In-Course Assessment

STAT0028: Statistical Models and Data Analysis

Group 69

Due: November 18, 2025

1 Problem 1

1.1 Data Analysis

We analyse the one-to-one relationships between the features of the dataset. The dataset is constituted of three features, wind (m.s^{-1}), humidity (g.kg^{1}) and traffic emission (sum of NOx emissions) and one target variable to be predicted, the air pollution (NOx ppb). Humidity features negative values which are nonsensical in the physical sense. The target variable, pollution, is well represented in its lower range, and observations becomes sparser for higher levels of pollution.

First, we try to identify correlations between our input variables wind, traffic emission, and pollution. Looking at the matrix plots in Figure 1, we cannot identify any correlations between wind, traffic emission and humidity. Analyzing the relationship of the input variables with pollution, we first notice that pollution seems linearly correlated with traffic emission. There does not seem to be any significant relationship between pollution and humidity. The relationship between pollution and wind seems inversely proportional. We can also see heteroscedasticity in the traffic emission, as its variance seems proportionally related to the pollution. At that stage, we cannot identify any outliers.

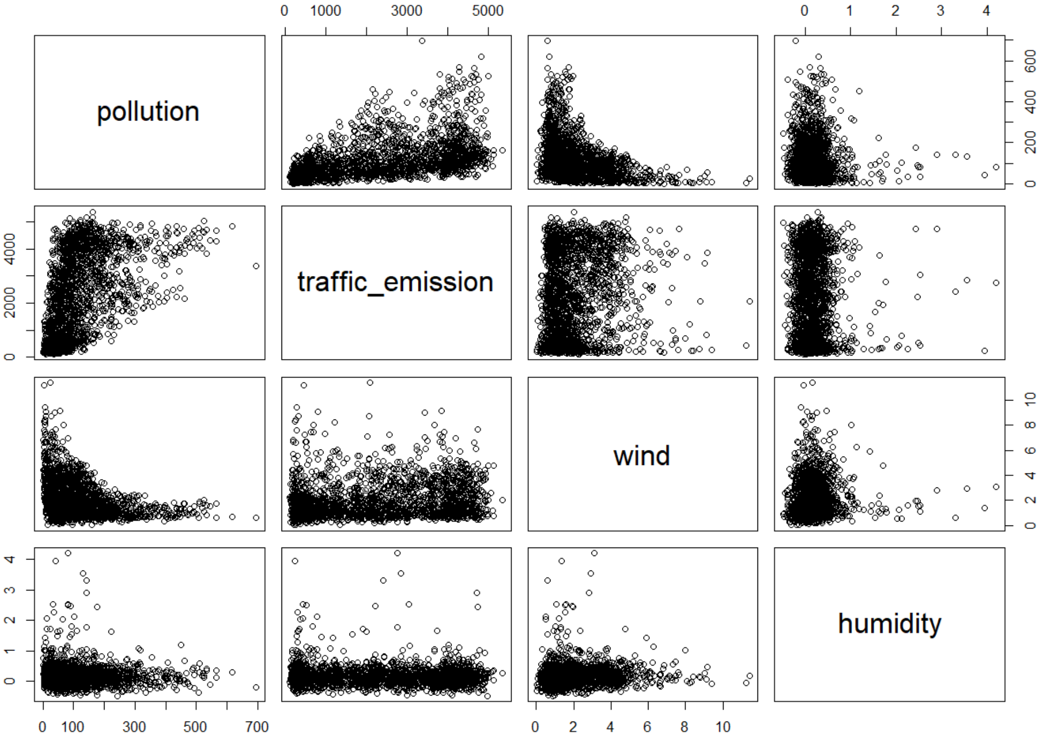

Figure 1: Data Pairplot of all variables in the dataset.

Listing 1: Code for matrix plots

```
> airquality <- read.table("airqualityCH.dat", header = TRUE)
> pairs(airquality)
```

1.2 Model 1: Simple OLS

We first fit a simple ordinary least squares (OLS) model, using all variables wind w, humidity h, and traffic emission TE, adding bias term β_0. We predict pollution P as per Equation 1:

$$P_i = \beta_0 + \beta_1 TE_i + \beta_2 w_i + \beta_3 h_i \tag{1}$$

The model obtained an R^2 value of 0.4373. Looking at the residual plots in Figure 3, we confirm a non-linear relationship with respect to wind. Humidity does not seem to be statistically relevant with a high p-value of 0.497. Both the residuals plot of traffic emission and the QQ-plot confirm the presence of heteroscedasticity in the data. We also identify 3 outliers looking at Cook's distance, though they do not seem relevant when looking at the fitted values.

The results confirm the hypotheses made during the exploratory analysis: the need for non-linear transformations on the features used to predict P_i, as well as a mean to mitigate heteroscedasticity. A low p-value for wind and traffic emission with a high F-statistics of 517.1 still suggest that a linear model would be appropriate to fit our data. Note that, the presence of heteroscedasticity means that the OLS model, originally assuming normal distribution of errors with constant variance, will lead to inaccurate inference, hence becoming inadequate in this situation.

Listing 2: Code and results for Model 1

```
> # Model 1: OLS, no transformations, use all variables
>
> lm1 <- lm(pollution~traffic_emission+wind+humidity,data=airquality)
> summary(lm1)

Call:
lm(formula = pollution ~ traffic_emission + wind + humidity,
    data = airquality)

Residuals:
    Min      1Q  Median      3Q     Max
-164.49  -41.73  -15.18   20.50  496.77

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       91.770542   3.680479   24.93   <2e-16 ***
traffic_emission   0.036352   0.001081   33.63   <2e-16 ***
wind             -27.678981   1.142490  -24.23   <2e-16 ***
humidity          -3.169578   4.661079   -0.68    0.497
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 73.86 on 1996 degrees of freedom
Multiple R-squared:  0.4373,    Adjusted R-squared:  0.4365
F-statistic: 517.1 on 3 and 1996 DF,  p-value: < 2.2e-16

> par(mfrow=c(2,2))
> plot(lm1,which=1:4,ask=FALSE)
> plot(airquality$traffic_emission,residuals(lm1))
> plot(airquality$wind,residuals(lm1))
> plot(airquality$humidity,residuals(lm1))
```

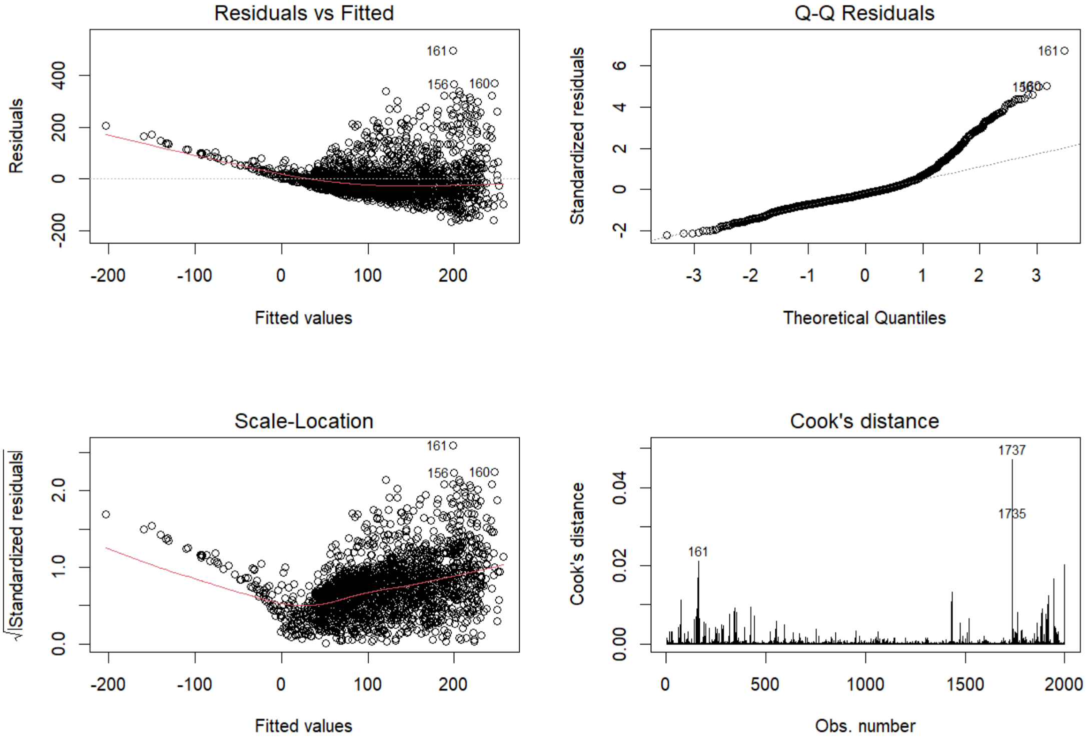

Figure 2: Result Plots of the Fitted Model 1

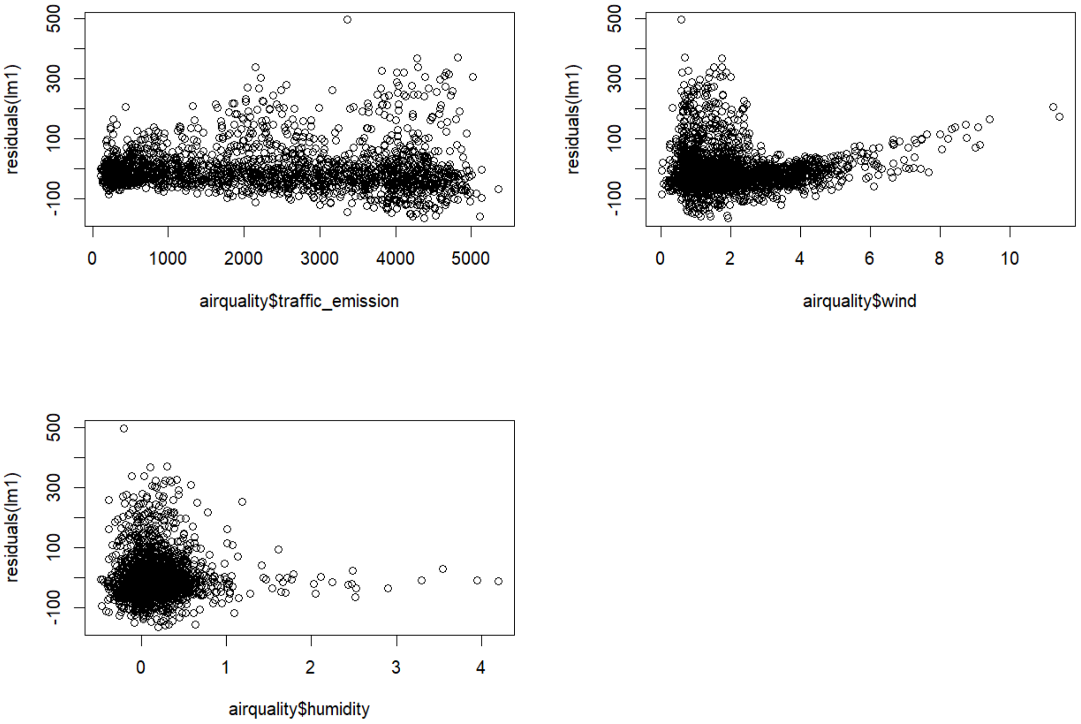

Figure 3: Residual Plots of the Fitted Model 1

1.3 Model 2: WLS with a transformation on the wind variable

To find an appropriate feature map for the wind variable, we look at the pair plot in Figure 1. For low wind speeds, we observe that pollution can take any value, with no relationship between the variables. As the wind speed increases, the maximum value that pollution takes decreases inversely proportional. Intuitively, we can model the wind as a dissipator of traffic emissions, as pollution is strongly linearly correlated with traffic emissions. To fix heteroscedasticity in traffic emissions, we suggest the method of Weighted Least Squares. We propose the following model:

$$P_i = \beta_0 + \beta_1 TE_i + \beta_2 \frac{TE_i}{1 + w_i} + \beta_3 h_i \tag{2}$$

For β_2, the idea is that low values of wind would draw the term closer to proportionality with traffic emission, while high values of wind would mute the effect of traffic emission. We can then see the first term β_1 as a baseline for traffic emission affecting pollution independently of wind.

The WLS model obtained an R^2 score of 0.4877. While this is an improvement over Model 1, we are still obtaining poor performance. The residuals of the transformed wind variable shown in Figure 5 have improved, but we still notice a non-linear relationship. The QQ-plot of residuals and the residuals plot of traffic emission and the transformed wind variable still show signs of heteroscedasticity. The F-statistics has also increased, and our transformed variable has a very low p-value, showing adequate fit.

Listing 3: Code and results for Model 2

```
> # Model 2: transformation on wind with WLS
>
> airquality$wind_transform <- airquality$traffic_emission/(1+airquality$wind)
> pairs( airquality[,c("pollution","traffic_emission","wind_transform","humidity")])
>
> lm2 <- lm(pollution~traffic_emission+wind_transform+humidity,data=airquality)
> res_lm <- lm(abs(residuals(lm2))~traffic_emission+wind_transform,data=airquality)
> weights <- 1/fitted(res_lm)^2
> wls_lm2 <- lm(pollution~traffic_emission+wind_transform+humidity,data=airquality,weights=weights)
>
> summary(wls_lm2)

Call:
lm(formula = pollution ~ traffic_emission + wind_transform +
    humidity, data = airquality, weights = weights)

Weighted Residuals:
    Min      1Q  Median      3Q     Max
-2.3989 -0.8843 -0.3038  0.5220  7.8441

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      29.486634   1.728130  17.063   <2e-16 ***
traffic_emission -0.011715   0.001250  -9.369   <2e-16 ***
wind_transform    0.136855   0.004192  32.649   <2e-16 ***
humidity         -3.629014   2.967916  -1.223    0.222
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 1.328 on 1996 degrees of freedom
Multiple R-squared:  0.4877,    Adjusted R-squared:  0.4869
F-statistic: 633.4 on 3 and 1996 DF,  p-value: < 2.2e-16

> plot(wls_lm2,which=1:4,ask=FALSE)
> plot(airquality$traffic_emission,residuals(wls_lm2))
> plot(airquality$humidity,residuals(wls_lm2))
> plot(airquality$wind_transform,residuals(wls_lm2))
```

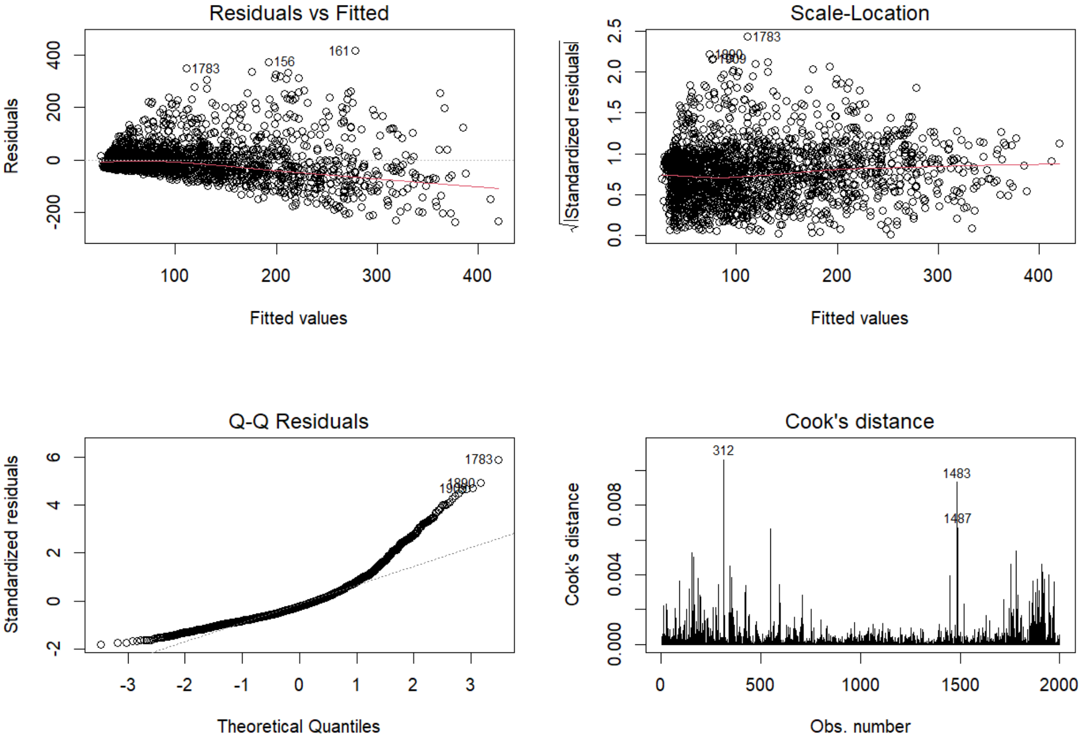

Figure 4: Result Plots of the Fitted Model 2

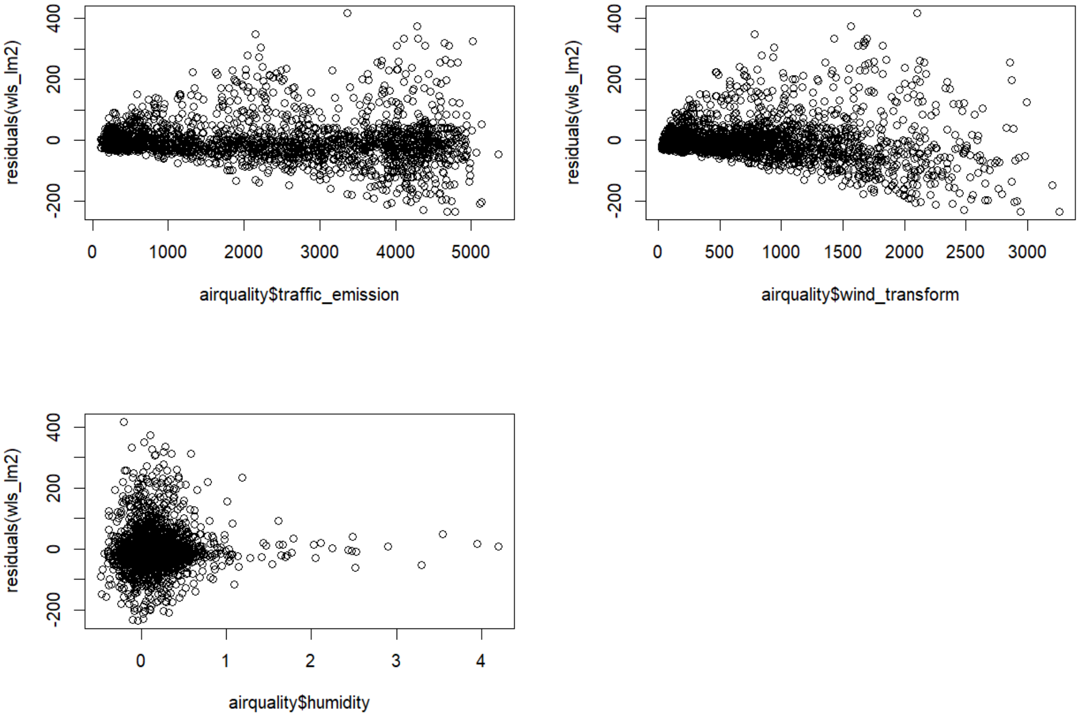

Figure 5: Residual Plots of the Fitted Model 2

1.4 Model 3: The Recommended Model

A log transformation of the predicted outcome is suggested to mitigate heteroscedasticity and recover a linear relationship between variables. Because traffic emission and pollution appear linear, traffic emission is also transformed logarithmically. The resultant model is shown in Equation 3:

$$\log P_i = \beta_0 + \beta_1 w_i + \beta_2 h_i + \beta_3 \log TE_i \tag{3}$$

The transformed data is presented in Figure 6. Compared with Figure 1, clearer linear relationships can be distinguished between log pollution and log traffic emission, and log pollution and wind.

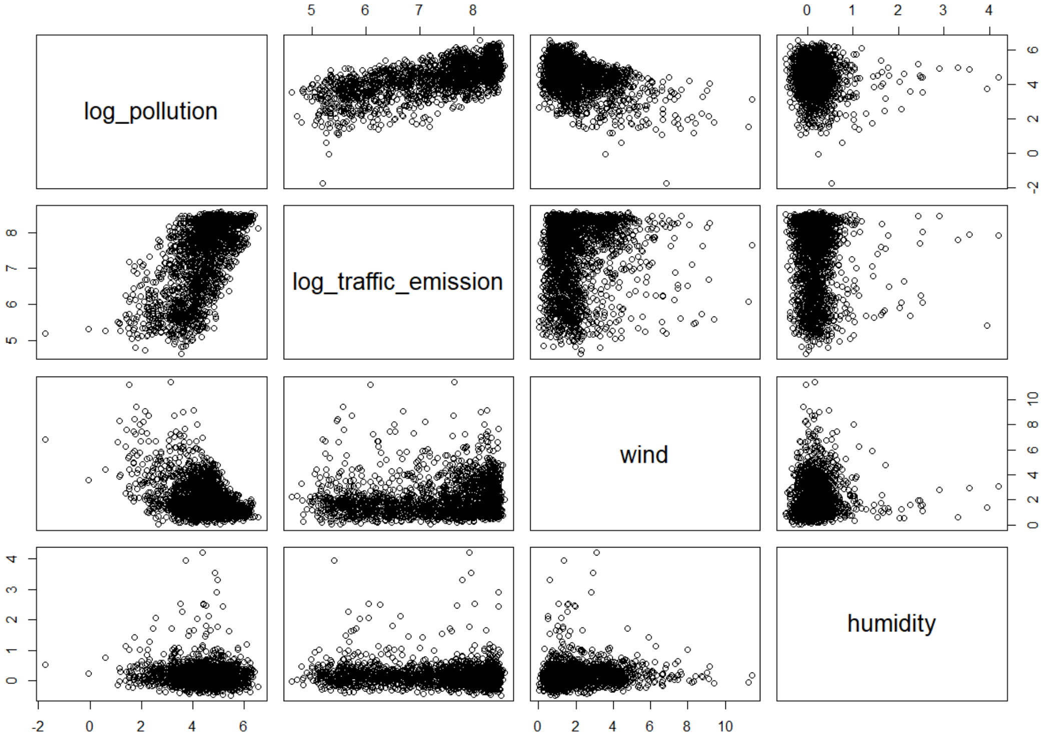

Figure 6: Data Pairplot with the log transformations.

The improvements made by the log transformation can be clearly seen by the plots in Figure 7. In the top left, the residuals versus fitted values plot shows a significant reduction in heteroscedasticity, when comparing to the same plot in Figure 4. The QQ-plot in the bottom left confirms this result, and shows a much improved normality of the residuals, outside of a few outliers in the tails. The Cook's distance plot in the bottom right shows 3 high influence points, but these do not correspond to the outliers seen in the QQ-plot, indicating that the outliers are not influential.

The models performance metrics are shown below, with the R^2 metric jumping to 0.6527. The P-value for the humidity parameter is still $0.2851 > 0.05$, which means that this parameter is not statistically significant at a 5% significance level. No significant difference in R^2 is seen by the removal of humidity as a regressor. We could thus suggest removing the humidity parameter, especially if it turns out difficult or expensive to measure.

Therefore, the final recommended model is shown below in Equation 4.

$$\log P_i = \beta_0 + \beta_1 w_i + \beta_2 \log TE_i \tag{4}$$

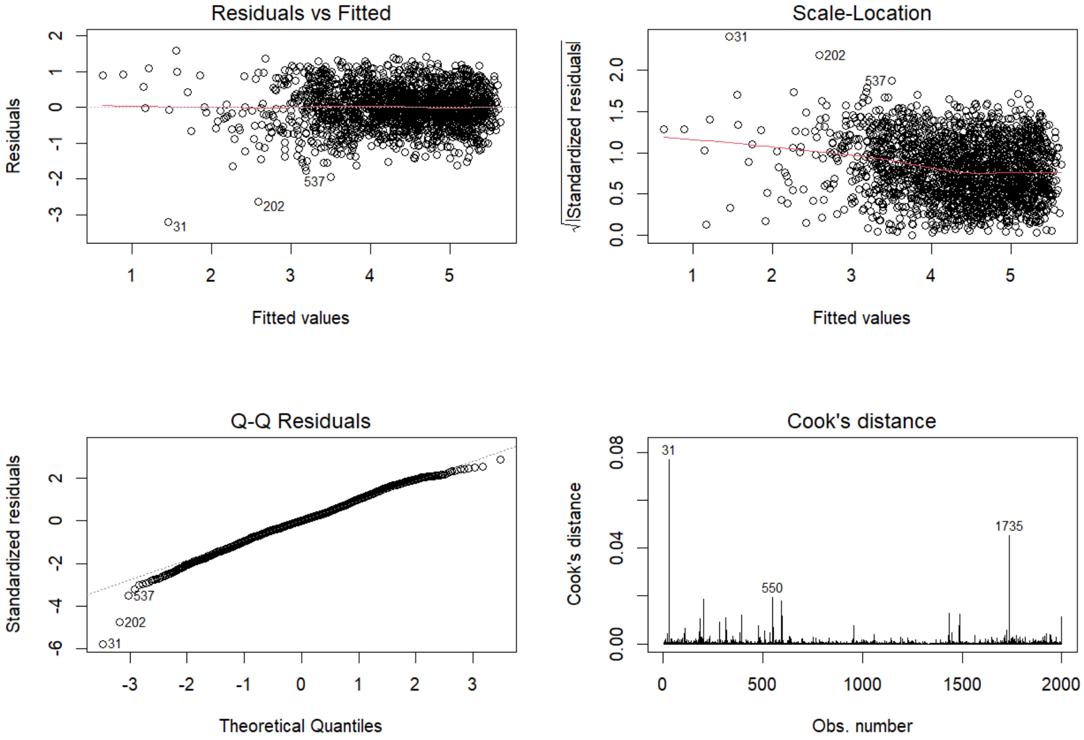

Figure 7: Result Plots of the Fitted Model 3

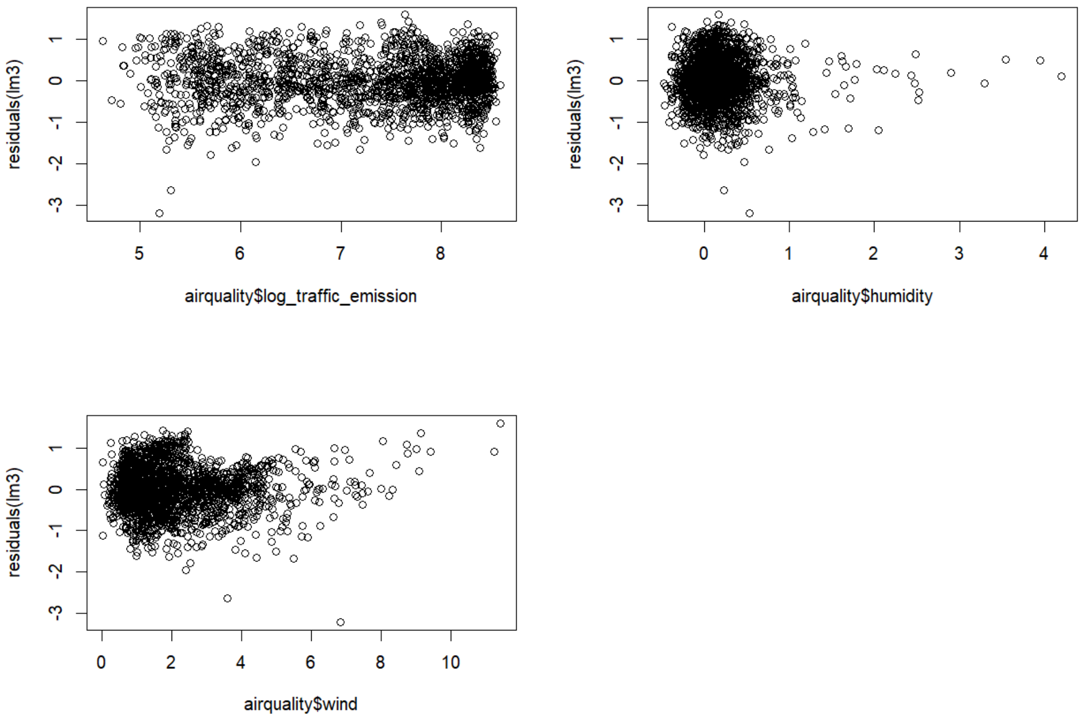

Figure 8: Residual Plots of the Fitted Model 3

Listing 4: Code and results for Model 3

```
> # Model 3: log_pollution and log_traffic_emission
>
> airquality$log_pollution <- log(airquality$pollution)
> airquality$log_traffic_emission <- log(airquality$traffic_emission)
>
> pairs( airquality[,c("log_pollution","log_traffic_emission","wind","humidity")])
> lm3 <- lm(log_pollution~log_traffic_emission+wind+humidity, data=airquality)
>
> summary(lm3)

Call:
lm(formula = log_pollution ~ log_traffic_emission + wind + humidity,
    data = airquality)

Residuals:
    Min      1Q  Median      3Q     Max
-3.2005 -0.3370  0.0020  0.3571  1.5907

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)           0.343734   0.092036   3.735 0.000193 ***
log_traffic_emission  0.642717   0.012382  51.908  < 2e-16 ***
wind                 -0.322704   0.008607 -37.491  < 2e-16 ***
humidity             -0.037626   0.035189  -1.069 0.285082
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.5574 on 1996 degrees of freedom
Multiple R-squared:  0.6527,    Adjusted R-squared:  0.6522
F-statistic:  1250 on 3 and 1996 DF,  p-value: < 2.2e-16

> plot(lm3,which=1:4,ask=FALSE)
> plot(airquality$log_traffic_emission,residuals(lm3))
> plot(airquality$wind,residuals(lm3))
> plot(airquality$humidity,residuals(lm3))
>
> # Recommended Model : log_pollution and log_traffic_emission without humidity
>
> recommended_model <- lm(log_pollution~log_traffic_emission+wind, data=airquality)
> summary(recommended_model)

Call:
lm(formula = log_pollution ~ log_traffic_emission + wind, data = airquality)

Residuals:
    Min      1Q  Median      3Q     Max
-3.2100 -0.3376  0.0026  0.3573  1.5954

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)           0.332969   0.091487    3.64  0.00028 ***
log_traffic_emission  0.643396   0.012366   52.03  < 2e-16 ***
wind                 -0.323179   0.008596  -37.59  < 2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.5574 on 1997 degrees of freedom
Multiple R-squared:  0.6525,    Adjusted R-squared:  0.6522
F-statistic:  1875 on 2 and 1997 DF,  p-value: < 2.2e-16
```

1.5 Testing the null-hypothesis

To test the null hypothesis (H_0) that the effects of the regression coefficients in Eq.4 are the same, the null hypothesis model (M_0) is constructed as shown by Eq. 5.

$$\log P_i = \beta_0 + \beta_1(w_i + \log TE_i) \tag{5}$$

The model M_0, as well as the recommended model (M_R) are fit to the data, and their respective residual sum of squares (RSS_0 and RSS_R) are computed, where $RSS = \sum_{i=1}^{n} (\mathbf{Y_i} - \mathbf{x_i^T}\hat{\beta})$. It can be shown that these are independent, and following Rice Section 6.2 the F-test can be obtained, shown in Eq.6.

$$\frac{(RSS_0 - RSS_R)/(p-q)}{RSS_R/(N-p)} \sim F_{p-q,N-p} \quad \text{under } H_0 \tag{6}$$

Where N is the number of samples, p is the number of parameters in the recommended model (M_R) and q is the number of parameters in the null hypothesis model (M_0).

The critical value is pulled from tables at the 5% significance level, $F_{crit,\alpha=0.05} = F_{1,1997} = F_{1,\infty} \approx 3.84$. In this test, the F-statistic is computed to be $F \approx 3749 \gg 3.84$. Therefore at a 5% significance level the null hypothesis can be rejected, and thus also the notion that the effect of the (non intercept) regression parameters in our recommended equation (Eq.4) are the same.

1.6 Report

A model was found to allow the prediction of pollution levels depending on both wind speed and traffic emission. The chosen model presented reasonable performance in predicting pollution, although shortcomings on the data's side may impinge on its performance, as detailed at the end of this report.

Statistical tests suggested that humidity was irrelevant to our model and could be discarded. Humidity measurements were sometimes negative, which is nonsensical in the physical sense. The hypothesis of humidity being an explanatory factor of pollution level based a corrected dataset is not excluded.

Traffic emission was found to be the largest contributor in predicting pollution, with an uptrend of pollution as a function of traffic emission. A corresponding unit change in traffic emission induces an equivalent 1.4 factor increase in pollution, if the effect of all other variables is disregarded. This relationship appears quite intuitive, the more traffic there is, the higher the emissions and the higher pollution levels are.

The second most influential variable was the wind. It was found to act as a dissipator of pollution, when in the presence of pollution. When we would observe high pollution levels from high traffic emission, a strong wind decreases the pollution. A corresponding unit change in wind induced a 0.72 factor on pollution levels disregarding the effect of other variables, effectively showing that wind is less influential than traffic emission in regards to pollution level predictions.

Although the model found achieved reasonable performance,the following recommendations are made to the institute in order to improve the accuracy of pollution level predictions in the future.

Firstly, no information is given about the wind orientation and the location of the measurements. This could provide us a better understanding on the role of wind in predicting pollution levels.

Secondly, having the dates and times at which the each measurement is taken could potentially allow better explainability to certain hours of the day or times of the week, by differentiating peak hour traffic from off-peak times.

2 Problem 2

Consider a linear regression model. Specifically, recall that $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$, as defined in the lecture notes, where $\mathbf{y}$, $\mathbf{X}$, $\boldsymbol{\beta}$ and $\mathbf{e}$ are the response vector, design matrix, parameter vector and model error term. The parameter estimator is $\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$ and the fitted value vector can be written as $\hat{\mathbf{y}} = \mathbf{H}\mathbf{y}$, where $\mathbf{H}$ is the hat matrix. Show that $\mathbf{y}^\top\mathbf{y} = \hat{\mathbf{y}}^\top\hat{\mathbf{y}} + \hat{\mathbf{e}}^\top\hat{\mathbf{e}}$

Solution

The vector of residuals $\hat{\mathbf{e}}$ is defined as $(\mathbf{I}_N - \mathbf{H})\mathbf{Y}$. From $\hat{\mathbf{y}} = \mathbf{H}\mathbf{y}$, we have

$$\begin{aligned}
\hat{\mathbf{y}}^\top\hat{\mathbf{y}} + \hat{\mathbf{e}}^\top\hat{\mathbf{e}} &= (\mathbf{H}\mathbf{y})^\top\mathbf{H}\mathbf{y} + \left[(\mathbf{I}_N - \mathbf{H})\mathbf{y}\right]^\top\left[(\mathbf{I}_N - \mathbf{H})\mathbf{y}\right] \\
&= \mathbf{y}^\top\mathbf{H}^\top\mathbf{H}\mathbf{y} + \mathbf{y}^\top(\mathbf{I}_N - \mathbf{H})^\top(\mathbf{I}_N - \mathbf{H})\mathbf{y} \\
&= \mathbf{y}^\top\left[\mathbf{H}^\top\mathbf{H} + (\mathbf{I}_N - \mathbf{H})^\top(\mathbf{I}_N - \mathbf{H})\right]\mathbf{y} \\
&= \mathbf{y}^\top\left[\mathbf{H}^\top\mathbf{H} + \mathbf{I}_N^\top\mathbf{I}_N + \mathbf{H}^\top\mathbf{H} - \mathbf{I}_N^\top\mathbf{H} - \mathbf{H}^\top\mathbf{I}_N\right]\mathbf{y} \\
&= \mathbf{y}^\top\left[\mathbf{H}^\top\mathbf{H} + \mathbf{I}_N + \mathbf{H}^\top\mathbf{H} - \mathbf{H} - \mathbf{H}^\top\right]\mathbf{y}
\end{aligned}$$

From Rice Section 14.4.3, Lemma A, we know $\mathbf{H} = \mathbf{H}^\top = \mathbf{H}^\top\mathbf{H}$. Hence,

$$\begin{aligned}
\hat{\mathbf{y}}^\top\hat{\mathbf{y}} + \hat{\mathbf{e}}^\top\hat{\mathbf{e}} &= \mathbf{y}^\top\left[\mathbf{H}^\top\mathbf{H} + \mathbf{I}_N + \mathbf{H}^\top\mathbf{H} - \mathbf{H} - \mathbf{H}^\top\right]\mathbf{y} \\
&= \mathbf{y}^\top\left[2\mathbf{H} - 2\mathbf{H} + \mathbf{I}_N\right]\mathbf{y} \\
&= \mathbf{y}^\top\left[\mathbf{I}_N\right]\mathbf{y} \\
&= \mathbf{y}^\top\mathbf{y}
\end{aligned}$$

Thus,

$$\boxed{\mathbf{y}^\top\mathbf{y} = \hat{\mathbf{y}}^\top\hat{\mathbf{y}} + \hat{\mathbf{e}}^\top\hat{\mathbf{e}}}$$

3 Problem 3

Consider the following simple linear model, where the ϵ_i are independent and have mean zero.

$$y_i = \beta_0 + \beta_1(x_1 - \bar{x}) + \epsilon_i, \quad i = 1, ..., n.$$

(a) Write down $\mathbf{X}^\top\mathbf{X}$, where $\mathbf{X}$ is the design matrix for the above model.
(b) Consider the case where all x_i are equal. What can you say about $(\mathbf{X}^\top\mathbf{X})^{-1}$?
(c) When all x_i are equal, what possible values can $\hat{\beta}_1$ take?

Solution

(a) The design matrix $\mathbf{X}$ of the linear model is

$$\mathbf{X} = \begin{bmatrix} 1 & (x_1 - \bar{x}) \\ 1 & (x_2 - \bar{x}) \\ \vdots & \vdots \\ 1 & (x_n - \bar{x}) \end{bmatrix}$$

for the vector parameter

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}.$$

Then,

$$\begin{aligned} \mathbf{X}^\top \mathbf{X} &= \begin{bmatrix} 1 & \cdots & 1 \\ (x_1 - \bar{x}) & \cdots & (x_n - \bar{x}) \end{bmatrix} \begin{bmatrix} 1 & (x_1 - \bar{x}) \\ 1 & (x_2 - \bar{x}) \\ \vdots & \vdots \\ 1 & (x_n - \bar{x}) \end{bmatrix} \\ &= \begin{bmatrix} 1 & \sum_{i=1}^n (x_i - \bar{x}) \\ \sum_{i=1}^n (x_i - \bar{x}) & \sum_{i=1}^n (x_i - \bar{x})^2 \end{bmatrix} \end{aligned}$$

(b) If all x_i are equal, then $x_i = \bar{x}$. This means

$$\sum_{i=1}^n (x_i - \bar{x}) = \sum_{i=1}^n (\bar{x} - \bar{x}) = 0$$

and similarly,

$$\sum_{i=1}^n (x_i - \bar{x})^2 = \sum_{i=1}^n (\bar{x} - \bar{x})^2 = 0.$$

$\mathbf{X}^\top \mathbf{X}$ then becomes

$$\mathbf{X}^\top \mathbf{X} = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}.$$

We notice that $\mathbf{X}^\top \mathbf{X}$ does not have full rank and is thus not invertible. $(\mathbf{X}^T \mathbf{X})^{-1}$ does not exist.

(c) If all x_i are equal, then $x_i = \bar{x}$ and the linear model becomes

$$\begin{aligned} y_i &= \beta_0 + \beta_1 (x_i - \bar{x}) + \epsilon_i \\ &= \beta_0 + \beta_1 (\bar{x} - \bar{x}) + \epsilon_i \\ &= \beta_0 + \beta_1 \cdot 0 + \epsilon_i \end{aligned}$$

β_1 can take any values. There does not exist a unique solution.