RNR / ENTO 613. Multiple Regression – Strategies for Variable Selection

There are 2 good reasons to minimize the subset of explanatory variables used in a regression model:

1) Simplicity is preferable to complexity (principle of parsimony)
2) Unnecessary terms in a model yield less precise inferences (multicollinearity)

The objective of the study guides the strategy used for paring down the explanatory variables. The final set of variables chosen is usually one of *several* good sets.

Four Types of Objectives for variable selection

1) *Well-defined questions*: The goal is to examine the effect of an explanatory variable (or a few) after accounting for other variables (covariates) that affect the responses.

Procedure: a) Use a variable selection technique to choose a subset of the covariates
b) Then add the variable of interest in the model
c) Only interpret the coefficient obtained for the variable of interest, which represents the association between that variable and the response *after accounting for* the effects of the other explanatory variables.

2) *Fishing for Explanation*: No well-defined question can be formulated. The goal is to search for variables that are associated with the responses, possibly after accounting for the effect of some important covariates.

Procedure: a) Use a variable selection technique to choose a subset of explanatory variables, with the restriction that this subset must contain the covariates.
b) add the *variables* of interest to the covariates (perform Extra-SS test to assess whether these variable improve fit of the model and by how much).
c) Interpretation of the coefficients associated with the explanatory variables in the final model has to be done with caution.

Caution is required because:

i) The explanatory variables chosen are just one of the possible sets (especially if sequential procedures for variable selection are used). Inclusion or exclusion of individual explanatory variables depends strongly on the correlation between them.
ii) Interpretation of the coefficients with *correlated* explanatory variables is difficult. For example, the sign of the coefficient associated with a particular explanatory variable may change depending on the other explanatory variables included in the model. It is possible to imagine a situation where a set of "good models" would differ qualitatively in their conclusions.

3) *Prediction*: No interpretation of coefficients is necessary. The variable selection techniques are used to select a model convenient for future predictions. Simple models are preferred to reduce potential *collinearity* problems that result in loss of precision.

4) *Regression for Adjustment or Ranking*: Multiple regression can be used for adjustment, i.e. rank a set of responses after removing the effect of some explanatory variables. To do this, a model is fitted to account for the variables which effect need to be removed, and the *residuals* are used for ranking.

Example: A model is fitted to describe the association between percentage of students taking the SAT test in different U.S. states (X_1), the median class score of these students (X_2), and the average SAT scores obtained for each state (Y). Residuals from that model (one for each state) are used to identify the states most likely to produce well-trained students.

<<Display 12.2, Sleuth>>

Multicollinearity

Multicollinearity describes the situation when two or more explanatory variables are highly correlated [i.e. $s_j^2 (1 - R_j^2)$ is small: see below], which results in inflated estimates of variance for the regression coefficients and a loss of precision of predictions.

The variance of the sampling distribution of a regression coefficient obtained by the method of least squares is:

$$Var\left(\hat{\beta}_j\right) = \frac{\sigma^2}{(n-1)s_j^2\left(1-R_j^2\right)}$$

Where s_j^2 is the sample variance of the explanatory variable X_j, and R_j^2 is the R-squared term from the regression of X_j on *the other explanatory variables* (i.e. X_j considered as a response variable). As additional explanatory variables are incorporated in a model, R_j^2 always increases.

Multicollinearity is likely to increase when *too many explanatory variables are included in the model*. Multicollinearity has many negative effects, including:

1) inflation of the SE of predicted values and of regression parameters
2) greater chance of having influential observations
3) intensification of the effect of measurement error in the explanatory variables.

A General Strategy for dealing with Many Explanatory variables

1. Identify the objectives of the study
2. Screen the explanatory variables, listing the ones that are relevant for the objectives, excluding redundancy (i.e. reduce potential for multicollinearity)
3. Perform exploratory analysis --scatterplots and correlations

4. Perform transformations as necessary
5. Examine residual plot after fitting a rich model --detection of important interactions is done here; consider further transformations and outliers
6. Use a computer-assisted technique to chose a subset of explanatory variables (= covariates), keeping in mind the questions of interest
7. Proceed with the analysis, using the selected covariates, to which you add the explanatory variables you are *particularly* interested in.

Example: Does pollution kill people?

Response variable: Death per 100,000 population in 60 U.S. cities

Explanatory variables:

Weather variables: 4 variables
Demographic variables: 8 variables
Pollution-related variables: 3 variables

<<Display 12.16, Sleuth>>

Steps:

1. Objective of the study: What is the role of the pollution-related variables on mortality, after accounting for the weather and demographic variables?

2. Screen the variables. All variables seem relevant and non-redundant.

3 and 4. Perform exploratory analysis with a scatterplot of response and explanatory variables. Check for nonlinear relationships, high correlation among explanatory variables, and outliers. Apply transformations if required.

Weather variables:

Correlations

Variable	PRECIP	HUMIDITY	JANTEMP	JULYTEMP
PRECIP	1.0000	-0.0773	0.0922	0.5033
HUMIDITY	-0.0773	1.0000	0.0679	-0.4528
JANTEMP	0.0922	0.0679	1.0000	0.3463
JULYTEMP	0.5033	-0.4528	0.3463	1.0000

Demographic variables:

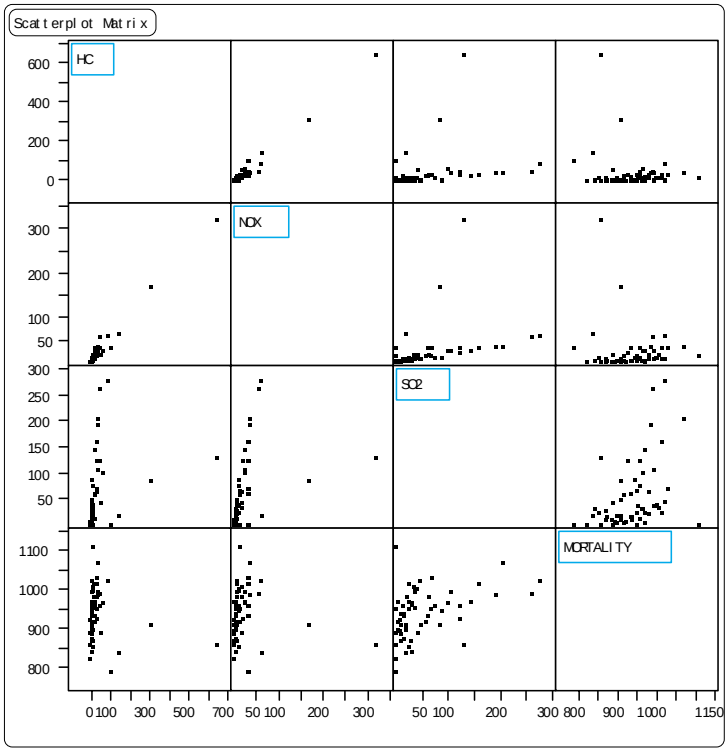
Correlations

Variable	OVER65	HOUSE	EDUC	SOUND	DENSITY	NONWHITE	WHITECOL	POOR
OVER65	1.0000	-0.5091	-0.1389	0.0650	0.1627	-0.6378	-0.1177	-0.3098
HOUSE	-0.5091	1.0000	-0.3951	-0.4106	-0.1857	0.4194	-0.4257	0.2599
EDUC	-0.1389	-0.3951	1.0000	0.5522	-0.2428	-0.2088	0.7032	-0.4033
SOUND	0.0650	-0.4106	0.5522	1.0000	0.1847	-0.4103	0.3387	-0.6807
DENSITY	0.1627	-0.1857	-0.2428	0.1847	1.0000	-0.0088	-0.0311	-0.1657
NONWHITE	-0.6378	0.4194	-0.2088	-0.4103	-0.0088	1.0000	-0.0044	0.7049
WHITECOL	-0.1177	-0.4257	0.7032	0.3387	-0.0311	-0.0044	1.0000	-0.1852
POOR	-0.3098	0.2599	-0.4033	-0.6807	-0.1657	0.7049	-0.1852	1.0000

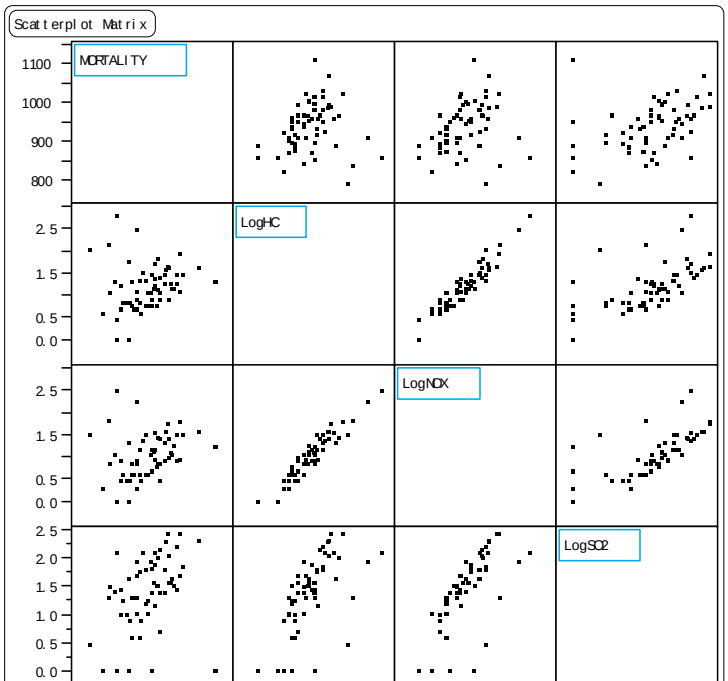
Pollution variables:

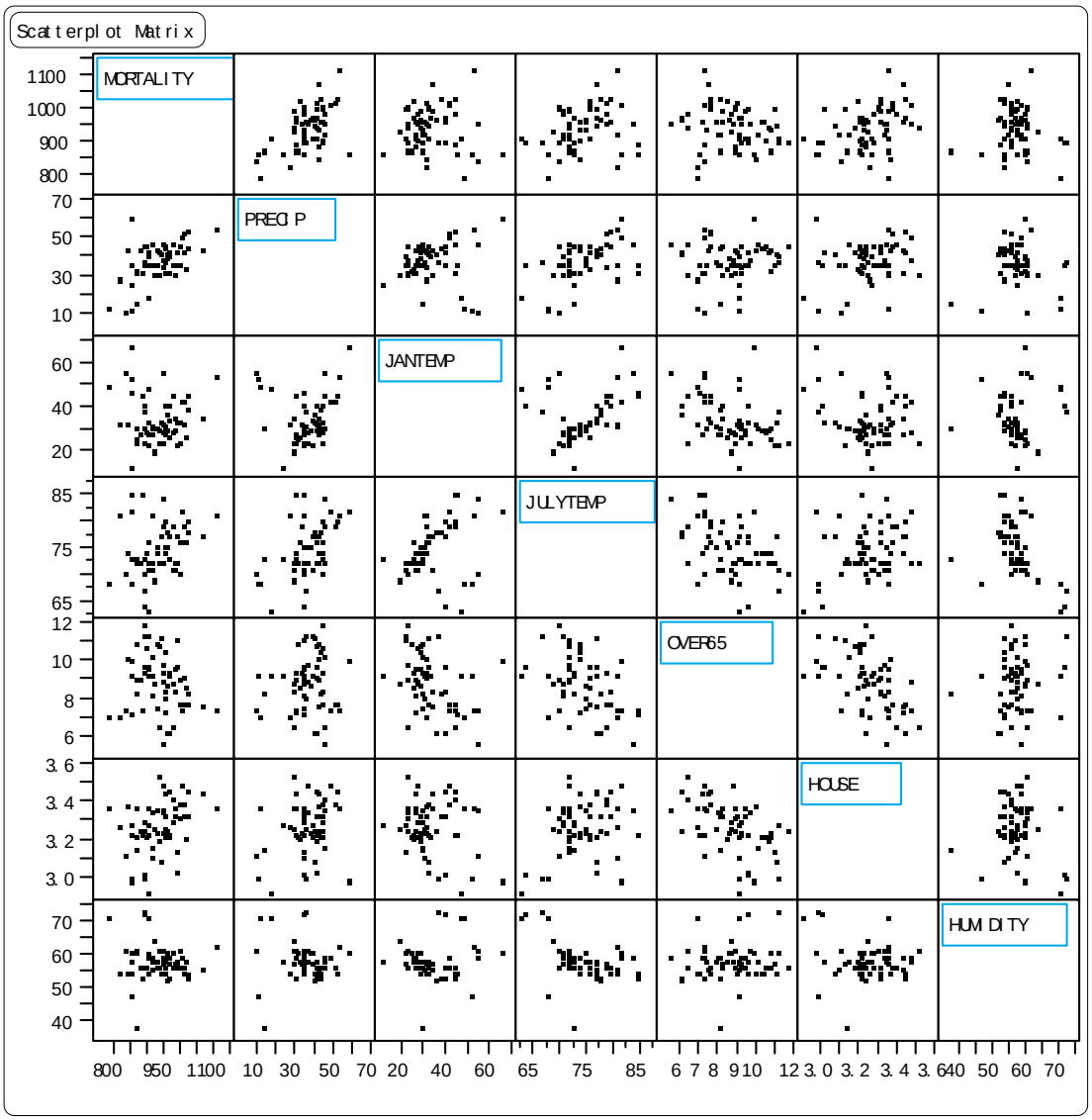
Correlations

Variable	HC	NOX	SO2
HC	1.0000	0.9838	0.2823
NOX	0.9838	1.0000	0.4094
SO2	0.2823	0.4094	1.0000



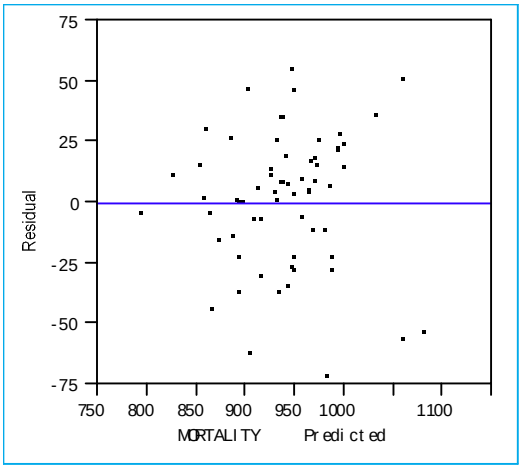
There is evidence of non-linearity and some observations are isolated. A log transformation of the pollution variables seems appropriate.





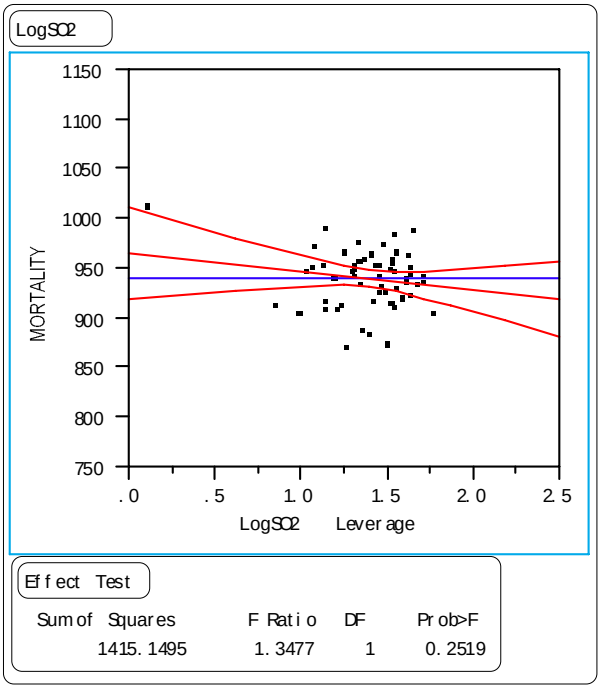
5. Residuals from a rich model to examine potential outliers and further transformations.

There are a relatively large number of explanatory variables compared to the number of observations, so we look first at the residuals from a Model only containing the simple effects:



There are no obvious outliers and the plot of the residuals versus the fitted values suggests no problem with unequal variance across combinations of values of the explanatory variables.

Some observations could have strong influences, as indicated for example by the leverage plot involving LogSO2. Those should be kept in mind in subsequent analysis.



6. Computer-assisted technique for choosing a subset of the explanatory variables.

Sequential or stepwise selection procedures are one option to explore some (but not all) models from the vast array of possible models.

In this problem, we want to select a suitable set of covariates. Once this set is chosen, we will add to the model the pollution-related variables.

Model Selection Procedures

A. Stepwise regression.

All procedures begin with a current model and advance by considering other models by either eliminating (backward elimination) or adding (forward selection) a variable.

Forward Selection: The procedure starts with a constant mean model (intercept only) and add explanatory variables one at a time until no further addition significantly improves the model fit.

Each step consists of:

1. Consider all models obtained by adding one more explanatory variable. Calculate the extra-sum-of-squares F-statistic ("F-to-enter") corresponding to each new model and identify the variable with the *largest* F-to-enter.
2. If the largest F-to-enter is larger than the user specified value (usually F = 4, which is equivalent to a t = 2), add that explanatory variable to the model.
3. Repeat until no additional variables can be added.

Backward Elimination: Start with a model containing all the covariates:

1. For each variable in the current model, calculate its extra-sum-of-squares F-statistic ("F-to-remove") and identify the variable with the *smallest* F.
2. If the smallest F-to-remove is smaller than the user specified value (F = 4), remove that explanatory variable from the model.
3. Repeat until no additional variables can be removed.

Stepwise Regression: Start with the constant mean model:

1. Do one step of forward selection
2. Do one step of backward elimination
3. Repeat until no variable can be added or removed.

Compounded Uncertainty in Stepwise Procedures:

At each step, the statistic considered to enter (or remove) a parameter is the largest (smallest) of *several* F-statistics. The significance levels on the statistics for the parameters entered in the model are in general *smaller* than they would be in a unique model, because of compounded uncertainty. Sequential procedures *tend to select models that have too many variables, i.e. they are too permissive*.

<<Display 12.7, Sleuth>>

Despite this limitation, Stepwise procedures are *convenient* and have been used for 30 years to eliminate unnecessary explanatory variables in models.

B. R-square and Adjusted R-square

Comparing 2 models with the same number of parameters is easy: the one with the larger R^2 (i.e. smaller residual mean square, σ^2) is preferred. R-square always increases with increased number of variables in a model.

On the other hand, the adjusted R-squared is discounted by the number of parameters in the model. Choosing the model with the highest *adjusted* R^2, is still similar to selecting the model with the smallest residual mean square, which is a criterion that generally favors models with *too many variables* if the set contains unimportant explanatory variables.

C. The Cp Statistics and other Information Criterion

An alternative approach for model selection involves fitting all possible subsets of explanatory variables, and then identifying those that best satisfy some criteria. JMP does not automatically fit all possible subsets of models.

There are many criteria available, such as the Schwarz's Bayesian Information Criterion (BIC) and the Cp statistic. (JMP provides the closely related Akaike's Information Criterion –the AIC, instead of the BIC).

The Cp criterion considers the *trade-off* between *bias* due to excluding important explanatory variables and *extra variance* in the coefficients (and the predicted response) due to including too many: Too few variables does not provide accurate prediction of the responses (bias); too many increases the risk of multicollinearity (large SE).

The Cp statistic compares the mean squares error of a reduced model to a model with all available explanatory variables, *assuming that the model with all the variables has no bias*. Cp is computed as:

$$C_p = p + (n - p)\frac{\left(\hat{\sigma}^2 - \hat{\sigma}^2_{full}\right)}{\hat{\sigma}^2_{full}}$$

where $\hat{\sigma}^2$ is the estimate of the population variance from the tentative model, $\hat{\sigma}^2_{full}$ is the estimate of variance from the model with all possible explanatory variables, and p the number of regression coefficients.

Models with *small Cp statistics* are more *favorable*:

1. If a model lacks important explanatory variables, it will show greater residual variability than the full model, so $\hat{\sigma}^2 - \hat{\sigma}^2_{full}$ will be large, and Cp will be large.
2. If the bias of the 2 models is the same (i.e. the estimates of $\hat{\sigma}^2$ are similar), including more explanatory variables in one of the model will increase the value of p, and thus incorporate a penalty for having more variables than necessary, i.e. increase Cp.

A model without bias is represented by Cp ≤ p (*the model with all the variables is assumed to have Cp = p)*, the number of regression parameters. Picking a single model among those with a small Cp is a mater of selecting the most convenient model that has all its coefficients different from zero.

Sequential Variable selection: Pollution Example.

In JMP, you can use the Cp criterion with a Sequential Variable Selection technique to choose a model without too many parameters:

If you use Mallow's Cp as a model selection criterion, *the model chosen is generally the one where Cp approaches p.*

With a *probability to enter = 0.25*, and a *probability to remove = 0.10*, these procedures select the following variables:

Step History: *Forward*

Step	Parameter	Action	"Sig Prob"	Seq SS	RSquare	Cp	(p)
1	NONWHITE	Entered	0.0000	94595.57	0.4144	43.366	(2)
2	EDUC	Entered	0.0000	33848.33	0.5627	20.206	(3)
3	JANTEMP	Entered	0.0011	17457.34	0.6392	9.2293	(4)
4	**HOUSE**	**Entered**	**0.0205**	**7722.099**	**0.6730**	**5.4893**	**(5)**
5	JULYTEMP	Entered	0.1173	3345.937	0.6876	5.0022	(6)
6	PRECIP	Entered	0.0685	4366.53	0.7068	4.7564	(7)
7	DENSITY	Entered	0.1254	2983.228	0.7198	3.5389	(8)

SSE	DFE	MSE	RSquare	RSquare Adj	Cp	AIC
63954.057	52	1229.886	0.7198	0.6821	3.538867	434.2945

Step History: *Backward*

Step	Parameter	Action	"Sig Prob"	Seq SS	RSquare	Cp	(p)
1	HUMIDITY	Removed	0.9179	14.46264	0.7229	11.011	(12)
2	SOUND	Removed	0.8442	51.42534	0.7227	9.049	(11)
3	POOR	Removed	0.9129	15.59842	0.7227	7.0606	(10)
4	WHITECOL	Removed	0.8378	53.60253	0.7224	5.1004	(9)
5	OVER65	Removed	0.4939	589.8486	0.7198	3.5389	(8)
6	**HOUSE**	**Removed**	**0.1549**	**2562.097**	**0.7086**	**3.4433**	**(7) ***

* The values of Cp are always smaller than p, so we could choose the model with the smallest Cp value.

SSE	DFE	MSE	RSquare	RSquareAdj	Cp	AIC
66516.153	53	1255.022	0.7086	0.6756	3.443346	434.6513

Step History: *Stepwise* (Mixed in JMP)

Step	Parameter	Action	"Sig Prob"	Seq SS	Rsquare	Cp	(p)
1	NONWHITE	Entered	0.0000	94595.57	0.4144	43.366	(2)
2	EDUC	Entered	0.0000	33848.33	0.5627	20.206	(3)
3	JANTEMP	Entered	0.0011	17457.34	0.6392	9.2293	(4)
4	**HOUSE**	**Entered**	**0.0205**	**7722.099**	**0.6730**	**5.4893**	**(5)**
5	JULYTEMP	Entered	0.1173	3345.937	0.6876	5.0022	(6)
6	JULYTEMP	Removed	0.1173	3345.937	0.6730	5.4893	(5)

SSE	DFE	MSE	RSquare	RSquare Adj	Cp	AIC
74649.752	55	1357.268	0.6730	0.6492	5.489281	437.5731

Summary of Sequential Methods (based on Cp criterion):

Parameter	Forward	Backward	Stepwise
Intercept	X	X	X
Precip		X	
JanTemp	X	X	X
JulyTemp		X	
Over65			
House	X		X
Educ	X	X	X
Sound			
Density		X	
NonWhite	X	X	X
WhiteCol			
Poor			
Humidity			

Knowing that Sequential Procedures tend to be too permissive, we could decide to retain a subset of 4 explanatory variables. The model with 4 explanatory variables also has a small AIC (the best models have the smallest values for the Akaike's Information Criterion):

JanTemp, House, Educ, NonWhite

We can now answer the question of interest:

Does pollution kill people?

We fit a reduced model containing the 4 selected covariates and compare it to the full model, which includes the 3 additional pollution variables. We then calculate an extra-sum-of-squares F-test.

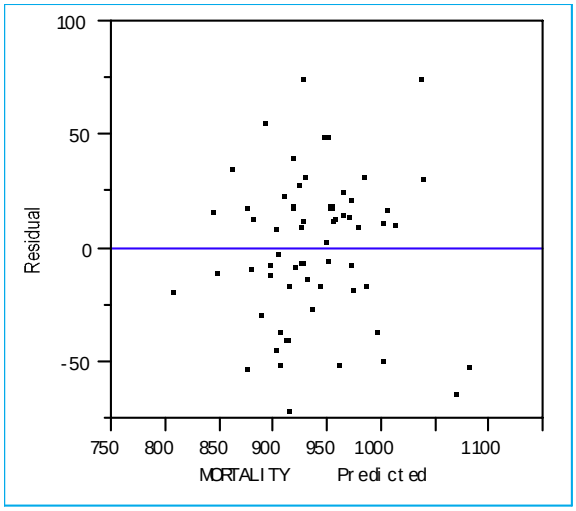
Full Model: Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	7	165254.85	23607.8	19.4802
Error	52	63018.24	1211.9	Prob>F
C Total	59	228273.09		<.0001

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	1529.2923	213.8991	7.15	<.0001
JANTEMP	-2.544714	0.801359	-3.18	0.0025
HOUSE	-93.70173	48.15509	-1.95	0.0571
EDUC	-24.35796	6.964375	-3.50	0.0010
NONWHITE	5.5194163	0.768159	7.19	<.0001

LogHC	-60.18314	30.7091	-1.96	0.0554
LogNOX	83.081118	34.09653	2.44	0.0183
LogSO2	-4.895639	15.87605	-0.31	0.7590



Reduced Model: Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	4	153623.34	38405.8	28.2964
Error	55	74649.75	1357.3	Prob>F
C Total	59	228273.09		<.0001

Extra SS $F_{3, 52}$ = [74649.75 – 63018.24 / 3] / 1211.9 = 3.20

$F_{3,52}$ = 3.20, P = 0.031, therefore we reject the null hypothesis and conclude that there is evidence for an association between pollution and mortality rate. The coefficients associated with the pollution variables should be interpreted cautiously, as always when we are fishing for an explanation (indeed, the coefficients seem difficult to interpret).