathbf{X}$ is nearly singular.

Perfect collinearity leads to rank-deficiency in $\mathbf{X}$ such that $\mathbf{X}^T\mathbf{X}$ is singular. In the case of perfect collinearity, two or more columns are linear-dependent.

An example of perfect collinearity

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + \beta_5 X_{i5} + \epsilon_i$$

Consider the case, where

- Y_i represents the amount of sales.
- $X_{i1}, X_{i2}, ..., X_{i4}$ are categorical that represent the quarter in which the sample is collected: $X_{ij} = \mathbf{1}(\text{sample } i \text{ collected in quarter } j)$.
- X_{i5} represents expense spent in advertising.

The dummy variable trap $X_{i4} = 1 - X_{i1} - X_{i2} - X_{i3}$. Recall that we need $m - 1$ dummy variables for m categories.

An example of high correlation between predictors

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i$$

Consider the case, where

- Y_i represents the salary of individual i.
- X_{i1} represents the age of individual i.
- X_{i2} represents the experience of individual i.

How to interpret β_1?

We expect high correlation between age and experience.

Problems caused by multicollinearity

1. large standard errors of the regression coefficients
   - small associated t-statistics
   - conclusion that truly useful explanatory variables are insignificant in explaining the regression
2. the sign of regression coefficients may be the opposite of what a mechanistic understanding of the problem would suggest
3. deleting a column of the predictor matrix will cause large changes in the coefficient estimates for other variables

However, multicollinearity does **not** greatly affect the **predicted values**.

Signs and detections of multicollinearity

Some signs for multicollinearity:

1. Simple correlation between a pair of predictors exceeds 0.9 or R^2.
2. High value of the multiple correlation coefficient with some high partial correlations between the explanatory variables.
3. Large F-statistics with some small t-statistics for individual regression coefficients

Some approaches for detecting multicollinearity:

1. Pairwise correlations among the explanatory variables
2. Variance inflation factor
3. Condition number

Variance inflation factor

For a multiple linear regression with k explanatory variables. We can regress X_j on the $(k-1)$ other explanatory variables and denote R_j as the coefficient of determination.

Then the variance inflation factor (VIF) is defined as

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

- $\mathrm{VIF}_j \in [1, +\infty)$
- A suggested threshold is 10
- May use the averaged $\overline{\mathrm{VIF}} = \sum_{j=1}^{k} \mathrm{VIF}_j \Big/ k$.

Condition index and condition number

We first scale the design matrix $\mathbf{X}$ into column-equilibrated predictor matrix $\mathbf{X}_E$ such that $\{X_E\}_{ij} = X_{ij}/\sqrt{\mathbf{X}_j^T\mathbf{X}_j}$.

Let $\mathbf{X}_E = \mathbf{U}\mathbf{D}\mathbf{V}^T$ be the singular-value decomposition (SVD) of the $n \times p$ matrix $\mathbf{X}_E$ where $\mathbf{U}^T\mathbf{U} = \mathbf{V}^T\mathbf{V} = \mathbf{I}_p$ and $\mathbf{D} = diag(d_1, d_2, ..., d_p)$ is a diagonal matrix with $d_j \geqslant 0$.

The j^{th} condition index is defined as

$$\eta(\mathbf{X}_E) = d_{\max}/d_j,\ \ j = 1, 2, ..., p$$

The condition number is defined as

$$C = d_{\max}/d_{\min}$$

$C \geqslant 1$, $d_{\max} = \max\limits_{1\leqslant j\leqslant p} d_j$ and $d_{\min} = \min\limits_{1\leqslant j\leqslant p} d_j$

Some properties of the condition number

- Large condition number indicates evidence of multicollinearity
- Typical cutoff values, 10, 15 to 30.

Some problems with the condition number

- practitioners have different opinions of whether $\mathbf{X}$ should be centered around their means for SVD.
  - centering may remove nonessential ill conditioning, e.g. $Cor(X, X^2)$
  - centering may mask the role of the constant term in any underlying near-dependencies

- the degree of multicollinearity with dummy variables may be influenced by the choice of reference category
- condition number is affected by the scale of the $\mathbf{X}$ measurements
  - By scaling down any column of $\mathbf{X}$, the condition number can be made arbitrarily large
  - Known as *artificial ill-conditioning*
  - The condition number of the scaled matrix $\mathbf{X}_E$ is also referred to as the *scaled condition number*

Recall that $\mathbf{X}_E = \mathbf{UDV}^T$ is the singular-value decomposition (SVD) of $\mathbf{X}_E$, where $\mathbf{U}^T\mathbf{U} = \mathbf{V}^T\mathbf{V} = \mathbf{I}_p$ and $\mathbf{D} = diag(d_1, d_2, ..., d_p)$ is a diagonal matrix with $d_j \geqslant 0$.

Then

$$\begin{aligned}\mathbf{X}_E^T\mathbf{X}_E &= \mathbf{VDU}^T\mathbf{UDV}^T \\ &= \mathbf{VD}^2\mathbf{V}^T\end{aligned}$$

is the spectral decomposition of the Gramian matrix $\mathbf{X}_E^T\mathbf{X}_E$ with $\{d_j^2\}$ being the eigenvalues and $\mathbf{V}$ being the corresponding eigen vector matrix. This relationship links the condition numbers to the eigen values of the Gramian matrix.

Variance decomposition method

The variance-covariance matrix of the coefficient

$$\begin{aligned}Cov(\hat{\beta}) &= \sigma^2(\mathbf{X}_E^T\mathbf{X}_E)^{-1} \\ &= \sigma^2\mathbf{VD}^{-2}\mathbf{V}^T\end{aligned}$$

Its j^{th} diagonal element is the estimated variance of the j^{th} coefficient, $\hat{\beta}_j$. Then

$$Var(\hat{\beta}_j) = \sigma^2 \sum_{h=1}^{p} \frac{v_{jh}^2}{d_h^2}$$

- Let $q_{jh} = \frac{v_{jh}^2}{d_h^2}$ and $q_j = \sum\limits_{h=1}^{p} q_{jh}$.
- The variance decomposition proportion is $\pi_{jh} = q_{jh}/q_j$.
- π_{jh} denotes the proportion of the variance of the j^{th} regression coefficient associated with the h^{th} component of its decomposition.
- The variance decomposition proportion matrix is $\mathbf{\Pi} = \{\pi_{jh}\}$.

Condition	Proportions of variance			
Index	$Var(\hat{\beta}_1)$	$Var(\hat{\beta}_2)$	...	$Var(\hat{\beta}_3)$
η_1	π_{11}	π_{12}	$\cdots$	π_{1p}
η_2	π_{21}	π_{22}	$\cdots$	π_{2p}
$\vdots$	$\vdots$	$\vdots$		$\vdots$
η_p	π_{p1}	π_{p2}	$\cdots$	π_{pp}

Table 1: Table of condition index and proportions of variance

In practice, it is suggested to combine condition index and proportions of variance for multicollinearity diagnostic. Identify multicollinearity if

- Two or more elements in the j^{th} row of matrix $\mathbf{\Pi}$ are relatively large
- And its associated condition index η_j is large too

Principal Components

The method of principal components, introduced by Karl Pearson (1901) and Harold Hotelling (1933), provides a useful representation of the correlational structure of a set of variables. Some advantages of the principal component analysis include

- more unified
- linear transformation of the original predictors into a new set of orthogonal predictors
- the new orthogonal predictors are called principal components

Principal components regression is an approach that inspects the sample data $(\mathbf{Y}, \mathbf{X})$ for directions of variability and uses this information to reduce the dimensionality of the estimation problem. The procedure is based on the observation that every linear regression model can be restated in terms of a set of orthogonal predictor variables, which are constructed as linear combinations of the original variables. The new orthogonal variables are called the principal components of the original variables.

Let $\mathbf{X}^T\mathbf{X} = \mathbf{Q}\mathbf{\Delta}\mathbf{Q}^T$ denote the spectral decomposition of $\mathbf{X}^T\mathbf{X}$, where $\mathbf{\Delta} = diag\{\lambda_1, \ldots, \lambda_p\}$ is a diagonal matrix consisting of the (real) eigenvalues of $\mathbf{X}^T\mathbf{X}$, with $\lambda_1 \geqslant \cdots \geqslant \lambda_p$ and $\mathbf{Q} = (\mathbf{q_1}, \ldots, \mathbf{q_p})$ denotes the matrix whose columns are the orthogonal eigenvectors of $\mathbf{X}^T\mathbf{X}$ corresponding to the ordered eigenvalues. Consider the transformation

$$\mathbf{Y} = \mathbf{X}\mathbf{Q}\mathbf{Q}^T\beta + \epsilon = \mathbf{Z}\theta + \epsilon,$$

where $\mathbf{Z} = \mathbf{X}\mathbf{Q}$, and $\theta = \mathbf{Q}^T\beta$.
The elements of θ are known as the regression parameters of the principal components. The matrix $\mathbf{Z} = \{\mathbf{z_1}, \ldots, \mathbf{z_p}\}$ is called the matrix of principal components of $\mathbf{X}^T\mathbf{X}$. $\mathbf{z}_j = \mathbf{X}\mathbf{q}_j$ is the jth principal component of $\mathbf{X}^T\mathbf{X}$ and $\mathbf{z}_j^T\mathbf{z}_j = \lambda_j$, the jth largest eigenvalue of $\mathbf{X}^T\mathbf{X}$.

Principal components regression consists of deleting one or more of the variables $\mathbf{z}_j$ (which correspond to small values of λ_j), and using OLS estimation on the resulting reduced regression model.

Derivation under standardized predictors, JF 13.1.1

Consider the vectors of standardized predictors, $\mathbf{x}_1^*, \mathbf{x}_2^*, \ldots, \mathbf{x}_p^*$ (obtained by subtracting the mean and divided by standard deviation of the original predictor vectors). Because the principal components are linear combinations of the original predictors, we write the first principal component as

$$\begin{aligned}\mathbf{w}_1 &= A_{11}\mathbf{x}_1^* + A_{21}\mathbf{x}_2^* + \cdots + A_{p1}\mathbf{x}_p^* \\ &= \mathbf{X}^*\mathbf{a}_1\end{aligned}$$

The variance of the first component becomes

$$\begin{aligned}S_{w_1}^2 &= \frac{1}{n-1}\mathbf{w}_1^T\mathbf{w}_1 \\ &= \frac{1}{n-1}\mathbf{a}_1^T\mathbf{X}^{*T}\mathbf{X}^*\mathbf{a}_1 \\ &= \mathbf{a}_1^T\mathbf{R}_{XX}\mathbf{a}_1\end{aligned}$$

where $\mathbf{R}_{XX} = \frac{1}{n-1}\mathbf{X}^{*T}\mathbf{X}^*$. We want to maximize $S_{w_1}^2$ under the normalizing constraint $\mathbf{a}_1^T\mathbf{a}_1 = 1$ (otherwise $S_{w_1}^2$ can be arbitrarily large by inflating $\mathbf{a}_1$). Consider

$$F_1 \equiv \mathbf{a}^T\mathbf{R}_{XX}\mathbf{a}_1 - L_1(\mathbf{a}_1^T\mathbf{a}_1 - 1)$$

where L_1 is a Lagrange multiplier. By differentiating this equation with respect to $\mathbf{a}_1$ and L_1,

$$\begin{aligned}\frac{\partial F_1}{\partial \mathbf{a}_1} &= 2\mathbf{R}_{XX}\mathbf{a}_1 - 2L_1\mathbf{a}_1 \\ \frac{\partial F_1}{\partial L_1} &= -(\mathbf{a}_1^T\mathbf{a}_1 - 1)\end{aligned}$$

Setting the partial derivatives to 0 produces

$$\begin{aligned}(\mathbf{R}_{XX} - L_1\mathbf{I}_p)\mathbf{a}_1 &= \mathbf{0} \\ \mathbf{a}_1^T\mathbf{a}_1 &= 1\end{aligned}$$

From the first equation, we see that L_1 is an eigenvalue of $\mathbf{R}_{XX}$ such that $\mathbf{R}_{XX}\mathbf{a}_1 = L_1\mathbf{a}_1$ such that

$$S_{w_1}^2 = \mathbf{a}_1^T\mathbf{R}_{XX}\mathbf{a}_1 = L_1\mathbf{a}_1^T\mathbf{a}_1 = L_1$$

To maximize $S_{w_1}^2$, we only need to pick the largest eigenvalue of $\mathbf{R}_{XX}$.