ISyE4031 Regression and Forecasting
Practice Problems 2
Spring 2016

1. A collector of antique clocks sold at auction believes that the price received for the clocks may be modeled as a linear function of the age of the clocks and the number of bidders at the auction. After a preliminary study and collection of 32 observations, the collector came up with the following second−order model: $y = \beta_0+\beta_1 x_1+\beta_2 x_2 + \beta_3 x_1x_2 +\beta_4 x_2^2 +\varepsilon$, where y: price of a clock; x_1: age of a clock; x_2: the number of bidders. Considering the output below and by setting α = 0.05, answer the following questions.

```
The regression equation is
Price = - 262 + 2.26 Age + 14.2 Bidder + 1.13 AgeBid - 4.20 Bid^2

Predictor     Coef   SE Coef    T      P
Constant   -261.7     404.4  -0.65  0.523
Age         2.260     2.052   1.10  0.280
Bidder      14.21     60.83   0.23  0.817
AgeBid     1.1301    0.2186   5.17  0.000
Bid^2      -4.196     1.344  -3.12  0.004

S = 84.5116   R-Sq = 96.0%   R-Sq(adj) = 95.4%

Analysis of Variance

Source           DF       SS       MS       F      P

Regression        4  4606950  1151737  161.26  0.000
Residual Error   27   192840     7142
Total            31  4799790
```

a. Is the model useful as a whole? Apply an appropriate F−test (state the hypothesis, critical F value, conclusion, etc.) and confirm your conclusion by stating the p−value.

b. What is the effect of one-year increase in Age on the mean value of Price when there are 10 bidders?

c. By considering the p−values, state the statistically significant predictors and state which ones should be removed from the model, i.e., fill in the table below.

Variable	Significant? (Yes or No)	Remove? (Yes or No)	Why?
Age			
Bidder			
AgeBid			
Bid^2			

2. The weekly sales (in $1000 per week) for fast-food outlets in each of four cities were collected. The objective is to model sales (y) as a function of traffic flow (in thousands of cars), adjusting for city-to-city variations that might be due to size or other market conditions. The linear regression model is therefore: $y = \beta_0 + \beta_1 C_1 +\beta_2 C_2 + \beta_3 C_3 + \beta_4 x +\varepsilon$, where x: traffic flow,

$C_1 = \begin{cases} 1, & \text{if city 1} \\ 0, & \text{otherwise,} \end{cases}$ $C_2 = \begin{cases} 1, & \text{if city 2} \\ 0, & \text{otherwise,} \end{cases}$ $C_3 = \begin{cases} 1, & \text{if city 3} \\ 0, & \text{otherwise,} \end{cases}$,

and City 4 is the base level.

The following output was obtained.

SALES = 1.08 - 1.22 CITY_1 - 0.531 CITY_2 - 1.08 CITY_3 + 0.104 TRAFFIC

Predictor	Coef	SE Coef	T	P
Constant	1.0834	0.3210	3.37	0.003
CITY_1	-1.2158	0.2054	-5.92	0.000
CITY_2	-0.5308	0.2848	-1.86	0.078
CITY_3	-1.0765	0.2265	-4.75	0.000
TRAFFIC	0.103673	0.004094	25.32	0.000

S = 0.362307 R-Sq = 97.9% R-Sq(adj) = 97.5%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	4	116.656	29.164	222.17	0.000
Residual Error	19	2.494	0.131		
Total	23	119.150			

Answer the following questions by using *p* values and $\alpha = 0.05$.

a. Does City 2 have more expected sales than City 4? State the hypothesis that you consider explicitly.

b. What is the expected sales in City 3 when the traffic flow is recorded as 60,000 cars (or $x = 60$)?

c. From ANOVA, since the *p*-value = 0 < 0.05, we reject H_0: $\beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$. Rejection of this hypothesis implies the following conclusions, except (circle the incorrect conclusion):

i. The model as a whole is significant.
ii. The expected sales in all cities are different.
iii. At least one of the independent variables is significant.
iv. Based on this test only, we cannot say whether expected sales in those cities are identical or not.

d. Suppose that you modeled and solved the problem as a simple linear regression model: $y = \beta_0 + \beta_1 x + \varepsilon$, where *y*: Sales and *x*: Traffic flow. In order to decide on the significance of the cities, compare this reduced model and the complete model by using a partial (nested) *F* test. State the hypothesis, test statistic, and critical *F* value explicitly.

SALES = 0.018 + 0.108 TRAFFIC

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	111.34	111.34	313.75	0.000
Residual Error	22	7.81	0.35		
Total	23	119.15			

3. Screening Techniques.

a. In a linear regression study five predictors are being evaluated by using a stepwise regression. In step 1, the predictor, X_5 was selected with $t = 7.73$ and *p*-value = 0. By considering the quantities given below, write down the selected variables in step 2. Assume $\alpha_{entry} = \alpha_{remove} = 0.10$.

Step 2	Predictors	$X_1,\ X_2$	$X_1,\ X_3$	$X_1,\ X_4$	$X_1,\ X_5$	$X_2,\ X_3$
	t-stat for each X_j	-0.13, 5.92	1.52, 4.28	0.94, -0.32	0.22, 7.77	4.43, 1.67
	p-value for each X_j	0.898, 0.000	0.138, 0.000	0.354, 0.755	0.824, 0.000	0.000, 0.104
	Predictors	$X_2,\ X_4$	$X_2,\ X_5$	$X_3,\ X_4$	$X_3,\ X_5$	$X_4,\ X_5$
	t-stat for each X_j	1.88, -1.69	6.37, 4.03	4.49, 7.73	3.68, 3.94	-1.95, 8.22
	p-value for each X_j	0.069, 0.101	0.000, 0.000	0.000, 0.000	0.001, 0.000	0.059, 0.000

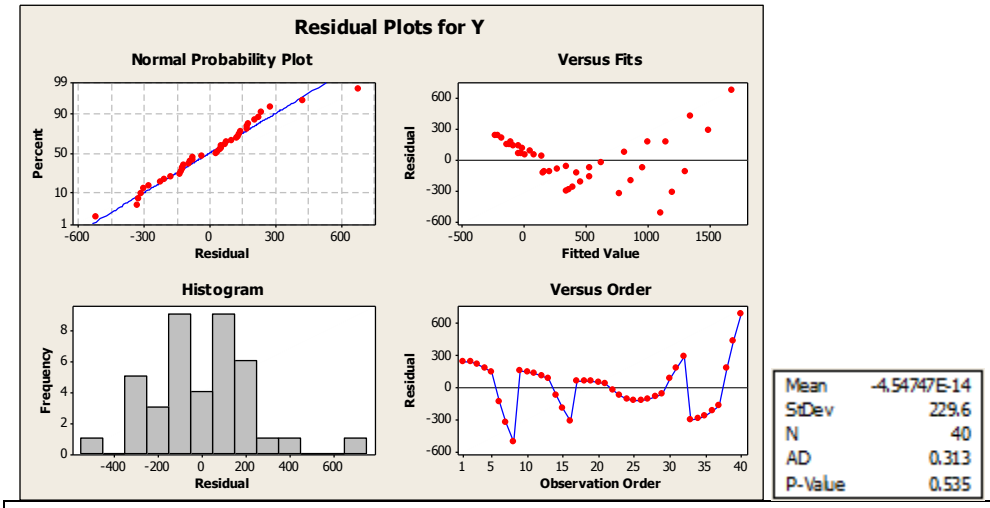
b. Suppose that you want to choose a subset among five independent variables, $X_1,\ X_2, X_3,\ X_4$, and X_5, by applying the best subsets technique. Consider the best subsets output given below. What is the best subset of variables? State your reasons.

Response is Y

Vars	R-Sq	R-Sq (adj)	R-Sq (pred)	Mallows Cp	S	X 1	X 2	X 3	X 4	X 5
1	62.4	61.4	58.7	9.0	1.2712				X	
1	50.0	48.6	45.3	23.2	1.4658		X			
2	66.1	64.2	60.6	6.8	1.2242				X	X
2	65.9	63.9	60.8	7.1	1.2288		X		X	
3	70.4	67.8	63.8	3.9	1.1613		X		X	X
3	68.0	65.2	59.3	6.6	1.2068	X			X	X
4	71.5	68.0	61.0	3.9	1.1618	X	X		X	X
4	70.5	66.9	61.0	5.8	1.1769		X	X	X	X
5	72.1	67.7	58.7	6.0	1.1625	X	X	X	X	X

4. Residual analysis and diagnostics.

a. A linear regression model, $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$, was fitted to 40 data points. After running the model, the model adequacy was investigated by performing a residual analysis.



Durbin-Watson statistic = 0.707338

Check whether the above results justify the basic random error term assumptions or not, i.e., $\varepsilon_i \sim$ *i.i.d.* $N(0, \sigma^2)$. Circle the correct answer (violated or not violated) and state the reasons explicitly by referring to the graph (name the plot(s)), apply statistical tests whenever possible (state the hypothesis and the critical value(s)), and use $\alpha = 0.10$.

i. $E[\varepsilon_i] = 0$: Violated / Not violated. Why?

ii. Each ε_i has a normal distribution: Violated / Not violated. Why?

iii. Each ε_i has an identical distribution: Violated / Not violated. Why?

iv. Each ε_i is independent: Violated / Not violated. Why?

b. A linear regression model, $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$, was fitted to $n = 25$ data points, and several test results for diagnostics were produced. The statistics to test whether an observation is unusual are given for 10 of those 25 observations below. Answer the following questions (circle the correct answer whenever required).

Obs.#	SRES	TRES	HI	COOK
1	−1.62768	−1.69563	0.101802	0.10009
2	0.36484	0.35754	0.070702	0.00338
3	−0.01609	−0.01572	0.098735	0.00001
4	1.57972	1.63916	0.085375	0.07765
5	−1.45000	−1.48962	0.391575	1.05105
6	−0.09081	−0.08874	0.042867	0.00012
7	0.27042	0.26465	0.081799	0.00217
8	0.36672	0.35939	0.063726	0.00305
9	3.21376	4.31078	0.498292	3.41932
10	0.81325	0.80678	0.196296	0.05385
⋮	⋮	⋮	⋮	⋮

i. Write down the observation number(s) of the high leverage point(s), if there are any. Which statistic(s) did you check?

ii. True or False? Only observation 5 is an influential point.

ii. True or False? Both observations 5 and 9 are outliers.

iii. True or False? Observation 9 is an outlier, high leverage point, and an influential point.

5. In a chemical experiment, the true relationship between yield (y) and reaction time (x) is assumed to be: $y = \theta_1 x^{\theta_2} e^{\varepsilon}$.

a. First, apply a transformation to the equation so that a simple linear regression solution can be found (write the additive form explicitly).

b. Then, consider the solution to the transformed model: $\hat{y}^* = -2.3 + 0.6\, x^*$. What is $\hat{y}$ when x is 5?

6. Answer the following short-answer questions (circle the correct answer whenever required).

a. True or False? If there are some independent variables having variance inflation factor (VIF) not less than 1, we can say that there exists multicollinearity between those independent variables.

b. What can we detect when t-tests for all (or nearly all) β parameters are nonsignificant whereas the F-test for overall model adequacy (H_0: $\beta_1 = \ldots = \beta_k = 0$) is significant?

c. True or False? Unusual observations can have dramatic effects on the least−squares estimates and may cause violation of random error assumptions.

d. True or False? If stepwise regression and best−subset selection produce different solutions, we must be suspicious of multicollinearity.