Variance Inflation Factor

Variance inflation factor is a measure of the amount of multicollinearity in a set of multiple regression variables. A multiple regression is used when a person wants to test the effect of multiple variables on a particular outcome. The dependent variable is the outcome that is being acted upon by the independent variables, which are the inputs into the model. Multicollinearity exists when there is a linear relationship, or correlation, between one and more of the independent variables or inputs. Multicollinearity creates a problem in the multiple regression because since the inputs are all influencing each other, they are not actually independent and it is difficult to test how much the combination of the independent variables affects the dependent variable, or outcome, within the regression model.

To ensure the model is properly specified and functioning correctly, there are tests that can be run for multicollinearity. Variance inflation factor is one such measuring tool. Using variance inflation factors helps to identify the severity of any multicollinearity issues so that the model can be adjusted. Variance inflation factor measures how much the behaviour (variance) of an independent variable is influenced, or inflated, by its interaction/correlation with the other independent variables.

A variance inflation factor (VIF) detects multicollinearity in regression analysis. Multicollinearity is when there’s correlation between predictors (i.e. independent variables) in a model; it’s presence can adversely affect your regression results. The VIF estimates how much the variance of a regression coefficient is inflated due to multicollinearity in the model.

VIFs are calculated by taking a predictor, and regressing it against every other predictor in the model. This gives you the R-squared values, which can then be plugged into the VIF formula. “i” is the predictor you’re looking at (e.g. x_1 or x_2):

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

Variance inflation factors (*VIF*) to help detect multicollinearity. 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.

Variance inflation factor (*VIF*) quantifies how much the variance is inflated. Variances — of the estimated coefficients are inflated when multicollinearity exists. So, the variance inflation factor for the estimated coefficient b_k —denoted VIF_k —is just the factor by which the variance is inflated.

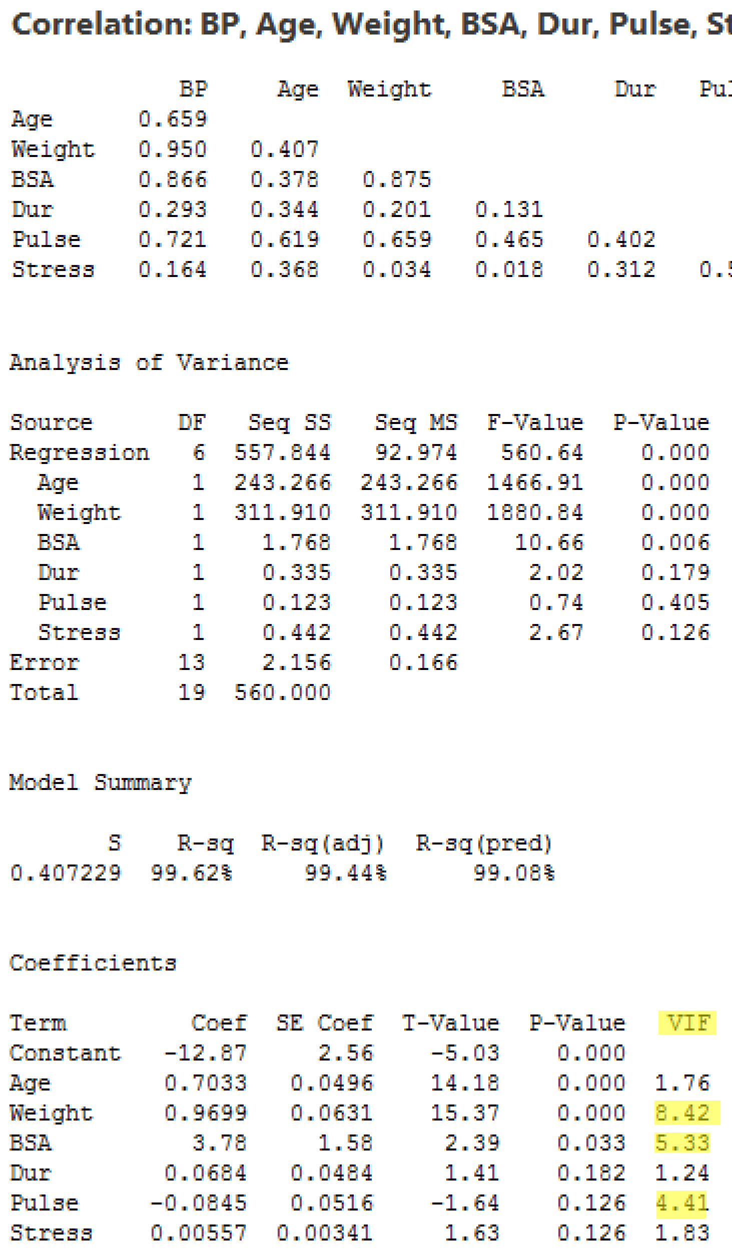
Using VIF we can see how much the variance of b_k is inflated when we add correlated predictors to our regression model.

$$VIFk=1/1-R2k$$

where $R2k$ is the R^2-value obtained by regressing the k^{th} predictor on the remaining predictors. Of course, the greater the linear dependence among the predictor x_k and the other predictors, the larger the $R2k$ value. And, as the above formula suggests, the larger the $R2k$ value, the larger the variance of b_k.

where *R2k* is the R^2-value obtained by regressing the k^{th} predictor on the remaining predictors. Note that a variance inflation factor exists for *each of the k predictors* in a multiple regression model.

How do we interpret the variance inflation factors for a regression model? Again, it is a measure of how much the variance of the estimated regression coefficient b_k is "inflated" by the existence of correlation among the predictor variables in the model. A VIF of 1 means that there is no correlation among the k^{th} predictor and the remaining predictor variables, and hence the variance of b_k 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.

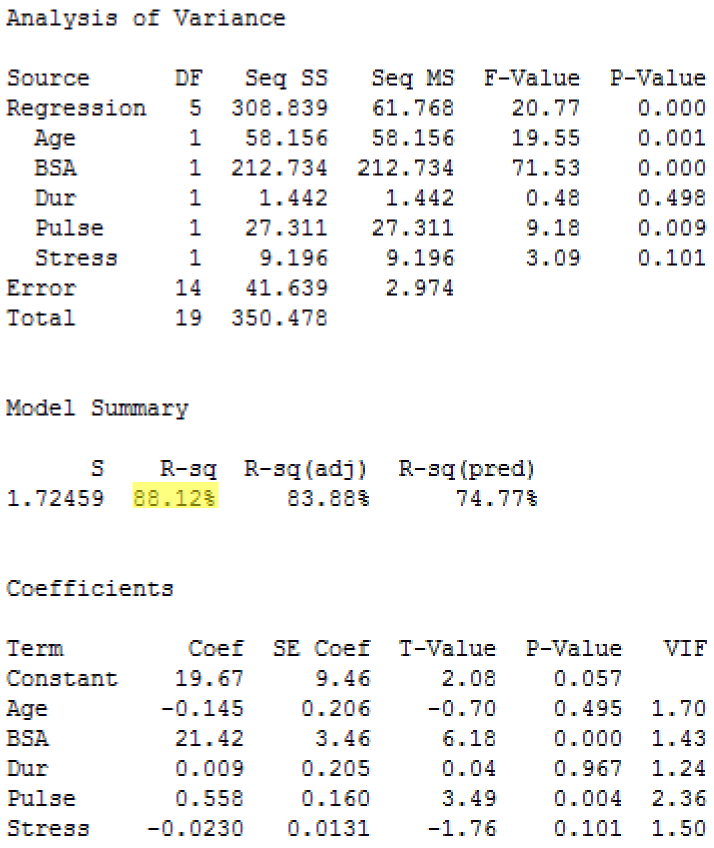


Correlation: BP, Age, Weight, BSA, Dur, Pulse, Stress

```
              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


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:



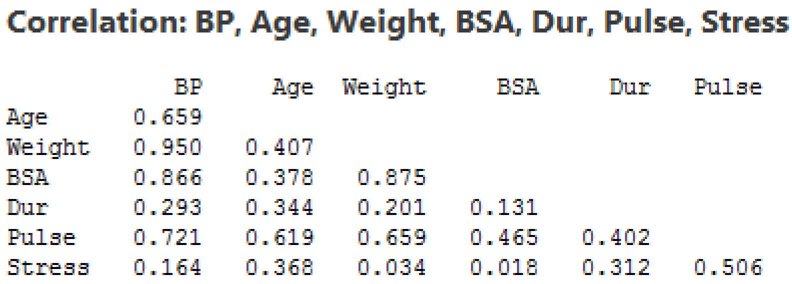
```
Analysis of Variance

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
```

R2Weight is 88.1% or, in decimal form, 0.881. Therefore, the variance inflation factor for the estimated coefficient *Weight* is by definition:

VIFWeight=*Var*(*bWeight*)/*Var*(*bWeight*)*min*=1/1−*R2Weight*=1/1−0.881=8.4

Again, this variance inflation factor tells us that the variance of the weight coefficient is inflated by a factor of 8.4 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:



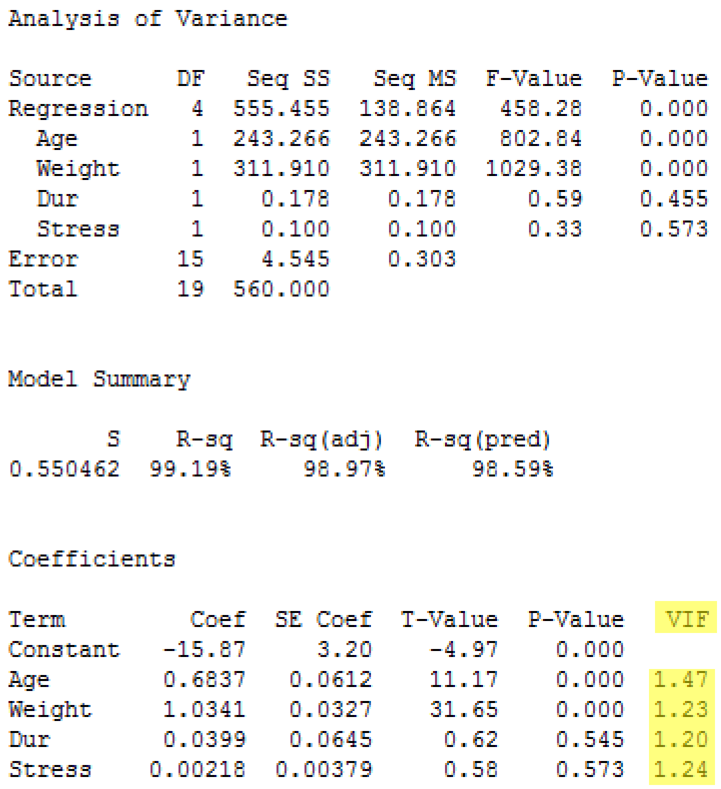
Correlation: BP, Age, Weight, BSA, Dur, Pulse, Stress

```
            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.

The ***variance inflation factor*** *(VIF)* quantifies the extent of correlation between one predictor and the other predictors in a model.

A VIF can be computed for each predictor in a predictive model. A value of 1 means that the predictor is not correlated with other variables. The higher the value, the greater the correlation of the variable with other variables. Values of more than 4 or 5 are sometimes regarded as being moderate to high, with values of 10 or more being regarded as very high. These numbers are just rules of thumb; in some contexts a VIF of 2 could be a great problem (e.g., if estimating *price elasticity*), whereas in straightforward predictive applications very high VIFs may be unproblematic.

If one variable has a high VIF it means that other variables must also have high VIFs. In the simplest case, two variables will be highly correlated, and each will have the same high VIF.

It's called the variance inflation factor because it estimates how much the variance of a coefficient is "inflated" because of linear dependence with other predictors. Thus, a VIF of 1.8 tells us that the variance (the square of the standard error) of a particular coefficient is 80% larger than it would be if that predictor was completely uncorrelated with all the other predictors.

If multicollinearity is found in the data, centering the data (that is deducting the mean of the variable from each score) might help to solve the problem. However, the simplest way to address the problem is to remove independent variables with high VIF values

Difference between Correlation Matrix and VIF:

The correlation matrix is not a reliable measurement for multicollinearity because it only considers the pairwise effects. Unfortunately, multicollinearity is defined as:

Phenomenon in which **two or more** predictor variables in a multiple regression model are highly correlated. You'll need to consider the correlation with all other variables in your data set, not just 1-to-1 pairwise comparison. VIF addresses the issue.

Assumptions of Regression:

- The Regression model is linear in its parameters (which are Coefficients and the error term).

  Linearity: The change in the response variable due to one-unit change in the predictor variable (Xk) is always constant irrespective of the current value of Xk. There should be a linear and additive relationship between dependent (response) variable and independent (predictor) variable(s). A linear relationship suggests that a change in response Y due to one-unit change in X^1 is constant, regardless of the value of X

- **Homoscedasticity** describes a situation in which the error term (that is, the "noise" or random disturbance in the relationship between the features and the target) is the same across all values of the independent variables. A scatter plot of residual values vs predicted values is a good way to check for homoscedasticity. There should be no clear pattern in the distribution and if there is a specific pattern, the data is heteroscedastic.
- There should be no correlation between the residual (error) terms. Absence of this phenomenon is known as Autocorrelation. **Autocorrelation:** The presence of correlation in error terms drastically reduces model's accuracy. This usually occurs in time series models where the next instant is dependent on previous instant. If the error terms are correlated, the estimated standard errors tend to underestimate the true standard error. If this happens, it causes confidence intervals and prediction intervals to be narrower.
- The independent variables should not be correlated. Absence of this phenomenon is known as multicollinearity.
- The error terms must have constant variance. This phenomenon is known as homoskedasticity. The presence of non-constant variance is referred to heteroskedasticity.
- The error terms must be normally distributed.

ROC Curve

Allows to create **ROC curve** and a complete sensitivity/specificity report. The ROC curve is a fundamental tool for diagnostic test evaluation.

In a ROC curve the true positive rate (Sensitivity) is plotted in function of the false positive rate (100-Specificity) for different cut-off points of a parameter. Each point on the ROC curve represents a sensitivity/specificity pair corresponding to a particular decision threshold. The area under the ROC curve (AUC) is a measure of how well a parameter can distinguish between two diagnostic groups (diseased/normal).

When you select a higher criterion value, the false positive fraction will decrease with increased specificity but on the other hand the true positive fraction and sensitivity will decrease.

When you select a lower threshold value, then the true positive fraction and sensitivity will increase. On the other hand, the false positive fraction will also increase, and therefore the true negative fraction and specificity will decrease.

When the variable under study cannot distinguish between the two groups, i.e. where there is no difference between the two distributions, the area will be equal to 0.5 (the ROC curve will coincide with the diagonal). When there is a perfect separation of the values of the two groups, i.e. there no overlapping of the distributions, the area under the ROC curve equals 1 (the ROC curve will reach the upper left corner of the plot).

The optimal criterion value takes into account not only sensitivity and specificity, but also disease prevalence, and costs of various decisions.

When a test is used either for the purpose of screening or to exclude a diagnostic possibility, a cut-off value with a higher sensitivity may be selected and when the test is used to confirm a disease, a higher specificity may be required.

In a ROC curve the true positive rate (Sensitivity) is plotted in function of the false positive rate (100-Specificity) for different cut-off points. Each point on the ROC curve represents a sensitivity/specificity pair corresponding to a particular decision threshold. A test with perfect discrimination (no overlap in the two distributions) has a ROC curve that passes through the upper left corner (100% sensitivity, 100% specificity). Therefore, the closer the ROC curve is to the upper left corner, the higher the overall accuracy of the test