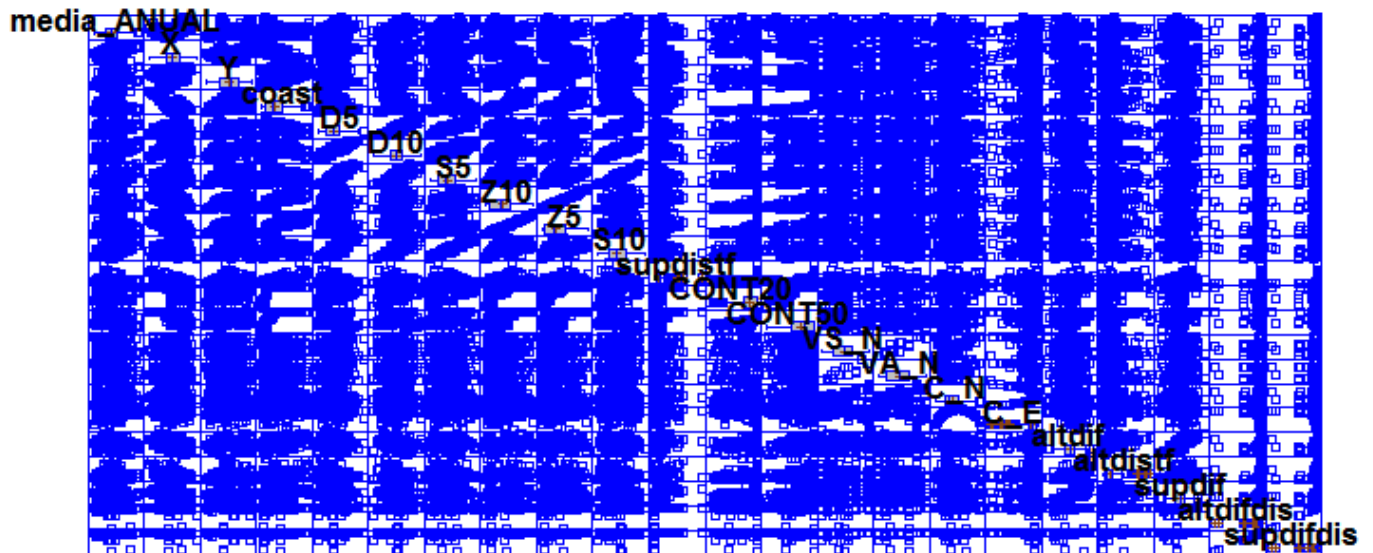


## Practical case 2: Regression models to predict annual mean rainfall

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

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



	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		<b>0,5053</b>	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 multicollineality, so the model cannot be solved.

Multiple Regression - media\_ANUAL

**Multiple Regression - media ANUAL**

Dependent variable: media\_ANUAL

Independent variables:

- X
- Y
- coast
- D5
- D10
- S5
- Z10
- Z5
- S10
- supdistf
- CONT20
- CONT30
- VS\_N
- VA\_N

No output due to data error.

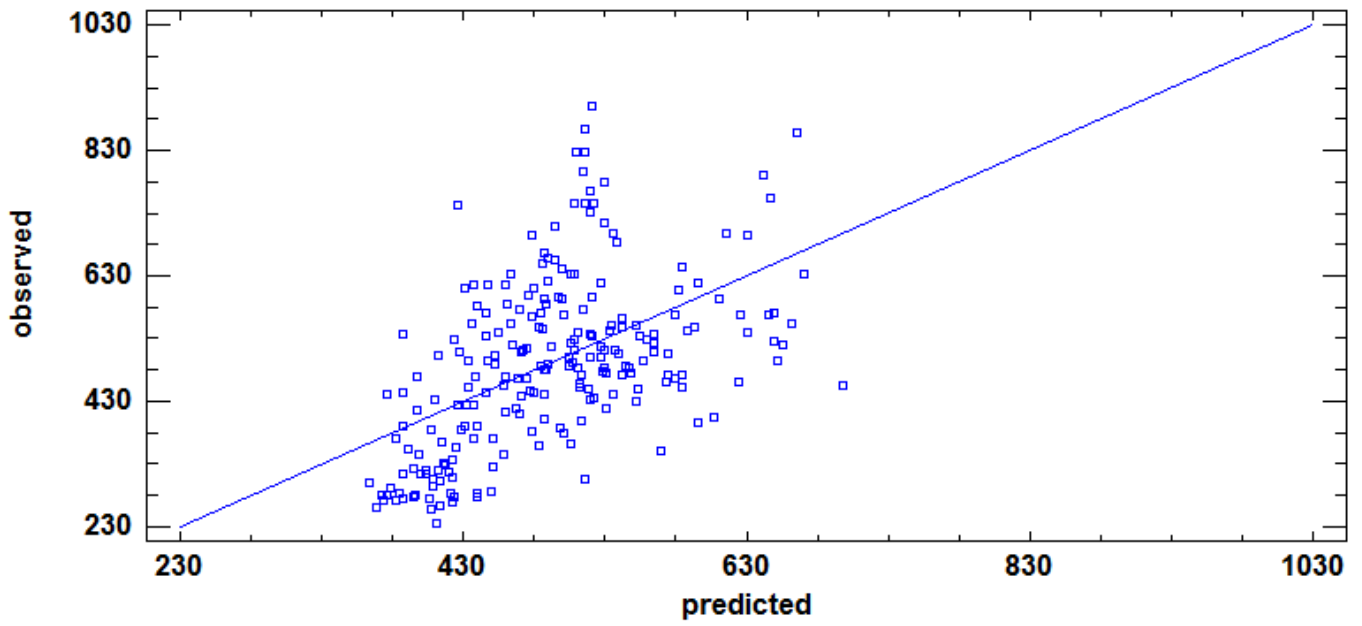
No output due to data error.

No output due to data error.

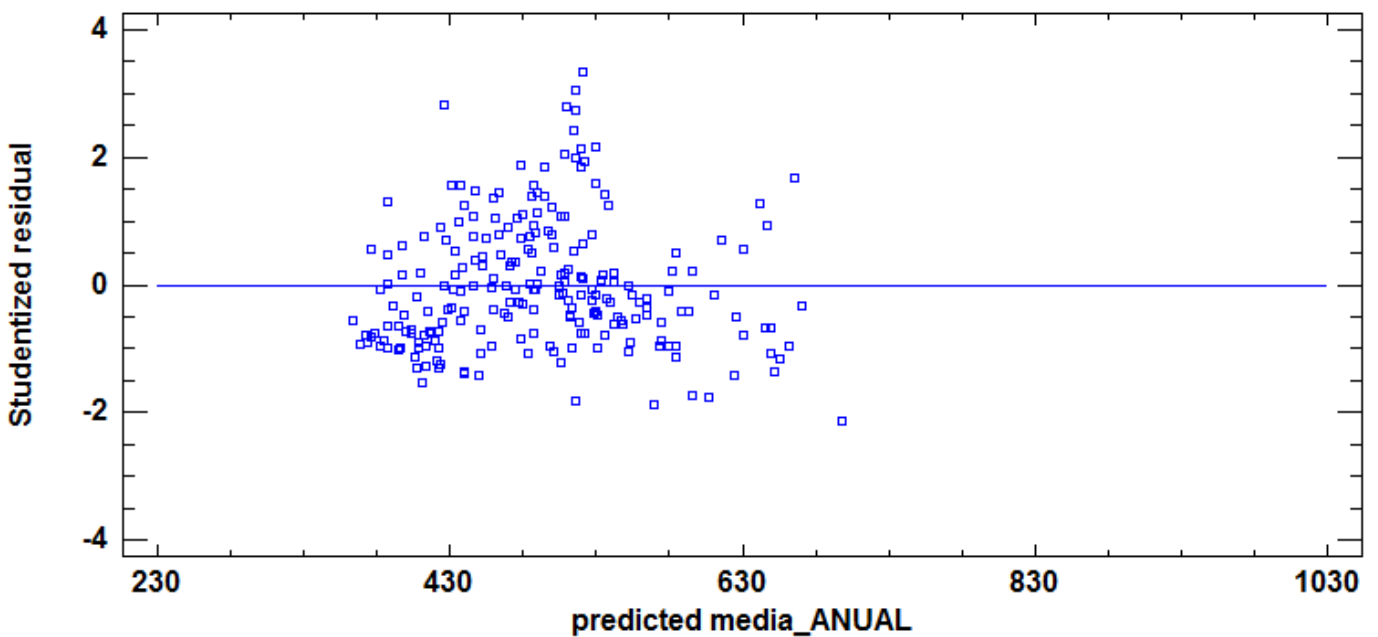
So colinearity were discovered.

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

Plot of media\_ANUAL



Residual Plot



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

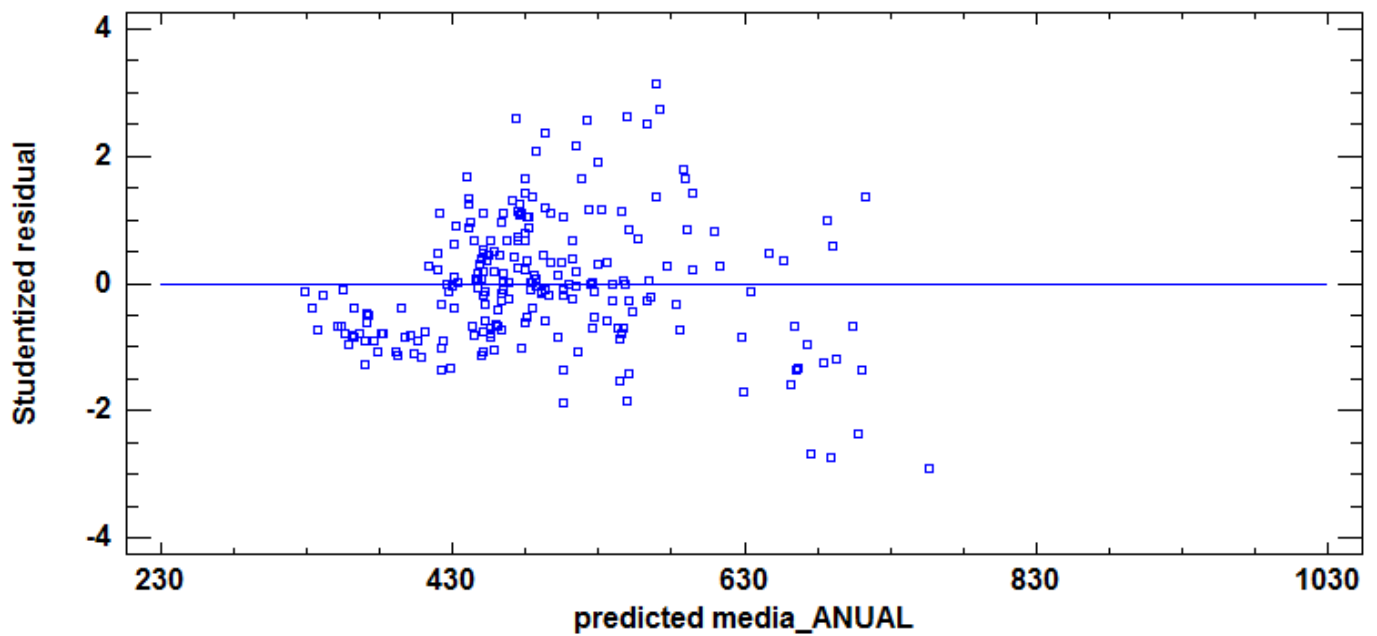
But adjusted r-square is still low (26,6) . So we need to add proprietary variables.

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.

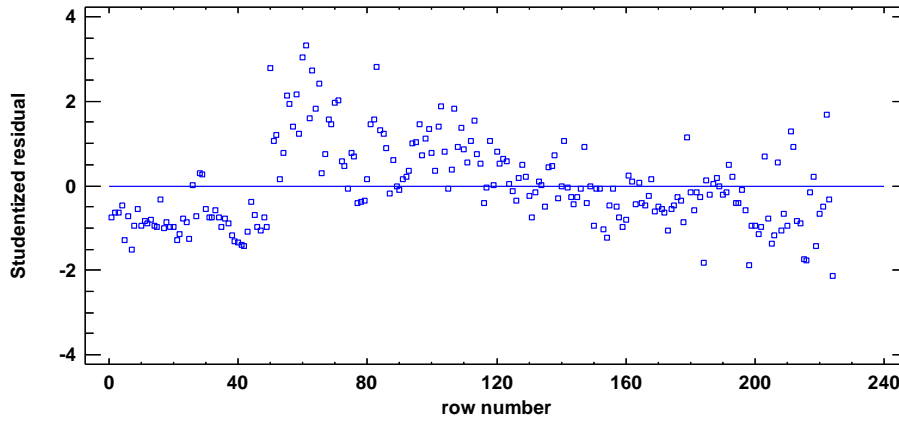
Residual Plot



are

### 3) MULTIPLE REGRESSION

Residual Plot



The screenshot shows the STATGRAPHICS Centurion interface. The main window displays the results of a Multiple Regression analysis for the dependent variable 'media\_ANUAL'. The regression statistics are as follows:

- 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

The 'Stepwise regression' section shows the method used is 'forward selection' with a P-to-enter of 0,05 and a P-to-remove of 0,05. The 'Step 0' section indicates that 0 variables are in the model and 223 d.f. are for error, with an R-squared of 0,00% and an Adjusted R-squared of 0,00%.

An 'Unusual Residuals' table is also visible, listing rows 60 through 224 with their predicted values, actual values, and studentized residuals.

Overlaid on the main window is a 'Multiple Regression' dialog box. It shows the dependent variable as 'media\_ANUAL' and the independent variables as 'X', 'coast', 'D5', 'S5', 'Z10', 'Z5', 'S10', 'supdist', 'CONT20', 'CONT50', 'VS\_N', 'VA\_N', 'C\_N', 'C\_E', 'aldist', 'supdif', 'aldifdis', and 'supdifdis'. The 'Sort column names' checkbox is unchecked.

On the right side of the screenshot, three residual plots are visible: a scatter plot of 'media\_ANUAL' vs 'predicted', a 'Residual Plot', and another 'Residual Plot'.

#### 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

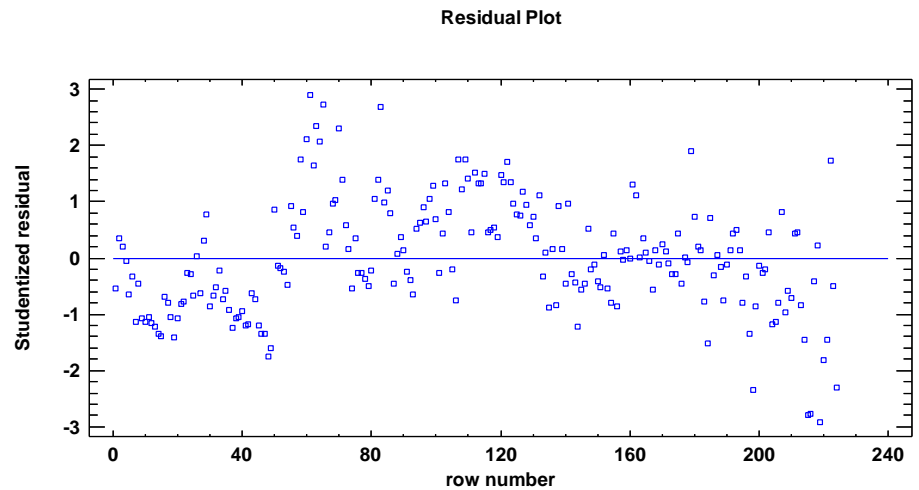
5 variables in the model. 218 d.f. for error.

R-squared = 50,11% Adjusted R-squared = 48,96% MSE = 9589,03

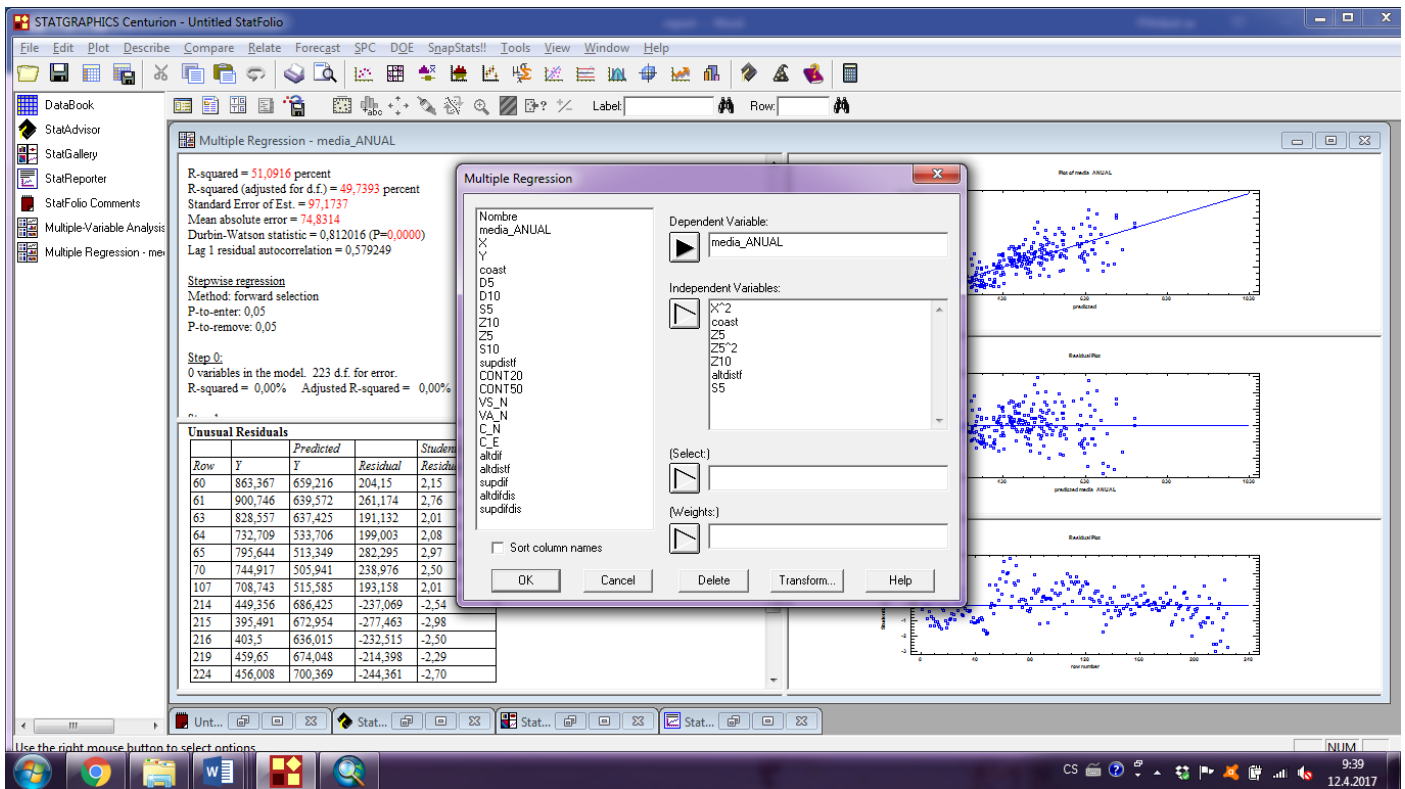
Final model selected.

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 homoedasticity effect.**

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

Residual Plot

