KING FAHD UNIVERSITY OF PETROLEUM & MINERALS
DEPARTMENT OF MATHEMATICS AND STATISTICS
Term 162

STAT 212 BUSINESS STATISTICS II
Final Exam
Sunday May 21, 2017

Name:_____ **ID #:**_____ **Srl #:**___

SECTION: 1 (E. AL-SAWI) 2 (R. ANABOSI)

Important Notes:

1) You must **show all work** to obtain full credit for questions on this exam.
2) **DO NOT round** your answers at each step. Round answers only if necessary at your final step to **4 decimal places**.

Question No	Full Marks	Marks Obtained
Q1	15	
Q2	6	
Q3	12	
Q4	8	
Q5	8	
Q6	12	
Q7	14	
Q8	10	
Total	**85**	

Question one (15 Marks) The table below shows the mean price for three types of energy products in the United States from1992 to 2008: Electricity (**$ per 500 kWH**), Natural Gas(**$ per 40 therms**), and Fuel Oil (**$ per gallon**).

Year	Electricity (ELC	Natural Gas (Gas)	Fuel Oil (Oil)	Price index (ELC)	Price index (Gas)	Price index (Oil)	unweighted aggregate price index
1992	44.501	26.376	0.985	100.00%	100.00%	100.00%	100.00%
1993	16.959	28.749	0.969	38.11%	109.00%	98.38%	64.95%
. . .	. . .	. . .	. . .				
2007	60.343	56.297	2.681				
2008	64.172	63.748	3.752				

For year 2008 Calculate: (using 1992 as the base year)

1) **(6 Marks)** The simple price index for electricity, natural gas, and fuel oil, using 1992 as the base year. **(Write the answer in the TABLE)**

(1 Mark) Interpretation for the Fuel Oil price index only:

2) **(2 Mark)** The unweighted aggregate price index for the group of three energy items. **(Write the answer in the TABLE)**

(1 Mark) Interpretation:

3) Laspeyres price index for the group of three energy items for a family that consumed 5,000 kWH of electricity (10 units), 960 therms of natural gas (24 units), and 400 gallons of fuel oil (400 units) in 1992.

(2 Marks) Calculation:

4) Paasche price index for the group of three energy items for a family that consumed 6,500 kWH of electricity, 1,040 therms of natural gas, and 235 gallons of fuel oil in 2008.

(3 Marks) **Calculation**:

Question Two (6 Marks) A study was recently performed in which it attempted to develop a regression model to explain variation in the mileage ratings of new cars. At one stage of the analysis, the estimated regression took the following form: $\hat{y} = 34.2 - .003X_1 + 4.56X_2$,where X_1 = Vehicle weight, and X_2 =0 if standard transmission and X_2 =1 if automatic transmission.

1. The Interpretation of the regression coefficient for
   (i) **(2 Marks)** X_1 is:

   (ii) **(2 Marks)** X_2 is:

2. **(2 Marks)** Cadillac's 2000 with automatic transmission weights 4012 pounds. Provide an estimate of the average highway mileage you would expect to obtain from this model.

Question Three (12 Marks) Are gasoline prices higher during the height of the summer vacation season than at other times? A study on the mean monthly prices (in dollars per gallon) for unleaded gasoline in the United States from January 2006 to April 2010 resulted in the following output.

The regression equation is:

Log(Price) = 0.886 - 0.00001 Months - 0.048 Jan - 0.021 Feb + 0.003 Mar + 0.078 Apr + 0.153 May + 0.234 Jun + 0.246 Jul + 0.226 Aug + 0.194 Sep + 0.140 Oct + 0.067 Nov

Predictor	Coef	SE Coef	T	P
Constant	0.8864	0.1107	8.01	0.000
Months	-0.000015	0.001815	-0.01	0.994
Jan	-0.0480	0.1310	-0.37	0.716
Feb	-0.0211	0.1309	-0.16	0.873
Mar	0.0030	0.1308	0.02	0.982
Apr	0.0776	0.1307	0.59	0.556
May	0.1533	0.1383	1.11	0.275
Jun	0.2339	0.1382	1.69	0.099
Jul	0.2459	0.1381	1.78	0.083
Aug	0.2262	0.1380	1.64	0.109
Sep	0.1944	0.1379	1.41	0.166
Oct	0.1402	0.1378	1.02	0.315
Nov	0.0670	0.1378	0.49	0.629

S = 0.194829 R-Sq = 27.1% R-Sq(adj) = 4.7%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	12	0.55069	0.04589	1.21	0.312
Residual Error	39	1.48038	0.03796		
Total	51	2.03107			

Source	DF	Seq SS
Months	1	0.00116
Jan	1	0.11895
Feb	1	0.10203
Mar	1	0.09089
Apr	1	0.02725
May	1	0.00015
Jun	1	0.02596
Jul	1	0.04770
Aug	1	0.05032
Sep	1	0.04700
Oct	1	0.03030
Nov	1	0.00899

I. **(2 Mark)** Unadjusted trend value for first quarter of first year is …….

II. **(2 Marks)** The estimated monthly compound growth rate is …….

III. **(1 Mark)** The estimated multiplier for August is ……

(2 Marks) **Interpretation:**

IV. **(3 Marks)** Predict the Gas price in December.

V. **(2 Marks)** Which months have lower prices on average? Why?

Question Four (8 Marks) A study on the mean monthly prices (in dollars per gallon) for unleaded gasoline in the United States from January 2006 to April 2010 is conducted. Part of the data is given in the table below

Price($)	2.197	2.357	2.254	2.498	2.914	2.867	2.869	3.004	2.845
Month	Jan-06	Feb-06	Mar-06	Apr-06	May-06	Jun-06	Jul-06	Aug-06	Sep-06

Four autoregressive models are shown in the following Minitab output.

Regression Analysis: Price($) versus Lag1

The regression equation is

Price($) = 0.343 + 0.879 Lag1

Predictor	Coef	SE Coef	T	P
Constant	0.3426	0.1817	1.89	0.065
Lag1	0.87881	0.06536	13.45	0.000

S = 0.254627 R-Sq = 78.7% R-Sq(adj) = 78.2%

Regression Analysis: Price($) versus Lag1, Lag2

The regression equation is

Price($) = 0.478 + 1.36 Lag1 - 0.529 Lag2

Predictor	Coef	SE Coef	T	P
Constant	0.4781	0.1625	2.94	0.005
Lag1	1.3554	0.1237	10.95	0.000
Lag2	-0.5285	0.1226	-4.31	0.000

S = 0.219836 R-Sq = 84.6% R-Sq(adj) = 83.9%

Regression Analysis: Price($) versus Lag1, Lag2, Lag3

The regression equation is

Price($) = 0.542 + 1.32 Lag1 - 0.436 Lag2 - 0.079 Lag3

Predictor	Coef	SE Coef	T	P
Constant	0.5423	0.1774	3.06	0.004
Lag1	1.3200	0.1473	8.96	0.000
Lag2	-0.4358	0.2355	-1.85	0.071
Lag3	-0.0790	0.1458	-0.54	0.591

S = 0.220499 R-Sq = 84.9% R-Sq(adj) = 83.9%

Regression Analysis: Price($) versus Lag1, Lag2, Lag3, Lag4

The regression equation is

Price($) = 0.578 + 1.33 Lag1 - 0.532 Lag2 + 0.127 Lag3 - 0.136 Lag4

Predictor	Coef	SE Coef	T	P
Constant	0.5781	0.1959	2.95	0.005
Lag1	1.3321	0.1507	8.84	0.000
Lag2	-0.5318	0.2491	-2.14	0.038
Lag3	0.1266	0.2470	0.51	0.611
Lag4	-0.1365	0.1474	-0.93	0.360

S = 0.221542 R-Sq = 85.4% R-Sq(adj) = 84.0%

a. **(2 Marks)** Which model is significant in predicting the Gas Prices? Explain.

b. **(2 Marks)** Using the 3-rd order autoregressive model, what is the estimated Gas Prices in Jun 2006?

c. **(2 Marks)** Compute the third moving average using a period of length 5.

d. **(2 Marks)** Compute the third exponentially smoothed Gas Price if the smoothing constant is 0.4.

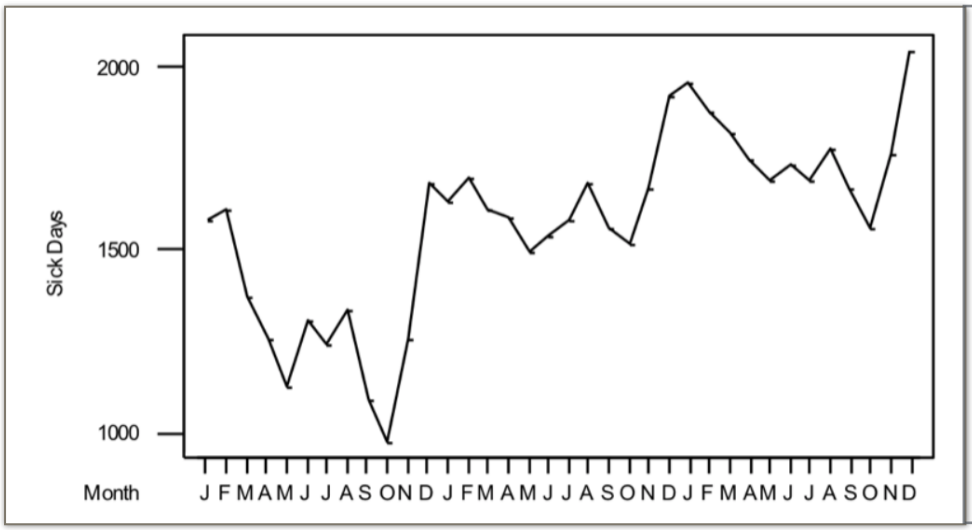
Question Five (8 Marks) A firm known to have a 56% market share wants to test whether this value is still valid in view of recent advertising campaigns of its competitors and its own increased level of advertising. A random sample of 500 consumers reveals 298 use the company's product. Is there evidence to conclude that the company's market share is no longer 56% at $\alpha = 0.01$?

Hypotheses (1 Mark):	H0:	H1:
Assumptions (1 Mark):		
Test Statistic (2 Marks):		
Critical Value(s) (1 Mark):		
Decision Rule (1 Mark):		
Conclusion (2 Mark):		

Question Six (3 pts*4=12 Marks) The Personal manger for a company has been asked to assist the production mangers in developing a staffing plan for the coming year. One factor that influences staffing is the amount of sick time taken by employees each month. **The data collected** *represent the number of sick days, by month f**or the past three years**.*

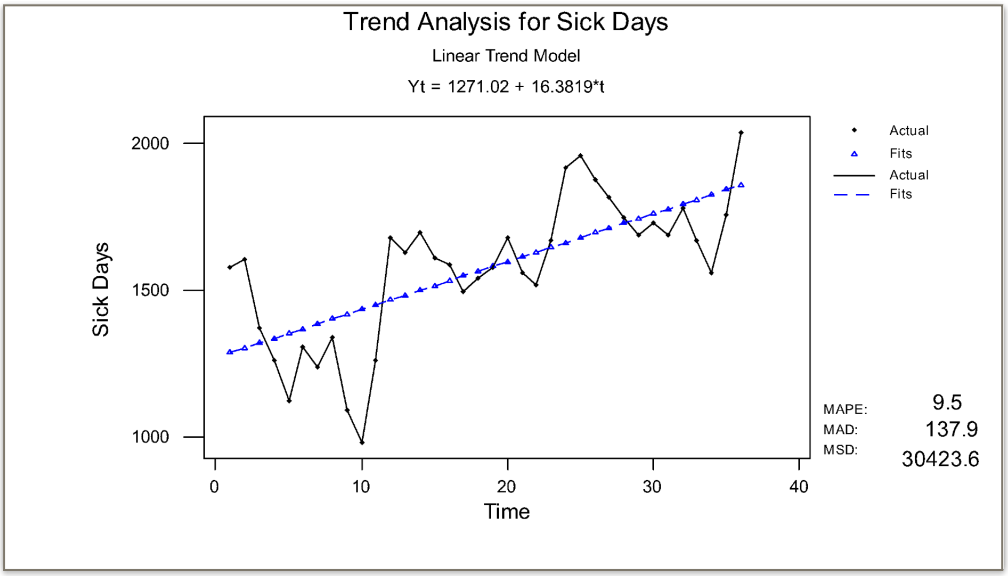
Use the given figures to answer the questions:

A) Identify the time series components exist.



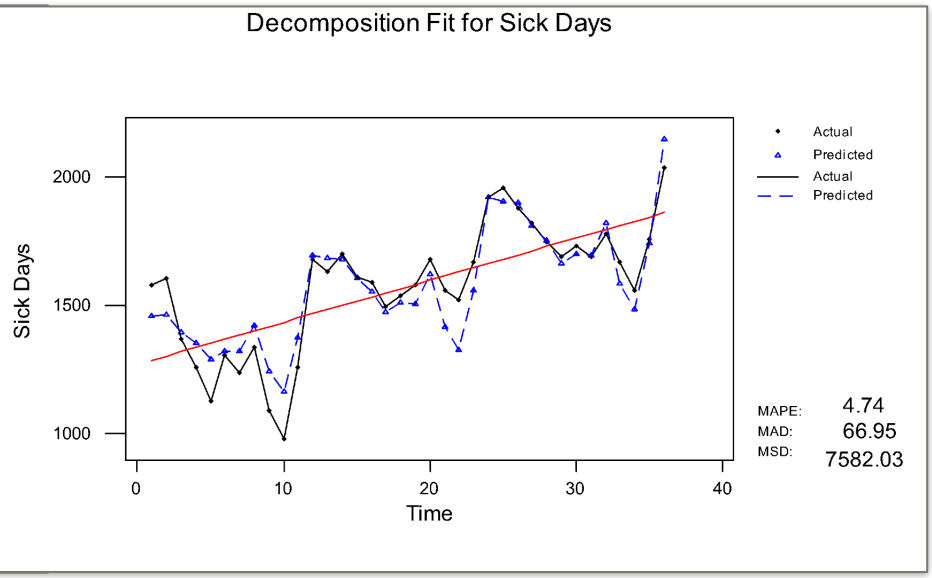
Answer:

B) Use the linear trend model to predict the amount of sick time for the next two months. Do you think this model is appropriate? Explain.



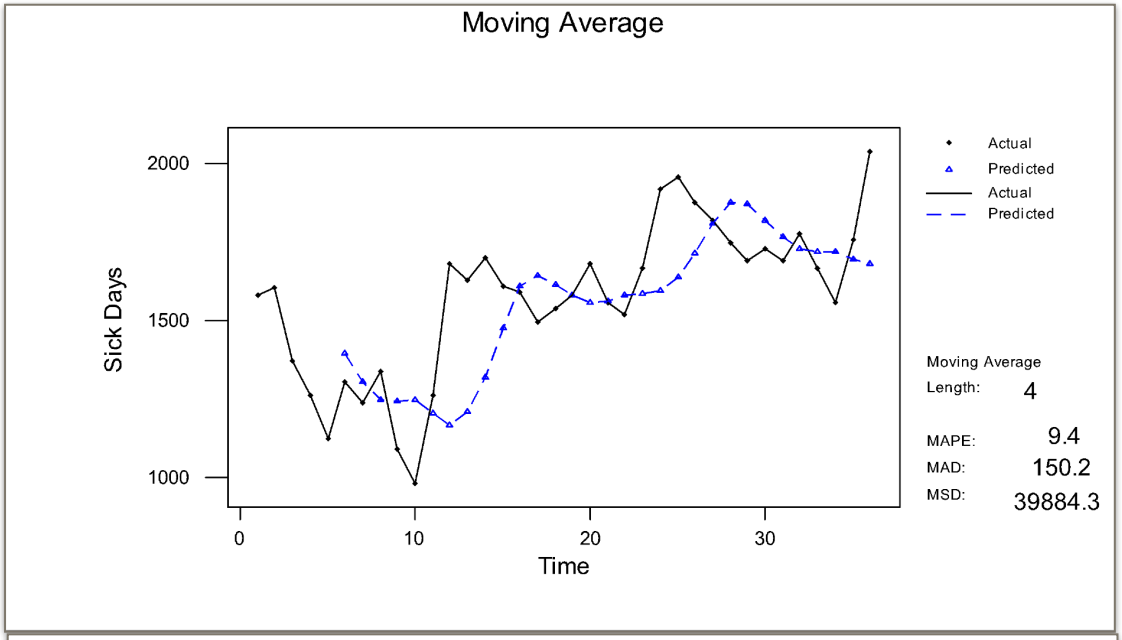
Seasonal Indexes

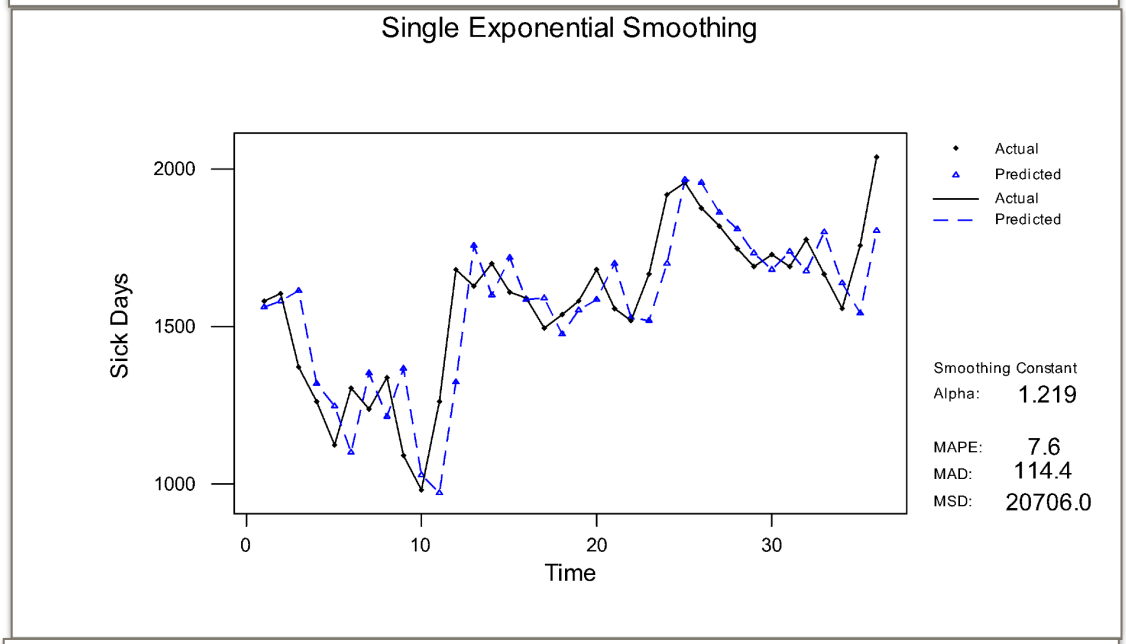
Period	Index
1	1.134210
2	1.121140
3	1.057700
4	1.013420
5	0.951899
6	0.966202
7	0.952811
8	**1.014250**
9	**0.875807**
10	0.812405
11	0.945842
12	1.154320

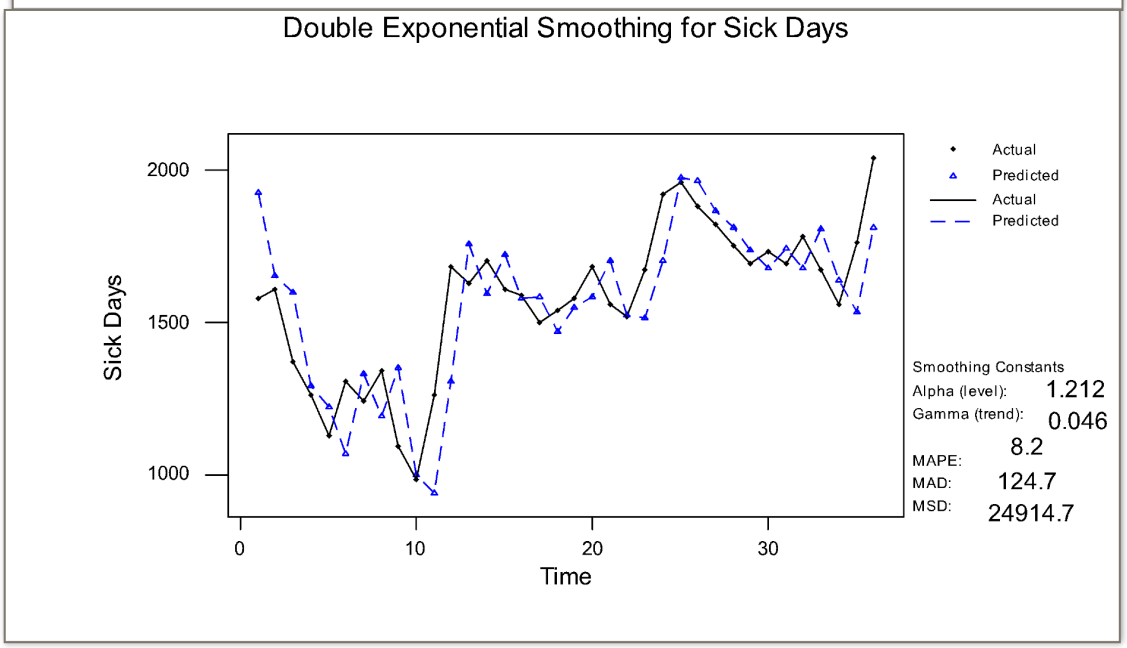


C) Comment on the developed monthly seasonal indexes **for period 8 and 9,** (Note that the seasonal length is 12). Compare the accuracy of the seasonal decomposition model with previous linear model. Which model we will suggest to be used to predict the sales for the coming seasons. (HINT: use MAPE, MAD and MSD)

D) We used the moving average, single exponential and double exponential smoothing (Graphs are in the next page), which smoothing you suggest that we should use? Explain. (HINT: use MAPE, MAD and MSD)


Moving Average
Sick Days
Time
2000
1500
1000
0
10
20
30
Actual
Predicted
Actual
Predicted
Moving Average
Length: 4
MAPE: 9.4
MAD: 150.2
MSD: 39884.3


Single Exponential Smoothing
Sick Days
Time
2000
1500
1000
0
10
20
30
Actual
Predicted
Actual
Predicted
Smoothing Constant
Alpha: 1.219
MAPE: 7.6
MAD: 114.4
MSD: 20706.0


Double Exponential Smoothing for Sick Days
Sick Days
Time
2000
1500
1000
0
10
20
30
Actual
Predicted
Actual
Predicted
Smoothing Constants
Alpha (level): 1.212
Gamma (trend): 0.046
MAPE: 8.2
MAD: 124.7
MSD: 24914.7

Question Seven (14 Marks) a well-known baseball analyst, wants to determine which variables are important in predicting a team's wins in a given baseball season. He has collected data related to wins (y), ERA (x1), saves (x2), runs scored (x3), hits allowed (x4), walks allowed (x5), and errors(x6) for the 2009 season. The coefficient of multiple determination (R_j^2) of each of the six predictors with all the other remaining predictors are, respectively, 91.013%, 27.431%, 91.038%, 62.990% , 39.247%, and 18.831%.

I. Regression using all independent variables.

Regression Analysis: Wins versus X1 (E.R.A.), X2(Runs Scored), ...

```
The regression equation is
Wins = 78.7 - 9.39 X1 (E.R.A.) + 0.0739 X2(Runs Scored)
       - 0.0176 X3 (Hits Allowed) - 0.0044 X4 (Walks Allowed)
       + 0.661 X5 (Saves) - 0.115 X6 (Errors)
```

Predictor	Coef	SE Coef	T	P	VIF
Constant	78.72	24.32	3.24	0.004	
X1 (E.R.A.)	-9.394	3.568	-2.63	0.015	_____
X2(Runs Scored)	0.073851	0.007288	10.13	0.000	1.378
X3 (Hits Allowed)	-0.01755	0.01912	-0.92	0.368	_____
X4 (Walks Allowed)	-0.00444	0.01717	-0.26	0.798	_____
X5 (Saves)	0.66072	0.08661	7.63	0.000	1.646
X6 (Errors)	-0.11471	0.03283	-3.49	0.002	1.232

```
S = 2.48411   R-Sq = 96.2%   R-Sq(adj) = 95.3%
```

1. (3 Marks) Compute the missing VIF numbers.

2. (2 Marks) Which of the predictors should first be dropped to remove collinearity? **Why?**

II. Best Subset Regression:

Best Subsets Regression: Y versus X1, X2, X4, X5, X6

Response is Y

Vars	R-Sq	R-Sq(adj)	Mallows Cp	S	X 1	X 2	X 4	X 5	X 6
1	61.4	60.0	212.0	7.2183				X	
1	40.6	38.5	340.5	8.9579	X				
2	82.9	81.6	81.4	4.8929	X	X			
2	75.7	73.9	126.0	5.8351		X		X	
3	94.3	93.6	13.4	2.8908	X	X		X	
3	84.4	82.6	74.3	4.7640	X	X			X
4	96.0	95.4	4.4	2.4443	X	X		X	X
4	94.3	93.3	15.4	2.9480	X	X	X	X	
5	96.1	95.3	6.0	2.4760	X	X	X	X	X

State the predictors that will be selected for the best subset regression models, based on

i) R-Sq criteria (1)

ii) R-Sq (adj) criteria (1)

iii) Mellows Cp criteria (2)

III. The analyst included X5_sq (the square of X5) to the list of independent variable. Use the MINITAB outputs to answer the following questions:

Stepwise Regression: Y versus X1, X2, X4, X5, X6, X5_sq

```
  Alpha-to-Enter: 0.15  Alpha-to-Remove: 0.15
Response is Y on 6 predictors, with N = 30

Step             1       2       3       4
Constant     52.39   14.73   67.01   76.99

X5_sq       0.0173  0.0144  0.0090  0.0090
T-Value       6.88    6.37    7.39    8.45
P-Value      0.000   0.000   0.000   0.000

X2                  0.0568  0.0757  0.0705
T-Value               3.57    9.67    9.98
P-Value              0.001   0.000   0.000

X1                          -13.3   -12.6
T-Value                     -9.65  -10.32
P-Value                     0.000   0.000

X6                                 -0.094
T-Value                             -3.01
P-Value                             0.006

S             7.09    5.95    2.83    2.47
R-Sq         62.80   74.72   94.48   95.96
R-Sq(adj)    61.47   72.85   93.85   95.31
Mallows Cp   194.0   125.5    10.6     3.9
```

(I) Which predictor will be chosen as the first variable entering in to the model (1)

(II) Write the prediction equation for the best model. (1)

(III) Which of the predictor(s) of the best selected model are significant at 0.005 significance level? (2)

(IV) Are there any predictors that have been removed from the model after selection? If yes, state them. (1)

Question Eight (2 pts*5=10 Marks)

1. If we are performing a two-tailed test of whether μ= 100, the probability of rejecting the H_0 when the true mean is 105 will be _____ the probability of rejecting the H_0 when the true mean is 110.
   A. less than
   B. greater than
   C. equal to
   D. not comparable to

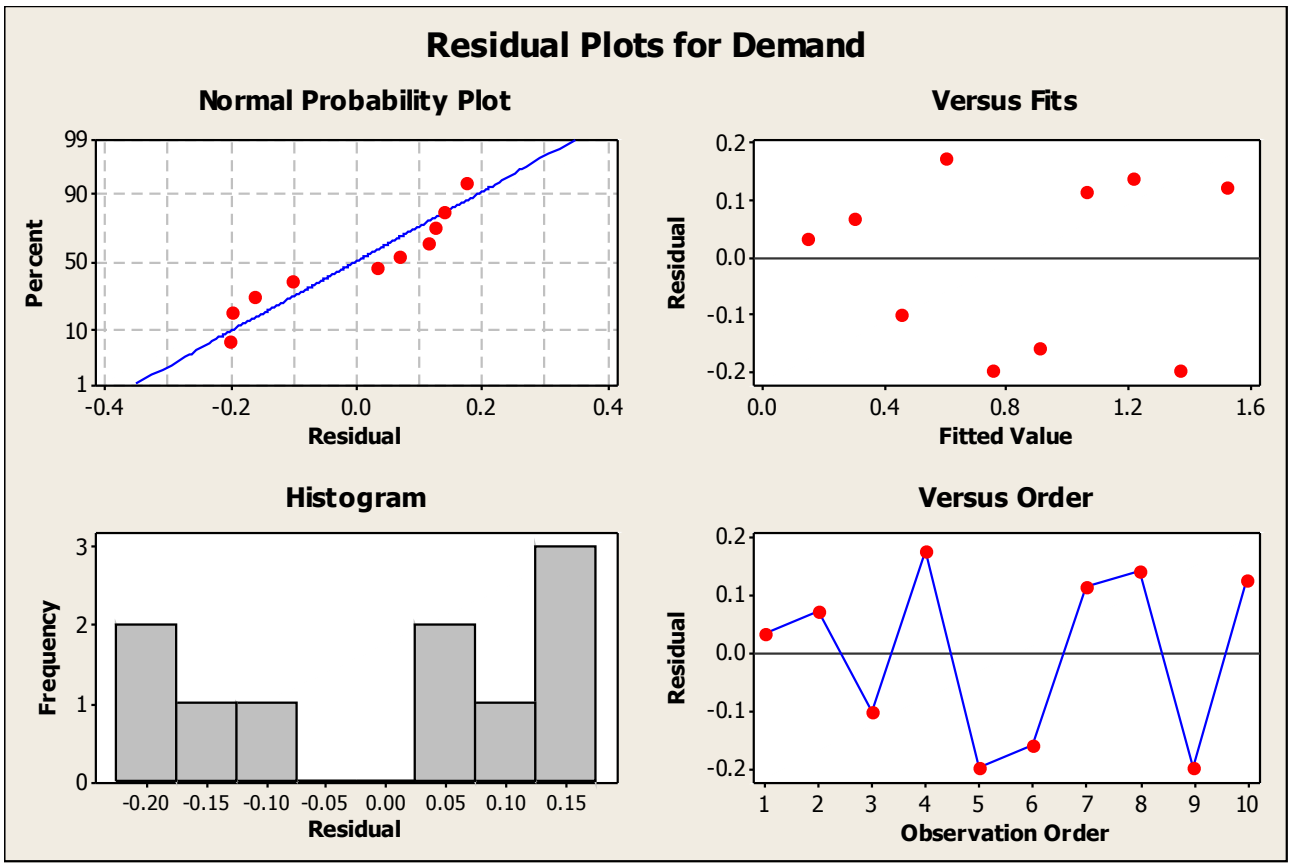
2. If we are testing for the difference between the means of two paired populations with samples of n_1 = 20 and n_2 = 20, the number of degrees of freedom is equal to:
   A. 39.
   B. 38.
   C. 19.
   D. 18.

3. When formulating a hypothesis test, which of the following statements is TRUE?
   A. The null hypothesis should never contain the equality.
   B. The null and alternative hypotheses should be stated in terms of the population parameters.
   C. If possible, the research hypothesis should be formed as the null hypothesis.
   D. The null hypothesis should be established such that the chance of making a Type I error is minimized.

4. A manufacturer of industrial plywood has a contract to supply a boat maker with a large amount of plywood. One of the specifications calls for the standard deviation in thickness to not exceed .03 inches. A sample of n = 30 sheets was sampled randomly from a recent production run. The mean thickness was right at the ¾ inch target thickness and the sample standard deviation was 0.05 inches. Testing at the 0.05 level of significance,

   Which of the following steps is **NOT correct** for this hypothesis testing?
   A. The test hypotheses are $H_0 : \sigma \leq 0.03 \; vs \; H_1 : \sigma > 0.03$
   B. The critical value is: $\chi^2_{.95,29} = 17.7084$
   C. The test statistic is $\chi^2_{cal} = 80.5556$
   D. The conclusion is that there is not enough evidence to indicate that the standard deviation in thickness to not exceed 0.03 inches.

5. Based on the following residual plot for simple linear regression model



which of the following is **NOT** true:

A. based on the histogram of the errors, the normality assumption is violated

B. based on Residual vs Fitted, the errors are independent

C. based on the normal probability plot, there is no linear relationship between the two variables.

D. the mean of the errors equal zero.