Topic 5 - Diagnostics and Remedial Measures

STAT 525 - Fall 2013

Outline

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

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 and Y

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

Proc Univariate

- 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

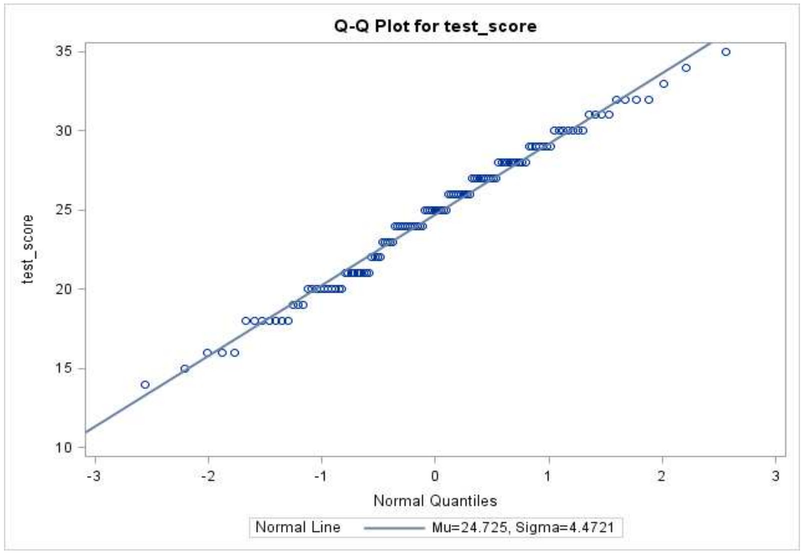
SAS Example - Grade Point

```
options nocenter;
goptions colors=(none);

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

proc print data=a1;
run;

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;
```

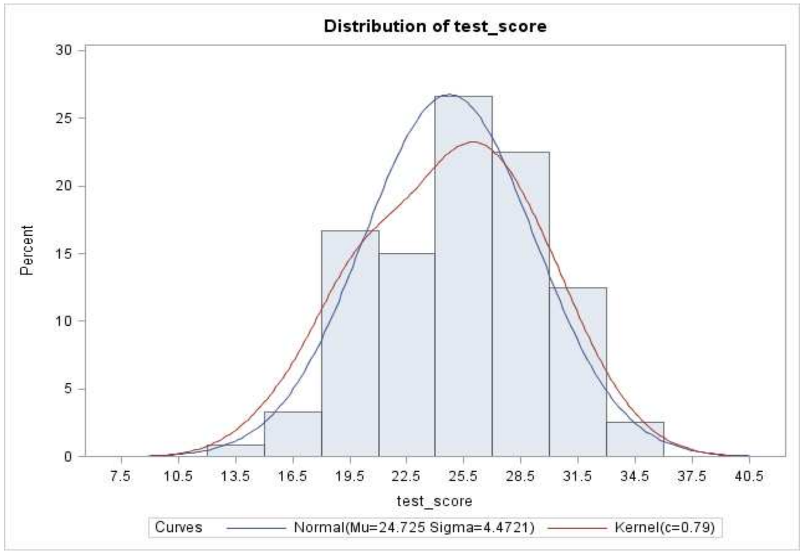


```
The UNIVARIATE Procedure
Variable:  test_score

                         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     24.725000      Std Deviation            4.47207
Median   25.000000      Variance                19.88837
Mode     24.000000      Range                   21.00000
                        Interquartile Range      7.00000
```


QQplot



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

Diagnostics for Residuals

- 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?

Getting at the Residuals

```
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;
 plot resid*lotsize;
 plot resid*seq;

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 Plots

- Plot e vs X can assess most questions
- Get same info from plot of e vs $\hat{Y}$ because X and $\hat{Y}$ 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 usually enough because looking for gross violations of assumptions (inference quite robust)

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

Other Formal Tests

- Durbin-Watson test for correlated errors
- Modified Levene / Brown-Forsythe test for constant variance
- Breusch-Pagan test for constant variance
- knnl106.sas contains SAS commands
- 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 $Y|X$ are independent $N(\beta_0 + \beta_1 X, \sigma^2)$
- $H_0 : E(Y) = \beta_0 + \beta_1 X$

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

Lack of Fit Test

- 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
  - $\hat{\mu}_j = \overline{Y}_{.j}$ and $s^2 = \sum\sum(Y_{ij} - \hat{\mu}_j)^2/(n-c)$
- Reduced Model: $Y_{ij} = \beta_0 + \beta_1 X_j + \varepsilon_{ij}$
  - Linear association
  - There are 2 parameters
  - $s^2 = \sum\sum(Y_{ij} - \hat{Y}_j)^2/(n-2)$

Lack of Fit Test

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

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

- 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

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

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

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

Nonconstant Variance

- Will discuss weighted analysis in Chpt 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

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

Box-Cox Transformation

- Special cases:

  $\lambda = 1 \longrightarrow$ no transformation

  $\lambda = .5 \longrightarrow$ square root

  $\lambda = 0 \longrightarrow$ natural log

- Can estimate λ using ML

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

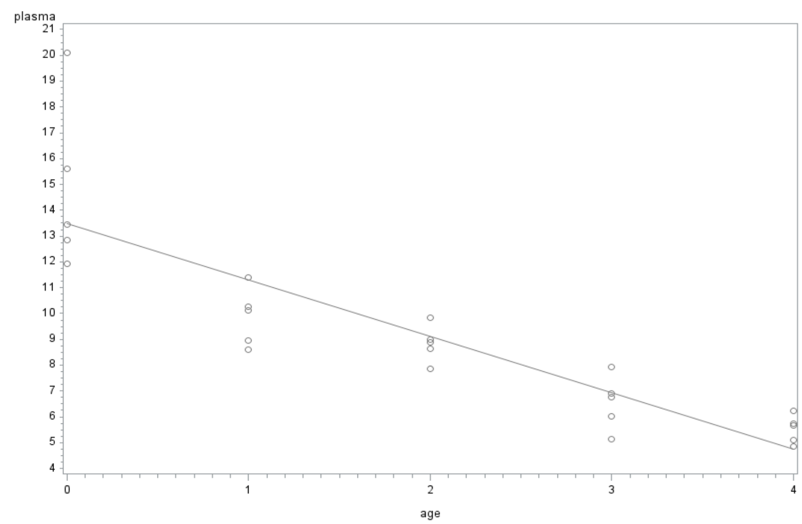
- Can also do a numerical search
- Proc Transreg will do this in SAS

Example

- Consider the plasma level example (pg 132)

```
data a1;
 infile 'u:\.www\datasets525\CH03TA08.dat';
 input age plasma lplasma;

symbol1 v=circle i=sm50;
proc gplot;
 plot plasma*age;
run;

proc reg;
 model plasma=age;
proc glm;
 class age;
 model plasma=age;
run;
```


Scatterplot



Lack of Fit Test

```
                 Analysis of Variance - Reduced Model
                            Sum of        Mean
Source             DF      Squares      Square  F Value  Pr > F
Model               1    238.05620   238.05620    70.21  <.0001
Error              23     77.98306     3.39057
Corrected Total    24    316.03926

                 Analysis of Variance - Full Model
                            Sum of        Mean
Source             DF      Squares      Square  F Value  Pr > F
Model               4    260.80498    65.20125    23.61  <.0001
Error              20     55.23428     2.76171
Corrected Total    24    316.03926
```

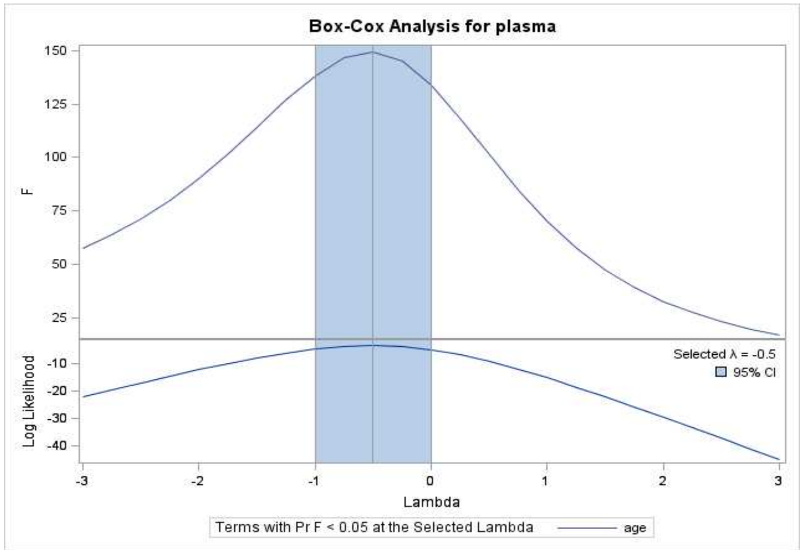
$$\begin{aligned} F^{\star} &= \frac{(77.98306 - 55.23428)/(23-20)}{2.76171} \\ &= 2.746 \\ &\downarrow \\ \text{P-value} &= 0.0674 \end{aligned}$$

Consider Box-Cox Transformation

- R^2 and F instead of SSE
- $\lambda = 0$ (log transform) is most convenient

```
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 *         < - Best Lambda
   -0.50          0.87      -3.5523 <         * - Confidence Interval
   -0.25          0.86      -3.9399 *         + - Convenient Lambda
    0.00 +        0.85      -5.0754 *
    0.25          0.84      -6.8988
    0.50          0.82      -9.2925
```

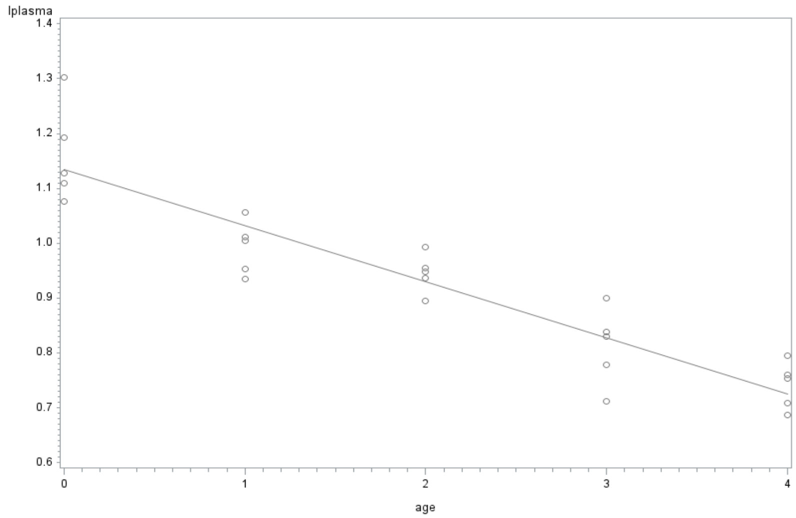

Visual for Box-Cox



Example, continued

```
proc gplot;
 plot lplasma*age;
run;

proc reg;
 model lplasma=age;
proc glm;
 class age;
 model lplasma=age;
run;
```

Scatterplot



Lack of Fit Test

Analysis of Variance - Reduced Model

Source	DF	Sum of Squares	Mean 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

Source	DF	Sum of Squares	Mean 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			

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

$$\downarrow$$

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

Assessing Correlation

- Nonconstant variance and/or correlated errors can be considered under a **linear mixed model**
- Assumes the errors $\mathbf{e} \sim \mathbf{MVN}(\mathbf{0}, \mathbf{\Sigma})$
- In standard regression model assumes $\Sigma = \sigma^2 I$
- These models are typically fit using **residual maximum likelihood** rather than ordinary least squares
- Can compare different choices of Σ using fit statistics
- More formal tests are possible if desired
- Typically choose the correlation structure first and then perform statistical inference

Assessing Correlation - Example

- Consider the Leaning Tower of Pisa example

```
data a1; input year lean @@;
cards;
75 642 76 644 77 656 78 667 79 673 80 688 81 696 82 698
83 713 84 717 85 725 86 742 87 757
;

proc mixed;
 model lean = year / ddfm=kr solution;
run;

proc mixed;
 model lean = year / ddfm=kr solution;
 repeated / subject=intercept type=ar(1) r=1;
run;
```

Standard Model

Covariance Parameter Estimates

Cov Parm	Estimate
Residual	17.4805

Fit Statistics

-2 Res Log Likelihood	70.5
AIC (smaller is better)	72.5
AICC (smaller is better)	72.9
BIC (smaller is better)	72.9

Solution for Fixed Effects

Effect	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept	-61.1209	25.1298	11	-2.43	0.0333
year	9.3187	0.3099	11	30.07	<.0001

AR(1) Correlation Structure

Covariance Parameter Estimates

Cov Parm	Subject	Estimate
AR(1)	Intercept	0.3219
Residual		20.7515

Fit Statistics

-2 Res Log Likelihood	70.0
AIC (smaller is better)	74.0
AICC (smaller is better)	75.5
BIC (smaller is better)	74.8

Solution for Fixed Effects

Effect	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept	-64.7894	27.6447	1.21	-2.34	0.2209
year	9.3689	0.3413	1.21	27.45	0.0122

***** Fit not as good here. Can stick with usual model *****

AR(1) Correlation Structure

Estimated R Matrix

20.7515	6.6791	2.1497	0.6919	0.2227	0.0717	0.0231	0.0074	0.0024	0.0008
6.6791	20.7515	6.6791	2.1497	0.6919	0.2227	0.0717	0.0231	0.0074	0.0024
2.1497	6.6791	20.7515	6.6791	2.1497	0.6919	0.2227	0.0717	0.0231	0.0074
0.6919	2.1497	6.6791	20.7515	6.6791	2.1497	0.6919	0.2227	0.0717	0.0231
0.2227	0.6919	2.1497	6.6791	20.7515	6.6791	2.1497	0.6919	0.2227	0.0717
0.0717	0.2227	0.6919	2.1497	6.6791	20.7515	6.6791	2.1497	0.6919	0.2227
0.0231	0.0717	0.2227	0.6919	2.1497	6.6791	20.7515	6.6791	2.1497	0.6919
0.0074	0.0231	0.0717	0.2227	0.6919	2.1497	6.6791	20.7515	6.6791	2.1497
0.0024	0.0074	0.0231	0.0717	0.2227	0.6919	2.1497	6.6791	20.7515	6.6791
0.0008	0.0024	0.0074	0.0231	0.0717	0.2227	0.6919	2.1497	6.6791	20.7515
0.0002	0.0008	0.0024	0.0074	0.0231	0.0717	0.2227	0.6919	2.1497	6.6791
0.0001	0.0002	0.0008	0.0024	0.0074	0.0231	0.0717	0.2227	0.6919	2.1497
0.0000	0.0001	0.0002	0.0008	0.0024	0.0074	0.0231	0.0717	0.2227	0.6919

Background Reading

- KNNL Chapters 3 and 4
- knnl101.sas, knnl106.sas, knnl134.sas