Collinearity Diagnostics

A brief introduction to empirical diagnostics to detect collinearity. Example taken from Chatterjee & Hadi (2012).

AUTHOR

Andrew Zieffler

PUBLISHED

Oct. 5, 2020

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 variables, 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 credentials and peer influence.

```r
# Load libraries
library(broom)
library(car)
library(corrr)
library(tidyverse)

# Read in data
eeo = read_csv("~/Documents/github/epsy-8264/data/equal-education-opportunity.csv")
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 assumptions
qqPlot(lm.1)
residualPlot(lm.1)
```

```
[1] 28 35
```

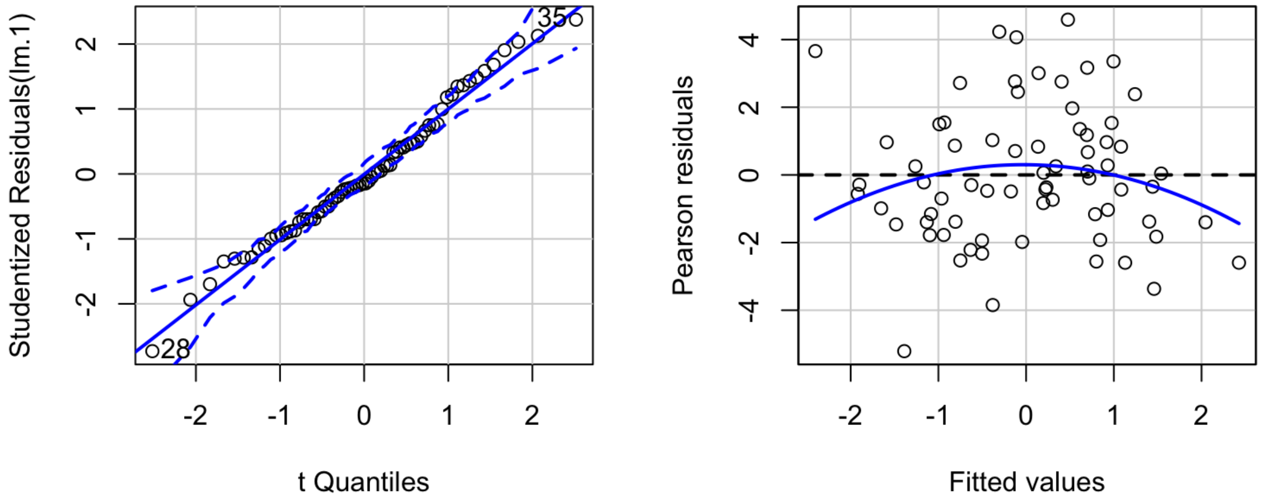

Figure 1: Residual plots for the model that includes the main effects of faculty credentials, influence of peer groups, and measure of school facilities to predict variation in student achievement.

```
# Index plots of several regression diagnostics
```

```
influenceIndexPlot(lm.1)
```

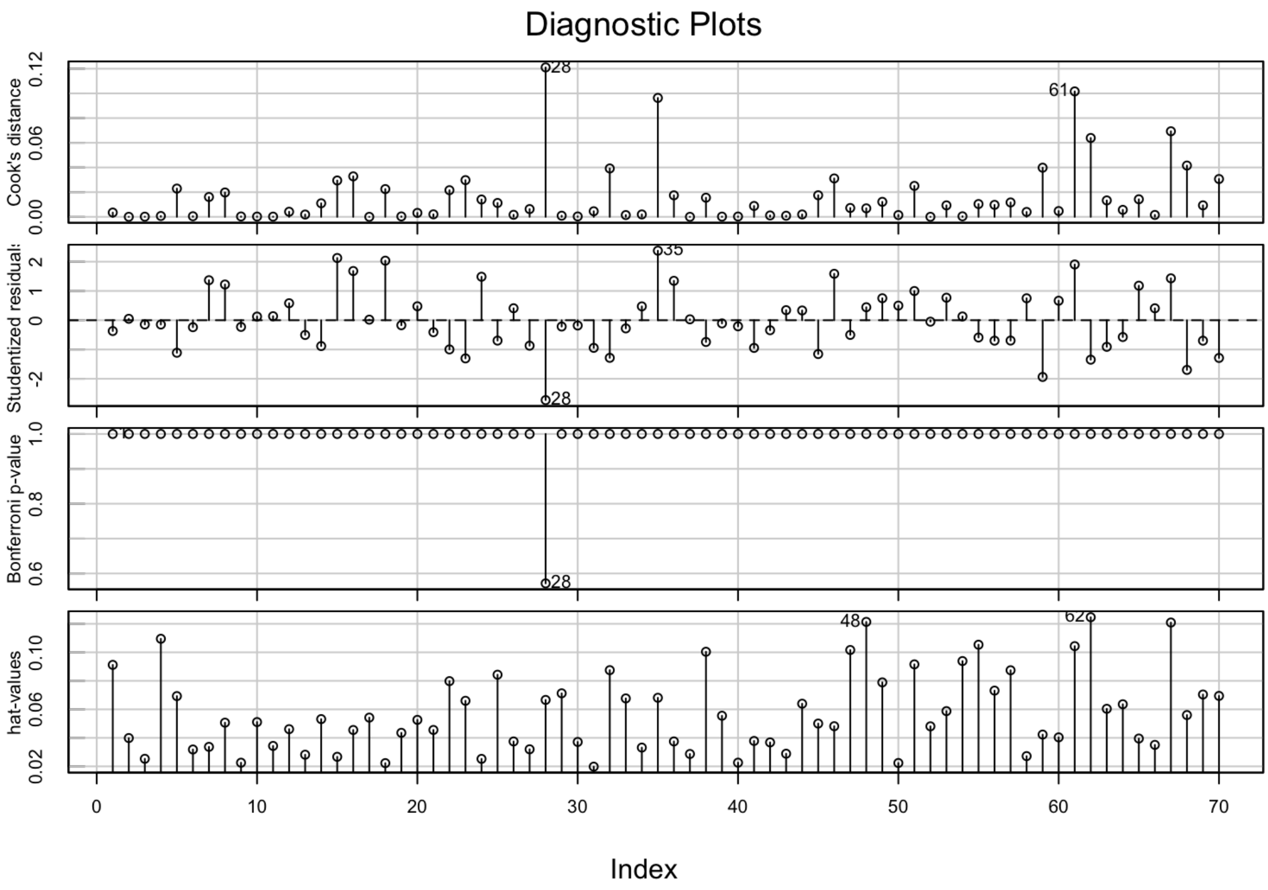

Figure 2: Diagnostic plots for the model that includes the main effects of faculty credentials, influence of peer groups, and measure of school facilities to predict variation in student achievement.

School 28 may be problematic, but removing this observation (work not shown) made little improvement in the residual plots. As such, School 28 was retained in the data. As the assumptions seem reasonably met, we next look to the model-level and coefficient-level output:

```
# Model-level information
print(glance(lm.1), width = Inf)
```

```
# A tibble: 1 x 12
  r.squared adj.r.squared sigma statistic p.value    df logLik   AIC
      <dbl>         <dbl> <dbl>     <dbl>   <dbl> <dbl>  <dbl> <dbl>
1     0.206         0.170  2.07      5.72 0.00153     3  -148.  306.
    BIC deviance df.residual  nobs
  <dbl>    <dbl>       <int> <int>
1  318.     283.          66    70
```

```
# Coefficient-level information
# Coefficient-level information
tidy(lm.1, conf.int = 0.95)
```

```
# A tibble: 4 x 7
  term        estimate std.error statistic p.value conf.low conf.high
  <chr>          <dbl>     <dbl>     <dbl>   <dbl>    <dbl>     <dbl>
1 (Intercept)  -0.0700     0.251    -0.279   0.781   -0.570     0.430
2 faculty       1.10       1.41      0.781   0.438   -1.72      3.92 
3 peer          2.32       1.48      1.57    0.122   -0.635     5.28 
4 school       -2.28       2.22     -1.03    0.308   -6.71      2.15 
```

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.

Collinearity

Mathematically, collinearity occurs when any of the columns of the design matrix, **X**, is a perfect linear combination of the other columns:

$$\mathbf{X_j} = c_0(\mathbf{1}) + c_1\mathbf{X_1} + c_2\mathbf{X_2} + c_3\mathbf{X_3} + \ldots + c_k\mathbf{X_k}$$

and the constants, $c_1, c_2, c_3, \ldots, c_k$ are not all 0. In this situation, **X** is not of full column rank, and the $\mathbf{X}^\intercal\mathbf{X}$ matrix is singular. In this situation, the OLS normal equations do not have a unique solution.

Moreover, the sampling variances for the coefficient are all infinitely large. To understand why this is the case, we can examine one formula for the sampling variance of a slope in a multiple regression:

$$\mathrm{Var}(B_j) = \frac{1}{1 - R^2_j} \times \frac{\sigma^2_\epsilon}{(n-1)S^2_j}$$

where

- R^2_j is the squared multiple correlation for the regression of X_j on the the other predictors;
- S^2_j is the sample variance of predictor X_j defined by $S^2_j = \dfrac{\sum(X_{ij} - \bar{X}_j)^2}{n-1}$;
- σ^2_ϵ is the variance of the residuals based on regressing Y on all the X's

Recall that the multiple correlation is the correlation between the outcome and the predicted values.

The first term in this product is referred to as the *variance inflation factor* (VIF). When one of the

The first term in this product is referred to as the variance inflation factor (VIF). When one of the predictors is perfectly collinear with the others, the value of R^2_j is 1 and the VIF is infinity. Thus the sampling variance of $B_j = \infty$.

Perfect Collinearity in Practice: Model Mis-specification

In practice, it is unlikely that you will have exact 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). As an example of this, imagine that you were creating the design matrix for a regression model that included occupational status (employed/not employed) to predict some outcome for 5 cases.

```
# Create design matrix
X = data.frame(
  b_0 = rep(1, 5),
  employed = c(1, 1, 0, 0, 1),
  not_employed = c(0, 0, 1, 1, 0)
)

# View design matrix
X
```

```
  b_0 employed not_employed
1   1        1            0
2   1        1            0
3   1        0            1
4   1        0            1
5   1        1            0
```

The columns in this design matrix are collinear because we can express any one of the columns as a linear combination of the others. For example,

$$b_0 = 1(\text{employed}) + 1(\text{not employed})$$

Checking the rank of this matrix, we find that this matrix has a rank of 2. Since there are three columns, **X** is not full column rank; it is rank deficient.

```
Matrix::rankMatrix(X)
```

```
[1] 2
attr(,"method")
[1] "tolNorm2"
attr(,"useGrad")
[1] FALSE
```

```
attr(,"tol")
[1] 1.110223e-15
```

Including all three coefficients in the model results in overparameterization. The simple solution here is to drop one of the predictors from the model. This is why we only include a single dummy variable in a model that includes an intercept for a dichotomous categorical predictor.

Including the intercept is not imperative, although it has a useful interpretation when using dummy coding. One could also include the two dummy-coded predictors and omit the intercept. This gives the means for the two groups, but does not provide a comparison of those means.

```r
# Create vector of outcomes
Y = c(15, 15, 10, 15, 30)

# Create data frame of Y and X
my_data = cbind(Y, X)
my_data
```

```
   Y b_0 employed not_employed
1 15   1        1            0
2 15   1        1            0
3 10   1        0            1
4 15   1        0            1
5 30   1        1            0
```

```r
# Coefficients (including all three terms)
coef(lm(Y ~ 1 + employed + not_employed, data = my_data))
```

```
 (Intercept)     employed not_employed 
        12.5          7.5           NA 
```

```r
# Coefficients (omitting intercept)
coef(lm(Y ~ -1 + employed + not_employed, data = my_data))
```

```
    employed not_employed 
        20.0         12.5 
```

If you overparameterize a model with `lm()`, one or more of the coefficients will not be estimated (the last parameters entered in the model).

Constraining some parameters is another way to produce a full rank design matrix. For example the ANOVA model has a constraint that the sum of the effect-coded variable is 0. This constraint ensures that the design matrix will be of full rank.

Non-Exact Collinearity

It is more likely, in practice, that you will have less-than-perfect collinearity, and that this will have an adverse effect on the computational estimates of the coefficients' sampling variances. Again, we look toward how the sampling variances for the coefficent's are computed:

$$\text{Var}(B_j) = \frac{1}{1 - R^2_j} \times \frac{\sigma^2_{\epsilon}}{(n-1)S^2_j}$$

When the predictors are completely independent, all of the columns of the design matrix will be orthogonal and the correlation between X_j and the other Xs will be 0. In this situation, the VIF is 1 and the second term in the product completely defines the sampling variance. This means that the sampling variance is a function of the model's residual variance, sample size, and the predictor's variance—the factors we typically think of affecting the sampling variance of a coefficient.

In cases where the columns in ths design matrix are not perfectly orthogonal, the correlation between X_j and the other Xs is larger than 0. (Perfect collinearity results in $R^2_j = 1$.) For these situations, the VIF has a value that is greater than 1. When this happens the VIF acts as a multiplier of the second term, inflating the the sampling variance and reducing the precision of the estimate (i.e., increasing the uncertainty).

How much the uncertainty in the estimate increases is a function of how correlated the predictors are. Here we can look at various multiple correlations (R_j) between X_j and the predicted values from using the other X's to predict X_j.

Table 1: Impact of various R_j values on the VIF and size of the CI for B_j.

R_j	VIF	CI Factor
0.0	1.00	1.00
0.1	1.01	1.01
0.2	1.04	1.02
0.3	1.10	1.05
0.4	1.19	1.09
0.5	1.33	1.15
0.6	1.56	1.25
0.7	1.96	1.40
0.8	2.78	1.67
0.9	5.26	2.29
1.0	Inf	Inf

For example, a multiple correlation of 0.7 results in a VIF of 1.96, which in turn means that the CI (which is based on the square root of the sampling variance) will increase by a factor of 1.4. This inflation increases the uncertainty of the estimate making it harder to make decisions or understand the effect of B_j.

To sum things up, while perfect collinearity is rare in practice, less-than-perfect collinearity is common. In these cases the VIF will be less than 1, but can still have an adverse effect on the sampling variances; sometimes making them quite large.

Identifying Collinearity

In our case study 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.

Collinearity Diagnostics

We can empirically diagnose problematic collinearity in the data (Belsley, 1991; Belsley et al., 1980). Before we do, however, it is important that the **functional form of the model** has been correctly specified. Since, a model needs to be specified before we can estimate coefficients or their sampling variances, and collinearity produces unstable estimates of these estimates, collinearity should only be investigated *after* the model has been satisfactorily specified.

Below we will explore some of the diagnostic tools available to an applied researcher.

Regress each Predcitor on the Other Predictors

Since collinearity is defined as linear dependence within the set of predictors, one way to diagnose collinearity is to regress each of the predictors on the remaining predictors and evaluate the R^2 value. If all the R^2 values are close to zero there is no collinearity problems. If one or more of the R^2 values are close to 1, there is a collinearity problem.

```r
# Use faculty as outcome; obtain R2
summary(lm(faculty ~ 1 + peer + school, data = eeo))$r.squared
```

```
[1] 0.9733906
```

```
# Use faculty as outcome; obtain R2
summary(lm(peer ~ 1 + faculty + school, data = eeo))$r.squared
```

```
[1] 0.9669002
```

```
# Use faculty as outcome; obtain R2
summary(lm(school ~ 1 + faculty + peer, data = eeo))$r.squared
```

```
[1] 0.9879743
```

All three R^2 values are quite high, which is indicative of collinearity.

One shortcoming with this method of diagnosing collinearity is that when the predictor space is large, you would need to look at the R^2 values from several models. And, while this could be automated in an R function, there are other methods that allow us to diagnose collinearity.

High Correlations among Predictors

Collinearity can sometimes be anticipated by examining the pairwise correlations between the predictors.

```
eeo %>%
  select(faculty, peer, school) %>%
  correlate()
```

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

In this example, all three of the predictors are highly correlated with one another. This is likely a good indicator that their may be problems in the estimation of coefficients, inflated standard errors, or both; especially given that the correlations are all very high. 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).

Beyond looking at the correlation matrix of the predictors, we will examine three common methods statisticians use to empirically detect collinearity: (1) computing variance inflation factors for the coefficients; (2) examining the eigenvalues of the correlation matrix; and (3) examining the condition indices of the correlation matrix.

Variance Inflation Factor

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

$$\mathrm{VIF} = \frac{1}{1 - R^2_j}$$

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. 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.

Remember, the VIF can range from 1 (independence among the predictors) to infinity (perfect collinearity). There is not consensus among statisticians about how high the VIF has to be to constitute a problem. Some references cite $\mathrm{VIF} > 10$ as problematic (which increases the size of the CI for the coefficient by a factor of over three); while others cite $\mathrm{VIF} > 4$ as problematic (which increases the size of the CI for the coefficient by a factor of two). As you consider what VIF value to use as an indicator of problematic inflation, it is more important to consider what introducing that much uncertainty would mean in your substantive problem. For example, would you be comfortable with tripling the uncertainty associated with the coefficient? What about doubling it? Once you make that decision, you can determine your VIF cutoff.

There are several situations in which high VIF values are expected and not problematic:

- **The variables with high VIFs are control variables, and the variables of interest do not have high VIFs.** Since we would not be interested in inference around the control variables, high VIF values on those variables would not
- **The high VIFs are caused by the inclusion of powers or products of other variables.** The *p*-

- **The high VIFs are caused by the inclusion of powers or products of other variables.** The *p* value for a product term is not affected by the multicollinearity. Centering predictors prior to creating the powers or the products will reduce the correlations, but the *p*-value the products will be exactly the same whether or not you center. Moreover the results for the other effects will be the same in either case indicating that multicollinearity has no adverse consequences.
- **The variables with high VIFs are indicator (dummy) variables that represent a categorical variable with three or more categories.** This is especially true when the reference category used has a small proportion of cases. In this case, *p*-values for the indicator variables may be high, but the overall test that all indicators have coefficients of zero is unaffected by the high VIFs. And nothing else in the regression is affected. To avoid the high VIF values in this situaton, just choose a reference category with a larger proportion of cases.

Eigenvalues of the Correlation Matrix

Recall that each square ($k \times k$) 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. If any of these eigenvalues is exactly equal to zero, it indicates a linear dependence among the predictors. In practice, perfect collinearity is rare, but near perfect collinearity can exist and is indicated when at least one of the eigenvalues is near zero, and is quite a bit smaller than the others.

As a diagnostic, rather than looking at the size of all the eigenvalues, we compute the sum of the reciprocals of the eigenvalues:

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

- If the predictors are orthogonal to one another (independent) then $\lambda_i = 1$ and the sum of the reciprocal values will be equal to the number of predictors, $\sum_{i=1}^{k} \frac{1}{\lambda_i} = k$.
- If the predictors are collinear with one another (dependent) then $\lambda_i = 0$ and the sum of the reciprocal values will be equal to infinity, $\sum_{i=1}^{k} \frac{1}{\lambda_i} = \infty$.
- When there is nonperfect collinearity then $0 < \lambda_i < 1$, and the sum of the reciprocal values will be greater than the number of predictors, $\sum_{i=1}^{k} \frac{1}{\lambda_i} > k$.

In an orthogonal matrix, the eigenvalues are all ± 1, but since the correlation matrix is positive semidefinite, the eigenvalues are all $+1$.

Larger sums of the reciprocal values of the eigenvalues is indicative of higher degrees of collinearity. In practice, we might use some cutoff to indicate when the collinearity is problematic. One such cutoff used is, if the sum is greater than five times the number of predictors, it is a sign of collinearity.

$$\text{IF} \quad \sum_{i=1}^{k} \frac{1}{\lambda_i} > 5k \quad \text{THEN} \quad \text{collnearity is a problem}$$

USING R TO COMPUTE THE EIGENVALUES OF THE CORRELATION MATRIX

We can use the `eigen()` function to compute the eigenvalues of a square matrix. We provide this function the correlation matrix for the model's predictors. To date, I have been using the `correlate()` function to produce correlation matrices. This function produces a formatted output that is nice for displaying the correlation matrix, but, because of its formatting, is not truly a matrix object. Instead, we will use the `cor()` function, which produces a matrix object, to produce the correlation matrix.

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

```
          faculty      peer    school
faculty 1.0000000 0.9600806 0.9856837
peer    0.9600806 1.0000000 0.9821601
school  0.9856837 0.9821601 1.0000000
```

Once we have the correlation matrix, we can use the `eigen()` function to compute the eigenvalues (and eigenvectors) of the inputted correlation matrix.

```r
# Compute eigenvalues and eigenvectors
eigen(r_xx)
```

```
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(r_xx)$values)
```

```
[1] 150.9477
```

We compare the sum of the reciprocal of the eigenvalues to five times the number of predictors; $5 \times 3 = 15$. Since this sum is greater than 15, we would conclude that there is a collinearity problem for this model.

Condition Indices

Another diagnostic measure of collinearity that relies on the eigenvalues of the correlation matrix are the *condition indices*. 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
# Sort eigenvalues from largest to smallest
lambda = sort(eigen(r_xx)$values, decreasing = TRUE)

# View eigenvalues
lambda
```

```
[1] 2.951993158 0.040047507 0.007959335
```

```r
# Compute condition indices
sqrt(max(lambda) / lambda)
```

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

The condition number of the correlation matrix, $\kappa = 19.26$, suggests strong collinearity among the predictors.

Fixing Collinearity in Practice

Although there are several solutions to “fix” collinearity in practice, none are a magic bullet. Here are three potential fixes:

- Re-specify the model
  - Drop one (or more) of the collinear predictors—This changes what you are controlling for;
  - Combine collinear predictors;
- 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. In upcoming notes, we will look at methods for combining collinear predictors and performing biased estimation.

References

References

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

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

Chatterjee, S., & Hadi, A. S. (2012). *Regression analysis by example*. 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*. U.S. Government Printing Office.

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