STOR 455
STATISTICAL METHODS I

Jan Hannig

Redo interactions in homework

- Import the body fat data set, Ch07ta01.txt in the extra data sets. Include all pair-wise interactions among three explanatory variables Use F-test to see if the interaction terms are significant. Also produce the variance inflation factors for the model with interactions.
- See the solutions!

Project

- I have decided to assign a project. It will be part of the midterm exam score (10% of the final grade).
- The two in-class midterm score make 40% of the grade.
- Due December 2.
- There are three files (info, more info and data)

Remedial Measures

- Discard outlier
  - or alternatively use robust procedure such as weighted least squares, Generalized linear models, nonparametric methods, …
- Transformation data to
  - Linearize mean response
  - Stabilize variance/Achieve normality

Box-Cox Transformations

- Also called power transformations

$$Y' = Y^{\lambda} \quad \text{or} \quad Y' = (Y^{\lambda} - 1)/\lambda$$

  - In the second form, the limit as λ approaches zero is the (natural) log

Important Special Cases

- λ = 1, Y' = Y^1, no transformation
- λ = .5, Y' = $Y^{1/2}$, square root
- λ = -.5, Y' = $Y^{-1/2}$, one over square root
- λ = -1, Y' = Y^{-1} = 1/Y, inverse
- λ = 0, (Y' = $(Y^{\lambda} - 1)/\lambda$), log is the limit

Box-Cox Details

- We can estimate λ by including it as a parameter in a non linear model
- $Y^{\lambda} = \beta_0 + \beta_1 X + \xi$
- Choose λ that give the best fit
- SAS code is in proc transreg

Plutonium Example

- Detecting plutonium 238 using alpha counts
- X: plutonium activity
- Y: observed alpha counts per second
- Relationship depend on measurement device
- Four standard aluminum/plutonium rods tested, each 4 to 10 times

Do it in SAS

```
proc transreg data=plu;
model boxcox(y)=identity(x);
run;
```

The TRANSREG Procedure

Box-Cox Transformation Information for y

Lambda	R-Square	Log Like
-3.00	0.09	-22.8128
-2.75	0.10	-10.5162
-2.50	0.12	1.5873
-2.25	0.14	13.4530
-2.00	0.17	25.0222
-1.75	0.20	36.2187
-1.50	0.25	46.9457
-1.25	0.31	57.0871
-1.00	0.38	66.5173
-0.75	0.47	75.1246
-0.50	0.57	82.8497
-0.25	0.67	89.7200
0.00	0.76	95.8371
0.25	0.84	101.2188
0.50 +	0.90	105.2423 *
0.75	0.92	105.7181 <
1.00	0.92	100.6558
1.25	0.89	91.9689
1.50	0.84	82.2649
1.75	0.79	72.5284
2.00	0.74	62.9473
2.25	0.69	53.5043
2.50	0.65	44.1498
2.75	0.61	34.8407
3.00	0.58	25.5459

< - Best Lambda

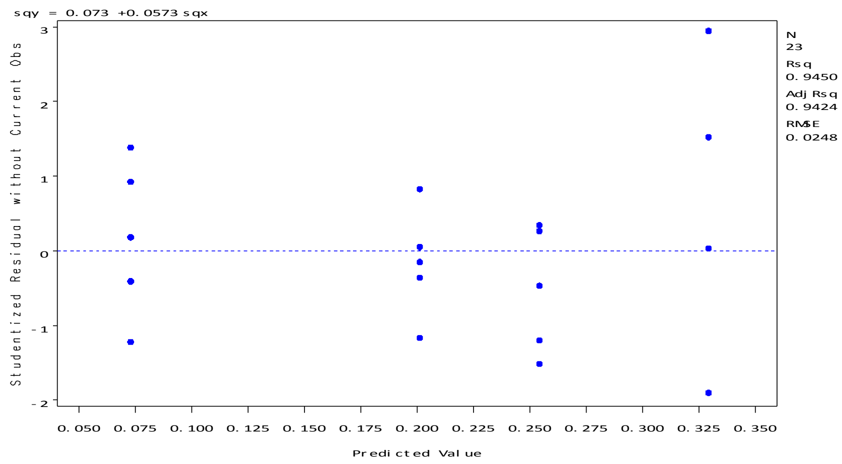
* - 95% Confidence Interval

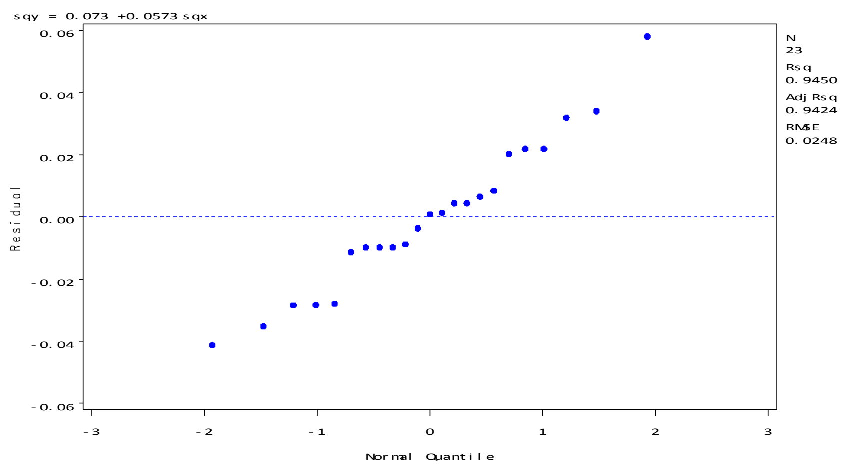
+ - Convenient Lambda

Do it in SAS

```
*Data transformation;
data plu1;
   set plu;
    where NOT ( x EQ 0 AND y GE
   0.09 );
    sqy=sqrt(y);
    sqx=sqrt(x);
    run;
proc print data=plu1;
run;

* transform y and x;
proc reg data=Plu1;
  model sqy = sqx;
  plot sqy*sqx rstudent.*p. r.*nqq.;
run;
```





Do it in SAS

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	0.22142	0.22142	360.92	<.0001
Error	21	0.01288	0.00061348		
Corrected Total	22	0.23430			

Root MSE	0.02477	R-Square	0.9450
Dependent Mean	0.18483	Adj R-Sq	0.9424
Coeff Var	13.40098		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	0.07301	0.00783	9.32	<.0001
sqx	1	0.05731	0.00302	19.00	<.0001

Do it in SAS

- The final model we fit is $\mathrm{sqrt}(y)=0.07301+0.05731*\mathrm{sqrt}(x)+\xi$
- Be careful when interpreting – the predicted values are sqrt(y), to get a predicted value for y need to square!

Steps in Data Analysis

- Plot and investigate the data!
  - %scatter.sas
  - proc univariate, proc means, proc print
- Run some exploratory models, look for outliers
  - Recode categorical variables, add interaction
  - Look for transformations
    residual plots & proc transreg
- Do a model selection
  - proc reg, %allsubsreg.sas

Steps in Data Analysis

- Fit the final model
  - Run tests, compute predictions, answer scientific questions
- Make sure that your results make sense and write a report.

SENIC Data

- Info about 113 hospitals between 1975 and 1976
- Variables: id, length of stay, age, infection risk, routine culturing ratio, routine chest X-ray ratio, number of beds, medical school affiliation, region (1=NE, 2=NC, 3=S, 4=W), average daily census, number of nurses, available facilities
- File AppendixC01.txt in the extra data sets

Recode the variable

```
data hospital1;
set hospital;
region1=0; region2=0;region3=0;
if region=1 then region1=1;
if region=2 then region2=1;
if region=3 then region3=1;
output;
run;
proc print data=hospital1; run;
```

Run various models

- Now the exploration begins. Start with the full model.
- proc GLM might simplify some of the notation

Proc GLMn (only use with caution)

```
ods graphics on;
proc glm data=hospital PLOTS=(DIAGNOSTICS
RESIDUALS);
class school region;
  model risk= stay|age|culture|xray|beds|
school|region|census|nurses|facilities@2;
run;
ods graphics off;
```