STP 530
Lecture 6: Multicollinearity

Yi Zheng, Ph.D.
yi.isabel.zheng@asu.edu

ARIZONA STATE UNIVERSITY
SCHOOL OF MATHEMATICAL AND STATISTICAL SCIENCES

October 3, 2023

Multicollinearity

Definition: When two or more of the independent variables used in the model are moderately or highly linearly dependent, we say that **multicollinearity** exits.

Consequence: Intuitively, when there are two predictors and the two predictors are correlated, it's difficult to define the regression surface.

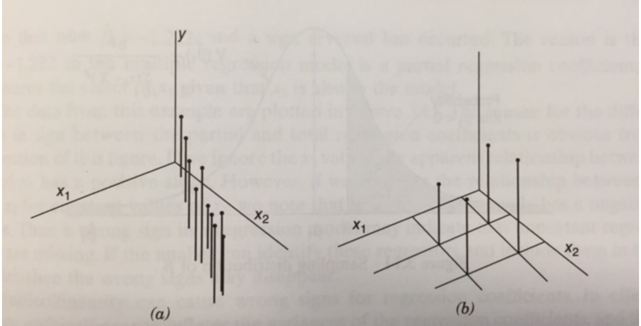


Figure 3.13 (*a*) A data set with multicollinearity. (*b*) Orthogonal regressors.

Consequences of multicollinearity

1. Inflated **standard errors** of the parameters, which means the coefficient estimates vary widely from sample to sample, and the estimates from your one sample are less trustworthy.
2. The coefficient estimates can swing wildly based on which other independent variables are in the model.
3. The regression results may be confusing and misleading, such as reversing the signs of the parameter estimates.
4. Multicollinearity makes it difficult to gauge the effect of one predictor on the dependent variable.
5. Extremely high correlations among the independent variables (i.e., **EXTREME** multicollinearity) increase the likelihood of rounding errors in the **estimation** of the model coefficients.

Multicollinearity may not reveal itself if you do not check on it proactively!

A few ways to detect multicollinearity:

1. Calculate the coefficient of correlation between each pair of independent variables. If one or more is close to 1 or -1, severely multicollinearity exists.

   Note: Examining the pairwise correlations is helpful in detecting near-linear dependence between **pairs of regressors** only. When more than two regressors are involved in a near-linear dependence, there is no assurance that any of the pairwise correlations will be large.

3. A **variance inflation factor (VIF)** for any β parameter greater than 10, where

$$\text{VIF}_i = \frac{1}{1 - R_i^2}, \qquad i = 1, 2, \cdots, k$$

R_i^2 is the coefficient of determination for regressing X_i on all the other $(p-2)$ predictors, and it is also the diagonal elements of the matrix $\mathbf{X}'\mathbf{X}$.

When all variables in the model are standardized, it can be shown that

$$\text{Var}\{b_i\} = s^2 \frac{1}{1 - R_i^2} = s^2 * \text{VIF}_i$$

where $s^2 = \text{SSE}/(n-p)$

Detect multicollinearity

4. Nonsignificant t-test for all (or nearly all) the individual β parameters but the F-test for overall model adequacy is significant.

5. Opposite signs from what is expected in the estimated parameters.

6. If adding or removing a predictor produces large changes in the estimates of the regression coefficients, multicollinearity is indicated.

Solutions to multicollinearity:

1. Drop one or more of the correlated independent variables.
2. If you decide to keep all independent variables in the model, avoid making inferences about individual β coefficients.
3. Use a designed experiment so that the levels of the X variables are uncorrelated.
4. For polynomial or interaction regression models with multicollinearity, center or standardize the independent variables.

Examples of causes of multicollinearity

Two causes of **non-extreme** multicollinearity:

- Cause 1: Correlated variables in observational studies
- Cause 2: Inclusion of polynomial or interaction terms

Three causes of **extreme** multicollinearity:

- Cause 3: Computed variable
- Cause 4: Subsetted data
- Cause 5: Poorly designed experiment with confounded factors

Cause 1: Correlated variables in observational studies

Example: Suppose we would like to identify factors that can explain variations in salaries among programmers and engineers in the Silicon Valley. We use data collected during the 2000 U.S. Census (R package `freqparcoord`, data `prgeng`).

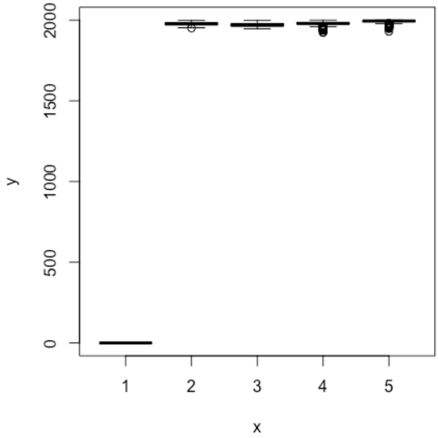
Two available variables, which come from two census questions, are:

- **cit:** Citizenship status
  - 1: Yes, born in the United States
  - 2: Yes, born in the U.S. islands
  - 3: Yes, born abroad of American parent or parents
  - 4: Yes, U.S. citizen by naturalization
  - 5: No, not a citizen of the United States
- **yrentry:** Year of entry to the U.S. (0 for natives)

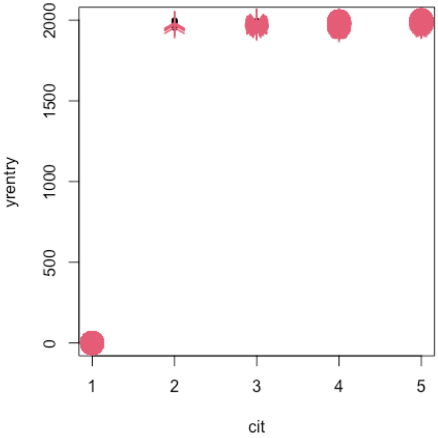
The two variables have multicollinearity because all $cit = 1$ individuals also have $yrentry = 0$.

Examples of causes of NON-EXTREME multicollinearity

All $cit = 1$ individuals also have $yrentry = 0$.



A sunflower plot shows the volume of cases sharing the same coordinates.



Examples of causes of NON-EXTREME multicollinearity

High VIF observed (but they are the **non-extreme** type):

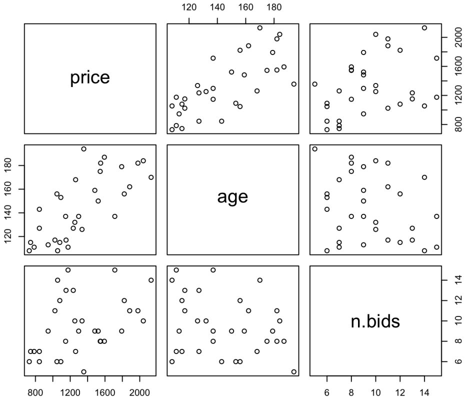
```
> m1 <- lm(wageinc ~ age + cit + educ + sex + wkswrkd + yrentry, data=prgeng)
> vif(m1)
                GVIF Df GVIF^(1/(2*Df))
age         1.171420  1        1.082322
cit     39215.481109  4        3.751303
educ        1.109502  1        1.053329
sex         1.007046  1        1.003517
wkswrkd     1.028155  1        1.013980
yrentry 37276.395709  1      193.070960
> summary(m1)

Call:
lm(formula = wageinc ~ age + cit + educ + sex + wkswrkd + yrentry,
    data = prgeng)

Residuals:
    Min      1Q  Median      3Q     Max
-103952  -20369   -4667   12750  291165

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       24593.18  120206.08   0.205    0.838
age                 455.69      29.16  15.627   <2e-16 ***
citBorn.US      -106715.21  120019.20  -0.889    0.374
citBorn.islands  -10206.54    6789.76  -1.503    0.133
citBorn.abroad    -2884.05    3179.92  -0.907    0.364
citNaturalized      627.99    1263.72   0.497    0.619
educ               5301.82     194.33  27.283   <2e-16 ***
sexfemale         -9957.31     706.14 -14.101   <2e-16 ***
wkswrkd            1304.29      21.00  62.117   <2e-16 ***
yrentry             -52.57      60.27  -0.872    0.383
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 42850 on 20080 degrees of freedom
Multiple R-squared:  0.2285,    Adjusted R-squared:  0.2281
F-statistic: 660.7 on 9 and 20080 DF,  p-value: < 2.2e-16
```

Removing *yrentry* from the model resolved multicollinearity. Coefficients for *cit* changed dramatically. Coefficients for other predictors and model predictions remain stable.

```
> m2 <- lm(wageinc ~ age + cit + educ + sex + wkswrkd, data=prgeng)
> vif(m2)
            GVIF Df GVIF^(1/(2*Df))
age     1.089244  1        1.043669
cit     1.166637  4        1.019452
educ    1.100723  1        1.049154
sex     1.007042  1        1.003515
wkswrkd 1.025048  1        1.012447
> summary(m2)

Call:
lm(formula = wageinc ~ age + cit + educ + sex + wkswrkd, data = prgeng)

Residuals:
    Min      1Q  Median      3Q     Max
-103021  -20372   -4655   12746  291203

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     -80226.59    2778.62 -28.873   <2e-16 ***
age                462.43      28.12  16.446   <2e-16 ***
citBorn.US       -2032.94     831.93  -2.444   0.0145 *
citBorn.islands  -9420.85    6729.70  -1.400   0.1616
citBorn.abroad   -1808.11    2930.90  -0.617   0.5373
citNaturalized    1287.65    1012.43   1.272   0.2034
educ              5286.75     193.56  27.313   <2e-16 ***
sexfemale        -9956.01     706.14 -14.099   <2e-16 ***
wkswrkd           1305.29      20.97  62.260   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 42850 on 20081 degrees of freedom
Multiple R-squared:  0.2284,    Adjusted R-squared:  0.2281
F-statistic: 743.2 on 8 and 20081 DF,  p-value: < 2.2e-16
```

Cause 2: Inclusion of polynomial or interaction terms

Example: Suppose we want to predict the price (Y) received for antique clocks sold at auction using the age of the clocks (X_1) and the number of bidders at the auction (X_2).



Scatter plot matrix: `pairs(clock)`

No associate between X_1 and X_2 — No multicollinearity seen here.

Fit the interaction model. Multicollinearity is found (non-extreme type).

```
> m1 <- lm(price ~ age + n.bids + age:n.bids)
> summary(m1)

Call:
lm(formula = price ~ age + n.bids + age:n.bids)

Residuals:
     Min       1Q   Median       3Q      Max 
-154.995  -70.431    2.069   47.880  202.259 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 320.4580   295.1413   1.086  0.28684    
age           0.8781     2.0322   0.432  0.66896    
n.bids      -93.2648    29.8916  -3.120  0.00416 ** 
age:n.bids    1.2978     0.2123   6.112 1.35e-06 ***
---
Signif. codes:  
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 88.91 on 28 degrees of freedom
Multiple R-squared:  0.9539,	Adjusted R-squared:  0.9489 
F-statistic:   193 on 3 and 28 DF,  p-value: < 2.2e-16

> vif(m1) # Found very big vif values
       age     n.bids age:n.bids 
  12.15313   28.25135   30.45770 
```

Center each predictor variable:

$$\text{age.c} = \text{age} - mean(\text{age}) \qquad \text{nbids.c} = \text{nbids} - mean(\text{nbids})$$

```
Call:
lm(formula = price ~ age.c + n.bids.c + age.c:n.bids.c)

Residuals:
     Min       1Q   Median       3Q      Max 
-154.995  -70.431    2.069   47.880  202.259 

Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)    1397.0169    16.6473  83.918  < 2e-16 ***
age.c            13.8566     0.6297  22.005  < 2e-16 ***
n.bids.c         94.9228     5.9964  15.830 1.69e-15 ***
age.c:n.bids.c    1.2978     0.2123   6.112 1.35e-06 ***
---
Signif. codes:  
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 88.91 on 28 degrees of freedom
Multiple R-squared:  0.9539,	Adjusted R-squared:  0.9489 
F-statistic:   193 on 3 and 28 DF,  p-value: < 2.2e-16

> vif(m2)
         age.c       n.bids.c age.c:n.bids.c 
      1.166929       1.136912       1.124721 
```

The standard errors of the first order terms have decreased noticeably, which indicates that centering did alleviate multicollinearity and recovered the previously inflated standard errors of the model coefficients.

The VIF values are now close to 1.

Two causes of **non-extreme** multicollinearity:

- Cause 1: Correlated variables in observational studies
- Cause 2: Inclusion of polynomial or interaction terms

Three causes of **extreme** multicollinearity:

- Cause 3: Computed variable
- Cause 4: Subsetted data
- Cause 5: Poorly designed experiment with confounded factors

Cause 3: Computed variable

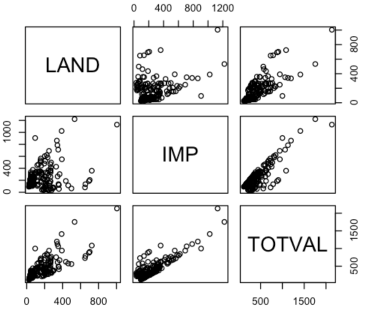
Example: Suppose we would like to predict the sale price of a house. Our dataset includes the following variables.

Variable Name	Description
SALES	Sales price (in thousands of dollars)
LAND	Land value (in thousands of dollars)
IMP	Value of improvments (in thousands of dollars)
TOTVAL	Total appraised value for the house (in thousands of dollars), which is the summation of LAND and IMP
NBHD	Neighborhood (HYDEPARK, DAVISISLES, CHEVAL, HUNTERSGREEN)

The model `SALES ~ LAND + IMP + TOTVAL` has **extreme** multicollinearity because $TOTVAL = LAND + IMP$.

However, you won't see this extreme multicollinearity in the pairwise correlations.

```
> cor(data.frame(LAND, IMP, TOTVAL))
            LAND       IMP    TOTVAL
LAND   1.0000000 0.3461659 0.7713288
IMP    0.3461659 1.0000000 0.8640958
TOTVAL 0.7713288 0.8640958 1.0000000
```



But you will see it in the VIFs:

```
> m <- lm(SALES ~ LAND + IMP + TOTVAL)
> vif(m)
      LAND        IMP     TOTVAL
1519009999 2428679102 5277466707
```

Cause 4: Subsetted data

Example: Suppose we would like to assess the impact of deregulation on the prices charged for trucking service. Our dataset includes the following variables:

- **PRICPTM:** Price per ton mile charged for trucking service
- **DISTANCE:** Miles traveled
- **WEIGHT:** Weight of product shipped
- **PCTLOAD:** Percent of truck load capacity
- **CARRIER:** Each truck is coded as a different carrier
- **DEREG:** Deregulation in effect (YES or NO)

There is no inherent extreme multicollinearity among the predictors. However, if you decide to focus on a single carrier, the subset data now has **extreme** multicollinearity because for the same truck, *PCTLOAD* is proportionate to *WEIGHT*.

Cause 5: Poorly designed experiment with confounded factors

Example: Suppose we would like to assess the effect of a new learning material on students' learning outcome, compared with the existing learning material.

We conduct an experiment where Teacher A teaches a class of students with the existing material, Teacher B teaches a class of students with the new material. We measure students' knowledge pre- and post-intervention.

We believe that the students' ability and teacher's teaching skill both correlate with the dependent variable, which constitute "nuisance factors" and should be controlled in the regression model. Thus we fit the following model:

$$post.score \sim pre.score + teacher + material$$

Examples of causes of EXTREME multicollinearity

$$post.score \sim pre.score + teacher + material$$

This model has **extreme** multicollinearity because *teacher* coincide with *material*. This experimental design fails to isolate the effects of the learning material and the teacher.

To improve the design, you may:

- Use the same teacher to teach both materials (hold the teacher factor constant), or
- Use more teachers to teach each material, and include a measure of teachers' teaching skill, or
- Use more teachers and cross the teacher factor with the material factor (completely factorial design).

Two good things to know when you have NON-EXTREME multicollinearity:

1. Multicollinearity affects only the specific predictors that are correlated. If multicollinearity is not present for the predictors that you are particularly interested in, you may not need to resolve it.

   Suppose your model contains the experimental variables of interest and some control variables. If non-extreme multicollinearity exists for the control variables but not the experimental variables, then you can interpret the experimental variables without problems.

2. Non-extreme multicollinearity affects the coefficients and their p-values, but it does not influence the predictions, precision of the predictions, and the model-fit statistics. If your primary goal is to make predictions, and you don't need to understand the role of each independent variable, you don't need to resolve multicollinearity.