1. Load the famous `iris` data set by typing `data(iris)` at the prompt. This data set contains measurements on 150 iris flowers of three different species.

(a) Produce a summary (like from the `summary` function) of the four length/width measurements for each species of iris.

(b) Draw a side-by-side boxplot of the petal lengths for the three species. By visual inspection of this plot, do you think the mean petal length is the same for each species?

(c) *Graduate only.* Use R to perform analysis of variance, or ANOVA for short, to precisely test the claim that the mean petal length is the same for each species; state your conclusion.

2. Write an R function `mymed` to compute the median of a numeric vector. Your function should allow the user to choose whether missing values (NA) are removed from the vector first before computing the median and, for this, the function `is.na` will be useful. Note that the median of a numeric vector is NA if at least one of its entries is NA.

3. A “regression through the origin” model may be used when specific knowledge about the problem at hand suggests that the response variable is zero if and only if the predictor variable is zero. For such problems, the model can be written as

\[ Y_i = \beta X_i + \varepsilon_i, \quad i = 1, \ldots, n, \]

where \( \varepsilon_1, \ldots, \varepsilon_n \) are iid \( \mathcal{N}(0, \sigma^2) \) random noise.

(a) Derive, by hand, the least-squares estimate of \( \beta \). That is, find the value of \( \beta \) that minimizes the function

\[ f(\beta) = \sum_{i=1}^{n} (y_i - \beta x_i)^2, \]

where \((x_i, y_i), i = 1, \ldots, n\), represents the given data. This is a relatively simple, single-variable calculus problem.

(b) One interesting application of this model is *Hubble’s law*, which states that the velocity at which a galaxy is moving away from earth is proportional to its distance from earth. In other words, if \( Y \) is the velocity of a galaxy and \( X \) is its distance from earth, then the relationship \( Y = \beta X \) should hold. Here \( \beta \) represents the unknown *Hubble Constant*. Measurements on distance (in Mpc) and recession velocity (in km/sec) are contained in the data set `451hw01.dat` on the website. Read this data file into R using `read.table` (file has a header), use the formula you derived in part (a) above to estimate \( \beta \) based on this data, and draw a scatterplot of the data with the fitted line overlaid.

(c) *Graduate only.* Read some help files online and explain how the “regression through the origin” can be done using the built-in R function `lm`.
4. Wigner’s semi-circle law appears in random matrix theory\(^1\) as a limiting distribution of eigenvalues of symmetric random matrices as dimension increases.

(a) Let \(X\) be a random matrix of dimension \(n \times n\) whose upper triangle is filled with iid \(N(0,1)\) random variables, and the lower triangle by symmetry. Let \(\lambda_1, \ldots, \lambda_n\) be the (random) eigenvalues of \(X\), and define the scaled eigenvalues \(\eta_i = n^{-1/2} \lambda_i\), \(i = 1, \ldots, n\). Use the R function `rnorm` to simulate two such matrices, one for \(n = 1000\) and the other for \(n = 5000\), and compute the corresponding sets of scaled eigenvalues.\(^2\) Use the R function `system.time` to measure how long it takes to find the eigenvalues of these two matrices.

(b) Draw two histograms, one for each \(n\). Overlay the density function of Wigner’s semi-circle law:

\[
    f(\eta) = (2\pi)^{-1/2} (4 - \eta^2)^{1/2}, \quad -2 \leq \eta \leq 2.
\]

Comment on the quality of the fit.

5. **Graduate only.** Let \(A\) be an invertible \(n \times n\) matrix with known inverse \(A^{-1}\). Let \(u\) and \(v\) be \(n \times 1\) vectors, and suppose our goal is to compute the inverse of \(A + uv^\top\), a rank-one perturbation of \(A\). Woodbury’s formula provides a rule for computing this inverse without any matrix inverse calculations. Show that

\[
    (A + uv^\top)^{-1} = A^{-1} - \frac{A^{-1}uv^\top A^{-1}}{1 + v^\top A^{-1}u}.
\]

(Hint: Just multiply the right-hand side above by \(A + uv^\top\) on either side and show that you get the identity matrix.)

\(^1\)Wigner’s result is old, but random matrix theory is still a hot topic. Lots of mathematicians, physicists, and statisticians are interested in various properties of the eigenvalues of large random matrices.

\(^2\)Read the documentation for the R function `eigen` for some tricks to help speed up the eigenvalue computation.