Least Squares Linear Regression

Jong-Han Kim

EE787 Machine learning
Kyung Hee University

Least squares linear regression

- linear predictor $\hat{y} = g_\theta(x) = \theta^\mathsf{T} x$
- $\theta \in \mathbf{R}^d$ is the model parameter
- we'll use square loss function $\ell(\hat{y}, y) = (\hat{y} - y)^2$
- empirical risk is MSE

$$\mathcal{L}(\theta) = \frac{1}{n} \sum_{i=1}^{n} (\theta^\mathsf{T} x^i - y^i)^2$$

- ERM: choose model parameter θ to minimize MSE
- called *linear least squares fitting* or *linear regression*

Least squares formulation

- express MSE in matrix notation as

$$\mathcal{L}(\theta) = \frac{1}{n}\sum_{i=1}^{n}(\theta^{\mathsf{T}}x^i - y^i)^2 = \frac{1}{n}\left\{(\theta^{\mathsf{T}}x^1 - y^1)^2 + \cdots (\theta^{\mathsf{T}}x^n - y^n)^2\right\}$$

$$= \frac{1}{n}\left\| \begin{bmatrix} (x^1)^{\mathsf{T}}\theta - y^1 \\ \vdots \\ (x^n)^{\mathsf{T}}\theta - y^n \end{bmatrix} \right\|^2$$

$$= \frac{1}{n}\left\| \underbrace{\begin{bmatrix} (x^1)^{\mathsf{T}} \\ \vdots \\ (x^n)^{\mathsf{T}} \end{bmatrix}}_{X}\theta - \underbrace{\begin{bmatrix} y^1 \\ \vdots \\ y^n \end{bmatrix}}_{y} \right\|^2 = \frac{1}{n}||X\theta - y||^2$$

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

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 θ

$$\theta^\star = (X^\mathsf{T} X)^{-1} X^\mathsf{T} y = X^\dagger y$$

- standard algorithm:
  - compute QR factorization $X = QR$ (*e.g.*, Gram-Schmidt) (with orthogonal Q and invertible upper triangular R)
  - compute $Q^\mathsf{T} y$
  - solve $R\theta^\star = Q^\mathsf{T} y$ by back substitution
- in Julia: `theta_opt = X\y`
- complexity is $2d^2 n$ flops

Data matrix

- the $n \times d$ matrix

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

  is called the *data matrix*

- ith row of X is ith feature vector, transposed
- jth column of X gives values of jth feature x_j across our data set
- X_{ij} is the value of jth feature for the ith 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 $X = \mathbf{1}_n$
- so $X^\dagger = (X^\mathsf{T} X)^{-1} X^\mathsf{T} = (1/n)\mathbf{1}^\mathsf{T}$ and

$$\theta^\star = X^\dagger y = \mathbf{1}^\mathsf{T} y/n = \mathrm{avg}(y)$$

- *the average of the outcome values is the best constant predictor* (for square loss)
- optimal RMSE is standard deviation of outcome values

$$\left(\frac{1}{n}\sum_{i=1}^{n}(\mathrm{avg}(y) - y^i)^2\right)^{1/2}$$

Regression

- with $u \in \mathbf{R}^{d-1}$: $x = \phi(u) = (1, u)$
- same as $x_1 = 1$ (the first feature is constant)
- predictor has form

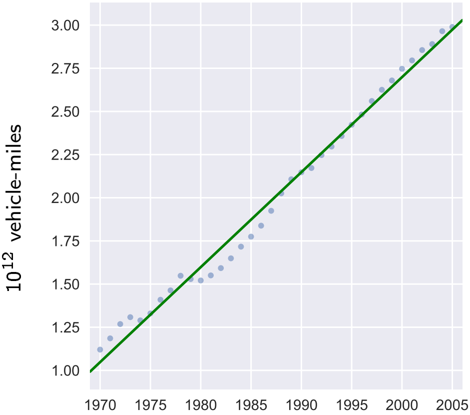
$$\hat{y} = \theta^{\mathsf{T}} x = \theta_1 + \theta_{2:d}^{\mathsf{T}} u$$

  an affine function of u

Straight line fit

- with $u \in \mathbf{R}$, $x = (1, u) \in \mathbf{R}^2$
- 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, θ_2 is called 'β'

Straight line fit



- 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 θ_1 to minimize

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

- for the straight-line model, we choose (θ_1, θ_2) to minimize

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

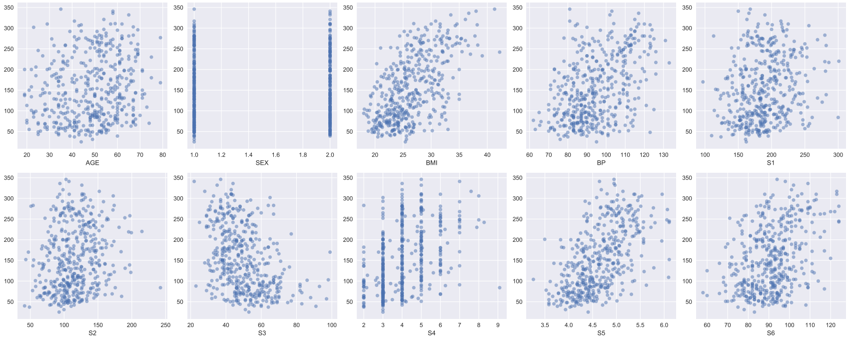
- for optimal choices, this value is less than or equal to the one above (since we can take $\theta_2 = 0$ in the straight-line model)

- 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 $x_1 = 1$, $x \in \mathbf{R}^{11}$
- 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

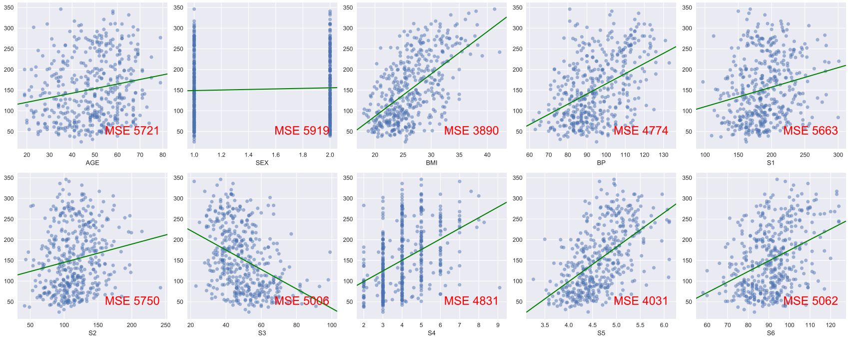
Example: Diabetes



- scatter plots of each explanatory variable versus y

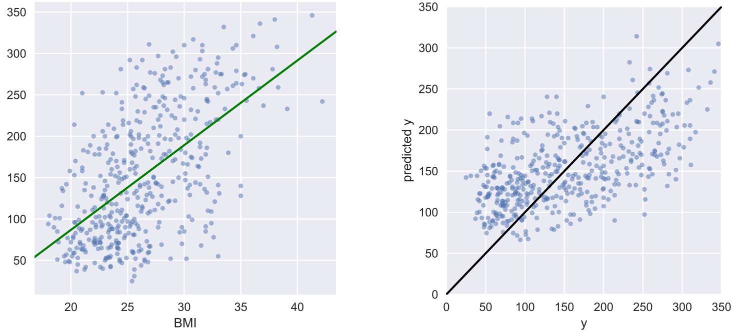
data from https://web.stanford.edu/~hastie/Papers/LARS/

Straight-line fits using each explanatory variable



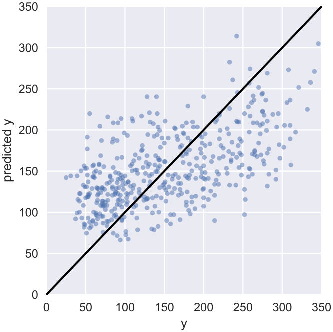
- a separate regression of each variable against y
- best single predictor is BMI, with MSE 3890

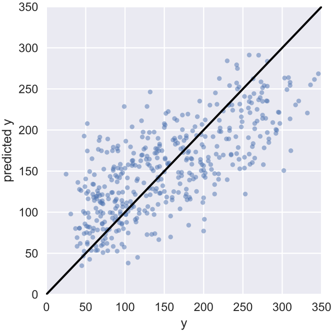
Straight-line fit with BMI



- left-hand plot shows optimal predictor $\hat{y} = -118 + 10.2\,\mathrm{bmi}$
- 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 ≈ 3890
- right-hand plot uses all features, achieves loss ≈ 2860
- model is

$$g(x) = -335 - 0.0364\,\mathrm{age} - 22.9\,\mathrm{sex} + 5.6\,\mathrm{bmi} + 1.12\,\mathrm{bp} - 1.09 s_1 \\ + 0.746 s_2 + 0.372 s_3 + 6.53 s_4 + 68.5 s_5 + 0.28 s_6$$