Lecture 19: (Multi)Collinearity

Ailin Zhang

2023-07-04

Agenda

1. What is collinearity/ multicollinearity (MC)?
2. How can MC be detected?
3. How does MC affect statistical inference and forecasting?
4. How can we resolve problems with MC?

The Problem of Multicollinearity

- Multicollinearity (MC) means predictors in an MLR model have a **close-to-exact linear relationship**, i.e., predictors are nearly linearly dependent
- When MC problem exists, the LS estimates for β_j's exists but would have large variability.
- Recall We interpret $\hat{\beta}_j$ as the mean response change when X_j increases by one unit holding all other predictors fixed If predictors are strongly correlated, we might not be able to alter X_j while holding other predictors fixed.

Example

- Ex:
  - $X_1 = \#$ of undergrads
  - $X_2 = \#$ of grads
  - $X_3 = \#$ of students

  then X_1, X_2, X_3 are linearly dependent since $X_1 + X_2 = X_3$
- No unique LS estimates for coefficients if predictors are linearly dependent
- $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon$
- $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 (X_1 + X_2) + \epsilon$
- the coefficients $(\beta_0, \beta_1, \beta_2, \beta_3)$ and $(\beta_0', \beta_1', \beta_2', \beta_3')$ give identical mean of Y if:
  - $\beta_1' = \beta_1 + \beta_3$
  - $\beta_2' = \beta_2 + \beta_3$
  - $\beta_3' = 0$.

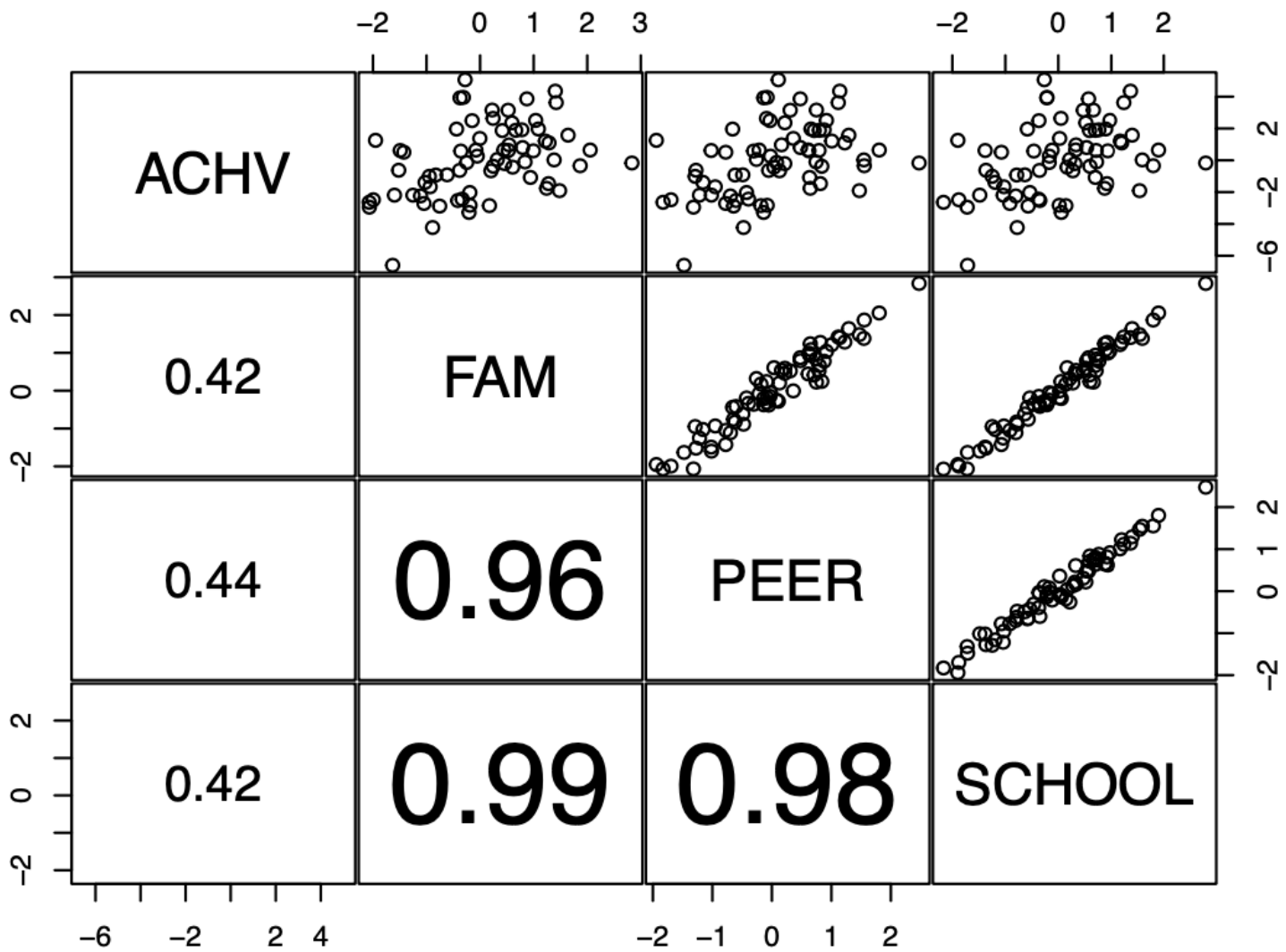
Example: Equal Educational Opportunity (EEO) Data

To examine the existence (or lack) of equal educational opportunities in public educational institutions, the following variables were measured for 70 schools selected at random in 1965.

- ACHV: Student achievement index (higher values are better)
- FAM: Faculty credentials index
- PEER: the influence of their peer group in the school
- SCHOOL: School facility/resource index Goal: to identify important determinants of student achievement

```
achv = read.table("achv.txt", h=T)
```

All 3 Predictors Are Highly Correlated!



```r
summary(lm(ACHV ~ FAM + PEER + SCHOOL, data=achv))
```

```
##
## Call:
## lm(formula = ACHV ~ FAM + PEER + SCHOOL, data = achv)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.2096 -1.3934 -0.2947  1.1415  4.5881
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.06996    0.25064  -0.279    0.781
## FAM          1.10126    1.41056   0.781    0.438
## PEER         2.32206    1.48129   1.568    0.122
## SCHOOL      -2.28100    2.22045  -1.027    0.308
##
## Residual standard error: 2.07 on 66 degrees of freedom
## Multiple R-squared:  0.2063, Adjusted R-squared:  0.1702
## F-statistic: 5.717 on 3 and 66 DF,  p-value: 0.001535
```

- None of the 3 predictors is significant but the overall model F-statistic is significant (i.e., at least one of the 3 predictor is significant)
- The coefficient of SCHOOL is negative! Normally, we expect higher student achievement if schools have more resources.

- Correlations between the 3 predictors are extremely high!
- Knowing any of 3 predictors, we can predict the other 2 very accurately.
- So it's almost like we have only one predictor. Only 1 of the 3 predictors is really needed.

However, multicollinearity prevents us from identifying the important predictors

Can not tell which predictor is more important

```
lmFPS = lm(ACHV ~ FAM+PEER+SCHOOL, data=achv)
lmF = lm(ACHV ~ FAM, data=achv)
lmP = lm(ACHV ~ PEER, data=achv)
lmS = lm(ACHV ~ SCHOOL, data=achv)

anova(lmF, lmFPS)
```

```
## Analysis of Variance Table
##
## Model 1: ACHV ~ FAM
## Model 2: ACHV ~ FAM + PEER + SCHOOL
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     68 293.68
## 2     66 282.87  2    10.803 1.2603 0.2903
```

```
anova(lmP, lmFPS)
```

```
## Analysis of Variance Table
##
## Model 1: ACHV ~ PEER
## Model 2: ACHV ~ FAM + PEER + SCHOOL
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     68 287.43
## 2     66 282.87  2    4.5593 0.5319   0.59
```

```
anova(lmS, lmFPS)
```

```
## Analysis of Variance Table
```

Can not tell which predictor is more important

```
lmFPS = lm(ACHV ~ FAM+PEER+SCHOOL, data=achv)
lmF = lm(ACHV ~ FAM, data=achv)
lmP = lm(ACHV ~ PEER, data=achv)
lmS = lm(ACHV ~ SCHOOL, data=achv)

anova(lmS, lmFPS)
```

```
## Analysis of Variance Table
##
## Model 1: ACHV ~ SCHOOL
## Model 2: ACHV ~ FAM + PEER + SCHOOL
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     68 294.08
## 2     66 282.87  2    11.208 1.3076 0.2774
```

Summary: Effects of Multicollinear Data

- Trouble in identifying important predictors
- Estimates are very sensitive to what other variables exist in the model
  - the estimated coefficient of SCHOOL changes from -2.281 to 0.928 when FAM and PEER is removed from the model.

```
lm(ACHV ~ FAM+PEER+SCHOOL, data=achv)$coef
```

```
## (Intercept)         FAM        PEER      SCHOOL
## -0.06995905  1.10125968  2.32205659 -2.28099511
```

```
lm(ACHV ~ SCHOOL, data=achv)$coef
```

```
## (Intercept)      SCHOOL
## -0.01042968  0.92834288
```

- Estimated coefficients are very sensitive to small changes in data.

Why Multicollinearity Making Predictors Insignificant?

The LS estimate β_1 for FAM in the MLR model:

$$\text{ACHV}=\beta_0 +\beta_1\text{FAM}+\beta_2\text{PEER}+\beta_3\text{SCHOOL}+\epsilon$$

would be identical to the slope for the SLR model below

1. Regress ACHV on PEER and SCHOOL
2. Regress FAM on PEER and SCHOOL
3. Fit a SLR model using the residuals from Step 1 as the response and the residuals from Step 2 as the predictor.

- $s.e.(\hat{\beta}_1) = \hat{\sigma^2}/\sqrt{\sum(x_i - \bar{x})^2}$ is inversely proportional to the SD of the predictor.
  - $SD \approx 0$.
  - So $s.e.(\hat{\beta}_1)$ would be huge $\Rightarrow$ small t-value $\Rightarrow$ insignificant predictor

A Solution to Tackle Multicollinearity

Problem: We do not have the necessary diversity of data to separate the effects of FAM, PEER, and SCHOOL.

- One solution to tackle multicollinearity is to obtain more data with the missing combinations of predictors
- Better to taking to obtain more combinations when the data are collected, though not always possible
- Moreover, obtaining more data is not always possible due to limitations on time or money
- Sometimes such observations doesn't exist

Multicollinearity is:

- associated with unstable coefficient estimates.
- a result of linear relationships between predictors.
- not due to model mis-specification.

Therefore, we do not worry about fixing multicollinearity until the model diagnostics of other assumptions are satisfactory.

- Nevertheless, some indications of multicollinearity arise during the process of adding and removing variables and altering or removing observations.

Detecting Multicollinearity (Qualitatively)

While detecting multicollineariy, pay attention to instability in estimated $\hat{\beta}_j$'s:

- Large changes in some $\hat{\beta}_j$'s when a variable is added or deleted.
- Large changes in some $\hat{\beta}_j$'s when a data point is altered or dropped.

Once the model fit is good, look for:

- Signs of some $\hat{\beta}_j$'s do not conform to prior expectations.
- Coefficients or variables that are expected to be important have large standard errors (small t values.)

Detecting Multicollinearity (Quantitatively): Variance Inflation Factor

Variance Inflation Factor (VIF) measures the extent to which the variance of an estimated regression coefficient is increased due to collinearity with other predictor variables in the model.

Consider two explanatory variables $Y = X\beta + \epsilon$:

$$Y = \beta_0 + \beta_1 x + \beta_2 z + \epsilon$$

$$X^T X = \begin{bmatrix} n & n\bar{x} & n\bar{z} \\ n\bar{x} & \sum_i x_i^2 & \sum_i x_i z_i \\ n\bar{z} & \sum_i x_i z_i & \sum_i z_i^2 \end{bmatrix} = n \begin{bmatrix} 1 & w^T \\ w & S + ww^T \end{bmatrix}$$

where $w = (\bar{x}, \bar{z})^T$, $S = n^{-1}\tilde{X}^T\tilde{X}$, $\tilde{X} = (x - \bar{x}, z - \bar{z})$

Variance Inflation Factor (VIF)

If all inverses exist,

$$\begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{pmatrix}^{-1} = \begin{pmatrix} \mathbf{A}_{11}^{-1} + \mathbf{B}_{12}\mathbf{B}_{22}^{-1}\mathbf{B}_{21} & -\mathbf{B}_{12}\mathbf{B}_{22}^{-1} \\ -\mathbf{B}_{22}^{-1}\mathbf{B}_{21} & \mathbf{B}_{22}^{-1} \end{pmatrix}$$
$$= \begin{pmatrix} \mathbf{C}_{11}^{-1} & -\mathbf{C}_{11}^{-1}\mathbf{C}_{12} \\ -\mathbf{C}_{21}\mathbf{C}_{11}^{-1} & \mathbf{A}_{22}^{-1} + \mathbf{C}_{21}\mathbf{C}_{11}^{-1}\mathbf{C}_{12} \end{pmatrix},$$

where $\mathbf{B}_{22} = \mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12}$, $\mathbf{B}_{12} = \mathbf{A}_{11}^{-1}\mathbf{A}_{12}$, $\mathbf{B}_{21} = \mathbf{A}_{21}\mathbf{A}$
$\mathbf{C}_{11} = \mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21}$, $\mathbf{C}_{12} = \mathbf{A}_{12}\mathbf{A}_{22}^{-1}$, and $\mathbf{C}_{21} = \mathbf{A}_{22}^{-1}\mathbf{A}_{21}$.

$$(X^T X)^{-1} = n^{-1} \begin{bmatrix} 1 + w^T S^{-1} w & -w^T S^{-1} \\ -S^{-1} w & S^{-1} \end{bmatrix}$$

- $\mathrm{Var}(\hat{\beta}_0) = \sigma^2(1 + w^T S^{-1} w)/n$
- $\mathrm{Var}(\hat{\beta}_1) = \sigma^2/\{s_{xx}(1 - r^2)\}$
- $\mathrm{Var}(\hat{\beta}_2) = \sigma^2/\{s_{zz}(1 - r^2)\}$

Variance Inflation Factor (VIF)

Center and scale x_j's: $Y = \gamma_0 + \gamma_1 x^* + \gamma_2 z^* + \epsilon$

where $x_i^* = (x_i - \bar{x})/\sqrt{s_{xx}}$, $z_i^* = (z_i - \bar{z})/\sqrt{s_{zz}}$

- $X_s = (1, x^*, z^*)$
- $X_s^T X_s = \begin{bmatrix} n & 0^T \\ 0 & 1 \end{bmatrix}$
- $R = \begin{bmatrix} 1 & r \\ r & 1 \end{bmatrix}$
- $\mathrm{Var}(\hat{\gamma}_0) = \sigma^2/n$
- $\mathrm{Var}(\hat{\gamma}_1) = \sigma^2/(1 - r^2)$
- $\mathrm{Var}(\hat{\gamma}_1) = \sigma^2/(1 - r^2)$

Variance Inflation Factor (VIF): General Case

- $\mathrm{Var}(\hat{\gamma}_j) = \sigma^2/(1 - R_j^2)$
- $VIF_j = \dfrac{1}{1 - R_j^2}$

Defined as jth variance inflation factor or VIF_j , where R_j^2 is the multiple R^2 for regressing X_j on the other predictors in the model.

- If X_j is orthogonal (has 0 correlation) to all other predictors, then $R_j^2 = 0$ and $VIF_j = 1$.
- Increasing values of VIF_j indicate departure from orthogonality toward multicollinearity.
- A rule of thumb: VIFs $> 5(10)$ suggest multicollinearity

Interpreting VIF_j

- $VIF_j = \mathrm{Var}(\hat{\gamma}_j)/\sigma^2 = \mathrm{Var}(\hat{\beta}_j s_j)/\sigma^2 = s_j^2[(X^T X)^{-1}]_{j+1,j+1}$
- VIF measures correlation among the predictors.
- VIF_j indicates the proportional increase in $\mathrm{Var}(\hat{\beta}_j)$ due to its collinearity with other predictors, relative to the orthogonal case.

VIF in R

The `vif()` function in the `car` library can calculate the VIF for us.

```
library(car)
vif(lm(ACHV ~ FAM + PEER + SCHOOL, data=achv))
```

```
##      FAM     PEER   SCHOOL
## 37.58064 30.21166 83.15544
```

```
summary(lm(SCHOOL ~ FAM + PEER, data=achv))$r.squared
```

```
## [1] 0.9879743
```

$$\frac{1}{1-R^2} = \frac{1}{1-0.987974} \approx 83.155$$

Variance Inflation Due to Multicollinearity

```
summary(lm(ACHV ~ SCHOOL, data=achv))$coef
```

```
##                Estimate Std. Error     t value     Pr(>|t|)
## (Intercept) -0.01042968  0.2486820 -0.04193982 0.966669576(
## SCHOOL       0.92834288  0.2445968  3.79540138 0.000316303
```

```
summary(lm(ACHV ~ FAM + PEER + SCHOOL, data=achv))$coef
```

```
##                Estimate Std. Error    t value  Pr(>|t|)
## (Intercept) -0.06995905  0.2506421 -0.2791193 0.7810260
## FAM          1.10125968  1.4105616  0.7807243 0.4377560
## PEER         2.32205659  1.4812872  1.5675937 0.1217584
## SCHOOL      -2.28099511  2.2204481 -1.0272679 0.3080446
```

The ratio of two s.e.'s of SCHOOL in the two models is $2.2204481/0.244597 \approx 9.08$, close to $\sqrt{VIF_{SCHOOL}} = \sqrt{83.33}$

Perturbation Theory (Extra)

- Condition number:
$$\kappa(X) = \sqrt{\lambda_{MAX}/\lambda_{MIN}}$$
- Consider projection matrix $P = I_n - n^{-1}1_n1_n^T$, $Px = \begin{bmatrix} x_1 - \bar{x} \\ x_2 - \bar{x} \\ \vdots \\ x_n - \bar{x} \end{bmatrix}$

$||Px||^2 = s_{xx} = n\hat{\sigma}^2$

- Centered and Scaled $x^* = Px/||Px||$
- In $Y = \gamma_0 + \gamma_1 x^* + \epsilon$, $\hat{\gamma_1} = \dfrac{Y^T Px}{||Px||}$

Perturbation Theory (Extra)

- $\hat{\gamma}_1 = \frac{Y^T P x}{||Px||}$
- If x are perturbed by δ

$$\hat{\gamma}_{\epsilon,1} = \frac{Y^T P(x+\delta)}{||P(x+\delta)||}$$

- Relative change: $|\frac{\hat{\gamma}_1 - \hat{\gamma}_{\epsilon,1}}{\hat{\gamma}_1}| \leq \frac{\kappa(X_c)\epsilon}{1-\kappa(X_c)\epsilon}(2 + \frac{||e||}{|\hat{\gamma}_1|})$
- $X_c = \begin{bmatrix} 1 & x_1 - \bar{x} \\ 1 & x_2 - \bar{x} \\ \vdots & \vdots \\ 1 & x_n - \bar{x} \end{bmatrix}$, $\kappa(X_c) = \frac{n}{s_{xx}}$
- This shows that provided the condition number is not too large and $|\hat{\gamma}|$ is not too small, a small change in the Xi's will not cause too great a change in the estimate

Summary

1. What is collinearity/ multicollinearity (MC)?

predictors are nearly linearly dependent

2. How can MC be detected?

bivariate: pairwise scatter plot, correlation matrix multivariate: statistical inference, VIF, Condition number,etc..

3. How does MC affect statistical inference and forecasting?

instability, insignificance, large variance, etc..

4. How can we resolve problems with MC?

Removing variables, More data, principal component analysis (PCA), LASSO and Ridge regression, Recenter the variables (structural MC) etc..