

 **THE  
POWER  
TO KNOW.**





## Modern Regression Analysis in SAS Software

Robert Cohen  
Software Manager  
Advanced Regression Research

**THE  
POWER  
TO KNOW.**

Copyright © 2007, SAS Institute Inc. All rights reserved.

 **THE  
POWER  
TO KNOW.**

## Good News and Bad News

### The Bad News

- I am NOT an expert in all the techniques and procedures I will present today

### The Good News

- I am NOT an expert in all the techniques and procedures I will present today

Copyright © 2007 SAS Institute Inc. All rights reserved.

## Two Themes – Robust and Nonparametric Modeling

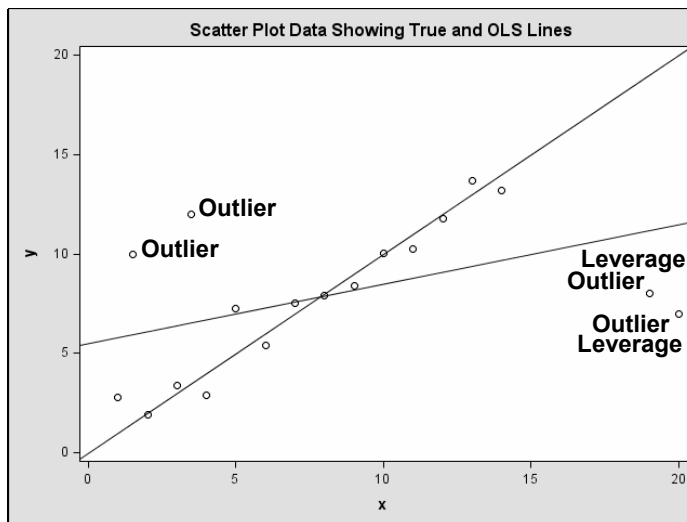
- Robust Modeling
  - PROC ROBUSTREG
  - PROC QUANTREG
  - PROC LOESS
- Non-Parametric Modeling
  - PROC LOESS
  - PROC GLIMMIX
  - PROC GLMSELECT

Copyright © 2007 SAS Institute Inc. All rights reserved.

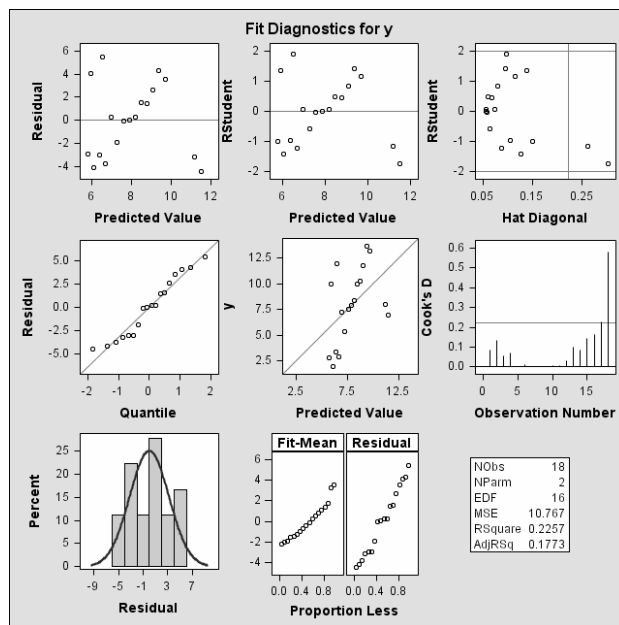
## Robust Regression – PROC ROBUSTREG

Copyright © 2007 SAS Institute Inc. All rights reserved.

## Why Use Robust Methods



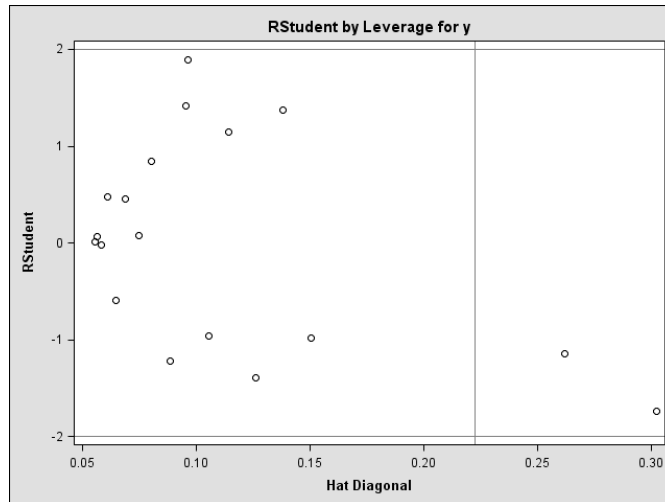
Copyright © 2007 SAS Institute Inc. All rights reserved.



Copyright © 2007 SAS Institute Inc. All rights reserved.

OLS  
Regression  
Diagnostics

## Catch 22: Non-Robust Diagnostics



Copyright © 2007 SAS Institute Inc. All rights reserved.

## Two Related Problems

### Identifying Points of Interest:

- Find outliers and leverage points that may be present in your data
- Often more important than modeling the data

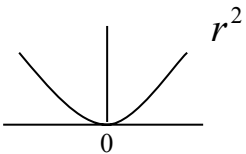
### Modeling:

- Fit a regression model resistant to outliers

Copyright © 2007 SAS Institute Inc. All rights reserved.

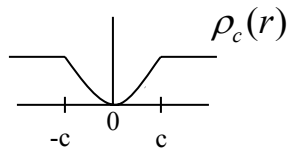
## M Estimation

OLS:  $\min_{\theta} \sum_{i=1}^n r_i^2$



$$r_i = y_i - x_i^T \theta$$

M:  $\min_{\theta} \sum_{i=1}^n \rho_c(r_i)$



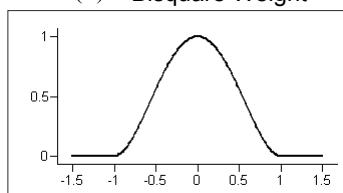
Copyright © 2007 SAS Institute Inc. All rights reserved.

## M Estimation via Iteratively Reweighted Least Squares (IRLS)

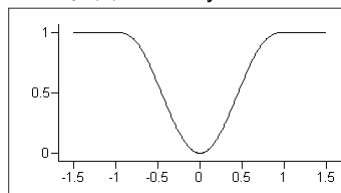
Iteration  $k$ :  $\min_{\theta} \sum_{i=1}^n w_i r_i^2$  where  $w_i = w(r_i^{(k-1)})$

and  $xw(x) = \rho'_c(x)$

$w(x)$  = Bisquare Weight

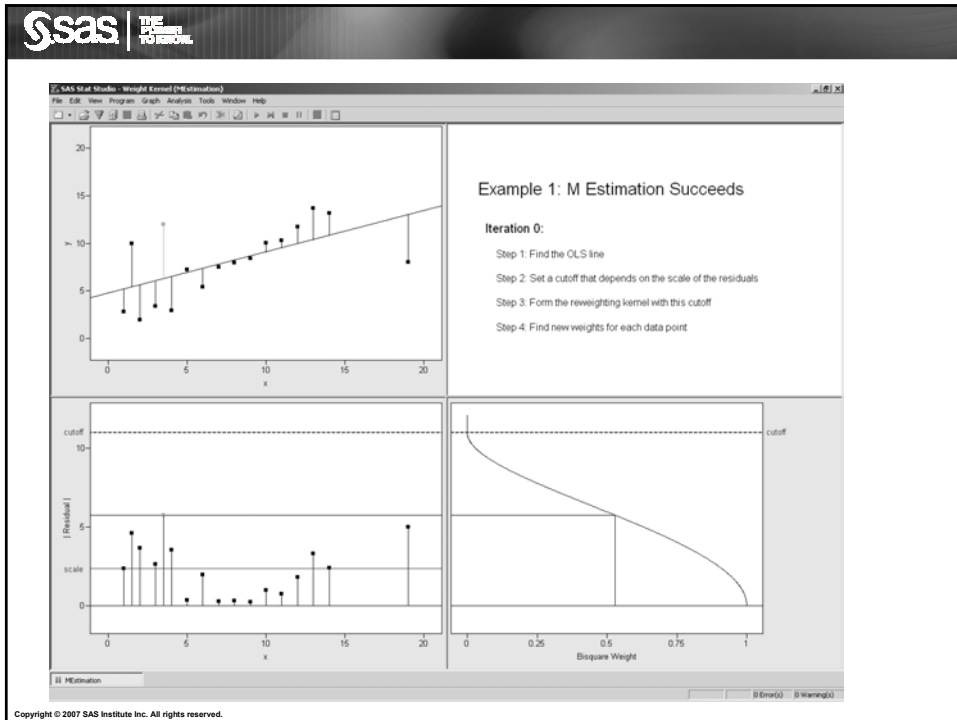


$\rho(x)$  = Tukey Function



Copyright © 2007 SAS Institute Inc. All rights reserved.

# Modern Regression Analysis



## M Estimation

### STRENGTHS

- fast - iteratively reweighted least squares
- can detect outliers
- extends to non-linear and nonparametric methods

### LIMITATIONS

- can fail if there are high leverage outliers
- can breakdown with “many” outliers

## High Breakdown Value (HBV) Methods

$$r_i = |y_i - x_i^T \theta| \quad r_{(1)} \leq r_{(2)} \leq r_{(3)} \cdots \leq r_{(n)}$$

HBV {

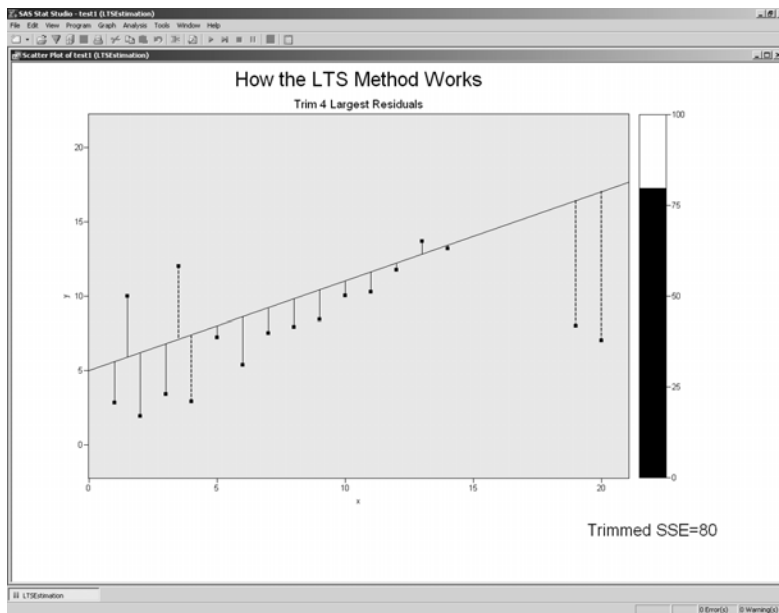
LMS:  $\min_{\theta} r_{(h)}^2$

LTS:  $\min_{\theta} \sum_{i=1}^h r_{(i)}^2$

S:  $\min_{\theta} S(\theta)$  where  $\frac{1}{n-p} \sum_{i=1}^n \chi \left( \frac{y_i - x_i \theta}{S} \right) = \beta$

MM: Hybrid of HBV and M

Copyright © 2007 SAS Institute Inc. All rights reserved.



Copyright © 2007 SAS Institute Inc. All rights reserved.

## Least Trimmed Squares (LTS) Estimation

### **STRENGTHS**

- robustly detect outliers and leverage points
- has high breakdown value

### **LIMITATIONS**

- computationally demanding
- low efficiency
- fit can be sensitive to small data variations

Copyright © 2007 SAS Institute Inc. All rights reserved.

## MM Estimation – (Two-stage M)

### **STRENGTHS**

- high efficiency of M with high breakdown of LTS
- identifies both high leverage points and outliers

### **LIMITATIONS**

- computationally demanding
- second stage (M) can depart from first stage (LTS) fit

Copyright © 2007 SAS Institute Inc. All rights reserved.

## Some Features of PROC ROBUSTREG

```

proc robustreg outest=est plots=all
    method=M (weightFunction=bisquare(c=6));
    M, LTS, S, MM
    class c1;
    model y = c1 x1 x1*x2
    / diagnostics leverage;
    test x1;
run;
    
```

Parameter Estimates      ODS Graphics

method suboptions

Classification variables (only with M and S)

GLM type model specification

identify outliers      identify leverage points (nonsingular only)

robust linear hypothesis tests (M and MM)

Copyright © 2007 SAS Institute Inc. All rights reserved.

## Robust Distance instead of Leverage

| Terminology                                                | Formula                                                            |
|------------------------------------------------------------|--------------------------------------------------------------------|
| $m$ = location (mean)                                      | $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$                           |
| $C$ = centered and scaled $X'X$                            | $\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^T (x_i - \bar{x})$     |
| $MD$ = Mahalanobis Distance                                | $MD_i^2 = (x_i - m)^T C^{-1} (x_i - m)$                            |
| $h$ = Leverage                                             | $h_i = (X(X'X)^{-1}X')_{i,i} = \frac{1}{n-1} MD_i^2 + \frac{1}{n}$ |
| <b>Robust Analogue of Mahalanobis distance</b>             |                                                                    |
| $RD$ = Robust Distance                                     | $RD_i^2 = (x_i - m_R)^T C_R^{-1} (x_i - m_R)$                      |
| $m_R$ and $C_R$ are robust estimates of location and scale |                                                                    |

Copyright © 2007 SAS Institute Inc. All rights reserved.

## Resistant Diagnostics

$$\text{OUTLIER} = \begin{cases} 0 & \text{if } |r| \leq 3\sigma \\ 1 & \text{otherwise} \end{cases}$$

$$\text{LEVERAGE} = \begin{cases} 0 & \text{if } RD \leq C(p) \\ 1 & \text{otherwise} \end{cases} \quad C(p) = \sqrt{\chi^2_{p;0.95}}$$

Copyright © 2007 SAS Institute Inc. All rights reserved.

## Running PROC ROBUSTREG

```
proc robustreg data=Ex1 plots=RDPlot method=LTS;  
    model y=x / diagnostics leverage;  
run;
```

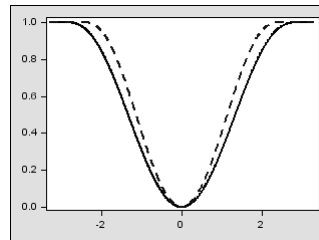
Run 1: method=LTS

Run 2: method=M

Run 3: method=MM

Run 4: method=MM(k<sub>0</sub>=2.5)

Tukey Function (MM)



Copyright © 2007 SAS Institute Inc. All rights reserved.

## Selected ROBUSTREG Output

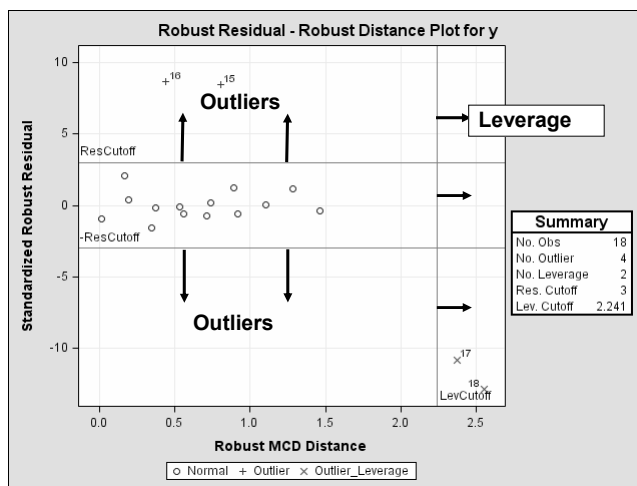
| LTS Profile                      |        |
|----------------------------------|--------|
| Total Number of Observations     | 18     |
| Number of Squares Minimized      | 14     |
| Number of Coefficients           | 2      |
| Highest Possible Breakdown Value | 0.2778 |

| Diagnostics Summary |            |        |
|---------------------|------------|--------|
| Observation Type    | Proportion | Cutoff |
| Outlier             | 0.2222     | 3.0000 |
| Leverage            | 0.1111     | 2.2414 |

| Diagnostics |                      |                     |          |                              |         |
|-------------|----------------------|---------------------|----------|------------------------------|---------|
| Obs         | Mahalanobis Distance | Robust MCD Distance | Leverage | Standardized Robust Residual | Outlier |
| 15          | 1.1840               | 0.8036              |          | 8.4482                       | *       |
| 16          | 0.8346               | 0.4407              |          | 8.6392                       | *       |
| 17          | 1.8730               | 2.3719              | *        | -10.8245                     | *       |
| 18          | 2.0477               | 2.5533              | *        | -12.8771                     | *       |

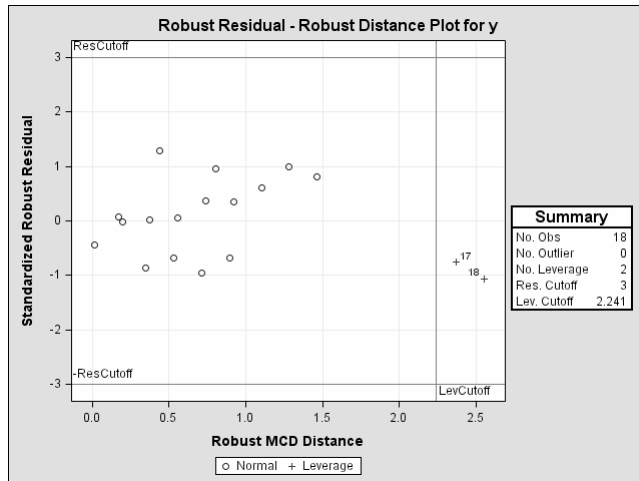
Copyright © 2007 SAS Institute Inc. All rights reserved.

## LTS Estimation – RD PLOT



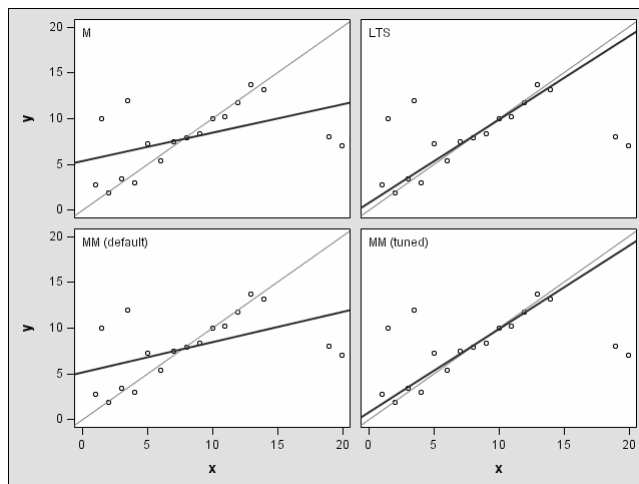
Copyright © 2007 SAS Institute Inc. All rights reserved.

## M Estimation RD Plot



Copyright © 2007 SAS Institute Inc. All rights reserved.

## Robust Fits with M, LTS, and MM methods

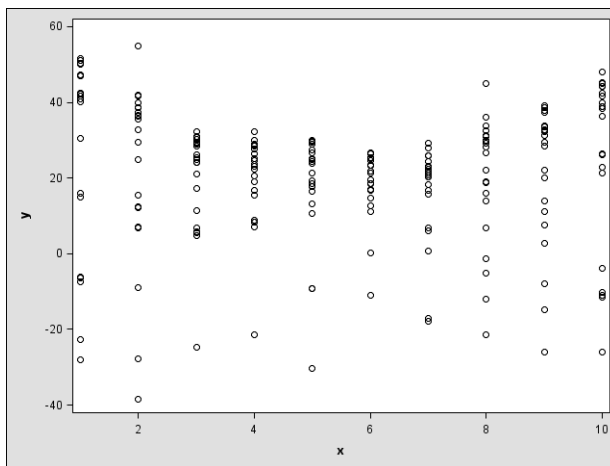


Copyright © 2007 SAS Institute Inc. All rights reserved.

## Quantile Regression – PROC QUANTREG

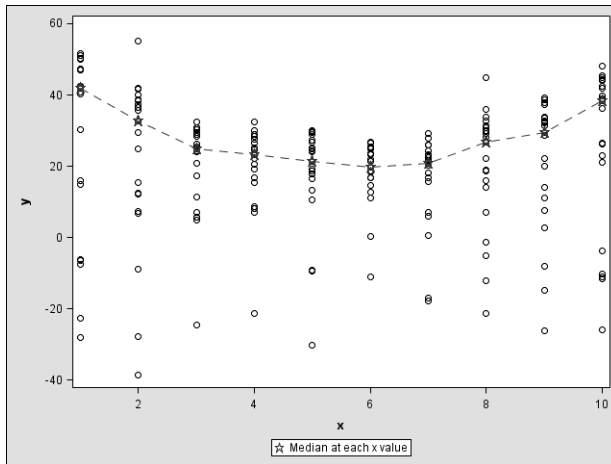
Copyright © 2007 SAS Institute Inc. All rights reserved.

## Predicting Quantiles



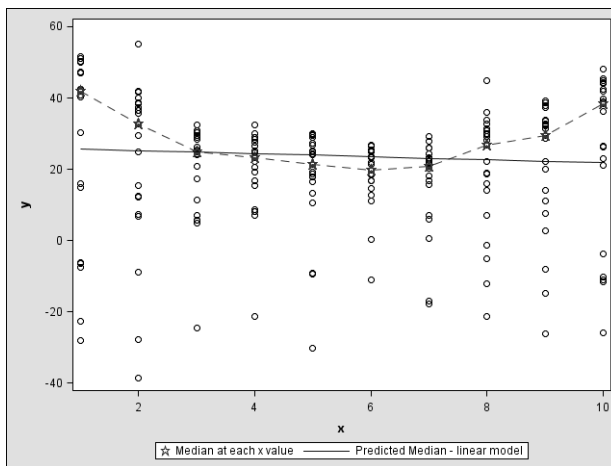
Copyright © 2007 SAS Institute Inc. All rights reserved.

## Predicting Quantiles



Copyright © 2007 SAS Institute Inc. All rights reserved.

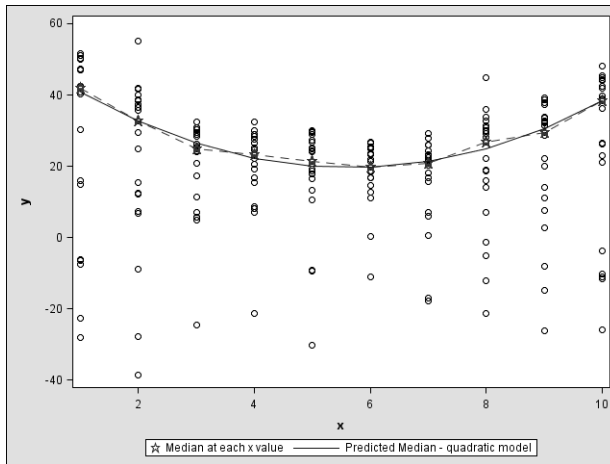
## Predicting Quantiles



```
proc quantreg;  
  model y=x;  
run;
```

Copyright © 2007 SAS Institute Inc. All rights reserved.

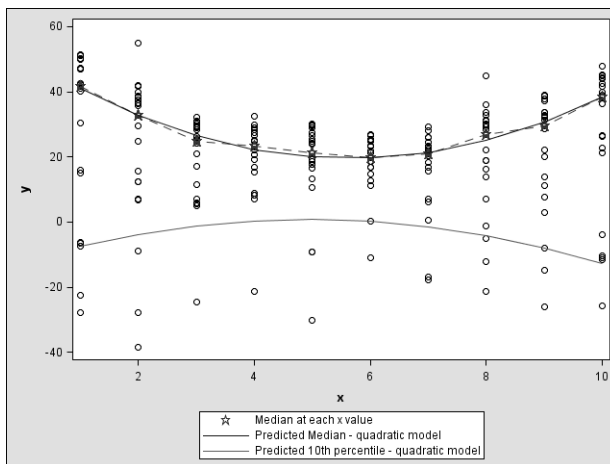
## Predicting Quantiles



```
proc quantreg;  
  model y=x x*x;  
run;
```

Copyright © 2007 SAS Institute Inc. All rights reserved.

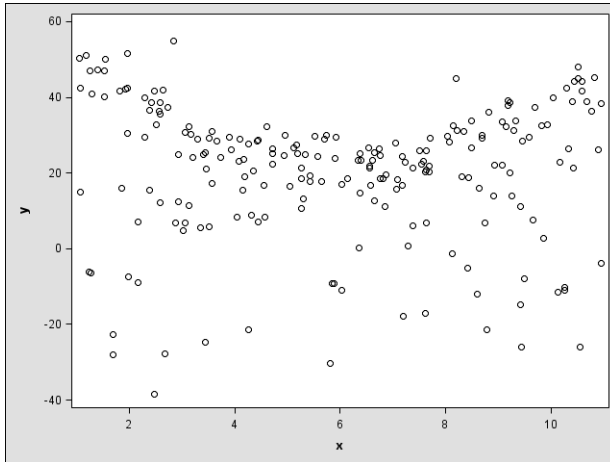
## Predicting Quantiles



```
proc quantreg;  
  model y=x x*x  
  /quantiles=0.1;  
run;
```

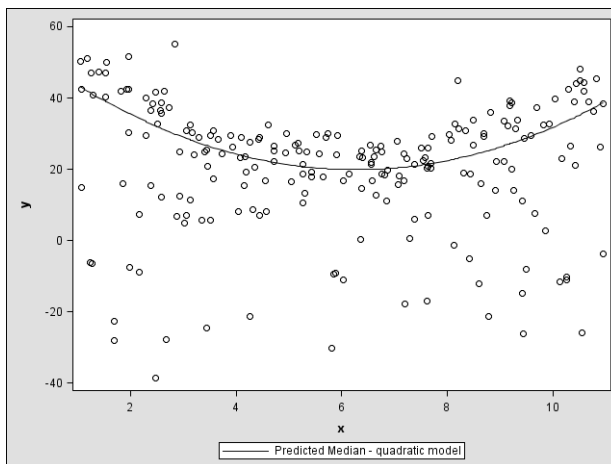
Copyright © 2007 SAS Institute Inc. All rights reserved.

## Predicting Quantiles



Copyright © 2007 SAS Institute Inc. All rights reserved.

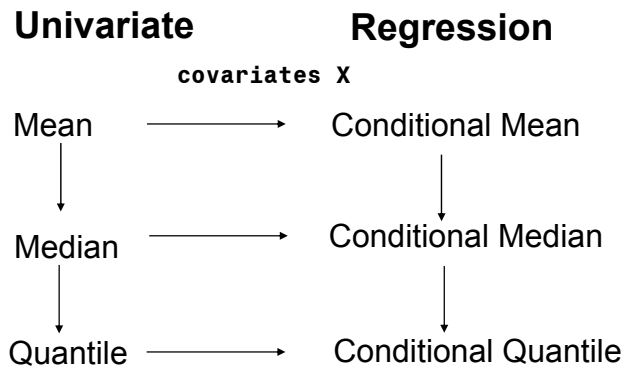
## Predicting Quantiles



```
proc quantreg;  
  model y=x x*x;  
run;
```

Copyright © 2007 SAS Institute Inc. All rights reserved.

## Moving from Univariate to Regression Context



Copyright © 2007 SAS Institute Inc. All rights reserved.

## Univariate Notions as Optimization Problems

$$y_1 \leq y_2 \leq y_3 \leq \cdots \leq y_n$$

mean:  $\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$        $\bar{y} = \arg \min_c \sum_{i=1}^n (y_i - c)^2$

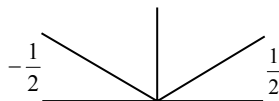
median:  $y^{(0.5)} = y_{\lfloor 0.5(n+1) \rfloor}$        $y^{(0.5)} = \arg \min_c \sum_{i=1}^n |y_i - c|$

quantile:  $y^{(\tau)} = y_{\lfloor \tau(n+1) \rfloor}$        $y^{(\tau)} = \arg \min_c \sum_{i=1}^n \rho_{\tau}(y_i - c)$

Copyright © 2007 SAS Institute Inc. All rights reserved.

## Loss Function

$$\rho_{\tau}(r) = \begin{cases} r\tau & r \geq 0 \\ r(\tau - 1) & r < 0 \end{cases} \quad \tau \in (0, 1)$$



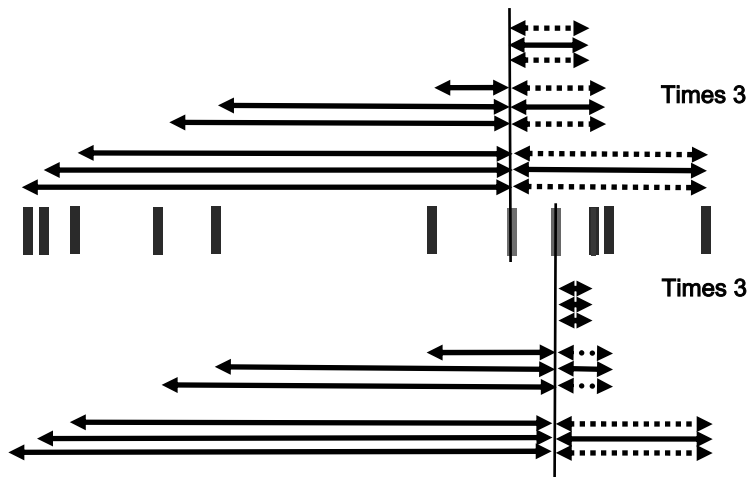
## Median Regression



## Quantile Regression

Copyright © 2007 SAS Institute Inc. All rights reserved

# Let's Play a Game



Copyright © 2007 SAS Institute Inc. All rights reserved.



THE POWER  
TO KNOW.

## Winning Strategy when distances to right multiplied by 3



**Number Ratio**    3 : 1

**Fraction**    3/4 : 1/4

**Winning Position**    75 percentile

Copyright © 2007 SAS Institute Inc. All rights reserved.



THE POWER  
TO KNOW.

## Finding Quantiles using Minimization



$$\begin{aligned}
 L &= \frac{1}{4} \{ (c - y_1) + \dots + (c - y_6) \} + \frac{3}{4} \{ (y_7 - c) + \dots + (y_9 - c) \} \\
 &= -\frac{1}{4} \sum_{y_i < c} (y_i - c) + \frac{3}{4} \sum_{y_i > c} (y_i - c) \\
 &= (\tau - 1) \sum_{y_i < c} (y_i - c) + \tau \sum_{y_i > c} (y_i - c) \quad \tau = \frac{3}{4} \\
 &= \sum_{i=1}^n \rho_\tau(y_i - c) \quad \rho_\tau(r) = \begin{cases} (\tau - 1) r, & r < 0 \\ \tau r, & r \geq 0 \end{cases}
 \end{aligned}$$

To minimize  $L$ ,  $c = \tau^{\text{th}}$  percentile of  $(y_1, y_2, y_3, \dots, y_n)$

Copyright © 2007 SAS Institute Inc. All rights reserved.

## Generalizing to Regression Context

### Univariate

$$\bar{y} = \arg \min_c \sum_{i=1}^n (y_i - c)^2$$

$$y^{(0.5)} = \arg \min_c \sum_{i=1}^n |y_i - c|$$

$$y^{(\tau)} = \arg \min_c \sum_{i=1}^n \rho_{\tau}(y_i - c)$$

### Regression

$$\beta^{\text{LS}} = \arg \min_{\beta} \sum_{i=1}^n (y_i - x_i' \beta)^2$$

$$\beta^{(0.5)} = \arg \min_{\beta} \sum_{i=1}^n |y_i - x_i' \beta|$$

$$\beta^{(\tau)} = \arg \min_{\beta} \sum_{i=1}^n \rho_{\tau}(y_i - x_i' \beta)$$

Copyright © 2007 SAS Institute Inc. All rights reserved.

## Terminology

$$\beta^{(\tau)} = \arg \min_{\beta \in \mathbb{R}^m} \sum_{i=1}^n \rho_{\tau}(y_i - x_i' \beta)$$

Quantile Regression Function:  $Q_{\tau}(Y | X) = X\beta^{(\tau)}$

Quantile Objective Function:  $g(\tau) = \sum_{i=1}^n \rho_{\tau}(y_i - x_i' \beta^{(\tau)})$

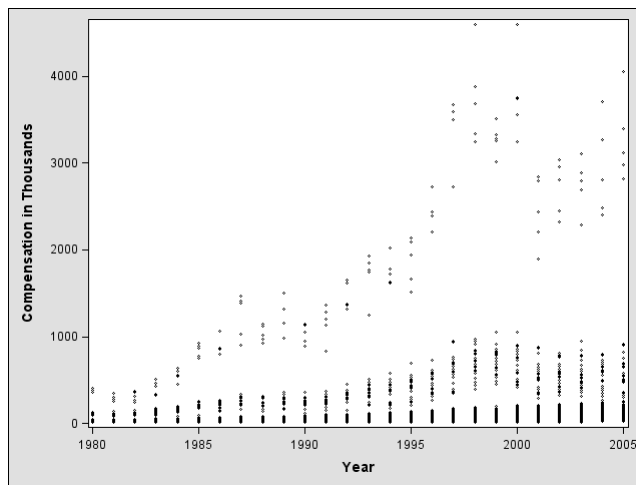
Copyright © 2007 SAS Institute Inc. All rights reserved.

## Benefits of Quantile Regression

- Predicts quantiles
- Makes no distributional assumptions
- Robust to outliers
- Robust to heterogeneity
- Provides a complete picture of the distribution

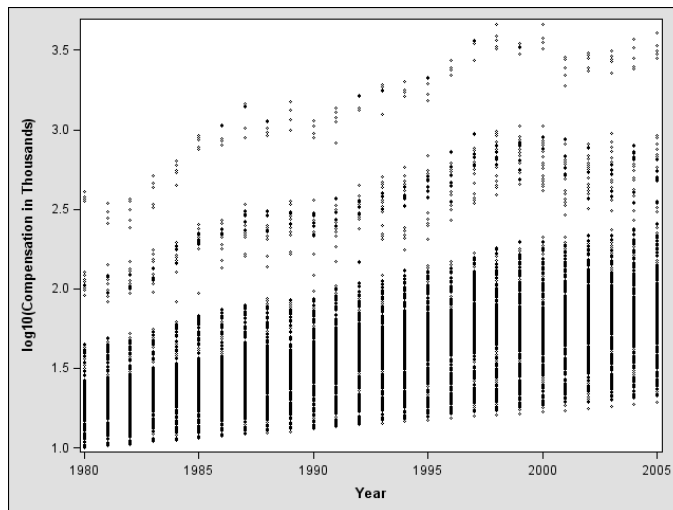
Copyright © 2007 SAS Institute Inc. All rights reserved.

## Employee Compensation



Copyright © 2007 SAS Institute Inc. All rights reserved.

## Employee Compensation



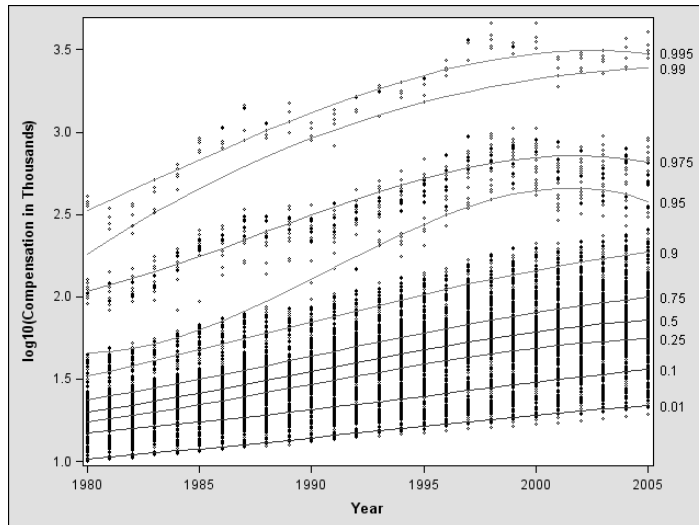
Copyright © 2007 SAS Institute Inc. All rights reserved.

## Running PROC QUANTREG

```
proc quantreg data=Compensation;  
  model LogCompensation=yearOffset|yearOffset|yearOffset  
    / quantiles = 0.75;  
  output out=myOut pred=p75;  
run;
```

Copyright © 2007 SAS Institute Inc. All rights reserved.

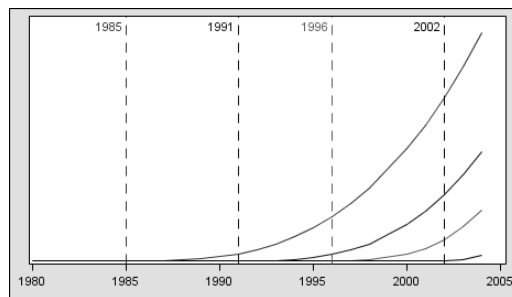
## Quantile Regression: Cubic Polynomial Model



Copyright © 2007 SAS Institute Inc. All rights reserved.

## Adding Regression Splines

- A cubic polynomial for upper quantiles is inadequate.
- Using higher order polynomials is ill-advised
- Instead use cubic splines with a few well-chosen knots



Copyright © 2007 SAS Institute Inc. All rights reserved.

## Regression Splines for the Compensation Data.

Place knots at years 1985, 1991, 1996, and 2001

```
data compensation;
  set compensation;
  array b{4} b1-b4;

  b{1}=yearOffset-5;
  b{2}=yearOffset-11;
  b{3}=yearOffset-16;
  b{4}=yearOffset-21;

  do i=1 to 4;
    if b{i}<0 then b{i}=0;
    else b{i}=b{i}**3;
  end;
run;
```

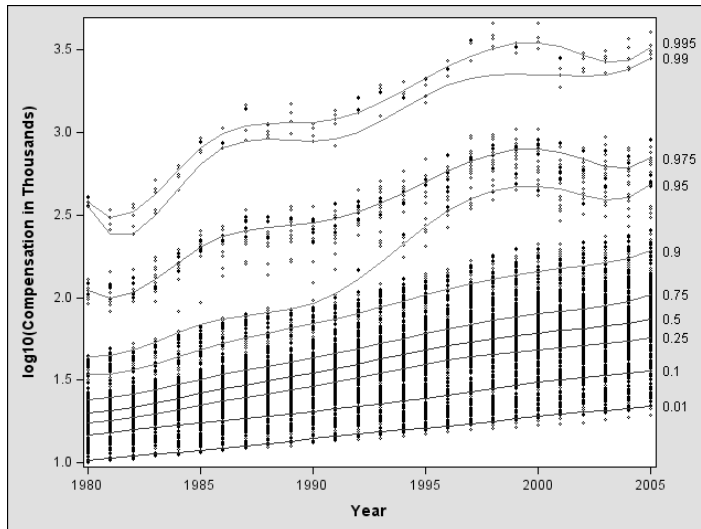
Copyright © 2007 SAS Institute Inc. All rights reserved.

## Include Regression Splines

```
proc quantreg data=Compensation;
  model LogCompensation = yearOffset|yearOffset|yearOffset
      b1-b4 / quantiles = 0.75;
  output out=myOut pred=pSpline75;
run;
```

Copyright © 2007 SAS Institute Inc. All rights reserved.

## Quantile Regression: Cubic Spline Model

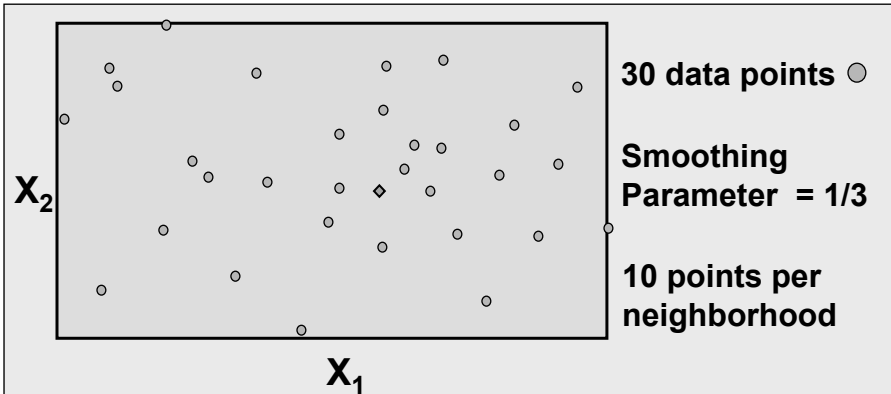


Copyright © 2007 SAS Institute Inc. All rights reserved.

## Robust Nonparametric Regression PROC LOESS

Copyright © 2007 SAS Institute Inc. All rights reserved.

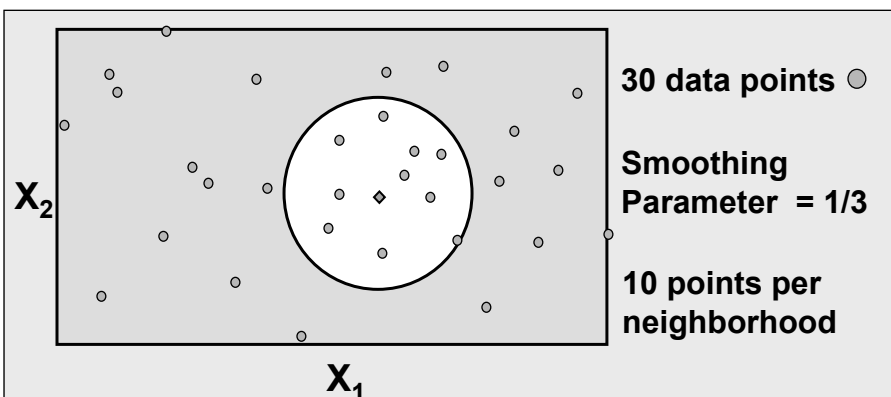
## A LOESS Primer – Two Predictors



**Step 1: Choose a point at which you want a loess fit. This need not be an original data point**

Copyright © 2007 SAS Institute Inc. All rights reserved.

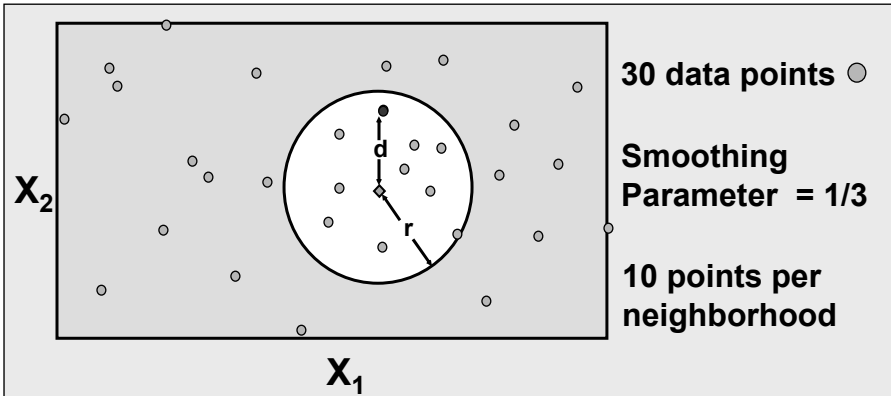
## A LOESS Primer – Two Predictors



**Step 2: Find the smallest circle about the fit point that encloses 10 data points. This circle defines the local neighborhood**

Copyright © 2007 SAS Institute Inc. All rights reserved.

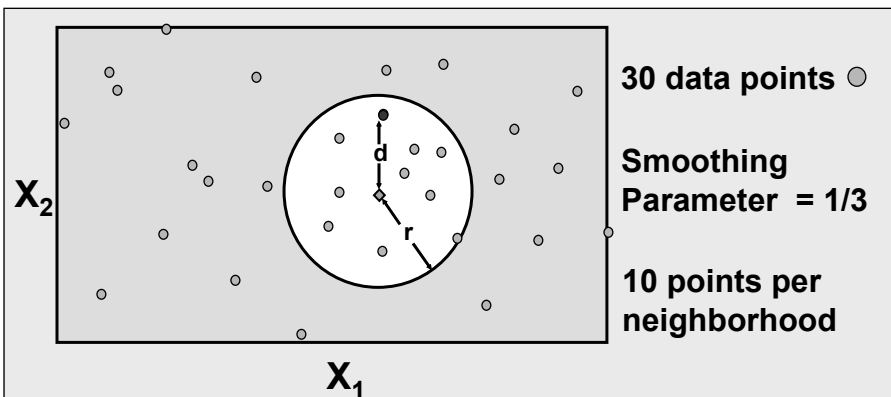
## A LOESS Primer – Two Predictors



**Step 3: Tricube local weight** =  $\left(1 - \frac{d^3}{r^3}\right)^3$

Copyright © 2007 SAS Institute Inc. All rights reserved.

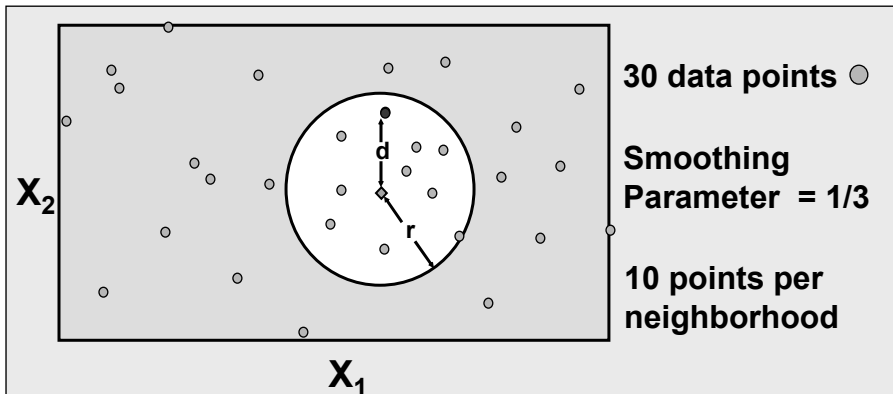
## A LOESS Primer – Two Predictors



**Step 4: Locally fit a surface using weighted linear or quadratic least squares.**

Copyright © 2007 SAS Institute Inc. All rights reserved.

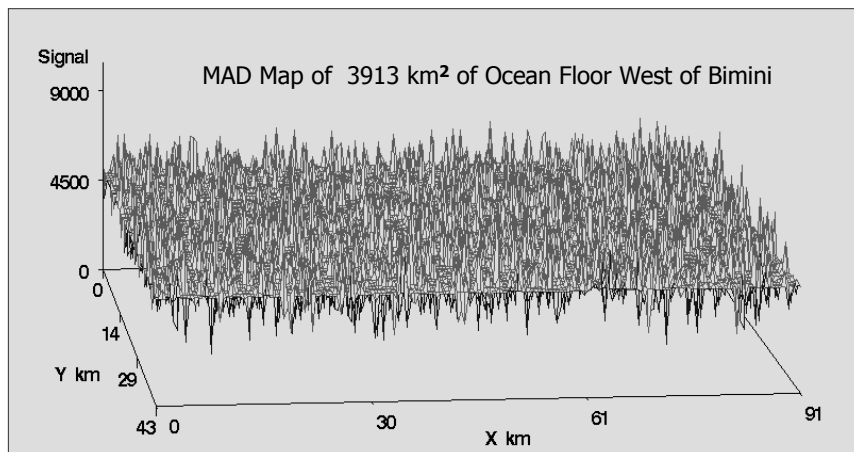
## A LOESS Primer – Two Predictors



**Step 5: Evaluate this regression surface at the fit point ♦ to get the loess prediction at this point.**

Copyright © 2007 SAS Institute Inc. All rights reserved.

## Magnetic Anomaly Detection Data



Reference: Kidding, I.M. (2006), to appear

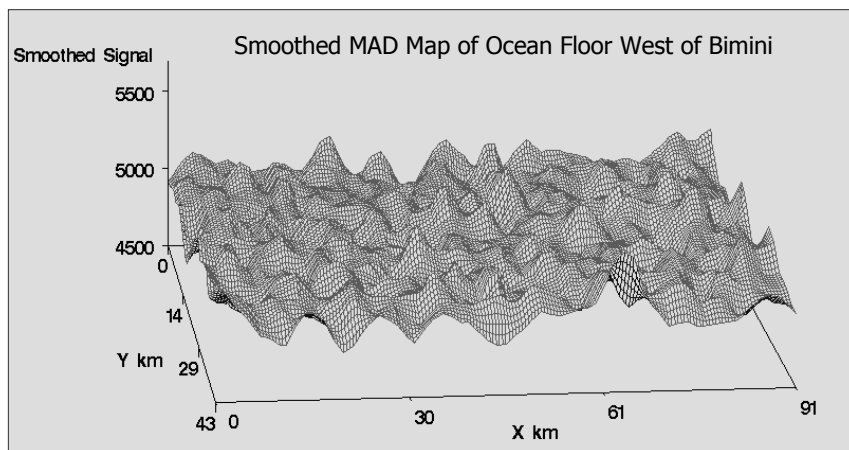
Copyright © 2007 SAS Institute Inc. All rights reserved.

## Using PROC LOESS

```
proc loess data=MAD;  
  model signal=x y/ smooth=0.0085;  
  ods output OutputStatistics=statOut;  
run;
```

Copyright © 2007 SAS Institute Inc. All rights reserved.

## Non-robust Loess Fit of MAD Data



Copyright © 2007 SAS Institute Inc. All rights reserved.

## Robust LOESS fitting – M Estimation

**Initialize:**  $w_i = 1 \quad i = 1, 2, \dots, n$

**Repeat until convergence (or a fixed number of times):**

- |                |                                                                                          |                                     |
|----------------|------------------------------------------------------------------------------------------|-------------------------------------|
| <b>Step 1:</b> | $\hat{y} = L_w y, \quad r = y - \hat{y}$                                                 | Perform weighted LOESS fit          |
| <b>Step 2:</b> | $c = \text{median}( r_i ), \quad t = \frac{r_i}{6c}$                                     | Estimate the scale of residuals     |
| <b>Step 3:</b> | $w_i = \begin{cases} (1-t^2)^2 & \text{if }  t  < 1 \\ 0 & \text{otherwise} \end{cases}$ | Bisquare kernel to form new weights |

**End Repeat:**

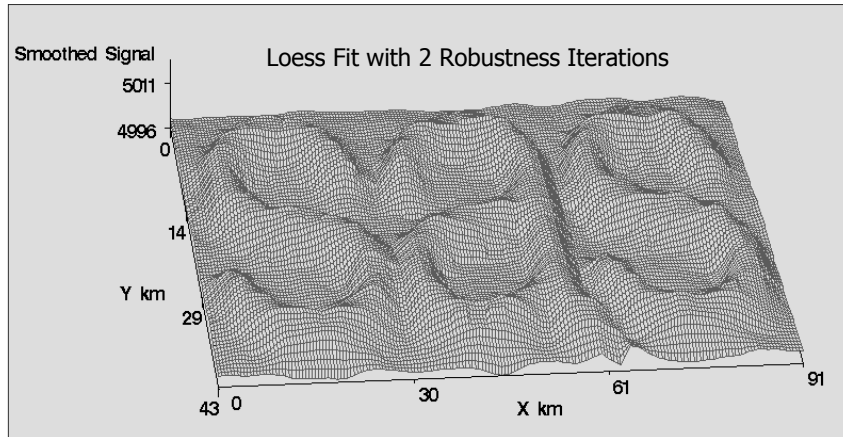
Copyright © 2007 SAS Institute Inc. All rights reserved.

## Robust LOESS Fitting

```
proc loess data=MAD;
  model signal=x y/
    smooth=0.0085
    iterations=3;
  ods output OutputStatistics=statOut;
run;
```

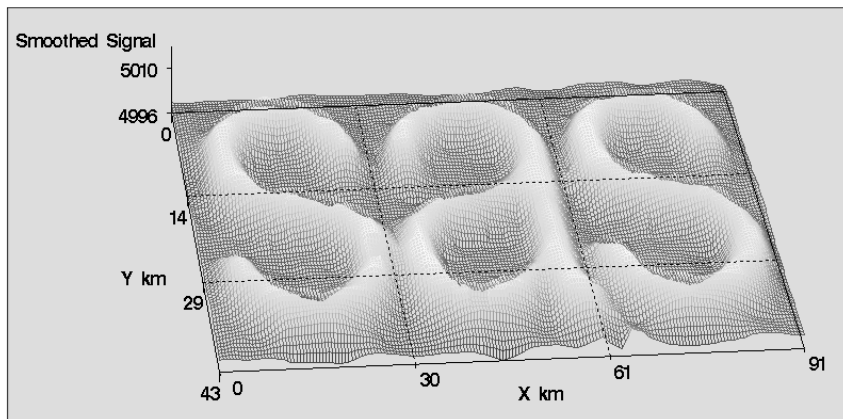
Copyright © 2007 SAS Institute Inc. All rights reserved.

## Robust Loess Fit



Copyright © 2007 SAS Institute Inc. All rights reserved.

## Atlantis Revealed



Copyright © 2007 SAS Institute Inc. All rights reserved.

## Smoothing with a Mixed Model Formulation PROC GLIMMIX

Copyright © 2007 SAS Institute Inc. All rights reserved.

## A Simple Mixed Model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma} + \mathbf{e}$$

$$\boldsymbol{\gamma} \sim N(\mathbf{0}, \sigma^2 \mathbf{I})$$

$$\mathbf{e} \sim N(\mathbf{0}, \phi \mathbf{I})$$

- Obtain  $\boldsymbol{\beta}$  and  $\boldsymbol{\gamma}$  by Henderson's method: For fixed  $\phi$  and  $\sigma^2$

$$\text{minimize } Q(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \phi^{-1} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\boldsymbol{\gamma}\|^2 + \sigma^{-2} \boldsymbol{\gamma}'\boldsymbol{\gamma}$$

- Mixed model approach simultaneously estimates  $\phi$  and  $\sigma^2$  and yields the above  $\boldsymbol{\beta}$  and  $\boldsymbol{\gamma}$  with these estimates

Copyright © 2007 SAS Institute Inc. All rights reserved.

## A Penalized (Linear) Spline

$$y_i = \beta_0 + \beta_1 x_i + \sum_{j=1}^K \gamma_j f(x_i; t_j) + e_i$$

$\mathbf{Y} = \mathbf{X}\beta + \mathbf{Z}\gamma + \mathbf{e}$   
 $\mathbf{e} \sim (0, \phi \mathbf{I})$

Spline coefficients:  $\gamma_j$   
 Spline knots:  $t_j$   
 Spline functions:  $f(x_i; t_j)$

- For a given  $\lambda$ , find  $\beta$  and  $\gamma$  that minimize
 
$$Q^*(\beta, \gamma) = \|\mathbf{y} - \mathbf{X}\beta - \mathbf{Z}\gamma\|^2 + \lambda^2 \gamma' \gamma$$
- Chose penalty  $\lambda$  by solving above many times and optimizing a fit criterion

Copyright © 2007 SAS Institute Inc. All rights reserved.

## Comparison

| Mixed Model                                                                                                       | Penalized Spline                                                                                        |
|-------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------|
| $Q(\beta, \gamma) = \phi^{-1} \ \mathbf{y} - \mathbf{X}\beta - \mathbf{Z}\gamma\ ^2 + \sigma^{-2} \gamma' \gamma$ | $Q^*(\beta, \gamma) = \ \mathbf{y} - \mathbf{X}\beta - \mathbf{Z}\gamma\ ^2 + \lambda^2 \gamma' \gamma$ |
| $\arg \min Q^*(\beta, \gamma) = \arg \min \phi^{-1} Q^*(\beta, \gamma) = \arg \min Q(\beta, \gamma)$              |                                                                                                         |

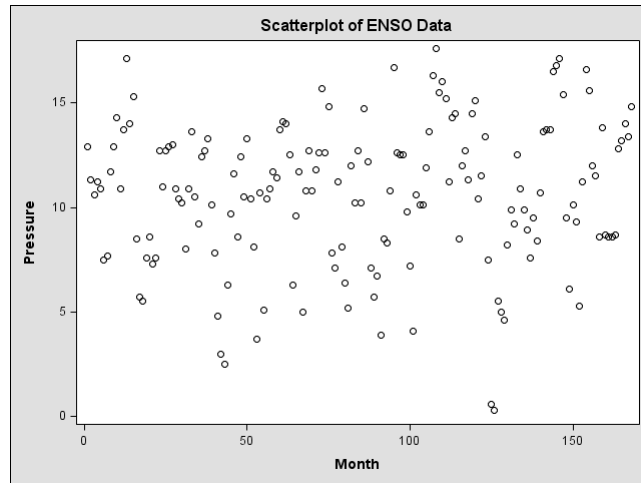
- A penalized spline can be fit with mixed model software

$$\sigma^2 = \phi / \lambda^2$$

Copyright © 2007 SAS Institute Inc. All rights reserved.

## Scatterplot Smoothing

Mean  
monthly  
pressure  
difference  
between  
Darwin  
and  
Easter  
Island



NIST (1998), "Statistical Reference Data Sets"

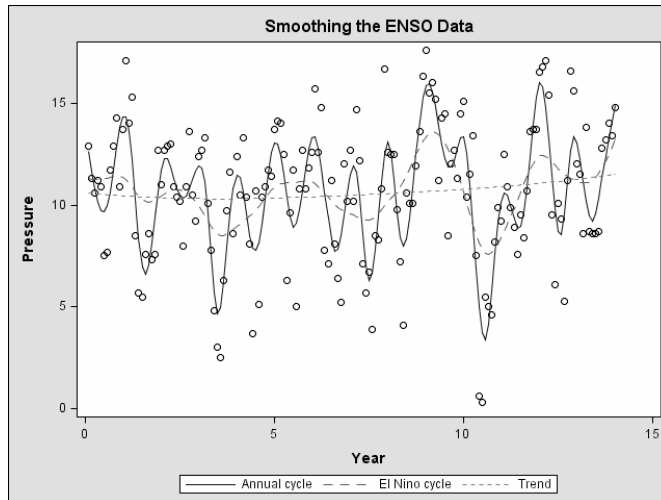
Copyright © 2007 SAS Institute Inc. All rights reserved.

## Scatterplot Smoothing with PROC GLIMMIX

```
proc glimmix data=ENSO noprofile;  
  model Pressure = Year;  
  random Year / type=rsmooth knotmethod=equal(50);  
  /* Starting Parameter Values for optimization */  
  parms (200) (10); /* yields Annual cycle */  
  parms (2) (10); /* yields El Nino cycle */  
  parms (0) (10); /* yields trend */  
run;
```

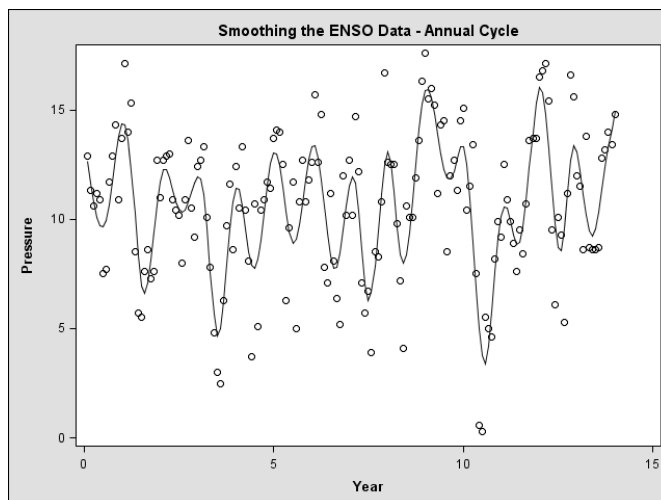
Copyright © 2007 SAS Institute Inc. All rights reserved.

## Smoothing Using Radial Basis Functions in a Mixed Model Formulation



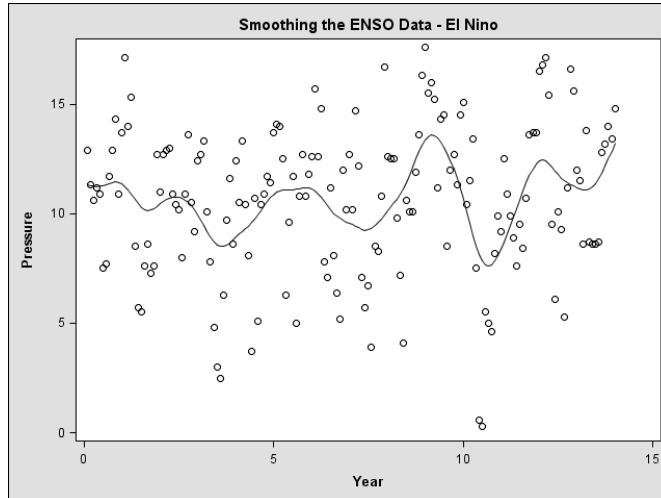
Copyright © 2007 SAS Institute Inc. All rights reserved.

## Mixed Model Smoothing



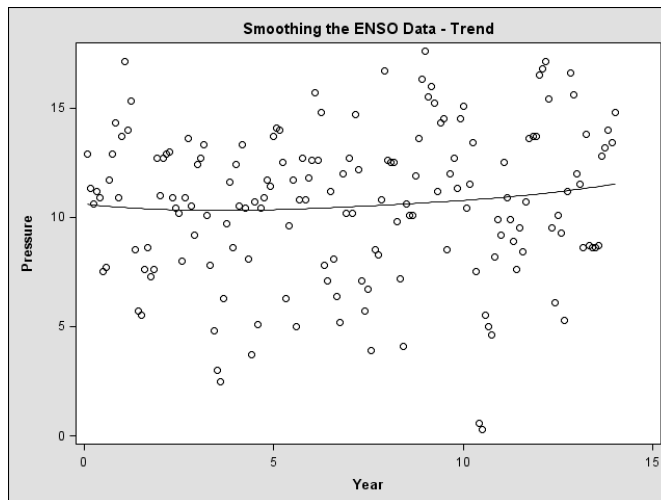
Copyright © 2007 SAS Institute Inc. All rights reserved.

## Mixed Model Smoothing



Copyright © 2007 SAS Institute Inc. All rights reserved.

## Mixed Model Smoothing



Copyright © 2007 SAS Institute Inc. All rights reserved.

## LASSO Selection – PROC GLMSELECT

Copyright © 2007 SAS Institute Inc. All rights reserved.

## Defining the LASSO

For a given tuning parameter  $t \geq 0$

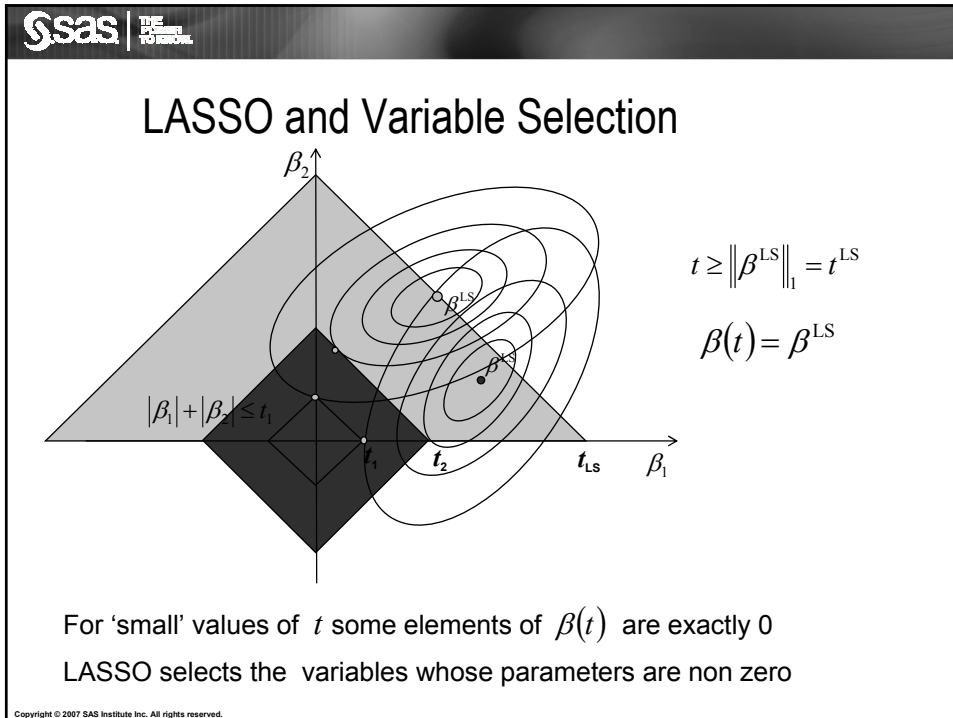
$$\beta(t) = \arg \min_{\beta} \sum_{i=1}^n (y_i - (X\beta)_i)^2$$

$$\text{subject to } \|\beta\|_1 = \sum_{j=1}^m |\beta_j| \leq t$$

Least Squares  
LASSO

Constraint

Copyright © 2007 SAS Institute Inc. All rights reserved.



**ssas** | THE POWER TO KNOW

## LASSO Viewed as a Stepwise Method

|                              |                                                                                                          |
|------------------------------|----------------------------------------------------------------------------------------------------------|
| <b>Sequence</b>              | $0 = t_0 < t_1 < t_2 < \dots < t_M = t_{LS}$                                                             |
| <b>Constraint</b>            | $\ \beta^{(i)}\ _1 \leq t_i$                                                                             |
| <b>LASSO Solutions</b>       | $0 = \beta^{(0)} \quad \beta^{(1)} \quad \beta^{(2)} \quad \dots \quad \beta^{(M)} = \beta^{LS}$         |
| <b>Add Step</b>              | $\beta^{(i)}$ has one more non-zero entry than $\beta^{(i-1)}$                                           |
| <b>Drop Step</b>             | $\beta^{(i)}$ has one fewer non-zero entry than $\beta^{(i-1)}$                                          |
| <b>Piecewise-linear path</b> | For $t_i \leq t \leq t_{i+1}$<br>$\beta(t)$ is a linear combination of $\beta^{(i)}$ and $\beta^{(i+1)}$ |

Copyright © 2007 SAS Institute Inc. All rights reserved.

## The Diabetes Data Set

```
data Diabetes;
  input Patient Age Sex BMI BP TC LDL HDL TCH LTG GLU Progression;
datalines;
  1 59 2 32.1 101 157 93.2 38 4 4.9 87 151
  2 48 1 21.6 87 183 103.2 70 3 3.9 69 75
  3 72 2 30.5 93 156 93.6 41 4 4.7 85 141
  .
  . more data lines
  .
  441 36 1 30.0 95 201 125.2 42 5 5.2 85 220
  442 36 1 19.6 71 250 133.2 97 3 4.6 92 57
;
```

### Source

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004),  
Least Angle Regression (with discussion),  
*Annals of Statistics*, 32, 407- 499.

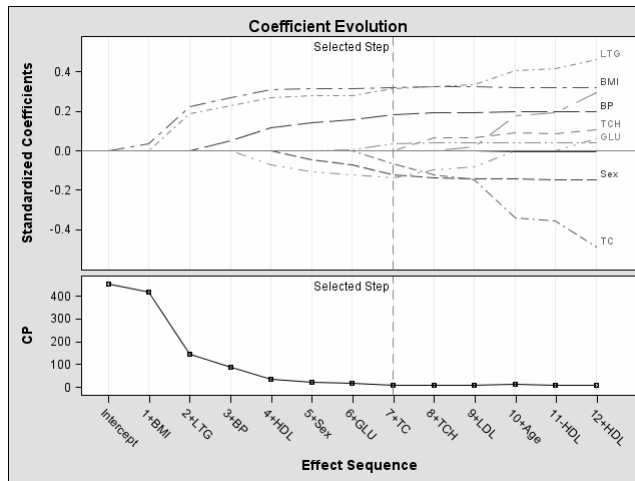
Copyright © 2007 SAS Institute Inc. All rights reserved.

## LASSO USING PROC GLMSELECT

```
ods graphics on;
proc glmselect data=diabetes plots=all;
  model Progression=Age Sex BMI BP TC LDL HDL TCH LTG GLU
    / selection=lasso(stop=none choose=cp);
run;
ods graphics off;
```

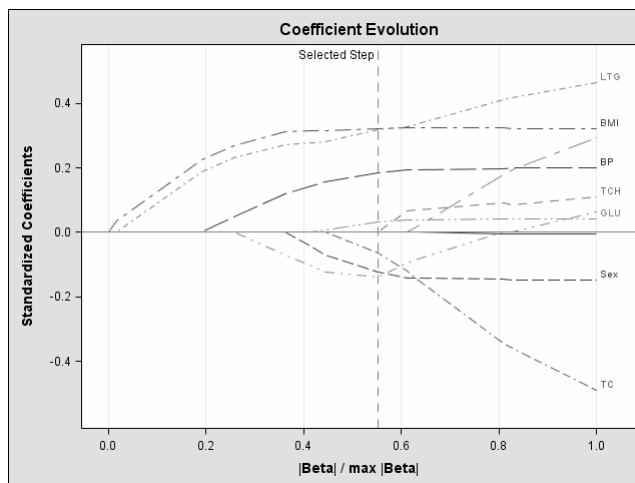
Copyright © 2007 SAS Institute Inc. All rights reserved.

## LASSO Coefficients



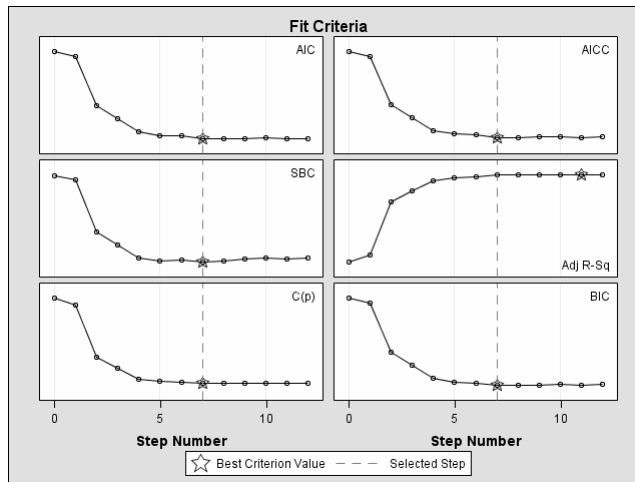
Copyright © 2007 SAS Institute Inc. All rights reserved.

## Lasso Solution Path



Copyright © 2007 SAS Institute Inc. All rights reserved.

## Fit Criteria for LASSO Selection



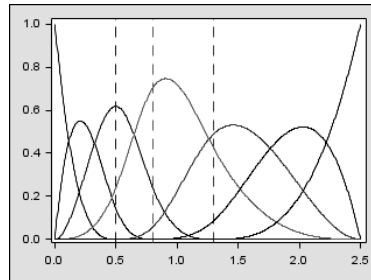
Copyright © 2007 SAS Institute Inc. All rights reserved.

## Nonparametric fitting using variable selection – PROC GLMSELECT

Copyright © 2007 SAS Institute Inc. All rights reserved.

## Cubic B-Spline Basis

- Numerically more stable than TPF Basis
- Each basis function non-zero over 5 adjacent knots
- Can be produced with PROC TRANSREG



```
proc transreg design data=one;  
  model bspline(x/knots=0.5 0.8 1.3);  
  output out=bsplineOut;  
run;
```

Copyright © 2007 SAS Institute Inc. All rights reserved.

## The EFFECT Statement – Release 9.2

- Defines collections of columns for the design matrix that can be referred to with a single label
- Generates the associated design matrix columns
- Can be used in complex effects
- Initial implementation includes regression splines

Copyright © 2007 SAS Institute Inc. All rights reserved.

## Nonparametric Fitting by Selecting Splines

### Idea:

- Generate “lots” of B-Spline basis functions
- Generate basis functions at multiple scales
- Use variable selection to choose an appropriate subset

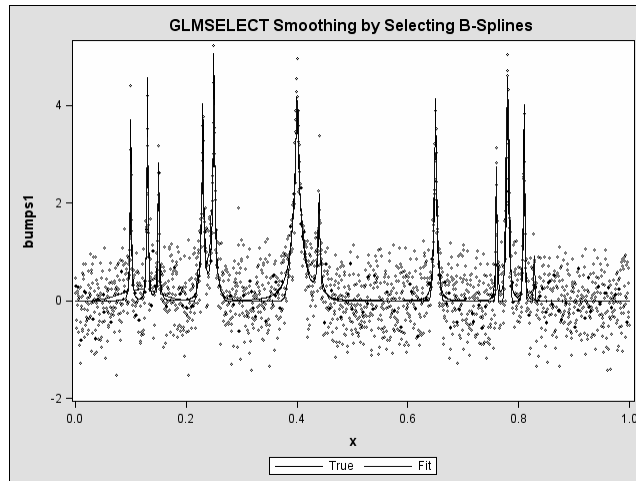
Copyright © 2007 SAS Institute Inc. All rights reserved.

## Scatterplot Smoothing – Release 9.2

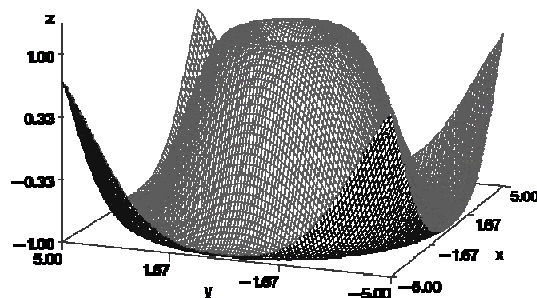
```
proc glmselect data=bumpsData;  
  
  partition fraction(validate=0.5);  
  
  effect spl_x=spline(x/knotmethod=multiscale split);  
  
  model y=spl_x/selection=stepwise(stop=sbc  
                                   choose=validate  
                                   drop=competitive);  
  
  output out=bumpsOut p=p_y;  
run;
```

Copyright © 2007 SAS Institute Inc. All rights reserved.

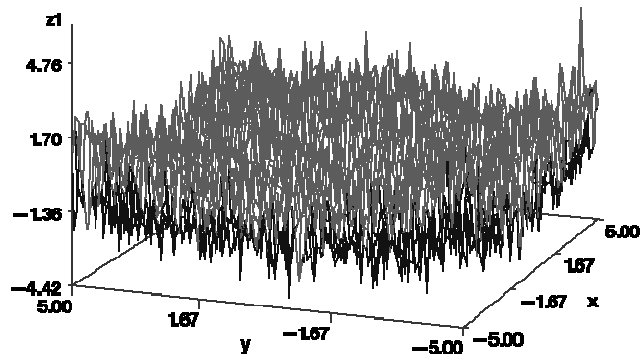
## Smoothing the “bumps” function



## Hat Surface



## Noisy Hat Surface



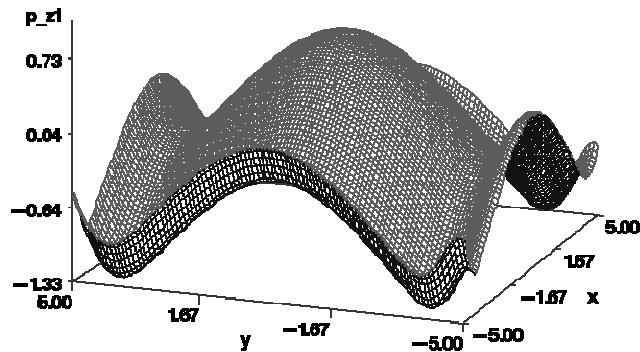
Copyright © 2007 SAS Institute Inc. All rights reserved.

## Fitting an Additive Model with GLMSELECT

```
proc glmselect data=hat;  
  effect spl_x=spline(x  
    /knotmethod=multiscale(endlevel=4) split);  
  effect spl_y = spline(y  
    /knotmethod=multiscale(endlevel=4) split);  
  model z1 = spl_y spl_x;  
  output out=out1;  
run;
```

Copyright © 2007 SAS Institute Inc. All rights reserved.

## Additive Model by Selecting B-Splines



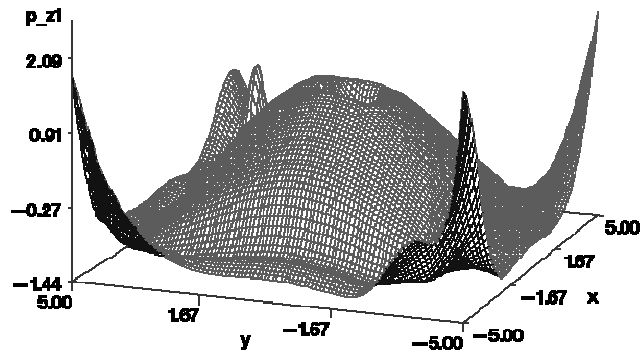
Copyright © 2007 SAS Institute Inc. All rights reserved.

## Surface Fitting with GLMSELECT

```
proc glmselect data=hat seed=1;
  partition fraction(validate=0.5);
  effect spl_x=spline(x
    /knotmethod=multiscale(endlevel=5) split);
  effect spl_y = spline(y
    /knotmethod=multiscale(endlevel=5) split);
  model z1 = spl_y|spl_x
    / selection=stepwise(stop=sbc choose=validate);
  output out=out1;
run;
```

Copyright © 2007 SAS Institute Inc. All rights reserved.

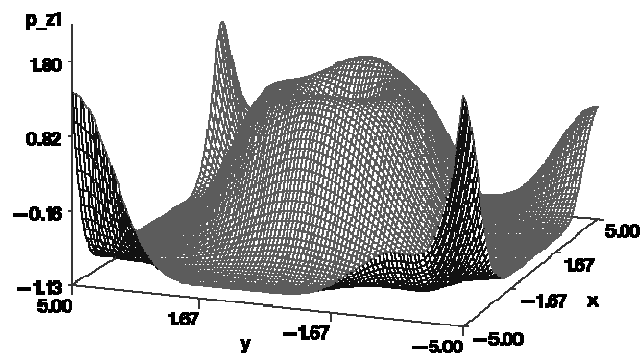
## Selecting from Tensor Products of Splines



Copyright © 2007 SAS Institute Inc. All rights reserved.

## Selecting from Tensor Products of Splines

```
effect spl_x = spline(x/knotmethod=equal(5) split);  
effect spl_y = spline(y/knotmethod=equal(5) split);
```



Copyright © 2007 SAS Institute Inc. All rights reserved.



## Availability

### SAS/STAT Procedures

- The following SAS/STAT procedures can be downloaded for Release 9.1 on Windows (32 bit). They do NOT ship with SAS/STAT
  - PROC GLIMMIX: Production
  - PROC GLMSELECT: Experimental
  - PROC QUANTREG: Experimental
- Download from <http://www.sas.com/statistics>

Copyright © 2007 SAS Institute Inc. All rights reserved.