International Journal of Multidisciplinary Research and Modern Education (IJMRME)
Impact Factor: 7.315, ISSN (Online): 2454 - 6119
(www.rdmodernresearch.org) Volume 5, Issue 1, 2019

PROBLEM OF AUTOCORRELATION IN LINEAR REGRESSION DETECTION AND REMEDIES

Ashis Kr. Mukherjee* & Moumi Laha**
* Department of Economics, Nistarini College, Purulia, West Bengal
** Student of J.K. College, (Economics Hons), Purulia, West Bengal

Cite This Article: Ashis Kr. Mukherjee & Moumi Laha, "Problem of Autocorrelation in Linear Regression Detection and Remedies", International Journal of Multidisciplinary Research and Modern Education, Volume 5, Issue 1, Page Number 105-110, 2019.

Abstract:

In the classical linear regression model we assume that successive values of the disturbance term are temporarily independent when observations are taken over time. But when this assumption is violated then the problem is known as Autocorrelation. When the disturbance term exhibits serial correlation, the value of the standard error of the parameter estimates are affected and the predictions based on ordinary least square estimates will be inefficient. So we cannot estimate the correct values of the parameters and the estimates are biased. In this study main focus is given on how one can detect the problem of autocorrelation and how the problem can be solved so as to we can estimate the values of the parameters correctly that are best, linear and unbiased. To explain the procedure of detection of autocorrelation and its remedial measure we take an example on indices of real compensation per hour and output per hour in the business sector of the U.S economy for the period 1960-2005. We use Run Test and Durbin-Watson test to detect the problem of autocorrelation, and then explain the procedure how the problem can be solved.

Key Words: Autocorrelation, Run Test, Durbin-Watson Test & Remedial Measure

Concept of Autocorrelation:

In the classical linear regression model we assume that the different error terms are independently distributed. Here Cov (U_iU_j) = E (U_iU_j)=0. This means that successive values of the disturbance term are temporarily independent when observations are taken over time. Now if the above assumption is not satisfied i.e. if the value of the disturbance term in any particular period is correlated with its own preceding value then there is a problem, known as autocorrelation. In the presence of autocorrelation Cov(U_iU_j) = E $(U_iU_j) \neq 0$.

Consequences of Autocorrelation:

When the disturbance term exhibits serial correlation, the value of the standard error of the parameter estimates are affected and the predictions based on ordinary least square estimates will be inefficient, in the sense that they will have a larger variance as compared with predictions based on estimates obtained from other econometric technique. Although the parameter estimates of OLS are statistically unbiased in the sense that their expected value is equal to the true parameter.

Detection of Autocorrelation:

To explain how to detect the problem of autocorrelation we take an example on indices of real compensation per hour (Y) and output per hour(X)in the business sector of the U.S economy for the period 1960-2005.

Year	Y	X	Year	Y	X
1960	60.8	48.9	1983	90.3	83
1961	62.5	50.6	1984	90.7	85.2
1962	64.6	52.9	1985	92	87.1
1963	66.1	55	1986	94.9	89.7
1964	67.7	56.8	1987	95.2	90.1
1965	69.1	58.8	1988	96.5	91.5
1966	71.7	61.2	1989	95	92.4
1967	73.5	62.5	1990	96.2	94.4
1968	76.2	64.7	1991	97.4	95.9
1969	77.3	65	1992	100	100
1970	78.8	66.3	1993	99.7	100.4
1971	80.2	69	1994	99	101.3
1972	82.6	71.2	1995	98.7	101.5
1973	84.3	73.4	1996	99.4	104.5
1974	83.3	72.3	1997	100.5	106.5
1975	84.1	74.8	1998	105.2	109.5
1976	86.4	77.1	1999	108	112.8
1977	87.6	78.5	2000	112	116.1

1978	89.1	79.3	2001	113.5	119.1
1979	89.3	79.3	2002	115.7	124
1980	89.1	79.2	2003	117.7	128.7
1981	89.3	80.8	2004	119	132.7
1982	90.4	80.1	2005	120.2	135.7

Run Test:

To detect the problem of autocorrelation in the above data we can use Run Test. A run is defined to be a succession of one or more identical symbols which are followed and proceeded by a different symbol or no symbol at all. We set two hypotheses namely Null Hypothesis (H_0) and Alternative Hypothesis (H_1).

H_0: There have no autocorrelation problem in the model.

H_1: There have a problem of autocorrelation in the model.

Let n_1 = Number of positive symbols in the residuals.

n_2 = Number of negative symbols in the residuals.

For sample size n >25 the appropriate test statistic is

$$Z=R-E(R) / [SD(R)]$$

Where $E(R) = (2n_1n_2/n_1+n_2) + 1$

and $Var(R) = 2n_1n_2 (2n_1n_2-n_1-n_2) / (/n_1+n_2)^2(n_1+n_2-1)$

We can accept the null hypothesis that there has no autocorrelation problem in the model if the value of test statistic lies inside the range $E(R)-1.96SD(R) < R < E(R)+1.96SD(R)$, otherwise the null hypothesis is rejected.

To detect the presence of autocorrelation in our example we regress Y on X using OLS technique. The regression result is shown in the following table:

	Coefficients	**Standard Error**	**t-Stat**	**P-value**	**Lower 95%**	**Upper 95%**
Intercept	32.741	1.394	23.487	0.000	29.932	35.551
Xt*	0.670	0.016	42.781	0.000	0.639	0.702

Regression Statistics	
Multiple R	0.988
R Square	0.976
Adjusted R Square	0.975
Standard Error	2.384
Observations	46

ANOVA					
	df	**SS**	**MS**	**F**	**Significance F**
Regression	1	10406.15	10406.15	1830.242	0.000
Residual	44	250.169	5.685		
Total	45	10656.31			

Residual Table:

Observations	**Residuals**	**Observations**	**Residuals**	**Observations**	**Residuals**
1	-4.72474	17	1.969818	33	0.217528
2	-4.16443	18	2.23125	34	-0.35063
3	-3.60636	19	3.194926	35	-1.654
4	-3.51422	20	3.394926	36	-2.08808
5	-3.12095	21	3.261966	37	-3.3993
6	-3.06176	22	2.389317	38	-3.64011
7	-2.07073	23	3.958601	39	-0.95133
8	-1.14226	24	1.914424	40	-0.36367
9	0.082849	25	0.839532	41	1.423996
10	0.981727	26	0.865761	42	0.912779
11	1.610199	27	2.022706	43	-0.17221
12	1.200104	28	2.054544	44	-1.32312
13	2.125212	29	2.415976	45	-2.70474
14	2.350319	30	0.312611	46	-3.51596
15	2.087765	31	0.1718		
16	1.211751	32	0.366191		

In our example n_1=27, n_2=19, R=5

$$E(R)= 24.3043, Var(R) = 10.5595, SD(R)= 3.249538.$$

The value of the test statistic is

$$Z= (5-24.3043)/3.249538$$
$$= -5.9406$$

The value of the critical regionis $E(R)-1.96SD(R) < R < E(R)+1.96SD(R)$

$$= (24.3043-1.96x3.249538) < R < (24.3043+1.96x3.249538)$$
$$= 17.9353 < R < 30.6733$$

Since the value of the test statistic (-5.940) lies outside the above range, we can accept the alternative hypothesis that there is a problem of autocorrelation in the data.

Durbin Watson Test:

One of the most popular test on Autocorrelation is Durbin Watson test. The test is used for first order autoregressive scheme.The test statistic is

$$d^*=\sum (e_t-e_{t-1})^2 / \sum e_t^2$$

Where e_t= OLSresidual term.

To test that the problem of autocorrelation exists or not let us set two hypothesis say Null hypothesis and Alternative hypothesis.

The null hypothesis is H_0:(p=0 or d*=2) (i.e., there is no autocorrelation)

The alternative hypothesis is H_1:(p≠ 0 or d*≠2) (i.e., there is aproblem of autocorrelation)

The acceptance-rejection rules in case of the test are as follows:

- If $d^*<d_L$we reject the null hypothesis of no autocorrelation and accept that there is a positive autocorrelation.
- If $d^*> (4-d_L)$ we reject the null hypothesis and conclude that there is a negative autocorrelation problem.
- If $d_u<d^*< (4-d_u)$ we accept the null hypothesis and conclude that there is no autocorrelation problem.
- If $d_L<d^*<d_u$ or if $(4-d_u)<d^*<(4-d_L)$ the test is inconclusive. We cannot say whether autocorrelation problem exist or not.

Now we use Durbin-Watson statistics for the above example. The steps are as follows:

Observations	e_t	e_t^2	e_{t-1}	e_t-e_{t-1}	$(e_t-e_{t-1})^2$	e_te_{t-1}	e_{t-1}^2
1	-4.7247	22.3232					
2	-4.1644	17.3425	-4.7247	0.5603	0.3139	19.6759	22.3232
3	-3.6064	13.0059	-4.1644	0.5581	0.3114	15.0185	17.3425
4	-3.5142	12.3497	-3.6064	0.0921	0.0085	12.6735	13.0059
5	-3.1209	9.74031	-3.5142	0.3933	0.1547	10.9677	12.3497
6	-3.0618	9.37436	-3.1209	0.0592	0.0035	9.55558	9.74031
7	-2.0707	4.28793	-3.0618	0.991	0.9821	6.34008	9.37436
8	-1.1423	1.30476	-2.0707	0.9285	0.8621	2.36531	4.28793
9	0.08285	0.00686	-1.1423	1.2251	1.5009	-0.0946	1.30476
10	0.98173	0.96379	0.08285	0.8989	0.808	0.08133	0.00686
11	1.6102	2.59274	0.98173	0.6285	0.395	1.58078	0.96379
12	1.2001	1.44025	1.6102	-0.41	0.1682	1.93241	2.59274
13	2.12521	4.51652	1.2001	0.9251	0.8558	2.55048	1.44025
14	2.35032	5.524	2.12521	0.2251	0.0507	4.99493	4.51652
15	2.08777	4.35876	2.35032	-0.263	0.0689	4.90691	5.524
16	1.21175	1.46834	2.08777	-0.876	0.7674	2.52985	4.35876
17	1.96982	3.88018	1.21175	0.7581	0.5747	2.38693	1.46834
18	2.23125	4.97848	1.96982	0.2614	0.0683	4.39516	3.88018
19	3.19493	10.2075	2.23125	0.9637	0.9287	7.12868	4.97848
20	3.39493	11.5255	3.19493	0.2	0.04	10.8465	10.2075
21	3.26197	10.6404	3.39493	-0.133	0.0177	11.0741	11.5255
22	2.38932	5.70884	3.26197	-0.873	0.7615	7.79387	10.6404
23	3.9586	15.6705	2.38932	1.5693	2.4627	9.45835	5.70884
24	1.91442	3.66502	3.9586	-2.044	4.1787	7.57844	15.6705
25	0.83953	0.70481	1.91442	-1.075	1.1554	1.60722	3.66502
26	0.86576	0.74954	0.83953	0.0262	0.0007	0.72683	0.70481
27	2.02271	4.09134	0.86576	1.1569	1.3385	1.75118	0.74954
28	2.05454	4.22115	2.02271	0.0318	0.001	4.15574	4.09134
29	2.41598	5.83694	2.05454	0.3614	0.1306	4.96373	4.22115
30	0.31261	0.09773	2.41598	-2.103	4.4241	0.75526	5.83694
31	0.1718	0.02952	0.31261	-0.141	0.0198	0.05371	0.09773

32	0.36619	0.1341	0.1718	0.1944	0.0378	0.06291	0.02952
33	0.21753	0.04732	0.36619	-0.149	0.0221	0.07966	0.1341
34	-0.3506	0.12294	0.21753	-0.568	0.3228	-0.0763	0.04732
35	-1.654	2.73571	-0.3506	-1.303	1.6988	0.57995	0.12294
36	-2.0881	4.36008	-1.654	-0.434	0.1884	3.45368	2.73571
37	-3.3993	11.5552	-2.0881	-1.311	1.7193	7.09801	4.36008
38	-3.6401	13.2504	-3.3993	-0.241	0.058	12.3738	11.5552
39	-0.9513	0.90502	-3.6401	2.6888	7.2296	3.46293	13.2504
40	-0.3637	0.13225	-0.9513	0.5877	0.3453	0.34596	0.90502
41	1.424	2.02777	-0.3637	1.7877	3.1957	-0.5179	0.13225
42	0.91278	0.83317	1.424	-0.511	0.2613	1.29979	2.02777
43	-0.1722	0.02966	0.91278	-1.085	1.1772	-0.1572	0.83317
44	-1.3231	1.75063	-0.1722	-1.151	1.3246	0.22785	0.02966
45	-2.7047	7.31561	-1.3231	-1.382	1.9089	3.57868	1.75063
46	-3.516	12.3619	-2.7047	-0.811	0.6581	9.50974	7.31561
Total		**250**			**43.5**	**211**	**238**

The value of Durbin-Watson statistics,

$$d^* = \sum (e_t - e_{t-1})^2 / \sum e_t^2$$

or, $d^* = 43.5/250$

or, $d^* = 0.174$

From the Durbin-Watson table at 5% level of significance and k=1 we get $d_L = 1.41$ and $d_u = 1.57$ Since $d^* < d_L$, we reject the null hypothesis of no autocorrelation and conclude that there is a positive autocorrelation exists in the data.When the disturbance term exhibits serial correlation the value as well as the standard errors of the parameter estimates are affected, and the predictions based on ordinary least square estimates will be inefficient, in the sense that they will have a larger variance as compared with predictions based on estimates obtained from other econometric technique.

Remedial Measure for Autocorrelation:

To remove the problem of autocorrelation from the data we have to transfer the original data and then apply OLS technique to estimate the parameters. At first we have to calculate autocorrelation coefficient (p). p can be calculated by using the formula

$$p = \sum e_t e_{t-1} / \sum e_{t-1}^2$$

Consider our original regression model

$$Y_t = \alpha + \beta X_t + U_t \quad \text{.....(1)}$$

where U_t follows first order autoregressive scheme, $U_t = pUt_{-1} + €_t$

$€_t$ is white noise term satisfy all the classical assumptions. Here /p/ <1

Now from (1) taking one period lag we get

$$Y_{t-1} = \alpha + \beta X_{t-1} + U_{t-1} \quad \text{.....(2)}$$

Multiplying (2) by p we have

$$pY_{t-1} = p\alpha + p\beta X_{t-1} + pU_{t-1} \quad \text{.....(3)}$$

Subtracting (3) from (1) we get

$$(Y_t - pY_{t-1}) = (\alpha - p\alpha) + (\beta X_t - p\beta X_{t-1}) + (U_t - pU_{t-1})$$

Or, $(Y_t - pY_{t-1}) = \alpha(1-p) + \beta(X_t - pX_{t-1}) + (U_t - pU_{t-1})$

Or, $Y_t^* = \alpha^* + \beta X_t^* + €_t$

Where $Y_t^* = (Y_t - pY_{t-1})$

$\alpha^* = \alpha(1-p)$

$X_t^* = (X_t - pX_{t-1})$

$€_t = (U_t - pU_{t-1})$

Now we can apply OLS to this transformed relation to obtain the estimated values of the parameters.

In our example,

$$p = \sum e_t e_{t-1} / \sum e_{t-1}^2 = 211/238 = 0.89$$

The data transformation procedure can be explained with the help of the following table:

Year	**Y_t**	**X**	**$.89Y_t$**	**$.89Y_{t-1}$**	**$Y_t^* = Y_t - 0.89Y_{t-1}$**	**$.89X_t$**	**$.89X_{t-1}$**	**$X_t^* = X_t - 0.89X_{t-1}$**
1960	60.8	48.9	54.112			43.521		
1961	62.5	50.6	55.625	54.112	8.388	45.034	43.521	7.079
1962	64.6	52.9	57.494	55.625	8.975	47.081	45.034	7.866
1963	66.1	55	58.829	57.494	8.606	48.95	47.081	7.919
1964	67.7	56.8	60.253	58.829	8.871	50.552	48.95	7.85
1965	69.1	58.8	61.499	60.253	8.847	52.332	50.552	8.248
1966	71.7	61.2	63.813	61.499	10.201	54.468	52.332	8.868

1967	73.5	62.5	65.415	63.813	9.687	55.625	54.468	8.032
1968	76.2	64.7	67.818	65.415	10.785	57.583	55.625	9.075
1969	77.3	65	68.797	67.818	9.482	57.85	57.583	7.417
1970	78.8	66.3	70.132	68.797	10.003	59.007	57.85	8.45
1971	80.2	69	71.378	70.132	10.068	61.41	59.007	9.993
1972	82.6	71.2	73.514	71.378	11.222	63.368	61.41	9.79
1973	84.3	73.4	75.027	73.514	10.786	65.326	63.368	10.032
1974	83.3	72.3	74.137	75.027	8.273	64.347	65.326	6.974
1975	84.1	74.8	74.849	74.137	9.963	66.572	64.347	10.453
1976	86.4	77.1	76.896	74.849	11.551	68.619	66.572	10.528
1977	87.6	78.5	77.964	76.896	10.704	69.865	68.619	9.881
1978	89.1	79.3	79.299	77.964	11.136	70.577	69.865	9.435
1979	89.3	79.3	79.477	79.299	10.001	70.577	70.577	8.723
1980	89.1	79.2	79.299	79.477	9.623	70.488	70.577	8.623
1981	89.3	80.8	79.477	79.299	10.001	71.912	70.488	10.312
1982	90.4	80.1	80.456	79.477	10.923	71.289	71.912	8.188
1983	90.3	83	80.367	80.456	9.844	73.87	71.289	11.711
1984	90.7	85.2	80.723	80.367	10.333	75.828	73.87	11.33
1985	92	87.1	81.88	80.723	11.277	77.519	75.828	11.272
1986	94.9	89.7	84.461	81.88	13.02	79.833	77.519	12.181
1987	95.2	90.1	84.728	84.461	10.739	80.189	79.833	10.267
1988	96.5	91.5	85.885	84.728	11.772	81.435	80.189	11.311
1989	95	92.4	84.55	85.885	9.115	82.236	81.435	10.965
1990	96.2	94.4	85.618	84.55	11.65	84.016	82.236	12.164
1991	97.4	95.9	86.686	85.618	11.782	85.351	84.016	11.884
1992	100	100	89	86.686	13.314	89	85.351	14.649
1993	99.7	100.4	88.733	89	10.7	89.356	89	11.4
1994	99	101.3	88.11	88.733	10.267	90.157	89.356	11.944
1995	98.7	101.5	87.843	88.11	10.59	90.335	90.157	11.343
1996	99.4	104.5	88.466	87.843	11.557	93.005	90.335	14.165
1997	100.5	106.5	89.445	88.466	12.034	94.785	93.005	13.495
1998	105.2	109.5	93.628	89.445	15.755	97.455	94.785	14.715
1999	108	112.8	96.12	93.628	14.372	100.392	97.455	15.345
2000	112	116.1	99.68	96.12	15.88	103.329	100.392	15.708
2001	113.5	119.1	101.015	99.68	13.82	105.999	103.329	15.771
2002	115.7	124	102.973	101.015	14.685	110.36	105.999	18.001
2003	117.7	128.7	104.753	102.973	14.727	114.543	110.36	18.34
2004	119	132.7	105.91	104.753	14.247	118.103	114.543	18.157
2005	120.2	135.7	106.978	105.91	14.29	120.773	118.103	17.597

Using OLS technique we regress Y_t^* on X_t^* to get the values of the parameters. The regression result are as follows;

	Coefficients	**Standard Error**	**t-Stat**	**P-value**	**Lower 95%**	**Upper 95%**
Intercept	4.803	0.509	9.434	0.000	3.776	5.830
Xt*	0.567	0.044	13.029	0.000	0.479	0.654

Regression Statistics	
Multiple R	0.893
R Square	0.797
Adjusted R Square	0.793
Standard Error	0.912
Observations	45

ANOVA					
	df	**SS**	**MS**	**F**	**Significance F**
Regression	1	141.125	141.125	169.773	0.000
Residual	43	35.744	0.831		
Total	44	176.869			

From the above result the estimated equation can be written as

$$Y_t^* = \alpha^* + \beta X_t^*$$

Or, $Y_t^* = 4.803 + 0.567 X_t^*$

But here $\alpha^* = \alpha(1-p)$

Or, $4.803 = \alpha(1-0.89)$

Or, $\alpha = 43.66$

Hence the original estimated equation is $Y_t = 43.66 + 0.567 X_t$

Durbin Two Step Method of Estimation of p:

Durbin has suggested the following steps for estimating p which is applicable for any order of autoregressive scheme.

Our original regression model is $Y_t = \alpha + \beta X_t + U_t$ .....(1)

where U_t follows first order autoregressive scheme, $U_t = pU_{t-1} + \epsilon_t$

ϵ_t is white noise term satisfy all the classical assumptions. Here /p/ <1

Now from (1), taking one period lag we get

$$Y_{t-1} = \alpha + \beta X_{t-1} + U_{t-1} \quad \text{.....(2)}$$

Multiplying (2) by p

$$pY_{t-1} = p\alpha + p\beta X_{t-1} + pU_{t-1} \quad \text{.....(3)}$$

Subtracting (3) from (1) we get

$$(Y_t - pY_{t-1}) = (\alpha - p\alpha) + (\beta X_t - p\beta X_{t-1}) + (U_t - pU_{t-1})$$

Or, $(Y_t - pY_{t-1}) = \alpha(1-p) + \beta(X_t - pX_{t-1}) + (U_t - pU_{t-1})$

Or, $Y_t = \alpha(1-p) + pY_{t-1} + \beta(X_t - pX_{t-1}) + (U_t - pU_{t-1})$

Or, $Y_t = \alpha_0 + pY_{t-1} + \beta X_t - p\beta X_{t-1} + V_t$.....(4)

Where $\alpha_0 = \alpha(1-p)$

Now consider the above example. Regressing Y_t on Y_{t-1}, Xt, and X_{t-1} we get,

$$Y_t = 5.156 + 0.879 Y_{t-1} + 0.567 X_t - 0.498 X_{t-1} \text{ .....(5)}$$

S.E (2.14) (0.12) (0.06) (0.15)

p-value (0.00) (0.00) (0.00) (0.00)

$$R^2 = 0.99$$

Comparing (4) and (5) we can say that p= 0.879.

Now we transform the data as follows

$$Y_t^* = (Y_t - pY_{t-1}) \text{ and } X_t^* = (X_t - pX_{t-1})$$

Applying OLSto the above transformed data we can get the estimated values of the parameter.

Conclusion:

When the successive values of the disturbance term are temporarily dependent then there is a problem known as autocorrelation. Autocorrelation can arise several reasons such as inertia or sluggishness of economic time series, specification bias resulting from excluding important variables from the model or due to incorrect functional form. Although in the presence of autocorrelation OLS estimators are unbiased, consistent and asymptotically normally distributed, they are no longer efficient. To make the estimators efficient we use different remedial measures such as Generalized Least Square, First Difference Method, Durbin two Step Method etc. After transforming the data, we can use OLS technique to get the values of the parameters, and these parameters are Best Linear and Unbiased (BLUE).

References:

1. Koutsoyiannis A. Theory of Econometrics, ELBS with Macmillan.
2. MadnaniG.M.K, Introductionto Econometrics, Oxford &IBH Publishingco.pvt.ltd
3. Damodar N. Gujarati, Basic Econometrics, Mc Graw Hill Education Private Limited.
4. Croxton and Cowden, Applied General Statistics (Prentice Hall, Inc , 1964)
5. N. G. Das, Statistical Method, M. Das & Co.
6. Goon Gupta and Dasgupta, Fundamentals of Statistics, The World Press.