Measuring Cost Avoidance in the Face of Messy Data

Jorge Romeu, PhD., Reliability Analysis Center, Rome
Joseph Ciccimaro, MA Mathematics, Naval Inventory Control Point, Philadelphia
John Trinkle, MBA, Advanced Information Engineering Services, Inc. , San Diego

Key Words: regression, non-parametric, confidence limits, heteroskedasticity, forecast

Summary and Conclusions

This paper presents alternative methods to forecast or predict failure trends when the data violates the assumptions associated with least squares linear regression. Simulations based on actual case studies validated that least squares linear regression may provide a biased model in the presence of messy data. Non-parametric regression methods provide robust forecasting models less sensitive to non-constant variability, outliers, and small data sets.

1. Introduction

In naval aviation, inventory management of systems, subsystems and components is an ever-challenging task. Due to the increased operational requirements, we are asking for better performance from our aircraft and the related components. In addition, aircraft and related systems must perform beyond their planned life cycles. Ensuring a sufficient supply of systems and components requires inventory managers, engineers, and logisticians to maintain an awareness of the system and component reliability.

Due to these demanding challenges the Naval Air Systems Command (NAVAIR) in partnership with the Naval Inventory Control Point (NAVICP) is continually seeking ways to improve the reliability of supported systems. These improvements are designed to increase operational readiness and reduce Total Ownership Cost (TOC). One such program is the Integrated In-Service Reliability Program (IISRP). Under this program the Naval Aviation Depots (NADEPs) analyze all relevant areas of engineering design, supply support, maintenance, repair and modification processes in an effort to identify potential reliability improvement opportunities. The NADEP engineers review the reliability metrics of selected Aviation Depot Level Repairable (AVDLR) components to identify poor performers. Beyond Capability of Maintenance Rate (BCMs/1000FH), is a measure that embodies the generally accepted premise that aircraft component failure rates are related to flight hours.
A change that is designed to improve a component's reliability is initiated and the resulting BCM rate is noted. This rate is then compared to the predicted BCM rate that did not consider the change. The difference between these two values is multiplied by the actual number of flight hours flown in the quarter, the components unit price, AVDLR, yielding the cost avoidance and, therefore, the value of the change in eq (1).

$$CA = (BCM/1000FH_{Proj} - BCM/1000FH_{Act})(FH_{Act})(AVDLR_{Act}) \quad (1)$$

Where,
CA = Cost Avoidance in dollars

The least squares linear regression model became the forecasting method of choice for the IISRP. However, evaluation of the forecast's validity revealed that this method could not apply to many of the data sets. The underlying data violated the least square regression assumptions of normality and homoskedasticity of the residuals. It was soon realized that the problem was in the difficulty of detecting failure trends in the face of very messy data. Messy data is defined as, data that does not fulfill the basic assumptions of parametric linear regression.

One proposed solution is the use of the non-parametric regression model. Non-parametric methods are well documented in statistical and operation research literature. It is a more robust method that has a lower sensitivity to data variance. In the following sections we discuss the effects of poorly behaved data sets on least squares linear regression followed by a comparison to non-parametric pair-wise regression. A real-world example is used to compare the results and show the difficulty associated with least squares regression in the presence of messy data.

2. Least Squares Approach In Predicting Future Failure Rates

The IISRP utilized the least squares linear regression model to predict future BCM Rates (BCMs/1000FH). Both the number of BCMs in a yearly quarter and associated operating hours are collected from various data sources. From this data, a quarterly BCM rate is calculated and the twenty most recent quarters of data is used in developing a linear regression line used to predict BCM rates. Although the initial analysis suggests using the least squares regression model, further investigations indicated that at least one of the required conditions of the random error variable ε was not satisfied. The conditions for random error variable ε are as follows:

1. The probability distribution of ε is normal with the mean of the distribution being zero; that is, $E(\varepsilon) = 0$.
2. The standard deviation of ε is σ_ε, which is constant no matter what the value of x is. In this study, time is the independent variable.
3. The errors are independent.

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.

Data was simulated to demonstrate the violations. A set of data with increasingly larger exponential variance (related to the quarter) was created using the following formula.

$$\text{Response} = 1 + 0.1 \cdot t + \varepsilon_{sim} \quad (2)$$

Where,
Response = the simulated datum
t= calendar quarters, 1, 2, …20
ε_{sim} = exponential increasing error, simulated

Selection of eq (2) was based on similarity to data sets observed by analysts supporting navy aviation reliability projects.

Refer to Table 1 for the results from the ten simulated data sets. For example, simulation set 1 least squares regression analysis indicates a valid regression model based on a p-value < 0.05, p-value = 0.007. However, upon examining the required assumptions it becomes apparent that the models are not adequate for forecasting.

Figures 1 and 2 show the results from simulation set 1. From the graphical analysis and by simulation model design:
1. The residuals versus the fitted plot show a 'funnel-like' form, this is a characteristic of variance that is related to the mean. This violates assumption number three indicating bias in the model that will result in fatal forecasting errors.
b. The normality plot of the residuals demonstrates a right skewed distribution that violates assumption number one. This violation has fatal impact as the resulting model is biased, and cannot be used for hypothesis testing.

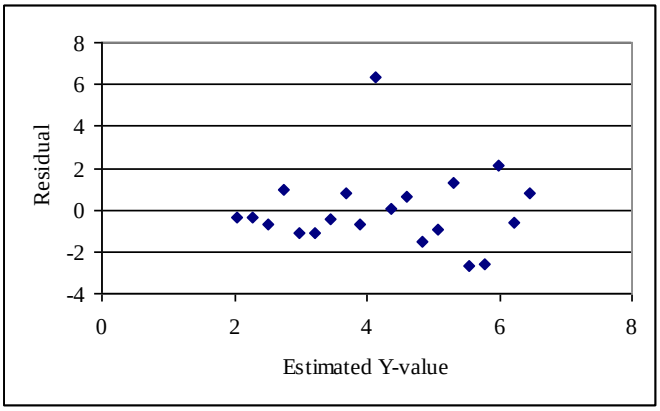

Figure 1. Simulation Set 1 Residual Plot

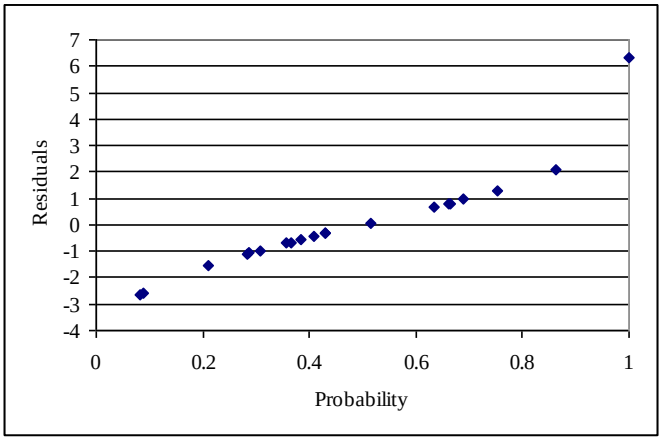

Figure 2. Simulation Set 1 Probability Plot

3. Non-Parametric Regression

Because parametric regression is dependent on the behavior and volume of data, inexperienced analyst can be challenged to determine if a model is unbiased and valid. Non-parametric regression provides a means to model a simple linear equation without consideration for the issues surrounding residual variability and normality, and for issues concerning small sample sizes. Specifically in this paper we refer to pair-wise slope regression that is simple to perform and confidence limits are easily determined using lookup tables. More importantly the method is robust and does not require additional analysis to validate once the analyst determines the slope, intercept and confidence limits.

Using the same simulated data sets from this paper's least squares regression section, we applied the pair-wise slope non-parametric regression. Refer to Table 1 for results. In this procedure for n data points we calculate slope between each (x_i, y_i) data pair and the (x_{i+m}, y_{i+m}) data pair, where $m = (i+1)$ *to* n. The median value of these slopes represents the slope value in the simple regression model. In a manner similar to least squares regression where the model will pass through the point $(\bar{x}, \bar{y})$, the non-parametric model will pass through the point (x_{median}, y_{median}). Thus, the intercept is determined by eq (3):

$$\text{Intercept} = y_{\text{median}} - \text{Slope} \cdot x_{\text{median}} \quad (3)$$

Once we developed the model, we test the slope. Confidence bounds derived from Kendall's Tau for n data samples and $\alpha/2$ confidence level. The test is based on the hypothesis that the slope is equal to zero, $H_0 : \text{Slope} = 0$ and $H_1 : \text{Slope} \neq 0$. If a zero value is between the slope upper confidence bound and the slope lower confidence bounds then we cannot reject the hypothesis. However if the slope is between the upper and lower confidence bounds and zero is not between those confidence bounds the slope is considered valid. Table 2 summarizes the test conditions.

Table 2. Non-Parametric Slope Tests

Upper Bound Sign	Lower Bound Sign	Action
Negative	Negative	Reject H_0; Accept slope
Positive	Negative	Fail to Reject H_0; Reject Slope, when slope is +
Negative	Positive	Fail to Reject H_0; Reject Slope, when slope is -
Positive	Positive	Reject H_0; Accept Slope

In practical use the non-parametric regression has greatest value in the confidence bounds for forecasting a range of future values. Point estimates have lesser value since there is no confidence bound around the point (x_{median}, y_{median}), all slopes (upper and lower confidence bound slopes, and the median slope) pass through this point.

Applying non-parametric methods to the simulated data set we arrive at a different result than the least square regression methods. The data set showed a higher degree of variability, residuals that varied with the independent variable, and some non-linearity. It is appropriate to question the validity of the model under these conditions and residual analysis confirms the uncertainty.

For simulation set 1, Figure 3 shows the resulting linear regression line with confidence bounds for the forecast region only. We display the confidence bounds in this manner because they only have value under these circumstances for forecasting. The IISRP team does not normally perform point estimating in the sample data range. However, using the non-parametric methods described above a point estimate is not appropriate for extrapolation beyond the data set. A more appropriate method is to use the upper and lower confidence bounds to determine a range of possible values at a specific x-value.

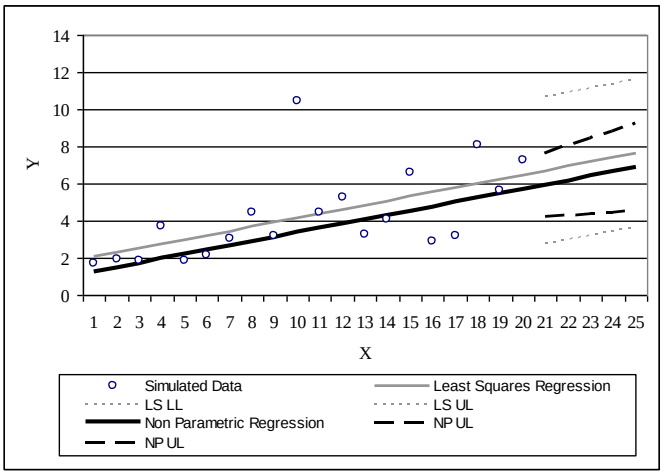

Figure 3. Simulation Set 1 Comparison of Least Squares Regression to Non-Parametric Regression, 20 Simulated Datum with 5 Forecast Datum

Note in Figure 3 that the confidence interval for the linear regression follows the usual hyperbolic curve indicating an increasing confidence interval as the forecast moves further from $\bar{x}$. Likewise the non-parametric regression confidence interval will increase as the forecast moves further from x_{median}.

The linear regression interval is much wider than the non-parametric interval. As previously stated the standard error of forecast is very dependent on the number of samples and the variability of the data. In the non-parametric case, the interval is mostly influenced by the sample size. The estimated slope is the median slope giving emphasis to the local effects and dampening the global effects of data variability. Therefore the non-parametric slopes are not as effected by the outliers as the larger slopes appear at the end of the ranked local slopes. Depending on the confidence interval selected, these large slope deviations are unlikely to appear. However if the

number of samples is very small then the large slope values are more likely to appear.

In Figure 4, simulation set 10 show a more dramatic difference between the non-parametric results and the least squares result. Because the distribution of residuals is right skewed the least squares model slope is pulled towards the outliers. The least squares model is biased despite a good p-value at 95% confidence level. The non-parametric slope is much closer to the expected value of 0.10 as defined by the simulation model.

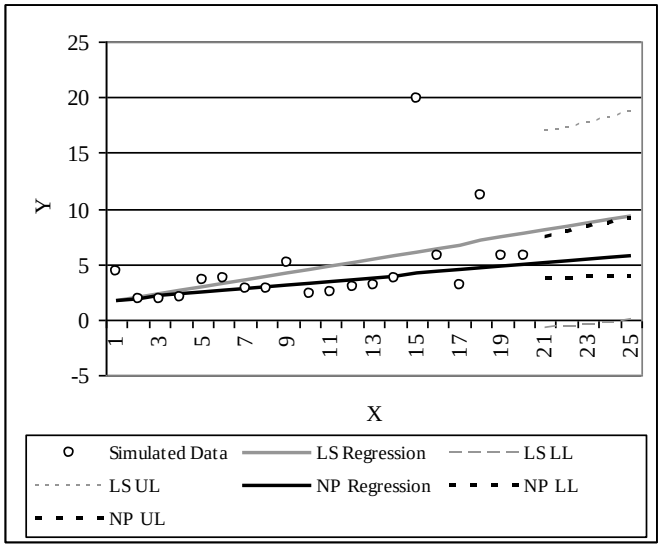

Figure 4. Simulation Set 10 Comparison of Least Squares Regression to Non-Parametric Regression, 20 Simulated Datum with 5 Forecast Datum

4. Aircraft Data Example

This example uses disguised data for a servo device on a Navy aircraft. Information concerning future demand for servo repairs or replacements was needed. Lacking any reliability data, the analyst used failures per 1000 flight hours (F/1000FH) from in-service maintenance databases as a general measure of the servo failure trends. Figure 4 shows the data and a least squares regression line. In addition the chart shows the confidence interval for a five-quarter forecast using the least squares regression model.

Figure 5 shows an increasing rate of failures. However, when we examined the residuals from the regression model we found that there was non-constant variability. Figure 6 shows the residuals spreading as the x variable, Calendar Quarters, increases. However the p-value << 0.05 indicates a good fit of the model to the data..

Figure 7 shows the non-parametric regression with 99% confidence limits. In Figure 7 the least squares regression is shown in grey for comparison. Note that despite the apparent good fit of the least squares regression, any forecasts using this model would be biased and faulty, greatly underestimating the failure rate. The non-parametric model indicates a much higher rate of failure and a narrower confidence interval. Therefore, at the 25th quarter the analyst would estimate between 4.5 and 9.2 F/1000FH at a 99% confidence interval

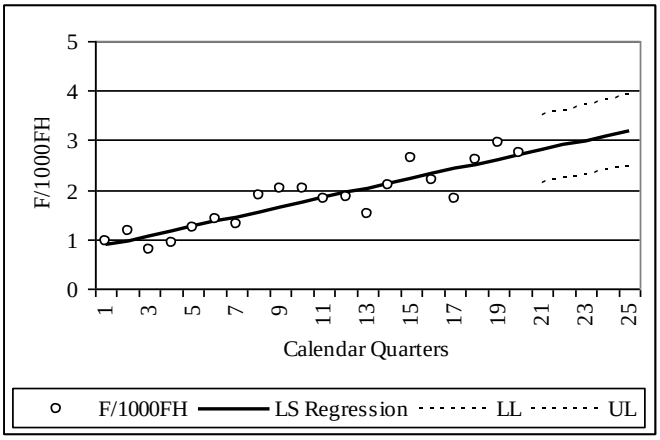

Figure 5. Servo Failures with Least Squares Regression and 95% Confidence Limits for Forecast.

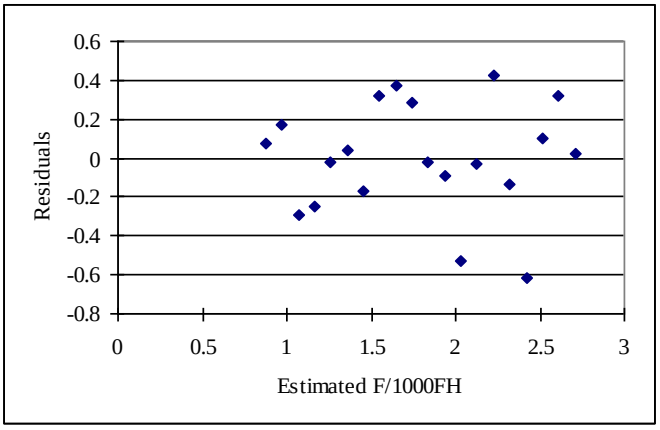

Figure 6. Servo Residual Plot From Least Squares Linear Regression.

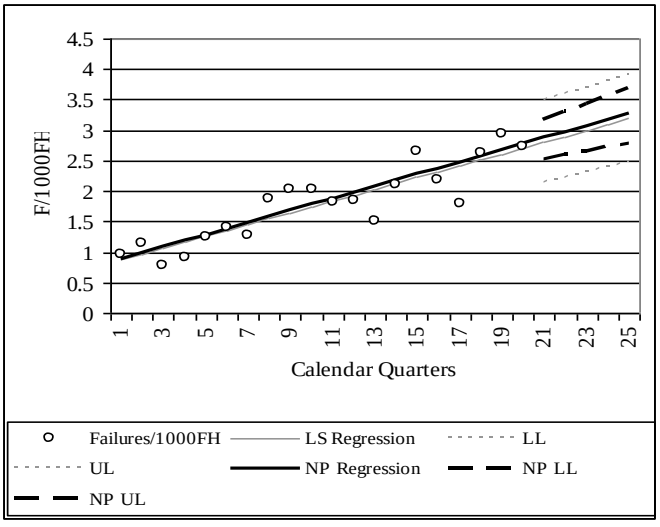

Figure 7. Servo Failures with Non-Parametric Regression and 99% Confidence Limits

A second example is also an aircraft component—a servo-cylinder that activates the aileron. As in the previous example the analyst is forecasting F/1000FH from 20 calendar-quarters of aircraft history. Figure 8 is the normal probability plot for the independent variable, F/1000FH. This plot shows that the data is not normally distributed being skewed to the right.

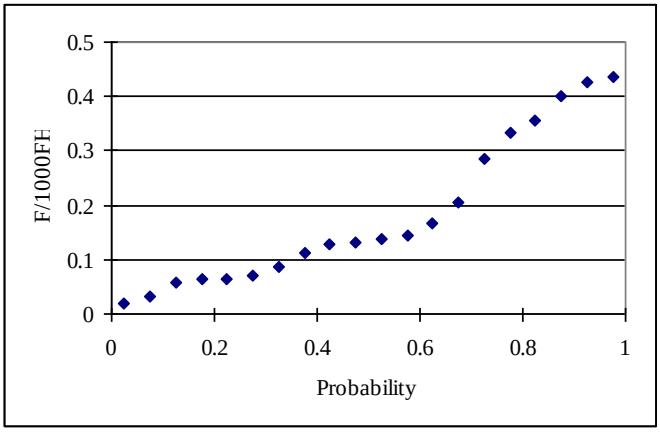

Figure 8. Servo-Cylinder Normal Probability Plot

The least squares regression model gives a high R2, 0.7249, and a low p-value, 0.000002, that indicates a very good fit to the resulting model. However examination of the residuals, as shown in Figure 9, tells us that the underlying assumption of constant variance is violated.

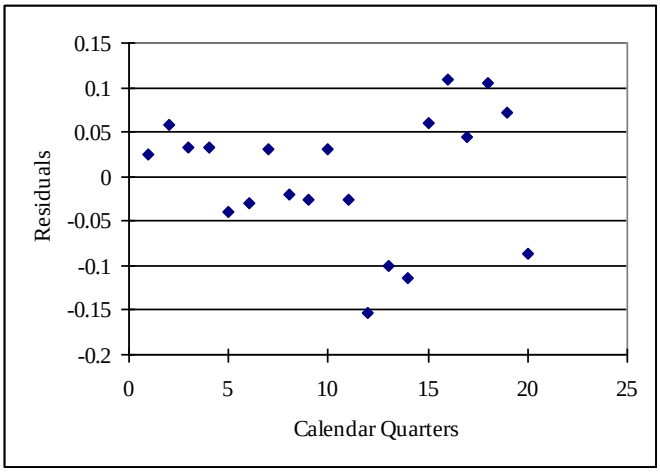

Figure 9. Servo-Cylinder Residual Plot from Least Squares Linear Regression Model

Figure 10 shows the results of the non-parametric regression compared to the least squares linear regression for the servo-cylinder F/1000FH data. The non-parametric slope is higher than the least squares regression however the intercept, which is based on the median of the F/100FH will shift the non-parametric model below the least squares model. The non-parametric model accounts for the skewed, non-normal distribution of the y-data.

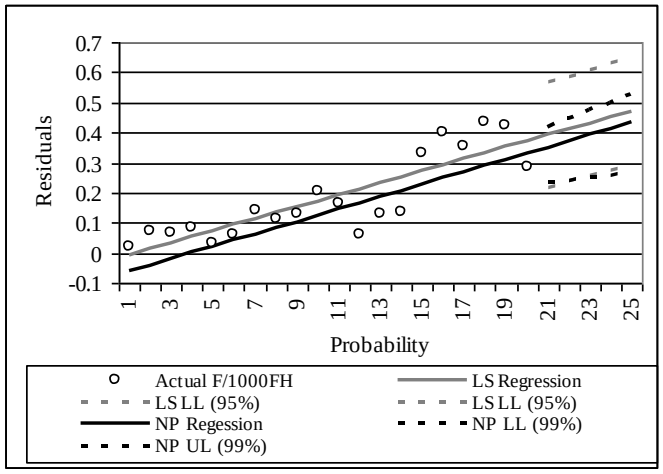

Figure 10. Servo-Cylinder Comparison of Least Squares Regression to Non-Parametric Regression, with Five Forecast Datum

It is important to note that the interpretation of confidence limits for non-parametric regression is different than for least squares regression. The non-parametric regression confidence limits mean that there is a 90% confidence, for example, that repeated measures of the data will result in a slope between the confidence limits. As compared to the least squares interpretation that the true forecast value lies between the confidence limits. The analyst is cautioned to be aware of the difference.

5. Conclusion

The non-parametric and the least squares regression methods provide analysts a method to predict trends based on historical data. The least squares model is greatly affected by the variability of the data. The least squares model minimizes the distance between the observed values and the predicted values, error. This characteristic also constrains the model to underlying assumptions that, if violated, can result in model bias and error. Data with outliers and high variability will have a large impact on the model. The non-parametric regression is less sensitive to variability in the data. Outliers have less influence on the model. It is the weighting of the pair-wise slopes that makes the non-parametric method less sensitive to slight non-linearity in the data, and therefore more robust.

Examination of over 100 data sets used by the Navy IISRP to forecast reliability trends, showed that the data suffers from a number of problematic issues: 1) high degrees of variability, 2) outlying data, and 3) small data sets. The IISRP is also challenged by the need to present long-term forecasts (five to ten years), and the need to present trends in the simplest possible manner. Time to conduct detailed analysis is at a premium and is reserved for a few high profile cases. Given these circumstances and challenges, the non-parametric methods offer a reasonable, defensible, and robust means to forecast aircraft component reliability trends.

5. *Future Work*

The non-parametric methods provides analysts with a valid and robust method to model data trends without consideration for data variance and distribution. However the least squares regression is also a valid and reliable method for modeling trends within constraining assumptions. Future work on forecasting using the least squares methods is focused on modification of the slope when the regression p-value is greater than 0.05 but less than 0.10 for 95% confidence level. The standard error of the slope provides the basis for adjusting the model towards a zero slope when the model proves inefficient based on the p-value. The hypothesis is that the estimate becomes more conservative. Using simulations we will compare parametric and non-parametric methods, and we will report findings at the Reliability and Maintainability Symposium.

References

1. W. J. Conover, " Practical Nonparametric Statistics", 2nd ed. , 1981.
2. H. M. Wadsworth, "Handbook of Statistical Methods for Engineers and Scientists", 1990.
3. J. L. Romeu and S Gloss-Soler, "Some Measurement Problems Detected in the Analysis of Software Productivity Data and Their Statistical Consequences", *Proceedings of the IEEE Computer Society's COMPSAC*, November 1983.

Biographies

Jorge Romeu
Reliability Analysis Center
201 Mill St.
Rome, NY 13440-6916 USA

e-mail: jlromeu@syr.edu

Jorge L. Romeu is a Senior Engineer with Alion (formerly IIT Research Institute) and works as statistical advisor to the Reliability Analysis Center (RAC), where he consults on statistical problems and teaches statistics training courses. Romeu, who is also a Professor at Syracuse University, retired Emeritus from SUNY where he taught statistics for fourteen years. He holds a Ph.D. in Operations Research and is a Chartered Statistician Fellow of the Royal Statistical Society. Romeu is the author of numerous papers and of the text book "Practical Guide to Statistical Analysis of Material Property Data". He won the Saaty Award for best statistics paper published in AJMMS in 1997.

Joseph Ciccimaro
Naval Inventory Control Point
700 Robbins Avenue
Philadelphia, PA 19111 USA

E-mail: joseph.ciccimaro@navy.mil

Joseph Ciccimaro is a Senior Operations Research Analyst at the Naval Inventory Control Point-Philadelphia. Prior to joining the Operations Research Department he worked as a Quality Assurance Specialist in the Engineering Department. In addition to his position at the NAVICP he is an Adjunct Instructor in the Statistics Department in the Fox School of Business at Temple University. He has been teaching various statistical courses for Temple University since 1997. He was granted a prestigious Naval Fellowship completing his Masters in Mathematics from Villanova University. He graduated Summa Cum Laude from Gwynedd-Mercy College with a B.S. in Mathematics. recognized with a certificate of achievement for Outstanding Public Service at the 2003 Excellence in Government Awards Program, and he received EEO Honorary Certificate of Achievement for his outstanding contributions to the Naval Inventory Control Point EEO and Diversity Programs. He is a member of the American Statistical Association.

John Trinkle
Veridian/General Dynamics
1550 Hotel Circle North
Suite 425
San Diego, CA 92108 USA

E-mail: john.trinkle@verdian.com

John Trinkle is an Operations Analyst with Advanced Information Engineering Services, Inc. (formerly Veridian). After joining Veridian's Engineering Division in 2002, he has provided the Navy with regression analysis training, system modeling and simulation, and has conducted reliability studies on aviation systems. He graduated with a BS in Electrical Engineering from San Diego State University in 1978 and received an MBA in 1994 from San Diego State University. For 25 years he worked for the U.S. Navy as an engineer and program manager in the automatic test equipment (ATE) and avionics field. From 1999 to 2002 he was project manager for the Naval Air Depot Component In-Service Reliability Program. . He is the co-author and recipient of the IEEE AUTOTESTCON David M. Goodman Best Paper on Management in 1998, and is a member of the American Society for Quality.