PennState Eberly College of Science

STAT 462

Applied Regression Analysis

10.7 - Detecting Multicollinearity Using Variance Inflation Factors

Okay, now that we know the effects that multicollinearity can have on our regression analyses and subsequent conclusions, how do we tell when it exists? That is, how can we tell if multicollinearity is present in our data?

Some of the common methods used for detecting multicollinearity include:

- The analysis exhibits the signs of multicollinearity — such as, estimates of the coefficients vary excessively from model to model.
- The t-tests for each of the individual slopes are non-significant ($P > 0.05$), but the overall F-test for testing all of the slopes are simultaneously 0 is significant ($P < 0.05$).
- The correlations among pairs of predictor variables are large.

Looking at correlations only among *pairs* of predictors, however, is limiting. It is possible that the pairwise correlations are small, and yet a linear dependence exists among three or even more variables, for example, if $X_3 = 2X_1 + 5X_2 + error$, say. That's why many regression analysts often rely on what are called **variance inflation factors** (VIF) to help detect multicollinearity.

What is a Variation Inflation Factor?

As the name suggests, a variance inflation factor (VIF) quantifies how much the variance is inflated. But what variance? Recall that we learned previously that the standard errors — and hence the variances — of the estimated coefficients are inflated when multicollinearity exists. A variance inflation factor exists for *each of the predictors* in a multiple regression model. For example, the variance inflation factor for the estimated regression coefficient b_j —denoted VIF_j —is just the factor by which the variance of b_j is "inflated" by the existence of correlation among the predictor variables in the model.

In particular, the variance inflation factor for the j^{th} predictor is:

$$VIF_j = \frac{1}{1 - R_j^2}$$

where R_j^2 is the R^2-value obtained by regressing the j^{th} predictor on the remaining predictors.

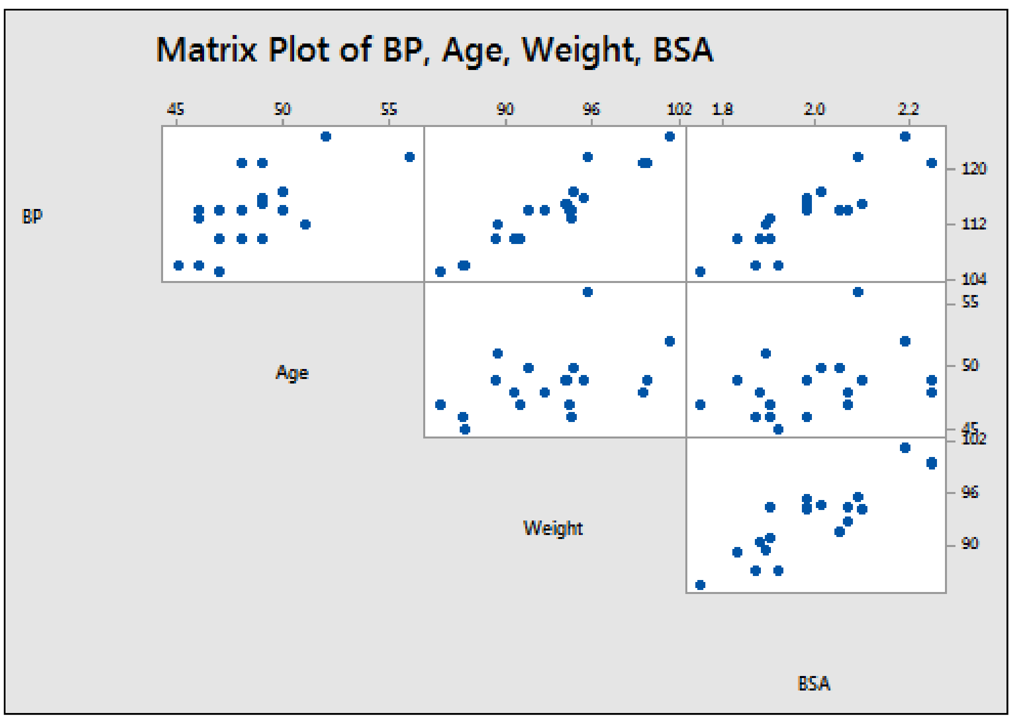
How do we interpret the variance inflation factors for a regression model? A VIF of 1 means that there is no correlation among the j^{th} predictor and the remaining predictor variables, and hence the variance of b_j is not inflated at all. The general rule of thumb is that VIFs exceeding 4 warrant further investigation, while VIFs exceeding 10 are signs of serious multicollinearity requiring correction.

An Example

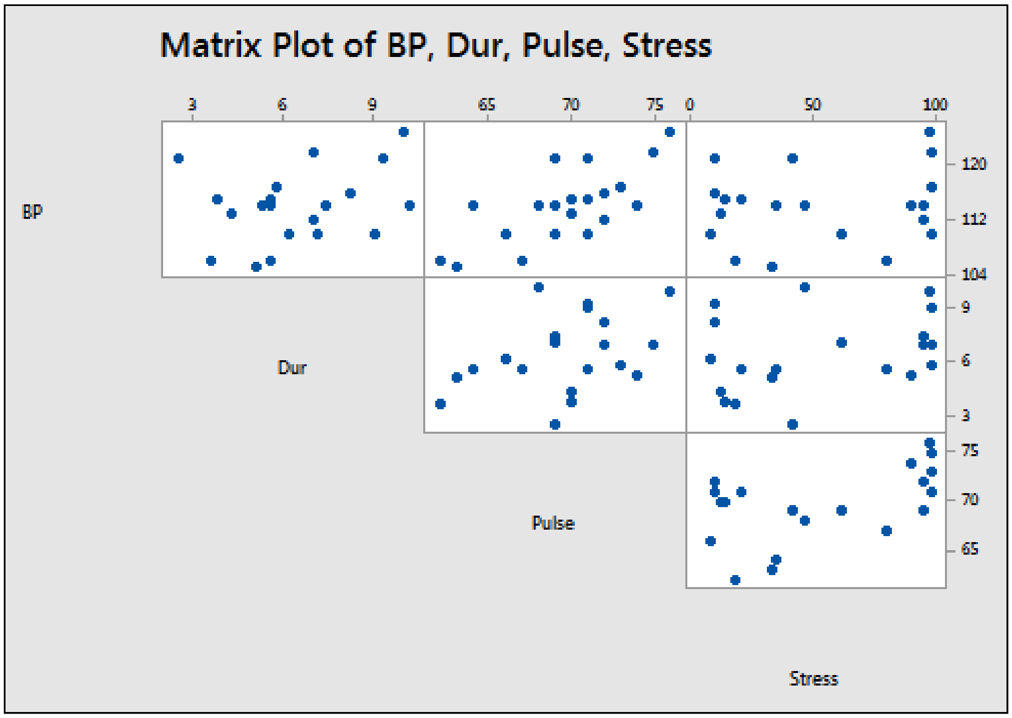
Let's return to the blood pressure data (bloodpress.txt (../../sites/onlinecourses.science.psu.edu.stat462/files/data/bloodpress/index.txt)) in which researchers observed the following data on 20 individuals with high blood pressure:

- blood pressure ($y = BP$, in mm Hg)
- age ($x_1 = Age$, in years)
- weight ($x_2 = Weight$, in kg)
- body surface area ($x_3 = BSA$, in sq m)
- duration of hypertension ($x_4 = Dur$, in years)
- basal pulse ($x_5 = Pulse$, in beats per minute)
- stress index ($x_6 = Stress$)

As you may recall, the matrix plot of *BP*, *Age*, *Weight*, and *BSA*:



the matrix plot of *BP*, *Dur*, *Pulse*, and *Stress*:



and the correlation matrix:

Correlation: BP, Age, Weight, BSA, Dur, Pulse, Stress

	BP	Age	Weight	BSA	Dur	Pulse
Age	0.659					
Weight	0.950	0.407				
BSA	0.866	0.378	0.875			
Dur	0.293	0.344	0.201	0.131		
Pulse	0.721	0.619	0.659	0.465	0.402	
Stress	0.164	0.368	0.034	0.018	0.312	0.506

suggest that some of the predictors are at least moderately marginally correlated. For example, body surface area (BSA) and weight are strongly correlated ($r = 0.875$), and weight and pulse are fairly strongly correlated ($r = 0.659$). On the other hand, none of the pairwise correlations among age, weight, duration and stress are particularly strong ($r < 0.40$ in each case).

Regressing y = BP on all six of the predictors, we obtain:

```
Analysis of Variance

Source      DF   Seq SS   Seq MS  F-Value  P-Value
Regression   6  557.844   92.974   560.64    0.000
  Age        1  243.266  243.266  1466.91    0.000
  Weight     1  311.910  311.910  1880.84    0.000
  BSA        1    1.768    1.768    10.66    0.006
  Dur        1    0.335    0.335     2.02    0.179
  Pulse      1    0.123    0.123     0.74    0.405
  Stress     1    0.442    0.442     2.67    0.126
Error       13    2.156    0.166
Total       19  560.000


Model Summary

       S    R-sq  R-sq(adj)  R-sq(pred)
0.407229  99.62%     99.44%      99.08%


Coefficients

Term         Coef  SE Coef  T-Value  P-Value   VIF
Constant   -12.87     2.56    -5.03    0.000
Age        0.7033   0.0496    14.18    0.000  1.76
Weight     0.9699   0.0631    15.37    0.000  8.42
BSA          3.78     1.58     2.39    0.033  5.33
Dur        0.0684   0.0484     1.41    0.182  1.24
Pulse     -0.0845   0.0516    -1.64    0.126  4.41
Stress    0.00557  0.00341     1.63    0.126  1.83
```

As you can see, three of the variance inflation factors —8.42, 5.33, and 4.41 —are fairly large. The VIF for the predictor *Weight*, for example, tells us that the variance of the estimated coefficient of *Weight* is inflated by a factor of 8.42 because *Weight* is highly correlated with at least one of the other predictors in the model.

For the sake of understanding, let's verify the calculation of the VIF for the predictor *Weight*. Regressing the predictor x_2 = *Weight* on the remaining five predictors:

Analysis of Variance

Source	DF	Seq SS	Seq MS	F-Value	P-Value
Regression	5	308.839	61.768	20.77	0.000
Age	1	58.156	58.156	19.55	0.001
BSA	1	212.734	212.734	71.53	0.000
Dur	1	1.442	1.442	0.48	0.498
Pulse	1	27.311	27.311	9.18	0.009
Stress	1	9.196	9.196	3.09	0.101
Error	14	41.639	2.974		
Total	19	350.478			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
1.72459	88.12%	83.88%	74.77%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	19.67	9.46	2.08	0.057	
Age	-0.145	0.206	-0.70	0.495	1.70
BSA	21.42	3.46	6.18	0.000	1.43
Dur	0.009	0.205	0.04	0.967	1.24
Pulse	0.558	0.160	3.49	0.004	2.36
Stress	-0.0230	0.0131	-1.76	0.101	1.50

R^2_{Weight} is 88.12% or, in decimal form, 0.8812. Therefore, the variance inflation factor for the estimated coefficient *Weight* is by definition:

$$VIF_{Weight} = \frac{Var(b_{Weight})}{Var(b_{Weight})_{min}} = \frac{1}{1 - R^2_{Weight}} = \frac{1}{1 - 0.8812} = 8.42.$$

Again, this variance inflation factor tells us that the variance of the weight coefficient is inflated by a factor of 8.42 because *Weight* is highly correlated with at least one of the other predictors in the model.

So, what to do? One solution to dealing with multicollinearity is to remove some of the violating predictors from the model. If we review the pairwise correlations again:

Correlation: BP, Age, Weight, BSA, Dur, Pulse, Stress

	BP	Age	Weight	BSA	Dur	Pulse
Age	0.659					
Weight	0.950	0.407				
BSA	0.866	0.378	0.875			
Dur	0.293	0.344	0.201	0.131		
Pulse	0.721	0.619	0.659	0.465	0.402	
Stress	0.164	0.368	0.034	0.018	0.312	0.506

we see that the predictors *Weight* and *BSA* are highly correlated ($r = 0.875$). We can choose to remove either predictor from the model. The decision of which one to remove is often a scientific or practical one. For example, if the researchers here are interested in using their final model to predict the blood pressure of future individuals, their choice should be clear. Which of the two measurements — body surface area or weight — do you think would be easier to obtain?! If indeed weight is an easier measurement to obtain than body surface area, then the researchers would be well-advised to remove *BSA* from the model and leave *Weight* in the model.

Reviewing again the above pairwise correlations, we see that the predictor *Pulse* also appears to exhibit fairly strong marginal correlations with several of the predictors, including *Age* (r = 0.619), *Weight* (r = 0.659) and *Stress* (r = 0.506). Therefore, the researchers could also consider removing the predictor *Pulse* from the model.

Let's see how the researchers would do. Regressing the response $y = BP$ on the four remaining predictors *Age*, *Weight*, *Duration*, and *Stress*, we obtain:

Analysis of Variance

Source	DF	Seq SS	Seq MS	F-Value	P-Value
Regression	4	555.455	138.864	458.28	0.000
Age	1	243.266	243.266	802.84	0.000
Weight	1	311.910	311.910	1029.38	0.000
Dur	1	0.178	0.178	0.59	0.455
Stress	1	0.100	0.100	0.33	0.573
Error	15	4.545	0.303		
Total	19	560.000			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.550462	99.19%	98.97%	98.59%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	-15.87	3.20	-4.97	0.000	
Age	0.6837	0.0612	11.17	0.000	1.47
Weight	1.0341	0.0327	31.65	0.000	1.23
Dur	0.0399	0.0645	0.62	0.545	1.20
Stress	0.00218	0.00379	0.58	0.573	1.24

Aha — the remaining variance inflation factors are quite satisfactory! That is, it appears as if hardly any variance inflation remains. Incidentally, in terms of the adjusted R^2-value, we did not seem to lose much by dropping the two predictors *BSA* and *Pulse* from our model. The adjusted R^2-value decreased to only 98.97% from the original adjusted R^2-value of 99.44%.

‹ 10.6 - Highly Correlated Predictors (../179/) up (../88/) 10.8 - Reducing Data-based Multicollinearity › (../181/)