Measuring Cost Avoidance in the Face of Messy Data

Jorge Romeu, Reliability Analysis Center
Joe Ciccimaro, Naval Inventory Control Point
John Trinkle, Advanced Information Engineering Services

Overview

- Background: The Customer Requirements
- Forecasting:
  - Least Squares Regression
  - Non-Parametric Regression
- Simulation Results
- Sample Cases

Customer Requirement

- Navy senior management
  - Needed to forecast failure rates for near-term and long-term planning.
    - Decision points for action to improve spares situation for Naval Air Craft
  - Forecast needed to be in simplest format
  - Forecast was basis for monetary cost avoidance.
    - Forecast became the *do nothing* model
    - Cost avoidance was based on difference between the actual and forecast post reliability implementation.
  - Lead time for analysis was short
  - Analysts with cursory knowledge of forecasting methods.

Forecasting Methods

- The agreed upon reliability measure was in the format of **Failures/(1000 Flight Hours)**
- Initial forecasting method was **Least Squares Linear Regression**
  - Occams Razor: “Use simplest model”
  - Tools are commonly available
  - Problems:
    - Data was often limited to small samples
    - Data showed non-constant variation
    - Data had large outliers

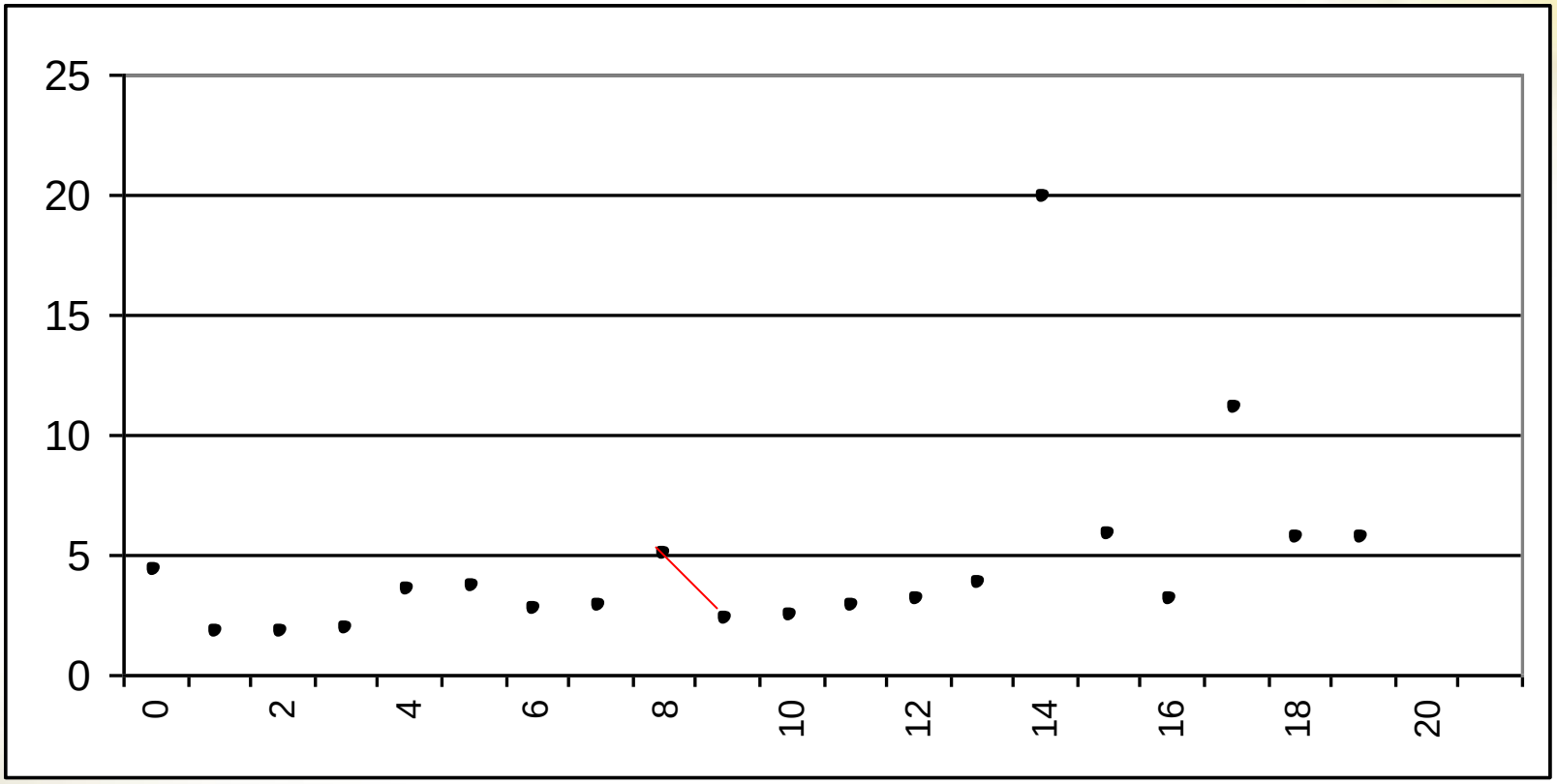
Forecasting Methods

- Examined over 100 data sets
  - Observations:
    - Residual analysis revealed heteroskedasticity
    - Occurrence of outliers
      - Many related to surge in military operations
    - Valid data sets could be small

Forecasting Methods

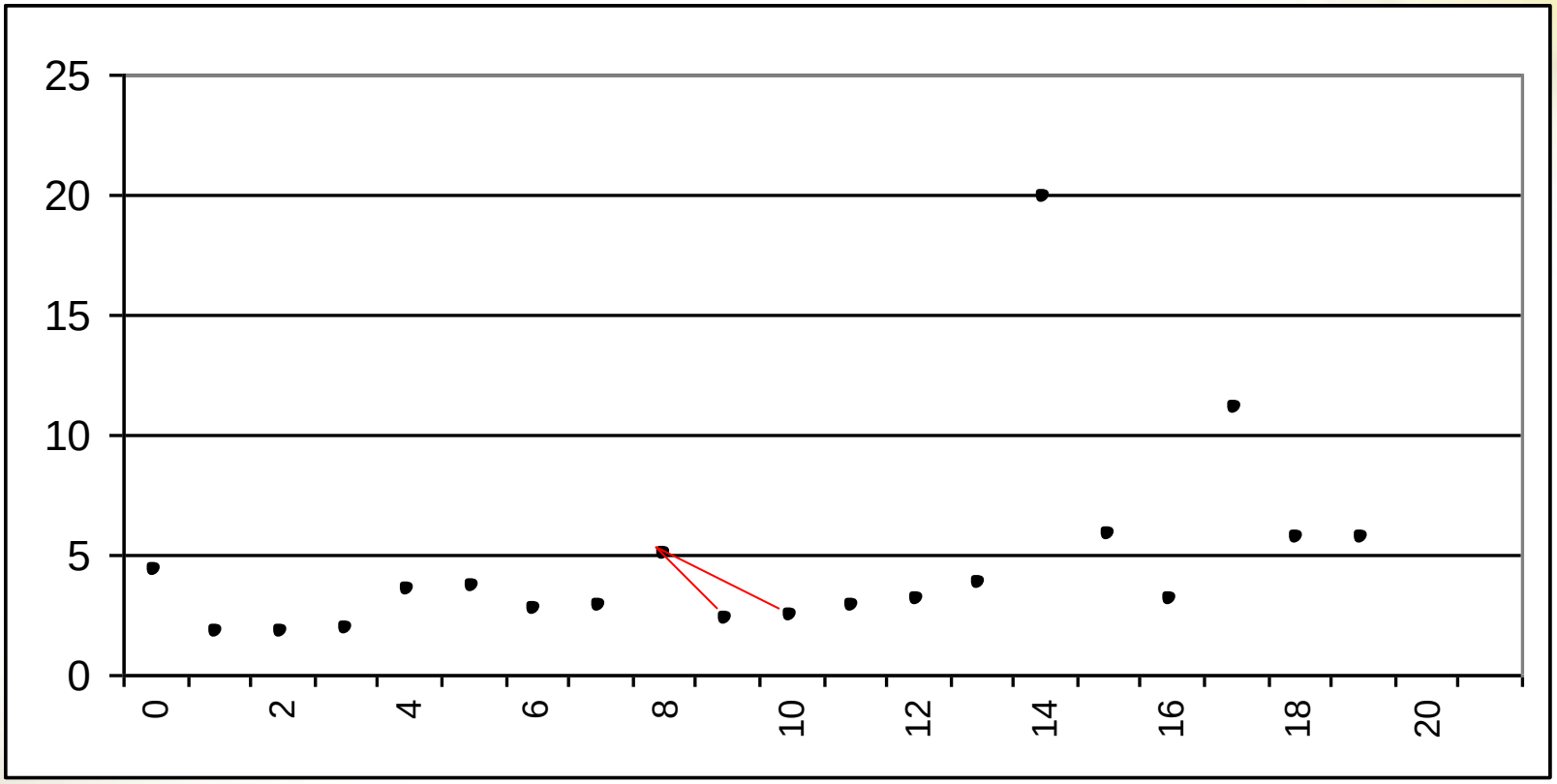
- Non-Parametric Linear Modeling
  - Because it is based on the median of pair-wise slopes
    - Less sensitive to skewed data
    - Less sensitive to outliers
    - Less sensitive to small data sets
  - Ranking of slopes moves disconcordant local-slopes to limits of ranks.
    - Effects from outliers and skewed residuals is reduced.

Pair-Wise Slopes



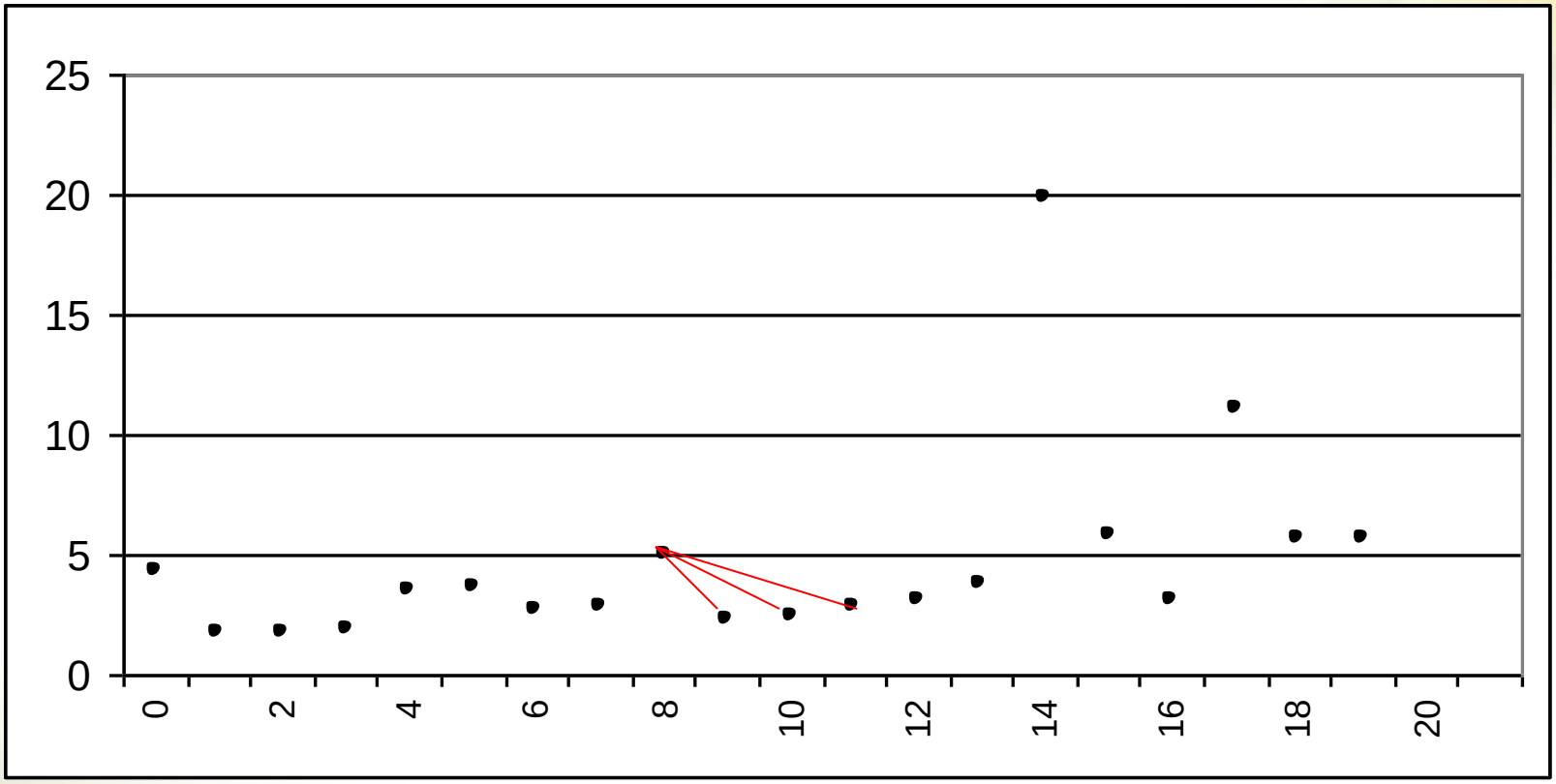
Shows pair-wise slopes between n=9 and n= 10, 11, 12…16

Pair-Wise Slopes



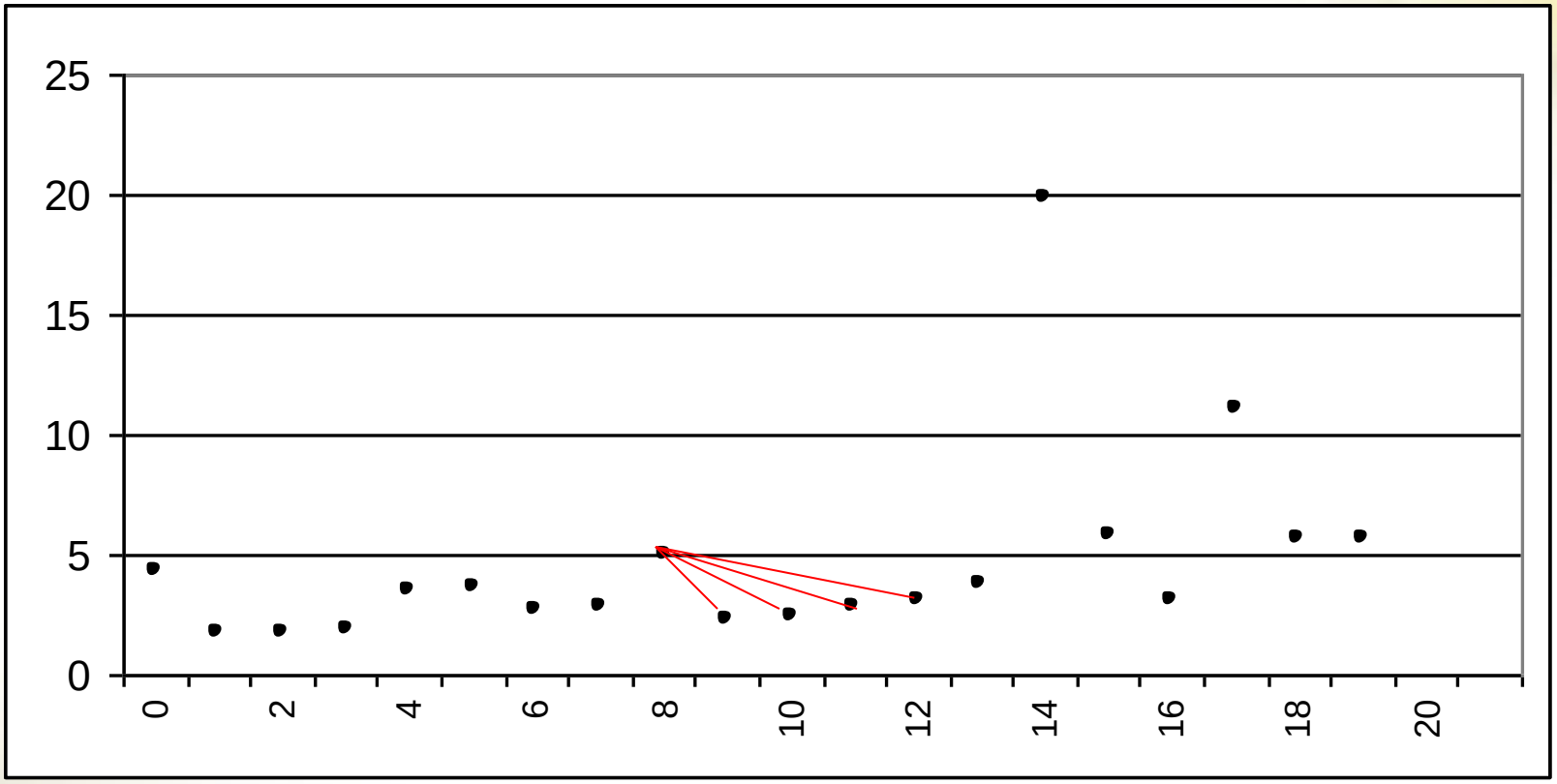
Shows pair-wise slopes between n=9 and n= 10, 11, 12…16

Pair-Wise Slopes



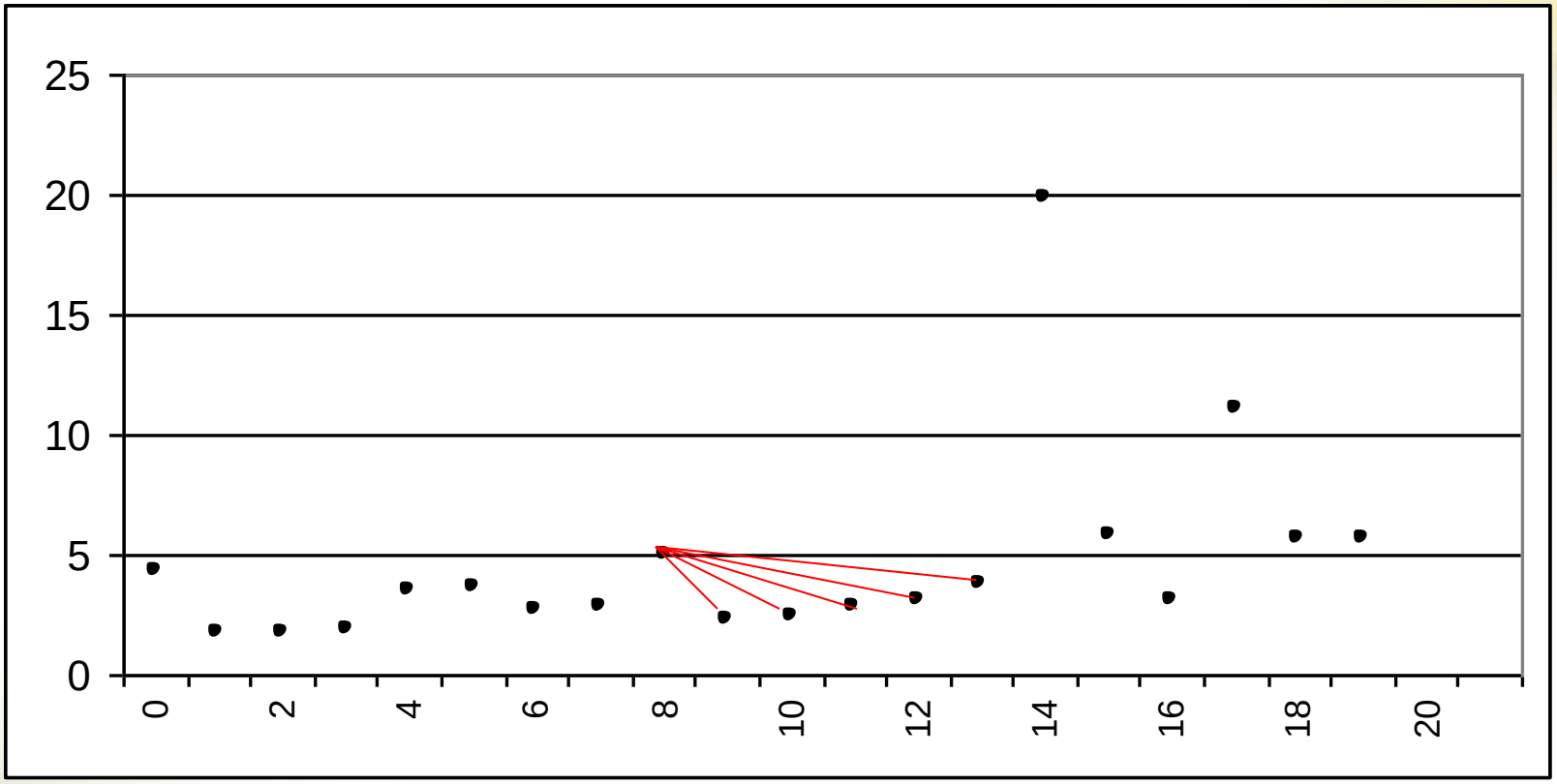
Shows pair-wise slopes between n=9 and n= 10, 11, 12…16

Pair-Wise Slopes



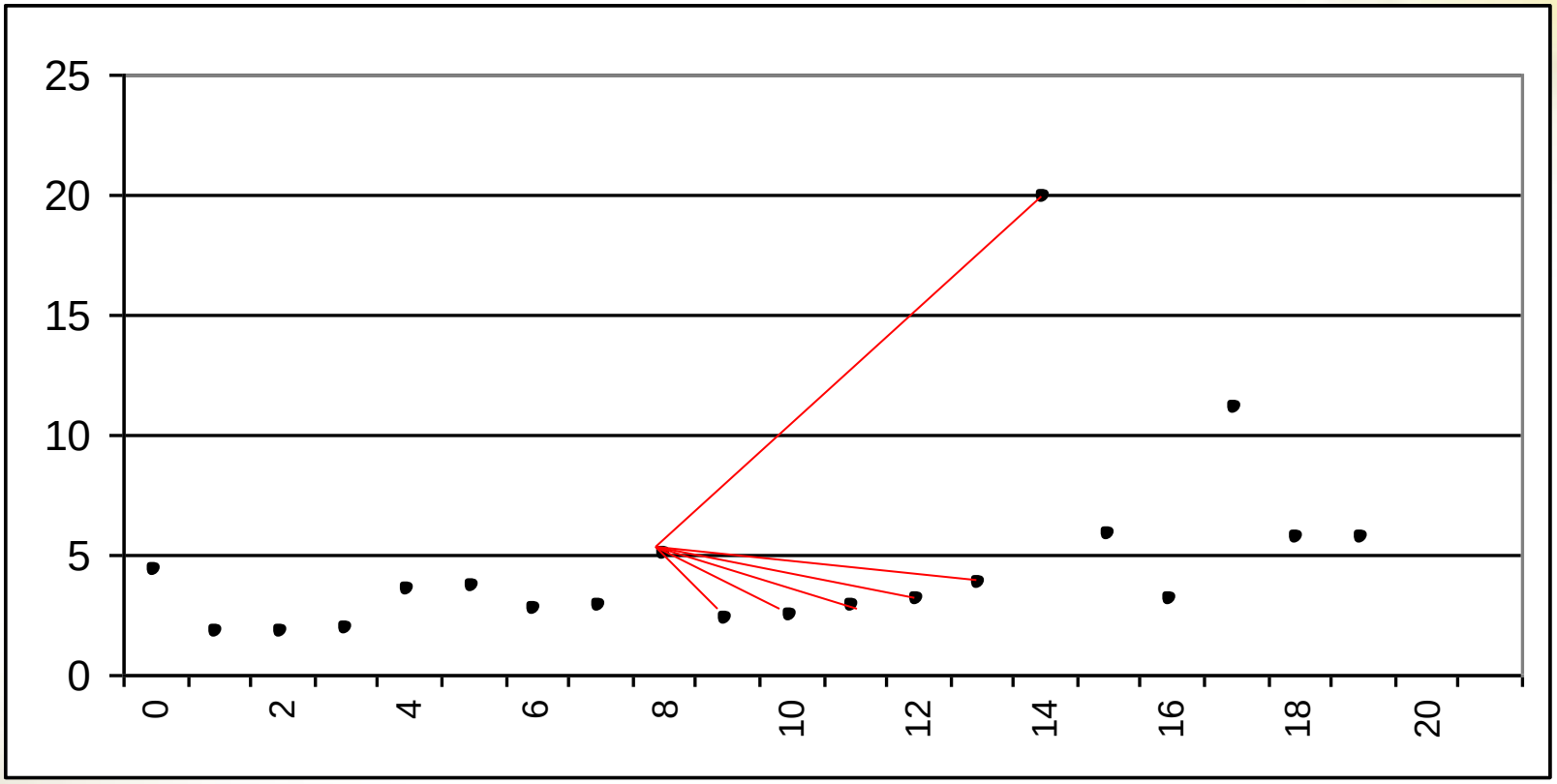
Shows pair-wise slopes between n=9 and n= 10, 11, 12…16

Pair-Wise Slopes



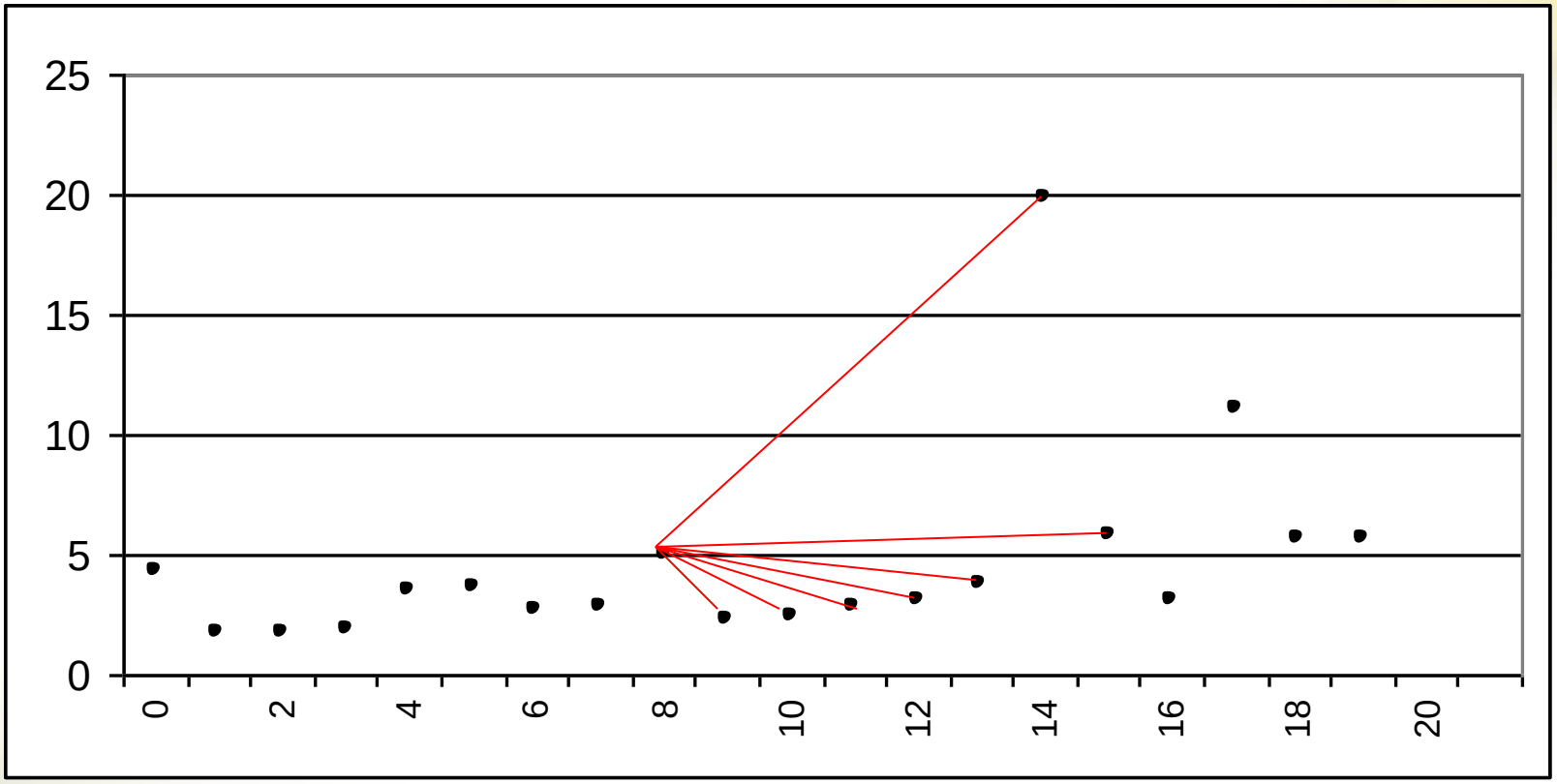
Shows pair-wise slopes between n=9 and n= 10, 11, 12…16

Pair-Wise Slopes



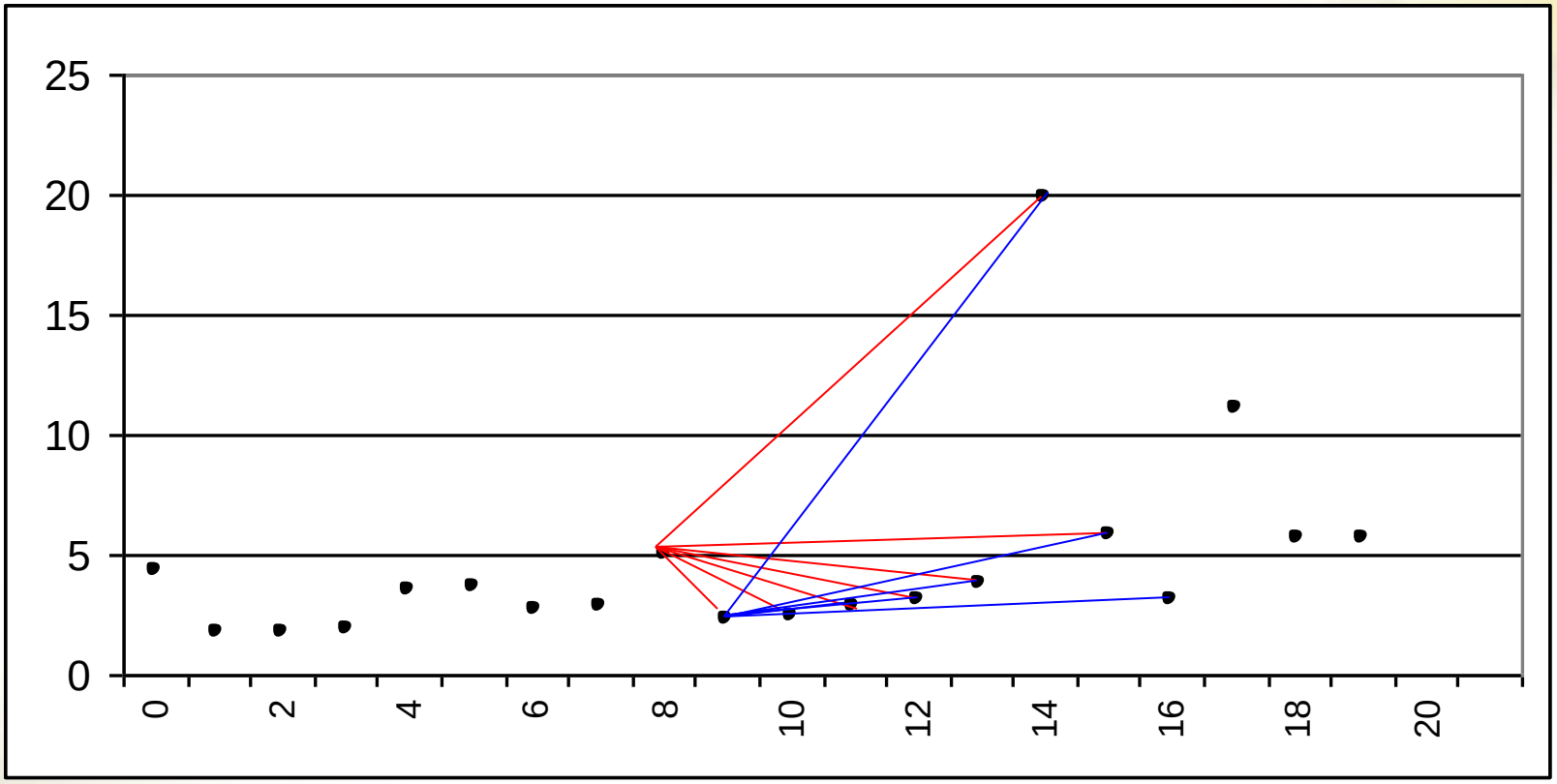
Shows pair-wise slopes between n=9 and n= 10, 11, 12…16

Pair-Wise Slopes



Shows pair-wise slopes between n=9 and n= 10, 11, 12…16

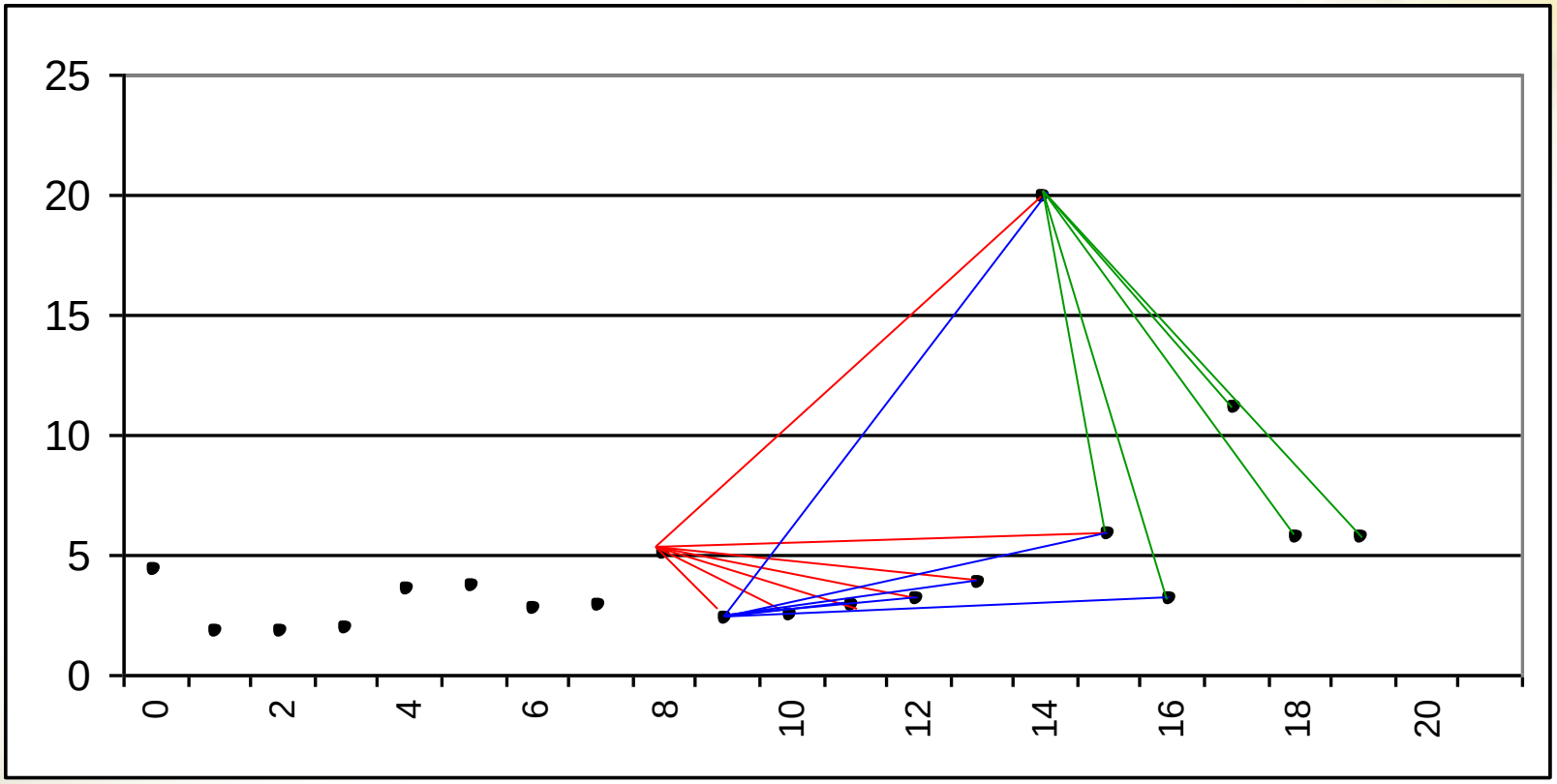
Pair-Wise Slopes



Shows pair-wise slopes between n=9 and n= 10, 11, 12…16

And between n=10 and n = 11, 12, 13…17.

Pair-Wise Slopes



Shows pair-wise slopes between n=9 and n= 10, 11, 12…16

And between n=10 and n = 11, 12, 13…17.

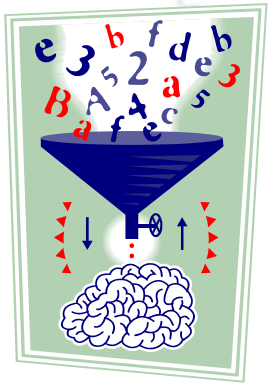
And between n= 15 and n = 16, 17, 18…20

Non-Parametric Method

- Pair-wise slopes are ranked from minimum to maximum.
- The extreme values are at the ends of the sorted slope set.
  - $Slope_{min}$ is the most negative slope and $Slope_{max}$ is the most positive slope.
  - Result from outliers and data in the long tails of the distribution.

Ordered Slopes

$slope_{1,2}$
$slope_{1,3}$
$slope_{1,4}$
.
.
.
.
.
.
.
.
.
$slope_{n-1,n}$

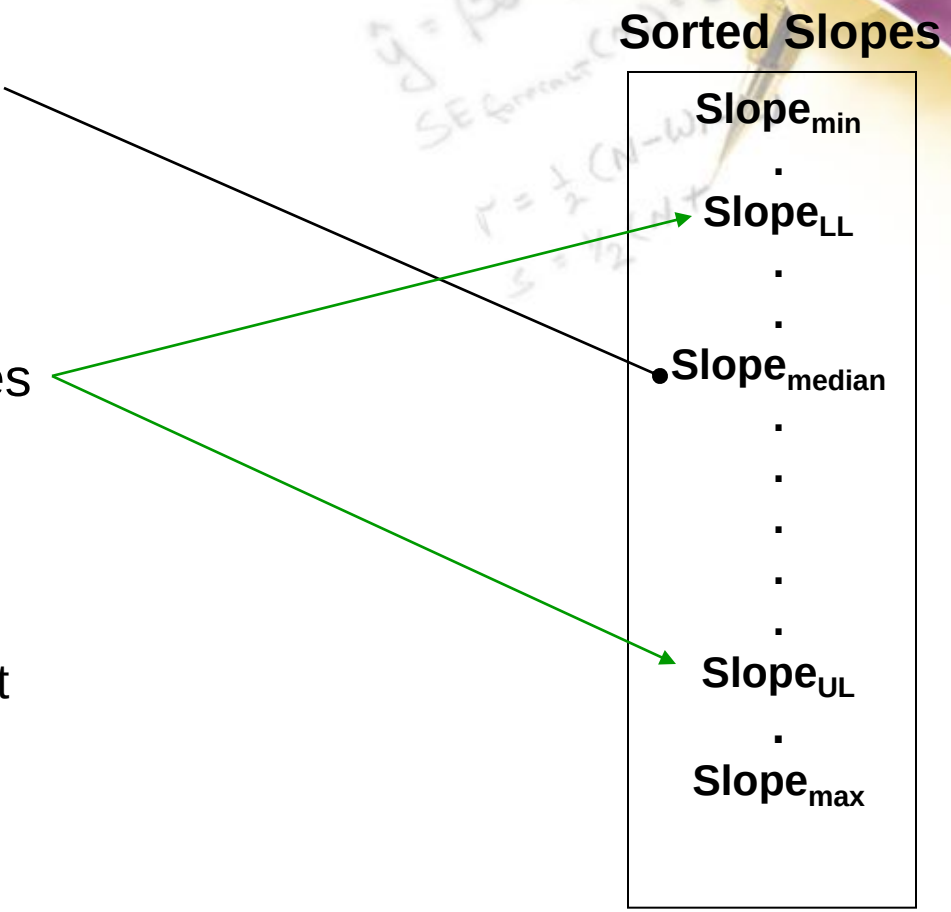


Sorted Slopes

$Slope_{min}$
.
.
.
.
$Slope_{median}$
.
.
.
.
.
.
$Slope_{max}$

Non-Parametric Method

- Median slope is the nominal slope.
- Slope confidence limits are determined from the quantiles of the Kendall-Tau test statistic.
- Slope Test: H_0: Slope = 0
  - If Slope = 0 is between the $Slope_{LL}$ and $Slope_{UL}$, cannot reject the hypothesis.
  - Otherwise, reject the hypothesis and accept the median slope.



Simulation

- Ten data set were simulated
  - Base model: $Y_{sim} = 1 + 0.1t + \mathcal{N}_{sim}$
    - $\mathcal{N}_{sim}$ = the error term
      - Randomly generated based on exponential distribution
      - Increases with dependent variable
      - Increased with each data set to simulate greater degrees of skewed variability
- Applied Least Squares Regression and Non-parametric regression to ten sets.

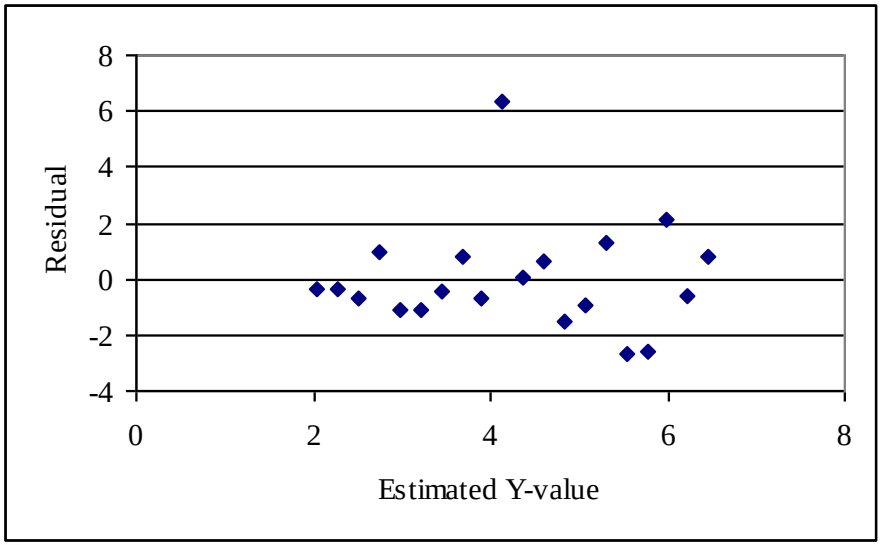
Simulation Results

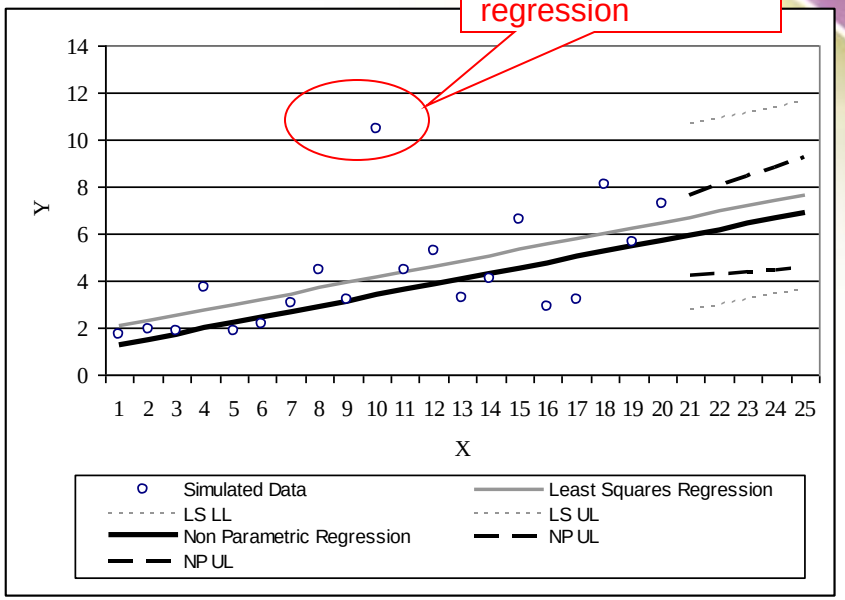
Table 1. Simulation Results

	Least Squares Regression			Non-Parametric Regression			
Simulation Set	**Slope**	**R^2**	**p-value**	**Slope**	**Confidence Level**	**Slope Lower Limit**	**Slope Upper Limit**
1	0.232	33.9%	0.007	0.234	99%	0.073	0.397
2	0.203	17.5%	0.066	0.203	99%	0.014	0.397
3	0.175	27.4%	0.018	0.124	95%	0.020	0.256
4	0.116	8.4%	0.215	0.116	95%	0.034	0.232
5	0.166	21.9%	0.037	0.167	95%	0.045	0.320
6	0.231	38.2%	0.004	0.191	99%	0.069	0.395
7	0.302	27.2%	0.018	0.229	99%	0.026	0.619
8	0.101	16.3%	0.077	0.119	95%	0.007	0.251
9	0.230	19.0%	0.055	0.161	95%	0.024	0.326
10	0.318	20.4%	0.046	0.169	99%	0.037	0.397

Note: n = 20 for all simulated data sets.

Simulation Results



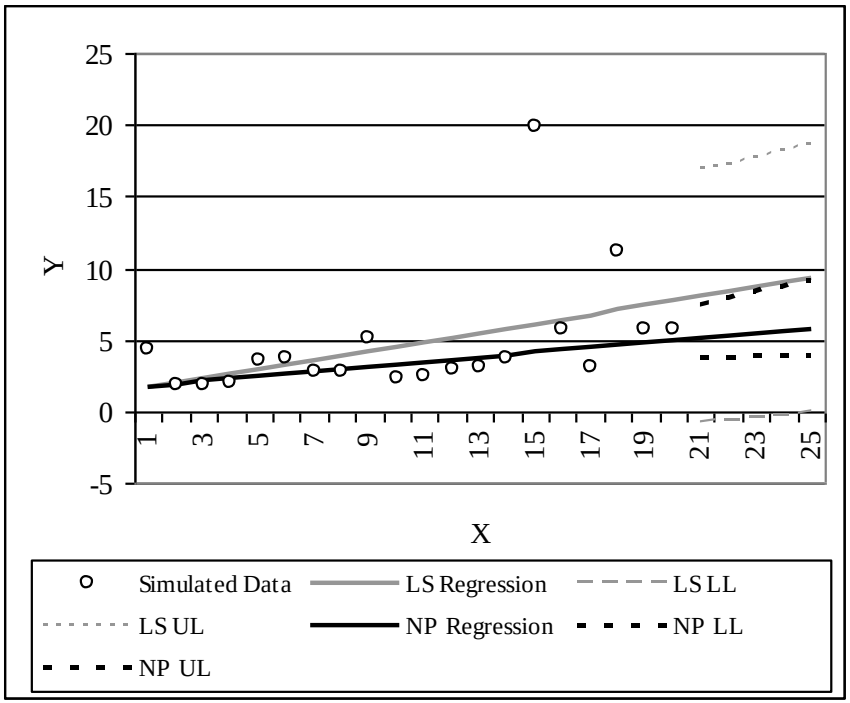


The residual plot for simulation set 1, shows a non-constant variance, typical of all the simulations set.

Least Squares p-value indicated a good slope value for this set.

When modeling methods are compared, Least squares will result in an over estimate, because the outlier is pulling the slope up.

Simulation Results

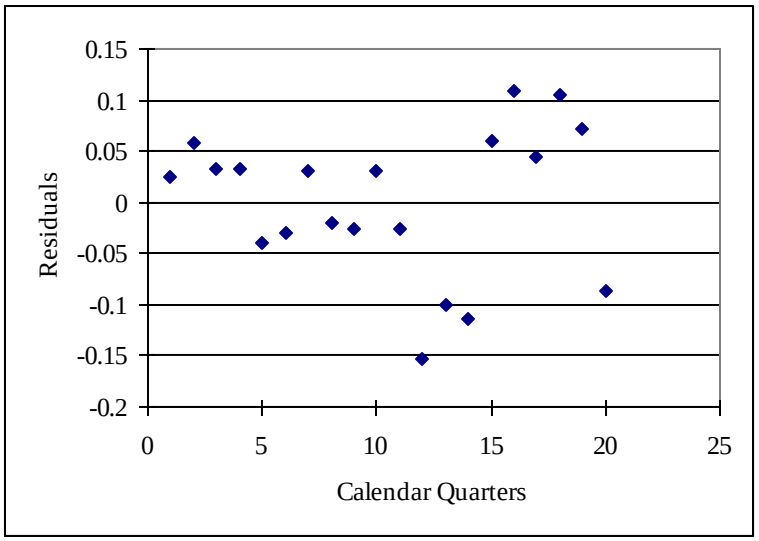


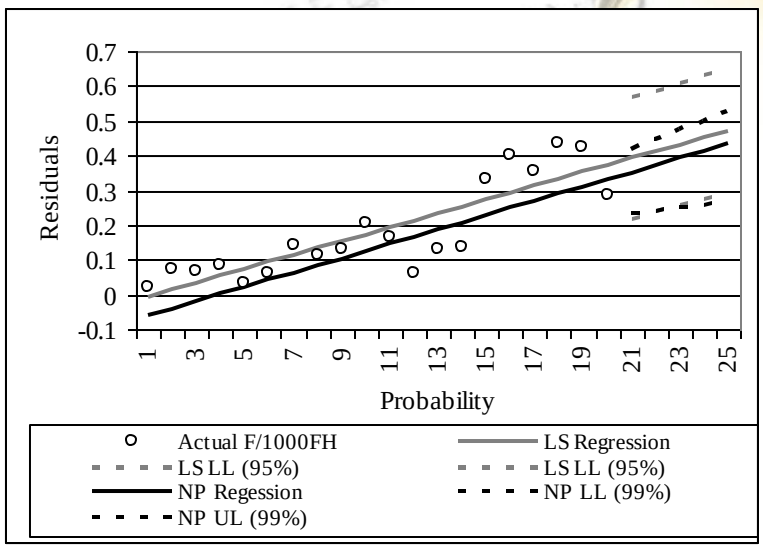
- Simulation Set 10
  - Least Squares Model
    - Pulled by outliers
    - Resulting in an over estimate in the forecast
  - Non-Parametric Model
    - More conservative slope
    - Accounts for skewed data.

Real World Examples

- Navy Aircraft Servo-Cylinder Device
- Experiencing increasing failure rates
- Data showed high variance
- Least Squares tended to over estimate forecast Failures/1000FH.
  - Non-constant variance
- Non-Parametric model
  - Had a higher rate of change, positive
  - But regression line shifted down

Real World Results: Navy Aircraft Servo-Cylinder





- Least Squares residuals showed non-constant variance.
- Violation of assumptions

- Non-Parametric model
  - Higher slope
  - Lower intercept
  - Closer confidence limits.

Comparison

- Least Squares Methods
  - Readily available in common software tools

- Non-Parametric
  - Less readily available
    - Must write your own programs or formulas
    - Must purchase a more sophisticated software package.

Conclusions

- Non-Parametric Methods
  - Less sensitive to high variability
  - Less sensitive to outliers
  - Less sensitive to small data sets
- Least Squares Methods
  - More sensitive to high variability
  - Very sensitive to small data sets
  - Must meet underlying assumptions
    - Often ignored by analysts
    - Can be subjective