Chapter 7
Case Studies with Regression

Jorge Luis Romeu
IIT Research Institute
June 24, 1999

Executive Summary

In this chapter we discuss the use of regression models through the development of four case studies in materials data analysis. We use real data sets taken from handbook [6] and the RECIPE program Users Guide [5]. We have also modified some of these sets to illustrate specific regression procedures of interest. We develop linear and non-linear regression models, compare them and select the one that best describes the problem at hand. All regression model assumptions are carefully checked, both graphically and analytically, via the residual analysis. An example in data transformation is also presented, for completeness. Finally, several important caveats regarding the correct implementation of regression models are presented and discussed.

Introduction

This is the last technical chapter of this SOAR. It deals with regression analysis of materials data. The study of this subject, jointly with that of analysis of variance developed in the previous chapter, constitute the main objective of our present work. For, both of these modeling approaches are extremely useful and widely utilized in data analysis in general and in materials data analysis in particular.

As done in the previous chapter, we follow the analysis road map established in the handbooks. This time it is the procedure in Figure 9.6.3 of [6] (herein Figure 7.1). This figure describes "General procedures for performing a regression analysis in order to calculate design allowables". We explain the figure procedures in detail, below.

At the start of any analysis, we should obtain the descriptive statistics and plot the data in several useful ways, via EDA (exploratory data analysis) techniques, in order to obtain an initial idea of their weak and strong points and to form some conjectures about them.

Then, we use these conjectures to guide our first analyses steps toward establishing the underlying statistical distribution and its parameters, as well as to assessing whether there are any outliers in the data set. We will use, as in previous chapters, the Anderson Darling (AD) GoF test to assess the Normal distribution of the data, which is the basic requirement of regression analysis models. We will use graphical (e.g. boxplots) and analytical (MNR) procedures to detect and assess the presence of outliers. These, in turn, should not be removed unless due cause (via the inspection of the data sources or other similar strong reason) is present. For, outliers usually convey a great deal of information.

Then, we proceed to fit a simple linear regression model (if there are three or more levels of X, the independent or predictor variable). We have explicitly dedicated one case study example to highlight the problem of lack of predictor levels in regression analysis.

After the linear regression is implemented, graphical and analytical residual analysis help assess whether the model assumptions have been met. In the affirmative case, we proceed to use the regression model results. Otherwise, if any model assumption has been rejected then we need to resort to data transformations or to alternative procedures.

If the (linear) regression is adequate and significant, we can use it to obtain parameters of interest (e.g. allowables). We can also use regression models to assess whether certain factors or predictor variables (say temperature, thickness) have an effect on the dependent (quantitative) variable (say, tensile strength) and to quantify this effect. Regression analysis results may thus be used in theoretical (study) or practical (prediction) work.

If there are four or more levels of the predictor variable(s) then it may be possible to fit higher order regression equations (say quadratic or cubic). After these regression models are fitted their assumptions must be assessed. If the assumptions are met, then we have an interesting situation: several valid models for the same problem. We then need to select the best between them, in the sense of better representing or explaining the problem.

Model selection is a complex problem and we will overview it via several case study examples. The most important caveat of this chapter is related to the selection problem. For example, there may be a model that fits very well the data but that has very little (if any) theoretical or experiential support. In this case, it may occur that we are modeling the data instead of the problem –which is a very expensive but possible mistake. The best advice is then, that practical and empirical models such as regression should follow and back the theory and experience –and not the other way around.

Finally, once a satisfactory, parsimonious and valid regression model is obtained, we proceed to use it to estimate parameters of interest. We can also use it to forecast, control, study or in any other way characterize a given set of data.

Case One. An example in stress-strain curves

The data for this case study are taken from Table 9.3.2.3 of the same numbered section, on page 9-67 of reference [6]. It deals with an example of the use of strain departures to establish typical stress-strain curves as described in handbook [6]. We have presented below three columns from the mentioned table: departure, average stress and total strain:

ROW	Depart	Avg-T	StrainT
1	0	42.59	4022
2	20	47.91	4544
3	40	50.17	4768
4	100	53.17	5121
5	500	59.21	6092
6	1000	61.66	6823

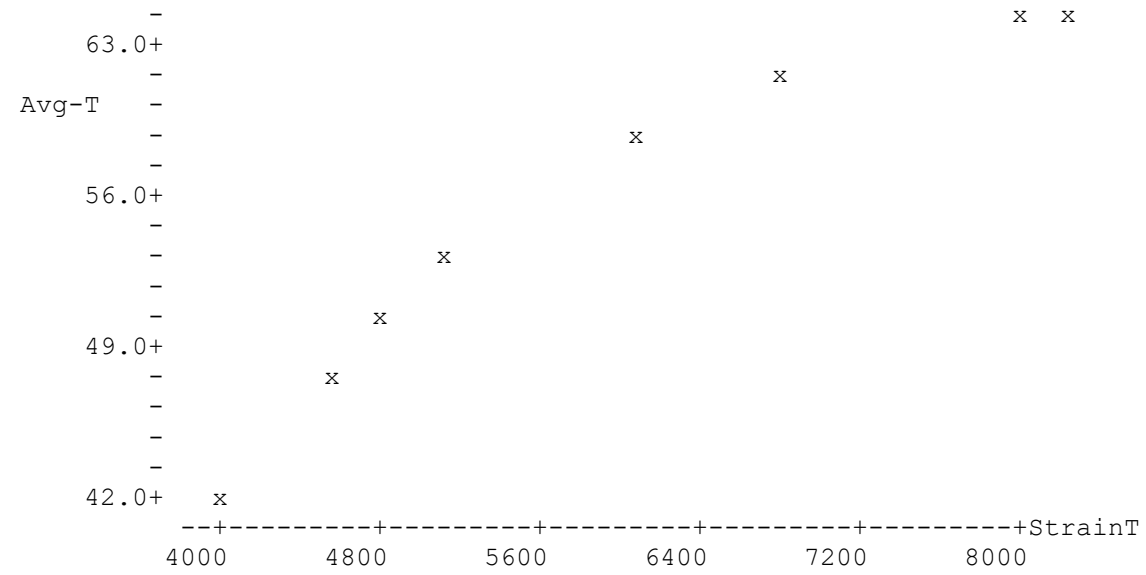
```
   7    2000   63.94     8038
   8    2200   64.25     8267
```

As usual, we first obtain the descriptive statistics and plot the data in several ways. This time, since we are interested in obtaining a regression model, we include the correlation. For, we want to establish a first diagnostic about the data that helps us better achieve our objective of deriving a regression function for tensile stress, based on strain.

```
                N     MEAN   MEDIAN   STDEV   MIN    MAX    Q1     Q3
Depart          8      733      300     911     0   2200    25   1750
Avg-T           8    55.36    56.19    8.08 42.59  64.25 48.47  63.37
StrainT         8     5959     5606    1616  4022   8267  4600   7734

          Depart    Avg-T
Avg-T     0.860
StrainT   0.971    0.958
```

We see there is a high correlation between average stress and total strain (0.95). We will explore further this relation graphically. But first, the distribution also has to be tested for bivariate Normality. This is a model requirement for the correct implementation of the Pearson correlation, a measure of linear association between two Normal variables (alternatively, we could use Spearmann's or other non-parametric correlation test).

We apply AD GoF test to each variable (e.g. stress and strain) individually. For, if the data are bivariate Normal, then their marginal distributions (each variable, individually) are also (univariate) Normal. The AD results for stress and strain above are, respectively, 0.29 and 0.32, with p-values of 0.51 and 0.45. We then assume each is, individually (marginally) Normal and also jointly bivariate Normal (there are specific GoF tests for bivariate Normality but they lie outside the scope of this SOAR). We then accept Pearson correlation results as a valid measure of linear association between the two variables and proceed to estimate the linear regression that describes this relation. We start with their bivariate plot, shown below:



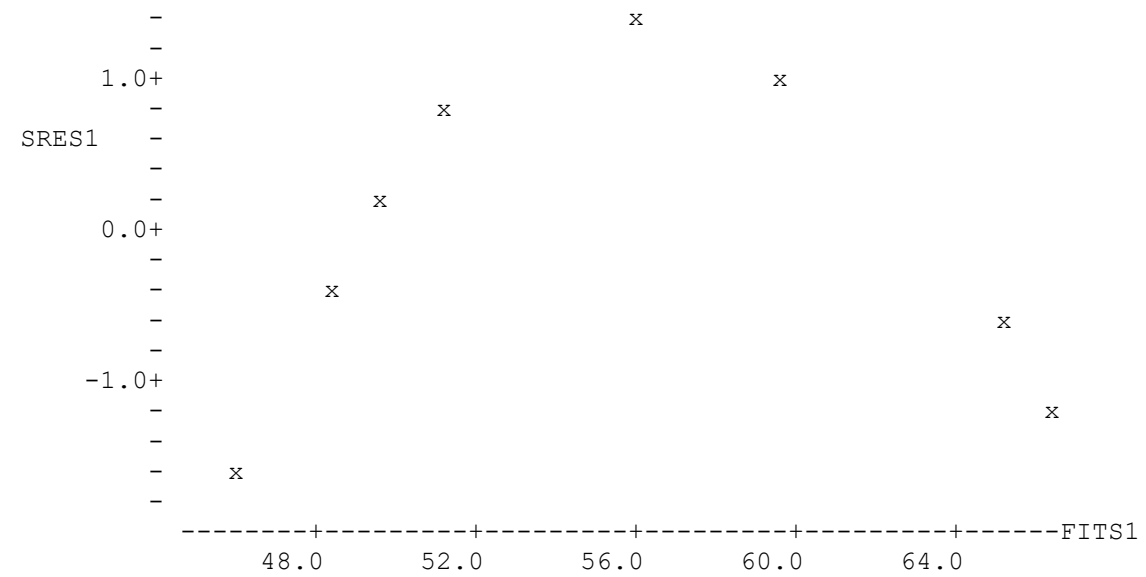
The bivariate plot shows a positive association between the two variables, that we will model via a simple linear regression. The estimated regression line is shown below:

```
                    Avg-T = 26.8 + 0.00479 StrainT

Predictor       Coef       Stdev    t-ratio        p
Constant      26.816       3.609       7.43    0.000
StrainT    0.0047902   0.0005871       8.16    0.000

s = 2.511       R-sq = 91.7%     R-sq(adj) = 90.4%
```

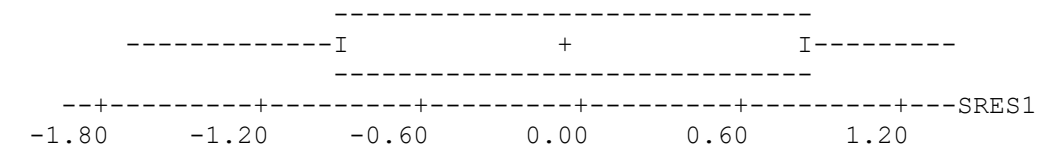
The above strain regression coefficient (predictor) is 4.79×10^{-3} highly significant (p-value is practically zero). The index of fit ($100R^2 = 91.7$) describes over 90% of the variation of the data. The model standard deviation is 2.51. The ANOVA table is:

```
SOURCE       DF          SS          MS         F        p
Regression    1      419.65      419.65     66.58    0.000
Error         6       37.82        6.30
Total         7      457.47
```

The regression equation has a residual sum of squares (SSR) of 37.82 and a very high F-statistic value (66.58) for the entire regression model, also highly significant. Since this is a simple a linear regression, predictor (coefficient) and model significance are equivalent. In the next analysis, using a quadratic model, we will realize the difference. However, before we implement the quadratic model, we need to check the validity of the current model assumptions. We start by plotting the standardized residuals vs. the fitted values.



The residual plot shows a clear concave down pattern. The reader can compare this one to the random patterns obtained in our previous regression and ANOVA residual analyses. This is characteristic of models that have not captured the totality of the problem structure and a first indication that this linear regression model is not yet satisfactory. However, residuals are symmetric about zero, as shown by the boxplot below:



This example shows how is it useful to look at all aspects of the residual analysis and not only to part of it. In this case, the Normality of the residuals is not suspect. The problem here is the lack of fit of the model. To try to solve this latter problem, we proceed by fitting a quadratic regression to the data. The quadratic regression equation is:

```
Avg-T = - 21.1 + 0.0212 StrainT -0.000001 StrTSq

Predictor       Coef       Stdev    t-ratio        p
Constant     -21.141       2.530      -8.36    0.000
StrainT    0.0212096   0.0008554      24.80    0.000
StrTSq   -0.00000132  0.00000007     -19.27    0.000

s = 0.3170      R-sq = 99.9%     R-sq(adj) = 99.8%
```

Comparing the above results with the simple linear regression ones presented before, we notice several improvements. First, the quadratic equation yields highly significant test statistics for all its coefficients. In addition, the new model index of fit (R-sq) describes over 99% of the problem. This improvement is also evident in the ANOVA table:

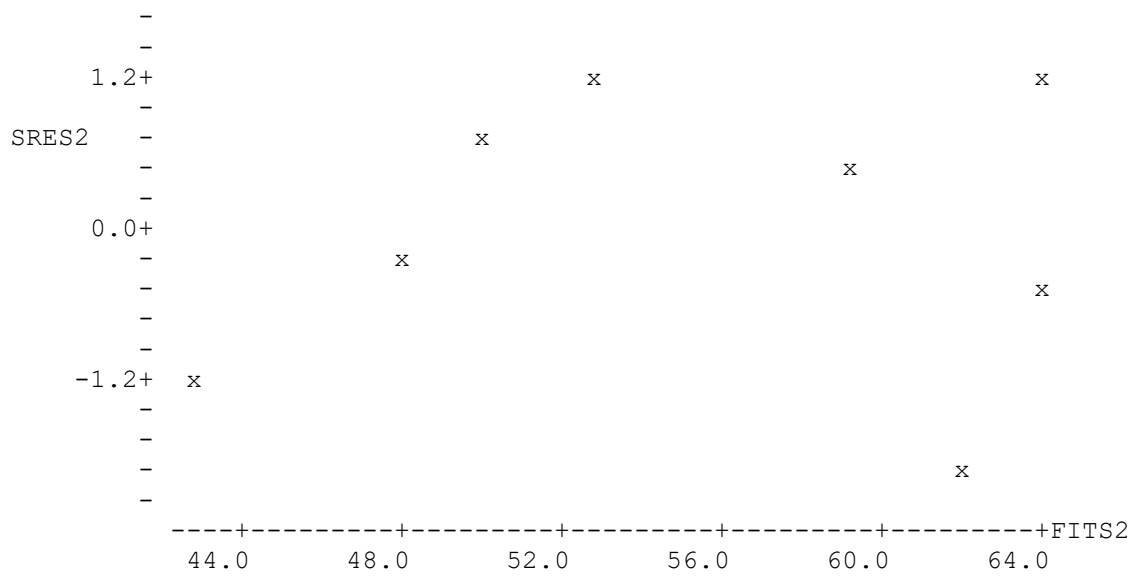
SOURCE	DF	SS	MS	F	p
Regression	2	456.97	228.48	2273.18	0.000
StrainT	1	419.65	419.65		
StrTSq	1	37.32	37.32		
Error	5	0.50	0.10		
Total	7	457.47			

Notice how the overall model F-statistic = 2273.18 is even more significant than before, as are also the two (linear and quadratic) strain coefficient terms. In addition, the residual sum of squares has been reduced to 0.5. Let's analyze this model improvement via the model comparison test, described in chapter five.

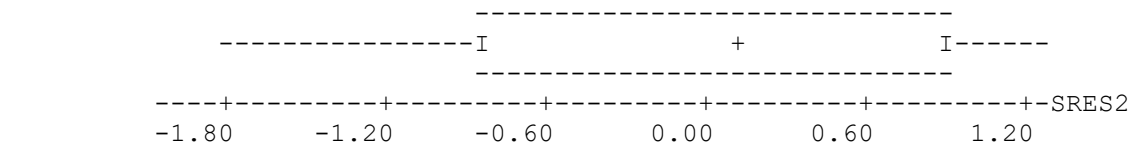
The full model is now the quadratic regression and the reduced model is the simple linear regression. The null hypothesis states that both models equivalently describe the problem. We compare the error sum of squares from the linear ($SSR_{L)}$ with that of quadratic (SSR_Q) regression. They have, respectively, DF_L=6 and DF_Q=5 d.f. Their difference (divided by the reduction in d.f. and divided by SSR_Q / DF_Q) provides a measure of the improvement obtained when moving from one model to the other. The test statistic is:

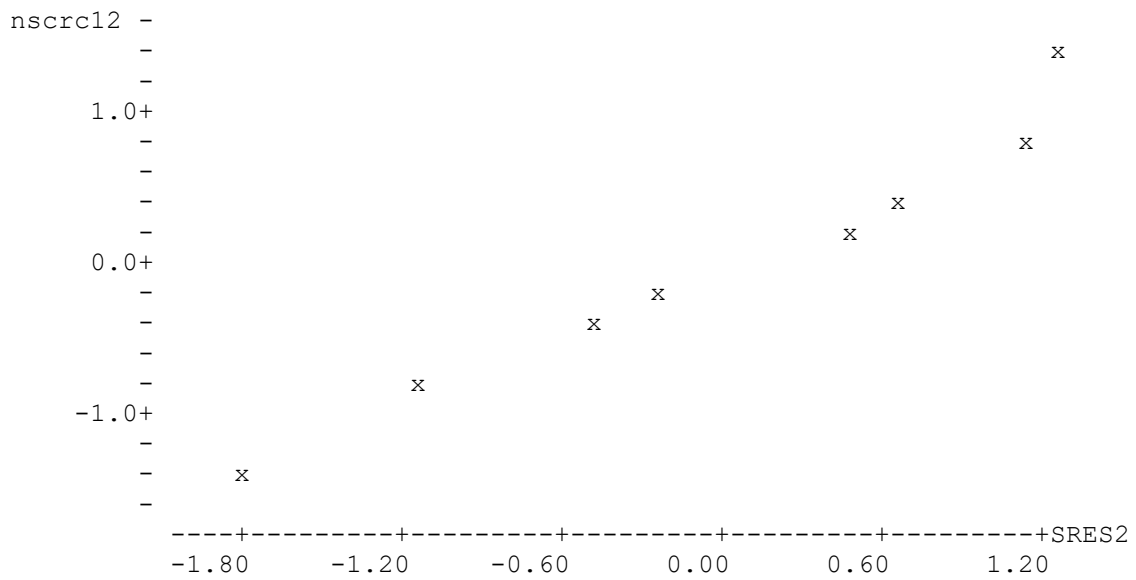
$$F = \frac{(SSR_L - SSR_Q)/(DF_L - DF_Q)}{(SSR_Q / DF_Q)} = \frac{(37.82-0.5)/(6-5)}{0.5/5} = 373.2$$

Comparing it with the F-Table (critical) value F(α=0.05, dfnum=1, dfden=5) = 6.61 we see that the F-test result is highly significant. Therefore, the quadratic equation improves significantly our model. However before adopting it, we still have to assess the residuals,.



This time, the pattern of residuals vs. model fits looks more random. In addition, the AD test for Normality yields 0.092 with p-value=0.99 and an almost perfect (straight-line) probability plot. The runs test for residual randomness yields p-value=0.12, too high for rejecting this assumption. The box and normal scores plots are also shown below.





The quadratic regression model has not only improved on the percent explanation of the problem, but also on the regression model assumptions, which are now better met. From the above plots the residuals appear random and Normally distributed about zero and their pattern does not suggest any variance problems. Thus, we adopt the quadratic model as an acceptable description of the structure of the problem under study.

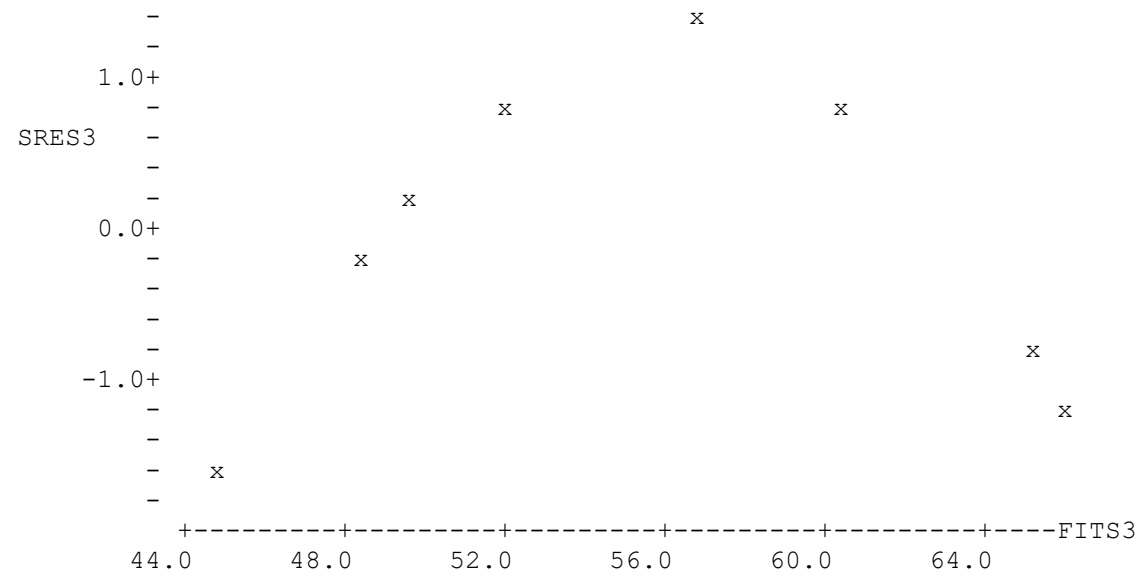
Next, and just for comparison, we implement a data transformation. We take the natural logarithm of the predictor variable total strain and then regress average stress on it.

The regression equation is: Avg-T = - 199 + 29.4 LnStrT

Predictor	Coef	Stdev	t-ratio	p
Constant	-199.36	21.00	-9.49	0.000
LnStrT	29.410	2.424	12.14	0.000

s = 1.728 R-sq = 96.1% R-sq(adj) = 95.4%

The regression on the transformed data has also improved on the index of fit (model explanation) and is highly significant. However, the residual analysis is still problematic, as shown by the pattern in the plot of residuals vs. regression fits, below:



The concave down residual pattern remains. The previous (linear regression) problem has been alleviated but not resolved with this (unsuccessful) data transformation. Therefore we select the quadratic regression model for modeling this problem and data set.

Case Two. Surface Damage Example revisited: modeling the data.

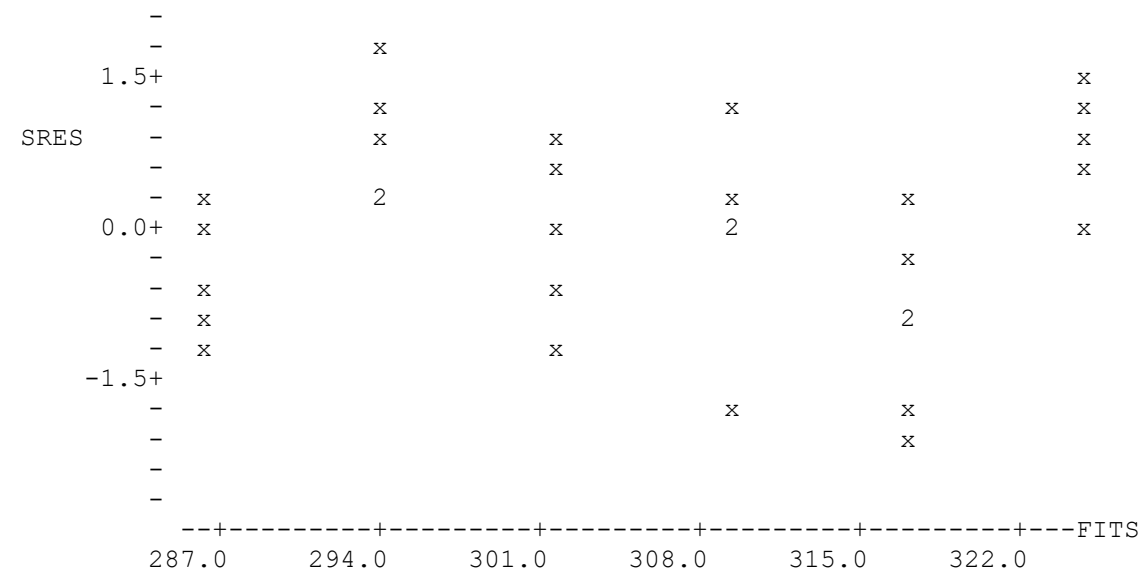
We briefly revisit the data set used in chapter five to introduce regression. For, it poses a frequent and difficult problem in statistical analysis: that of modeling the problem vs. modeling the data. This (fictitious) data set is composed of variables surface damage and

tensile strength (matstr). We have also obtained the squares (dam-2) and cubes (dam-3) of variable "damage". We describe them again, below, for completeness:

	N	MEAN	MEDIAN	STDEV	MIN	MAX	Q1	Q3
matstr	31	305.66	307.72	19.67	270.02	344.78	291.04	319.60
damage	31	3.452	3.000	1.72	1.00	6.00	2.00	5.00
dam-2	31	14.81	9.00	12.38	1.00	36.00	4.00	25.00
dam-3	31	71.4	27.0	77.00	1.00	216.00	8.00	125.00

In chapter five we plotted and analyzed them, fitting two regression models: one linear and the second quadratic. Then, we performed a model comparison and found that there was no significant difference between the explanations provided ($100R^2$=45%) by either of the regression models. Hence, we selected the most parsimonious: linear regression.

In this section we will fit a cubic regression to the data. A cubic function is suggested by the residual plot, whose undulating pattern reminds us of such a sinusoidal form.



The results of fitting a cubic regression to these data are shown below:

matstr = 406 - 96.7 damage + 28.7 dam-2 - 2.67 dam-3

Predictor	Coef	Stdev	t-ratio	p
Constant	406.04	18.94	21.44	0.000
damage	-96.74	21.33	-4.53	0.000
dam-2	28.662	6.812	4.21	0.000
dam-3	-2.6736	0.6444	-4.15	0.000

s = 11.89 R-sq = 67.1% R-sq(adj) = 63.4%

Notice how all regression coefficients are now highly significant (the p-value is practically zero) and the index of fit, or model explanation, has increased to 67%. This improvement is also evident from the ANOVA table for the regression, shown below:

SOURCE	DF	SS	MS	F	p
Regression	3	7786.0	2595.3	18.35	0.000
damage	1	5277.6	5277.6		
dam-2	1	73.7	73.7		
dam-3	1	2434.7	2434.7		
Error	27	3818.4	141.4		
Total	30	11604.4			

The overall (now Full Model) cubic regression F-statistic (18.35) is highly significant (p-value is practically zero) and the residual sum of squares SSRc=3818.4. We now test for the best model fit by comparing the previously selected (linear) and the current (cubic) regressions. We compare the linear regression residual sum of squares (SSR_L=6326.7) with that of the cubic SSRc regression. Their d.f. are respectively, DF_L=29 and DF_Q=27. Their difference and ratio provide a measure of improvement gained by moving from one model to the other. In our case, the F-test statistic to assess such improvement is:

$$F = \frac{(SSR_L - SSR_c)/(DF_L - DF_Q)}{(SSR_c / DF_c)} = \frac{(6326.7-3818.4)/(29-27)}{3818.4/27} = 8.87$$

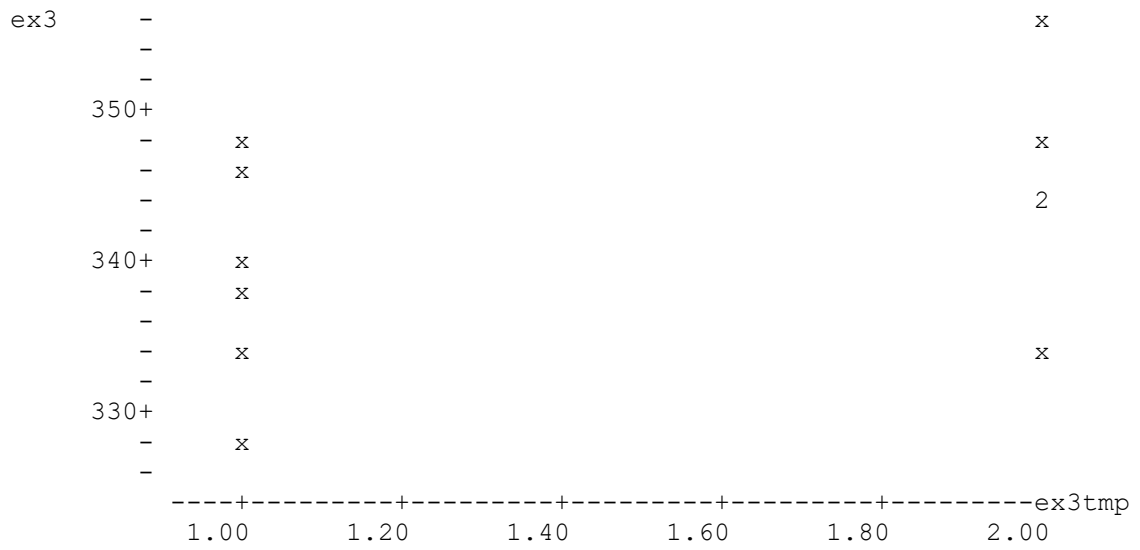
Comparing 8.87 with the F-Table (critical) value F (α=0.05, dfnum=2, dfden=27) = 4.21 we see that the F-test result is highly significant. Therefore, the cubic equation provides a better explanation than the linear model. In principle the cubic model would be accepted.

However, is there a fundamental reason for this cubic model to describe this problem? Or is it just an artifact of the data on hand? If the second answer were correct, then we would be modeling the data and not the problem. The objective of statistical data analysis is not to model a particular data set but the problem mechanism or structure that generates the data. Hence, if there is no good justification for a specific functional form to fit a set of data, we must be extremely careful when postulating it as a model.

This is an important criterion that helps in assessing and selecting which model to adopt. Therefore, in the present case, before adopting the cubic model a thorough engineering analysis that justifies this cubic regression should be successfully undertaken.

Case Three. Ex3.dat revisited: regression data constrains

We again discuss Ex3.dat data set, which we analyzed in chapter six using ANOVA. This data set was also analyzed in the RECIPE program Users Guide [5] and in the handbook [6] (section 8.3.7.7, page 8-61) using regression. The data is composed of 11 tensile strength observations, taken at two different temperatures, 75 and –67 degrees Farenheit, all from the same batch. For completeness, we present again below, the scatter plot of tensile strength vs. temperature.



In both of the above-mentioned references, regression analyses were implemented. They clearly and correctly illustrate the use of the RECIPE program for the derivation of allowables as well as the step-by-step implementation of regression analysis and of the allowable calculations. In [6] (page 8-61) a note clarifies that "a linear relationship between strength and temperature is not appropriate for all temperature ranges". We would like to expand on this important caveat, regarding such regression analyses.

We must repeat what we said in the previous case study example regarding statistical modeling. In statistical analysis, there is always a risk of modeling the data and not the problem. This is especially dangerous in a situation where we have only two predictor measurement levels, as is the case in the present example.

For illustration and comparison, we refer the reader to the scatter plot of the raw data in the previous case study (surface damage) example, which was initially presented in our introduction to regression of chapter five. If we had only taken the two end-point values in the abscissas, from this surface damage example, we would observe a downward linear trend. When we add the middle range of the abscissa observations, we then see how an oscillating trend is present, instead.

This is the main caveat we want to raise with the present example. The two temperatures (75 and –67) are widely separated. Therefore, one can question whether these two points may be signaling a linear trend, or rather a concave up (or concave down) or oscillating trend. The safest way to resolve this important question is to include at least a third intermediate temperature (this is why handbook [6] suggests three or more levels of the predictor). But most importantly yet, is to always keep in mind that the statistical (regression) model is empirical and should follow a logical or theoretical explanation – not lead it. If there is no basis for assuming that a linear (or cubic or other) trend is legitimate, then we run the risk of modeling the data and not the problem, as in the previous case study. This can lead to inefficient conclusions since, if we model the data but not the problem, another data set may have a completely different functional form.

Case four. A small but more complex data set.

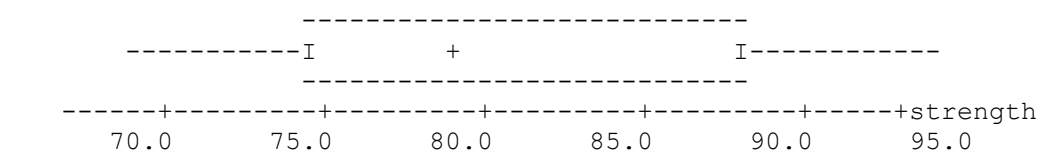
We now present Ex5.dat, from the RECIPE program Users Guide [5], also discussed in section 8.3.7.9 (page 8-68) of the handbook [6]. This data, presented below, consists of 15 tensile strength observations, from five batches and two different manufacturers:

ROW	manufac	batch	strength
1	1	1	75.8
2	1	1	78.4
3	1	1	82.0
4	1	2	68.8
5	1	2	70.9
6	1	2	73.5
7	1	3	74.5
8	1	3	74.8
9	1	3	78.8
10	2	4	81.3
11	2	4	87.7
12	2	4	89.0
13	2	5	88.2
14	2	5	91.2
15	2	5	94.2

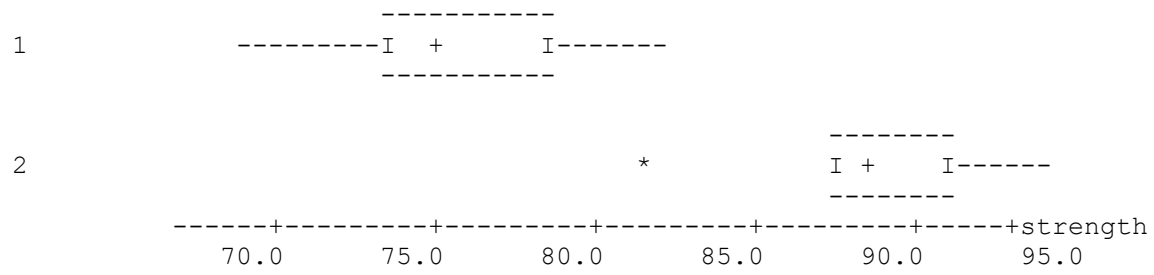
The statistical problem consists in determining whether this is a homogeneous data set and what are its measures of central tendency, dispersion and others, that characterize it. We also want to know what is the underlying distribution and its parameters –and if there are possible outliers in the set. If the data are not homogeneous, then we want to know if they vary by manufacturer or by batch or both. If such variation exists, we then want to know if there are reasons for this (e.g. they are caused by a trend on some other factor). We can later use this additional information, say, for validating or forecasting one tensile value, given the ancillary information. This case study summarizes everything we have seen. As usual, the first thing we do with a data set is to obtain its descriptive statistics:

	N	MEAN	MEDIAN	STDEV	MIN	MAX	Q1	Q3
strength	15	80.61	78.80	7.860	68.80	94.20	74.50	88.20

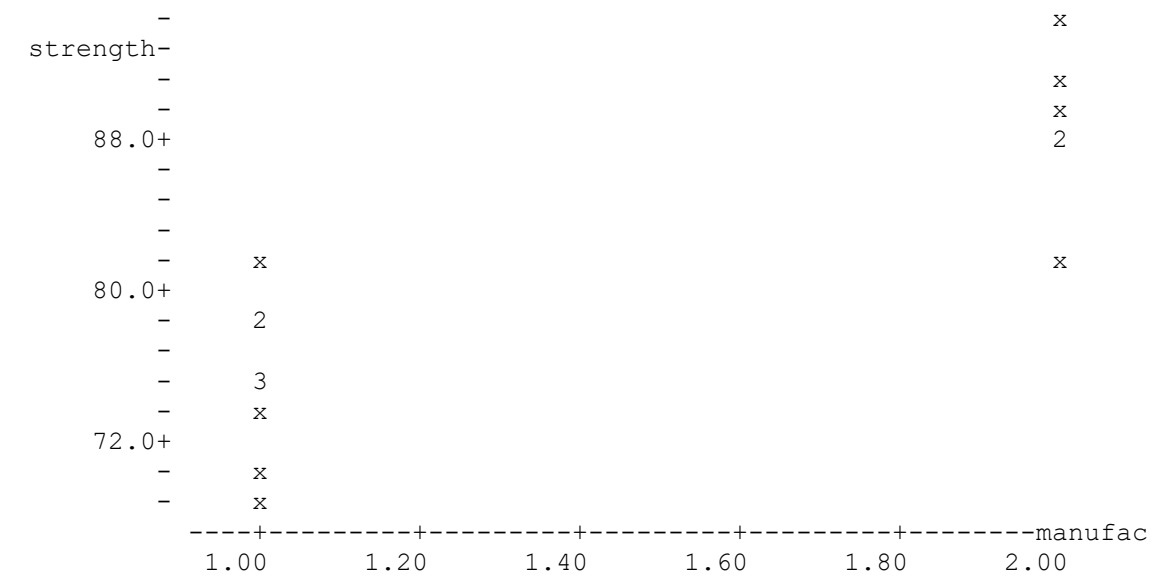
Then, we plot the data in various useful ways (pooled, by groups, etc.) to obtain a first diagnostic about how they are similar or about how they differ:



The boxplot of the combined data set shows a flat and symmetric population, with heavy tails. The median and mean are close and the data are spread out, as shown by the extended upper/lower quartiles. We then break down the data by manufacturer and some reasons for the data variability becomes apparent:



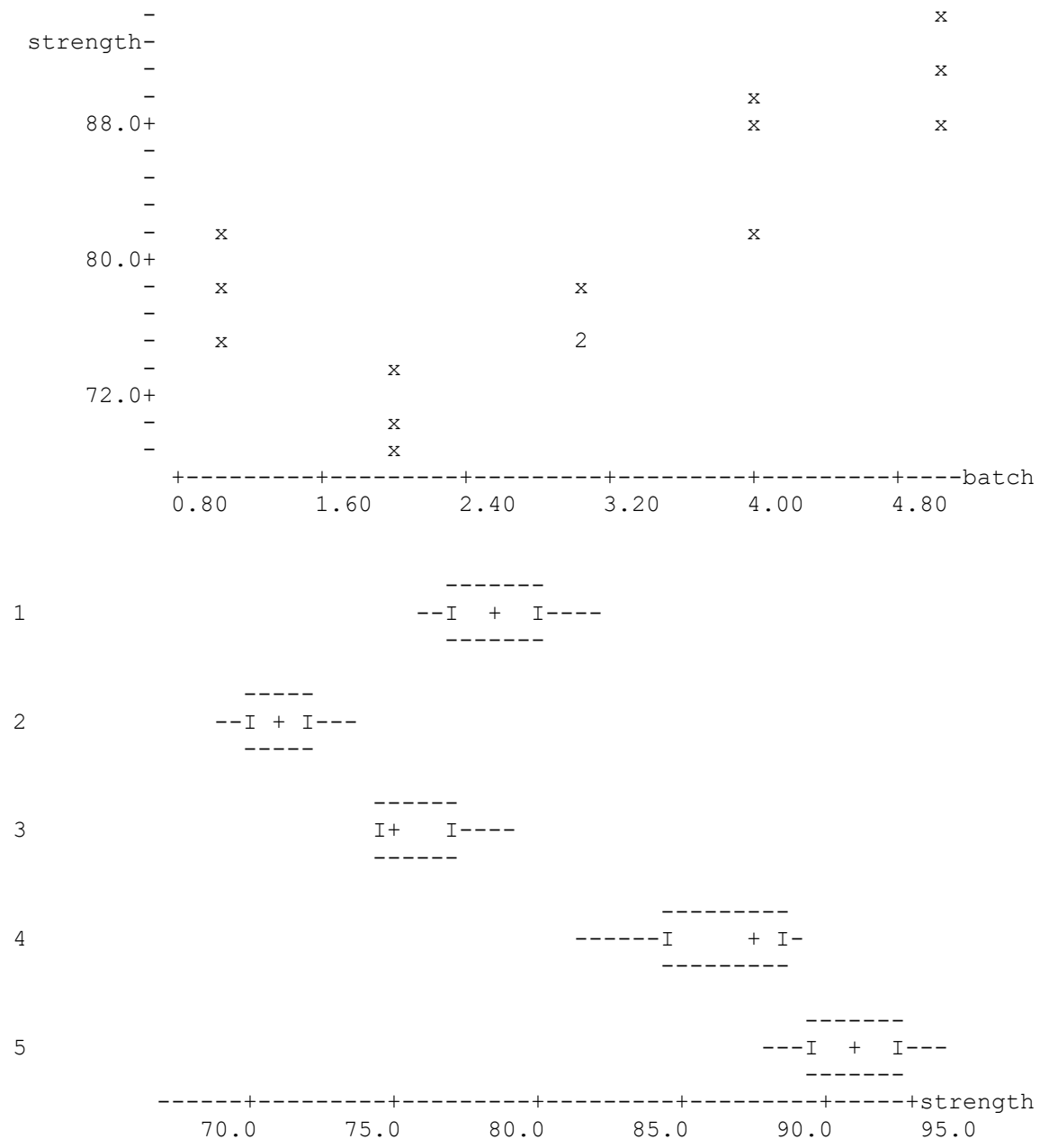
The above boxplots signal that there are differences in the tensile strengths of the two (manufacturers) groups, which is also clearly apparent in the scatter plot below. We need to explore further this manufacturer's difference:



The descriptive statistics, obtained by manufacturer's group, confirm such difference:

	N	MEAN	MEDIAN	STDEV	MIN	MAX	Q1	Q3
manuf-1	9	75.28	74.80	4.07	68.80	82.00	72.20	78.60
manuf-2	6	88.60	88.60	4.30	81.30	94.20	86.10	91.95

We want, in addition, to investigate if there are also batch differences, within the two manufacturers –or whether they are internally homogeneous. We present the batch scatter and box plots, below.



It is apparent from these plots that batches also differ by manufacturer. We will explore this situation analytically after establishing the underlying distributions of the two groups.

We perform the AD GoF tests for Normality for the entire data set obtaining AD=0.33 with a p-value=0.47, too high for rejecting Normality as a plausible data distribution. The boxplots do not suggest the presence of outliers in the combined group, either. We then implement the AD GoF test for each manufacturer, obtaining AD values of 0.15 and 0.28 respectively, with p-values of 0.93 and 0.50. With such results we don't reject Normality, a convenient distribution for the implementation of (two-sample and ANOVA) comparison tests between groups, that require the Normality assumption.

We then compare the two (manufacturers) groups via two-sample t-tests. For, the two group variances are (see descriptive statistics) very similar and, hence, assumed equal.

```
          N      MEAN     STDEV   SE MEAN
manuf-1   9     75.28      4.07       1.4
manuf-2   6     88.60      4.30       1.8
POOLED STDEV =       4.16
95 PCT CI FOR MU manuf-1 - MU manuf-2: ( -18.1,  -8.6)
TTEST MU manuf-1 = MU manuf-2 (VS NE): T= -6.07  P=0.0000  DF=  13
```

Results show that the two groups differ. The second manufacturer has a tensile strength mean between 8.6 and 18.1 units higher, than that of the first, with 95% confidence. For illustration, we also perform the ANOVA test for (only) two (manufacturers) groups. Notice below how the ANOVA results are equivalent to those of the two-sample t-test above. This is no surprise, since the distribution of the (ANOVA) F-test, for only two groups, is the square of that of the t-test for the two-sample case (i.e. $F(1, m) = [\ t(m)\ ]^2$) where m are the degrees of freedom of the t-test statistic or the d.f. of the denominator of the F-statistic (here, the F-test statistic value 36.87 is the square of the t-test -6.07).

```
ANALYSIS OF VARIANCE ON strength
SOURCE     DF        SS        MS        F        p
manufac     1     638.9     638.9    36.87    0.000
ERROR      13     225.3      17.3
TOTAL      14     864.2
                                    INDIVIDUAL 95% CI'S FOR MEAN
                                    BASED ON POOLED STDEV
 LEVEL      N      MEAN     STDEV  ----------+---------+---------+-----
     1      9    75.278     4.073  (----*----)
     2      6    88.600     4.302                      (-----*-----)
                                    ----------+---------+---------+-----
POOLED STDEV =    4.163                  78.0      84.0      90.0
```

Within each of the two manufacturers groups we already observed some differences. We will explore them now, analytically, via ANOVA. For manufacturer #1 we detect that there is a statistical difference between batches. ANOVA results are presented below:

```
SOURCE     DF        SS        MS        F        p
bat-1       2     90.74     45.37     6.48    0.032
ERROR       6     42.00      7.00
TOTAL       8    132.74
                                    INDIVIDUAL 95% CI'S FOR MEAN
                                    BASED ON POOLED STDEV
 LEVEL      N      MEAN     STDEV  ------+---------+---------+--------+
     1      3    78.733     3.113                (------*-------)
     2      3    71.067     2.354   (------*-------)
     3      3    76.033     2.401            (------*-------)
                                    ------+---------+---------+--------+
POOLED STDEV =    2.646             70.0      75.0      80.0     85.0
```

For manufacturer #2, however, the two batches appear to come from the same population. This result may indicate that their production process is more homogeneous (controlled) than that of manufacturer #1. Further investigation, with more batches, is suggested.

```
SOURCE      DF        SS        MS        F        p
bat-2        1      40.6      40.6     3.12    0.152
ERROR        4      52.0      13.0
TOTAL        5      92.5

                                      INDIVIDUAL 95% CI'S FOR MEAN
                                      BASED ON POOLED STDEV
 LEVEL       N      MEAN     STDEV  ----------+---------+---------+-----
     4       3    86.000     4.122  (-----------*-----------)
     5       3    91.200     3.000             (----------*-----------)
                                      ----------+---------+---------+-----
POOLED STDEV =    3.605                      85.0      90.0      95.0
```

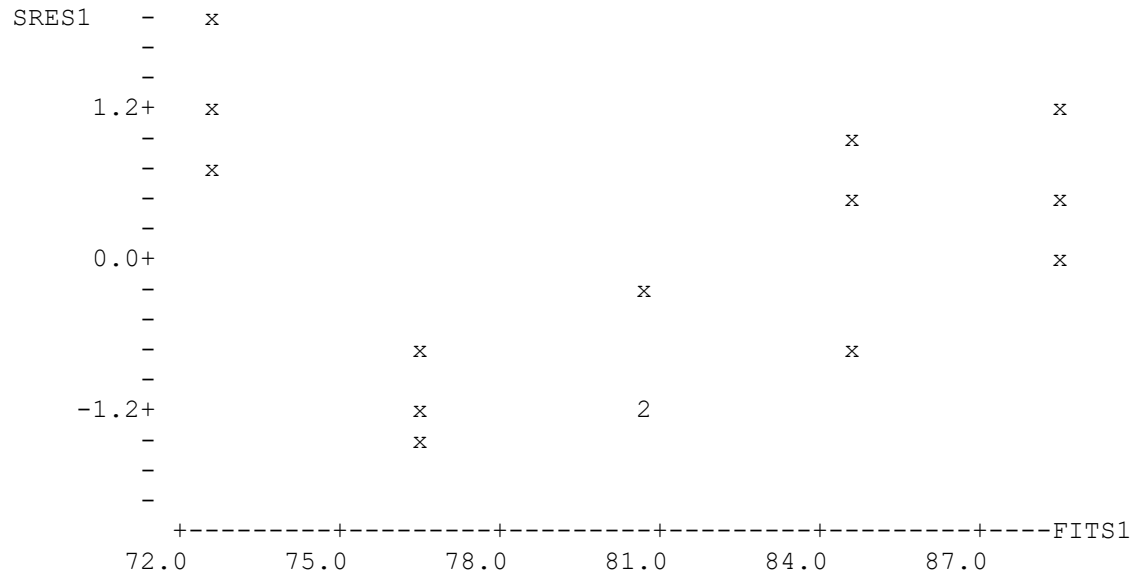
For illustration also, and since the scatter plots show an increasing trend among tensile strength batches, we will perform a regression analysis on the combined data set. We will regress tensile strength on the five batches. Results are presented below:

The regression equation is: strength = 68.6 + 3.99 batch

```
Predictor       Coef       Stdev    t-ratio        p
Constant      68.647       3.306      20.77    0.000
batch         3.9867      0.9967       4.00    0.002
s = 5.459       R-sq = 55.2%     R-sq(adj) = 51.7%

                    Analysis of Variance
SOURCE        DF          SS          MS          F        p
Regression     1      476.81      476.81     16.00    0.002
Error         13      387.40       29.80
Total         14      864.21
```

There is an increasing effect of batch. The index of fit ($100R^2$) is 55%; the model explains over half of the data variation. The t and F regression tests are highly significant (p-values are practically zero). Again, there are two important caveats. First, look at the residual plot. Second (and more important) does this result have a sound basis?

The residual plot (below) shows a distinct pattern as opposed to the randomness expected from a well-fitted regression model. Such pattern signals that our model has not yet captured the structure of the problem. In addition, unless there is some specific reason in both manufacturers processes (e.g. they have been using sequentially aged raw material, that produces an increasing effect on tensile strength) there is no reason to suspect that batches should produce such trend in the response. This example shows how, in addition to serving for purely statistical purposes (e.g. checking the validity of the assumptions) the residuals also help in the complex process of validating the conceptual models.

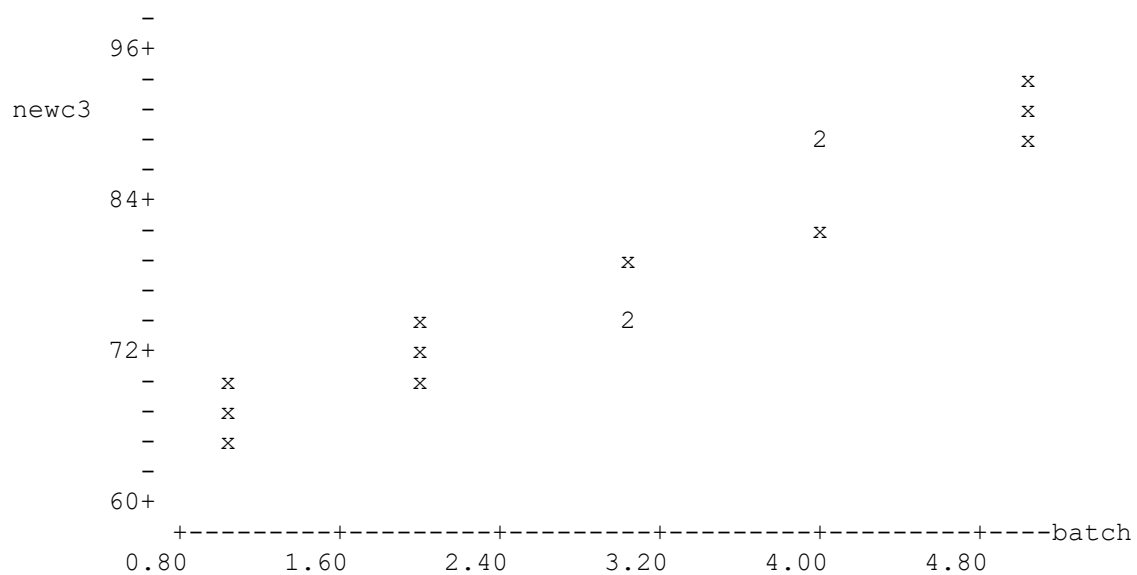


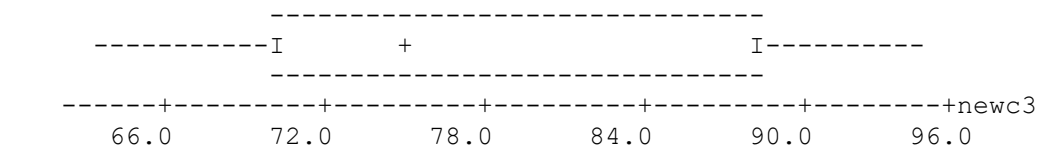
Finally, and also for the purpose of illustrating some problems of applying regression models to materials data, we transform the Ex5.dat data set into the set "newc3.dat". We do it in the following way: we decrease in twelve units all tensile strengths coming from the first batch of the first manufacturers. The resulting (newc3) data set is shown below:

```
63.8   66.4   70.0   68.8   70.9   73.5   74.5   74.8   78.8   81.3
87.7   89.0   88.2   91.2   94.2
```

We obtain the descriptive statistics and basic plots for the new tensile data set (newc3):

	N	MEAN	MEDIAN	STDEV	MIN	MAX	Q1	Q3
newc3	15	78.21	74.80	9.80	63.80	94.20	70.00	88.20





Let's assume now, for illustration, that the factor "batch" is instead a material "thickness" level specification, to which the product has been manufactured. This last assumption tries to provide some physical meaning to the stress problem that we will now analyze:

The regression equation is: newc3 = 59.0 + 6.39 thickness

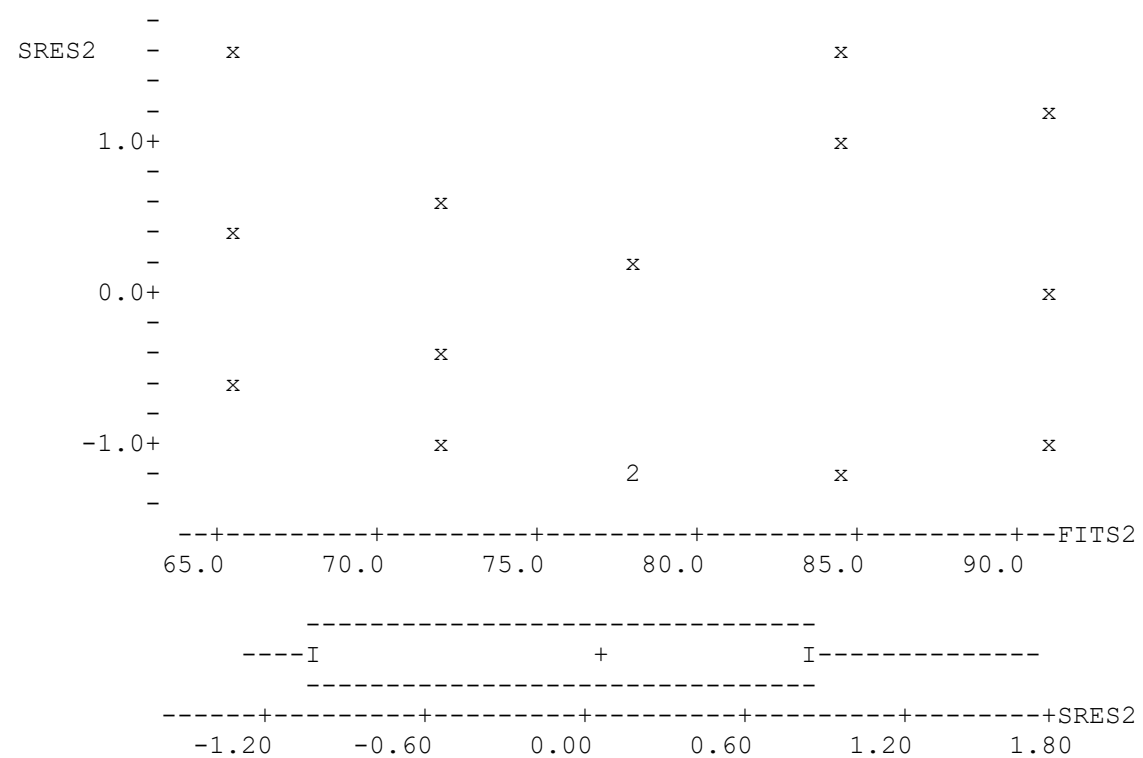
Predictor	Coef	Stdev	t-ratio	p
Constant	59.047	1.847	31.96	0.000
thickness	6.3867	0.5570	11.47	0.000

s = 3.051 R-sq = 91.0% R-sq(adj) = 90.3%

Analysis of Variance

SOURCE	DF	SS	MS	F	p
Regression	1	1223.7	1223.7	131.47	0.000
Error	13	121.0	9.3		
Total	14	1344.7			

Both the regression and ANOVA tables above indicate that this new regression is highly significant (the t and F statistics have p-values practically zero). The new model explains (index of fit) over 90% of the variation in the data. We still need to check the residuals to assess the validity of the model assumptions. Residual and box plots are shown below:



The scatter plot of residuals vs. fits seems plausibly random and the residual boxplot seems symmetric about zero. The AD GoF test for Normality yields a value AD=0.45 with a p-value of 0.24. In addition, the runs test yields 10 runs out of an expected 8.47, with a p-value of 0.40. We thus assume the Normality of the residuals. Since all other regression assumptions have been met, we proceed to use the regression model results, namely that material "thickness" does increase tensile strength, as per the equation above.

However, thickness levels were implemented by two different manufacturers. Therefore we also want to assess whether there is a manufacturers effect in this problem, too. There are three ways in which such an effect may appear in the regression model. First, the level (independent term of the regression) may differ by group. Secondly, the rate (the slope of the regression) may be different. Finally, both level and rate effects may be different (in which case, two totally different regression models apply).

We first test the hypothesis that the two manufacturers do not differ in level. We compare the above regression model with another one where we introduce E_i as an extra "dummy variable" (full model). This dummy variable has value 0, if the observation comes from the first manufacturer and 1, if it comes from the second. The model functional form is:

$$Y_i = \beta_{00} + \beta_{11} E_i + \beta_1 X_{i1} + \varepsilon_i \;\; ; \;\; 1 \le i \le n$$

The null hypothesis is H_0 $\beta_{11}=0$. If it is true, both regressions (and both manufacturer processes) are statistically equivalent. However, if we reject H_0 then all variables from the above regression are statistically significant (different from zero). Then, an estimation from the first manufacturer would be obtained using the regression equation:

$$Y_i = \beta_{00} + \beta_1 X_{i1} + \varepsilon_i \;\; ; \;\; 1 \le i \le n$$

Whereas, an estimation for the second manufacturer would be obtained using the second regression equation, that now has a different (sum of) independent term(s) shown below:

$$Y_i = (\beta_{00} + \beta_{11}) + \beta_1 X_{i1} + \varepsilon_i \;\; ; \;\; 1 \le i \le n$$

The regression results for the current fictitious example are given below:

```
            newc3 = (61.8 + 5.42 dumlevel) + 4.76 thickness

Predictor       Coef       Stdev    t-ratio        p
Constant      61.758       2.253      27.41    0.000
thickness      4.760       1.025       4.64    0.000
dumlevel       5.422       2.959       1.83    0.092

s = 2.807       R-sq = 93.0%     R-sq(adj) = 91.8%
```

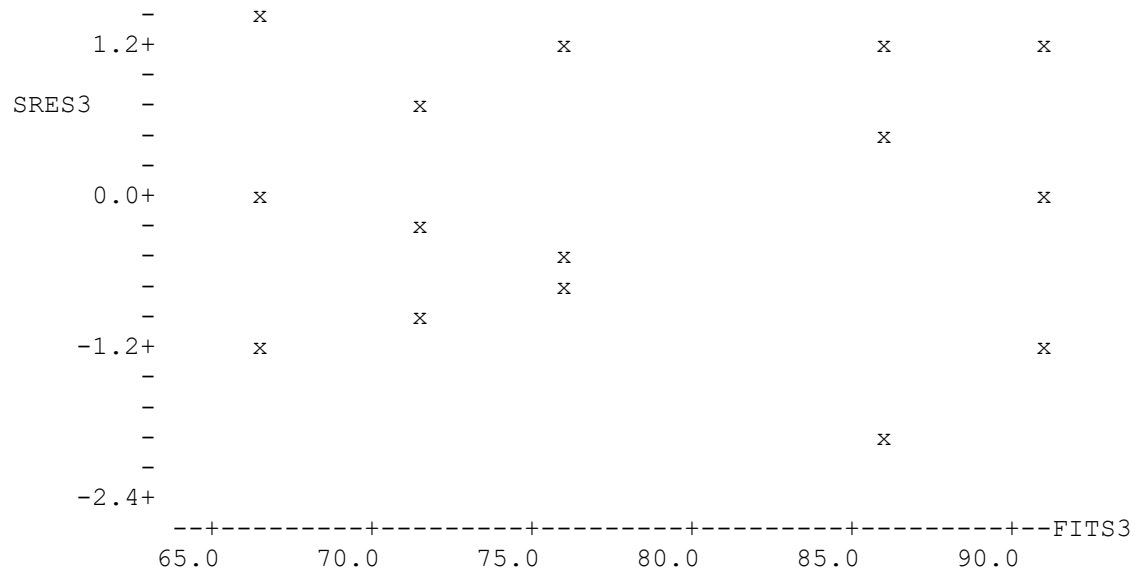
Notice how the model explanation (index of fit) has barely increased a couple of percentage points (from 91 to 93%). The coefficient of the "thickness" factor (4.76) remains highly significant. But the coefficient for the dummy variable (dumlevel) has a

p-value=0.092. This result is statistically significant only if we were willing to assume an error α (risk of wrong decision) of 10%. This α may be too high, especially for such a small (n=15) data set. The ANOVA table for the regression is shown below.

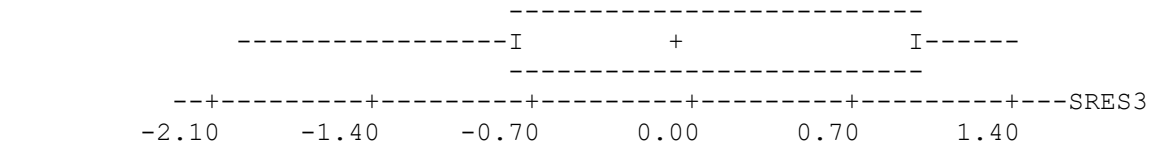
Analysis of Variance

SOURCE	DF	SS	MS	F	p
Regression	2	1250.15	625.07	79.34	0.000
thickness	1	1223.69	1223.69		
dumlevel	1	26.46	26.46		
Error	12	94.54	7.88		
Total	14	1344.69			

Below we present the plot of residuals vs. fits. Notice the random pattern that supports the validity of the regression model assumptions of Normality and equality of variance.



The residual boxplot (below) and the AD GoF test results (AD=0.45 with p-value=0.24) also point toward Normality and randomness. Therefore, we accept the above regression model as valid and proceed to use its results, namely that variable E_i is not significant.



Therefore, since we do not reject the (null) hypothesis, both above regressions have the same and unique β_{00} intercept (since β_{11}=0). Hence, we can use the pooled regression results. Had we opted to reject the null (at 10% risk of stating, erroneously, that β_{11} was not zero) then we would have to use the two different regressions above: one for the first manufacturer and the second for the other. Their difference would be in 5.42 units, corresponding to the new $\beta_{00} + \beta_{11}$ intercept (due to the added coefficient from the dummy variable "dumlevel" to the second equation).

Finally, and also for illustration, we investigate the alternative that both manufacturers differ, in level and in rate. This implies that the alternative (full) regression model is:

$$Y_i = \beta_{00} + \beta_{01} E_i + \beta_{10} X_{i1} + \beta_{11} (E_i X_{i1}) + \varepsilon_i \;\; ; \;\; 1 \le i \le n$$

The null hypothesis that both regression models (manufacturers) do not differ is now expressed as: H_0 $\beta_{01}=\beta_{11}=0$. The alternative hypothesis (H_1) is that at least one of the two (dummy variable) coefficients, β_{01} or β_{11} differs from zero, i.e. that both regressions differ in level (intercept), or in rate (slope) or in both.

The above regression model results are shown below. Variable "interdum" corresponds to the interaction (or product) of the dummy variable and the material thickness ($E_i X_{i1}$).

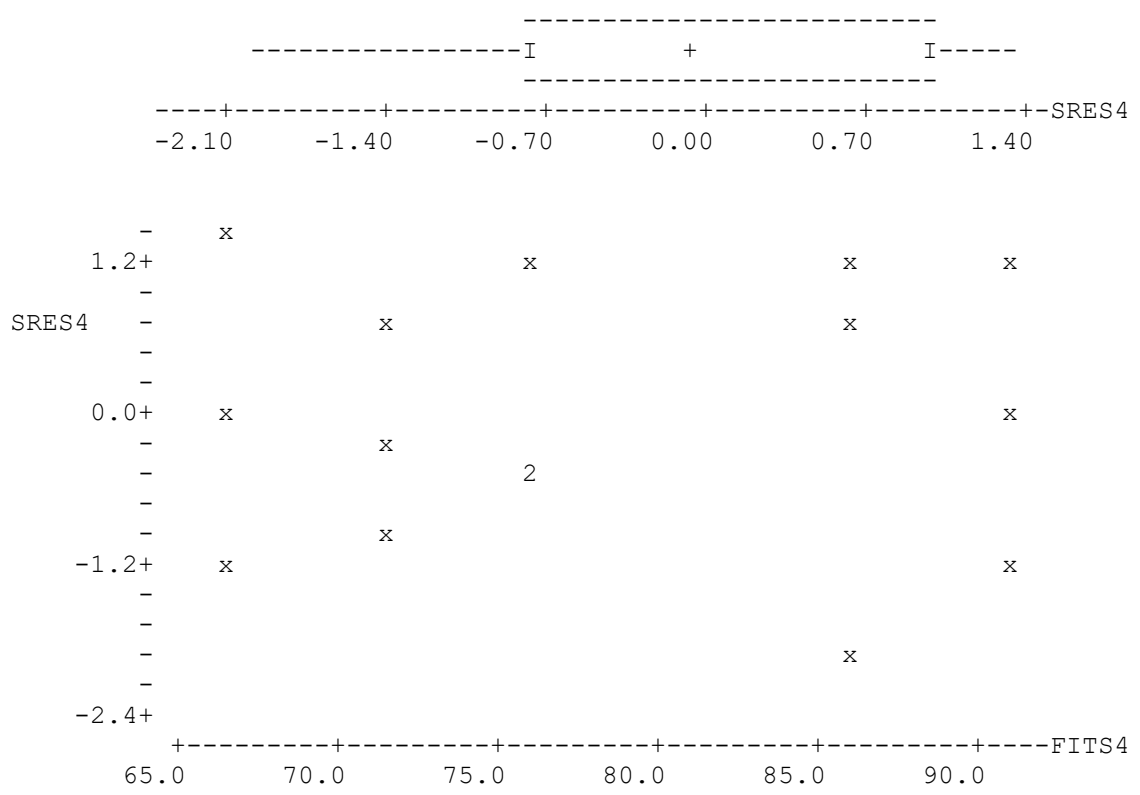
```
      newc3 = 62.0 + 4.65 thickness + 3.2 dumlevel + 0.55 interdum

Predictor       Coef       Stdev    t-ratio        p
Constant      61.978       2.581      24.02    0.000
thickness      4.650       1.195       3.89    0.003
dumlevel        3.22       11.12       0.29    0.777
interdum       0.550       2.671       0.21    0.841
s = 2.926       R-sq = 93.0%     R-sq(adj) = 91.1%
```

Notice here too, how the index of fit (model explanation) has barely increased by 2% (from 91 to 93%) and that only the variable "thickness" is now statistically significant. The coefficients of the remaining two variables ("dumlevel" and "interdum") have unacceptably high p-values (0.77 and 0.84), something also suggested by the regression ANOVA table results, shown below. Hence, these two variables are assumed zero.

Such a situation of having a significant full regression equation (F=48.69) with some non-significant individual coefficients (the t-ratio statistics) suggests that there are redundant variables in the model.

```
                          Analysis of Variance
SOURCE       DF          SS          MS         F        p
Regression    3     1250.51      416.84     48.69    0.000
batch         1     1223.69     1223.69
dumlevel      1       26.46       26.46
interdum      1        0.36        0.36
Error        11       94.18        8.56
Total        14     1344.69
```

As usual, before using the regression model results, we check the validity of the model assumptions via the residual analysis. Below, we show several plots. The plot of residuals vs. fits look random and show no variance problems. The residual boxplot and the AD GoF tests for Normality are also acceptable. We thus assume that the regression model is valid and use the results, namely that we do not reject the above stated null hypothesis.



Since we accept as valid the regression model results, it means that we do not have enough grounds to reject the null hypothesis. Hence the regression of tensile strength on thickness, for both manufacturers, does not differ either in level or in rate or in both.

Consequently, we assume that a single regression model (the first one above) acceptably represents the problem situation for both manufacturers. Therefore, we can use a single regression to estimate tensile strengths on material thickness, for both manufacturers.

Summary and Conclusions

In this chapter we have developed four materials analysis case studies that illustrate how regression models are used and some times even misused. Some of these regression models were linear and others were quadratic and cubic. We presented ways of comparing them in order to select the model that best captures the structure of the problem under study. Finally, and most important, we presented detailed ways of checking, via the residual analysis, graphically and analytically, the validity of the three main regression model assumptions. These are that residuals are Normal, independent and homoscedastic (have equal variance).

Several important caveats regarding statistical modeling in general and regression modeling in particular were discussed. The most important caveat is that the statistical model should always follow reality and not the other way around. If care is not taken, then we may end up modeling the data and not the problem. The only thing we will have

accomplished, at that stage, is to make things worse. For, the next data set, from the same problem area, may have little in common with the work we have performed, nor with the results we have obtained before.

Additional Suggested Readings

Box, G.E.P., W. G. Hunter and J. S. Hunter. Statistics for Experimenters. John Wiley, NY. 1978.
Chatterjee, S. and B. Price. Regression Analysis by Example. John Wiley, NY. 1977.
Draper, N. and H. Smith. Applied Regression Analysis. John Wiley, NY. 1980.
Dixon, W. J. and F. J. Massey. Introduction to Statistical Analysis. McGraw Hill, NY. 1983.