STAT 525 FALL 2018

Chapter 3
Diagnostics and Remedial Measures

Professor Min Zhang

Diagnostics

- Procedures to determine appropriateness of the model and check assumptions used in the standard inference
- If there are violations, inference and model may not be reasonable thereby resulting in faulty conclusions
- Always check before any inference!!!!!!!!
- Procedures involve both graphical methods and formal statistical tests

Diagnostics for X

- Scatterplot of Y vs X common diagnostic
  - Fit smooth curve $\longrightarrow$ I=SM## (e.g., I=SM70 in slide 1-5)
  - Is linear trend reasonable?
  - Any unusual/influential (X, Y) observations?
- Can also look at distribution of X alone
  - Skewed distribution
  - Unusual or outlying values?
  - Recall model does **not** state $X \sim$ Normal
  - Does X have pattern over time (order collected)?
- If Y depends on X, looking at Y alone may be deceiving (i.e., mixture of normal dists)

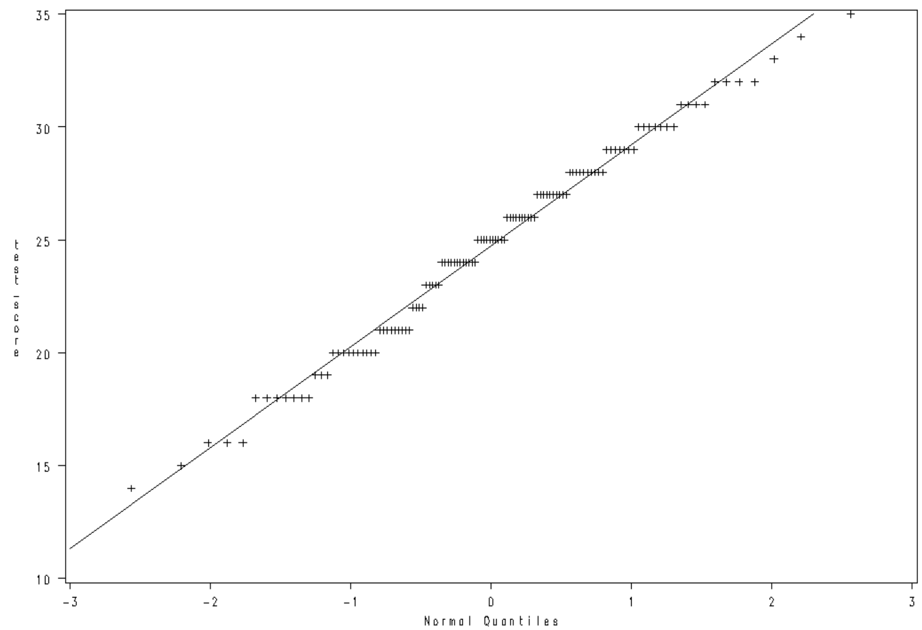
`PROC UNIVARIATE` in SAS

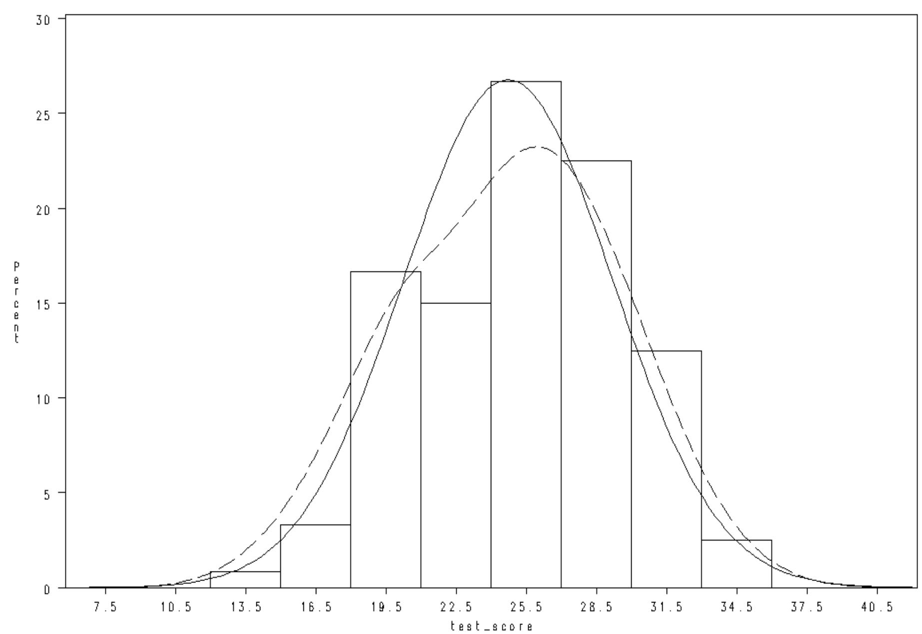
- Provides numerous graphical and numerical summaries
  - Mean, median
  - Variance, std dev, range, IQR
  - Skewness, kurtosis
  - Tests for normality
  - Histograms
  - Box plots
  - QQ plots
  - Stem-and-leaf plots

Example: Grade Point Average

```
options nocenter; /* output layout: not centerized */
goptions colors=(none); /* graphics display: black/white */

data a1;
    infile 'U:\.www\datasets525\CH01PR19.txt';
    input grade_point test_score;

/* Line printer plots: stem-and-leaf, horizontal bar chart
                       box plot, normal probability plot */
/* Graphics display: histogram, probplot, qqplot */
proc univariate data=a1 plot;
    var test_score;
    qqplot test_score / normal (L=1 mu=est sigma=est);
    histogram test_score / kernel(L=2) normal;
run; quit;
```

```
The UNIVARIATE Procedure
Variable:  test_score

                            Moments

N                         120    Sum Weights                 120
Mean                   24.725    Sum Observations           2967
Std Deviation      4.47206549    Variance             19.9993697
Skewness           -0.1363553    Kurtosis             -0.5596968
Uncorrected SS          75739    Corrected SS           2379.925
Coeff Variation    18.0872214    Std Error Mean       0.40824186

              Basic Statistical Measures

    Location                    Variability

Mean     24.72500     Std Deviation            4.47207
Median   25.00000     Variance                19.99937
Mode     24.00000     Range                   21.00000
                      Interquartile Range      7.00000

...
```





Upper – QQ Plot Lower – Histogram

Diagnostics for Residuals

- If model is appropriate, residuals should reflect assumptions on error terms

$$\varepsilon_i \sim \text{i.i.d. } N(0, \sigma^2)$$

- Recall properties of residuals
  - $\sum e_i = 0 \longrightarrow$ Mean is zero
  - $\sum (e_i - \overline{e})^2 = \text{SSE} \longrightarrow$ Variance is MSE
  - e_i's not independent (derived from same fitted regression line)
  - When sample size large, the dependency can basically be ignored

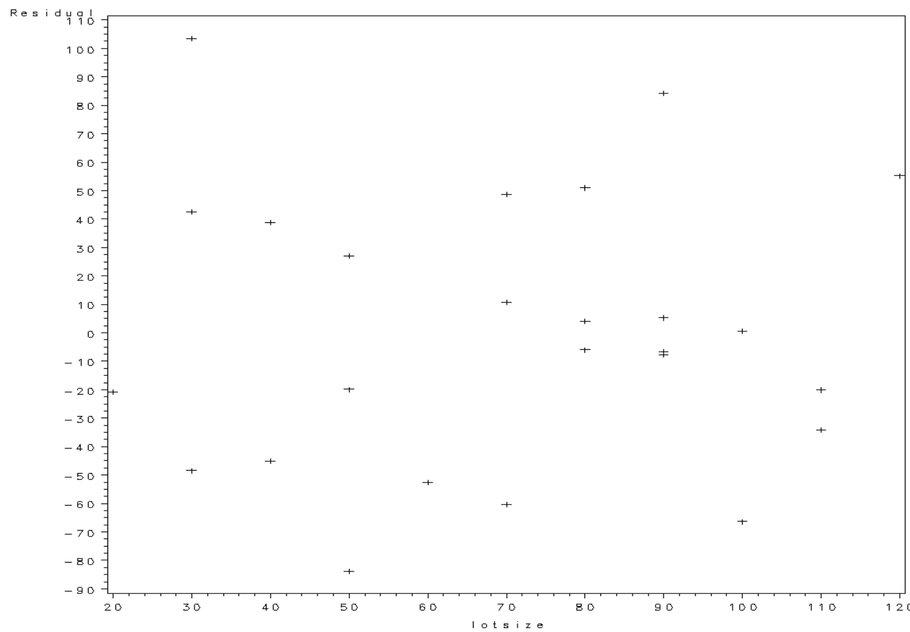
- Questions addressed by diagnostics
  - Is the relationship linear?
  - Does the variance depend on X?
  - Are there outliers?
  - Are error terms not independent?
  - Are the errors normal?
  - Can other predictors be helpful?

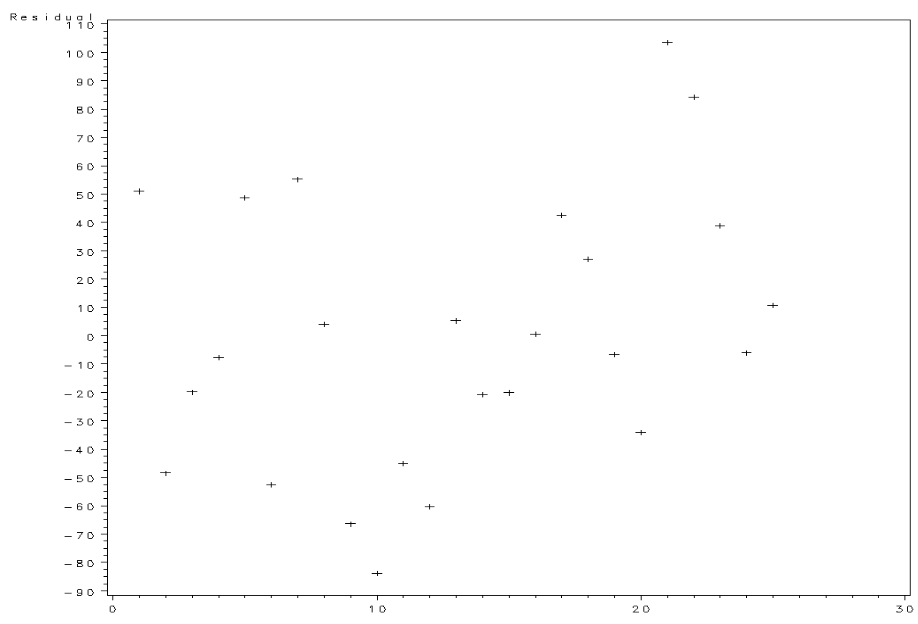
Residual Plots

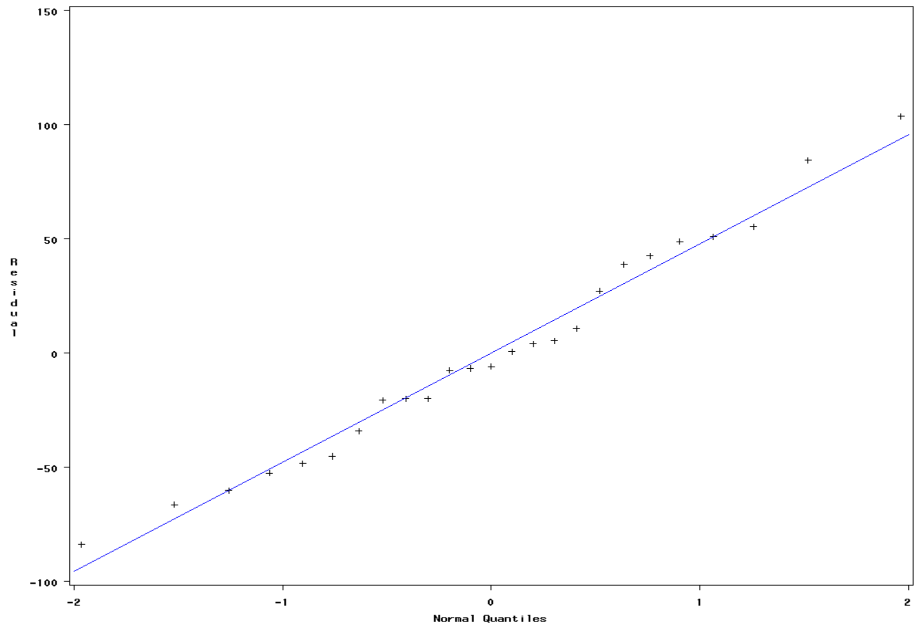
- Plot e vs X can assess most questions
- Get same info from plot of e vs $\widehat{Y}$ because X and $\widehat{Y}$ linearly related
- Other plots include e vs time/order, a histogram or QQplot of e, and e vs other predictor variables
- See pages 102-113 for examples
- Plots are usually enough for identifying gross violations of assumptions (since inferences are quite robust)

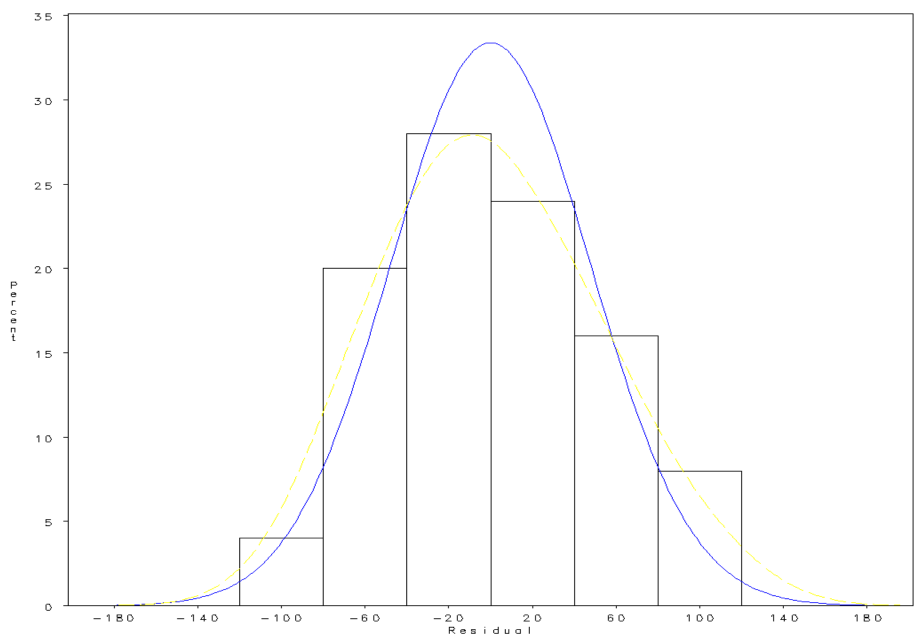
Example: Toluca Campany

```
data a1;
   infile 'U:\.www\datasets525\CH01TA01.txt';
   input lotsize workhrs;
   seq = _n_;

proc reg data=a1;
   model workhrs=lotsize;
   output out=a2 r=resid;

proc gplot data=a2;
   plot resid*lotsize;
   plot resid*seq;
run;

/* Line type: L=1 for solid line; L=2 for dashed line */
proc univariate data=a2 plot normal;
   var resid;
   histogram resid / normal kernel(L=2);
   qqplot resid / normal (L=1 mu=est sigma=est);
run;
```


Residual
lotsize


Residual
seq





Upper – QQ Plot   Lower – Histogram

Tests for Normality

- Test based on the correlation between the residuals and their expected values under normality proposed on page 115
- Requires table of critical values
- SAS provides four normality tests

```
proc univariate normal;
  var resid;
```

- Shapiro-Wilk most commonly used

Example: Plasma Level (p. 132)

```
The UNIVARIATE Procedure
Variable:  resid  (Residual)

                   Tests for Normality
Test                  --Statistic---    -----p Value------

Shapiro-Wilk          W     0.839026    Pr < W      0.0011
Kolmogorov-Smirnov    D     0.167483    Pr > D      0.0703
Cramer-von Mises      W-Sq  0.137723    Pr > W-Sq   0.0335
Anderson-Darling      A-Sq   0.95431    Pr > A-Sq   0.0145
```

Other Formal Tests

- Durbin-Watson test for correlated errors (assuming AR(1) for errors as in Chapter 12)
- Modified Levene / Brown-Forsythe test for constant variance (Chapter 18)
- Breusch-Pagan test for constant variance
- Plots vs Tests

  Plots are more likely to suggest a remedy. Also, test results are very dependent on n. With a large enough sample size, we can reject most null hypotheses even if the deviation is slight

Lack of Fit Test

- More formal approach to fitting a smooth curve through the observations
- Requires repeat observations of Y at one or more levels of X
- Assumes $Y|X \overset{ind}{\sim} N(\mu(X), \sigma^2)$
- $H_0 : \mu(X) = \beta_0 + \beta_1 X$

  $H_a : \mu(X) \neq \beta_0 + \beta_1 X$
- Will use full/reduced model framework

- Notation
  - Define X levels as $X_1, X_2, \ldots, X_c$
  - There are n_j replicates at level X_j ($\sum n_j = n$)
  - Y_{ij} is the i^{th} replicate at X_j
- Full Model: $Y_{ij} = \mu_j + \varepsilon_{ij}$
  - No assumption on association : $E(Y_{ij}) = \mu_j$
  - There are c parameters
  - $\widehat{\mu}_j = \overline{Y}_{.j}$ and $s^2 = \sum\sum(Y_{ij} - \widehat{\mu}_j)^2/(n-c)$
- Reduced Model: $Y_{ij} = \beta_0 + \beta_1 X_j + \varepsilon_{ij}$
  - Linear association
  - There are 2 parameters
  - $s^2 = \sum\sum(Y_{ij} - \widehat{Y}_j)^2/(n-2)$

- SSE(F)=$\sum\sum(Y_{ij} - \widehat{\mu}_j)^2$
- SSE(R)=$\sum\sum(Y_{ij} - \widehat{Y}_j)^2$

$$F^\star = \frac{(\text{SSE(R)} - \text{SSE(F)})/((n-2) - (n-c))}{\text{SSE(F)}/(n-c)}$$

- Is variation about the regression line substantially bigger than variation at specific level of X?
- Approximate test can be done by grouping similar X values together

Example: Plasma Level (p. 132)

```
/* Analysis of Variance - Reduced Model */
proc reg;
    model lplasma=age;
run;
                         Sum of       Mean
Source            DF    Squares     Square   F Value   Pr > F
Model              1    0.52308    0.52308    134.03   <.0001
Error             23    0.08976    0.00390
Corrected Total   24    0.61284
-----------------------------------------------
/* Analysis of Variance - Full Model */
proc glm;
    class age;
    model lplasma=age;
run;
                         Sum of       Mean
Source            DF    Squares     Square   F Value   Pr > F
Model              4    0.53854    0.13463     36.24   <.0001
Error             20    0.07430    0.00372
Corrected Total   24    0.61284
-----------------------------------------------
```

$$F^{\star} = \frac{(.08976 - .07430)/(23 - 20)}{.00372} = 1.387$$

$$\downarrow$$

$$\text{P-value} = 0.2757$$

Remedies

- Nonlinear relationship
  - Transform X or add additional predictors
  - Nonlinear regression
- Nonconstant variance
  - Transform Y
  - Weighted least squares
- Nonnormal errors
  - Transform Y
  - Generalized Linear model
- Nonindependence
  - Allow correlated errors
  - Work with first differences

Nonlinear Relationships

- Can model many nonlinear relationships with linear models, some with several explanatory variables

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \varepsilon_i \\ Y_i &= \beta_0 + \beta_1 \log(X_i) + \varepsilon_i \end{aligned}$$

- Can sometimes transform nonlinear model into a linear model

$$\begin{aligned} Y_i &= \beta_0 \exp(\beta_1 X_i)\varepsilon_i \\ &\downarrow \\ \log(Y_i) &= \log(\beta_0) + \beta_1 X_i + \log(\varepsilon_i) \end{aligned}$$

- Have altered our assumptions about error
- Can perform nonlinear regression (`PROC NLIN`)

Nonconstant Variance

- Will discuss weighted analysis in Chapter 11
- Nonconstant variance often associated with a skewed error term distribution
- A transformation of Y often remedies both violations
- Will focus on Box-Cox transformations

$$Y' = Y^{\lambda}$$

Box-Cox Transformation

- Special cases:

  $\lambda = 1 \longrightarrow$ no transformation

  $\lambda = .5 \longrightarrow$ square root

  $\lambda = 0 \longrightarrow$ natural log (by definition)

- Can estimate λ using ML

$$f_i = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{1}{2\sigma^2}(Y_i^\lambda - \beta_0 - \beta_1 X_i)^2 \right\}$$

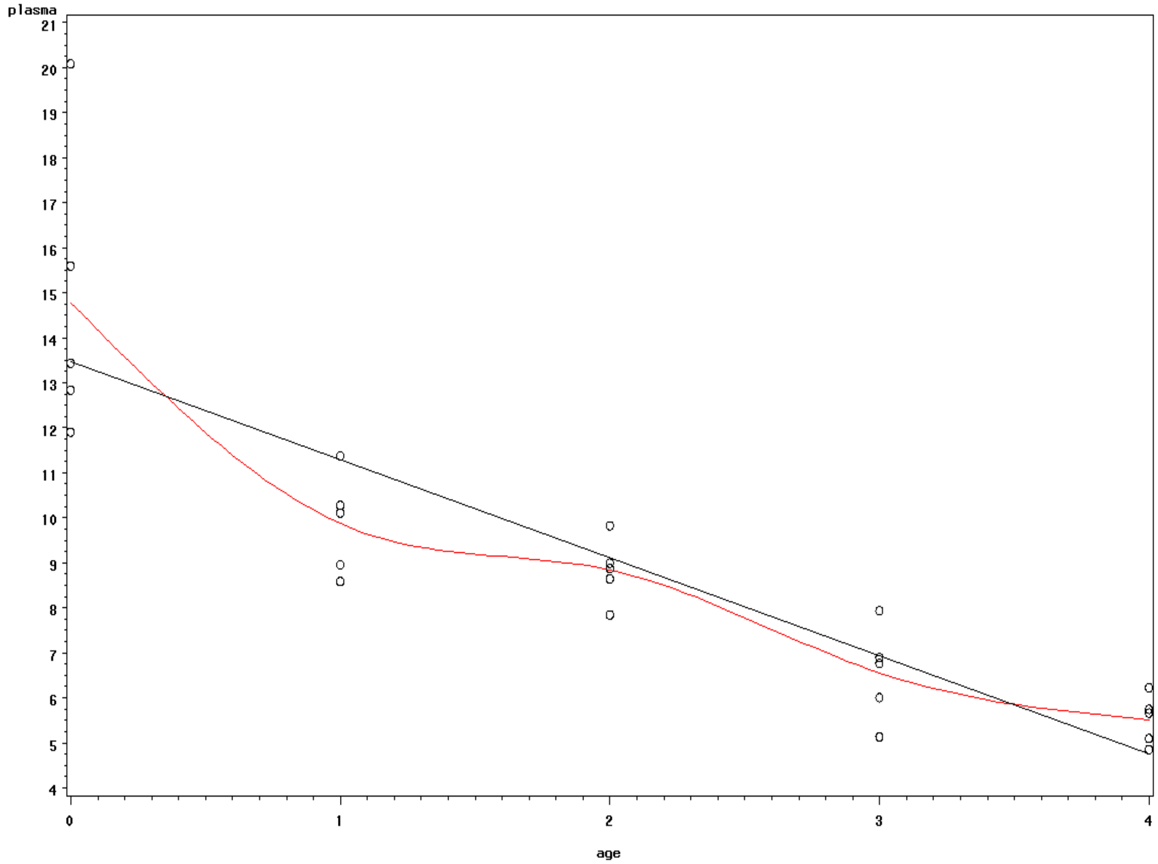
  - λ_{ML} minimizes SSE

- Can also do a numerical search

- `PROC TRANSREG` will do this in SAS

Example: Plasma Level (p. 132)

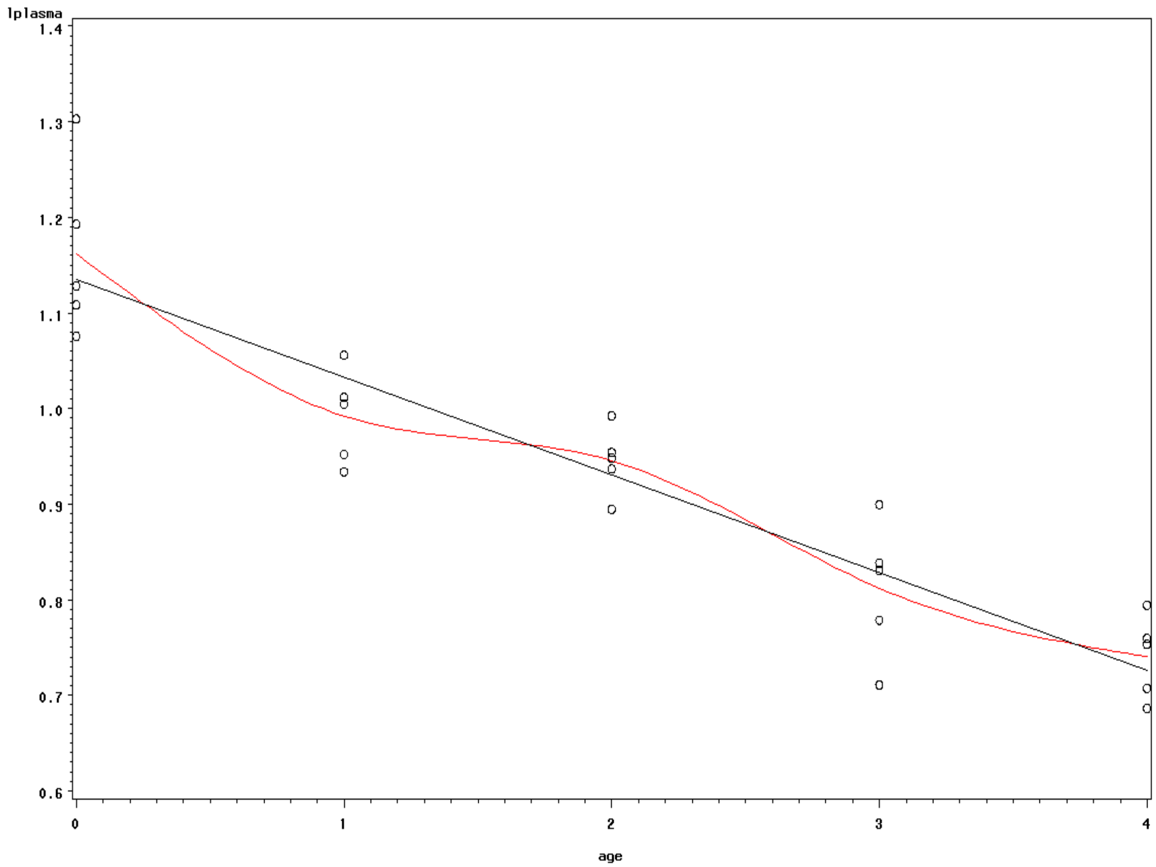
```
data a1;
    infile 'd:\nobackup\tmp\CH03TA08.txt';
    input age plasma lplasma;

symbol1 v=circle i=sm50 c=red; symbol2 v=circle i=rl c=black;
proc gplot;
    plot plasma*age=1 plasma*age=2/overlay; run;
```



```
proc transreg data=a1;
    model boxcox(plasma)=identity(age);
run;

The TRANSREG Procedure

  Lambda      R-Square    Log Like
   -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

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

- R^2 instead of SSE is given
- $\lambda = 0$ (log transform) is the most convenient value

```
proc gplot;
    plot lplasma*age=1 lplasma*age=2/overlay;
run; quit;
```



Chapter Review

- Diagnostics
  - Graphical methods
  - Statistical tests
- Remedies
  - Nonlinearity
  - Nonconstant variance