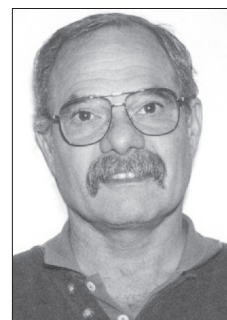

VISUAL REVELATIONS

*Howard Wainer,
Column Editor*



Old Mother Hubbard and the United Nations: An Adventure in Exploratory Data Analysis

Danielle Vasilescu and Howard
Wainer

**WEALTH IS NOT WITHOUT ITS
ADVANTAGES AND THE CASE
TO THE CONTRARY, ALTHOUGH
IT HAS OFTEN BEEN MADE,
HAS NEVER PROVED WIDELY
PERSUASIVE.**

John Kenneth Galbraith,
The Affluent Society, 1958

Our tale begins innocently enough as a homework assignment for Statistics 112--a second course in statistics for undergraduates at the University of Pennsylvania. Students were asked to find a publicly available data display that could be modified to serve its purpose better. One student found a table on the United Nations web site (<http://unstats.un.org/unsd/demographic/products/socind/housing.htm>) describing an aspect of the crowdedness of housing in 63 countries (Table 1). The table is arranged alphabetically and contains the year that the data were gathered and the average number of persons per

Column Editor: Howard Wainer, Distinguished Research Scientist, National Board of Medical Examiners, 3750 Market Street, Philadelphia, PA 19104; hwainer@nbme.org



room in that country. This latter figure was also broken down into its rural and urban components. Although the table adequately archived the information, it was not an evocative display that illuminated the character of housing crowdedness. The path toward that goal is the subject of this essay.

Revision was an iterative process of emails and conversations between

the student and the instructor, which pointed toward re-analyses and supplemental data gathering. The first email concerned the wide range of years during which the data were gathered. Did it make sense to compare Cameroon's crowdedness in 1976 with Brazil's in 1998? Two options were considered: The first was only to consider those countries whose data were no more than

Table 1—Average Number of Persons per Room*

Country or area	Year	Total	Urban	Rural
Aruba	1991	0.7
Austria	1997	0.7	0.7	0.7
Azerbaijan	1998	2.1	1.9	2.3
Bahamas	1990	1.3	1.3	1.1
Belgium	1991	0.6	...	0.6
Bermuda	1991	0.6	0.6	...
Bolivia	1988	...	1.7 a	...
Brazil	1998	0.7	0.7	0.8
Bulgaria	1992	1.0	1.2	0.8
Cameroon	1976	1.2	1.2	1.3
Canada	1996	0.5	0.5	0.5
China, Macao SAR	1996	1.1
Colombia	1993	1.4	1.3	1.7
Costa Rica	1997	0.9	0.8	1.0
Croatia	1991	1.2
Cuba	1981	1.0	1.0	1.0
Cyprus	1992	0.6 b	0.6 b	0.7 b
Czech Republic	1991	1.0	1.1	1.0
Egypt	1996	1.3	1.3	1.4
Finland	1998	0.8	0.8	1.0
France	1990	0.7	0.7	0.7
French Guiana	1990	1.1	1.1	1.5
Gambia	1993	1.5	1.3	1.6
Germany	1987	0.5 c	0.5 c	0.5 c
Guadeloupe	1990	0.9 d	0.9 d	0.9 d
Guam	1990	0.8	0.7	0.8
Honduras	1988	2.2 e	1.8 e	2.6 e
Hungary	1990	0.8	0.8	0.8
India	1981	2.7 d f g	2.6 d f g	2.8 d f g
Iraq	1987	1.5	1.8	1.0
Israel	1983	1.2
Japan	1978	0.8 h	0.8 h	0.7 h
Korea, Republic of	1995	1.1	1.2	0.9
Kuwait	1985	1.7 i
Lesotho	1996	2.1
Martinique	1990	0.9 d	0.9 d	0.9 d
Mauritius	1990	1.2	1.1	1.3
Mexico	1995	...	1.4 j	...
Netherlands	1998	0.7	0.6	0.6
New Caledonia	1989	1.2 d	1.1 d	1.4 d
New Zealand	1991	0.5 k	0.5 k	0.5 k
Nicaragua	1995	2.6	2.2	3.1
Norway	1990	0.6	0.6	0.6
Pakistan	1998	3.0 l	2.8 l	3.1 l
Panama	1990	1.6	1.4	1.9
Peru	1990	2.0	1.9	2.4
Poland	1995	1.0	0.9	1.1
Portugal	1991	0.7	0.7	0.7
Puerto Rico	1990	0.7	0.6	0.7
Reunion	1990	1.0	1.0	0.9
Romania	1992	1.3 k m	1.3 k m	1.2 k m
San Marino	1991	0.7	0.7	0.8
Serbia and Montenegro	1991	1.2	3.1	3.3
Slovakia	1991	1.2	1.2	1.2
Sri Lanka	1981	2.2	2.3	2.1
Sweden	1990	0.5	0.5	0.5
Switzerland	1990	0.6	0.6	0.6
Syrian Arab Republic	1994	2.0	1.8	2.3
Turkey	1994	1.3	1.3	1.4
United Kingdom	1996	0.5	0.5	0.4
United States	1997	0.5	0.5	0.5
US Virgin Islands	1990	0.6 n
Uruguay	1996	1.0	1.0	1.0

Source: Statistics Division of the United Nations Secretariat and United Nations Centre for Human Settlements (Habitat)

- *— Excluding bathrooms and toilet rooms.
- ...— Not available.
- a— La Paz.
- b— Data refer to Government controlled areas only.
- c— Data do not include the former German Democratic Republic.
- d— Based on households in housing units.
- e— Based on the number of occupied housing units with people present at the time of the census.
- f— Excluding Assam and some areas where 1981 census could not be conducted. Also, not included are institutional households for which the number of living rooms in occupation of household was not canvassed.
- g— Including the Indian-held part of Jammu and Kashmir, the final status of which has not yet been determined.
- h— Kitchens not counted as rooms.
- i— Based on occupied conventional and unconventional dwellings and occupied collective living quarters.
- j— Ciudad de Mexico.
- k— Based on conventional dwellings only.
- l— Excluding Jammu and Kashmir, the final status of which has not yet been determined.
- m— The definition of room does not cover kitchen even if it suits the definition of a room.
- n— Island of St. Thomas. Total number of rooms was estimated by the UN.

Sources:

United Nations Secretariat and United Nations Centre for Human Settlements (Habitat), Compendium of Human Settlement Statistics 2001 (United Nations publication, Sales No. E.01.XVII.5), Compendium of Human Settlement Statistics 1995 (United Nations publication, Sales No. E.95.XVII.11) and United Nations, Compendium of Human Settlements Statistics 1983 (United Nations publication, Sales No. E/F.84.XVII.5).

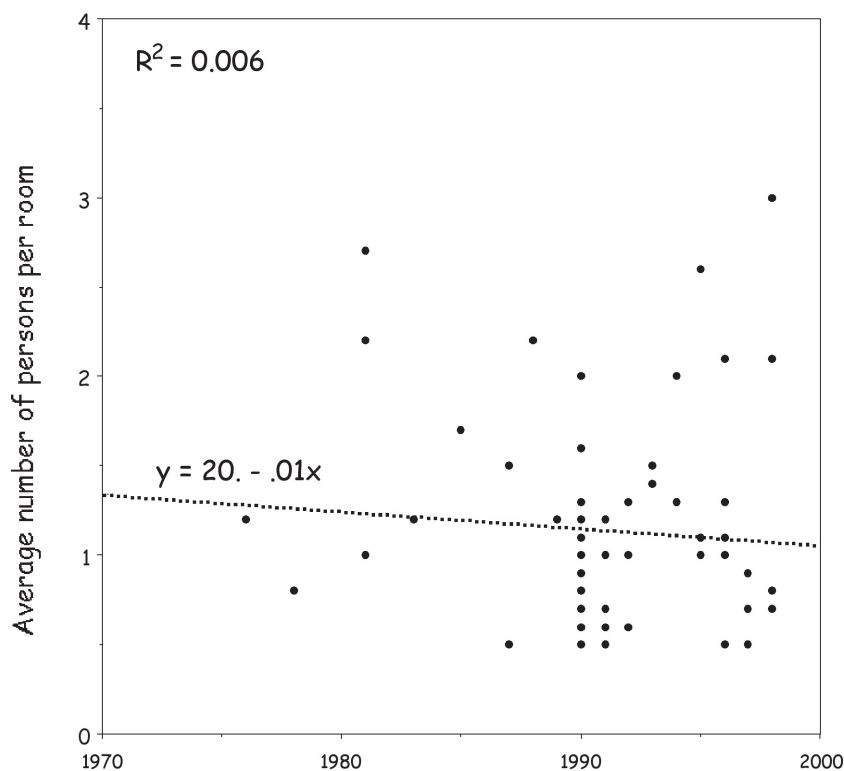


Figure 1. There is essentially no relationship between housing crowdedness and the year the data were collected.

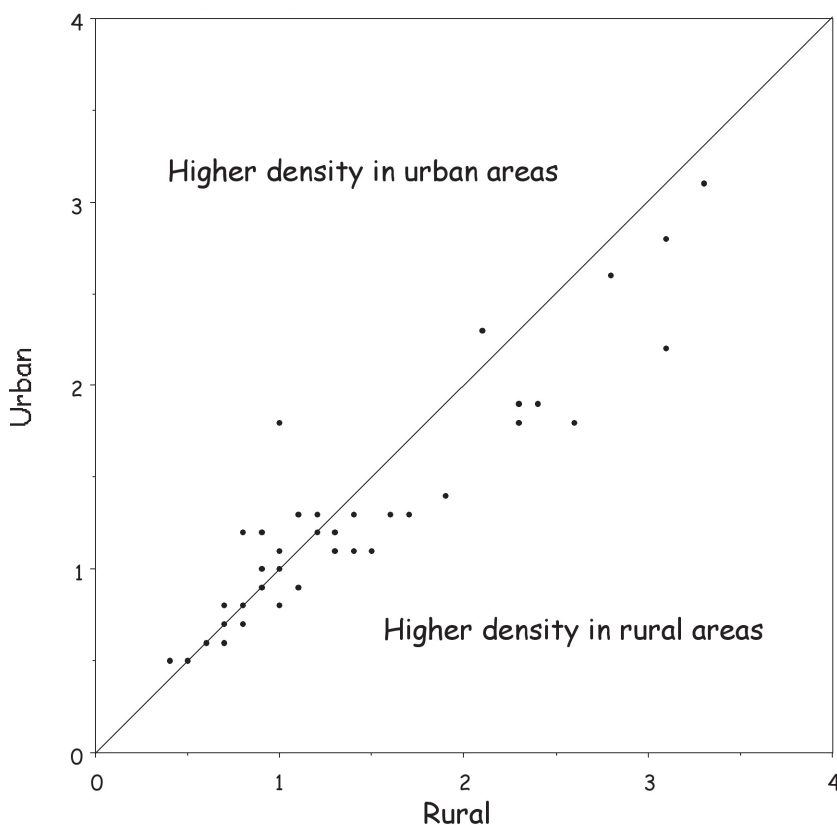


Figure 2. Although there is a high correlation between urban and rural housing density, there is a tendency for more people per room in rural areas than urban.

10 years old, this would have resulted in trimming off 11 countries. The second option was more empirical; it examined a plot of crowdedness against the year the data were gathered. The notion was that, if there was a substantial overall trend (either increasing or decreasing), we could consider it as evidence that the year mattered and we would indeed be erring if we considered all the data together. This, of course, assumes that when the U.N. chose to gather data, that choice was not dependent on how crowded that country was thought to be. The investigatory plot (Figure 1) indicated no such trend.

Once this evidence was in hand, the next question was whether each country was well characterized by its total crowdedness, or was it important to preserve, at least initially, the rural and urban breakdown. It seemed plausible to believe that there is likely to be more crowdedness in urban areas than in rural ones, and hence countries that are less developed should have less crowdedness. Happily, this too was subject to empirical verification. So the student made a plot (Figure 2) comparing each country's rural crowdedness with its urban crowdedness; if they were highly related, we would do no harm characterizing a country by its total figure. What we discovered was that, although rural and urban crowding was, in fact, highly related ($r = 0.94$), there was a tendency for rural areas to be more crowded than urban ones. This was an important clue for subsequent investigations.

Once these preliminaries were out of the way the student felt justified in devising an alternative display showing the total crowdedness of all the countries, knowing that the structure observed was unlikely to misrepresent any country. The goal of the display would be to show the distribution of crowdedness as well as provide a tentative grouping of countries on this important variable. The resulting display (Figure 3), ordered by crowdedness not the alphabet, provides an evocative view.

By the time the student completed this part of the investigation, the class had moved on from the topic of graphic display to regression, which provided a new opportunity for further analysis. It was natural to look at the order of the countries in Figure 3 and ask if there was some underlying variable that

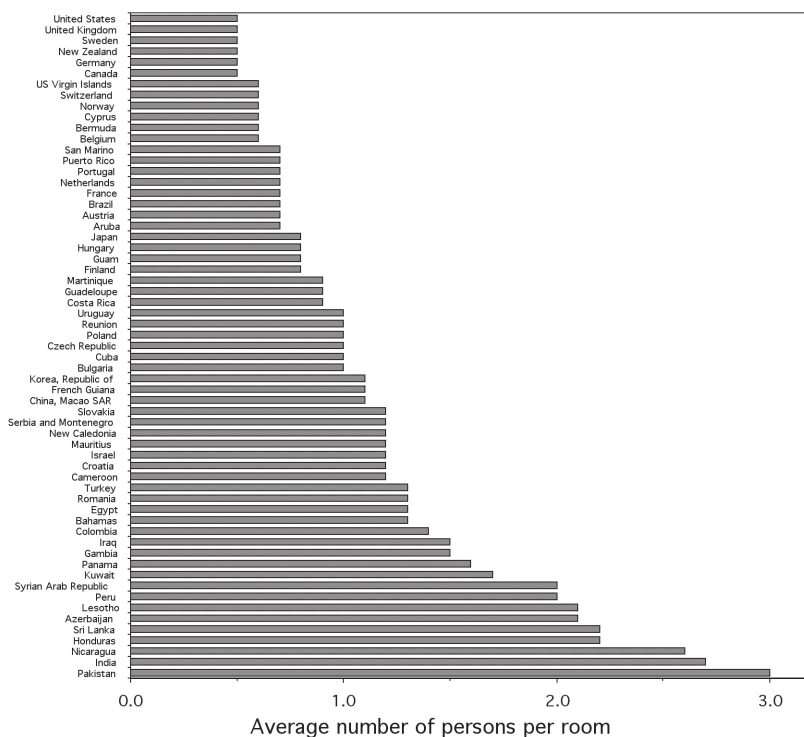


Figure 3. U.N. housing data indicate that the number of people per room varies by a factor of six over the 62 countries surveyed.

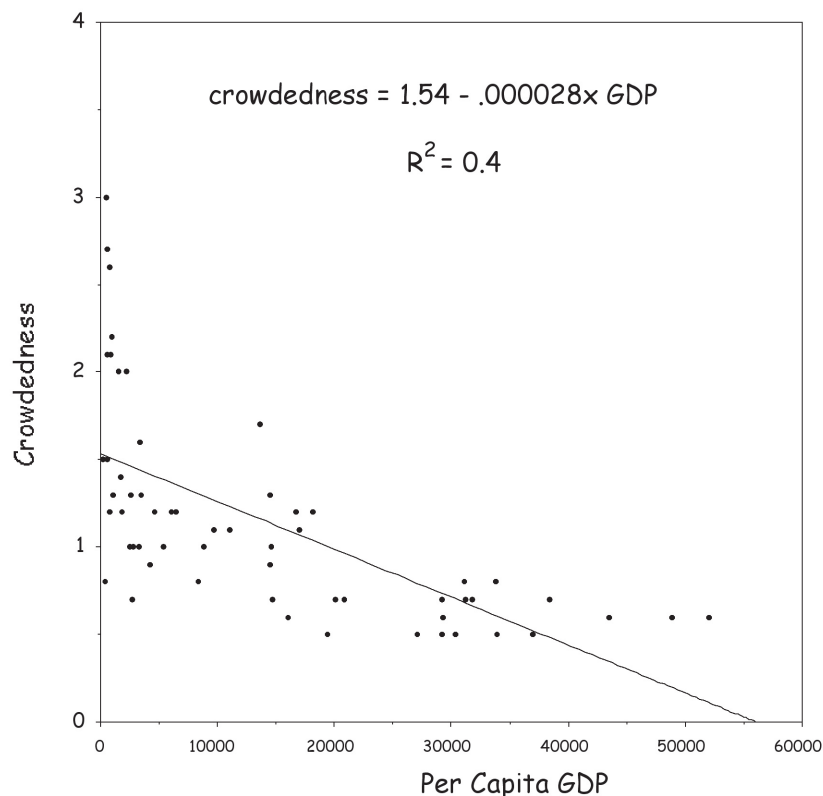


Figure 4. There is a pile-up of data at the low end of GDP.

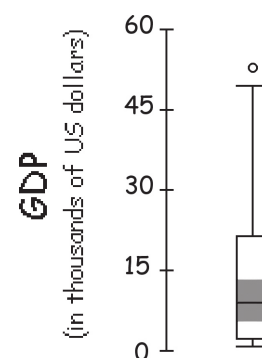


Figure 5. Stem-and-leaf diagram of GDP.

might “explain” why some nations were so much more crowded than others. The results in Figure 2 had allowed us to rule out urban/rural as a major explanation, so what else might it be? Guessing that wealth was a plausible differential explanation, the student searched the same U.N. web site and found a table of per capita gross domestic product (GDP) for the same set of countries. Using that, the student created a plot (Figure 4) purporting to show a strong relationship between GDP and crowdedness; as GDP increased, crowdedness decreased.

At about the same time that this aspect of the investigation was completed, the class was learning about threats to the validity of regression and how re-expression of variables can help. The accumulation of data points at the low end of GDP “screamed” for re-expression. A stem-and-leaf diagram of GDP (Figure 5), which we should have looked at before doing any bivariate analyses, clearly shows a long, thin tail upwards. A log transformation repairs this (Figure 6), and the student was able to return to the regression to try to understand better the plausible causes of crowdedness.

Plotting crowdedness against log (GDP) (Figure 7) then revealed a non-linear relationship. At the same time that this plot was completed, the class was involved in bivariate re-expression, that is, trying to transform the dependent variable to yield linearity by moving up or down Tukey’s (1977) ladder of transformations. Since the relationship was concave upward, it suggested moving down the ladder of transformations. The resulting plot of square root of

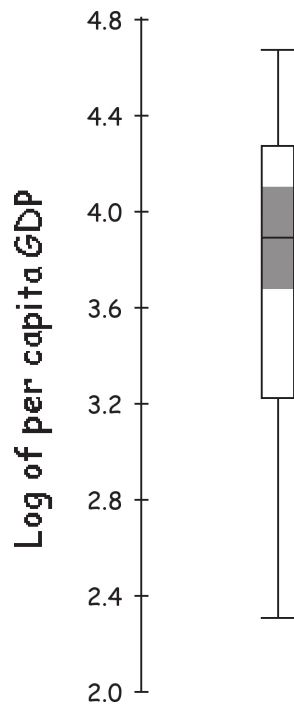


Figure 6. Log transformation of stem-and-leaf diagram of GDP.

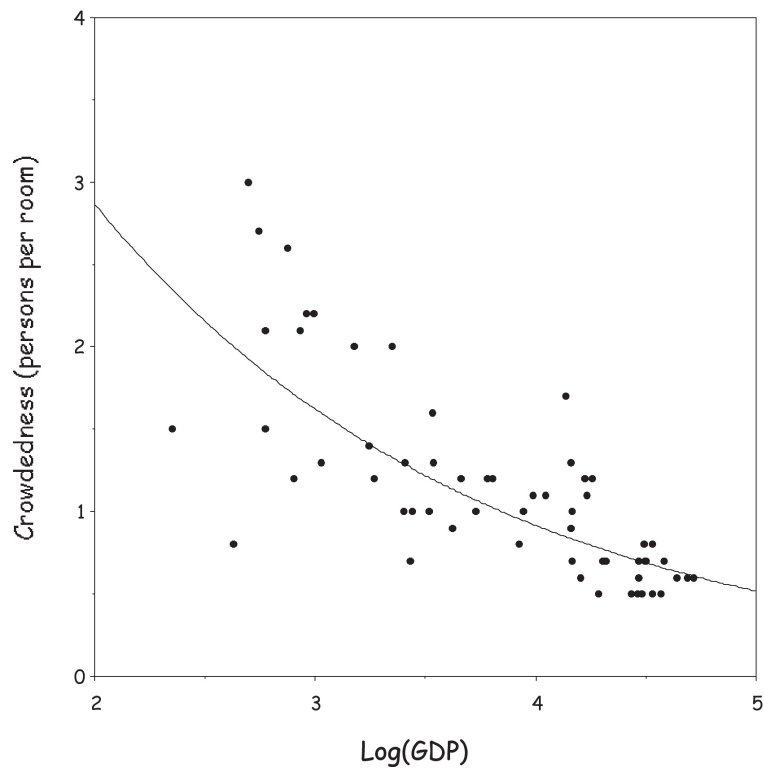


Figure 7. Taking logs of GDP opens up the plot and reveals a nonlinearity.

crowdedness versus log (GDP) (Figure 8) seemed to provide a neat ending to this exercise.

Fortunately, for the student's education, this was not to be the case. The next topic of class discussion was multiple regression, and with it the examination of residuals in the hopes of better understanding and, perhaps, finding other plausible explanatory variables. The student prepared a stem-and-leaf diagram of the residuals, using each country's name as the stem (Figure 9). Both student and instructor pored over this plot looking for plausible explanations. Initially, climate was among the variables considered as an influence on crowdedness. Countries with tropical climates, for example, might appear more crowded per room given that a great deal of living is likely to be outdoors throughout the year, whereas in colder climates people are forced to stay indoors and comfort would require more space. This possibility was quickly dismissed when we observed that tropical countries were seen on both ends of the distribution (Kuwait and India with large positive residuals and Brazil, Gambia,

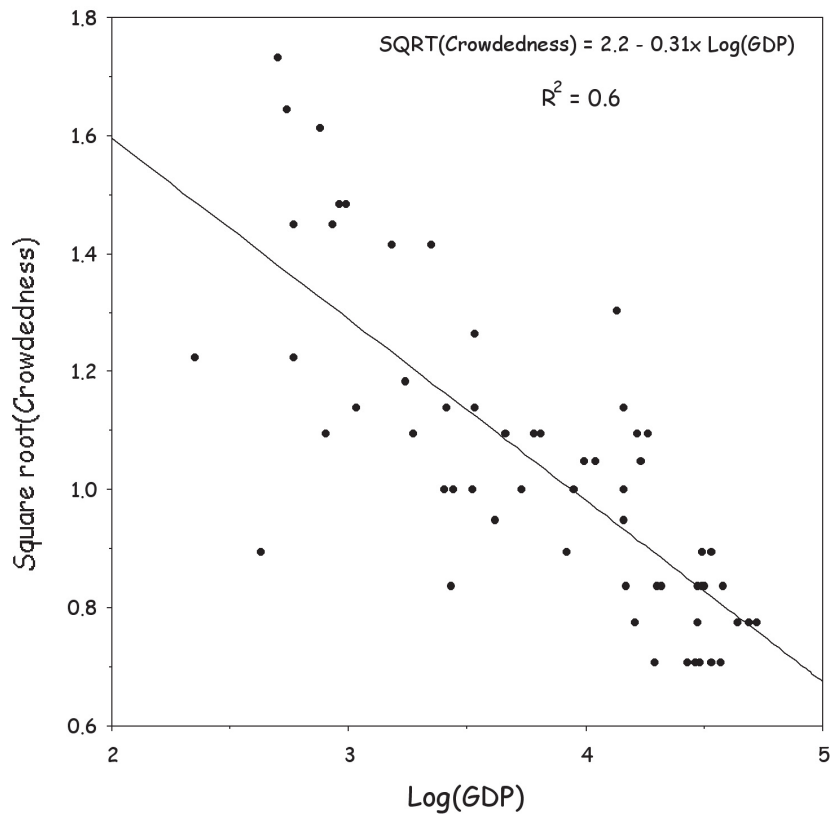


Figure 8. The square root of crowdedness is linear in log(GDP)..

Residuals	Country/Region
0.36	Kuwait, Pakistan
0.34	
0.32	
0.30	
0.28	Nicaragua
0.26	India
0.24	Peru
0.22	Bahamas, Israel, New Caledonia
0.20	Honduras
0.18	Sri Lanka, Syria
0.16	
0.14	Azerbaijan, China, Macao, Panama
0.12	
0.10	Lesotho, Republic of Korea, Japan
0.08	Reunion, French Guiana
0.06	Croatia, Finland, Slovakia
0.04	San Marino, Guadeloupe, Martinique, Turkey, Mauritius, Netherlands, Bermuda
0.02	
0.00	Austria, Czech, France, Norway, Switzerland
-0.02	Romania
-0.04	Columbia, Puerto Rico
-0.06	Aruba
-0.08	Poland
-0.10	Belgium, Portugal,
-0.12	United States
-0.14	Serbia, Montenegro, Sweden, Hungary,
-0.16	United Kingdom
-0.18	Urguay, Germany, Iraq, Egypt
-0.20	Canada, Cyprus, Costa Rica,
-0.22	Bulgaria
-0.24	
-0.26	New Zealand
-0.28	
-0.30	
-0.32	Cameroon
-0.34	
-0.36	Gambia
-0.38	
-0.40	
-0.42	
-0.44	
-0.46	
-0.48	
-0.50	
-0.52	
-0.54	
-0.56	Brazil

Figure 9. Stem-and-leaf diagram with each country as the stem.

and Cameroon with large negative ones). This confirmed that climate was unlikely to be a major explanatory factor.

At this point the student posited that perhaps the structure of the residuals could be, at least partially, explained by the way that women were treated in each society, with countries that limited women's rights being more crowded than would be expected from their GDP. She immediately began seeking indicator variable that might capture this and came up with two candidates: fertility rate and unemployment among women.

From lessons learned earlier, the student first examined the distribution of each variable to see if some re-expression was necessary. Figure 10 shows three box plots for fertility and two re-expressions of it (square root and log). It was clear that a log transform was the most sensible, and the same was true for unemployment.

With the end of the trail seemingly in sight a scatter plot was drawn showing the residuals against log (fertility) (Figure 11). But, as luck would have it, the topic under discussion in class at this time was the sensitivity of least squares regression to influential outliers, and various robust alternatives to least squares regression were illustrated. So, when the plot was made the least squares line (shown as the solid line in the plot) was clearly affected by a few influential points and a more robust alternative (the dashed line) was substituted as a better representative of the bivariate structure.

We can summarize progress up until this point with an equation and a graph. The equation – square root (crowdedness) = $1.85 - .24 \log(\text{GDP}) + .38 \log(\text{fertility})$ – can be interpreted qualitatively in the not very surprising way, that is, that a country's housing becomes more crowded as the wealth of the country decreases and the fertility rate of its women increases. A helpful graph is a plot of the residuals from this model (Figure 12), which indicates that Mexico is less crowded than this model would suggest (as are, to a lesser extent, Guinea, Gambia, Brazil, and Cameroon), whereas Kuwait, India, and Pakistan are more crowded than the model predicts. Further investigations showed that these residuals are shrunk still further when log (unemployment among women) is added to the model, but the lessons learned from this aspect of the

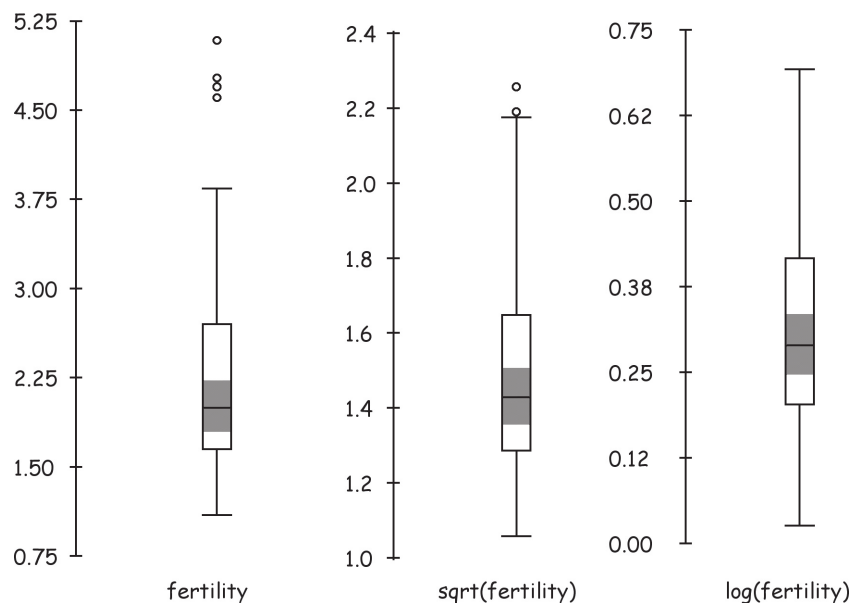


Figure 10. Three box plots for fertility and two re-expressions of it.

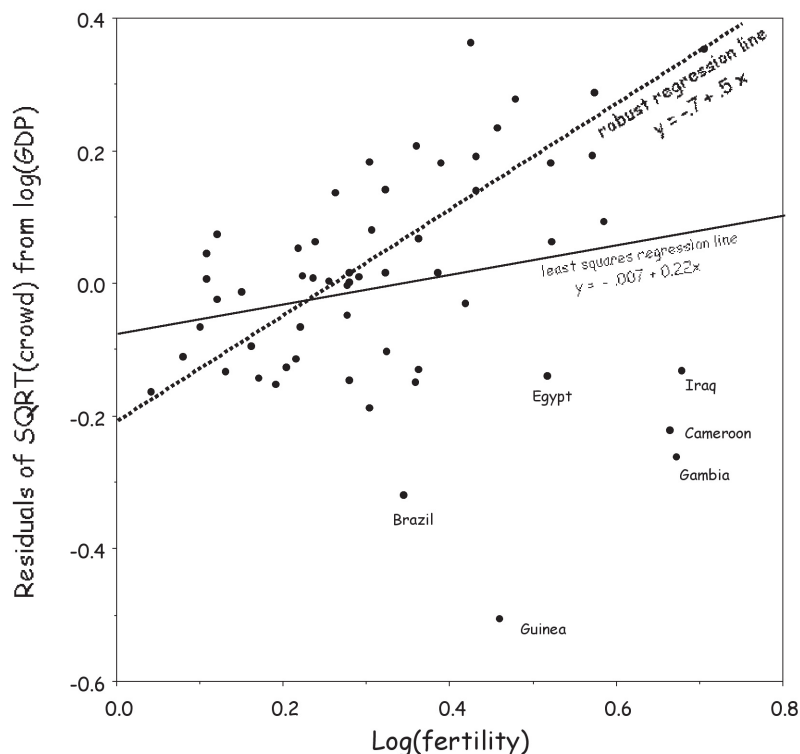


Figure 11. Four highly influential points make the least-squared regression undesirable, and a more robust version is preferred.

analysis do not justify the space that its inclusion would require.

This narrative shows how a small assignment can grow and thus be used

to illustrate a broad range of data analytic methods as well as provide an example of how statistical thinking helps us to understand our environment. The par-

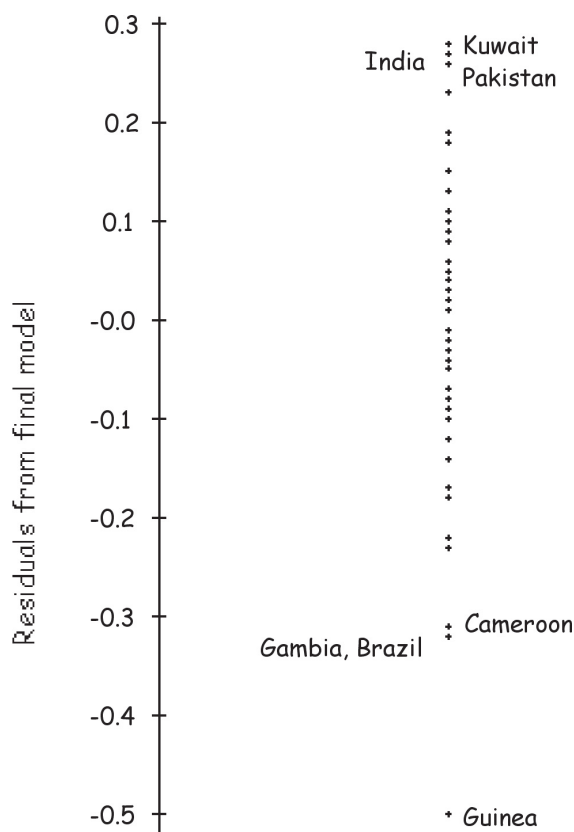


Figure 12. Final plot of the residuals.

particulars of the initial dataset are of little import, so long as it has the potential for the requisite complexity. But even this is a minor issue, because experience has shown that the complexity of the natural world always seems to intrude on all data collections; there will always be an opportunity to re-express data, and puzzling over what covariates to gather is a game that everyone can play. Of course, using complex, adaptive, individualized assignments like this are more time consuming for the instructor, but the potential gains are very great indeed. It also illustrates the omnipresent, existential issue, “How much is enough?” We ended this investigation with three independent variables (only two of which we reported), but we could have gone further. When are you done? Doing term-length exercises like this one helps to teach students the important lesson: You are never done, they just take it away.

Obesa cantavit ☐

References

Tukey, J. W. 1977. Exploratory data analysis. Reading, Mass.: Addison-Wesley.

Comment

Sanford Weisberg

Statistics courses have a reputation for being dull. The article by Vasilescu and Wainer shows that this need not be so. An antidote for a dull course is getting students, on their own, to figure out that the material covered is interesting. Most statistics courses stress the *how*—the methodology needed to convert a set of numbers to useful ideas, conclusions, and summaries. We often fail to get to the *answer*—application of the *how* to problems people care about. Some students get excited by the *how*, but many more will be engaged by the *answer*, which leads inevitably to learning more about the *how*. I do not know what Vasilescu thinks about the *how*, but it is clear that the *answer* was a motivation for her work.

Wainer kept the *how* to an intuitive level, using graphs and summary statistics applied to the graphs to find both intermediate and final answers. Intermediate answers lead to new questions, which is, I think, the paradigm that drives all

of scientific inquiry. At the end of their article, there are a few answers, but many questions could lead in new directions. So, a good course would be one that allows students either to ask or answer interesting questions or, even better, do both.

The course Wainer was teaching was progressing from topic to topic, and he and Vasilescu would apply the topics to the questions posed by the previous answers, reminiscent of “just in time” manufacturing so popular in industry. They used simple graphical ideas suggested by Tukey’s (1977) book, which defined exploratory data analysis.

Tukey did his graphical magic, at least metaphorically, on the back of an envelope, but data analysis today is inevitably connected with computers. Different tools can lead to looking at data in different ways. The remainder of this comment uses different graphs to look at the regression part of Wainer’s article. Expanded presentations of this approach are outlined in Cook and Weisberg (1999) and Weisberg (2005a). The graphs in this article were drawn using *Arc*, a regression computer package described in the first of these two references and also discussed in Weisberg (2005b).

I will take as given the goal of understanding how the average number of persons per room (crowdedness) varies as the predictors GDP (the per-person gross domestic product)

and fertility (the fertility rate) change. Wainer kindly provided me with their data on 57 localities, mostly U.N. member countries, but a few other localities like Macao, Bermuda, Aruba, French Guiana, Guadeloupe, and Reunion are also included.

A good place to start looking for dependence is with a scatterplot matrix like the one shown in Figure 1. This scatterplot matrix consists of two parts, an array of scatterplots at the right and a list of plot controls at the left. The array of scatterplots at the right includes simple scatterplots of each variable versus each other variable. The labels for the plots are given along the diagonal. For example, since the label in the upper-right corner of the array is crowdedness, the variable crowdedness appears on the vertical axis for all plots in the top row and on the horizontal axis for all plots in the last column. The two numbers shown with the label, 0.5 and 3, give the range for this variable, that is, between 0.5 and 3 persons per room. Similarly, the range for GDP is \$224 to \$51,991 per person per year. Similarly, the range for fertility is 1.1 to 5.08.

The scatterplot matrix is like a visual version of a correlation matrix. Whereas a correlation matrix summarizes each pairwise relationship by one number, the scatterplot matrix summarizes via a graph. Graphs have much more information than do single numbers. In particular, the correlation coefficient is only a useful summary when the scatterplot is dominated by a linear relationship with scatter. We cannot tell by looking at the correlation if it is a good summary or not, but we can tell by looking at the scatterplots. Most of the plots in Figure 1 are clearly not dominated by straight lines with scatter, and so correlations will not be relevant here, at least without transformation.

Scatterplot matrices can also be connected to linear regression problems like the one under consideration here. The plot in the upper-left corner has the response crowdedness on the vertical axis and the predictor fertility on the horizontal axis. This is the appropriate summary graph for learning about the regression with crowdedness as the response and fertility as

the only predictor, ignoring GDP. The plot in the lower-right corner reverses the role of these variables and is appropriate for a different regression problem, with fertility as the response and crowdedness as the predictor. The top row says something about the marginal regressions of crowdedness on each of the two predictors separately, but these graphs does not say much about the joint regression of crowdedness on both predictors. Are there conditions under which the plots in the scatterplot matrix can tell us about the multiple regression problem?

Theory can come to the rescue here. If the plots of the predictors are dominated by straight lines with scatter, then straight lines plus scatter in the plots of the response versus each of the predictors is indicative of an appropriate multiple regression model. This ideal case is clearly not appropriate for Figure 1, which is nothing like a set of straight lines with scatter. First, the two predictors (fertility and GDP) appear to be related, but the graphs are strongly curved. Localities where fertility is very high have very low GDP, but areas with low fertility can apparently have either high or low GDP.

The utility of a multiple linear regression model applied to these data is doubtful. However, by transforming scales we may be able to get straight, or at least nearly straight, lines. This can change a complicated problem with nonlinear relationships into a simpler one with straight-line relationships.

Once a plot like Figure 1 gets printed, it becomes an example of a “dead” graph, although the less evocative term “static” graph might be more appropriate. The best graphs for analysis invite interaction from the user and are viewed on the computer screen, not on the printed page. Interactive plots should have intuitive controls. Clicking on a point, for example, should tell about the point, perhaps giving labels, case numbers, values for that point for the plotted variables or other variables, or highlight corresponding points for that case in other graphs. The plot controls at the left of the graph provide more tools for interaction. The palettes at the top of the column of plot controls allow the viewer to color and mark points in all the frames of the scatterplot matrix. The sliders

allow the user to replace the plotted variables by transformations of them. So, for example, moving the slider for GDP to the value 0.5 will replace the variable GDP by the square root of GDP in all graphs in which GDP appears; the value of zero is interpreted as a log transformation. We can quickly try Tukey’s ladder of transformations (or use the similar advice in Cook and Weisberg 1999) to view many choices for transforming for a straight mean function. Numerical methods to help choose transformations can also be made available, and some of these are accessed with popup menus summoned by clicking on the little triangles in the plot controls.

We used a two-step approach to transform the variables in this problem so that fitting a multiple linear regression model in the transformed scale would be the right approach to summarizing these data: first, transforming the predictors

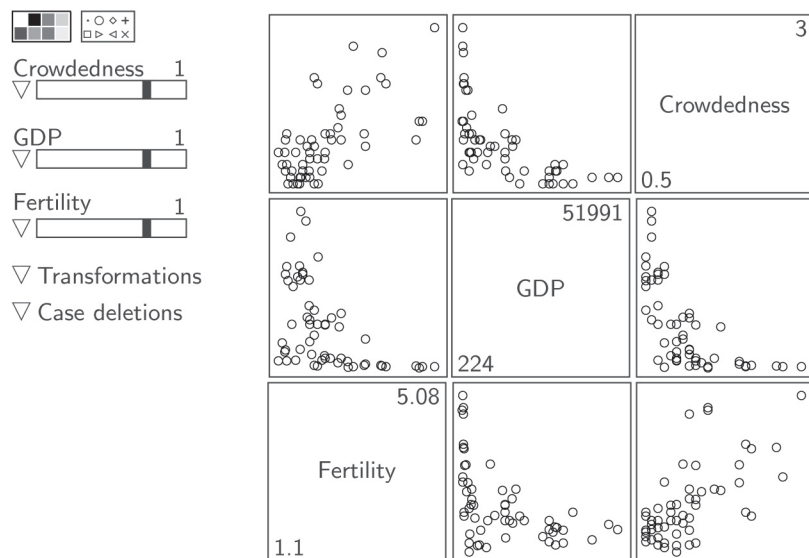


Figure 1. Scatterplot matrix of the three variables in the original scale.

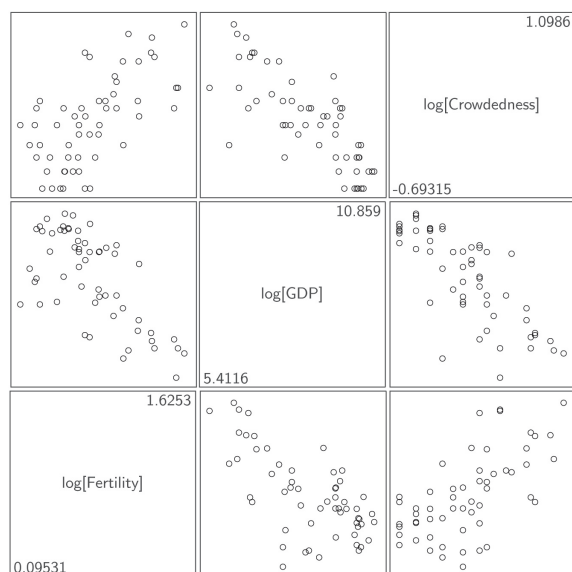


Figure 2. Scatterplot matrix of the three variables in logarithmic scale.

while ignoring the response to make the predictors as close to linearly related as possible; then, transforming the response, so that the regression of the transformed response given the predictors is as close to linear as possible. The first step is different from the approach taken by Vasilescu and Wainer because it explicitly looks at the relationship between the predictors; this approach agrees with the approach outlined by Cook and Weisberg (1999).

Interactive graphics do not translate well to the printed page, and the thrill of changing a complex picture to a simple one using interactive tools requires active participation, not passive viewing. Figure 2 shows the end product with the transformed predictors. The plot controls do not appear in this graph. Although plot controls are desirable, even required, for an analytical graph, where we would like to learn about a problem, they are reduced to chart junk in a presentation graph, which is designed to summarize what we have found. This latter type of graph invites inference—not interaction—on the part of the reader.

Figure 2 suggests that transforming all three variables to a logarithmic scale would greatly simplify the problem of understanding how crowdedness depends on the two predictors. All the plots look more or less like straight lines with scatter. Although not a surprise, we see that $\log(\text{fertility})$ and $\log(\text{GDP})$ are reasonably closely related; at least to a first approximation, both are measuring the same thing. The transformations found by Vasilescu and Wainer are slightly different but would lead to similar conclusions. With everything in log-scale, the fitted model,

$$\log(\text{crowdedness}) = 1.60 - 0.20\log(\text{GDP}) + 0.27\log(\text{fertility}) \quad (1)$$

has the form of a Cobb-Douglas production function and so may have additional interpretability that might appeal to econometricians.

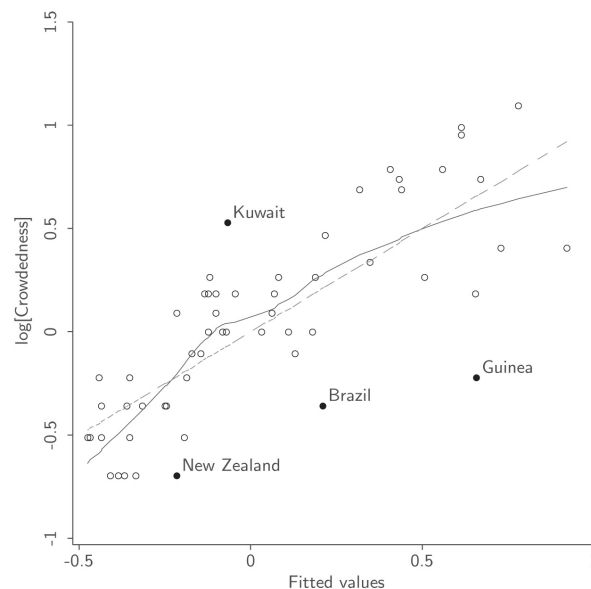



Figure 3. Model checking plot.

In conclusion, a model checking plot (Cook and Weisberg 1999, Chapter 17; Weisberg 2005a, Section 8.4) is designed to let the analyst see how well the data match the fitted model. On the horizontal axis we plotted the fitted values determined by (1); on the vertical axis we plotted the transformed response, $\log(\text{crowdedness})$. If the model matched the data perfectly, then all the points in this graph would fall exactly on a straight line of slope one, represented by a dashed line in Figure 3. The solid line is a smoother fit to the points in the graph, really without any reference to the model. If the model matches the data, then these two lines should match; if the model does not work, then these two lines will not match. The match here is reasonably good, suggesting that the fitted model does indeed match the data with reasonable accuracy. Of course, the fit is not perfect, as represented by the scatter of points about the line in the graph. Kuwait seems to be more crowded than expected according to the model, whereas New Zealand, Brazil, and Guinea are less crowded.

Try It Yourself

You can get a copy of *Arc* for free from www.stat.umn.edu/arc. You can also get a copy of the data used in this article at that site. 

References

- Cook, R. D., and Weisberg, S. 1999. *Applied Regression Including Computing and Graphics*. New York: Wiley.
- Tukey, J. W. 1977. *EDA*. Reading, Mass.: Addison-Wesley.
- Weisberg, S. 2005a. *Applied Linear Regression*, third edition. New York: Wiley.
- Weisberg, S. 2005b. Lost opportunities: Why we need a variety of statistical languages. *Journal of Statistical Software*, www.jstatsoft.org, 13(1).