# Example G  Cost of construction of nuclear power plants

## Description of data

Table G.1 gives data, reproduced by permission of the Rand Corporation, from a report (*Mooz, 1978*) on 32 light water reactor (LWR) power plants constructed in USA. It is required to predict the capital cost involved in the construction of further LWR power plants. The notation used in Table G.1 is explained in Table G.2. The final 6 lines of data in Table G.1 relate to power plants for which there were partial turnkey guarantees and for which it is possible that some manufacturer's subsidies might be hidden in the quoted capital costs.

Table G.1 Data on thirty-two LWR power plants in USA

| | C | D | T1 | T2 | S | PR | NE | CT | BW | N | PT |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 460.05 | 68.58 | 14 | 46 | 687 | 0 | 1 | 0 | 0 | 14 | 0 |
| 2 | 452.99 | 67.33 | 10 | 73 | 1065 | 0 | 0 | 1 | 0 | 1 | 0 |
| 3 | 443.22 | 67.33 | 10 | 85 | 1065 | 1 | 0 | 1 | 0 | 1 | 0 |
| 4 | 652.32 | 68.00 | 11 | 67 | 1065 | 0 | 1 | 1 | 0 | 12 | 0 |
| 5 | 642.23 | 68.00 | 11 | 78 | 1065 | 1 | 1 | 1 | 0 | 12 | 0 |
| 6 | 345.39 | 67.92 | 13 | 51 | 514 | 0 | 1 | 1 | 0 | 3 | 0 |
| 7 | 272.37 | 68.17 | 12 | 50 | 822 | 0 | 0 | 0 | 0 | 5 | 0 |
| 8 | 317.21 | 68.42 | 14 | 59 | 457 | 0 | 0 | 0 | 0 | 1 | 0 |
| 9 | 457.12 | 68.42 | 15 | 55 | 822 | 1 | 0 | 0 | 0 | 5 | 0 |
| 10 | 690.19 | 68.33 | 12 | 71 | 792 | 0 | 1 | 1 | 1 | 2 | 0 |
| 11 | 350.63 | 68.58 | 12 | 64 | 560 | 0 | 0 | 0 | 0 | 3 | 0 |
| 12 | 402.59 | 68.75 | 13 | 47 | 790 | 0 | 1 | 0 | 0 | 6 | 0 |
| 13 | 412.18 | 68.42 | 15 | 62 | 530 | 0 | 0 | 1 | 0 | 2 | 0 |
| 14 | 495.58 | 68.92 | 17 | 52 | 1050 | 0 | 0 | 0 | 0 | 7 | 0 |
| 15 | 394.36 | 68.92 | 13 | 65 | 850 | 0 | 0 | 0 | 1 | 16 | 0 |
| 16 | 423.32 | 68.42 | 11 | 67 | 778 | 0 | 0 | 0 | 0 | 3 | 0 |
| 17 | 712.27 | 69.50 | 18 | 60 | 845 | 0 | 1 | 0 | 0 | 17 | 0 |
| 18 | 289.66 | 68.42 | 15 | 76 | 530 | 1 | 0 | 1 | 0 | 2 | 0 |
| 19 | 881.24 | 69.17 | 15 | 67 | 1090 | 0 | 0 | 0 | 0 | 1 | 0 |
| 20 | 490.88 | 68.92 | 16 | 59 | 1050 | 1 | 0 | 0 | 0 | 8 | 0 |
| 21 | 567.79 | 68.75 | 11 | 70 | 913 | 0 | 0 | 1 | 1 | 15 | 0 |
| 22 | 665.99 | 70.92 | 22 | 57 | 828 | 1 | 1 | 0 | 0 | 20 | 0 |
| 23 | 621.45 | 69.67 | 16 | 59 | 786 | 0 | 0 | 1 | 0 | 18 | 0 |
| 24 | 608.80 | 70.08 | 19 | 58 | 821 | 1 | 0 | 0 | 0 | 3 | 0 |
| 25 | 473.64 | 70.42 | 19 | 44 | 538 | 0 | 0 | 1 | 0 | 19 | 0 |
| 26 | 697.14 | 71.08 | 20 | 57 | 1130 | 0 | 0 | 1 | 0 | 21 | 0 |
| 27 | 207.51 | 67.25 | 13 | 63 | 745 | 0 | 0 | 0 | 0 | 8 | 1 |
| 28 | 288.48 | 67.17 | 9 | 48 | 821 | 0 | 0 | 1 | 0 | 7 | 1 |
| 29 | 284.88 | 67.83 | 12 | 63 | 886 | 0 | 0 | 0 | 1 | 11 | 1 |
| 30 | 280.36 | 67.83 | 12 | 71 | 886 | 1 | 0 | 0 | 1 | 11 | 1 |
| 31 | 217.38 | 67.25 | 13 | 72 | 745 | 1 | 0 | 0 | 0 | 8 | 1 |
| 32 | 270.71 | 67.83 | 7 | 80 | 886 | 1 | 0 | 0 | 1 | 11 | 1 |

Table G.2 Notation for data of Table G.2

| | |
|---|---|
| C | Cost in dollars $\times 10^{-6}$, adjusted to 1976 base |
| D | Date construction permit issued |
| T1 | Time between application for and issue of permit |
| T2 | Time between issue of operating license and construction permit |
| S | Power plant net capacity (MWe) |
| PR | Prior existence of an LWR on same site (=1) |
| NE | Plant constructed in north-east region (=1) |
| CT | Use of cooling tower (=1) |
| BW | Nuclear steam supply system manufactured by Babcock-Wilcox (=1) |
| N | Cumulative number of power plants constructed by each architect-engineer |
| PT | Partial turnkey plant (=1) |

## General considerations

One of the most common problems in advanced applied statistics is the study of the relation between a single continuous response variable and a number of explanatory variables. When the expected response can be represented as a linear combination of unknown parameters, with coefficients determined by the

explanatory variables, and when the error structure is suitably simple, the techniques of multiple regression based on the method of least squares are applicable. The formal theory of multiple regression, and the associated significance tests and confidence regions, have been extensively developed; see, for example, Draper and Smith (1981) and Seber (1977). Further, computer programs for implementing the methods are widely available.

Nevertheless, there can be difficulties, partly of technique but more important of interpretation, in applying methods, especially to observational data with fairly large number of explanatory variable. We now mention briefly some commonly occurring points. Of course, in any particular application many of the potential difficulties may be absent and indeed the present example seems relatively well behaved.

Some issues that arise fairly commonly are the following:

(i) What is the right general form of model to fit?
(ii) Are there aspects of error structure that seriously affect the analysis?
(iii) Are there outliers or anomalous observations that need to be isolated?
(iv) What can be done if a subset of observations is isolated, possibly not following the same model as the main body of data?
(v) Is it feasible to simplify the model, normally by reducing the number of explanatory variables?
(vi) What are the limitations on the interpretation and application of the final relation achieved?

All these points, except (vi), can to some extent be dealt with formally, for instance, by comparing the fits of numerous competing models. Often, though, this would be a ponderous way to proceed.

Considerations of point (i), choice of form of relation, involves a possible transformation of response variable, in the present instance `cost` and `log(cost)` being two natural variables for analysis, and a choice of the nature and form of the explanatory variables. For instance, should the explanatory variables, where quantitative, be transformed? Should derived explanatory variables be formed to investigate interactions? Frequently in practice, any transformations are settled on the basis of general experience: the need for interaction terms may be examined graphically or, especially with large numbers of explanatory variables, may be checked by looking only for interactions between variables having large '*main effects*'. In the present example, `log(cost)` has been taken as response variable and the explanatory variables `S`, `T1`, `T2` and `N` have also taken in `log` form, partly to lead to unit-free parameters whose values can be interpreted in terms of power-law relations between the original variables. It is plausible that random variations in cost should increase with the value of cost and this is another reason for log transformation.

Complexities of error structure, point (ii), can arise via systematic changes in variance, via notable non-normality of distribution and, particularly importantly, via correlation in errors for different individuals. All these effects may be of intrinsic interest, but more commonly have to be considered either because a modification of the method of least squares is called for or because, while the least-squares estimates and fit may be satisfactory, the precision of the least-squares estimates may be different from that indicated under standard assumptions. In particular, substantial presence of positive correlations can mean that the least-squares estimates are much less precise than standard formulae suggest. A special form of correlated error structure is that of clustering of individuals into groups, the regression relations between and within groups being quite different. There is no sign that any of these complications are important in the present instance.

Somewhat related is point (iii), occurrence of outliers. Where interest is focused on particular regression coefficients, the most satisfactory approach is to examine informally or formally whether there is any single observation or a small set of observations whose omission would greatly change the estimate in question; see also point (iv).

In the present example, there is a group of 6 observations distinct from the main body of 26 and there is some doubt whether the 6 should be included. This is quite a common situation; the possibly anomalous group may, for example, have extreme values of certain explanatory variables. The most systematic approach is to fit additional linear models to test consistency. Thus one extra parameter can be fitted to allow for a constant displacement of the anomalous group and the significance of the resulting estimate tested. A more searching analysis is provided by allowing the regression coefficients also to be different in the anomalous group; in the present instance this has been done one variable at a time, because with 10 explanatory variables and only 6 observations in the separate group there are insufficient observations to examine for anomalous slopes simultaneously.

Point (v), the simplification of the fitted model, is particularly important when the number of explanatory variables is large, and even more particularly when there is an *a priori* suspicion that many of the explanatory variables are measuring essentially equivalent things. The need for such simplification arises particularly, although by no means exclusively, in observational studies. More explicitly, the reasons for seeking simplification are that:

(a) estimation of parameters of clear interest can be degraded by including unnecessary terms in the model;
(b) prediction of response of new individuals is less precise if unnecessary terms are included in the predictor;
(c) it is often reasonable to expect that when explanatory variables are available only a few will have a major effect on response and it may be of primary interest to isolate these important variables and to interpret their effects;
(d) it may be desirable to simplify future work by recording a smaller number of explanatory variables.

Techniques for the retention of variables are, as explained in Section 3.4 of Part I, forward, backward or some mixture. Where some of the parameters represent effects of direct interest they should be included regardless of operation of a selection procedure. It is entirely possible that forward selection leads to a different equation from backward selection, although this has not happened in the present example. It is therefore important, especially where interpretation of the particular form of equation is central to the analysis, that if there are several simple equations that fit almost equally well, all should be isolated for consideration and not one chosen somewhat arbitrarily.

Suppose now that a representation, hopefully quite a simple one, has been obtained for expected response as a function of certain explanatory variables. What are the principal aspects in using and interpreting such an equation? This is point (vi) of the list above. There are at least five rather different possibilities.

Firstly, an equation such as that summarized in Table G.4, including the residual standard deviation, provides a concise description of the data, as regards the dependence of cost on the other variables. Such a description can be useful in thinking about the data quantitatively and in comparing different, somewhat related, sets of data.

A second descriptive use is in the study of the individual cases. The residual from the fitted model is an index for each power station assessing its cost relative to what might have been anticipated given the explanatory variables.

Thirdly, the equation can be used for prediction. A new individual has given (or sometimes predicted) values of the relevant explanatory variables and the equation, and the associated measures of variability,

are used to forecast cost, preferably with confidence limits. In such prediction the main assumption, in addition to the technical adequacy of the model in the region of explanatory variables required for prediction, is that any unmeasured variable affecting response keeps the same statistical relationship with the measured explanatory variables as obtains in the data. Thus, in particular, if the new individual to be predicted differs in some way from the reference data, other than is directly or indirectly accounted for in the explanatory variables, a modification of the regression predictor is worth consideration. For example, a major technological innovation between the data analyzed and the individual to be predicted would call for such modification of the predictor.

Fourthly, the equation may be used to predict for a new individual, or sometimes for one of original individuals, the consequences of changes in one or more of the explanatory variables. For example, one might wish to predict not so much the cost for a new individual as the change in cost for that individual as size changes. The relevant regression coefficient predicts the change, provided that the other explanatory variables are held fixed and that any important unobserved explanatory variables change appropriately with change in size. The prediction of changes in uncontrolled observational systems, e.g. the social sciences, needs particularly careful specification of the changes in explanatory variables envisaged.

Finally, and in some ways most importantly, one may hope to gain insight into the system under study by careful inspection of which explanatory variables contribute appreciably to the response and of the signs and the magnitude of the associated regression coefficients. Thus in the present example, why do certain variables appear not to contribute appreciably, why is the regression coefficient on `log(size)` appreciably less than `1`, the value for proportionality, and so on? As indicated in the previous paragraph, the regression coefficients estimate changes in response under perturbations of the system whose precise specification needs care.

The last two applications of the regressions need considerable thought, especially if there is any possibility that an important explanatory variable has been overlooked.

**The analysis**

As explained in the preceding section, we take `log(C)`, `log(S)`, `log(N)`, `log(T1)` and `log(T2)`; throughout natural logs are used.

A regression of `log(C)` on all `10` explanatory variables gives a residual mean square of `0.5680/21=0.0271` with `21` degree of freedom. Elimination of insignificant variables successively one at a time removes `BW`, `log(T1)`, `log(T2)` and `PR` (`Table G.3`), leaving `6` variables and a residual mean square of `0.0253` with `25` degrees of freedom; the residual standard deviation is `0.159`.

Table G.3 Elimination of variables

| No. variables included | variables eliminated | Residuals ss | df | ms |
|---|---|---|---|---|
| 10 | - | 0.56803 | 21 | 0.02705 |
| 9 | BW | 0.57091 | 22 | 0.02595 |
| 8 | log(T1) | 0.57300 | 23 | 0.02491 |
| 7 | log(T2) | 0.61654 | 24 | 0.02569 |
| 6 | PR | 0.63374 | 25 | 0.02535 |

None of the eliminated variables is significant if re-introduced. The estimated coefficients and standard errors for the six-variable regression are given in `Table G.4`. The variable `PT`, denoting partial turnkey

guarantee, has a coefficient of -0.0261, with a standard error of 0.01135(25df), suggesting that cost tends to be reduced on average by about 20% for these 6 plants.

Table G.4 Multiple regression: full and reduced models

| Variables | Regression coefficient | | | |
|---|---|---|---|---|
| | Reduced model | | Full model | |
| | Estimate | Std. Error | Estimate | Std. Error |
| Constant | -13.26031 | 3.13950 | -14.24198 | 4.22880 |
| PT | -0.22610 | 0.11355 | -0.22429 | 0.12246 |
| CT | 0.14039 | 0.06042 | 0.12040 | 0.06632 |
| log(N) | -0.08758 | 0.04147 | -0.08020 | 0.04596 |
| log(S) | 0.72341 | 0.11882 | 0.69373 | 0.13605 |
| D | 0.21241 | 0.04326 | 0.20922 | 0.06526 |
| NE | 0.24902 | 0.07414 | 0.25807 | 0.07693 |
| log(T1) | - | - | 0.09187 | 0.24396 |
| log(T2) | - | - | 0.28553 | 0.27289 |
| PR | - | - | -0.09237 | 0.07730 |
| BW | - | - | 0.03303 | 0.10112 |
| Residual st.dev | 0.1592 (25 df) | | 0.1645 (21 df) | |

To check whether these 6 plants and the 26 others can be fitted by a model with common coefficients for each of the variables CT, log(N), log(S) and D, we include in turn in the regression the interaction of each variable with PT. This cannot be done for the variable NE since all 6 PT plants were constructed in the same region. Table G.5 summarizes the results. None of the interaction coefficients is significant.

Table G.5. Regression including interaction with PT

| | Z = CT | | Z = lon(N) | | Z = log(S) | | Z = D | |
|---|---|---|---|---|---|---|---|---|
| Variable | Estimate | s.e. | Estimate | s.e. | Estimate | s.e. | Estimate | s.e. |
| Constant | -13.23435 | 3.19296 | -13.258896 | 3.225077 | -13.08645 | 3.23858 | -13.22438 | 3.23096 |
| PT | -0.24289 | 0.12210 | -0.229274 | 0.826544 | -2.18759 | 5.85357 | -1.52852 | 15.17047 |
| CT | 0.13123 | 0.06515 | 0.140440 | 0.062872 | 0.13998 | 0.06154 | 0.14120 | 0.06237 |
| log(N) | -0.08680 | 0.04221 | -0.087574 | 0.042334 | -0.08683 | 0.04229 | -0.08749 | 0.04234 |
| log(cap) | 0.72291 | 0.12083 | 0.723359 | 0.121937 | 0.71761 | 0.12222 | 0.72217 | 0.12210 |
| D | 0.21213 | 0.04399 | 0.212398 | 0.044348 | 0.21044 | 0.04444 | 0.21201 | 0.04440 |
| NE | 0.24899 | 0.07539 | 0.249020 | 0.075679 | 0.24841 | 0.07551 | 0.24889 | 0.07567 |
| PT*Z | 0.07976 | 0.18867 | 0.001427 | 0.368278 | 0.29159 | 0.87002 | 0.01928 | 0.22459 |

We note that the coefficients of the 6 common variables in the Table G.5 remain fairly stable, except for PT which, in two cases, is estimated very imprecisely. A model with common coefficients as given in Table G.4 seems reasonable. With this model the predicted cost increases with size, although less rapidly than proportionally to size, is further increased if a cooling tower is used or if constructed in the NE region, but decreases with experience of architect-engineer.

Fitted values and residuals are given in Table G.6. The residuals give no evidence of outliers or of any systematic departure from the assumed model; this can be checked by plotting in the usual ways, against the explanatory variables, for example, against D (Fig G.1) and log(S) (Fig G.2), against fitted values (Fig G.3), and normal order statistics (FigG.4).

The setimated standard error of predicted log(cost) for a new power plant, provided conditions are fairly close to the avarage of the the observed 32 plants, is approximately $0.159(1+1/32)^{1/2}=0.161$ with 25 degrees of freedom. Thus there is a 95% chance that the actual cost for the new plant will lie within about $\pm$39% of the predicted cost.

Table G.6 Comparison of observed and fitted values based on 6-variable regression of Table G.4 fitted to log(C)

| | Observed | Fitted | Residual | | Observed | Fitted | Residual |
|---|---|---|---|---|---|---|---|
| 1 | 6.13134 | 6.05053 | 0.08081 | 17 | 6.56846 | 6.37869 | 0.18976 |
| 2 | 6.11587 | 6.22464 | -0.10877 | 18 | 5.66871 | 5.89063 | -0.22192 |
| 3 | 6.09407 | 6.22464 | -0.13057 | 19 | 6.78133 | 6.49187 | 0.28946 |
| 4 | 6.48054 | 6.39836 | 0.08218 | 20 | 6.19620 | 6.22961 | -0.03341 |
| 5 | 6.46495 | 6.39836 | 0.06659 | 21 | 6.34175 | 6.17770 | 0.16405 |
| 6 | 5.84467 | 5.97577 | -0.13109 | 22 | 6.50128 | 6.65139 | -0.15011 |
| 7 | 5.60716 | 5.93437 | -0.32721 | 23 | 6.43206 | 6.24881 | 0.18325 |
| 8 | 5.75956 | 5.70374 | 0.05583 | 24 | 6.41149 | 6.38393 | 0.02756 |
| 9 | 6.12495 | 5.98747 | 0.13748 | 25 | 6.16045 | 6.12914 | 0.03131 |
| 10 | 6.53697 | 6.41112 | 0.12585 | 26 | 6.54699 | 6.79742 | -0.25043 |
| 11 | 5.85973 | 5.78855 | 0.07119 | 27 | 5.33518 | 5.40053 | -0.06535 |
| 12 | 5.99792 | 6.26190 | -0.26398 | 28 | 5.66463 | 5.60589 | 0.05873 |
| 13 | 6.02146 | 5.89063 | 0.13083 | 29 | 5.65207 | 5.62123 | 0.03084 |
| 14 | 6.20573 | 6.24130 | -0.03558 | 30 | 5.63607 | 5.62123 | 0.01484 |
| 15 | 5.97726 | 6.01604 | -0.03878 | 31 | 5.38165 | 5.40053 | -0.01888 |
| 16 | 6.04813 | 5.99241 | 0.05572 | 32 | 5.60105 | 5.62123 | -0.02018 |

*Fig G.1 Residuals of log(C) from6-variable model vs D,date construction permit issued*

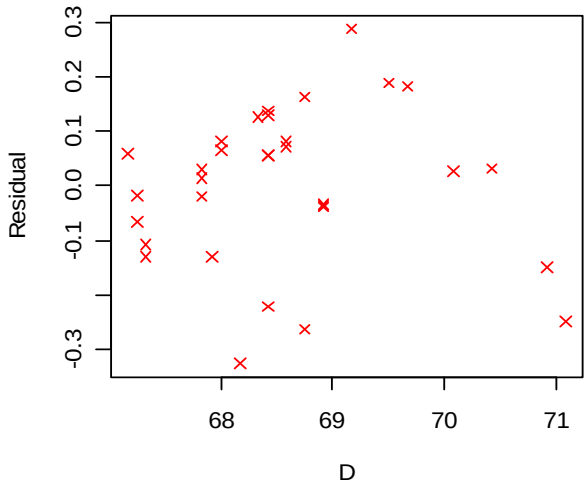



*Fig G.2 Residuals of log(C) from6-variable model vs log(S),power plant net capacity*

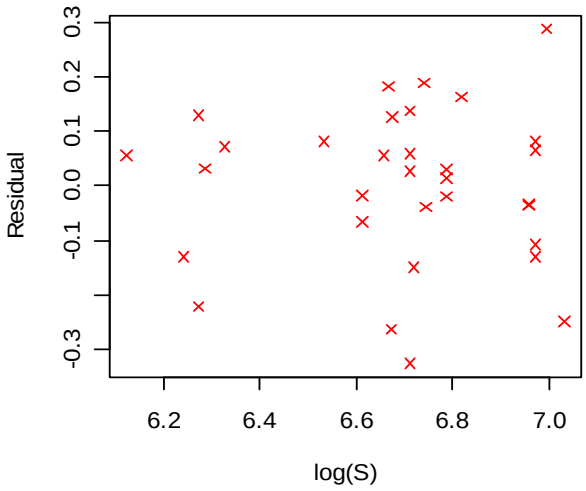
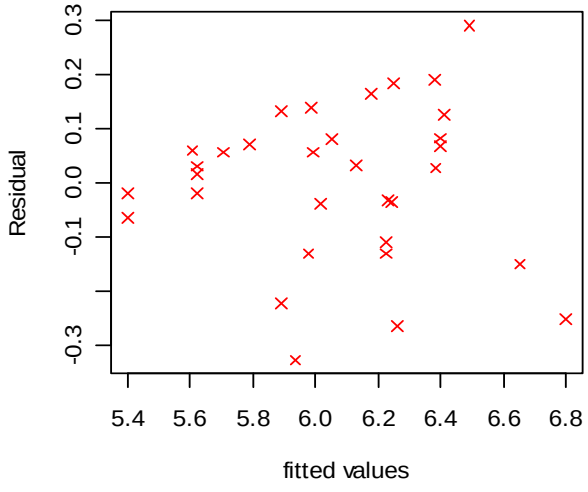



*Fig G.3 Residuals of log(C) from 6-variable model vs fitted values*



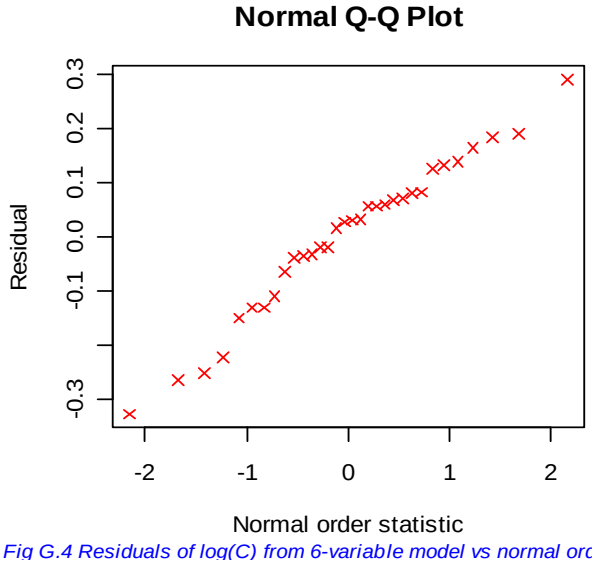




*Fig G.4 Residuals of log(C) from 6-variable model vs normal order sta*