Typical Problems Estimating Econometric Models

When we are using the cross sectional and time series data we come across with a few econometric problems

- Heteroscedasticity
- Multicollinearity
- Autocorrelation

MULTICOLLINAIRITY

If we have several regressors in a regression model, how do we find out that we do not have the problem of Multicollinearity? If we have that problem, what are the consequences? And how do we deal with them? We discuss this topic

In this we examined the problem of Multicollinearity, a problem commonly encountered in empirical work, especially if there are several correlated explanatory variables in the model. As long as collinearity is not perfect, we can work within the framework of the classical linear regression model, provided the other assumptions of the CLRM are satisfied.

If collinearity is not perfect, but high, several consequences ensue. The OLS estimators are still BLUE, but one or more regression coefficients have large standard errors relative to the values of the coefficients, thereby making the t ratios small. Therefore, one may conclude (misleadingly) that the true values of these coefficients are not different from zero. Also, the regression coefficients may be very sensitive to small changes in the data, especially if the sample is relatively small

There are several diagnostic tests to detect collinearity, but there is no guarantee that they will yield satisfactory results. It is basically a trial and error process.

The best practical advice is to do nothing if you encounter collinearity, for very often we have no control over the data. However, it is very important that the variables included in the model are chosen carefully. As our illustrative example shows, redefining a model by excluding variables that may not belong in the model may attenuate the collinearity problem, provided we do not omit variables that are relevant in a given situation. Otherwise, in reducing collinearity we will be committing model specification errors. So, think about the model carefully before you estimate the regression model.

There is one caveat. If there is Multicollinearity in a model and if your objective is forecasting, Multicollinearity may not be bad, provided the collinear relationship observed in the sample continues to hold in the forecast period.

Finally, there is a statistical technique, called principal components analysis, which will “resolve” the problem of near-collinearity. In PCA we construct artificial variables in such a way that they are orthogonal to each other. These artificial variables, called principal components (PC), are extracted from the original X regressors.

We can then regress the original regressand on the principal components. We showed how the PCs are computed and interpreted, using our illustrative example.

One advantage of this method is that the PCs are usually smaller in number than the original number of regressors. But one practical disadvantage of the PCA is that the PCs very often do not have viable economic meaning, as they are (weighted) combinations of the original variables which may be measured in different units of measurement. Therefore, it may be hard to interpret the PCs. That is why they are not much used in economic research, although they are used extensively in psychological and education research.

Getting a grasp on perfect Multicollinearity, which is uncommon, is easier if you can picture an econometric model that uses two independent variables, such as the following:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i$$

Suppose that, in this model,

$$X_{i2} = \alpha_0 + \alpha_1 X_{i1}$$

where the alphas are constants. By substitution, you obtain

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 \left(\alpha_0 + \alpha_1 X_{i1}\right) + \varepsilon_i$$

which indicates that the model collapses and can’t be estimated as originally specified.

Remember- *Perfect Multicollinearity occurs when two or more independent variables in a regression model exhibit a deterministic (perfectly predictable or containing no randomness) linear relationship.*

The result of perfect Multicollinearity is that you can’t obtain any structural inferences about the original model using sample data for estimation. In a model with perfect Multicollinearity, your regression coefficients are indeterminate and their standard errors are infinite.

Warning- *Perfect Multicollinearity usually occurs when data has been constructed or manipulated by the researcher. For example, you have perfect Multicollinearity if you include a dummy variable for every possible group or category of a qualitative characteristic instead of including a variable for all but one of the groups.*

The 2 Types of Multicollinearity

M*ulticollinearity* arises when a linear relationship exists between two or more independent variables in a regression model. In practice, you rarely encounter perfect Multicollinearity, but high Multicollinearity is quite common and can cause substantial problems for your regression analysis.

Two types of Multicollinearity exist:

- **Perfect Multicollinearity** occurs when two or more independent variables in a regression model exhibit a *deterministic* (perfectly predictable or containing no randomness) linear relationship. When perfectly collinear variables are included as independent variables, you can’t use the OLS technique to estimate the value of the parameters. Perfect Multicollinearity, therefore, violates one of the classical linear regression model (CLRM) assumptions.
- **High Multicollinearity** results from a linear relationship between your independent variables with a high degree of correlation but aren’t completely deterministic (in other words, they don’t have perfect correlation). It’s much more common than its perfect counterpart and can be equally problematic when it comes to estimating an econometric model.

Remember- *In practice, perfect Multicollinearity is uncommon and can be avoided with careful attention to the model’s independent variables. However, high Multicollinearity is quite common and can create severe estimation problems. For this reason, when* *econometricians point to a Multicollinearity issue, they're typically referring to high Multicollinearity rather than perfect Multicollinearity.*

The problem of Multicollinearity occurs in a multiple regression where pair wise correlation exists between independent variables

F (X_1, X_2, X_3, X_4…., X_n)

Where,

Y=dependent variable

Xi…..=independent variables

Suppose we find a high degree of correlation in the pair of independent variables

i.e. (x1, x2), (x1, x3), (x3, x2) ...

Are highly correlated in the data set there could be a perfect or near Multicollinearity

If the correlation is 1 or near to 1 between the pairs of variables such as between (x1,x2) or (x2,x3) it is known as near perfect Multicollinearity.

X1	X2	X3	X4
5	2	5	20
8	3	7	29
1	4	2	30
3	9	4	45
5	5	6	20
4	6	5	19

Here x1 and x3 are Multicollinearity as they are near by

Such situation may occur when we use more than 1 independent variable in a multiple regression model

Datatype-Multicollinearity

These Multicollinearity problem exist when there is high correlation between the pair of independent variables in multiple regression.it can be observed both cross sectional data and time series data.

Particularly when we are using macro data. The time series data it will be more.

<u>Consequences of Multicollinearity:</u>

1. t-rations are satisfactory level is low. Statistically insignificant

2. R^2 which is the explanatory power is very high

3. The OLS (ordinary least square) estimator have large variance and covariance making precious estimation difficult.

4. the larger variance leads to a situation when the confidence interval tends to be much wider and therefore the zero null hypothesis is accepted since t – ratio of one or more coefficients tends to be statistically insignificant.

Detection of Multicollinearity:

1. When we observe high r^2 but few significant t-values in an estimated equation we may conclude the presence of Multicollinearity in the data set.
2. Examination of correlation matrix
3. Examination of partial correlation
4. Examine the auxiliary regression
5. Computation of conditional index

Remedial measure in the presence of the Multicollinearity:

Since there may be some problem faced during estimation and fore casting in the presence of collinearity

Method 1: Dropping a highly collinear variable

1. We can check the correlation matrix and drop one variable out of the paid that has a highly correlation.

2. A variable out of the pale can be dropped and the equation estimated. Again the variables could be dropped and then equation resituated these process can be re-estimated and the best fitted equation could be selected on the basis of the explanatory power

Method 2: Transformation of data

If the researcher is unable to exclude any of the variable due to theoretical consideration

A remedial measure like the first difference method is used in time series data

In these method the values of all variables for time period t is subtracted from the value of time period (t+1) and fresh data set is generated

Method3: A researcher can change the independent variable by deleting or adding variable in case of empirical research in a time series data Multicollinearity in normally found, while in case of Crossectional data these problem is found.

Example: Determinates of agriculture production:

Determinants of Agricultural Production

A consultancy firm is interested in examining the determinants of agricultural production In India for the last few decades and identifying the best-fitted model for explanation and forecasting. It collects data on agricultural production (the dependent variable), gross sown area, gross irrigated area, consumption of fertilizers, consumption of pesticides, and annual tractor sales (the independent variables). It uses a multiple regression model with the time series data. The regression model is used to estimate the relationship and is given as follows

$$\mathbf{Y_1 = b_1 + b_2X_2 + b_3X_3 + b_4X_4 + b_5X_5 + b_6X_6 + u}$$

Where Y = Total food grain production; X_2 = Gross sown area; X_3 = Gross irrigated area; X_4 = Consumption of fertilizers; X_5 = consumption of pesticides; X_6 = Annual tractor sales; and u = Error term.

The experts in the firm, suggest that since time series data for six independent variables are used, it may be worth a while to check for the problem of multicollinearity and address it before selecting an odd model. Table - 1 summarizes a set of time series data on the aforementioned variables. The data reported in the table have been used in the Excel OR SPSS package to estimate the regression equation, and the output is given in Table - 2

Table 1 Time series data

year	Total food grain (million tonnes)	Gross sown area (million hectares)	Gross irrigated area (million hectares)	Consumption of fertilisers (N+P+K) (Lakh tonnes)	Consumption of pesticides (technical grade material) (‘000 tonnes)	Annual tractor sales
1982-1983	129.52	172.75	51.83	63.88	50.00	63,073
1983-1984	152.37	179.56	53.82	77.10	55.00	74,318
1984-1985	145.54	176.33	54.53	82.11	56.00	80,317
1985-1986	150.44	178.46	54.28	84.74	52.00	76,886
1986-1987	143.42	176.41	55.76	86.45	50.00	80,164
1987-1988	140.35	170.74	56.04	87.84	66.90	93,157
1988-1989	169.92	182.28	61.13	110.40	75.89	1,10,323
199891990	171.04	182.27	61.85	115.68	72.00	1,22,098
1990-1991	176.39	185.74	63.20	125.46	75.00	1,39,831
1991-1992	163.38	182.24	65.68	127.28	72.13	1,50,582
1992-1993	179.48	185.70	66.76	121.55	70.79	1,44,330
1993-1994	184.26	186.58	68.26	123.66	63.65	1,38,879
1994-1995	191.50	188.05	70.65	135.64	61.36	1,64,841
1995-1996	180.42	187.47	71.35	138.76	61.26	1,91,329
1996-1997	199.43	189.50	76.03	143.08	56.11	2,20,937
1997-1998	193.12	189.99	75.67	161.88	52.24	2,51,198
1998-1999	203.61	191.65	78.67	167.98	49.16	2,62,322
1999-2000	209.80	188.40	79.22	180.70	46.20	2,73,181
2000-2001	196.81	185.34	76.19	167.02	43.58	2,54,825
2001-2002	212.85	188.29	78.42	173.60	47.02	2,25,280
2002-2003	174.78	175.58	73.41	160.94	48.30	1,73,098
2003-2004	213.19	190.08	78.15	167.99	41.00	1,90,336
2004-2005	198.36	191.55	81.18	183.98	40.67	2,47,531
2005-2006	208.59	193.05	83.94	203.40	39.77	2,96,080
2006-2007	217.28	193.23	86.50	216.51	37.95	3,52,781
2007-2008	230.78	195.83	87.26	225.70	36.29	3,46,501
2020-2021						

*Source: Indiastat.com

Table 2 (a) Output: Model Summary

Model	R	R^2	Adjusted R^2	SE of the estimate
1	**0.980**[a]	0.961	**0.951**	5.98485

*Predictors: (Constant), tractor sales, pesticide consumption, gross sown area, gross irrigated area, fertilizer consumption.

Table 2 (b) Output: ANOVA

Model	Sum of squares	df	Mean square	F	Sig.
1 Regression	17,510.023	5	3,502.005	97.771	**0.000**[b]
Residual	716.369	20	35.818		
Total	18,266.392	25			

Table 2 (c) Output: Coefficients

Models	Unstandardized coefficients		Standardized coefficients	t	Sig.
	B	SE	Beta		
(Constant)	-225.234	59.399		-3.792	**0.001 (OK)**
Gross_sown_area	**1.603**	**0.397**	**0.400**	**4.039**	**0.001(OK)**
Gross_irrigated_area	**1.387**	**0.688**	**0.563**	**2.017**	**0.057(OK)**
Fertilizer_Consumption	0.206	0.172	0.339	1.202	**0.244(NS)**
Pesticide_Consumption	0.042	0.314	0.019	0.315	0.756(NS)
Tractor_Sales	-9.036E-5	0.000	-0.286	-1.477	0.155(NS)

Dependent variable: Total food grain production; predictors: (constant); tractor sales, pesticides consumption, gross sown area, gross irrigated area, fertilizer consumption.

The estimated equation from Table 2 (c) can be written as follows:

Total food Production = -225.234 + 1.603 Gross sown area + 1.387 Gross irrigated area + 0.206 fertilizer consumption + 0.042 pesticide consumption – 0.0000903 Tractor sales

Using this equation, we can have short-run forecast by giving the future values of the independent variable in the equation and finding the value of the dependent variable. However, before going to accept this equation as one of the best-fitted model, it may be necessary to examine whether the problem of multicollinearity is present in the model. As mentioned earlier, there are different ways to detect multicollinearity. However, we will use the method of examinations of t-ratios, explanatory power, and correlation matric to conclude on the presence or absence of multicollinearity.

It may be observed that the explanatory power of the equation is very high as seen from adjusted R^2, which is 0.95. However, when we examine the significance levels of the coefficients of the explanatory variables, we find that only gross sown area and irrigated area are statistically significant at the acceptable limit of 0.10. this may be due to the pairwise high correlation between the explanatory variables since, we may recall our class discussion , high explanatory power and few significant t-ratios are symptoms of multicollinearity.

Therefore, the estimated equation will have less generalizability. This means that we are not sure about the effect of each of the variables on the dependent variables since the respective t-values are not statistically significant. Moreover, in such a situation, the coefficients are not BLUE. Such a situation arises when some of the explanatory variables are highly correlated, which is known as near-perfect collinearity.

Let us examine the correlation matrix that gives the pairwise correlation between the explanatory variables with respect to this example. The correlation matric is reported in Table 3.

Table 3 Correlation Matrix

	Gross sown area	Gross irrigated area	Fertilizer consumption	Pesticide consumption	Tractor sales
Gross sown area	1.000	0.874	0.845	-0.396	0.850
Gross irrigated area	0.874	1.000	0.984	-0.590	0.964
Fertilizer consumption	0.845	0.984	1.000	-0.601	**0.969**
Pesticide consumption	-0.396	-0.590	-0.601	1.000	-0.619
Tractor sales	0.850	0.964	0.969	-0.619	1.000

From the table, it can be observed that some of the pairs of the explanatory variables have high correlation amongst them, and therefore, there could be a possibility of multicollinearity in the data set. In such a situation, the estimated coefficient in the equation are not expected to possess the characteristic of BLUE.

Example from Gary Coop: For this question, use data set which is attached FOREST.XLS with Y=deforestation.
X= population density, X =% change in cropland and X: =change in pasture. Carry out a multiple regression analysis of this data set addressing the issues raised in this. For instance, you may want to:

(a) Regress Y on X1, X2 and X3 and verbally interpret the coefficient estimates you obtain.
(b) Discuss the statistical significance of the coefficients. Are there explanatory variables that can be dropped?

(c) Discuss the fit of the regression.

(d) Calculate a correlation matrix. Through consideration of this and regression results, discuss the issue of multicollinearity.

The data set FOREST XLS Contains data on Y=deforestation,
X = population density, W= change in cropland and Z = change in pasture land

Forest loss	Pop density	Crop ch	Pasture ch
0.7	357	27.9	0
0.7	48	1.7	0
0.8	932	14.5	0
0.7	366	17.9	0
0.8	83	2.2	0
0	22	5.1	0
0	67	4	-6.6
0.6	413	0	0
0.3	496	0.4	-1.1
0.5	458	6.5	0
0.4	152	3.9	0
1	115	3.9	12.2
0.9	964	18.3	0
1.2	459	3.9	0
1	421	19	0
1.3	723	-2.6	0
1.1	256	3.4	0
0.5	294	0.5	0
0.7	1310	6.2	0
0.6	627	8.6	0
1.4	714	7.3	0
0.6	270	1.2	0
0.4	52	3.4	0
0.2	71	13.7	0
0.4	135	0	0
0.6	50	0.9	0
0.6	182	3.5	0
0.7	82	0.5	0
0.5	24	2	0
1.3	1137	25.2	0
0.7	195	1.3	0
1.2	325	1	0
1	120	3.1	0
0.6	282	8.8	0
0.8	228	3.3	0
0.6	351	8.5	2.4
1	1541	14.3	6
2.9	1661	4	0

1.3	2769	1.5	0
1	510	0.4	0
0.9	200	3.6	0
1.2	678	0.2	-0.8
2.9	1113	25.5	29.7
1.4	2178	-2.9	18.3
1	1074	12.3	-1.5
1.8	586	1.5	0
2.9	2232	2.5	24
0.3	92	8.3	-20.1
2.6	640	4.5	15.6
2	2663	1.1	0
1.6	925	7.9	7.7
2	503	3.7	6.2
1.2	472	0.7	0
1.7	346	2.1	10.7
1.7	337	16.7	13.9
0.2	0.89	9	9.1
0.9	993	4	15.3
2.5	1575	2.4	0
0.1	41	0.1	0.8
3.9	2501	1.7	-2.4
5.3	2304	1.8	-8.1
0.1	29	39.7	1.7
1.9	2493	3.4	0
1.1	71	12.9	-1.7
0.6	185	23.1	7.5
0.6	327	4.1	5.8
1.7	409	9.4	29.2
2.4	117	26.7	33.5
0.4	179	6.1	0
1.2	234	4.3	2.9

Multicollinearity 2B

The Klein-Goldberger data on the relationship between aggregate US Consumption(C), as a function of wage income (W), nonwage-nonfarm income (p), and farm income (A) for the period 1921-50

Year	C	W	P	A
1928	58.2	39.21	17.73	4.39
1929	62.2	42.31	20.29	4.60
1930	58.6	40.37	18.83	3.25
1931	56.6	39.15	17.44	2.61
1932	51.6	34.00	14.76	1.67
1933	51.1	33.59	13.39	2.44

1934	54.0	36.88	13.93	2.39
1935	57.2	39.27	14.67	5.00
1936	62.8	45.51	17.20	3.93
1937	65.0	46.06	17.15	5.48
1938	63.9	44.16	15.92	4.47
1939	67.5	47.68	17.59	4.51
1940	71.3	50.79	18.49	4.90
1941	76.6	57.78	19.18	6.37
1945	86.3	78.97	19.12	8.42
1946	95.7	73.54	19.76	9.27
1947	98.3	71.92	17.55	8.87
1948	100.3	74.01	19.17	9.30
1949	103.2	75.51	20.20	6.95
1950	108.9	80.97	22.12	7.15

Multicollinearity 3C

Tata Motors is a major automobile manufacturer in India. The company is interested to know how octane level present in gasoline and weight of the car impact the mileage. A random sample of 12 cars were taken for studying this relationship. The following table gives data on mileage (Y), weight of the car (X_1) and octane level (X_2). Answer the following questions.

Y	**X_1(Tons)**	**X_2**
16.5	3.4	88
14.6	4.1	90
21.8	2.5	94
15.4	2.8	86
18.4	5.6	86
20.2	8.2	95
25.4	4.0	98
16.6	6.5	92
13.5	4.5	84
18.7	4.8	96
14.8	3.2	81
21.6	5.4	92

1. Fit a multiple linear regression model to the above data.
2. Compute coefficient of determination and adjusted R^2.
3. Calculate the standard error of slope coefficients.
4. Construct confidence interval for slope coefficients at 5% significance level.
5. Test the hypothesis that β_2=1 at 1% significance level.
6. Test the hypothesis of overall fit of the model.
7. Test the marginal contribution of each independent variables in the model.
8. Report the regression results in a standard format.

Multicollinearity

Determinants of BP

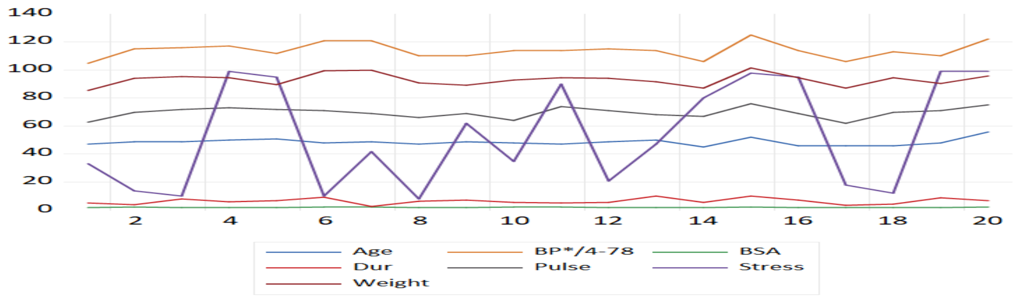
Some researchers observed the following data on 20 individuals with high blood pressure:

- blood pressure (y = *BP*, in mm Hg)
- age (x_1 = *Age*, in years)
- weight (x_2 = *Weight*, in kg)
- body surface area (x_3 = *BSA*, in sq m)
- duration of hypertension (x_4 = *Dur*, in years)
- basal pulse (x_5 = *Pulse*, in beats per minute)
- stress index (x_6 = *Stress*)

The researchers were interested in determining if a relationship exists between blood pressure and age, weight, body surface area, duration, pulse rate and/or stress level.

	AGE	BP	BSA	PULSE	DUR	STRESS	WEIGHT
Mean	48.6	114	1.998	69.6	6.43	53.35	93.09
Median	48.5	114	1.98	70	6	44.5	94.15
Maximum	56	125	2.25	76	10.2	99	101.3
Minimum	45	105	1.75	62	2.5	8	85.4
Std. Dev.	2.500526	5.428967	0.136482	3.803046	2.145276	37.08635	4.294905
Skewness	1.215166	0.206531	0.306786	-0.383815	0.163171	0.11172	0.085864
Kurtosis	4.957521	2.498469	2.405399	2.529865	2.261233	1.315228	2.441721
Jarque-Bera	8.115335	0.351794	0.608351	0.675236	0.543562	2.406984	0.284305
Probability	0.017289	0.838704	0.737731	0.713468	0.762021	0.300144	0.867489
Sum	972	2280	39.96	1392	128.6	1067	1861.8
Sum Sq. Dev.	118.8	560	0.35392	274.8	87.442	26132.55	350.478
Observations	20	20	20	20	20	20	20

B

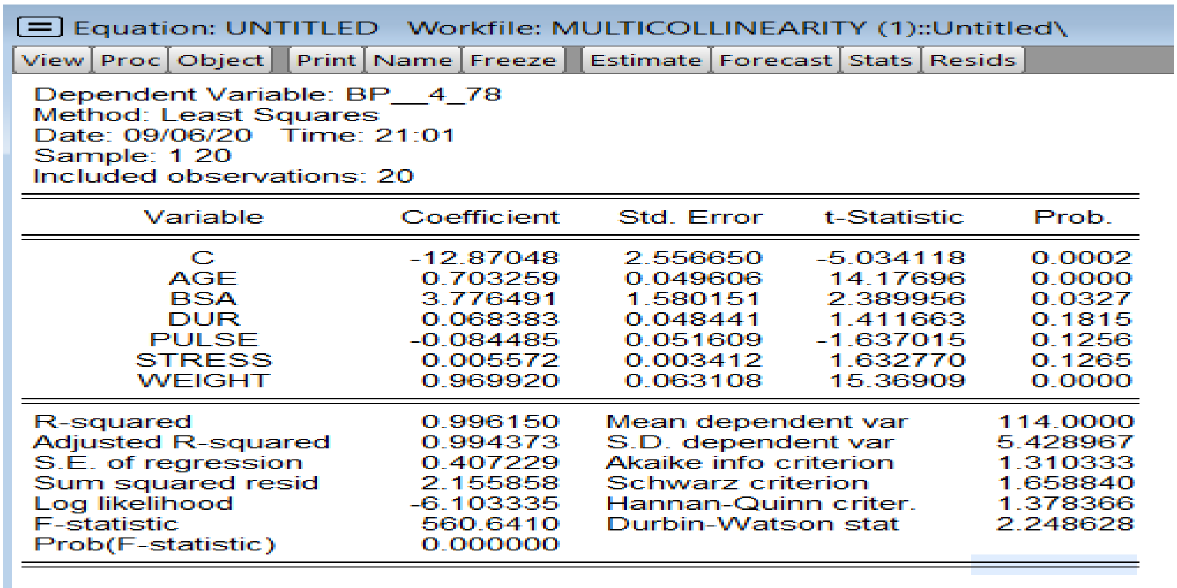


121	49	99.8	2.25	2.5	69	42
110	47	90.9	1.9	6.2	66	8
110	49	89.2	1.83	7.1	69	62
114	48	92.7	2.07	5.6	64	35
114	47	94.4	2.07	5.3	74	90
115	49	94.1	1.98	5.6	71	21
114	50	91.6	2.05	10.2	68	47
106	45	87.1	1.92	5.6	67	80
125	52	101.3	2.19	10	76	98
114	46	94.5	1.98	7.4	69	95
106	46	87	1.87	3.6	62	18
113	46	94.5	1.9	4.3	70	12
110	48	90.5	1.88	9	71	99
122	56	95.7	2.09	7	75	99

Solution:

	AGE	BSA	DUR	PULSE	STRESS	WEIGHT
AGE	1	0.378454507	0.343792	0.618764	0.368224	0.407349
BSA	0.3784546	1	0.13054	0.464819	0.018446	0.875305
DUR	0.3437921	0.130540013	1	0.401514	0.31164	0.20065
PULSE	0.6187643	0.464818807	0.401514	1	0.50631	0.65934
STRESS	0.3682237	0.018446338	0.31164	0.50631	1	0.034355
WEIGHT	0.4073493	0.875304815	0.20065	0.65934	0.034355	1

On analysing the data, we obtain the following results with regards to its basic descriptive statistics:


Equation: UNTITLED   Workfile: MULTICOLLINEARITY (1)::Untitled\
View
Proc
Object
Print
Name
Freeze
Estimate
Forecast
Stats
Resids

```
Dependent Variable: BP__4_78
Method: Least Squares
Date: 09/06/20   Time: 21:01
Sample: 1 20
Included observations: 20

     Variable        Coefficient    Std. Error    t-Statistic    Prob.

        C             -12.87048      2.556650     -5.034118     0.0002
       AGE             0.703259      0.049606      14.17696     0.0000
       BSA             3.776491      1.580151      2.389956     0.0327
       DUR             0.068383      0.048441      1.411663     0.1815
      PULSE           -0.084485      0.051609     -1.637015     0.1256
      STRESS           0.005572      0.003412      1.632770     0.1265
      WEIGHT           0.969920      0.063108      15.36909     0.0000

R-squared               0.996150    Mean dependent var      114.0000
Adjusted R-squared      0.994373    S.D. dependent var      5.428967
S.E. of regression      0.407229    Akaike info criterion   1.310333
Sum squared resid       2.155858    Schwarz criterion       1.658840
Log likelihood         -6.103335    Hannan-Quinn criter.    1.378366
F-statistic             560.6410    Durbin-Watson stat      2.248628
Prob(F-statistic)       0.000000
```

On further analysis i.e. running a regression analysis using LS method in Eviews, the following results are obtained:

Dependent Variable: BP

Independent Variables: Age, BSA, Pulse, Dur, Stress, Weight

Estimated Model:

BP= -12.8704760217 + 0.703259394357*AGE + 3.77649100344*BSA + 0.0683830403318*DUR - 0.0844846868374*PULSE + 0.00557150018635*STRESS + 0.969919778179*WEIGHT

Here we can see that the model has a high explanatory power since the Adjusted R^2 value is 0.994. However on further analysis, it can be seen that only three of the independent variable I.e. Age, BSA and Weight is statistically significant (0.000, 0.0327, 0.000 respectively) while Duration, Pulse and Weight are statistically insignificant (0.1815, 0.1256, 0.1265respectively).

These majority statistically insignificant variables along with the high explanatory power implies that there may be high pair wise correlation among the independent variables. To test this, we check the correlation among the variables.

Here as it can be seen that high correlation exists between a pair of the independent variable, hence we may conclude that there is a problem of Multicollinearity in the data.

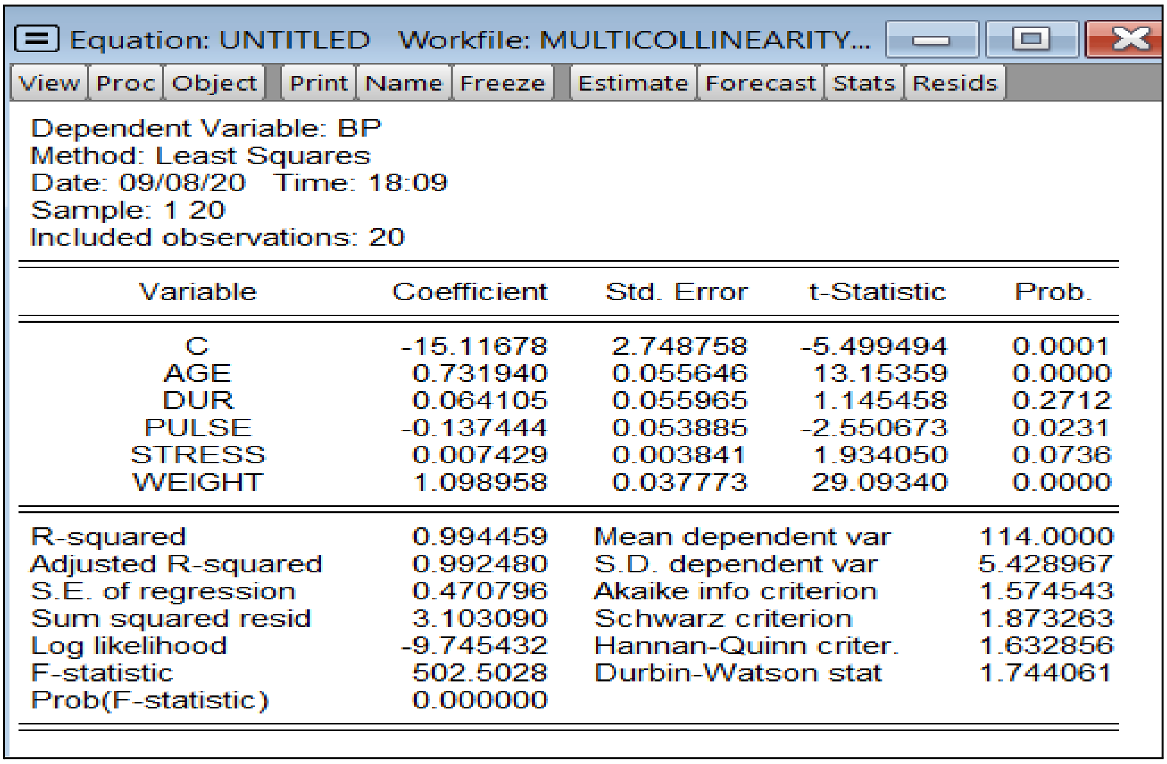
Remedial Measures

In order to remove the problem of Multicollinearity one remedial measure taken is to remove one of the independent variables. To do this the following process is undertaken.

Step 1:

Find the pair of independent variables with the highest correlation. In the given data, it is as following:

	AGE	BSA	DUR	PULSE	STRESS	WEIGHT
AGE	1	0.378454597	0.343792	0.618764	0.368224	0.407349
BSA	0.3784546	1	0.13054	0.464819	0.018446	0.875305
DUR	0.3437921	0.130540013	1	0.401514	0.31164	0.20065
PULSE	0.6187643	0.464818807	0.401514	1	0.50631	0.65934
STRESS	0.3682237	0.018446338	0.31164	0.50631	1	0.034355
WEIGHT	0.4073493	0.875304815	0.20065	0.65934	0.034355	1



```
Dependent Variable: BP
Method: Least Squares
Date: 09/08/20   Time: 18:09
Sample: 1 20
Included observations: 20

      Variable          Coefficient    Std. Error    t-Statistic    Prob.

         C               -15.11678     2.748758      -5.499494     0.0001
        AGE               0.731940     0.055646      13.15359      0.0000
        DUR               0.064105     0.055965      1.145458      0.2712
       PULSE             -0.137444     0.053885      -2.550673     0.0231
       STRESS             0.007429     0.003841      1.934050      0.0736
       WEIGHT             1.098958     0.037773      29.09340      0.0000

R-squared                 0.994459     Mean dependent var       114.0000
Adjusted R-squared        0.992480     S.D. dependent var       5.428967
S.E. of regression        0.470796     Akaike info criterion    1.574543
Sum squared resid         3.103090     Schwarz criterion        1.873263
Log likelihood           -9.745432     Hannan-Quinn criter.     1.632856
F-statistic               502.5028     Durbin-Watson stat       1.744061
Prob(F-statistic)         0.000000
```

Here the pair BSA and weight has the highest correlation (0.875) among all the other pairs of variables.

Step 2:

Among these variables, the variable with the highest p-value is removed and a new model is estimated.

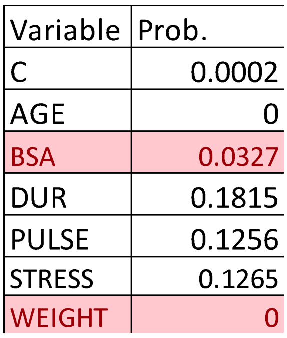
In the model estimated previously we obtained the following p-values:

Here among the variables BSA and Weight, BSA has a higher p-value (0.0327) than BSA (0.000).

Hence we remove the variable BSA and another regression analysis is done to estimate the new model.

On running the regression analysis after removing BSA from the list of independent variables, we obtain the following results:

BP = -15.1167805865 + 0.73194045*AGE + 0.064105*DUR - 0.137443*PULSE + 0.00742915*STRESS + 1.098958*WEIGHT



Variable	Prob.
C	0.0002
AGE	0
BSA	0.0327
DUR	0.1815
PULSE	0.1256
STRESS	0.1265
WEIGHT	0

Here it can be seen that the significance level are comparatively better than the last time. This process can be repeated till majority of the variables are statistically significant.

Summary of All – Multicollinearity – Heteroscedasticity - Autocorrelation

Problem	**Definition**	**Consequences**	**Detection**	**Solution**
High Multicollinearity	Two or more independent variables in a regression model exhibit a close linear relationship.	Large standard errors and insignificant *t*-statistics Coefficient estimates sensitive to minor changes in model specification Nonsensical coefficient signs and magnitudes	Pairwise correlation coefficients Variance inflation factor (VIF)	1. Collect additional data. 2. Re-specify the model. 3. Drop redundant variables.
Heteroscedasticity	The variance of the error term changes in response to a change in the value of the independent variables.	Inefficient coefficient estimates Biased standard errors Unreliable hypothesis tests	Park test Goldfeld-Quandt test Breusch-Pagan test White test	1. Weighted least squares (WLS) 2. Robust standard errors
Autocorrelation	An identifiable relationship (positive or negative) exists between the values of the error in one period and the values of the error in another period.	Inefficient coefficient estimates Biased standard errors Unreliable hypothesis tests	Geary or runs test Durbin-Watson test Breusch-Godfrey test	1. Cochrane-Orcutt transformation 2. Prais-Winsten transformation 3. Newey-West robust standard errors

MULTICOLLINAIRITY

The problem of Multicollinearity occurs in a multiple regression when pair wise correlation exists between independent variables.

Consider F (X1, X2, X3, X4…., Xn) Where Y=dependent variable and Xi….. =independent variables Suppose we find a high degree of correlation in the pair of independent variables i.e. (x1, x2), (x1, x3), (x3, x2) ... are highly correlated in the data set, then there could be a perfect or near Multicollinearity.

Q. The Klein-Goldberger data on the relationship between aggregate US Consumption(C), as a function of wage income (W), nonwage-nonfarm income (p), and farm income (A) for the period 1921-50 is given in the table below. Check if there is problem of multicollinearity present in the data.

C	W	P	A
58.20	39.21	17.73	4.39
62.20	42.31	20.29	4.60
58.60	40.37	18.83	3.25
56.60	39.15	17.44	2.61
51.60	34.00	14.76	1.67
51.10	33.59	13.39	2.44
54.00	36.88	13.93	2.39
57.20	39.27	14.67	5.00
62.80	45.51	17.20	3.93
65.00	46.06	17.15	5.48
63.90	44.16	15.92	4.47
67.50	47.68	17.59	4.51
71.30	50.79	18.49	4.90
76.60	57.78	19.18	6.37
86.30	78.97	19.12	8.42
95.70	73.54	19.76	9.27
98.30	71.92	17.55	8.87
100.30	74.01	19.17	9.30
103.20	75.51	20.20	6.95
108.90	80.97	22.12	7.15

Solution: Here it is given that US Consumption(C) is the independent variable. Thus, the regression model equation can be given as –

$$C = \beta_0 + \beta_1 W + \beta_2 A + \beta_3 P + e$$

Where W = Wage income ; A = Farm Income ; P = Nonwage-Nonfarm income ; e = Error terms

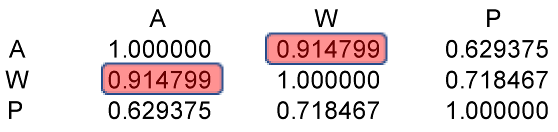
In order to test for multicollinearity, first the correlation matrix of independent variables is obtained.

Steps:

1. Open E-Views software
2. Go to *File -> Open -> Foreign data as Workfile*
3. Select the required data and import
4. Once the data is imported, go to *Quick -> Group Statistics -> Correlations*

The correlation matrix obtained for the current problem is given below –

Output 1



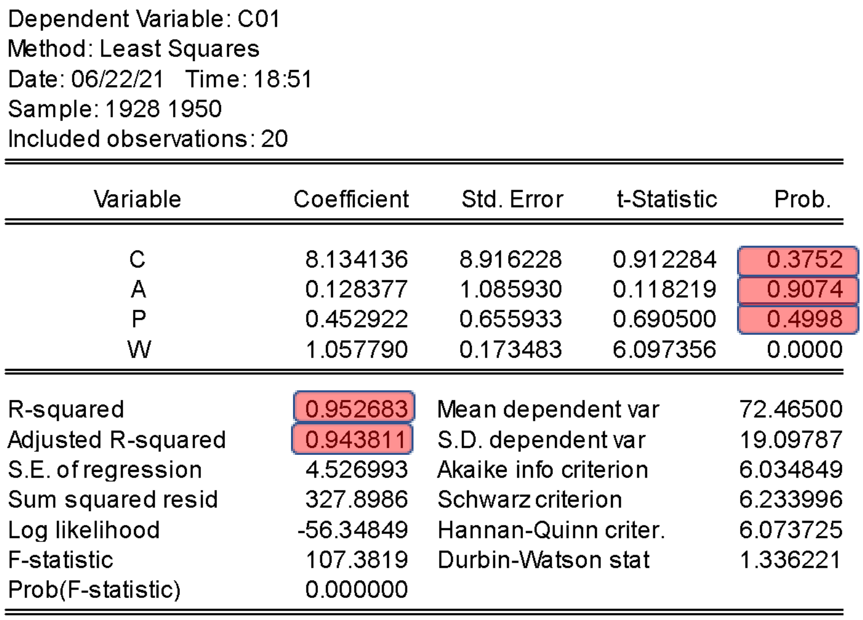
	A	W	P
A	1.000000	0.914799	0.629375
W	0.914799	1.000000	0.718467
P	0.629375	0.718467	1.000000

In the above correlation matrix, it can be observed that Farm income and Wage income are highly correlated. So, there is linear relationship of correlation between the independent variables. Thus, there could be possibility of multicollinearity in the data set. In such a situation, the estimated coefficient in the equation is not expected to possess the characteristic of BLUE.

Next the regression output is obtained for the data.
Steps:

1. Go to *Quick -> Estimate Equation*
2. In the window that pops up, enter the variables of regression - starting with dependent variable then type “C” followed by independent variables.
3. Then click *ok*

Output 2



```
Dependent Variable: C01
Method: Least Squares
Date: 06/22/21   Time: 18:51
Sample: 1928 1950
Included observations: 20
==========================================================================
     Variable        Coefficient    Std. Error    t-Statistic    Prob.
==========================================================================
        C             8.134136      8.916228      0.912284      0.3752
        A             0.128377      1.085930      0.118219      0.9074
        P             0.452922      0.655933      0.690500      0.4998
        W             1.057790      0.173483      6.097356      0.0000
==========================================================================
R-squared              0.952683    Mean dependent var       72.46500
Adjusted R-squared     0.943811    S.D. dependent var       19.09787
S.E. of regression     4.526993    Akaike info criterion    6.034849
Sum squared resid      327.8986    Schwarz criterion        6.233996
Log likelihood        -56.34849    Hannan-Quinn criter.     6.073725
F-statistic            107.3819    Durbin-Watson stat       1.336221
Prob(F-statistic)      0.000000
==========================================================================
```

Thus, the regression equation can be written as –

Consumption = 8.134 + 0.128 *(Farm Income) + 0.4529 *(Nonfarm-Nonwage Income) + 1.0577 *(Wage Income)

In the regression output obtained the intercept, Farm Income(A) and Nonfarm-Nonwage Income(P) are insignificant as they have prob value less than 0.05 at that specified level of significance but the R^2 (Goodness of fit) and adjusted R^2 value is too high, around 0.95 and 0.94 respectively. Thus, we can confirm that there is presence of multi collinearity in the dataset. In such a situation, the estimated coefficient in the equation is not expected to possess the characteristic of BLUE.

1) The GDP in India is increasing gradually over the years. It is hypothesized that food grains production, foreign direct investment(FDI), Exports, and Import may influence

the growth in GDP. Data for the Indian economy on these variables have been collected for 1982-2006 and an exercise is being conducted to find whether these variables or groups out of these are explaining the variation in GDP in India. A multiple regression analysis is being conducted to test the hypothesis. Since the exercise uses time series data, the problem of **Multicollinearity** may be addresses and a decision is taken on an appropriate model. **(Source: BRM Prahlad Mishra, Page: 421)**

Description of the variable is as follows

Dependent Variable (Y) = GDP(₹ in billion), Independent Variables= Total food grains

(in million tons), FDI(₹ in crore), Exports (₹ in crore), Imports(₹ in crore)

S #	**Years**	**Total food grains (in million tons)**	**FDI(₹ in crore)**	**Exports (₹ in crore)**	**Imports (₹ in crore)**	**GDP (₹ in billion)**
1	1981-1982	133.3	91	7,805.90	13,607.60	190
2	1982-1983	129.52	98	8,803.40	14.292.7	198
3	1983-1984	152.37	103	9,770.70	15,831.50	215
4	1984-1985	145.54	109	11,743.70	17,134.20	210
5	1985-1986	150.44	116	10,894.60	19,657.70	230
6	1986-1987	143.42	173	12,452.00	20,095.80	246
7	1987-1988	140.35	196	15,673.70	22,243.70	276
8	1988-1989	169.92	205	20,231.50	28,235.20	293
9	1989-1990	171.04	289	27,658.40	35,328.40	293
10	1990-1991	176.39	313	32,557.60	43,193.00	317
11	1991-1992	168.38	409	44,041.80	47,851.00	268
12	1992-1993	179.48	1,094	53,688.30	63,375.00	246
13	1993-1994	184.26	2,018	69,674.40	73,101.00	276
14	1994-1195	191.5	4,312	82,674.10	89,971.00	324
15	1995-1996	180.42	6,916	1,06,353.30	1,22,678.00	356
16	1996-1997	199.43	9,654	1,18,817.10	1,38,920.00	388
17	1997-1998	193.12	13,548	1,30,100.60	1,54,176.00	411
18	1998-1999	203.61	12,343	1,39,753.10	1,78,332.00	416
19	1999-2000	209.8	10,311	1,59,561.40	2,15,236.00	450
20	2000-2001	196.81	12,645	2,03,571.00	2,30,873.00	460

21	2001-2002	212.85	19,361	2,09,018.00	2,45,200.00	478
22	2002-2003	174.78	14,932	2,55,137.30	2,97,206.00	507
23	2003-2004	213.19	12,117	2,93,366.80	3,59,108.00	599
24	2004-2005	198.36	17,138	3,75,339.50	5,01,065.00	701
25	2005-2006	208.59	24,613	4,56,417.90	6,60,409.00	810
26	2006-2007	217.28	70,630	5,71,779.30	8,40,506.00	915
27	2007-2008	230.78	98,664	6,55,863.50	10,12,312.00	1,180
28	2008-2009	233.88	1,23,025	7,66,935.00	13,74,436.00	1,160