Multicollinearity

Lecture 20

Ravleen Bajaj

Today:

- What is multicollinearity?
- Why does it cause problems?
- Numerical instability of $(X'X)^{-1}$
- Example with simulated data
- Remedy and interpretation

What is Multicollinearity?

- "A set of predictors exhibits multicollinearity when at least one predictor is highly correlated with another predictor or linear combination of predictors in the set."
- Causes:
  - Redundant information among predictors.
  - Linear dependence of columns of the design matrix X.
- Consequences:
  - Large (inflated) standard errors.
  - Unstable coefficient estimates and SEs.

Remember:

$$\tilde{\beta} = (X'X)^{-1}X'Y$$
$$\text{Var}(\tilde{\beta}) = \sigma^2(X'X)^{-1}$$

If predictors are highly correlated:

- $X'X$ is close to or approximately singular.
- $(X'X)^{-1}$ has very large entries and is numerically unstable.
- $\Rightarrow$ Parameter estimates and SEs become unstable.

Numerical Example:

Setup: Two nearly identical predictors x_1 and x_2.

$$X = \begin{bmatrix} 1 & 1 & 1.01 \\ 1 & 2 & 1.99 \\ 1 & 3 & 3.02 \\ 1 & 4 & 3.98 \\ 1 & 5 & 5.01 \end{bmatrix}$$

Compute: $(X'X)^{-1}$

$$(X'X)^{-1}_{\text{before}} = \begin{bmatrix} 1.13 & -0.38 & -0.37 \\ -0.38 & 0.14 & 0.14 \\ -0.37 & 0.14 & 0.14 \end{bmatrix}$$

Interpretation: nearly singular, unstable coefficients.

After Adding One Distinct Data Point

Add: $(x_1, x_2) = (8, 8.2)$

$$(X'X)^{-1}_{\text{after}} = \begin{bmatrix} 0.33 & -0.06 & -0.06 \\ -0.06 & 0.02 & 0.01 \\ -0.06 & 0.01 & 0.02 \end{bmatrix}$$

Result:

- Variance of $\tilde{\beta}$ (and SEs) shrink dramatically.
- These quantities are highly unstable and results could vary with adding just a single point.

Example:

Consider a situation where we want to:

- Predict BMI using left and right arm skinfold measurements of individuals.
- Both measures are very similar $\Rightarrow$ highly correlated.

Model 1:

$$\text{BMI}_i = 20 + x_{1i} + x_{2i} + \varepsilon_i$$

and ε_i are independent and $\varepsilon_i \sim N(0, \sigma^2)$

where:

$$x_{1i} = \text{LeftArmMeasure}$$
$$x_{2i} = \text{RightArmMeasure}$$

Output:

```
Call:
lm(formula = BMI ~ x1 + x2)

Residuals:
   Min     1Q Median     3Q    Max
-3.796 -1.433 -0.235  1.206  6.316

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  20.1388     0.9095  22.144   <2e-16 ***
x1            0.2429     0.2610   0.931    0.354
x2            0.2393     0.2583   0.926    0.357
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.951 on 97 degrees of freedom
Multiple R-squared:  0.5619,    Adjusted R-squared:  0.5529
F-statistic: 62.21 on 2 and 97 DF,  p-value: < 2.2e-16
```

Interpretation: Strong linear relationship between left and right arm skinfolds, SE of the estimated effect could be inflated.

A Possible Remedy:

average the left and right arm measurements (because they are quantifying roughly the same quantity):

Let $x_{3i} = (x_{1i} + x_{2i})/2$

Model 2:

$$\text{BMI} = 20 + 0.5x_{3i} + \varepsilon$$

where:

$$x_{1i} = \text{LeftArmMeasure}$$
$$x_{2i} = \text{RightArmMeasure}$$

Output:

```
Call:
lm(formula = BMI ~ x3)

Residuals:
    Min      1Q  Median      3Q     Max
-3.7974 -1.4307 -0.2337  1.2062  6.3156

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 20.13929    0.90162   22.34   <2e-16 ***
x3           0.48220    0.04301   11.21   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.941 on 98 degrees of freedom
Multiple R-squared:  0.5619,    Adjusted R-squared:  0.5574
F-statistic: 125.7 on 1 and 98 DF,  p-value: < 2.2e-16
```

Interpretation: parameter estimates and SEs are likely more stable now! Also, notice the decrease in SE and the now significant p-value

R Code

```
set.seed(123)
n <- 100

x <- rnorm(n, mean = 20, sd = 5)
BMI <- 20 + 0.5 * x + rnorm(n, 0, 2)

x1 <- x + rnorm(n, 0, 0.5)
x2 <- x + rnorm(n, 0, 0.55)

mod1 <- lm(BMI ~ x1 + x2)

x3 <- (x1 + x2)/2
mod2 <- lm(BMI ~ x3)

summary(mod1)
summary(mod2)
```

Variance Inflation Factor

$$R^2 = 0.5619$$

Variance Inflation Factor (VIF):

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

For this example:

$$\text{VIF}_j = 36.32744 \quad \text{where } j = 1, 2$$

For these predictors, we have large VIFs $\Rightarrow$ multicollinearity.

Interpretation: High correlation $\Rightarrow$ unstable coefficient estimates and SEs.

Thank You!

Questions?