Team Number: #5

Team Captain: Sriharsha Aitharaju

Team Members: Daniel Rodriguez

Fernando Sosa

Kzzy Centeno

Robert Law

Activity on

PART I: (15 Points):

Problem 1.1 (5 Points) The response variable of the observed data and the fitted prediction are listed in the following table.

Response (Y)	Model I Prediction ($\hat{Y}_1$)	Model II Prediction ($\hat{Y}_2$)
3	3.2	3.3
4	4.3	4.2
5	4.9	4.8
6	5.7	5.9
7	6.9	7.1

1. Calculate the sum squared of error of Model I and Model II.

Response (Y)	Model I Prediction ($\hat{Y}_1$)	Difference	(Difference)^2
3	3.2	-0.2	0.04
4	4.3	-0.3	0.09
5	4.9	0.1	0.01
6	5.7	0.3	0.09
7	6.9	0.1	0.01
		SSE1	0.06

$$SSE = \frac{(\sum(Y-\hat{Y}_1))^2}{N-1}$$

Response (Y)	Model II Prediction ($\hat{Y}_1$)	Difference	(Difference)^2
3	3.3	-0.3	0.09
4	4.2	-0.2	0.04
5	4.8	0.2	0.04
6	5.9	0.1	0.01
7	7.1	-0.1	0.01
		SSE2	0.0475

2. Calculate the average squared error of Model I and Model II.

Response (Y)	Model I Prediction ($\hat{Y}_1$)	Difference	(Difference)^2	
3	3.2	-0.2	0.04	
4	4.3	-0.3	0.09	
5	4.9	0.1	0.01	
6	5.7	0.3	0.09	
7	6.9	0.1	0.01	$ASE = \frac{(\Sigma(Y-\hat{Y}_1))^2}{N}$
		ASE1	0.048	

Response (Y)	Model II Prediction ($\hat{Y}_1$)	Difference	(Difference)^2
3	3.3	-0.3	0.09
4	4.2	-0.2	0.04
5	4.8	0.2	0.04
6	5.9	0.1	0.01
7	7.1	-0.1	0.01
		ASE2	0.038

3. Calculate both R_I^2 and R_{II}^2.

Response (Y)	Model I Prediction ($\hat{Y}_1$)	Difference	(Difference)^2	Y-Ybar	
				4	
3	3.2	-0.2	0.04	1	
4	4.3	-0.3	0.09	0	
5	4.9	0.1	0.01	1	
6	5.7	0.3	0.09	4	
7	6.9	0.1	0.01	10	0.976
5		(R^2)1	0.24		

Response (Y)	Model II Prediction ($\hat{Y}_1$)	Difference	(Difference)^2	Y-Ybar	
				4	
3	3.3	-0.3	0.09	1	
4	4.2	-0.2	0.04	0	
5	4.8	0.2	0.04	1	
6	5.9	0.1	0.01	4	
7	7.1	-0.1	0.01	10	0.981
5		(R^2)2	0.19		

4. Calculate both $MAPE_I$ and $MAPE_{II}$

Response (Y)	Model I Prediction ($\hat{Y}_1$)	Difference	Difference/Response
3	3.2	-0.2	-0.066666667
4	4.3	-0.3	-0.075
5	4.9	0.1	0.02
6	5.7	0.3	0.05
7	6.9	0.1	0.014285714
5		MAPE1	-0.01147619

Response (Y)	Model II Prediction ($\hat{Y}_1$)	Difference	Difference/Response
3	3.3	-0.3	-0.1
4	4.2	-0.2	-0.05
5	4.8	0.2	0.04
6	5.9	0.1	0.016666667
7	7.1	-0.1	-0.014285714
5		MAPE2	-0.02152381

5. Calculate both MAE_I and MAE_{II}

Measure	Model I	Model II
SSE	**0.06**	**0.0475**
ASE	**0.048**	**0.038**
R^2	**0.976**	**0.981**
MAPE	**1.15%**	**2.15%**
MAE	**0.2**	**0.18**

Problem 1.2 (10 Points) Work on Problem 1, Problem 2, and Problem 3 in the Textbook (Chapter 5 on Page 219)

1. Using basic statistical properties of the variance, as well as single-variable calculus, derive (5.6). In other words, prove that α given by (5.6) does indeed minimize $\mathrm{Var}(\alpha X + (1-\alpha)Y)$.

$$\alpha = \frac{\sigma_Y^2 - \sigma_{XY}}{\sigma_X^2 + \sigma_Y^2 - 2\sigma_{XY}},$$

$\sigma^2_X = \mathrm{Var}(X)$; $\sigma^2_Y = \mathrm{Var}(Y)$; $\sigma_{XY} = (X,Y)$

$\mathrm{Var}(\alpha X + (1-\alpha)Y)$

$\rightarrow \alpha^2\sigma^2_X + (1-\alpha)^2\sigma^2_y + 2\alpha(1-\alpha)\sigma_{XY}$

$\rightarrow \frac{d}{d\alpha}\left(\alpha^2\sigma^2_X + (1-\alpha)^2\sigma^2_y + 2\alpha(1-\alpha)\sigma_{XY}\right) = 0$

$\rightarrow 2\alpha\sigma^2 - 2\sigma^2_y + 2\alpha\sigma^2_y + 2\sigma_{XY} - 4\alpha\sigma_{XY} = 0$

$\rightarrow \alpha = \dfrac{\sigma^2_y - \sigma_{XY}}{\sigma^2_X + \sigma_Y - 2\sigma_{XY}}$

2. We will now derive the probability that a given observation is part of a bootstrap sample. Suppose that we obtain a bootstrap sample from a set of n observations.

   (a) What is the probability that the first bootstrap observation is *not* the jth observation from the original sample? Justify your answer.

   (b) What is the probability that the second bootstrap observation is *not* the jth observation from the original sample?

   (c) Argue that the probability that the jth observation is *not* in the bootstrap sample is $(1 - 1/n)^n$.

(d) When $n = 5$, what is the probability that the jth observation is in the bootstrap sample?

(e) When $n = 100$, what is the probability that the jth observation is in the bootstrap sample?

(f) When $n = 10,000$, what is the probability that the jth observation is in the bootstrap sample?

(g) Create a plot that displays, for each integer value of n from 1 to $100,000$, the probability that the jth observation is in the bootstrap sample. Comment on what you observe.

(h) We will now investigate numerically the probability that a bootstrap sample of size $n = 100$ contains the jth observation. Here $j = 4$. We repeatedly create bootstrap samples, and each time we record whether or not the fourth observation is contained in the bootstrap sample.

```
> store <- rep(NA, 10000)
> for(i in 1:10000){
    store[i] <- sum(sample(1:100, rep=TRUE) == 4) > 0
}
> mean(store)
```

Comment on the results obtained.

2.

a) WE USE THE FORMULA

$$1-\frac{1}{n}$$

n = # OF OBSERVATIONS

b) SAME AS ABOVE

$$1-\frac{1}{n}$$

c) NO $\left(1-\frac{1}{n}\right)^n$ WOULD NOT BE

AS $1-\frac{1}{n}$ IS THE CORRECT WAY

d) $n = 5$

$$1-\left(1-\frac{1}{5}\right)^5 = \boxed{0.67}$$

e) $n = 100$

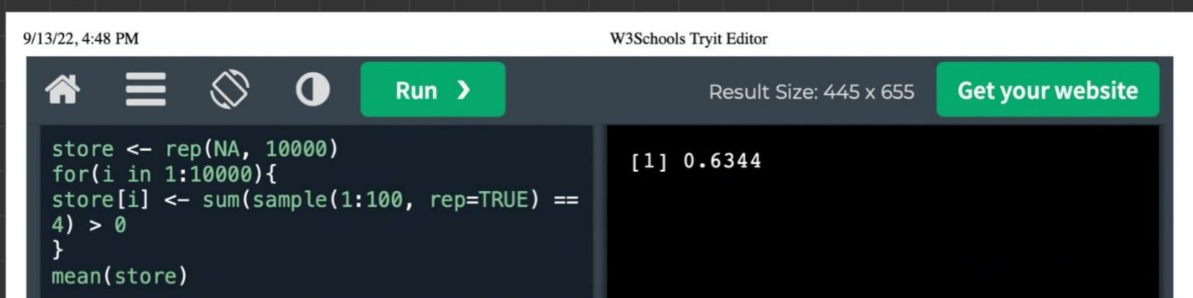
$$1-\left(1-\frac{1}{100}\right)^{100} = \boxed{0.63}$$

f) $n = 10000$

$$1-\left(1-\frac{1}{10000}\right)^{10000} = \boxed{0.63}$$

g)

h)



RESULTS SIMILAR TO f)

3. We now review k-fold cross-validation.

   (a) Explain how k-fold cross-validation is implemented.

   (b) What are the advantages and disadvantages of k-fold cross-validation relative to:

      i. The validation set approach?

      ii. LOOCV?

a) Using *n* for observations and *k* for observations randomly split up in the data set we use the equation *n* / *k*. We can now use the average of *k* to obtain the MSE estimate. The observations we used to split up is used as a validation set.

b) i.

- K-Fold Validation can be applied to almost all learning algorithms
- The computation time for K-Fold Validation is approximately K times of the computation time of "Validation Set Approach" that is acceptable for most learning algorithms
- 5-Fold or 10-fold validation have been shown to yield the test error rate that suffer neither from excessively high bias nor from very high variance.
- Disadvantages are that there may be overestimation when based on a few observations

ii. The LOOCV

- The LOOCV method can have a shorter computation time but in special cases

PART II Programming (15 Points)

Data Used: "House_Prices_PRED.CSV" with three variables: ID, House_Price (observed value), and P_House_Price (Model Predicted Value).

Problem 2.1 (0 Points) Read the CSV file "House_Prices_PRED.CSV"

Problem 2.2 (3 Points) Write a program to calculate the sum squared of error and the average squared error of the Model (i.e., P_House_Price).

Problem 2.3 (3 Points) Write a program to calculate the R^2 of the Model (i.e., P_House_Price).

Problem 2.4 (3 Points) Write a program to calculate the MAPE of the Model (i.e., P_House_Price).

Problem 2.5 (3 Points) Write a program to calculate the MAE of the Model (i.e., P_House_Price).

Problem 2.6 (3 Points) Write a program to produce a residual plot with residual on the Y-axis and observed value (House_Price) and to impose a loess line on the graph.

PART II Programming (15 Points)

Problem 2.1 (0 Points) Read the CSV file "House_Prices_PRED.CSV"

In [ ]:
```
import pandas as pd
import statsmodels.formula.api as smf
houses = pd.read_csv('House_Prices_PRED.csv')
houses = houses.iloc[: , 1:]
houses.head()
```

Out[ ]:

	P_SalePrice	SalePrice
0	206307.7360	208500
1	179044.5328	181500
2	217258.4337	223500
3	161547.6322	140000
4	272594.2471	250000

Problem 2.2 (3 Points) Write a program to calculate the sum squared of error and the average squared error of the Model (i.e., P_House_Price).

In [ ]:
```
import numpy as np

#sum of differences squared
sumDif2 = np.sum((houses['SalePrice'] - houses['P_SalePrice'])**2)
#sum of differences squared divided by # of rows
print("SSE = ",sumDif2)
print("ASE = ",sumDif2/houses.shape[0])
```

```
SSE =  740014639177.1643
ASE =  506859341.9021673
```

Problem 2.3 (3 Points) Write a program to calculate the R2 of the Model (i.e., P_House_Price).

In [ ]:
```
#Y minus YBar Squared
ySubMean2 = np.sum((houses['SalePrice'] - houses['SalePrice'].mean())**2)
#1 minus sum of differences squared divided by y minus ybar squared
print("R^2 = ", (1 - sumDif2/ySubMean2))
```

```
R^2 =  0.9196327362106914
```

Problem 2.4 (3 Points) Write a program to calculate the MAPE of the Model (i.e., P_House_Price).

In [ ]:
```
dif = np.abs(houses['SalePrice'] - houses['P_SalePrice'])
DifdivY = np.sum(np.divide(dif, houses['SalePrice']))
# sum of differences divided by actual value times, divided by number of rows, times 1
print("MAPE = ", (DifdivY*(1/houses.shape[0]) * 100), "%")
```

```
MAPE =  7.026392138631052 %
```

Problem 2.5 (3 Points) Write a program to calculate the MAE of the Model (i.e., P_House_Price).

In [ ]:
```
dif = np.sum(np.abs(houses['SalePrice'] - houses['P_SalePrice']))
# sum of differences divided by number of rows
print("MAE = ", dif*(1/houses.shape[0]))
```

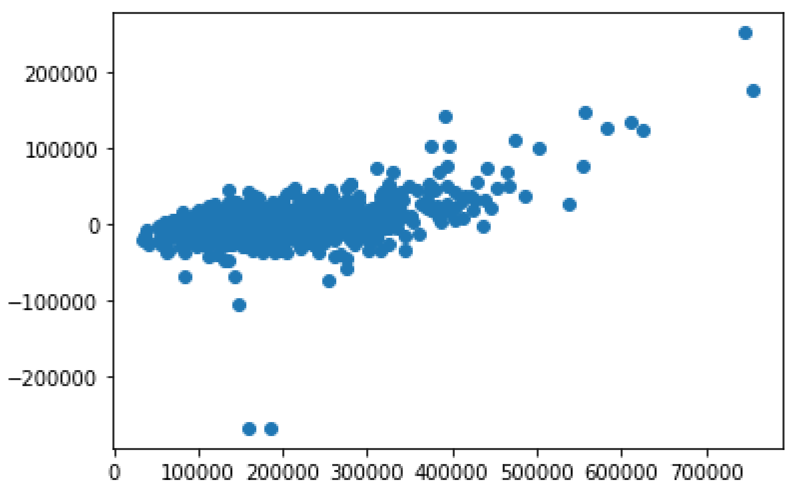
```
MAE =  12470.833673842466
```

Problem 2.6 (3 Points) Write a program to produce a residual plot with residual on the Y-axis and observed value (House_Price) and to impose a loess line on the graph.

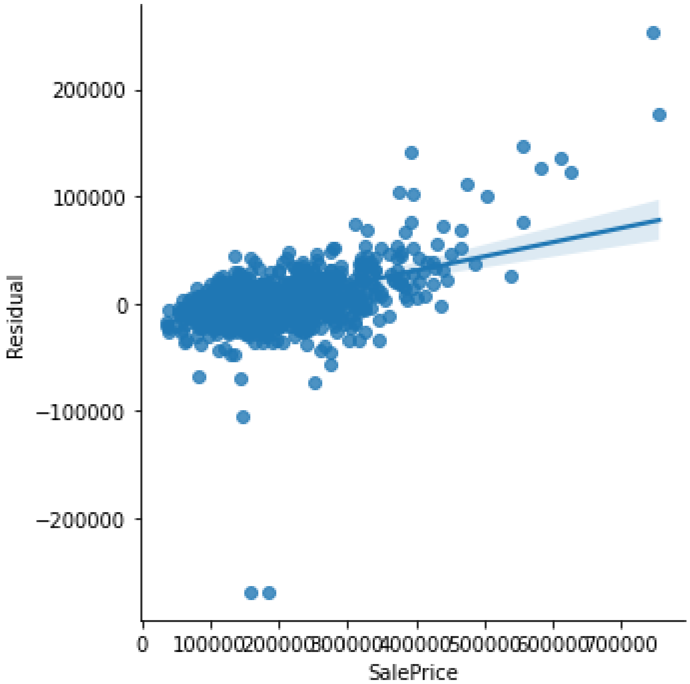
In [17]:
```
from matplotlib import pyplot as plt
import seaborn as sns

houses['Residual'] = houses['SalePrice'] - houses['P_SalePrice']

plt.scatter(houses['SalePrice'], houses['Residual'])
plt.show()
sns.lmplot(x='SalePrice', y='Residual', data=houses)
```



Out[17]:
```
<seaborn.axisgrid.FacetGrid at 0x7f88ae09b290>
```


200000
100000
0
−100000
−200000
Residual
0
100000
200000
300000
400000
500000
600000
700000
SalePrice