STA 2101/442 Assignment Nine[1]

Please bring printouts of your SAS log and list files to the quiz. Note that the log and list files *must be from the same run of SAS*. Marks will be deducted for errors or warnings. **Do not write anything on your log and list files in advance, except possibly your name and student number.** The non-computer questions are just practice for the quiz, and are not to be handed in.

1. For the usual fixed effects multiple regression model, let $\mathbf{W} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{e}$.
   (a) Simplify this expression for $\mathbf{W}$.
   (b) What is the probability distribution of $\mathbf{W}$?
   (c) Now you know whether $V(\mathbf{e})$ has an inverse. Why?

2. This question uses data from the furnace study described in Assignment 8. The data file is `furnace3.data`. There is a link from the course home page in case the one in this document does not work.

   You will see that the description of the data file in Assignment 8 is not completely accurate. This is typical. Furthermore, the client is spear fishing off a coral reef in Samoa, and is unavailable to answer questions. Please use common sense and do the best you can.

   Using SAS, fit a regression model in which the response variable is the average of energy consumption with vent damper in and vent damper out, and the explanatory variables are age of house, chimney height and type of chimney liner (3 categories). Use indicator dummy variables to represent type of chimney liner, and make Unlined the reference category. Please use `proc reg simple` instead of just `proc reg`. This way, you will get simple descriptive statistics including the means of house age and chimney height, which will be useful below.

   (a) Allowing for type of chimney liner and age of house, is chimney height related to energy consumption?
      i. Give the value of the test statistic, a number from your printout.
      ii. Give p-value, a number from your printout.
      iii. Do you reject the null hypothesis at $\alpha = 0.05$? Answer Yes or No.
      iv. If the answer is Yes, what do you conclude? Use plain, non-statistical language.
   (b) Controlling for for type of chimney liner and chimney height, is age of house related to energy consumption?
      i. Give the value of the test statistic, a number from your printout.
      ii. Give p-value, a number from your printout.
      iii. Do you reject the null hypothesis at $\alpha = 0.05$? Answer Yes or No.
      iv. If the answer is Yes, what do you conclude? Use plain, non-statistical language.
   (c) Taking chimney height and age of house into account, is type of chimney liner related to energy consumption?
      i. Give the value of the test statistic, a number from your printout.
      ii. Give p-value, a number from your printout.
      iii. Do you reject the null hypothesis at $\alpha = 0.05$? Answer Yes or No.

[1]Copyright information is at the end of the last page.

(d) Looking at the estimated regression coefficients and disregarding hypothesis tests for the moment, when you control for chimney height and age of house, for which type of chimney liner is estimated energy consumption greatest? For which type of liner is it least?

(e) Now it's important to decide which of these differences are real, and which ones might be due to chance. Still guided by the $\alpha = 0.05$ significance level and for the present not worrying about the problem of multiple testing, carry out tests of all pairwise comparisons of the chimney liner types, correcting for chimney height and age of house. Two of these tests are already part of the default output; you'll need to request only one custom test. Express your conclusion *briefly*, using non-statistical language.

(f) When you do an analysis like this, it's really helpful to present numbers for the average energy consumption of houses with different types of chimney liner. But you don't want to just give sample means, because these ignore chimney height and age of house, rather than controlling for them. A good solution is to calculate three $\widehat{Y}$ values, one for each liner type, with chimney height and age of house set to their *sample mean values*. That's the mean of the entire sample. Go ahead and do this. Please use `proc iml`, so that the numbers appear on your printout. Are they consistent with your answer to Question 2f?
