## ECMI Class 7

- More diagnostics for residuals
- Remedial measures

## Outliers

- An outlier can be *influential* for the estimation of some model parameters
- And not influential for others
- Outliers usually inflate the variance and decrease our chances of finding statistically significant results

## Other plots

- You can also look at
  - the stem plot and
  - the boxplot
  - that are part of the output file generated by the plot option with proc univariate

## More plots

- Plot residuals vs
  - Time (order)
  - Other candidate explanatory variables
- Look for
  - Overall patterns
  - outliers

## Significance tests for normality?

- Test has high power when the sample size is large
- Test is sensitive to departures from normality that we might not care about

| Test | | |
|---|---|---|
| (Shapiro-Wilk | | W ) |
| Kolmogorov-Smirnov | D | |
| Cramer-von Mises | W-Sq | |
| Anderson-Darling | A-Sq | |

| statistic | -----p Value------ | |
|---|---|---|
| (0.978 | Pr < W | 0.8626) |
| 0.095 | Pr > D | >0.1500 |
| 0.033 | Pr > W-Sq | >0.2500 |
| 0.207 | Pr > A-Sq | >0.2500 |

## Comments on plots vs significance tests for model assumptions

- Plots are more likely to lead to a remedy
- Significance tests results are very dependent on the sample size; with sufficiently large samples we can reject most null hypotheses

## Nonlinear relationships

- We can model many nonlinear relationships with linear models, some have several explanatory variables (multiple linear regression)
  - Quadratic $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \xi$
  - $Y = \beta_0 + \beta_1 \log(X) + \xi$

## Nonlinear (2)

- Sometimes we transform a nonlinear problem into a linear form
- eg if $Y = \beta_0 \exp(\beta_1 X) + \xi$
- we could consider the linear model
- $\log(Y) = \log(\beta_0) + \beta_1 X + \xi$
- Note that we have changed our assumption about the error

## Nonlinear (3)

- We can perform a nonlinear regression analysis
- SAS PROC NLIN

## Non constant error variance

- Sometimes we model the way in which the error variance changes (eg it may be linearly related to X)
- We can use a weighted analysis
- Use a weight statement in PROC REG

## Non normal errors

- Transformations often help
- Use a procedure that allows different distributions for the error term
- SAS PROC GENMOD

## GENMOD (1)

- Possible distributions of Y:
- Binomial (binary data)
- Poisson
- Gamma (exponential)
- Inverse gaussian
- Negative binomial
- multinomial

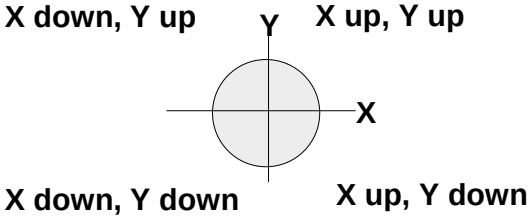
## GENMOD (2)

- Allows you to specify the link function g in the equation
- $g(\mu)=b_0+ b_1x_1+ \ldots+b_kx_k$
- where $\mu=E(Y)$.

## Ladder of Reexpression (transformations)

| p | Transformation is $x^p$ |
|---|---|
| 1.5 | |
| 1.0 | |
| 0.5 | |
| 0.0 | |
| -0.5 | |
| -1.0 | |

## Circle of Transformations




## Box-Cox Transformations

- Also called power transformations
- $Y' = Y^{\lambda}$
- or $Y' = (Y^{\lambda} - 1)/\lambda$
- In the second form, the limit as $\lambda$ approaches zero is the (natural) log

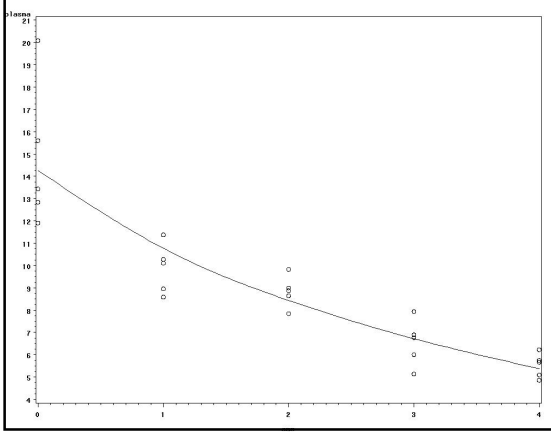
## Important Special Cases

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

## Box-Cox Details

- We can estimate $\lambda$ by including it as a parameter in a non linear model
- $Y^{\lambda} = \beta_0 + \beta_1 X + \xi$
- and using the method of maximum likelihood

```
data a1; input age plasma @@;
cards;
0 13.44 0 12.84 0 11.91 0 20.09 0 15.60
1 10.11 1 11.38 1 10.28 1 8.96 1 8.59
2 9.83 2 9.00 2 8.65 2 7.85 2 8.88
3 7.94 3 6.01 3 5.14 3 6.90 3 6.77
4 4.86 4 5.10 4 5.67 4 5.75 4 6.23
;
```




## PROC TRANSREG

**proc transreg** data=a1;
model Box(plasma)=identity(age) /alpha=0.05;
**run**;

| Lambda | R-Square | Log Like | |
|---|---|---|---|
| -3.00 | 0.71 | -21.8996 | |
| -2.75 | 0.73 | -19.4576 | |
| -2.50 | 0.76 | -17.0444 | |
| -2.25 | 0.78 | -14.6732 | |
| -2.00 | 0.80 | -12.3665 | |
| -1.75 | 0.82 | -10.1608 | |
| -1.50 | 0.83 | -8.1127 | |
| -1.25 | 0.85 | -6.3056 | |
| -1.00 | 0.86 | -4.8523 | * |
| -0.75 | 0.86 | -3.8891 | * |
| -0.50 | 0.87 | -3.5523 | < |
| -0.25 | 0.86 | -3.9399 | * |
| 0.00 | + 0.85 | -5.0754 | * |
| 0.25 | 0.84 | -6.8988 | |
| 0.50 | 0.82 | -9.2925 | |
| 0.75 | 0.79 | -12.1209 | |
| 1.00 | 0.75 | -15.2625 | |
| 1.25 | 0.72 | -18.6233 | |
| 1.50 | 0.67 | -22.1378 | |
| 1.75 | 0.63 | -25.7629 | |
| 2.00 | 0.59 | -29.4720 | |
| 2.25 | 0.54 | -33.2490 | |
| 2.50 | 0.50 | -37.0844 | |
| 2.75 | 0.46 | -40.9728 | |
| 3.00 | 0.42 | -44.9109 | |

< - Best Lambda
* - Confidence Interval
+ - Convenient Lambda

## BOX-COX - Przykład

```
data a2; set a1;
ynew=1/sqrt(plasma);
proc gplot data=a2;
plot ynew*age;
run;
```