

POL 213 – Spring 2024
Quantitative Analysis in Political Science II
Lecture 5
Multicollinearity

Lauren Peritz

U.C. Davis

`lperitz@ucdavis.edu`

May 2, 2024

Last time...some more details

Last time, we discussed variable transformations. The goal was to make our data and the model specification as close as possible to satisfying the Gauss Markov assumptions so that we can make reliable inferences using OLS.

A few useful transformations that are readily available in R:

- ▶ Log transformation (select appropriate base)
- ▶ Standardization ($Z = \frac{x_i - \bar{x}}{\sigma}$)
- ▶ Box-cox transformation = power transformation of Y so that error distribution is normalized and error variance is stabilized.
- ▶ Polynomial terms (e.g. x^2)

Today, we will pivot to a tricky problem with the relationship among our explanatory variables.

Table of Contents

(Multi) collinearity is a problem

Detecting and addressing collinearity

Concluding remarks on linear regression

Collinearity is a Problem

(Multi) collinearity is when two or more regressors in a multiple regression model are strongly correlated. It a problem for regression analysis.

- ▶ When there is a perfect linear relationship among regressors in a linear model, the LS coefficients are not uniquely defined (singularities).
- ▶ A strong, imperfect linear relationship among the Xs causes the LS coefficients to be unstable
 - ▶ coefficient standard errors are large reflecting imprecision
 - ▶ broad confidence intervals
 - ▶ hypothesis test have low power
- ▶ Even small changes in the data can produce huge swings in LS coefficients.

Example of *Perfect* Multicollinearity Problem

```
library(AER)
library(MASS)
data(CASchools)

# define variables
CASchools$STR <- CASchools$students/CASchools$teachers
CASchools$score <- (CASchools$read + CASchools$math)/2

# define the fraction of English learners
CASchools$FracEL <- CASchools$english / 100

# estimate the model
mult.mod <- lm(score ~ STR + english + FracEL, data = CASchools)

# obtain a summary of the model
summary(mult.mod)
```

Source: <https://www.econometrics-with-r.org/rmwmr.html>

Perfect collinearity means that the computation of OLS model simply fails.

```
##
## Call:
## lm(formula = score ~ STR + english + FracEL, data = CASchools)
##
## Residuals:
```

| | Min | 1Q | Median | 3Q | Max |
|--|---------|---------|--------|-------|--------|
| | -48.845 | -10.240 | -0.308 | 9.815 | 43.461 |

```
##
## Coefficients: (1 not defined because of singularities)
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 686.03224    7.41131  92.566 < 2e-16 ***
## STR          -1.10130    0.38028  -2.896  0.00398 **
## english      -0.64978    0.03934 -16.516 < 2e-16 ***
## FracEL                NA          NA      NA      NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.46 on 417 degrees of freedom
## Multiple R-squared:  0.4264, Adjusted R-squared:  0.4237
## F-statistic: 155 on 2 and 417 DF, p-value: < 2.2e-16
```

Another example: if you use include an intercept and all possible dummy variable categories. Suppose we include an indicator variable for whether a school is located in one of four regions of the country.

```
# set seed for reproducibility
set.seed(1)

# generate artificial data on location
CASchools$direction <- sample(c("West", "North", "South", "East"),
420, replace = T)

# estimate the model
mult.mod <- lm(score ~ STR + english + direction, data = CASchools)

# obtain a model summary
summary(mult.mod)
```

R solves this problem on its own by dropping one of the dummy variable categories. Another solution would be to exclude the intercept and include all dummies instead.

```
## lm(formula = score ~ STR + english + direction, data = CASchools)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -48.018 -10.098  -0.556   9.304  42.596
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   685.67356    7.43308  92.246 < 2e-16 ***
## STR           -1.12246    0.38231  -2.936  0.00351 **
## english       -0.65096    0.03934 -16.549 < 2e-16 ***
## directionNorth  1.60680    1.92476   0.835  0.40431
## directionSouth -1.17013    2.07665  -0.563  0.57342
## directionWest   2.44340    2.05191   1.191  0.23442
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.45 on 414 degrees of freedom
## Multiple R-squared:  0.4315, Adjusted R-squared:  0.4246
## F-statistic: 62.85 on 5 and 414 DF, p-value: < 2.2e-16
```


Imperfect multicollinearity

More often, we have imperfect multicollinearity in which there is strong dependence among explanatory (X) variables. Then we have unstable and imprecise coefficient estimates.

Take regression model:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$$

Suppose you are interested in estimating β_1 , the effect on Y_i of a one unit change in X_{1i} while holding X_{2i} constant. You included X_2 because you want to avoid potential omitted variable bias. If X_1 and X_2 are highly correlated, OLS struggles to precisely estimate β_1 .

Let's simulate this problem.

Simulation of imperfect multicollinearity

1. Make fake data. Choose $\beta = [5, 2.5, 3]$ and error term $u_i \sim N(0, 5)$. Start by sampling regressor data from a bivariate normal distribution in which the correlation is low:

$$(X_{1i}, X_{2i}) \stackrel{d}{\sim} N \left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 10, 2.5 \\ 2.5, 10 \end{pmatrix} \right]$$

the correlation between X s is low: $\rho_{X_1, X_2} = 0.25$

2. Estimate the model, save the estimates $\hat{\beta}_1$ and $\hat{\beta}_2$. Repeat 1000 times in order to obtain sampling distributions for the parameters.
3. Repeat steps 1 and 2 but with a higher covariance such that $\rho_{X_1, X_2} = 0.85$.
4. In order to assess the effect on the precision of the estimators of increasing collinearity between X_1 and X_2 , we estimate the variances of $\hat{\beta}_1$ and $\hat{\beta}_2$ in each scenario and compare.

```

library(MASS)
library(mvtnorm)
set.seed(1)

# initialize vectors of coefficients
n <- 50
coefs1 <- cbind("hat_beta_1" = numeric(10000), "hat_beta_2" = numeric(10000))
coefs2 <- coefs1

# loop sampling and estimation
for (i in 1:10000) {
  # for cov(X_1,X_2) = 0.25
  X <- rmvnorm(n, c(50, 100), sigma = cbind(c(10, 2.5), c(2.5, 10)))
  u <- rnorm(n, sd = 5)
  Y <- 5 + 2.5 * X[, 1] + 3 * X[, 2] + u
  coefs1[i, ] <- lm(Y ~ X[, 1] + X[, 2])$coefficients[-1]

  # for cov(X_1,X_2) = 0.85
  X <- rmvnorm(n, c(50, 100), sigma = cbind(c(10, 8.5), c(8.5, 10)))
  Y <- 5 + 2.5 * X[, 1] + 3 * X[, 2] + u
  coefs2[i, ] <- lm(Y ~ X[, 1] + X[, 2])$coefficients[-1]
}

```

```
# obtain variance estimates
diag(var(coefs1))

# hat_beta_1 hat_beta_2
0.05674375 0.05712459

diag(var(coefs2))

# hat_beta_1 hat_beta_2
0.1904949 0.1909056
```

Due to the high collinearity, the variances of $\hat{\beta}_1$ and $\hat{\beta}_2$ have more than tripled, meaning it is more difficult to precisely estimate the true coefficients.

Table of Contents

(Multi) collinearity is a problem

Detecting and addressing collinearity

Concluding remarks on linear regression

Detecting multicollinearity

When there is imperfect collinearity, the sampling variance of the LS slope coefficient B_j is

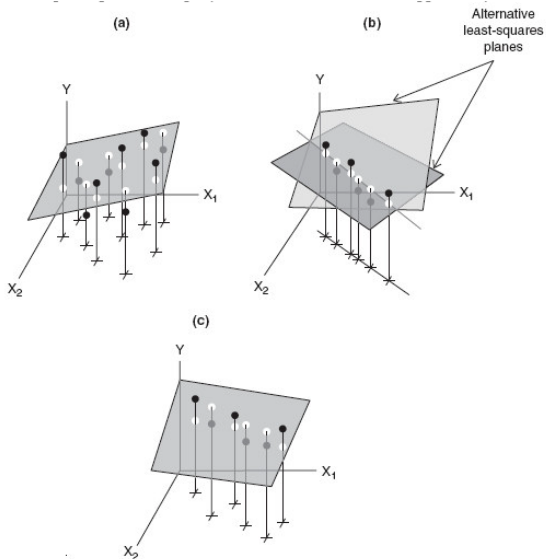
$$V(B_j) = \frac{1}{1 - R_j^2} \times \frac{\sigma_\varepsilon^2}{(n - 1)S_j^2}$$

The equation has the following quantities:

- ▶ R_j^2 is the squared multiple correlation for regression of X_j on other X s
- ▶ $S_j^2 = \sum(X_{ij} - \bar{X}_j)^2 / (n - 1)$ is the variance of X_j .
- ▶ The first term is the variance inflation factor which indicates the impact of collinearity on the precision of B_j

$$VIF = 1/(1 - R_j^2)$$

Impact of multicollinearity



Source: Fox (2015) fig 13-2

Impact of multicollinearity

The preceding figure shows the impact of collinearity on the stability of the least-squares regression plane.


- (a) the correlation between X_1 and X_2 is small, and the regression plane therefore has a broad base of support
- (b) X_1 and X_2 are perfectly correlated; the least-squares plane is not uniquely defined
- (c) there is a strong, but less-than-perfect, linear relationship between X_1 and X_2 ; the least-squares plane is uniquely defined, but it is not well supported by the data

No Easy Fix

There aren't great options for simultaneously addressing multicollinearity and reducing the risk of omitted variable bias.

1. Variable selection approaches can sometimes work, if used with caution. Sometimes referred to as *stepwise regression* on the basis of fit criterion like BIC.¹ However, can easily worsen model problems ('snake oil').
2. In social science, data scarcity (small n), insufficient variation in explanatory variables, large error variance are usually more serious problem (except time series regression)
3. Increasing sample size or decreasing error variance in measurement can help

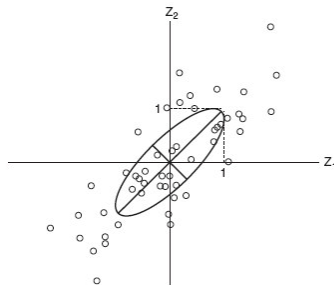
One useful approach is implementing a dimension reduction technique such as a *principal components analysis*. The idea is to distill a bunch of multicollinear X 's down to their constituent dimensions.

¹BIC can be helpful when collinearity is not a serious issue. 

Principal Components Analysis - Visualize

This is a more advanced topic but let's introduce the general idea.

Figure 13.6 The principal components for two standardized variables Z_1 and Z_2 are the principal axes of the standard data ellipse $\mathbf{z}'\mathbf{R}_{XX}^{-1}\mathbf{z} = 1$. The first eigenvalue L_1 of \mathbf{R}_{XX} gives the half-length of the major axis of the ellipse; the second eigenvalue L_2 gives the half-length of the minor axis. In this illustration, the two variables are correlated $r_{12} = .8$, so L_1 is large and L_2 is small.



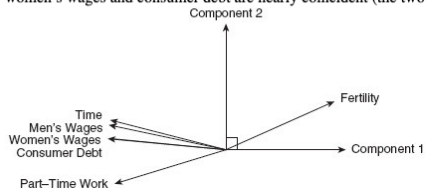
Source: Fox (2015) fig 13-6

Principal Components Analysis - Visualize

Table 13.3 Principal-Components Coefficients for the Explanatory Variables in B. Fox's Regression

| Variable | Principal Component | | | | | |
|-----------------------|---------------------|---------|---------|---------|---------|---------|
| | W_1 | W_2 | W_3 | W_4 | W_5 | W_6 |
| Fertility | 0.3849 | 0.6676 | 0.5424 | 0.2518 | -0.1966 | -0.0993 |
| Men's Wages | -0.4159 | 0.3421 | -0.0223 | 0.1571 | 0.7055 | -0.4326 |
| Women's Wages | -0.4196 | 0.1523 | -0.2658 | 0.7292 | -0.2791 | 0.3472 |
| Consumer Debt | -0.4220 | 0.1591 | -0.0975 | -0.2757 | -0.6188 | -0.5728 |
| Part-Time Work | -0.3946 | -0.4693 | 0.7746 | 0.1520 | -0.0252 | -0.0175 |
| Time | -0.4112 | 0.4106 | 0.1583 | -0.5301 | 0.0465 | 0.5951 |
| Eigenvalue | 5.5310 | 0.3288 | 0.1101 | 0.0185 | 0.0071 | 0.0045 |
| Cumulative percentage | 92.18 | 97.66 | 99.50 | 99.81 | 99.93 | 100.00 |

Figure 13.7 Orthogonal projections of the six explanatory variables onto the subspace spanned by the first two principal components. All the variables, including the components, are standardized to common length. The projections of the vectors for women's wages and consumer debt are nearly coincident (the two vectors are essentially on top of one another).



Summary

Source: Fox (2015) fig 13-7

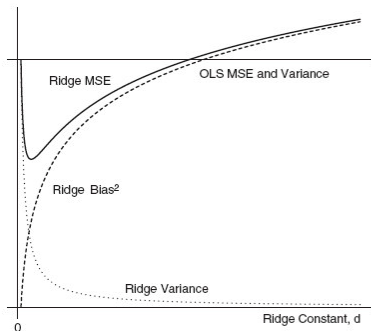
Ridge regression for biased estimation

Another approach to multicollinear data is biased estimation.

- ▶ Key idea is to trade a small amount of bias in the coefficient estimates for a large reduction in coefficient sampling variance.
- ▶ The hoped for result is a smaller mean-squared error in the estimation of the β s.
- ▶ A common technique is *ridge regression*. Because it requires choosing an arbitrary “ridge constant” tolerance parameter, it requires some knowledge about the true values of the unknown β s we are trying to estimate.
- ▶ Use biased estimation with caution!!

Ridge Regression - Bias Variance Tradeoff

Figure 13.9 Trade-off of bias and against variance for the ridge-regression estimator. The horizontal line gives the variance of the least-squares (OLS) estimator; because the OLS estimator is unbiased, its variance and mean-squared error are the same. The broken line shows the squared bias of the ridge estimator as an increasing function of the ridge constant d . The dotted line shows the variance of the ridge estimator. The mean-squared error (MSE) of the ridge estimator, given by the heavier solid line, is the sum of its variance and squared bias. For some values of d , the MSE error of the ridge estimator is below the variance of the OLS estimator.



Source: Fox (2015) fig 13-9

Table of Contents

(Multi) collinearity is a problem

Detecting and addressing collinearity

Concluding remarks on linear regression

Linear Regression Best Practices

1. Know your data. Outliers, leverage, clusters, influence.
2. Consider variable transformations, interaction effects, and collinearity when choosing model specification.
3. Run diagnostics to evaluate distribution of residuals and ensure key distributional assumptions hold
4. **GOAL: Satisfy the Gauss Markov Assumptions** and if they are not fully met, understand the degree of violation.
5. We conduct research with imperfect research designs, measurement strategies, and data availability. These approaches have to be considered in relation to the the sampling process.