

Multicollinearity Diagnostics

- Variance Inflation Factor (VIF)

Grady Keene and Yuka Chen

What is Multicollinearity?

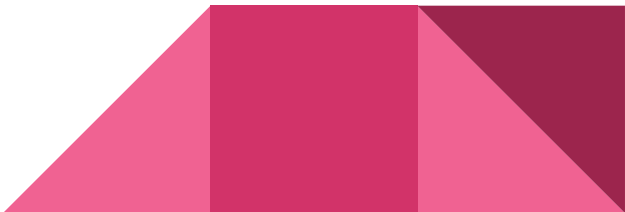
Multicollinearity is the condition by which the set of explanatory variables contains some redundancies, causes inflation of standard errors of estimated regression coefficients and makes it difficult to evaluate partial effects (Agresti, 2018).

- multicollinearity exists whenever two or more of the predictors in a regression model are moderately or highly correlated



Review...

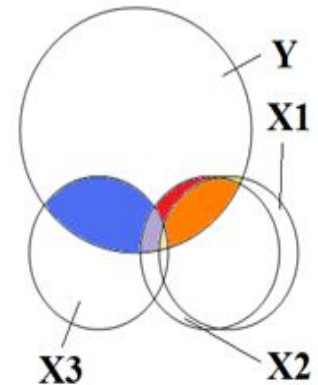
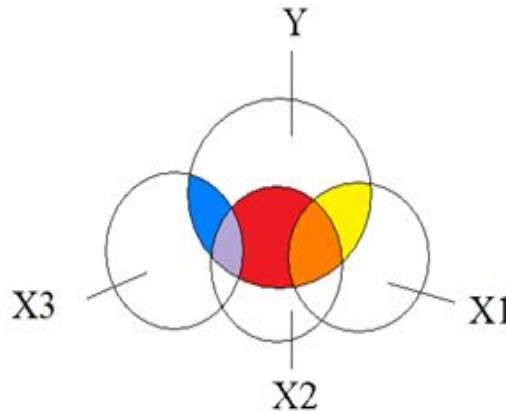
Common issues arise when the predictor variables beings considered for the regression model are highly correlated:

1. Adding or deleting a predictor variable changes the regression coefficients.
 2. The extra sum of squares associated with a predictor variable varies, depending upon which other predictor variables are already included in the model.
 3. The estimated standard deviation of the regression coefficients become large when the predictor variables in the regression model are highly correlated with each other.
 4. The estimated regression coefficient individually may not be statistically significant even though a definite statistical relation exists between the response variable and the set of predictor variables.
- 

Observational Indicators

Informal diagnostics may indicate the presence of multicollinearity through noticing:

1. Large changes in the estimated regression coefficients when a predictor variable is added or deleted, or when an observation is altered or deleted.
2. Nonsignificant results in individual tests on the regression coefficients for important predictor variables.
3. Estimated regression coefficients with an algebraic sign that is the opposite of that expected from theoretical considerations or prior experience.



Observational Indicators (cont.)

4. Large coefficients of simple correlation between pairs of predictor variables in the correlation matrix r_{xx} .
5. Wide confidence intervals for the regression coefficients representing important predictor variables.

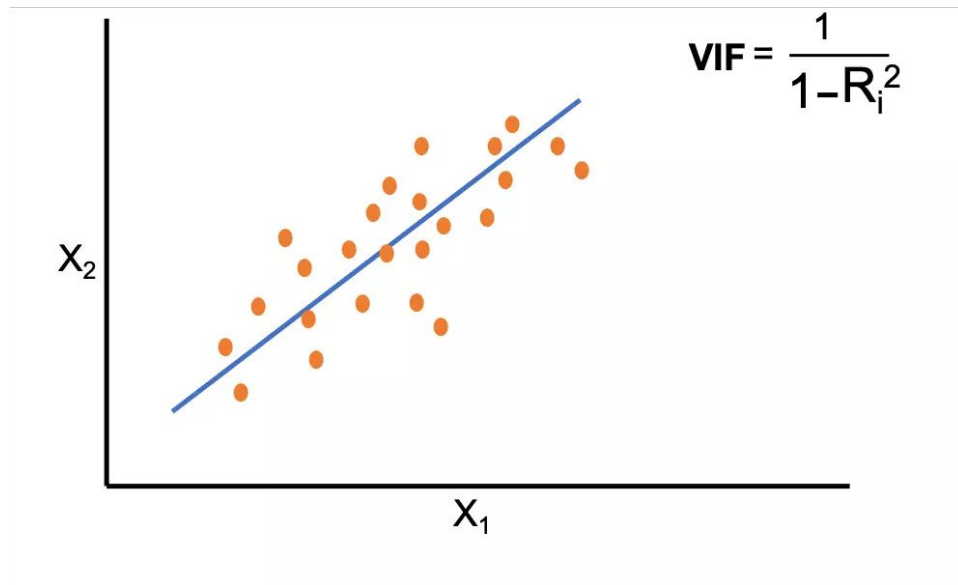
Note:

These informal diagnostics do not provide quantitative measurements of the impact that multicollinearity may have on the model.

Ex. The predictor variables may have low pairwise correlation among each other, but these correlations may not express grouped linear combinations of predictor variables effect on another singular variable, such as: $X_3 = (X_1 + X_2)/2$.

What is a Variation Inflation Factor(VIF)?

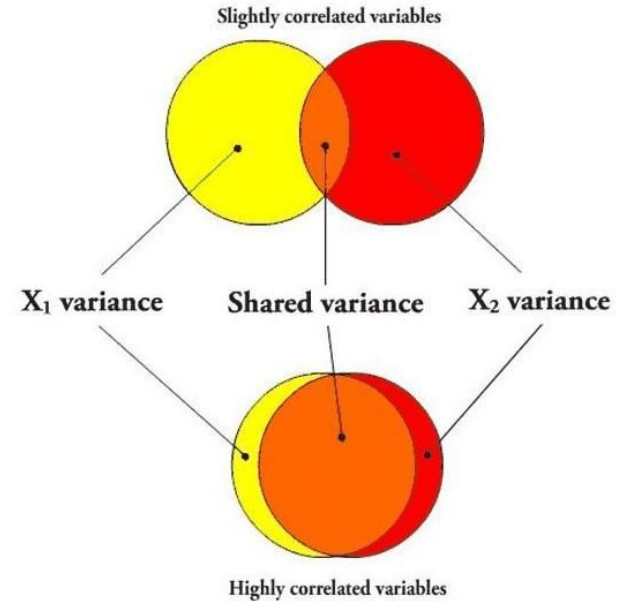
A **formal** method to approaching the detection of multicollinearity makes use of variation inflation factors. The VIF quantifies how much variance of the estimated regression coefficients are inflated as compared to when these predictor variables are not linearly related.



How Multicollinearity Alters the Model

If multicollinearity is present in the model, or if two or more predictor variables contain “overlapping” information, the standard error of some or all of the regression coefficients will be inflated.

Using a multivariate regression model to predict an outcome when two or more variables are correlated will cause problems where a change in one variable may shift the value of another.



Classical Multiple Linear Regression:

We have a sample of n observations on the response variable Y and the $p-1$ predictor variables $X_1, X_2, X_3 \dots X_{p-1}$

- Model: $Y_i = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \dots + \beta_{p-1} X_{i,p-1} + \varepsilon_i, \quad i = 1, 2, 3, \dots, n$

We can use `lm()` function in R $lm(Y \sim X_1 + X_2 + X_3 \dots X_{p-1})$

When we try to see how well our predictor variable is explained by the explanatory variables: R^2

- In general, a larger value of R^2 is considered to indicate a “better fit” of the regression model. The more variability the regression explains, the “better” the model is.
- Low values R^2 of indicate that the model doesn't do a very well explaining the variability in the response variable.
- R^2 explains how much of the variation in the Y is explained by X in this model.

What VIF Means

VIFs are calculated by taking a predictor variable and regressing it against every other predictor in present in the model, whereas:

$(VIF)_k = 1$ when $R_k^2 = 0$, where X_k is not linearly related to other X variables.

When $R_k^2 \neq 0$, then (VIF) is greater than one, which indicates an inflated variance for b_k^* as a result of intercorrelations between X variables.

$$VIF_k = \frac{1}{1 - R_k^2}$$

k being the predictor variable you are looking at



Importance of R^2_k and Where it Comes From

Using a predictor variable X_k , run an OLS regression that has X_k as a function for all other explanatory variables in the equation


ie. X_1 :

$$X_1 = \alpha_1 + \alpha_2 X_2 + \alpha_3 X_3 + \dots + \alpha_k X_k + v$$

Calculating the VIF of $\hat{\beta}_k$:

$$\text{VIF}(\hat{\beta}_k) = \frac{1}{(1 - R_k^2)}$$

Where R_k^2 is the unadjusted R^2 from the **above** created OLS regression



Importance of R^2_k and Where it Comes From cont.

ie. Running a given OLS regression model...

$$X_1 = \alpha_1 + \alpha_2 X_2 + \alpha_3 X_3 + \dots + \alpha_k X_k + v$$

We find an unadjusted R^2 value of 0.75, therefore:

$$\text{VIF}(\hat{\beta}_{k=l}) = \frac{1}{(1 - 0.75)} = 4$$

As a predictor variable's variation can be explained by other predictor variables in the model more and more, their R^2_k will continue to increase. As the R^2_k grows larger, as does the value of their variation inflation factor.



Interpretation of VIF Value

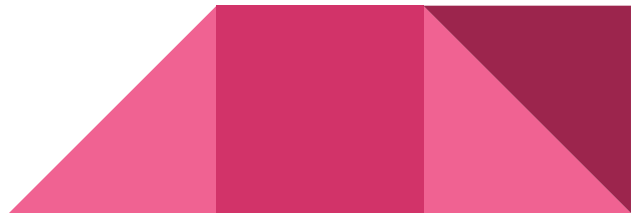
Ex1. The variable X_1 in our model has a $VIF = 4$, this value can be interpreted in 2 ways:

1. The variance of β_1 is 4 times greater than it would have been if X_1 had been entirely non-related to other variables in our model
2. The variance of β_1 is 300% greater than it would be if there were no collinearity effect at all between X_1 and other variables in our model

This percentage is calculated by subtracting 1 (value of VIF if there were no collinearity) from the actual value of VIF:

$$4 - 1 = 3 \longrightarrow 3 * 100 / 100 = 300\%$$

Ex2. A VIF value = 1 for any X predictor variable indicates total absence of collinearity between the given X and other predictors in the model.



Use of Standardized Regression Model

Often, staticians use the standardized regression model to create exact computation for different predictor variables for VIF to detect multicollinearity, which is obtained by transforming the variables by means of the correlation transformation.

The correlation transformation helps with controlling roundoff errors and by expressing the regression coefficients in the same units, done by ratios of standard deviations.

Since `lm()` in R would do it for us, we don't need to really go through the same process, instead we can understand it with OLS model.

The correlation transformation is a simple function of the standardized variables

$$Y_i^* = \frac{1}{\sqrt{n-1}} \left(\frac{Y_i - \bar{Y}}{s_Y} \right)$$

$$X_{ik}^* = \frac{1}{\sqrt{n-1}} \left(\frac{X_{ik} - \bar{X}_k}{s_k} \right) \quad (k = 1, \dots, p-1)$$

$$Y_i^* = \beta_1^* X_{i1}^* + \dots + \beta_{p-1}^* X_{i,p-1}^* + \varepsilon_i^*$$

Given standardized regression model



Diagnostic Uses of the VIF

The largest VIF value reported is used as an indicator of the influence of multicollinearity in the model. In general, a VIF value that exceeds 10 indicates high correlation and is regarded as a cause for concern and that multicollinearity may be influencing the least squares estimate.

Using the mean of the VIF

The mean of the VIF values gives information on the severity of present multicollinearity in terms of the measured distance of the estimated standardized regression coefficients to the true values of β_k^* .

$$(\overline{VIF}) = \frac{\sum_{k=1}^{p-1} (VIF)_k}{p - 1}$$

Provides useful information on the effect of multicollinearity on the sum of squared errors

Calculating VIF in R Manually

Transferring each explanatory variable as predictor variables to find the r-squared value associated with X

Using the lm() function...

```
lm(Y ~ X1 + X2 + X3 + X4 + X5 , data)
```

And we want to know the X₁ VIF

```
X1 VIF = 1 / 1 - glance(lm(X1 ~ 1 + X2 + X3 + X4 + X5)$r.squared)
```

```
X2 VIF = 1 / 1 - glance(lm(X2 ~ X1 + 1 + X3 + X4 + X5)$r.squared)
```

⋮

```
Xn VIF = 1 / 1 - glance(lm(Xn ~ X1 + X2 + X3 + X4 + X5....+1)$r.squared)
```

Demonstrating Multicollinearity Detection in R

```
library(tidyverse)
library(broom)
library(GGally)
library(fastDummies)
library(car)
```

The variables for this data set are violent crime rate (number of violent crimes per 100,000 population), murder rate, percent in metropolitan areas, percent white, percent high school graduates, percent below the poverty level, and percent of families headed by a single parent. The data are from Statistical Abstract of the United States for 2005.

```
Crime <- read_table("https://users.stat.ufl.edu/~aa/smss/data/Crime2.dat",
  col_types = cols(X9 = col_skip()))
```

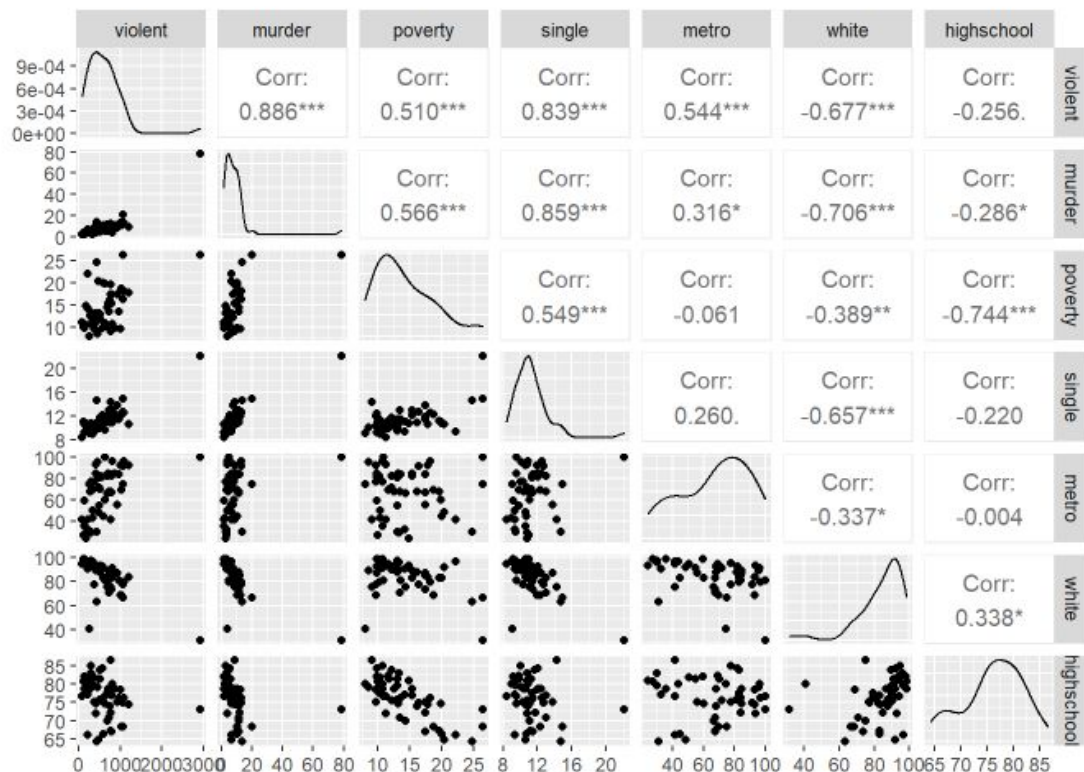
```
## Warning: Missing column names filled in: 'X9' [9]
```

```
glimpse(Crime)
```

```
## Rows: 51
## Columns: 8
## $ State      <chr> "AK", "AL", "AR", "AZ", "CA", "CO", "CT", "DE", "FL", "GA", ~
## $ violent    <dbl> 761, 780, 593, 715, 1078, 567, 456, 686, 1206, 723, 261, 32~
## $ murder     <dbl> 9.0, 11.6, 10.2, 8.6, 13.1, 5.8, 6.3, 5.0, 8.9, 11.4, 3.8, ~
## $ poverty    <dbl> 9.1, 17.4, 20.0, 15.4, 18.2, 9.9, 8.5, 10.2, 17.8, 13.5, 8.~
## $ single     <dbl> 14.3, 11.5, 10.7, 12.1, 12.5, 12.1, 10.1, 11.4, 10.6, 13.0, ~
## $ metro      <dbl> 41.8, 67.4, 44.7, 84.7, 96.7, 81.8, 95.7, 82.7, 93.0, 67.7, ~
## $ white      <dbl> 75.2, 73.5, 82.9, 88.6, 79.3, 92.5, 89.0, 79.4, 83.5, 70.8, ~
## $ highschool <dbl> 86.6, 66.9, 66.3, 78.7, 76.2, 84.4, 79.2, 77.5, 74.4, 70.9, ~
```


Pairwise Correlation Matrix

```
ggpairs(Crime[2:8])
```



Creating the Model With All Variables

```
model2 <- lm(violent ~ murder + poverty + single + metro + white + highschool, Crime)
glance(model2)
```

r.squared	adj.r.squared	sigma	statistic	p.value	df	logLik	AIC	BIC	deviance
<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
0.8950076	0.8806905	152.3614	62.51302	6.523194e-20	6	-324.9402	665.8803	681.3349	1021415

1 row | 1-10 of 12 columns

```
tidy(model2)
```

term	estimate	std.error	statistic	p.value
<chr>	<dbl>	<dbl>	<dbl>	<dbl>
(Intercept)	-1143.7458232	585.070199	-1.9548865	5.696789e-02
murder	19.3338669	4.443685	4.3508636	7.941724e-05
poverty	15.0033959	9.721858	1.5432642	1.299291e-01
single	54.8546234	21.306597	2.5745370	1.347893e-02
metro	6.6218687	1.118556	5.9200160	4.423125e-07
white	-0.6956544	2.506081	-0.2775866	7.826299e-01
highschool	4.7876066	6.677033	0.7170260	4.771473e-01

7 rows

$$\widehat{violent} = -1143.8 + 19.33 \cdot murder + 15 \cdot poverty + 54.85 \cdot single + 6.62 \cdot metro - 0.70 \cdot white + 4.79 \cdot highschool$$

Finding Individual VIF Values

Calculating murder VIF

```
murder <- lm(murder ~ 1 + poverty + single + metro + white + highschool, data = Crime)
murder_g <- glance(murder)
murder_g
```

r.squared <dbl>	adj.r.squared <dbl>	sigma <dbl>	statistic <dbl>	p.value <dbl>	df <dbl>	logLik <dbl>	AIC <dbl>	BIC <dbl>	deviance <dbl>
0.7953084	0.7725648	5.111229	34.96858	1.983664e-14	5	-152.3776	318.7553	332.2781	1175.61

1 row | 1-10 of 12 columns

```
murdevif <- 1/(1 - murder_g[[1]])
murdevif
```

```
## [1] 4.885397
```

$$VIF_k = \frac{1}{1 - R_k^2}$$

We can use `vif()` function from package `{car}` to see all variables' VIF

```
vif(model2)
```

```
##      murder      poverty      single      metro      white highschool
##  4.885397  4.278128  4.400805  1.299233  2.375882  3.002861
```

Creating a Correlation Matrix

```
x_vari2 <- Crime[ , c("murder", "poverty", "single", "metro", "white", "highschool")]
cor(x_vari2)
```

```
##           murder    poverty    single    metro    white
## murder      1.0000000  0.5658711  0.8589106  0.316114166 -0.7062589
## poverty     0.5658711  1.0000000  0.5485890 -0.060538499 -0.3891346
## single      0.8589106  0.5485890  1.0000000  0.259810085 -0.6567078
## metro       0.3161142 -0.0605385  0.2598101  1.000000000 -0.3374351
## white       -0.7062589 -0.3891346 -0.6567078 -0.337435120  1.0000000
## highschool -0.2860708 -0.7439382 -0.2197829 -0.003977358  0.3381212
##           highschool
## murder      -0.286070828
## poverty     -0.743938249
## single      -0.219782892
## metro       -0.003977358
## white       0.338121236
## highschool  1.000000000
```

Variable *murder* is highly correlated with *single* and *white*, expressing its larger R^2_k value.

Redesigning Model without *murder*

```
model2_1 <- lm(violent ~ poverty + single + metro + white + highschool, Crime)
glance(model2_1)
```

r.squared <dbl>	adj.r.squared <dbl>	sigma <dbl>	statistic <dbl>	p.value <dbl>	df <dbl>	logLik <dbl>	AIC <dbl>	BIC <dbl>	deviance <dbl>
0.849837	0.8331523	180.1762	50.93489	2.048114e-17	5	-334.0649	682.1298	695.6526	1460857

1 row | 1-10 of 12 columns

```
tidy(model2_1)
```

term <chr>	estimate <dbl>	std.error <dbl>	statistic <dbl>	p.value <dbl>
(Intercept)	-1795.904479	668.788462	-2.685310	1.011057e-02
poverty	26.244163	11.083273	2.367907	2.224639e-02
single	109.466605	20.359892	5.376581	2.601006e-06
metro	7.608807	1.295273	5.874290	4.796832e-07
white	-4.482908	2.779073	-1.613094	1.137159e-01
highschool	8.646443	7.826015	1.104833	2.751042e-01

6 rows

Change in Regression Estimates

With variable “murder”

$$\widehat{violent} = -1143.8 + 19.33 \cdot murder + 15 \cdot poverty + 54.85 \cdot single + 6.62 \cdot metro - 0.70 \cdot white + 4.79 \cdot highschool$$

Without Variable “murder”

$$\widehat{violent} = -1795.9 + 26.2 \cdot poverty + 109.5 \cdot single + 7.6 \cdot metro - 4.48 \cdot white + 8.65 \cdot highschool$$