Least Squares Linear Regression

Sanjay Lall and Stephen Boyd

EE104
Stanford University
Least squares linear regression

- linear predictor $\hat{y} = g_{\theta}(x) = \theta^T x$
- $\theta \in \mathbb{R}^d$ is the model parameter
- we’ll use square loss function $\ell(\hat{y}, y) = (\hat{y} - y)^2$
- empirical risk is MSE
  \[ L(\theta) = \frac{1}{n} \sum_{i=1}^{n} (\theta^T x^i - y^i)^2 \]
- ERM: choose model parameter $\theta$ to minimize MSE
- called *linear least squares fitting* or *linear regression*
Least squares formulation

- express MSE in matrix notation as

\[
\frac{1}{n} \sum_{i=1}^{n} (\theta^T x^i - y^i)^2 = \frac{1}{n} \|X \theta - y\|^2
\]

where \( X \in \mathbb{R}^{n \times d} \) and \( y \in \mathbb{R}^n \) are

\[
X = \begin{bmatrix}
(x^1)^T \\
\vdots \\
(x^n)^T
\end{bmatrix} \quad y = \begin{bmatrix}
y^1 \\
\vdots \\
y^n
\end{bmatrix}
\]

- ERM is a least squares problem: choose \( \theta \) to minimize \( \|X \theta - y\|^2 \)
  (factor 1/n doesn’t affect choice of \( \theta \))
Least squares solution

(see *Introduction to Applied Linear Algebra: Vectors, Matrices, and Least Squares*)

- assuming \( X \) has linearly independent columns (which implies \( n \geq d \)), there is a unique optimal \( \theta \)

\[
\theta^* = (X^T X)^{-1} X^T y = X^\dagger y
\]

- standard algorithm:
  - compute QR factorization \( X = QR \)
  - compute \( Q^T y \)
  - solve \( R\theta^* = Q^T y \) by back substitution

- in Julia: \( \text{theta}_{\text{opt}} = X \backslash y \)

- complexity is \( 2d^2 n \) flops
Data matrix

- the $n \times d$ matrix

$$X = \begin{bmatrix} (x^1)^T \\ \vdots \\ (x^n)^T \end{bmatrix}$$

is called the data matrix

- $i$th row of $X$ is $i$th feature vector, transposed
- $j$th column of $X$ gives values of $j$th feature $x_j$ across our data set
- $X_{ij}$ is the value of $j$th feature for the $i$th data point
Constant fit

- the simplest feature vector is constant: \( x = \phi(u) = 1 \) (doesn’t depend on \( u \)!)  

- corresponding predictor is a constant function: \( g(x) = \theta_1 \)  

- data matrix is \( X = 1_n \)  

- so \( X^\dagger = (X^T X)^{-1} X^T = (1/n) 1^T \) and \( \theta^* = X^\dagger y = 1^T y/n = \text{avg}(y) \)  

- the average of the outcome values is the best constant predictor (for square loss)  

- optimal RMSE is standard deviation of outcome values

\[
\left( \frac{1}{n} \sum_{i=1}^{n} (\text{avg}(y) - y^i)^2 \right)^{1/2}
\]
Regression

- with $u \in \mathbb{R}^{d-1}$: $x = \phi(u) = (1, u)$
- same as $x_1 = 1$ (the first feature is constant)
- predictor has form
  $$\hat{y} = \theta^T x = \theta_1 + \theta_{2:d}^T u$$
- an affine function of $u$
Straight line fit

- with $u \in \mathbb{R}$, $x = (1, u) \in \mathbb{R}^2$
- model is $\hat{y} = g(x) = \theta_1 + \theta_2 u$
- this model is called straight-line fit
- when $u$ is time, it’s called the trend line
- when $u$ is the whole market return, and $y$ is an asset return, $\theta_2$ is called ‘$\beta$’
data from Federal Highway Administration road monitoring stations

- total number of vehicle-miles traveled per year in U.S.
Constant versus straight-line fit models

- for the constant model, we choose $\theta_1$ to minimize

$$\frac{1}{n} \sum_{i=1}^{n} (\theta_1 - y^i)^2$$

- for the straight-line model, we choose $(\theta_1, \theta_2)$ to minimize

$$\frac{1}{n} \sum_{i=1}^{n} (\theta_1 + \theta_2 u^i - y^i)^2$$

- for optimal choices, this value is less than or equal to the one above (since we can take $\theta_2 = 0$ in the straight-line model)

- so the RMS error of the straight-line fit is no more than the standard deviation
Example: Diabetes

- $u$ consists of 10 explanatory variables (age, bmi, ...)
- with constant feature $x_1 = 1$, $x \in \mathbb{R}^{11}$
- outcome $y$ is measure of diabetes progression over after 1 year
- we’d like to predict $y$ given the features
- constant model (mean) is $g(x) = 152$, with MSE 5930, RMS error 77
Example: Diabetes

scatter plots of each explanatory variable versus $y$

data from https://web.stanford.edu/~hastie/Papers/LARS/
a separate regression of each variable against $y$

best single predictor is BMI, with MSE 3890
Straight-line fit with BMI

- left-hand plot shows optimal predictor $\hat{y} = -118 + 10.2 \text{ bmi}$
- right-hand plot shows $y$ versus $\hat{y}$
- ideal plot would have all points on the diagonal
Regression with all explanatory variables

- left-hand plot uses only BMI to predict $y$, achieves loss $\approx 3890$
- right-hand plot uses all features, achieves loss $\approx 2860$
- model is

$$g(x) = -335 - 0.0364 \text{age} - 22.9 \text{sex} + 5.6 \text{bmi} + 1.12 \text{bp} - 1.09s_1$$

$$+ 0.746s_2 + 0.372s_3 + 6.53s_4 + 68.5s_5 + 0.28s_6$$