STAT 400 Data Analysis Nov 3, 2003

Today: Multiway Plots (sec 6.1and 6.2)
Intro to 1-1 Plots of Multivariate Data

Typical data: one quantitative response and several categorical predictor variables.
Regression is a logical possibility but limited use when many categories.
pp 303 –319 concern an "animals" data set

In this example, the quantitative response is "count" data. Count data often requires transformation before plotting or analysis will work well. Log_{10} in this example.

Multipanel plots are useful here. The panels can be determined by either categorical variable. Effectively, the variable used for the panels is the variable we condition upon.

There are two characteristics of the plot on p 302 that should be selected to enhance the readability of the plot: the order of the panels, and the order of the countries within the panels. Both are ordered according to the overall median in this plot. Note the order of the panels is bottom left to top right (increasing median counts). Note also that this ordering counts each animal as 1, no matter what kind. 1 chicken = 1 horse!

What anomalies can be read from this display?

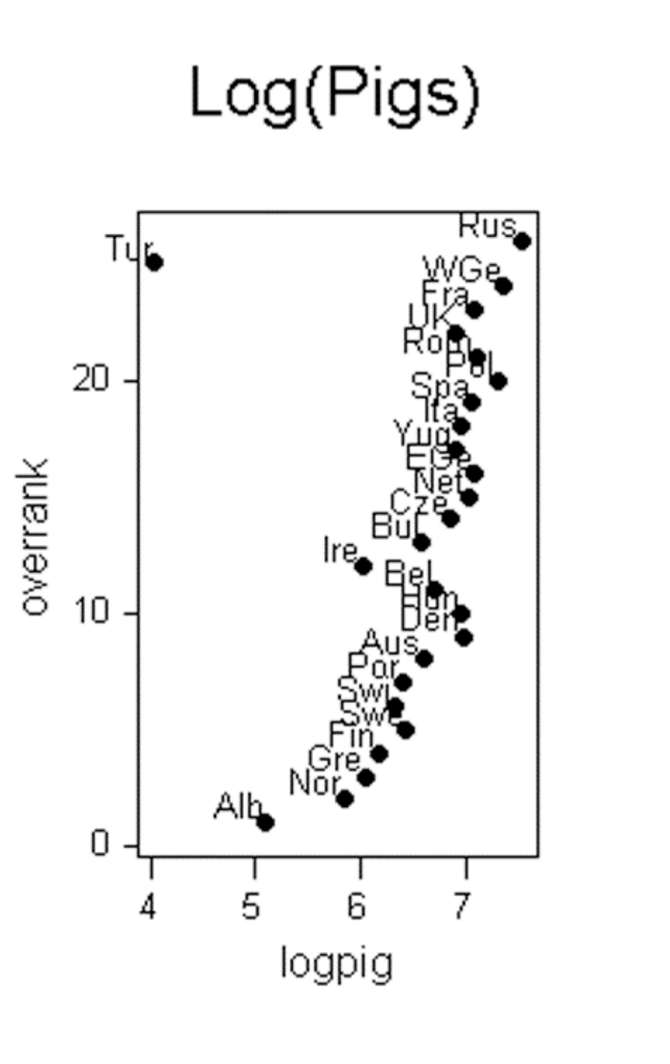
Pigs in Turkey, Ireland, Albania
Chickens in Albania
Horses in Greece and Poland
Sheep in West Germany, Belgium, Denmark, and Finland

The similar graph on p 305 suggests some other anomalies:

Horses in UK, France, Italy

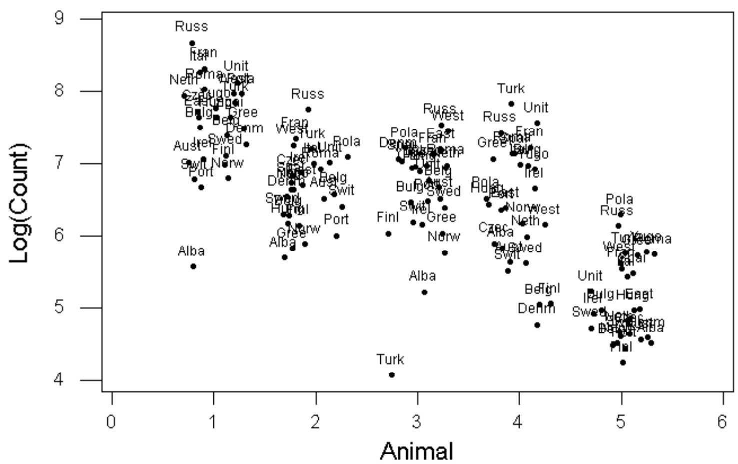
However, a slight improvement might be to find a way to look at pair-wise ratios of animals. But this is getting too complex for our current purposes.

The Splus Trellis graphs are best for this kind of plot. MINITAB can't do the multipanel plots without a lot of work. Here is a MINITAB attempt at a plot which is to accomplish the same information display as the graph on p 302:



Four more of these would be similar, but not so automatic as in

Another attempt (see below) is too messy:



1=Poultry   2=Cattle   3=Pigs   4=Sheep   5=Horses

It is quite easy to get Splus graphs like the ones in the text p 302 or p 305.

Residual Analysis is important but no new technology so we will skip over, except …

Last topic from Cleveland: Level Plot on p 319:

Part of the context of this data is that countries have a geographic location – likely some spatial correlations.   How can data display make use of this feature?

The graph on p 319 shows a special kind of level plot, where the levels are levels of the residual. The focus in this plot is the number of sheep, since this number is the most variable as a proportion of total animals, as suggested by the graph on p 302. Residuals are low, medium, or high depending on whether the number of sheep is low, medium or high relative to the number of animals. The graph invites speculation about why certain countries have more or fewer sheep relative to other farm animals. This kind of data-motivated speculation is often the birthplace of theory.

Next topic: 1-1 plots of multivariate data

We have seen various ways to get around the 2-D limitation of flat displays – coplots, level plots, contour plots, etc. However, these are most effective for small numbers of observations (less than 5 variables, perhaps).   What strategies are available for larger numbers of variables?

Formation of Indexes
Regression Models (usually Linear Models) along with residual plots

Kernel Models (extensions of the techniques we have tried)
Icon plots (1-1 plots are an example)

Consider the animals data again:

Row	C12	cattle	pigs	poultry	horses	sheep
1	Alba	478000	125000	484000	43000	914000
2	Aust	2536000	3920000	11371000	35000	326000
3	Belg	3246000	5108000	29448000	31000	123000
4	Bulg	1796000	3808000	41096000	120000	11271000
5	Czec	5131000	7126000	49212000	44000	990000
6	Denm	2873000	9317000	16296000	41000	58000
7	West	15098000	22478000	83033000	369000	1172000
8	Finl	1719000	1475000	7763000	20000	104000
9	Fran	23599000	11711000	234131008	278000	14346000
10	East	5690000	12107000	51356000	81000	2220000
11	Gree	831000	1107000	29846000	420000	12669000
12	Hung	1922000	9035000	45397000	120000	3183000
13	Irel	6908000	1031000	9903000	68000	3323000
14	Ital	8734000	8928000	137999008	273000	10497000
15	Neth	5241000	10254000	88174000	59000	1210000
16	Norw	971000	687000	5571000	25000	1999000
17	Pola	11912000	19471000	67244000	1734000	3899000
18	Port	1173000	2448000	4143000	23000	2811000
19	Roma	6303000	12464000	109244000	566000	17748000
20	Spai	4495000	11263000	42824000	242000	16543000
21	Swed	1902000	2677000	11393000	57000	437000
22	Swit	1954000	2071000	6188000	45000	336000
23	Turk	16983000	11000	59660000	784000	70093000
24	Unit	13155000	7975000	130018000	140000	32888000
25	Russ	42200000	33100000	385000000	1556000	21000000
26	Yugo	5474000	7867000	65690000	573000	7384000

This looks like 26 rows of 5-variate data.

Suppose we are interested in the relative proportions of these counts. What does "relative" mean here. Not necessarily relative to total animal count …

Try standardized data:

Row	C12	cat	pig	poult	hors	shep
1	Alba	-0.75632	-1.01557	-0.78149	-0.57106	-0.55083
2	Aust	-0.53136	-0.52512	-0.65212	-0.58898	-0.59023
3	Belg	-0.45375	-0.37159	-0.43731	-0.59794	-0.60383

4	Bulg	-0.61225	-0.53959	-0.29890	-0.39860	0.14301
5	Czec	-0.24770	-0.11079	-0.20245	-0.56882	-0.54574
6	Denm	-0.49453	0.17237	-0.59359	-0.57554	-0.60818
7	West	0.84179	1.87326	0.19944	0.15911	-0.53355
8	Finl	-0.62067	-0.84110	-0.69499	-0.62257	-0.60510
9	Fran	1.77103	0.48177	1.99493	-0.04471	0.34901
10	East	-0.18660	0.53294	-0.17698	-0.48595	-0.46334
11	Gree	-0.71774	-0.88866	-0.43258	0.27334	0.23666
12	Hung	-0.59848	0.13593	-0.24779	-0.39860	-0.39883
13	Irel	-0.05346	-0.89848	-0.66956	-0.51506	-0.38945
14	Ital	0.14614	0.12210	0.85260	-0.05591	0.09116
15	Neth	-0.23568	0.29347	0.26053	-0.53522	-0.53100
16	Norw	-0.70243	-0.94294	-0.72104	-0.61138	-0.47815
17	Pola	0.49353	1.48464	0.01182	3.21642	-0.35086
18	Port	-0.68035	-0.71536	-0.73801	-0.61586	-0.42375
19	Roma	-0.11959	0.57908	0.51090	0.60035	0.57692
20	Spai	-0.31723	0.42387	-0.27836	-0.12534	0.49619
21	Swed	-0.60067	-0.68576	-0.65186	-0.53970	-0.58279
22	Swit	-0.59498	-0.76408	-0.71371	-0.56658	-0.58956
23	Turk	1.04784	-1.03031	-0.07830	1.08862	4.08365
24	Unit	0.62940	-0.00106	0.75776	-0.35380	1.59119
25	Russ	3.80431	3.24601	3.78771	2.81774	0.79478
26	Yugo	-0.21021	-0.01502	-0.00665	0.61603	-0.11739

Is this a proper data set to analyze further with graphics?

What about human population?

Nevertheless, we can plot these data is several ways:

Profile plots
Stars
Chernoff Faces

Next topic – the bootstrap.