**Problem Set 1** **Harshita Jhavar**
The Elements of Statistical Learning, WS 2017/18 Matriculation Number: 2566267
Prof. Dr. Dr. Thomas Lengauer s8hajhav@stud.uni-saarland.de
TA: Michael Scherer Due Date: 08/11/2017 by 10.00 am

## Solution 1 : Main Principles of Statistical Learning (T, 10 Points)

A machine is said to be 'Learning' a task from recorded set of experiences if its performance eventually improves for given task following experiences. This experiences can be recorded either in categorical data format or numerical data format. Entire copus is divided into training data and test data. Learning is performed on training data while accuracy of learned model is obtained from test data. If training data involves both input and output labels, it is called as supervised learning while if there is only input data and no output data labels in training corpus, it is called as unsupervised learning. Supervised learning ideally aims at obtaining a model function which gives same response or outcome as that of label associated with input in gold corpus. Based on type of task, it can be classified into two types of algorithms: Regression and Classification. In classification, independent class variables or predictors are categorized based on a set of categorical variables using binary form of representation. These categorical variables are representation of some qualitative property of predictors. In regression, predictors and dependent variables or outcomes are quantitative variables whose relationship can be represented by a family of function: ex- Linear function, Polynomial. Form of model is either assumed to be from a particular family of function (parametric modelling) or form of function is learnt without any prior assumptions (non parametric modelling). In latter case, there is more chance of overfitting as we have to derive perfect trade-off between different parameters where all of these parameters are unset. Interestingly, a learnt model aids in predictions of correct class labels and helps in inferencing relations between different features. Through inferencing, one can identify which feature(s) played a dominating role and thus, perform some exploratory analysis to make predictions and conclusions about nature of data.

**Solution 2 : Proof (T, 8 Points)**

Q-2) To prove: $E(Y) = \text{argmin}_c E\left[(Y-c)^2\right]$.

Proof: Let $f(c) = E\left[(Y-c)^2\right]$.

$= E\left[Y^2 + c^2 - 2Yc\right]$

$= E[Y^2] + E[c^2] + E[-2Yc]$ (By property of linearity).

Finding $f'(c)$ to minimize and see at what value of 'c' we get minimum value;

$f'(c) = 2E[Y] - 2E[c]$

$= 2E[Y] - 2c \overset{\text{set}}{=} 0$.

Solve for c:

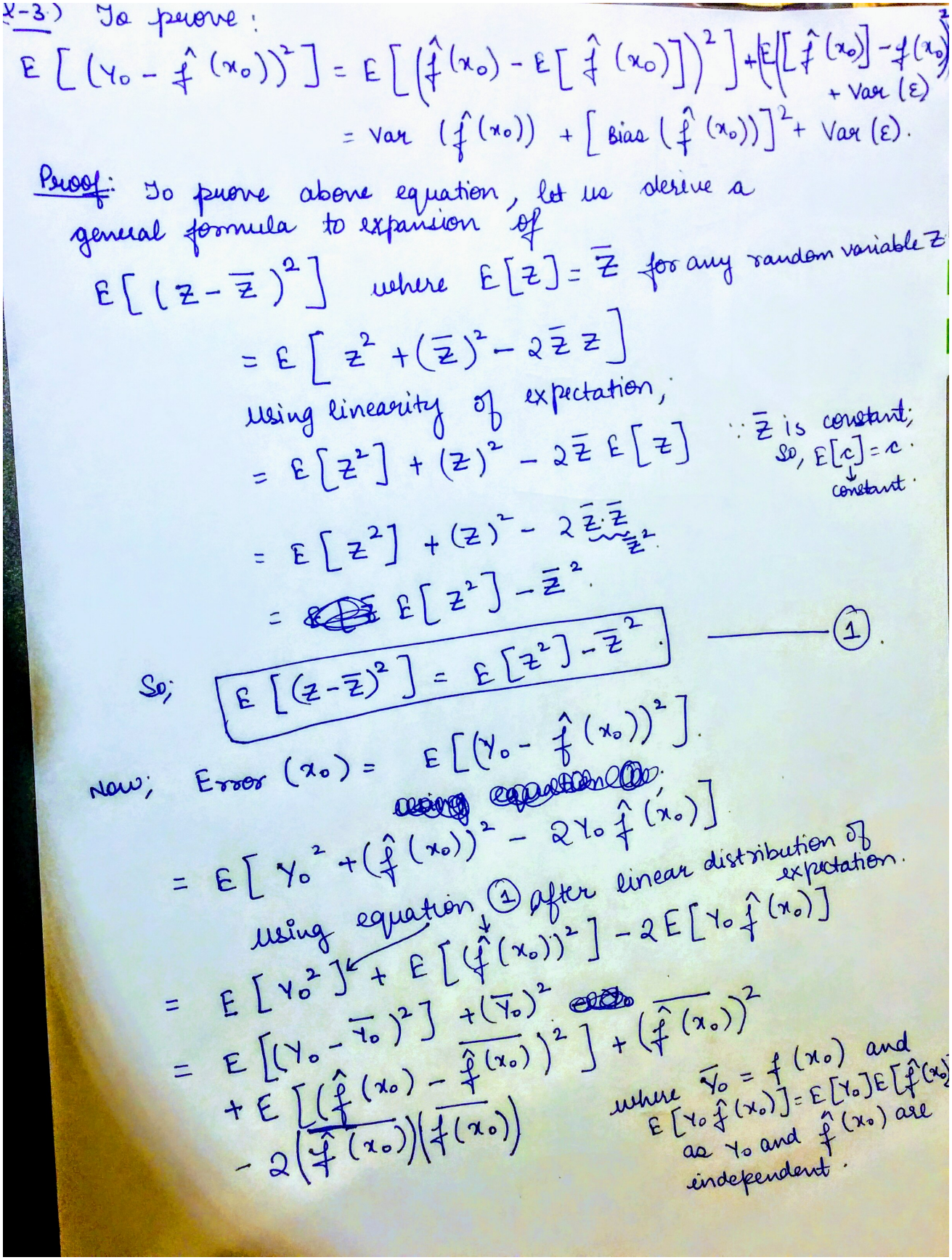
$$\boxed{E[Y] = c}$$

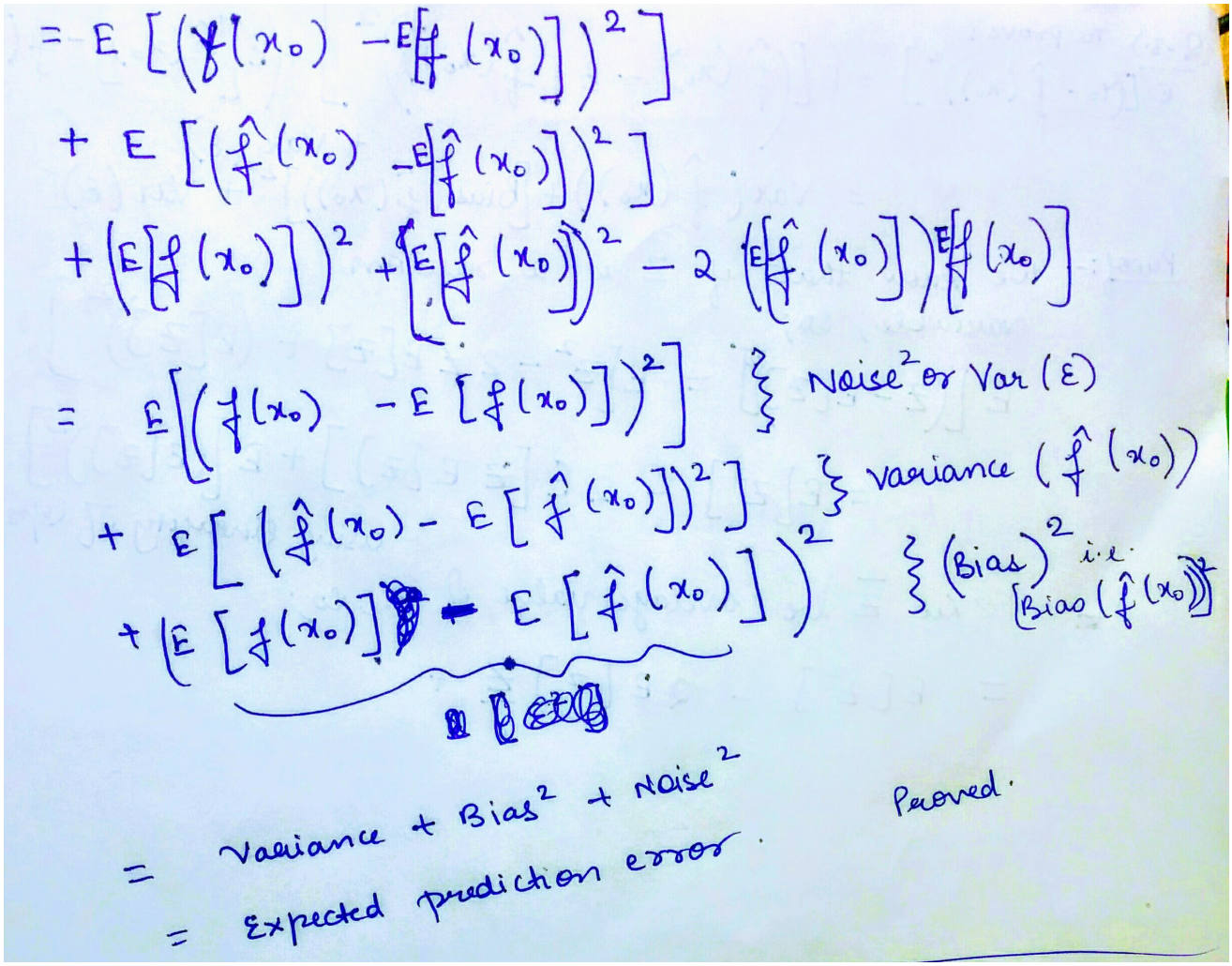
Finding $f''(c)$ to check if its minima,

$f''(c) = 2 > 0$, so, $E[Y]$ is value of c that minimizes $E\left[(Y-c)^2\right]$ with 'c' constant

Proved.

**Solution 3: Proof (T, 12 Points)**






## Solution 4: Predict the ozone concentration (P, 20 Points)

Part b, c: The 'ozone.Rdata' file contains 3 objects. Let us discuss the structure of each of the three objects: Complete Ozone datatable, Training set and Test set.

**1. Complete Ozone dataset** : The ozone dataset contains 111 observations for 4 variables. It contains 4 columns.

Table 1: A table describing the structure of the complete 'Ozone' dataset which has dimensions of the order 111 X 4 and was recorded in New York(NY)

| Column's Name | Data Type | Range | Mean | SD | Column Description |
|---|---|---|---|---|---|
| ozone | num | 1-168 | 42.0991 | 33.27597 | Ozone concentration in NY |
| radiation | int | 7-334 | 184.8018 | 91.1523 | Solar Radiation in NY |
| temperature | int | 57-97 | 77.79279 | 9.529969 | Daily maximum temperature in NY |
| wind | num | 2.3-20.7 | 9.938739 | 3.559218 | Wind speed in NY |

**2. Training Set** : There are 80 observations in the trainset. The 'trainset' contains 80 row-indices values where the index values are obtained from the ozone dataset. So, the training dataset has dimensions 80 X 4.

Table 2: A table describing the structure of the training set

| Column's Name | Data Type | Range | Mean | SD | Column Description |
|---|---|---|---|---|---|
| ozone | num | 1-168 | 41.2 | 35.97235 | Ozone concentration in NY |
| radiation | int | 8-334 | 180.225 | 92.9983 | Solar Radiation in NY |
| temperature | int | 57-94 | 76.7625 | 9.594624 | Daily maximum temperature in NY |
| wind | num | 2.3-20.7 | 9.93125 | 3.613077 | Wind speed in NY |

**3. Test Set** : There are 31 observations in the testset. The 'testset' contains 31 row-indices values where the index values are obtained from the ozone dataset. So, the test dataset has dimensions 31 X 4.

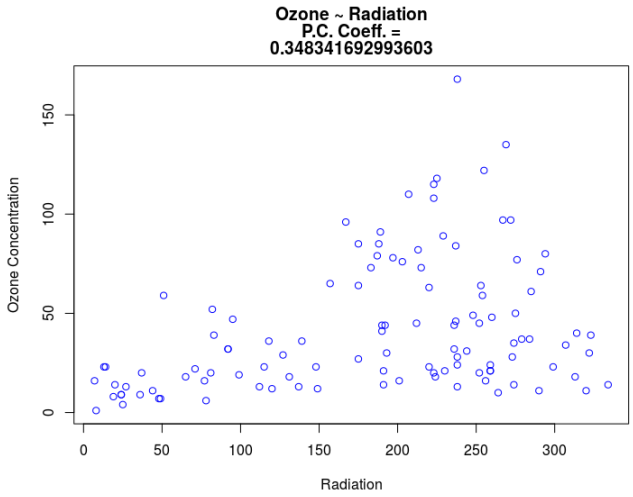
Table 3: A table describing the structure of the Test set

| Column's Name | Data Type | Range | Mean | SD | Column Description |
|---|---|---|---|---|---|
| ozone | num | 8-97 | 44.41935 | 25.39262 | Ozone concentration in NY |
| radiation | int | 7-323 | 196.6129 | 86.54774 | Solar Radiation in NY |
| temperature | int | 61-97 | 80.45161 | 8.969722 | Daily maximum temperature in NY |
| wind | num | 4.6-20.1 | 9.958065 | 3.474553 | Wind speed in NY |

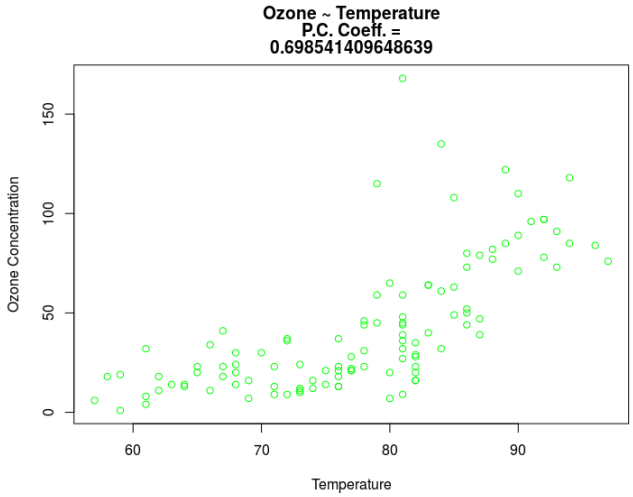
Observation: We observe that the mean of the training-set and the test set is alomost similar to that of the comple ozone data table. Thus, the division of the data table in a ratio of 80:20 for training to test is sufficient.

Part d.

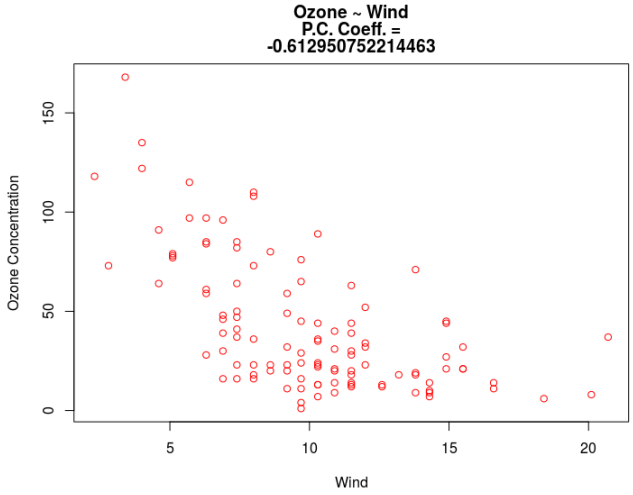
If there are two continuous quantitative variables, we can calculate the Pearson's Correlation Coefficient to comment on the strength of association between the two variables. The range of this coefficient value is from -1 to +1. The Pearson's Correlation Coefficient should be calculated only when the scatter plot of the two variables look linear otherwise not. If there is no linear relationship between the two variables, the Pearson's Correlation Coefficient will be 0. A positive Pearson's Correlation Coefficient value indicate that both variables either increase or decrease together whereas a negative Pearson's Correlation Coefficient value indicate that as value of one variable increase, the value of another variable decreases. I have also used t-test to confirm if the Pearson's Correlation Coefficient is significant or not by calculating the p-value and using the threshold of 0.05. With significance test, we may be able to conclude if the indication o fassociation between the two variables is significant or not. The nearer the scatter plot of points to the straight line with slope +/-1, the higher the strength of association between the variables. Below are the scatter plot from the 'ozone' data-set for each pair of variables along with the corresponding Pearson's Correlation Coefficient Values:



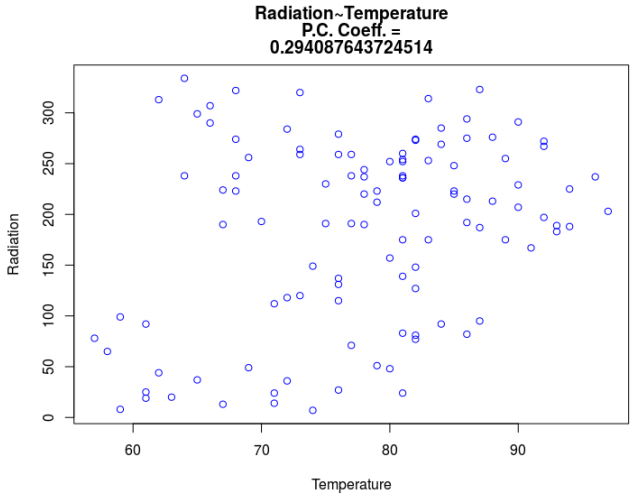

Observation: From the above plot, it is visible that values of ozone concentration and values of radiation increase together but the increase in ozone concentration is much more than that of radiation.



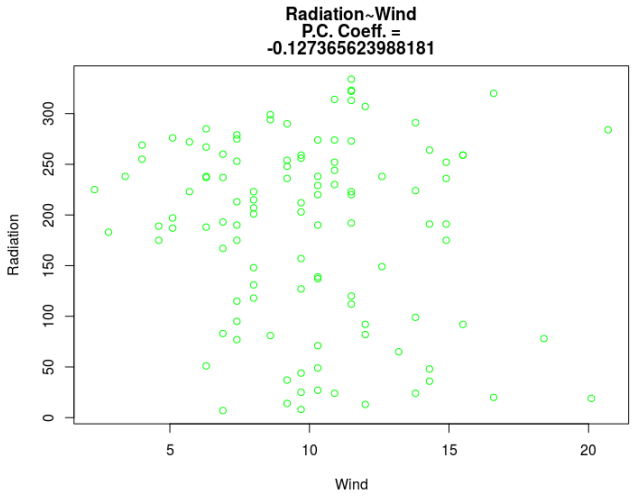

Observation: From the above plot, it is visible that values of ozone concentration and values of temperature increase together but the association is much more closer as compared to the plot between Ozone Concentration $\sim$ Radiation. Thus, the Pearson's Correlation Coefficient is more positive here as that of Ozone $\sim$ Radiation. The growth scale value is almost similar for the two variables here.



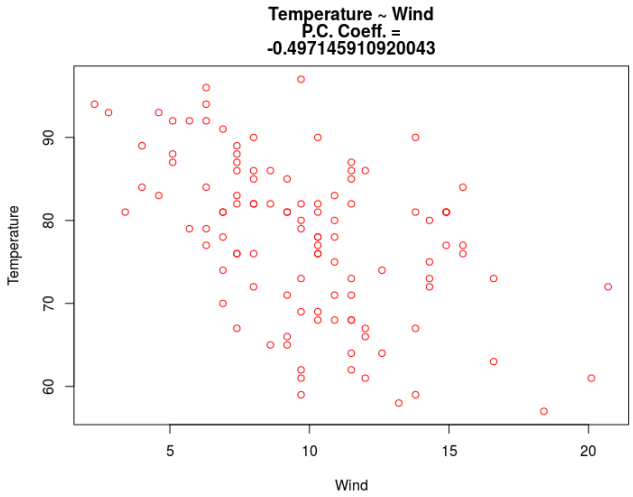

Observation: From the above plot, it is visible that values of ozone concentration increases while the values of wind decreases. However, this negative correlation is much more closer as compared to ozone $\sim$ radiation.




Observation: There is no linear relation between the radiation and temperature and that is why, the Pearson's Correlation Coefficient is almost 0.




Observation: There is no linear relation between the radiation and wind and that is why, the Pearson's Correlation Coefficient is almost 0.

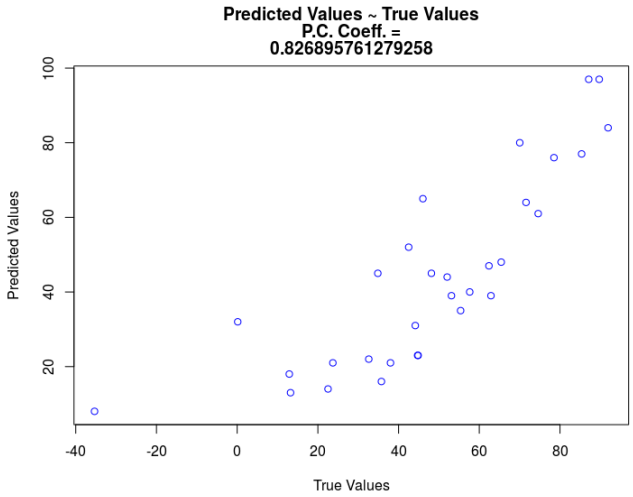



Observation: The Pearson's Correlation Coefficient value between Temperature $\sim$ Wind is -0.49 which means the value of Temperature increases while the value of wind decreases and vice-versa.

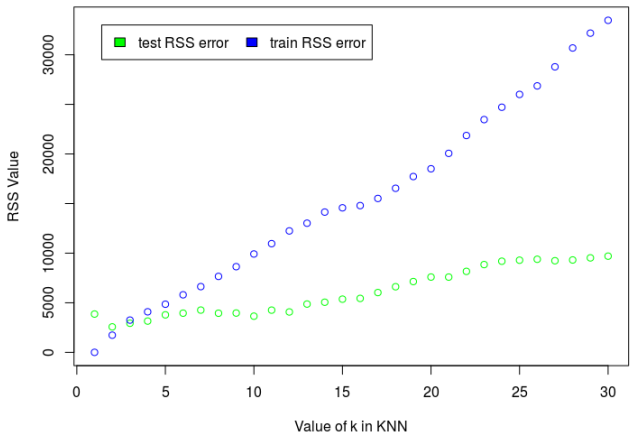
So, we can conclude that on the ozone concentration, there is association with temperature, radiation and wind however, the last three variables are independent of each other. So, if I were to predict the ozone concentration, then, I can use temperature, radiation and wind as my features or predictors if ozone concentration is my target variable. In particular, the features are not associated with each other. However, in the plot between temperature $\sim$ wind, there is a strong negative correlation between the two features.
Note: Correlation does not imply causality. So, I cannot comment on what factor causes what. Also, here the p-value from the Student-T test is less than 0.05 for all the above variables, so, the result is significant in all the above cases. However, I have normalised the training set to perform Student t-test.

Part e: For Function for Residual sum of squares, please see the attached code file.

Part f:
To train the linear model, it is important to normalize the features column-wise The value of RSS for the model ozone $\sim$ radiation + temperature + wind is 8208.509 (The linear model was trained on column-wise normalized training data) and the value of Pearson's coefficient is 0.8266351 as shown in plot below. Clearly, the true value and the predicted value have value close positive correlation.




Part g: KNN classification
The plot obtained by plotting the RSS values for training and testing data using the KNN algorithm is given below:




On which side of the graph do you have the most complex models? Argue with the bias-variance tradeoff.
The model is more complex on left side of the plot for lower value of k. For lower value of k, there will be higher variance and less bias while for larger value for k, there will be lower variance but high bias. With low value of k, each example of the training set is potentially the center of an area predicting classes with most of its neighbors the center of an area predicting the other classes. Thus, the complexity is high as there are more center of an area predicting the classes. If we increase k, the areas predicting each class will be more "smoothed", since it's

the majority of the k-nearest neighbours which decide the class of any point. Thus the areas will be of lesser number, larger sizes and probably simpler shapes, thus "less complexity".
Our target is to lessen the test error. As per the plot, we can see that for k¿3, the performance is giving higher test error i.e. the green plot keeps on showing the increase in the test error. So, I will take k=3 as the value of k.
No, the KNN algorithm does not assume any distribution on the underlying data as it is just based on the calculation of the distance metric. We are not training any model here rather, relying on the distance metric. Thus, no distribution is assumed here which is an advantage of using KNN.

Part h: The value for RSS for the linear model is 8208.509 and the value for RSS for 3-nearest neighbour model is 2931.889 which is much lesser than that of the linear model. The value of correlation coefficient for linear model is 0.8268958 and for 3-nearest neighbour model is 0.9268312 which is very close to the perfect value of 1. I will prefer using 3NN model rather than linear model as the value of RSS is much less for 3NN than that of the linear model, so the prediction quality is better here. Also, there are no underlying assumption for distribution for implementing KNN which is an advantage over linear model in which we assume that the different features share a linear relationship. 3NN is purely based on distance metric used. The degrees of freedom of linear model is 2 while that of the KNN model is approximately N/K i.e $31/3 \sim 13$. Thus, 3NN model has higher degree of freedom which makes it more preferable.