**Stat 5100 Handout 3.3.1 – SAS: Influential Observations and Outliers**

Example: Data collected on 50 countries relevant to a cross-sectional study of a life-cycle savings hypothesis, which states that the response variable

* SavRatio: aggregate personal saving divided by disposable income

can be explained by the following four predictor variables:

* AvIncome: per-capita disposable income, in USD (yearly average over decade)
* GrowRate: percentage growth rate in per-capita disposable income (over decade)
* PopU15: percentage of the population less than 15 years old (yearly average over decade)
* PopO75: percentage of the population over 75 years old (yearly average over decade)

The decade is 1960-1970. These data are published in section 2.2 of *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity* (1980) by Belsley, Kuh, and Welsch (limited excerpt available through Google books).

**/\* Define options \*/**

**ods html image\_dpi=300 style=journal;**

**/\* Read in the data \*/**

**proc import out=work.savings dbms=csv replace**

**datafile =**

**"<file path here>”**

**getnames=yes;**

**datarow=2;**

**run;**

**/\* Look at a regression model to predict SavRatio,**

**with diagnostics for influential obs. and outliers \*/**

**proc reg data = savings**

**plots(label)=(CooksD RStudentByLeverage DFFITS DFBETAS);**

**id Country;**

**model SavRatio = PopU15 PopO75 AvIncome GrowRate /**

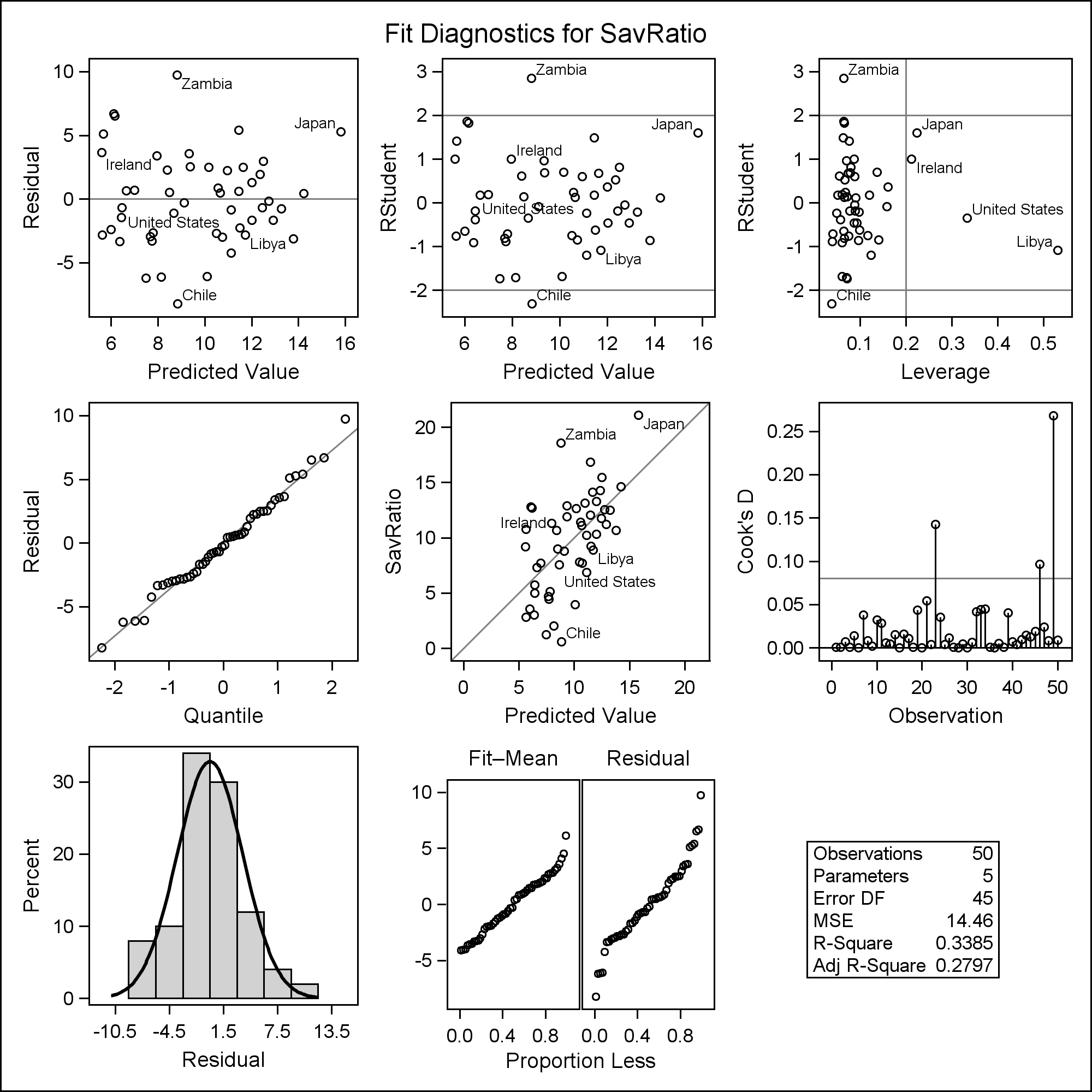
**partial partialdata;**

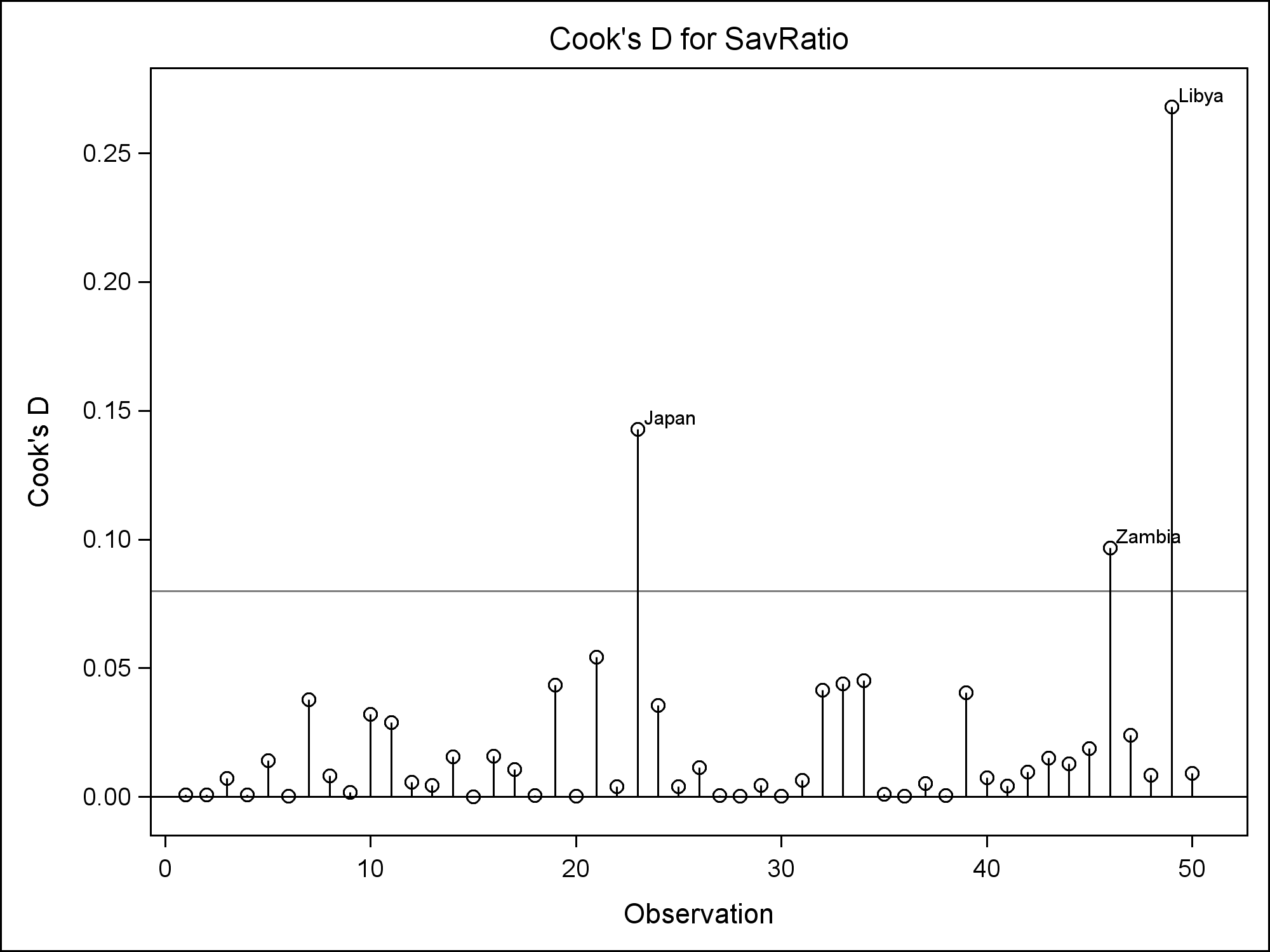
**output out=out1 r=resid p=pred;**

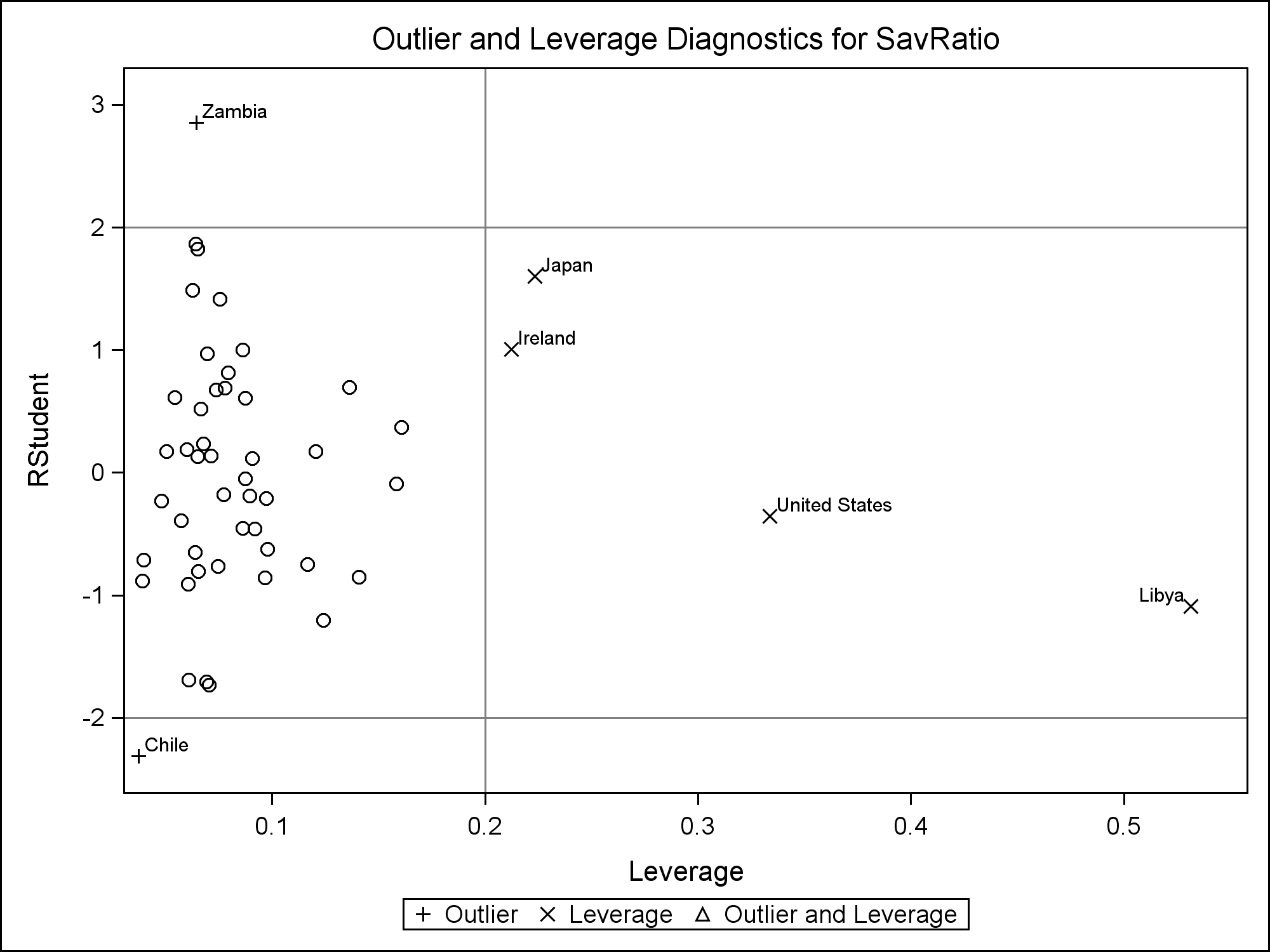
**title1 'Predict SavRatio';**

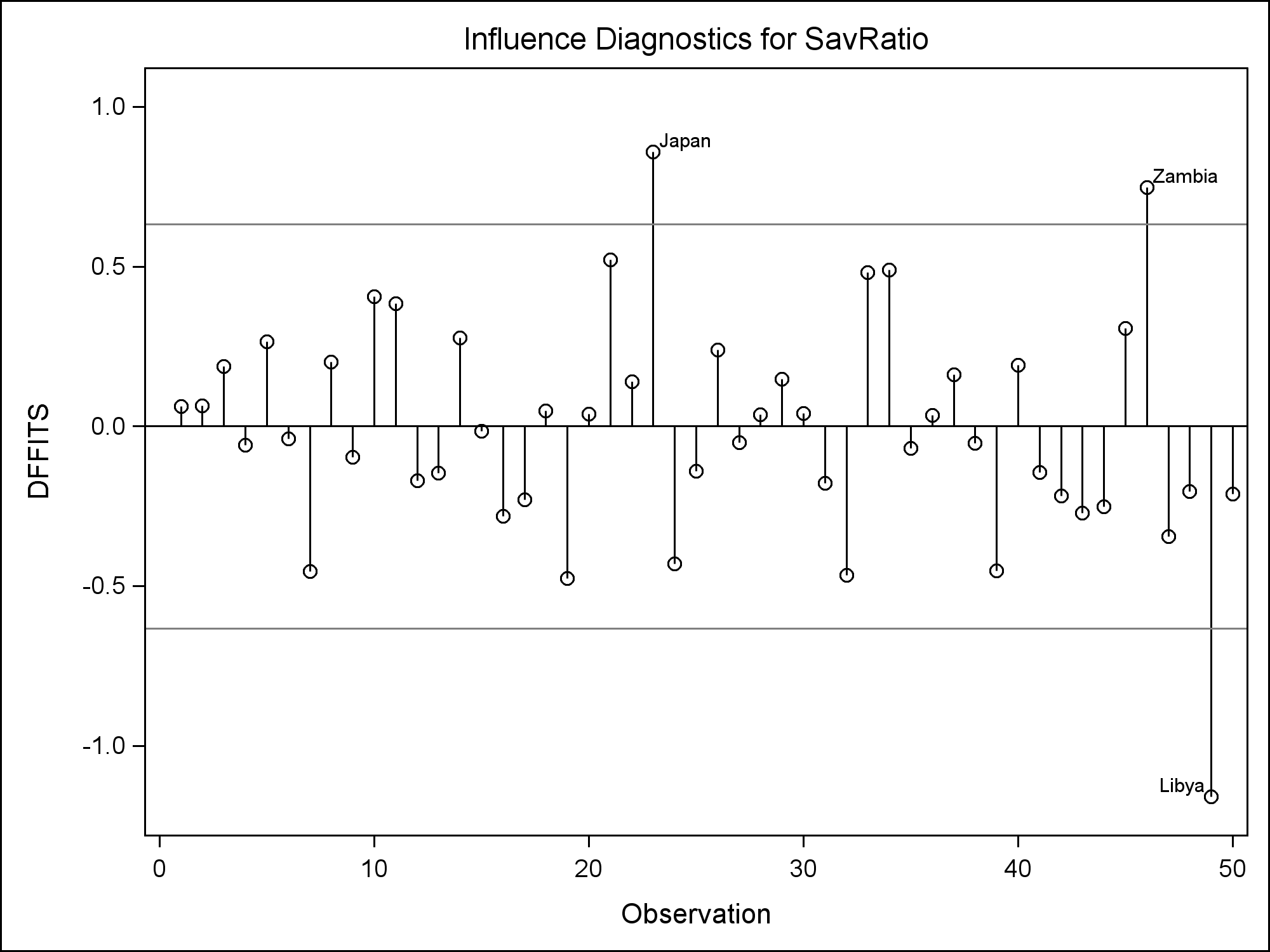
**run;**

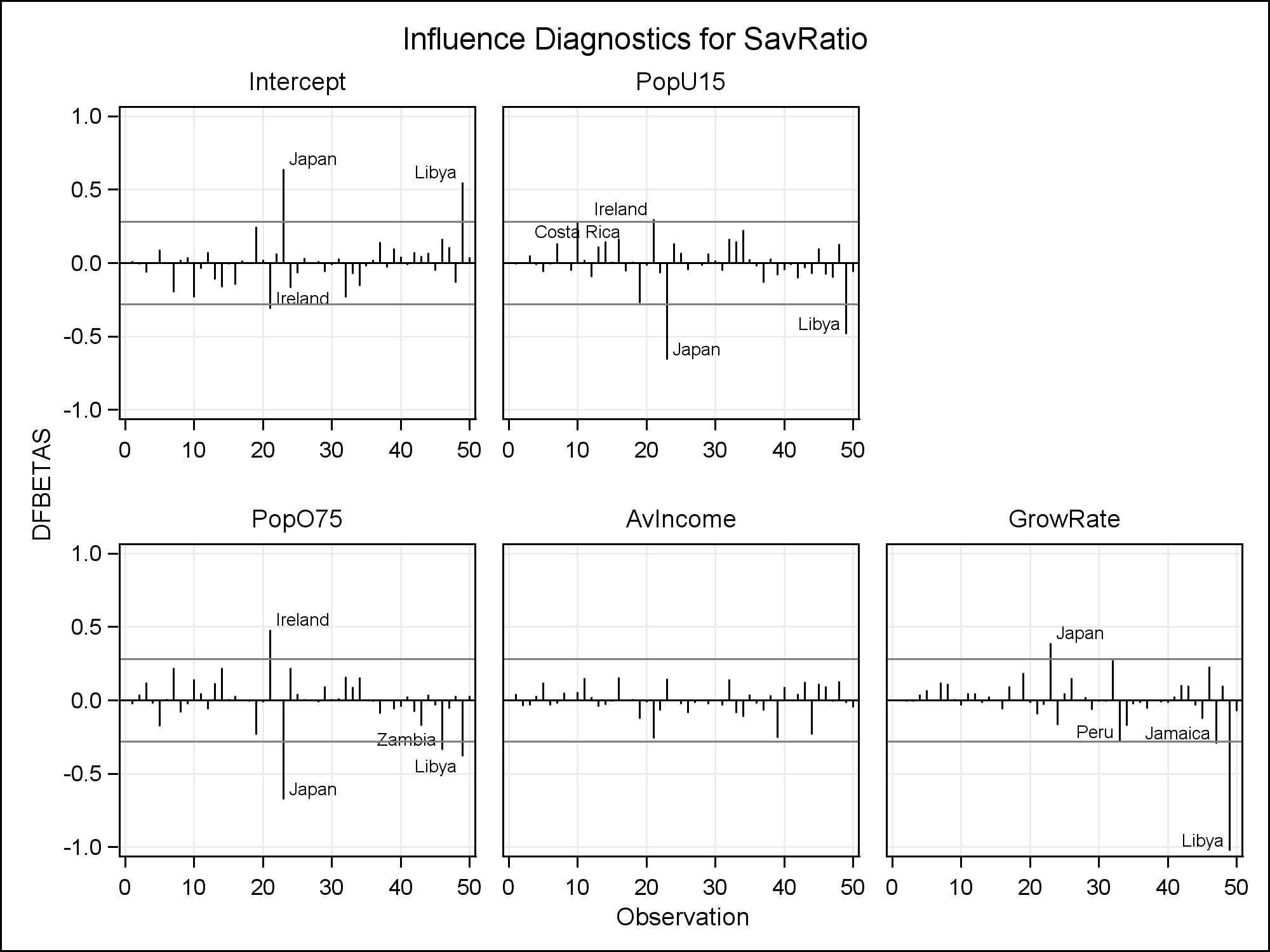
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  | | --- | | ***Predict SavRatio*** |      |  |  | | --- | --- | | **Number of Observations Read** | 50 | | **Number of Observations Used** | 50 |      | **Analysis of Variance** | | | | | | | --- | --- | --- | --- | --- | --- | | **Source** | **DF** | **Sum of Squares** | **Mean Square** | **F Value** | **Pr > F** | | **Model** | 4 | 332.91525 | 83.22881 | 5.76 | 0.0008 | | **Error** | 45 | 650.71300 | 14.46029 |  |  | | **Corrected Total** | 49 | 983.62825 |  |  |  |      |  |  |  |  | | --- | --- | --- | --- | | **Root MSE** | 3.80267 | **R-Square** | 0.3385 | | **Dependent Mean** | 9.67100 | **Adj R-Sq** | 0.2797 | | **Coeff Var** | 39.32033 |  |  |      | **Parameter Estimates** | | | | | | | --- | --- | --- | --- | --- | --- | | **Variable** | **DF** | **Parameter Estimate** | **Standard Error** | **t Value** | **Pr > |t|** | | **Intercept** | **1** | 28.56609 | 7.35452 | 3.88 | 0.0003 | | **PopU15** | **1** | -0.46119 | 0.14464 | -3.19 | 0.0026 | | **PopO75** | **1** | -1.69150 | 1.08360 | -1.56 | 0.1255 | | **AvIncome** | **1** | -0.00033690 | 0.00093111 | -0.36 | 0.7192 | | **GrowRate** | **1** | 0.40969 | 0.19620 | 2.09 | 0.0425 | |

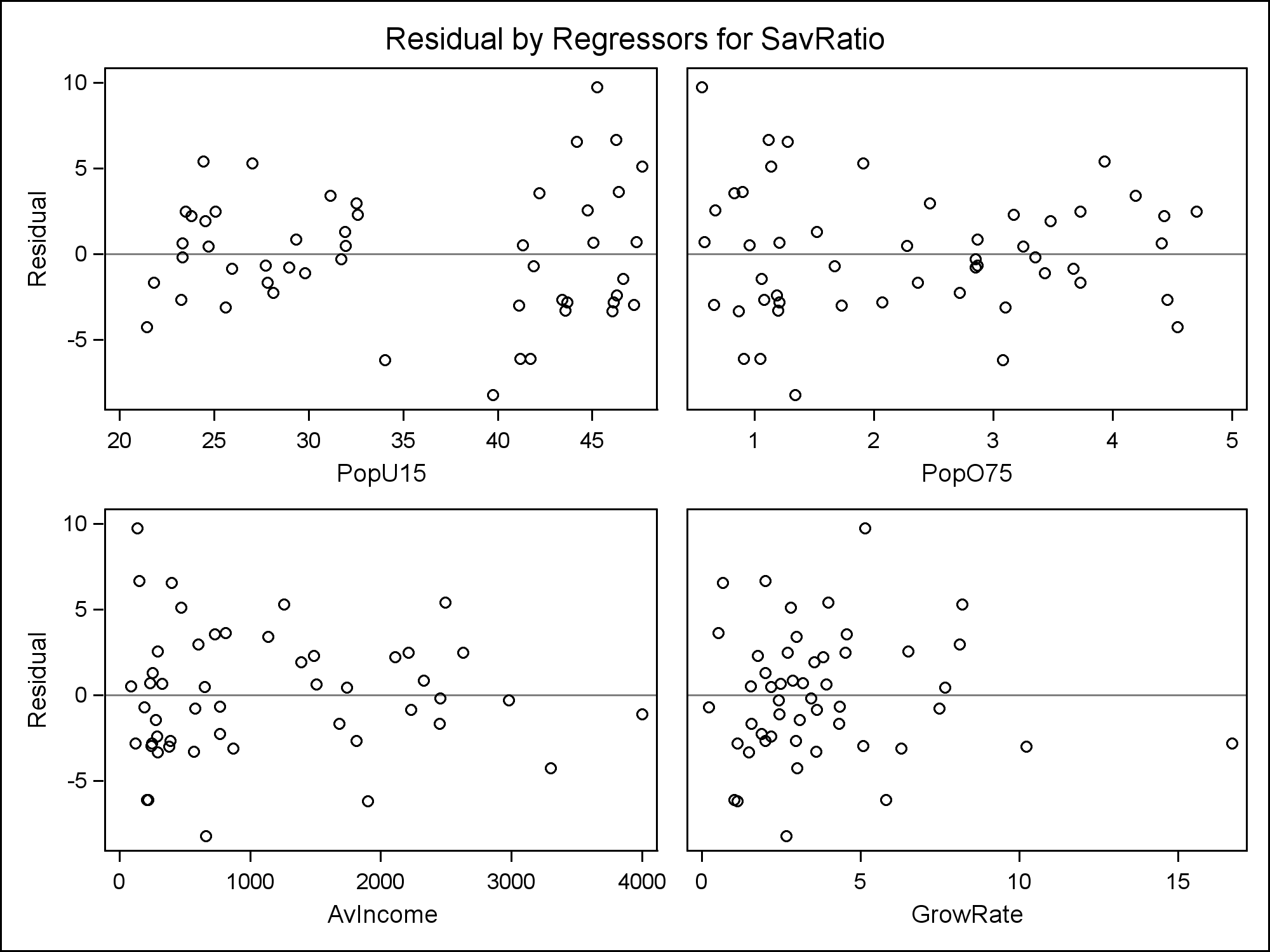


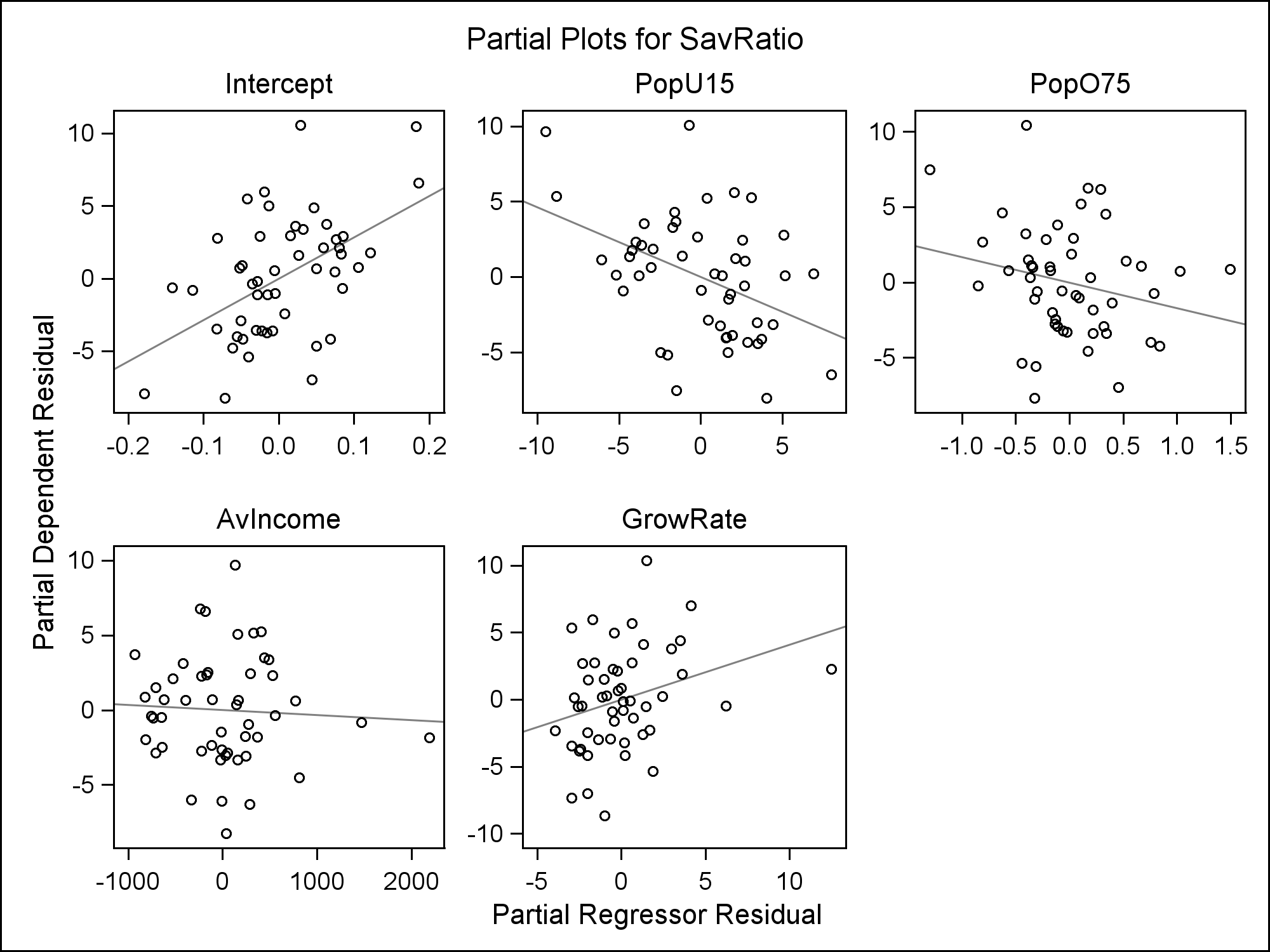












| **Obs** | **Country** | **partial PopU15** | **SavRatio partial PopU15** | **partial PopO75** | **SavRatio partial PopO75** | **partial AvIncome** | **SavRatio partial AvIncome** | **partial GrowRate** | **SavRatio partial GrowRate** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **1** | **Australia** | -1.13831 | 1.38856 | -0.38628 | 1.51696 | 768.25943 | 0.60475 | -0.01280 | 0.85833 |
| **…** |  |  |  |  |  |  |  |  |  |
| **21** | **Ireland** | 6.87903 | 0.21857 | 1.49268 | 0.86626 | -928.27477 | 3.70387 | -1.59754 | 2.73663 |
| **22** | **Italy** | -3.46588 | 3.52520 | 0.02001 | 1.89291 | -530.90583 | 2.10562 | -1.03136 | 1.50421 |
| **23** | **Japan** | -9.48247 | 9.65473 | -1.30002 | 7.48046 | 328.00826 | 5.17098 | 4.14032 | 6.97775 |
| **24** | **Korea** | -2.03533 | -5.16830 | -0.44035 | -5.36212 | -11.96560 | -6.10295 | 1.88266 | -5.33567 |
| **…** |  |  |  |  |  |  |  |  |  |
| **44** | **United States** | 4.41082 | -3.14583 | -0.30252 | -0.59987 | **2191.50614** | -1.84991 | 1.46020 | -0.51335 |
| **45** | **Venezuela** | 2.53497 | 2.46341 | -0.11300 | 3.82366 | 444.00821 | 3.48293 | -2.30789 | 2.68698 |
| **46** | **Zambia** | -0.70557 | 10.07632 | -0.40325 | 10.43302 | 130.21746 | 9.70704 | 1.49950 | 10.36525 |
| **…** |  |  |  |  |  |  |  |  |  |
| **49** | **Libya** | 7.98336 | -6.51140 | 0.83739 | -4.24597 | 49.71961 | -2.84628 | **12.47740** | 2.28240 |
| **50** | **Malaysia** | 1.93033 | -3.86112 | -0.13636 | -2.74022 | 243.12949 | -3.05278 | 1.68312 | -2.28130 |

**/\* Check other assumptions \*/**

**/\* Define shortcut macro, using line copied from**

**Course Canvas Page \*/**

**%macro resid\_num\_diag(dataset, ...**

**/\* Call shortcut macro \*/**

**%*resid\_num\_diag*(dataset=out1, datavar=resid,**

**label='Residual', predvar=pred, predlabel='Predicted');**

**run;**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  | | --- | | ***P-value for Brown-Forsythe test of constant variance*** | | ***in Residual vs. Predicted Value*** |  | **Obs** | **t\_BF** | **BF\_pvalue** | | --- | --- | --- | | **1** | 2.40263 | 0.020193 |  |  | | --- | | ***Output for correlation test of normality of Residual*** | | ***(Check text Table B.6 for threshold)*** |  | **Pearson Correlation Coefficients, N = 50  Prob > |r| under H0: Rho=0** | | | | --- | --- | --- | |  | **resid** | **expectNorm** | | |  | | --- | | **resid** | | Residual | | |  | | --- | | 1.00000 | |  | | |  | | --- | | 0.99252 | | <.0001 | | | |  | | --- | | **expectNorm** | |  | | |  | | --- | | 0.99252 | | <.0001 | | |  | | --- | | 1.00000 | |  | | |

**/\* Alternative thresholds for influential obs.**

**and outlier diagnostics \*/**

**data temp;**

**p=5; /\* p = # beta's (incl. intercept \*/**

**n = 50; /\* n = sample size \*/**

**CooksDsimple = 4/n;**

**CooksD10 = finv(.10,p,n-p);**

**CooksD20 = finv(.20,p,n-p);**

**CooksD50 = finv(.50,p,n-p);**

**RStudent95 = tinv((1-.05/2),(n-p));**

**RStudent95Bonf = tinv((1-.05/2/n),(n-p));**

**Leverage2 = 2\*p/n;**

**Leverage3 = 3\*p/n;**

**DFBETAS = 2/n\*\*0.5; if (n <= 30) then DFBETAS = 1;**

**DFFITS = 2\*(p/n)\*\*0.5; if (n <= 30) then DFFITS = 1;**

**;**

**proc print data=temp;**

**var CooksDsimple CooksD10 CooksD20 CooksD50 RStudent95**

**RStudent95Bonf Leverage2 Leverage3 DFBETAS DFFITS;**

**title1 'Alternative thresholds';**

**run;**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  | | --- | | ***Alternative thresholds*** |  | **Obs** | **CooksDsimple** | **CooksD10** | **CooksD20** | **CooksD50** | **RStudent95** | **RStudent95Bonf** | **Leverage2** | **Leverage3** | **DFBETAS** | **DFFITS** | | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | | **1** | 0.08 | 0.31729 | 0.46527 | 0.88349 | 2.01410 | 3.52025 | 0.2 | 0.3 | 0.28284 | 0.63246 | |

**/\* Now look more closely at distribution of predictors,**

**and suspect observations \*/**

**proc univariate data=savings noprint;**

**var PopU15 PopO75 AvIncome GrowRate;**

**histogram PopU15 PopO75 AvIncome GrowRate;**

**title1;**

**run;**

|  |  |
| --- | --- |
|  |  |
|  |  |

**proc print data=savings;**

**where country = 'Ireland' | country = 'Japan'**

**| country = 'United States' | country = 'Libya'**

**| country = 'Zambia';**

**var country SavRatio PopU15 PopO75 AvIncome GrowRate;**

**title1 'Suspect observations';**

**run;**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  | | --- | | ***Suspect observations*** |  | **Obs** | **Country** | **SavRatio** | **PopU15** | **PopO75** | **AvIncome** | **GrowRate** | | --- | --- | --- | --- | --- | --- | --- | | **21** | Ireland | 11.34 | 31.16 | 4.19 | 1139.95 | 2.99 | | **23** | Japan | 21.1 | 27.01 | 1.91 | 1257.28 | 8.21 | | **44** | United States | 7.56 | 29.81 | 3.43 | 4001.89 | 2.45 | | **46** | Zambia | 18.56 | 45.25 | 0.56 | 138.33 | 5.14 | | **49** | Libya | 8.89 | 43.69 | 2.07 | 123.58 | 16.71 | |

**/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\***

**Possible Remedial Measures for these data:**

**Drop Japan**

**-- PopU15 and PopO75 don't match profile**

**(influential but not outliers)**

**Take log of AvIncome and log of GrowRate**

**-- their distributions are skew right**

**-- the extreme observation in each is a suspect obs.**

**(United States for AvIncome,**

**Libya for GrowRate)**

**\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/**

**/\* Create new data set and fit regression model;**

**check assumptions \*/**

**data newsavings; set savings;**

**if country ne 'Japan';**

**logAvIncome = log(AvIncome);**

**logGrowRate = log(GrowRate);**

**run;**

**proc reg data = newsavings**

**plots(label)=(CooksD RStudentByLeverage DFFITS DFBETAS);**

**id Country;**

**model SavRatio = PopU15 PopO75 logAvIncome logGrowRate**

**/ partial;**

**output out=out2 r=resid p=pred;**

**title1 'Predict SavRatio';**

**title2 '(after remedial measures)';**

**run;**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  | | --- | | ***Predict SavRatio*** | | ***(after remedial measures)*** |      | **Parameter Estimates** | | | | | | | --- | --- | --- | --- | --- | --- | | **Variable** | **DF** | **Parameter Estimate** | **Standard Error** | **t Value** | **Pr > |t|** | | **Intercept** | **1** | 26.25118 | 10.52632 | 2.49 | 0.0165 | | **PopU15** | **1** | -0.33837 | 0.15791 | -2.14 | 0.0377 | | **PopO75** | **1** | -0.68558 | 1.13571 | -0.60 | 0.5492 | | **logAvIncome** | **1** | -0.71860 | 0.97492 | -0.74 | 0.4650 | | **logGrowRate** | **1** | 1.33042 | 0.72528 | 1.83 | 0.0734 | |

|  |  |  |
| --- | --- | --- |
|  |  |  |
|  |  |  |
|  |  |  |

**/\* Check model assumptions \*/**

**%*resid\_num\_diag*(dataset=out2, datavar=resid,**

**label='New Residual', predvar=pred,**

**predlabel='New Predicted Value');**

**run;**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  | | --- | | ***P-value for Brown-Forsythe test of constant variance*** | | ***in New Residual vs. New Predicted Value*** |  | **Obs** | **t\_BF** | **BF\_pvalue** | | --- | --- | --- | | **1** | 2.43339 | 0.018815 |  |  | | --- | | ***Output for correlation test of normality of New Residual*** | | ***(Check text Table B.6 for threshold)*** |      | **Pearson Correlation Coefficients, N = 49  Prob > |r| under H0: Rho=0** | | | | --- | --- | --- | |  | **resid** | **expectNorm** | | |  | | --- | | **resid** | | New Residual | | |  | | --- | | 1.00000 | |  | | |  | | --- | | 0.99516 | | <.0001 | | | |  | | --- | | **expectNorm** | |  | | |  | | --- | | 0.99516 | | <.0001 | | |  | | --- | | 1.00000 | |  | | |

**/\* Look at final model \*/**

**proc reg data = newsavings;**

**model SavRatio = PopU15 logGrowRate;**

**title1 'Final Model';**

**run;**

| **Parameter Estimates** | | | | | |
| --- | --- | --- | --- | --- | --- |
| **Variable** | **DF** | **Parameter Estimate** | **Standard Error** | **t Value** | **Pr > |t|** |
| **Intercept** | **1** | 14.27955 | 2.40166 | 5.95 | <.0001 |
| **PopU15** | **1** | -0.18046 | 0.05915 | -3.05 | 0.0038 |
| **logGrowRate** | **1** | 1.45209 | 0.71058 | 2.04 | 0.0468 |

**/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\***

**What if want to add your own reference lines?**

**\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/**

**proc reg data = savings ;**

**id country;**

**model SavRatio = PopU15 PopO75 AvIncome GrowRate /**

**influence;**

**ods output outputstatistics=out3;**

**run;**

**proc print data=out3;**

**run;**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| *Observation* | *Country* | *Residual* | *RStudent* | *HatDiagonal* | *CovRatio* | *DFFITS* | *…* | *DFB\_GrowRate* |
| 1 | Australi | 0.8636 | 0.2327 | 0.0677 | 1.1928 | 0.0627 |  | -0.0002 |
| 2 | Austria | 0.6164 | 0.1710 | 0.1204 | 1.2678 | 0.0632 |  | -0.0082 |
| 3 | Belgium | 2.2190 | 0.6066 | 0.0875 | 1.1762 | 0.1878 |  | -0.0073 |
| … |  |  |  |  |  |  |  |  |

**proc sgplot data=out3;**

**scatter x=HatDiagonal y=RStudent / markerchar=country;**

**xaxis label='Leverage';**

**yaxis label='Studentized Deleted Residual';**

**refline 2.01 / axis=Y lineattrs=(pattern=2);**

**refline 3.52 / axis=Y;**

**refline .2 / axis=X lineattrs=(pattern=2);**

**refline .3 / axis=X;**

**run;**

