Applied Statistics Comprehensive Examination
Regression Methods & Linear Models

Calculators are permitted on this part of the examination.

1. (25 points) An experiment was run to assess the impact of two types of fertilizer (1 and 2) and three types of soil (1, 2, and 3) on corn yield. Four plots were assigned to each of the six treatments, and the mean yield for each treatment was as tabled below.

	Soil Type 1	Soil Type 2	Soil Type 3
Fertilizer Type 1	55	65	70
Fertilizer Type 2	45	75	80

   a) (15 points) Fitting an effects model with interaction gives MSE=100. Using an appropriate contrast, test at level 0.05 for the presence of an interaction involving soil types 1 and 2 and fertilizer types 1 and 2.

   b) (10 points) Assuming an effects model with interaction, determine whether each expression is estimable or not. Here α_1 and α_2 are main effects for fertilizer type, and γ_{11} and γ_{22} are interaction effects. Justify your answers. (i) $\alpha_1 - \alpha_2$ (ii) $\alpha_1 - \alpha_2 + \gamma_{11} - \gamma_{22}$.

2. (25 points) An experiment was run to compare the yields for three wheat varieties. At each of four randomly selected locations in Ohio, each variety was assigned at random to four different plots. Assume a full model that includes a variety effect, a location effect, and an interaction term, and let y_{ijk} be the yield from the kth plot at the jth location using the ith variety.

   a) (10 points) State the mathematical model for the values $\{y_{ijk}\}$, making sure to include all assumptions.

   b) (10 points) Provide an ANOVA table that lists all factors, together with the degrees of freedom and the expected mean square for each factor.

   c) (5 points) For each variance component in your model from part (a), specify a function of the mean squares that gives an unbiased estimate of that variance component.

3. (50 points) Consider a regression equation on 18 observations where the dependent variable is the amount of suspended solids (mg/L) in a coal cleansing system. The independent variables are the pH of the cleansing tank (x_1), which is measured on a numeric scale, as well as the type of polymer used, which could be one of three types being examined. Let the indicator variables x_2 be 1 when polymer 1 is used and 0 otherwise, the indicator variable x_3 be 1 when polymer 2 is used and 0 otherwise, and the indicator variable x_4 be 1 when polymer 3 is used and 0 otherwise. Consider the model

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \varepsilon_i \text{ where } \varepsilon_i \sim iidN(0, \sigma_2)$$

For this model, $SS_{Reg} = 80182$ and $SS_{Total} = 85260$. The following table summarizes the regression coefficients:

Variable	Estimate	p-value
Intercept	-161.9	0.0007
x_1	54.3	0.0001
x_2	90.0	0.0001
x_3	27.2	0.0271

a) (5 points) State and conduct the global hypothesis test for the model at the 5% level of significance.

b) (5 points) State and conduct the hypotheses associated with the p-values in the table of regression coefficients above and interpret the results.

c) (5 points) What is the value of R-squared? Interpret this value.

d) (20 points) Below are residual plots for this model.

    i.   Examine the assumptions for this model using the residual plots and comment on whether the assumption seems appropriate.

    ii.   Which assumptions cannot be examined from these plots?

Residual Plots for y
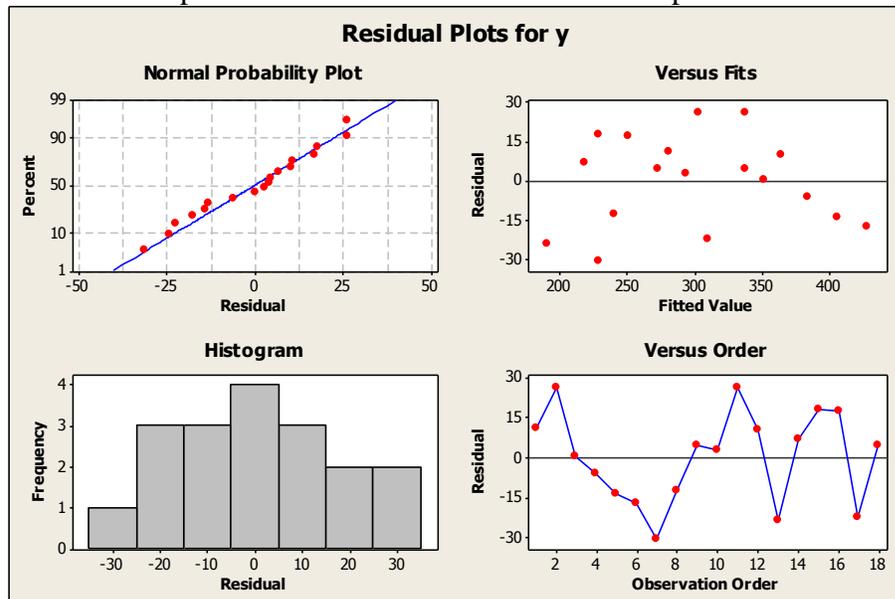
e) (10 points) State the model that includes appropriate interaction effects. Describe how you would determine whether the interactions are necessary to include.

f) (5 points) Someone suggested adding x_4 to the model. In one or two sentences, comment on this suggestion.