Linear Regression with Regression Diagnostics

in S-Plus

An application for
Outliers detection in the accuracy examination of a pK_a prediction

Karel Kupka
kupka@trilobyte.cz, www.trilobyte.cz

January 2007

User's instructions

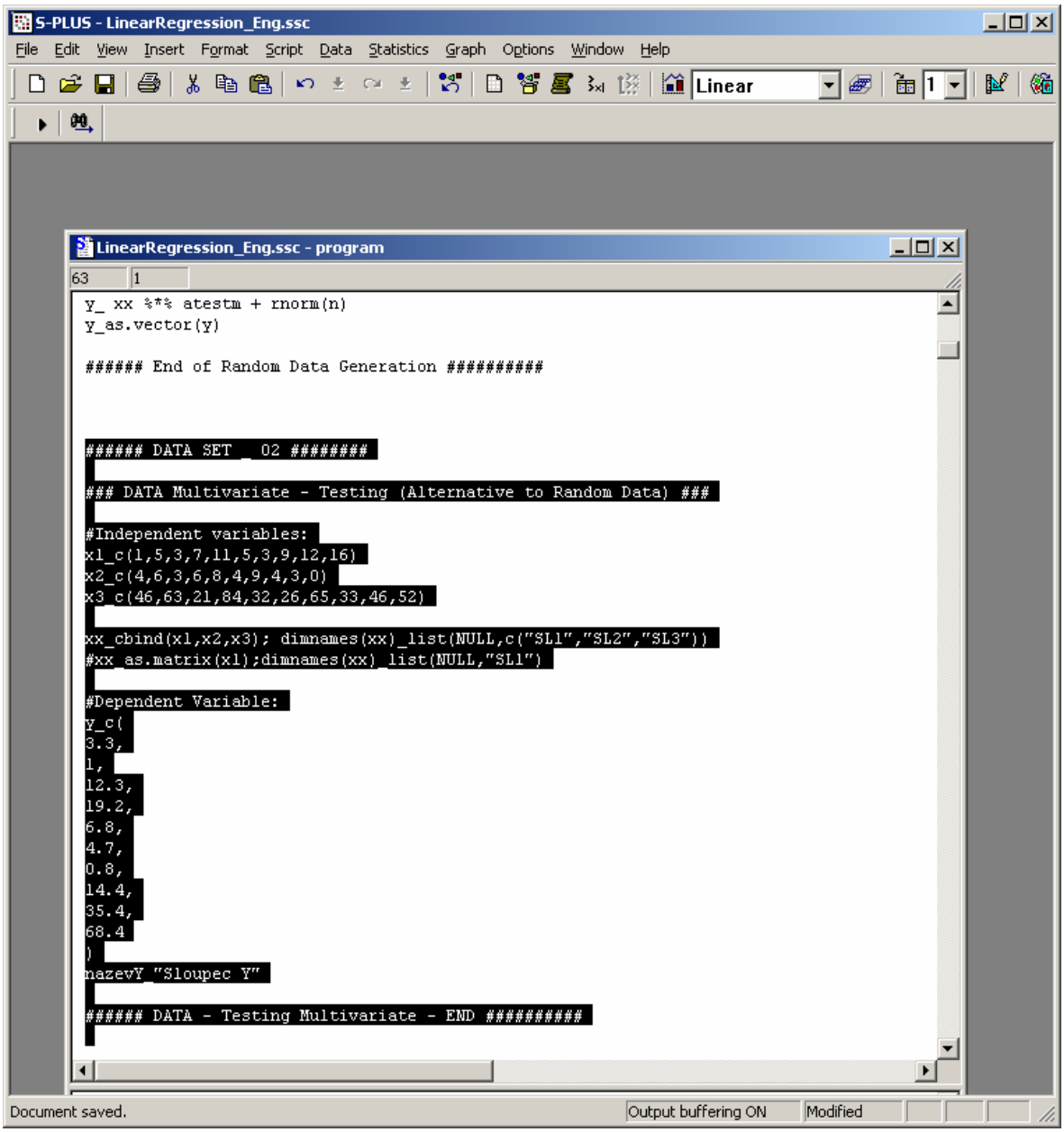
Code: LinearRegression_Eng.ssc

To run this code you need the statistical software S-Plus by Insightful Inc., version 5 or higher (www.insightful.com). (With minor changes it will run also on version 4.) The code can be run by parts by selecting apropriate part of the source text and pressing F10. An example use of the code is given below.

General detailed example use of the code (for pK PALLAS example – see page 16):

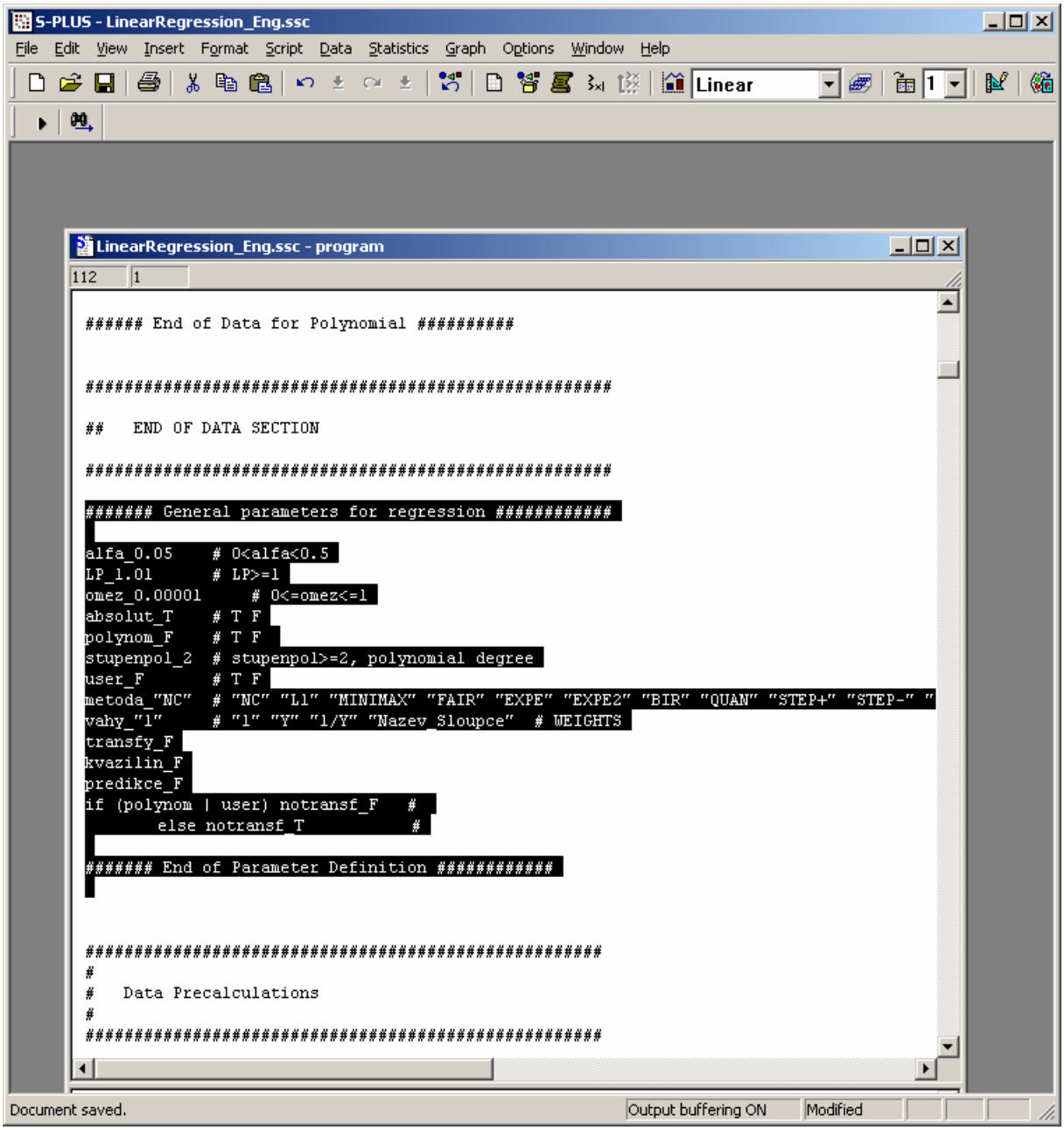
Step 1

Data are expected in a matrix „xx“ (n x m) – independent variable and vector „y“ (n x 1) – dependent variable. You can generate sample data using code at the beginning of the file. Select part of code DATASET_02 and press F10.



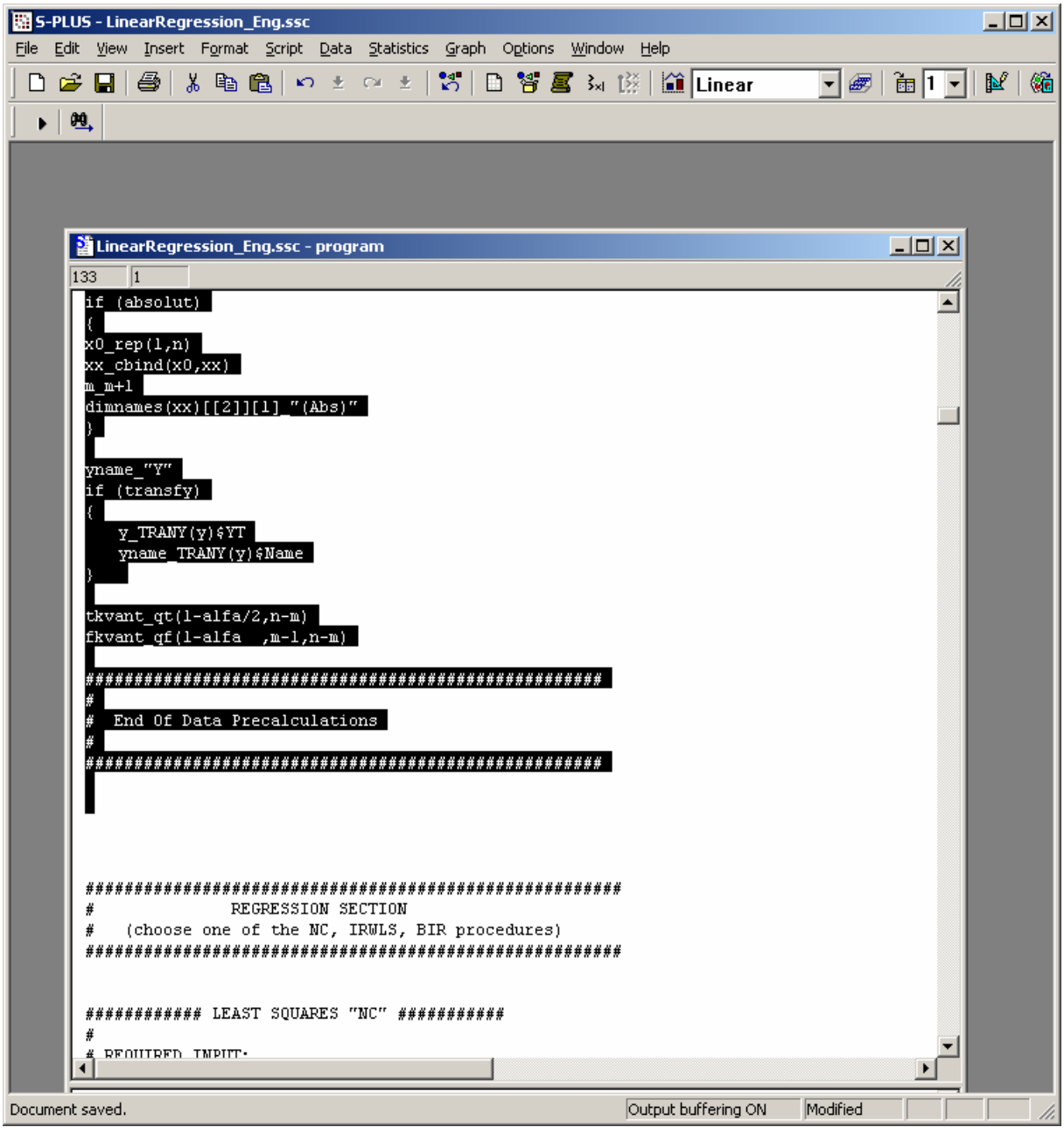
Step 2.

Select code General parameters for regression to define initial settings for regression and press F10.



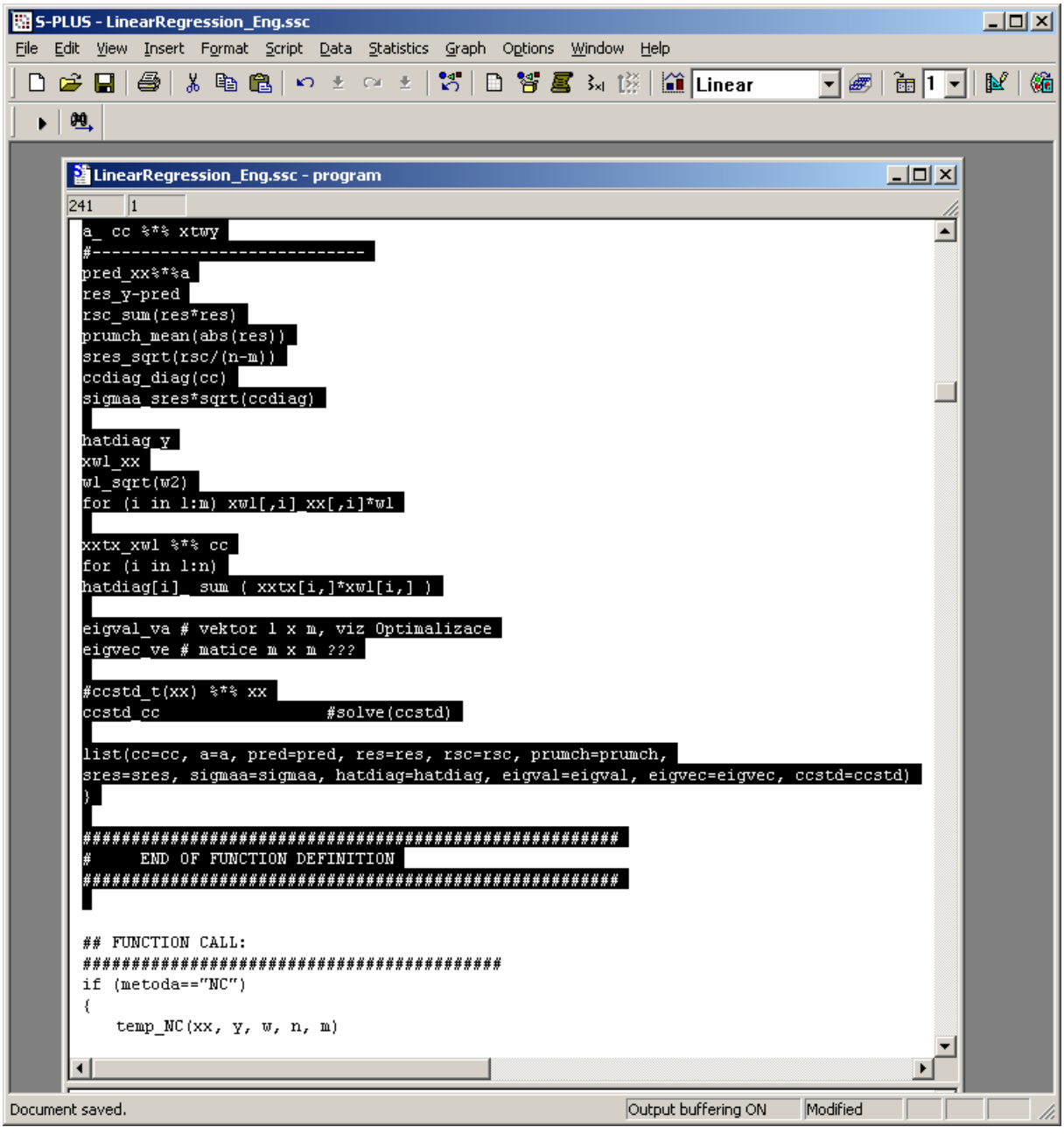
Step 3.

Select the section Data Precalculations and press F10. This code prepares the data matrices and calculates some critical quantiles. Number of rows is stored in variable *n*, number of columns of matrix X si stored in *m*.



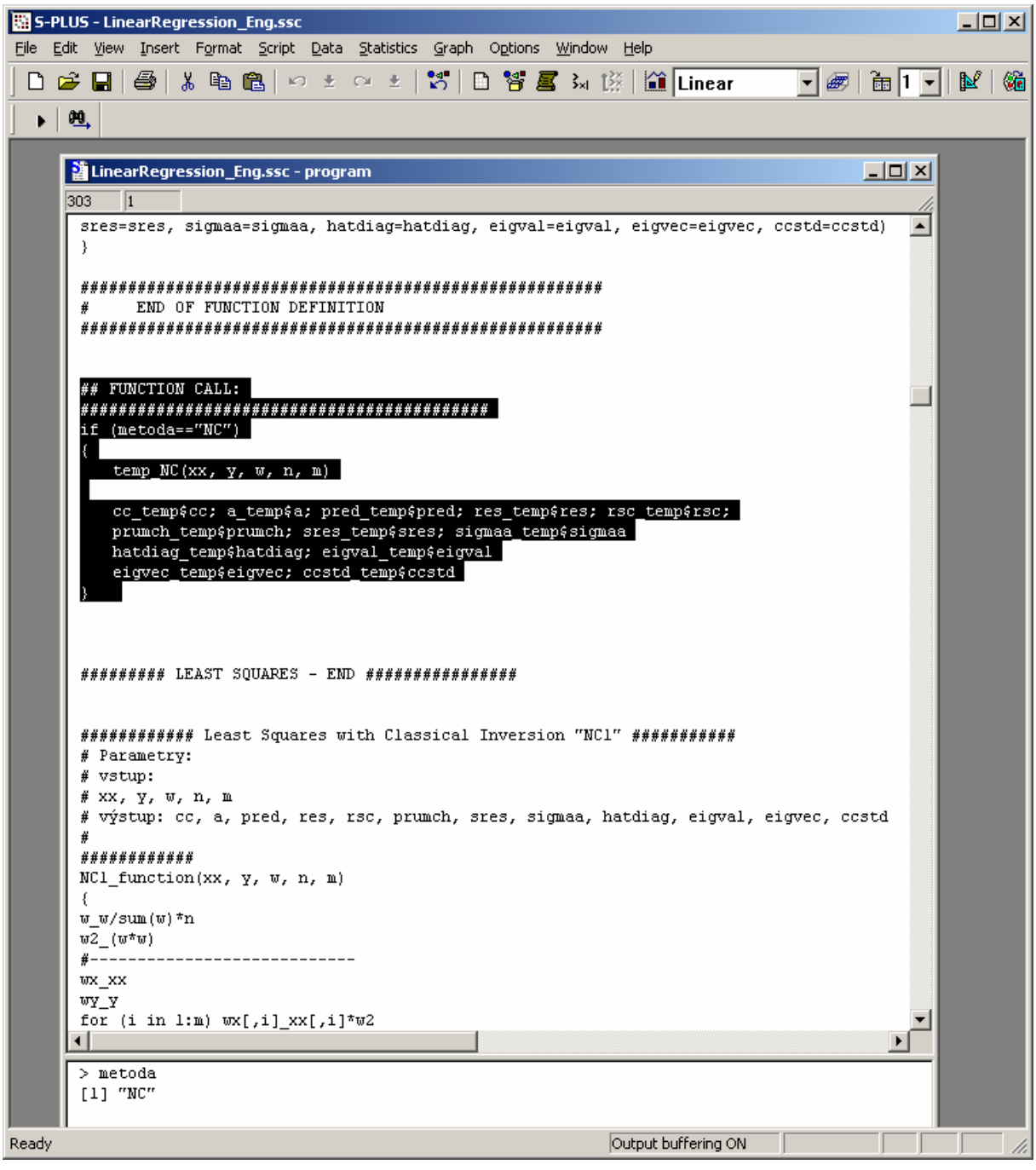
Step 4.

Select the part of code NC - LEAST SQUARES FUNCTION DEFINITION and press F10. This defines the least squares algorithm, no calculations are performed in this step. The name „NC“ is derived (as much of other variable names) from Czech.



Step 5.

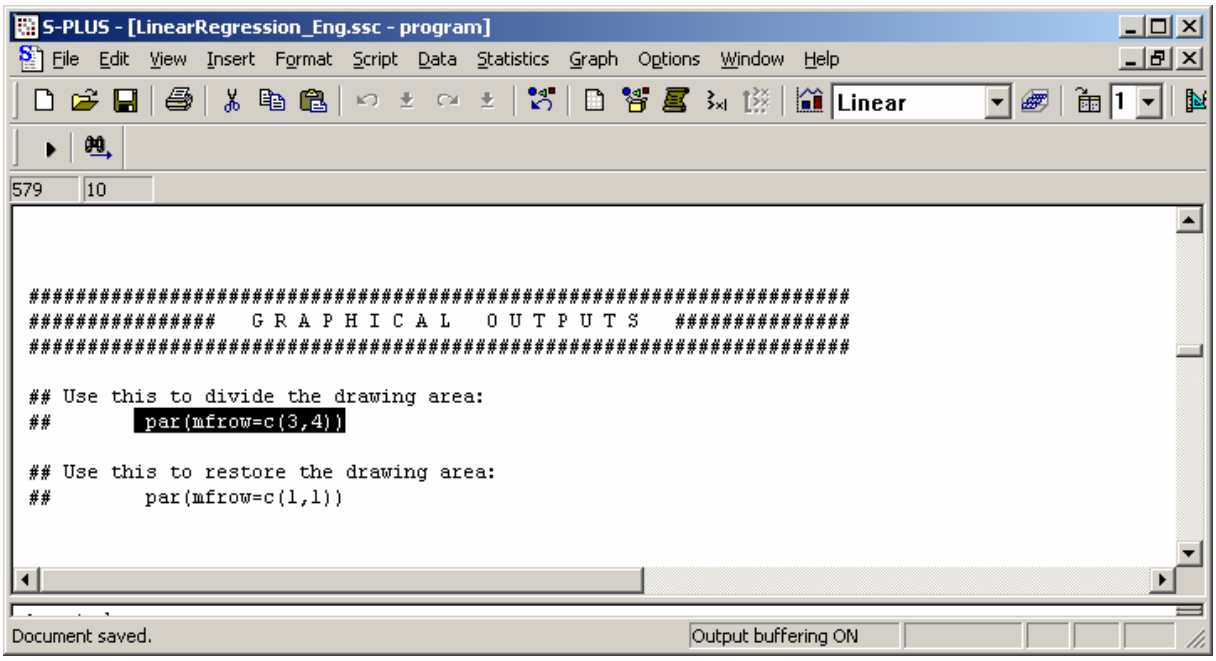
Select the function call NC and press 10. This performs the least squares regression defined in the previous step.



Jump over following code for alternative robust regression algorithms and go to GRAPHICAL OUTPUTS.

Step 6.

Select code for preferred plots and press F10. If you want to create more plots you can divide the plotting area by command *par(mfrow=c(3,4))*. (3 rows by 4 columns)



The plot of the regression curve is available only for data with one independent variable. Since we have 3 independent variables in our example, we cannot create this plot. Any of the following plots can be used. Available plots are as follows:

Regression curve Plot

Residuals Plots Group:
Y-Prediction
Residuals vs. prediction
Abs. residuals
Squared Residuals
Q-Q plot of residuals
Autocorrelation
Heteroscedasticity
Jack-Knife residuals
Predicted Residuals

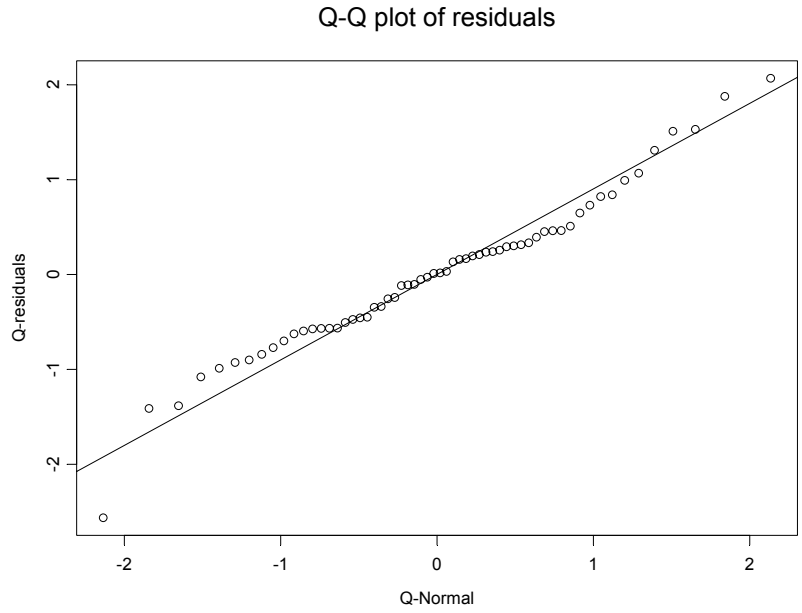
Partial regression and residual plots Group:
Partial regression plot
Partial Residual plot

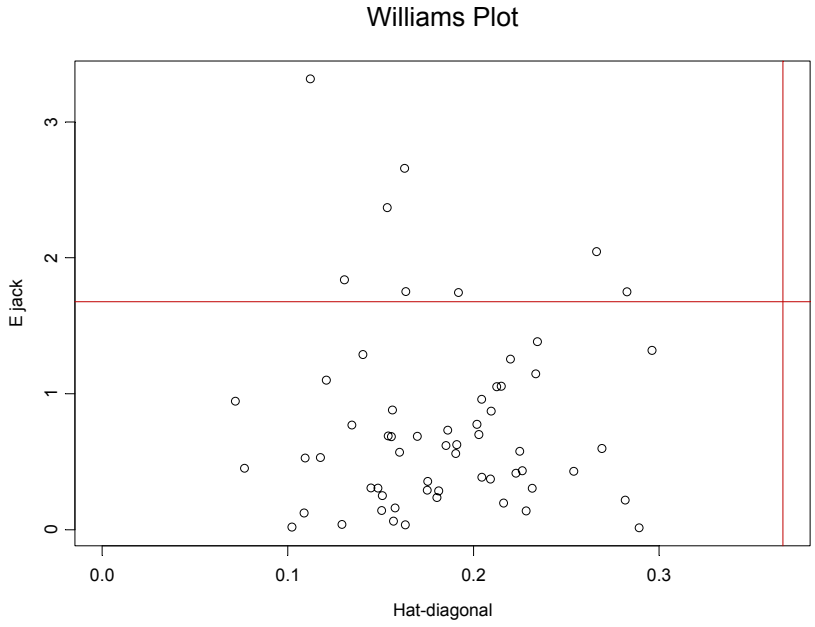
Influential Points Diagnostics plots Group:
Projection Matrix Plot
Prediction of Residuals
Pregibon Plot
Williams Plot
McCulloh-Meter plot

L-R plot
Cook Distance Plot
Atkinson Distance Plot
Likelihood Distance Plots

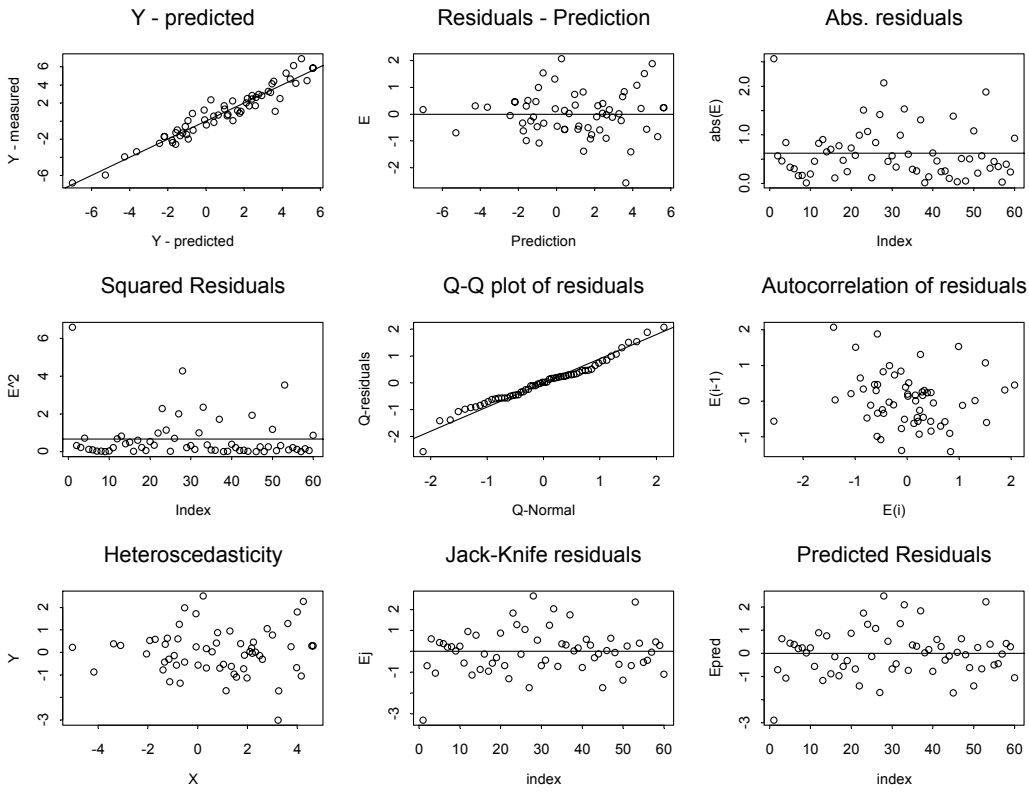
Rankit Plots Group:
Q-Q for Normalized Residuals
Q-Q Predicted Residuals
Q-Q Jack-Knife Residuals

Example plots:

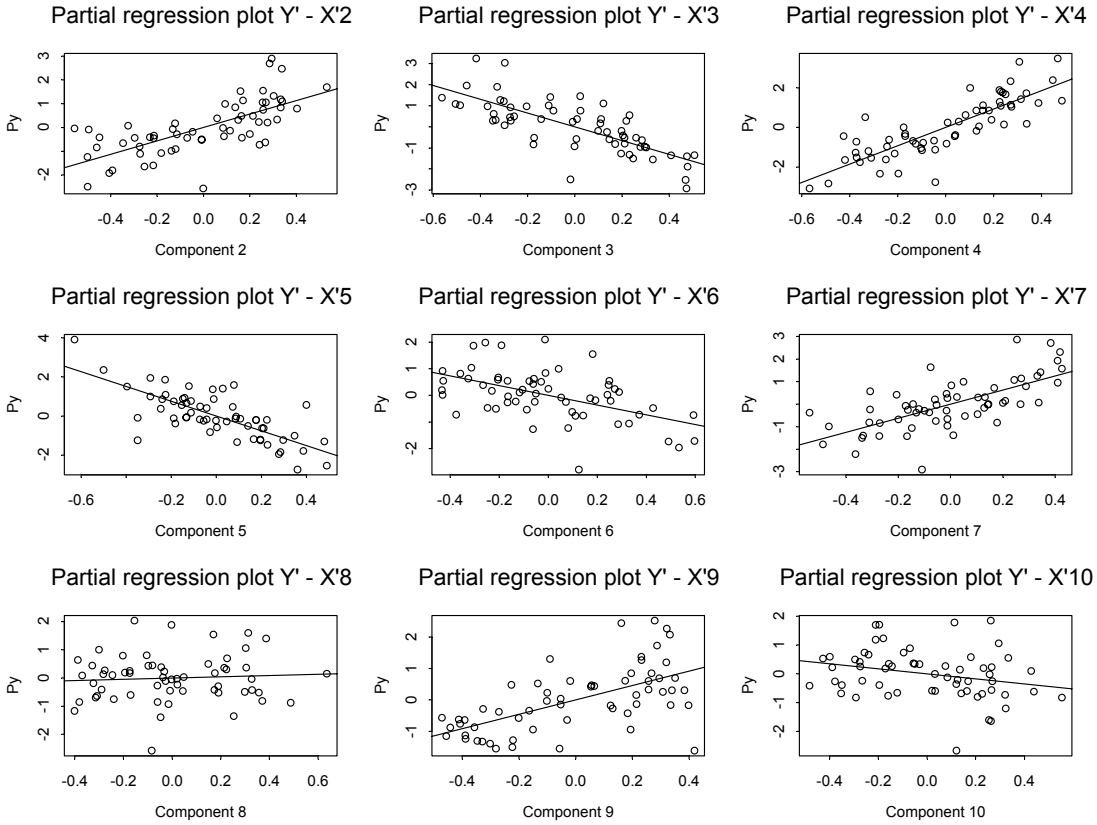




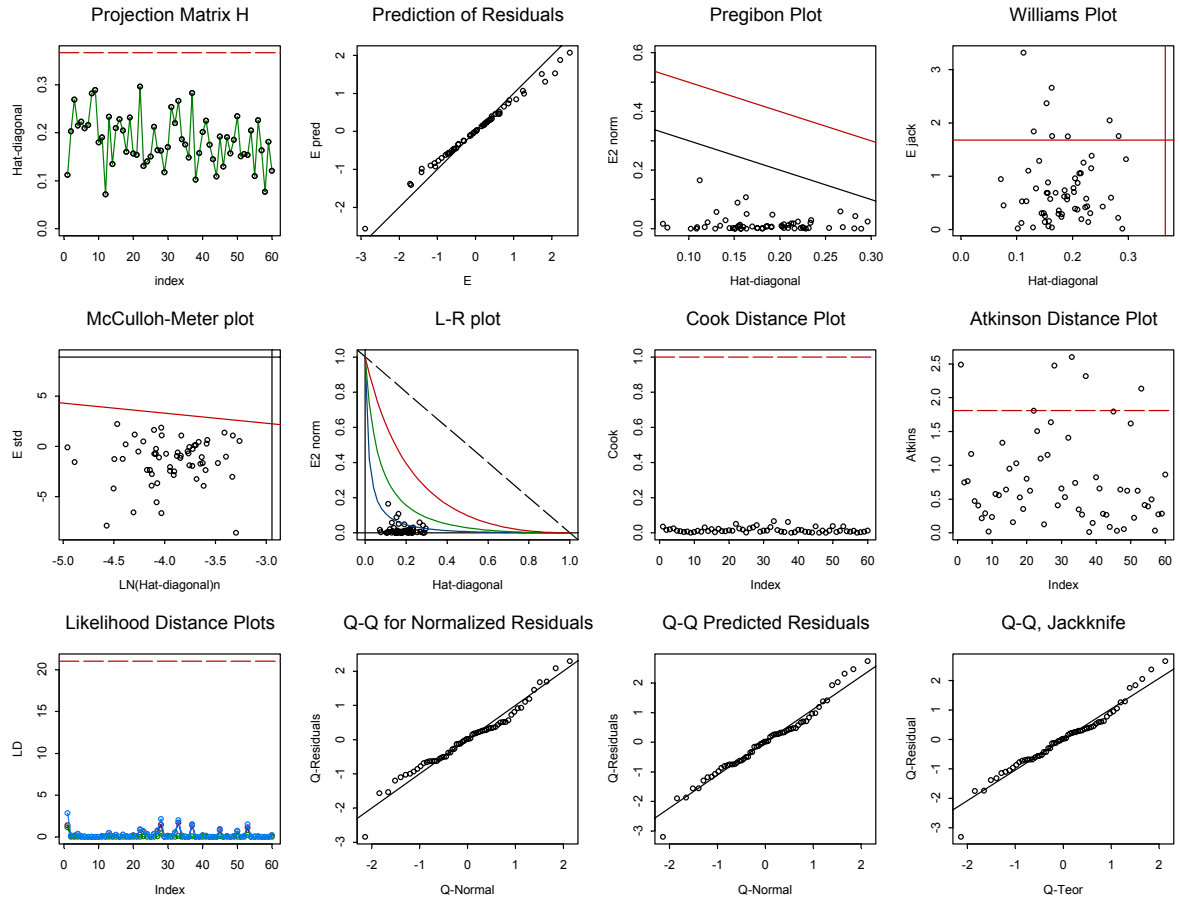
Plot of all Residuals Group Plots:



Plot of all Partial Regression Plots



Plot of all Influence Diagnostic and Rankit plots:



Step 7.
Text outputs.

Select and run the „TEXT OUTPUT / RESULT TABLES“ code. The results will apprear in the output window of S-Plus. From here it may be pasted into Excel in the form of the following table divided into 9 sections:

Section 1. Basic Analysis
Section 2. Pair Correlations (Xi, Xj)
Section 3. Multicollinearity Detection
Section 4. Analysis of Variance
Section 5. Parameter Estimates
Section 6. Statistical Characteristics
Section 7. Analysis of Classical Residuals
Section 8. Regression Triplet Analysis
Section 9. Influential Points Diagnostics
- Subsection A. Residual Analysis
- Subsection B. Influence Analysis

Linear Regression Results

Significance Level	0.05					
Quantile t(1-alfa/2,n-m)	2.446911851					
Quantile F(1-alfa,m,n-m)	4.757062663					
Absolute Term:	YES					
No of Rows	10					
No of Parameters	4					
Method	Classical Least Squares					
Columns						
	Column 1					
	Column 2					
	Column 3					
Transformation	No transormation					
Section 1. Basic Analysis						
Variable Characteristics						
Variable	Mean	StdDev	Corr. vs. Y	Significance		
SL1	7.2	4.732863826	0.817487563	0.003870068		
SL2	4.7	2.626785107	-0.736945571	0.015033604		
SL3	46.8	19.75854246	0.095069186	0.793905594		
Section 2. Pair Correlations (Xi, Xj)						
Variable	Corr.Coeff	Significance				
(Abs) - SL1	0	1				
(Abs) - SL2	0	1				
(Abs) - SL3	0	1				
SL1 - SL2	-0.414693202	0.233422623				
SL1 - SL3	0.000475268	0.998960351				
SL2 - SL3	0.309132904	0.384773263				
Section 3. Multicollinearity Detection						
Variable	Eigenvalue cor.	Condition number kappa	VI factor	Multiple corr.		
(Abs)	1.517008839	3.143827236	1	0		
SL1	1.000455473	2.07332949	1.234988555	0.436206255		
SL2	1	2.072385575	1.365477622	0.517354346		
SL3	0.482535688	1	1.130656073	0.339937864		
Section 4. Analysis of Variance						
Mean Y	16.63					
Source	Square Sum	Mean Square	Variance			
Overall Variability	3974.301	397.4301	441.589			
Explained Variability	3682.113726	368.2113726	409.1237473			
Residual variability	292.187274	29.2187274	32.46525267			
F-statistic	25.20379259					
Critical F-quantile (1-alfa, m-1, n-m)	4.757062663					
p-value	0.000844931					
Conclusion	Model is Significant					

Section 5. Parameter Estimates						
Variable	Estimate	StdDev.	Conclusion	p-value	Lower limit	Upper limit
(Abs)	6.474243433	8.292998992	Insignificant	0.464662766	-13.81799408	26.76648095
SL1	2.554956891	0.546186832	Significant	0.003403678	1.218485858	3.891427923
SL2	-4.666888885	1.034788317	Significant	0.004060519	-7.19892468	-2.13485309
SL3	0.29261634	0.125182645	Insignificant	0.058040863	-0.013694558	0.598927238
Section 6. Statistical Characteristics						
Multiple Correlation Coefficient R	0.962538747					
Determination Ceofficient R^2	0.926480839					
Predicted Correlation Coefficient Rp	0.766496792					
Mean Error of Prediction MEP	92.80120324					
Akaike Information Criterion	41.74809853					
Section 7. Analysis of Classical Residuals						
Index	Y measured	Y predicted	StdDev. Y	Residual	Rel. Residual [%Y]	
1	3.3	3.821996424	4.343007263	-0.521996424	-15.81807346	
2	1	9.682523998	3.113775963	-8.682523998	-868.2523998	
3	12.3	6.283390591	4.594826023	6.016609409	48.91552365	
4	19.2	20.93738092	4.913361548	-1.737380919	-9.048858956	
5	6.8	6.607381033	5.991566876	0.192618967	2.832631874	
6	4.7	8.189497188	3.487801957	-3.489497188	-74.24462101	
7	0.8	-8.842823759	4.569319327	9.642823759	1205.35297	
8	14.4	20.45763913	2.894543277	-6.057639131	-42.06693841	
9	35.4	36.59341111	3.283668827	-1.193411109	-3.371217821	
10	68.4	62.56960337	5.733888587	5.830396633	8.523971685	
Residual Sum of Squares	292.187274					
Mean of Absolute Residuals	4.336489754					
Residual StdDev	6.978386562					
Residual Variance	48.697879					
Residual Skewness	0.059799837					
Residual Variance	2.143652778					
Section 8. Regression Triplet Analysis						
Fisher-Snedecor test of Significance						
F-Statistic	25.20379259					
Critical F- (1-alfa, m-1, n-m)	4.757062663					
p-value	0.000844931					
Conclusion	Model is Significant					
Scott Criterion of Multicollinearity						

SC Statistic	16.09845951					
Conclusion	Multicollinearity is too high, model is incorrect!					
Cook-Weisberg test of heteroscedasticity						
CW-statistic	0.152125821					
Critical quantile Chi^2(1-alfa,1)	3.841458821					
p-value	0.696512069					
Conclusion	Residuals are homoscedastic.					
Jarque-Berra test of normality						
JB-statistic	0.405220796					
Critical quantile Chi^2(1-alfa,2)	5.991464547					
p-value	0.816596327					
Conclusion	Residuals are normally distributed.					
Wald test of autocorrelation						
WA-statistic	1.346001087					
Critical quantile Chi^2(1-alfa,1)	3.841458821					
p-value	0.245978431					
Conclusion	Autocorrelation is insignificant					
Durbin-Watson test of autocorrelation						
DW-statistic	2.91601665					
Conclusion	Residuals are negatively correlated!					
Sign test of fit						
Sgn-statistic	1.896586268					
Critical Quantile N(1-alfa/2)	1.959963985					
p-value	0.057882564					
Conclusion	There is no trend in residuals.					
Section 9. Influential Points Diagnostics						
Subsection A. Residual Analysis						
Index	Standard	Jackknife	Predicted	Diag(Hii)	Diag(H*ii)	Cook Dist.
1	-0.095564373	-0.087304406	-0.851990096	0.387321018	0.388253572	-0.015103378
2	-1.390275931	-1.541493961	-10.84091807	0.199096982	0.457103506	-0.086402391
3	1.14554396	1.183083812	10.62139958	0.433538926	0.557430663	0.219184478
4	-0.350598084	-0.323380407	-3.445355783	0.495732508	0.506063185	-0.086166008
5	0.05384073	0.049161515	0.732880009	0.737175301	0.737302282	0.037753355
6	-0.57732384	-0.542299858	-4.651426651	0.249800664	0.291474589	-0.048059186
7	1.828234549	2.507699554	16.8798911	0.428738983	0.746973411	0.343028579
8	-0.953994429	-0.945532464	-7.316414898	0.172048166	0.297635401	-0.049559946
9	-0.193812715	-0.177482436	-1.53279651	0.221415823	0.226290197	-0.013779243
10	1.465849425	1.670211647	17.94695062	0.675131628	0.791473193	0.761570988

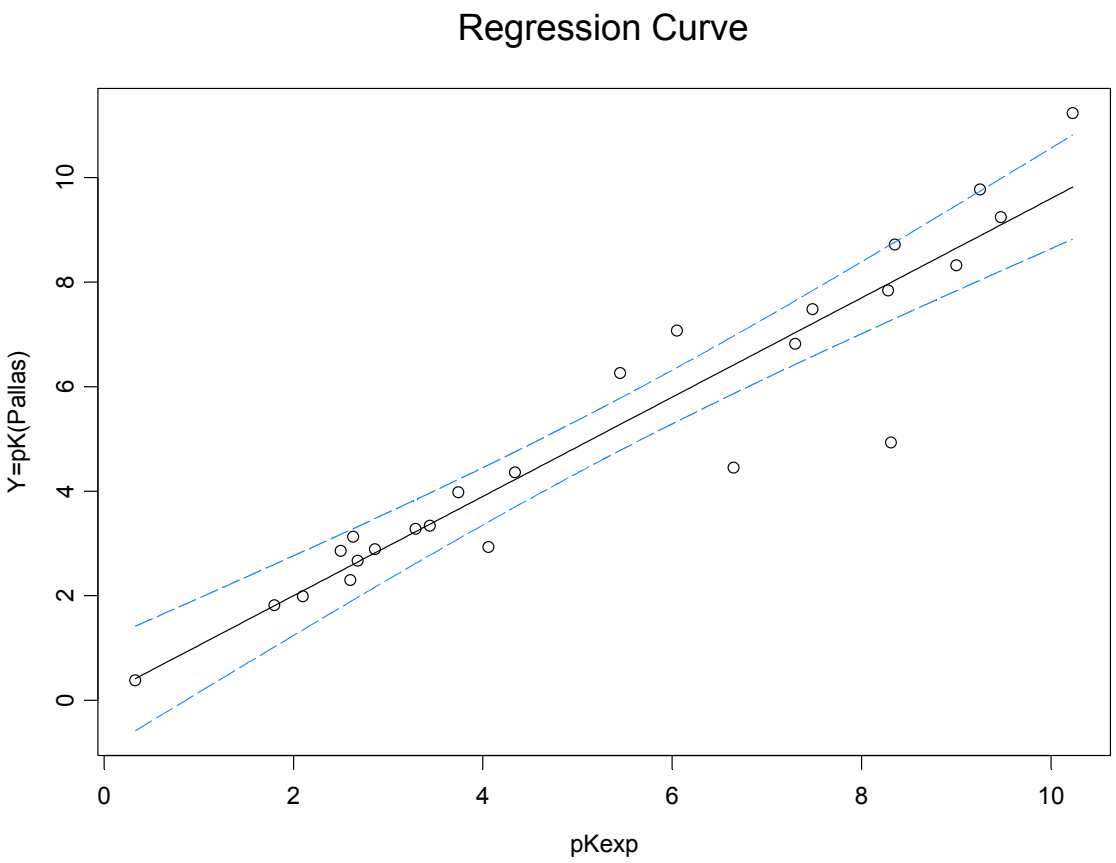
Subsection B. Influence Analysis						
Index	Atkinson Dist.	Andrews-Pregibon stat	Infl on Y^	Infl on parameters LD(b)	Infl on variance LD(s)	Total Influence LD(b,s)
1	0.085016032	0.611746428	-0.0694153	0.009617673	0.052092352	0.060765623
2	0.941303841	0.542896494	-0.768571368	0.770370656	0.442559886	1.505823391
3	1.267623129	0.442569337	1.035009951	1.547707474	0.104930124	2.033172201
4	0.392692445	0.493936815	-0.320632039	0.199396358	0.034847334	0.219896162
5	0.100837759	0.262697718	0.082333685	0.013541948	0.053122954	0.065324858
6	0.383260072	0.708525411	-0.312930538	0.183281512	0.011435317	0.187701256
7	2.660727047	0.253026589	2.172474537	3.493116951	3.22949135	11.72485345
8	0.52789137	0.702364599	-0.431021499	0.310333858	0.017838708	0.352241866
9	0.115918446	0.773709803	-0.094647015	0.01778812	0.047502774	0.063627278
10	2.948883722	0.208526807	2.407753476	5.563144948	0.641372759	11.076471

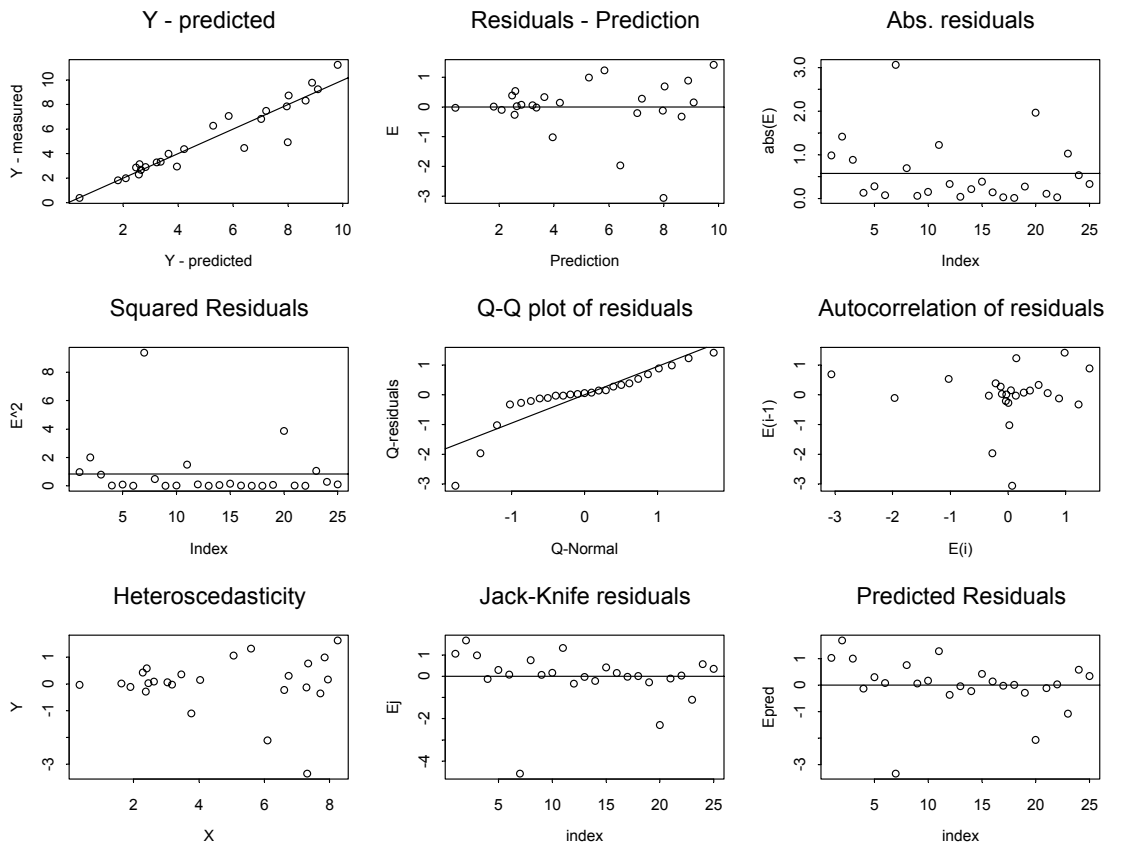
pK_a PALLAS example

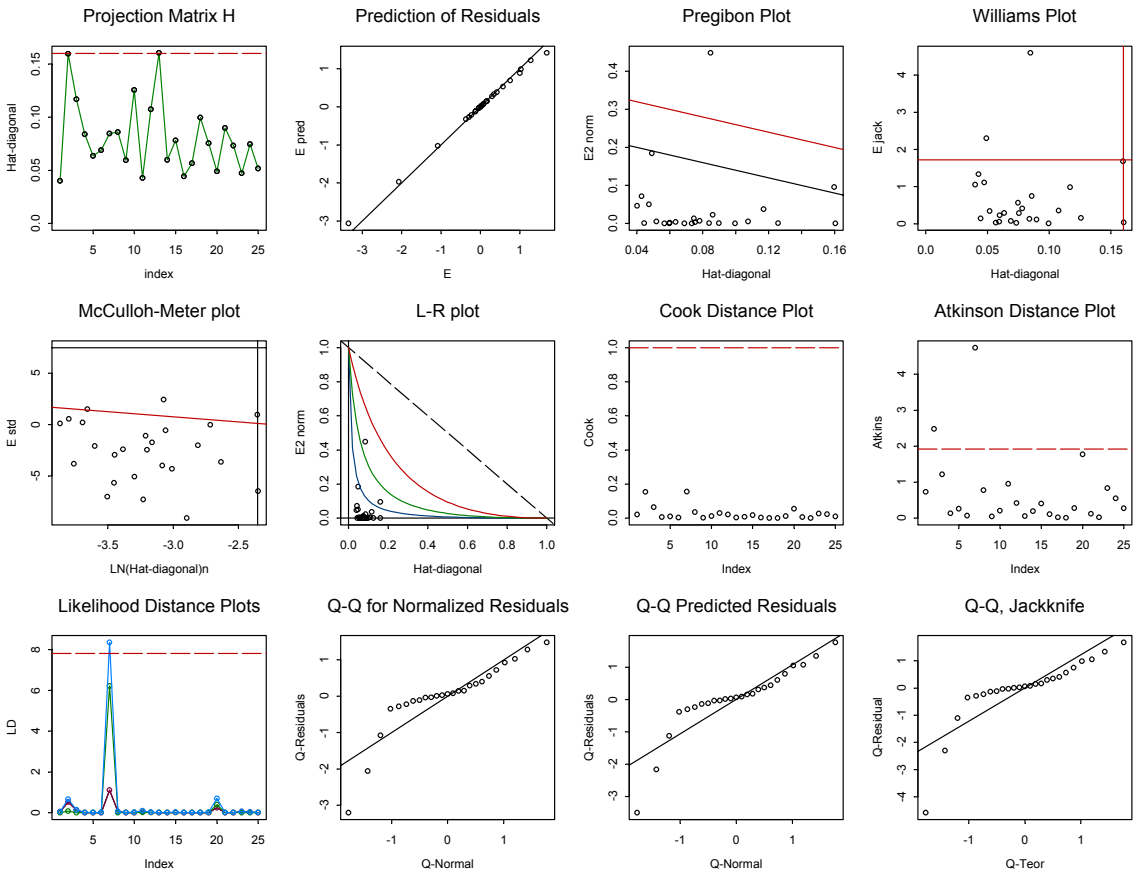
Data:

```
###### DATA SET    pK PALLAS ########

xx_as.matrix(c(5.45,10.23,9.25,8.28,7.48,2.86,8.31,8.35,3.29,9.47,6.05,9,0.33,7.3,2
    .5,4.34,3.44,1.8,2.6,6.65,2.1,2.68,4.06,2.63,3.74))
dimnames(xx)_list(NULL,"pKexp")
y_c(6.26,11.23,9.77,7.84,7.48,2.89,4.93,8.72,3.28,9.24,7.07,8.32,0.38,6.82,2.86,4.3
    6,3.34,1.82,2.3,4.45,1.99,2.67,2.93,3.13,3.98)

yname_"pK(Pallas)"

###### END OF DATA SET    pK PALLAS ########
```

Graphical Output and Diagnostics:




Y - predicted
Y - measured
Y - predicted
Residuals - Prediction
E
Prediction
Abs. residuals
abs(E)
Index
Squared Residuals
E^2
Index
Q-Q plot of residuals
Q-residuals
Q-Normal
Autocorrelation of residuals
E(i-1)
E(i)
Heteroscedasticity
Y
X
Jack-Knife residuals
Ej
index
Predicted Residuals
Epred
index


Projection Matrix H
Hat-diagonal
index
Prediction of Residuals
E pred
E
Pregibon Plot
E2 norm
Hat-diagonal
Williams Plot
E jack
Hat-diagonal
McCulloh-Meter plot
E std
LN(Hat-diagonal)n
L-R plot
E2 norm
Hat-diagonal
Cook Distance Plot
Cook
Index
Atkinson Distance Plot
Atkins
Index
Likelihood Distance Plots
LD
Index
Q-Q for Normalized Residuals
Q-Residuals
Q-Normal
Q-Q Predicted Residuals
Q-Residuals
Q-Normal
Q-Q, Jackknife
Q-Residual
Q-Teor

Linear Regression Results

Significance Level	0.05
Quantile t(1-alfa/2,n-m)	2.06866
Quantile F(1-alfa,m,n-m)	4.27934
Absolute Term:	YES
No of Rows	25
No of Parameters	2
Method	Classical Least Squares
Columns	
	NazevSloupce 1
Transformation	No transormation

Section 1. Basic Analysis

Variable Characteristics

Variable	Mean	StdDev	Corr. vs. Y	Significance
pKexp	5.2876	2.91580	0.94754	0.00000

Section 2. Pair Correlations (Xi, Xj)

Variable	Corr.Coeff	Significance
(Abs) - pKexp	0	1

Section 3. Multicollinearity Detection

Variable	Eigenvalue kor. m.	Condition number kappa	VI factor	Multiple corr.
(Abs)	1	1	1	0
pKexp	1	1	1	0

Section 4. Analysis of Variance

Mean Y	5.1224		
Source	Square Sum	Mean Square	Variance
Overall Variability	204.95126	8.19805	8.53964
Explained Variability	184.01286	7.36051	7.66720
Residual variability	20.93840	0.83754	0.87243
F-statistic	202.13080		
Critical F-quantile (1-alfa, m-1, n-m)	4.27934		
p-value	0.00000		
Conclusion	Model is Significant		

Section 5. Parameter Estimates

Variable	Estimate	StdDev.	Conclusion	p-value	Lower limit	Upper limit
(Abs)	0.10107	0.40144	Insignificant	0.80346	-0.72938	0.93151
pKexp	0.94964	0.06680	Significant	0.00000	0.81147	1.08782

Section 6. Statistical Characteristics

Multiple Correlation Coefficient R	0.94754
Determination Ceofficient R^2	0.89784
Predicted Correlation Coefficient Rp	0.87901
Mean Error of Prediction MEP	0.99190
Akaike Information Criterion	-0.43227

Section 7. Analysis of Classical Residuals

Index	Y measured	Y predicted	StdDev. Y	Residual	Rel. Residual [%Y]

1	6.26	5.27662	0.19113	0.98338	15.70891
2	11.23	9.81592	0.38131	1.41408	12.59202
3	9.77	8.88527	0.32629	0.88473	9.05562
4	7.84	7.96411	0.27634	-0.12411	-1.58306
5	7.48	7.20440	0.24054	0.27560	3.68453
6	2.89	2.81705	0.25042	0.07295	2.52434
7	4.93	7.99260	0.27780	-3.06260	-62.12173
8	8.72	8.03059	0.27974	0.68941	7.90611
9	3.28	3.22539	0.23285	0.05461	1.66485
10	9.24	9.09419	0.33832	0.14581	1.57806
11	7.07	5.84641	0.19750	1.22359	17.30682
12	8.32	8.64785	0.31290	-0.32785	-3.94056
13	0.38	0.41445	0.38219	-0.03445	-9.06566
14	6.82	7.03346	0.23342	-0.21346	-3.12994
15	2.86	2.47517	0.26662	0.38483	13.45542
16	4.36	4.22252	0.20105	0.13748	3.15325
17	3.34	3.36784	0.22725	-0.02784	-0.83352
18	1.82	1.81042	0.30114	0.00958	0.52611
19	2.3	2.57014	0.26200	-0.27014	-11.74519
20	4.45	6.41619	0.21141	-1.96619	-44.18413
21	1.99	2.09532	0.28592	-0.10532	-5.29235
22	2.67	2.64611	0.25836	0.02389	0.89473
23	2.93	3.95662	0.20770	-1.02662	-35.03816
24	3.13	2.59863	0.26063	0.53137	16.97672
25	3.98	3.65273	0.21703	0.32727	8.22280

Residual Sum of Squares	20.93840
Mean of Absolute Residuals	0.57269
Residual StdDev	0.95413
Residual Variance	0.91037
Residual Skewness	2.54619
Residual Variance	6.39714

Section 8. Regression Triplet Analysis

Fisher-Snedecor test of Significance	
F-Statistic	202.13080
Critical F- (1-alfa, m-1, n-m)	4.27934
p-value	0.00000
Conclusion	Model is Significant

Scott Criterion of Multicollinearity	
SC Statistic	202.19418
Conclusion	Multicollinearity is too high, model is incorrect!

Cook-Weisberg test of heteroscedasticity	
CW-statistic	6.60005
Critical quantile Chi^2(1-alfa,1)	3.84146
p-value	0.01020
Conclusion	Residuals are heteroscedastic!

Jarque-Berra test of normality	
JB-statistic	22.63055
Critical quantile Chi^2(1-alfa,2)	5.99146
p-value	0.00001
Conclusion	Residual are NOT normally distributed!

Wald test of autocorrelation	
WA-statistic	0.00564
Critical quantile Chi^2(1-alfa,1)	3.84146
p-value	0.94014
Conclusion	Autocorrelation is insignificant

Durbin-Watson test of autocorrelation	
DW-statistic	1.91736
Conclusion	Residuals are not correlated.

Sign test of fit	
Sgn-statistic	0.21320
Critical Quantile N(1-alfa/2)	1.95996
p-value	0.83117
Conclusion	There is no trend in residuals.

Section 9. Influential Points Diagnostics

Subsection A. Residual Analysis

Index	Standard	Jackknife	Predicted	Diag(Hii)	Diag(H*ii)	Cook Dist.
1	1.0520	1.0545	1.0245	0.0401	0.0863	0.0220
2	1.6168	1.6796	1.6829	0.1597	0.2552	0.1537
3	0.9868	0.9862	1.0019	0.1169	0.1543	0.0653
4	-0.1359	-0.1330	-0.1355	0.0839	0.0846	-0.0062
5	0.2985	0.2925	0.2943	0.0636	0.0672	0.0101
6	0.0792	0.0775	0.0784	0.0689	0.0691	0.0029
7	-3.3552	-4.5925	-3.3463	0.0848	0.5327	-0.1554
8	0.7558	0.7485	0.7542	0.0860	0.1087	0.0355
9	0.0590	0.0577	0.0581	0.0596	0.0597	0.0019
10	0.1634	0.1599	0.1668	0.1257	0.1267	0.0118
11	1.3108	1.3327	1.2784	0.0428	0.1144	0.0293
12	-0.3637	-0.3568	-0.3674	0.1075	0.1127	-0.0219
13	-0.0394	-0.0385	-0.0410	0.1605	0.1605	-0.0038
14	-0.2307	-0.2259	-0.2271	0.0598	0.0620	-0.0073
15	0.4201	0.4124	0.4174	0.0781	0.0852	0.0178
16	0.1474	0.1442	0.1439	0.0444	0.0453	0.0034
17	-0.0300	-0.0294	-0.0295	0.0567	0.0568	-0.0009
18	0.0106	0.0103	0.0106	0.0996	0.0996	0.0006
19	-0.2944	-0.2885	-0.2922	0.0754	0.0789	-0.0120
20	-2.1132	-2.3024	-2.0677	0.0491	0.2337	-0.0546
21	-0.1157	-0.1132	-0.1157	0.0898	0.0903	-0.0057
22	0.0260	0.0254	0.0258	0.0733	0.0734	0.0010
23	-1.1024	-1.1078	-1.0777	0.0474	0.0977	-0.0274
24	0.5789	0.5704	0.5742	0.0746	0.0881	0.0233
25	0.3522	0.3454	0.3451	0.0517	0.0569	0.0096

Subsection B. Influence Analysis

Index	Atkinson Dist.	Andrews-Pregibon stat	Infl on Y^	Infl on parameters LD(b)	Infl on variance LD(s)	Total Influence LD(b,s)
1	0.73120	0.91369	0.21562	0.05024	0.00090	0.05162
2	2.48318	0.74478	0.73225	0.53431	0.08181	0.66675
3	1.21703	0.84567	0.35888	0.13977	0.00007	0.14058
4	0.13645	0.91538	-0.04024	0.00184	0.01975	0.02152
5	0.25841	0.93282	0.07620	0.00657	0.01685	0.02318
6	0.07149	0.93086	0.02108	0.00050	0.02028	0.02076
7	4.73965	0.46727	-1.39765	1.10838	6.22193	8.35293

8	0.77843	0.89134	0.22955	0.05832	0.00306	0.06054
9	0.04926	0.94030	0.01453	0.00024	0.02040	0.02063
10	0.20569	0.87326	0.06065	0.00418	0.01940	0.02342
11	0.95625	0.88565	0.28198	0.08347	0.01716	0.10390
12	0.41998	0.88732	-0.12385	0.01732	0.01518	0.03191
13	0.05714	0.83949	-0.01685	0.00032	0.02048	0.02079
14	0.19330	0.93798	-0.05700	0.00368	0.01830	0.02184
15	0.40702	0.91484	0.12002	0.01624	0.01356	0.02928
16	0.10543	0.95470	0.03109	0.00110	0.01962	0.02067
17	0.02444	0.94323	-0.00721	0.00006	0.02051	0.02057
18	0.01167	0.90038	0.00344	0.00001	0.02055	0.02056
19	0.27940	0.92111	-0.08239	0.00768	0.01694	0.02435
20	1.77411	0.76627	-0.52316	0.24938	0.40642	0.70499
21	0.12056	0.90967	-0.03555	0.00144	0.01997	0.02135
22	0.02427	0.92665	0.00716	0.00006	0.02052	0.02058
23	0.83790	0.90228	-0.24708	0.06562	0.00228	0.06888
24	0.54924	0.91190	0.16196	0.02936	0.00847	0.03709
25	0.27362	0.94315	0.08069	0.00736	0.01549	0.02259