22S:166
Computing in Statistics

Data validation and description
Proc format

Lecture 18
Nov. 9, 2007

Kate Cowles
374 SH, 335-0727
kcowles@stat.uiowa.edu

Data checking and screening

- important to do prior to any multivariate analyses
- purpose: to identify incorrect, invalid, or otherwise suspect data
- begin with simple descriptive statistics and plots for each variable
- types of checks for binary, nominal, and ordinal data
  - frequency and proportion of invalid categories
  - frequency and proportion of missing classifications
  - adequate representation of categories of interest?

- types of checks for continuous data
  - range screens
  - consistency screens
  - accuracy of measurement

Primary question for the applied statistician:
Does this make sense?

Example: the Berkeley Guidance Study

```
options linesize = 70 pagesize = 60 nodate nonumber ;

data berkboy ;
infile '/group/ftp/pub/kcowles/datasets/berkboy.dat' ;
input wt2 ht2 wt9 ht9 lg9 st9 wt18 ht18 lg18 st18 soma ;
run ;

proc corr ;
run ;

proc reg data = berkboy ;
model soma = ht2 wt2 ht9 wt9 st9 ;
run ;

proc reg data = berkboy ;
model soma = ht9 wt9 st9 ;
run ;
```

The CORR Procedure

11 Variables: wt2 ht2 wt9 ht9 lg9 st9
wt18 ht18 lg18 st18 soma

Simple Statistics

Variable	N	Mean	Std Dev	Sum
wt2	26	214.53846	8.37726	5578
ht2	26	13.59231	1.61862	353.40000
wt9	26	88.40000	3.03592	2298
ht9	26	31.58462	4.35850	821.20000
lg9	26	136.54615	5.31603	3550
st9	26	27.53077	1.89626	715.80000
wt18	26	71.30769	10.69119	1854
ht18	26	71.58077	11.56509	1861
lg18	26	180.03846	6.39619	4681
st18	26	36.33846	2.72882	944.80000
soma	26	210.42308	25.26210	5471

Simple Statistics

Variable	Minimum	Maximum
wt2	201.00000	228.00000
ht2	11.30000	17.20000
wt9	81.30000	92.20000
ht9	24.50000	43.10000
lg9	125.40000	146.00000
st9	24.20000	32.40000
wt18	45.00000	98.00000
ht18	50.30000	110.20000
lg18	169.40000	195.10000
st18	31.00000	44.10000
soma	152.00000	252.00000

Pearson Correlation Coefficients, N = 26
Prob > |r| under H0: Rho=0

	wt2	ht2	wt9	ht9	lg9	st9
wt2	1.00000	0.09354 0.6495	0.28546 0.1575	-0.07875 0.7022	0.08547 0.6781	-0.07209 0.7264
ht2	0.09354 0.6495	1.00000	0.49768 0.0097	0.57922 0.0019	0.38230 0.0539	0.58066 0.0019
wt9	0.28546 0.1575	0.49768 0.0097	1.00000	0.53122 0.0052	0.77583 <.0001	0.28446 0.1590
ht9	-0.07875 0.7022	0.57922 0.0019	0.53122 0.0052	1.00000	0.62049 0.0007	0.90553 <.0001
lg9	0.08547 0.6781	0.38230 0.0539	0.77583 <.0001	0.62049 0.0007	1.00000	0.35332 0.0766

.
.
.

The REG Procedure
Model: MODEL1
Dependent Variable: soma

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	2295.78296	459.15659	0.67	0.6491
Error	20	13659	682.92816		
Corrected Total	25	15954			

Root MSE	26.13289	R-Square	0.1439
Dependent Mean	210.42308	Adj R-Sq	-0.0701
Coeff Var	12.41921		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	51.66778	293.72149	0.18	0.8621
ht2	1	0.42206	4.47469	0.09	0.9258
wt2	1	-0.22891	0.70601	-0.32	0.7491
ht9	1	0.29498	4.16397	0.07	0.9442
wt9	1	1.20330	3.00337	0.40	0.6929
st9	1	3.13973	8.73447	0.36	0.7230

The REG Procedure
Model: MODEL1
Dependent Variable: soma

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	2216.26058	738.75353	1.18	0.3390
Error	22	13738	624.45844		
Corrected Total	25	15954			

Root MSE	24.98917	R-Square	0.1389
Dependent Mean	210.42308	Adj R-Sq	0.0215
Coeff Var	11.87568		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	33.23337	259.54880	0.13	0.8993
ht9	1	0.68933	3.65283	0.19	0.8520
wt9	1	0.90587	2.32088	0.39	0.7001
st9	1	2.73651	7.41980	0.37	0.7158

- SAS procedures for describing binary, ordinal, and nominal data
  - proc freq
  - proc chart (or gchart)
  - proc tabulate
- SAS procedures for describing quantitative data
  - proc means
  - proc univariate
    - most thorough description
    - also does one-sample t-tests
  - proc tabulate

Example: AIDS Clinical Trials Group (ACTG) Protocol 320

- randomized, double-blind, placebo-controlled clinical trial
- eligibility criteria
  - HIV-infected adults
  - CD4 counts <= 200 and at least 3 months of prior zidovudine therapy
- two treatment groups
  - 3-drug regimen: indinavir, lamivudine, and either zidovudine or stavudine
  - 2-drug regimen: zidovudine and lamivudine
- 1156 patients randomized

- patients stratified according to their CD4 count at study entry
  - ≤ 50 cells/mm^3
  - 50-200 cells/mm^3
- primary endpoint: occurrence of an AIDS-defining event or death