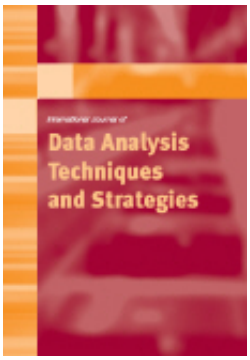


International Journal of Data Analysis Techniques and Strategies

ISSN online: 1755-8069 - ISSN print: 1755-8050
https://www.inderscience.com/ijdats

Abnormal accrual estimation: an automation data analysis technique

Francesca Rossignoli, Nicola Tommasi

DOI: 10.1504/IJDATS.2025.10069523

Article History:

Received:	01 July 2024
Last revised:	20 September 2024
Accepted:	23 September 2024
Published online:	21 February 2025

Abnormal accrual estimation: an automation data analysis technique

Francesca Rossignoli*

Department of Management,
University of Verona, Italy,
Via Cantarane 24, 37129, Verona, Italy
Email: francesca.rossignoli@univr.it
*Corresponding author

Nicola Tommasi

Interdepartmental Centre of Economic Documentation (C.I.D.E.),
University of Verona, Italy,
Via Cantarane 24, 37129, Verona, Italy
Email: nicola.tommasi@univr.it

Abstract: Accounting studies rely on predictive analytics to estimate abnormal accruals as indicators of managerial opportunism. Abnormal accruals are estimated by running predictive models and manually imposing a combination of conditions to select the control sample. This process is executed using loops where the estimation is repeated over the control observations meeting the combined conditions. The recursive estimation generates several inefficiencies. We provide a technique to estimate abnormal measures by automatising: i) the estimation of the predictive model; and ii) the selection of the control sample according to multiple procedures. The command offers a unique information set about the estimation results and process. We illustrate the use of `abnormalest` through empirical applications. We compare the accuracy of predictions under different approaches and models. The command `abnormalest` allows to overcome the inefficiencies, provides a unique set of information about the estimation, and is extendible to every social science.

Keywords: abnormal estimation; abnormal accrual; earnings management; prediction model; financial accounting.

Reference to this paper should be made as follows: Rossignoli, F. and Tommasi, N. (2025) 'Abnormal accrual estimation: an automation data analysis technique', *Int. J. Data Analysis Techniques and Strategies*, Vol. 17, No. 5, pp.1–18.

Biographical notes: Francesca Rossignoli, PhD, is an Associate Professor of Accounting at the Department of Management, University of Verona (Italy). Her research interests include financial accounting, corporate governance, and family business. Her papers are published in journals like the *Journal of Management and Governance, Accounting, Auditing and Accountability Journal, Meditari Accountancy Research, Journal of Applied Accounting Research*, and *Journal of*

International Accounting Research. She is a Member of the European Accounting Association and the Italian accounting associations AIDEA, SIDREA, and SISR.

Nicola Tommasi is a Data Technician at the Interdepartmental Centre of Economic Documentation (C.I.D.E.), University of Verona, Italy. His main research interests are applied econometrics and statistical programming.

1 Introduction

Predicting business performance from corporate information is a major stream of data analytics application for managers and practitioners (Corlu et al., 2021). Data analytics are used to identify the relevance of specific organisational factors that could impact corporate performance (Jr. and Stanley, 2012) and to predict future performance (Bouasabah, 2024). Predictive analytics includes a variety of analytical techniques to make predictions by unveiling patterns found in historical and transactional data. The predictive techniques are used to forecast the likelihood of enterprise survival (Sujatha et al., 2022) and of financial statement fraud (Ravisankar et al., 2011). Financial accounting studies rely on predictive analytics to detect managerial opportunism (Dechow et al., 1995). Specifically, accounting studies use abnormal accrual to estimate the likelihood of managers' opportunistic behaviour of extracting private benefits. Multiple estimations are used to predict the normal accruals. Empirical evidence shows that alternative measures and regression procedures are important in evaluating the performance of specific abnormal accruals models (Agnes Cheng et al., 2012). The procedure used to estimate abnormal accruals has drawn the attention of recent studies (McMullin and Schonberger, 2020); (McMullin and Schonberger, 2022) that propose innovative preprocessing procedures to select the control sample.

This paper's primary contribution is to provide a technique to estimate abnormal measures by applying predictive analytics. We propose a data analysis technique to automatise the estimation of abnormal accruals. The technique's flexibility allows the researchers to easily employ different prediction models and multiple procedures to identify the control sample. Furthermore, the technique provides a set of information of essence for researchers.

1.1 Research gap and motivation

In the data analysis practice, abnormal cases are detected as outliers (Karthikeyan and Balasubramanie, 2020) using several techniques (Seelammal and Devi, 2018). However, these techniques do not provide metrics to measure the magnitude of abnormality. Financial accounting studies use abnormal measures of accrual. The abnormal measures suggest as to whether unusual accruals arise from the predicted accruals. The abnormal accruals are the differences between reported and normal accruals.

The normal accruals are predicted using several estimating approaches relying on the observable accounting numbers. The accrual is to be qualified as "normal" to the extent to which it is estimated based on the predictive characteristics observable on the control observations that are identified as a reference for regularity. To the extent that observed accruals deviate from their normal amount, abnormal accrual occurs. The procedure to estimate the abnormal accrual implies two methodological steps:

i the estimation of the normal accrual using a predictive model;

ii the control sample selection.

In the first step, the normal accrual is estimated at the firm level in the control sample by running a regression model using the observable data drawn from the companies' annual reports. Several predictive approaches are used in the practice (see Section 2.2), each including various explanatory variables. Time series data are essential in this step, and missing even one data in the explanatory variables makes that observation unusable for the estimation. Although the software automatically removes the observations showing missing data, a problem referring to the degrees of freedom emerges as the second step conditions the first step. In the second step, the researcher imposes the condition of selecting the control sample for each "treated" observation where the normal measure is to be estimated. In addition, multiple procedures are applied to select the control sample (see Section 2.3). Resulting of this step, the number of observations where the estimation is run differs and may lead to insufficient observations needed to meet the degrees of freedom. As a result, the number of observations needed to run the estimation meeting the degree of freedom differs depending on the combination of the choices made in the two steps.

Existing research tackles the two steps separately by running recursive commands as loops. These procedures are long, time-consuming, and inefficient. For example, insufficient observations prevent the estimation, and researchers need to refine the control sample selection. Still, the manual procedure does not provide the information to identify where the estimation fails and, consequently, which additional aggregations to impose to run the estimation. An integrated approach with extreme flexibility in the two steps is needed to optimise the estimation and the control sample selection, to prevent inefficiencies, and to provide a set of information that helps the researcher to optimise the estimation. Additionally, this set of information can provide details about the estimation process.

Furthermore, data analysis literature offers several preprocessing techniques (Cagnoni et al., 2021; Harshe and Kulkarni, 2018) as a data reduction. The proposed data reduction techniques aim to select a reduced number of subset features with high predictive information and remove irrelevant features with minimal predictive information (Sarkar et al., 2016). However, in the abnormal accrual estimation, there is the need for a preprocessing technique to identify a control sample that is nearly identical to the treated sample with respect to observable covariates. The technique we propose integrates this preprocessing step in the estimation.

1.2 Main contributions

The paper contribution to the social sciences is threefold. Firstly, the paper offers an overview of the options every researcher needs to deal with when approaching a procedure of estimation measures departing from the "expected ones". Multiple procedures have been applied, combining alternative conditions to select the control observations to predict the normal measure. Recently, preprocessing methods such as entropy balance have become popular to adjust the covariate distribution of the control group data by reweighting or discarding of units such that it becomes more similar to the covariate distribution in the treatment group (Jann, 2021).

Secondly, the data analysis technique allows the researchers to employ different prediction models and multiple procedures to select the control sample, thus automatising and optimising the procedures suggested in the literature. The automatisation reduces inefficiencies generated by running recursive commands as loops. Furthermore, the technique provides a set of information of essence for researchers to further investigate the estimations during the process. In particular, the command provides these additional metrics: the abnormal measure, the estimated normal measure, the degrees of freedom theoretically computed, the number of the control observations iteratively identified, and the reasons for the failure of the estimation. This set of information is not obtainable when running recursive estimations.

Lastly, the data analysis technique presented here is suitable to be extended to the estimation of any measures, thus providing researchers with a useful tool for detecting unusual measures in any field of social sciences.

We illustrate the `abnormalest` technique using financial statements data retrieved from Compustat. We compare the accuracy of predictions under different approaches commonly used in the literature (Jones, Jones modified and Kothari), as well as under different estimation models, including OLS regression, the quantile regression model, and the entropy balance preprocessing method.

1.3 Paper organisation

The rest of the paper is structured as follows. Section 2 provides a brief illustration of abnormal accrual measures in financial accounting. Section 2.2 summarises the main predictive approaches used in financial accounting studies to predict normal accrual while Section 2.3 illustrates the procedures used to select the control sample. Section 3 illustrates the technique that integrates the optimisation of the two steps in a Stata command that we programmed. Section 4 provides an empirical application of the data technique, comparing the accuracy of predictions under different approaches commonly used in the literature and under different estimation models. Section 5 concludes by illustrating the contributions.

2 Abnormal accrual measures

2.1 Abnormal accrual in financial accounting studies

In financial accounting studies, a key concern pertaining agency conflicts consists in detecting earnings management practices in reporting financial performance wherein managers detain information advantages with respect to the outsider stakeholders. Earnings management practices are usually aimed at increasing reported earnings, with the goal to ultimately raise managers' compensation through higher bonuses. Schipper (1989) defines earnings management as "a purposeful intervention on the external financial reporting process, with the content of obtaining some private gain" (Schipper, 1989, p.92). Under this perspective, opportunistic earnings management negatively impacts the quality of earnings.

In the seminal paper that paved the way for the abnormal accrual measures, Jones (1991) claims that earnings management can be achieved using accruals. Accruals are earnings

components subjected to estimation which makes them inherently uncertain (Guay et al., 1996; Dechow and Dichev, 2002). Such discretion in the estimation offers management the option to opportunistically manipulate reported earnings (Watts and Zimmerman, 1986; Healy and Wahlen, 1999). Misestimation of accruals by management, regardless the intentionality, generates noise in reported earnings (Healy, 1996). Therefore, the accrual component of earnings is a challenge to financial reporting users.

Jones (1991) defines abnormal accrual as "an estimate of the discretionary component of total accruals" (Jones, 1991, p.194). While the discretionary portion of a single accrual account has been used in the past (McNichols and Wilson, 1988), the discretionary portion of total accruals is preferable since it is expected to capture a larger portion of managers' manipulations (Jones, 1991, p.206).

According to the agency theory (Michael and William, 1976), detecting the likelihood of abnormal behaviour is essential for the uninformed party to monitor risks subsiding contracts between parties where there is information asymmetry and potential conflict of interests. Theoretically, accruals that diverge from their "normal" amount may signal lower-quality earnings and mislead financial statement users (Dechow et al., 1996). Empirically, this implies that the abnormal accrual estimation moves from the normal accrual prediction. The accrual model that best captures abnormal accruals is a function of not only the variables used to predict normal accruals but also the estimation procedure employed (Agnes Cheng et al., 2012). This is extendible to every attempt in social science to detect systematic anomalies with respect to the expected ones.

The normal accruals are predicted using several estimating approaches relying on the observable accounting numbers. An accrual measure is to be qualified as "normal" to the extent to which it is estimated based on the predictive characteristics observable on the control observations that are selected as reference for regularity. Therefore, the procedure to estimate the abnormal accrual implies two methodological steps:

i the estimation of the normal measure using a predictive model;

ii the control sample selection.

The observed accruals deviating from their normal amount indicate abnormal accrual. The abnormal accruals are the excess amounts of accruals reported by a company with respect to the normal accrual estimated in the control sample.

2.2 The predictive approaches for estimating normal accrual

In the Jones approach (Jones, 1991), the total accrual (TA) is calculated as the change in non-cash current assets minus the change in current liabilities, excluding the current portion of long-term debt minus depreciation and amortisation. Jones (1991) suggests that total accrual is expected to vary with the change in revenues and gross property, plant, and equipment in the current period. Therefore, the total accrual is predicted using the equation (1).

$$\frac{TA_{i,t}}{Asset_{i,t-1}} = \beta_1 \frac{1}{Asset_{i,t-1}} + \beta_2 \frac{\Delta Rev}{Asset_{i,t-1}} + \beta_3 \frac{PPE_{i,t}}{Asset_{i,t-1}} + \varepsilon_{i,t} \tag{1}$$

where

$TA_{i,t}$ is total accruals in year t for firm i

$Asset_{i,t-1}$ is total assets in year $t-1$ for firm i

ΔRev is revenues in year t less revenues in year $t-1$ for firm i

$PPE_{i,t}$ is gross property, plant and equipment in year t for firm i

$\varepsilon_{i,t}$ is error term in year t for firm i.

The parameter estimates obtained from the above equation are used to estimate firm-specific normal accruals (NA) as shown in Equation (2):

$$NA_{i,t} = \hat{\beta}_1 \frac{1}{Asset_{i,t-1}} + \hat{\beta}_2 \frac{\Delta Rev}{Asset_{i,t-1}} + \hat{\beta}_3 \frac{PPE_{i,t}}{Asset_{i,t-1}} + \varepsilon_{i,t} \tag{2}$$

The Jones model relies on the assumption that revenues are non-discretionary. However, earnings are manageable through discretionary revenues. For example, managers might use their discretion to accrue revenues at year-end when the cash has not yet been received and it is highly questionable whether the revenues have been earned. This managerial discretion will result in an increase in revenues and total accruals (through an increase in receivables). This approach causes the estimate of earnings management to be biased toward zero (Dechow et al., 1995). To eliminate such bias, Dechow et al. (1995) propose a modification of the Jones model to measure discretionary accruals with error when discretion is exercised over revenues. In the modified Jones model the change in revenues is adjusted for the change in receivables. Equation 3 shows the modified Jones model.

$$TA_{i,t} = \beta_1 \frac{1}{Asset_{i,t-1}} + \beta_2 \frac{(\Delta Rev_{i,t} - \Delta Rec_{i,t})}{Asset_{i,t-1}} + \beta_3 \frac{PPE_{i,t}}{Asset_{i,t-1}} + \varepsilon_{i,t} \tag{3}$$

where $\Delta Rec_{i,t}$ is the firm's i change in account receivables between year $t-1$ and t.

The parameter estimates obtained from the above equation are used to estimate firm-specific normal accruals (NA) as shown in Equation (4):

$$NA_{i,t} = \hat{\beta}_1 \frac{1}{Asset_{i,t-1}} + \hat{\beta}_2 \frac{(\Delta Rev_{i,t} - \Delta Rec_{i,t})}{Asset_{i,t-1}} + \hat{\beta}_3 \frac{PPE_{i,t}}{Asset_{i,t-1}} + \varepsilon_{i,t} \tag{4}$$

The modified Jones model assumes that all changes in credit sales in the event period result from earnings management. This is based on the reasoning that it is easier to manage earnings by exercising discretion over the recognition of revenue on credit sales than it is to manage earnings by exercising discretion over the recognition of revenue on cash sales.

The Kothari approach (Kothari et al., 2005) controls for earnings performance when estimating abnormal accruals including in the Jones model the previous year's return on asset (ROA).

$$\begin{aligned} \frac{TA_{i,t}}{Asset_{i,t-1}} = {} & \beta_1 \frac{1}{Asset_{i,t-1}} + \beta_2 \frac{\Delta Rev}{Asset_{i,t-1}} + \beta_3 \frac{PPE_{i,t}}{Asset_{i,t-1}} \\ & + \beta_4 ROA_{i,t-1} + \varepsilon_{i,t} \end{aligned} \tag{5}$$

where $ROA_{i,t-1}$ is the return on assets for firm i in period $t-1$.

Whatever the predictive approach used, the abnormal accrual (AA) is calculated as the difference between the firm-specific observed total accrual and the estimated normal accrual, as shown in equation (6):

$$AA_{i,t} = x_1 \frac{TA}{Asset_{i,t-1}} - NA_{i,t} \tag{6}$$

2.3 The procedures to select the control sample

Multiple procedures have been applied referring to alternative conditions imposed to select the control observations whereon to predict the normal measure.

The approaches move from the assumption that even without intentional earnings management, accrual quality will be systematically related to characteristics that are observable at a certain level of aggregation. Such characteristics are likely to be both observable and recurring compared to the determinants of managerial opportunism, which are often unobservable. Therefore, the estimation of the normal measures is predicted by relying on these observable characteristics.

The most used procedures estimate the normal accrual aggregating the control observations at the firm level and at the industry-level. The firm-specific estimation implies that the observable characteristics predicting normal accrual are constant along all years for a given firm. The industry-specific estimation imposes the constraint that the observable characteristics predicting normal accrual are constant across all firms within the same industry.

The superiority of one procedure over the others nourished a wide debate among accounting scholars.

Some suggest the firm-specific method is more appropriate (e.g., Dechow and Dichev, 2002; DeFond and Park, 2001), while others suggest the industry-specific method is preferable (e.g., Bartov et al., 2000; Kothari et al., 2005).

Dechow and Dichev (2002) report that the industry-specific procedure yields weaker results. They state: "we believe that a firm level specification is superior. However, we also present industry-specific and pooled results because our firm-specific time-series is short, and we are concerned about noisy estimation at the firm level" (Dechow and Dichev, 2002, p.44).

DeFond and Park (2001, p.399) mention: "firm-specific measures are likely to be superior to industry-wide estimates. For example, within a given industry, a firm's size, age, and accounting choices may affect the normal level of working capital required to sustain current sales levels, but Jones model measures of normal accruals only reflect the average effects of these factors".

Bartov et al. (2000) show that for firms with qualified auditor reports the industry-specific estimations perform better than the firm-specific model in detecting earnings management.

Kothari et al. (2005) point out that the industry-specific procedure has the advantage of avoiding small sample and survival bias stemming from requiring a long time series of data for each firm that occurs when the firm-specific procedure is used.

Further aggregations are commonly used according to emerging evidence demonstrating cross-sectional variations in the determinants of financial accounting choices. For example, the country-level aggregation is widely used, relying on the evidence that institutional characteristics qualifying the investor protection are observable at country-level (Leuz et al., 2003).

Recent studies (McMullin and Schonberger, 2020) claim that the distributions of the accrual determinants depart from linearity, thus questioning the adequacy of the accrual-generating processes employing linear models.

To address this issue, these studies suggest employing the entropy balancing technique (Jann, 2021; Hainmueller, 2012; Hainmueller and Xu, 2013) in estimating the abnormal accrual. Entropy balancing is a statistical method for identifying a control sample that is nearly identical to the treated sample with respect to observable covariates.

The control sample selection using the entropy balance allows to incorporate non-linear relations, thus addressing covariate differences in one or more distributional moments. This procedure identifies continuous weights for all the control observations to equalise the distribution for the accrual determinants across the sample and control observations, thus avoiding assuming linearity for the full set of the accrual predictors. The accrual is estimated on the control sample balancing the covariates of the predictors.

The entropy balance approach (McMullin and Schonberger, 2020) relies on the disentanglement between the treatment sample and control sample. The combination identifies the treatment sample of the multiple conditions imposed in the estimation procedure, while the control sample include all the rest of the observations that are weighted equalising the distribution for the accrual determinants across the sample and control observations.

3 The data analysis technique: abnormalest

The `abnormalest` command is a Stata command that automates the estimation of the abnormal measure as the difference between the reported and predicted measures. The technique includes the automation of:

i the sample selection:

ii the normal accrual estimation

iii the calculation of the difference between the reported and predicted measures.

Multiple estimation models are available to predict the measure: OLS, quantile regression, and regression preprocessed by entropy balance. The command is intended to detect unusual measures with respect to the predicted one. In the accounting field, all the estimation models are employable to implement the alternative approaches proposed by the literature and the mentioned (Jones, Jones modified and Khotari), with the only exception of quantile regression that does not apply to the Jones modified approach because it does not allow the omit the constant term.

3.1 Installation

The package can be installed via GitHub. Copy in Stata command bar the following command:

net install abnormalest, from("https://raw.githubusercontent.com/NicolaTommasi8/abnormalest/master/https://raw.githubusercontent.com/NicolaTommasi8/abnormalest/master/") replace

The package is also published on CodeOcean (https://codeocean.com/capsule/6827601/tree/v1https://codeocean.com/capsule/6827601/tree/v1) (Rossignoli and Tommasi, 2024)

3.2 Syntax

The syntax of `abnormalest` is as follows:

```
abnormalest depvar indepvars [if] [in][weight], condvar(varlist)
  abnvar(varname) [ estvar(varname) model(string) minobs(#)
  noconstant quantile(#) targets(options) iterate(integer)
  btolerance(#) ptolerance(#) difficult fix(varlist2) ]
```

where `depvar` is the depended variable and `indepvars` the list of independent variables.

3.3 Options

`condvar(varlist)` variables identifying the control sample observations whereon to predict the normal measure

`abnvar(varname)` variable generated as the difference between the reported measure in the treated sample `depvar` and the predicted measure in the control sample `estvar`

`estvar(varname)` estimated variable for `depvar` obtained from the estimation model used

`model(ststring)` estimation model. Available alternatives: `model(ols|qreg|ebal)`. Default is `model(ols)`

`minobs(#)` imposes the minimum number of observations to execute the conditional estimation. Default minimum is equal to the estimation model degrees + 1 e(df_m)+1 in the `models(ols)` or `model(qreg)`, and to the minimum number of observations in the treated sample for the `model(ebal)`

`noconstant` omits constant term in the conditional estimation model

`quantile(#)` estimate # quantile; default is `quantile(.5)`. `model(qreg)` is required

`targets(options)` specify types of moments to be balanced; default is `targets(mean)`. Possible options are: mean (the default), variance (implies mean), skeweness (implies mean and variance) and covariance (implies mean). `model(ebal)` is required

iterate(integer) specifies the maximum number of iterations; default is `iterate(300)`. `model(ebal)` is required

btolerance(#) sets the balancing tolerance. Balance is achieved if the balancing loss is smaller than the balancing tolerance. The default is `btolerance(1e-6)`. `model(ebal)` is required

ptolerance(#) specifies the convergence tolerance for the coefficient vector. The default is `ptolerance(1e-6)`. `model(ebal)` is required

difficult use a different stepping algorithm in nonconcave regions. `model(ebal)` is required

fix(varlist2) allows selecting the control units fixing one or more conditions to be met. Variables in `varlist2` must be a subgroup of `condvar(varlist)` variables. `model(ebal)` is required

Note on `fix()` option. By default, the treated observations are those meeting each of the combinations of the imposed conditions. The control units are selected as the residual observations discarding those meeting each of the combinations of the imposed conditions. Given a treated unit meeting a combination of the imposed conditions, the control units are the residual observations. Using `fix()` option, the control units are the non-treated observations showing in correspondence to the fixed conditions the same values as the treated observations. For example, being the treated observations those located in Thaiwan, operating in industry 7 in year 2019 by option `condvar(country year industry1d)`, fixing the condition `fix(country)` means selecting all the control units located in Thaiwan discarding the treated units identified by the combination of the imposed conditions.

4 Empirical application

We extract from Compustat all financial data required to calculate the total accrual and to estimate the abnormal accruals for the three named approaches (Jones, Jones modified and Khotari). The full sample obtained contains 63,279 observations. The full sample is available along with the command package. For the sake of conciseness we employed a reduced sample to perform the examples illustrated in this paper. The reduced sample includes observations across 4 countries, 4 industries and 3 years for a total of 4,301 observations. The variables used to identify the control sample whereon to predict the normal measure of accrual are: `country`, `year`, and `industry1d`. The variable `country` is the codification of the Current ISO Country Code indicating the country where each company is located. It is equal to CYM if the company is located in Cayman, KOR if the company is located in Korea, TWN if the company is located in Taiwan and CHN if the company is located in China. The variable `industry1d` is the first digit of the Standard Industry Classification Code identifying the industry wherein the company is operating.

It is equal to 5 for the industry "Wholesale trade", 7 for "Commercial services", 8 for "Social and health services", and 9 for the industry "Public administration". The variable `year` indicates the year of reporting, respectively being 2019, 2020 and 2021. The variable `total_accrual_scaled` is the total accrual reported in year t for firm i scaled by total assets in year $t-1$ for firm i. The composition of total accruals is as follows:

$$\begin{aligned}\text{Total accruals} \;=\; & \Delta\text{Current Assets Cash} - \Delta\text{Current Liabilities} \\ & -\Delta\text{Current Maturities of Long-Term Debt} \\ & -\Delta\text{Income Taxes Payable} \\ & -\text{Depreciation and Amortisation Expense}\end{aligned}$$

where the change Δ is computed between time t and time $t-1$.

The variable Intercept scaled (`intercept_scaled`) for firm i is calculated as

$$\text{Intercept scaled} = \frac{1}{\text{total asset}_{t-1}}$$

The variable Delta revenues scaled (`delta_REV_scaled`) for firm i is calculated as

$$\text{Delta revenues scaled} = \frac{\text{revenues}_t - \text{revenues}_{t-1}}{\text{total asset}_{t-1}}$$

The variable PPE scaled (`PPE_scaled`) for firm i is calculated as

$$\text{PPE scaled} = \frac{\text{value of property, plant and equipment}_t}{\text{total asset}_{t-1}}$$

The variable Net receivables (`REV_REC`) for firm i is calculated as

$$\text{Net receivables} = \frac{\Delta\text{revenues}_t - \Delta\text{receivables}_t}{\text{total asset}_{t-1}}$$

The variable Return on asset (`ROA`) for firm i is calculated as

$$\text{Return on asset} = \frac{\text{net income}_t}{\text{total asset}_{t-1}}$$

To ensure sufficient degrees of freedom and enhance the quality of these measures, we limit our sample to companies in those industry-country-year groups that had at least 7 or more observations.

In the first example, we apply the command to Jones' approach (see equation (1)) using OLS as the estimation model:

```
. abnormalest total_accruals_scaled intercept_scaled delta_REV_scaled PPE_scaled, ///
>   condvars(industry1d year country) estvar(est1a) abnvar(abn1a) minobs(7)


Conditional var #1: industry1d
Levels of industry1d:
5 7 8 9


Conditional var #2: year
Levels of year:
2019 2020 2021


Conditional var #3: country
Levels of country:
CHN CYM KOR TWN


I'm performing regressions... please wait!
Looping until 45 regressions
----+--- 1 ---+--- 2 ---+--- 3 ---+--- 4 ---+--- 5
.....................................x..x.xxx
x means Insuf. obs, n fail in regression

Minimum number of obs: 7
Theoric minimum number of obs: 4
Conditional vars: industry1d year country


OK regressions: 40
No reg. by insuf obs.: 5
Failed regressions: 0
```

Conditional vars	Valid obs.	Reg. outcome	Adj Rsq	Prob > F
5 2019 CHN	157	OK	0.0199	0.1087
5 2019 CYM	105	OK	0.2497	0.0000
5 2019 KOR	57	OK	0.1363	0.0129
5 2019 TWN	107	OK	0.4166	0.0000
5 2020 CHN	154	OK	0.3507	0.0000
5 2020 CYM	100	OK	0.1137	0.0022
5 2020 KOR	60	OK	0.0186	0.2604
5 2020 TWN	106	OK	0.3754	0.0000
5 2021 CHN	104	OK	0.2109	0.0000
5 2021 CYM	65	OK	0.1405	0.0065
5 2021 KOR	54	OK	-0.0058	0.4489
5 2021 TWN	101	OK	0.0994	0.0043
7 2019 CHN	263	OK	0.0651	0.0001
7 2019 CYM	139	OK	0.0783	0.0029
7 2019 KOR	154	OK	0.0863	0.0009
7 2019 TWN	103	OK	0.0940	0.0051
7 2020 CHN	251	OK	0.0574	0.0005
7 2020 CYM	136	OK	0.0377	0.0447
7 2020 KOR	159	OK	-0.0087	0.6523
7 2020 TWN	104	OK	-0.0102	0.5836
7 2021 CHN	182	OK	0.0508	0.0064
7 2021 CYM	87	OK	0.1892	0.0001
7 2021 KOR	145	OK	0.0084	0.2432
7 2021 TWN	89	OK	0.0039	0.3482
8 2019 CHN	103	OK	-0.0083	0.5424
8 2019 CYM	47	OK	0.2601	0.0011
8 2019 KOR	27	OK	-0.0345	0.5551
8 2019 TWN	21	OK	0.0945	0.2057
8 2020 CHN	97	OK	-0.0176	0.7199
8 2020 CYM	48	OK	0.0081	0.3482
8 2020 KOR	29	OK	0.6382	0.0000
8 2020 TWN	21	OK	-0.1013	0.7641
8 2021 CHN	70	OK	0.2105	0.0003
8 2021 CYM	26	OK	0.0332	0.3039
8 2021 KOR	28	OK	0.1552	0.0715
8 2021 TWN	20	OK	-0.0157	0.4617
9 2019 CHN	8	OK	0.8584	0.0119
9 2019 CYM	2	Insuff. obs	.	.
9 2019 KOR	7	OK	-0.0950	0.5604
9 2020 CHN	7	OK	-0.4497	0.7763

```
9 2020 CYM         |     2               Insuff. obs                  .          .
9 2020 KOR         |     7               OK                      0.5274     0.1805
9 2021 CHN         |     2               Insuff. obs                  .          .
9 2021 CYM         |     2               Insuff. obs                  .          .
9 2021 KOR         |     4               Insuff. obs                  .          .
-------------------------------------------------------------------------------
```

The output shows the levels of the conditional variables and, subsequently, the looping progress and the estimation results. Alternative results of each regression are: "Ok regressions", indicating the regressions correctly estimated, "No reg. by insuf. obs.", indicating the number of not-executed regressions because of insufficient number of observations and "Failed regressions", indicating the number of regressions executed but failed to produce the estimated parameters. The final table shows for each combination of the conditional variables, the number of observations meeting the specific combination of the conditional variables, the regression outcomes, and, for the OLS models, the adjusted R-square and the Prob. F statistic, while for the qreg model only the Pseudo R-square is available.

In the second example, we apply the command to Jones' approach (see equation (1)) using the quantile regression as the estimation model:

```
. abnormalest total_accruals_scaled intercept_scaled delta_REV_scaled PPE_scaled, ///
>   condvars(industry1d year country) estvar(est2a) abnvar(abn2a) minobs(7) ///
>   model(qreg)

(output omitted)
```

In the third example, we apply the command to Jones' approach (see equation (1)) using the ebalance as estimation model:

```
. abnormalest total_accruals_scaled intercept_scaled delta_REV_scaled PPE_scaled, ///
>   condvars(industry1d year country) estvar(est3a) abnvar(abn3a) minobs(7) ///
>   model(ebal)

(output omitted)
```

In case of ebalance model, the output is similar to the one of the OLS model; additionally the number of observations counted in the control group is shown in brackets. In this example some estimations fail in estimating the regression (Insuff. obs) because of insufficient observations in the control sample. The researcher might be willing to force the estimation aggregating some observations. For example, aggregating by industry and year, relaxing the country condition would allow to pull together the observations in industry and year regardless the country. In this case, the aggregation by industry and year would allow for treated observations in industry 9 and in year 2019 to be matched with a control sample including observations in CYM and KOR. As well as for treated observation in industry 9 and in year 2021 to be matched with a control sample including observations in CHN, CYM and KOR.

Fails in regression might be resolved by increasing maximum number of iterations (option `iterate()`) or increasing tolerance level (`ptolerance()`).

```
. clonevar country_aggr=country

. replace country_aggr="KOR-CYM" if inlist(country,"CYM","KOR") & year==2019 ///
>    & industry1d==9
variable country_aggr was str3 now str7
(10 real changes made)

. replace country_aggr="KOR-CYM-CHN" if inlist(country,"CYM","KOR","CHN") ///
>   & year==2021 & industry1d==9
variable country_aggr was str7 now str11
(13 real changes made)
.
. abnormalest total_accruals_scaled intercept_scaled delta_REV_scaled PPE_scaled, ///
>   condvars(industry1d year country_aggr) estvar(est3a_aggr) abnvar(abn3a_aggr) ///
>   minobs(7) model(ebal)

(output omitted)
```

Table 1 shows the descriptive statistics of the dependent variable `total_accrual_scaled` (total accrual reported) and the estimated variables (normal accrual) according to the different approaches. In particular, `est1a` is the predicted accrual estimated with the OLS model according to the Jones approach, `est1b` according to the modified Jones model and `est1c` according to the Kothari model.

Table 1 Estimated accruals

	N. obs	*Mean*	*Dev. Std.*	*Median*	*Min*	*Max*
Observed accruals	3698	–0.024	0.249	–0.029	–4.415	4.345
Estimated accrual Jones model (est1a)	3548	–0.022	0.093	–0.015	–1.568	0.920
Estimated accrual Modified Jones model (est1b)	3544	–0.027	0.093	–0.017	–1.503	0.856
Estimated accrual Kothari model (est1c)	3103	–0.039	0.098	–0.029	–1.683	1.050
Estimated accrual Jones model quantile regression (est2a)	3527	–0.034	0.069	–0.024	–0.891	1.011
Estimated accrual Kothari model quantile regression (est2c)	3096	–0.038	0.078	–0.029	–1.476	1.039
Estimated accrual Jones model ebalance (est3a)	3548	–0.022	0.075	–0.010	–0.987	0.508
Estimated accrual Modified Jones model ebalance (est3b)	3530	–0.022	0.082	–0.013	–1.031	0.485
Estimated accrual Kothari model ebalance (est3c)	3103	–0.039	0.078	–0.027	–1.501	0.800
Estimated accrual Jones model ebalance (est3abis)	3548	–0.022	0.076	–0.012	–1.039	0.565
Estimated accrual Modified Jones model ebalance (est3bbis)	3544	–0.022	0.083	–0.013	–1.048	0.567
Estimated accrual Kothari model ebalance (est3cbis)	3103	–0.039	0.078	–0.028	–1.588	0.767

Table 2 shows the descriptive statistics of the abnormal accruals estimated executing the command `abnormalest` according to the different approaches and models. In particular, `abn1a` is the abnormal accrual estimated with the OLS model according to the Jones

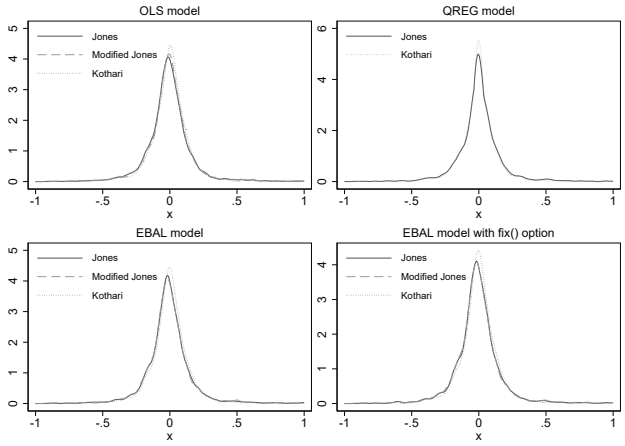
approach, `abn1b` according to the modified Jones model and `abn1c` according to the Kothari model.

Table 2 Abnormal accruals

	N. obs	*Mean*	*Dev. Std.*	*Median*	*Min*	*Max*
Abnormal accrual estimated Jones model (abn1a)	3548	0.000	0.217	–0.008	–1.604	4.198
Abnormal accrual estimated Modified Jones model (abn1b)	3544	0.005	0.217	–0.004	–1.690	4.251
Abnormal accrual estimated Kothari model (abn1c)	3103	0.000	0.165	0.001	–1.365	3.979
Abnormal accrual estimated Jones model quantile regression (abn2a)	3527	0.012	0.223	0.000	–2.117	4.394
Abnormal accrual estimated Kothari model quantile regression (abn2c)	3096	–0.001	0.173	0.000	–2.089	4.273
Abnormal accrual estimated Jones model ebalance (abn3a)	3548	0.000	0.221	–0.012	–2.043	4.277
Abnormal accrual estimated Modified Jones model ebalance (abn3b)	3530	–0.000	0.223	–0.013	–2.130	4.325
Abnormal accrual estimated Kothari model ebalance (abn3c)	3103	–0.000	0.171	–0.001	–1.847	4.198
Abnormal accrual estimated Jones model ebalance (abn3abis)	3548	–0.000	0.221	–0.011	–2.005	4.275
Abnormal accrual estimated Modified Jones model ebalance (abn3bbis)	3544	0.000	0.222	–0.011	–2.110	4.320
Abnormal accrual estimated Kothari model ebalance (abn3cbis)	3103	–0.000	0.173	–0.001	–1.916	4.212

Lastly, Figure 1 shows the kernel density for each approach, divided by the estimation model. For the ebal, the model results are shown for the method with no restrictions and the method with the option `fix()` for the conditional variable country.

Figure 1 The kernel densities across the models



5 Conclusion

The paper moves from an overview about the options every researcher needs to deal with when approaching a procedure to estimate abnormal values departing from the expected ones. In this paper, we describe `abnormalest`: a new Stata command that automatises the estimation of abnormal measures, overcoming the inefficiencies generated by running recursive commands as loops. The command's flexibility allows the researchers to employ different prediction models. Furthermore, the command allows users to impose multiple conditions to select the control observations, including the preprocessing data according to the entropy balancing. The command offers unique output about the estimation results and process that are not obtainable when running recursive estimations.

We illustrate the use of `abnormalest` through several empirical applications in financial accounting. We compare the accuracy of predictions under different approaches commonly used in the literature, as well as under different estimation models.

The data analysis technique we propose opens the way for further research. For the sake of accounting studies, future development of the data analysis technique shall incorporate as an option to the command the formula to compute the most common measures of abnormal accruals: Jones (1991), Dechow et al. (1995), and Kothari et al. (2005). They are acknowledged as milestones and referred to in every study employing abnormal accrual as a performance measure to detect managerial opportunism.

Beyond the application in accounting studies, the command can be extended to a wide plethora of measures in social sciences to detect observable behaviours departing from the expected normal, indicating systematic anomalies. While in accounting studies, the estimation of the predicted measures relies on widely acknowledged models (Jones, 1991; Dechow et al., 1995; Kothari et al., 2005), extending the approach to non-accounting studies implies the need to identify the explanatory variables to estimate the predicted measures. Further developments in the data analysis technique shall contemplate alternative preprocessing options for data reductions to identify the relevant measures for estimating the predicted measures.

Acknowledgement

The authors gratefully acknowledge that the publication in Open Access was funded by the University of Verona by means of the "Fondo straordinario di Ateneo per la pubblicazione in Open Access".

References

Agnes Cheng, C., Zishang Liu, C. and Thomas, W. (2012) 'Abnormal accrual estimates and evidence of mispricing', *Journal of Business Finance and Accounting*, Vol. 39, Nos. 1–2, pp.1–34.

Bartov, E., Gul, F.A. and Tsui, J.S. (2000) 'Discretionary-accruals models and audit qualifications', *Journal of Accounting and Economics*, Vol. 30, No. 3, pp.421–452.

Bouasabah, M. (2024) 'Analysis of machine learning's performance in stock market prediction, compared to traditional technical analysis indicators', *International Journal of Data Analysis Techniques and Strategies*, Vol. 16, No. 1, pp.32–46.

Cagnoni, S., Ferrari, L., Fornacciari, P., Mordonini, M., Sani, L. and Tomaiuolo, M. (2021) 'Improving sentiment analysis using preprocessing techniques and lexical patterns', *International Journal of Data Analysis Techniques and Strategies*, Vol. 13, No. 3, pp.171–185.

Corlu, C.G., Goyal, A., Lopez-Lopez, D., Torre, R. D.L. and Juan, A.A. (2021) 'Ranking enterprise reputation in the digital age: a survey of traditional methods and the need for more agile approaches', *International Journal of Data Analysis Techniques and Strategies*, Vol. 13, No. 4, pp.265–290.

Dechow, P.M. and Dichev, I.D. (2002) 'The quality of accruals and earnings: The role of accrual estimation errors', *The Accounting Review*, Vol. 77, pp.35–59.

Dechow, P.M., Sloan, R.G. and Sweeney, A.P. (1995) 'Detecting earnings management', *The Accounting Review*, Vol. 70, No. 2, pp.193–225.

Dechow, P.M., Sloan, R.G. and Sweeney, A.P. (1996) 'Causes and consequences of earnings manipulation: An analysis of firms subject to enforcement actions by the sec', *Contemporary Accounting Research*, Vol. 13, No. 1, pp.1–36.

DeFond, M.L. and Park, C.W. (2001) 'The reversal of abnormal accruals and the market valuation of earnings surprises', *The Accounting Review*, Vol. 76, No. 3, pp.375–404.

Guay, W.R., Kothari, S.P. and Watts, R.L. (1996) 'A market-based evaluation of discretionary accrual models', *Journal of Accounting Research*, Vol. 34, pp.83–105.

Hainmueller, J. (2012) 'Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies', *Political Analysis*, Vol. 20, No. 1, pp.25–46.

Hainmueller, J. and Xu, Y. (2013) 'ebalance: A stata package for entropy balancing', *Journal of Statistical Software*, Vol. 54, No. 7, pp.1–18.

Harshe, M.V. and Kulkarni, R.H. (2018) 'Outlier detection using weighted holoentropy with hyperbolic tangent function', *International Journal of Data Analysis Techniques and Strategies*, Vol. 10, No. 2, pp.182–203.

Healy, P. (1996) 'Discussion of a market-based evaluation of discretionary accrual models', *Journal of Accounting Research*, Vol. 34, pp.107–115.

Healy, P.M. and Wahlen, J.M. (1999) 'A review of the earnings management literature and its implications for standard setting', *Accounting Horizons*, Vol. 13, No. 4, pp.365–383.

Jann, B. (2021) *Entropy Balancing as an Estimation Command*, University of Bern Social Sciences Working Papers 39, University of Bern, Department of Social Sciences.

Jones, J.J. (1991) 'Earnings management during import relief investigations', *Journal of Accounting Research*, Vol. 29, No. 2, pp.193–228.

Jr., O.P.H. and Stanley, D.J. (2012) 'A comparative modelling analysis of firm performance', *International Journal of Data Analysis Techniques and Strategies*, Vol. 4, No. 1, pp.43–56.

Karthikeyan, G. and Balasubramanie, P. (2020) 'A novel attribute-based dynamic clustering with schedule-based rotation method for outlier detection', *International Journal of Business Intelligence and Data Mining*, Vol. 16, No. 2, pp.214–230.

Kothari, S.P., Leone, A.J. and Wasley, C.E. (2005) 'Performance matched discretionary accrual measures', *Journal of accounting and economics*, Vol. 39, No. 1, pp.163–197.

Leuz, C., Nanda, D. and Wysocki, P.D. (2003) 'Earnings management and investor protection: an international comparison', *Journal of Financial Economics*, Vol. 69, No. 3, pp.505–527.

McMullin, J. and Schonberger, B. (2022) 'When good balance goes bad: A discussion of common pitfalls when using entropy balancing', *Journal of Financial Reporting*, Vol. 7, No. 1, pp.167–196.

McMullin, J.L. and Schonberger, B. (2020) 'Entropy-balanced accruals', *Review of Accounting Studies*, Vol. 25, No. 1, pp.84–119.

McNichols, M. and Wilson, G.P. (1988) 'Evidence of earnings management from the provision for bad debts', *Journal of Accounting Research*, Vol. 26, No. 1, pp.1–31.

Michael, C.J. and William, H.M. (1976) 'Theory of the firm: Managerial behavior, agency costs and ownership structure', *Journal of financial economics*, Vol. 3, No. 4, pp.305–360.

Ravisankar, P., Ravi, V., Raghava Rao, G. and Bose, I. (2011) 'Detection of financial statement fraud and feature selection using data mining techniques', *Decision Support Systems*, Vol. 50, No. 2, pp.491–500.

Rossignoli, F. and Tommasi, N. (2024) *Abnormalest: Abnormal Accrual Estimation*, https://codeocean.com/capsule/6827601/tree/v1

Sarkar, A., Sahoo, G. and Sahoo, U. (2016) 'Feature selection in accident data: an analysis of its application in classification algorithms', *International Journal of Data Analysis Techniques and Strategies*, Vol. 8, No. 2, pp.108–121, PMID: 77484.

Schipper, K. (1989) 'Earnings management', *Accounting horizons*, Vol. 3, No. 4, pp.91.

Seelammal, C. and Devi, K.V. (2018) 'Multi-criteria decision support for feature selection in network anomaly detection system', *International Journal of Data Analysis Techniques and Strategies*, Vol. 10, No. 3, pp.334–350.

Sujatha, R., Maheswari, B.U. and Mansurali, A. (2022) 'Application of machine learning algorithms to predict survival of micro small and medium enterprises in India', *International Journal of Data Analysis Techniques and Strategies*, Vol. 14, No. 4, pp.317–335.

Watts, R.L. and Zimmerman, J.L. (1986) 'Positive accounting theory', *The Accounting Review*, Vol. 65, No. 1, pp.131–156.