1. **ISCHEMIC HEART DISEASE** – A health insurance company collected information on 788 of its subscribers who had made claims resulting from ischemic (coronary) heart disease. Data include total costs of the services provided during 1998 and 1999 and some summaries of the types of services provided during that period. There is one line of data for each patient. There are 3 pages of output from a linear regression analysis of these data. The variables are denoted on the output as:

- obs = observation number
- cost = total cost in dollars of services provided
- age = age of subscriber in years
- gender = gender of subscriber (1=male, 0=female)
- intervnt = number of interventions/procedures
- drugs = number of drugs prescribed
- ervisits = number of emergency room visits
- complic = number of complications
- comorbid = number of other diseases
- duration = number of days treatment condition lasted.

(a) From the summary statistics provided on the top half of the first page of output:
- i. Do you expect the distribution of these variables to be approximately normal? Explain.
- ii. Do you expect collinearity to be a problem? Explain.

(b) The first page of output also includes results for Model 1 regressing cost on all of the predictors.
- i. Which predictors have coefficients significantly different from zero at the .05 level?
- ii. Based on the two residual plots do you believe the usual linear regression assumptions are satisfied? Explain.
- iii. The residual plot on the right appears to have strict border beyond which points can’t go (at the bottom left of the plot). Explain this.

(c) Model 2 (second page of output) uses the logarithm of cost as the response variable. (I actually took the logarithm of (cost + 1) to handle the few cases with zero reported cost.)
- i. Interpret the coefficient of “intervnt” in Model 2, that is, what is the expected effect on cost of an additional intervention/procedure?
- ii. Comment on the residual plots for Model 2.

(d) Model 3 was obtained by adding quadratic terms for each predictor, identifying which variables appeared to have a nonlinear relationship with log(cost), and finally searching for a suitable transformations to explain the nonlinear relationships. The residual plot is much improved (though still not perfect).
- i. There are 41 (internally) studentized residuals bigger than two in absolute value and 13 of these are bigger than three. What is your reaction to this many “outliers”?
- ii. The case with largest leverage has $h_{ii} = 2049$ and this case also has the highest Cook’s distance. This is the only patient with 3 complications (all others have zero or one). Is the leverage value unusually high? What would you recommend be done with this high leverage case.
- iii. A prediction interval is provided for the first 10 cases. Interpret the prediction interval for case 1.
2. SENIC HOSPITAL INFECTION – The Study on the Efficacy of Nosocomial Infection Control (SENIC) evaluated hospital-acquired infection rates in U.S. hospitals. Data are analyzed here from 113 hospitals across the country. One question concerns comparing the infection rates across the four regions of the country (east, central/midwest, south, and west... abbreviated as e, c, s, w here). There are two pages of output for this example (pages 4 and 5 of the attached). The output includes data from a number of models; not all are addressed in this question.

(a) The first bit of output provides the ANOVA table for a one factor analysis of variance that compares the mean infection rates across the four regions; a table of means is also provided. Test the hypothesis that all 4 regions have the same mean infection rate. Be sure to state your conclusion.

(b) The analysis of variance is followed by 5 regression models. These regression models include two important covariates: cult = the number of cultures performed per 100 patients (this is a measure of the number of infections in the hospital), and stay = average length of stay in the hospital (with longer stays presumed to increase the risk of infection). Some of the models included interactions, for example Model 1 includes c*stay = c * stay and similar terms for other interactions. The eastern region is taken as the reference group.

i. Interpret Model 2, i.e., what does the fitted model say about infection rates in the four regions? (You can answer by describing the model carefully or by giving examples about what the models says regarding infection rates in two or more of the four regions.)

ii. Test Model 3 vs Model 1 to determine whether the analysis of covariance model provides an adequate fit to these data. Be sure to state your inclusion.

iii. Model 5 eliminates some of the regional indicator variables. How do we interpret the coefficient of “w” in this model?

iv. The table of means accompanying the analysis of variance suggests that hospitals in the west are in the middle in terms of infection rate. However the regression models 1,2,3 and 5 all indicate that the west has the most positive coefficient indicating higher rates of infection. Explain this apparent paradox.

3. DRUGS EFFECT EXPERIMENT – This question considers data from an experiment that explores the effect of a particular drug on behavior in rats. The rats in the experiment were trained to press a lever for water when they are thirsty. The question is whether the drug leads to impaired performance on this task. The response variable is the number of lever presses per second (called “response” in the output). Before being given drugs each rat was characterized during the training process as being slow, moderate or fast in pressing the lever. This factor is referred to as “initrate” in the output (with 1=slow, 2=moderate, 3=fast). Four doses of the drug were used in the experiment (0, 0.5, 1, 1.8 where 0 refers to a control); these dosages are coded as 1,2,3,4 in the “dosage” factor. There are 8 measurements per combination of initrate and dosage. The results of an analysis of these data is provided on the final page of the output.

(a) This is an example of a two factor analysis of variance. Remember that even though there are two factors, you can think about these data as being just like a one-factor analysis of variance with 12 treatment levels (3 initrates x 4 dosages). Compute a 95% confidence interval for the linear contrast in dosage (i.e., the contrast with weights (3, 1, -1, -3)). Interpret the result.

(b) In one published analysis two contrasts were considered with weight vectors (1,1,1,-3) and (1,1,-1,-1). Both were significantly different from zero. The two contrasts suggest very different hypotheses about the mean response for the four dosages. Explain how both can be significant at the same time.

(c) It turns out that the whole experiment was carried out with only 12 rats. Each rat was trained and categorized as slow/moderate/fast; then the rat was used for two runs at each of the four doses. This means that each rat gives us 8 observations.

i. Describe a graphical approach that you might use to determine if there is correlation among the observations due to the multiple observations from each rat.

ii. It is possible to develop a statistical approach to assessing whether the rate effect is significant. The MS(Error) on the output includes the rat effect as well as the usual individual variation/error. Recall that there are two runs for each rat at each of four doses. Tell how you might be able to test whether there is a rat effect using these replicate data.

iii. Finally, assume that there is a rat effect (there is!), then explain how you might include the rat effect in the model. (There is more than one correct answer.)