Collinearity and Biased Estimation

EPsy 8264

2018-10-06

Collinearity is defined mathematically as when any of the predictors (X) in a regression model is a perfect linear combination of the other predictors:

$$X_{k+1} = c_0(1) + c_1X_1 + c_2X_2 + c_3X_3 + \ldots + c_kX_k$$

When this happens, the matrix $\mathbf{X}^\intercal\mathbf{X}$ is singular, and the OLS normal equations do not have a unique solution. Recall that the sampling variance for a predictor, B_j is

$$\text{Var}(B_j) = \frac{1}{1 - R_j^2} \times \frac{\sigma_\epsilon^2}{(n-1)S_j^2}$$

The first term in this product is referred to as the *variance inflation factor* (VIF). Recall that R_j^2 in this term is the squared multiple correlation between X_j on the other Xs. - When the correlation between X_j and the other Xs is 0 (they are independent; *orthogonal*) then the VIF becomes 1. - When the correlation between X_j and the other Xs is high then the VIF becomes larger than 1; it becomes a multiplier of the variance. - When there is perfect correlation between X_j and the other Xs then the VIF approaches ∞; the sampling variances (and SEs) are infinitely large.

Collinearity in Practice

In practice it is rare to have perfect collinearity. When it does happen it is often the result of mis-formulating the model (e.g., including dummy variables in the model for all levels of a categorical variable, as well as the intercept).

It is, however, common to have strong less-than-perfect collinearity in practice. In thes cases the VIF will be less than 1, but can still have an adverse effect on the sampling variances; making them quite large.

Detecting Collinearity

We can empirically diagnose problematic collinearity in the data using several methods. Before we do, however, it is important to first determine the functional form of the model (model specification). Collinearity produces unstable estimates of the coefficients and sampling variances which result from strong linear relationships between the predictors. In applied work, the model needs to be specified before we can estimate coefficients or their sampling variances; hence, collinearity should only be investigated after the model has been satisfactorily specified.

Data Set

In 1964, the US Congress passed the Civil Rights Act and also ordered a survey of school districts to evaluate the availability of equal educational opportunity in public education. The results of this survey were reported on in Coleman et al. (1966) and Mosteller & Moynihan (1972). The data in *equal-educational-opportunity.csv* consist of data taken from a random sample of 70 schools in 1965. The varibales, which have all been mean-centered and standardized, include:

- `achievement`: Measurement indicating the student achievement level
- `faculty`: Measurement indicating the faculty's credentials
- `peer`: Measurement indicating the influence of peer groups in the school
- `school`: Measurement indicating the school facilities (e.g., building, teaching materials)

We will use these data to mimic one of the original regression analyses performed; examining whether the level of school facilities was an important predictor of student achievement after accounting for the variation in faculty credientials and peer influence.

```
# Load libraries
library(broom)
library(car)
```

```
## Loading required package: carData
```

```
library(corrr)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following object is masked from 'package:car':
##
##     recode

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)
library(readr)
library(sm)
```

```
## Package 'sm', version 2.2-5.5: type help(sm) for summary information
```

```
# Read in data
eeo = read_csv("~/Dropbox/epsy-8264/data/equal-education-opportunity.csv")
```

```
## Parsed with column specification:
## cols(
##   achievement = col_double(),
##   faculty = col_double(),
##   peer = col_double(),
##   school = col_double()
## )
```

```
head(eeo)
```

```
## # A tibble: 6 x 4
##   achievement faculty    peer school
##         <dbl>   <dbl>   <dbl>  <dbl>
```

```
## 1      -0.431   0.608  0.0351  0.166
## 2       0.800   0.794  0.479   0.534
## 3      -0.925  -0.826 -0.620  -0.786
## 4      -2.19   -1.25  -1.22   -1.04
## 5      -2.85    0.174 -0.185   0.142
## 6      -0.662   0.202  0.128   0.273
```

Regression Analysis

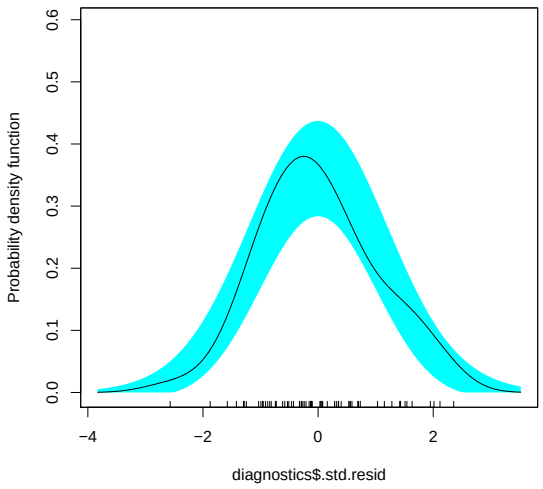
To examine the RQ, the following model was posited:

$$\text{Achievement}_i = \beta_0 + \beta_1(\text{Faculty}_i) + \beta_2(\text{Peer}_i) + \beta_3(\text{School}_i) + \epsilon_i$$

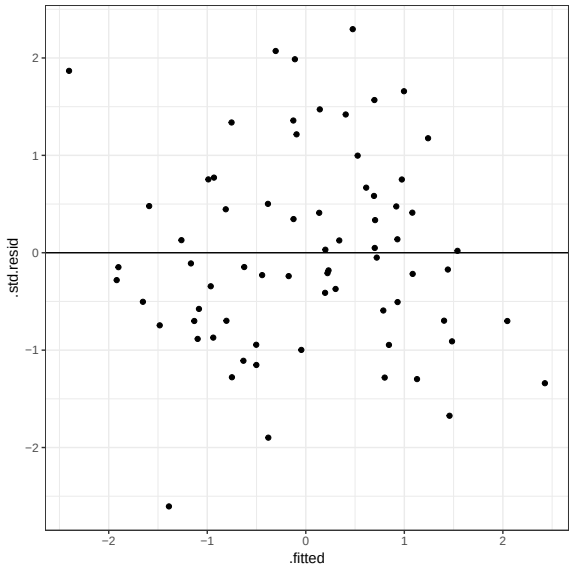
```
# Fit the regression model
lm.1 = lm(achievement ~ faculty + peer + school, data = eeo)

# Examine residuals
diagnostics = augment(lm.1)

# Check Normality
sm.density(diagnostics$.std.resid, model = "normal")
```




```
# Check linearity and constant variance
ggplot(data = diagnostics, aes(x = .fitted, y = .std.resid)) +
  geom_point() +
  geom_hline(yintercept = 0) +
  theme_bw()
```



The school that has a fitted value of -1.38 and studentized residual of -2.6 may be problematic.

```
# Model-level information
glance(lm.1)
```

```
## # A tibble: 1 x 11
##   r.squared adj.r.squared sigma statistic p.value    df logLik   AIC   BIC
## *     <dbl>         <dbl> <dbl>     <dbl>   <dbl> <int>  <dbl> <dbl> <dbl>
## 1     0.206         0.170  2.07      5.72 0.00153     4  -148.  306.  318.
## # ... with 2 more variables: deviance <dbl>, df.residual <int>
```

```
# Coefficient-level information
tidy(lm.1)
```

```
## # A tibble: 4 x 5
##   term        estimate std.error statistic p.value
##   <chr>          <dbl>     <dbl>     <dbl>   <dbl>
## 1 (Intercept)  -0.0700     0.251    -0.279   0.781
## 2 faculty       1.10       1.41      0.781   0.438
## 3 peer          2.32       1.48      1.57    0.122
## 4 school       -2.28       2.22     -1.03    0.308
```

Examining this information we find:

- 20% of the variation in student achievement is explained by the model; which is statistically significant $F(3, 66) = 5.72$; $p = 0.002$.
- However, none of the individual coefficients are statistically significant!

These results are typical when there is collinearity in the model. In this example, the collinearity could have been anticipated by examining the pairwise correlations between the predictors.

```
eeo %>%
  correlate()
```

```
##
## Correlation method: 'pearson'
## Missing treated using: 'pairwise.complete.obs'
```

```
## # A tibble: 4 x 5
##   rowname     achievement faculty   peer school
##   <chr>             <dbl>   <dbl>  <dbl>  <dbl>
## 1 achievement      NA       0.419  0.440  0.418
## 2 faculty           0.419  NA      0.960  0.986
## 3 peer              0.440   0.960 NA      0.982
## 4 school            0.418   0.986  0.982 NA
```

All three of the predictors are highly correlated with one another.

Signs of Potential Collinearity

In our example we were alerted to the possible collinearity by finding that the predictors jointly were statistically significant, but that each of the individual predictors were not. Other signs that you may have collinearity problems are:

- Large changes in the size of the estimated coefficients when variables are added to the model;
- Large changes in the size of the estimated coefficients when an observation is added or deleted;
- The signs of the estimated coefficients do not conform to their prior substantively hypothesized directions;
- Large SEs on variables that are expected to be important predictors.

Detecting Collinearity

Unfortunately the source of collinearity may be due to more than just the simple relationships among the predictors. As such, just examining the pairwise correlations is not enough to detect collinearity (although it is a good first step). There are two common methods statisticians use to detect collinearity, (1) computing variance inflation factors for the coefficients, and (2) examining the eigenvalues of the correlation matrix.

Variance Inflation Factor

The variance inflation factor (VIF) is an indicator of the degree of collinearity, where VIF is:

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

The VIF impacts the size of the variance estimates for the regression coefficients, and as such, can be used as a diagnostic of collinearity. In practice, since it is more conventional to use the SE to measure uncertainty, it is typical to use the square root of the VIF as a diagnostic of collinearity in practice. The square root of the VIF expresses the proportional change in the CI for the coefficients.

R_j	R_j^2	VIF	$\sqrt{\text{VIF}}$
0.5	0.25	1.33	1.15
0.6	0.36	1.56	1.25
0.7	0.49	1.96	1.40
0.8	0.64	2.78	1.67
0.9	0.81	5.26	2.29

For example, if the correlation between X_j and the other Xs is 0.9, then the CIs for the coefficients would increase by a factor of of 2.29. The uncertainty in the estimates would more than double!

In our example, we can use the `vif()` function from the **car** package to compute the variance inflation factors for each coefficient.

```
# VIF
vif(lm.1)
```

```
##  faculty     peer   school
## 37.58064 30.21166 83.15544
```

```
# Square root of VIF
sqrt(vif(lm.1))
```

```
##  faculty     peer   school
## 6.130305 5.496513 9.118960
```

All three coefficients are impacted by VIF. The SEs for these coefficients are all more than five times as large as they would be if the predictors were independent.

Eigenvalues of the Correlation Matrix

Each $k \times k$ matrix has a set of k scalars, called eigenvalues (denoted λ) associated with it. These eigenvalues can be arranged in descending order such that,

$$\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \ldots \geq \lambda_k$$

Because the correlation matrix of the predictors is a square matrix, we can find a corresponding set of eigenvalues for this correlation matrix. It turns out that if any of these eigenvalues is exactly equal to zero, there would be a linear dependence among the predictors. In practice, if one of the eigenvalues is quite a bit smaller than the others (and near zero), there is collinearity.

Empirically, we compute the sum of the reciprocals of the eigenvalues

$$\sum_{i=1}^{k} \frac{1}{\lambda_i}$$

If the sum is greater than a given criterion, say, five times the number of predictors, it is a sign of collinearity.

```
# Correlation matrix of predictors
x = cor(eeo[c("faculty", "peer", "school")])

# Compute eigenvalues and eigenvectors
eigen(x)
```

```
## eigen() decomposition
## $values
## [1] 2.951993158 0.040047507 0.007959335
##
## $vectors
```

```
##            [,1]        [,2]       [,3]
## [1,] -0.5761385  0.67939712 -0.4544052
## [2,] -0.5754361 -0.73197527 -0.3648089
## [3,] -0.5804634  0.05130072  0.8126687
```

```r
# Sum of reciprocal of eigenvalues
sum(1 / eigen(x)$values)
```

```
## [1] 150.9477
```

Since this sum is greater than 15 (five times the number of predictors) then we would conclude that there is a collinearity problem.

Condition Indices

One related diagnostic measure of collinearity are the *condition indices* of the correlation matrix; see D.A. Belsley (1991) and D. Belsley, Kuh, & Welsch (1980). The *j*th condition index is defined as

$$\kappa_j = \sqrt{\frac{\lambda_1}{\lambda_j}}$$

for $j = 1, 2, 3, \ldots, k$, where λ_1 is the first (largest) eigenvalue and λ_j is the *j*th eigenvalue.

The first condition index, κ_1, will always be equal to 1, and the other condition indices will be larger than one. The largest condition index, which will be,

$$\kappa_k = \sqrt{\frac{\lambda_1}{\lambda_k}}$$

where λ_k is the smallest eigenvalue, is known as the *condition number* of the correlation matrix. If the condition number is small, it indicates that the predictors are not collinear, whereas large condition numbers are evidence supporting collinearity.

From empirical work, condition numbers that exceed 15 are typically problematic (this indicates that the maximum eigenvalue is more than 225 times greater than the maximum eigenvalue). When the condition number exceeds 30, corrective action will almost surely need to be taken.

```r
# Compute condition indices
sqrt(max(eigen(x)$values) / eigen(x)$values)
```

```
## [1]  1.000000  8.585586 19.258359
```

The condition number of 19.26, suggests strong collinearity among the predictors.

Fixing Collinearity in Practice

Although there are several solutions in practice, none are a magic bullet. Here are three potential fixes:

- Re-specify the model
  - Combine collinear predictors,

  - Drop one (or more) of the collinear predictors—This changes what you are controlling for.
- Biased estimation
  - Trade small amount of bias for a reduction in coefficient variability
- Introduce prior information about the coefficients
  - This can be done formally in the analysis (e.g., Bayesian analysis)
  - It can be used to give a different model specification

Note that although collinearity is a data problem, the most common fixes in practice are to change the model. For example, we could alleviate the collinearity by dropping any two of the predictors and re-fitting the model with only one predictor.

This would fix the problem, but would be unsatisfactory because the resulting model would not allow us to answer the research question. The highly correlated relationships between the predictors is an inherent characteristic of the data generating process we are studying. This makes it difficult to estimate the individual effects of the predictors. Instead, we could look for underlying causes that would explain the relationships we found among the predictors and perhaps re-formulate the model using these underlying causes.

References

Belsley, D. (1991). *Conditioning diagnostics, collinearity and weak data in regression*. New York: John Wiley & Sons.

Belsley, D., Kuh, E., & Welsch, R. (1980). *Regression diagnostics*. New York: Wiley.

Coleman, J. S., Cambell, E. Q., Hobson, C. J., McPartland, J., Mood, A. M., Weinfield, F. D., & York, R. L. (1966). *Equality of educational opportunity*. Washington, DC: U.S. Government Printing Office.

Mosteller, F., & Moynihan, D. F. (1972). *On equality of educational opportunity*. New York: Random House.