Exploration librairie swirl : Variation Inflation Factors

> swirl()

Welcome to swirl! Please sign in. If you've been here before, use the same name as you did then. If you are new, call yourself something unique.

What shall I call you? ***jlbellier***

Would you like to continue with one of these lessons?

1: Regression Models Least Squares Estimation
2: No. Let me start something new.

Selection: ***2***

Please choose a course, or type 0 to exit swirl.

1: Regression Models
2: Statistical Inference
3: Take me to the swirl course repository!

Selection: ***1***

Please choose a lesson, or type 0 to return to course menu.

1: Introduction	*2: Residuals*
3: Least Squares Estimation	*4: Residual Variation*
5: Introduction to Multivariable Regression	*6: MultiVar Examples*
7: MultiVar Examples2	*8: MultiVar Examples3*
9: Residuals Diagnostics and Variation	*10: Variance Inflation Factors*
11: Overfitting and Underfitting	*12: Binary Outcomes*
13: Count Outcomes	

Selection: ***10***

Attempting to load lesson dependencies...

0%

Variance Inflation Factors. (Slides for this and other Data Science courses may be found at github https://github.com/DataScienceSpecialization/courses. If you care to use them, they must be downloaded as a zip file and viewed locally. This lesson corresponds to Regression_Models/02_04_residuals_variation_diagnostics.)

==== ***4%***

In modeling, our interest lies in parsimonious, interpretable representations of the data that enhance our understanding of the phenomena under study. Omitting variables results in bias in the coefficients of interest - unless their regressors are uncorrelated with the omitted ones. On the other hand, including any new variables increases (actual, not estimated) standard errors of other regressors. So we don't want to idly throw variables into the model. This lesson is about the second of these two issues, which is known as variance inflation.

======== ***8%***

We shall use simulations to illustrate variance inflation. The source code for these simulations is in a file named vifSims.R which I have copied into your working directory and tried to display in your source code editor. If I've failed to display it, you should open it manually.

============= ***12%***

Find the function, makelms, at the top of vifSims.R. The final expression in makelms creates 3 linear models. The first, lm(y ~ x1), predicts y in terms of x1, the second predicts y in terms of x1 and x2, the third in terms of all three regressors. The second coefficient of each model, for instance coef(lm(y ~ x1))[2], is extracted and returned in a 3-long vector. What does this second coefficient represent?

1: The coefficient of x2.
2: The coefficient of x1.
3: The coefficient of the intercept.

Selection: ***2***

You are really on a roll!

================= ***17%***

In makelms, the simulated dependent variable, y, depends on which of the regressors?

1: x1
2: x1, x2, and x3
3: x1 and x2

Selection: ***1***

All that practice is paying off!

===================== ***21%***

In vifSims.R, find the functions, rgp1() and rgp2(). Both functions generate 3 regressors, x1, x2, and x3. Compare the lines following the comment Point A in rgp1() with those following Point C in rgp2(). Which of the following statements about x1, x2, and x3 is true?

1: x1, x2, and x3 are uncorrelated in both rgp1() and rgp2().
2: x1, x2, and x3 are correlated in rgp1(), but not in rgp1().
3: x1, x2, and x3 are correlated in both rgp1() and rgp2().
4: x1, x2, and x3 are uncorrelated in rgp1(), but not in rgp2().

Selection: ***4***

You are amazing!

==========================   25%

In the line following Point B in rgp1(), the function maklms(x1, x2, x3) is applied 1000 times. Each time it is applied, it simulates a new dependent variable, y, and returns estimates of the coefficient of x1 for each of the 3 models, y ~ x1, y ~ x1 + x2, and y ~ x1 + x2 + x3. It thus computes 1000 estimates of the 3 coefficients, collecting the results in 3x1000 array, beta. In the next line, the expression, apply(betas, 1, var), does which of the following?

1: Computes the variance of each row.
2: Computes the variance of each column.

Selection: ***1***

Your dedication is inspiring!

==============================   29%

The function rgp1() computes the variance in estimates of the coefficient of x1 in each of the three models, y ~ x1, y ~ x1 + x2, and y ~ x1 + x2 + x3. (The results are rounded to 5 decimal places for convenient viewing.) This simulation approximates the variance (i.e., squared standard error) of x1's coefficient in each of these three models. Recall that variance inflation is due to correlated regressors and that in rgp1() the regressors are uncorrelated. Run the simulation rgp1() now. Be patient. It takes a while.

```
> rgp1()
[1] "Processing. Please wait."
     x1      x1      x1
0.00110 0.00111 0.00112
```

Nice work!

==================================   33%

The variances in each of the three models are approximately equal, as expected, since the other regressors, x2 and x3, are uncorrelated with the regressor of interest, x1. However, in rgp2(), x2 and x3 both depend on x1, so we should expect an effect. From the expressions assigning x2 and x3 which follow Point C, which is more strongly correlated with x1?

1: x3
2: x2

Selection: ***1***

That's correct!

======================================   38%

Run rgp2() to simulate standard errors in the coefficient of x1 for cases in which x1 is correlated with the other regressors

```
> rgp2()
[1] "Processing. Please wait."
```

```
    x1      x1      x1
0.00110 0.00240 0.00981
```

Nice work!

=== 42%

In this case, variance inflation due to correlated regressors is clear, and is most pronounced in the third model, y ~ x1 + x2 + x3, since x3 is the regressor most strongly correlated with x1.

== 46%

In these two simulations we had 1000 samples of estimated coefficients, hence could calculate sample variance in order to illustrate the effect. In a real case, we have only one set of coefficients and we depend on theoretical estimates. However, theoretical estimates contain an unknown constant of proportionality. We therefore depend on ratios of theoretical estimates called Variance Inflation Factors, or VIFs.

== 50%

A variance inflation factor (VIF) is a ratio of estimated variances, the variance due to including the ith regressor, divided by that due to including a corresponding ideal regressor which is uncorrelated with the others. VIF's can be calculated directly, but the car package provides a convenient method for the purpose as we will illustrate using the Swiss data from the datasets package.

== 54%

According to its documentation, the Swiss data set consists of a standardized fertility measure and socioeconomic indicators for each of 47 French-speaking provinces of Switzerland in about 1888 when Swiss fertility rates began to fall. Type head(swiss) or View(swiss) to examine the data.

```
> View(swiss)
```

That's correct!

== 58%

Fertility was thought to depend on five socioeconomic factors: the percent of males working in Agriculture, the percent of draftees receiving the highest grade on the army's Examination, the percent of draftees with Education beyond primary school, the percent of the population which was Roman Catholic, and the rate of Infant Mortality in the province. Use linear regression to model Fertility in terms of these five regressors and an intercept. Store the model in a variable named mdl.

```
> mdl <- lm(Fertility~., swiss)
```

All that practice is paying off!

=== 62%

Calculate the VIF's for each of the regressors using vif(mdl).

```
> vif(mdl)
    Agriculture     Examination      Education       Catholic Infant.Mortality
       2.284129        3.675420       2.774943       1.937160         1.107542
```

Keep up the great work!

*=== **67%***

These VIF's show, for each regression coefficient, the variance inflation due to including all the others. For instance, the variance in the estimated coefficient of Education is 2.774943 times what it might have been if Education were not correlated with the other regressors. Since Education and score on an Examination are likely to be correlated, we might guess that most of the variance inflation for Education is due to including Examination.

*=== **71%***

Make a second linear model of Fertility in which Examination is omitted, but the other four regressors are included. Store the result in a variable named mdl2.

> ***mdl2 <- lm(Fertility~Agriculture+Education+Catholic+Infant.Mortality, swiss)***

That's the answer I was looking for.

*=== **75%***

Calculate the VIF's for this model using vif(mdl2).

> ***vif(mdl2)***

```
   Agriculture       Education       Catholic Infant.Mortality
      2.147153        1.816361       1.299916        1.107528
```

You got it!

*=== **79%***

As expected, omitting Examination has markedly decreased the VIF for Education, from 2.774943 to 1.816361. Note that omitting Examination has had almost no effect the VIF for Infant Mortality. Chances are Examination and Infant Mortality are not strongly correlated. Now, before finishing this lesson, let's review several significant points.

*===**83%***

A VIF describes the increase in the variance of a coefficient due to the correlation of its regressor with the other regressors. What is the relationship of a VIF to the standard error of its coefficient?

1: VIF is the square of standard error inflation.
2: They are the same.
3: There is no relationship.

Selection: ***3***

Not quite, but you're learning! Try again.

Variance is the square of standard deviation, and standard error is the standard deviation of an estimated coefficient.

1: There is no relationship.
2: VIF is the square of standard error inflation.
3: They are the same.

Selection: ***2***

That's a job well done!

== 88%

If a regressor is strongly correlated with others, hence will increase their VIF's, why shouldn't we just exclude it?

1: We should never exclude anything.
2: Excluding it might bias coefficient estimates of regressors with which it is correlated.
3: We should always exclude it.

*Selection: **2***

You nailed it! Good job!

== 92%

The problems of variance inflation and bias due to excluded regressors both involve correlated regressors. However there are methods, such as factor analysis or principal componenent analysis, which can convert regressors to an equivalent uncorrelated set. Why then, when modeling, should we not just use uncorrelated regressors and avoid all the trouble?

1: We should always use uncorrelated regressors.
2: Factor analysis takes too much computation.
3: Using converted regressors may make interpretation difficult.

*Selection: **3***

Keep working like that and you'll get there!

== 96%

That completes the exercise in variance inflation. The issue of omitting regressors is discussed in another lesson.

== 100%