6.5 Multiple Linear Regression - Variance Inflation Factor - Part 2

Monday, 07 November 2022 18:57

Summary	• An example where explanatory variables on their own are significant (SLR), but when taken together become insignificant (MLR). • Signs and Remedies of Collinearity
	Interpreting Multiple Regression • Example: Estimating the price of a house • Response variable: Price of the house (INR). • Three explanatory variables: Size of the house (in square foot), number of bedrooms and the number of parking lots provided.
	Apartment data is given:

Area (Sq ft)	# of bedrooms	Parking lot	Price
9.5	2	2	5.68
10	3	2	8.9
8.7	2	1	7.6
10	3	3	10
11.45	2	2	8
20	2	3	9.8
9	2	2	8.1
8.34	2	2	7.1
11	3	2	9.1
13	3	1	5
14.5	4	3	12
16	3	3	11.5
8.19	1	2	6.4
7.9	1	2	7
8.6	2	3	8
10	2	2	7.9
11.45	3	1	9
12	3	3	9.35
15	4	3	10.3
11	3	2	14

- We're trying to predict the price of an apartment.
- Price given in the data is in 10 Lakhs.
- Area is given in 100 sq ft.

Here's the correlation coefficient matrix:

	Area (Sq ft)	*# of bedrooms*	*Parking lot*	*Price*
Area (Sq ft)	1			
# of bedrooms	0.503295389	1		
Parking lot	0.434453051	0.274318858	1	
Price	0.467910037	0.605720798	0.501392503	1

- Price seems to be affecting by all of the three factors.

SLR on Area-Price

Response Variable: Price
Explanatory Variable: Area

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.467910037
R Square	0.218939803
Adjusted R Square	0.17554757
Standard Error	1.972955837
Observations	20

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	19.64026979	19.64026979	5.045598878	0.037477004
Residual	18	70.06598521	3.892554734		

Total	19	89.706255

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	5.040157858	1.703678785	2.958396796	0.008412107	1.46086155	8.619454166
Area (Sq ft)	0.327646336	0.145864281	2.246241055	0.037477004	0.021196854	0.634095819

- $Significance\ F$ values tells us that the regression is significant ($p - value < 0.05$).

SLR on Number of Bedroom-Price

Response Variable: Price
Explanatory Variable: Number of Bedrooms

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.605720798
R Square	0.366897685
Adjusted R Square	0.331725334
Standard Error	1.776282599
Observations	20

ANOVA

	df	SS	MS	F	Significance F
Regression	1	32.91301731	32.91301731	10.43142345	0.004647911
Residual	18	56.79323769	3.155179872		
Total	19	89.706255			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	4.758615385	1.294091339	3.677186642	0.00172406	2.039830369	7.477400401
# of bedrooms	1.591153846	0.492652153	3.229771424	0.004647911	0.556130079	2.626177613

- $Significance\ F$ values tells us that the regression is significant ($p - value < 0.05$).

SLR on Parking lot-Price

Response Variable: Price
Explanatory Variable: Parking lot

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.501392503
R Square	0.251394442
Adjusted R Square	0.209805244
Standard Error	1.931530785
Observations	20

ANOVA

	df	SS	MS	F	Significance F
Regression	1	22.55165391	22.55165391	6.04470526	0.024307596
Residual	18	67.15460109	3.730811171		
Total	19	89.706255			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	5.292065217	1.466039445	3.609770008	0.002003183	2.212030636	8.372099798
Parking lot	1.565652174	0.636806842	2.458598231	0.024307596	0.227770645	2.903533703

- $Significance\ F$ values tells us that the regression is significant ($p - value < 0.05$).

- So, by themselves, each explanatory variable does help us to predict the prices.

MLR

Response Variable: Price
Explanatory Variable: Area, Number of bedrooms, Parking lot

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.702259246
R Square	0.493168049
Adjusted R Square	0.398137058
Standard Error	1.685711946
Observations	20

ANOVA

	df	SS	MS	F	Significance F
Regression	3	44.24025875	14.74675292	5.189549687	0.010764764
Residual	16	45.46599625	2.841624766		
Total	19	89.706255			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	2.667893129	1.67836639	1.589577309	0.131492473	-0.890084674	6.225870932
Area (Sq ft)	0.059750485	0.154380203	0.387034628	0.70383001	-0.267520926	0.387021895
# of bedrooms	1.236821718	0.542442699	2.280096533	0.036650141	0.086894565	2.38674887
Parking lot	1.046580675	0.618622039	1.691793388	0.110065941	-0.264839463	2.358000813

- R^2 indicates that the fitted equation explains % of the variation in price.

- The overall $Significance\ F$ values of the MLR is significant ($p - value < 0.05$).

- But,

	P-value
Intercept	0.131492473
Area (Sq ft)	0.70383001
# of bedrooms	0.036650141
Parking lot	0.110065941

	Lower 95%	Upper 95%
Intercept	-0.890084674	6.225870932
Area (Sq ft)	-0.267520926	0.387021895
# of bedrooms	0.086894565	2.38674887
Parking lot	-0.264839463	2.358000813

$p - value$ for Area

$H_0 : \beta_1 = 0$
$H_0 : \beta_1 \neq 0$

$\because p - value = 0.7 > 0.05$
null hypothesis cannot be rejected.

Also, look at the confidence interval. b_1 can be anywhere between $[-0.26,\ 0.38]$. It includes 0, we cannot reject it.

$p - value$ for # of bedrooms

$H_0 : \beta_2 = 0$
$H_0 : \beta_2 \neq 0$

$\because p - value = 0.03 < 0.05$
null hypothesis is rejected.

$\therefore\ b_2 = 1.23$ is a good estimate of β_2.

$p - value$ for Parking lot

$H_0 : \beta_3 = 0$
$H_0 : \beta_3 \neq 0$

$\because p - value = 0.11 > 0.05$
null hypothesis cannot be rejected.

Also, the confidence interval for b_3 is $[-0.26,\ 2.35]$.

- Individually, area and parking lot were significant.

- It happened because there was a strong correlation among the explanatory variables.

	Area (Sq ft)	# of bedrooms	Parking lot	Price
Area (Sq ft)	1			
# of bedrooms	0.503295389	1		
Parking lot	0.434453051	0.274318858	1	
Price	0.467910037	0.605720798	0.501392503	1

- That's what's making partial slopes insignificant even when the marginal slopes were larger.

	Marginal Slope	Partial Slope
Area	0.327	0.059
# of bedroom	1.59	1.23
Parking lot	1.56	1.04

- Also the standard errors for the partial slopes are larger than the marginal slope.

	Marginal Slope	Partial Slope
Area	0.145	0.154
# of bedroom	0.12	0.54
Parking lot	0.63	0.628

VIF

- From our example, the explanatory variables in the MLR are turning to be insignificant.
- The explanatory variables aren't significant once we have taken account of the other explanatory variables.

	• Partial slope conveys the unique variation explained by that particular explanatory variable. • However once you take account of the parking lot and the area, now the number of bedrooms do not offer anything unique that has already not been explained by these two variables. • Similarly, once you take account of the area and the number of bedrooms, now the parking lot does not offer anything unique that has not already been explained by these two variables. • This is why the explanatory variables are turning out to be insignificant in MLR, but they are significant in the SLR. • This is the impact of explanatory variables being correlated. b R-Square VIF VIF_SQRT Area 0.05975 0.348302 1.534452 1.23873 # of bedrooms 1.236822 0.257125 1.346122 1.160225 Parking lot 1.046581 0.192899 1.239002 1.113104
• What signs do these parameters show of collinearity: 1. R^2 2. Marginal and partial slopes 3. F-statistic 4. Standard errors for partial and marginal slopes 5. VIF	Collinearity Signs of Collinearity ▪ R^2 increases less than we'd expect. ▪ Slopes of correlated explanatory variables in the model change dramatically. ▪ The *F*-statistic is more impressive than individual *t*-statistics. ▪ Standard errors for partial slopes are larger than those for marginal slopes. ▪ Variance inflation factors increase. • We know that whenever we add an explanatory variable, R^2 is supposed to go up. • If there's a multi-collinearity, R^2 does not go up drastically, it goes up only fractionally. • Value of Marginal slopes > Partial slopes • Standard errors for Partial slopes > Marginal slopes
• What are some remedies for collinearity?	Collinearity • Remedies for Collinearity ▪ Remove redundant explanatory variables. ▪ Re-express explanatory variables (e.g., use the average of *Market % Change* and *Dow % Change* as an explanatory variable). ▪ Do nothing if the explanatory variables are significant with sensible estimates. • Re-expressing explanatory variables mean we can combine the correlated explanatory variables to create another explanatory variable.
	Removing Explanatory Variables • Issues ▪ After adding several explanatory variables to a model, some of those added and some of those originally present may not be statistically significant. ▪ Remove those variables for which both statistics and substance indicate removal (e.g., remove *Dow % Change* rather than *Market % Change*).