Introduction to Regression Analysis with R

The purpose of this exercise is to introduce you to the basics of doing regression analysis with R. Here we will use a 'toy' data set containing n=10 observations on the following variables:

```
      sex perstest therapy IE
John    M       26      32  3
Susan   F       24      40  4
Mary    F       22      44  8
Paul    M       33      44  4
```

Here, `therapy` is the outcome measure—improvement after therapy; `perstest` is a personality test, `IE` is a scale of internal-external locus of control and `sex` is a factor variable.

1. The data is stored in `the matlib` package. Read it into R using

```
library(matlib)
data(therapy)
```

2. It is useful to get an overview of the data by plotting. One simple way is a scatterplot matrix, either the default from `plot(therapy),` or a more informative version from the car package.

```
# reorder columns, for convenience in plots
therapy <- therapy[,c(3,2,4,1)]
# default pairs plot
plot(therapy)

library(car)
# add regression lines and data ellipses
scatterplotMatrix(therapy, smooth=FALSE, ellipse=TRUE, levels=0.68)

# another version, using a formula and conditioning on sex
scatterplotMatrix(~therapy+perstest+intext|sex, data=therapy,
    smooth=FALSE, ellipse=TRUE, levels=0.68)
```

3. Let's start using `perstest` as a single predictor. We use lm() and save the model object. These statements illustrate a standard way to examine the output, and make diagnostic plots.

```
mod1 <- lm(therapy ~ perstest, data= therapy)
print(mod1)
summary(mod1)
# model diagnostic plots
op <- par(mfrow=c(2,2))
plot(mod1)
par(op)
```

Q: Is there evidence that the `perstest`, by itself, significantly predicts `therap`y?

4. We can fit additional predictors simply by adding them to the right of the model

```
mod2 <- lm(therapy ~ perstest + IE, data=therapy)
summary(mod2)

mod3 <- lm(therapy ~ perstest + IE + sex, data=therapy)
summary(mod2)
```

Examine the coefficients for PERSTEST and INTEXT, as well as the R^2 for each model. What has changed from the 1-predictor to the 2- to the 3-predictor model?

5. Use `anova()` to compare `mod1, mod2, mod3`. Each line of output tests the additional fit of the current model over the previous one.

6. Make a table containing the results of your analyses for these three models

Model	b: perstest	b: intext	b: sx	MSE	R^2
1		xxx	xxx		
2			xxx		
3					

In the ‘b:’ columns, enter the parameter estimates and their Pr > \|t\| values. You can do some of this in R as shown below

```
# make a table of model statistics
models <- rbind(c(coef(mod1),intext=NA,sex=NA), c(coef(mod2),NA), coef(mod3))
rownames(models) <- paste0('mod', 1:3)
# extract R^2 values, and add to models summary table
Rsq <- unlist(lapply(list(mod1, mod2, mod3), function(x) summary(x)$r.squared))
models <- cbind(models, Rsq=Rsq)
models
```

Q: Why do you think the effect of the `perstest` as a predictor of `therapy` changes?

We’ll explore other ways of fitting models with multiple predictors starting next week.