ESI 6247 Statistical Design Models

Hui Yang, PhD

Assistant Professor

Department of Industrial and Management
Systems Engineering

University of South Florida

Spring 2010

Unit 1 : Basic Concepts and Introductory Regresssion Analysis

Sources : Chapter 1.

- Historical perspectives and basic definitions (Section 1.1).
- Planning and implementation of experiments (Section 1.2).
- Fisher's fundamental principles (Section 1.3).
- Simple linear regression (Sections 1.4-1.5).
- Multiple regression, variable selection (Sections 1.6-1.7).
- Example: Air Pollution Data (Section 1.8).

Historical perspectives

- **Agricultural Experiments :** Comparisons and selection of varieties (and/or treatments) in the presence of uncontrollable field conditions, Fisher's pioneering work on design of experiments and analysis of variance (ANOVA).
- **Industrial Era :** Process modeling and optimization, Large batch of materials, large equipments, Box's work motivated in chemical industries and applicable to other processing industries, regression modeling and response surface methodology.

Historical perspectives (Contd.)

- **Quality Revolution :** Quality and productivity improvement, variation reduction, total quality management, Taguchi's work on robust parameter design, Six-sigma movement.
- A lot of successful applications in manufacturing (cars, electronics, home appliances, etc.)
- **Current Trends and Potential New Areas :** Computer modelling and experiments, large and complex systems, applications to biotechnology, nanotechnology, material development, etc.

Types of Experiments

- **Treatment Comparisons :** Purpose is to compare several treatments of a factor (have 4 rice varieties and would like to see if they are different in terms of yield and draught resistence).
- **Variable Screening :** Have a large number of factors, but only a few are important. Experiment should identify the important few.
- **Response Surface Exploration :** After important factors have been identified, their impact on the system is explored; regression model building.

Types of Experiments (Contd.)

- **System Optimization :** Interested in determining the optimum conditions (e.g., maximize yield of semiconductor manufacturing or minimize defects).
- **System Robustness :** Wish to optimize a system and also reduce the impact of uncontrollable (noise) factors. (e.g., would like cars to run well in different road conditions and different driving habits; an IC fabrication process to work well in different conditions of humidity and dust levels).

Some Definitions

- **Factor :** variable whose influence upon a response variable is being studied in the experiment.
- **Factor Level :** numerical values or settings for a factor.
- **Trial** (or **run**) : application of a treatment to an experimental unit.
- **Treatment or level combination :** set of values for all factors in a trial.
- **Experimental unit :** object to which a treatment is applied.
- **Randomization :** using a chance mechanism to assign treatments to experimental units or run order.

Systematic Approach to Experimentation

- State the objective of the study.
- Choose the response variable . . . should correspond to the purpose of the study.
    - Nominal-the-best, larger-the-better or smaller-the-better.
- Choose factors and levels.
    - Use flow chart or cause-and-effect diagram (see Figure 1).
- Choose experimental design (i.e., plan).
- Perform the experiment (use a planning matrix to determine the set of treatments and the order to be run).
- Analyze data (design should be selected to meet objective so that the analysis is efficient and easy).
- Draw conclusions.

Cause-and Effect Diagram

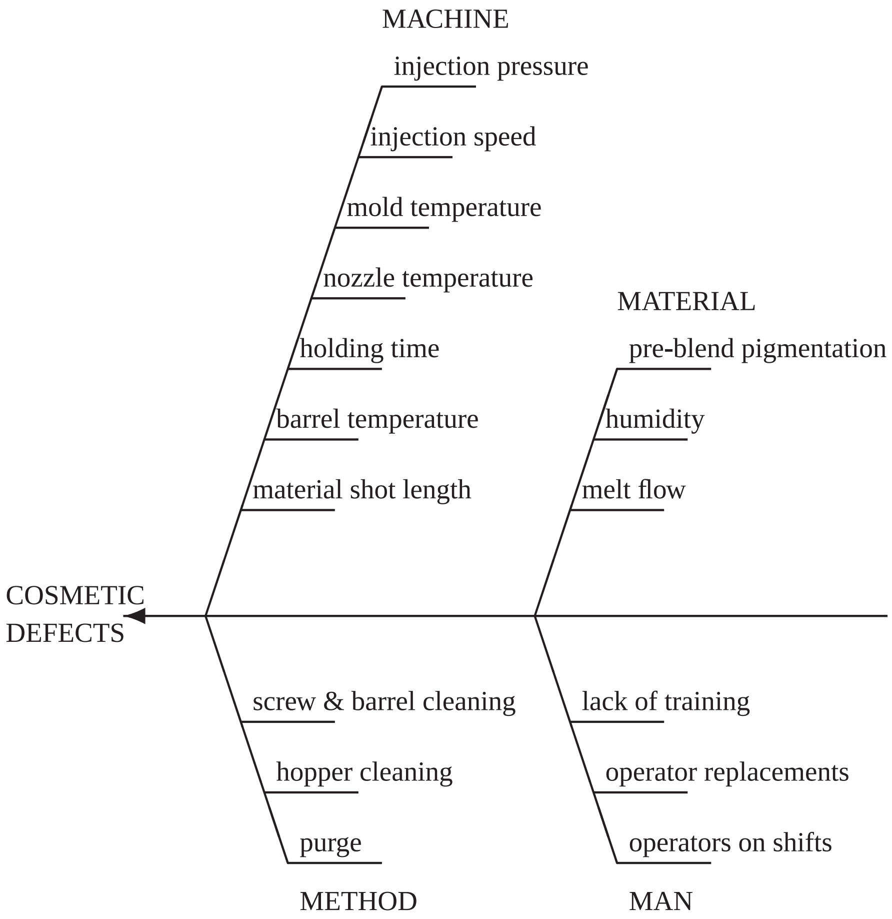


Figure 1: Cause-and-Effect Diagram, Injection Molding Experiment

Choice of Response : An Example

- To improve a process that often produces underweight soap bars. Obvious choice of response, y = soap bar weight.
- There are two sub-processes : (i) mixing, which affects soap bar density ($=y_1$), (ii) forming, which affects soap bar dimensions ($=y_2$).
- Even though y is a function of y_1 and y_2, better to study y_1 and y_2 separately and identify factors important for each of the two sub-processes.

Fundamental Principles : Replication, randomization, and blocking

Replication

- Each treatment is applied to units that are representative of the population (example : measurements of 3 units vs. 3 repeated measurements of 1 unit).
- Replication vs Repetition (i.e., repeated measurements).
- Enable the estimation of experimental error. Use sample standard deviation.
- Decrease variance of estimates and increase the power to detect significant differences : for independent y_i's,

$$Var(\frac{1}{N}\sum_{i=1}^{N} y_i) = \frac{1}{N} Var(y_1).$$

Randomization

Use of a chance mechanism (e.g., random number generators) to assign treatments to units or to run order. It has the following advantages.

- Protect against latent variables or “lurking” variables (give an example).
- Reduce influence of subjective bias in treatment assignments (e.g., clinical trials).
- Ensure validity of statistical inference (This is more technical; will not be discussed in the book. See Chapter 4 of “*Statistics for Experimenters*” by Box, Hunter, Hunter for discussion on randomization distribution.)

Blocking

A **block** refers to a collection of homogeneous units. Effective blocking : larger between-block variations than within-block variations.
(Examples: hours, batches, lots, street blocks, pairs of twins.)

- Run and compare treatments within the same blocks. (Use randomization within blocks.) It can eliminate block-block variation and reduce variability of treatment effects estimates.
- **Block what you can and randomize what you cannot.**
- Discuss **typing experiment** to demonstrate possible elaboration of the blocking idea. See next page.

Illustration: Typing Experiment

- To compare two keyboards A and B in terms of typing efficiency. Six manuscripts 1-6 are given to the same typist.
- Several designs (i.e., orders of test sequence) are considered:

1.

$$1.\ A, B,\ \ 2.\ A, B,\ \ 3.\ A, B,\ \ 4.\ A, B,\ \ 5.\ A, B,\ \ 6.\ A, B.$$

   (A always followed by B, why bad ?)

2. Randomizing the order leads to a new sequence like this

$$1.\ A, B,\ \ 2.\ B, A,\ \ 3.\ A, B,\ \ 4.\ B, A,\ \ 5.\ A, B,\ \ 6.\ A, B.$$

   (an improvement, but there are four with A, B and two with B, A. Why is this not desirable? Impact of *learning effect*.)

3. *Balanced randomization*: To mitigate the learning effect, randomly choose three with A, B and three with B, A. (Produce one such plan on your own).
4. Other improved plans?

Simple Linear Regression : Mortality Data

The data, taken from certain regions of Great Britain, Norway, and Sweden contains the mean annual temperature (in degrees F) and mortality index for neoplasms of the female breast.

Mortality rate (M)	102.5	104.5	100.4	95.9	87.0	95.0	88.6	89.2
Temperature (T)	51.3	49.9	50.0	49.2	48.5	47.8	47.3	45.1
Mortality rate (M)	78.9	84.6	81.7	72.2	65.1	68.1	67.3	52.5
Temperature (T)	46.3	42.1	44.2	43.5	42.3	40.2	31.8	34.0

Objective : Obtaining the relationship between mean annual temperature and the mortality rate for a type of breast cancer in women.

Getting Started

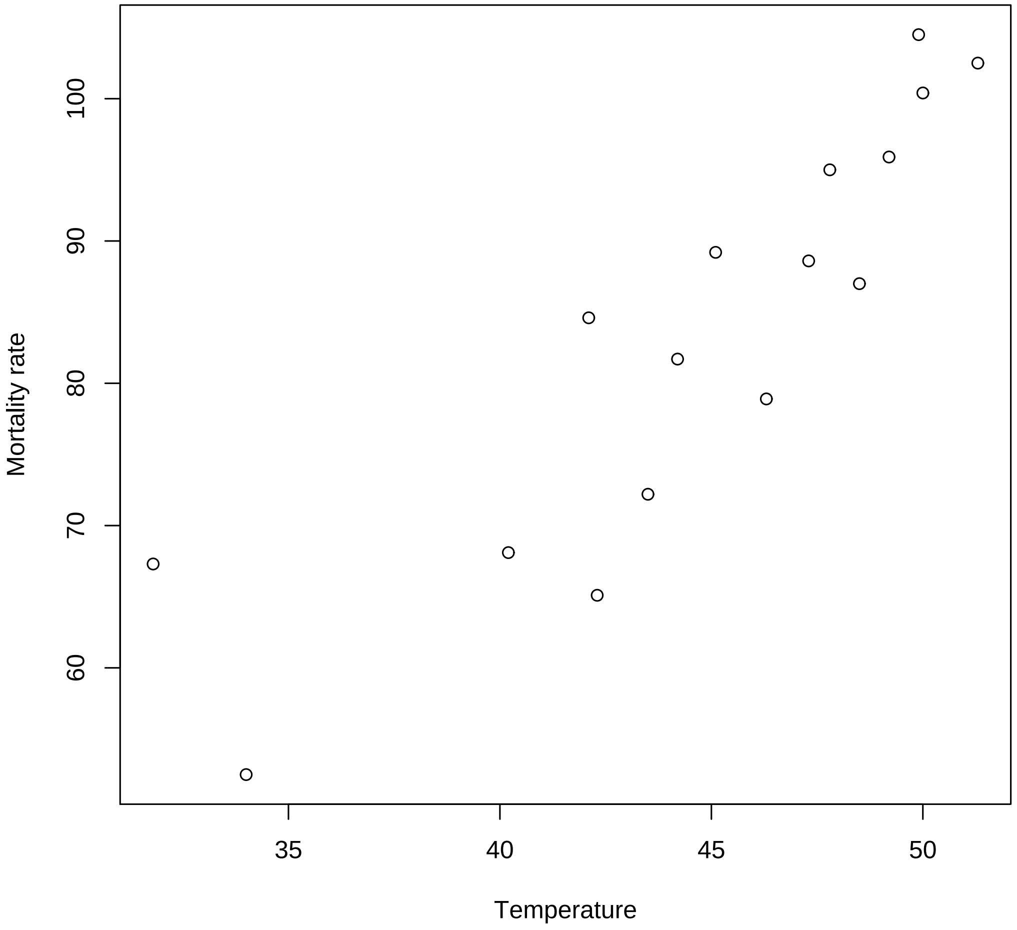


Figure 2: Scatter Plot of Temperature versus Mortality Rate, Breast Cancer Data.

Fitting the Regression Line

- Underlying Model :

$$y = \beta_0 + \beta_1 x + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2).$$

- Coefficients are estimated by minimizing

$$\sum_{i=1}^{N} \Big(y_i - (\beta_0 + \beta_1 x_i)\Big)^2.$$

- **Least Squares Estimates**

  Estimated Coefficients :

$$\hat{\beta}_1 = \frac{\sum(x_i - \overline{x})(y_i - \overline{y})}{\sum(x_i - \overline{x})^2} \quad , \quad var\left(\hat{\beta}_1\right) = \frac{\sigma^2}{\sum(x_i - \overline{x})^2},$$

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x} \quad , \quad var\left(\hat{\beta}_0\right) = \sigma^2 \left(\frac{1}{N} + \frac{\overline{x}^2}{\sum(x_i - \overline{x})^2)}\right),$$

$$\overline{x} = \frac{1}{N} \sum x_i \quad , \quad \overline{y} = \frac{1}{N} \sum y_i.$$

Explanatory Power of the Model

- The total variation in y can be measured by corrected total sum of squares $CTSS = \sum_{i=1}^{N}(y_i - \overline{y})^2$.

- This can be decomposed into two parts (Analysis of Variance (ANOVA)):

$$CTSS = RegrSS + RSS,$$

where

$$RegrSS = \text{Regression sum of squares} = \sum_{i=1}^{N}(\hat{y}_i - \overline{y})^2,$$

$$RSS = \text{Residual sum of squares} = \sum_{i=1}^{N}(y_i - \hat{y}_i)^2.$$

$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ is called the predicted value of y_i at x_i.

- $R^2 = \frac{RegrSS}{CTSS} = 1 - \frac{RSS}{CTSS}$ measures the proportion of variation in y explained by the fitted model.

ANOVA Table for Simple Linear Regression

ANOVA Table for Simple Linear Regression

	Degrees of	Sum of	Mean
Source	Freedom	Squares	Squares
regression	1	$\hat{\beta}_1 \sum (x_i - \overline{x})^2$	$\hat{\beta}_1 \sum (x_i - \overline{x})^2$
residual	$N-2$	$\sum_{i=1}^{N} (y_i - \hat{y}_i)^2$	$\frac{\sum_{i=1}^{N} (y_i - \hat{y}_i)^2}{(N-2)}$
total (corrected)	$N-1$	$\sum_{i=1}^{N} (y_i - \overline{y})^2$	

ANOVA Table for Breast Cancer Example

	Degrees of	Sum of	Mean
Source	Freedom	Squares	Squares
regression	1	2599.53	2599.53
residual	14	796.91	56.92
total (corrected)	15	3396.44	

t-Statistic

- To test the null hypothesis $H_0 : \beta_j = 0$ against the alternative hypothesis $H_0 : \beta_j \neq 0$, use the test statistic

$$t_j = \frac{\hat{\beta}_j}{s.d.(\hat{\beta}_j)}.$$

- The higher the value of t, the more significant is the coefficient.
- For 2-sided alternatives, p-value $=$ Prob$\left(|t_{df}| > |t_{obs}|\right)$, df = degrees of freedom for the t-statistic, t_{obs} = observed value of the t-statistic. If p-value is very small, then either we have observed something which rarely happens, or H_0 is not true. In practice, if p-value is less then $\alpha = 0.05$ or 0.01, H_0 is rejected at level α.

Confidence Interval

$100(1-\alpha)\%$ confidence interval for β_j is given by

$$\hat{\beta}_j \pm t_{N-2,\frac{\alpha}{2}} \times s.d.(\hat{\beta}_j),$$

where $t_{N-2,\frac{\alpha}{2}}$ is the upper $\alpha/2$ point of the t distribution with $N-2$ degrees of freedom.

If the confidence interval for β_j does not contain 0, then H_0 is rejected.

Predicted Values and Residuals

- $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ is the predicted value of y_i at x_i.
- $r_i = y_i - \hat{y}_i$ is the corresponding residual.
- Standardized residuals are defined as $\frac{r_i}{s.d.(r_i)}$.
- Plots of residuals are extremely useful to judge the “goodness” of fitted model.
  - Normal probability plot (will be explained in Unit 2).
  - Residuals versus predicted values.
  - Residuals versus covariate x.

Analysis of Breast Cancer Data

```
The regression equation is
M = - 21.79 + 2.36 T

Predictor        Coef     SE Coef          T        P
Constant       -21.79       15.67      -1.39    0.186
T              2.3577      0.3489       6.76    0.000

S = 7.54466     R-Sq = 76.5%      R-Sq(adj) = 74.9%

Analysis of Variance

Source            DF          SS          MS         F        P
Regression         1      2599.5      2599.5     45.67    0.000
Residual Error    14       796.9        56.9
Total             15      3396.4

Unusual Observations
Obs          T          M         Fit      SE Fit    Residual    St Resid
 15       31.8      67.30       53.18        4.85       14.12        2.44RX

R denotes an observation with a large standardized residual.
X denotes an observation whose X value gives it large leverage.
```

Outlier Detection

- Minitab identifies two types of outliers denoted by R and X:
  R: its standardized residual $(y_i - \hat{y}_i)/se(\hat{y}_i)$ is large.
  X: its X value gives large leverage (i.e., far away from majority of the X values).

- For the mortality data, the observation with T = 31.8, M = 67.3 (i.e., left most point in plot on p. 16) is identified as both R and X.

- After removing this outlier and refitting the remaining data, the output is given on p. 27. There is still an outlier identified as X but not R. This one (second left most point on p.16) should not be removed (why?)

- Residual plots on p. 28 show no systematic pattern.

Notes: Outliers are not discussed in the book, see standard regression texts. Residual plots will be discussed in unit 2.

Prediction from the Breast Cancer Data

- The fitted regression model is $Y = -21.79 + 2.36X$, where Y denotes the mortality rate and X denotes the temperature.
- The predicted mean of Y at $X = x_0$ can be obtained from the above model. For example, prediction for the temperature of 49 is obtained by substituting $x_0 = 49$, which gives $y_{x_0} = 93.85$.
- The standard error of y_{x_0} is given by

$$S.E.(y_{x_0}) = \hat{\sigma}\sqrt{\frac{1}{N} + \frac{(\bar{x} - x_0)^2}{\sum_{i=1}^{N}(x_i - \bar{x})^2}}.$$

- Here $x_0 = 49$, $1/N + (\bar{x} - x_0)^2 / \sum_{i=1}^{N}(x_i - \bar{x})^2 = 0.1041$, and $\hat{\sigma} = \sqrt{MSE} = 7.54$. Consequently, $S.E.(y_{x_0}) = 2.432$.

Confidence interval for mean and prediction interval for individual observation

- A 95% confidence interval for the mean response $\beta_0 + \beta_1 x_0$ of y at $x = x_0$ is

$$\hat{\beta}_0 + \hat{\beta}_1 x_0 \pm t_{N-2,0.025} \times S.E.(y_{x_0}).$$

- Here the 95% confidence interval for the mean mortality corresponding to a temperature of 49 is [88.63, 99.07].

- A 95% prediction interval for an individual observation y_{x_0} corresponding to $x = x_0$ is

$$\hat{\beta}_0 + \hat{\beta}_1 x_0 \pm t_{N-2,0.025} \hat{\sigma} \sqrt{1 + \frac{1}{N} + \frac{(\bar{x} - x_0)^2}{\sum_{i=1}^{N} (x_i - \bar{x})^2}},$$

  where 1 under the square root represents σ^2, variance of the *new* observation y_{x_0}.

- The 95% prediction interval for the predicted mortality of an individual corresponding to the temperature of 49 is [76.85, 110.85].

Regression Results after Removing the Outlier

```
The regression equation is
M = - 52.62 + 3.02 T

Predictor        Coef     SE Coef          T       P
Constant       -52.62       15.82      -3.33    0.005
T              3.0152      0.3466       8.70    0.000

S = 5.93258     R-Sq = 85.3%      R-Sq(adj) = 84.2%

Analysis of Variance

Source            DF          SS          MS         F        P
Regression         1      2664.3      2664.3     75.70    0.000
Residual Error    13       457.5        35.2
Total             14      3121.9

Unusual Observations
Obs          T          M         Fit      SE Fit    Residual   St Resid
 15       34.0      52.50       49.90        4.25        2.60        0.63 X

X denotes an observation whose X value gives it large leverage.
```

Residual Plots After Outlier Removal

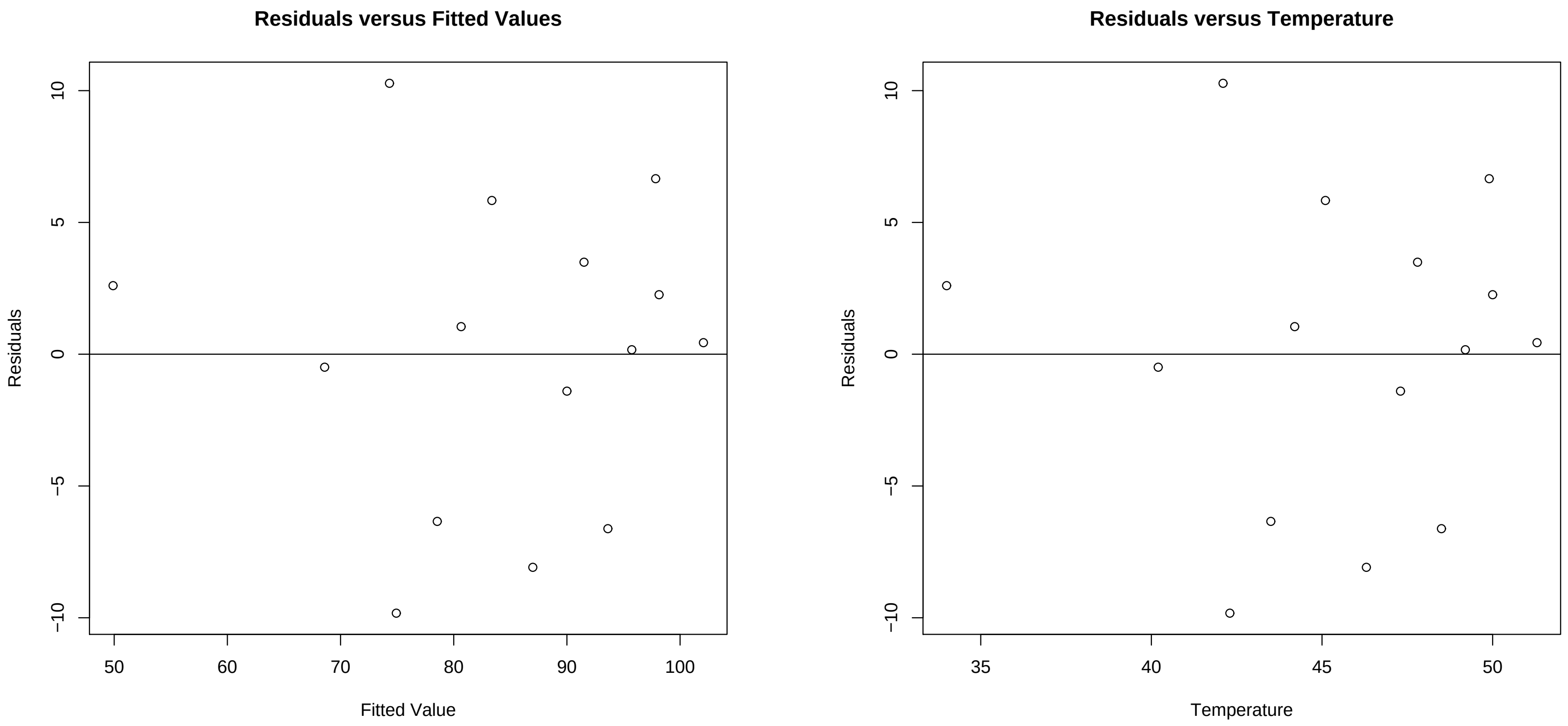


Figure 3: Residual Plots

Comments : No systematic pattern is discerned.

Multiple Linear Regression : Air Pollution Data

`http://lib.stat.cmu.edu/DASL/Stories/AirPollutionandMortality.html`

- Data collected by General Motors.
- Response is age-adjusted mortality.
- Predictors :
  - Variables measuring demographic characteristics.
  - Variables measuring climatic characteristics.
  - Variables recording pollution potential of 3 air pollutants.
- Objective : To determine whether air pollution is significantly related to mortality.

Predictors

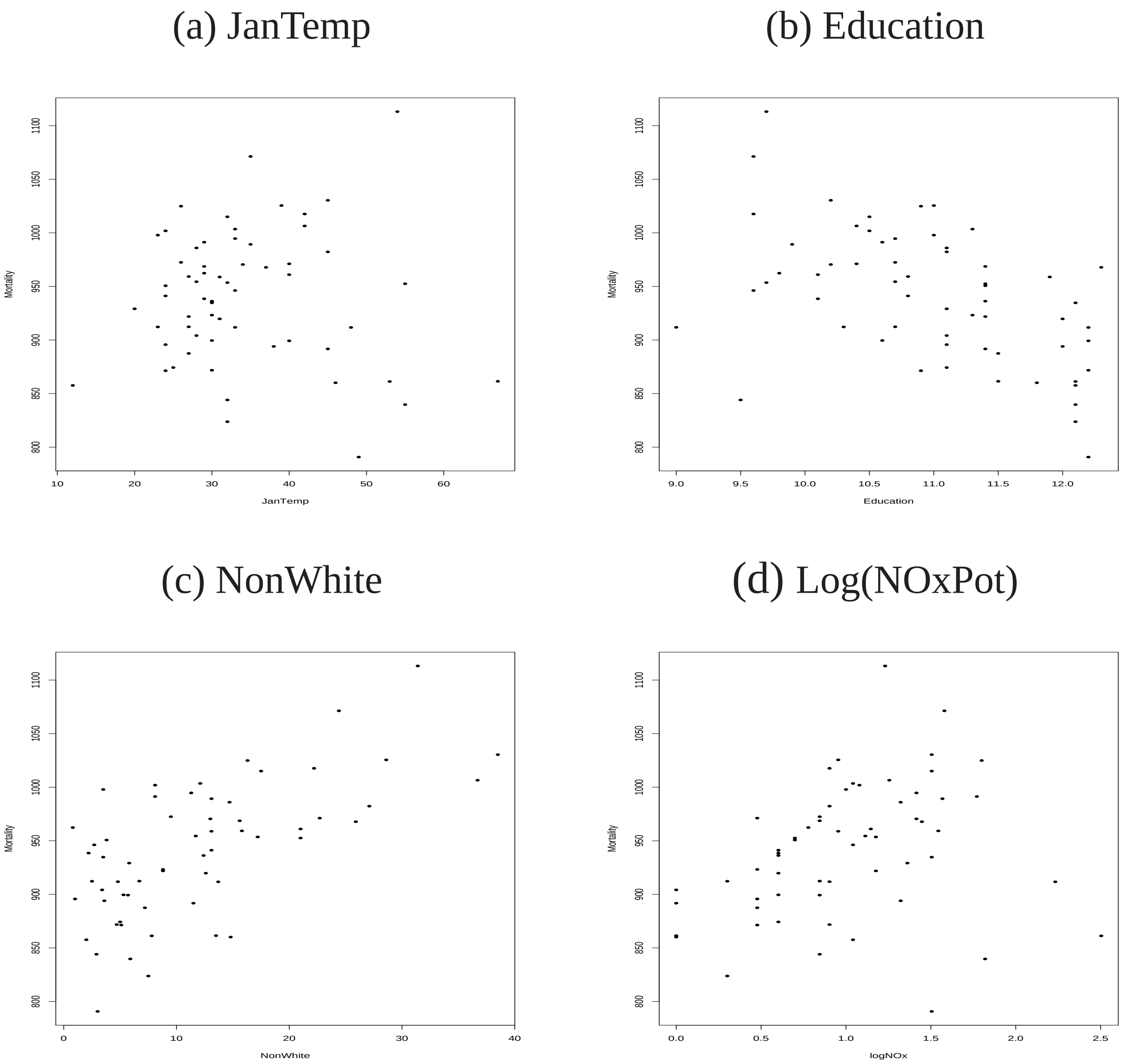
1. **JanTemp :** Mean January temperature (degrees Farenheit)
2. **JulyTemp :** Mean July temperature (degrees Farenheit)
3. **RelHum :** Relative Humidity
4. **Rain :** Annual rainfall (inches)
5. **Education :** Median education
6. **PopDensity :** Population density
7. **%NonWhite :** Percentage of non whites
8. **%WC :** Percentage of white collar workers
9. **pop :** Population
10. **pop/house :** Population per household
11. **income :** Median income
12. **HCPot :** HC pollution potential
13. **NOxPot :** Nitrous Oxide pollution potential
14. **SO2Pot :** Sulphur Dioxide pollution potential

Getting Started

- There are 60 data points.
- Pollution variables are highly skewed, log transformation makes them nearly symmetric. The variables HCPot, NOxPot and SO2Pot are replaced by log(HCPot), log(NOxPot) and log(SO2Pot).
- Observation 21 (Fort Worth, TX) has two missing values, so this data point will be discarded from the analysis.

Scatter Plots

Figure 4: Scatter Plots of mortality against selected predictors



(a) JanTemp

(b) Education

(c) NonWhite

(d) Log(NOxPot)

Fitting the Multiple Regression Equation

- Underlying Model :

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2).$$

- Coefficients are estimated by minimizing

$$\sum_{i=1}^{N} \Big(y_i - (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik}) \Big)^2 = (\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta).$$

- **Least Squares estimates :**

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\mathbf{X}'\mathbf{y}.$$

- Variance-Covariance matrix of $\hat{\beta}$: $\Sigma_{\hat{\beta}} = \sigma^2 (\mathbf{X}'\mathbf{X})^{-1}$.

Analysis of Variance

- The total variation in y, i.e., corrected total sum of squares, $CTSS = \sum_{i=1}^{N}(y_i - \bar{y})^2 = \mathbf{y}^T\mathbf{y} - N\bar{y}^2$, can be decomposed into two parts (Analysis of Variance (ANOVA)):

$$CTSS = RegrSS + RSS,$$

where $RSS =$ Residual sum of squares $= \sum(y_i - \hat{y}_i)^2 = (\mathbf{y} - \mathbf{X}\hat{\beta})^{\mathbf{T}}(\mathbf{y} - \mathbf{X}\hat{\beta})$,
$RegrSS =$ Regression sum of squares $= \sum_{i=1}^{N}(\hat{y}_i - \bar{y})^2 = \hat{\beta}^T\mathbf{X}^T\mathbf{X}\hat{\beta} - N\bar{y}^2$.

ANOVA Table

	Degrees of	Sum of	Mean
Source	Freedom	Squares	Squares
regression	k	$\hat{\beta}^T\mathbf{X}^T\mathbf{X}\hat{\beta} - N\bar{y}^2$	$(\hat{\beta}^T\mathbf{X}^T\mathbf{X}\hat{\beta} - N\bar{y}^2)/k$
residual	$N-k-1$	$(\mathbf{y} - \mathbf{X}\hat{\beta})^T(\mathbf{y} - \mathbf{X}\hat{\beta})$	$(\mathbf{y} - \mathbf{X}\hat{\beta})^T(\mathbf{y} - \mathbf{X}\hat{\beta})/(N-k-1)$
total	$N-1$	$\mathbf{y}^T\mathbf{y} - N\bar{y}^2$	
(corrected)			

Explanatory Power of the Model

- $R^2 = \frac{RegrSS}{CTSS} = 1 - \frac{RSS}{CTSS}$ measures of the proportion of variation in y explained by the fitted model. R is called the multiple correlation coefficient.

- **Adjusted R^2 :**

$$R_a^2 = 1 - \frac{\frac{RSS}{N-(k+1)}}{\frac{CTSS}{N-1}} = 1 - \left(\frac{N-1}{N-k-1}\right)\frac{RSS}{CTSS}.$$

- When an additional predictor is included in the regression model, R^2 always increases. This is not a desirable property for model selection. However, R_a^2 may decrease if the included variable is not an informative predictor. Usually R_a^2 *is a better measure of model fit*.

Testing significance of coefficients : t-Statistic

- To test the null hypothesis $H_0 : \beta_j = 0$ against the alternative hypothesis $H_0 : \beta_j \neq 0$, use the test statistic

$$t_j = \frac{\hat{\beta}_j}{s.d.(\hat{\beta}_j)}.$$

- The higher the value of t, the more significant is the coefficient.
- In practice, if p-value is less then $\alpha = 0.05$ or 0.01, H_0 is rejected.
- **Confidence Interval :** $100(1-\alpha)\%$ confidence interval for β_j is given by

$$\hat{\beta}_j \pm t_{N-(k+1),\frac{\alpha}{2}} \times s.d.(\hat{\beta}_j),$$

  where $t_{N-k-1,\frac{\alpha}{2}}$ is the upper $\alpha/2$ point of the t distribution with $N-k-1$ degrees of freedom.

  If the confidence interval for β_j does not contain 0, then H_0 is rejected.

Analysis of Air Pollution Data

```
Predictor          Coef      SE Coef          T        P
Constant         1332.7        291.7       4.57    0.000
JanTemp         -2.3052       0.8795      -2.62    0.012
JulyTemp         -1.657        2.051      -0.81    0.424
RelHum            0.407        1.070       0.38    0.706
Rain             1.4436       0.5847       2.47    0.018
Educatio         -9.458        9.080      -1.04    0.303
PopDensi       0.004509     0.004311       1.05    0.301
%NonWhit          5.194        1.005       5.17    0.000
%WC              -1.852        1.210      -1.53    0.133
pop          0.00000109   0.00000401       0.27    0.788
pop/hous         -45.95        39.78      -1.16    0.254
income        -0.000549     0.001309      -0.42    0.677
logHC            -53.47        35.39      -1.51    0.138
logNOx            80.22        32.66       2.46    0.018
logSO2            -6.91        16.72      -0.41    0.681

S = 34.58       R-Sq = 76.7%     R-Sq(adj) = 69.3%

Analysis of Variance

Source            DF          SS           MS          F        P
Regression        14      173383        12384      10.36    0.000
Residual Error    44       52610         1196
Total             58      225993
```

Variable Selection Methods

- **Principle of Parsimony (Occam's razor):** Choose fewer variables with sufficient explanatory power. This is a desirable modeling strategy.
- The goal is thus to identify the smallest subset of covariates that provides good fit. One way of achieving this is to retain the significant predictors in the fitted multiple regression. This may not work well if some variables are strongly correlated among themselves or if there are too many variables (e.g., exceeding the sample size).
- Two other possible strategies are
  - Best subset regression using Mallows' C_p statistic.
  - Stepwise regression.

Best Subset Regression

- For a model with p regression coefficients, (i.e., $p-1$ covariates plus the intercept β_0), define its C_p value as

  $$C_p = \frac{RSS}{s^2} - (N - 2p),$$

  where RSS = residual sum of squares for the given model, s^2 = mean square error = $\frac{RSS \text{ (for the complete model)}}{\text{df (for the complete model)}}$, N = number of observations.

- If the model is true, then $E(C_p) \approx p$. Thus one should choose p by picking models whose C_p values are low and close to p. For the same p, *choose a model with the smallest C_p value* (i.e., the smallest RSS value).

AIC and BIC Information Criteria

- The Akaike information criterion (AIC) is defined by

$$AIC = Nln(\frac{RSS}{N}) + 2p$$

- The Bayes information criterion (BIC) is defined by

$$BIC = Nln(\frac{RSS}{N}) + pln(N)$$

- In choosing a model with the AIC/ BIC criterion, we choose the model that minimizes the criterion value.

- Unlike the C_p criterion, the AIC criterion is applicable even if the number of observations do not allow the complete model to be fitted.

- The BIC criterion favors smaller models more than the AIC criterion.

Stepwise Regression

- This method involves adding or dropping one variable at a time from a given model based on a *partial F*-statistic.

  Let the smaller and bigger models be Model I and Model II, respectively. The partial F-statistic is defined as

$$\frac{RSS(Model\ I) - RSS(Model\ II)}{RSS(Model\ II)/\nu},$$

  where ν is the degrees of freedom of the *RSS* (residual sum of squares) for Model II.

- There are three possible ways

  1. **Backward elimination :** starting with the full model and removing covariates.
  2. **Forward selection :** starting with the intercept and adding one variable at a time.
  3. **Stepwise selection :** alternate backward elimination and forward selection.

  Usually stepwise selection is recommended.

Criteria for Inclusion and Exclusion of Variables

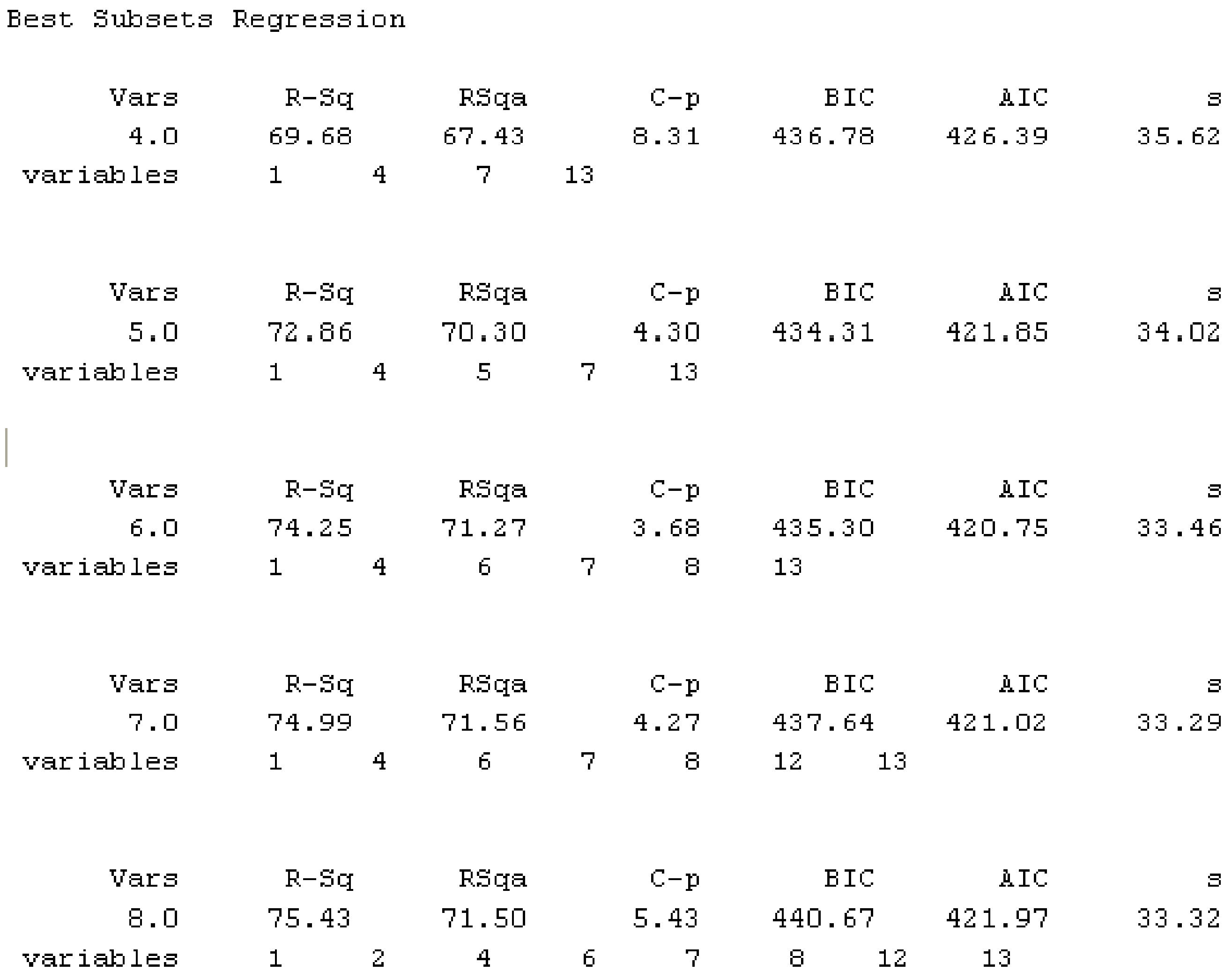
- ***F*-to-remove :** At each step of backward elimination, compute the partial F value for each covariate being considered for removal. The one with the lowest partial F, provided it is smaller than a preselected value, is dropped. The procedure continues until no more covariates can be dropped. The preselected value is often chosen to be $F_{1,\nu,\alpha}$, the upper α critical value of the F distribution with 1 and ν degrees of freedom. Typical α values range from 0.1 to 0.2.

- ***F*-to-enter :** At each step of forward selection, the covariate with the largest partial F is added, provided it is larger than a preselected F critical value, which is referred to as an *F-to-enter* value.

- For stepwise selection, the F-to-remove and F-to-enter values should be chosen to be the same.

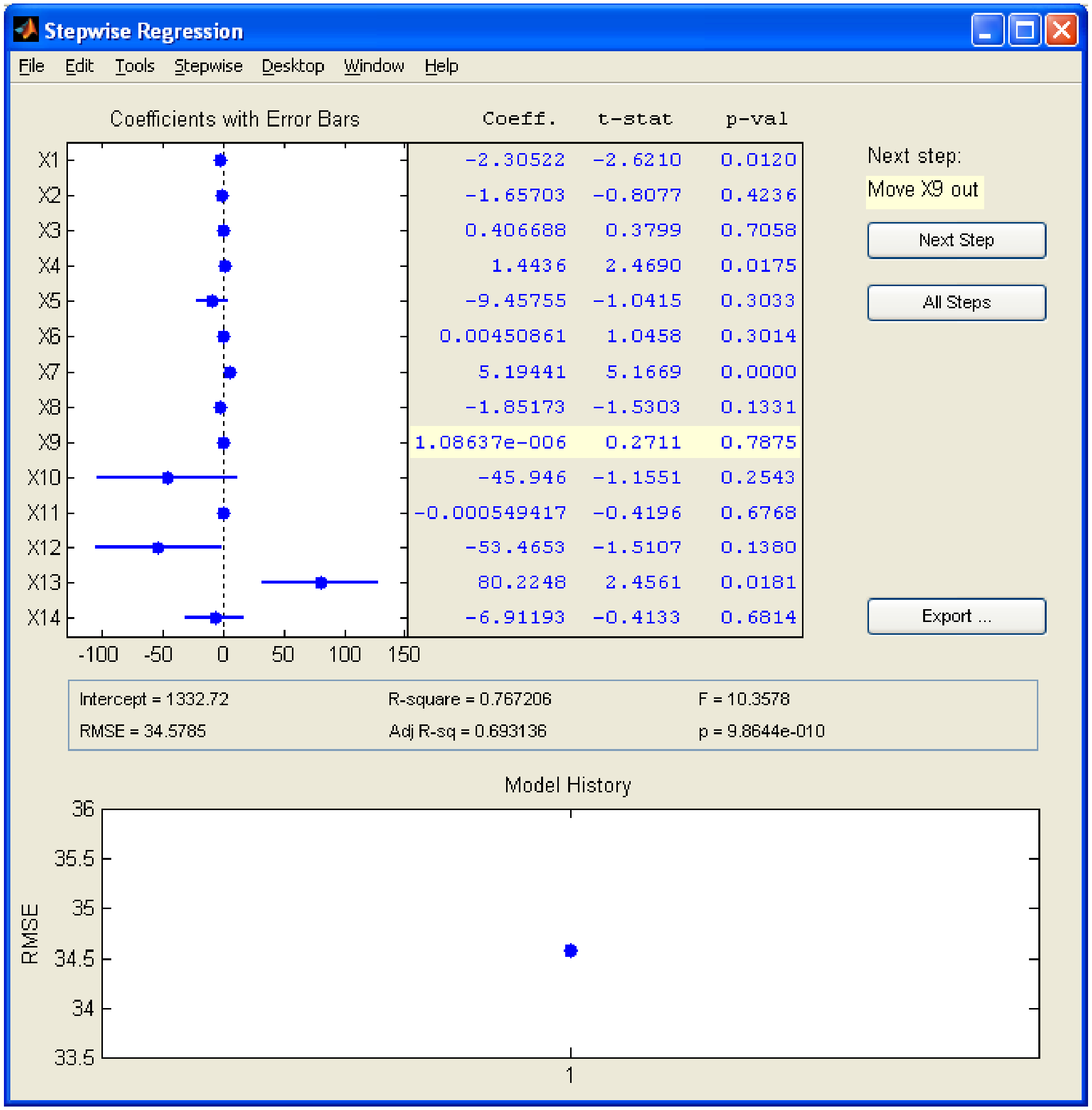
(See Section 1.7)

Air Pollution Example: Best Subsets Regression



```
Best Subsets Regression

          Vars       R-Sq       RSqa        C-p        BIC        AIC          s
           4.0      69.68      67.43       8.31     436.78     426.39      35.62
     variables      1     4      7     13


          Vars       R-Sq       RSqa        C-p        BIC        AIC          s
           5.0      72.86      70.30       4.30     434.31     421.85      34.02
     variables      1     4      5     7     13


          Vars       R-Sq       RSqa        C-p        BIC        AIC          s
           6.0      74.25      71.27       3.68     435.30     420.75      33.46
     variables      1     4      6     7     8     13


          Vars       R-Sq       RSqa        C-p        BIC        AIC          s
           7.0      74.99      71.56       4.27     437.64     421.02      33.29
     variables      1     4      6     7     8     12     13


          Vars       R-Sq       RSqa        C-p        BIC        AIC          s
           8.0      75.43      71.50       5.43     440.67     421.97      33.32
     variables      1     2      4     6     7      8     12     13
```

Pollution Data Analysis - Stepwise Regression (F-to-remove: demo)



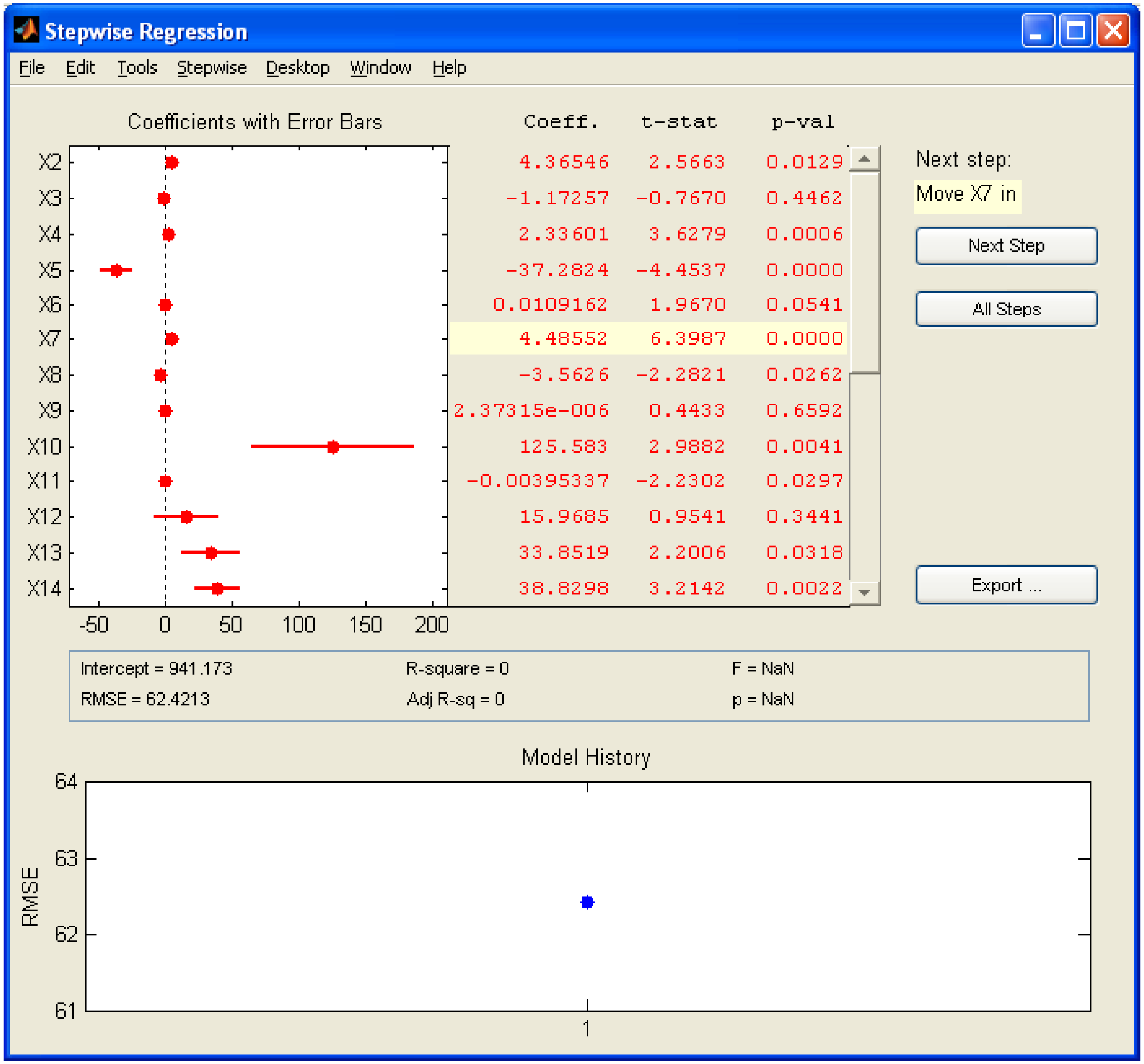
	Coeff.	t-stat	p-val
X1	-2.30522	-2.6210	0.0120
X2	-1.65703	-0.8077	0.4236
X3	0.406688	0.3799	0.7058
X4	1.4436	2.4690	0.0175
X5	-9.45755	-1.0415	0.3033
X6	0.00450861	1.0458	0.3014
X7	5.19441	5.1669	0.0000
X8	-1.85173	-1.5303	0.1331
X9	1.08637e-006	0.2711	0.7875
X10	-45.946	-1.1551	0.2543
X11	-0.000549417	-0.4196	0.6768
X12	-53.4653	-1.5107	0.1380
X13	80.2248	2.4561	0.0181
X14	-6.91193	-0.4133	0.6814

Next step:
Move X9 out

Intercept = 1332.72
R-square = 0.767206
F = 10.3578

RMSE = 34.5785
Adj R-sq = 0.693136
p = 9.8644e-010

Pollution Data Analysis - Stepwise Regression (F-to-enter: demo)



	Coeff.	t-stat	p-val
X2	4.36546	2.5663	0.0129
X3	-1.17257	-0.7670	0.4462
X4	2.33601	3.6279	0.0006
X5	-37.2824	-4.4537	0.0000
X6	0.0109162	1.9670	0.0541
X7	4.48552	6.3987	0.0000
X8	-3.5626	-2.2821	0.0262
X9	2.37315e-006	0.4433	0.6592
X10	125.583	2.9882	0.0041
X11	-0.00395337	-2.2302	0.0297
X12	15.9685	0.9541	0.3441
X13	33.8519	2.2006	0.0318
X14	38.8298	3.2142	0.0022

Next step:
Move X7 in

Intercept = 941.173 R-square = 0 F = NaN
RMSE = 62.4213 Adj R-sq = 0 p = NaN

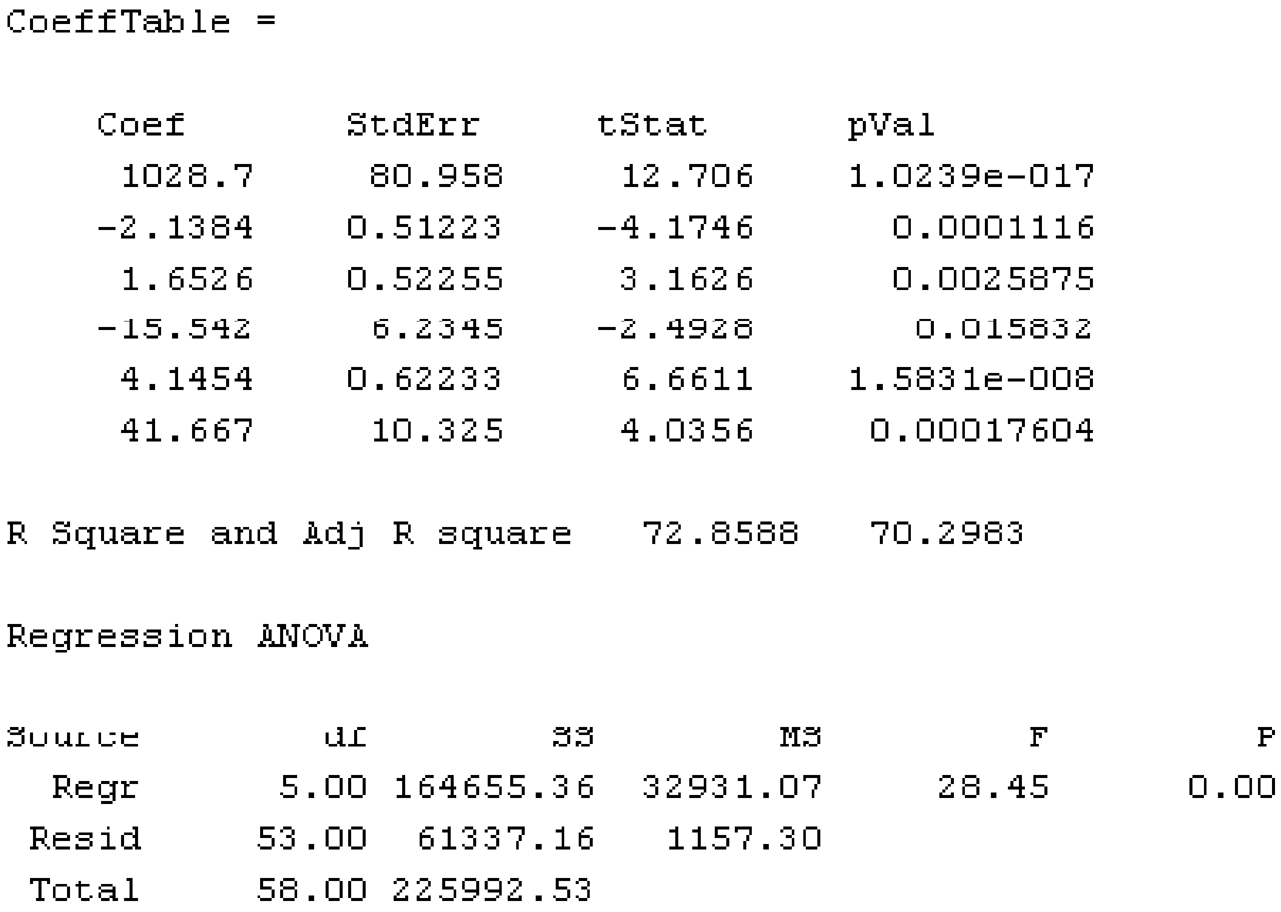
Final Model

Rival Models

	Variables	C_p	BIC	Remarks
Model 1	1,4,6,7,8,13	3.7	607	Minimum C_p
Model 2	1,4,5,7,13	4.3	606	Minimum BIC and chosen by stepwise

We shall analyze data with Model 2. (Why? Refer to the rules on page 38 and use the principle of parsimony.)

Analysis of Model 2



```
CoeffTable =

    Coef        StdErr      tStat       pVal
     1028.7      80.958      12.706     1.0239e-017
    -2.1384     0.51223     -4.1746       0.0001116
     1.6526     0.52255      3.1626       0.0025875
    -15.542      6.2345     -2.4928        0.015832
     4.1454     0.62233      6.6611     1.5831e-008
     41.667      10.325      4.0356      0.00017604

R Square and Adj R square   72.8588   70.2983

Regression ANOVA

Source        df         SS          MS           F          P
  Regr      5.00  164655.36    32931.07       28.45       0.00
 Resid     53.00   61337.16     1157.30
 Total     58.00  225992.53
```

Residual Plot

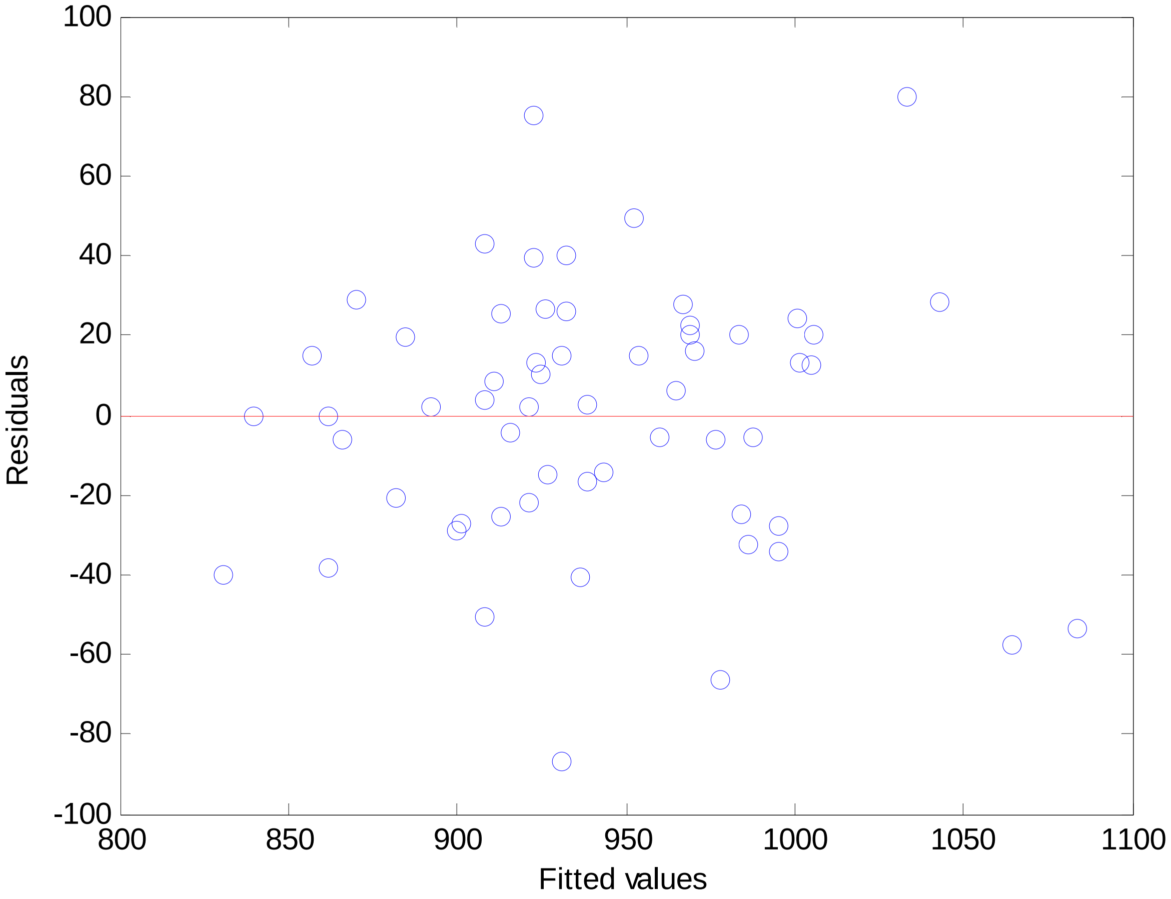


Figure 6: Plot of Residuals

Comments on Board