RIAC
A PUBLICATION OF THE RELIABILITY
INFORMATION ANALYSIS CENTER
VOLUME 1, NUMBER 1

RELTIQUES
POWERFUL TECHNIQUES IN RMQSI FROM THE RIAC

ON REGRESSION ANALYSIS

TABLE OF CONTENTS

INTRODUCTION

Regression modeling is probably the most widely used (and misused) statistical tool in the analysis of data. Assume that we have two or more variables, that they are associated (correlated), and that the variable of interest is difficult to assess. However, the other variables are cheaper, easier or faster to obtain. Then, it is appropriate to use regression to find a model that expresses the variable of interest, as a function of all the others.

On the other hand, regression is often misused because people forget it is based on a two-part methodology. The first part is purely deterministic. The regression line (or surface, if it is multivariate) is obtained through an optimization process, minimizing the sums of squares of the distances to every data point in the data set. There is no statistics work in this part. Its results, the point estimators, are therefore always valid.

The second part, however, is purely stochastic. Once the minimization is implemented and the line (or surface) is obtained, we add three statistical assumptions to the errors (also called residuals, or distances from the data points to such line or surface). We have to assume that such errors are independent, Normally distributed and have equal variance (i.e. homoskedastic). Only when these three model assumptions are met, can we validly implement statistical tests on the regression coefficients and obtain confidence intervals for them, as well as for the estimations and forecasts done using the regression model.

Unfortunately, many regression users either ignore or overlook the second part. They do not perform tests on the regression residuals to assess whether model assumptions are met. Thence, in many cases, statistical results are used without knowing whether the regression meets its assumptions, or not. The results obtained with invalid regressions are not only also invalid, but often far from correct. As a result, the regression method credibility suffers, when in fact the real problem lies in its improper application.

The objective of the present RelTIQUE is to review the application of the regression methodology, emphasizing the verification of model assumptions. We begin with an example illustrating the complete implementation for simple linear regression. We then move to multiple regression (more than one independent variable) and non-linear or polynomial regression. And we then compare several models to select the "best".

Finally, we discuss the consequences of using regression when assumptions are violated. In such cases we suggest some adaptive solutions that circumvent those problems and provide technical bibliography for further reading.

DEVELOPING A SIMPLE LINEAR REGRESSION

We review the theory behind regression modeling via the implementation of a Simple Linear Regression example. Here, we have only one dependent and one independent variable. The extension of simple linear regression to multivariate (several independent variables) and non linear (i.e. polynomial) regressions, follows directly from this case.

Assume a manufacturer has a processing machine that works at six different speeds (say 1 through 6). The machine starts overheating after some time (given in minutes) and thus has to be stopped. To better understand and manage such machine failures, management collects some data after running the machine at its six different speeds, until it overheats (and the operation has to be terminated). The data is given in Table 1 of Spreadsheet 1.

We begin our analysis obtaining descriptive statistics for variable Time:

```
Variable        N    Mean  Median StDev   Min     Max      Q1       Q3
Time           31  314.53  312.97 17.14 283.12  347.78  302.03   327.40

                            ---------------------
            ----------------I        +          I------------------
                            ---------------------
          ------+---------+---------+---------+---------+---------+Time
              288       300       312       324       336       348
```

Lacking any further explanation, variable "machine Time to overheating" has a mean of 314.53 and a standard deviation of 17.14 minutes. However, when we plot Time vs. the machine speed (Figure 1), we immediately see an alternative explanation: there is a decreasing and linear trend of the variable *Time to overheat*, on *machine Speed*:

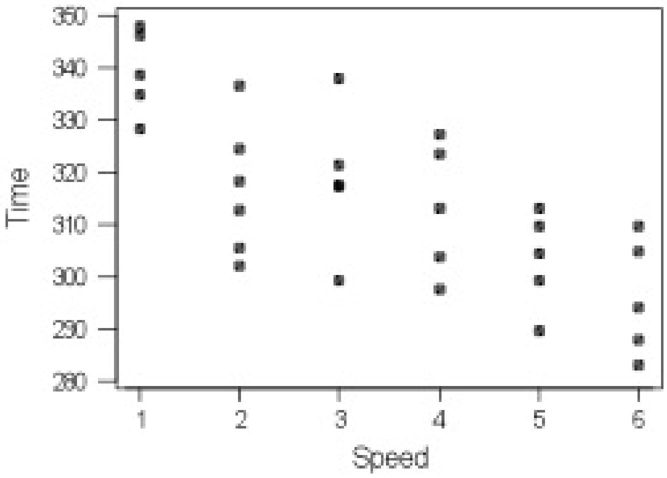


Figure 1. Time vs. Machine Speed

Since *machine Speed* (X) is easy to measure (it is also a controllable factor), we can use it to estimate the *processing Time* (Y) at which such machine overheats. This knowledge allows us to take some appropriate action to counter or diminish the failures brought on by such problem. We accomplish this by implementing a Simple Linear Regression:

$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$ (where "i" runs from 1 to n, the number of pairs of data points)

Regression is simply the line of best fit that runs through the "n" pairs of points (Y_i , X_i) given in Figure 1. Obtaining the point estimators for parameters β_0 and β_1 corresponds to Part I of the regression implementation (the deterministic optimization) discussed in the Introduction (ε_i is the error, or distance from point to line).

Now, in addition, let's assume that these "n" values constitute a random sample of the overheating Times, for all possible runs of the machine, at all its speeds. Hence, what we obtain from the data are the "estimators" b_0 and b_1 of the unknown regression parameters β_0 and β_1. From our example data, such a regression line is:

```
                  Time = 340 - 7.32 Speed

Predictor        Coef       StDev          T        P
Constant      339.791       4.779      71.10    0.000
Speed          -7.319       1.242      -5.89    0.000

S = 11.76        R-Sq = 54.5%      R-Sq(adj) = 52.9%
```

Point estimators for slope and intercept (b_0 and b_1) are obtained by the formulas derived in the minimization process mentioned above. For mathematical details, see Reference 1.

$$b_1 = \frac{\sum_{i=1,n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2}; b_0 = \bar{y} - b_1 \bar{x}$$

Let's now assume that these distances, or differences between each point and the line of best fit ε_i (= Y_i - β_0 + $\beta_1 X_i$), are also independently and Normally distributed random variables (r.v.), with mean Zero and the same variance σ^2. These assumptions constitute Part II of the regression model, already explained in the Introduction. The above formulas for "statistics" b_0 and b_1 are obtained by minimizing the sums of squares of the data point distances: $z = f(x,y) = \sum e_i^2$, where ($e_i = Y_i - b_0 + b_1 X_i$) . Additional mathematical details are provided in References 2, 4 and 5.

When the Errors or Residuals (ε_i) are distributed Normal, regression slope estimator b_1 is distributed as a Student t, with (n-2) degrees of freedom (DF). Only then can we correctly test if such regression slope b_1 is Zero (and we say that there is no regression) or different from Zero (and we say there is a regression line). From here, the importance of checking the three assumptions of Normality, equal variance and independence of the Errors.

In the example given, the Student t statistic for slope b_1 = -7.319 is T = -5.89, and has n-2 = 31 – 2 = 29 DF. We can compare T with the Student t percentile (with DF=29 and $\alpha/2$ = 0.025) which, from the t table is -2.045. Since |T| > 2.045 we reject the assertion (H_0) that true but unknown regression slope β_1 is Zero. However, it is easier to compare the computer generated p-value (= 0.000) with the nominal test error α (= 0.05). When the p-value is less than α, we reject the null hypothesis that the regression slope is Zero.

The step-by-step procedure for testing a linear regression is:

- State the Null H_0: $\beta_1 = 0$ (The slope is zero)
- State the Alternative hypotheses: H_1: $\beta_1 \neq 0$ (It is not)
- Statistic Distribution (under H_0) is $t_{(n-2)}$
- Test Significance level α=0.05;
- Degrees of Freedom = 31-2 = 29
- Student t(n-2, $\alpha/2$) table value = 2.045
- Test p-value = 0.000 (almost zero, very highly significant)
- Model Explanation (R^2): 54% of the problem
- Decision: the time to overheat is well explained by the regression

To verify that everything that we have done is correct, we **must** check all three regression assumptions, which is done via the analysis of Residuals or Errors (e_i). There exist several theoretical tests for Normality (Ref. 9), and for equality of Variance (Ref. 6), useful for the experienced practitioners.

But, at the very minimum, everyone should graphically check the residual plots for the standardized residuals (e_i/s), where "s" (equal to 11.76 in the example) is the regression model standard deviation, which is an estimator of the unknown, theoretical variance σ^2. We illustrate the graphical procedure below.

In our example, the model estimations (Fits) and the standardized residuals (e_i / s) are given in Table 2 of Spreadsheet 1. Using these Residuals and Fits, we check Normality via a Histogram (Figure 2), or a computer-based Anderson-Darling Goodness of Fit test (Figure 3) (Ref. 9).

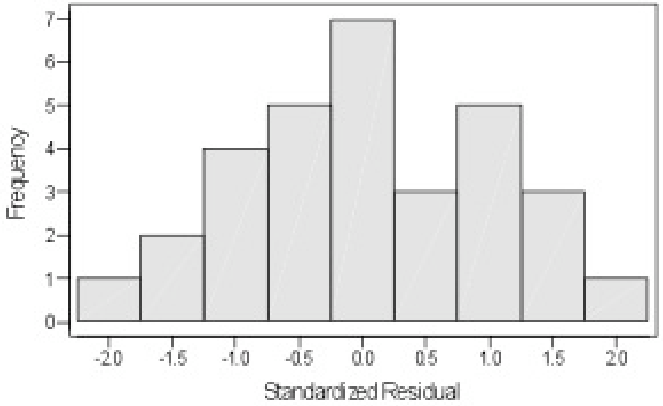


Figure 2. Histogram of the Residuals (Response is Time)

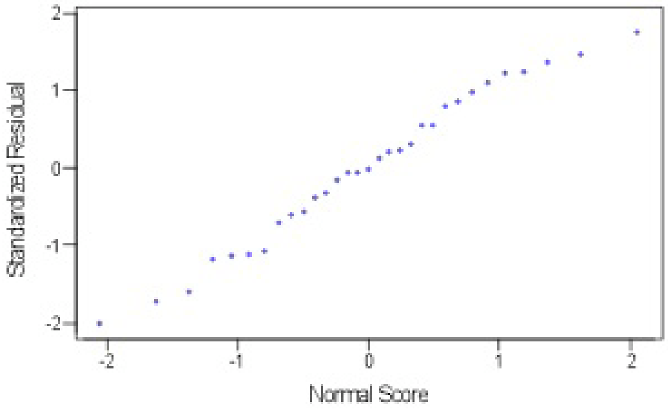


Figure 3. Normal Probability Plots of the Residuals (Response is Time)

The above Histogram, as well as the Normal Plot (which should be close to a straight line), suggest the Normality of the data. The production runs were also selected at random from many different runs done by the manufacturer. This suggests the independence of the data. The Error equality of the variances, which is usually the most important of the three model assumptions, is checked via the plot of standardized residuals vs. regression fits (Figure 4).

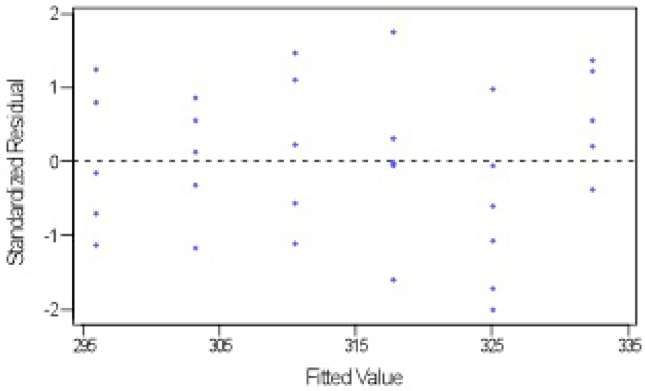


Figure 4. Residuals vs. the Fitted Values (Response is Time)

When variances are equal, most plot data points above lie between two parallel lines at ± 2. When the plot presents a funnel-like pattern instead of homogeneous, there are serious variance problems. In addition, the apparent randomness (absence of patterns) also points to the independence of the data. Hence, in our case, the three model assumptions appear to be met. Hence, we assume that the regression model results are correct.

The Regression Fit is assessed by the R^2 index. In our case $R^2 = 54.5\%$ (it is given in percentages, even when strictly speaking it is a decimal, between zero and unit). This means that 54% of the problem variation (i.e. the variation in times to overheating) can be "explained" by the speed at which them machine is running. However, the remaining 45% of the problem variation is explained by factors other than machine speed (e.g. material, humidity, temperature, etc.). Such factors may later be included in the experimentation, and may help refine the model.

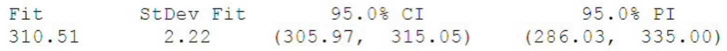
The regression model can also be used to estimate specific values of the dependent variable Y (time to overheat). Formulas for the variance of a given forecast:

$$\hat{y}_0 = b_0 + b_1 x_0 .; V(\hat{y}_o) = \sigma^2 \left[\frac{1}{n} + \frac{\left(x_0 - \bar{x}\right)^2}{\sum \left(x_i - \bar{x}\right)^2} \right]$$

They can be used to obtain the corresponding *confidence intervals* CI, for a specific value of machine speed. Fortunately, most regression packages compute these values for us. For example, with 95% confidence, the time to overheat at speed Four will be:

$$\hat{y}_o \pm t(df, \alpha/2) \sqrt{V(y_o)} = 310.51 \pm 2.04 \times 11.76 \times \left(\frac{1}{31} + \frac{(4 - 3.45)^2}{89.79} \right)^{1/2}$$

That is, a 95% CI for the variable in question will be given by:



```
Fit          StDev Fit         95.0% CI              95.0% PI
310.51          2.22      (305.97,  315.05)     (286.03,  335.00)
```

The first CI above is for an "average" value time to overheat, at speed Four. The second CI is for any *single or individual* value. For example, with 95% confidence, the time to overheat of any run at speed Four will not be less than 286 or greater than 335 minutes.

We can use this information to plan corrective actions and avoid possible problems. For example, if we assess that a processing job will take less than 286 minutes, we can be very certain (with about 98% confidence) that it will be completed before the machine overheats. If the jobs take over 335 minutes, it is very certain the machine will overheat before completing them. If time estimated is between 286 and 335 minutes, then chances are that some of them may have problems and others may not.

Finally, we can obtain a CI for the true but unknown slope and intercept of the regression model. We can do so by using the formulas below, in developing these CI:

$$V(b_1) = \frac{\sigma^2}{\sum (x_i - x)^2} = \frac{218.2}{89.79} = (1.24)^2$$

$$V(b_0) = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum (x_i - \bar{x})^2} \right] = (4.78)^2$$

For example, we can obtain a CI for the true regression slope, because its estimator is distributed as a Student t:

$$\hat{b}_1 \pm t(df, \alpha/2)\sqrt{V(b_1)} = -7.32 \pm 2.045 \times 1.242 = (-9.86, -4.78)$$

This procedure allows us to establish lower and upper bounds for the (linear) effect of machine speed on time to overheat. With 95% probability, for example, the slope (rate) of change for these variables is between -9.86 and -4.78. Hence, these CI lower and upper limits can now be used to provide optimistic, pessimistic and average *estimates* of the times to overheat, given the machine speed.

DEVELOPING MULTIVARIATE REGRESSIONS

Let's now enhance the regression model to more than one independent variables or factors. All assumptions discussed for Linear Regression still hold here. Such multiple regression model, for "k" components, is mathematically expressed as:

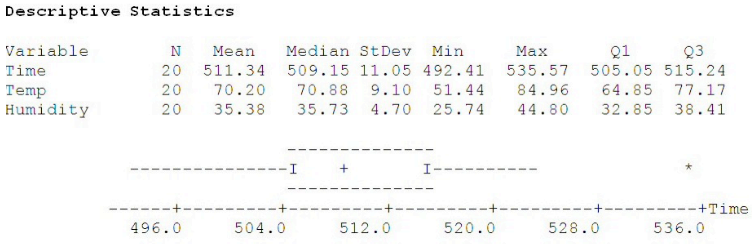
$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots \ \beta_k X_{ik} + \varepsilon_i \ ; \ 1 \le i \le n$$

Now, we have "n" data vectors, (Y_i, X_{i1}, X_{i2} ... X_{ik}) each with (k+1) elements. There are "k" regression coefficients (β_1, ... β_k) one for each independent factor in the equation, in addition to the independent term β_0. Each coefficient must be tested as being equal to, or different from Zero, using an individual t-test. In addition, the equation-wide F-test will determine whether the whole equation is significant (i.e. at least one of the β_i coefficients, of the "k" regressors or factors, is non-zero). If there is "no regression", all variable or factor coefficients are zero. That is, the response Y is simply equal to the general average (the constant β_0). Let's see all this through a numerical example.

A processing machine is operated in one of three possible speeds (namely 1, 2, 3) until it overheats and has to be stopped. To study this problem, an experiment is performed. For each run of the machine, the average temperature and humidity are recorded. These, plus speed, constitute the three regressors, factors or independent variables (X_1, X_2, X_3). Time to overheating (Y) is the dependent variable or response, recorded in hours of operation. Management wants (requirement) the machine to operate up to 520 minutes continuously, at least 90% of the time, before overheating. Regression can be useful in this problem.

An experiment, where the machine operates under all possible scenarios, is implemented, and the resulting data (Y_i, X_{i1}, X_{i2}, X_{i3}; i = 1, … 20) is given in Table 3 of Spreadsheet 1:

Lacking other explanation, the machine processing time to overheat (Y) is distributed Normally (Ref. 9), with the sample statistics and Box Plot:

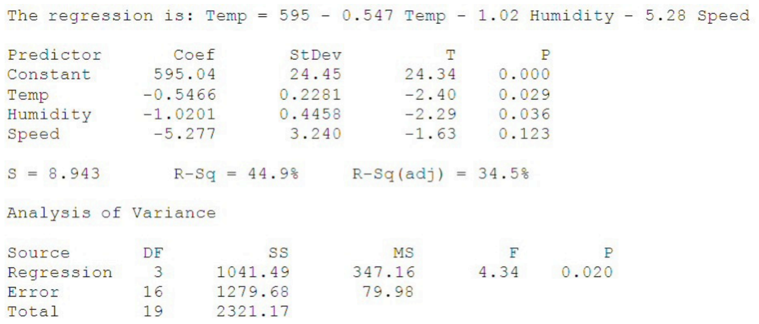


```
Descriptive Statistics

Variable        N    Mean    Median StDev  Min      Max       Q1      Q3
Time           20  511.34   509.15 11.05 492.41   535.57   505.05  515.24
Temp           20   70.20    70.88  9.10  51.44    84.96    64.85   77.17
Humidity       20   35.38    35.73  4.70  25.74    44.80    32.85   38.41

                           --------------
          ---------------I    +        I----------            *
                           --------------
       ------+---------+---------+---------+---------+---------+Time
           496.0     504.0     512.0     520.0     528.0     536.0
```

Under present conditions, the probability that a machine job has to be aborted (a failure), is the probability that the Time to overheat Y, during a run, is less than the specified goal of 500 minutes (i.e. Mission Time). In statistical terms, we state:

P (Abort/Failure) = P(Y < 500) = P{Z < (500-511.34)/11.05} = 0.1524

Through multiple regression we assess if any of the factors contribute to such overheating and by how much. This lets us select better alternatives to improve the current situation:



```
The regression is: Temp = 595 - 0.547 Temp - 1.02 Humidity - 5.28 Speed

Predictor        Coef       StDev          T        P
Constant       595.04       24.45      24.34    0.000
Temp          -0.5466      0.2281      -2.40    0.029
Humidity      -1.0201      0.4458      -2.29    0.036
Speed          -5.277       3.240      -1.63    0.123

S = 8.943       R-Sq = 44.9%      R-Sq(adj) = 34.5%

Analysis of Variance

Source         DF          SS          MS         F        P
Regression      3     1041.49      347.16      4.34    0.020
Error          16     1279.68       79.98
Total          19     2321.17
```

The regression is, overall, significant (F statistic is 4.34, with a p-value of 0.02). In particular, the first two individual t-tests, for the coefficients of independent variables X_1, X_2, are significant (their respective p-values are 0.029 and 0.036). The third coefficient, for variable "speed" is non-significant (p-value = 0.123). If there is still concern about variable X_3 having an effect on the equipment (current p-value = 0.12 is low) additional experiments should be carried out (i.e. increase sample size). Otherwise, we just drop X_3 from the analysis (equation), as it has no significant effect on time.

The three factors, together, explain (R^2) about 44.9% of the problem. Other factors not present in the current model (e.g. operator, material, day of the week) would have to be investigated, if the current explanation (R^2) is to be increased.

Now, just like in the previous section, we need to investigate whether the regression model fulfills its assumptions and, hence, the above test results are valid –or not. We do this graphically (see Figures 5 and 6, below), as well as analytically (formal statistical tests).

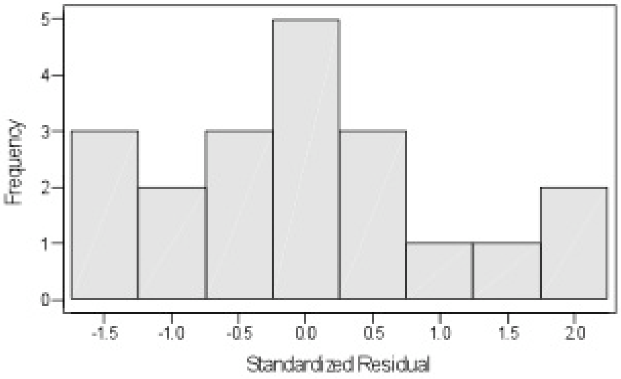


Figure 5. Histogram of the Residuals (Response is Time)

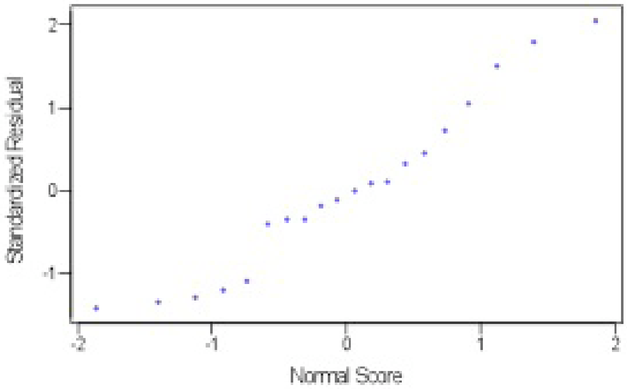


Figure 6. Normal Probability Plot of the Residuals (Response is Time)

The Histogram of the multiple regression residuals peaks around zero, and the probability plot is close to a straight line. There is no strong indication that non-Normality is present.

In practice, ideal conditions seldom exist. When analyzing residuals, we appraise strong vs. the weak violations of each model assumption and strive for a working balance. When some violation occurs, we should report it up front, using analysis results with caution and making others aware that our results have to be taken with care.

We show next the plot of residuals vs. fits. It confirms residual equal variance when it shows most data points within two (imaginary) parallel lines at levels ±2. This plot also shows no apparent pattern that may lead us to suspect that non-randomness is present (Figure 7).

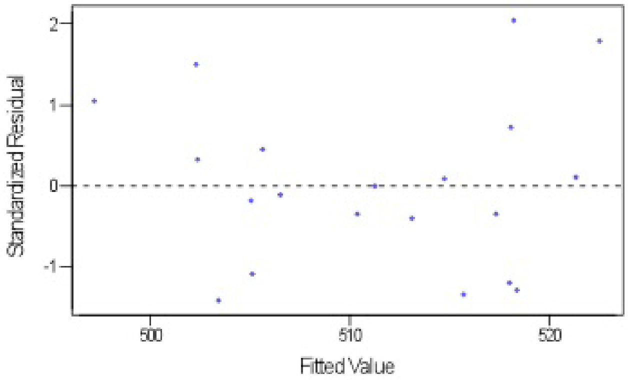


Figure 7. Residuals vs. the Fitted Values (Response is Time)

Therefore we will assume that all regression assumptions have been reasonably met, and we can proceed to use the analyses results in solving our reliability problem.

For example, the average temperature and humidity are 70° F, and 35%, with standard deviations of 9.1° F and 4.7%. Finding ways to reduce the temperature and humidity, at least one standard deviation below their respective averages, can help increase response "Y". By pushing up the mean of the distribution of Y (i.e. Time to overheat) to say, 530 minutes, we may obtain a significant improvement:

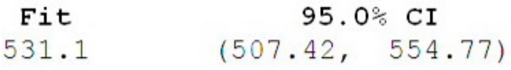
$$Y = Temp = 595 - (0.547*(70-9.1)) - (1.02*(35-4.7)) = 530.782$$

Assuming variance of Y remains constant, the probability of failure to complete a job is:

$$P\ (Abort/Failure) = P\ (Y < 500) = P\ \{Z < (500-530.78)/11.05\} = 0.0027$$

If more conservative estimates are desired, instead of using regression coefficient point estimators, we can obtain their CI and use their upper/lower limits as comparison values.

For example, we show below the CI for time to completion Y of any individual job, done under controlled temperature (less than 61° F) and humidity (less than 30 %). We use the multiple regression equation obtained, giving for variables X_1, X_2 the above-mentioned values. Variable X_3 (speed) is considered zero, since the coefficient of this third variable resulted non-significant (zero) in the regression tests. Results are given below:



```
 Fit              95.0% CI
531.1         (507.42,   554.77)
```

DEVELOPING NONLINEAR REGRESSIONS

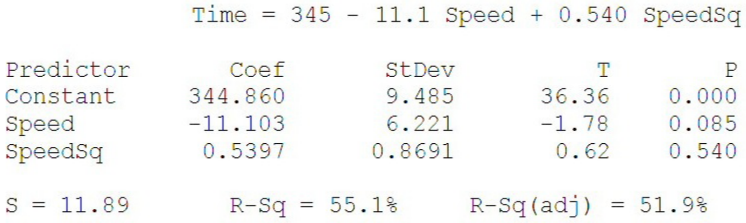
Some times, instead of multiple variables (multivariate regression), model response "Y" depends on **powers** of the same, single independent variable "X". If such is the case, the resulting equation does not define a line, but a non-linear function (polynomial). If more than a single independent variable, then it does not define a plane, but a general surface. We illustrate below, the first case by reprocessing the data used for the linear model in our first example:

Assume we want to include a quadratic term in our regression. The new model becomes:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_{12} X_1^{\,2} + \varepsilon_i \ ; \quad 1 \le i \le n$$

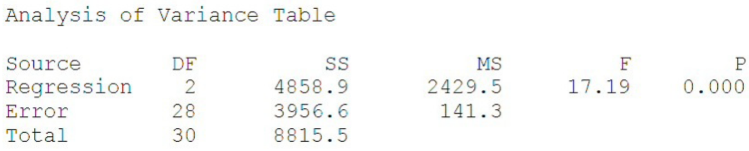
We want to assess whether such polynomial regression improves on the original solution.

Reusing the data set in Table 1 of Spreadsheet 1, we fit a second degree parabola. The quadratic regression equation is:



```
                Time = 345 - 11.1 Speed + 0.540 SpeedSq

Predictor        Coef        StDev          T        P
Constant      344.860        9.485      36.36    0.000
Speed         -11.103        6.221      -1.78    0.085
SpeedSq        0.5397       0.8691       0.62    0.540

S = 11.89        R-Sq = 55.1%      R-Sq(adj) = 51.9%
```

The regression p-values show how the quadratic term introduced is not statistically significant (p-val = 0.540). The addition of a quadratic term does not improve the model:



```
Analysis of Variance Table

Source        DF        SS        MS       F       P
Regression     2    4858.9    2429.5   17.19   0.000
Error         28    3956.6     141.3
Total         30    8815.5
```

The above ANOVA table is used whenever we have *more than a single* regressor "X". For, in such cases, we use the statistic F only to assess the overall regression equation. In addition we use individual t-statistics, to assess each individual regression coefficient.

MODEL COMPARISONS

For completion, we *compare* the linear and quadratic models. For, when we develop more than one regression model for the same problem, we need to select the "best" one.

The best regression model, just as the best engineering design, is usually the simplest model: one that does the work with the minimum complexity. In regression, this means the function with the fewest terms (most parsimonious) and the best explanation (R^2).

To compare the two (linear and quadratic) regression models, we use their respective R^2 values, as well as the degrees of freedom (DF) for the residuals of the Full and Reduced (FM, RM) Models. We implement such comparison using the formula below:

$$F = \frac{\left[\dfrac{R^2_{FM} - R^2_{RM}}{\Delta_{DF}}\right]}{\dfrac{1 - R^2_{FM}}{DF_{FM}}} = \frac{\dfrac{0.551 - 0.545}{1}}{\dfrac{1 - 0.551}{28}} = \frac{0.006}{0.016} = 0.374$$

In this example, we gain nothing by using the second degree parabola (F=0.37). Even if the above F comparison showed an advantage (i.e., F test is significant and FM or larger equation is better than RM or simpler one) the main condition remains that all regression model assumptions (as per graphical analysis at least) are not strongly violated.

DISCUSSION

For an introductory but complete treatment of regression modeling, the reader can consult Reference 7. For a more in-depth, still mainly practical approach, see References 2 and 4.

There are two additional regression modeling important topics, among many other ones that we have not had space to cover in a single RELTIC. We want, at least, to mention them, even in passing. For, they both deal with the important issue of what to do when regression model assumptions are violated.

If the violations consist of lack of Normality or heterogeneous variance, then a variable transformation may provide a solution. For example, if regression residuals are Binomial, with parameters "n" and "p", it is known that the Mean (Expected Value) is "np" and the Variance is "np(1-p)". In such cases, the residual variance is a function of the mean and the *residual plot* yields the characteristic *funnel-like* pattern, resulting from:

$$\text{Variance} = np \times (1\text{-}p) = \text{Mean} \times (1\text{-}p)$$

A square root, logarithmic or arctangent transformation may resolve the issue. For additional information, the reader should consult the advanced References 1, 5 and 6.

If transformations are not feasible, or do not resolve model violations, and the model is a simple linear regression, then non-parametric regression procedures may prove useful.

Non parametric methods (also known as distribution free) do not depend on assuming a specific distribution. Hence, the need for residual Normality is no longer an issue (and therefore, no longer we require equal variances). For an introduction to the use of such linear non-parametric regression in a reliability problem, see Reference 8.

REFERENCES FOR FURTHER STUDY

1. Anderson, T. W., An Introduction to Multivariate Statistical Analysis (2nd Ed.), Wiley, NY, 1984
2. Chatterjee, S. and B. Price, Regression Analysis by Example, John Wiley, NY, 1977
3. Coppola, A., Practical Statistical Tools for Reliability Engineers, Reliability Analysis Center, 1999
4. Draper, N. and H. Smith, Applied Regression Analysis, John Wiley, NY, 1980
5. Dixon, W. J. and F. J. Massey, Introduction to Statistical Analysis, McGraw Hill, NY, 1983
6. Flury, B. and H. Riedwyl, Multivariate Statistical Analysis: a Practical Approach, Chapman Hall, NY, 1988
7. Romeu, J. L. and C. Grethlein, A Practical Guide to Statistical Analysis of Material Property Data, AMPTIAC, 2000
8. Romeu, J. L., Ciccimaro, J. and J. Trinkle, "Measuring Cost Avoidance in the Face of Messy Data", Proceedings of the 2003 RAMS Symposium, January 2003
9. Romeu, J. L., Anderson-Darling Goodness-of-Fit Tests, START, RIAC, 2004

ABOUT THE AUTHOR

Jorge Luis Romeu is a Senior Science Advisor with Quanterion Solutions Incorporated, and a Research Professor at Syracuse University. He holds a Ph.D. in operations research and Quality and Reliability ASQ Certifications. Romeu is a Chartered Statistician Fellow of the Royal Statistical Society and a Senior Member of ASQ. He has over thirty years experience in teaching, research and consulting in industrial statistics.