* <u>Multicollinearity</u>

* <u>Exact multicollinearity</u>

Consider a three-variable regression:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i \qquad (1)$$

And suppose we have exact collinearity in:

$$x_{i3} = \alpha_1 x_{i1} + \alpha_2 x_{i2} \qquad (\ddot{\frown})$$

Then, model (1) is not well-determined or estimable. The reason for this is if we swap $(\beta_1, \beta_2, \beta_3)$ with any other set $(\beta_1^*, \beta_2^*, \beta_3^*)$ where

$$\left.\begin{aligned} \beta_1 + \alpha_1 \beta_3 &= \beta_1^* + \alpha_1 \beta_3^* \quad \text{and} \\ \beta_2 + \alpha_2 \beta_3 &= \beta_2^* + \alpha_2 \beta_3^* \end{aligned}\right. \qquad (A)$$

then model (1) is unchanged!

* In such situations, $X^T X$ is a singular matrix and $\hat{\beta}_{OLS}$ does not exist.

* <u>Approximate multicollinearity</u>

In most cases, we don't have perfect linear relationships like $(\ddot{\frown})$, but instead have approximate linear dependence.

- In this case, $X^T X$ will be <u>ill-conditioned</u> – its inverse is numerically unstable.

- It follows that some of the parameter estimates, or ~~at least~~ some combinations of them, will have large variances.

  <u>Implication</u>: large perturbations of the parameters will have only a small effect on the fit of the model.

* <u>Detecting Multicollinearity</u>

* I. <u>Variance Inflation Factors</u>

- For all columns of X except the intercept, we must first standardize the columns to have mean 0 and variance 1.

  <u>R code</u>: use scale(X)

- The j^{th} <u>variance inflation factor</u> is the j^{th} diagonal entry of $(X^T X)^{-1}$ once X has been standardized as above. <u>Notation</u>: V_j

- <u>Idea</u>: if there were no multicollinearity, the cross-products of each pair of variables would be 0 and $X^T X$ would be the identity matrix $\Rightarrow V_j \equiv 1 \ \forall j$

* There are no formal tests for large VIF's, but it is accepted that significant multicollinearity is present if

$$V_j > 10 \quad \text{for any} \quad j = 1, \dots, p.$$

* <u>Note</u>: this does <u>not</u> indicate <u>where</u> collinearity occurs, only that it is present

II. <u>Singular Value Criteria for MC</u>

* We start with the singular value decomposition of the standardized version of X:

$$X = UDV^T, \quad \text{where} \qquad (\ddot{u})$$

U is $n \times p$, V is $p \times p$ and $U^TU = V^TV = I_p$ and D is a $p \times p$ ~~mat~~ diagonal matrix with entries $\mu_1, \dots \mu_p$, which are the <u>singular values</u> of X.

$(\ddot{u})$ implies

$$X^TX = VD^2V^T \Rightarrow (X^TX)^{-1} = VD^{-2}V^T \qquad (**)$$

$(**)$ $\Rightarrow$ small singular values imply large entries in $(X^TX)^{-1}$ (and hence large variance).

* Define

$$\tau_k := \frac{\mu_{max}}{\mu_k}$$

as the k^{th} condition index of X, where

$\mu_{max} = \max\{\mu_1, \dots, \mu_p\}$.

Large values of τ_k indicate multicollinearity.
Rule of thumb:

$\tau_k \in (30, 100) \Rightarrow$ moderate to strong association

$\tau_k < 10 \Rightarrow$ weak association.

* Again, τ_k does not show where association occurs. To measure this, we define

$$\pi_{kj} := \frac{v_{jk}^2 / \mu_k^2}{\sum_{k'} (v_{jk'}^2 / \mu_{k'}^2)}$$

= proportion of variance of the j^{th} parameter estimate that is accounted for by the k^{th} singular value.

* Rule of thumb: Suspect variables are those w/ high τ_k and at least two large values of π_{kj}.

~~Monday~~ ~~→ review~~

Remedies for MC

I. Ridge regression → reduces the effects of MC through regularization

II. Principal components regression:

Main idea:

1. Identify k principal components of the columns of X. These represent the k directions of most variability in X.
2. Regress Y on the principal components. Since the directions are orthogonal, we have no issues w/ MC but we generally lose interpretability.
3. Removal of suspicious variables from MC diagnostics.
4. Partial Least Squares (not covering here, but it is an option).