Identifying Outliers When Creating an Imputation Base for the Quarterly Financial Report

Melvin McCullough, Terry L Pennington
U.S. Census Bureau, Washington DC

Abstract

The Quarterly Financial Report Survey (QFR) collects income and balance sheet data for most manufacturing corporations and for large mining, wholesale, and retail corporations. Unit non-respondents are imputed using a combination of ratio and mean imputation. In order to enhance the imputation process by eliminating influential cases from the base, we investigated an iterative regression approach of outlier detection. The approach utilizes a combination of two regression diagnostics, leverage and studentized deleted residuals. We compared the effectiveness of the “regression fits” approach to the Hidiriglou-Berthelot method of outlier detection for several positive valued QFR items. To evaluate the effectiveness of the approaches, we created plots of inliers and outliers. The “regression fits” approach can also detect outliers for negative valued QFR items.

Key Words: imputation, outlier

Disclaimer: This report is released to inform interested parties of research and to encourage discussion. The views expressed are those of the author and not necessarily those of the U.S. Census Bureau.

1. Background

The Quarterly Financial Report (QFR) program publishes aggregate statistics on the financial results and position of U.S. corporations. Based upon a sample survey, the QFR presents estimated statements of income and retained earnings, balance sheets, and related financial and operating ratios for manufacturing corporations with assets of $250,000 and over, and mining, wholesale trade and retail trade corporations with assets of $50 million and over. The statistical data are classified by industry and by asset size. Industry classification is determined based on the North American Classification System (NAICS), which provides a two-digit to six-digit classification depending on level of detail.

1.1 A Description of the Data Collected on the Form

The questionnaire is divided into three sections: (1) income and retained earnings; (2) assets; and (3) liabilities and stockholders equity. The characteristics of each of the questionnaire items differ greatly. Survey items in the income section have a mixture of strictly positive items along with items that can be negative. Several asset items have a large proportion of reported zeros while a number of other asset items are strictly positive. In addition, there are several items from the liabilities section that can be negative.

1.2 An Overview of QFR Imputation Methodology

QFR imputation methods are based on unit non-response and the imputation cell is defined by stratum and three-digit NAICS. Ratio imputation is the primary imputation method utilized when prior quarter data is available. When prior quarter data is unavailable, or a new company is added to the survey, weighted means imputation handles non-response. Ratio imputation operates by adjusting a prior quarter reported data value $X_{it\text{-}k}$, or ‘auxiliary,’ by the current quarter imputation cell trend ratio to obtain the imputed value X_{it} via $X_{it-k} * R = X_{it}$. The cell trend, or ratio-of-identicals R, represents the amount of growth or decline in the cell for the current quarter relative to a prior quarter. This ratio comprises the imputation base for the ratio method of imputation and is defined as $R = \sum W_i X_{it} / \sum W_i X_{it-k}$ where W_i is the weight, X_{it} represents the current quarter item value, and $X_{it\text{-}k}$ represents the item value in a prior quarter. Note that the current quarter is represented by t while the prior quarter is represented by t-k, k=1, 2, 3, 4. The ratio-of-identicals includes only cases that responded in both current and prior quarter t-k. The method assumes that trends for the non-responding units are similar to those of the responding units.

The reasonableness of the imputed data depends on the effectiveness of identifying and removing influential cases (outliers) from the imputation base.

1.3 Research and Development of the Regression Fits Approach

In order to obtain adequate cell counts, we chose to pursue the outlier investigation by defining group by two digit NAICS (NAICS2)*STRATUM**.** As a preliminary step, we investigated several items from the income, assets, and liabilities sections of the questionnaire. For each item and group (NAICS2*STRATUM), we regressed the prior quarter version of the data against current quarter. We determined if an intercept term in the model was statistically significant. The strength of the correlations and the estimates for the slope and intercepts were noted. For most items and groups, the intercepts were not statistically significant. The estimates for the regression coefficients usually ranged from 0.8 to 1.2. For most items by group, the R-squared statistic was reasonably strong --usually greater than 0.8. We concluded that a reasonable way to predict current quarter data by item*group would be to utilize a linear regression no intercept model. The prediction would be based on the most recent available prior data for that item and group. Residual plots were investigated to get an indication of how to stabilize the error variance for each of the survey items. Generally, we found that either using the predictor or the square root of the predictor worked fairly well in obtaining a standard deviation function that explained the increasing variance as a function of the predictor.

We developed an algorithm that involved executing multiple regression runs in order to obtain estimated coefficients with desirable properties. The first step involved running an un-weighted regression with a no intercept model between the current to prior survey item saving the absolute residuals. The absolute residuals were then regressed against the predictor or a function of the predictor to obtain a set of fitted values. The fitted values from the second regression were in turn utilized to obtain a set of weights used to stabilize the error variance. The third regression run, this time weighted, yielded error terms that had a more nearly constant variance. Based on our regression algorithm, we came up with a strategy to identify outliers. The general strategy involved running a few iterations of weighted regressions in which the weights were computed in such a way to stabilize the variances over the range of the predictor. At each step, outliers were identified and removed, the diagnostics recomputed, and the regression line refit. The formula for weighted regression in matrix form follows:

$b_w = (X^{'}WX)^{-1}X^{'}WY$ where X is the predictor, Y is the response, and W are regression weights derived from a standard deviation function. The W_i are obtained by $W_i = \frac{1}{\hat{S}^2_i}$ and $\hat{S}_i = b_0 + b_1 Z_i$ where Z is chosen to estimate the error term in an ordinary least squares regression modeled as: $Y = b_1 X + e$

1.4 Description of the Regression Fits Approach

Three criteria were utilized to detect outliers by the regression fits approach. If a given case met any one of the criteria, the case was identified as an outlier and removed from subsequent regression runs. “Criteria1” was based on the computed leverage value (hat), where leverage is a measure of distance of a reporting unit from the average (weighted mean) reporting unit in that group, and the computed value of the studentized deleted residual relative (RSTD) to respective specified upper limits of these diagnostics. “Criteria2” was based on the RSTD compared to a specified upper limit of this diagnostic. “Criteria3” was based on hat relative to a specified upper limit of this diagnostic. Note that our approach relies on a combination of two regression diagnostics. Our rationale for using two regression diagnostics for outlier identification was to minimize the number of non-outliers identified as outliers (TYPE1 ERROR).

Let “hatcrit1” and “rstdcrit1” be the respective specified upper limits of the hat criteria and RSTD criteria based on a combination of leverage and studentized deleted residual respectively. Let “hatcrit2” be the upper limit of hat criteria and “rstdcrit2” be the upper limit of RSTD criteria based on each diagnostic alone. The constraints are imposed that hatcrit2>hatcrit1 and rstdcrit2>rstdcrit1. The three criteria to check for outliers are as follows:

Criteria1: hat > hatcrit1/num and RSTD> rstdcrit1
Criteria2: RSTD > rstdcrit2
Criteria3: hat > hatcrit2/num: where num is the number of cases per group.

Note that leverage-- the diagonal elements of the hat matrix H are obtained from:

$h_{ii} = X_i^{'}(X^{'}X)^{-1}X_i$ where X has dimension NX1, X transpose has dimension 1XN, and N is the number of cases per group. Studentized deleted residuals are based on the following formula[1] $RSTD_i = \frac{e_i}{\sqrt{MSE_i(1-h_{ii})}}$

Where e_i is the ordinary residual and MSE_i is the Mean Squared Error with the ith case omitted and h_{ii} is the ith diagonal element from the hat matrix.

1.5 Limitations of the Regression Fits Approach

Since the outlier detection procedure is based on a weighted linear regression model, it assumes that the two survey items are correlated. The regression fits method does not work as well for survey items that have a large proportion of reported values of zero. The iterative approach of excluding outliers and re-computing the resulting diagnostics presumes that the numbers of cases per group are moderate to large. The user should consider choosing less iteration for items that have relatively small respondent counts by group. The regression fits approach relies on having an available item for which the regression error term can be utilized to obtain a standard deviation function to obtain regression weights.

1.6 Setting Limits for Leverage and Studentized Deleted Residuals

We reviewed a few established guidelines[2] to identify cases with high leverage that influence the fitting of a regression equation. One rule frequently used is to consider a case to have high leverage if hat> (2/num) where num is the number of cases in that group. Using this rule, a case is considered influential if hat is greater than twice the average leverage for the group. Another rule considers as very high leverage those cases having hat>0.5. The rule considers cases with 0.2< hat<0.5 as having moderately large values of leverage. Based on empirical results to flag cases as extreme outliers, we set limits for outlier detection based on the "hat criteria" of at least 12/num. We desired to keep to an absolute minimum the possibility of incorrectly identifying cases as outliers (TYPE1 error). Setting guidelines for influential cases using criteria based on studentized deleted residuals (RSTD) is straightforward because the (RSTD) follow a t distribution with n-2 degrees of freedom[3]. One approach to testing RSTD is to utilize a Bonferroni testing procedure that adjusts the TYPE1 error rate based on multiple tests. For the data we utilized, the group sizes were usually between 50 and 300. An example of joint limits for outlier detection based on criteria1above and that minimize the TYPE1 error rate are: (hat>(12/num) and RSTD>4.0).

1.7 Evaluation of Regression Fits for Selected Negative Valued Survey Items

In order to evaluate the regression fits outlier detection approach, we chose three real valued income items. Income-Loss from Operations (E104), Income Loss Before income taxes (E111), and Net Income-Loss for Quarter (E118) were three survey items containing a fair proportion of negative valued data. Moreover these three items contained few cases of reported zeros. For each group (NAICS2*STRATUM) we created plots of prior to current quarter by item. For positive values, we transformed the data to obtain the square root of the original data value. For negative data values, we transformed the data creating the negative of the absolute value of the square root of the original data. The transformed data made comparisons by graphing easier due to the large range in the original data values. Plot 1, as an example, depicts three outliers chosen in group 31*18 for item E104 using regression fits. Outliers are denoted by the "*' symbol. Note that the outlier corresponding closely to the transformed order pair (1200, 1200) is a point of high leverage.

For each group, we recorded the total number of cases and number of outliers identified. Generally we found that by group the regression fits approach detected from 0.5% to 3% of the cases as outliers. A review of the plots showed a definite trend between prior to current quarter data values by group. Usually there was definite separation geometrically between the chosen outliers and non-outliers. Occasionally, we noticed that two points appearing equally spaced from the trend line and "center of mass" would be treated differently-- one as an outlier and the other not. We attributed these anomalies to the fact that criteria1 utilizes a combination of leverage and studentized deleted residual values for outlier detection. Overall, the regression fits approach was effective in identifying outliers. The effectiveness of regression fits depends on making a good choice of a variable from which to obtain a standard deviation function that is in turn utilized to obtain weights for a regression.

2. Discussion of the HB and Regression Fits Outlier Detection Methods

2.1 The Hidiroglou-Berthelot Method (HB)

The Hidiroglou_Berthelot (HB) is a reliable method of outlier detection for positive valued survey items often used by periodic business surveys at the Census Bureau. One of the strengths of HB is that it utilizes two parameters U and C, in determining outliers to control for the size of the reporting unit and width of the acceptance region respectively. The size parameter U may assume any value in the region 0<=U<=1. When U is 0, size has no effect on the HB statistic that is used to detect outliers but as U increases toward 1 the term in the maximization function below increases so that the HB statistic is more extreme—placing greater emphasis on size as a criteria for outlier detection[4]. HB uses a transformed ratio that results in S_i that has a symmetric distribution centered at 0. Let X_i be the survey item to be evaluated, Y_i the auxiliary item (prior quarter), and R_i the ratio of X_i to Y_i. The transformed ratio, S_i is given by the following relation:

$$S_i = \frac{R_i - R_{med}}{R_i} \quad 0 < R_i < R_{med} \qquad S_i = \frac{R_i - R_{med}}{R_{med}} \quad R_i > R_{med}$$ where R_{med} is the median ratio of X to Y.

The HB statistic along with the parameter U is given by the following formula: $E_i = S_i * \{\max(W_i X_i , W_i Y_i)\}^U$

2.2 Comparison of HB and Regression Fits

We compared the HB and the regression fits approaches to outlier detection for a few positive valued items. We desired an approach to outlier detection that effectively flags a relatively small number of extreme cases for each group and properly separates the outliers from the non-outliers. Both HB and Regression Fits rely on a moderate to strong correlation between the survey item tested and the auxiliary item. Again HB is a transformed ratio edit that provides a parameter U that allows the user to place more emphasis on flagging larger reporting units as outliers. HB always chooses outliers symmetrically about the trend line. HB uses a parameter C such that when C increases, the acceptance region is increased resulting in fewer cases flagged as outliers.

In contrast, the regression fits approach equally evaluates units based on the distance from a trend line expressed in units of standard deviations irrespective of the size of the reporting unit. Neither method of outlier detection works effectively on pairs of items that are weakly correlated or on data pairs that contain a high proportion of reported zeros. Regression fits, unlike HB, may flag cases that are “close” to a trend line but have high leverage values. Pairs of units that are consistent but have very large leverage may need to be identified. Imputation procedures rely on the removal of influential cases from the base in order to minimize the chances of obtaining unusually small or large imputes.

2.3 Comparisons of Graphical Analysis Between HB and Regression fits

In order to compare and contrast the differences in HB and regression fits results; we selected the positive valued survey items sales (E101) from the income section and total assets (E223) from the asset section. For each of the item*group combinations, we recorded the total number of cases, the number of cases flagged as outliers by HB, the number of cases flagged as outliers by regression fits, and the number of cases for which there were differences in classification. The comparisons were based on HB parameters C=40 and U=0.5. We used hatcrit1=16 and hatcrit2=32 as well as rstdcrit1=4.0 and rstdcrit2=6.0 for regression fits in making the comparisons.

We found complete agreement between HB and regression fits in outlier detection for 5 of the 12 groups investigated for item E101. Three of the five groups for which there was agreement in outlier detection had no outliers selected for either method. As shown in Table A, the NAICS2*stratum groups 32*18 and 44*16 had the same three outliers detected for each of the two methods. Two of the remaining seven groups had ‘substantial agreement’ between the two methods. For one group, HB selected three outliers below the trend line whereas regression fits selected a set of three outliers above the trend line. The six cases classified as outliers for this group were all ‘borderline cases’, therefore adjusting the parameter settings for either method would have likely yielded different results. The results of the comparisons in outlier detection between HB and Regression Fits for item E101 are summarized in Table A. A graphical comparison is illustrated for the group 21*18 by a review of Plot 2 and Plot 3.

We found complete agreement in outlier detection between HB and regression fits for 7 of the 12 groups investigated for item E223. For 3 of the 7 groups for which there was complete agreement between the two methods no outliers were detected. Of the 5 groups for which there was disagreement in outlier detection, 4 groups differed by one case classified differently between the two methods. Overall, the agreement in the outlier sets chosen from the two methods was remarkably good given the differences in approach. Please refer to Table A columns 7-9 for a summary of counts of outliers detected based on the two approaches.

We evaluated the effectiveness of the regression fits outlier approach for both real and positive valued survey items. In evaluating the approach, we considered the relative number of outliers chosen to the total count by item*group. We also considered the relationship of outliers to non-outliers within the group as shown by a review of plots. Investigation of plots was crucial in determining whether or not the procedure was properly differentiating outliers from non-outliers. The results indicated that regression fits approach was effective in distinguishing outliers for the negative valued survey items chosen from the questionnaire. In making a comparison between HB and regression fits, for positive valued items, we experimented with a few different combinations of parameter settings for the two approaches. An overall objective comparison is difficult since one can obtain fewer counts of outliers from one of the approaches by simply adjusting the parameters for that approach. We settled on parameter settings that yielded comparable results for the two methods of outlier detection. Based on our comparisons that included counts and review of plots we concluded that regression fits yielded similar results to HB for positive valued survey items.

Acknowledgements

We wish to acknowledge Donald Luery for his contribution to the outlier research and his part in creating a SAS macro outlier program based on Regression Fits as described in this paper. The SAS outlier program is available upon request

Table A: Comparison of Outlier Results Between HB and Regression Fits for Items E101 and E223

			Item E101			**Item E223**		
NAICS2 (1)	Stratum (2)	Number Of Cases In Group (3)	Outliers Detected By HB (4)	Outliers Detected By Regression Fits (5)	Classification Differences Between Methods (6)	Outliers Detected By HB (7)	Outliers Detected By Regression Fits (8)	Classification Differences (9)
21	16	43	0	0	0	0	0	0
31	16	119	2	0	2	0	0	0
32	16	155	3	4	1	1	2	1
33	16	293	1	2	1	0	2	2
42	16	383	3	1	2	0	0	0
44	16	132	3	3	0	1	1	0
21	18	149	4	3	1	0	1	1
31	18	215	0	0	0	2	2	0
32	18	350	3	3	0	1	2	1
33	18	568	0	3	3	1	1	0
42	18	444	3	3	3	2	3	1
44	18	191	0	0	0	1	1	0

NAICS2 Descriptions

21	Mining	31	Food, Beverage, Textiles, Apparel, and Leather Manufacturing
32	Paper, Printing, Petroleum and Coal Products, Chemical, Plastic	33	Primary Metal, Fabricated Metal, Machinery, Computer and Electronic, Electrical, Transportation Equipment, Furniture, and Miscellaneous
42	Wholesale Trade	44	Retail Trade

References

[1] SAS/STAT Guide for Personal Computers, Version 6 Edition Page 837

[2] Regression Diagnostics: Identifying Influential Data and Sources of Collinearity: Belsley D.A.,E Kuh, and R.E. Welsh

[3] Applied Linear Statistical Models: Neter, Kutner, Nachtssheim, and Wasserman

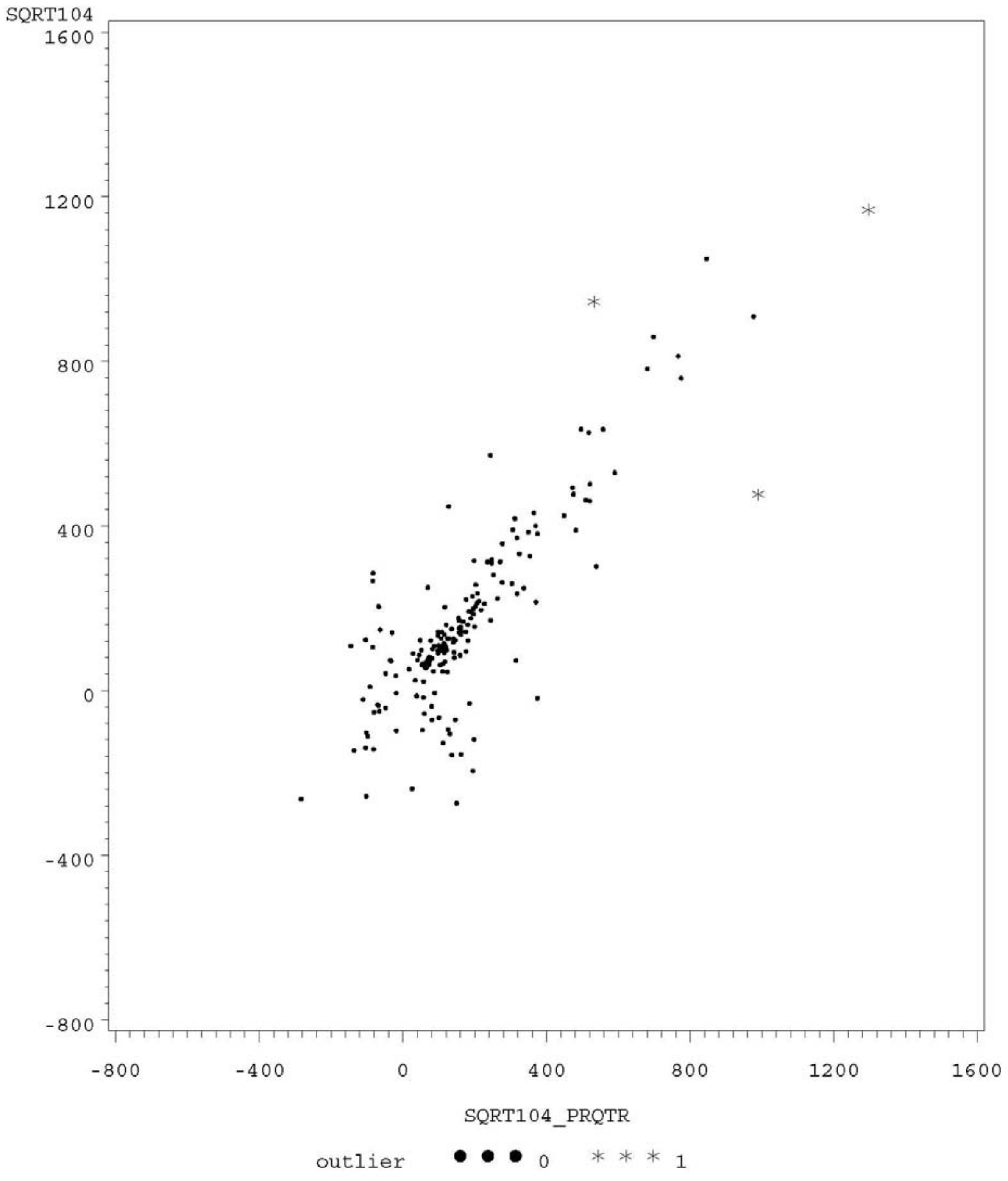
[4] Application of the Hidiroglou-Berthelot Method of Outlier Detection For Periodic Business Surveys: Richard Belchar

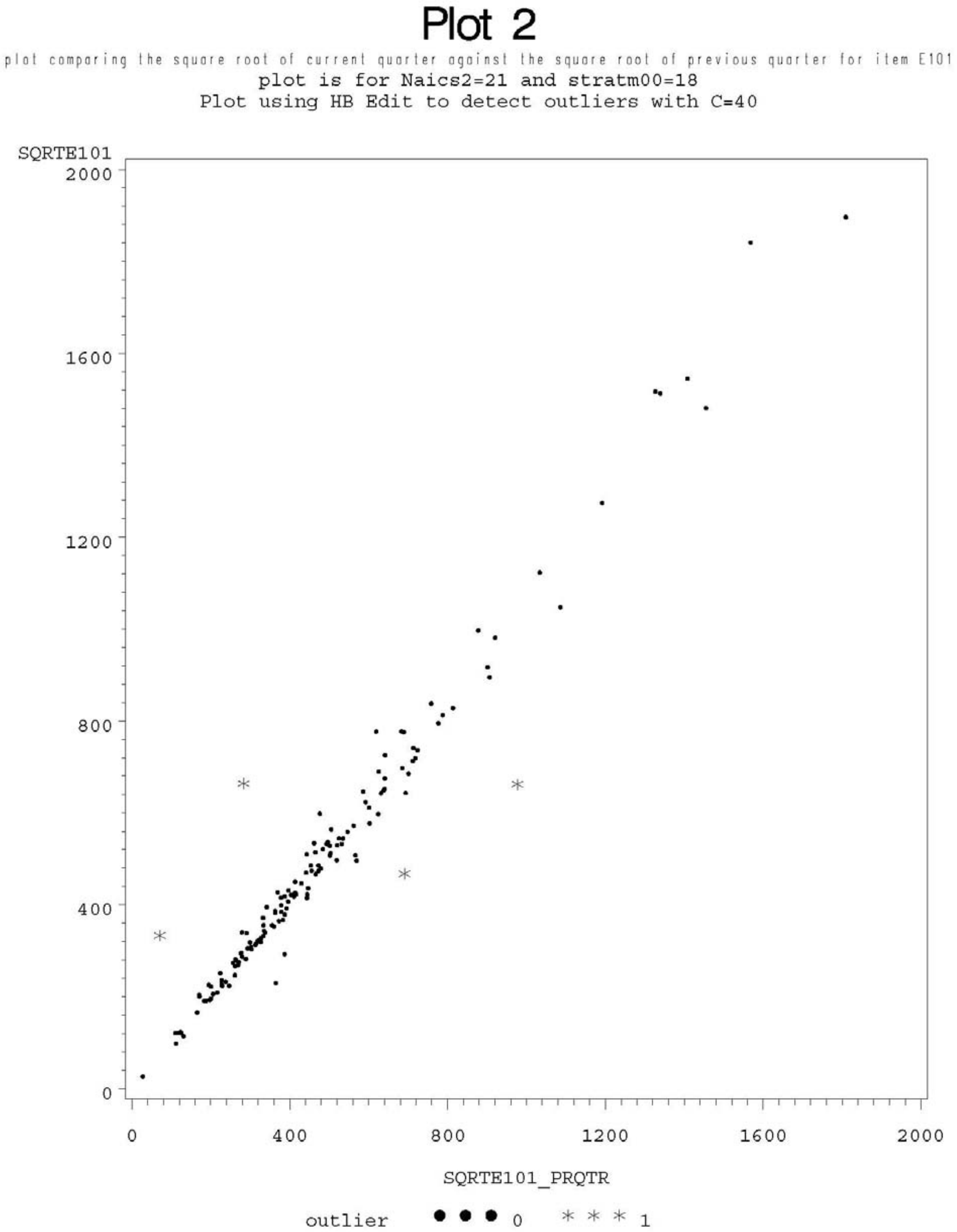
Plot 1

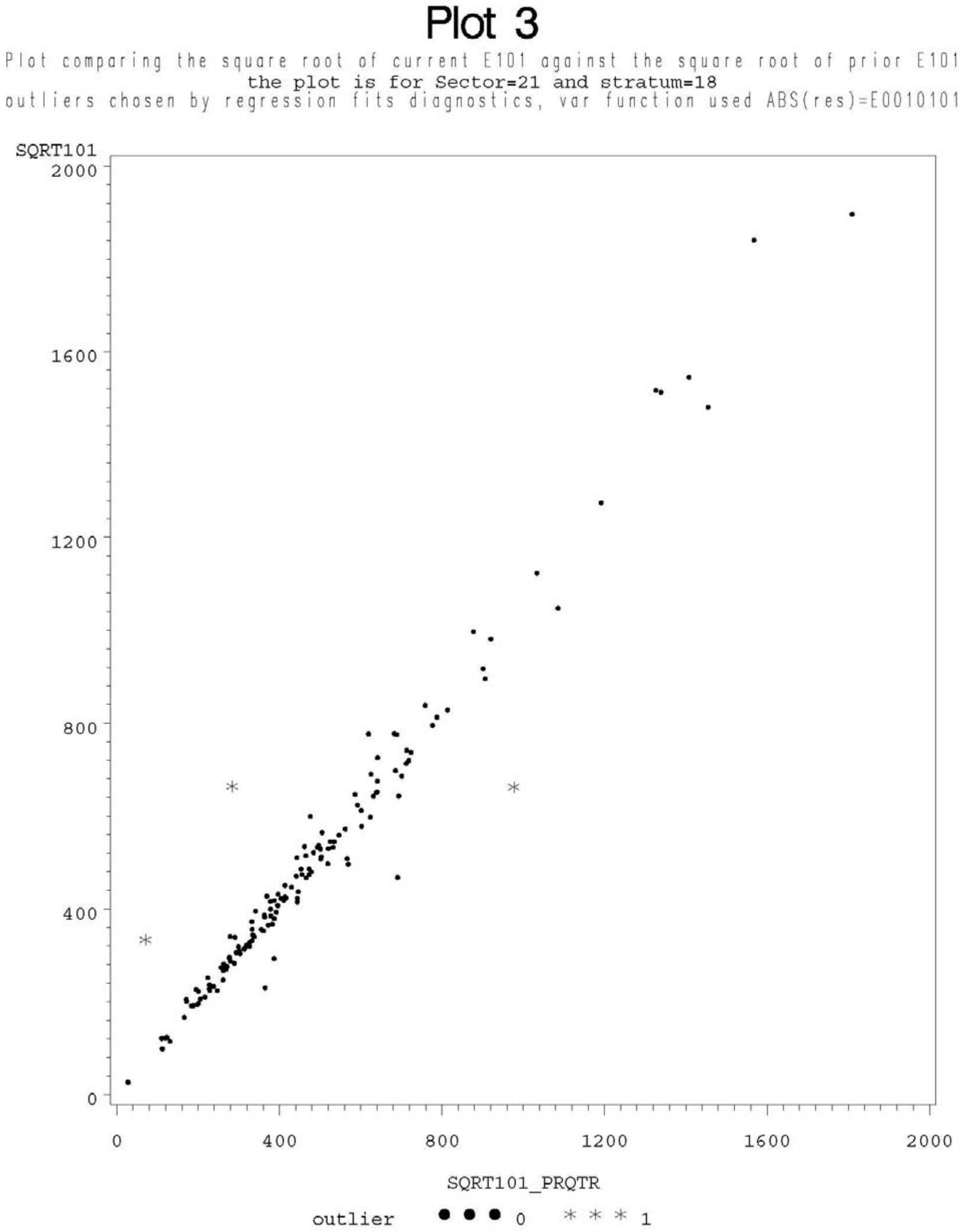
Plot comparing the square root of current E104 against the square root of prior E104

The plot is for Naics2=31 and stratum=18

outliers chosen by regression fits diagnostics, var function used ABS(res)=E00104




Plot 2
plot comparing the square root of current quarter against the square root of previous quarter for item E101
plot is for Naics2=21 and stratm00=18
Plot using HB Edit to detect outliers with C=40
SQRTE101
2000
1600
1200
800
400
0
0
400
800
1200
1600
2000
SQRTE101_PRQTR
outlier ● ● ● 0 * * * 1


Plot 3
Plot comparing the square root of current E101 against the square root of prior E101
the plot is for Sector=21 and stratum=18
outliers chosen by regression fits diagnostics, var function used ABS(res)=E0010101
SQRT101
2000
1600
1200
800
400
0
0
400
800
1200
1600
2000
SQRT101_PRQTR
outlier ● ● ● 0 * * * 1