Regression Analysis Tutoring6

Seung Bong Jung

Seoul National University

November 1, 2021

Multicollinearity

- A set of predictors $x_1, \ldots, x_p$ is said to have multicollinearity if there exist linear or near-linear dependencies among the predictors.
- In case there exists a linear dependency among the predictors, the columns of $\mathbf{X} = (\mathbf{1}, \mathbf{x}_1, \ldots, \mathbf{x}_p)$ are linearly dependent, or equivalently the centered columns $\mathbf{x}_1 - \bar{x}_1\mathbf{1}, \ldots, \mathbf{x}_p - \bar{x}_p\mathbf{1}$ are linearly dependent, so that the matrices $\mathbf{X}$ and $\mathbf{X}^t\mathbf{X}$ are not of full rank.
- Multicollinearity not only makes the computation of the parameter estimates erratic, but also increases the variance of the estimates.

$$\sum_{j=0}^{p} \mathrm{var}(\hat{\beta}_j) = \mathrm{trace}(\mathrm{var}(\hat{\boldsymbol{\beta}})) = \sigma^2 \cdot \mathrm{trace}((\mathbf{X}^t\mathbf{X})^{-1}) = \sigma^2 \sum_{j=0}^{p} \frac{1}{\kappa_j},$$

where κ_j's are eigenvalues of $\mathbf{X}^t\mathbf{X}$.

Effect of Multicollinearity

- Let $S_{jj} = \sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2$ and R_j^2 denote the coefficient of determination in regressing the jth predictor x_j on the remaining $(x_k : k \neq j)$. Then,

$$\text{var}(\hat{\beta}_j) = \frac{1}{1 - R_j^2} \cdot \frac{\sigma^2}{S_{jj}}, 1 \leq j \leq p,$$

- Proof of the identity: Take $j = 1$ without loss of generality. Let $\mathbf{x}_{1,\perp} = \mathbf{x}_1 - \Pi(\mathbf{x}_1 | \mathcal{C}_{\mathbf{1},\mathbf{x}_2,\ldots,\mathbf{x}_p})$. Then, var$(\hat{\beta}_1) = \sigma^2(\mathbf{x}_{1,\perp}^t \mathbf{x}_{1,\perp})^{-1}$. Think of fitting the regression model

$$x_{i1} = \alpha_0 + \alpha_1 x_{i2} + \cdots + \alpha_{p-1} x_{ip} + \epsilon_i, 1 \leq i \leq n.$$

The scalar value $\mathbf{x}_{1,\perp}^t \mathbf{x}_{1,\perp}$ is nothing else than the residual sum of squares. The total sum of squares in this case is S_{11}, and $R_1^2 = \frac{S_{11} - \mathbf{x}_{1,\perp}^t \mathbf{x}_{1,\perp}}{S_{11}}$ or $\mathbf{x}_{1,\perp}^t \mathbf{x}_{1,\perp} = S_{11}(1 - R_1^2)$.

Diagnostics of Multicollinearity

- Variance inflation factor: $\text{VIF}_j := (1 - R_j^2)^{-1}$
- VIF_j is simply the inflation rate of $\text{var}(\hat{\beta}_j)$ in comparison with the case where x_j is not correlated with other predictors, i.e.,
$$S_{jk} = \sum_{i=1}^{n}(x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k) = 0 \text{ for all } k \neq j.$$
- Note that large VIF_j for one or multiple j's indicates multicollinearity. The inspection of all pairwise correlations between two predictors is not sufficient for detecting multicollinearity in general.
- Typically the existence of $\text{VIF}_j > 10$ is considered as an indiction of severe multicollinearity.

Ridge Regression

- One useful method for dealing with multicollinearity is ridge regression.
- Note that $\mathbf{X}\boldsymbol{\beta} = \mathbf{1}\beta_0' + \mathbf{X}_{1,\perp}\boldsymbol{\beta}_1$ with $\boldsymbol{\beta}^t = (\beta_0, \boldsymbol{\beta}_1^t)$ and that β_0 is estimated by $\bar{Y}$.
- Adding a positive constant $k > 0$ to the diagonal entries of $\mathbf{X}_{1,\perp}^t \mathbf{X}_{1,\perp}$:

$$\hat{\boldsymbol{\beta}}_{1,R}(\lambda) = (\mathbf{X}_{1,\perp}^t \mathbf{X}_{1,\perp} + \lambda \mathbf{I})^{-1} \mathbf{X}_{1,\perp}^t \mathbf{Y}$$

- The ridge estimator $\hat{\boldsymbol{\beta}}_{1,R}$ is biased, but has smaller total variance. Though it does not give best fit, it may do better job in out-of-sample prediction.
- Penalized least squares estimation: The ridge estimator may be defined to be the minimizer of $||\mathbf{Y} - \mathbf{X}_{1,\perp}\boldsymbol{\beta}_1||^2 + \lambda||\boldsymbol{\beta}_1||^2$.

Geometry of Ridge Regression

- In fact, the given optimization problem is equivalent to following optimization problem:

$$\text{Minimize } ||\mathbf{Y} - \mathbf{X}_{1,\perp}\boldsymbol{\beta}_1||^2 \text{ subject to } ||\boldsymbol{\beta}_1||^2 \leq d$$

,where $d = \hat{\boldsymbol{\beta}}_1^t[\mathbf{I} + \lambda(\mathbf{X}_{1,\perp}^t\mathbf{X}_{1,\perp})^{-1}]^{-2}\hat{\boldsymbol{\beta}}_1$

- Note that

$$||\mathbf{Y} - \mathbf{X}_{1,\perp}\boldsymbol{\beta}_1||^2 = ||\mathbf{Y} - \mathbf{X}_{1,\perp}\hat{\boldsymbol{\beta}}_1||^2 + (\boldsymbol{\beta}_1 - \hat{\boldsymbol{\beta}}_1)^t\mathbf{X}_{1,\perp}^t\mathbf{X}_{1,\perp}(\boldsymbol{\beta}_1 - \hat{\boldsymbol{\beta}}_1)$$

- $\hat{\boldsymbol{\beta}}_{1,R}(\lambda)$ as a shrinkage stimator: The penalty term $k||\boldsymbol{\beta}_1||^2$ in the penalized least squares criterion shrinks $\hat{\boldsymbol{\beta}}_1$ toward $\mathbf{0}$.

Bayesian interpretation of Ridge Regression

- Let $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^t$ and $x_i = (1, x_{i1}, \ldots, x_{ip})^t$.
- Let $Y_i|\boldsymbol{\beta} \sim \mathcal{N}(x_i^t\boldsymbol{\beta}, \sigma^2)$ for $i = 1, 2, \ldots, n$ and $\beta_j \overset{\text{i.i.d}}{\sim} \mathcal{N}(0, \sigma^2/\lambda)$ for $j = 0, 1, \ldots, p$.
- Denoting $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)^t$, by Bayes Rule, one can see that the ridge estimator $\hat{\boldsymbol{\beta}}_{1,\mathsf{R}}(\lambda)$ is mode of the distribution of $\boldsymbol{\beta}|Y$.

Principal Component Analysis

- Recall that the entries of $\mathbf{X}_{1,\perp}^t \mathbf{X}_{1,\perp}$ are
$$S_{jk} = \sum_{i=1}^{n} (x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k).$$
- Principal component analysis of the predictors:
$$\mathbf{X}_{1,\perp}^t \mathbf{X}_{1,\perp} = \mathbf{P}\boldsymbol{\Lambda}\mathbf{P}^t,\ \boldsymbol{\Lambda} = \mathrm{diag}(\lambda_j),\ \mathbf{P} = (\mathbf{v}_1, \ldots, \mathbf{v}_p),$$
where $\mathbf{v}_j$'s are the orthonormal eigenvectors of $\mathbf{X}_{1,\perp}^t \mathbf{X}_{1,\perp}$ ordered in terms of the respective eigenvalues $\lambda_1 \geq \cdots \geq \lambda_p$.

Principal Component Analysis

- Write $\mathbf{v}_j = (v_{j1}, \ldots, v_{jp})^t$. Then, $\mathbf{z}_j := \mathbf{X}_{1,\perp}\mathbf{v}_j$ are the observed vector of new variables $z_j = v_{j1}(x_1 - \bar{x}_1) + \cdots + v_{jp}(x_p - \bar{x}_p)$, called principal components.
- Since $\mathbf{X}_{1,\perp}\mathbf{P} = (\mathbf{z}_1, \ldots, \mathbf{z}_p)$, we obtain

$$\lambda_j = ||\mathbf{z}_j||^2, \ 1 \leq j \leq p.$$

- Identification of sources of multicollinearity: $\lambda_j = 0$ if and only if the observed values of the predictors satisfy the equation

$$v_{j1}(x_1 - \bar{x}_1) + \cdots + v_{jp}(x_p - \bar{x}_p) = 0$$

Principal Component Analysis

- Recall the definition of $\mathbf{z}_j$ in the PCA of the predictors: $\mathbf{z}_j = \mathbf{X}_{1,\perp}\mathbf{v}_j$, where $\mathbf{v}_j$'s are the orthonormal eigenvectors of $\mathbf{X}_{1,\perp}^t\mathbf{X}_{1,\perp}$ ordered in terms of the respective eigenvalues $\lambda_1 \geq \cdots \geq \lambda_p$.
- One may rewrite the regression equation using the centered predictors:

$$\begin{aligned}
\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p &= \beta_0' + \beta_1(x_1 - \bar{x}_1) + \cdots + \beta_p(x_p - \bar{x}_p), \\
\mathbf{X}\boldsymbol{\beta} &= \mathbf{1}\beta_0' + \mathbf{X}_{1,\perp}\boldsymbol{\beta}_1 \\
&= \mathbf{1}\beta_0' + \mathbf{X}_{1,\perp}\mathbf{P}\cdot\mathbf{P}^t\boldsymbol{\beta}_1 \\
&\overset{\text{let}}{=} \mathbf{1}\beta_0' + \mathbf{Z}\cdot\boldsymbol{\alpha}
\end{aligned}$$

- The intercept parameter β_0' is estimated by $\bar{Y}$ in least squares regression.

Principal Component Analysis

- Choose the first q principal components $Z_1, \ldots, z_q$ with $q < p.$, set $z_j \equiv 0$ for all $q+1 \leq j \leq p$, and fit the resulting (reduced) model

$$\mathbf{Y} = \mathbf{1}\beta_0' + \mathbf{Z}_q \cdot \boldsymbol{\alpha}_q + \boldsymbol{\epsilon},$$

  where $\mathbf{Z}_q = (\mathbf{z}_1, \ldots, \mathbf{z}_q)$.

- Since each $\mathbf{z}_j$ is orthogonal to $\mathbf{1}$, it follows that

$$\hat{\boldsymbol{\alpha}}_q = (\mathbf{Z}_q^t \mathbf{Z}_q)^{-1} \mathbf{Z}_q^t \mathbf{Y} = \boldsymbol{\Lambda}_q^{-1} \mathbf{Z}_q^t \mathbf{Y}, \hat{\beta}_0' = \bar{Y}.$$

- Letting $\hat{\boldsymbol{\alpha}}^t = (\hat{\boldsymbol{\alpha}}_q^t, \mathbf{0}_{p-q}^t)$,

$$\hat{\boldsymbol{\beta}}_{1,\text{pcr}} = \mathbf{P}\hat{\boldsymbol{\alpha}} = \sum_{j=1}^{q} \hat{\alpha}_j \mathbf{v}_j, \ \hat{\beta}_0' = \bar{Y}.$$