Imperial College London
Department of Mechanical Engineering
MECH60017 Statistics A Coursework

Statistical Modelling, and Analysis on Data Collected Regarding Air-Pollution In a Town In Europe

---

Oluwaseyi Ayodeji Adeniyi

Word Count:

Date: 07/03/2024

CID: 01858040

Table of Contents

Introduction

Daily data collected over 150 days by a team studying outdoor air-pollution, and consequently the air quality in a town in Europe is to be used to make inferences on the variability of Carbon Monoxide (CO) concentration, with respect to Temperature (T) , and Relative Humidity (%) of the surrounding environment. The aim of this study is to model the CO concentration as a function of the more easily obtained measurements of T, and RH. Additionally, the inclusion of a time index as a covariate is to be investigated, and remarks concerning the quality of data collected, as well as critiques of data analysis methods used will follow.

Exploratory Data Analysis

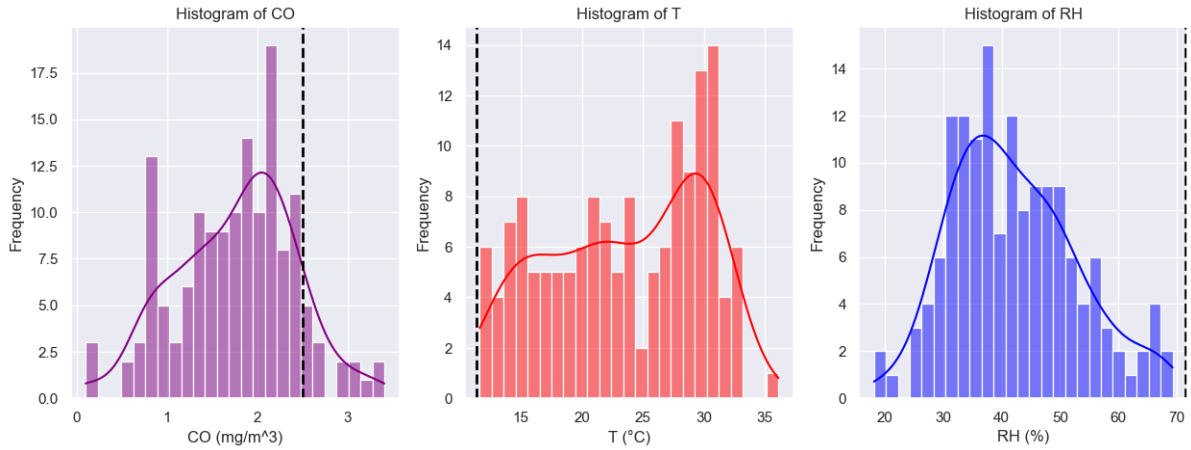

Figure 1: Histograms showing distributions of (i) *CO conc., (ii)* Temperature *°C , and (iii) RH% (L – R)*

From the samples collected, some preliminary analysis was undertaken on the distributions shown by the 3 quantities of interest; see Fig. 1 above. The raw data was firstly collated into a set of histograms, bin size 25, with kernel density lines shown to aid distribution inspection. This data is matched against expected values[A] for the quantities of interest depicted using the black dashed line; see Appendix A.

The leftmost plot, shows Carbon Monoxide (CO) concentration in mg/mm^3 over the course of the study. It is slightly negatively skewed and asymmetrical with a high variance of frequencies. As for the plot in the centre, showing variations in Temperature (°C), which is similarly negatively skewed but much more asymmetrical as the left tail shows there was a large range of temperatures experienced during the 150 day time frame. The rightmost plot is a histogram of the distribution of Relative Humidity. From its kernel density line estimate, it can be noticed that it is positively skewed and although asymmetrical, it is the least asymmetrical of the 3 plots. In comparison to the expected values, none of the plots perfectly match expectations from prior research. The CO concentration being the closest exhibits its peak at circa 2 mg/mm^3 while the expected value was 2.5 mg/mm^3. According to the EEE[1], and another source, the annual mean temperature is between 2 - 17 °C. This was overshot greatly by the measurements which exhibit a peak frequency of 13 in the 26 – 28 °C range. Almost double the expected average. As for the RH, the humidity levels were expected to range from 60.5 % – 82.5%[2], but the population mean lies much lower at circa 37%. From the onset, these discrepancies suggest that there could be some fault in the methods, and/or apparatus used for measurement.

Scatter Plots were configured to gain even better insight into possible relationships within the dataset. The plots are for CO concentration against Temperature, CO concentration against RH, and Temperature against RH. Which yielded correlation coefficients of -0.2363, 0.1288, and -0.6659 respectively; see Fig. 2. 2 out of 3 of

[1] European Environment Agency

[2] CoolWeather.net

these plots show negative correlation, which implies that they move in opposite directions when their values are slightly altered, with the inverse correlation between T and RH being the best. The scatter diagram clearly shows that there could be an approximately linear relationship between T and RH, and this is expected since taking a glance at a psychometry chart shows an exponential distribution which can be accurately assumed linear for some sections of the curve.

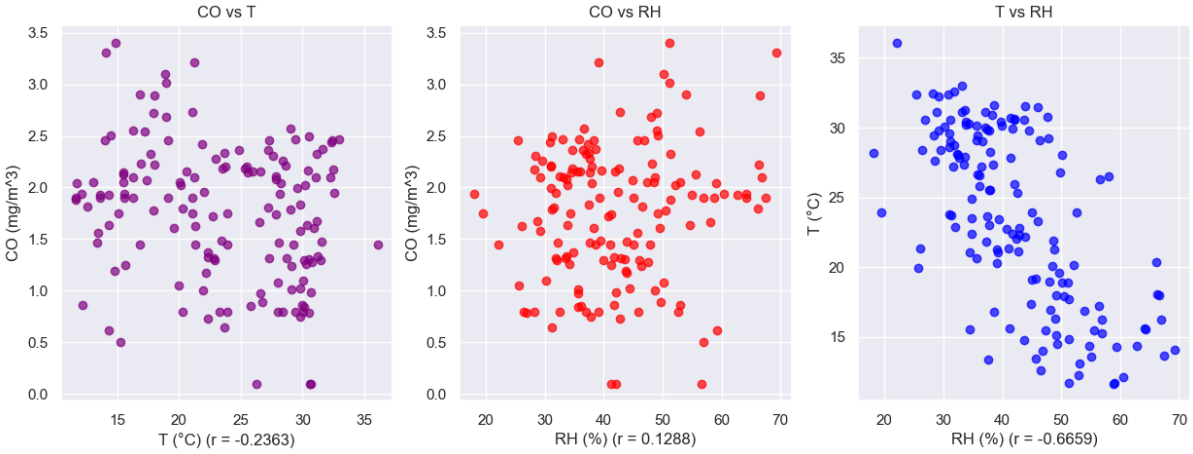

Figure 2: Scatter *Plots of Certain* Relationships *(i) CO vs T, (ii) CO vs RH, (iii) T vs RH (L - R)*

Modelling

The scatter diagrams in Fig. 2 had linear regression models fit over i, and ii the resulting coefficients of determination are 0.0558, and 0.0166 respectively. This implied that the model suggests the variation in Carbon Monoxide Concentration was only about 5.6% due to changes in Temperature , and 1.7% due to changes in the Relative Humidity of the surrounding Atmosphere. According to textbook heuristic, since both values of r^2 obtained are below 0.5 i.e., 50%, the correlation must be weak. Another consideration would be that the relationship is non-linear, but from the scatter chart, no higher order polynomial is immediately obvious, and therefore the inference that the relationships between these quantities are weak must be true.

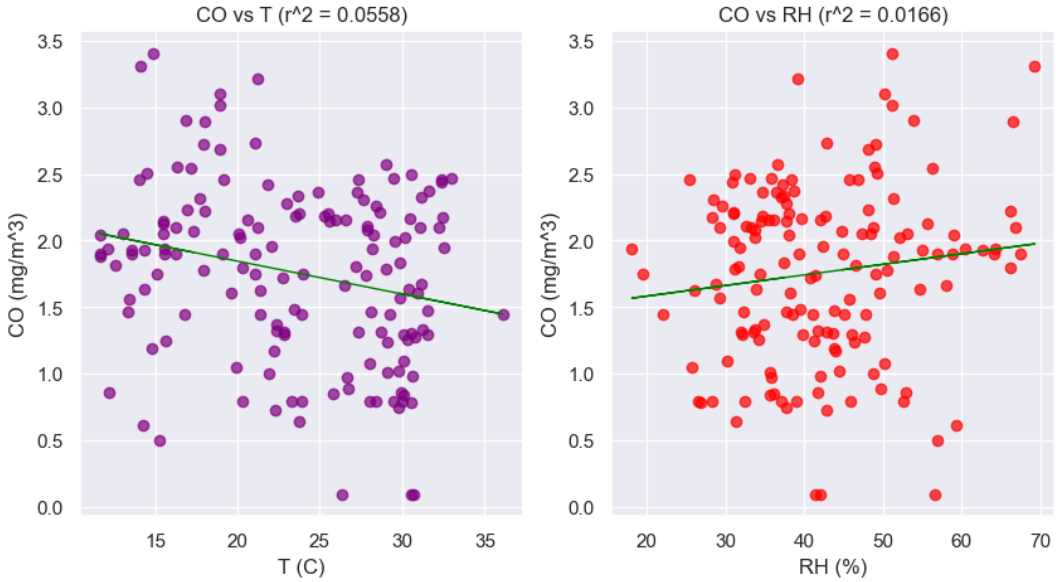

Figure 3: Scatter Plots with Linear Regression Fit onto them

As is evident from Figs. 2, and 3 above, the lack of any trends in the plotted points, even when a best fit line is drawn shows that the linear regression model is not a great fit for the data sample available. Further investigation is needed, and given access to the entire population the next logical thing is to apply a linear

model to a scatter of CO against T and RH making use of a polynomial regression instead . This resulted in a slight improvement in coefficient of determination but too minimal to recharacterize the coefficient as moderate; increased r^2 from .

Changing from linear to a polynomial regression however, shows instant improvements to the calculated coefficient of determination. The outcome of increasing the order of the polynomial regression from linear to quartic is shown in the form of a bar chart in Fig. 4 below. Here we can see that a quartic regression yields a coefficient corresponding to > 20%, implying that increasing the order of polynomials, is a step in the right direction. Increasing the order of the polynomial continues this relationship until n = 10 is reached where there is a steep drop off to circa r^2 = -1.5 and at n = 11 the curve immediately bounces back and continues on its upward trend; see Fig. 4 right.

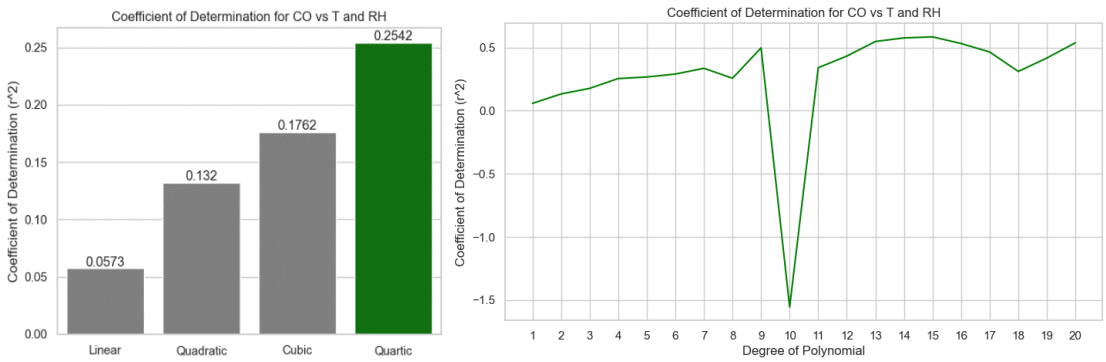

Figure 4: Variations *in Coefficient of Determination as n increases (i) Bar Chart n = 1 – 4 (ii) Line Diagram n = 1 – 20; where n is the order of the regression polynomial (L-R)*

Fitting the scatterplot with multiple regression models showed that the

Prediction

Furthering the investigation into the relationships present, we employ the 2-fold cross-validation method, in which we use one half of the sample to instantiate a model, and the other half of the sample to test said model, recording the residuals that occur for each prediction, and repeating vice versa. The 2-fold cross-validation method was employed on the 3 models being considered. CO vs T, CO vs RH, and CO vs T and RH.

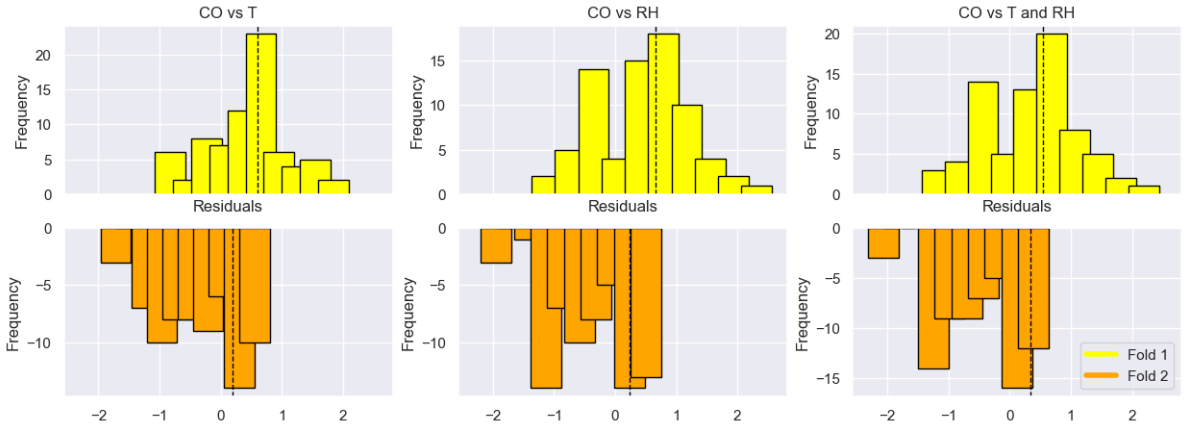

Figure 5: Bar Charts *to Assess Normality of Residuals from 2-fold cross-validation*

After performing 2-fold cross-validation on the models available, the residuals of both folds were collated and plot on a pseudo – butterfly chart[3]. The fold 1, and 2 residuals are plot as yellow and orange bars respectively, and their means are plot as dashed black lines. Overall the fold 1 Residuals exhibit more normality in terms of frequency distribution all distributions being slightly negatively skewed (excluding the distribution of CO vs RH) and centred roughly around the 0-1 bin. For fold 2, the Residuals are all heavily negatively skewed where a steep drop off at the right tail would be noticed if the kernel density estimates were plot.

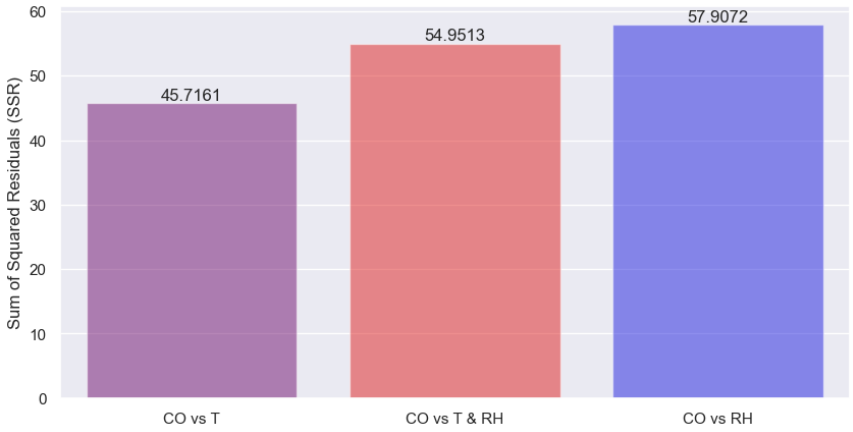

Figure 6: Sum of Squared Residuals for the 3 models considered

Assessing the models using the sum of the squares of the residuals, the CO vs T model has the lowest value (SSR = 45.72), with the CO vs RH model having the highest value (SSR = 57.9). This implies that they have the best and poorest fits to the given data respectively.

Including Time as a Feature

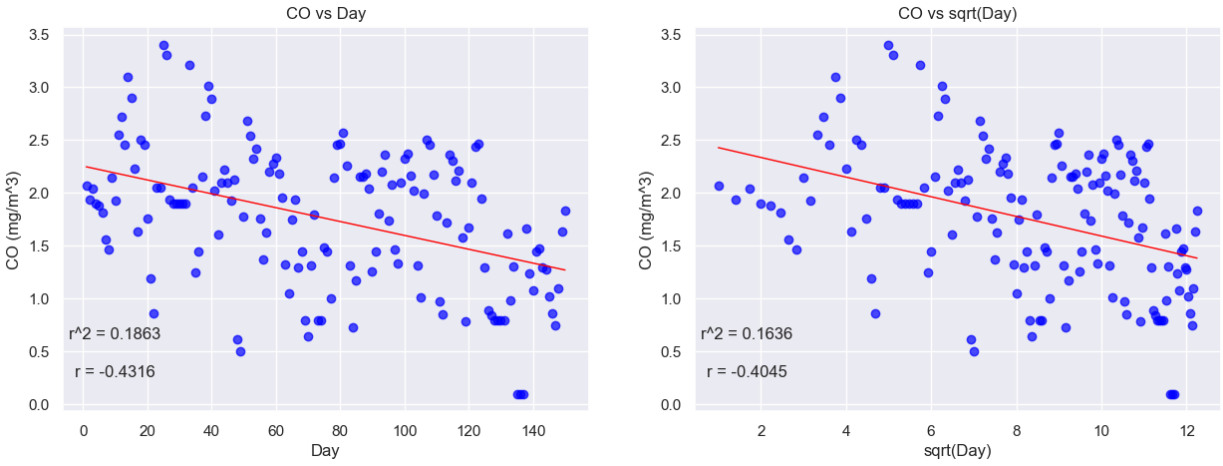

Figure 7: (i) Carbon Monoxide Conc. with time *as a covariate (i) Carbon Monoxide Conc. with squared root of time as a covariate*

Further inspection of the carbon monoxide concentration with respect to time in the form of number of days and squared root of the number of days was carried out. These relationships were pot on a scatter graph and yielded correlation coefficients of 0.1863, and 0.1636 respectively showing weak correlation once again (< 0.5).

[3] Note that the frequency on the bottom part of the chart is not actually representative of real values, they are negative to allow the chart construction; negative frequencies are not possible in reality.

The linear regression model built for the CO vs Day relationship shows a better overall fit with the line of best fit, as there is a more balanced distribution of points above and below said line, when compared to the CO vs sqrtDay plot.

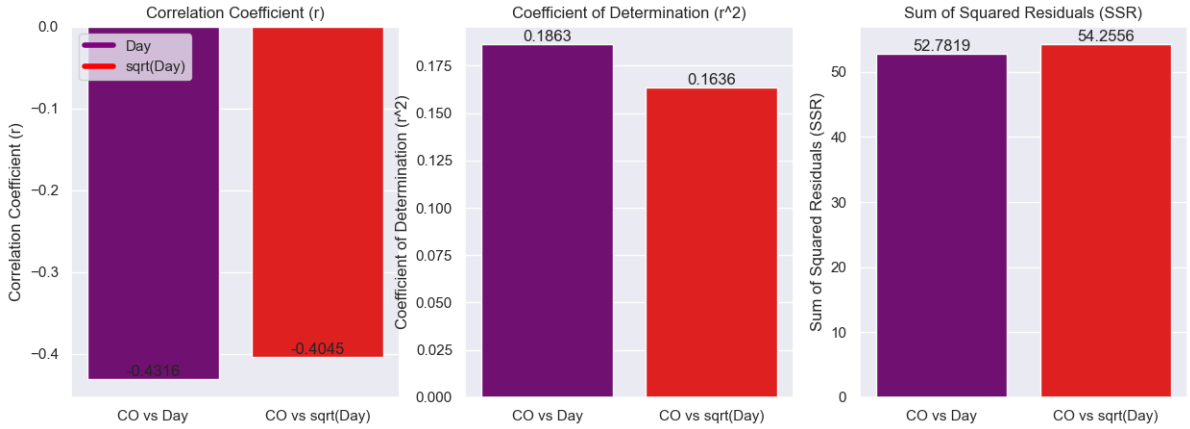

Figure 8: 3 Quantity 2-Way Model Comparison

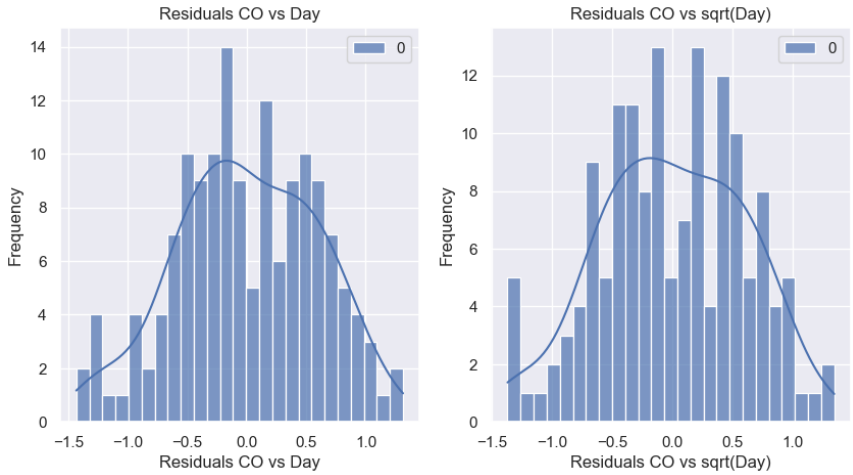

Figure 9: Histograms Showing Distribution of Residuals for 2 Models Considered

The better model is evidently, CO vs Day, due to its larger correlation coefficient, coefficient of determination, and lower sum of squared residuals albeit minute differences between the two models considered see Fig. 8 above where a bar chart comparing r, r^2, and the ssr of both models is plot.

Assessing the normality of errors, both distributions are asymmetric but roughly centred close to bin -0.1 with the sqrtDay residuals having higher variance evident by the wider peak on the kernel density estimate line diagram.

An even better model could be obtained by using polynomial regression as outlined in the modelling section of the report. Slightly outside the scope of the coursework, the models can be ranked using AICs, and BICs (Akaike Information criterion, and Bayesian Information Criterion) to rank the models more explicitly, only making improvements to the best models from each polynomial level.