1. The file airquality contains 111 observations taken from an environmental study that measured the four variables: ozone (surface concentration of ozone in New York, in parts per million), radiation (solar radiation), temperature (observed temperature, in degrees Fahrenheit) and wind (wind speed, in miles per hour) for 111 consecutive days. The study investigated the influence of solar radiation, temperature and wind speed on concentration of ozone.

   a. Fit a linear model to the data. Does it seem adequate?

   b. Fit an additive nonparametric model and compare it with the linear one.

   c. Perform projection pursuit regression trying various smoothing methods and different number of terms (one, two, three). Choose the most reasonable projection pursuit model. What are the resulting explanatory variables? Compare the results with those obtained on previous steps.

   d. Perform ACE to find transformations of variables that maximize the correlation. Do the resulting transformations hint on some parametric model?

   e. Summarize the results.

**Computational Note:** To get the data set air issue

attach(airquality)

and then you can call the variables simply by their names. NB R is case sensitive. The function lm(...) fits a linear model. The function gam performs the backfitting algorithm for additive models. To plot the resulting estimated functions and to print summary statistics issue

```
gam.obj<-gam(...)
plot(gam.obj, se = T)
print(summary(gam.obj))
```

The function anova provides an ANOVA table for comparison between nested models (but be careful to use its statistics for nonparametric models). To perform a projection pursuit regression use the function `ppr` in library `modreg`. Read carefully help comments about these functions.
The function ace performs the ACE algorithm and provides estimated values of transformed variables (a matrix tx and a vector ty respectively). You have to download the package **acepack** to be able to fit ACE. Note that to plot the resulting transformation of any variable you should first sort its original values. You can also get the correlation coefficients between transformed variables (see help for details).

2. The file iris (available in R) contains the data on 50 flowers from each of 3 species of iris: Setosa, Versicolor and Virginica (totally, 150 flowers). 4 measurements have been made on each flower: sepal length and width, and petal length and width. The data were collected by Edgar Anderson in 1935, Fisher was the first statistician to study it in 1936 and from that on it became a famous test case for various classification procedures. Fit classification trees to these data and comment on the results. Load the data by entering: `attach(iris)` or `data(iris)`. Download the package **rpart** from the CRAN site.

3. **Confidence Bands:** Refer to section 3.8.2 in Hastie and Tibshirani. Consider the following function (from Homework 2)

   \[ m_2(x) = 1.741[2x^{10}(1 - x)^2 + x^2(1 - x)^{10}] \]

   for \(0 \leq x \leq 1\). Assume the model

   \[ y = m(x) + \epsilon \]

   where \(\epsilon \sim N(0, \sigma^2)\). Generate data \(y, x\) for \(n = 150\), and \(\sigma = 1/218\) and design points \(x_i = (i - .5)/n\), \(i = 1, \ldots, n\). Compute 95% global confidence bands for \(m_2\) using the bootstrap method to estimate the distribution of \(\nu(g)\). The bootstrap replications should be at least 500.