TMATH 390 Spring 2015 **Homework Set 4**

Chapter 3 questions: all but 1 should be done in R.

1. **(Interpreting output from statistical software)** Here is a relatively standard-looking output from a statistical software. The data deals with predicting concrete strength from its modulus of elasticity.

Predictor	Coeff	Stdev	t-ratio	p
Constant	3.2925	0.6008	5.48	0.000
mod elas	0.10748	0.01280	8.40	0.000

s = 0.8657 R-sq = 73.8% R-sq (adj) = 72.8%

Analysis of Variance

SOURCE	DF	SS	MS	F	p
Regression	1	52.870	52.870	70.55	0.000
Error	25	18.736	0.749		
Total	26	71.605			

Use your knowledge of regression and the various quantities that arise within regression to

(a) Identify the estimated intercept and slope in the regression equation, and interpret them.
(b) Identify SS_total, SS_explained, and SS_unexplained.
(c) Identify R^2 and state what it means (it is a percentage of. . .).

2. **(Transforming data)** The article "Reduction in Soluble Protein and Chlorophyll Contents in a Few Plants as Indicators of Automobile Exhaust Pollution" (Intl. J. of Environ. Studies, 1983: 239-244) reported the accompanying data on x distance from a highway (meters) and y lead content of soil at that distance (parts per million, or ppm):

$x:$	0.3	1	5	10	15	20	25	30	40	50	75	100
$y:$	62.75	37.15	29.70	20.71	17.65	15.41	14.15	13.50	12.11	11.40	10.85	10.85

(a) Construct scatter plots of y versus x, y versus $\ln(x)$, $\ln(y)$ versus $\ln(x)$ and $1/y$ versus $1/x$.
(b) Based on the results of part (a), which transformation does the best job of producing an approximate linear relationship?
(c) Use the selected transformation to predict lead content when distance is 45 meters.

3. **(Polynomial regression)** One frequently encountered problem in crop production is deciding when to harvest to maximize yield. Data on the time to harvesting (number of days after flowering) and the yield (kg/ha) of paddy—a grain farmed in India—appeared in the article "Determination of Biological Maturity and Effect of Harvesting and Drying Conditions on Milling Quality of Paddy" (J. of Agric. Engr., 1975: 353-361), and appears below.

(time to harvest) :	16	18	20	22	24	26	28	30
(paddy yield) :	2508	2518	3304	3423	3057	3190	3500	3883

(time to harvest) :	32	34	36	38	40	42	44	46
(paddy yield) :	3823	3646	3708	3333	3517	3241	3103	2776

(a) Is it possible to transform this data as described in this section so that there is an approximate linear relationship between the transformed variables? Why or why not? (Think about the goal of the researchers: to maximize yield.)

(b) Use a statistical computer package to fit a quadratic function to this data, and then predict yield when time to harvesting is 25 days. Assess the fit of the quadratic data, i.e., interpret the R^2 and the standard deviation about regression, s_e. Remember if you want to do quadratic regression of y versus x you should use this: `lm( y ~ x + I(x^2))` in R.

4. **(Polynomial regression two independent variables)** The article "The Undrained Strength of Some Thawed Permafrost Soils" (Canadian Geotech. J., 1979: 420-427) contained the accompanying data on y shear strength of sandy soil (kPa), x_1 depth (m), and x_2 water content (%).

Obs	Depth	Water	Strength
1	8.9	31.5	14.7
2	36.6	27.0	48.0
3	36.8	25.9	25.6
4	6.1	39.1	10.0
5	6.9	39.2	16.0
6	6.9	38.3	16.8
7	7.3	33.9	20.7
8	8.4	33.8	38.8
9	6.5	27.9	16.9
10	8.0	33.1	27.0
11	4.5	26.3	16.0
12	9.9	37.0	24.9
13	2.9	34.6	7.3
14	2.0	36.4	12.8

(a) Perform regression to predict y from x_1, x_2, x_1^2, x_2^2, and $x_1 \cdot x_2$. Remember to put $I()$ around any terms you're squaring. You don't need it around "`x1 * x2`". Write down the coefficients of the various terms.
(b) Compute R^2 and explain what it says about goodness-of-fit.
(c) Now perform regression to predict y from x_1 and x_2 only.
(d) Compute R^2 and explain what it says about goodness-of-fit.
(e) Compare the above two R^2 values. Does the comparison suggest that at least one of the higher order terms in the regression equation provides useful information about strength?

Note: the data for this problem are posted. Below is code that load the file.

```
dat = read.table("Hwk4_prob_4.dat", sep="&", header=T)
y = dat$Strength
x1 = dat$Depth
x2 = dat$Water
```

5. **(Interpreting more output)** An experiment carried out to study the effect of the mole contents of cobalt (x_1) and the calcination temperature (x_2) on the surface area of an iron cobalt hydroxide catalyst (y) resulted in the following data ("Structural Changes and Surface Properties of CoxFe3-xO4 Spinels," J. of Chemical Tech. and Biotech., 1994: 161-170):

x_1 :	0.6	0.6	0.6	0.6	0.6	1.0	1.0	1.0	1.0	1.0
x_2 :	200	250	400	500	600	200	250	400	500	600
y :	90.6	82.7	58.7	43.2	25.0	127.1	112.3	19.6	17.8	9.1

x_1 :	2.6	2.6	2.6	2.6	2.6	2.8	2.8	2.8	2.8	2.8
x_2 :	200	250	400	500	600	200	250	400	500	600
y :	53.1	52.0	43.4	42.4	31.6	40.9	37.9	27.5	27.3	19.0

A request to the SAS package to fit $\alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$ yielded the following output:

Dependent Variable: SURFAREA

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Prob > F
Model	3	15223.52829	5074.50943	18.924	0.0001
Error	16	4290.53971	268.15873		
Total	19	19514.06800			

Root MSE 16.37555 R-square 0.7801
Dep Mean 48.06000 Adj R-sq 0.7389
C.V. 34.07314

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T	Prob > |T|
INTERCEP	1	185.485740	21.19747682	8.750	0.0001
COBCON	1	-45.969466	10.61201173	-4.332	0.0005
TEMP	1	-0.301503	0.05074421	-5.942	0.0001
CONTEMP	1	0.088801	0.02540388	3.496	0.0030

(a) Interpret the value of the coefficient of determination R^2.

(b) Predict the value of surface area when cobalt content is 2.6 and temperature is 250.

(c) Since β_1 is about -46.0, is it legitimate to conclude that if cobalt content increases by 1 unit while the values of the other predictors remain fixed, surface area can be expected to decrease by 46 units? Explain your reasoning.