This report aims to analyse data an on the daily air quality measurements in New York City for 30 consecutive days. The four variables studied are: `ozone` (Y) – the surface concentration of ozone ; `radiation` (X_1) – the solar radiation; `temperature` (X_2) – the temperature in degrees Fahrenheit(°F); `wind` (X_3) – wind speed measured in miles per hour(mph). With $n = 30$ observations, these variables are explored further starting with a plot of each variables against each other.

(A) Before making further assumptions on the model relating these variables to each other, a plot of the data helps visualise the relationship each variable may have with one another. To do this, the command `pairs` in R is used.

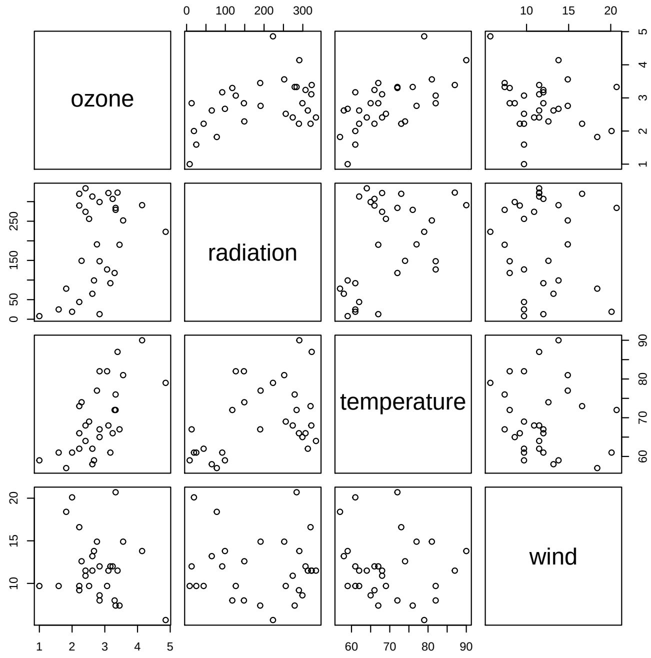

Figure 1: Scatterplot of X_1, X_2, X_3 and Y against each other

In the first row, we see plots of `radiation`, `temperature` and `wind` against the `ozone` concentration. `radiation` and `temperature` data against `ozone`, seems to have a linear pattern with a positive-valued slope, with the slope of `temperature` vs `ozone` being higher than `radiation` vs `ozone`. It is also noticeable in the latter case, that the slope seems to be very close to zero. In contrast, the plot of `wind` vs `ozone` seems to show a negative, linear relationship between the two variables, where a lot of the data seems to be clustered between 0mph to 15mph of the wind speed and 2 to 4 concentration of the `ozone` surface. The plots in the first column show similar patterns to the first row since the variables are flipped. There are no observable patterns shown in the rest of the plots in the diagram.

(B) In this part of the report, we will be assuming that the variables relate to each other through a multiple linear regression model, with the `ozone` as the response and the other three variables as the regressors.

(i) The model to be fitted is

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

$$\boldsymbol{Y} = \boldsymbol{\mu} + \boldsymbol{\epsilon}$$

where $\boldsymbol{Y}$ and $\boldsymbol{\epsilon}$ are $n \times 1$ matrices, containing $n = 30$ observations of variable `ozone` and n residuals, $\boldsymbol{Y} - \boldsymbol{\mu}$ respectively. $\boldsymbol{X}$ is an $n \times p$ matrix where the first column only contain 1, and the rest of the columns contain observations of the regressors- `radiation` (X_1), `temperature` (X_2), `wind` (X_3) in that order. $\boldsymbol{\beta}$ is a $p \times 1$ matrix containing all the parameters where in this case, $p = 4$. The matrix multiplication $\boldsymbol{X\beta}$ makes $\boldsymbol{\mu}$ become an $n \times 1$ matrix.

The errors in this model, $\boldsymbol{\epsilon}$ are assumed to be independent of each other and normally distributed with mean 0 and constant variance σ^2. The fitted values, $\hat{\boldsymbol{Y}}$ is the model fitted without the errors, namely $\hat{\boldsymbol{Y}} = \boldsymbol{X\beta}$.

(ii) Using R, we have

a. inverse of information matrix,

$$\boldsymbol{G}^{-1} = (\boldsymbol{X}^T\boldsymbol{X})^{-1} = \begin{pmatrix} 2.8559225951 & 6.173356e-04 & -3.535069e-02 & -4.090973e-02 \\ 0.0006173356 & 3.341412e-06 & -1.807261e-05 & 1.409908e-08 \\ -0.0353506868 & -1.807261e-05 & 5.338541e-04 & 1.439104e-04 \\ -0.0409097295 & 1.409908e-08 & 1.439104e-04 & 2.613921e-03 \end{pmatrix}$$

b. $\boldsymbol{X}^T\boldsymbol{Y} = \begin{pmatrix} 83.840 \\ 17097.380 \\ 5953.470 \\ 973.641 \end{pmatrix}$

c. $SST = \boldsymbol{Y}^T\boldsymbol{Y} = 251.0472$

(iii) The first two components in part (B)(ii) are used to calculate the estimate the model parameters, vector $\hat{\boldsymbol{\beta}}$ using the least squares (LS) method. In summary, the LS estimators are obtained by minimising the sum of squares of the errors,

$$SSE = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \boldsymbol{\epsilon}^T\boldsymbol{\epsilon}$$

through differentiation with respect to each element in $\boldsymbol{\beta}$, equating $p = 4$ equations to zero and solving for $\hat{\boldsymbol{\beta}}$ we get

$$\hat{\boldsymbol{\beta}} = \boldsymbol{G}^{-1}\boldsymbol{X}^T\boldsymbol{Y} = \begin{pmatrix} -0.295271027 \\ 0.001305816 \\ 0.045605744 \\ -0.027843496 \end{pmatrix}$$

The first element in the vector, $\hat{\beta}_0$ is the estimate for the intercept while the other three are the estimates of the slopes of each variable in $\boldsymbol{X}$. Looking at the signs of the slope parameter estimates, we see that they match with the comments made in part (A) about their possible relationship. We also see how close to zero the second parameter, $\hat{\beta}_1$ estimate is, as we noted when looking at the plots.

Source of Variation	Df	Sum of Squares	Mean Sum of Squares
Radiation	1	3.1469	3.1469
Remp	1	4.2250	4.2250
Wind	1	0.2966	0.2966
Residuals	26	9.0738	0.3490
Total	30	16.7423	

Table 1: Analysis of Variance Table

(iv)

(v) An F-test at 5% significance level for overall significance of the regression model can be done, testing if *all* the parameters are equal to zero.

$$H_0 : \boldsymbol{\beta} = \mathbf{0} \text{ vs } H_A : \beta_i \neq 0 \text{ for at least one value of } i = 0, 1, 2, 3$$

where the test statistic can be obtained as follows

$$F = \frac{(SST - SSE)/p}{SSE/(n-p-1)} \sim F(p-1, n-p)$$

The p-value 0.001026 is computed for F = 7.324408. Since the value is smaller than 0.05, it is safe to say that there is significant evidence that at least one of the parameter values in $\hat{\boldsymbol{\beta}}$ is not equal to zero. This means that at least one of the regressors, `radiation`, `temperature` or `wind` does affect the `ozone` linearly.

(vi) The coefficient of determination, R^2 is the proportion of the variation in the dependent variables explained by the model. In other words, it is a measure of goodness of fit of a linear regression model.

$$R^2 = \frac{SST_c - SSE}{SST_c} = 0.458$$

where $SST_c = SST - n\bar{Y}^2 = 16.74235$. This tells us that 45.8% of the variation in the model is explained by the multiple linear regression model. In general, a higher (closer to 1) R^2 value is preferred, however, the 45.8% is a good proportion to start off with when model building.

(vii) Assuming that the observations from the two observed days are independent, to predict the difference in `ozone` concentration in the two days, we get the observed difference in the variables which in this case is (50, -10, 0). These values can then be substituted into $\boldsymbol{f}_0$ in the formula for fitted values to get the predicted difference in `ozone`,

$$\begin{aligned}\hat{Y} &= \boldsymbol{f}_0^T \hat{\boldsymbol{\beta}} \\ &= \begin{pmatrix} 1 & 50 & -10 & 0 \end{pmatrix} \begin{pmatrix} -0.295271027 \\ 0.001305816 \\ 0.045605744 \\ -0.027843496 \end{pmatrix} \\ &= -0.6860377\end{aligned}$$

which says that the `ozone` concentration dropping by 0.686 on the second day compared to the first based on the fitted model. The $(100 - \alpha)\%$ prediction interval is defined as follows

$$\boldsymbol{f}_0^T \hat{\boldsymbol{\beta}} \pm t_{n-p,\alpha/2} \hat{\sigma} \sqrt{1 + \boldsymbol{f}_0^T \boldsymbol{G}^{-1} \boldsymbol{f}_0}$$

Substituting all the terms in the formula, the interval for the prediction made is

$$[-1.333036, -0.03903891]$$

The interval obtained using the formula tells we have 95% confidence that the difference of the `ozone` surface concentration between two days where the `radiation` rises by 50, the `temperature` is drops by 10°F and there is no change in the `wind` speed, will drop between 0.039 and 1.333.

(viii) The residuals plot against the fitted values and each of the covariates in Figure 2 help verify the assumptions that were made about the errors in part (B)(ii). The constant variability can be confirmed when there are no observable patterns in the plot which we can see in all four plots. The data points are also concentrated about the red line, when the residuals are zero which validates the mean assumed. We can also tell that the plots are fairly symmetrically distributed about the mean, which confirms its Normality.

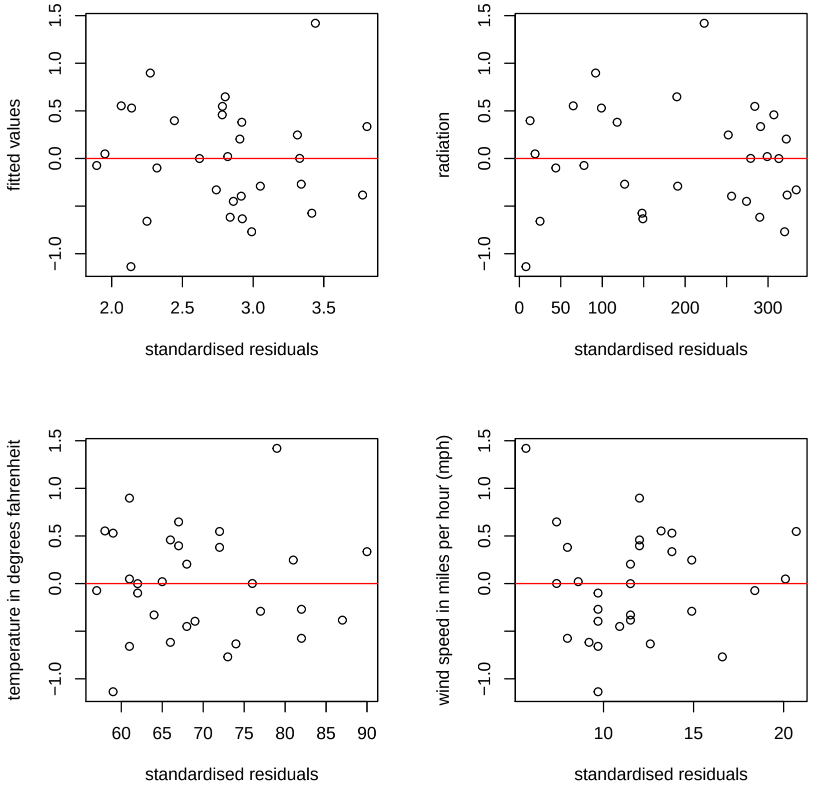

Figure 2: Scatterplot of standardised residuals against fitted values and each regressor

(C) (i) To test if `temperature`, X_2 can be removed from the regression model, the following null hypothesis is used

$$H_0 : \boldsymbol{C}\boldsymbol{\beta} = 0$$

where $\boldsymbol{C} = \begin{pmatrix} 0 & 0 & 1 & 0 \end{pmatrix}$ or in simpler terms,

$$H_0 : \beta_2 = 0 \text{ vs } H_A : \beta_2 \neq 0$$

(ii) The $100\gamma\%$ confidence interval for individual parameters can be defined as follows,

$$\hat{\beta}_i \pm t_{n-p,\frac{1-\gamma}{2}}\hat{\sigma}\sqrt{(\boldsymbol{G}^{-1})_{ii}}$$

where $\hat{\sigma}^2 = SSE/(n-p)$ and $(\boldsymbol{G}^{-1})_{ii}$ is the ith diagonal element in the inverse of information matrix, $\boldsymbol{G}^{-1}$. Plugging in all the terms in the formula($t = 2.056$, $\hat{\sigma} = 0.5907$, $(G^{-1})_{ii} = 0.0005339$), we have the confidence interval

$$[0.0175485843, \quad 0.073662904]$$

Based on the interval, we are 95% confident that the true parameter value for how the `temperature` affects the `ozone` in the model, β_2 will fall between 0.0175 and 0.07366. It is also important to note that the confidence interval does not include 0, and since they complement hypothesis tests, we expect that the test in part (C)(ii) will conclude that the variable *cannot* be removed from the model.

(iii) To conclude the test, we must first calculate the test statistic,

$$\frac{\hat{\beta}_i - c_i}{\hat{\sigma}\sqrt{(\boldsymbol{G}^{-1})_{ii}}} \sim t_{n-p}$$

The test statistic obtained is 3.341177. Under the null hypothesis, it follows t distribution with 26 degrees of freedom. The quantile at 0.975 is 2.055529. Since the test statistic is larger than this value, there *is* significant evidence to reject the null hypothesis that the variable `temperature` has no effect on the surface concentration of the `ozone`. The test conducted does correspond to the aforementioned confidence interval.

(D) (i) In this final part of the report, we introduce a model that has model (B) nested in it. This model is defined similarly to model (B) but now with interaction terms included. Therefore, now $p = 7$ and matrix $\boldsymbol{X}$ has more columns of observations to it

$$\boldsymbol{X} = \begin{pmatrix}\mathbf{1}_n & \boldsymbol{X_1} & \boldsymbol{X_2} & \boldsymbol{X_3} & \boldsymbol{X_4 = X_1X_2} & \boldsymbol{X_5 = X_1X_3} & \boldsymbol{X_6 = X_2X_3}\end{pmatrix}$$

Using the same formula mentioned in part (B)(iii), the least squares estimate of this full model can be calculated;

$$\hat{\boldsymbol{\beta}} = \begin{pmatrix} -3.624373e+00 \\ -2.983466e-04 \\ 9.302588e-02 \\ 2.760159e-01 \\ 2.559709e-05 \\ 2.205472e-05 \\ -4.468543e-03 \end{pmatrix}$$

(ii) The proportion of explained variability to the total variability in the model is 0.4821. Similar to the previous nested model, this value is a good step towards a better model. With that said, not all regression models need an R^2 value that is close to 1 since there are other factors to be considered when evaluating if a model fits the data to explain current ones and predict the future outcome.

(iii) Using the same formula stated in part (C)(ii), The 95% confidence intervals for the model parameters are compiled in the following table

ith parameter	confidence interval
$\hat{\beta}_0$	[-1.257773e+01, 5.3289822545]
$\hat{\beta}_1$	[-2.491653e-02, 0.0243198330]
$\hat{\beta}_2$	[-4.600561e-02, 0.2320573786]
$\hat{\beta}_3$	[-4.403388e-01, 0.9923705637]
$\hat{\beta}_4$	[-2.930591e-04, 0.0003442532]
$\hat{\beta}_5$	[-8.270448e-04, 0.0008711543]
$\hat{\beta}_6$	[-1.623413e-02, 0.0072970402]

Looking at the signs of the upper and lower bound of all of the confidence intervals in the table, we see that they all include 0. This means that with this full model, all the parameter values may take value 0. We also note that the interaction terms' upper and lower bounds are both close to zero which means that even if the small sample size $n = 30$ is increased for the intervals to grow smaller, it is possible that the true parameter still take 0 as its value.

(iv) The main comparison we can make between full model (D) and reduced model (B) is by comparing the R^2 values of each model. We see that the coefficient of determinant for model (D) is slightly higher than that of model (B). However, this tells us little about the proportion of explained variation to the total variation since it is natural for the R^2 determinant to increase when more variables are added to the model. It would be more useful to compare both of the models using the adjusted R^2 since it takes that into account.

(v) Finally, we conduct a hypothesis test at 5% significance level

$$H_0 : \boldsymbol{\beta}_2 = \mathbf{0}$$
$$F = \frac{(SSE_{(B)} - SSE_{(D)})/q}{\hat{\sigma}^2_{(D)}} \sim F_{q,n-p}$$

where $\boldsymbol{\beta}_2$ is a $(q = 3) \times 1$ matrix containing the parameters of the interaction terms. Using the formula, the test statistic obtained is 0.3568617. We compare this value to quantile 0.95 of the assumed distribution under the null hypothesis, which is 3.027998. Since the F-statistic is a lot smaller than the quantile, there is *not* enough evidence to reject the null hypothesis that the true parameter values for the interaction terms are equal to zero. Then it *is* possible that how `radiation`, `temperature` and `wind` interact with each other does not affect the `ozone` concentration.