STA 6207 – Practice Problems – Model Diagnostics

Q.1. A simple regression model is fit, relating a dependent variable Y, to an independent variable X. You are given the following data and summary statistics.

X	Y	Y-hat_j	Y-bar_j			sum(X)	sum(Y)
0	15					90	225
0	13						
0	17					SS_XX	SS_YY
10	25					600	642
10	27						
10	23					SS_XY	
20	34					600	
20	32						
20	39						

p.1.a. Give the fitted equation, based on the simple linear regression model:

$\hat{\beta}_1 =$ ____________________________ $\hat{\beta}_0 =$ ____________________________ $\hat{Y} =$ ________________________

p.1.b. Fill in the 3rd and 4th columns of the table above.

p.1.c. Compute the Pure Error Sum of Squares and degrees of freedom:

SSPE = ____ df_{PE} = ___________________________

p.1.d. Compute the Lack-of-Fit Sum of Squares and degrees of freedom:

SSLF = ____ df_{LF} = ___________________________

p.1.e. Test H_0: Model is Linear vs H_A: Model is not Linear at $\alpha = 0.05$ significance level.

Test Statistic ________________________________ Rejection Region ____________________________

Q.2. An analysis reports that a studentized deleted residual is 5.0

a) Conclude this case is an outlier b) Conclude this case is not an outlier c) Need more information

Q.3. A regression through the origin is to be fit, relating Y to X for n=5 observations. The X levels are (1,2,3,4,10).

p.3.a. Obtain X'X, $(X'X)^{-1}$, the P matrix, and the diagonal elements of P (v_{ii}).

p.3.b. What do the v_{ii} elements sum to? Which, if any, elements are potentially influential?

p.3.c. Below is the fit for all cases (top), and with observation 5 dropped (bottom) $b_1 = \hat{\beta}_1 = \frac{\sum_{i=1}^{n} X_i Y_i}{\sum_{i=1}^{n} X_i^2}$ Note: $\sum_{i=1}^{n} e_i^2 \neq 0$

X	Y	Y-hat	e		b1
1	2	2.75	-0.75		2.75
2	5	5.51	-0.51		SS(Res)
3	6	8.26	-2.26		
4	7	11.02	-4.02		
10	30	27.54	2.46		
Observation 5 dropped					
X	Y	Y-hat	e		b1
1	2	1.93	0.07		1.93
2	5	3.87	1.13		SS(Res)
3	6	5.80	0.20		
4	7	7.73	-0.73		

All Cases:

SS(Res) =

s =

Observation 5 dropped:

p.3.d. Compute $\hat{Y}_{5(5)}$, and using $s_{(5)}$ as estimate of σ: $DFFITS_5$, and $DFBETAS_{1(5)}$

Q.4. A simple linear regression model is fit, with n = 12 observations (3 each at 4 levels of X). The residual sum of squares from the Regression model is SSResidual = 9042. The 4 fitted values at the distinct X levels are: (40, 80, 120, and 160). The 4 sample means at the distinct X levels are: (60, 70, 80, and 190).

p.4.a. Complete the following ANOVA table (degrees of freedom and sums of squares

ANOVA		
Source	**df**	**SS**
Regression		
Residual		
Lack of Fit		
Pure Error		
Total Corrected		

p.4.b. Conduct the F-test for Lack-of Fit

$$H_0: E\{Y_{ij}\} = \mu_j = \beta_0 + \beta_1 X_j \quad j = 1,\ldots,c; \quad i = 1,\ldots,n_j \qquad H_A: E\{Y_{ij}\} = \mu_j \neq \beta_0 + \beta_1 X_j$$

Test Statistic: ____________________________________ Rejection Region: ___________________________

Do you conclude that the relationship between $E\{Y\}$ and X is linear? **Yes** or **No**

Q.5. A simple linear regression model is fit, based on n=5 individuals. The data and the projection matrix are given below:

X		Y	P				
1	2	8	0.344	0.303	-0.129	0.262	0.221
1	4	12	0.303	0.274	-0.035	0.244	0.215
1	25	24	-0.129	-0.035	0.953	0.059	0.153
1	6	14	0.262	0.244	0.059	0.226	0.209
1	8	18	0.221	0.215	0.153	0.209	0.203

p.5.a. Give the leverage values for each observation. Do any exceed twice the average of the leverage values?

Observation 1 ________ Obs 2 ______________ Obs 3 ___________ Obs 4 ____________ Obs 5 ______________

p.5.b. Give $\boldsymbol{\beta}$, based on the following results

X'X		X'Y	INV(X'X)	
5	45	76	0.438	-0.026
45	745	892	-0.026	0.003

p.5.c. Compute SSE and S_e (Note: $\mathbf{Y'Y} = 1304$ $\mathbf{Y'PY} = 1282.45$)

p.5.d. The following table contains the fitted values with and without each observation, residual standard deviation when that observation was not included in the regression, and the regression coefficients when that observation was not included in the regression.

Y-hat	Y-hat(-i)	S_i	beta0_i	beta1_i
10.92	12.45	2.07	11.41	0.52
12.14	12.19	3.28	9.76	0.61
24.99	45.00	0.63	5.00	1.60
13.36	9.46	3.24	9.46	0.62
14.59	13.72	1.86	8.72	0.62

p.5.d.i. Compute DFFITS for the fifth observation

p.5.d.ii. Compute DFBETAS0 for the first observation

p.5.d.iii. Compute DFBETAS1 for the third observation

Q.6. A linear regression model was fit, relating energy consumption (Y) to population (X) for n = 183 nations. The following results are for models with and without the United States. The leverage value for the US is $P_{ii} = .0275$, and the Population is X = 310.0

INV(X'X)				
0.00590971	-0.00001109			
-0.00001109	0.00000030			
	With US	With US	Without US	Without US
	Coefficients	*Standard Error*	*Coefficients*	*Standard Error*
Intercept	0.6257	0.5912	0.4231	0.3687
popMill	0.0571	0.0042	0.0505	0.0026
MSResidual	59.14		22.98	

p.6.a. Compute the fitted values for the US based on each model.

With US ____ Without US _____________________________________

p.6.b. Compute DFFITS for the US

p.6.c. Compute each of the DFBETAS for the US.

Intercept ____ Population ____

p.6.d. What is the average leverage value among the 183 nations?

Q.7. An experiment was conducted to measure the subsoil pressure of a steel ground roller. **There were 3 replicates at each of 4 depths (X=5, 10, 15, 20 cm).** The response was measured force (100s of Newtons).

The fitted regression equation is $\hat{Y} = 49.371 - 2.036X$

We want to test $H_0: E\{Y_j\} = \beta_0 + \beta_1 X_j$ $H_A: E\{Y_j\} = \mu_j \neq \beta_0 + \beta_1 X_j$

j	X_j	Ybar_j	SD_j	Yhat_j	Pure Error	Lack of Fit
1	5	40.38	4.32			
2	10	28.87	6.83			
3	15	16.23	3.76			
4	20	11.15	4.48			

p.7.a. Compute the Pure Error Sum of Squares and Degrees of Freedom. Hint: What is SD_j equal to?

SSPE = ____________________________________ df_{PE} = ___________________________________

p.7.b. Compute the Lack-of-Fit Sum of Squares and Degrees of Freedom.

SSLF = ____________________________________ df_{LF} = ___________________________________

p.7.c. Conduct the F-test for Lack-of-Fit

Test Statistic: _________________________ Rejection Region: _______________ Reject H_0? **Yes** / **No**

Q.8. A simple linear regression model is fit, based on n=5 individuals. The data and the projection matrix are given below:

X		Y	P				
1	0	4	0.402	0.330	-0.175	0.258	0.186
1	2	6	0.330	0.284	-0.041	0.237	0.191
1	16	32	-0.175	-0.041	0.897	0.093	0.227
1	4	7	0.258	0.237	0.093	0.216	0.196
1	6	10	0.186	0.191	0.227	0.196	0.201

p.8.a. Give the leverage values for each observation. Do any exceed twice the average of the leverage values?

Observation 1 ________ Obs 2 ______________ Obs 3 ___________ Obs 4 ____________ Obs 5 ______________

p.8.b. Give $\boldsymbol{\beta}$, based on the following results

X'X		X'Y	INV(X'X)	
5	28	59	0.402	-0.036
28	312	612	-0.036	0.006

p.8.c. Compute SSE and S_e (Note: $\mathbf{Y'Y} = 1225$ $\mathbf{Y'PY} = 1207.144$)

p.8.d. The following table contains the fitted values with and without each observation, residual standard deviation when that observation was not included in the regression, and the regression coefficients when that observation was not included in the regression.

Y-hat	Y-hat(-i)	S_i	beta0_i	beta1_i
1.64	12.45	2.07	11.41	0.52
5.27	12.19	3.28	9.76	0.61
30.67	45.00	0.63	5.00	1.60
8.90	9.46	3.24	9.46	0.62
12.53	13.72	1.86	8.72	0.62

p.8.d.i. Compute DFFITS for the fifth observation

p.8.d.ii. Compute DFBETAS0 for the first observation

p.8.d.iii. Compute DFBETAS1 for the third observation

Q.9. For the F-test for Lack-of-Fit, where:

$$H_0: E\{Y_{ij}\} = \beta_0 + \beta_1 X_j \qquad H_A: E\{Y_{ij}\} = \mu_j \neq \beta_0 + \beta_1 X_j \quad j = 1,\ldots,c; \quad i = 1,\ldots,n_j$$

with: Residual $\equiv$ $Y_{ij} - \hat{Y}_j$ Pure Error $\equiv$ $Y_{ij} - \overline{Y}_j$ Lack of Fit $\equiv$ $\overline{Y}_j - \hat{Y}_j$ $\quad \hat{Y}_j = \hat{\beta}_0 + \hat{\beta}_1 X_j \quad \overline{Y}_j = \frac{1}{n_j}\sum_{i=1}^{n_j} Y_{ij}$

Show: $$\sum_{j=1}^{c}\sum_{i=1}^{n_j}\left(Y_{ij} - \hat{Y}_j\right)^2 = \sum_{j=1}^{c}\sum_{i=1}^{n_j}\left(Y_{ij} - \overline{Y}_j\right)^2 + \sum_{j=1}^{c}\sum_{i=1}^{n_j}\left(\overline{Y}_j - \hat{Y}_j\right)^2$$

Q.10. A simple linear regression model was fit, relating Weight (Y, in pounds) to Height (X, in inches) among a sample of n=9 Women's NBA basketball players. There were 3 distinct height levels, with 3 players per height. Use the table below to conduct the F-test for Lack of Fit. $H_0: E\{Y_{ij}\} = \beta_0 + \beta_1 X_j$ vs $H_A: E\{Y_{ij}\} = \mu_j \neq \beta_0 + \beta_1 X_j$

Height	Weight	Ybar(Grp)	Yhat(Grp)	PE	LF
68	130				
68	165				
68	155				
72	180				
72	165				
72	150				
76	200				
76	185				
76	230				

	Coefficients
Intercept	-321.667
Height	6.875

Q.11. A random sample of n = 20 cricket players from India was obtained, and a simple linear regression was fit relating the number of runs (Y, in 1000s) and number of completed innings (X, in 100s) in international matches. Player 20 is Sachin Tendulakar who played in 411 completed innings (X = 4.11) and accounted for 18426 runs (Y = 18.426). Regression models were fit with and without Sachin, and results are given below.

X'X			X'Y
20.0000	22.6600		77.1320
22.6600	44.7782		172.7793
INV(X'X)			Beta-hat
0.1172	-0.0593		-1.2074
-0.0593	0.0523		4.4696
Y'Y	Y'PY		
693.2798	679.1187		

X'X_(20)			X'Y_(20)	
19.0000	18.5500		58.7060	
18.5500	27.8861		97.0484	
INV(X'X_(20))			Beta-hat_(20)	
0.1501	-0.0999		-0.8785	
-0.0999	0.1023		4.0646	
Y'Y_(20)	Y'PY_(20)			
353.7623	342.8852			

p.11.a. Obtain SSE and MSE and the estimated error standard deviation for each model.

SSE_1 = __________________ MSE_1 = ______________________________ s = __________________________

SSE_2 = __________________ MSE_2 = ______________________________ $s_{(20)}$ = ________________________

p.11.b. Obtain the fitted value and residual for Satchin for each model.

Fitted_1 = _________________ Residual_1 = _______________ Fitted_2 = _________________ Residual_2 = _______________

p.11.c. Obtain the leverage value for Satchin from the full data model, and his studentized residual.

$v_{20,20} =$ ____________________________ $r_{20}^{*} =$ ______________________________

p.11.d. Compute $DFFITS_{20}$ and $DFBETAS_{1(20)}$ for Satchin. Note that both DFFITS and DFBETAS make use of $s_{(20)}$ to estimate σ.

$DFFITS_{20}$ = ____________________________ $DFBETAS_{1(20)}$ = ____

p.11.e. What is the average of the leverage values for the full data model?

Q.12. An experiment was conducted relating springiness in berries (Y, in mm) to sugar equivalent (X, in g/L). There were c = 4 distinct sugar equivalent groups, with n_j = 5 berries per group. The lack-of-fit test for a linear relation is:

$$H_0: E\left\{Y_{ij}\right\} = \mu_j = \beta_0 + \beta_1 X_j \quad i = 1,\ldots,5;\ j = 1,\ldots,4 \quad H_A: E\left\{Y_{ij}\right\} = \mu_j \neq \beta_0 + \beta_1 X_j$$

ANOVA												
	df	*SS*	*MS*	*F*	*gnificance F*		j	X_j	n_j	Yhat_j	Ybar_j	s^2_j
Regressio	1	245.74	245.74	57.40	0.0000		1	176.5	5	21.89	21.77	0.38
Residual	18	77.06	4.28				2	209.3	5	18.04	18.77	2.90
Total	19	322.80					3	225	5	16.20	15.41	11.30
							4	259.5	5		12.32	3.18
	Coefficients	*andard Err*	*t Stat*	*P-value*								
Intercept	42.6025	3.4019	12.5233	0.0000								
sugCont	-0.1174	0.0155	-7.5763	0.0000								

Note: $s_j^2 = \dfrac{\sum_{i=1}^{n_j}\left(Y_{ij} - \overline{Y}_j\right)^2}{n_j - 1} \quad n_j > 1,\ 0 \text{ otherwise}$

p.12.a. Give the fitted value for the linear regression for the 4th group (X_4 = 259.5).

p.12.b. Compute the Pure Error Sum of Squares, degrees of freedom and Mean Square.

SS_{PE} = ______________________________ df_{PE} = _______________ MS_{PE} = _____________________________

p.12.c. Compute the Lack-of-Fit Sum of Squares, degrees of freedom and Mean Square.

SS_{LF} = ______________________________ df_{LF} = _______________ MS_{LF} = _____________________________

p.12.d. Give the Test Statistic, Rejection Region, and P-value relative to .05 for the Lack-of-Fit test.

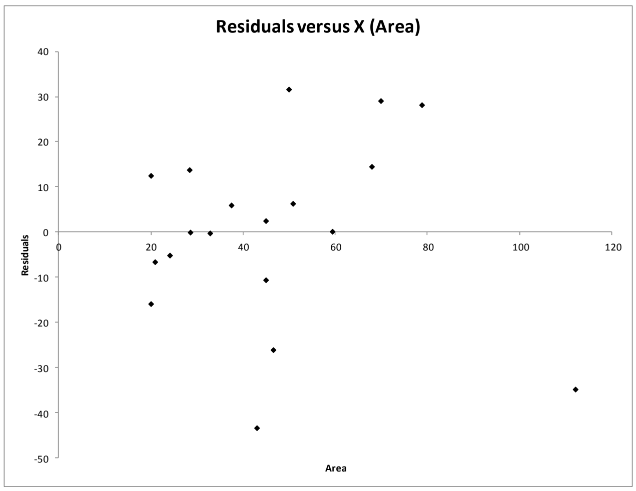
Test Statistic: ___________________ Rejection Region: ___________________ P-value **>** or **<** .05

Q.13. A regression model is fit, relating energy consumption (Y) to total area (X) for a sample of n = 19 luxury hotels in Hainan Province, China. The Analysis of Variance for the simple linear regression model is given below.

ANOVA					
	df	*SS*	*MS*	*F*	*gnificance*
Regressio	1	25521.76	25521.76	57.75	0.0000
Residual	17	7512.94	441.94		
Total	18	33034.70			

p.13.a. A plot of the residuals versus area is given below. It demonstrates which possible violations of assumptions (circle all that apply).

Non-normal Errors   Unequal Variance   Serial Correlation of Errors   Non-linear Relation between Y and X



p.13.b. A second regression model is fit, relating the squared residuals (Y) to area (X). Conduct the Breusch-Pagan test to test whether the equal variance assumption is reasonable. The sums of squares are given below.

ANOVA		
	df	*SS*
Regressio	1	1239379
Residual	17	3871645
Total	18	5111024

Test Statistic: ___________________ Rejection Region: ___________________ P-value **>** or **<** .05

Q.14. A random sample of n = 15 Bollywood movies was obtained, and a simple linear regression was fit relating the log Revenues (Y) the log Budget (X) in international matches. Movie 15 was **Sultan** with log Revenues Y_{15} = 5.705 and log Budget X_{15} = 4.500 . Regression models were fit with and without **Sultan**, and results are given below.

X'X			X'Y
15.000	55.474		53.796
55.474	214.736		213.087
INV(X'X)			Beta-hat
1.495	-0.386		-1.871
-0.386	0.104		1.476
Y'Y	Y'PY		
218.04	213.79		

X'X_(15)			X'Y_(15)	
14.000	50.974		48.091	
50.974	194.488		187.417	
INV((X'X)_(15))			Beta-hat_(15)	
1.563	-0.410		-1.609	
-0.410	0.112		1.385	
Y'Y_15	Y'PY_15			
185.50	182.26			

p.14.a. Obtain SSE and MSE and the estimated error standard deviation for each model.

SSE_1 = __________________ MSE_1 = ______________________________ s = __________________________

SSE_2 = __________________ MSE_2 = _______________________________ $s_{(15)}$ = ________________________

p.14.b. Obtain the fitted value for **Sultan** for each model and its leverage value based on the full data set.

Fitted_1 = _________________ Fitted_2 = _________________ Leverage = _______________

p.8.c. . Compute $DFFITS_{15}$ for **Sultan**. Note that DFFITS makes use of $s_{(15)}$ to estimate σ .

$DFFITS_{15}$ = ______________________________