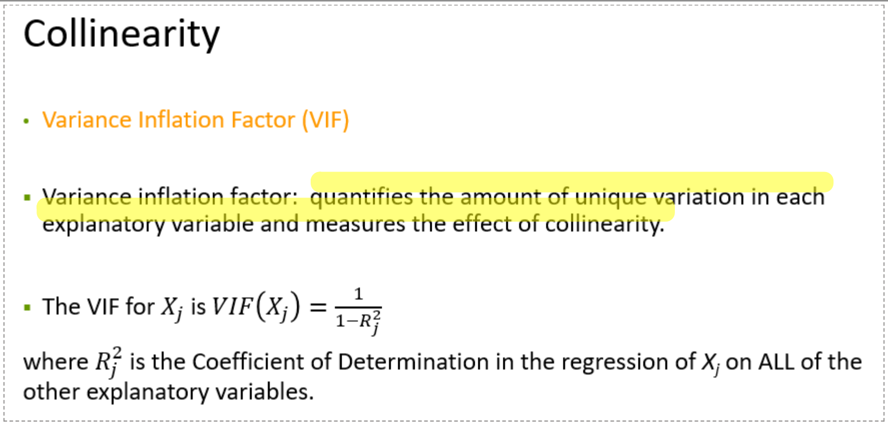
6.4 Multiple Linear Regression - Variance Inflation Factor - Part 1

Monday, 07 November 2022 16:34

Summary

- Variance Inflation Factor

- Define VIF.
    - Formula
    - How do you calculate it?
- What is R_j^2? or How do you calculate it?
- If $R_j^2 \uparrow \ \Rightarrow VIF$?



Collinearity

- Variance Inflation Factor (VIF)
- Variance inflation factor: quantifies the amount of unique variation in each explanatory variable and measures the effect of collinearity.
- The VIF for X_j is $VIF(X_j) = \frac{1}{1-R_j^2}$

where R_j^2 is the Coefficient of Determination in the regression of X_j on ALL of the other explanatory variables.

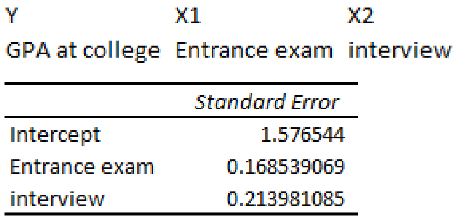
- R_j^2 is the coefficient of determination on a regression where that particular j^{th} variable is the response variable and all the other explanatory variables are the explanatory variables.
- R_j^2 will be a large value if the regression is significant, that is, if the response variable is fairly correlated with the explanatory variables.
- If R_j^2 is a large value, then VIF will be a large value.

  $R_j^2 \uparrow \ \Rightarrow \ VIF \uparrow$

  How?:

Let $R_j^2 = 0.9$	Let $R_j^2 = 0.1$
then $VIF(X_j) = \frac{1}{1-0.9} = \frac{1}{0.1} = 10$	then $VIF(X_j) = \frac{1}{1-0.1} = \frac{1}{0.9} = 1.11$

In MLR:

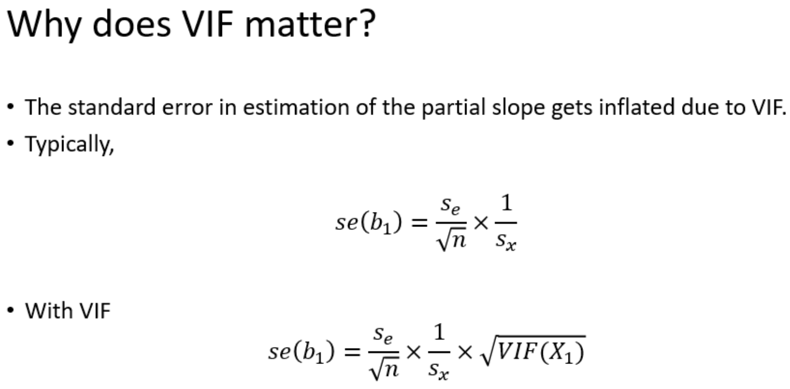


Y	X1	X2
GPA at college	Entrance exam	interview

	Standard Error
Intercept	1.576544
Entrance exam	0.168539069
interview	0.213981085

- Standard error in estimating b_1 is : $s_e(b_1) = 0.168$
- Standard error in estimating b_2 is : $s_e(b_2) = 0.213$

- Standard error in estimation of partial slopes, $se(b_1) = ?$
- What is s_e?
- What is s_{X_1}?
- With VIF, $se(b_1) = ?$



Why does VIF matter?

- The standard error in estimation of the partial slope gets inflated due to VIF.
- Typically,

$$se(b_1) = \frac{s_e}{\sqrt{n}} \times \frac{1}{s_x}$$

- With VIF

$$se(b_1) = \frac{s_e}{\sqrt{n}} \times \frac{1}{s_x} \times \sqrt{VIF(X_1)}$$

- s_e → Standard error, estimate of σ_ε
- s_{X_1} → Standard deviation in X_1

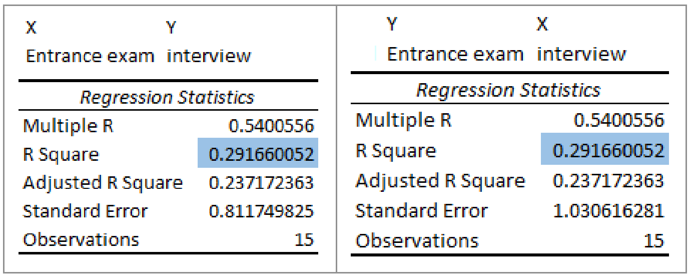
- If s_{X_1} is quite large, it helps us in understading the variation in Y.
- So if s_{X_1} is large, $se(b_1)$ will be small. Meaning we get high precision in estimating β_1.

- If the explanatory variables are uncorrelated, then $VIF =$?
- For correlated explanatory variables, VIF?
- Larger the VIF, larger the _____?
- What effect does VIF have on $se(b_1)$?

VIF

- If the explanatory variables are uncorrelated, then $R_j^2 = 0$, and $VIF = 1$.
- However, if the explanatory variables are correlated, then $VIF > 1$. Larger the VIF, larger is collinearity.
- Large VIF also substantially increases the standard error in predicting the partial slopes (*se(b)*). Thereby, making those predictions unreliable.

e.g., Take the example where we treated one explanatory variable as response variable and the other remained explanatory

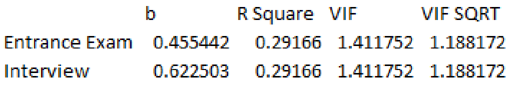


X Entrance exam, Y interview

Regression Statistics	
Multiple R	0.5400556
R Square	0.291660052
Adjusted R Square	0.237172363
Standard Error	0.811749825
Observations	15

Y Entrance exam, X interview

Regression Statistics	
Multiple R	0.5400556
R Square	0.291660052
Adjusted R Square	0.237172363
Standard Error	1.030616281
Observations	15



	b	R Square	VIF	VIF SQRT
Entrance Exam	0.455442	0.29166	1.411752	1.188172
Interview	0.622503	0.29166	1.411752	1.188172

- Here, since explanatory variables are correlated, $VIF > 1$.
- There is going to be an 18% of increase in $se(b)$.
- Note: Here, b values are taken from MLR, and the R Square is square of coefficient of correlation between both of the explanatory variable.

Inference in Multiple Regression

Inference for One Coefficient

- The *t*-statistic is used to test each slope using the null hypothesis H_0: $\beta_j = 0$.
- The *t*-statistic is calculated as

$$t_j = \frac{b_j - 0}{se(b_j)}$$

- In the example that we considered, $\sqrt{VIF}$ turned out to be very small.
- If [Equation] were high, it would inflate the standard errors, and then the t-stats would come down.
- If t-stats is very small, it will impact the [Equation] , and we may not be able to reject the null hypothesis.
- It will mean that that particular explanatory variable may be statistically insignificant for the regression.
- So, we want VIF value small, and it will happen only when the explanatory variables don't have too much correlation.