36-401 Homework #5 Due 10/5/00

1. (**10 pts**) This problem demonstrates variable selection for the wages dataset.

   Last homework we defined xlist_c("Educ","South","Sex","Exper","Union","Age","Marr"). For these 7 variables run stepwise() with method="forward":

   ```
   stepwise(wages[,xlist], wages$logWage, method="forward")
   ```

   To choose a model from the stepwise() results, look at $which and $f.stat. The f.stat column gives the F statistic to compare each model to the previous one. The models considered are in $which in the order that they are examined. The model being fit is the one with all the "T" covariates included and the "F" ones excluded. A common F statistic cutoff is 4. Using this cutoff find the last model with f.stat$\geq$ 4 (last model for which it was worthwhile to add another covariate). Which covariates are in this chosen model?

   Repeat for the "backward" method, finding the last model for which f.stat$<$ 4 (last model for which it was OK to drop a covariate). In this case, both methods give the same model, but that is not the case generally. (There is nothing to turn in for the "backward" part.)

2. (**10 pts**) This problem continues with variable selection for the wages dataset. Use source("sum.step.q") to load the sum.step() function.

   Carry out the stepwise() procedure using the seven covariates, logWage as the outcome, method="exhaustive", nbest=2, and plot=F.

   Use the result of this stepwise() function call as the input to sum.step(), referring to the 9/26 class notes or the header in the sum.step.q file. Find the best 2 models based on the BIC model selection criterion. Fit the models (just run lm() without doing graphs to save time since this is only a homework assignment, and not a real data analysis). Write out the regresson equations for the 2 models, using numbers, not, e.g. b_0. Explain why the intercepts are different.

Turn over

3. (**10 pts**) For the previous problem, find a 95% CI for the expected value of logWage when the covariates are

```
newdata_data.frame(Educ=13, South=0, Sex=0, Exper=18, Union=0, Age=36, Marr=1)
```

for both models. (Hints: Use predict with se=T. Use a t-table or the qt() formula in the 9/28 notes under CI's to get the multiplier for the standard error.) Convert the CI's to dollars.

4. (**10 pts**) Write a paragraph for your boss, a high ranking official in the Bureau of Labor Statistics, explaining how you might do further research and/or analysis to choose a model to be implemented on a web site that will allow citizens to see if they are being paid a fair wage.

5. (**10 pts, extra credit**) Assume that you will not have any contact with the concept of simple regression for a couple of years after you graduate. Then you will be required to carry out a simple Y on X regression for your boss. Write an approximately 1/2 page summary of the most important aspects of simple (not multiple) regression analysis that will allow you to quickly re-learn about this subject. (The audience for this writing assignment is yourself.)