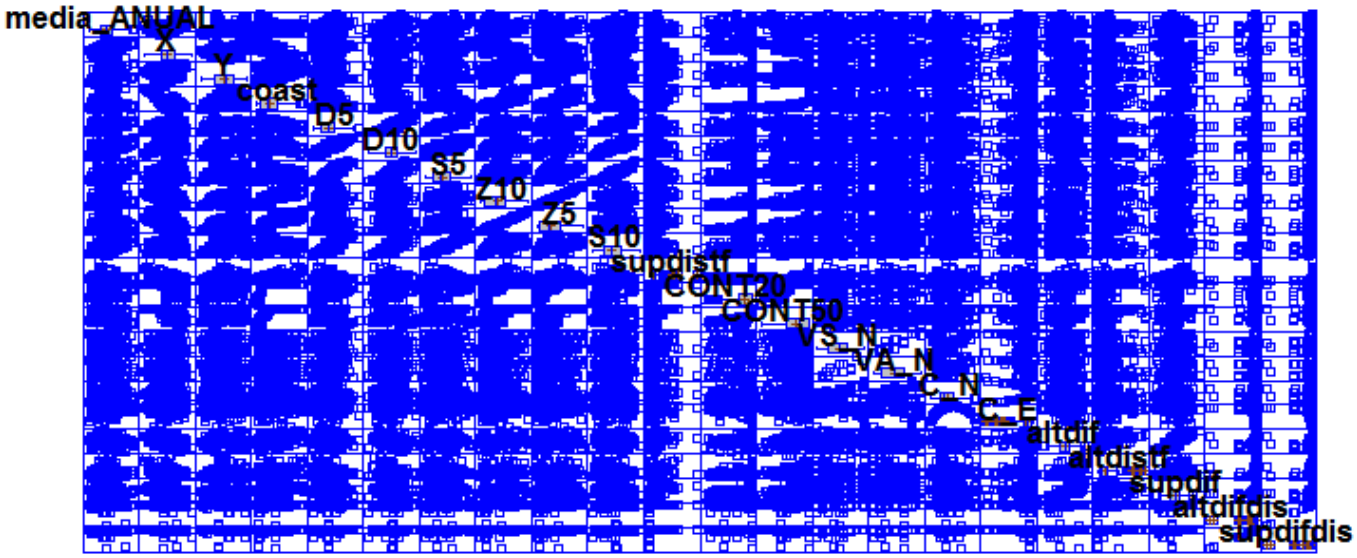
Practical case 2: Regression models to predict annual mean rainfall

1) ANALYSIS OF RELATIONSHIP BETWEEN THE DIFFERENT VARIABLES

- Multiple variable analysis



Correlations

	media_ANUAL	X	Y	coast	D5	D10	S5	Z10
media_ANUAL		0,4215	0,3875	-0,0645	0,3511	0,3441	0,4312	0,0974
		(224)	(224)	(224)	(224)	(224)	(224)	(224)
		0,0000	0,0000	0,3369	0,0000	0,0000	0,0000	0,1461
X	0,4215		0,2014	-0,7981	0,1211	0,1530	0,0664	-0,5178
	(224)		(224)	(224)	(224)	(224)	(224)	(224)
	0,0000		0,0025	0,0000	0,0704	0,0220	0,3223	0,0000
Y	0,3875	0,2014		0,3063	0,0341	0,1279	0,2318	0,4114
	(224)	(224)		(224)	(224)	(224)	(224)	(224)
	0,0000	0,0025		0,0000	0,6122	0,0560	0,0005	0,0000
coast	-0,0645	-0,7981	0,3063		0,0650	0,0487	0,2078	0,8263
	(224)	(224)	(224)		(224)	(224)	(224)	(224)
	0,3369	0,0000	0,0000		0,3327	0,4683	0,0018	0,0000
D5	0,3511	0,1211	0,0341	0,0650		0,8720	0,8841	0,4169
	(224)	(224)	(224)	(224)		(224)	(224)	(224)
	0,0000	0,0704	0,6122	0,3327		0,0000	0,0000	0,0000
D10	0,3441	0,1530	0,1279	0,0487	0,8720		0,7963	0,4282
	(224)	(224)	(224)	(224)	(224)		(224)	(224)
	0,0000	0,0220	0,0560	0,4683	0,0000		0,0000	0,0000
S5	0,4312	0,0664	0,2318	0,2078	0,8841	0,7963		0,5442
	(224)	(224)	(224)	(224)	(224)	(224)		(224)
	0,0000	0,3223	0,0005	0,0018	0,0000	0,0000		0,0000
Z10	0,0974	-0,5178	0,4114	0,8263	0,4169	0,4282	0,5442	
	(224)	(224)	(224)	(224)	(224)	(224)	(224)	
	0,1461	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	
Z5	0,1042	-0,5248	0,3988	0,8251	0,4123	0,4076	0,5383	0,9948
	(224)	(224)	(224)	(224)	(224)	(224)	(224)	(224)
	0,1198	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
S10	0,4536	0,1111	0,2927	0,2015	0,8637	0,8557	0,9495	0,5716
	(224)	(224)	(224)	(224)	(224)	(224)	(224)	(224)
	0,0000	0,0973	0,0000	0,0025	0,0000	0,0000	0,0000	0,0000
supdistf	-0,0198	-0,3972	0,0404	0,5378	-0,0174	-0,0612	0,0195	0,3217
	(224)	(224)	(224)	(224)	(224)	(224)	(224)	(224)
	0,7680	0,0000	0,5472	0,0000	0,7960	0,3620	0,7719	0,0000
CONT20	-0,0379	-0,5153	0,1649	0,6188	0,2932	0,3594	0,3743	0,6111
	(224)	(224)	(224)	(224)	(224)	(224)	(224)	(224)
	0,5729	0,0000	0,0135	0,0000	0,0000	0,0000	0,0000	0,0000

CONT50	-0,2021	-0,6847	0,2866	0,8008	0,0767	0,1287	0,2093	0,7294
	(224)	(224)	(224)	(224)	(224)	(224)	(224)	(224)
	0,0024	0,0000	0,0000	0,0000	0,2528	0,0544	0,0016	0,0000
VS_N	-0,0417	-0,1029	-0,1355	0,0310	0,0129	-0,0062	0,0188	0,0426
	(224)	(224)	(224)	(224)	(224)	(224)	(224)	(224)
	0,5347	0,1246	0,0427	0,6448	0,8481	0,9262	0,7801	0,5263
VA_N	-0,0027	-0,0987	-0,1475	0,0292	0,0991	0,0913	0,0691	0,0674
	(224)	(224)	(224)	(224)	(224)	(224)	(224)	(224)
	0,9675	0,1410	0,0273	0,6642	0,1394	0,1732	0,3034	0,3150
C_N	-0,0505	-0,0970	-0,0273	0,0716	-0,0956	-0,0408	-0,0532	-0,0137
	(224)	(224)	(224)	(224)	(224)	(224)	(224)	(224)
	0,4522	0,1477	0,6844	0,2862	0,1538	0,5439	0,4283	0,8389
C_E	0,0037	0,0501	0,0415	-0,0402	-0,0464	0,0100	-0,0820	-0,0285
	(224)	(224)	(224)	(224)	(224)	(224)	(224)	(224)
	0,9560	0,4558	0,5367	0,5497	0,4894	0,8821	0,2216	0,6712
altdif	-0,0590	-0,2766	0,0159	0,3022	-0,2211	-0,3161	-0,1205	0,1088
	(224)	(224)	(224)	(224)	(224)	(224)	(224)	(224)
	0,3794	0,0000	0,8124	0,0000	0,0009	0,0000	0,0719	0,1042
altdistf	0,0697	-0,2834	0,0699	0,3663	-0,1704	-0,2391	-0,0793	0,1322
	(224)	(224)	(224)	(224)	(224)	(224)	(224)	(224)
	0,2992	0,0000	0,2976	0,0000	0,0106	0,0003	0,2374	0,0482
supdif	-0,1177	-0,3170	-0,0113	0,3513	-0,1322	-0,2168	-0,0702	0,1859
	(224)	(224)	(224)	(224)	(224)	(224)	(224)	(224)
	0,0788	0,0000	0,8660	0,0000	0,0482	0,0011	0,2957	0,0053
altdifdis	-0,0629	-0,1297	-0,0217	0,1331	0,0361	0,0179	0,0254	0,1241
	(224)	(224)	(224)	(224)	(224)	(224)	(224)	(224)
	0,3489	0,0526	0,7470	0,0466	0,5908	0,7900	0,7059	0,0638
supdifdis	-0,0629	-0,1297	-0,0217	0,1331	0,0361	0,0179	0,0254	0,1241
	(224)	(224)	(224)	(224)	(224)	(224)	(224)	(224)
	0,3489	0,0526	0,7470	0,0466	0,5908	0,7900	0,7059	0,0638

	Z5	S10	supdistf	CONT20	CONT50	VS_N	VA_N	C_N
media_ANUAL	0,1042	0,4536	-0,0198	-0,0379	-0,2021	-0,0417	-0,0027	-0,0505
	(224)	(224)	(224)	(224)	(224)	(224)	(224)	(224)
	0,1198	0,0000	0,7680	0,5729	0,0024	0,5347	0,9675	0,4522
X	-0,5248	0,1111	-0,3972	-0,5153	-0,6847	-0,1029	-0,0987	-0,0970
	(224)	(224)	(224)	(224)	(224)	(224)	(224)	(224)
	0,0000	0,0973	0,0000	0,0000	0,0000	0,1246	0,1410	0,1477
Y	0,3988	0,2927	0,0404	0,1649	0,2866	-0,1355	-0,1475	-0,0273
	(224)	(224)	(224)	(224)	(224)	(224)	(224)	(224)
	0,0000	0,0000	0,5472	0,0135	0,0000	0,0427	0,0273	0,6844
coast	0,8251	0,2015	0,5378	0,6188	0,8008	0,0310	0,0292	0,0716
	(224)	(224)	(224)	(224)	(224)	(224)	(224)	(224)
	0,0000	0,0025	0,0000	0,0000	0,0000	0,6448	0,6642	0,2862
D5	0,4123	0,8637	-0,0174	0,2932	0,0767	0,0129	0,0991	-0,0956
	(224)	(224)	(224)	(224)	(224)	(224)	(224)	(224)
	0,0000	0,0000	0,7960	0,0000	0,2528	0,8481	0,1394	0,1538
D10	0,4076	0,8557	-0,0612	0,3594	0,1287	-0,0062	0,0913	-0,0408
	(224)	(224)	(224)	(224)	(224)	(224)	(224)	(224)
	0,0000	0,0000	0,3620	0,0000	0,0544	0,9262	0,1732	0,5439
S5	0,5383	0,9495	0,0195	0,3743	0,2093	0,0188	0,0691	-0,0532
	(224)	(224)	(224)	(224)	(224)	(224)	(224)	(224)
	0,0000	0,0000	0,7719	0,0000	0,0016	0,7801	0,3034	0,4283
Z10	0,9948	0,5716	0,3217	0,6111	0,7294	0,0426	0,0674	-0,0137
	(224)	(224)	(224)	(224)	(224)	(224)	(224)	(224)
	0,0000	0,0000	0,0000	0,0000	0,0000	0,5263	0,3150	0,8389
Z5		0,5531	0,3121	0,6064	0,7244	0,0488	0,0740	-0,0165
		(224)	(224)	(224)	(224)	(224)	(224)	(224)
		0,0000	0,0000	0,0000	0,0000	0,4670	0,2703	0,8057
S10	0,5531		0,0214	0,3893	0,2206	0,0225	0,0662	-0,0378
	(224)		(224)	(224)	(224)	(224)	(224)	(224)
	0,0000		0,7506	0,0000	0,0009	0,7377	0,3238	0,5738
supdistf	0,3121	0,0214		0,2936	0,3457	0,0584	-0,0051	0,0896
	(224)	(224)		(224)	(224)	(224)	(224)	(224)
	0,0000	0,7506		0,0000	0,0000	0,3843	0,9394	0,1816
CONT20	0,6064	0,3893	0,2936		0,8459	-0,0098	0,0391	0,0260
	(224)	(224)	(224)		(224)	(224)	(224)	(224)
	0,0000	0,0000	0,0000		0,0000	0,8844	0,5604	0,6986
CONT50	0,7244	0,2206	0,3457	0,8459		-0,0058	0,0106	0,0548

	(224)	(224)	(224)	(224)		(224)	(224)	(224)
	0,0000	0,0009	0,0000	0,0000		0,9315	0,8745	0,4146
VS_N	0,0488	0,0225	0,0584	-0,0098	-0,0058		0,8627	0,4500
	(224)	(224)	(224)	(224)	(224)		(224)	(224)
	0,4670	0,7377	0,3843	0,8844	0,9315		0,0000	0,0000
VA_N	0,0740	0,0662	-0,0051	0,0391	0,0106	0,8627		0,3296
	(224)	(224)	(224)	(224)	(224)	(224)		(224)
	0,2703	0,3238	0,9394	0,5604	0,8745	0,0000		0,0000
C_N	-0,0165	-0,0378	0,0896	0,0260	0,0548	0,4500	0,3296	
	(224)	(224)	(224)	(224)	(224)	(224)	(224)	
	0,8057	0,5738	0,1816	0,6986	0,4146	0,0000	0,0000	
C_E	-0,0231	-0,0450	-0,0235	-0,0663	-0,0591	-0,3145	-0,2458	0,1088
	(224)	(224)	(224)	(224)	(224)	(224)	(224)	(224)
	0,7305	0,5025	0,7266	0,3229	0,3786	0,0000	0,0002	0,1045
altdif	0,1023	-0,1650	0,2806	0,2218	0,2943	0,0163	0,0030	-0,0127
	(224)	(224)	(224)	(224)	(224)	(224)	(224)	(224)
	0,1267	0,0134	0,0000	0,0008	0,0000	0,8083	0,9648	0,8499
altdistf	0,1359	-0,1173	0,3779	0,2600	0,2838	-0,0722	-0,1190	0,0389
	(224)	(224)	(224)	(224)	(224)	(224)	(224)	(224)
	0,0421	0,0799	0,0000	0,0001	0,0000	0,2817	0,0756	0,5626
supdif	0,1751	-0,0946	0,3863	0,2438	0,3111	0,0506	0,0401	-0,0226
	(224)	(224)	(224)	(224)	(224)	(224)	(224)	(224)
	0,0086	0,1583	0,0000	0,0002	0,0000	0,4509	0,5509	0,7363
altdifdis	0,1147	0,0367	0,2168	0,1700	0,0959	0,0755	0,0347	-0,0420
	(224)	(224)	(224)	(224)	(224)	(224)	(224)	(224)
	0,0868	0,5844	0,0011	0,0108	0,1525	0,2602	0,6057	0,5313
supdifdis	0,1147	0,0367	0,2168	0,1700	0,0959	0,0755	0,0347	-0,0420
	(224)	(224)	(224)	(224)	(224)	(224)	(224)	(224)
	0,0868	0,5844	0,0011	0,0108	0,1525	0,2602	0,6057	0,5313

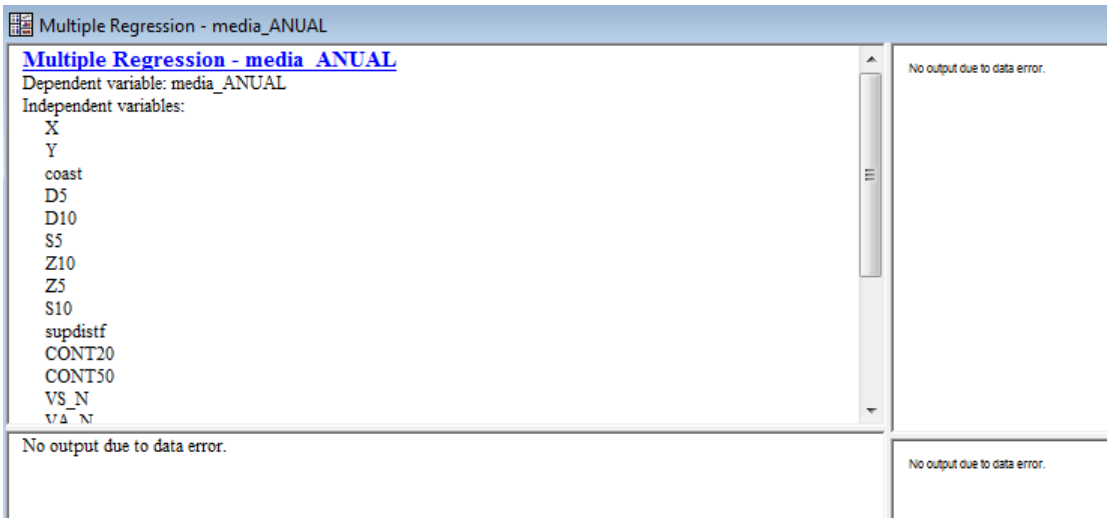
	C_E	altdif	altdistf	supdif	altdifdis	supdifdis
media_ANUAL	0,0037	-0,0590	0,0697	-0,1177	-0,0629	-0,0629
	(224)	(224)	(224)	(224)	(224)	(224)
	0,9560	0,3794	0,2992	0,0788	0,3489	0,3489
X	0,0501	-0,2766	-0,2834	-0,3170	-0,1297	-0,1297
	(224)	(224)	(224)	(224)	(224)	(224)
	0,4558	0,0000	0,0000	0,0000	0,0526	0,0526
Y	0,0415	0,0159	0,0699	-0,0113	-0,0217	-0,0217
	(224)	(224)	(224)	(224)	(224)	(224)
	0,5367	0,8124	0,2976	0,8660	0,7470	0,7470
coast	-0,0402	0,3022	0,3663	0,3513	0,1331	0,1331
	(224)	(224)	(224)	(224)	(224)	(224)
	0,5497	0,0000	0,0000	0,0000	0,0466	0,0466
D5	-0,0464	-0,2211	-0,1704	-0,1322	0,0361	0,0361
	(224)	(224)	(224)	(224)	(224)	(224)
	0,4894	0,0009	0,0106	0,0482	0,5908	0,5908
D10	0,0100	-0,3161	-0,2391	-0,2168	0,0179	0,0179
	(224)	(224)	(224)	(224)	(224)	(224)
	0,8821	0,0000	0,0003	0,0011	0,7900	0,7900
S5	-0,0820	-0,1205	-0,0793	-0,0702	0,0254	0,0254
	(224)	(224)	(224)	(224)	(224)	(224)
	0,2216	0,0719	0,2374	0,2957	0,7059	0,7059
Z10	-0,0285	0,1088	0,1322	0,1859	0,1241	0,1241
	(224)	(224)	(224)	(224)	(224)	(224)
	0,6712	0,1042	0,0482	0,0053	0,0638	0,0638
Z5	-0,0231	0,1023	0,1359	0,1751	0,1147	0,1147
	(224)	(224)	(224)	(224)	(224)	(224)
	0,7305	0,1267	0,0421	0,0086	0,0868	0,0868
S10	-0,0450	-0,1650	-0,1173	-0,0946	0,0367	0,0367
	(224)	(224)	(224)	(224)	(224)	(224)
	0,5025	0,0134	0,0799	0,1583	0,5844	0,5844
supdistf	-0,0235	0,2806	0,3779	0,3863	0,2168	0,2168
	(224)	(224)	(224)	(224)	(224)	(224)
	0,7266	0,0000	0,0000	0,0000	0,0011	0,0011
CONT20	-0,0663	0,2218	0,2600	0,2438	0,1700	0,1700
	(224)	(224)	(224)	(224)	(224)	(224)
	0,3229	0,0008	0,0001	0,0002	0,0108	0,0108
CONT50	-0,0591	0,2943	0,2838	0,3111	0,0959	0,0959
	(224)	(224)	(224)	(224)	(224)	(224)

	0,3786	0,0000	0,0000	0,0000	0,1525	0,1525
VS_N	-0,3145	0,0163	-0,0722	0,0506	0,0755	0,0755
	(224)	(224)	(224)	(224)	(224)	(224)
	0,0000	0,8083	0,2817	0,4509	0,2602	0,2602
VA_N	-0,2458	0,0030	-0,1190	0,0401	0,0347	0,0347
	(224)	(224)	(224)	(224)	(224)	(224)
	0,0002	0,9648	0,0756	0,5509	0,6057	0,6057
C_N	0,1088	-0,0127	0,0389	-0,0226	-0,0420	-0,0420
	(224)	(224)	(224)	(224)	(224)	(224)
	0,1045	0,8499	0,5626	0,7363	0,5313	0,5313
C_E		-0,0943	0,0130	-0,0552	-0,0811	-0,0811
		(224)	(224)	(224)	(224)	(224)
		0,1594	0,8471	0,4110	0,2267	0,2267
altdif	-0,0943		**0,5053**	0,8721	0,1603	0,1603
	(224)		(224)	(224)	(224)	(224)
	0,1594		0,0000	0,0000	0,0163	0,0163
altdistf	0,0130	0,5053		0,2916	0,0848	0,0848
	(224)	(224)		(224)	(224)	(224)
	0,8471	0,0000		0,0000	0,2061	0,2061
supdif	-0,0552	0,8721	0,2916		0,2686	0,2686
	(224)	(224)	(224)		(224)	(224)
	0,4110	0,0000	0,0000		0,0000	0,0000
altdifdis	-0,0811	0,1603	0,0848	0,2686		1,0000
	(224)	(224)	(224)	(224)		(224)
	0,2267	0,0163	0,2061	0,0000		0,0000
supdifdis	-0,0811	0,1603	0,0848	0,2686	1,0000	
	(224)	(224)	(224)	(224)	(224)	
	0,2267	0,0163	0,2061	0,0000	0,0000	

Typically, you use the coefficient p-values to determine which terms to keep in the regression model.

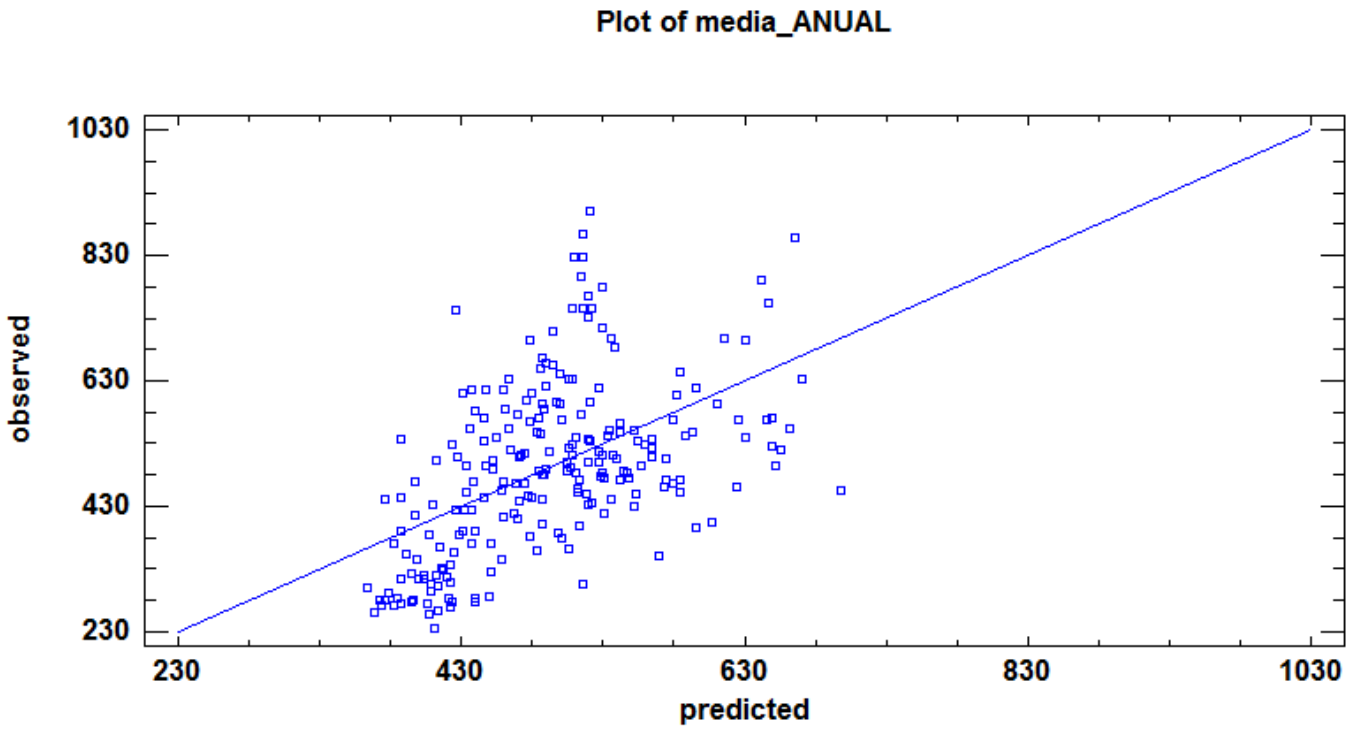
2) SELECTION OF THE REGRESSION MODEL

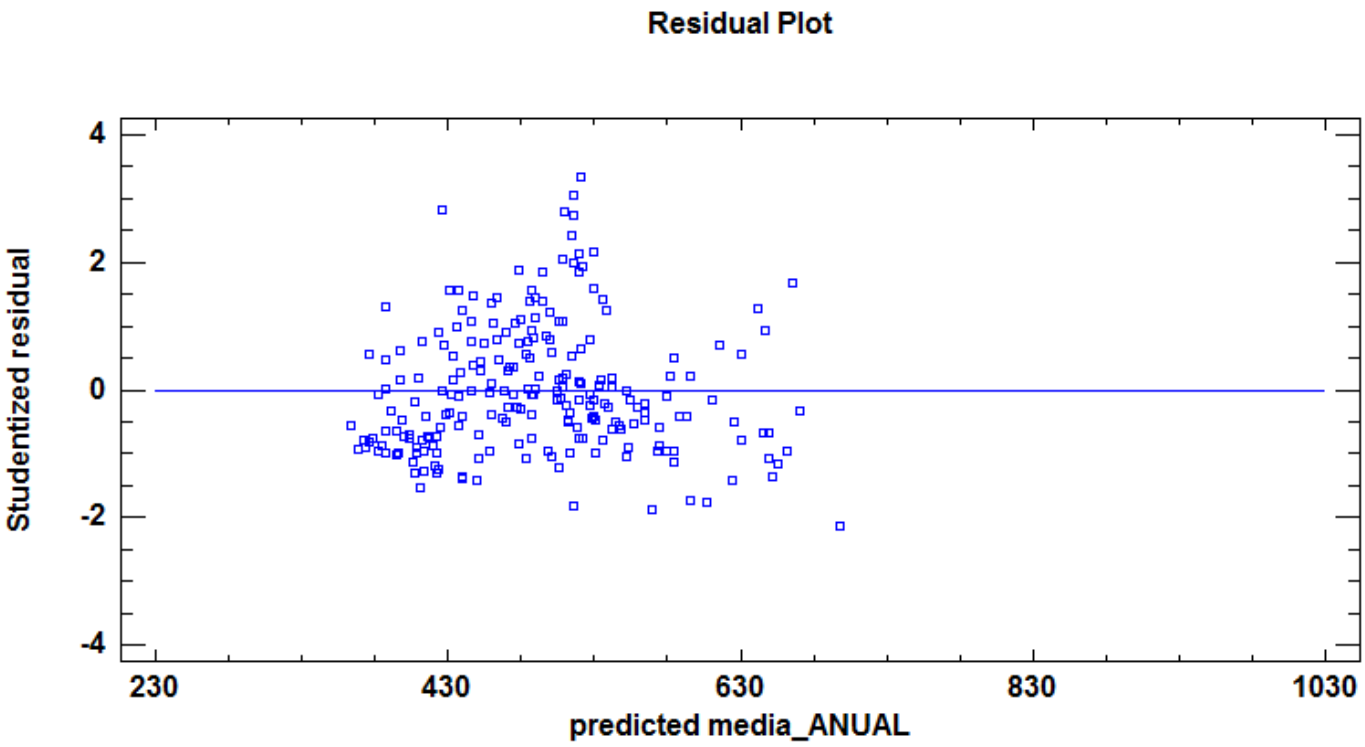
As you can see, after adding all variable - Some variables are highly correlated, generating multicolineality, so the model cannot be solved.



So colinearity were discovered.

We will try the same proces, but adding only X and Y





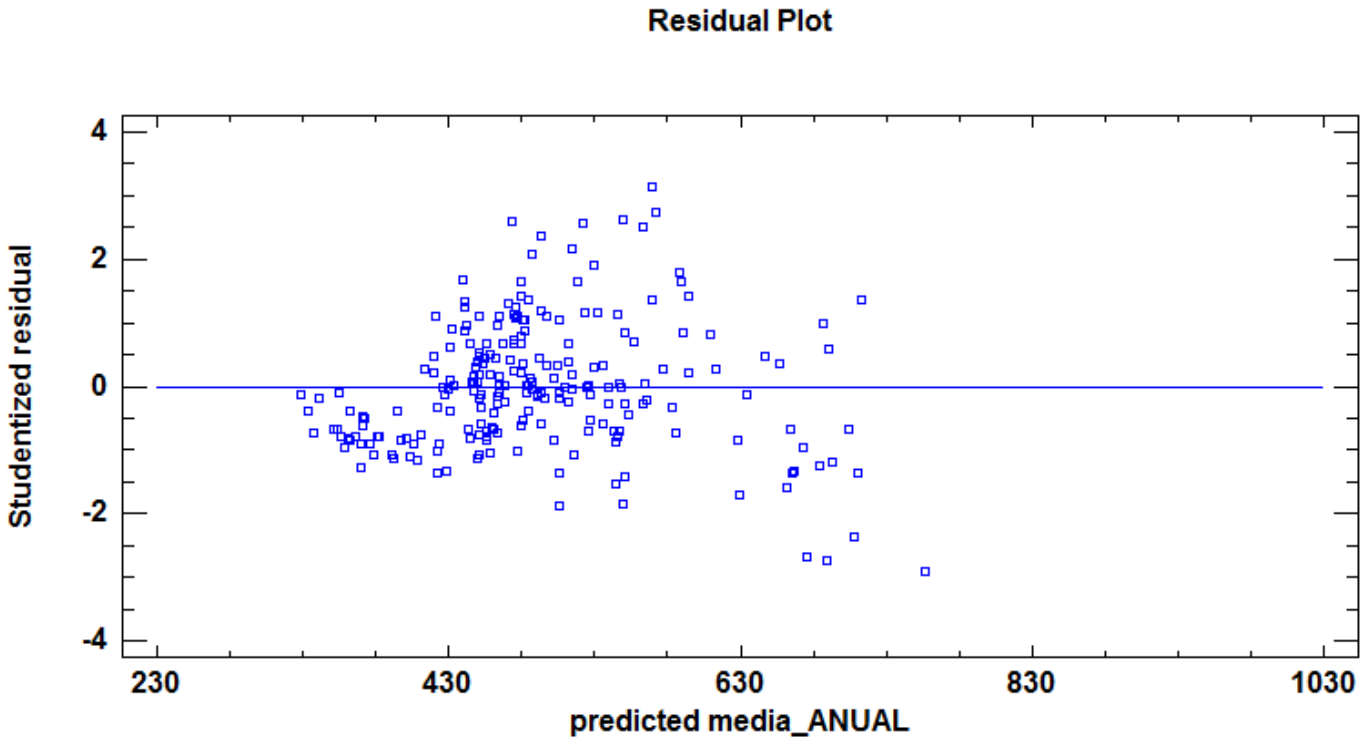
R-squared = 27,3107 percent
R-squared (adjusted for d.f.) = 26,6529 percent
Standard Error of Est. = 117,389
Mean absolute error = 93,0005
Durbin-Watson statistic = 0,627432 (P=0,0000)
Lag 1 residual autocorrelation = 0,675401

But adjusted r-square is still low (26,6) . So we need to add proprietary variables.

For X and Cost it is much more better, but still not enough

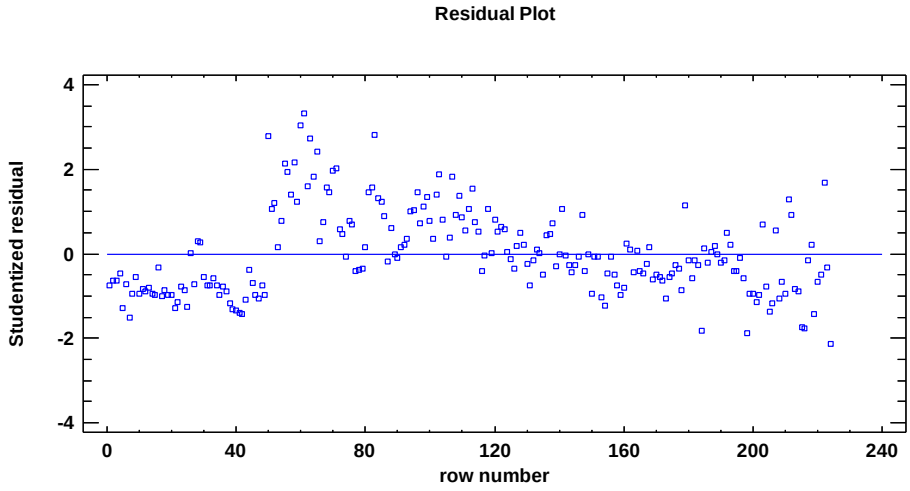
R-squared = 38,1306 percent
R-squared (adjusted for d.f.) = 37,5707 percent
Standard Error of Est. = 108,3
Mean absolute error = 83,3469
Durbin-Watson statistic = 0,712685 (P=0,0000)
Lag 1 residual autocorrelation = 0,625866

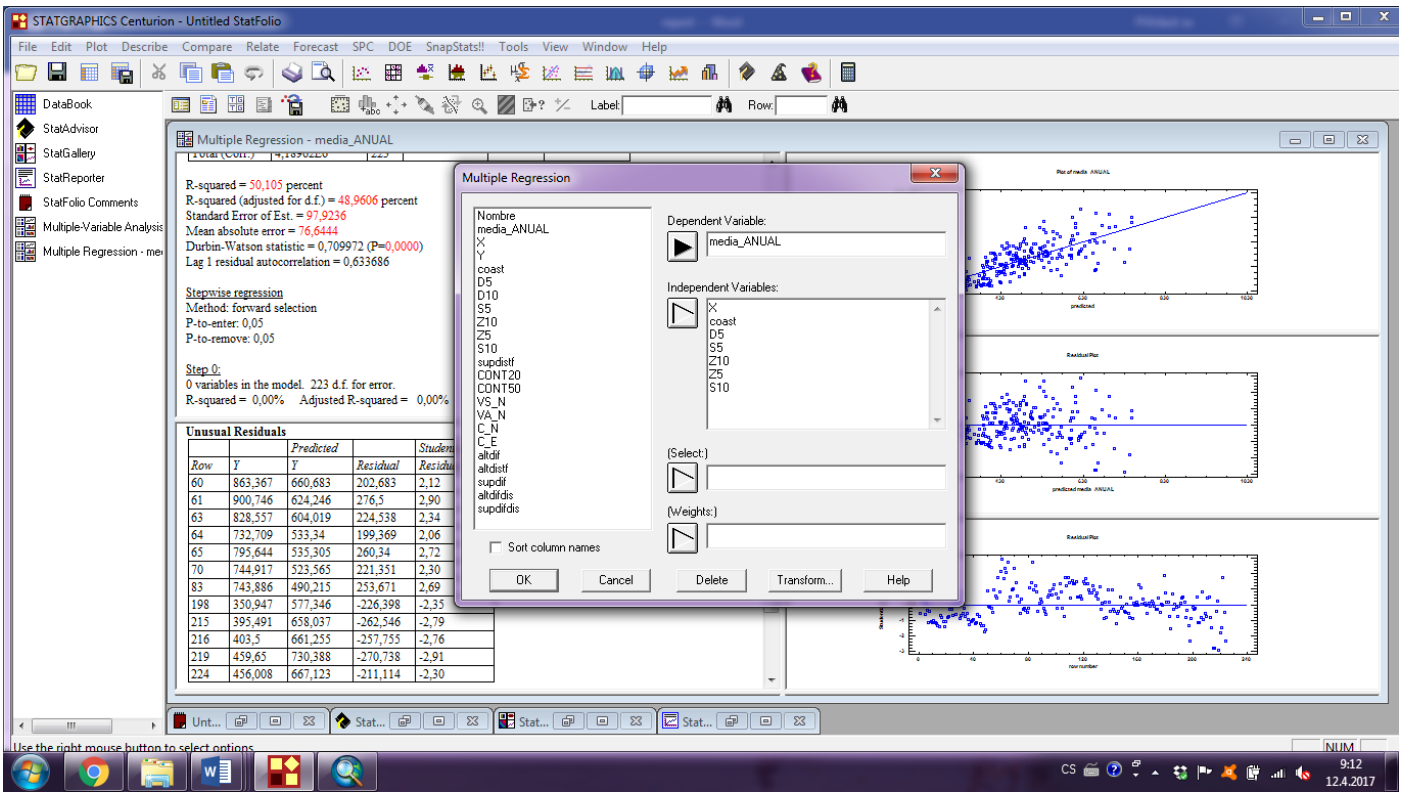
R-squared is a statistical measure of how close the data are to the fitted regression line. It is also known as the coefficient of determination, or the coefficient of multiple determination for multiple regression.



are

3) MULTIPLE REGRESSION





Analysis of Variance

Source	*Sum of Squares*	*Df*	*Mean Square*	*F-Ratio*	*P-Value*
Model	2,09921E6	5	419842,	43,78	0,0000
Residual	2,09041E6	218	9589,03		
Total (Corr.)	4,18962E6	223			

R-squared = 50,105 percent
R-squared (adjusted for d.f.) = 48,9606 percent
Standard Error of Est. = 97,9236
Mean absolute error = 76,6444
Durbin-Watson statistic = 0,709972 (P=0,0000)

Lag 1 residual autocorrelation = 0,633686

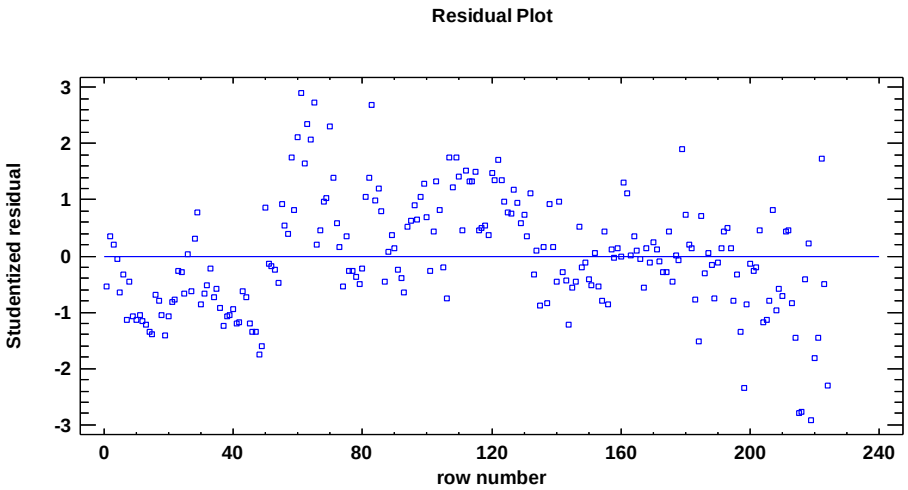
it is not close to 2 so as we can see on this graph there is small correlation(dependence).

Stepwise regression
Method: forward selection
P-to-enter: 0,05
P-to-remove: 0,05

Step 0:
0 variables in the model. 223 d.f. for error.
R-squared = 0,00% Adjusted R-squared = 0,00%
MSE = 18787,5

Step 1:
Adding variable S10 with P-to-enter =0
1 variables in the model. 222 d.f. for error.
R-squared = 20,57% Adjusted R-squared = 20,22% MSE = 14989,6

Step 2:
Adding variable X with P-to-enter =1,81502E-7
2 variables in the model. 221 d.f. for error.
R-squared = 34,52% Adjusted R-squared = 33,92% MSE = 12414,0



Step 3:
Adding variable coast with P-to-enter =4,32723E-8
3 variables in the model. 220 d.f. for error.
R-squared = 42,88% Adjusted R-squared = 42,10% MSE = 10877,5

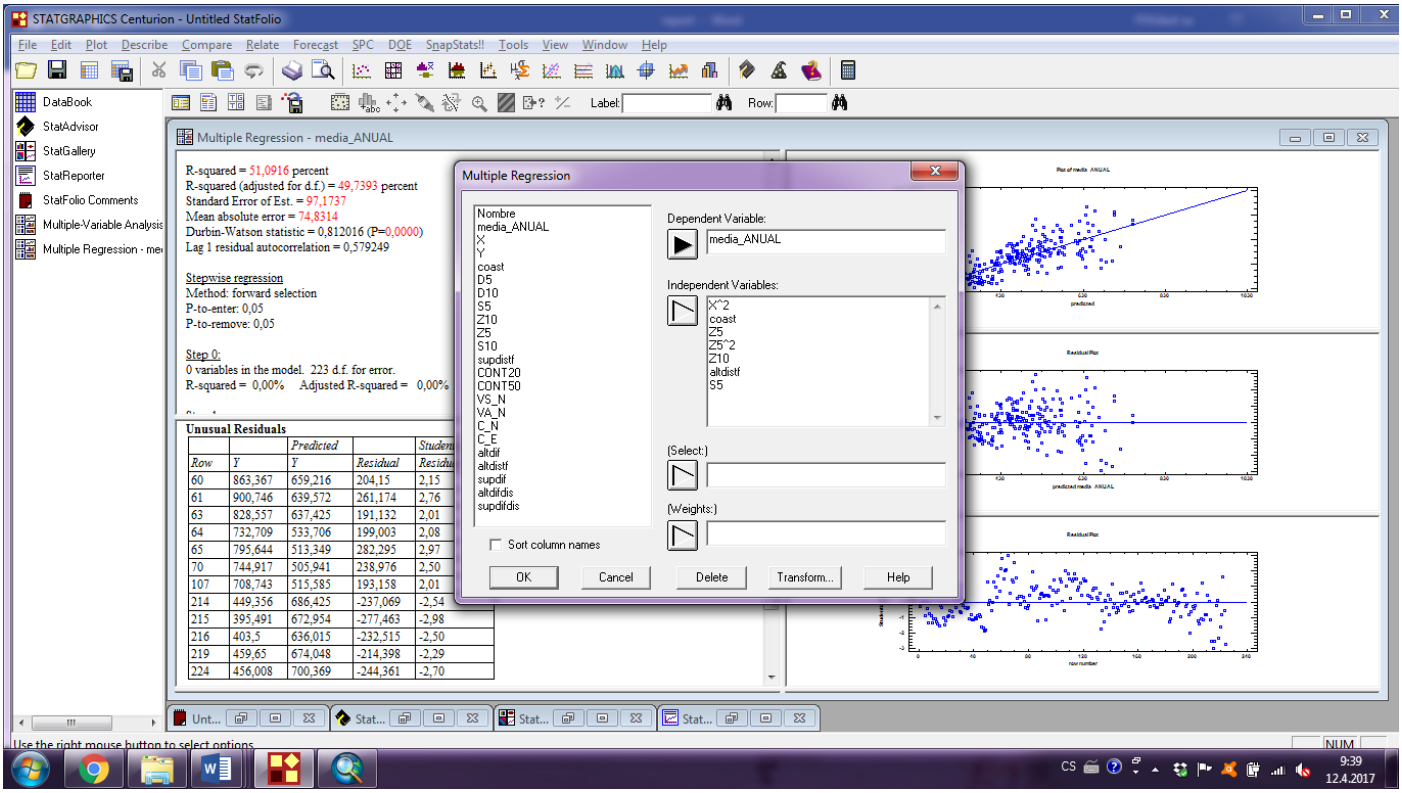
Step 4:
Adding variable Z10 with P-to-enter =0,000323532
4 variables in the model. 219 d.f. for error.
R-squared = 46,16% Adjusted R-squared = 45,18% MSE = 10299,3

Step 5:
Adding variable Z5 with P-to-enter =0,0000476752
5 variables in the model. 218 d.f. for error.
R-squared = 50,11% Adjusted R-squared = 48,96% MSE = 9589,03

Final model selected.

Trying to obtain as high r-square as possible

X2, coast, z5, z52, z10, s10, alrdistf



X2, coast, S10, z10, z5

R-squared = 40,3486 percent
R-squared (adjusted for d.f.) = 38,1186 percent
Standard Error of Est. = 66,677
Mean absolute error = 50,2786
Durbin-Watson statistic = 2,25496
Lag 1 residual autocorrelation = -0,136602

There is a modification of the data in order to reduce homoedasciticity effect.

R-squared were rapidly go down, but Durbin-Watson statistic is better and we managed to fix the problém… at all