Computer Workshop 1

- You are expected to come prepared to the computer workshop. This means that you have to make sure that you understand the theoretical concepts behind the questions.
- Solutions will NOT be published on Blackboard. You are expected to raise your questions in the workshops.

This computer session uses the data set `nba.csv` which contains salary information and career statistics for 267 players in the National Basketball Association (NBA). The variables are:

Variable	Description
marr	=1 if married
wage	annual salary, thousands $
exper	years as a professional player
age	age in years
coll	years playing at college
games	average games per year
minutes	minutes per season
guard	=1 if guard
forward	=1 if forward
center	=1 if center
points	points per game
rebounds	rebounds per game
assists	assists per game
draft	draft number
allstar	all-star player
avgmin	minutes per game
black	=1 if black
children	=1 if has children

In this session, you should carry out an empirical analysis of this data set. The following questions should help structure this process:

(a) Download the data set from Blackboard and save it on your `h:` drive. Then open the data set in R and make it the default data set.

```
  nba <- read.csv( "h:\\nba.csv" )
  attach( nba )
```

(b) Have a look at the summary statistics of the data set.

```
summary( nba )
```

What is the average age of the players? How many play forwards?

(c) Plot a histogram of points-per-game.

```
hist( points )
```

(d) Produce a scatterplot of points-per-game versus years in league.

```
plot( exper, points )
```

Discuss what you find.

(e) Run a regression of points-per-game on years in league, age, years played in college and position dummies.

```
model1 <- lm( points~exper+age+coll+center+forward )
summary( model1 )
```

How would you interpret the coefficient estimates? Are all the explanatory variables statistically significant at the 5% level? Does the interpretation of the dummy variables differ from the other explanatory variables? What is the R^2? How would you interpret this number?

(f) Why do you think `coll` has a negative and statistically significant coefficient? *Hint: NBA players can be drafted before finishing their college careers and even directly out of high schools.*

(g) Look at the correlation matrix.

```
cor( cbind(exper,age,coll,center,forward) )
```

Do you need to worry about multicollinearity?

(h) Now consider an extension of the basic model. Generate a new variable which is experience squared and include it in the regression.

```
expersq <- exper * exper
model2 <- lm( points~exper+expersq+age+coll+center+forward )
summary( model2 )
```

Holding `age`, `coll`, `center` and `forward` fixed, at what value of experience does the next year of experience reduce points-per-game? Does this make sense?

(i) Now you want to explain the log(`wage`).

```
logwage <- log(wage)
model3 <- lm( logwage~points+exper+expersq+age+coll )
summary( model3 )
```

How do you interpret the results?

(j) Test whether `age` and `coll` are jointly significant in the regression from (i). What does this imply about whether age and education have a separate effect on wage, once productivity and senority are controlled for? *(Hint: You need to estimate both the unrestricted and the restricted model. Then you can use the anova() command to get the sums of squared residuals for the two models.)*