204: Regression Analysis

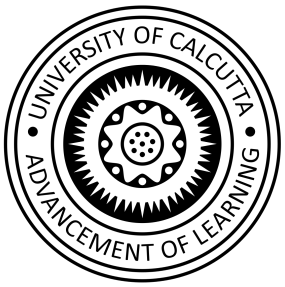
Instructor: Prof. Sugata Sen Roy

Report Card
Data Analysis with Regression

Dealing with Multicollinearity

Submitted by: Soumarya Basak

Submitted on: June 23, 2022



University of Calcutta

Department of Statistics

Contents

Problem 1

1.1 About the Practical

We have a data on consumption expenditure of family along with their wealth and income. We need to fit a suitable regression model on the data set, considering whether there may be presence of multicollinearity.

1.2 Analysis

In the given problem the "Cons. Exp (y)" is our response variable with two explanatory variables "income (X_1)" , and "wealth(X_2)" . For in depth analysis we need to fit a linear regression first based on the OLS estimators.
In practical scenario the variables and `income` and `wealth` are correlated variables, so the presence of multicollinearity is a feasible case here.

Analysis Required:

1. Fit a linear regression model.
2. Check for multicollinearity by different methods.
3. Add two observation to resolve the problem of multicollinearity.

1.2.1 Fitting Linear Regression

Output:

Parameters	***Estimate***	***Std. Error***	$t-value$	$Pr(>\|t\|)$
Intercept	18.16553	11.361	1.599	0.154
X_1	0.632614	1.384489	0.457	0.662
X_2	-0.009891	0.135714	-0.073	0.944

Residual standard error: 11.45 on 7degrees of freedom
Multiple R-squared: 0.9104
Adjusted R-squared: 0.8848
F-statistic: 35.57 on 2 and 7 DF
p-value: 0.00021

Observation:

The above is a very interesting result as though the parameters are comes out to be insignificant, the model is fitted with a higher Adjusted R^2 value. Which indicates that, though the variables are not at necessary to build the model still the model explains almost 88% of the response variability.

Conclusion 1:

A reason of such a result is that the explanatory variables may be highly correlated, that makes the Adjusted R^2 of the model high. So, the OLS estimator will not perform well for the model, we need for some remedy.

We are assuming the dependency between the variables, so to be sure we will use some measure to verify our assumption.

1.2.2 Check for Multicollinearity

If we consider the correlation matrix for the explanatory variables–

```
          x1          x2
x1  1.0000000  0.9989624
x2  0.9989624  1.0000000
```

It can be easily shown that, we variables are highly related.

VIF of the Variables

On calculating the VIFs of the regressor variables it come out to be —

```
      x1       x2
482.1275 482.1275
```

Since the VIFs are larger than 10, we may assume there may be multicollinearity between the variables.

CN for the model

The CN for the model comes out to be **5786.226** which is very large and indicates presence of Multicollinearity.

Conclusion 2:

Its now clear that here is a problem of multicollinearity in the model, so we need some further treatment to remove this problem.

1.2.3 Adding Observations

Some times adding new observation resolve the problem of Multicollinearity. Here we will add two observations to the data set to resolve the problem.
We add two observations one having — expenditure 160, income 120 and wealth 3000 and another one having expenditure 85, income 255 and wealth 920.

On fitting a regression model using the new data set we have the following result,
Output:

Parameters	***Estimate***	***Std. Error***	$t-value$	$Pr(>\|t\|)$
Intercept	17.852113	10.173655	1.755	0.1132
X_1	0.100630	0.054652	1.841	0.0987
X_2	0.042541	0.004735	8.984	8.66e-06

Residual standard error: 10.28 on 9 degrees of freedom
Multiple R-squared: 0.9289
Adjusted R-squared: 0.9131
F-statistic: 58.8 on 2 and 9 DF
p-value: 6.809e-06

VIF of the Variables

On calculating the VIFs of the regressor variables it come out to be —

```
      x1       x2
1.203552 1.203552
```

Correlation Matrix

If we consider the correlation matrix for the explanatory variables–

```
       x1      x2
x1 1.00000 0.41125
x2 0.41125 1.00000
```

Observation

The VIFs of the variables becomes very low for the new model, also the correlation between the variables also drops down, there is no such high relation between the variables. So we can say that addition of two new data point somewhat resolve our problem.

Conclusion

Since the model is free from multicollinearity so we need to drop the insignificant parameters from the model and refit it to get the final model.

Output:

Parameters	***Estimate***	***Std. Error***	$t-value$	$Pr(>\|t\|)$
X_1	0.157099	0.048548	3.236	0.00893
X_2	0.046237	0.004661	9.921	1.71e-06

Residual standard error: 11.3 on 10 degrees of freedom
Multiple R-squared: 0.9921
Adjusted R-squared: 0.9905
F-statistic: 624.3 on 2 and 10 DF
p-value: 3.167e-11

Observation

On removing the intercept term from the model both the parameters become significant with Adjusted R^2 0.99.
Also note this model is free from multicollinearity and the residuals follows the necessary assumption.

Conclusion

So adding 2 new observation to the data set make the model free from multicollinearity the final model is,

$$y = 0.157099 \times X_1 + 0.046237 \times X_2$$

Which signifies keeping X_2 fix a unit change in X_1 will affect y 0.157099 units, and keeping x_1 fixed, unit change in X_2 will affect y by 0.046237 units.

1.2.4 Principal Component Analysis

Another approach to get remedy from the problem of multicollinearity is performing `Principal Component Analysis` on the data set.

Main Idea Of PCA:

In Principal Component Analysis we mainly make orthogonal transformation to our explanatory variables, that is instead of taking the explanatory variable individually we try to predict response by some linear combination of regressor variables that contains almost all the information of the data.
By using principal component analysis we may drop few of our variables (Dimensionality Reduction). So the variables which causes multicollinearity may be avoided easily by this method without losing information.

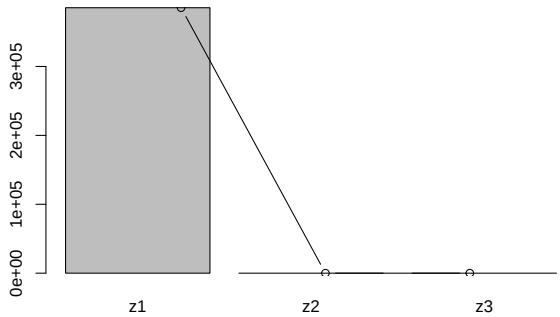
Principal Component Analysis:

Let X be our design matrix and we take a Orthogonal transformation, $Z = XL$, where L is a orthogonal matrix consist of the eigen vectors of $X'X$.
So, Columns of Z is nothing but some linear combination of regressor variables which are often called **Principal Components**.
The usefulness of such a transformation is that the first few columns of Z matrix will contain most of the information of the data, so we may only work with only those variables and may ignore other.

In the context of the given problem if we perform the transformation we will get a matrix Z with 3 columns Z_1, Z_2, Z_3

And among these 3 variables Z_1, Z_2, Z_3 the first variable explains much of the variability of the data whereas other 2 variables explains very little variability. So we may drop this 2 variables and fit the regression model with Z_1 only.

The plot shows how much the variability that 3 Principal Component explains of the data set



On fitting a regression model using the first principle component we get a model with insignificant intercept, so we after refitting the model without intercept term we have the following result,
Output:

Parameters	***Estimate***	***Std. Error***	$t-value$	$Pr(>\|t\|)$
Z_1	-0.061016	0.002038	-29.93	2.53e-10

Residual standard error: 11.89 on 9 degrees of freedom
Multiple R-squared: 0.9901
Adjusted R-squared: 0.989
F-statistic: 896.1 on 1 and 9 DF
p-value: 2.532e-10

Conclusion

The above model is based on only one regressor so the problem of multicollinearity doesn't arises here so our main problem is resolved, also the model gives a good Adjusted R^2 with low residual standard error. And the residuals follows the assumption of constant variance and without having autocorrelation leads to a good model to work with.

Final Model

$$y = -0.061016 \times Z_1$$

On transforming back to the our original variables the model becomes

$$y = 0.00003 + 0.00593 \times X_1 + 0.06073 \times X_2$$

Which signifies keeping X_2 fix a unit change in X_1 will affect y 0.00593 units, and keeping x_1 fixed, unit change in X_2 will affect y by 0.06073 units.

1.2.5 Ridge Regression

Ridge regression maybe use as a remedy of multicollinearity also. Rich regression is useful when the determinant of $X'X$ is zero or near to zero which may happen for dependency of regressor variables.

Main Idea Of Ridge Regression:

Mainly by ridge regression instead of finding a unbiased estimator of parameters we try to find a better biased estimator of the parameters. i.e instead of finding $\widehat{\beta}$ by

$$\widehat{\beta} = (X'X)^{-1}X'y$$

we find $\widehat{\beta}$ by

$$\widehat{\beta_R} = (X'X + cI_p)^{-1}X'y$$

The choice of c would be such that the $\widehat{\beta_R}$ has less MSE than $\widehat{\beta}$.

Performing Ridge Regression:

The first task before performing Ridge regression is to find the choice of c. In context of our problem, we consider different values of c and fit a model and choose that value for which the regression model gives minimum RSS with a good Adjusted R^2.

By the method of grid searching, c comes out to be **1.995262**. Now based on this $c = 1.995262$, we will find the ridge estimates of the parameters by,

$$\widehat{\beta_R} = (X'X + cI_p)^{-1}X'y$$

which comes out to be,

	Estimates
Intercept	6.129046444
x1	0.547649530
x2	0.004625073

Final Model

So, the final model after Ridge Regression is ,

$$y = 6.129046444 + 0.547649530 \times x_1 + 0.004625073 \times x_2$$

Problem 2

2.1 About the Practical

We have a data on 20 individuals with high blood pressure. The data consists of several video bills like Blood pressure, Age, Weight, Body surface area, Duration of hypertension, Basal Pulse, stress index etc. We are interested to develop a relation between blood pressure and other variables. We have to check whether the data further poses multicollinearity or not and try to resolve if there is the presence.

2.2 Analysis

In the given problem we have response variable `Blood pressure(y)` and the others are explanatory variables.

Analysis Required:

1. To fit a model based on all the explanatory variables to predict response.
2. To check whether multicollinearity is there.
3. To drop variables to resolve the problem of multicollinearity

2.2.1 Fitting Linear Regression

Here we have several explanatory variables so a correlation plot between the variables will give us an idea that which of the variables are needed to predict response.

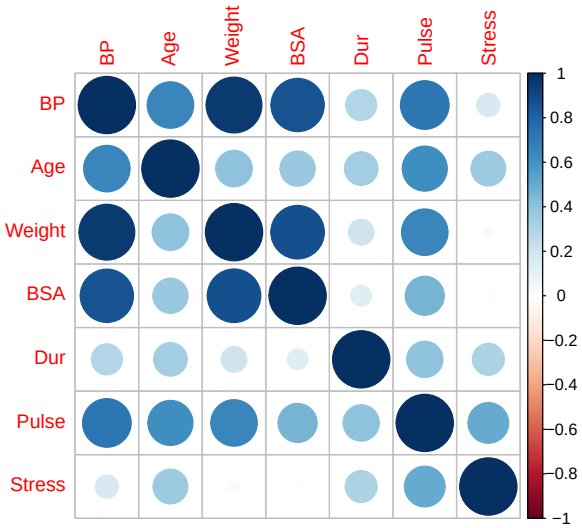


Figure 2.1: Correlation Plot of the variable

From the first row or the first column of the plot it is clear that all the variables has a good positive correlation with "BP (response variable)". So need to fit a full model.

Output:

Parameters	***Estimate***	***Std. Error***	$t - value$	$Pr(>\|t\|)$
Intercept	-12.870476	2.556650	-5.034	0.000229
Age	0.703259	0.049606	14.177	2.76e-09
Weight	0.969920	0.063108	15.369	1.02e-09
BSA	3.776491	1.580151	2.390	0.032694
Dur	0.068383	0.048441	1.412	0.181534
Pulse	-0.084485	0.051609	-1.637	0.125594
Stress	0.005572	0.003412	1.633	0.126491

Residual standard error: 0.4072 on 13 degrees of freedom
Multiple R-squared: 0.9962
Adjusted R-squared: 0.9944
F-statistic: 560.6 on 6 and 13 DF
p-value: 6.395e-15

Observation

Though we get a significant model the last 3 variables namely pulse, Dur and stress are insignificant and the Adjusted R^2 is very high.

Conclusion

Insignificance of 3 variables maybe due to several reasons but here we will focus on dependency of explanatory variable which may be a reason behind such high Adj R^2 even after 3 insignificant variables.
The correlation plot gives us a notion that the explanatory variables are correlated within themselves, which may cause multicollinearity.

VIFs of the Variable

As we suspect multicollinearity in the model, it will more clear if we see the VIFs of the Variables.

```
      Age    Weight       BSA       Dur     Pulse    Stress
 1.762807 8.417035 5.328751 1.237309 4.413575 1.834845
```

The variance inflation factor for `weight` and `BSA` variables are higher than the others. So the multicollinearity may be arises due to those variables.

Our Further Observation

In real life scenario `Blood Pressure` is much like to depends on `weight` rather than `Body surface area`. So, we will keep the `weight` variable and remove the the other one from the model.
Our correlation plot gives us a support to this belief, as then correlation between `weight` and `Blood Pressure` are more higher than between `BP` and `BSA`.

2.2.2 Get rid of Multicollinearity

Removing BSA

Removing the BSA variable leads us the following result,

Output:

Parameters	***Estimate***	***Std. Error***	$t-value$	$Pr(>\|t\|)$
Intercept	-15.116781	2.748758	-5.499	7.83e-05
Age	0.731940	0.055646	13.154	2.85e-09
Weight	1.098958	0.037773	29.093	6.37e-14
Dur	0.064105	0.055965	1.145	0.2712
Pulse	-0.137444	0.053885	-2.551	0.0231
Stress	0.007429	0.003841	1.934	0.0736

Residual standard error: 0.4708 on 14 degrees of freedom
Adjusted R-squared: 0.9925
F-statistic: 502.5 on 5 and 14 DF
p-value: 2.835e-15

Observation and Conclusion

The above model have 2 insignificant parameters. But the model is free from multicollinearity as the VIFs become low. So, we may get a better fit by dropping the insignificant variables.

It has to be noted that the `Pulse` variable has a good correlation with `Age` and `weight` which is a problem of real life situation, as a aged person or a heavy weighted person can have a high pulse rate. Lets keep this thought aside for now.

Removing 'Stress' on the last Model

Among all the variables `stress` has less correlation with response variable so we'll drop the `stress` variable from the model first, it may has least effect on response.

Output:

Parameters	***Estimate***	***Std. Error***	$t-value$	$Pr(>\|t\|)$
Intercept	-15.96851	2.95072	-5.412	7.21e-05
Age	0.74032	0.06033	12.271	3.19e-09
Weight	1.06556	0.03654	29.164	1.26e-14
Dur	-0.08165	0.04950	-1.650	0.120
Pulse	0.07448	0.06058	1.229	0.238

Residual standard error: 0.512 on 15 degrees of freedom
Adjusted R-squared: 0.9911
F-statistic: 530.3 on 4 and 15 DF
p-value: 5.957e-16

Observation and Conclusion

The above model is free from multicollinearity as the VIFs of the variables are low, but still

there are 2 insignificant parameters in the model which we will remove further.
Among the insignificant variables `Dur` has a lower correlation with response variable, so next we will drop `Dur`.

Removing 'Dur' on the last Model

Among the insignificant variables `Dur` has a lower correlation with response variable, so next we will drop `Dur`.

Output:

Parameters	***Estimate***	***Std. Error***	$t-value$	$Pr(>\|t\|)$
Intercept	-16.69000	2.93761	-5.681	3.40e-05
Age	0.75018	0.06074	12.350	1.36e-09
Weight	1.06135	0.03695	28.722	3.40e-15
Pulse	-0.06566	0.04852	-1.353	0.195

Residual standard error: 0.5201 on 16 degrees of freedom
Adjusted R-squared: 0.9908
F-statistic:684.7 on 3 and 16 DF
p-value: < 2.2e-16

Observation and Conclusion

Still the `Pulse` variable is insignificant for the model. The low VIFs of the variables indicates that the model is free of multicollinearity. So the dependency of the `Pulse` Variable on `Age` and `weight` do not causes multicollinearity, though we have to remove it from the model due to insignificance.

Removing 'Pulse' on the last Model

The last insignificant variable of the is need to remove further.

Output:

Parameters	***Estimate***	***Std. Error***	$t-value$	$Pr(>\|t\|)$
Intercept	-16.57937	3.00746	-5.513	3.80e-05
Age	0.70825	0.05351	13.235	2.22e-10
Weight	1.03296	0.03116	33.154	< 2e-16

Residual standard error:0.5327 on 17 degrees of freedom
Adjusted R-squared: 0.9904
F-statistic:978.2 on 2 and 17 DF
p-value: < 2.2e-16

Observation and Conclusion

Now the above model can be use as good model — as all the parameters are significant with low VIFs and the model almost explains almost 99% of the response variability, with a low RSS.

On performing the Residual Analysis, the residuals satisfy the assumption of constant variability and absence of auto-correlation.

Argument

In Real life scenario `Blood pressure ` of a person mostly depends on their age and weight , though in this data set there are several other variables, those are again somehow dependent upon these two major variables `Age` and `Weight`. For example for a heavy weighted person the body surface area and Pulse rate are high, for a aged person the stress level, hypertension are generally high. So it was likely for the regression model to be affected by multicollinearity. So dropping the variables with suitable reasons we come to a model which only consist of `Age` and `weight` which backs our prior beliefs on the problem.

2.2.3 Final Model

After performing lots of analysis we come to a model like below to predict `Blood Pressure` of a person.

$$\textbf{BP= -16.57937 + 0.70825} \times \textbf{Age + 1.03296} \times \textbf{Weight}$$

i.e. for persons of same age a unit increase in their weight will increase their BP 1.03296 units and for persons of same weight a unit increase in their Age will increase their BP 0.7082 units. so among age and weight, weight affects the BP much.

Appendix

The overall analysis is performed in R software, and the necessary output is discussed here. The data set and the codes can be found in the following Github repository.

Source code : https://github.com/SoumaryaBasak/Regression_Analysis_1.git