ST505: Fall 2012 Homework 5. Due: Monday, Oct. 22 (START Of CLASS)

1. This problem works with the gypsy moth/virus data introduced in class (data courtesy of Greg Dwyer). This examining the proportion of gypsy moth larvae surviving as a function of the density of a virus which they are exposed to. Each observation corresponds to a bag in a tree, with variables (in the order given) of:
   tree= tree number
   vden = virus density
   totno = number in bag
   inf = number infected

   NOTE that all of the calculations for parts b) to g) will just mimic those done for the cholesterol and kishi examples done out in class, some Friday and coming next Monday.

   (a) First read in the data and create two new variables, psurv = 1 - (inf/totno) = proportion surviving and $x = 1/vden$. Working in R you'll also want to create a group variable associated with vden (which has replicates at different levels) using $group < -factor(vden)$ (see class examples also). In SAS you can specify vden as a class variable in proc means (getting descriptive statistics for each level of vden) and in proc anova; see example on page 63 of notes. Throughout psurv is the outcome variable (y):

   (b) Get the sample size, mean and standard deviation for psurv for each of the groups. In SAS you can use proc means with vden as a class variable (see similar example in notes). In R, you can use tapply, as done in the example in handout in class. I added a little piece to the R code (which is not in the class handout) that gives you summary measures for each group in a concise fashion; I'll point his out on Monday.

   (c) Get a one-way analysis of variance table with psurv as outcome and grouped on the level of vden (so proc anova in SAS with vden as a class variable and anova(lm(psurv)) in R. What hypothesis is being tested by the F statistic in the anova table? What is the conclusion?

   (You should be able to explain exactly how the sums of squares, degrees of freedom and means squares, F statistics and p-value are calculated here. You should try to do this but don't need to hand it in)

   (d) Without explicitly modeling psurv as a function of vden (i.e., in the model that just allows different means at each level of the virus density, test the hypothesis that the variance is constant across the vden groups. In R, you can use the bartlett.test (see example) since it is automatically available. It typically agrees with the other tests. In SAS you can use the hovtest=levene option in proc anova (there is also a hovtest=bf option which does Brown-Forsythe). State your conclusion.

   (e) Use the plot of psurv versus vden (or the residuals versus vden) AND a test for lack of fit to assess whether a simple linear regression model of psurv on vden is adequate here. Note that you need to run the simple linear regression of psurv on vden to construct the lack of fit test. Show how the lack of fit test is getting constructed and obtain the P-value for it. State your conclusion.

   (f) Repeat the previous problem but now considering regressing psurv on x = 1/vden.

   (g) Assuming the linear model for psurv on x = 1/vden is good, test for constant variance, by running Levene's test. Note, as done in class, with the examples, you need to run the regression and then get the residuals to feed into Levene's test. State your conclusion.

   (h) Using the residuals from the fit of psurv on x = 1/den, plot them versus tree number. Does it look like the model should account for tree effects in some manner?

2. Suppose we have a linear regression model $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$ where the ϵ_i follow the usual assumptions. Suppose we wanted to test $H_0 : \beta_0 = 0$ and $\beta_1 = 1$ (simultaneously). (This was of interest to the experiments in the acid rain monitoring project since they wanted to see if the assumption that the

measurements were unbiased, that is $E(Y_i) = X_i$, was reasonable.) To do this use the general linear test approach in Section 2.8 (I talked about this some on Friday, I'll follow up with more discussion on Monday). Note that the reduced (null) model here has no parameters in the model for $E(Y_i)$ and so $\hat{Y}_i$ will just be X_i. So, the only difference with respect to section 2.8 is that SSE(R) will be different and there is n degrees of freedom associated with it.