KING FAHD UNIVERSITY OF PETROLEUM & MINERALS
DEPARTMENT OF MATHEMATICS & STATISTICS

Term 181

STAT 310: Linear Regression

Final Exam

Thursday, December 20, 2018 @7:00 PM

Name:_____

ID #:_____

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 **5 decimal places**.

Question No	Full Marks	Marks Obtained
Q1	20	
Q2	30	
Q3	29	
Q4	15	
Q5	26	
Total	**120**	

Quest. One.

1. (4 pts) Prove that the matrices **H** and **I** − **H** are idempotent, that is, **HH** = **H** and (**I**−**H**)(**I**−**H**)= **I** − **H**. (**Where H** = $\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ **)**

2. (3 pts) **Show that:** $\hat{\mathbf{e}} = (\mathbf{I} - \mathbf{H})\mathbf{Y}$

3. (6 pts) Show that E ($\hat{\mathbf{e}}$ | $\mathbf{X}$) =0

4. (7 pts) Show that Var($\hat{\mathbf{e}}$ | $\mathbf{X}$) = σ^2 ($\mathbf{I}-\mathbf{H}$)

Quest. Two. **Computer outputs for a sample of size 196 observations and 7 independent variables.**

Regression analysis: Log Y vs x1, x2, x3, x4, x5, x6, x7

Forward Selection of Terms

Candidate terms: x1, x2, x3, x4, x5, x6, x7

	-----Step 1-----		-----Step 2-----		-----Step 3-----		-----Step 4-----	
	Coef	P	Coef	P	Coef	P	Coef	P
Constant	-1.47		28.19		15.81		0.39	
x2	0.1818	0.000	0.2777	0.000	0.2454	0.000	0.1935	0.000
x7			-1.230	0.000	-0.848	0.000	-0.378	0.103
x4					0.1145	0.000	0.1659	0.000
x6							0.0628	0.008
x5								
S	0.848270		0.702898		0.673847		0.663218	
R-sq	25.49%		49.10%		53.47%		55.16%	
R-sq(adj)	25.11%		48.58%		52.74%		54.22%	
R-sq(pred)	23.58%		47.32%		51.37%		52.63%	
Mallows' Cp	124.70		26.33		9.79		4.60	

	------Step 5-----	
	Coef	P
Constant	-0.58	
x2	0.1970	0.000
x7	-0.350	0.132
x4	0.1628	0.000
x6	0.0496	0.046
x5	0.01552	0.113
S	0.660563	
R-sq	55.75%	
R-sq(adj)	54.59%	
R-sq(pred)	52.74%	
Mallows' Cp	4.09	

α to enter = 0.15

Backward Elimination of Terms

Candidate terms: x1, x2, x3, x4, x5, x6, x7

	------Step 1-----		------Step 2-----		------Step 3-----		------Step 4-----	
	Coef	P	Coef	P	Coef	P	Coef	P
Constant	0.19		0.06		-0.58		-11.20	
x1	-0.0035	0.765	-0.0033	0.770				
x2	0.1993	0.000	0.2013	0.000	0.1970	0.000	0.1633	0.000
x3	-0.47	0.946						
x4	0.1573	0.000	0.1586	0.000	0.1628	0.000	0.1996	0.000
x5	0.01517	0.126	0.01519	0.124	0.01552	0.113	0.01667	0.089
x6	0.0515	0.107	0.0502	0.045	0.0496	0.046	0.0753	0.000
x7	-0.343	0.470	-0.371	0.128	-0.350	0.132		
S		0.663908		0.662158		0.660563		0.662794
R-sq		55.77%		55.77%		55.75%		55.22%
R-sq(adj)		54.12%		54.37%		54.59%		54.28%
R-sq(pred)		51.73%		52.31%		52.74%		52.63%
Mallows' Cp		8.00		6.00		4.09		4.36

α to remove = 0.1

Stepwise Selection of Terms

Candidate terms: x1, x2, x3, x4, x5, x6, x7

	-----Step 1-----		-----Step 2-----		-----Step 3-----		-----Step 4-----	
	Coef	P	Coef	P	Coef	P	Coef	P
Constant	-1.47		28.19		15.81		0.39	
x2	0.1818	0.000	0.2777	0.000	0.2454	0.000	0.1935	0.000
x7			-1.230	0.000	-0.848	0.000	-0.378	0.103
x4					0.1145	0.000	0.1659	0.000
x6							0.0628	0.008
x5								
S	0.848270		0.702898		0.673847		0.663218	
R-sq	25.49%		49.10%		53.47%		55.16%	
R-sq(adj)	25.11%		48.58%		52.74%		54.22%	
R-sq(pred)	23.58%		47.32%		51.37%		52.63%	
Mallows' Cp	124.70		26.33		9.79		4.60	

	-----Step 5-----		------Step 6-----	
	Coef	P	Coef	P
Constant	-11.08		-11.20	
x2	0.1566	0.000	0.1633	0.000
x7				
x4	0.2062	0.000	0.1996	0.000
x6	0.0918	0.000	0.0753	0.000
x5			0.01667	0.089
S	0.666111		0.662794	
R-sq	54.53%		55.22%	
R-sq(adj)	53.82%		54.28%	
R-sq(pred)	52.45%		52.63%	
Mallows' Cp	5.28		4.36	

α to enter = 0.1, α to remove = 0.1

I. **(3*4=12 pts) Write the optimal model based on**

1) forward selection

2) backward selection

3) stepwise selection

4) are the models in 1 and 2 identical? If not, explain why.

Best Subsets Regression: log y versus x1, x2, x3, x4, x5, x6, x7

Vars	variables	R-Sq	R-Sq (adj)	PRESS	R-Sq (pred)	Mallows Cp	S	BIC	AIC
1	x2	25.5	25.1	143.2	23.6	124.7	0.84827	-55.960113	-62.176974
2	x2,x7	49.1	(A)2 pts _____	98.7	47.3	(B) 3 pts _____	0.7029	(C) 2 pts _____	(D) 2 pts _____
3	x2,x4,x6	54.5	53.8	89.1	52.5	5.3	0.66611	-142.1987	-154.47794
4	x2,x4,x5,x6	55.2	54.3	88.7	52.6	4.4	0.66279	-139.90276	-155.4449
5	x2,x4,x5,x6,x7	55.7	54.6	88.5	52.7	4.1	0.66056	-136.97466	-155.78292
6	x1,x2,x4,x5,x6,x7	55.8	54.4	89.4	52.3	6	0.66216	-131.7825	-153.86843
7	All variables	55.8	54.1	90.4	51.7	8	0.66391	-126.50953	-151.87356

$$\mathrm{BIC}_{\mathrm{Sch}} = n\ln\left(\frac{SS_{\mathrm{Res}}}{n}\right) + p\ln(n).$$

$$\mathrm{AIC} = n\ln\left(\frac{SS_{\mathrm{Res}}}{n}\right) + 2p.$$

Where

P=k+1

$$\mathrm{PRESS}_p = \sum_{i=1}^{n}[y_i - \hat{y}_{(i)}]^2 = \sum_{i=1}^{n}\left(\frac{e_i}{1-h_{ii}}\right)^2$$

$$C_p = \frac{SS_{\mathrm{Res}}(p)}{\hat{\sigma}^2} - n + 2p$$

$$R^2_{\mathrm{Adj},p} = 1 - \frac{n-1}{n-p}(1-R^2_p) = 1 - \frac{n-1}{n-p}\frac{SS_{\mathrm{Res}}(p)}{SS_{\mathrm{T}}} = 1 - \frac{MS_{\mathrm{Res}}(p)}{SS_{\mathrm{T}}/(n-1)}$$

II. (9 pts) Compute the missing values. (Show your work.)

III. (3*3=9 pts) Using best subset outputs, Identify the optimal model or models based on

	1) R^2_{adj}	*2)* **AIC**	*3)* **BIC**
Rule			
Model(s)			

Quest. Three. A professor of accounting wanted to develop a multiple regression model to predict the students' grades in her fourth-year accounting course. She decides that the two most important factors are the student's grade point average in the first three years and the student's major. She proposes the model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$$

where

y = Fourth-year accounting course mark (out of 100)
x_1 = G.P.A. in first three years (range 0 to 12)
x_2 = 1 if student's major is accounting
= 0 if not
x_3 = 1 if student's major is finance
= 0 if not

The analysis for a sample of size 100 students showed that the mean squares for regression is 5699.333 and the sum of squares for error is 21553. The table of coefficients is shown below.

Table of coefficients

Predictor	*Coef*	**SE Coef**	*T*
Constant	9.14	7.10	1.287
x_1	6.73	1.91	3.524
x_2	10.42	4.16	2.505
x_3	5.16	3.93	1.313

1. (9 pts) Develop the ANOVA table.

ANALYSIS OF VARIANCE

Source of Variation	***df***	***SS***	***MS***	***F***
Regression	_____	_____	_____	_____
Error	_____	_____	_____	
Total	_____			

2. (5 pts) Do these results allow us to conclude at the 1% significance level that the model is useful in predicting the fourth-year accounting course mark?

Hypotheses	H_0: H_1:
C.V.	
D.R.	
Decision	
Conclusion	

3. (5 pts) Do these results allow us to conclude at the 1% significance level that on average accounting majors outperform those whose majors are not accounting or finance?

Hypotheses	H_0:	H_1:
C.V.		
D.R.		
Decision		
Conclusion		

4. (5 pts) Do these results allow us to conclude at the 1% significance level that on average finance majors outperform those whose majors are not accounting or finance?

Hypotheses	H_0:	H_1:
C.V.		
D.R.		
Decision		
Conclusion		

5. (5 pts) Do these results allow us to conclude at the 1% significance level that grade point average in first three years is linearly related to fourth-year accounting course mark?

Hypotheses	H_0:	H_1:
C.V.		
D.R.		
Decision		
Conclusion		

Quest. Four. A sample of 30 companies was randomly selected for a study investigating what factors affect the size of company bonuses. Data were collected on the number of employees at the company and whether or not the employees were unionized (1 = yes, 0 = no). Below are the multiple regression results.

Dependent Variable is Average Annual Bonus

Predictor	***Coef***	**SE Coef**	***T***	**P**
Constant	347.9	872.2	0.40	0.693
Employees	0.6547	0.1105	5.92	0.000
Union	1259.5	605.8	2.08	0.047

S = 1631.56 R-Sq = 62.4% R-Sq(adj) = 59.6%

Analysis of Variance

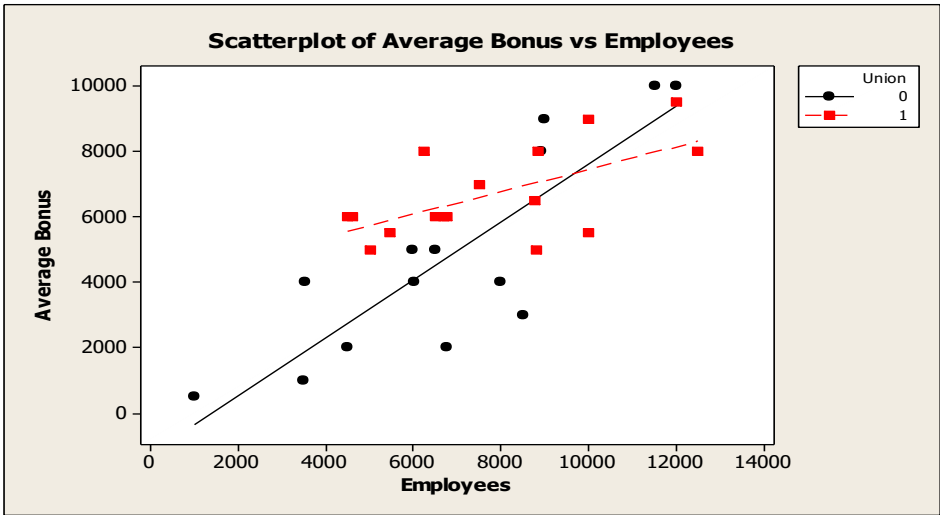
Source of Variation	*df*	*SS*	*MS*	*F*	**P**
Regression	2	119368382	59684191	22.42	0.000
Error	27	71873285	2661974		
Total	29	191241667			

1. (2 pts) Write out the estimated regression equation.

2. (3 pts) Are all of the independent variables significant in this regression equation (using $\alpha = .05$)? Explain.

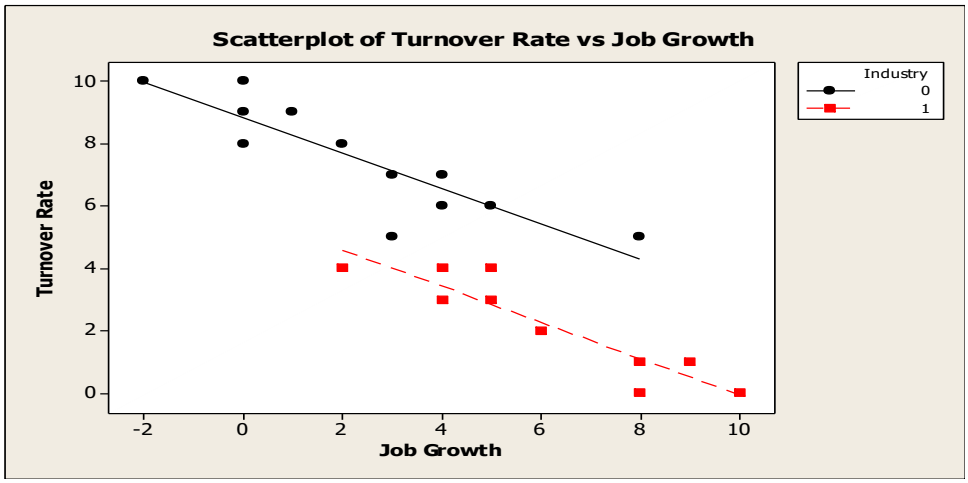
3. (4 pts) Interpret the coefficient of the *Union*.

4. (3 pts) Based on the scatterplot below, suggest a variable to be included in this model? **Justify your answer**.



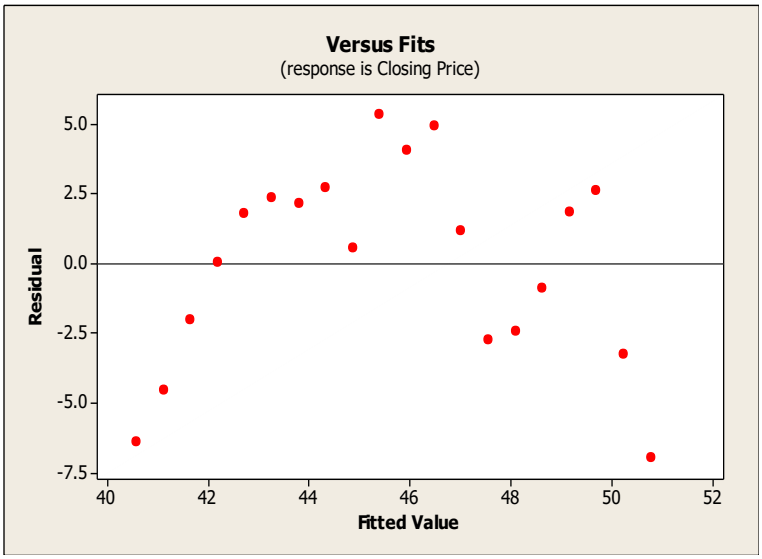
5. (3 pts) we Defined a new categorical variable "Industry" as follow
   Industry = 1, If the sector is high tech industry and
   Industry = 0, if the financial services sector.

Based on the scatterplot below, is it appropriate to use the variable *Industry* as an indicator variable in this regression model? **Justify your answer**.



Quest. Five. (2*13=26 pts) MCQ

1. Monthly closing stock prices for a utility company were obtained from January 2007 through August 2008. A regression model was estimated to describe the trend in closing stock prices over time. What does the plot of residuals below suggest?



A. An outlier is present in the data set.

B. The linearity condition is not satisfied.

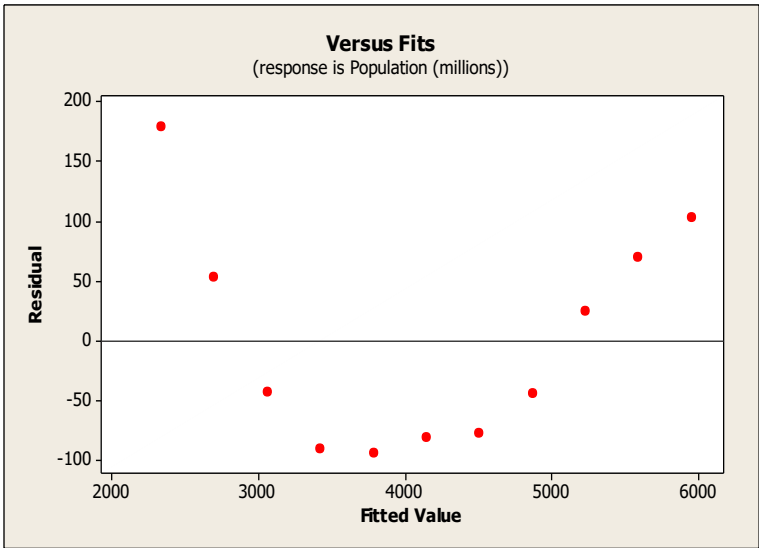
C. A high leverage point is present in the data set.

D. The data are not normal.

E. The equal spread condition is not satisfied.

2. A linear regression model was estimated to describe the trend in world population over time. Below is the plot of residuals versus predicted values.

What does the plot of residuals suggest?



A. An outlier is present in the data set.

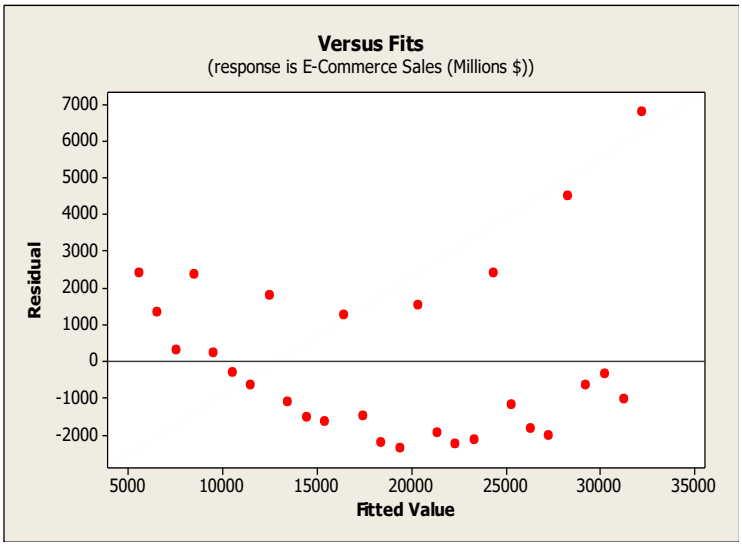
B. The linearity condition is not satisfied.

C. A high leverage point is present in the data set.

D. The data are not normal.

E. The equal spread condition is not satisfied.

3. Quarterly figures for e-commerce retail sales were obtained from the first quarter of 2001 through the fourth quarter of 2007. A regression model was estimated to describe the trend in e-commerce retail sales over time. What does the plot of residuals versus predicted values suggest?



A. The data are not normal.

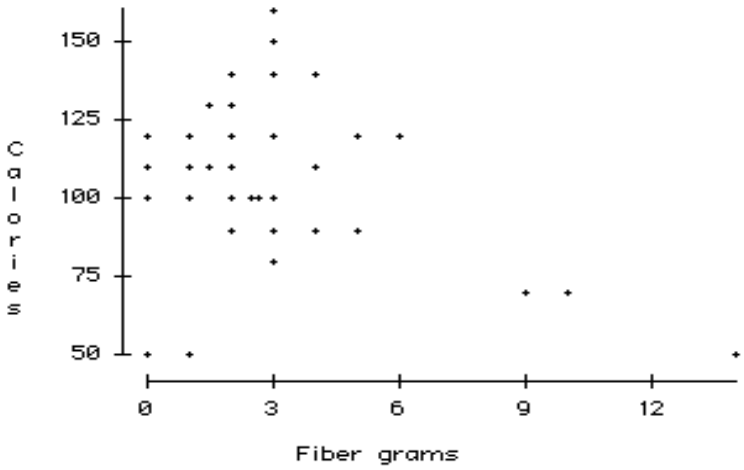
B. The linearity condition is not satisfied.

C. The equal spread condition is not satisfied.

D. Both A and B.

E. Both B and C.

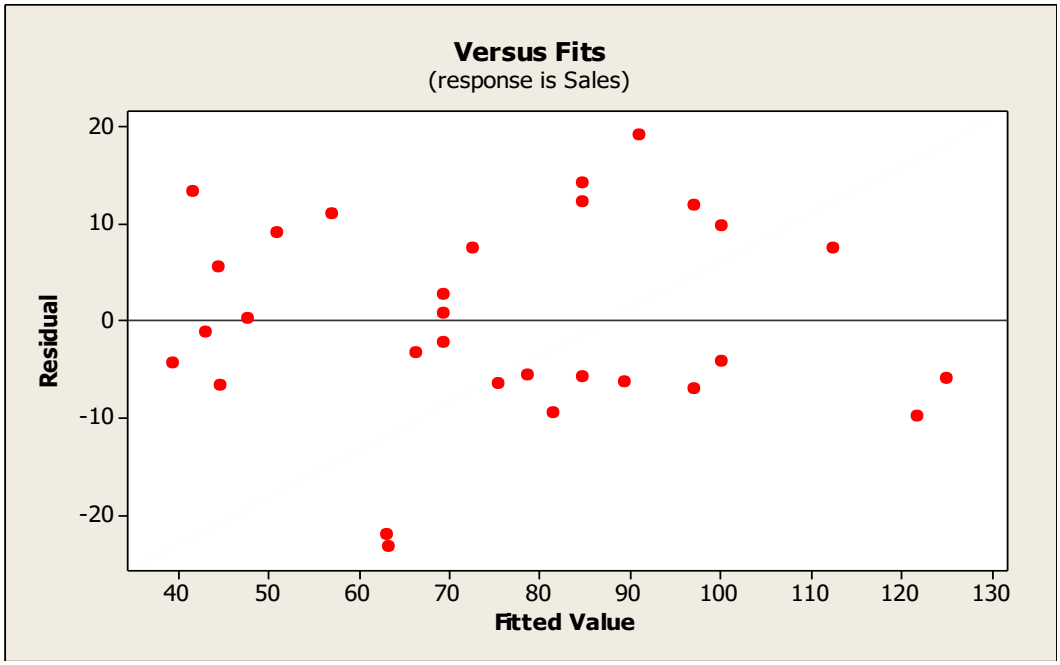
4. The advertising campaign for a high fiber cereal wants to claim that high fiber cereals are lower in calories. In order to research this claim, they obtain nutritional information for 77 breakfast cereals including the amount of fiber (in grams) and the number of calories per serving. The data resulted in the following scatterplot.



Which statement is true?

A. Some high leverage points are evident.
B. There is a strong positive association between amount of fiber and number of calories per serving.
C. As the amount of fiber increases so does the number of calories per serving.
D. Both A and B.
E. All of the above.

5. Below is the plot of residuals versus predicted values for this estimated multiple regression model. What does the residual plot suggest?



A. The Linearity condition is not satisfied.
B. There is an extreme departure from normality.
C. The variance is not constant.
D. The presence of a couple of outliers.
E. The plot thickens from left to right.

A sample of 8 households was asked about their monthly income (X) and the number of hours they spend connected to the internet each month (Y). The data yield the following statistics:

$$\sum x = 324, \sum y = 393, \sum (x-\bar{x})^2 = 1720.875, \sum (y-\bar{y})^2 = 1150, \sum (x-\bar{x})(y-\bar{y}) = 1090.5$$

6. What is the sample correlation coefficient between X and Y?
7. What is the slope of the regression line of hours on income?
8. What is the y-intercept of the regression line of hours on income?
9. What is the regression sum of squares?
10. What is the value of the coefficient of determination?
11. What is the estimate of the variance of the population model error?
12. What is the standard error of the slope of the regression line of hours on income?
13. What is the value of the test statistic for testing $H_0 : \beta_1 = 0$ vs. $H_1 : \beta_1 \neq 0$?