1. Time-Series Forecasting (Adapted from Problem 2.2 in Snyder & Shen) (20 points)

The demand for a new brand of dog food has been steadily rising at the local PetMart store. The previous 26 weeks' worth of demand (number of bags) are given in the file 'dog_food.xlsx'

Use the following methods to forecast next week's demand: (2 points) Naïve forecast

Week	Demand	Forecast
1	646	
2	683	646
3	708	683
4	761	708
5	787	761
6	809	787
7	856	809
8	892	856
9	944	892
10	991	944
11	1034	991
12	1091	1034
13	1123	1091
14	1144	1123
15	1164	1144
16	1186	1164
17	1231	1186
18	1255	1231
19	1298	1255
20	1337	1298
21	1389	1337
22	1436	1389
23	1490	1436
24	1528	1490
25	1555	1528
26	1613	1555
27		1613

(2 points) Average forecast

Week	Demand	Forecast
1	646	
2	683	646.00
3	708	664.50
4	761	679.00
5	787	699.50
6	809	717.00
7	856	732.33
8	892	750.00
9	944	767.75
10	991	787.33
11	1034	807.70
12	1091	828.27
13	1123	850.17
14	1144	871.15
15	1164	890.64
16	1186	908.87
17	1231	926.19
18	1255	944.12
19	1298	961.39
20	1337	979.11
21	1389	997.00
22	1436	1015.67
23	1490	1034.77
24	1528	1054.57
25	1555	1074.29
26	1613	1093.52
27		1113.50

(3 points) Moving average method with time window n = 3

Week	Demand	Forecast
1	646	
2	683	
3	708	
4	761	679.00
5	787	717.33
6	809	752.00
7	856	785.67
8	892	817.33
9	944	852.33
10	991	897.33
11	1034	942.33
12	1091	989.67
13	1123	1038.67
14	1144	1082.67
15	1164	1119.33
16	1186	1143.67
17	1231	1164.67
18	1255	1193.67
19	1298	1224.00
20	1337	1261.33
21	1389	1296.67
22	1436	1341.33
23	1490	1387.33
24	1528	1438.33
25	1555	1484.67
26	1613	1524.33
27		1565.33

(3 points) Weighted average method with weights (0.5, 0.3, 0.2)

Week	Demand	Forecast
1	646	
2	683	
3	708	
4	761	669.50
5	787	706.10
6	809	739.70
7	856	778.40
8	892	807.40
9	944	839.70
10	991	884.40
11	1034	927.40
12	1091	976.10
13	1123	1023.90
14	1144	1068.90
15	1164	1111.20
16	1186	1137.50
17	1231	1158.40
18	1255	1184.00
19	1298	1213.30
20	1337	1251.60
21	1389	1284.30
22	1436	1327.90
23	1490	1372.40
24	1528	1423.30
25	1555	1470.60
26	1613	1514.40
27		1553.10

(5 points) Exponential Smoothing with α = 0.4

Week	Demand	Forecast
1	646	646
2	683	646.00
3	708	660.80
4	761	679.68
5	787	712.21
6	809	742.12
7	856	768.87
8	892	803.72
9	944	839.03
10	991	881.02
11	1034	925.01
12	1091	968.61
13	1123	1017.56
14	1144	1059.74
15	1164	1093.44
16	1186	1121.67
17	1231	1147.40
18	1255	1180.84
19	1298	1210.50
20	1337	1245.50
21	1389	1282.10
22	1436	1324.86
23	1490	1369.32
24	1528	1417.59
25	1555	1461.75
26	1613	1499.05
27		1544.63

(5 points) Report the MSE, MAD and MAPE for the predictions in (a) made for week 4 – 26.

{Kindly refer to attached Excel file.}

2. Linear Trend Model for US Gasoline Expenditure (20 points)

The US gasoline expenditure from 1954 – 2004 has been recorded in 'gas.csv'. You can find the description of each data column in the following table. Using a simple linear regression (assuming $D=\beta 0+\beta 1 YEAR+\epsilon$), forecast the gasoline expenditure. Use data from years 1954 – 1993 as training data and years after 1993 as test data.

(5 points) Report your forecast for the gasoline expenditure in 2005.

{Kindly refer to atached R file titled Question2.R}

(5 points) Report the estimated values for $\beta 0$ and $\beta 1$. How do you interpret these two values?

$\beta 0$ = 6098.0385

$\beta 0$ represents the intercept of the regression line. It is the value of the dependent variable (gasoline expenditure) when the independent variable (Year) is zero.

$\beta 1$ = 3.1147

$\beta 1$ is the slope of regression line.

(5 points) Report the R2 and OSR2. How do you interpret their meaning?

R2 = 0.8898 (The coefficient of determination for the training data.)

OSR2 = 0.896 (The performance of the model on the test data.)

(5 points) Report the p-value associated with $\beta 1$. How do you interpret this p-value?

$\beta 1$ P-Value = <2e-16. P-value is the level of statistical significance.

Smaller p-values are better($p<0.05$ to be significant).

3. Multiple Linear Regression for US Gasoline Expenditure (40 points)

The US gasoline expenditure from 1954 – 2004 has been recorded in 'gas.csv'. You can find the description of each data column in the table above. Use a multiple linear regression, assuming $D=\beta 0+\beta 1YEAR+\beta 2POP+\beta 3GASP+\beta 4INCOME+\beta 5PNC+\beta 6PUC+\epsilon$,

forecast the gasoline expenditure. Use data from years 1954 – 1993 as training data and years after 1993 as test data.

(5 points) Report the R2 and OSR2. What can you imply from the R2 and OSR2?

R2 = 0.987 (The coefficient of determination for the training data.)

OSR2 = 0.871 (The performance of the model on the test data.)

(5 points) Report the estimated coefficient associated with POP. Does the sign of the coefficient make sense to you? What can you imply from it?

Estimated coefficient = -1.404e-03

A negative sign on the coefficient of POP would suggest that as the population increases, gasoline expenditure decreases.

Implication: A negative coefficient could imply that the rise in population has been accompanied by shifts in behavior or technology that reduce gasoline dependence.

(5 points) Report the VIFs for all the variables. What can you imply from the VIFs?

YEAR	POP	GASP	INCOME	PNC	PUC
3838.48138	1887.96693	13.19133	335.44178	68.99352	55.49170

Variance Inflation Factor (VIF) is a measure used in regression analysis to detect multicollinearity, which occurs when independent variables are highly correlated with each other.

High multicollinearity can make it difficult to determine the individual effect of each variable on the dependent variable, leading to unreliable estimates of the coefficients.

(20 points) Build a better model using feature selection. Show, step by step, which variable you removed and why you chose to remove it.

{Kindly refer to the attached R file titled Question3.R}

(5 points) Report your final model and the final R2 and OSR2.

R2 = 0.973

OSR2 = 0.948