Kiraci, Arzdar

Article

Confirmation, Correction and Improvement for Outlier Validation Using Dummy Variables: t-Statistics or F-Incremental Statistics is not enough in OLS

International Econometric Review (IER)

Provided in Cooperation with:
Econometric Research Association (ERA), Ankara

Suggested Citation: Kiraci, Arzdar (2013) : Confirmation, Correction and Improvement for Outlier Validation Using Dummy Variables: t-Statistics or F-Incremental Statistics is not enough in OLS, International Econometric Review (IER), ISSN 1308-8815, Econometric Research Association (ERA), Ankara, Vol. 5, Iss. 2, pp. 43-52

This Version is available at:
https://hdl.handle.net/10419/238806

Confirmation, Correction and Improvement for Outlier Validation using Dummy Variables

Arzdar Kiraci®

Siirt University

ABSTRACT

Dummy variables can be used to detect, validate and measure the impact of outliers in data. This paper uses a model to evaluate the effectiveness of dummy variables in detecting outliers. While generally confirming some findings in the literature, the model refutes the presumption that the *t*-statistic or the *F*-incremental statistic is enough to validate an observation as an outlier. In order to rectify this fallacy, this paper recommends an easily-calculable robust standardized residual statistic that is more compatible with the definition of outliers.
The robust standardized residual statistic suggested herein is still used in many robust regression methods and is more effective than the *t*-statistic or the *F*-incremental statistic in validating outliers with dummy variables. The results of this study suggest some practical recommendations for dealing with outliers and improvements in maintaining the integrity of data. We recommend all previous studies using this statistics be revised in light of the findings presented in this paper.

Key words: *Dummy Variable, t-Statistic, Outlier, Robust Dummy Statistic, Robust Standardized Residual*
JEL Classifications: C2, C20, C51, C52
AMS 2000 subject classifications: Primary 60K35, 60K35; secondary 60K35

1. INTRODUCTION

Outliers are defined as observations that do not obey the (linear) pattern formed by the majority of the observations, and if they are influential they give rise to misleading results (Rousseeuw and Van Aelst, 1999). During regression analysis, a dummy variable (DV) or indicator variable is introduced for an observation that is suspected to be affected by different variables other than the ones in the model. In some cases these suspicious observations may turn out to be significant outliers.

The standard in the literature has been to assign a DV to each outlier observation, which is then validated as an outlier using the *t*-statistic of the DV. If a search is performed in academic journal databases (for example, in Business Source Complete) thousands of studies are found in 2010 that identify outliers using the *t*-statistic of the DV.

This paper investigates the following questions: If there are many reasons for outliers like an economic shock, a technological breakthrough, a natural disaster, even an incorrect recording of observation, or, alternatively stated, if an outlying observation is created under the influence of a rare occurring and unpredictable incident (variable) then can a DV represent

® Arzdar Kiraci, Siirt University, Faculty of Economics and Administrative Sciences, Gures Caddesi 56100 Siirt/Turkey, (email: arzdarkiraci@siirt.edu.tr, arzdar.kiraci@gmail.com), Tel: +90 (484) 223 12 24 - 223 17 39 - 224 11 38, Fax: +90 (484) 223 19 98.

this event? If a DV is used for each outlier, then how are the statistics in Ordinary Least Squares (OLS) affected? Is the *t*-statistic of DVs or the *F*-incremental statistic successful in identifying these outliers? Is there a statistic that is easier to calculate? This paper answers these questions.

As noted by Greene (2002:117), if a DV is used for an observation it has the effect of deleting that observation from the computation of OLS parameters and standard error (SE). Studenmund (2002:224) has, in accordance, advanced evidence that the dummy variable coefficient is equal to the studentized residual for that observation, which is also proven in the following section. This is an advantage because many robust regression techniques identify outliers by deleting suspicious observations from the data according to a selected criterion. One important point that is not emphasized in the literature is that the ratio of the dummy coefficient to the SE of the regression is the robust standardized residual (*RSR*). *RSR* is also referred to as the deleted studentized residual, externally studentized residual or jackknifed residual (Rousseeuw and Leroy, 1987:226), which is still used in many robust regression methods. However, the *t*-statistic, which will be shown to be unsuccessful, is still widely used in the literature. As an alternative to the *t*-statistic this paper recommends the use of the dummy coefficient to SE ratio as a new robust *RSR* statistic. This statistic is easier to calculate than Cook's distance (which measures the change in the estimates that result from deleting each observation, Cook, 1985), DFFITS (which is the predicted value for a point, obtained when that point is left out of the regression, Billingsley et.al., 1980) or DFBETAS (which is the scaled measure of the change in each parameter estimate and is calculated by deleting the i^{th} observation, Billingsley et.al., 1980).

The next section explains the notation and the model used in this paper to derive an important statistic based on DVs. The third section introduces this new statistic, and the fourth section presents an example, through which the insufficiency of the *t*-statistic of the DV or the *F*-incremental statistic is illustrated. Finally, the conclusion part summarizes the findings of this paper with suggestions for future research.

2. THE MODEL

For the given data let there be n observations with $v+1$ independent variables and $n-m$ suspicious observations be represented by $n-m$ different DVs. Let $\mathbf{Y}'=[y_1 \dots y_m]_{1\times m}$ be the dependent variable vector for the observations without DVs, $\mathbf{Y}'_d=[y_{m+1} \dots y_n]_{1\times(n-m)}$ be the dependent variable vector for the observations with DVs, $\mathbf{x}_i=[1\, x_{i,1} \dots x_{i,v}]_{1\times(v+1)}$ be the independent variable vector for i^{th} observation, $\mathbf{X}'=[\mathbf{x}'_1 \dots \mathbf{x}'_m]_{(v+1)\times m}$ be the independent variable matrix for the observations without DVs, and $\mathbf{X}'_d=[\mathbf{x}'_{m+1} \dots \mathbf{x}'_n]_{(v+1)\times(n-m)}$ be the independent variable matrix for the observations with DVs. Let $\mathbf{0}_{m\times(n-m)}$ be the zero matrix, $\boldsymbol{\varepsilon}_{m\times 1}$ be the error vector, $\boldsymbol{\varepsilon}_d$ be the $(n-m)\times 1$ error vector that contains the outlier information, and $\mathbf{I}_{(n-m)\times(n-m)}$ be the unit matrix. Then the regression model $\overline{\mathbf{Y}}=\overline{\mathbf{X}}\,\overline{\boldsymbol{\beta}}+\overline{\boldsymbol{\varepsilon}}$ can be written as follows:

$$\begin{bmatrix}\mathbf{Y}\\ \mathbf{Y}_d\end{bmatrix}=\begin{bmatrix}\mathbf{X} & \mathbf{0}\\ \mathbf{X}_d & \mathbf{I}\end{bmatrix}\begin{bmatrix}\boldsymbol{\beta}\\ \boldsymbol{\delta}\end{bmatrix}+\begin{bmatrix}\boldsymbol{\varepsilon}\\ \boldsymbol{\varepsilon}_d\end{bmatrix} \tag{2.1}$$

Proposition 2.1. The OLS estimator for $\boldsymbol{\beta}_{(v+1)\times 1}$ and $\boldsymbol{\delta}_{(n-m)\times 1}$ is given by:

$$\begin{bmatrix}\mathbf{b}\\ \mathbf{d}\end{bmatrix}=\begin{bmatrix}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}\\ \mathbf{Y}_d-\mathbf{X}_d\mathbf{b}\end{bmatrix} \tag{2.2}$$

The residuals and other important statistics can be calculated as follows:

$$\begin{bmatrix} \mathbf{e} \\ \mathbf{e}_d \end{bmatrix} = \begin{bmatrix} \mathbf{Y} \\ \mathbf{Y}_d \end{bmatrix} - \begin{bmatrix} \mathbf{X} & \mathbf{0} \\ \mathbf{X}_d & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{b} \\ \mathbf{d} \end{bmatrix} = \begin{bmatrix} \mathbf{Y - Xb} \\ \mathbf{0} \end{bmatrix} \tag{2.3}$$

In equation (2.2) **b** is the OLS estimator of **X**, i.e.; it is not affected by the suspicious observations represented by DVs. Therefore, the DVs have the effect of omitting these observations from the computation of OLS parameters (Greeen, 2002:117). If residual values for the dummy variables are required, they can be calculated as $\mathbf{Y}_d - \mathbf{X}_d\mathbf{b} = \mathbf{d}$, namely as the coefficients of DVs in (2.2), but not as the **0** vector in (2.3) (Studenmund, 2002:224). Hence, the DV coefficients are successful in representing the information contained in suspicious observations, and their coefficient values are a measure of their influence on the OLS regression results.

The DVs have the same effect of deleting these observations from the computations of the SE ($\hat{\sigma}$) and cov(**b**). These estimators can be calculated as follows:

$$\text{var} = \hat{\sigma}^2 = \frac{\begin{bmatrix} \mathbf{e}' & \mathbf{e}_d' \end{bmatrix} \begin{bmatrix} \mathbf{e} \\ \mathbf{e}_d \end{bmatrix}}{n - (v + n - m + 1)} = \frac{\mathbf{e}'\mathbf{e}}{m - v - 1} = \frac{\mathbf{(Y - Xb)}'\mathbf{(Y - Xb)}}{m - v - 1}$$

$$\text{cov}\left(\begin{bmatrix} \mathbf{b} \\ \mathbf{d} \end{bmatrix}\right) = \hat{\sigma}^2 \begin{bmatrix} \mathbf{(X'X)}^{-1} & \mathbf{-(X'X)}^{-1}\mathbf{X}_d' \\ \mathbf{-X}_d\mathbf{(X'X)}^{-1} & \mathbf{(I + X}_d\mathbf{(X'X)}^{-1}\mathbf{X}_d') \end{bmatrix}$$

For the coefficient of determination R^2, adding a DV does not have the same effect of deleting the suspected observations from the computation as proven in the following proposition.

Proposition 2.2. Let R^2 be the coefficient of determination in a regression with DVs and R_0^2 be in a regression without any suspicious observations. Then, $R^2 > R_0^2$.[1]

3. PROPOSAL OF AN ALTERNATIVE ROBUST DUMMY STATISTIC TO THE *T*-STATISTIC OR *F*-INCREMENTAL STATISTIC

In the literature on robustness, an observation can be identified as an outlier by the robust standardized residual (*RSR*) value, which is the ratio of the residual value of that observation to the SE of the regression. Both the residual value and the SE are calculated using robust estimators that are not affected by any outliers. As mentioned previously, for a regression with DVs, the ratio of the dummy coefficient to SE of the regression gives the *RSR* for that observation. Alternatively, for observation *j* this value can be written as follows:

$$RSR_j = \frac{d_j}{\hat{\sigma}} \tag{3.4}$$

This RSR_j is the robust dummy statistic suggested by this paper. For an outlier observation this RSR_j of a DV is compared with the t_j value of a DV in the following proposition.

Proposition 3.3. The *t*-statistic of a DV is unable to validate outliers with large **X** independent variable values.

[1] While equality is possible, it would not make the observations suspicious.

According to the proof of proposition 3.3, provided in the Appendix, it is impossible to assign critical values to the calculated *t*-values of DVs, because they will depend on $\mathbf{X}_d$ or outliers. The same conclusion can be stated for the *F*-incremental statistic as shown in the following proposition.

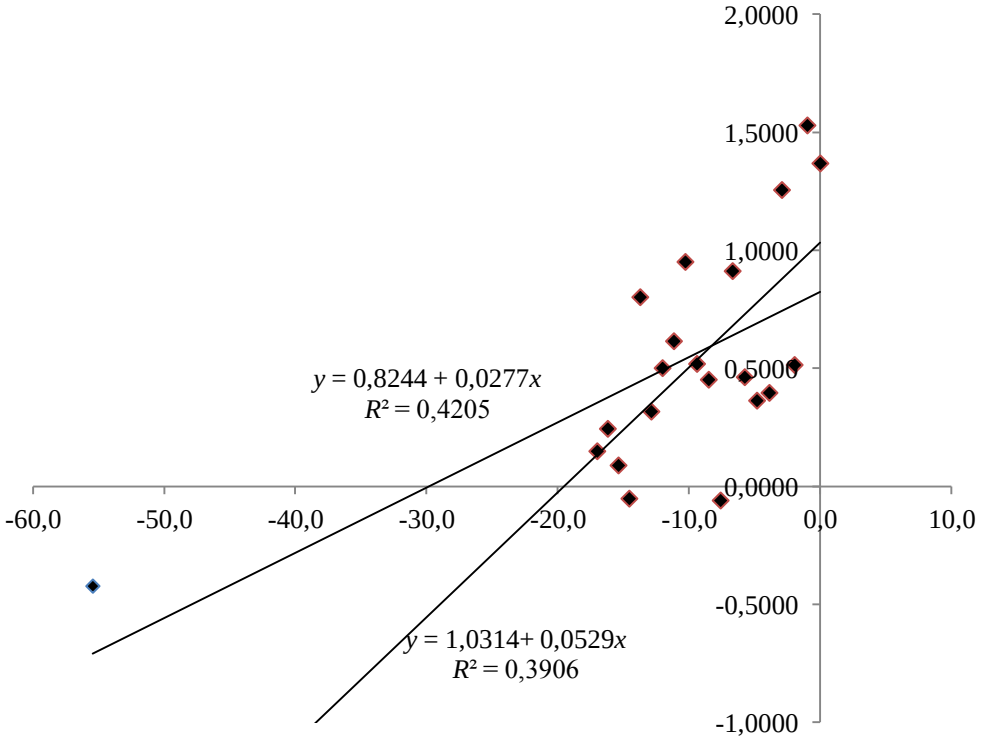
Proposition 3.4. The *F*-incremental statistic is unable to validate outliers with the average of $\mathbf{Y}_d$ dependent variable values of outliers close to the average of **Y** dependent variable values of other observations.

In the following part, the problems that may arise as a result of using the *t*-statistic or the *F*-incremental statistic instead of the *RSR* value are illustrated with an example.

4. ILLUSTRATIVE EXAMPLE

For the simple regression model $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ ($\varepsilon_i \sim N(0,\sigma^2)$), using a software program written in Octave, different values for the parameters β_0, β_1, σ^2 are generated, and an outlier is placed.

Figure 4.1 Plot for the data in Table 4.1.



Obs.	y	x	Obs.	y	x	Obs.	y	x
1	1.368872	0.005	8	0.912649	-6.680	15	0.317508	-12.875
2	1.529872	-0.980	9	-0.059299	-7.595	16	0.802124	-13.720
3	0.514864	-1.955	10	0.452557	-8.500	17	-0.051358	-14.555
4	1.256283	-2.920	11	0.519466	-9.395	18	0.089404	-15.380
5	0.396734	-3.875	12	0.951551	-10.280	19	0.244413	-16.195
6	0.363789	-4.820	13	0.615652	-11.155	20	0.149377	-17.000
7	0.463096	-5.755	14	0.501619	-12.020	21	-0.422100	-55.450

Table 4.1 Data for the example.

The data in Table 4.1 is generated using $y_i = 1 + 0.049307299\, x_i + \varepsilon_i\, (\varepsilon_i \sim N(0,0.3^2))$. The last ($21^{st}$) observation is an outlier as illustrated in Figure 4.1, where the OLS estimator deviates considerably from the linearity indicated by the majority of observations. Important statistics are presented in Table 4.2 along with the outlier, in Table 4.3 without the outlier, and in Table 4.4 where a DV is used for the outlier observation.

For the regression results with the outlier included in the data in Table 4.2, all of the coefficients are statistically significant. In Table 4.2 the Durbin-Watson statistic is 1.93,thus there is no sign of autocorrelation or model specification error. However, the results are biased, because the data contains an outlier. The 99% confidence interval for the slope estimator in Table 4.2 is (0.006349, 0.048997), and it does not contain the true population parameter 0.049307, while the 99% confidence interval for the slope estimator in Table 4.3 is (0.008339, 0.097426). For the regression results without the outlier in the data in Table 4.3, again, all of the coefficients are statistically significant. The Durbin-Watson statistic is 2.17, hence there may not be an autocorrelation or model specification error.

	Coefficient	**SE**	***t* stat**	***p*-value**
Intercept	0.824398	0.116728	7.0625483	1.01E-06
Slope	0.027673	0.007454	3.7127689	1.47E-03
R^2	0.420461		DW *d*	1.9342573
Std. Error	0.380587			
Obs.	21			
F	13.784653			

Table 4.2 Regression results with the outlier.

	Coefficient	**SE**	***t* stat**	***p*-value**
Intercept	1.031396	0.1586173	6.5024186	4.09E-06
Slope	0.052882	0.0155696	3.3965079	3.22E-03
R^2	0.390580		DW *d*	2.1729233
Std. Error	0.359492			
Obs.	20			
F	11.5362659			

Table 4.3 Regression results without the outlier.

	Coefficient	**SE**	***t* stat**	***p*-value**
Intercept	1.031396	0.1586173	6.5024185	4.09E-06
Slope	0.052882	0.0155696	3.3965079	3.22E-03
Dummy	1.478817	0.8146362	1.8153101	0.086175
R^2	0.510141			
Std. Error	0.359492		*RSR*	4.113634
Obs.	21			
F	9.372650			

Table 4.4 Regression results with a dummy variable for the outlier.

Unfortunately, the OLS regression results are not immune to masking effect of outliers (Rousseeuw and Leroy, 1987). The situation does not change if a DV is added for the suspicious observation. If the outlier is represented with a DV, the *t*-statistic of the DV is (1.82), and it is statistically insignificant at a 5% significance level. In addition, the *F*-incremental statistic is 4.39 ($F_{cr} = 4.41$ at a 5% significance level), which does not indicate the presence of an outlier. However, the *RSR* statistic, which is used to identify outliers in the literature on robust regression where an observation is identified as outlier if $RSR > 2.5$ in value (Rousseeuw and Leroy, 1987), indicates an outlier with an *RSR* value of 4.11.

As theoretically proven and as illustrated in Table 4.3 and Table 4.4, when a DV is used for an outlier, it has the effect of deleting that observation from the computation of the OLS parameters. Furthermore, adding a DV for an outlier increases the R^2 value, while the *F*-value may increase or decrease.

Using a software program written in Octave, different data are generated, and one outlier is placed in each data. In some of data the *t*-statistic of a DV or *F*-incremental statistic failed to validate outliers that were highly influential, which in certain cases even caused the slope coefficients to change from positive significant to negative significant or vice versa.

4. COMMENTS AND CONCLUSION

This paper proves that the *t*-statistic of a DV or the *F*-incremental statistic are not always effective in validating outliers; this argument is both theoretically proven and illustrated with an empirical example. Therefore, these statistics should not be used. Proposed in place of these statistics is an alternative statistic, which is consistent with the standards for outlier detection methods in the literature on robustness.

If a DV is used for an outlier observation it has the effect of deleting that observation from computation of OLS parameters and standard error (Greene, 2002). However, if the dummy variable is used in the model, then the coefficient of determination, R^2, always increases. In addition, according to proposition 3.3, it is impossible to assign critical values to the calculated *t*-value of a DV, because they will depend on $\mathbf{X}_d$ or outliers.

In the literature on robust regression the debate about whether to keep outliers in data or to remove them continues. According to the findings in this paper, instead of using outlier observations with DVs, removing the outlier (and its DV) might be more appropriate for scientific inference, because using DVs for outliers has the effect of deleting that observation from computation of OLS parameters and SE. However, as shown in the proof of proposition 2.2, included in the Appendix, using DVs for outliers always increases the R^2 statistic.

As mentioned in Studenmund (2002), the dummy's coefficient equals the residual for that observation. In the literature on robustness, *RSR*, which is used to detect outliers, is calculated by dividing the residual of the observation by the SE of the regression. This paper suggests that the ratio of the DVs coefficient to the SE of the regression be used as the method for outlier detection, because this ratio is the *RSR* of that observation. This *RSR* value, which is the robust dummy statistic suggested by this paper is easier to calculate than Cook's distance (Cook, 1985), DFFITS (Billingsley et.al., 1980), or DFBETAS (Billingsley et.al., 1980). In addition, *RSR* of a DV is a residual value that gives the magnitude of the information contained in the suspicious observations and their values are a measure of their influence on OLS results when they are included in the data without DVs.

The illustrative example provided in this paper proves that the *t*-statistic of a DV or the *F*-incremental statistics are not always successful in identifying outliers. This paper aims to provide scientists with a better alternative for outlier validation, namely the *RSR* statistic, derived from the techniques/ideas in robust regression methods. The findings of this paper both theoretically and empirically demonstrate the superiority of the *RSR* statistic to the *t*-statistic and the *F*-incremental statistic.

If there is prior information that correctly identifies a group of observations in which all of the outliers are contained then using the findings of this paper the *RSR* statistic of DV will enable the accurate detection of all outliers. Using the *RSR* statistic of DV is not an advanced technique, and is in fact easier and faster to calculate than most other robust regression techniques.

One important warning must be given about the use of DVs with outliers: due to the masking effect, the detection/validation of outliers using the *RSR* statistic of DVs works correctly only if all possible outliers are in the group of observations represented by DVs. If there is no prior information on which observations are outliers, all of the outliers have to first be identified using any of the other robust techniques for outlier detection, but these advanced techniques require a lot of computer calculation[2]. Only then can outliers be validated and their effects be analyzed or measured using the *RSR* statistic of DV.

APPENDIX

Proof of Proposition 2.1:

Without loss of generality, the places of $\mathbf{Y}$ and $\mathbf{Y}_d$ in equation (2.1) can be interchanged to make matrix inversion easier, thus the OLS estimator becomes:

$$\begin{bmatrix}\mathbf{d}\\ \mathbf{b}\end{bmatrix}=\left(\begin{bmatrix}\mathbf{I} & \mathbf{X}_d\\ \mathbf{0} & \mathbf{X}\end{bmatrix}'\begin{bmatrix}\mathbf{I} & \mathbf{X}_d\\ \mathbf{0} & \mathbf{X}\end{bmatrix}\right)^{-1}\begin{bmatrix}\mathbf{I} & \mathbf{X}_d\\ \mathbf{0} & \mathbf{X}\end{bmatrix}'\begin{bmatrix}\mathbf{Y}_d\\ \mathbf{Y}\end{bmatrix}$$

$$\begin{bmatrix}\mathbf{d}\\ \mathbf{b}\end{bmatrix}=\left(\begin{bmatrix}\mathbf{I} & \mathbf{X}_d\\ \mathbf{X}_d' & \mathbf{X}'\mathbf{X}+\mathbf{X}_d'\mathbf{X}_d\end{bmatrix}\right)^{-1}\begin{bmatrix}\mathbf{Y}_d\\ \mathbf{X}'\mathbf{Y}+\mathbf{X}_d'\mathbf{Y}_d\end{bmatrix}$$

From the property of inverses of partitioned matrices in Frees (2004:420) or Timm (2002:46)

$$\left(\begin{bmatrix}\mathbf{I} & \mathbf{X}_d\\ \mathbf{X}_d' & \mathbf{X}'\mathbf{X}+\mathbf{X}_d'\mathbf{X}_d\end{bmatrix}\right)^{-1}=\begin{bmatrix}(\mathbf{I}+\mathbf{X}_d(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_d') & -\mathbf{X}_d(\mathbf{X}'\mathbf{X})^{-1}\\ -(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_d' & (\mathbf{X}'\mathbf{X})^{-1}\end{bmatrix} \tag{A.5}$$

$$\begin{bmatrix}\mathbf{d}\\ \mathbf{b}\end{bmatrix}=\begin{bmatrix}(\mathbf{I}+\mathbf{X}_d(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_d') & -\mathbf{X}_d(\mathbf{X}'\mathbf{X})^{-1}\\ -(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_d' & (\mathbf{X}'\mathbf{X})^{-1}\end{bmatrix}\begin{bmatrix}\mathbf{Y}_d\\ \mathbf{X}'\mathbf{Y}+\mathbf{X}_d'\mathbf{Y}_d\end{bmatrix}$$

$$\begin{bmatrix}\mathbf{d}\\ \mathbf{b}\end{bmatrix}=\begin{bmatrix}\mathbf{Y}_d-\mathbf{X}_d(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}\\ (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}\end{bmatrix}$$

QED.

[2] Zaman et al. (2001) is a good example for the use and explanation of these advanced techniques.

Proof of Proposition 2.2:

Let R^2 be the coefficient of determination in the regression with DV and R_0^2 be the one without suspicious observations in data, which can be formulated as follows:

$$R^2 = 1 - \frac{\begin{bmatrix}\mathbf{e}' & \mathbf{e}_d'\end{bmatrix}\begin{bmatrix}\mathbf{e}\\ \mathbf{e}_d\end{bmatrix}}{\sum_{i=1}^{n}(y_i-\bar{y})^2} = 1 - \frac{\mathbf{e}'\mathbf{e}}{\sum_{i=1}^{n}(y_i-\bar{y})^2}$$

$$R_0^2 = 1 - \frac{\mathbf{e}'\mathbf{e}}{\sum_{i=1}^{m}(y_i-\bar{y}_0)^2}$$

where $\bar{y} = (\sum_{i=1}^{n} y_i / n)$ and $\bar{y}_0 = (\sum_{i=1}^{m} y_i / m)$;

The relation between $\bar{y}$ and $\bar{y}_0$ is $\bar{y} = \bar{y}_0 + (\frac{m-n}{n}\bar{y}_0 + \sum_{i=m+1}^{n} y_i) = \bar{y}_0 + \eta$

Then $\sum_{i=1}^{n}(y_i-\bar{y})^2 = \sum_{i=1}^{m}\left((y_i-\bar{y}_0)-\eta\right)^2 + \sum_{i=m+1}^{n}(y_i-\bar{y})^2$

$$= \sum_{i=1}^{m}(y_i-\bar{y}_0)^2 + 2\eta\sum_{i=1}^{m}(y_i-\bar{y}_0) + m\eta^2 + \sum_{i=m+1}^{n}(y_i-\bar{y})^2$$

$$= \sum_{i=1}^{m}(y_i-\bar{y}_0)^2 + m\eta^2 + \sum_{i=m+1}^{n}(y_i-\bar{y})^2 > \sum_{i=1}^{m}(y_i-\bar{y}_0)^2$$

$$\sum_{i=1}^{n}(y_i-\bar{y})^2 > \sum_{i=1}^{m}(y_i-\bar{y}_0)^2$$

This implies that the denominator of the negative term in R^2 is larger than the one in R_0^2. Hence $R^2 > R_0^2$. QED.

Proof of Proposition 3.3:

For observation *j* the *t*-statistic of DV can be represented as follows:

$$t_j = \frac{d_j}{\hat{\sigma}\cdot\sqrt{(\overline{\mathbf{X}}'\overline{\mathbf{X}})^{-1}_{jj}}} = \frac{RSR}{\sqrt{(\overline{\mathbf{X}}'\overline{\mathbf{X}})^{-1}_{jj}}}$$

It should be noted that the *t*-statistic contains a robust term (*RSR*) and a non-robust term, namely, $(\overline{\mathbf{X}}'\overline{\mathbf{X}})^{-1}_{jj}$. This can be shown as follows:

From equation (A.5) terms for the dummies can be calculated with the suitable element in the diagonal of $\mathbf{(I + X}_d\mathbf{(X'X)^{-1}X}_d'\mathbf{)}$. Let $\mathbf{(X'X)^{-1}}$ be:

$$\mathbf{(X'X)^{-1}} = \begin{bmatrix} c_{11} & \cdots & c_{1(v+1)} \\ \vdots & \ddots & \vdots \\ c_{(v+1)1} & \cdots & c_{(v+1)(v+1)} \end{bmatrix}$$

then the diagonal of $(\overline{\mathbf{X}}'\overline{\mathbf{X}})^{-1}_{jj}$ becomes:

$$(\bar{\mathbf{X}}'\bar{\mathbf{X}})^{-1}_{jj} = (\mathbf{I} + \mathbf{X}_d(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'_d)_{ii} = 1 + \sum_{g=0}^{v} \bar{x}_{ig} \sum_{h=0}^{v} c_{(h+1)(g+1)} \bar{x}_{ih}$$

where $i = (j–m)$, $\bar{x}_{ab} = x_{(a+m)b}$, $a = 1...(n–m)$, $b = 1...v$, $\bar{x}_{a0} = 1$ if the regression with a constant is considered and $\bar{x}_{a0} = 0$ if the regression without a constant is considered. From this equation even one outlier in x direction, i.e. a large $\bar{x}_{ab}$ in one observation can make $(\bar{\mathbf{X}}'\bar{\mathbf{X}})^{-1}_{jj} >> 1$, which is possible in situations where $RSR_j > 2.5$ and $t_j < 2$. QED

Proof of Proposition 3.4:

Assume that the same notation is used as the one in Proposition 2.2. Let R^2 be the coefficient of determination in the regression with DVs and R_0^2 be the one without the suspicious observations in the data, then *F*-incremental can be expressed as:

$$F = \frac{(R^2 - R_0^2)/(n-m)}{(1-R^2)/(m-v-1)} =$$

$$F = \frac{(m-v-1)}{(n-m)} \frac{(1 - \frac{\begin{bmatrix}\mathbf{e}' & \mathbf{e}'_d\end{bmatrix}\begin{bmatrix}\mathbf{e} \\ \mathbf{e}_d\end{bmatrix}}{\sum_{i=1}^{n}(y_i - \bar{y})^2}) - (1 - \frac{\mathbf{e}'\mathbf{e}}{\sum_{i=1}^{m}(y_i - \bar{y}_0)^2})}{(1 - (1 - \frac{\begin{bmatrix}\mathbf{e}' & \mathbf{e}'_d\end{bmatrix}\begin{bmatrix}\mathbf{e} \\ \mathbf{e}_d\end{bmatrix}}{\sum_{i=1}^{n}(y_i - \bar{y})^2}))}$$

$$F = \frac{(m-v-1)}{(n-m)} \left[\frac{(\frac{\mathbf{e}'\mathbf{e}}{\sum_{i=1}^{m}(y_i - \bar{y}_0)^2} - \frac{\mathbf{e}'\mathbf{e}}{\sum_{i=1}^{n}(y_i - \bar{y})^2})}{\frac{\mathbf{e}'\mathbf{e}}{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \right] = \frac{(m-v-1)}{(n-m)} \left[\frac{(\sum_{i=1}^{n}(y_i - \bar{y})^2)}{\sum_{i=1}^{m}(y_i - \bar{y}_0)^2} - 1 \right]$$

It should be noted that this *F*-value does not depend on independent variable values. If the $(n–m)$ suspicious observations have their y_i average values around $\bar{y}_0$, then *F* will be around 0 (insignificant) without depending on independent variable values. QED

REFERENCES

Billingsley, P., D.A. Belsley, E. Kuh and R.E. Welsch (1980). *Regression diagnostics: identifying influential data and sources of collinearity*. Wiley Series in Probability and Mathematical Statistics. New York-Chichester-Brisbane: Wiley. MR0576408

Cook, R.D. (1985). Detection of influential observation in linear regression. *Technometrics*, 19, 15–18. MR0436478

Frees, E.W. (2004). *Longitudinal and Panel Data: Analysis and Applications in the Social Sciences*. Cambridge: Cambridge University Press. MR2087947

Greene, W. H. (2002). *Econometric Analysis.* Fifth edition, NJ: Prentice Hall.

Rousseeuw, P.J. and A.M. Leroy (1987). *Robust regression and outlier detection*. New York: John Wiley & Sons, Inc. MR0914792

Rousseeuw, P.J. and S. Van Aelst (1999). Positive-breakdown robust methods in computer vision. *Computing Science and Statistics*, 31, 451-460.

Studenmund, A.H. (2002). *Using Econometrics: A Practical Guide*. London: Addison Wesley.

Timm, N.H. (2002). *Applied Multivariate Analysis*. New York: Springer-Verlag. MR1908225

Zaman, A., P.J. Rousseeuw and M. Orhan (2001). Econometric applications of high-breakdown robust regression techniques. *Economics Letters*, 71 (1), 1-8.