Chapter 8

Statistical Techniques -- I

Chair: Alan Estes, Federal Reserve Board

Michael Scrim

Linda Simpson ◆ Henry Chiang ◆ Cathy Tomczak

Adeline J. Wilcox

Rethinking the Editing Algorithm for the Survey of Employment Payrolls and Hours

Michael Scrim, Statistics Canada

Chapter 8

Abstract

The original editing systems and methodology of 11 years was costly, rigid, and required much human intervention. As a part of the redesign of the Survey of Employment Payrolls and Hours (SEPH), a revamped editing system was created. It has two major components, the first is that it takes into account the factors of industrial detail, firm size and seasonality. How it does this is with "curved bounds" (Hidiroglou-Bertholot Bounds), allowing more variation in data for larger firms. The second part of the change, equally as important, was the use of a Score function, a tool used to rank all records with errors to allow resources to focus on the most severe cases. The end results have been a cost saving of around $260,000 annually, and the number of human interventions reduced to around 1,500 records a month versus 30,000 with the old system while maintaining data quality. ■

A Statistical Edit for Livestock Slaughter Data

Linda Simpson, Henry Chiang, and Cathy Tomczak,
National Agricultural Statistics Service

8
Chapter

Abstract

For the past four years, the National Agricultural Statistics Service has edited data from its weekly survey of livestock slaughter plants using a PC-based statistical edit system. The data consist of daily numbers of cattle, hogs, calves and sheep inspected by U. S. Department of Agriculture meat inspectors, as well as weekly live-weight and dressed-weight totals.

This interactive edit is based on a robust estimator, called Tukey's Biweight. Each plant's historical data are used to flag outliers, determine which species are normally inspected, determine if a pattern is typically followed (i.e., slaughter only on Saturdays), and impute for missing data. This allows a "custom" edit for specialty (i.e., veal) or very large plants, and frees up time to reconcile data problems not possible with the previous mainframe edit.

The previous system, which was keyed on a PC, but edited on a mainframe in batch mode, used a generalized edit system. The edit, available several hours later, flagged values if they differed more than a given percent from the plant's previous three week average or outside some predetermined range; and it did not impute for missing data. ■

Chapter 8

A CSFII Data User's Principal Components Analysis for Outlier Detection

*Adeline J. Wilcox, Beltsville Agricultural Research Center**

Abstract

In my analysis of Continuing Survey of Food Intakes by Individuals 1989-1991 respondents whose dietary folate intake exceeded the safe upper limit recommended by the Centers for Disease Control and the Food and Drug Administration, I found seven of 277 high intakes which appeared to me to be due to coding errors. These apparent errors are of two types; one intake where 217 grams of *Tang*-powdered concentrate was reported, evidently 1 cup beverage prepared with water miscoded as 1 cup powdered concentrate, and six intakes where *Kool-Aid* appears to have been miscoded as *Tang*. *Tang* is a folate-fortified food.

I investigated the usefulness of principal components analysis for detecting these outliers. I ran the SAS procedure PROC PRINCOMP on two sets of variables. First, intake of energy, cholesterol, carbohydrate, vitamins A and C, folate and iron, and age. Second, intake of energy, carbohydrate, vitamin C and folate and age.

All seven of these outliers could have been discovered by examining the top percentile of folate intake. Using the second, smaller set of variables, six of the seven outliers could have been found in the 0.16 percent of the data with the largest positive first principal components. The only female among these outliers cannot be detected in this principal components analysis combining data from both sexes.

*Present affiliation: U.S. Bureau of the Census

A CSFII Data User's Principal Components Analysis for Outlier Detection

*Adeline J. Wilcox, Beltsville Agricultural Research Center**

Introduction

I investigated the usefulness of principal components analysis (PCA) for detecting suspect observations in nutrient intake data from the U.S. Department of Agriculture's Continuing Survey of Food Intakes by Individuals (CSFII) 1989-91. This work is related to my effort to estimate the proportion of the U. S. population whose intake of dietary folates on any given day exceeded the safe upper limit recommended by the Centers for Disease Control (CDC) and the Food and Drug Administration (FDA). Among those with extreme intake of dietary folates, there are a few whose high values may be due to coding error. When the proper variables are selected, PCA can identify observations known to be suspect.

In the analysis I plan, the weighted numerator of the proportion will comprise those with dietary folate intake of at least 1000 µg. The denominator will be a weighted total of all survey respondents. For a conservative estimate of this proportion, I will remove observations with apparent coding errors from the numerator.

Dietary Folate Intake Assessed by USDA Food Consumption Surveys

The CSFII 89-91 endeavored to collect three consecutive days of food consumption data from persons sampled in the 48 coterminous states. USDA obtained daily nutrient intake totals for each respondent by using food composition data to summarize the information collected on the kinds and quantities of food individuals consumed. Nutrient intake from dietary supplements such as multivitamin tablets was not included in these totals.

I used final USDA in-house CSFII data. Public use data are available (U. S. Department of Agriculture, 1996a, 1996b). After I did this work I discovered the in-house CSFII 89-91 nutrient intake data have more digits right of the decimal than the public use data.

*Present affiliation: U. S. Bureau of the Census

In CSFII 89-91, 15,398 respondents provided a total of 39,696 days of dietary intake data over the three days they were surveyed. Of the 15,398 respondents, 232 reported dietary folate intakes of at least 1000 µg on at least one day. Both the CDC (Public Health Service, 1992) and the FDA (Department of Health and Human Services, 1993) have advised limiting total folate intake to less than 1000 µg per day. Because some of these 232 respondents exceeded this safe upper limit on more than one day they were surveyed, 277 of the 39,696 dietary intakes exceeded the safe upper limit for folate. See Tables 1 and 2.

Table 1.--Intakes Judged Complete by Survey and Day Surveyed

		Intakes Judged Complete		
		Dietary Folates		
Survey	Day	>= 1000 µg	< 1000 µg	Total
1989-91	One	116	15,076	15,192
	Two	89	12,281	12,370
	Three	72	12,062	12,134
	Total	27	39,419	39,696
1994	One	53	5,536	5,589
	Two	38	5,273	5,311
	Total	91	10,809	10,900

Table 2.--Days Excess Dietary Folate Intake Consumed by Survey

	Number of Days Respondents Reported Dietary Folate Intake of at Least 1000 µg				
Survey	Zero	One	Two	Three	Any
1989-91	15,166	195	29	8	232
1994	5,502	83	4	*	87

*Only two days surveyed in 1994.

The histogram in Figure 1 shows an estimate of the distribution of dietary folates intake among the U.S. population for 1989-91. Day One refers to the first of the three days individuals were surveyed. Observations from individual respondents have been weighted up to the population total. The dotted line at 1000 µg of dietary folate intake marks the safe upper limit recommended by CDC and FDA. The dietary folate intake distribution is right-skewed with a long tail extending well past the 1000µg mark. The maximum value came from a respondent who enjoyed fried chicken livers. In my estimate of the proportion of survey respondents with dietary folate intakes of at least 1000 µg, the numerator consists entirely of extreme values. Dietary folate intake exceeding 1000 µg is also a feature of the CSFII 94 data displayed in Figure 2.

Two Types of Apparent Coding Errors

To study the source of this excess dietary folate, I listed the food codes which contributed the most folate to each diet and caused total daily folate intake to exceed 1000 µg. Among these, I found two types of apparent coding errors.

The first food code, 925-4200 FRT FLVRD DRNK, HI VIT C, FROM DRY MIX, caused six of the 277 dietary intakes to exceed 1000 µg folate. Although the description for this food code indicates the fruit-flavored drink mix contains added vitamin C, it does not reveal that it is also fortified with folic acid. See box below. *Tang* is a folic acid-fortified food. It appears that fruit-flavored drink mix without added folate (*Kool-Aid*) may have been coded as fruit-flavored drink mix fortified with folic acid (*Tang*) by coding some beverages as 925-4200 instead of 925-4101.

925-4101	Fruit-flavored drink, made from powdered mix, with sugar and vitamin C added (Include Kool-Aid, Wylers, NS as to sweetner)
925-4200	Fruit-flavored drink, made from powdered mix, mainly sugar, with high vitamin C added (Include Borden's Instant Breakfast Drink, Keen, Tang Instant Breakfast Juice Drink)

For one intake, 217 grams of food coded as 292-0010 TANG, DRY CONCENTRATE, was reported. One cup of dry Tang powdered concentrate weighs 217 grams. While it is humanly possible to consume this quantity of Tang powder, it seems likely that 1 cup beverage prepared with water was miscoded as 1 cup powdered concentrate.

In all, seven intakes appear affected by these two types of apparent coding errors. The records with apparent coding errors are listed in Table 3. If I had access to the original data collection forms, I would have been able to verify whether these items were coded correctly or not.

Figure 1.--USDA CSF11 89-91 Day One -- Weighted Sample of 15,192 Males and Females, All Ages

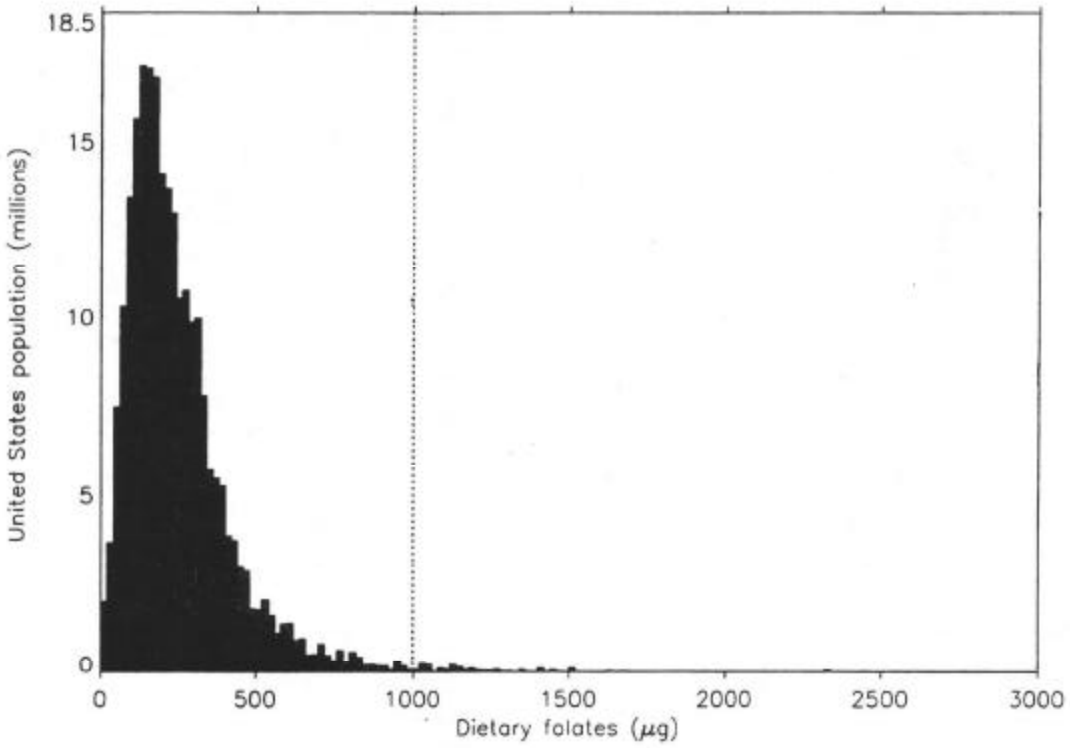


Figure 2.--USDA CSF11 94 Day One -- Weighted Sample of 5,589 Males and Females, All Ages

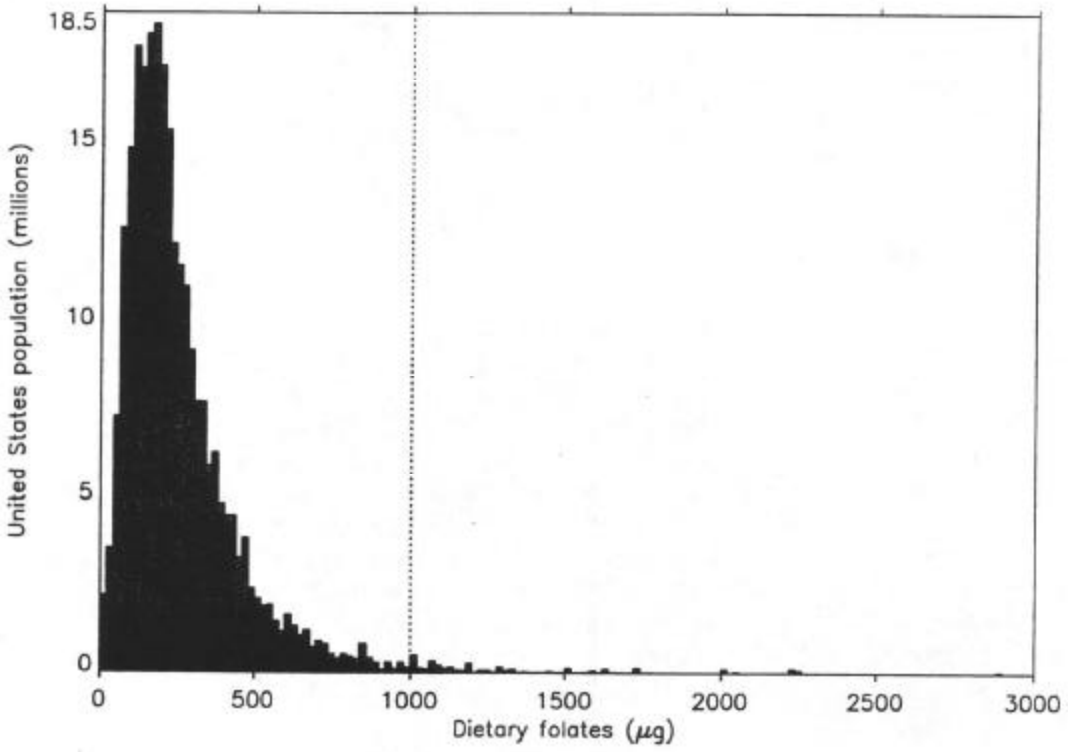


Table 3.--Suspect Observations, Their Principal Components and Ranks

ID	RESPNUM	DAYCODE	AGE	SEX	SOURCE	PRIN1OF8	RKPRN1_8	PRIN1OF5	RKPRN1_5	PRIN1OF4	RKPRN1_4	RANKFOLA
1116100	1	1	26	M	koolaid?	4.85739	229	6.9355	18	6.9360	18	33
1226079	4	3	13	M	powder	2.83278	576	4.7775	85	4.7728	85	137
2260877	1	1	65	M	koolaid?	4.00732	313	7.1605	17	7.1752	17	27
2260877	1	2	65	M	koolaid?	3.98032	317	7.4110	13	7.4260	12	22
2260877	1	3	65	M	koolaid?	7.21401	106	10.7103	2	10.7278	2	2
3117346	2	2	28	F	koolaid?	2.75879	606	3.3352	354	3.3348	355	262
3117346	3	1	11	M	koolaid?	2.65234	672	6.1076	28	6.1018	28	79

Principal Components Analysis

I investigated the usefulness of principal components analysis for detecting these seven outliers with PCA, a known statistical method for detecting outliers. In the Current Index to Statistics 1975-1993, I didn't find any entries on the use of PCA for editing nutrient intake data. Since the nutrient data base used to convert food consumption data to nutrient intake values contains no missing values (imputed values are used where data are not available), there are no missing values for respondents' daily nutrient intake totals, a great convenience for PCA. Nutrient intake data are particularly suitable for PCA because they are really collinear. I will explain collinearity in nutrient intake data later.

Selection of Variables for PCA

Inspecting my list of food codes contributing excess folate to respondents' diets, I learned that liver, folic acid-fortified cereal, and *Tang* were among the foods contributing to dietary folate intake in excess of 1000 µg. In my first attempt to search for outliers with PCA, I decided, based on my subject-matter knowledge, to use nutrients and food components for which liver and *Tang* are rich sources as variables. I didn't consider cereal because many brands are fortified with several vitamins and minerals. Liver is a rich source of iron, vitamin A, folate, and cholesterol, and a good source of vitamin C. Besides contributing carbohydrate to the diet, *Tang* is fortified with vitamin C and folic acid. Because nutrient intake is related to energy intake and age is related to energy intake, I also used energy measured in kilocalories and age measured in years as variables in my first look at the data with PCA. I selected eight variables for PCA: age, energy intake (KCAL), cholesterol (CHOL), carbohydrate (CHO), vitamin A (VITA), vitamin C (VITC), folates (FOLA), and iron.

Collinearity

The correlation matrix of variables selected for PCA reveals values for correlation coefficients greater than 0.5 between cholesterol and energy, carbohydrate and energy, iron and energy, folate and carbohydrate, iron and carbohydrate, and folate and iron. No relation between age and nutrient intake is apparent from the correlation matrix below.

	AGE	KCAL	CHOL	CHO	VITA	VITC	FOLA	IRON
AGE	1.00	-.03	0.02	-.05	0.09	0.02	0.04	0.02
KCAL	-.03	1.00	0.55	0.89	0.23	0.29	0.47	0.56
CHOL	0.02	0.55	1.00	0.32	0.24	0.09	0.21	0.26
CHO	-.05	0.89	0.32	1.00	0.23	0.35	0.51	0.56
VITA	0.09	0.23	0.24	0.23	1.00	0.24	0.45	0.36
VITC	0.02	0.29	0.09	0.35	0.24	1.00	0.49	0.28
FOLA	0.04	0.47	0.21	0.51	0.45	0.49	1.00	0.71
IRON	0.02	0.56	0.26	0.56	0.36	0.28	0.71	1.00

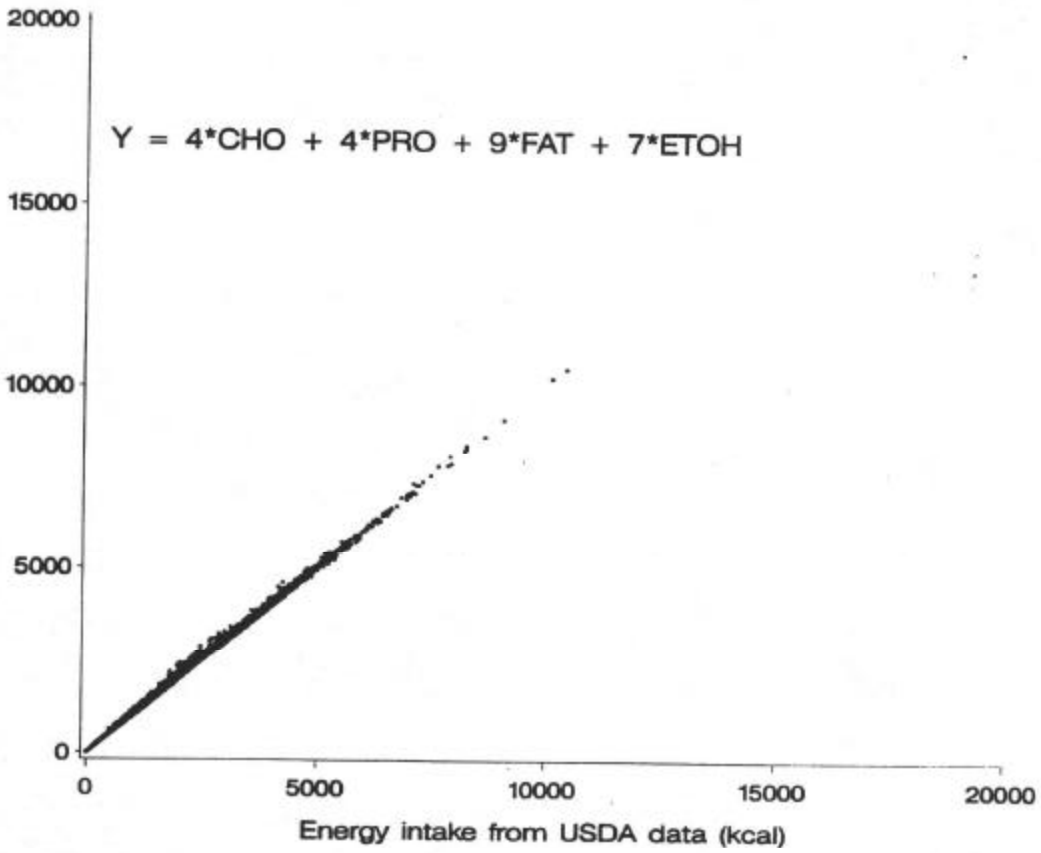
Nearly exact collinearity exists between energy intake and a linear combination of carbohydrate, protein, fat and ethanol intake. Figure 3 shows a linear combination of the macronutrient intake variables plotted against energy intake measured in kilocalories. This linear combination, Y, is energy intake computed from macronutrient intake, measured in grams. The units of the coefficients of this linear combination are kilocalories per gram. Most of the 39,696 points plotted in Figure 3 fall along a straight line at a 45 degree angle to the abscissa, indicating nearly exact collinearity. Note the outlying energy intake value of 18,955 kilocalories.

Principal components are orthogonal and rid me of collinearity.

Choice of Principal Components

I used the first principal component (PC) for detecting outliers, following the method I learned from Robert M. Hamer in a short course sponsored by SAS® Institute Inc. (Hamer, 1995). Johnson and Wichern recommend using the last few PCs for outlier detection (Johnson and Wichern, 1992). They illustrate use of scatter plots and Q-Q plots for finding an outlier in a small data set. This proved useless for my large data set. Plotting the seventh PCs against the eighth PCs left all seven suspect observations

Figure 3.--Evidence of True and Nearly Exact Collinearity



buried deep in the cloud of 39,696 points. With PCA on five variables, described later, I plotted the fourth PCs against the fifth PCs, but none of the seven suspect observations fell outside the dense scatter of points.

PCA Versus Checking the Top Percentile

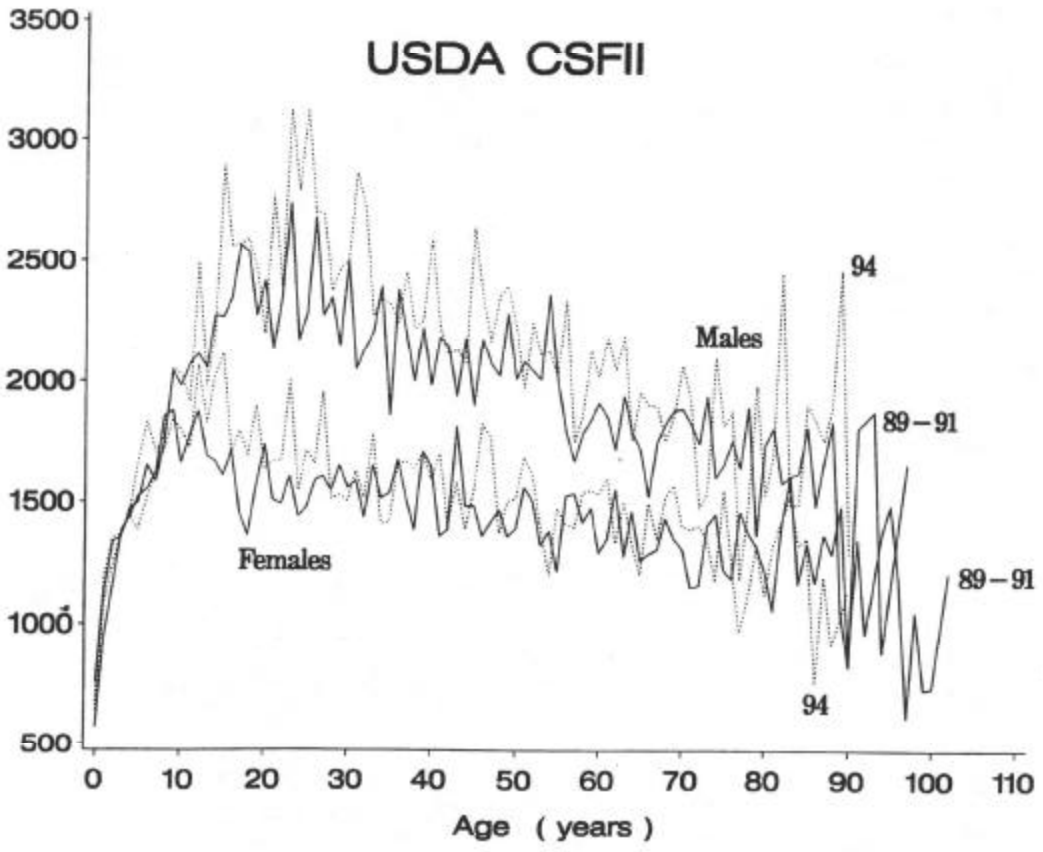
To be more useful than simply examining the top percentile of dietary folate intake, PCA should rank the seven suspect observations above the ranks obtained by sorting the data by descending dietary folate intake. In a column headed RANKFOLA, Table 3 shows the ranks assigned when the 39,696 records, sorted by descending dietary folate intake, are ranked. All seven of these outliers could have been discovered by examining the top percentile of folate intake.

Using the SAS procedure PROC PRINCOMP with the eight variables listed above, I got the first PCs listed in Table 3 under the heading PRIN1OF8. I ranked these by descending value as shown in the column headed RKPRN1_8. If one started checking data with the largest positive first PC and continued until all seven suspect observations had been examined, one would have looked at more than the top percentile of the data. Thus, PCA with these eight variables is not more efficient than checking the top percentile of dietary folate intake.

Since I was trying to identify outliers involving *Tang*, not liver, I eliminated nutrients or food components which liver, but note *Tang* is a rich source of, that is, iron, cholesterol and vitamin A. PCA with five variables results in PCs, PRIN1OF5, ranking the suspect observations in the 100 largest PCs except for the only female among them. Her first PC is ranked 354th. See the column headed RKPRN1_5 in Table 3. That this female outlier doesn't rise to the top of the data as well as the male outliers do when PCA is applied, suggests performing PCA separately on males and females.

At this point, I took another look at the correlation matrix and realized that age was not correlated with intake of any nutrient or cholesterol. However, I knew from experience with nutrient intake data that age and energy intake are related. Figure 4 shows median energy intake plotted against age for each age in years for both the CSFII 89-91 and CSFII 94. The median values for each gender are connected by solid lines for 1989-91. Dotted lines connect the median values for 1994. There is a relationship between age and energy intake but it is not a linear one. This plot shows the data used for detecting outliers, not the data used for assessing dietary intake. The difference between these data sets is the inclusion of the known energy intake of breast-fed infants. The medians for infants on this plot should not be used for nutritional assessment.

Figure 4.--USDA CSFII Median Day One Energy Intake by Age and Sex



Given the lack of a linear relationship between age and energy intake, I performed another PCA without the age variable. As shown by the columns headed PRIN1OF4 and RKPRIN1_4 in Table 3, neither the PCs nor their ranks changed much, indicating that age is not an important variable for detecting folate outliers with PCA.

Specificity of PCA

Laura Gillis inquired if I had looked at the specificity of PCA for outlier detection. I did not. In this analysis I studied only sensitivity. However, it is reassuring to note that, in the PCA using four variables, the largest PC belonged to the individual who reported consuming 18,955 kilocalories.

Conclusion

PCA shows promise for editing nutrient intake and food consumption data.

Acknowledgments

I wish to thank Will Potts and Gordon Marten for giving me the opportunity to learn more about multivariate methods of statistical analysis, Robert M. Hamer and Julie A. Smith for looking over this paper and making helpful comments, and my statistician friends for their nontechnical support.

References

Department of Health and Human Services, Public Health Service (1992). Recommendations for the Use of Folic Acid to Reduce the Number of Cases of Spina Bifida and Other Neural Tube Defects, *Morbidity and Mortality Weekly Report*, 41/No. RR-14.

Department of Health and Human Services, Food and Drug Administration (1993). Food Labeling: Health Claims and Label Statements: Folate and Neural Tube Defects, 58 *Federal Register* 53254-75.

Hamer, R. M. (1995). *Multivariate Statistical Methods: Practical Applications Course Notes*, Cary, NC: SAS Institute, Inc.

Johnson, R. A., and Wichern, D. W. (1992). *Applied Multivariate Statistical Analysis*, 3rd ed., Englewood Cliffs, NJ: Prentice-Hall, Inc.

U. S. Department of Agriculture (1996a). *Continuing Survey of Food Intakes by Individuals and the Diet and Health Knowledge Survey 1989-91* (on CD-ROM), Accession No. PB96-501747. Springfield, VA: National Technical Information Service.

U. S. Department of Agriculture (1996b). *Continuing Survey of Food Intakes by Individuals 1994* (on CD-ROM), Accession No. PB96-501010. Springfield, VA: National Technical Information Service. ■

This paper has not been approved through any clearance mechanism of the U. S. Department of Agriculture or any other agency of the United States Government. The views and findings presented are those of the author and not necessarily those of the U. S. Department of Agriculture.