You are not allowed to discuss the exam problems with anyone except the instructor. You may pick any two questions from the following three questions for your take home exam for full credit.

Pr 3: Federal laws require that local units of government find alternatives to disposing of solid waste (also known as garbage) in landfills. Some local units of government have considered using solid waste as a fuel for energy plants. The idea is that the solid waste would be collected, then processed into a fuel, and finally the fuel would be used to produce electricity. Although the garbage is free, the cost of transportation and processing makes it much more costly than usual fuels (hydroelectric, coal, etc.). Consequently, burning garbage requires a subsidy.

This problem uses a set of data collected by a local unit of government to try understanding how the amount of garbage produced by a business varies as a function of business characteristics for 147 randomly selected business properties. The eventual goal is to get a prediction equation for all businesses, and these businesses will be taxed in proportion to their estimated garbage production.

The data are contained in the file waste.txt, which can be obtained from the class web site; go to the handouts and codes. Here is a listing of the variables:

```
FTE   Variate 147  number of full time equivalent employees
ImprV Variate 147  Value of the improvements to the parcel, in dollars
LandV Variate 147  Land value in dollars
Size  Variate 147  Total size of all buildings on the parcel, sq. ft.
Use   Variate 147  Type of commercial use
Waste Variate 147  waste production in tons per year
```

Use is a coded variable with Use = 2 if manfacturing, 3 if warehouse or storage, 4 if office building, 5 if retail, 6 if restaurant or entertainment. Your goal is to produce (and defend) (1) an understanding of how Waste depends on the other variables and (2) a prediction method.

What to Turn In. Your solution should consist of two parts, a ``Summary" and ``Supporting Evidence."

The **summary** will consist of: (1) a statement of your conclusions, with relevant summary statistics and probability statements. This should be at most 300 words. Your conclusions may be equivocal: for example, they might depend on whether or not a specific case is treated as an outlier. (2) AT MOST two graphical or numerical displays that are designed to convince someone familiar with statistical analysis that your analysis is sound, and that your conclusions are justified. Just giving a graph is NOT enough: you must explain what the graph shows and why it is interesting. *The summary should be understandable by an intelligent public official.*

Your **supporting evidence** will consist of AT MOST 500 words explaining how you got your answer, with as much computer output (text/figures) that you think is necessary to support your text. Unlabeled or unreferenced computer output will count against you.

Problem 1. A recent published paper concerning a response surface experiment included the following data set, which was named as "cp21.txt" and you can also find on our course web site under the **data** subdirector:

$x1$	$x2$	$x3$	y
-1.00000	-1.00000	-1.00000	0.926
-1.00000	-1.00000	1.00000	0.998
-1.00000	1.00000	-1.00000	1.072
-1.00000	1.00000	1.00000	1.091
1.00000	-1.00000	-1.00000	0.926
1.00000	-1.00000	1.00000	1.007
1.00000	1.00000	-1.00000	1.009
1.00000	1.00000	1.00000	1.058
-1.68170	0.00000	0.00000	1.232
1.68179	0.00000	0.00000	0.997
0.00000	-1.68179	0.00000	0.945
0.00000	1.68179	0.00000	1.231
0.00000	0.00000	-1.68179	0.927
0.00000	0.00000	1.68179	1.234
0.00000	0.00000	0.00000	1.245
0.00000	0.00000	0.00000	1.232
0.00000	0.00000	0.00000	1.212
0.00000	0.00000	0.00000	1.201
0.00000	0.00000	0.00000	1.222
0.00000	0.00000	0.00000	1.213

Here $x1, x2, x3$ were three variables whose optimal values the experimenter was trying to find, and y was the response variable.

(a) Fit the alternative models

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_{11} x_{i1}^2 + \beta_{22} x_{i2}^2 + \beta_{33} x_{i3}^2 + \epsilon_i \tag{1}$$

and

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_{11} x_{i1}^2 + \beta_{22} x_{i2}^2 + \beta_{33} x_{i3}^2 \\ &\quad \beta_{12} x_{i1} x_{i2} + \beta_{13} x_{i1} x_{i3} + \beta_{23} x_{i2} x_{i3} + \epsilon_i \end{aligned} \tag{2}$$

where x_{i1}, x_{i2}, x_{i3} are the observed values of $x1, x2, x3$ in row i. Which model do you prefer? Find the values of the coefficients, and their standard errors, and verify that each of $\hat{\beta}_{11}$, $\hat{\beta}_{22}$, $\hat{\beta}_{33}$ is negative, under either model.

(b) In the published paper, the researchers favored model 1. Assuming this model, let (x_1^*, x_2^*, x_3^*) be the point at which the expected response is maximized. Find point estimates for $x_j^*, j = 1, 2, 3$.

(c) Is this a sound analysis? What things might be wrong with it and how might they be corrected?

[*Note:* Run the standard regression diagnostics.]

Problem 2. This question is about an analysis of variance experiment, reinterpreted as a linear regression. It does not assume detailed knowledge about analysis of variance.

A recent paper discussed the following experiment related to the extraction of juice from blueberries. Three control variables were considered: temperature, level of sulfur dioxide (SO_2) and citric acid (coded as 0 or 1). Two response variables were measured: ACY (anthocynanin) and TP (total phenolics), both of which are considered to have beneficial health effects. The data were named as "cp22.txt" on the same web address as in Problem 1, and they were as follows:

Number	Temp	SO2	Citric	ACY	TP
	(deg C)	(ppm)	Acid		
1	50	0	0	27.5	55.9
2	50	0	1	42.6	62.6
3	80	0	0	50.2	71.4
4	80	0	1	62.4	88.8
5	50	50	0	92.2	307.3
6	50	50	1	96.5	316.4
7	80	50	0	97.5	420.6
8	80	50	1	102.2	413.8
9	50	100	0	90.6	386.0
10	50	100	1	82.2	337.5
11	80	100	0	92.1	641.0
12	80	100	1	91.4	684.3

Consider the model

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_k + \delta_{ij} + \eta_{ik} + \zeta_{jk} + \epsilon_{ijk}, \tag{3}$$

where α_i, $i = 1, 2$, β_j, $j = 1, 2, 3$, γ_k, $k = 1, 2$ are main effects due to temperature, SO_2 and citric acid respectively, δ_{ij}, η_{ik}, ζ_{jk} are interaction terms, and ϵ_{ijk} are independent $N(0, \sigma^2)$ errors. To make the model identifiable, assume any of α_i, β_j, γ_k, δ_{ij}, η_{ik}, ζ_{jk} is 0 when any of i, j, k is 1.

(a) Write the model (3) in the form $Y = X\beta + \epsilon$, where Y is the vector of responses (of dimension 12), the vector β consists of all the non-zero unknown parameters, and X is a design matrix of zeros and ones. (You should find that X is 12×10).

(b) Fit the model (3) to the data, where temperature, SO_2, and citric acid are the three factor variables and *ACY* is the response. Also consider both the square root and the log transformatins of the response and indicate which you prefer. (It is not necessary to give detailed tables of parameter values, but state the value of the residual sum of squares or the estimated s, and any other statistics that are directly relevant to the question).

(c) Now using whatever transformation you selected in (b), decide which of the main effects and interactions is significant. (Use the model selection criteria and your own judgement. Keep in mind that the goal is to find a setting that both ACT and TP are high.)

(d) Repeat the steps of (b) and (c) for TP response variable. (It's not necessary that the transformation of TP be the same as that for ACY).

(e) Write a short report on your conclusion for the company. Recall that the company's objective is to choose *one* setting of the three control variables so that both ACT and TP are high. Your report should indicate which settings you recommend, but should also make clear to what extent the differences among different possible settings are statistically significant, and whether you would recommend further experimentation.