Stats 500 Homework 1

Online Submission to Canvas, Due date: 11:59pm, September 12, 2025

Problem 1. The dataset `teengamb` concerns a study of teenage gambling in Britain. After you install R (if it is not already installed on the system you are using) and the `faraway` package, you can type `library(faraway)` to load the package and `data(teengamb)` to load the data. This dataset is from a survey conducted to study teenage gambling in Britain and contains five variables:

- `sex`: $0 = \text{male}, 1 = \text{female}$
- `status`: socioeconomic status score based on parents' occupation
- `income`: in pounds per week
- `verbal`: verbal score in words out of 12 correctly defined
- `gamble`: expenditure on gambling in pounds per year

More details about the dataset can be found in Ide-Smith & Lea (1988), Journal of Gambling Behavior, 4, 110-118.

1. Make a numerical and graphical summary of the data, commenting on any features that you find interesting. For the variable "sex", assign the label "male" and "female" and ask R to treat it as a categorical variable before you compute the summary.
   Solution to this question should be no longer than 1.5 pages.

2. In your summary, report the means and medians of variables "income" and "gamble". Please comment on the relative locations of their means and medians, and explain why the means are larger than the medians.

3. How many different values are there for the variable "verbal"? (hint: help(unique))

   Based on the boxplot (and any other ways you can define and explain) of the variable "verbal", what could be the possible values of outlying verbal scores?

4. Suppose you are interested in how variables, such as "verbal", "income" and "gamble" differ for different "sex". Use numerical and/or graphical tool(s) to summarize the data for this purpose, commenting on any features that you find interesting. Limit the output you present to a quantity that a busy reader would find sufficient to get a basic understanding of your answer.

Hints: Some useful R functions for this homework can be found in the course documents that are available on Canvas. You can always type `help(subject)` to get detailed help on the `subject`, e.g. `help(plot)`. Or you can type `help.start()` to get interactive help with a search engine.

Problem 2. Simple linear regression

Consider a simple linear regression $Y_i = \beta_0 + \beta_1 X_i + e_i$, for $i = 1, \ldots, n$. Show that the OLS estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ have the following forms

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}, \quad \hat{\beta}_1 = \frac{\sum_{i=1}^n \left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{\sum_{i=1}^n \left(X_i - \bar{X}\right)^2}$$

where $\bar{X} = n^{-1} \sum_{i=1}^n X_i$ and $\bar{Y} = n^{-1} \sum_{i=1}^n Y_i$.

Problem 3. Multi-task regression (by Andrew Ng)

Thus far, we only considered regression with scalar-valued responses. In some applications, the response is itself a vector: $\mathbf{y}_i \in \mathbb{R}^{m \times 1}$. We posit the relationship between the features/predictors ($\mathbf{x}_i \in \mathbb{R}^{d \times 1}$) and the vector-valued response $\mathbf{y}_i$ is linear:

$$\mathbf{y}_i^T = \mathbf{x}_i^T B^* + error, \text{ for } i = 1, \cdots, n$$

where $B^* \in \mathbb{R}^{d \times m}$ is a matrix of regression coefficients. Here note that for the linear regression model in class, the dimension of response variable $\mathbf{y}_i$ is $m = 1$.

1. Express the sum of squared residuals (also called residual sum of squares, RSS) in matrix notation (*i.e.* without using any summations). Similarly to the linear regression model, the RSS is defined as
$$RSS(B) = \sum_{i=1}^n (\mathbf{y}_i^T - \mathbf{x}_i^T B)(\mathbf{y}_i^T - \mathbf{x}_i^T B)^T.$$
Hint: *work out how to express the RSS in terms of the data matrices*
$$\mathbf{X} = \begin{bmatrix} - & \mathbf{x}_1^T & - \\ & \vdots & \\ - & \mathbf{x}_n^T & - \end{bmatrix} \in \mathbb{R}^{n \times d}, \quad \mathbf{Y} = \begin{bmatrix} - & \mathbf{y}_1^T & - \\ & \vdots & \\ - & \mathbf{y}_n^T & - \end{bmatrix} \in \mathbb{R}^{n \times m}.$$
Also note that for a matrix $A = (a_{ij})_{n \times m}$ with its ith row vector denoted by $\mathbf{a}_i$, we have $tr(AA^\top) = \sum_{i=1}^n \mathbf{a}_i \mathbf{a}_i^\top = \sum_{1 \le i,j \le n} a_{ij}^2$.
2. Find the matrix of regression coefficients that minimizes the RSS.
3. Instead of minimizing the RSS, we break up the problem into m regression problems with scalar-valued responses. That is, we fit m linear models of the form
$$(\mathbf{y}_i)_k = \mathbf{x}_i^T \beta_k + error,$$
where $(\mathbf{y}_i)_k$ denotes the kth element in the vector $\mathbf{y}_i$ and $\beta_k \in \mathbb{R}^d$. How do the regression coefficients from the m separate regressions compare to the matrix of regression coefficients that minimizes the SSR in question (2).