MATH 567: Mathematical Techniques in Data Science
Lab 1

Dominique Guillot

Departments of Mathematical Sciences
University of Delaware

February 8, 2016

Exercise 1

1. Matrix/vectors

1. Construct two 4×4 random matrices A, B with entries uniformly distributed in $[0, 1]$.
2. Compute the matrix product of A and B, and the entrywise product of A and B.
3. Compute the determinant of A.
4. Compute the eigenvalues and the associated eigenvectors of A.
5. Construct a random vector $b \in \mathbb{R}^4$ with $N(0, 1)$ entries.
6. Solve the linear system $Ax = b$.
7. Compute A^{-1}. Verify your previous solution by computing $A^{-1}b$ explicitly.

Exercise 2

2. Cars data

1. Load the `ISLR` library (`library(ISLR)`). (Install the `ISLR` package first if necessary).
2. Load the `Auto` dataset (`data(Auto)`).
3. Read the documentation (`?Auto`).
4. Use the `fix` function to look at the data.
5. Extract the first row from the table.
6. Extract the “mpg” column from the table.
7. Compute summary statistics for the data (`summary(Auto)`). Do you understand the output?
8. Make a plot of “mpg” as a function of “weight”.
9. Construct a histogram for the “mpg” values.
10. Use the command `pairs` to produce scatter plots of all pairs of variables. Save the plot in pdf to better visualize it.
11. Examine the relation between a subset of the variables: `pairs(~ mpg + horsepower + weight)`.

Exercise 3

3. Linear regression

Let's try to identify *linear relationships* between variables.

mpg	horsepower	weight
18	130	3504
15	165	3693
18	150	3436
⋮	⋮	⋮

Exercise 3

3. Linear regression

Let's try to identify *linear relationships* between variables.

mpg	horsepower	weight
18	130	3504
15	165	3693
18	150	3436
⋮	⋮	⋮

General case: $Y = \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon$.

3. Linear regression

Let's try to identify *linear relationships* between variables.

mpg	horsepower	weight
18	130	3504
15	165	3693
18	150	3436
⋮	⋮	⋮

General case: $Y = \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon$.
Vector form: $\mathbf{Y} = \mathbf{X}\beta + \epsilon$ with $\mathbf{Y}, \epsilon \in \mathbb{R}^n$, $\mathbf{X} \in \mathbb{R}^{n \times p}$, $\beta \in \mathbb{R}^p$.
Goal: Find the coefficients $\beta_1, \ldots, \beta_p$ that minimize the “error” ϵ.

Least squares approach

We measure the *error* in the fit

$$Y = \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon$$

by the *mean squared error*:

$$\mathrm{MSE}(\beta) = \frac{1}{n}\|\mathbf{Y} - \mathbf{X}\beta\|_2^2 = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2 .$$

Least squares approach

We measure the *error* in the fit

$$Y = \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon$$

by the *mean squared error*:

$$\mathrm{MSE}(\beta) = \frac{1}{n}\|\mathbf{Y} - \mathbf{X}\beta\|_2^2 = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2.$$

We solve:

$$\hat{\beta}_{\mathrm{LS}} := \operatorname*{argmin}_{\beta\in\mathbb{R}^p} \|\mathbf{Y} - \mathbf{X}\beta\|_2^2.$$

Least squares approach

We measure the *error* in the fit

$$Y = \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon$$

by the *mean squared error*:

$$\mathrm{MSE}(\beta) = \frac{1}{n}\|\mathbf{Y} - \mathbf{X}\beta\|_2^2 = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2.$$

We solve:

$$\hat{\beta}_{\mathrm{LS}} := \operatorname*{argmin}_{\beta \in \mathbb{R}^p} \|\mathbf{Y} - \mathbf{X}\beta\|_2^2.$$

In R:

```
model <- lm(Auto$mpg ~ Auto$horsepower + Auto$weight)
```