A STATISTICAL EDIT FOR LIVESTOCK SLAUGHTER DATA

Cathy Mazur, USDA-NASS
Rm 4168 South Building, Washington DC 20250

KEY WORDS: Tukey's Biweight, Robust Estimation, Outlier, Inlier, Double Root Residual.

INTRODUCTION. A cooperative program currently exists between the Food Safety and Inspection Service (FSIS), National Agricultural Statistics Service (NASS), and Agricultural Marketing Service (AMS) for collecting livestock slaughter data in federally inspected plants. The joint effort involves data collection, editing, summarization, and public dissemination of data. In addition to being published, data on the number of head is currently used by NASS as check data for the Quarterly Agriculture Survey (QAS). Current livestock numbers are validated by adding births and subtracting deaths from the previous livestock figure in a balance sheet approach.

In the current edit system, data are entered on the personal computer using a software package called KeyEntry III and uploaded at the end of the day to a leased mainframe. The data are then edited using a Generalized Edit System (a parameter driven program run in batch mode). The results of the edit are available several hours later at higher cost, or the next morning at a lower cost. Analysts must pore over printouts in order to resolve errors, and corrections must be rekeyed on the personal computer. An outlier for head data is a value which differs more than a given percent from the plant's previous 3 week average (calculated using positive and zero kill days), and an outlier for weight data is a value which is outside some predetermined weight range for each class of livestock.

One problem with this edit is that some head data values are incorrectly identified as outliers during holiday weeks. A reason for this is that the current edit does not take the non kill days into account. Therefore, plants which do not kill the same number of days each week (as occurs during holidays) are not being edited reasonably. A second problem is that plants slaughtering specialty weight animals (e.g. lower weight veal calves) are incorrectly flagged as errors. The reason for this is that the same edit limits are used for all plants. These as well as other problems compelled Livestock Branch (who runs the survey in NASS) to request improved editing techniques for livestock slaughter data.

Consequently, a research project was initiated to develop specifications for a statistical edit for livestock slaughter data, by utilizing each plant's historic data. The problem with head data during holidays could be solved by basing the edit only on positive kill days using a robust estimator. Plants which slaughter 5 or 6 days a week provide enough positive data in a few weeks to calculate a daily average, but plants which only slaughter 1 or 2 days a week would not supply enough data. Therefore, more weeks must be used in these cases. A way to handle the problem with specialty weight plants is to use a statistical approach, by editing each plant based on that plant's historic data. In addition, plants with a lower coefficient of variance in head counts and weights would be edited more accurately, checks on a plant's weekly slaughter pattern (head data) would be made, and the edit would take place interactively on the personal computer.

DATA. A census of federally inspected slaughter plants occurs each week using a one page mail questionnaire. Plants report daily numbers of head kill (Monday through Saturday), and weekly dressed and live weight totals. The species of livestock include Cattle, Calves, Hogs, Sheep, Goats and Equine. The class of livestock refers to animals within species, for example, steer, heifer, cows, and bulls and stags are classes within the species cattle. Long term historical information for each plant is available.

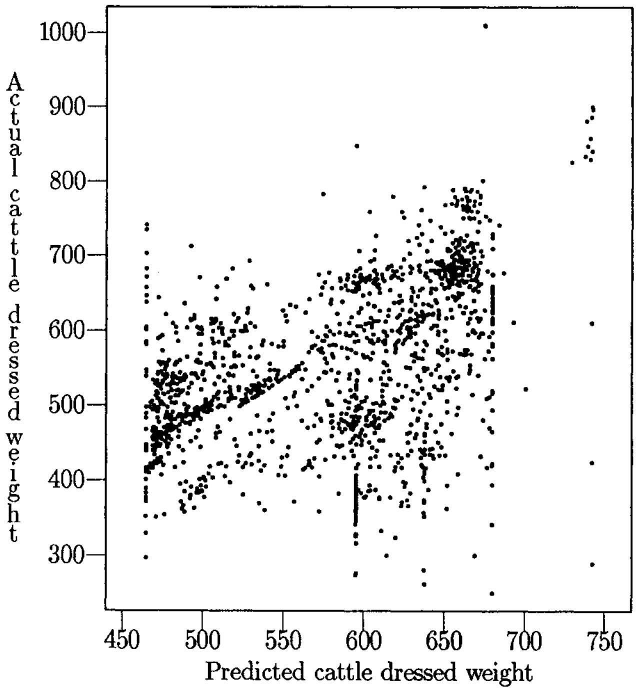
A basic understanding of the plants in the universe is fundamental. As is typical of establishments, plant size (equal to the number of animals in a class) has a skewed distribution (many small plants and a few large plants). Although small plants make up the majority, data from large plants dominate the summary. For example, 68.9% of all plants had between 1 and 1,000 head of cattle in 1987, but represent only 0.8% of the total number of cattle. However, 49.6% of all cattle are found in plants which had over 500,000 head in 1987 (1.4% of all plants). One reason that a statistical edit is appropriate is that the plants are so different from one another, both in average weight per animal, and in the number of head slaughtered (totals per week, number of days per week, and consistency of pattern). For example, small steer plants have varying average dressed weights (300-700) and CVs (0-35%), but large steer plants tend to have weights between 650-750 and CVs <10. The significance of these plant differences are shown by comparing Graphs 1 (which uses universe means) and Graph 2 (which uses individual plant means). The predicted cattle weight was calculated for each plant (represented by a dot) by multiplying the number of animals (for steer, heifer, cows, and bulls and stags) by their mean weights (using 1 of the 2 methods). The method shown in Graph 2 predicts the weights better than the method shown in Graph 1.

To facilitate research, 64 weeks of data from all plants in 5 states were obtained. The states of CT,

Graph 1

Actual *vs.* Predicted Cattle Dressed Weight

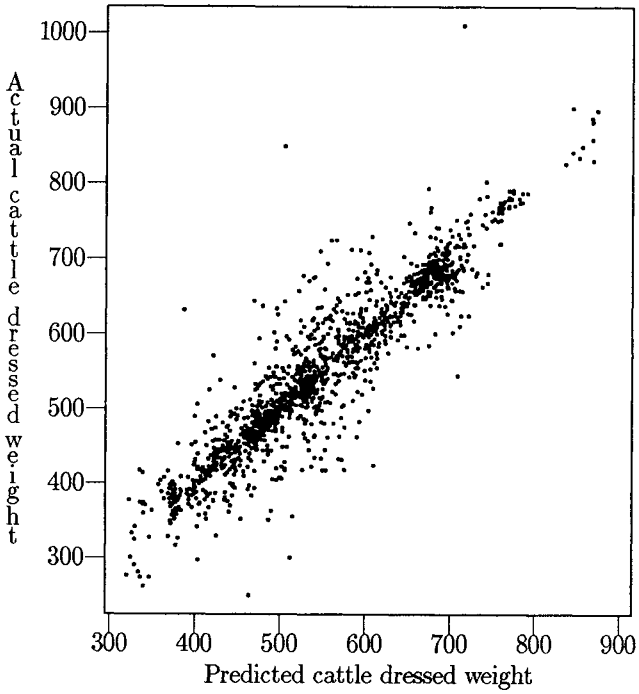
(Using First Half Universal Means)



Graph 2

Actual *vs.* Predicted Cattle Dressed Weight

(Using Biweight Means)



ME, MD, MA and TX were selected to ensure that some animals of each class and species were available. Large hog operations were not well represented in this data. A subsequent sample for all plants in one week was obtained. Using this data, we could determine how well our 5 state sample represented one typical slaughter week.

GENERAL METHODOLOGY. The purpose of the livestock edit is to determine whether plant data is reasonable (that is, to check for reporting errors and keying mistakes). Errors can be in the form of outliers (a value which is outside the edit limits) as well as inliers (a value which does not change, or changes very little over time).

1. Identification of Outliers.

The first step in constructing a statistical edit was to determine which statistical estimator to use. The goal was to choose a measure of center and spread that would quickly stabilize to new levels when true changes did occur in the data, or return to old levels in the presence of outliers.

As to time series models, the exponentially weighted moving average was considered, and a few data sets were analyzed using a time series analysis package. However, the results were not overwhelming. Also, a plot of weekly average steer dressed weights by size group showed no obvious trend. An examination of many time plots for steer dressed weights showed many plants with no trend, and the plants with a trend were large. Since only 64 weeks of data were available, research on time series models and any seasonality effect was postponed.

Robust measures were considered, as they work well in many distributions of data, and are less susceptible to outliers; whereas, the standard statistical method (mean and standard deviation), works best only in the Normal distribution, and is affected by outliers. Robustness can be "thought of as an insensitivity to underlying assumptions. In practice, we never know the underlying conditions precisely, due to data disturbances. Therefore, we seek estimates that do well for a variety of underlying conditions." (See reference 1 page 283 and 297)

In the **initial analysis**, classical types and simple robust measures were examined. This included 4 measures of center and 4 measures of spread over 4 time periods (6, 10, 13, and 26 weeks). Weight data were used to evaluate these measures. The definitions of these measures follow.

Measure of Center

a) Mean - sum all values (X_i) and divide by n (the number of values). $\overline{X} = \Sigma X_i$ / n. Note: We now reorder the data from smallest to largest, where $X_{[1]}$ is the smallest, $X_{[k]}$ is the k^{th} order statistic, and $X_{[n]}$ is the largest of the n observations.

b) Median - M = $X_{[m]}$, where

$$X_{[m]}=\begin{cases}(X_{[n/2]}+X_{[(n/2)+1]})/2 & \text{if n is even}\\ X_{[(n+1)/2]} & \text{if n is odd}\end{cases}$$

c) Trimmean - T1 = (Q1+2M+Q3)/4, where Q1 and Q3 are the 25th and 75th percentiles. If m is noninteger, then let m=m-0.5.

$$Q1=\begin{cases}X_{[(m+1)/2]} & \text{if m is odd}\\ (X_{[m/2]}+X_{[(m/2)+1]})/2 & \text{if m is even}\end{cases}$$

$$Q3=\begin{cases}X_{[n-((m-1)/2)]} & \text{if m is odd}\\ (X_{[n-(m/2)]}+X_{[n-(m/2)+1]})/2 & \text{if m is even}\end{cases}$$

d) 20% Trimmed Mean (T2) - the lowest n*0.20 values and the highest n*0.20 values are dropped, then T2 is the average of the center n*0.60 values.

Measure of Spread

e) Standard Deviation (SD) - sum the squares of the deviations of each value from the mean, and divide by n-1 (one less than the number of values).

$$SD=\sqrt{\Sigma(X_i-\overline{X})^2/(n-1)}$$

f) Inter-Quartile Range (IQ), IQ = Q3 - Q1.

g) Median Absolute Difference (MAD) - tranform each value by subtracting the median (M), and taking the absolute values. Then obtain the new median of the transformed values.

$$MAD = \text{median}\ \{|\,X_i - M\,|\}$$

h) 20% Trimmed Standard Deviation (TSD) is the standard deviation of the center n*0.60 values.

The different measures of center can be distinguished by the weights which the values receive. The mean gives equal weights of 1/n to each value. The 20% trimmed mean gives equal weights of 1/(0.6*n) to the center 60% values and zero to the lower 20% and upper 20% of the values. The trimean gives a weight of 1/4 to the 25th and 75th percentiles, and 1/2 to the median. The median gives a weight of 1 to the center value (or 1/2 to the center two values) and zero to all other values. The measures of spread provide weights in a similar manner.

Several conclusions were made following the analysis with regards to the measures of center. When outliers were present, the mean changed considerably, as all values (good and bad) were included. The trimean was dropped early in the analysis, as the mean, median, and 20% trimmed mean seemed sufficient. The median and 20% trimmed mean were inadequate as good values were being excluded (e.g. the upper and lower 20% in the trimmed mean, and all but the center values in the median).

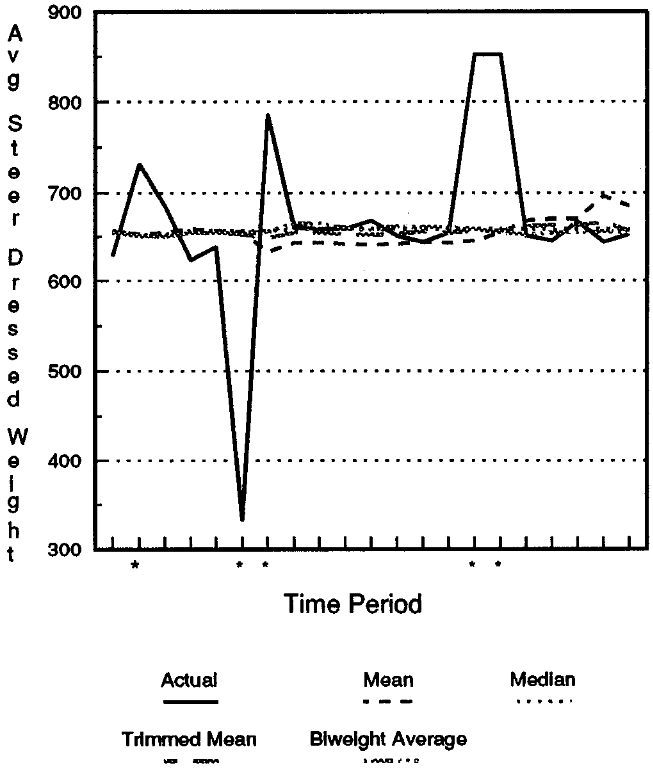
A visual comparison of these measures of center is shown in Graph 3, using one particular data set with several outliers (represented by an '*'). The actual data (the solid line) is the average steer dressed weight for a week, where the term "average" refers to the total steer dressed weight for a given week divided by the total number of steer for that week. The first value represents one week in a long series. Therefore, the 13 values prior to that week were used to calculate the corresponding measures of center. The measures of center are close, but the mean does tend to lag a bit.

As to the measures of spread, the standard deviation is greatly affected by outliers. The 20% TSD, the IQ range, and the MAD (although robust) are also inadequate due to the exclusion of good data.

A visual comparison of the measures of spread are given in Graph 4, using the same data set from Graph 3. The SD increases drastically due to the outliers in the 6th and 7th time periods. In fact these outliers may cause the system to miss the outliers in the 15th and 16th time periods, since the prior 13 values are used. The IQ range and MAD wer normalized to make a better comparison with the SD (represented as SIQ and SMAD). Although the SIQ, SMAD, and TSD are not nearly as affected by the outliers, the concern is that they underestimate

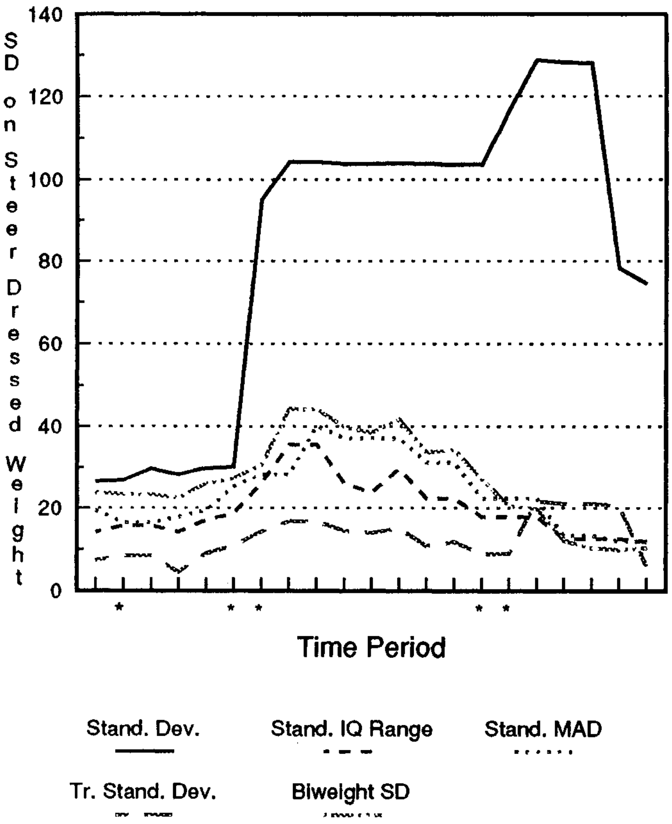
GRAPH 3

PLOT OF 5 MEASURES OF CENTER OVER TIME (13 wks)



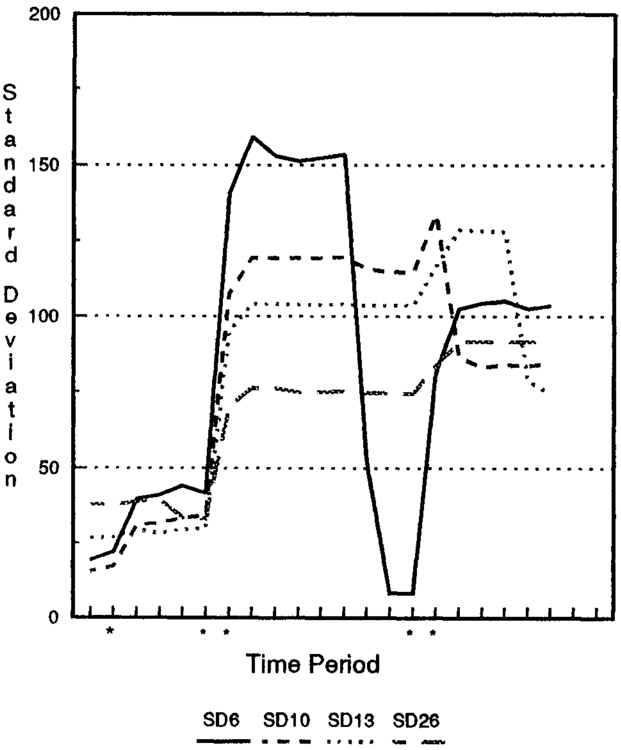
GRAPH 4

PLOT OF 5 MEASURES OF SPREAD OVER TIME (13 wks)



GRAPH 5

PLOT OF STANDARD DEVIATION vs. TIME PERIOD



As to the number of values used in the calculations, the 6 and 10 week time periods provided unreliable measures of spread. The 26 week period required too much data (half a year), and took longer to detect changes. Graph 5 displays the standard deviation calculated using the four time periods. The 6 week SD ranges from 19 to 153, and the outlier at week 6 affects the SD for 6 weeks. The 10 and 13 week calculations peak at subsequently lower values, but the effect of the outlier is felt over more weeks. The 26 week SD is much more constant with gradual (but minimal) increases due to the outliers.

These shortcomings were a motivation to do a literature search to find other estimators. The set of statistical measures from the first analysis were L estimators, or linear combinations of order statistics. One characteristic of these estimators is that the same weights are used for all data sets, that is, the weight is independent of the data set. For example, the median of any data set (where n is odd) is calculated by giving the center value a weight of one, and all other values a weight of zero. In the next analysis, we looked at an estimate from the class of W and M estimators called Tukey's biweight. This class of estimators differs from L estimates, in that weights differ for different data sets, that is the weight is dependent on the data set.

In the **second analysis**, Tukey's biweight was calculated using a 13 week time period. Head data was also used to evaluate this measure (using the number of whole weeks so that at least 13 positive values occurred).

The biweight mean (BiAv) and biweight standard deviation (BiSd) incorporate unequal weights, where reasonable values are given weights close to 1, and unreasonable values (outliers) are given very small weights or are excluded altogether (by giving a zero weight). The BiAv has the advantage of the mean if the data is normal (all good values are included), but has the advantage of the median if outliers are present (it excludes them).

The biweight has many interesting properties. It is flexible, yet computationally simple. The basic form is iterative (M estimator), but a single step form is available (W estimator). [Note: The first step uses the median and IQ range (or MAD) to calculate the weight which is used in calculating BiAv and BiSd. The second step then uses BiAv and BiSd to recalculate the weight, which is used to calculate the new BiAv and BiSd.] Also, the biweight takes into account the grouping and rounding effect, where livestock weights may be rounded (for example) to the nearest 25 pounds. [Note: This refers to how changes in values near the center of the distribution can affect the estimator.] The only assumptions for the biweight are that outliers are symmetric, and that the percent of outliers is less than 50 percent (see the discussion of the breakdown bound on pages

357-8 in reference 1). In symmetric distributions, the measures of location almost coincide. In skewed distributions, the targets differ, and a bias must be considered (see pages 287-9 in reference 1).

The calculation of BiAv and BiSd requires each X_i value to be assigned a weight (i.e. a measure of distance from the center of the distribution). The user determines how this weight is calculated, as shown below. In our case, the median was chosen as the measure of center, and both the IQ range and the MAD were considered as measures of spread. The parameter "c" represents the number of measures of spread a value must be away from the measure of center before the value (X_i) is excluded entirely (6 and 9 were used in this analysis).

$$Wt_i = \frac{X_i - M}{c * IQ} \quad \text{or} \quad \frac{X_i - M}{c * MAD}$$

Note that the MAD is equal to 0.6745 times the SD, and the IQ range is equal to 1.349 times the SD. Thus, the distance away from the median (M) a value (X_i) must be to be excluded entirely (c*IQ or c*MAD) can be normalized as follows.

6 MAD = 4.05 SD
9 MAD = 6.07 SD
6 IQ = 8.09 SD
9 IQ = 12.14 SD

Using the weight, the BiAv and BiSd are defined as follows, where values having a weight greater than the absolute value of 1 are excluded from both calculations. Note, one problem with the BiSd is that it is possible for a value to have a negative term in the denominator. (See the discussion on redescending estimators on pages 397-8 of reference 1).

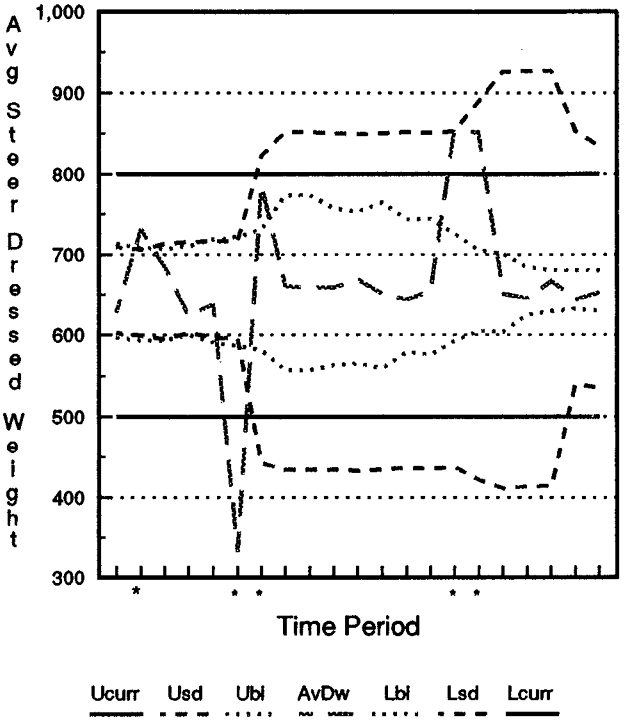
$$\text{BiAv} = \frac{\sum [X_i * (1 - Wt_i^2)^2]}{\sum (1 - Wt_i^2)^2}$$

$$\text{BiSd} = \frac{\{ n * \sum [(X_i - M)^2 * (1 - Wt_i^2)^4]\}}{[\sum (1 - Wt_i^2) * (1 - 5 * Wt_i^2))]}$$

Graphs 3 and 4 include BiAv and BiSd in their comparisons of measures of center and spread using IQ as the initial measure of spread, and c=6. The analysis identified two problem cases. The first occurred when the IQ range was used in a data set with greater than 25% outliers (BiSd was too large). The second problem occurred when the MAD was used in a data set with two similar size clusters (BiSd was too small as the second cluster was ignored). Therefore, we decided to use the biweight with c=6 and IQ as the measure of spread, with a test for cases where the proportion of outliers is greater than 25 percent. In this case, the MAD will be used.

GRAPH 6

CONFIDENCE INTERVALS on AVERAGE DRESSED WEIGHT
(using current method, mean/standard deviation, and biweight)



Edit limits are obtained by calculating a confidence interval from BiAv and BiSd. For the livestock edit, the calculation of the confidence interval will require that the Coefficient of Variance be at least 1% (if not, BiSd will be set to 1% of BiAv). The t distribution with 0.7*(n-1) degrees of freedom is recommended for biweight intervals. However, an additional factor (given below) should be used for sample sizes less than 20 (see reference 1, page 423).

n	factor	df	n	factor	df
13	1.071	8.4	17	1.044	11.2
14	1.068	9.1	18	1.036	11.9
15	1.063	9.8	19	1.023	12.6
16	1.055	10.5	20	1.009	13.3

Graph 6 compares 3 edit ranges using the current method, a confidence interval using $\bar{X}$ and SD, and a confidence interval using BiAv and BiSd.

Since each value is an average (total weight divided by the number of animals for a week), a weighted approach will be used to account for the different numbers of animals each week. Additional discussions of the biweight and its properties are found in chapters 9-12 of reference 1.

2. Identification of Inliers

An investigation of data from the analysis, showed that the weights for some plants do not change much over time. A few explanations for this are plants with the same imputed value over time, plants without the proper scales that report the same average weight over time, and plants with a low coefficient of variation. The Double Root Residual (DRR) is a measure which provides a way to identify inliers in a distribution. The DRR is calculated each week. SDRR is then the sum of the DRR over time (where w is the week).

$$DRR = \sqrt{(2+(4*obs.))} - \sqrt{(1+(4*pred.))}$$

$$SDRR = \sum |DRR_w|$$

A value is flagged as an inlier when the SDRR is below a certain value, that is, the biweight (pred.) too closely predicts the observed value (obs.). (See reference 5.) As an example, week 1 in graph 3 has an actual value of 628, and a predicted value (BiAv using IQ and c=6) of 655. Therefore DRR_1 is

$\sqrt{(2+(4*628))} - \sqrt{(1+((4*655))}$, which equals -1.056.

FEATURES OF THE SYSTEM. The discussion above provides a statistical framework. The following is a discussion of methods for implementing these techniques into the system. A validation edit is basically a within record check. These include identification code checks, checks that certain rows and columns sum to the appropriate totals, checks that the number of head in the head section corresponds to the number of head in the weight section, and that dressed weight is less than live weight (or %DW/LW is between 0 and 1). A statistical edit is a between-record check (in our case, using historical data within plant across time). The general features of the system are provided below.

1. **Stratification/Imputation**
   Slaughter plants will be stratified based on size (the number of animals) for each class.
   a) A biweight from the strata will be used to edit plants with not enough data to calculate their own biweight (<13 values in the last year). It can also be used for new, changed or seasonal plants.
   b) A biweight from the strata will be used to edit small plants (i.e., under 20 animals per day).
   c) The plant's biweight will be used to impute missing weight data. A strata biweight will be used for plants with too little data (<13 values in the last year).
2. **Journal**
   The journal file will be used to identify errors. The ability to sort the errors on some measure (such as plant size) will be available. Also, a way to determine the effect of the edit on the summary, through an audit trail will be available.
3. **Master ID File**
   This file can be used to identify plants which are closed for some reason (strike, holiday or other), but it can also be used to verify id codes and protect against duplication.
4. **Missing Analysis Routine**
   This routine will enable the user to determine the number of plants not yet reported for a week, and the effect on the summary.
5. **User Interaction**
   The user will be able to set the necessary parameters, and the strata definitions.
6. **Interactive Microcomputer-based Edit**
   The integrated system will use DBase III+ on the PC to enter, edit, and summarize the data. One reason for using a database package was the ability to updata (correct) records at any time. Currently, updates are only done once per year due to the high cost of processing a sequential file on the mainframe. A modular program will allow changes or other data series to be incorporated.

COSTS. The new edit system will result in substantial cost savings. The yearly leasing cost of processing and storing data on the mainframe (federally inspected plants) will be exchanged for microcomputer equipment which will be purchased initially (network), but require only maintenance charges thereafter. Equipment purchases will be low, as several PCs are already available. The non-federally inspected plants (used in the summary) will be on the mainframe, but will be downloaded to the PC for summary. Roughly speaking, a 75% savings will result the first year, and a 81% savings the following years (compared to what it would have cost on the current edit).

FUTURE RESEARCH.
1. Apply technique to other data series.
2. Incorporate graphics of historic time series plots.
3. Look at potential seasonality effects.

REFERENCES:

1. David Hoaglin, Frederick Mosteller, and John Tukey, editors, *Understanding Robust and Exploratory Data Analysis*, (New York: John Wiley & Sons, 1983).
   (Note the Reference to 1981 Personal Communication by John Tukey on pg. 387.)
2. Andrews, et al. *Robust Estimates of Location: Survey and Advances* (Princeton, NJ: Princeton University Press,1972).
3. Charles Du Mond and Russell Lenth, *A Robust Confidence Interval*, Abstract from the 1986 ASA Proceedingsof the Statistical Computing Section, pgs. 139-143.
4. Mark Pierzchala, *A Review of the State of the Art in Automated Data Editing and Imputation,* NASS Staff Report No. SRB-88-10, 1988.
5. Paul F. Velleman and David Hoaglin, *Applications, Basics, and Computing of Exploratory Data Analysis* (Boston, MA: Duxbury Press, 1981) pgs. 265-266.