36-401 Homework #4 Due 9/28/00

1. (**10 pts**) Get data from wages.dat. Don't forget to examine wages.txt.

```
wages_read.table("wages.dat", T)
wages$logWage_log10(wages$Wage)
names(wages); dim(wages); wages[1:4,]
```

   Perform a multiple regression of logWage on Educ+Age+Sex+South. Write out the regression equation. Interpret the R^2, coefficients and p-values.

2. (**10 pts**) Create a vector over a reasonable range of Education that contains 100 evenly spaced values as follows:

```
educvec_seq(min(wages$Educ), max(wages$Educ), len=100)
```

   Verify for yourself that the median age is 35.

   Now we will make conditional effects plots to help us understand the model. Note that the regression equation for males from the South for any given Education level and Age is logWage=0.3001+0.0356Educ+0.0054Age-0.0549 which simplifies to logWage=0.2452+0.0356Educ+0.0054Age. Therefore, the prediction for logWage for various levels of Education for 35 year old Southern males can be obtained in either of these ways:

```
y_0.2452+0.0356*educvec+0.0054*35
newdata_data.frame(Age=35,Educ=educvec,South=rep(1,100),Sex=rep(0,100))
y_predict(wage.lm, newdata)
```

   Now you can create a nice "conditional effects plot"

```
educvec_seq(min(wages$Educ), max(wages$Educ), len=50)
# Southern Males
newdata_data.frame(Age=35, Educ=educvec, South=rep(1,50), Sex=rep(0,50))
y_predict(wage.lm, newdata)
plot(educvec, 10^y, type="l", ylim=c(1.5,3.2), xlab="Education (years)",
     ylab="Wage: Dollars per Hour")
title("1985 CPS: logWages on Age+Education+Southern+Sex", cex=1.8)
```

Add the other 3 curves: Southern Females, Non-Southern Males, and Non-Southern Females. You will use lines(educvec, 10∧y, lty=i) to add lines to the existing plot. Let i go from 2 to 4 for these three added plots. Of course y is different each time. Add a legend as follows (click to place it on the plot):

```
legend(locator(1), c("35 y/o Southern Males", "35 y/o Southern Females",
       "35 y/o Non-Southern Males", "35 y/o Non-Southern Females"), lty=1:4)
```

Turn in the plot and explain what the plot tells us. Also explain why we get curves, not straight lines. Can you suggest a better order for the legends to improve the ease of interpretation of the plot?

3. (**10 pts**) Download and source collin.q. Carry out the following commands. (Be sure you know what the commands do.)

```
source("collin.q")
xlist_c("Educ","South","Sex","Exper","Union","Age","Marr")
pairs(wages[, c(xlist,"logWage")])
round(cor(wages[,c(1:7,11)]),3)
collin(wages[,xlist], cutoff=0.7)
```

Turn in the answers to these questions: What do the pairs plot and correlation matrix tell you about collinearity in this problem? What additional information is given by the multiple regressions of collin()? Also, use multiple regression, e.g. stepwise(), to find 2 covariates that explain more than 99% of the variability in age. What is happening between age and these 2 variables?