Collinearity

- Consider a MLR model with a design matrix $\mathbf{X}_{n\times p}$ including the intercept. If each column of $\mathbf{X}$ is orthogonal to each other (i.e., the sample correlation of any two predictors is equal to 0), then the LS problem is greatly simplified.

$$\hat{\beta}_j = \left[(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}\mathbf{y}\right]_j = \frac{\mathbf{X}_{\cdot j}^t\mathbf{y}}{\|\mathbf{X}_{\cdot j}\|^2},$$

where $\mathbf{X}_{\cdot j}$ denotes the j-th column of $\mathbf{X}$. That is, the regression coefficient for the j-th predictor does not depend on whether other predictors are included in the model or not.

- However, we often encounter problems in which many of the predictors are highly correlated. In this case, the values and sampling variance of regression coefficients can be highly dependent on the particular predictors chosen for the model.

- If there exists a set of constants $c_1, \ldots, c_p$ (at least one of them is non-zero), such that the corresponding linear combination of the columns of $\mathbf{X}$ is zero, i.e.,

$$\sum_{j=1}^{p} c_j \mathbf{X}_{\cdot j} = \mathbf{0},$$

  then the columns of $\mathbf{X}$ are called linearly dependent or exactly collinear. That is, at least one column in the design matrix $\mathbf{X}$ can be expressed as a linear combination of the other columns.

When the columns of $\mathbf{X}$ are collinear,

1. $(\mathbf{X}^t\mathbf{X})^{-1}$ does not exist;

2. the LS estimate $\hat{\boldsymbol{\beta}}$ is not unique, and

3. the corresponding linear model is not identifiable.

For example, suppose the 1st column of $\mathbf{X}$ is the intercept, and the 2nd column of $\mathbf{X}$ is $(2, 2, \ldots, 2)^t$. Then if $(\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \ldots)^t$ is one LS estimate of $\boldsymbol{\beta}$, so is $(\hat{\beta}_1 - c, \hat{\beta}_2 + c/2, \hat{\beta}_3, \ldots)^t$ where c is any real number.

(Approximate) Collinearity

- We generally do not need to worry about exact collinearity[a] , but (approximate) collinearity. That is, at least one column $\mathbf{X}_{\cdot j}$ can be approximated by the others,

$$\mathbf{X}_{\cdot k} \approx -\sum_{j \neq k} c_j \mathbf{X}_{\cdot j} / c_k.$$

  A simple diagnostic for this is to obtain the regression of $\mathbf{X}_{\cdot k}$ on the remaining predictors, and if the corresponding R_k^2 is close to 1, we would diagnose approximate collinearity.

[a] R can detect it and also can fix it automatically.

Why is collinearity a problem?

- In a multiple regression $Y = \beta_0 + \beta_1 X_1 + \cdots \beta_p X_p + e$, the LS estimate $\hat{\beta}_k$ is unbiased with

$$\text{var}(\hat{\beta}_k) = \sigma^2 \Big(\frac{1}{1 - R_k^2} \Big) \Big(\frac{1}{\sum_{i=1}^n (x_{ik} - \bar{x}_{\cdot k})^2} \Big),$$

  where R_k^2 is the R-square from the regression of $\mathbf{X}_{\cdot k}$ on the remaining predictors. When R_k^2 is close to 1, the variance of $\hat{\beta}_k$ is large.

  Consequently: 1) large MSE and 2) large p-value, i.e., we could miss a significant variable.

- The quantity $\frac{1}{1-R_k^2}$ is called the k-th variance influential factor (VIF).

- A global measure of collinearity is given by examining the eigenvalues of $\mathbf{X}^t\mathbf{X}$. A popular measure is the condition number of $\mathbf{X}^t\mathbf{X}$, denoted by

$$\kappa = (\text{largest eigenvalue}/\text{smallest eigenvalue})^{1/2}.$$

  An empirical rule for declaring collinearity is $\kappa \geq 30$.

- Note that κ is not scale invariant, so we should standardize each column of X (i.e., now each column of X has mean 0 and sample variance 1) before calculating the condition number.

Symptoms and Remedy

- Possible symptoms of collinearity: high pair-wise (sample) correlation between predictors, high VIF, high condition number, R^2 is relatively large but none of the predictor is significant.
- What to do with collinearity? Remove some predictors.