MSCA31010: Linear & Non-Linear Models

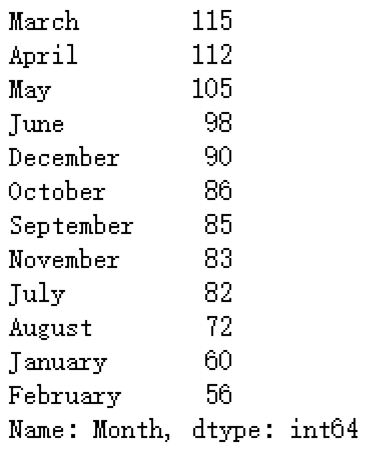
Winter 2022 Assignment 1

Question 1 (50 points)

You will use this linear model Weight ~ Intercept + Month + DayOfWeek to study the effect of Month and DayOfWeek on Weight. You are required to implement the SWEEP Operator in Python.

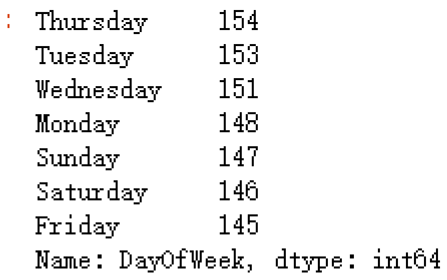
(a) (5 points). Provide a frequency table for the Month, and another frequency table for the DayOfWeek.

frequency table for the Month



```
March        115
April        112
May          105
June          98
December      90
October       86
September     85
November      83
July          82
August        72
January       60
February      56
Name: Month, dtype: int64
```

frequency table for the DayOfWeek



```
: Thursday     154
  Tuesday      153
  Wednesday    151
  Monday       148
  Sunday       147
  Saturday     146
  Friday       145
  Name: DayOfWeek, dtype: int64
```

(b) (5 points). What is the Residual Sum of Squares for this model Weight ~ Intercept? Give your answer using the ".7E" scientific notation.

Residual Sum of Squares = 22360.2295019=2.2360230e+04

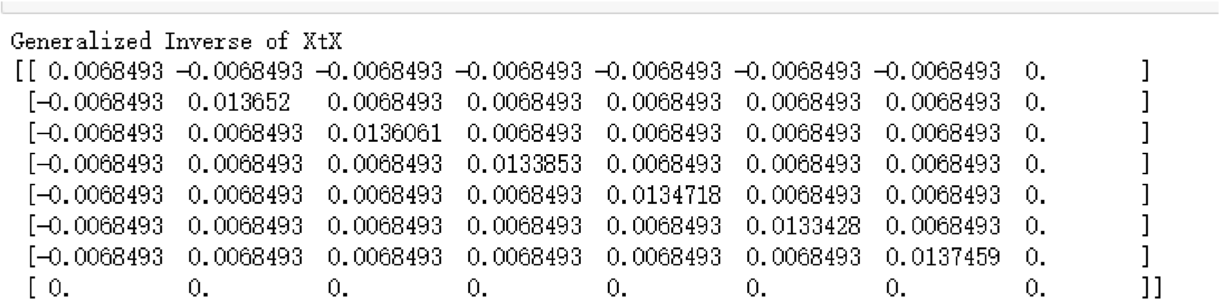
(c) (5 points). What is the Residual Sum of Squares for this model Weight ~ Intercept + Month? Give your answer using the ".7E" scientific notation.

Residual Sum of Squares = 17776.054172=1.7776054e+04

(d) (5 points). What is the Residual Sum of Squares for this model Weight ~ Intercept + DayOfWeek? Give your answer using the ".7E" scientific notation.

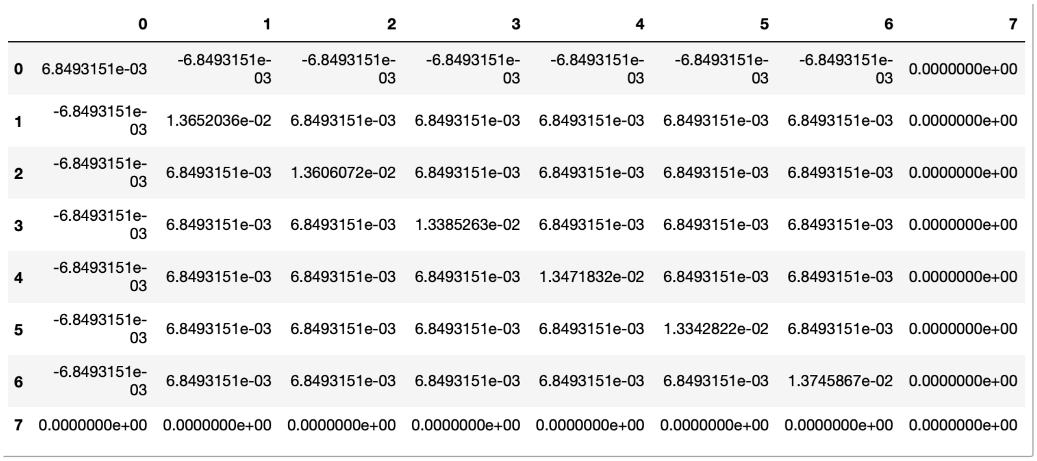
Residual Sum of Squares = 22239.1704454=2.2239170e+04

(e) (5 points). What is the generalized inverse that the SWEEP Operator gives for this model Weight ~ Intercept + DayOfWeek? Give your answer using the ".7E" scientific notation.



```
Generalized Inverse of XtX
[[ 0.0068493 -0.0068493 -0.0068493 -0.0068493 -0.0068493 -0.0068493 -0.0068493  0.       ]
 [-0.0068493  0.013652   0.0068493  0.0068493  0.0068493  0.0068493  0.0068493  0.       ]
 [-0.0068493  0.0068493  0.0136061  0.0068493  0.0068493  0.0068493  0.0068493  0.       ]
 [-0.0068493  0.0068493  0.0068493  0.0133853  0.0068493  0.0068493  0.0068493  0.       ]
 [-0.0068493  0.0068493  0.0068493  0.0068493  0.0134718  0.0068493  0.0068493  0.       ]
 [-0.0068493  0.0068493  0.0068493  0.0068493  0.0068493  0.0133428  0.0068493  0.       ]
 [-0.0068493  0.0068493  0.0068493  0.0068493  0.0068493  0.0068493  0.0137459  0.       ]
 [ 0.         0.         0.         0.         0.         0.         0.         0.       ]]
```

Transform to 7E dataframe：



	0	1	2	3	4	5	6	7
0	6.8493151e-03	-6.8493151e-03	-6.8493151e-03	-6.8493151e-03	-6.8493151e-03	-6.8493151e-03	-6.8493151e-03	0.0000000e+00
1	-6.8493151e-03	1.3652036e-02	6.8493151e-03	6.8493151e-03	6.8493151e-03	6.8493151e-03	6.8493151e-03	0.0000000e+00
2	-6.8493151e-03	6.8493151e-03	1.3606072e-02	6.8493151e-03	6.8493151e-03	6.8493151e-03	6.8493151e-03	0.0000000e+00
3	-6.8493151e-03	6.8493151e-03	6.8493151e-03	1.3385263e-02	6.8493151e-03	6.8493151e-03	6.8493151e-03	0.0000000e+00
4	-6.8493151e-03	6.8493151e-03	6.8493151e-03	6.8493151e-03	1.3471832e-02	6.8493151e-03	6.8493151e-03	0.0000000e+00
5	-6.8493151e-03	6.8493151e-03	6.8493151e-03	6.8493151e-03	6.8493151e-03	1.3342822e-02	6.8493151e-03	0.0000000e+00
6	-6.8493151e-03	6.8493151e-03	6.8493151e-03	6.8493151e-03	6.8493151e-03	6.8493151e-03	1.3745867e-02	0.0000000e+00
7	0.0000000e+00	0.0000000e+00	0.0000000e+00	0.0000000e+00	0.0000000e+00	0.0000000e+00	0.0000000e+00	0.0000000e+00

(f) (5 points). What is the Residual Sum of Squares for this model Weight ~ Intercept + Month + DayOfWeek? Give your answer using the “.7E” scientific notation.

Residual Sum of Squares = 17665.5659525=1.7665566e+04

(g) (5 points). Which model yields the smallest Residual Sum of Squares?

Weight ~ Intercept + Month + DayOfWeek

(h) (5 points). How many regression parameters (including the aliased parameters) are in this model Weight ~ Intercept + Month + DayOfWeek?

len(aliasParam)+len(nonAliasParam)=20

There are 20 regression parameters.

(i) (10 points). What are the regression coefficients (including the aliased parameters) of this model Weight ~ Intercept + Month + DayOfWeek? Give your answer using the “.7E” scientific notation.

Parameter Estimates

Intercept	2.1122573e+02
Month_January	-4.4054252e+00
Month_February	-4.8239659e+00
Month_March	-2.8253761e+00

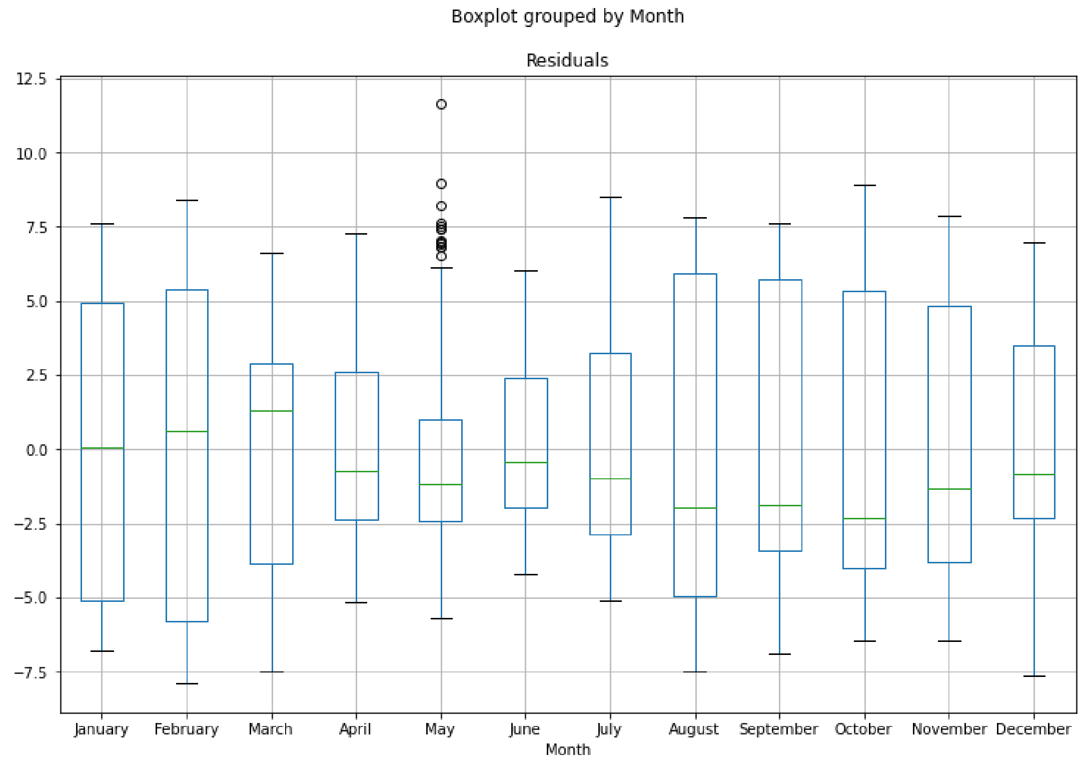
```
Month_April         -4.2731582e+00
Month_May           -6.4393219e+00
Month_June          -7.1583571e+00
Month_July          -7.1115251e+00
Month_August        -4.8223766e+00
Month_September     -4.0327408e+00
Month_October       -3.3379674e+00
Month_November      -1.5751820e+00
Month_December       0.0000000e+00
DayOfWeek_Sunday     3.9628664e-01
DayOfWeek_Monday     6.6924303e-01
DayOfWeek_Tuesday    7.3832484e-02
DayOfWeek_Wednesday -1.9775855e-01
DayOfWeek_Thursday  -3.0878554e-01
DayOfWeek_Friday    -1.7015053e-01
DayOfWeek_Saturday   0.0000000e+00
dtype: float64
```

Question 2 (50 points)

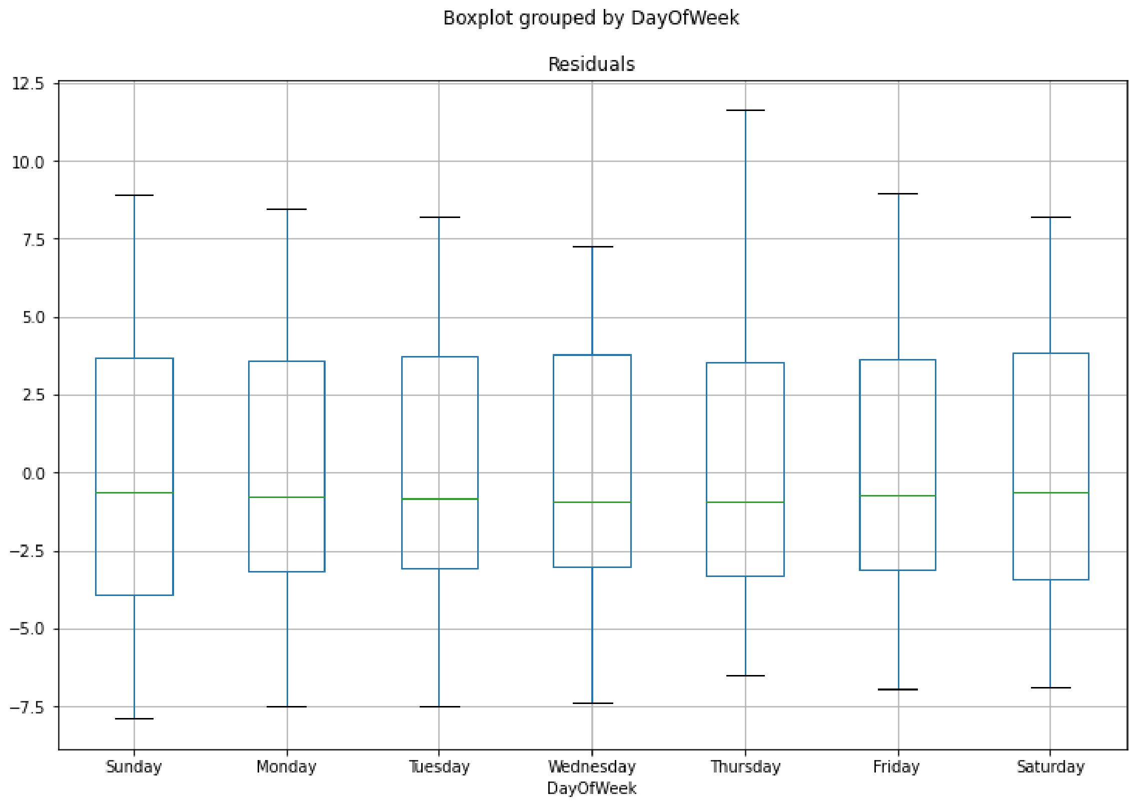
Let us focus on this model Weight ~ Intercept + Month + DayOfWeek.

a) (10 points). Generate a Boxplot of the residuals versus Month. The residuals are on the vertical axis and the Month categories are on the horizontal axis. Also, generate another Boxplot of the residuals versus DayOfWeek. Comment on the evidence of heteroskedasticity of the residuals.

Boxplot of the residuals versus Month

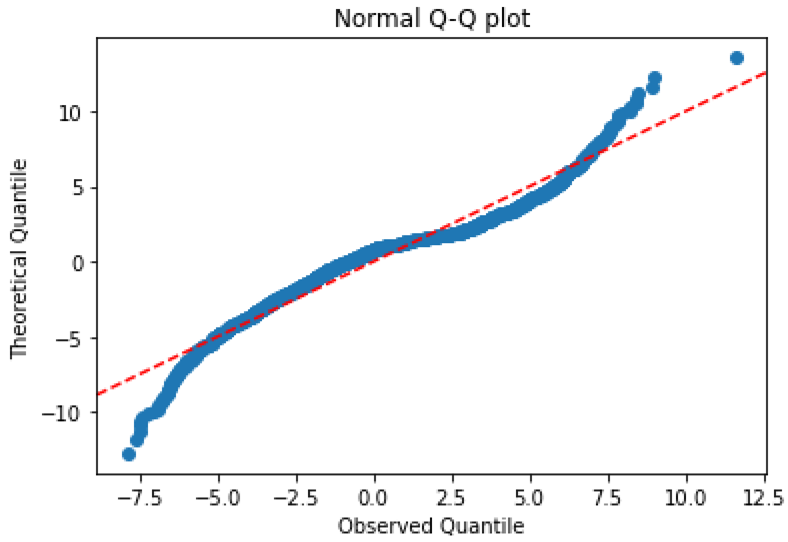


Boxplot of the residuals versus DayOfWeek



From Month graph, there is evidence of heteroskedasticity since the variance are not equal. From Dayofweek graph, there is no evidence of heterokedasticity since the variance are similar.

b) (10 points). Calculate the Anderson-Darling Test statistic and generate a Normality Q-Q Plot for the residuals. Comment on the evidence of normality (or non-normality) of the residuals.



Anderson Test = 15.636691540376432

Critical Values = [0.574 0.654 0.784 0.915 1.088]

p-values = [0.15 0.1   0.05 0.025 0.01]

The anderson test statistic is greater than the critical values, so we can reject the normality assumption. Meanwhile, from the q-q plot, we can see that the points are not lying on the line.There is evidence of non normality.

c) (10 points). Perform the Breusch-Pagan Test and the White Test of Heteroskedasticity. Provide the Chi-square statistics, the degrees of freedom, and the significance values. Comment on the evidence of non-homogenous variance.

Breusch-Pagan Test

test statistic: 209.35170433060833

df: 17

p-value: 3.7450745860152025e-35<0.05

The p-value is close to zero, so there is evidence of non-homogenous variance

White Test of Heteroskedasticity

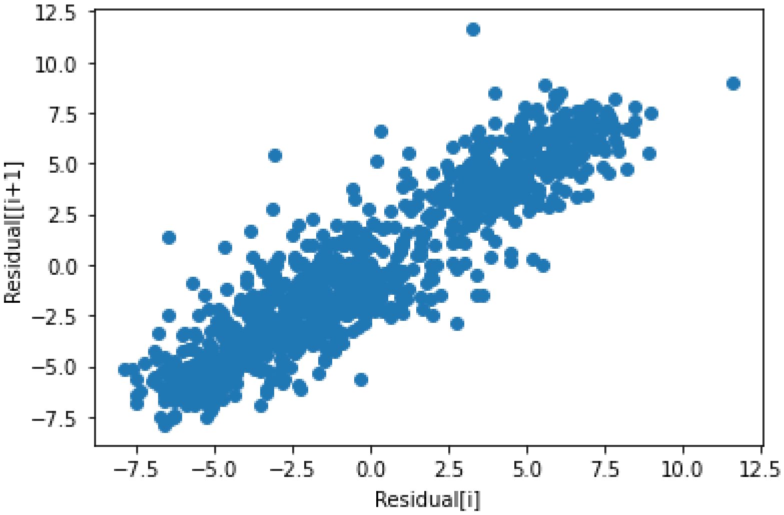
test statistic: 233.2818239759793

df: 85

p-value: 3.3282382921650404e-16<0.05

The p-value is close to zero, so there is evidence of non-homogenous variance

d) (10 points). Calculate the Durbin-Watson Test statistic. Comment on the evidence of autocorrelation among observations.



Autocorrelation = 0.9199539086078394

Durbin-Watson Test 0.16007798648747495

There is positive autocorrelation among observations since Durbin Watson test is in the range of 0 and 1

e) (10 points). Calculate the Shapley values of the two predictors Month and DayOfWeek. Also, provide the Percent Shapley values of the two predictors. Among these two predictors, which one influences the weight more?

Shapley value of Month=0.20477831=2.0477831e-01

Shapley value of DayofWeek=0.0051776588=5.1776588e-03

Percent Shapley value of Month=97.5339%

Percent Shapley value of DayofWeek=2.46607%

Months influence the weight more.