# Generalized Linear Models for Predicting the Levels of Nitrogen Oxides (NOX) using the Boston Housing Dataset

## Contents

# I. Introduction

## i. Description and Origin of Data

The dataset used in this analysis was originally compiled from multiple sources by David Harrison Jr. and Daniel L. Rubinfeld in their article titled "Hedonic Housing Prices and the Demand for Clean Air" which was published in the Journal of Environmental Economics and Management in 1978. The dataset contains 506 observations across 14 variables. The following table provides a summary of each of these variables and their sources.

| Variable | Description of Variable | Source |
|---|---|---|
| CRIM | Crime rate per capita by town. | FBI (1970) |
| ZN | Proportion of residential land zoned for lots over 25,000 square feet. | Metropolitan Area Planning Commission (1972) |
| INDUS | Proportion of non-retail business acres per town. | Vogt, Ivers, and Associates (1965) |
| CHAS | Dummy variable for proximity to the Charles River. 0 means houses are far from the river. 1 means houses are close to the river. | 1970 U.S. Census |
| NOX | Concentration of Nitrogen oxide in pphm. | TASSIM |
| RM | Average number of rooms. | 1970 U.S. Census |
| AGE | Percentage of houses built prior to 1940. | 1970 U.S. Census |
| DIS | Weighted distance to major employment center. | Schnare (1973) |
| RAD | Index value of accessibility to highways. | MIT Boston Project |
| TAX | Property tax rate. | Mass. Taxpayers Foundation (1970 |
| PTRATIO | Ratio of number of students to teachers by district. | Massachusetts Dept. of Education (1971-1972) |
| BL | Proportion of population that is African American | 1970 U.S. Census |

| | | |
|---|---|---|
| LSTAT | Log of the percentage of adults that are either classified as having not graduated high school or men whose profession is manual labor | 1970 U.S. Census |
| MEDV | Median value of homes as reported by homeowners (reported in units of $1000) | 1970 U.S. Census |

## ii. A question of interest

The question investigated in this report is how the level of air pollution, represented by the nitric oxide levels (NOX), is affected as a function of other variables.

According to Wikipedia, Nitrogen oxides (NOx) refers to a mixture of compounds such as nitric oxide (NO), nitrogen dioxide (NO2) and nitrous oxide (N2O). Nitrogen oxides occur naturally and also are produced by man's activities. The primary source of man-made nitrogen oxides is from the burning of fossil fuels. Nitrogen oxides help form acid rain and can also cause a wide range of health and environmental damage. By identifying what factors might be associated with the nitric oxide levels, proactive steps can be planned to limit nitrogen oxides emissions, thus to improve the air quality.

## iii. Dependent and Independent Variables

Six variables that might not be relevant to air pollution were removed from consideration when modeling the nitric oxide levels (NOX), which include the crime rate (CRIM), the proportion of residential land (ZN), average number of rooms per house (RM), tax rate (TAX), the student-teacher ratio (PTRATIO), and the proportion of black population (B). Seven independent variables were kept, which include four continuous variables and two categorical variables as listed below. The reasons why

| Variable | Description of Variable | Type of Variable | Reason for Inclusion |
|---|---|---|---|
| Dependent Variable | | | |
| NOX | nitric oxides concentration (parts per 10 million) | Continuous | |
| Independent Variables | | | |
| INDUS | Proportion of non-retail business acres per town | Continuous | Non-retail business might be a source of NOX emission. |
| AGE | Proportion of owner-occupied units built prior to 1940 | Continuous | Old-style stoves, ovens or heaters used in these houses might give off more air pollution. |
| DIS | weighted distances to five Boston employment centers | Continuous | Areas close to the major employment centers might have higher level of air pollution. |
| MEDV | Median value of owner-occupied homes in $1000's | Continuous | Houses with higher value should be associated with lower level of air pollution. |

| | | | |
|---|---|---|---|
| CHAS | Charles River dummy variable (= 1 if tract bounds river or 0 otherwise) | Categorical (Dichotomous) | Charles River might be a source of pollution which can be converted to air pollution. |
| RAD | Index of accessibility to radial highways (9 levels, 1-8 and 24) | Categorical (Nominal) | Areas close to radial highways might have higher level of air pollution. |

### iv. Basic Descriptive Analysis

The Proc univariate procedure was used to get basic statistics such as mean, median, standard deviation, range, and skewness as well as descriptive plots such as histograms and probability plots for each continuous variable. The purpose is to have a preliminary exploration on each variable before modeling. Detailed information can be found in Appendix A. All the histograms are skewed and don't follow a bell-shape which indicate the continuous variables don't follow a normal distribution.

## II Method

### i. Initial Modeling with General Linear Model

Since the independent variables involve both continuous variables and categorical variables, general linear model was used in initial modeling. First the GLMSELECT procedure was used to perform model selection in the framework of general linear model. Similar to REG procedure, GLMSELECT has forward, backward, and stepwise selection methods with a variety of fit criteria to specify.

The CLASS statement was put before the MODEL statement to specify the two categorical variables to be used in the analysis, CHAS and RAD. By specifying the selection =stepwise (select=SL) stats=all option in the MODEL statement, all the six variables were kept in the model. Even though CHAS would be excluded if based on some of the criteria, I decide to temporarily keep it for the GLM model in the next step.

**Stepwise Selection Summary**

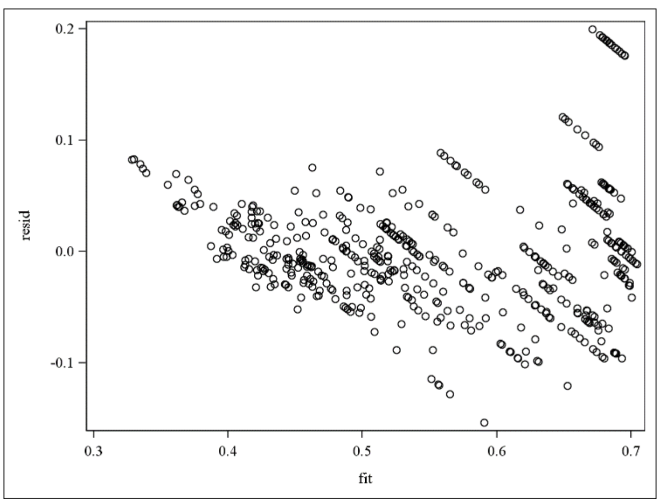
| Step | Effect Entered | Effect Removed | Number Effects In | Number Parms In | Model R-Square | Adjusted R-Square | AIC | AICC | BIC | CP |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | Intercept | | 1 | 1 | 0.0000 | 0.0000 | -1672.0838 | -1672.0599 | -2180.7949 | 1668.5677 |
| 1 | dis | | 2 | 2 | 0.5917 | 0.5909 | -2123.3534 | -2123.3056 | -2631.4407 | 385.0269 |
| 2 | indus | | 3 | 3 | 0.6879 | 0.6867 | -2257.2883 | -2257.2085 | -2764.9396 | 178.0556 |
| 3 | rad | | 4 | 11 | 0.7448 | 0.7397 | -2343.1898 | -2342.5569 | -2851.1248 | 70.3758 |
| 4 | age | | 5 | 12 | 0.7666 | 0.7614 | -2386.3988 | -2385.6589 | -2892.4458 | 24.9929 |
| 5 | medv | | 6 | 13 | 0.7721 | 0.7666 | -2396.4777 | -2395.6223 | -2901.9027 | 15.0337 |
| 6 | chas | | 7 | 14 | 0.7735 | 0.7676* | -2397.5881* | -2396.6085* | -2902.7929* | 14.0000* |

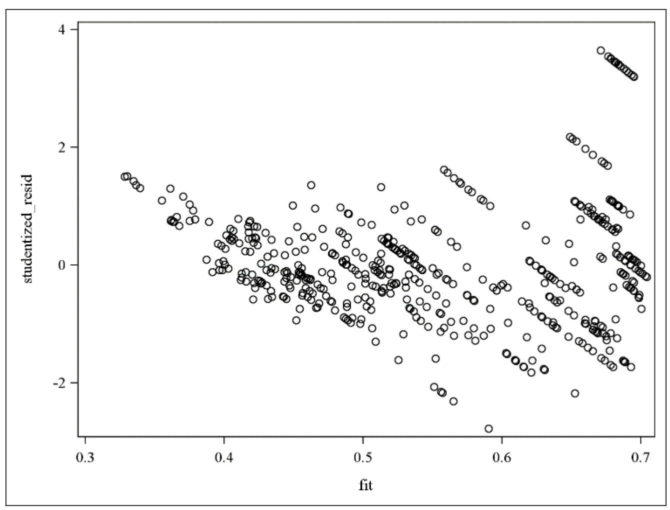
*** Optimal Value Of Criterion**

| Stepwise Selection Summary | | | | | | | |
|---|---|---|---|---|---|---|---|
| **Step** | **Effect Entered** | **Effect Removed** | **SBC** | **PRESS** | **ASE** | **F Value** | **Pr > F** |
| **0** | **Intercept** | | -2175.8573 | 6.8078 | 0.0134 | 0.00 | 1.0000 |
| **1** | **dis** | | -2622.9003 | 2.7951 | 0.0055 | 730.43 | <.0001 |
| **2** | **indus** | | -2752.6087 | 2.1460 | 0.0042 | 155.02 | <.0001 |
| **3** | **rad** | | -2804.6979 | 1.7869 | 0.0034 | 13.80 | <.0001 |
| **4** | **age** | | -2843.6803 | 1.6381 | 0.0031 | 46.17 | <.0001 |
| **5** | **medv** | | -2849.5327* | 1.6051* | 0.0031 | 11.91 | 0.0006 |
| **6** | **chas** | | -2846.4165 | 1.6096 | 0.0030 | 3.03 | 0.0822 |
| *** Optimal Value Of Criterion** | | | | | | | |

The six variables were used to fit the model by using PROC GLM procedure. Again, the CLASS statement was used before the MODEL statement since RAD and CHAS are categorical variables. The model is significantly better than an error-only model. However, the diagnostic plots show a clear trend that residuals increase as the predicted values increase. So generalized linear models may be a good option for the next step.

Detailed results on General Linear Model can be found in Appendix B.




## ii. Generalized Linear Model with Gamma Distribution

First a generalized linear model with gamma distribution and log link function was used to build the model. The GENMOD procedure was performed.

The type 1 analysis shows only one variable CHAS is not significant at a 0.05 level when the terms are added sequentially. The type 3 analysis shows that given all the other variables, only one variable CHAS is not significant at a 0.05 level. Based on the analyses of Type 1 and Type 3, the variable CHAS should be removed and the model should be refit. Examining the residuals and Cook's distances with the criterion as larger than .5, no observations should be removed.

The model was refit with all predictors except CHAS. All the parameters are significant at a 0.05 level in the estimation table except one level of RAD (RAD=5). Both the type 1 and type 3 analyses indicate that all the five terms should be retained in the model. No observations should be removed by using the Cook's distance to check the influential points. The residual plots look fine and there are no more issues.

Detailed results on the GENMOD procedure with Gamma distribution and log link function can be found in Appendix C.

### iii. Generalized Linear Model with Inverse Gaussian Distribution

Then a generalized linear model with a different distribution, the Inverse Gaussian distribution and log link function was also performed.

The type 1 analysis shows five variables INDUS, AGE, DIS, MEDV and RAD are significant at a level of 0.05. The type 3 analysis shows that given all the other variables, INDUS, AGE, DIS, MEDV and RAD are significant at a level of 0.05. The Cook's distance was used to check the influential points. All of the points have Cook's distance less than 0.5. Therefore the model needs to be refit by removing the variable CHAS.

The model was refit with these five variables identified above. All the parameters are significant at a 0.05 level in the estimation table except RAD=5. The type 1 and type 3 analyses both indicate that all these five terms are significant at a level of 0.05. All the observations have Cook's distance less than 0.5 and the diagnostic plots look good.

Detailed results on the GENMOD procedure with the Inverse Gaussian distribution and log link function can be found in Appendix D.

### iv. Comparison between Gamma Model and Inverse Gaussian Model

When comparing the Gamma model with Inverse Gaussian model, both have the same set of significant terms and similar parameter estimates. But the Inverse Gaussian model has smaller AIC, AICC and BIC, so it was chosen as the final model.

| Model Information | |
|---|---|
| Data Set | WORK.HOUSE |
| Distribution | Gamma |
| Link Function | Log |
| Dependent Variable | nox |

| Criteria For Assessing Goodness Of Fit | | | |
|---|---|---|---|
| Criterion | DF | Value | Value/DF |
| Deviance | 493 | 3.7069 | 0.0075 |
| Scaled Deviance | 493 | 506.6171 | 1.0276 |
| Pearson Chi-Square | 493 | 3.8160 | 0.0077 |
| Scaled Pearson X2 | 493 | 521.5200 | 1.0578 |
| Log Likelihood | | 834.2130 | |
| Full Log Likelihood | | 834.2130 | |
| AIC (smaller is better) | | -1640.4261 | |

| Model Information | |
|---|---|
| Data Set | WORK.HOUSE |
| Distribution | Inverse Gaussian |
| Link Function | Log |
| Dependent Variable | nox |

| Criteria For Assessing Goodness Of Fit | | | |
|---|---|---|---|
| Criterion | DF | Value | Value/DF |
| Deviance | 493 | 6.2905 | 0.0128 |
| Scaled Deviance | 493 | 506.0000 | 1.0264 |
| Pearson Chi-Square | 493 | 6.5577 | 0.0133 |
| Scaled Pearson X2 | 493 | 527.4967 | 1.0700 |
| Log Likelihood | | 855.0649 | |
| Full Log Likelihood | | 855.0649 | |
| AIC (smaller is better) | | -1682.1298 | |
| AICC (smaller is better) | | -1681.2744 | |
| BIC (smaller is better) | | -1622.9583 | |

### v. Final Generalized Linear Model with Inverse Gaussian Distribution and Log Link Function

Based on the parameter estimation table, the final model tells us that given all the other variables, one unit increase in the proportion of non-retail business acres per town (INDUS) corresponds to NOX increasing by 0.0075 (parts per 10 million); given all the other variables, one unit increase in the proportion of owner-occupied units built prior to 1940 (AGE) corresponds to NOX increasing by 0.0016 (parts per 10 million); given all the other variables, one unit increase in weighted distances to five Boston employment centers (DIS) corresponds to NOX decreasing by a multiple of -0.0302(parts per 10 million); given all the other variables, one unit increase in the median value of owner-occupied homes in $1000's (MEDV) corresponds to NOX decreasing 0.0018 (parts per 10 million). The index of accessibility to radial highways (RAD) is negatively associated with NOX, with the parameter for different levels of RAD ranging from 0 to -0.1437.

| Analysis Of Maximum Likelihood Parameter Estimates | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **Parameter** | | **DF** | **Estimate** | **Standard Error** | **Wald 95% Confidence Limits** | | **Wald Chi-Square** | **Pr > ChiSq** |
| **Intercept** | | 1 | -0.5885 | 0.0335 | -0.6542 | -0.5228 | 308.17 | <.0001 |
| **indus** | | 1 | 0.0075 | 0.0010 | 0.0056 | 0.0094 | 59.94 | <.0001 |
| **age** | | 1 | 0.0016 | 0.0002 | 0.0012 | 0.0020 | 65.30 | <.0001 |
| **dis** | | 1 | -0.0302 | 0.0029 | -0.0358 | -0.0246 | 112.23 | <.0001 |
| **rad** | 1 | 1 | -0.0801 | 0.0213 | -0.1217 | -0.0384 | 14.16 | 0.0002 |
| **rad** | 2 | 1 | -0.1437 | 0.0194 | -0.1818 | -0.1056 | 54.61 | <.0001 |
| **rad** | 3 | 1 | -0.1113 | 0.0178 | -0.1462 | -0.0764 | 39.07 | <.0001 |
| **rad** | 4 | 1 | -0.1102 | 0.0128 | -0.1352 | -0.0852 | 74.69 | <.0001 |
| **rad** | 5 | 1 | -0.0237 | 0.0129 | -0.0490 | 0.0017 | 3.35 | 0.0674 |
| **rad** | 6 | 1 | -0.0867 | 0.0190 | -0.1239 | -0.0495 | 20.82 | <.0001 |
| **rad** | 7 | 1 | -0.0845 | 0.0224 | -0.1284 | -0.0405 | 14.19 | 0.0002 |
| **rad** | 8 | 1 | -0.0852 | 0.0203 | -0.1250 | -0.0454 | 17.61 | <.0001 |
| **rad** | 24 | 0 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | . | . |
| **medv** | | 1 | -0.0018 | 0.0005 | -0.0027 | -0.0008 | 12.37 | 0.0004 |
| **Scale** | | 1 | 0.1115 | 0.0035 | 0.1048 | 0.1186 | | |

## III. ANOVA Analysis on RAD

Since categorical variable RAD has been identified as a significant term in both Gamma model and Inverse Gaussian model, a one-way ANOVA analysis was performed to compare the NOX differences in these regions with different RAD index.

This ANOVA model is statistically significant at a 5% level, along with the individual terms in the model. Based on the R-square value, about 46.8% of the variance in NOX is explained by RAD.

By testing the equal variance assumption using Levene's test. We see that Levene's test has a p-value less than 0.0001, so there are significant differences between the sample variances and a Welch correction would be appropriate.

By using Tukey's test, we get all pairwise comparisons. RAD 24 has the significantly higher NOX level than other regions. RAD 5 is the second highest one.

Detailed results can be found in Appendix E.

## IV. Possible Explanations and Remained Issues

The positive relationship between non-retail business proportion and NOX can be explained by the possibility that non-retail business sometimes does cause pollution. According to the Small Business Environmental Assistance Program set up by the Environmental Protection Agency (EPA), there are a variety of businesses activities that can generate air pollution.

The proportion of structures built before 1940 (AGE) is also positively related to NOX. A possible explanation is that maybe some appliances used in these houses such as the stove or the heating systems are old-fashioned and may cause some air pollution.

The other three terms are negatively related to air pollution.

The quality of air is associated with housing values. As the median house value increases, the air pollution level is expected to decrease as the air quality tends to be a factor to take into consideration when people buy houses.

The distance to employment centers is surprisingly negative related to NOX. Maybe because most air pollution come from industrial plants which usually are not located in downtown? Or maybe because the skyscrapers serve as a shelter from the pollution? No solid explanation has been found yet due to my limit knowledge on this subject matter.

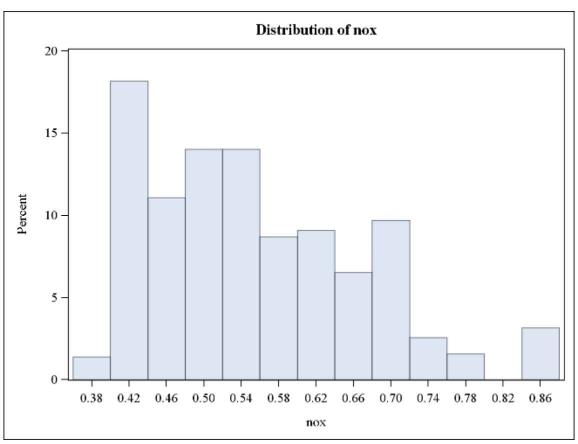
The distance from roadways appears to have an impact on nitrogen dioxide levels. But surprisingly, a negative one, in this case. Vehicles emission is always a big cause of air pollution so again it is not clear why the accessibility to highway is negatively associated with NOX levels.

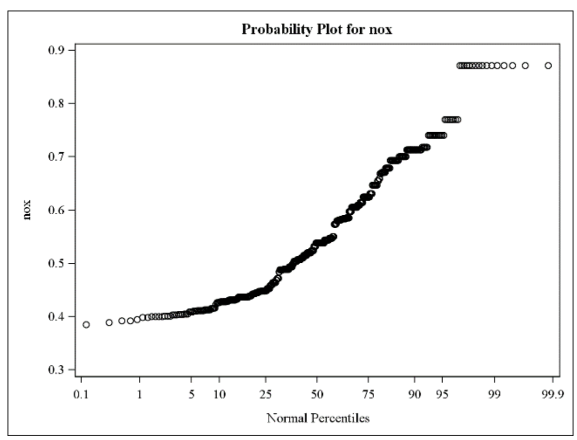
## Appendix A: Basic Descriptive Information on Continuous Variables

**Variable NOX:** The mean and median of NOX are 0.5547 and 0.5380, respectively, with standard deviation 0.1159. The range is 0.486. The skewness is positive as 0.7293 so the distribution has a long tail to the right. The histogram does not match a bell curve and the probability plot shows deviation from a straight line so the data is not normally distributed.

| Moments | | | |
|---|---|---|---|
| **N** | 506 | **Sum Weights** | 506 |
| **Mean** | 0.55469506 | **Sum Observations** | 280.6757 |
| **Std Deviation** | 0.11587768 | **Variance** | 0.01342764 |
| **Skewness** | 0.72930792 | **Kurtosis** | -0.0646671 |
| **Uncorrected SS** | 162.47038 | **Corrected SS** | 6.78095604 |
| **Coeff Variation** | 20.8903385 | **Std Error Mean** | 0.00515139 |

| Basic Statistical Measures | | | |
|---|---|---|---|
| **Location** | | **Variability** | |
| **Mean** | 0.554695 | **Std Deviation** | 0.11588 |
| **Median** | 0.538000 | **Variance** | 0.01343 |
| **Mode** | 0.538000 | **Range** | 0.48600 |
| | | **Interquartile Range** | 0.17500 |

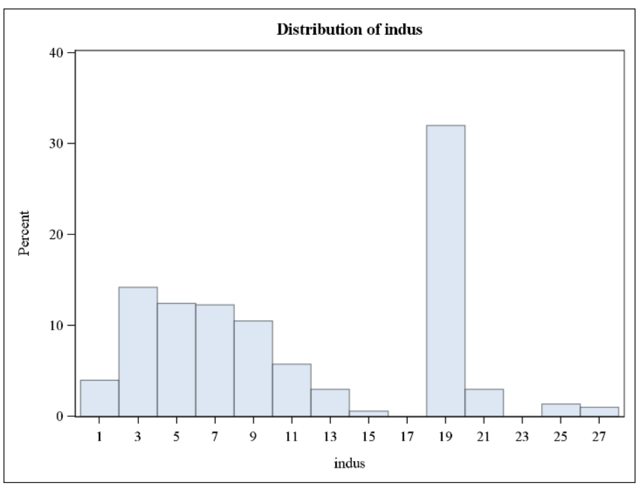


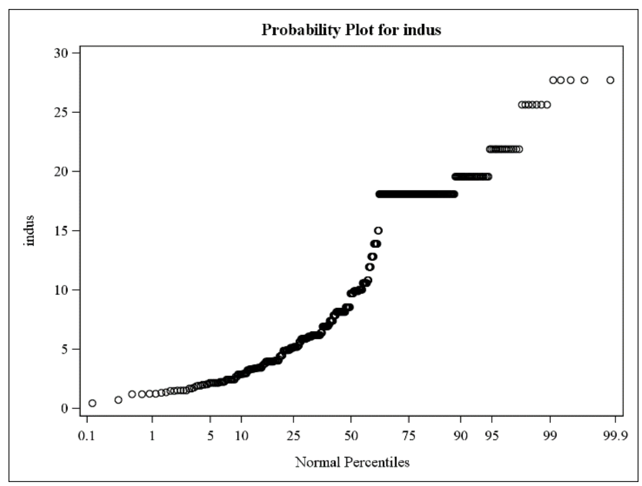




**Variable INDUS:** The mean and median of INDUS are 11.1367 and 9.6900, respectively, with standard deviation 6.8604. The range is 27.2800. The skewness is positive as 0.2950. The histogram does not match a bell curve and the probability plot shows deviation from a straight line so the data is not normally distributed.

| Moments | | | |
|---|---|---|---|
| **N** | 506 | **Sum Weights** | 506 |
| **Mean** | 11.1367787 | **Sum Observations** | 5635.21 |
| **Std Deviation** | 6.86035294 | **Variance** | 47.0644425 |
| **Skewness** | 0.29502157 | **Kurtosis** | -1.2335396 |
| **Uncorrected SS** | 86525.6299 | **Corrected SS** | 23767.5434 |
| **Coeff Variation** | 61.6008736 | **Std Error Mean** | 0.30497989 |

| Basic Statistical Measures | | | |
|---|---|---|---|
| **Location** | | **Variability** | |
| **Mean** | 11.13678 | **Std Deviation** | 6.86035 |
| **Median** | 9.69000 | **Variance** | 47.06444 |
| **Mode** | 18.10000 | **Range** | 27.28000 |
| | | **Interquartile Range** | 12.91000 |



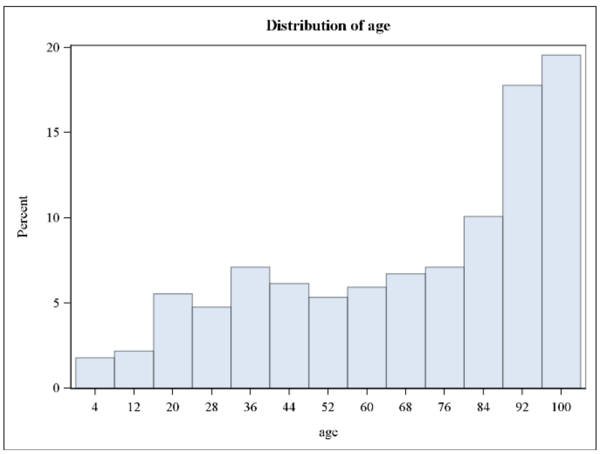



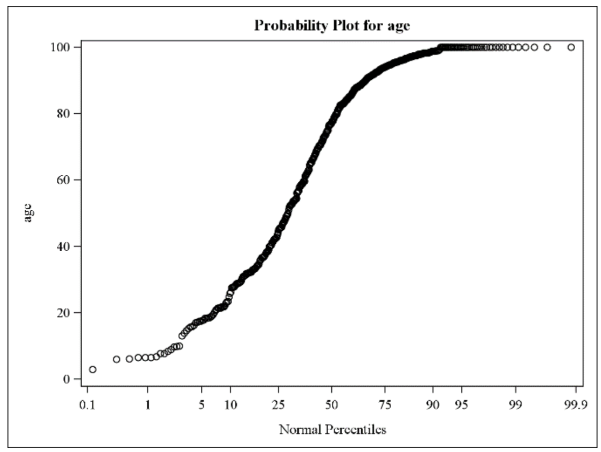

**Variable AGE:** The mean and median of AGE are 68.5749 and 77.5000, respectively, with standard deviation 28.1489. The range is 97.1000. The skewness is negative as -0.5990 so the distribution has a long tail to the left. The histogram does not match a bell curve and the probability plot shows deviation from a straight line so the data is not normally distributed.

| Moments | | | |
|---|---|---|---|
| **N** | 506 | **Sum Weights** | 506 |
| **Mean** | 68.5749012 | **Sum Observations** | 34698.9 |
| **Std Deviation** | 28.1488614 | **Variance** | 792.358399 |
| **Skewness** | -0.5989626 | **Kurtosis** | -0.9677156 |

| Moments | | | |
|---|---|---|---|
| **Uncorrected SS** | 2779614.63 | **Corrected SS** | 400140.991 |
| **Coeff Variation** | 41.0483441 | **Std Error Mean** | 1.25136953 |

| Basic Statistical Measures | | | |
|---|---|---|---|
| **Location** | | **Variability** | |
| **Mean** | 68.5749 | **Std Deviation** | 28.14886 |
| **Median** | 77.5000 | **Variance** | 792.35840 |
| **Mode** | 100.0000 | **Range** | 97.10000 |
| | | **Interquartile Range** | 49.10000 |

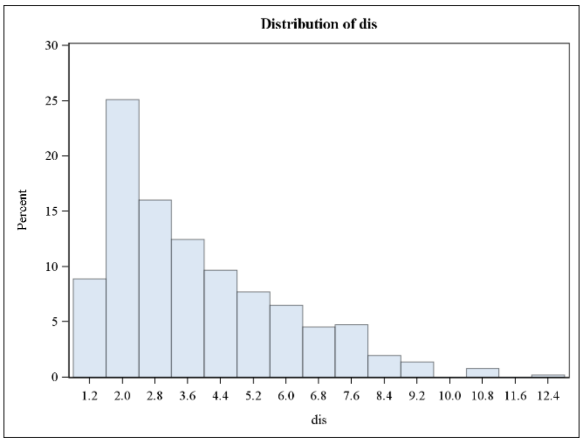


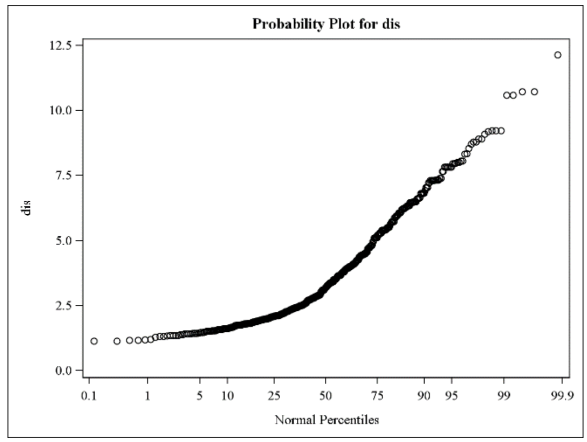




**Variable DIS:** The mean and median of DIS are 3.7950 and 3.2075, respectively, with standard deviation 2.1057. The range is 10.9969. The skewness is positive as 1.0118 so the distribution has a long tail to the right. The histogram does not match a bell curve and the probability plot shows deviation from a straight line so the data is not normally distributed.

| Moments | | | |
|---|---|---|---|
| **N** | 506 | **Sum Weights** | 506 |
| **Mean** | 3.79504269 | **Sum Observations** | 1920.2916 |
| **Std Deviation** | 2.10571013 | **Variance** | 4.43401514 |
| **Skewness** | 1.01178058 | **Kurtosis** | 0.48794112 |
| **Uncorrected SS** | 9526.76624 | **Corrected SS** | 2239.17764 |
| **Coeff Variation** | 55.4858087 | **Std Error Mean** | 0.09361023 |

| Basic Statistical Measures | | | |
|---|---|---|---|
| **Location** | | **Variability** | |
| **Mean** | 3.795043 | **Std Deviation** | 2.10571 |
| **Median** | 3.207450 | **Variance** | 4.43402 |
| **Mode** | 3.495200 | **Range** | 10.99690 |
| | | **Interquartile Range** | 3.11190 |

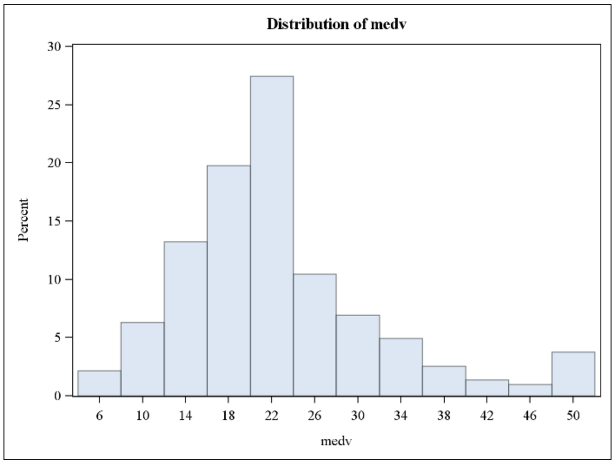


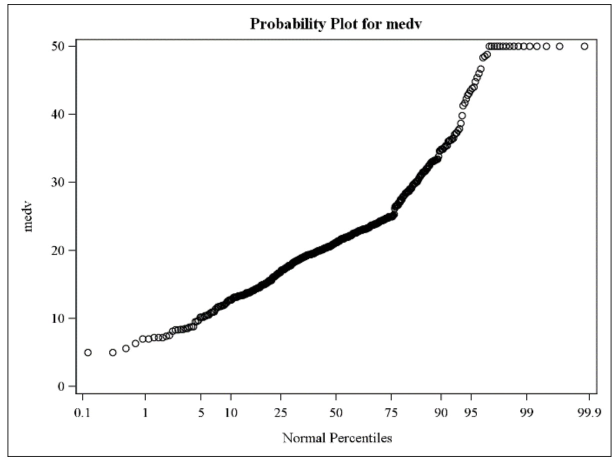




**Variable MEDV:** The mean and median of MEDV are 22.5328 and 21.2000, respectively, with standard deviation 9.1971. The range is 45.0000. The skewness is positive as 1.1081. The histogram does not match a bell curve and the probability plot shows deviation from a straight line so the data is not normally distributed.

| Moments | | | |
|---|---|---|---|
| **N** | 506 | **Sum Weights** | 506 |
| **Mean** | 22.5328063 | **Sum Observations** | 11401.6 |
| **Std Deviation** | 9.19710409 | **Variance** | 84.5867236 |
| **Skewness** | 1.10809841 | **Kurtosis** | 1.49519694 |
| **Uncorrected SS** | 299626.34 | **Corrected SS** | 42716.2954 |
| **Coeff Variation** | 40.8165053 | **Std Error Mean** | 0.40886115 |

| Basic Statistical Measures | | | |
|---|---|---|---|
| **Location** | | **Variability** | |
| **Mean** | 22.53281 | **Std Deviation** | 9.19710 |
| **Median** | 21.20000 | **Variance** | 84.58672 |
| **Mode** | 50.00000 | **Range** | 45.00000 |
| | | **Interquartile Range** | 8.00000 |




Categorical variable CHAS is a dummy variable. Among 506 observations, 471 observations for CHAS=0 which means tract does not bound river with mean of NOX as 0.55 and standard deviation 0.11. Only 35 observations for CHAS=1 with mean of NOX as 0.59 and standard deviation 0.14. Area close to the Charles River has higher level of NOX.

Categorical variable RAD has 9 levels, from 1-8 and 24. RAD 24 has more observations (n=132) than others. It also has the highest mean value of NOX among all the RAD levels (mean=0.67 with standard deviation 0.06). RAD 4 and 5 also have large number of observations (n=110, 115, respectively). The rest have observations from 17 to 38.

| | nox | | |
|---|---|---|---|
| | **Mean** | **Std** | **N** |
| **chas** | | | |
| **0** | 0.55 | 0.11 | 471 |
| **1** | 0.59 | 0.14 | 35 |
| **rad** | | | |
| **1** | 0.46 | 0.08 | 20 |
| **2** | 0.48 | 0.07 | 24 |
| **3** | 0.45 | 0.03 | 38 |
| **4** | 0.50 | 0.07 | 110 |
| **5** | 0.57 | 0.14 | 115 |
| **6** | 0.51 | 0.07 | 26 |
| **7** | 0.44 | 0.02 | 17 |
| **8** | 0.49 | 0.02 | 24 |
| **24** | 0.67 | 0.06 | 132 |

## Appendix B: General Linear Model

The GLMSELECT procedure was used to perform stepwise selection and identify significant terms, with all the selection criteria. Even though CHAS would be excluded if based on some of the criteria, I decide to temporarily keep it for the GLM model in the next step.

| Stepwise Selection Summary | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Step | Effect Entered | Effect Removed | Number Effects In | Number Parms In | Model R-Square | Adjusted R-Square | AIC | AICC | BIC | CP |
| 0 | Intercept | | 1 | 1 | 0.0000 | 0.0000 | -1672.0838 | -1672.0599 | -2180.7949 | 1668.5677 |
| 1 | dis | | 2 | 2 | 0.5917 | 0.5909 | -2123.3534 | -2123.3056 | -2631.4407 | 385.0269 |
| 2 | indus | | 3 | 3 | 0.6879 | 0.6867 | -2257.2883 | -2257.2085 | -2764.9396 | 178.0556 |
| 3 | rad | | 4 | 11 | 0.7448 | 0.7397 | -2343.1898 | -2342.5569 | -2851.1248 | 70.3758 |
| 4 | age | | 5 | 12 | 0.7666 | 0.7614 | -2386.3988 | -2385.6589 | -2892.4458 | 24.9929 |
| 5 | medv | | 6 | 13 | 0.7721 | 0.7666 | -2396.4777 | -2395.6223 | -2901.9027 | 15.0337 |
| 6 | chas | | 7 | 14 | 0.7735 | 0.7676* | -2397.5881* | -2396.6085* | -2902.7929* | 14.0000* |
| * Optimal Value Of Criterion | | | | | | | | | | |

| Stepwise Selection Summary | | | | | | | |
|---|---|---|---|---|---|---|---|
| Step | Effect Entered | Effect Removed | SBC | PRESS | ASE | F Value | Pr > F |
| 0 | Intercept | | -2175.8573 | 6.8078 | 0.0134 | 0.00 | 1.0000 |
| 1 | dis | | -2622.9003 | 2.7951 | 0.0055 | 730.43 | <.0001 |
| 2 | indus | | -2752.6087 | 2.1460 | 0.0042 | 155.02 | <.0001 |
| 3 | rad | | -2804.6979 | 1.7869 | 0.0034 | 13.80 | <.0001 |
| 4 | age | | -2843.6803 | 1.6381 | 0.0031 | 46.17 | <.0001 |
| 5 | medv | | -2849.5327* | 1.6051* | 0.0031 | 11.91 | 0.0006 |
| 6 | chas | | -2846.4165 | 1.6096 | 0.0030 | 3.03 | 0.0822 |
| * Optimal Value Of Criterion | | | | | | | |

The six variables were used to fit the model by using PROC GLM procedure. The model is significantly better than an error-only model. However, the diagnostic plots show a clear trend that residuals increase as the predicted values increase.

| Class Level Information | | |
|---|---|---|
| **Class** | **Levels** | **Values** |
| **chas** | 2 | 0 1 |
| **rad** | 9 | 1 2 3 4 5 6 7 8 24 |

| | |
|---|---|
| **Number of Observations Read** | 506 |
| **Number of Observations Used** | 506 |

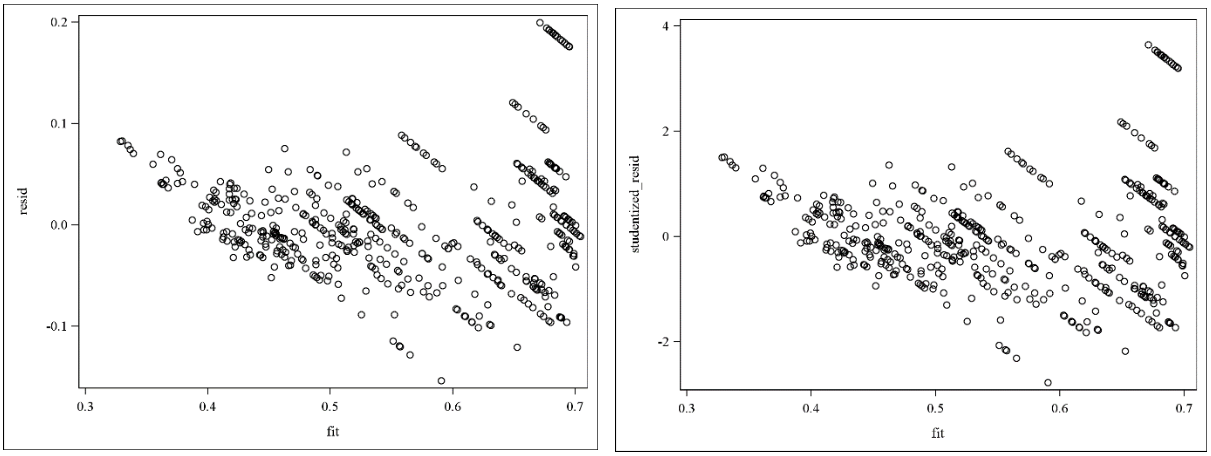
***Dependent Variable: nox***

| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| **Model** | 13 | 5.24533971 | 0.40348767 | 129.27 | <.0001 |
| **Error** | 492 | 1.53561632 | 0.00312117 | | |
| **Corrected Total** | 505 | 6.78095604 | | | |

| R-Square | Coeff Var | Root MSE | nox Mean |
|---|---|---|---|
| 0.773540 | 10.07174 | 0.055867 | 0.554695 |

| Source | DF | Type I SS | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| **indus** | 1 | 3.95440628 | 3.95440628 | 1266.96 | <.0001 |
| **rad** | 8 | 0.63897444 | 0.07987180 | 25.59 | <.0001 |
| **chas** | 1 | 0.01095290 | 0.01095290 | 3.51 | 0.0616 |
| **age** | 1 | 0.44753753 | 0.44753753 | 143.39 | <.0001 |
| **dis** | 1 | 0.14951062 | 0.14951062 | 47.90 | <.0001 |
| **medv** | 1 | 0.04395796 | 0.04395796 | 14.08 | 0.0002 |

| Source | DF | Type III SS | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| **indus** | 1 | 0.15999151 | 0.15999151 | 51.26 | <.0001 |
| **rad** | 8 | 0.34332722 | 0.04291590 | 13.75 | <.0001 |
| **chas** | 1 | 0.00946855 | 0.00946855 | 3.03 | 0.0822 |
| **age** | 1 | 0.10209335 | 0.10209335 | 32.71 | <.0001 |
| **dis** | 1 | 0.18102222 | 0.18102222 | 58.00 | <.0001 |
| **medv** | 1 | 0.04395796 | 0.04395796 | 14.08 | 0.0002 |



resid
fit
studentized_resid
fit

# Appendix C: Gamma Models

The type 1 analysis shows only one variable CHAS is not significant at a 0.05 level when the terms are added sequentially. The type 3 analysis shows that given all the other variables, only one variable CHAS is not significant at a 0.05 level. Based on the analyses of Type 1 and Type 3, the variable CHAS should be removed and the model should be refit. Examining the residuals and Cook's distances with the criterion as larger than .5, no observations should be removed.

| Model Information | |
|---|---|
| **Data Set** | WORK.HOUSE |
| **Distribution** | Gamma |
| **Link Function** | Log |
| **Dependent Variable** | nox |

| Criteria For Assessing Goodness Of Fit | | | |
|---|---|---|---|
| **Criterion** | **DF** | **Value** | **Value/DF** |
| **Deviance** | 492 | 3.6936 | 0.0075 |
| **Scaled Deviance** | 492 | 506.6149 | 1.0297 |
| **Pearson Chi-Square** | 492 | 3.7964 | 0.0077 |
| **Scaled Pearson X2** | 492 | 520.7189 | 1.0584 |
| **Log Likelihood** | | 835.1249 | |
| **Full Log Likelihood** | | 835.1249 | |
| **AIC (smaller is better)** | | -1640.2498 | |
| **AICC (smaller is better)** | | -1639.2703 | |
| **BIC (smaller is better)** | | -1576.8518 | |

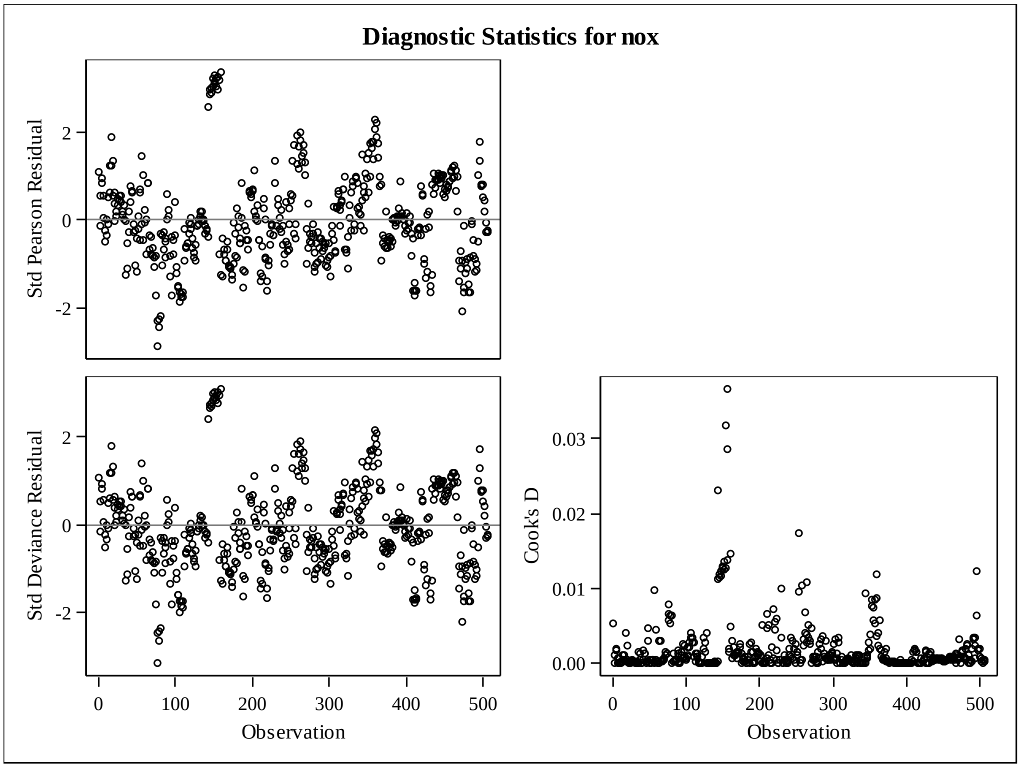
| Analysis Of Maximum Likelihood Parameter Estimates | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **Parameter** | | **DF** | **Estimate** | **Standard Error** | **Wald 95% Confidence Limits** | | **Wald Chi-Square** | **Pr > ChiSq** |
| **Intercept** | | 1 | -0.5549 | 0.0399 | -0.6331 | -0.4768 | 193.62 | <.0001 |
| **indus** | | 1 | 0.0075 | 0.0010 | 0.0056 | 0.0094 | 57.75 | <.0001 |
| **chas** | 0 | 1 | -0.0210 | 0.0156 | -0.0516 | 0.0095 | 1.82 | 0.1776 |
| **chas** | 1 | 0 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | . | . |
| **age** | | 1 | 0.0016 | 0.0002 | 0.0012 | 0.0020 | 52.10 | <.0001 |
| **dis** | | 1 | -0.0323 | 0.0032 | -0.0386 | -0.0260 | 101.10 | <.0001 |
| **rad** | 1 | 1 | -0.0700 | 0.0230 | -0.1151 | -0.0249 | 9.24 | 0.0024 |
| **rad** | 2 | 1 | -0.1420 | 0.0204 | -0.1820 | -0.1021 | 48.66 | <.0001 |

| Analysis Of Maximum Likelihood Parameter Estimates | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **Parameter** | | **DF** | **Estimate** | **Standard Error** | **Wald 95% Confidence Limits** | | **Wald Chi-Square** | **Pr > ChiSq** |
| **rad** | 3 | 1 | -0.1051 | 0.0190 | -0.1422 | -0.0679 | 30.74 | <.0001 |
| **rad** | 4 | 1 | -0.1074 | 0.0127 | -0.1323 | -0.0826 | 71.82 | <.0001 |
| **rad** | 5 | 1 | -0.0128 | 0.0127 | -0.0377 | 0.0122 | 1.01 | 0.3157 |
| **rad** | 6 | 1 | -0.0789 | 0.0198 | -0.1177 | -0.0402 | 15.92 | <.0001 |
| **rad** | 7 | 1 | -0.0764 | 0.0248 | -0.1250 | -0.0279 | 9.53 | 0.0020 |
| **rad** | 8 | 1 | -0.0817 | 0.0216 | -0.1241 | -0.0394 | 14.33 | 0.0002 |
| **rad** | 24 | 0 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | . | . |
| **medv** | | 1 | -0.0021 | 0.0005 | -0.0031 | -0.0011 | 15.87 | <.0001 |
| **Scale** | | 1 | 137.1596 | 8.6127 | 121.2764 | 155.1228 | | |

**Note:** The scale parameter was estimated by maximum likelihood.

| LR Statistics For Type 1 Analysis | | | | |
|---|---|---|---|---|
| **Source** | **2*LogLikelihood** | **DF** | **Chi-Square** | **Pr > ChiSq** |
| **Intercept** | 789.5166 | | | |
| **indus** | 1272.6639 | 1 | 483.15 | <.0001 |
| **chas** | 1274.1119 | 1 | 1.45 | 0.2288 |
| **age** | 1452.9114 | 1 | 178.80 | <.0001 |
| **dis** | 1539.8670 | 1 | 86.96 | <.0001 |
| **rad** | 1654.7475 | 8 | 114.88 | <.0001 |
| **medv** | 1670.2498 | 1 | 15.50 | <.0001 |

| LR Statistics For Type 3 Analysis | | | |
|---|---|---|---|
| **Source** | **DF** | **Chi-Square** | **Pr > ChiSq** |
| **indus** | 1 | 54.98 | <.0001 |
| **chas** | 1 | 1.82 | 0.1769 |
| **age** | 1 | 49.45 | <.0001 |
| **dis** | 1 | 91.30 | <.0001 |
| **rad** | 8 | 115.50 | <.0001 |
| **medv** | 1 | 15.50 | <.0001 |




The model was refit with all predictors except CHAS. All the parameters are significant at a 0.05 level in the estimation table except one level of RAD (RAD=5). Both the type 1 and type 3 analyses indicate that all the five terms should be retained in the model. No observations should be removed by using the Cook's distance to check the influential points. The residual plots look fine and there are no more issues.

| Model Information | |
|---|---|
| **Data Set** | WORK.HOUSE |
| **Distribution** | Gamma |
| **Link Function** | Log |
| **Dependent Variable** | nox |

| Criteria For Assessing Goodness Of Fit | | | |
|---|---|---|---|
| **Criterion** | **DF** | **Value** | **Value/DF** |
| **Deviance** | 493 | 3.7069 | 0.0075 |
| **Scaled Deviance** | 493 | 506.6171 | 1.0276 |
| **Pearson Chi-Square** | 493 | 3.8160 | 0.0077 |
| **Scaled Pearson X2** | 493 | 521.5200 | 1.0578 |
| **Log Likelihood** | | 834.2130 | |
| **Full Log Likelihood** | | 834.2130 | |

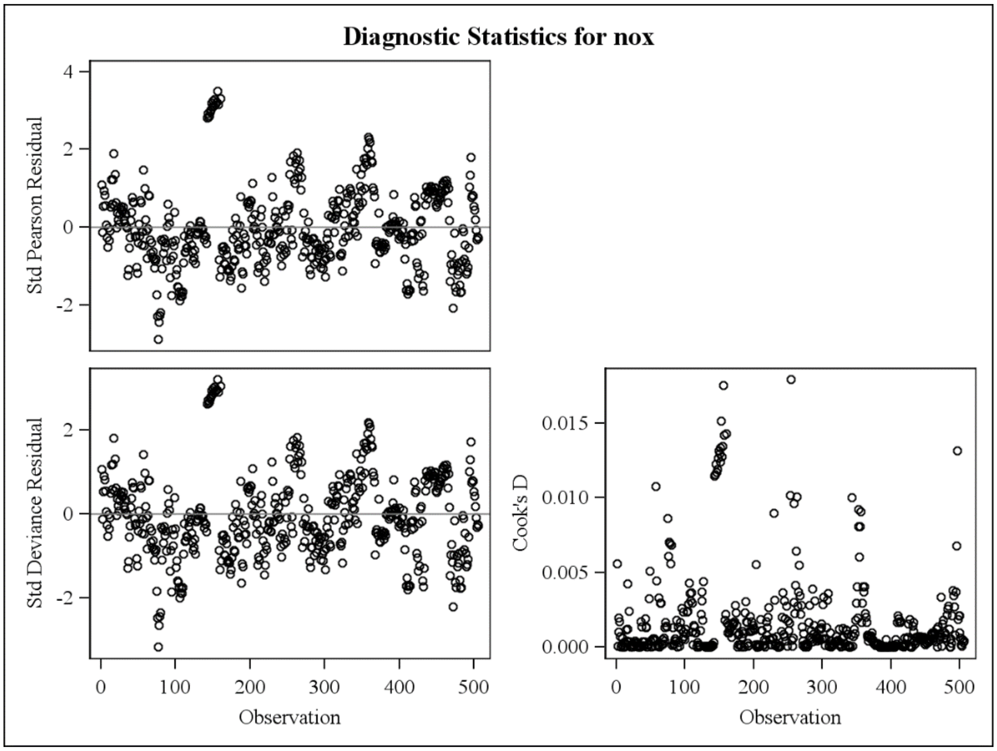
| Criteria For Assessing Goodness Of Fit | | | |
|---|---|---|---|
| **Criterion** | **DF** | **Value** | **Value/DF** |
| **AIC (smaller is better)** | | -1640.4261 | |
| **AICC (smaller is better)** | | -1639.5707 | |
| **BIC (smaller is better)** | | -1581.2546 | |

| Analysis Of Maximum Likelihood Parameter Estimates | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **Parameter** | | **DF** | **Estimate** | **Standard Error** | **Wald 95% Confidence Limits** | | **Wald Chi-Square** | **Pr > ChiSq** |
| **Intercept** | | 1 | -0.5810 | 0.0350 | -0.6495 | -0.5125 | 276.27 | <.0001 |
| **indus** | | 1 | 0.0076 | 0.0010 | 0.0057 | 0.0096 | 60.43 | <.0001 |
| **age** | | 1 | 0.0016 | 0.0002 | 0.0012 | 0.0020 | 53.08 | <.0001 |
| **dis** | | 1 | -0.0323 | 0.0032 | -0.0386 | -0.0260 | 100.74 | <.0001 |
| **rad** | 1 | 1 | -0.0689 | 0.0231 | -0.1141 | -0.0237 | 8.93 | 0.0028 |
| **rad** | 2 | 1 | -0.1433 | 0.0204 | -0.1832 | -0.1033 | 49.42 | <.0001 |
| **rad** | 3 | 1 | -0.1045 | 0.0190 | -0.1417 | -0.0673 | 30.29 | <.0001 |
| **rad** | 4 | 1 | -0.1065 | 0.0127 | -0.1314 | -0.0817 | 70.53 | <.0001 |
| **rad** | 5 | 1 | -0.0119 | 0.0127 | -0.0369 | 0.0130 | 0.87 | 0.3497 |
| **rad** | 6 | 1 | -0.0791 | 0.0198 | -0.1179 | -0.0402 | 15.92 | <.0001 |
| **rad** | 7 | 1 | -0.0767 | 0.0248 | -0.1253 | -0.0281 | 9.56 | 0.0020 |
| **rad** | 8 | 1 | -0.0788 | 0.0215 | -0.1210 | -0.0366 | 13.41 | 0.0002 |
| **rad** | 24 | 0 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | . | . |
| **medv** | | 1 | -0.0019 | 0.0005 | -0.0030 | -0.0009 | 14.30 | 0.0002 |
| **Scale** | | 1 | 136.6673 | 8.5817 | 120.8412 | 154.5660 | | |

**Note:** The scale parameter was estimated by maximum likelihood.

| LR Statistics For Type 1 Analysis | | | | |
|---|---|---|---|---|
| **Source** | **2*LogLikelihood** | **DF** | **Chi-Square** | **Pr > ChiSq** |
| **Intercept** | 789.5166 | | | |
| **indus** | 1272.6639 | 1 | 483.15 | <.0001 |
| **age** | 1452.3690 | 1 | 179.71 | <.0001 |
| **dis** | 1539.7638 | 1 | 87.39 | <.0001 |
| **rad** | 1654.4478 | 8 | 114.68 | <.0001 |
| **medv** | 1668.4261 | 1 | 13.98 | 0.0002 |

| LR Statistics For Type 3 Analysis | | | |
|---|---|---|---|
| **Source** | **DF** | **Chi-Square** | **Pr > ChiSq** |
| **indus** | 1 | 57.37 | <.0001 |
| **age** | 1 | 50.33 | <.0001 |
| **dis** | 1 | 91.01 | <.0001 |
| **rad** | 8 | 115.14 | <.0001 |
| **medv** | 1 | 13.98 | 0.0002 |



| Obs | nox | indus | age | dis | rad | medv | cd |
|---:|---:|---:|---:|---:|---:|---:|---:|
| **1** | 0.5380 | 2.31 | 65.2 | 4.0900 | 1 | 24.0 | 0.005605 |
| **2** | 0.4690 | 7.07 | 78.9 | 4.9671 | 2 | 21.6 | 0.000059 |
| **3** | 0.4690 | 7.07 | 61.1 | 4.9671 | 2 | 34.7 | 0.001011 |
| **4** | 0.4580 | 2.18 | 45.8 | 6.0622 | 3 | 33.4 | 0.001930 |
| **5** | 0.4580 | 2.18 | 54.2 | 6.0622 | 3 | 36.2 | 0.001699 |
| **6** | 0.4580 | 2.18 | 58.7 | 6.0622 | 3 | 28.7 | 0.000713 |
| **7** | 0.5240 | 7.87 | 66.6 | 5.5605 | 5 | 22.9 | 0.000002 |
| **8** | 0.5240 | 7.87 | 96.1 | 5.9505 | 5 | 27.1 | 0.000143 |
| **9** | 0.5240 | 7.87 | 100.0 | 6.0821 | 5 | 16.5 | 0.000623 |
| **10** | 0.5240 | 7.87 | 85.9 | 6.5921 | 5 | 18.9 | 0.000000 |

## Appendix D: Inverse Gaussian Models

Inverse Gaussian distribution and log link function was used to build the generalized linear model. The type 1 analysis shows five variables INDUS, AGE, DIS, MEDV and RAD are significant at a level of 0.05. The type 3 analysis shows that given all the other variables, INDUS, AGE, DIS, MEDV and RAD are significant at a level of 0.05. The Cook's distance was used to check the influential points. All of the points have Cook's distance less than 0.5. Therefore the model needs to be refit by removing the variable CHAS.

| Model Information | |
|---|---|
| **Data Set** | WORK.HOUSE |
| **Distribution** | Inverse Gaussian |
| **Link Function** | Log |
| **Dependent Variable** | nox |

| Criteria For Assessing Goodness Of Fit | | | |
|---|---|---|---|
| **Criterion** | **DF** | **Value** | **Value/DF** |
| **Deviance** | 492 | 6.2823 | 0.0128 |
| **Scaled Deviance** | 492 | 506.0000 | 1.0285 |
| **Pearson Chi-Square** | 492 | 6.5409 | 0.0133 |
| **Scaled Pearson X2** | 492 | 526.8252 | 1.0708 |
| **Log Likelihood** | | 855.3912 | |
| **Full Log Likelihood** | | 855.3912 | |
| **AIC (smaller is better)** | | -1680.7824 | |
| **AICC (smaller is better)** | | -1679.8028 | |
| **BIC (smaller is better)** | | -1617.3843 | |

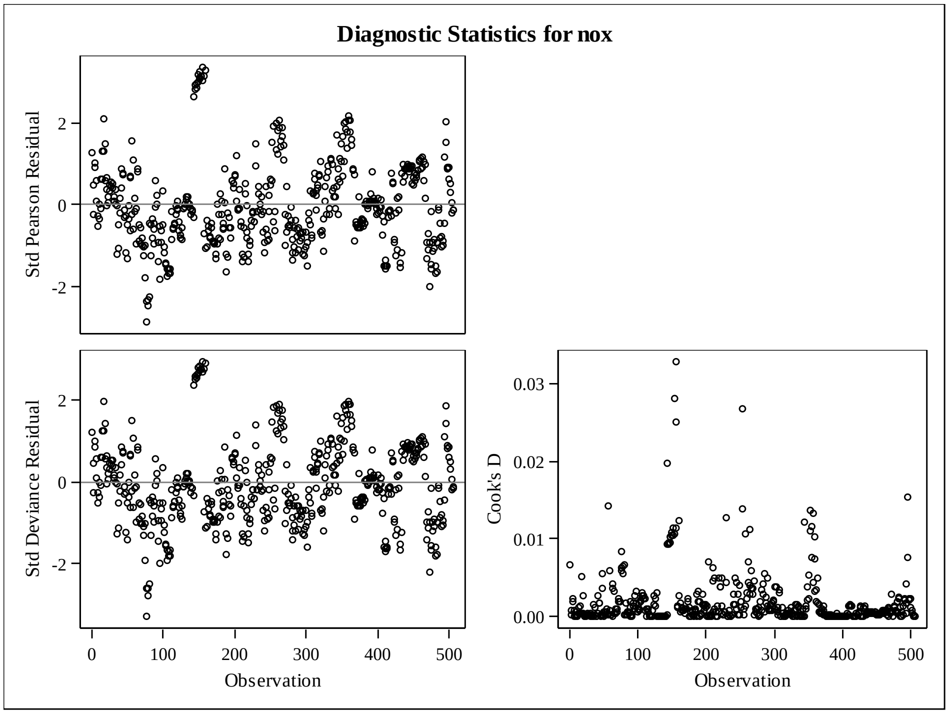
| Analysis Of Maximum Likelihood Parameter Estimates | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **Parameter** | | **DF** | **Estimate** | **Standard Error** | **Wald 95% Confidence Limits** | | **Wald Chi-Square** | **Pr > ChiSq** |
| **Intercept** | | 1 | -0.5734 | 0.0384 | -0.6486 | -0.4983 | 223.49 | <.0001 |
| **indus** | | 1 | 0.0074 | 0.0010 | 0.0055 | 0.0093 | 57.94 | <.0001 |
| **chas** | 0 | 1 | -0.0123 | 0.0153 | -0.0423 | 0.0177 | 0.65 | 0.4211 |
| **chas** | 1 | 0 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | . | . |
| **age** | | 1 | 0.0016 | 0.0002 | 0.0012 | 0.0020 | 64.61 | <.0001 |
| **dis** | | 1 | -0.0302 | 0.0029 | -0.0358 | -0.0246 | 112.38 | <.0001 |
| **rad** | 1 | 1 | -0.0808 | 0.0213 | -0.1225 | -0.0391 | 14.43 | 0.0001 |
| **rad** | 2 | 1 | -0.1431 | 0.0194 | -0.1812 | -0.1050 | 54.20 | <.0001 |

| Analysis Of Maximum Likelihood Parameter Estimates | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **Parameter** | | **DF** | **Estimate** | **Standard Error** | **Wald 95% Confidence Limits** | | **Wald Chi-Square** | **Pr > ChiSq** |
| **rad** | 3 | 1 | -0.1117 | 0.0178 | -0.1466 | -0.0768 | 39.35 | <.0001 |
| **rad** | 4 | 1 | -0.1108 | 0.0128 | -0.1358 | -0.0858 | 75.39 | <.0001 |
| **rad** | 5 | 1 | -0.0242 | 0.0130 | -0.0495 | 0.0012 | 3.48 | 0.0621 |
| **rad** | 6 | 1 | -0.0866 | 0.0190 | -0.1238 | -0.0494 | 20.82 | <.0001 |
| **rad** | 7 | 1 | -0.0844 | 0.0224 | -0.1283 | -0.0404 | 14.17 | 0.0002 |
| **rad** | 8 | 1 | -0.0868 | 0.0204 | -0.1268 | -0.0468 | 18.12 | <.0001 |
| **rad** | 24 | 0 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | . | . |
| **medv** | | 1 | -0.0018 | 0.0005 | -0.0028 | -0.0008 | 13.01 | 0.0003 |
| **Scale** | | 1 | 0.1114 | 0.0035 | 0.1048 | 0.1185 | | |

**Note:** The scale parameter was estimated by maximum likelihood.

| LR Statistics For Type 1 Analysis | | | | |
|---|---|---|---|---|
| **Source** | **2*LogLikelihood** | **DF** | **Chi-Square** | **Pr > ChiSq** |
| **Intercept** | 805.1130 | | | |
| **indus** | 1300.3746 | 1 | 495.26 | <.0001 |
| **chas** | 1300.8944 | 1 | 0.52 | 0.4709 |
| **age** | 1495.3083 | 1 | 194.41 | <.0001 |
| **dis** | 1584.8857 | 1 | 89.58 | <.0001 |
| **rad** | 1698.1876 | 8 | 113.30 | <.0001 |
| **medv** | 1710.7824 | 1 | 12.59 | 0.0004 |

| LR Statistics For Type 3 Analysis | | | |
|---|---|---|---|
| **Source** | **DF** | **Chi-Square** | **Pr > ChiSq** |
| **indus** | 1 | 55.96 | <.0001 |
| **chas** | 1 | 0.65 | 0.4192 |
| **age** | 1 | 60.30 | <.0001 |
| **dis** | 1 | 98.87 | <.0001 |
| **rad** | 8 | 114.91 | <.0001 |
| **medv** | 1 | 12.59 | 0.0004 |




The model was refit with these five variables identified above. All the parameters are significant at a 0.05 level in the estimation table except RAD=5. The type 1 and type 3 analyses both indicate that all these five terms are significant at a level of 0.05. All the observations have Cook's distance less than 0.5 and the diagnostic plots look good.

| Model Information | | |
|---|---|---|
| **Data Set** | WORK.GAMMA | Predicted Values and Diagnostic Statistics |
| **Distribution** | Inverse Gaussian | |
| **Link Function** | Log | |
| **Dependent Variable** | nox | |

| Criteria For Assessing Goodness Of Fit | | | |
|---|---|---|---|
| **Criterion** | **DF** | **Value** | **Value/DF** |
| **Deviance** | 493 | 6.2905 | 0.0128 |
| **Scaled Deviance** | 493 | 506.0000 | 1.0264 |
| **Pearson Chi-Square** | 493 | 6.5577 | 0.0133 |
| **Scaled Pearson X2** | 493 | 527.4967 | 1.0700 |
| **Log Likelihood** | | 855.0649 | |
| **Full Log Likelihood** | | 855.0649 | |
| **AIC (smaller is better)** | | -1682.1298 | |

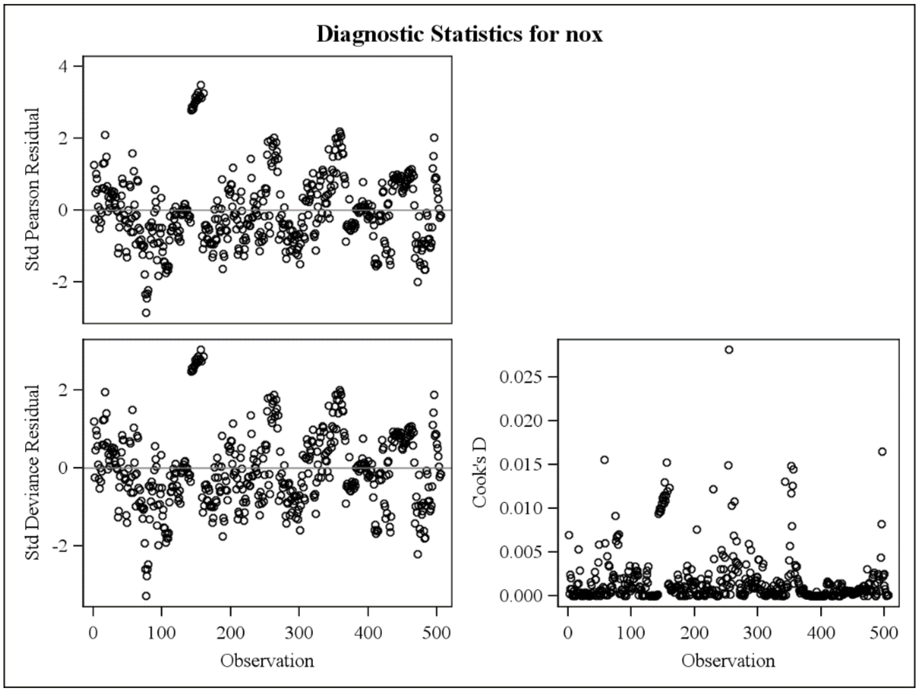
| Criteria For Assessing Goodness Of Fit | | | |
|---|---|---|---|
| **Criterion** | **DF** | **Value** | **Value/DF** |
| **AICC (smaller is better)** | | -1681.2744 | |
| **BIC (smaller is better)** | | -1622.9583 | |

| Analysis Of Maximum Likelihood Parameter Estimates | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **Parameter** | | **DF** | **Estimate** | **Standard Error** | **Wald 95% Confidence Limits** | | **Wald Chi-Square** | **Pr > ChiSq** |
| **Intercept** | | 1 | -0.5885 | 0.0335 | -0.6542 | -0.5228 | 308.17 | <.0001 |
| **indus** | | 1 | 0.0075 | 0.0010 | 0.0056 | 0.0094 | 59.94 | <.0001 |
| **age** | | 1 | 0.0016 | 0.0002 | 0.0012 | 0.0020 | 65.30 | <.0001 |
| **dis** | | 1 | -0.0302 | 0.0029 | -0.0358 | -0.0246 | 112.23 | <.0001 |
| **rad** | 1 | 1 | -0.0801 | 0.0213 | -0.1217 | -0.0384 | 14.16 | 0.0002 |
| **rad** | 2 | 1 | -0.1437 | 0.0194 | -0.1818 | -0.1056 | 54.61 | <.0001 |
| **rad** | 3 | 1 | -0.1113 | 0.0178 | -0.1462 | -0.0764 | 39.07 | <.0001 |
| **rad** | 4 | 1 | -0.1102 | 0.0128 | -0.1352 | -0.0852 | 74.69 | <.0001 |
| **rad** | 5 | 1 | -0.0237 | 0.0129 | -0.0490 | 0.0017 | 3.35 | 0.0674 |
| **rad** | 6 | 1 | -0.0867 | 0.0190 | -0.1239 | -0.0495 | 20.82 | <.0001 |
| **rad** | 7 | 1 | -0.0845 | 0.0224 | -0.1284 | -0.0405 | 14.19 | 0.0002 |
| **rad** | 8 | 1 | -0.0852 | 0.0203 | -0.1250 | -0.0454 | 17.61 | <.0001 |
| **rad** | 24 | 0 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | . | . |
| **medv** | | 1 | -0.0018 | 0.0005 | -0.0027 | -0.0008 | 12.37 | 0.0004 |
| **Scale** | | 1 | 0.1115 | 0.0035 | 0.1048 | 0.1186 | | |

**Note:** The scale parameter was estimated by maximum likelihood.

| LR Statistics For Type 1 Analysis | | | | |
|---|---|---|---|---|
| **Source** | **2*LogLikelihood** | **DF** | **Chi-Square** | **Pr > ChiSq** |
| **Intercept** | 805.1130 | | | |
| **indus** | 1300.3746 | 1 | 495.26 | <.0001 |
| **age** | 1495.2108 | 1 | 194.84 | <.0001 |
| **dis** | 1584.8693 | 1 | 89.66 | <.0001 |
| **rad** | 1698.1661 | 8 | 113.30 | <.0001 |
| **medv** | 1710.1298 | 1 | 11.96 | 0.0005 |

| LR Statistics For Type 3 Analysis | | | |
|---|---|---|---|
| **Source** | **DF** | **Chi-Square** | **Pr > ChiSq** |
| **indus** | 1 | 57.73 | <.0001 |
| **age** | 1 | 60.89 | <.0001 |
| **dis** | 1 | 98.75 | <.0001 |
| **rad** | 8 | 114.62 | <.0001 |
| **medv** | 1 | 11.96 | 0.0005 |




| Obs | nox | indus | age | dis | rad | medv | cd |
|---|---|---|---|---|---|---|---|
| **1** | 0.5380 | 2.31 | 65.2 | 4.0900 | 1 | 24.0 | 0.006982 |
| **2** | 0.4690 | 7.07 | 78.9 | 4.9671 | 2 | 21.6 | 0.000216 |
| **3** | 0.4690 | 7.07 | 61.1 | 4.9671 | 2 | 34.7 | 0.000820 |
| **4** | 0.4580 | 2.18 | 45.8 | 6.0622 | 3 | 33.4 | 0.002421 |
| **5** | 0.4580 | 2.18 | 54.2 | 6.0622 | 3 | 36.2 | 0.002034 |
| **6** | 0.4580 | 2.18 | 58.7 | 6.0622 | 3 | 28.7 | 0.000777 |
| **7** | 0.5240 | 7.87 | 66.6 | 5.5605 | 5 | 22.9 | 0.000007 |
| **8** | 0.5240 | 7.87 | 96.1 | 5.9505 | 5 | 27.1 | 0.000147 |
| **9** | 0.5240 | 7.87 | 100.0 | 6.0821 | 5 | 16.5 | 0.000562 |
| **10** | 0.5240 | 7.87 | 85.9 | 6.5921 | 5 | 18.9 | 0.000000 |

## Appendix E: One-way ANOVA

This ANOVA model is statistically significant at a 5% level, along with the individual terms in the model. Based on the R-square value, about 46.8% of the variance in NOX is explained by RAD.
By testing the equal variance assumption using Levene's test. We see that Levene's test has a p-value less than 0.0001, so there are significant differences between the sample variances and a Welch correction would be appropriate.
By using Tukey's test, we get all pairwise comparisons. RAD 24 has the significantly higher NOX level than other regions. RAD 5 is the second highest one.

| Class Level Information | | |
|---|---|---|
| **Class** | **Levels** | **Values** |
| **rad** | 9 | 1 2 3 4 5 6 7 8 24 |

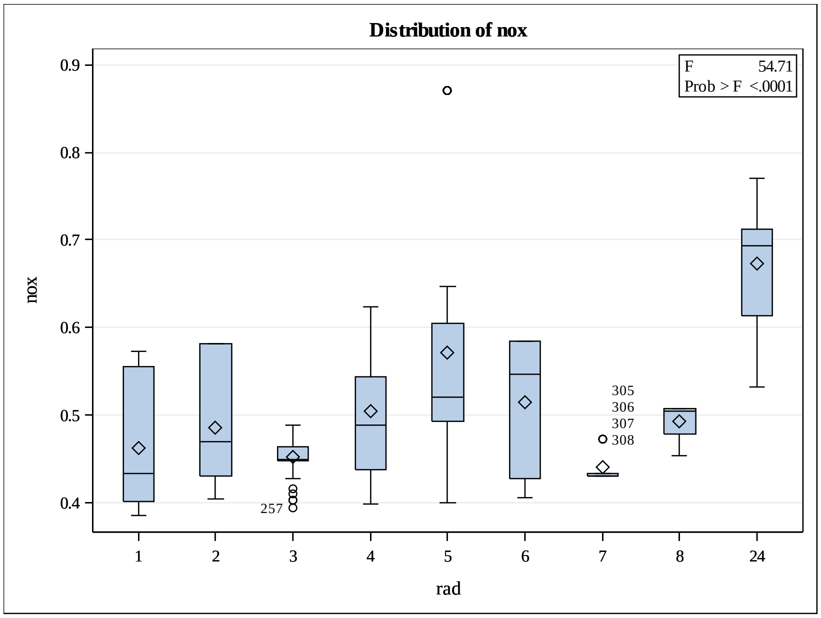
| | |
|---|---|
| **Number of Observations Read** | 506 |
| **Number of Observations Used** | 506 |

***Dependent Variable: nox***

| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| **Model** | 8 | 3.17543890 | 0.39692986 | 54.71 | <.0001 |
| **Error** | 497 | 3.60551713 | 0.00725456 | | |
| **Corrected Total** | 505 | 6.78095604 | | | |

| R-Square | Coeff Var | Root MSE | nox Mean |
|---|---|---|---|
| 0.468288 | 15.35505 | 0.085174 | 0.554695 |

| Source | DF | Anova SS | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| **rad** | 8 | 3.17543890 | 0.39692986 | 54.71 | <.0001 |




| Levene's Test for Homogeneity of nox Variance<br>ANOVA of Squared Deviations from Group Means | | | | | |
|---|---|---|---|---|---|
| **Source** | **DF** | **Sum of Squares** | **Mean Square** | **F Value** | **Pr > F** |
| **rad** | 8 | 0.0210 | 0.00262 | 12.19 | <.0001 |
| **Error** | 497 | 0.1067 | 0.000215 | | |

| Welch's ANOVA for nox | | | |
|---|---|---|---|
| **Source** | **DF** | **F Value** | **Pr > F** |
| **rad** | 8.0000 | 170.04 | <.0001 |
| **Error** | 114.5 | | |

## *Tukey's Studentized Range (HSD) Test for nox*

**Note:** This test controls the Type I experimentwise error rate.

| | |
|---|---|
| **Alpha** | 0.05 |
| **Error Degrees of Freedom** | 497 |
| **Error Mean Square** | 0.007255 |
| **Critical Value of Studentized Range** | 4.40601 |

| **Comparisons significant at the 0.05 level are indicated by ***.** | | | | |
|---|---|---|---|---|
| **rad Comparison** | **Difference Between Means** | **Simultaneous 95% Confidence Limits** | | |
| **24 - 5** | 0.10153 | 0.06768 | 0.13538 | *** |
| **24 - 6** | 0.15757 | 0.10063 | 0.21451 | *** |
| **24 - 4** | 0.16811 | 0.13385 | 0.20236 | *** |
| **24 - 8** | 0.17992 | 0.12103 | 0.23880 | *** |
| **24 - 2** | 0.18750 | 0.12861 | 0.24639 | *** |
| **24 - 1** | 0.20953 | 0.14585 | 0.27320 | *** |
| **24 - 3** | 0.21999 | 0.17114 | 0.26884 | *** |
| **24 - 7** | 0.23142 | 0.16304 | 0.29979 | *** |
| **5 - 24** | -0.10153 | -0.13538 | -0.06768 | *** |
| **5 - 6** | 0.05604 | -0.00159 | 0.11366 | |
| **5 - 4** | 0.06657 | 0.03118 | 0.10196 | *** |
| **5 - 8** | 0.07838 | 0.01883 | 0.13793 | *** |
| **5 - 2** | 0.08597 | 0.02642 | 0.14552 | *** |
| **5 - 1** | 0.10799 | 0.04370 | 0.17228 | *** |
| **5 - 3** | 0.11846 | 0.06881 | 0.16811 | *** |
| **5 - 7** | 0.12988 | 0.06093 | 0.19884 | *** |
| **6 - 24** | -0.15757 | -0.21451 | -0.10063 | *** |
| **6 - 5** | -0.05604 | -0.11366 | 0.00159 | |
| **6 - 4** | 0.01054 | -0.04733 | 0.06840 | |
| **6 - 8** | 0.02235 | -0.05277 | 0.09746 | |
| **6 - 2** | 0.02993 | -0.04519 | 0.10504 | |
| **6 - 1** | 0.05196 | -0.02697 | 0.13088 | |

| Comparisons significant at the 0.05 level are indicated by ***. | | | | |
|---|---|---|---|---|
| rad Comparison | Difference Between Means | Simultaneous 95% Confidence Limits | | |
| **6 - 3** | 0.06242 | -0.00512 | 0.12996 | |
| **6 - 7** | 0.07385 | -0.00892 | 0.15661 | |
| **4 - 24** | -0.16811 | -0.20236 | -0.13385 | *** |
| **4 - 5** | -0.06657 | -0.10196 | -0.03118 | *** |
| **4 - 6** | -0.01054 | -0.06840 | 0.04733 | |
| **4 - 8** | 0.01181 | -0.04797 | 0.07160 | |
| **4 - 2** | 0.01939 | -0.04039 | 0.07918 | |
| **4 - 1** | 0.04142 | -0.02308 | 0.10593 | |
| **4 - 3** | 0.05189 | 0.00196 | 0.10182 | *** |
| **4 - 7** | 0.06331 | -0.00584 | 0.13246 | |
| **8 - 24** | -0.17992 | -0.23880 | -0.12103 | *** |
| **8 - 5** | -0.07838 | -0.13793 | -0.01883 | *** |
| **8 - 6** | -0.02235 | -0.09746 | 0.05277 | |
| **8 - 4** | -0.01181 | -0.07160 | 0.04797 | |
| **8 - 2** | 0.00758 | -0.06902 | 0.08419 | |
| **8 - 1** | 0.02961 | -0.05073 | 0.10995 | |
| **8 - 3** | 0.04008 | -0.02911 | 0.10926 | |
| **8 - 7** | 0.05150 | -0.03262 | 0.13562 | |
| **2 - 24** | -0.18750 | -0.24639 | -0.12861 | *** |
| **2 - 5** | -0.08597 | -0.14552 | -0.02642 | *** |
| **2 - 6** | -0.02993 | -0.10504 | 0.04519 | |
| **2 - 4** | -0.01939 | -0.07918 | 0.04039 | |
| **2 - 8** | -0.00758 | -0.08419 | 0.06902 | |
| **2 - 1** | 0.02203 | -0.05832 | 0.10237 | |
| **2 - 3** | 0.03249 | -0.03670 | 0.10168 | |
| **2 - 7** | 0.04392 | -0.04020 | 0.12804 | |
| **1 - 24** | -0.20953 | -0.27320 | -0.14585 | *** |
| **1 - 5** | -0.10799 | -0.17228 | -0.04370 | *** |
| **1 - 6** | -0.05196 | -0.13088 | 0.02697 | |
| **1 - 4** | -0.04142 | -0.10593 | 0.02308 | |

| Comparisons significant at the 0.05 level are indicated by ***. | | | | |
|---|---|---|---|---|
| **rad Comparison** | **Difference Between Means** | **Simultaneous 95% Confidence Limits** | | |
| **1 - 8** | -0.02961 | -0.10995 | 0.05073 | |
| **1 - 2** | -0.02203 | -0.10237 | 0.05832 | |
| **1 - 3** | 0.01047 | -0.06284 | 0.08377 | |
| **1 - 7** | 0.02189 | -0.06565 | 0.10943 | |
| **3 - 24** | -0.21999 | -0.26884 | -0.17114 | *** |
| **3 - 5** | -0.11846 | -0.16811 | -0.06881 | *** |
| **3 - 6** | -0.06242 | -0.12996 | 0.00512 | |
| **3 - 4** | -0.05189 | -0.10182 | -0.00196 | *** |
| **3 - 8** | -0.04008 | -0.10926 | 0.02911 | |
| **3 - 2** | -0.03249 | -0.10168 | 0.03670 | |
| **3 - 1** | -0.01047 | -0.08377 | 0.06284 | |
| **3 - 7** | 0.01142 | -0.06600 | 0.08885 | |
| **7 - 24** | -0.23142 | -0.29979 | -0.16304 | *** |
| **7 - 5** | -0.12988 | -0.19884 | -0.06093 | *** |
| **7 - 6** | -0.07385 | -0.15661 | 0.00892 | |
| **7 - 4** | -0.06331 | -0.13246 | 0.00584 | |
| **7 - 8** | -0.05150 | -0.13562 | 0.03262 | |
| **7 - 2** | -0.04392 | -0.12804 | 0.04020 | |
| **7 - 1** | -0.02189 | -0.10943 | 0.06565 | |
| **7 - 3** | -0.01142 | -0.08885 | 0.06600 | |