Course Project: Regression Model for Predicting Mortality Rate from Air Pollution Data

David Chung

Statistical Models and Regression

James Hung

5/12/2019

Table of Contents

Introduction ..... *- 2 -*

Assumptions ..... *- 2 -*

Model Construction with Variable Selection ..... *- 3 -*

All-Possible-Regressions Procedure ..... **- 3 -**

Model Adequacy Check – All-Possible-Regressions Model ..... **- 6 -**

Regression Model with Transformed Variables ..... **- 11 -**

Review of Model Construction and Variable Selection Procedure ..... **- 17 -**

Assessing Statistical Significance of Regression Coefficients ..... *- 17 -*

Significance of Regression (Full Model) ..... **- 17 -**

Statistical Significance of Individual Regression Coefficients ..... **- 18 -**

Measuring Predictive Power of Regression Model ..... *- 19 -*

PRESS Statistic for Entire Dataset of Observations ..... **- 19 -**

Cross-Validation Using Training and Test Data ..... **- 20 -**

Conclusions and Future Recommendations ..... *- 21 -*

APPENDIX: MATLAB Code ..... *- 23 -*

Introduction

Air pollution data for 60 cities in the United States from 1960 are collected. The dataset contains information on the total age-adjusted mortality rate (y), mean annual precipitation in inches (x_1), median number of school years completed for those over 25 in 1960 SMSA (x_2), percentage of urbanized area population that is nonwhite (x_3), relative pollution potential of oxides of nitrogen (NOx) (x_4), and relative pollution potential of sulfur dioxide (SO_2) (x_5).

The United States Environmental Protection Agency (EPA) and the National Institute of Health (NIH) assembled a team of scientists and statisticians to construct and determine a regression model for predicting a city's age-adjusted mortality rate using air pollution data. This report conveys the team's findings and is divided into four sections.

The first section explains a procedure used to select a subset of the five possible regressors for the regression model. For each candidate regression model, model adequacy checks were also conducted to verify if the model assumptions are validated by the data. Once a regression model is selected based on a pre-defined set of selection criteria, the significance of each regression coefficient is evaluated. The second section summarizes the computations and results of the regression analysis (i.e. significance of regression of each regressor). After individually examining each regression coefficient, the selected regression model's predictive power is assessed using cross-validation techniques. The third section details and interprets the outcomes of the cross-validation procedure. The last section of the report summarizes the major conclusions of the study and suggests future considerations to improve the selected regression model.

Assumptions

For the model construction and regression analysis, the following assumptions were made:

1. The random errors ε of the 60 observations are independent and normally distributed as $\varepsilon \sim NIND(0, \sigma^2)$.

2. The mean of the random errors is zero while the variance σ^2 of the random errors is constant.

3. The five regressors (x_1, …, x_5) are fixed variables when they are present in the regression model.

4. Only multiple linear regression models are considered during the model construction and variable selection procedure.

Model Construction with Variable Selection

The mortality and air pollution regression model was constructed by considering all possible regression models. For this model construction and variable selection procedure, only linear regression models were considered. Furthermore, interactions between regressors were not considered in the model construction.

Once a regression model was determined, model adequacy checks were conducted to determine if a variable transformation or revision to the model structure was necessary. This sequence of variable selection and model adequacy checks was repeated until the procedure selected a regression model that was statistically sufficient for the data and satisfied all of its model assumptions.

All-Possible-Regressions Procedure

Initially, the dataset of n=60 observations was fitted to every possible regression model that could be constructed from the five regressors. The coefficient of determination $R_p^2, p = 2, \dots, 6$, the residual mean sum of squares $MS_{res}(p), p = 2, \dots, 6$, and Mallow's C_p statistic were computed for each regression model. For each model size $p = 2, \dots, 6$, the maximum coefficients of determination $R_{p,max}^2 = \frac{SS_R(p),max}{SS_T}, p = 2, \dots, 6$ are plotted in Figure 1:

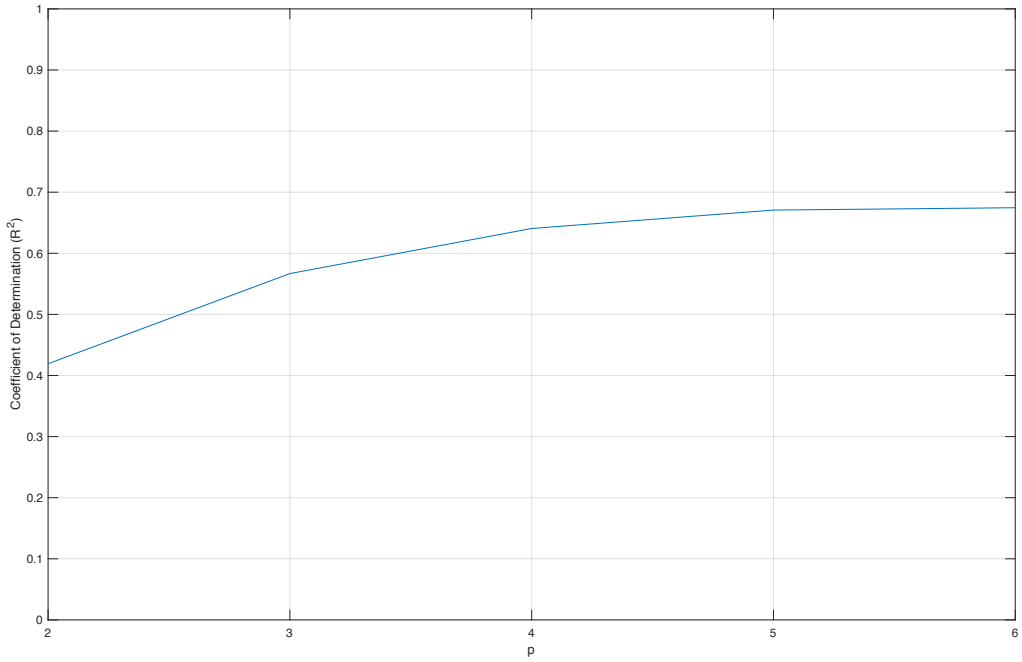

Figure 1: Plot of the maximum coefficients of determination R_p^2 for all model sizes

As shown in Figure 1, the maximum coefficient of determination R_p^2 increases from 0.419 (p=2 regression coefficients) to 0.671 (p=6 regression coefficients). In other words, the percent of the total variance in the response y explained by the regression model increases from 41.9% to 67.1% when the subset model size increases from p=2 to p=6 regression coefficients. In addition, the maximum coefficient of determination for p=5 regression coefficients ($R_p^2 = 0.641$) is approximately equal to the coefficient of determination for the full model (0.671).

Based on the maximum R_p^2 plot, it appears that subset models of sizes p=5, 6 regression coefficients achieve maximum coefficients of determination R_p^2 that are comparable to the full model coefficient of determination $R_p^2 = 0.671$.

After examining the plot of the maximum R_p^2 versus the model subset size p, the residual mean squares are determined for all regression models. Figure 2 plots the minimum residual mean squares $MS_{res}(p) = \frac{SS_{res}(p)}{n-p} = \frac{SS_{res}(p)}{60-p}, p = 2, \ldots, 6$ versus model size:

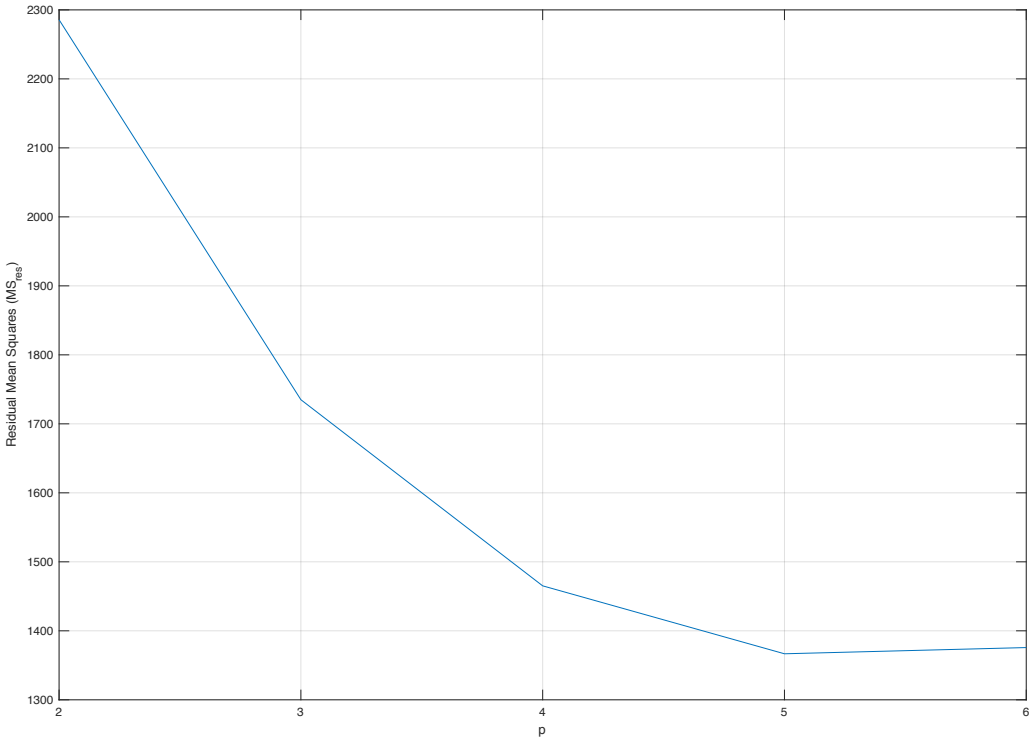

Figure 2: Plot of the minimum residual mean squares $MS_{res}(p)$ for all regression model sizes

Based on Figure 2, the minimum residual mean squares $MS_{res}(p)$ decreases as the model size increases until it reaches its minimum value when the model size is p=5 regression coefficients (i.e. the minimum residual mean squares $MS_{res}(p)$ for the full regression model is actually greater than one of the subset models of size p=5). This result occurs when the decrease

in the residual sum of squares by adding a regressor is not sufficient to overcome the loss of one degree of freedom in the computation of $MS_{res}(p) = \frac{SS_{res}(p)}{n-p}$.

Lastly, the Mallow's C_p statistic is computed for all possible linear regression models:

$$C_p = \frac{SS_{res}(p)}{\hat{\sigma}^2} - n + 2p$$

where $\hat{\sigma}^2 = MS_{res}(p = 6) = \frac{SS_{res}(p=6)}{n-p} = \frac{SS_{res}(p=12)}{60-6} = \frac{SS_{res}(p=6)}{54} = 25.476$ is the residual mean squares of the regression model constructed with all five regressors. Figure 3 plots the maximum C_p statistic for all regression models versus model size:

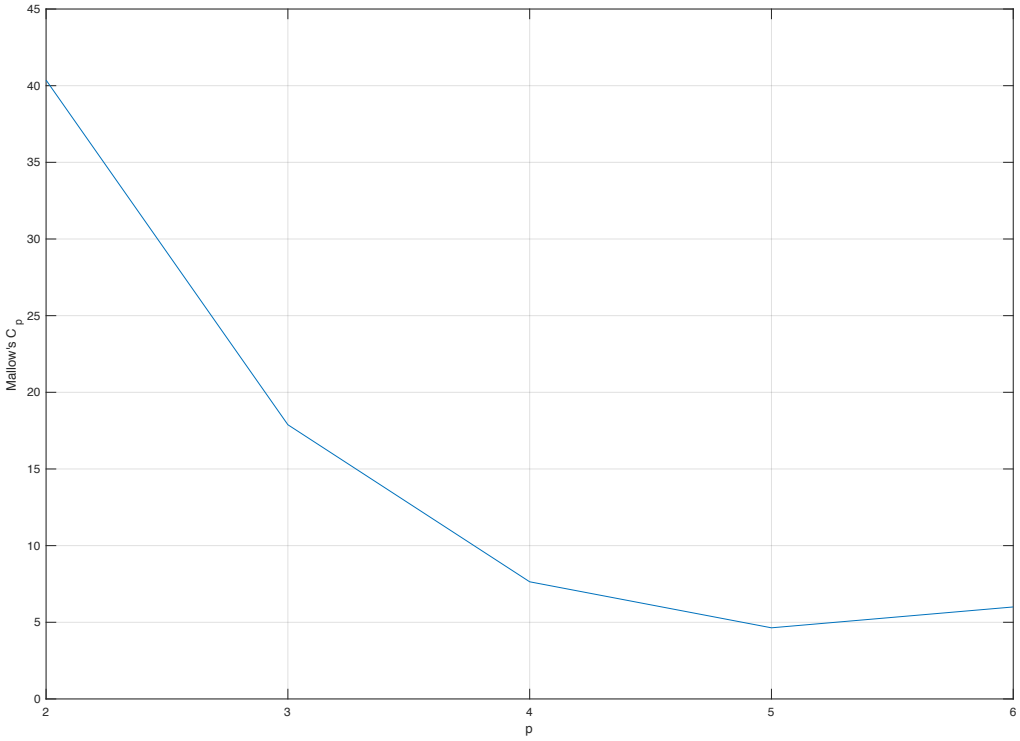

Figure 3: Plot of minimum Mallow's C_p statistic for all regression model sizes

Analogous to the plot of the minimum residual mean sum of squares, the minimum Mallow's C_p statistic is achieved when the subset regression model size is p=5 regression coefficients. The corresponding subset regression model of size p=5 is $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_5 x_5 + \varepsilon$. The regressor x_4 (relative pollution potential of oxides of nitrogen) is omitted from this regression model.

Therefore, the all-possible-regressions approach concludes that an appropriate regression model for the n=60 sampled observations is $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_5 x_5 + \varepsilon$. This

regression model of size p=5 achieves a coefficient of determination $R_p^2 = 0.641$ that is comparable to the coefficient of determination for the full regression model while also producing the lowest attainable residual mean squares $MS_{res}(p = 5) = 1.377 * 10^3$ and exhibiting the minimum Mallow's C_p test statistic $C_p(p = 5) = 4.639$ among all possible linear regression models.

Model Adequacy Check – All-Possible-Regressions Model

After the all-possible-regressions procedure converged to the selected regression model, model adequacy checks were conducted to determine if the model assumptions are satisfied. A normal probability plot of the externally studentized residuals (*R*-student residuals) is constructed to evaluate the validity of the normality assumption made for the random error ε.

For the *i*th observation (*i*=1, 2, …, 60), the variance of the random error σ^2 is estimated with the *i*th observation removed:

$$S_{(i)}^2 = \frac{(n-p)MS_{res} - \frac{e_i^2}{(1-h_{ii})}}{n-p-1} = \frac{(60-5)\left(\frac{SS_{res}}{(60-5)}\right) - \frac{e_i^2}{(1-h_{ii})}}{(60-5-1)}, i = 1,2,\dots,60$$

where p=5, $e_i = y_i - \hat{y}_i, i = 1, 2, \dots, 60$ is the ordinary residual of the *i*th observation (non-standardized) and $h_{ii}, i = 1, 2, \dots, 60$ is the *i*th diagonal entry of the hat matrix $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$.

The externally studentized residuals are computed for the 60 observations:

$$t_i = \frac{e_i}{\sqrt{S_{(i)}^2(1-h_{ii})}}, i = 1, 2, \dots, 60$$

The *R*-student residuals are ordered $t_{(1)} < \cdots < t_{(60)}$ and plotted against the cumulative probability $P_i = \frac{i-0.5}{60}, i = 1, \dots, 60$:

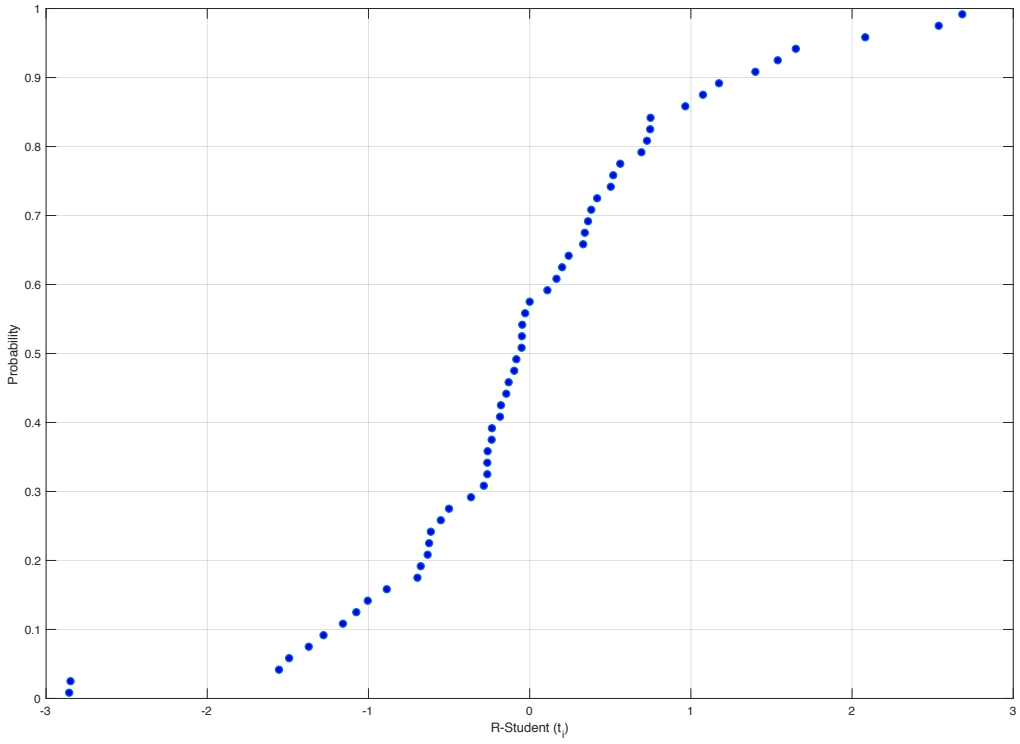

Figure 4: Normal probability plot of externally studentized residuals (t_i)

The normal probability plot shows flattening at extreme values of the externally studentized residuals, which suggests that the distribution of the random error ε contains heavier tails than the normal distribution. Thus, the normality assumption of the random errors is questionable. However, the following statistical inferences will be conducted under the assumption that the residuals are nearly normal.

In addition, a plot of the externally studentized residuals versus the predicted response $\hat{y}$ is constructed to determine if there are possible correlations between the residuals and response for each observation:

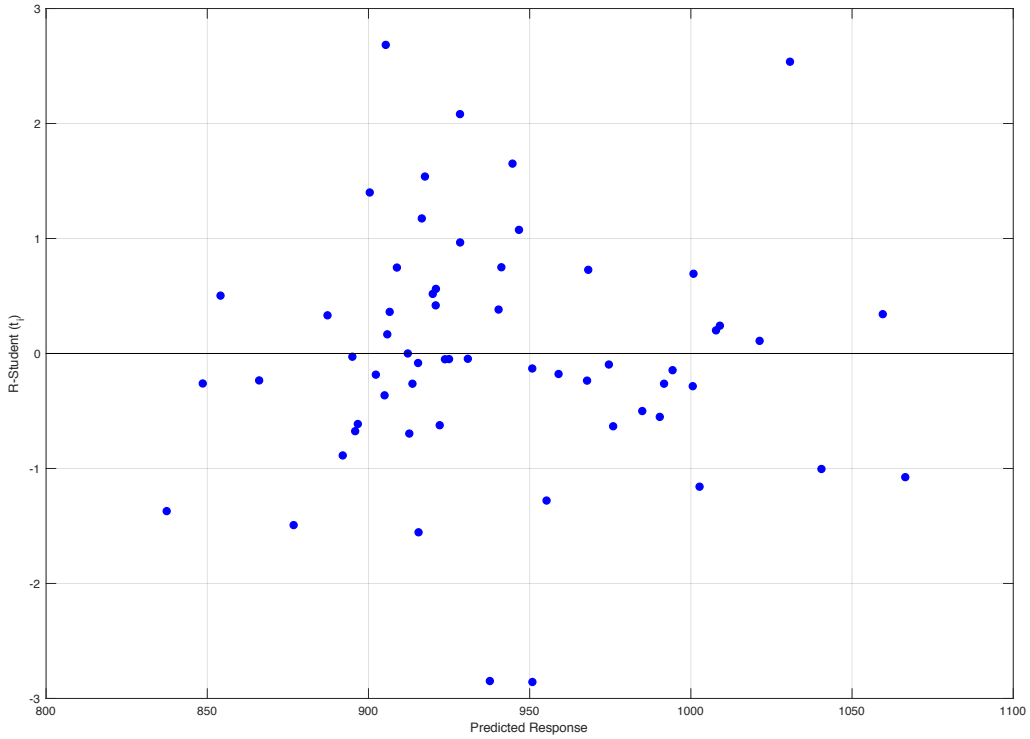

Figure 5: Plot of externally studentized residuals (t_i) versus predicted response $\hat{y}_i$

As shown in Figure 5, the externally studentized residuals appear to be contained in a horizontal band, which suggests that the residuals are uncorrelated with the predicted responses (i.e. there are no obvious model inadequacies). However, four of the *R*-student residuals are unusually large (i.e. approaching ± 3), which suggests that there are possible outliers and/or influential points in the dataset of 60 observations.

Likewise, plots of the externally studentized residuals versus the four regressors (x_1, x_2, x_3, x_5) are constructed to determine if there are possible correlations between the residuals and the regressors for each observation. Figures 6-9 plot the externally studentized residuals versus regressors x_1 and x_2:

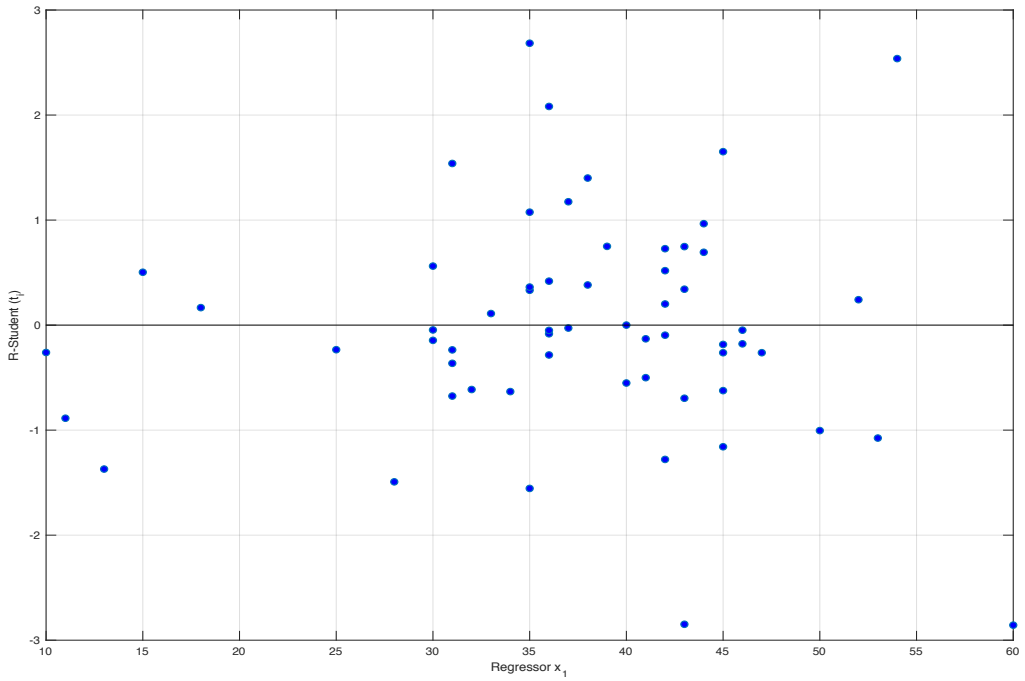

Figure 6: Plot of externally studentized residuals (t_i) versus regressor x_1

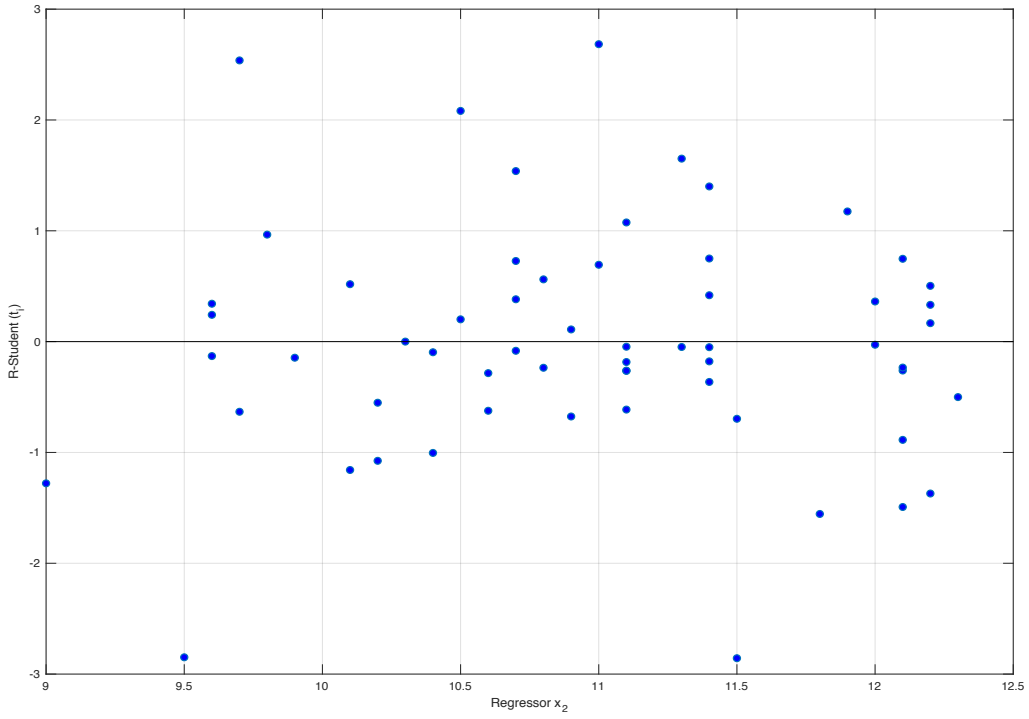

Figure 7: Plot of externally studentized residuals (t_i) versus regressor x_2

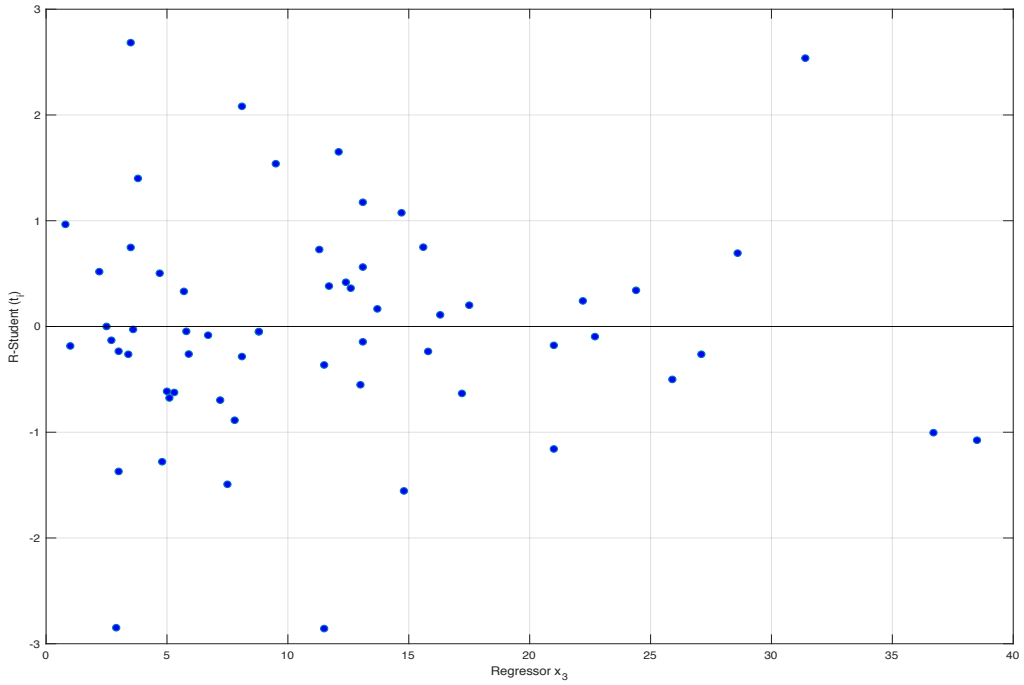

Figure 8: Plot of externally studentized residuals (t_i) versus regressor x_3

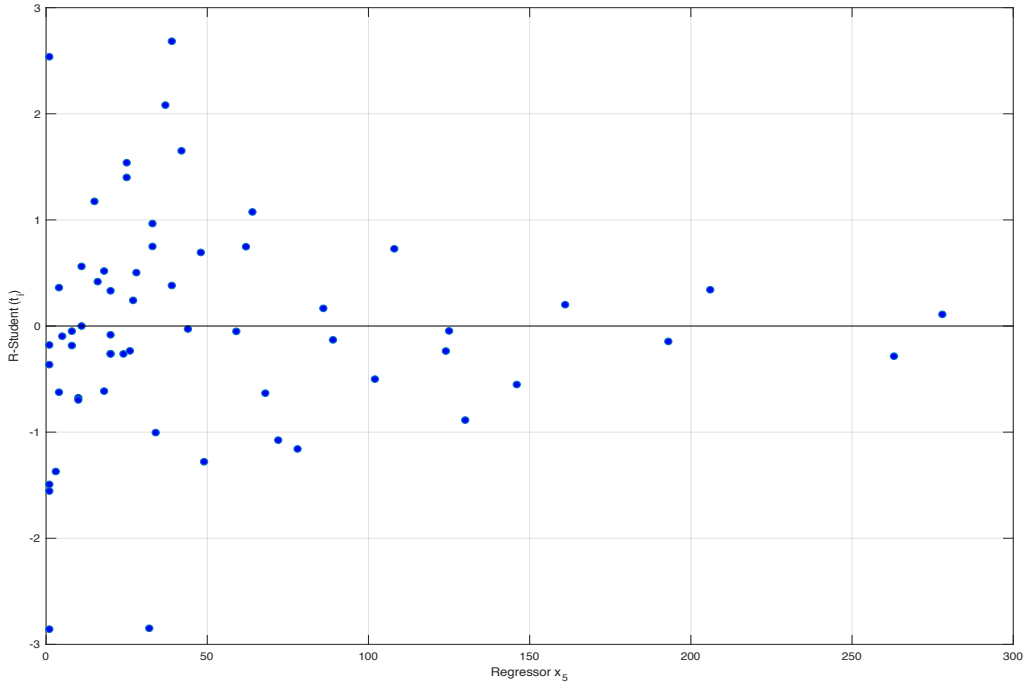

Figure 9: Plot of externally studentized residuals (t_i) versus regressor x_5

The plots of the externally studentized residuals versus regressors x_1 and x_2 appear to be contained in a horizontal band, which conveys that the residuals are also uncorrelated with regressors x_1 and x_2. Similar to the residual plot for the predicted responses, four of the R-student residuals are unusually large (i.e. approaching ± 3), which suggests that there are possible outliers and/or influential points in the dataset of 60 observations.

On the other hand, the plots of the externally studentized residuals versus regressors x_3 and x_5 appear to be contained in a tapering (inward-opening) funnel, which conveys that the variance of the random error $Var(\varepsilon)$ is decreasing as regressors x_3 and x_5 increase. Consequently, the constant variance assumption of the random error is violated for this regression model, and the model structure must be modified.

Regression Model with Transformed Variables

In order to determine an appropriate variable transformation so that the regression model satisfies its model assumptions, the mortality rate observations (response y) are plotted versus the percentage of nonwhites in urban areas (regressor x_3) and the relative pollution potential of sulfur dioxide (regressor x_5):

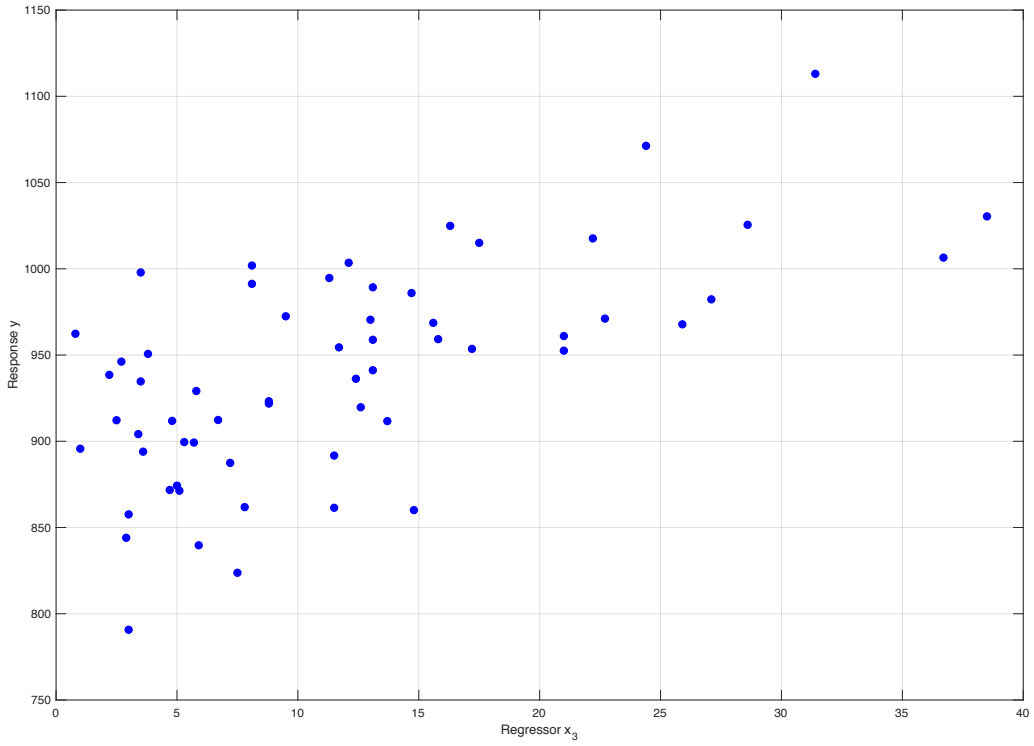

Figure 10: Plot of mortality rate (y) versus percentage of nonwhites in urban areas (x_3)

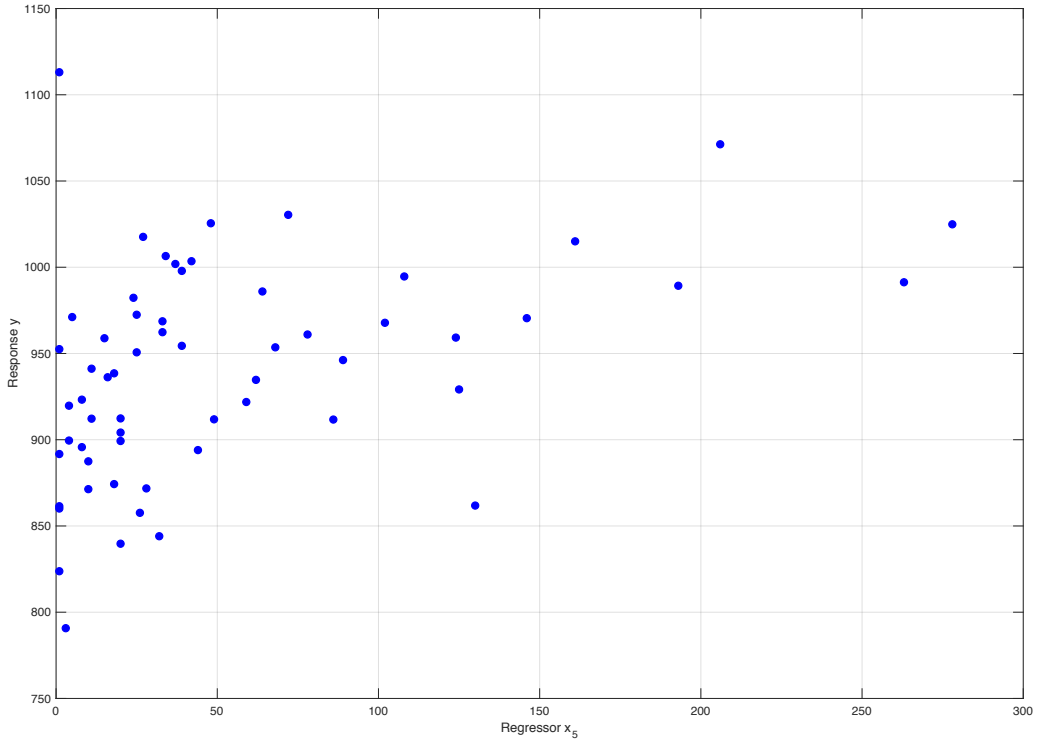

Figure 11: Plot of mortality rate (y) versus relative pollution potential of sulfur dioxides (x_5)

The plots in Figures 10-11 suggest that the relationship between the response y and regressors x_3 and x_5 is nonlinear. Specifically, the response y appears to be equal to the natural logarithm of regressors x_3 and x_5. Thus, the data for regressors x_3, x_5 will be transformed and fitted to an alternative multiple linear regression model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 \ln(x_3) + \beta_5 \ln(x_5) + \varepsilon$$

Once again, model adequacy checks are conducted to verify if the pre-defined model assumptions are satisfied. Figure 12 displays the normal probability plot of the externally studentized residuals (*R*-student residuals) constructed from the alternative multiple linear regression model versus the cumulative normal probability:

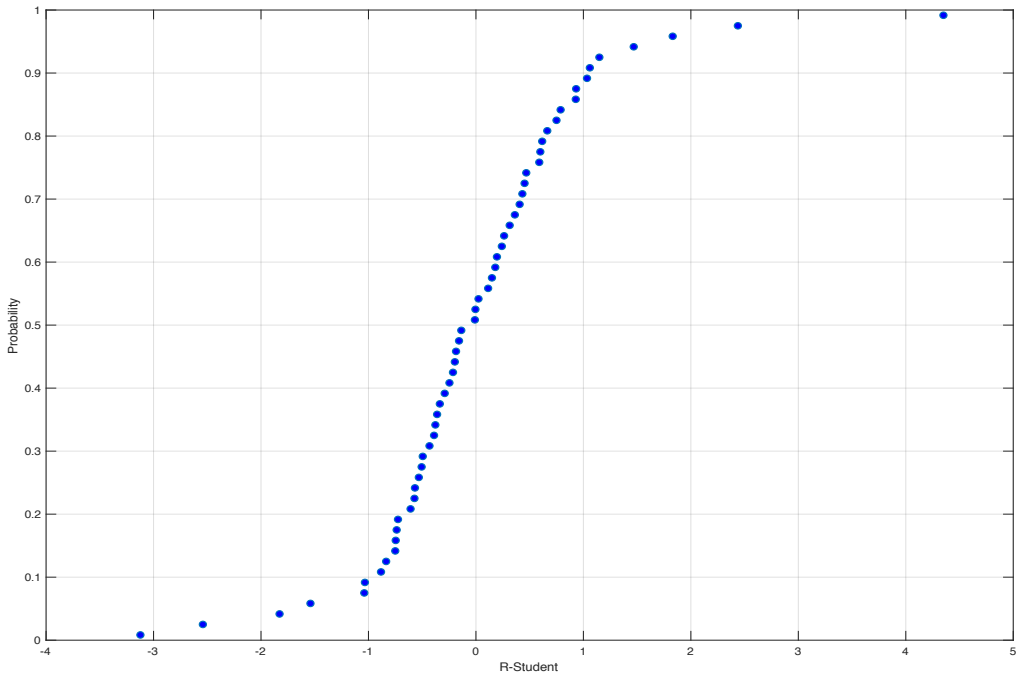

Figure 12: Normal probability plot of externally studentized residuals (t_i) from alternative regression model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 \ln(x_3) + \beta_5 \ln(x_5) + \varepsilon$

The normal probability plot for the alternative regression model remains unchanged from its predecessor. The plot also conveys that the distribution of the random error ε is heavily-tailed distribution (i.e. diverges from the normality assumption). However, the remainder of the statistical inferences will be conducted under the notion that the normality assumption was satisfied.

Figures 13-15 plot the externally studentized residuals versus the predicted response $\hat{y}$, and the two original regressors x_1 and x_2. In the original regression model, the response and the regressors x_1 and x_2 were shown to be uncorrelated with the residuals (and random error ε). These residual plots will determine if the transformation of regressors x_3 and x_5 introduced any new correlations between the random error and y, x_1, or x_2:

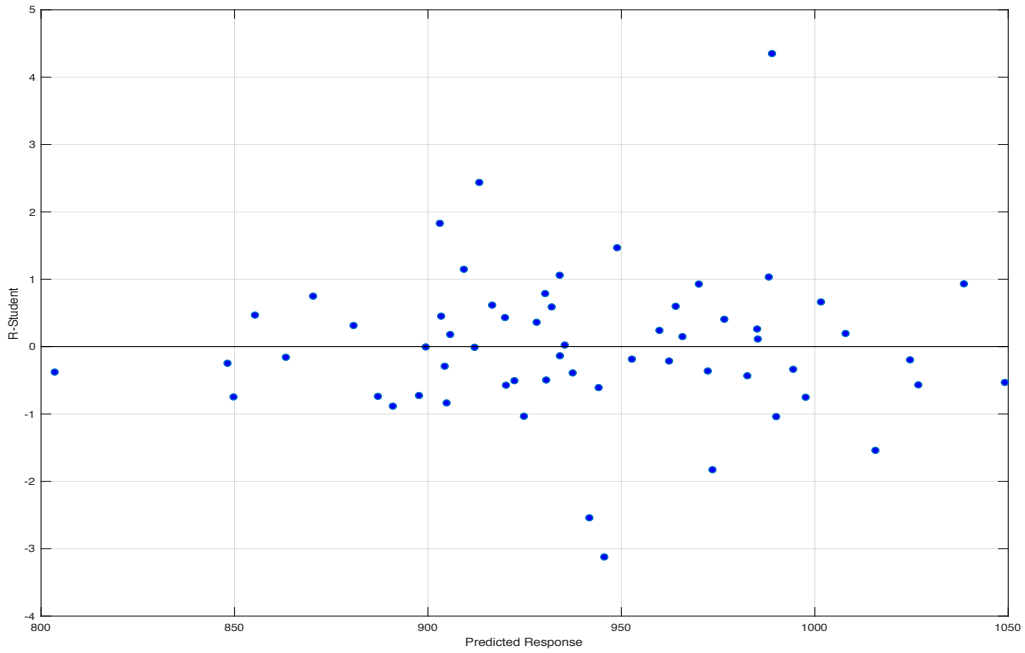

Figure 13: Plot of externally studentized residuals (t_i) versus predicted response $\hat{y}_i$ for alternative regression model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 \ln(x_3) + \beta_5 \ln(x_5) + \varepsilon$

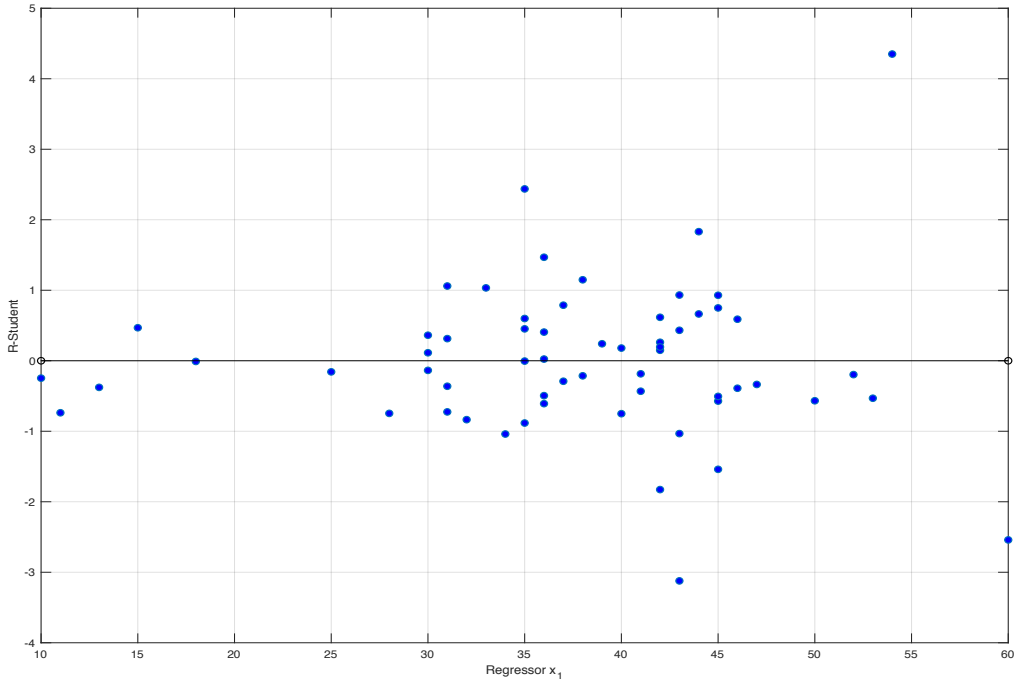

Figure 14: Plot of externally studentized residuals (t_i) versus regressor x_1 for alternative regression model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 \ln(x_3) + \beta_5 \ln(x_5) + \varepsilon$

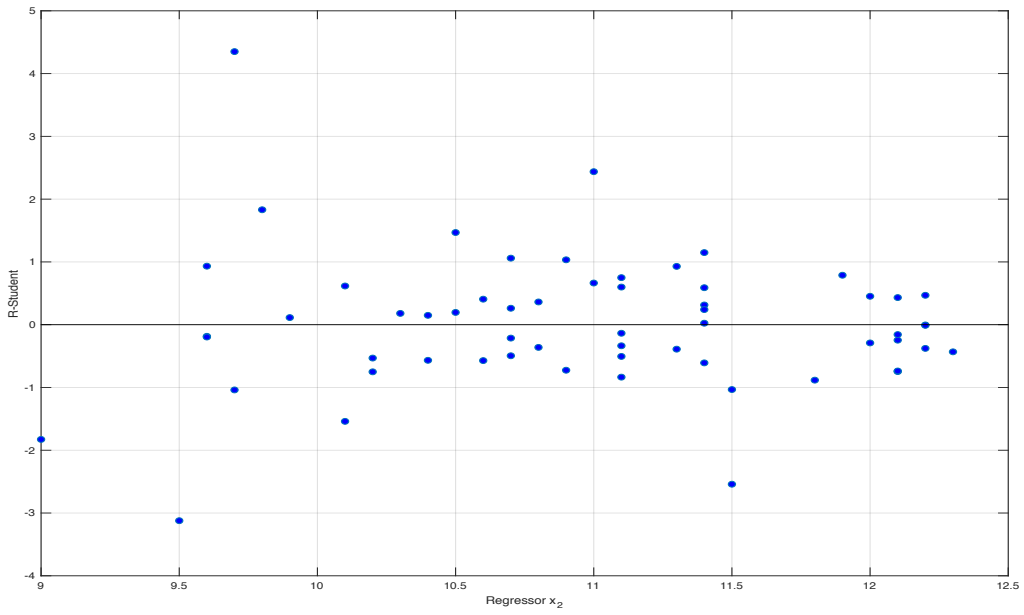

Figure 15: Plot of externally studentized residuals (t_i) versus regressor x_2 for alternative regression model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 \ln(x_3) + \beta_5 \ln(x_5) + \varepsilon$

The externally studentized residuals in all three plots (Figures 13-15) are contained in horizontal bands, which illustrates that the transformations applied to regressors x_3 and x_5 preserved the relationship between the random error and y, x_1, x_2 (i.e. the random error remains uncorrelated with response y and regressors x_1, x_2).

Figures 16-17 plot the externally studentized residuals of the n=60 observations versus the transformed regressors $\ln(x_3)$, $\ln(x_5)$. These residual plots will determine if the transformed variables satisfy the constant random error variance $Var(\varepsilon)$ assumption or if another transformation is required:

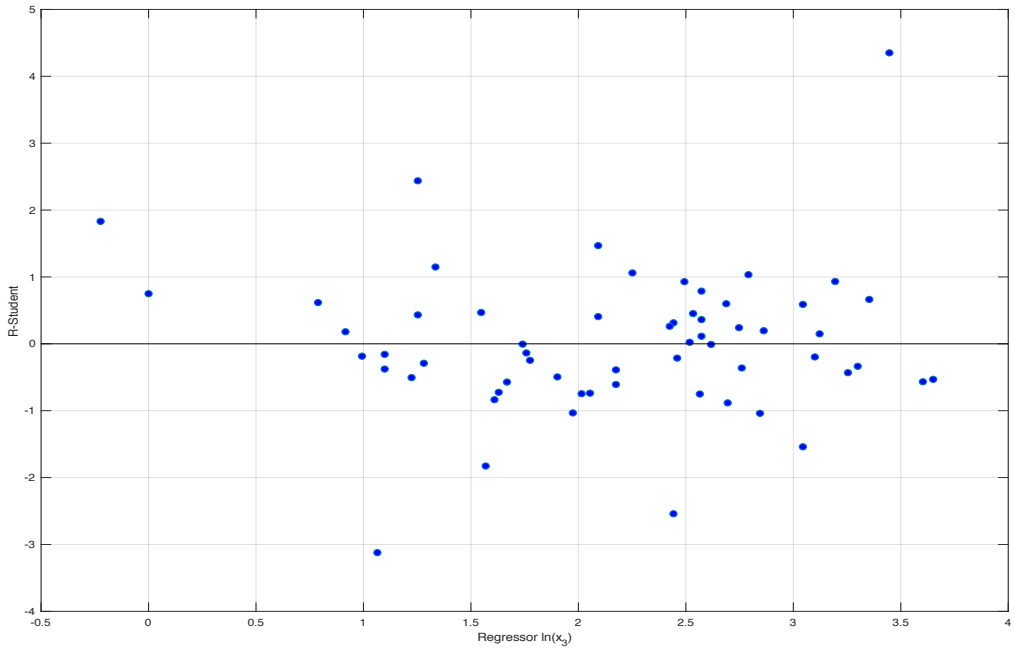

Figure 16: Plot of externally studentized residuals (t_i) versus regressor $\ln(x_3)$ for alternative regression model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 \ln(x_3) + \beta_5 \ln(x_5) + \varepsilon$

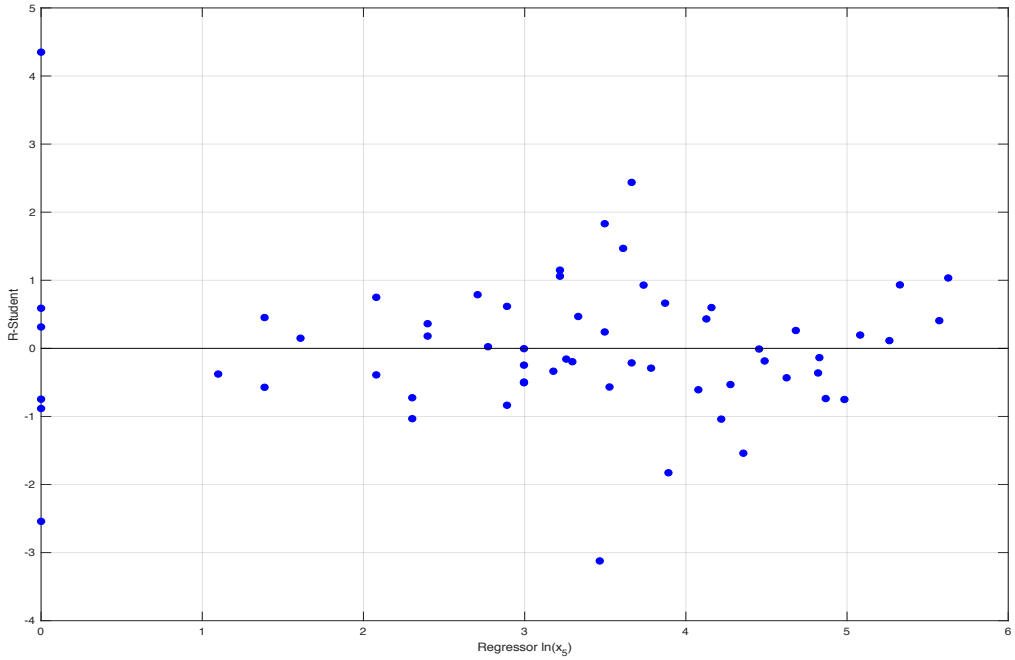

Figure 17: Plot of externally studentized residuals (t_i) versus regressor $\ln(x_5)$ for alternative regression model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 \ln(x_3) + \beta_5 \ln(x_5) + \varepsilon$

In contrast from the inward-opening funnels shown in the residual plots of x_3 and x_5 for the previous regression model, the externally-studentized residuals computed using the transformed regressors appear to be contained in horizontal bands. Thus, the residual plots suggest that the random error variance is approximately constant for all regressors in the updated regression model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 \ln(x_3) + \beta_5 \ln(x_5) + \varepsilon$.

Review of Model Construction and Variable Selection Procedure

These two steps conclude the model construction and variable selection procedure. After considering all possible regression models, the candidate regression model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_5 x_5 + \varepsilon$ was selected based on a set of three model selection criteria (i.e. coefficient of determination, residual mean sum of squares, Mallow's C_p statistic).

Model adequacy checks identified a violation of one of the model assumptions, so a transformation was applied to obtain an updated model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 \ln(x_3) + \beta_5 \ln(x_5) + \varepsilon$. Model adequacy checks were repeated for this transformed model and confirmed that no flagrant model inadequacies were detected. Therefore, the regression model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 \ln(x_3) + \beta_5 \ln(x_5) + \varepsilon$ will be used in the next two sections for statistical inference and prediction.

Assessing Statistical Significance of Regression Coefficients

Based on the assumptions made for the regression model, the estimates of the regression coefficients are determined via the method of ordinary least squares:

$$\widehat{\boldsymbol{\beta}} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{\beta}_3 \\ \hat{\beta}_5 \end{pmatrix} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} = \begin{pmatrix} 919.088 \\ 2.157 \\ -15.939 \\ 31.328 \\ 14.948 \end{pmatrix}$$

where $\boldsymbol{X}$ is the 60 x 5 design matrix containing the transformed regressors determined from the all-possible-regression model approach, and $\boldsymbol{y}$ is the 60 x 1 vector of response observations (mortality rate). Therefore, the fitted regression model for predicting mortality rate is $\hat{y} = 919.088 + 2.157x_1 - 15.939x_2 + 31.328\ln(x_3) + 14.948 + \ln(x_5)$.

Significance of Regression (Full Model)

Before evaluating the predictive power of this model for deployment, a test for significance of regression is conducted to identify any regressors whose impact on the response may be statistically insignificant in comparison to the impact of the other regressors. For significance of regression, the null hypothesis is $H_0: \beta_1 = \beta_2 = \beta_3 = \beta_5 = 0$. An assumed level of significance $\alpha = 0.05$ is used for hypothesis testing.

The total, regression, and residual sum of squares are computed from the observations and the fitted model:

$$SS_T = \boldsymbol{y}'\boldsymbol{y} - \frac{(\sum_{i=1}^{60} y_i)^2}{60} = 2.283 * 10^5$$

$$SS_R = \widehat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{y} - \frac{(\sum_{i=1}^{60} y_i)^2}{60} = 1.531 * 10^5$$

$$SS_{res} = \boldsymbol{y}'\boldsymbol{y} - \widehat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{y} = 7.514 * 10^4$$

The regression and residual mean sum of squares are $MS_R = \frac{SS_R}{4} = 3.828 * 10^4$ and $MS_{res} = \frac{SS_{res}}{60-5} = 1.366 * 10^3$. The F-test statistic is computed:

$$F_0 = \frac{MS_R}{MS_{res}} = \frac{3.828 * 10^4}{1.366 * 10^3} = 28.020$$

The rejection criterion for this F-test is $F_{\alpha,4,55} = F_{0.05,4,55} = 2.540$. Since $F_0 > F_{0.05,4,55}$, the null hypothesis is rejected at $\alpha = 0.05$ level of significance. There is statistically sufficient evidence to believe that at least one of the regression coefficients $\beta_j, j = 1,2,3,5$ is nonzero.

Statistical Significance of Individual Regression Coefficients

After confirming that at least one of the regression coefficients is nonzero, hypothesis tests are conducted individually on the regression coefficients $\beta_j, j = 1,2,3,5$ to determine if they significantly contribute to the prediction of mortality rate y (i.e. conduct hypothesis tests on $H_0: \beta_J = 0, j = 1,2,3,5$). The four test statistics are computed:

$H_0: \beta_1 = 0$

$$F_0 = \frac{MS_R(\beta_1|\beta_2,\beta_3,\beta_5)}{MS_{res}} = \frac{\big(SS_R(\beta_1,\beta_2,\beta_3,\beta_5) - SS_R(\beta_2,\beta_3,\beta_5)\big)/1}{MS_{res}} = \frac{1.778 * 10^4}{1.366 * 10^3} = 13.017$$

$H_0: \beta_2 = 0$

$$F_0 = \frac{MS_R(\beta_2|\beta_1,\beta_3,\beta_5)}{MS_{res}} = \frac{\big(SS_R(\beta_1,\beta_2,\beta_3,\beta_5) - SS_R(\beta_1,\beta_3,\beta_5)\big)/1}{MS_{res}} = \frac{6.994 * 10^3}{1.366 * 10^3} = 5.119$$

$H_0: \beta_3 = 0$

$$F_0 = \frac{MS_R(\beta_3|\beta_1,\beta_2,\beta_5)}{MS_{res}} = \frac{\big(SS_R(\beta_1,\beta_2,\beta_3,\beta_5) - SS_R(\beta_1,\beta_2,\beta_5)\big)/1}{MS_{res}} = \frac{3.944 * 10^4}{1.366 * 10^3} = 28.870$$

$H_0: \beta_5 = 0$

$$F_0 = \frac{MS_R(\beta_5|\beta_1,\beta_2,\beta_3)}{MS_{res}} = \frac{\big(SS_R(\beta_1,\beta_2,\beta_3,\beta_5) - SS_R(\beta_1,\beta_2,\beta_3)\big)/1}{MS_{res}} = \frac{2.485 * 10^4}{1.366 * 10^3} = 18.189$$

The rejection criterion for these F-tests are $F_{\alpha,1,55} = F_{0.05,1,55} = 4.016$. For all four hypothesis tests, $F_0 > F_{0.05,1,55}$, and all four null hypotheses are rejected at $\alpha = 0.05$ level of significance. There is statistically sufficient evidence to believe that each of the individual regression coefficients $\beta_j, j = 1, 2, 3, 5$ is nonzero. After computationally verifying that each of the four regression coefficients are statistically significant parameters for the n=60 observations, the predictive capability of the regression model will be assessed.

Measuring Predictive Power of Regression Model

Before deploying this regression model into practice, its predictive performance is evaluated using cross-validation in order to establish an expectation of the accuracy and potential limitations of the regression model's predictive power. Two cross-validation techniques are conducted to test the model's predictive power. First, the predicted residual error sum of squares (PRESS) statistic is computed for the fitted observations to measure the fit of the regression model to observations that were not used to estimate the regression coefficients. After this initial test, the original dataset is randomly partitioned into two datasets of equal size: a training set (i.e. dataset used to fit the regression model) and a test set (i.e. dataset used to evaluate the model's effectiveness in predicting mortality rates from unfitted observations).

PRESS Statistic for Entire Dataset of Observations

The PRESS statistic is a global measure of how well the model predicts the response observations of an entire dataset. It is a summation of the squared PRESS residuals, which are determined when one of the observations in the original dataset is not fitted to the regression model. The PRESS residuals are computed from the ordinary residuals $e_i = y_i - \hat{y}_i, i = 1, \dots, 60$ and the hat matrix $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$:

$$e_{(i)} = y_i - \hat{y}_{(i)} = \frac{e_i}{1 - h_{ii}} = \frac{y_i - \hat{y}_i}{1 - h_{ii}}, i = 1, \dots, 60$$

where h_{ii} is the ith diagonal entry of the hat matrix $\boldsymbol{H}$. The PRESS statistic and the prediction coefficient of determination are calculated:

$$PRESS = \sum_{i=1}^{60} \left(y_i - \hat{y}_{(i)}\right)^2 = \sum_{i=1}^{60} \left(\frac{e_i}{1 - h_{ii}}\right)^2 = 9.858 * 10^4$$

$$R^2_{pred} = 1 - \frac{PRESS}{SS_T} = \frac{9.858 * 10^4}{2.283 * 10^5} = 0.568$$

As evidenced by $R^2_{pred} = 0.568$, the regression model accounts for 56.8% of the variance of the predicted observations. This prediction coefficient of determination is adequate, but most likely not sufficient enough for model deployment.

Cross-Validation Using Training and Test Data

The PRESS statistic computed in the previous section measured the model's predictive power when only one observation in the original dataset was omitted from model fitting. In this section, a different PRESS statistic will be computed that measures the model's predictive power when many observations (i.e. half of the original dataset) are omitted from the original dataset. This model validation procedure is intended to replicate the model's performance in its operating environment (i.e. predicting mortality rates for new data points (or cities) based on their air pollution data).

The original dataset is randomly partitioned into a training set (n=30 observations) that is used to refit the regression model and a test set (n=30 observations) that will be used to evaluate the refitted regression model's predictive performance. The estimates of the regression coefficients for the model fitted to the training set are computed:

$$\widehat{\boldsymbol{\beta}}_{\boldsymbol{tr}} = \begin{pmatrix} \hat{\beta}_{0,tr} \\ \hat{\beta}_{1,tr} \\ \hat{\beta}_{2,tr} \\ \hat{\beta}_{3,tr} \\ \hat{\beta}_{5,tr} \end{pmatrix} = (\boldsymbol{X}_{\boldsymbol{tr}}'\boldsymbol{X}_{\boldsymbol{tr}})^{-1}\boldsymbol{X}_{\boldsymbol{tr}}'\boldsymbol{y}_{\boldsymbol{tr}} = \begin{pmatrix} 788.762 \\ 3.029 \\ -6.998 \\ 24.654 \\ 19.573 \end{pmatrix}$$

where $\boldsymbol{X}_{\boldsymbol{tr}}$ is the 30 x 5 design matrix for the training set and $\boldsymbol{y}_{\boldsymbol{tr}}$ is a 30 x 1 vector of response observations for the training set.

In comparison to the regression coefficient estimates for the model fitted to all n=60 observations, the regression coefficient estimates for the model fitted to the training set are identical in sign. However, the magnitudes of the estimates for the intercept and regression coefficient β_2 are considerably different between the model fitted to the entire dataset and the model fitted to the training set. This deviation can potentially be explained by the F-test statistic $F_0 = 5.119$ for testing H_0: $\beta_2 = 0$ that marginally exceeded the rejection criterion $F_{\alpha,1,55} = F_{0.05,1,55} = 4.016$ (i.e. the regressor x_2 may be marginally significant in predicting response y).

The fitted values of the test set responses $\widehat{\boldsymbol{y}}_{\boldsymbol{test}} = \boldsymbol{X}_{\boldsymbol{test}}\widehat{\boldsymbol{\beta}}_{\boldsymbol{tr}}$ are determined from the fitted regression coefficients, where $\boldsymbol{X}_{\boldsymbol{test}}$ is the 30 x 5 design matrix for the test set. The PRESS residuals $e_{(i)} = y_{test,i} - \hat{y}_{test,i}, i = 1, \dots, 30$ are computed for the 30 observations in the test set. The PRESS statistic and prediction coefficient of determination for the test set observations are computed:

$$PRESS = \sum_{i=1}^{30} \left(y_{test,i} - \hat{y}_{test,i}\right)^2 = 6.164 * 10^4$$

$$R^2_{pred} = 1 - \frac{PRESS}{SS_{T,test}} = \frac{6.164 * 10^4}{1.285 * 10^5} = 0.521$$

where $SS_{T,test} = \boldsymbol{y'_{test}y_{test}} - \frac{\left(\sum_{i=1}^{30} y_{test,i}\right)^2}{30} = 1.285 * 10^5$ is the total sum of square of the response observations in the test set.

For the test set observations, the regression model fitted to the training set data accounts for 52.1% of the variance of the predicted observations. This prediction coefficient of determination is approximately equal to the coefficient of determination computed from the PRESS statistic of the entire dataset of *n*=60 observations. Therefore, these results suggest that the predictive performance of the regression model during deployment is expected to be comparable to its performance on the training set data of *n*=60 observations.

Conclusions and Future Recommendations

This report concludes that a multiple linear regression model for predicting a city's adjusted mortality rate is:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 \ln(x_3) + \beta_5 \ln(x_5) + \varepsilon$$

where x_1 is the mean annual precipitation in inches, x_2 is the median number of school years completed for those over 25 in 1960 SMSA, $\ln(x_3)$ is the percentage of urbanized area population that is nonwhite (log-scale), and $\ln(x_5)$ is the relative pollution potential of sulfur dioxide (SO_2) (log-scale). Model adequacy checks for this multiple linear regression model identified deviations from the normality assumption made for the random errors (i.e. heavier-tailed sampling distribution of random error) and four potential influential points among the *n*=60 observations.

Tests for significance of regression determined that there is statistically significant evidence that all regression coefficients $\beta_1, \beta_2, \beta_3, \beta_5$ are nonzero (i.e. contribute to the prediction of the mortality rate *y*). However, it is worth noting that the regression coefficient β_2 for regressor x_2 exhibited a *F*-test statistic that marginally exceed the rejection criterion for an assumed level of significance of $\alpha = 0.05$.

In the two cross-validation studies that were conducted, the regression model demonstrated that it is an adequate predictor of age-adjusted mortality rate; however, its predictive power is likely not sufficient enough for the team of statisticians and data scientists to recommend model deployment. In light of these results, the team has compiled a list of future studies to improve upon this multiple linear regression model:

1. Repeat the tests for significance of regression using a stricter level of significance, such as $\alpha = 0.01$. The regression coefficient β_2 exhibited a *F*-test statistic that was marginally deemed as significant. Thus, it is plausible that a lower level of significance could

classify β_2 as statistically insignificant. Therefore, a repeated test for significance using a stricter rejection criterion and accompanying cross-validation studies would verify if the regression coefficient β_2 improves or weakens the predictive power of the regression model.

2. Implement non-parametric regression analysis methods (i.e. robust estimation, regression on ranked data) to determine estimates of the regression coefficients. From the model adequacy checks, there is strong evidence that the sampling distribution of the random error is heavily-tailed. Consequently, outliers and extreme values are more observable from a heavily-tailed sampling distribution than the standard normal distribution. Therefore, parameter estimation methods that dampen the impact of outliers or extreme values would improve the accuracy and stabilize the estimates of the regression coefficients.

3. Examine the four potential outliers in the dataset. The four studentized residuals for these observations are -3.1223, -2.5411, 2.4375, and 4.3503. These four points correspond to the air pollution and mortality rate data for Lancaster, PA, Miami, FL, Albany, NY, and New Orleans, LA, respectively. The data collection team can determine if there is any sufficient non-statistical evidence to believe that these observations could be erroneous. If so, then the model construction and regression analysis procedure can be repeated on the dataset, excluding the erroneous measurements. On the other hand, if a surplus in funds for data collection are available, more data on other air pollution measurables can be collected to determine if there are other variables besides the original five candidate regressors that could explain the observed variance in the response of these four outliers.

As a result, it is advised that other forms of regression models, such as nonlinear and nonparametric regression, be constructed and compared to this multiple linear regression model. In addition, the normality assumption and significance of the regression coefficients for these candidate nonlinear and nonparametric regression models should also be evaluated for model adequacy. This report summarizes the findings for only one multiple linear regression model; therefore, the implementation of other model construction methods would confirm if a more powerful predictive model for mortality rate is attainable.

APPENDIX: MATLAB Code

```
clear;
cd '/Users/davidchung/Documents/ACM/Spring 2019/Statistical Models and
Regression/Textbook Datasets/Appendices'
data = importdata('data-table-B-15_csv.csv', ',');
y = data.data(:, 1);
X = data.data(:, [2:6]);
n = size(X,1)
X = [ones(n,1) X];
cd ..; cd ..;
K = size(X, 2) - 1
```

Construct regression model with all regressors:

```
beta_hat_full = inv(X'*X)*X'*y
SS_tot = y'*y - (sum(y))^2/n
SS_R_full = beta_hat_full'*X'*y - (sum(y))^2/n
R2_full = SS_R_full/SS_tot
SS_res_full = SS_tot - SS_R_full
MS_res_full = SS_res_full/(n-K-1)

y_hat = X*beta_hat_full;
e = y-y_hat;
H = X*inv(X'*X)*X';
```

Construct model with all possible regressions method:

```
% 1 regressor (p=2)
p = 2;
R2_2 = [];
MS_res_2 = [];
Cp_2 = [];
for i=1:K
    X_test = [ones(n,1) X(:,i+1)];
    beta_hat = inv(X_test'*X_test)*X_test'*y;
    SS_R = beta_hat'*X_test'*y - (sum(y))^2/n;
    SS_res = SS_tot - SS_R;
    R2_2 = [R2_2 SS_R/SS_tot];
    MS_res_2 = [MS_res_2 SS_res/(n-p)];
    Cp_2 = [Cp_2 SS_res/MS_res_full - n + 2*p];

    if SS_R/SS_tot >= max(R2_2)
        R2_2_max = i;
    end
    if SS_res/(n-2) <= min(MS_res_2)
        MS_res_2_min = i;
    end
    if SS_res/MS_res_full - n + 2*p <= min(Cp_2)
        Cp_2_min = i;
    end
end % x3

% 2 regressors (p=3)
p = 3;
R2_3 = [];
```

```
MS_res_3 = [];
Cp_3 = [];
for i=1:K
    for j=i+1:K
        X_test = [ones(n,1) X(:,i+1) X(:,j+1)];
        beta_hat = inv(X_test'*X_test)*X_test'*y;
        SS_R = beta_hat'*X_test'*y - (sum(y))^2/n;
        SS_res = SS_tot - SS_R;
        R2_3 = [R2_3 SS_R/SS_tot];
        MS_res_3 = [MS_res_3 SS_res/(n-p)];
        Cp_3 = [Cp_3 SS_res/MS_res_full - n + 2*p];

        if SS_R/SS_tot >= max(R2_3)
            R2_3_max = [i j];
        end
        if SS_res/(n-p) <= min(MS_res_3)
            MS_res_3_min = [i j];
        end
        if SS_res/MS_res_full - n + 2*p <= min(Cp_3)
            Cp_3_min = [i j];
        end
    end
end % x2, x3

% 3 regressors (p=4)
p = 4;
R2_4 = [];
MS_res_4 = [];
Cp_4 = [];
for i=1:K
    for j=i+1:K
        for l=j+1:K
            X_test = [ones(n,1) X(:,i+1) X(:,j+1) X(:,l+1)];
            beta_hat = inv(X_test'*X_test)*X_test'*y;
            SS_R = beta_hat'*X_test'*y - (sum(y))^2/n;
            SS_res = SS_tot - SS_R;
            R2_4 = [R2_4 SS_R/SS_tot];
            MS_res_4 = [MS_res_4 SS_res/(n-p)];
            Cp_4 = [Cp_4 SS_res/MS_res_full - n + 2*p];

            if SS_R/SS_tot >= max(R2_4) && SS_R/SS_tot <=1
                R2_4_max = [i j l];
            end
            if SS_res/(n-p) <= min(MS_res_4)
                MS_res_4_min = [i j l];
            end
            if SS_res/MS_res_full - n + 2*p <= min(Cp_4)
                Cp_4_min = [i j l];
            end
        end
    end
end % x1, x3, x5

% 4 regressors (p=5)
p = 5;
```

```
R2_5 = [];
MS_res_5 = [];
Cp_5 = [];
for i=1:K
    for j=i+1:K
        for l=j+1:K
            for m=l+1:K
                X_test = [ones(n,1) X(:,i+1) X(:,j+1) X(:,l+1) X(:,m+1)];
                beta_hat = inv(X_test'*X_test)*X_test'*y;
                SS_R = beta_hat'*X_test'*y - (sum(y))^2/n;
                SS_res = SS_tot - SS_R;
                R2_5 = [R2_5 SS_R/SS_tot];
                MS_res_5 = [MS_res_5 SS_res/(n-p)];
                Cp_5 = [Cp_5 SS_res/MS_res_full - n + 2*p];

                if SS_R/SS_tot >= max(R2_5) && SS_R/SS_tot <=1
                    R2_5_max = [i j l m];
                end
                if SS_res/(n-p) <= min(MS_res_5)
                    MS_res_5_min = [i j l m];
                end
                if SS_res/MS_res_full - n + 2*p <= min(Cp_5)
                    Cp_5_min = [i j l m];
                end
            end
        end
    end
end % x1, x2, x3, x5

% Plot R2(p), MS_res(p), Cp
figure(1) % R2_p
plot([2 3 4 5 6], [max(R2_2), max(R2_3), max(R2_4), max(R2_5), R2_full]);
figure(2) % MS_res(p)
plot([2 3 4 5 6], [min(MS_res_2) min(MS_res_3) min(MS_res_4) min(MS_res_5)
MS_res_full]);
figure(3) %C_p
plot([2 3 4 5 6], [min(Cp_2) min(Cp_3) min(Cp_4) min(Cp_5) 6]);
```

Select model with x1, x2, x3, x5. Assess normality assumption and model adequacy:

```
p=5;
X_ap = X(:,[1:4 6]);
beta_hat_ap = inv(X_ap'*X_ap)*X_ap'*y
SS_R_ap = beta_hat_ap'*X_ap'*y - (sum(y))^2/n
SS_res_ap = SS_tot - SS_R_ap
MS_res_ap = SS_res_ap/(n-p)

y_hat_ap = X_ap*beta_hat_ap;
e_ap = y - y_hat_ap;
H_ap = X_ap*inv(X_ap'*X_ap)*X_ap';

% Normality assumption

for i=1:n
    S_2i = ((n-p)*MS_res_ap - (e_ap(i))^2/(1-H_ap(i,i)))/(n-p-1);
    t_ap(i) = e_ap(i)/sqrt(S_2i*(1-H_ap(i,i)));
```

```
    r_ap(i) = e_ap(i)/sqrt(MS_res_ap*(1-H_ap(i,i)));
    P_ap(i) = (i-0.5)/n;
end
figure(1)
t_sorted = sort(t_ap)';
plot(t_sorted, P_ap);

% Residual plots
figure(2)
plot(y_hat_ap, t_ap);
figure(3)
plot(X_ap(:,2), t_ap); % x1
figure(4)
plot(X_ap(:,3), t_ap); % x2
figure(5)
plot(X_ap(:,4), t_ap); % x3
figure(6)
plot(X_ap(:,5), t_ap) % x5
```

Response appears to have nonlinear relationship with x3, x5 - plot scatterplots of x3, x5 vs y - apply natural log transformation:

```
figure(7)
plot(X_ap(:,4), y)
figure(8)
plot(X_ap(:,5), y)

X_ap_t = [ones(n,1) X_ap(:,[2 3]) log(X_ap(:,4)) log(X_ap(:,5))];
beta_hat_ap_t = inv(X_ap_t'*X_ap_t)*X_ap_t'*y
SS_R_ap_t = beta_hat_ap_t'*X_ap_t'*y - (sum(y))^2/n
SS_res_ap_t = SS_tot - SS_R_ap_t
MS_res_ap_t = SS_res_ap_t/(n-p)

y_hat_ap_t = X_ap_t*beta_hat_ap_t;
e_ap_t = y - y_hat_ap_t;
H_ap_t = X_ap_t*inv(X_ap_t'*X_ap_t)*X_ap_t';

% Normality assumption

for i=1:n
    S_2i = ((n-p)*MS_res_ap_t - (e_ap_t(i))^2/(1-H_ap_t(i,i)))/(n-p-1);
    t_ap_t(i) = e_ap_t(i)/sqrt(S_2i*(1-H_ap_t(i,i)));
    r_ap_t(i) = e_ap_t(i)/sqrt(MS_res_ap_t*(1-H_ap_t(i,i)));
    P_ap_t(i) = (i-0.5)/n;
end
figure(9)
t_sorted = sort(t_ap_t)';
plot(t_sorted, P_ap_t);

% Residual plots (SAVE PLOTS)
figure(10)
plot(y_hat_ap_t, t_ap_t);
figure(11)
plot(X_ap_t(:,2), t_ap_t); % x1
figure(12)
plot(X_ap_t(:,3), t_ap_t); % x2
```

```
figure(13)
plot(X_ap_t(:,4), t_ap_t); % ln(x3)
figure(14)
plot(X_ap_t(:,5), t_ap_t); % ln(x5)
```

Test for significance of regression:

```
alpha = 0.05;
MS_R_ap_t = SS_R_ap_t/(p-1);
MS_res_ap_t = SS_res_ap_t/(n-p);

F_0 = MS_R_ap_t/MS_res_ap_t
F_crit = finv(1-alpha, p-1, n-p)

% test for x1 coefficient

X_test_sig = [ones(n,1) X_ap_t(:,[3:5])];
beta_hat_sig = inv(X_test_sig'*X_test_sig)*X_test_sig'*y;
SS_R_sig = beta_hat_sig'*X_test_sig'*y - (sum(y))^2/n
SS_R_diff = SS_R_ap_t - SS_R_sig

F_0_sig = SS_R_diff/MS_res_ap_t
F_crit_sig = finv(1-alpha, 1, n-p)

% test for x2 coefficient

X_test_sig = [ones(n,1) X_ap_t(:,[2 4:5])];
beta_hat_sig = inv(X_test_sig'*X_test_sig)*X_test_sig'*y;
SS_R_sig = beta_hat_sig'*X_test_sig'*y - (sum(y))^2/n
SS_R_diff = SS_R_ap_t - SS_R_sig

F_0_sig = SS_R_diff/MS_res_ap_t
F_crit_sig = finv(1-alpha, 1, n-p)

% test for ln(x3) coefficient

X_test_sig = [ones(n,1) X_ap_t(:,[2:3 5])];
beta_hat_sig = inv(X_test_sig'*X_test_sig)*X_test_sig'*y;
SS_R_sig = beta_hat_sig'*X_test_sig'*y - (sum(y))^2/n
SS_R_diff = SS_R_ap_t - SS_R_sig

F_0_sig = SS_R_diff/MS_res_ap_t
F_crit_sig = finv(1-alpha, 1, n-p)

% test for ln(x5) coefficient

X_test_sig = [ones(n,1) X_ap_t(:,[2:4])];
beta_hat_sig = inv(X_test_sig'*X_test_sig)*X_test_sig'*y;
SS_R_sig = beta_hat_sig'*X_test_sig'*y - (sum(y))^2/n
SS_R_diff = SS_R_ap_t - SS_R_sig

F_0_sig = SS_R_diff/MS_res_ap_t
F_crit_sig = finv(1-alpha, 1, n-p)
```

Test for model validation with transformed model:

```
SS_PRESS = 0;
```

```
for i=1:n
    SS_PRESS = SS_PRESS + (e_ap_t(i)/(1-H_ap_t(i,i)))^2;
end

SS_PRESS % PRESS (prediction residual) sum of squares
R2_pred = 1- (SS_PRESS/SS_tot)
```

Now delete half of observations (randomly) and refit regression model (have outliers in the test set iobservations 4, 7, 50)

```
X_train = X_ap_t;
X_test = X_ap_t;
y_train = y;
y_test = y;

X_train([3:8 11 13 16 18 23:25 27:28 30 32 38 42:43 47:51 53 55:56 58 60],
:) = []; % set up training/test sets
y_train([3:8 11 13 16 18 23:25 27:28 30 32 38 42:43 47:51 53 55:56 58 60])
= [];
X_test([1:2 9:10 12 14:15 17 19:22 26 29 31 33:37 39:41 44:46 52 54 57
59], :) = [];
y_test([1:2 9:10 12 14:15 17 19:22 26 29 31 33:37 39:41 44:46 52 54 57
59]) = [];

beta_hat_train = inv(X_train'*X_train)*X_train'*y_train
y_pred = X_test*beta_hat_train;

e_pred = y_test - y_pred; % prediction errors

PRESS_test = sum(e_pred.^2)
SS_T_pred = y_test'*y_test - (sum(y_test))^2/30;
R2_pred_test = 1 - PRESS_test/SS_T_pred
```