Least Squares Linear Regression

Discussion 7

Japjot Singh   Raguvir Kunani

Data 100

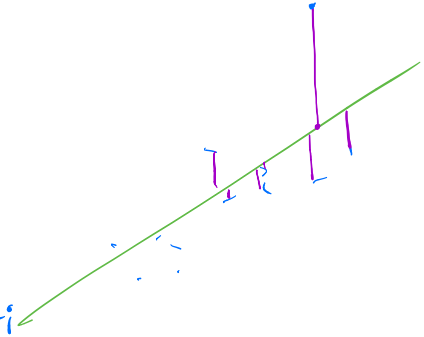
October 13, 2020

Recap

$\frac{1}{n}\sum |y_i - \hat{y}_i|$ $\quad \frac{1}{n}\sum (y_i - \hat{y}_i)^2$ $\quad [1, 2, 3, -3, 4]$ $\quad \theta_0 \cdot 1$

- If we have a constant model then our model is $\hat{y} = \theta_0$ and it only captures the distribution of a single variable (summary statistic like mean, median depending on loss function)
- If our model is linear in X then $\hat{Y} = \theta_0 + \theta_1 X = a + bX$ for some $a, b \in \mathbb{R}$
- Pearson's correlation coefficient r measures strength of linear association between two variables
  - $r \in [-1, 1]$
  - if $r = 0$ then our two variables are uncorrelated
    - correlation does NOT mean causation
    - correlation gives no information about non-linear association
  - $r = \frac{1}{n} \sum_{i=1} (\frac{x_i - \overline{x}}{\sigma_X})(\frac{y_i - \overline{y}}{\sigma_Y})$
    - with some manipulation we see $\sigma_{X,Y} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y}) = r\sigma_X \sigma_Y$

Simple Linear Regression



Starts with a simple regression model

$$\hat{y} = a + bX$$

$\hat{y}_i = a + b x_i$

Choose squared loss (L2 loss)

$$(y_i - \hat{y}_i)^2 = (y_i - (a + bx_i))^2$$

Average across the entire dataset (MSE)

$$L\underbrace{(a, b)}_{\theta} = \frac{1}{n}\sum_{i=1}^{n}(y_i - (a + bx_i))^2$$

Solving for optimal model parameters we have

$$\hat{b} = r\frac{\sigma_y}{\sigma_x} \qquad \hat{a} = \overline{y} - \hat{b}\overline{x}$$

Multiple Linear regression

The multiple linear regression model is linear in its features

$[\theta_0 \quad \theta_1 \quad \theta_2 \quad \cdots \quad \theta_p]$

$$\hat{y} = \underset{\substack{= \\ 1}}{\theta_0 x_0} + \theta_1 x_1 + \theta_2 x_2 + \ldots + \theta_p x_p = \theta_0 x_0 + \sum_{j=1}^{p} \theta_j x_j$$

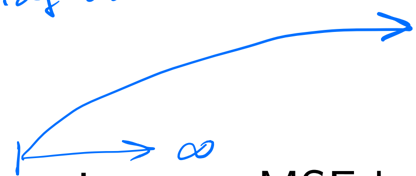
$[x_0 \quad x_1 \quad x_2 \quad \cdots \quad x_p]$

SLR → Multiple

RMSE stay same or ↓

R^2 stay same or ↑

- This model has p features $x_{1:p}$
- The weight of feature x_j is θ_j
- if we let $x_0 = 1$ then $\hat{y} = \sum_{j=0}^{p} \theta_j x_j$



Root Mean Square (RMSE) is just $\sqrt{\text{MSE}}$

- We do this because RMSE has the same units as y, MSE has units of y^2
- adding features cannot increase RMSE

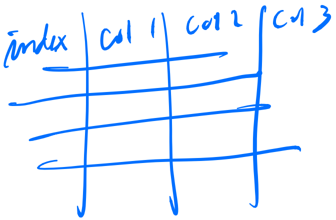
Multiple R^2 is the square correlation between true y and predicted $\hat{y}$, tells us the proportion of variance (information) of y that our fitted features (model) explains

$$R^2 = [r(y, \hat{y})]^2 = \frac{\sigma^2_{\hat{y}}}{\sigma^2_y}$$

Vectorized Multiple Regression

$[-4, 7623, 5] \in \mathbb{R}^2$

$\mathbb{R}^n$ is a vector space, we can think of it as the set of all lists of length n of elements of $\mathbb{R}$, the dot product is a function $(\mathbb{R}^n, \mathbb{R}^n) \to \mathbb{R}$, Then our multiple regression model is just



$$\hat{y} = f_\theta(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_p x_p = x^T\theta$$

$$\begin{bmatrix} 1 & 2 \end{bmatrix} \begin{bmatrix} 3 \\ 4 \end{bmatrix}$$

$= x \cdot \theta$

$= \theta \cdot x$

where $\theta = \begin{pmatrix} \theta_0 \\ \theta_1 \\ \vdots \\ \theta_p \end{pmatrix}$ and $x = \begin{pmatrix} 1 \\ x_1 \\ \vdots \\ x_p \end{pmatrix}$ if we do this with different x vectors, each corresponding to a different observation allows then our model $\hat{\mathbb{Y}} = \mathbb{X}\theta$ where

$$\mathbb{X} = \begin{pmatrix} — & \vec{x_1} & — \\ — & \vec{x_2} & — \\ & \vdots & \\ — & \vec{x_n} & — \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & x_{13} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & x_{23} & \cdots & x_{2p} \\ 1 & x_{31} & x_{32} & x_{33} & \cdots & x_{3p} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & x_{n3} & \cdots & x_{np} \end{pmatrix} \begin{pmatrix} \theta_0 \\ \theta_1 \\ \vdots \\ \theta_p \end{pmatrix}$$

$$\hat{y}_1 = x_1^T \theta$$
$$\hat{y}_2 = x_2^T \theta$$
$$= x_3^T \theta$$
$$\vdots$$
$$x_n^T \theta$$

$$\hat{y} = \begin{pmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \hat{y}_3 \\ \vdots \end{pmatrix} \qquad y \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

$$y - \hat{y} = \vec{e} = \begin{pmatrix} y_1 - \hat{y}_1 \\ y_2 - \hat{y}_2 \\ \vdots \end{pmatrix}$$

More terminology

$(\mathbb{R}^n) \to \mathbb{R}$

For vector $v \in \mathbb{R}^n$ we denote the p-norm of v as $||v||_p$, in this class we will work with $p = 1, 2$ corresponding to the L_1 and L_2 vector norms, for this class a norm is an operator which tells us the size of a vector

$$||v||_2 = \sqrt{v_1^2 + v_2^2 + \ldots + v_n^2} = \sqrt{\sum_{i=1}^{n} x_i^2}$$

$$||v||_1 = |v_1| + |v_2| + \ldots + |v_n| = \sum_{i=1}^{n} |x_i|$$

If we let $e_i = y_i - \hat{y}_i$ then we can reformulate MSE as $\frac{1}{n}\sum_{i=1}^{n}(e_i)^2$
If we stack these values we construct the residual vector $e = \mathbb{Y} - \hat{\mathbb{Y}}$

Vectorize MSE

Let us vectorize MSE loss under model $\theta = \begin{pmatrix} \theta_0 & \theta_1 & \dots & \theta_n \end{pmatrix}^T$

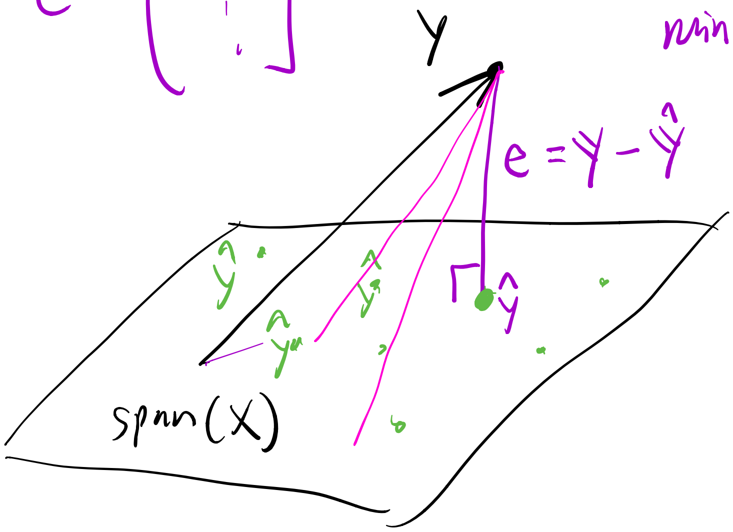
$$
\begin{aligned}
\min \quad L(\theta) &= \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \underbrace{(\mathbb{X}_i \cdot \theta)}_{\hat{y}_i})^2 \\
&= \frac{1}{n}\left(\sqrt{\underbrace{(y_1 - \hat{y}_1)^2}_{e^2} + \underbrace{\dots}_{e_2^2} + \underbrace{(y_n - \hat{y}_n)^2}_{e_n^2}}\right)^2 \\
&= \frac{1}{n}||\mathbb{Y} - \hat{\mathbb{Y}}||_2^2 \\
\min \qquad &= \frac{1}{n}||\mathbb{Y} - \mathbb{X}\theta||_2^2
\end{aligned}
$$

$$
\implies \hat{\theta} = \min_{\theta} L(\theta) = \min_{\theta} ||\mathbb{Y} - \mathbb{X}\theta||_2^2 = \min_{\theta} ||e||_2^2
$$

Analogously to the scalar-case we can analytically solve the vector-case using matrix calculus (out of scope) or geometrically (very in scope)

$e = \begin{bmatrix} \vdots \\ \vdots \end{bmatrix}$

$\min \|\vec{e}\|_2^2$



$\vec{y} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{bmatrix}$

lists of len n

$\text{span}(X) \in \mathbb{R}^n$

$X_{n \times p} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & & & \ddots & \vdots \\ 1 & x_{n1} & \cdots & \cdots & x_{np} \end{pmatrix} \begin{pmatrix} \vdots \end{pmatrix}$

$X\theta = [\ ]_{n \times \_}$

Geometric derivation

- Our prediction is a linear combination of the columns of $\mathbb{X}$, thus our prediction lives in span$(\mathbb{X}) \in \mathbb{R}^n$
- Our goal is to find some vector $\hat{\mathbb{Y}}$ in span$(\mathbb{X})$ closest to $\mathbb{Y}$
  - This is the same as finding $\hat{\mathbb{Y}}$ which minimizes e
  - This is achieved if you set $\hat{\mathbb{Y}}$ to orthogonal projection of $\mathbb{Y}$ onto span$(\mathbb{X})$
    - two vectors are orthogonal if and only if their dot product is 0
    - we want e to be orthogonal to span$(\mathbb{X})$ so we want $\mathbb{X}^T e = 0$

$$\begin{pmatrix} - & x_1 & - \\ - & x_2 & - \\ - & x_3 & - \\ & \vdots & \end{pmatrix} \vec{e} \quad (n \times 1)$$

$$Xe = \begin{pmatrix} x_1 \cdot e \\ x_2 \cdot e \\ \vdots \end{pmatrix}$$

$$\mathbb{X}^T e = \mathbb{X}^T(\mathbb{Y} - \mathbb{X}\hat{\theta}) = \mathbb{X}^T\mathbb{Y} - \mathbb{X}^T\mathbb{X}\hat{\theta} = 0$$

$$\mathbb{X}^T\mathbb{X}\hat{\theta} = \mathbb{X}^T\mathbb{Y}$$

$$\begin{pmatrix} | & | & & | \\ col_1 & col_2 & \cdots & col_p \\ | & | & & | \end{pmatrix}^T = \begin{pmatrix} - & col_1 & - \\ - & col_2 & - \\ & \vdots & \\ - & col_p & - \end{pmatrix} e = \begin{pmatrix} col_1 \cdot e \\ col_2 \cdot e \\ \vdots \end{pmatrix}$$

If $\mathbb{X}^T\mathbb{X}$ is full rank (implies invertibility) then $\hat{\theta} = (X^T\mathbb{X})^{-1}\mathbb{X}^T\mathbb{Y}$

Invertibility of $\mathbb{X}^T\mathbb{X}$

$X^T e = 0$

$\begin{pmatrix} | & \\ col_1 & \cdots \\ | & \end{pmatrix}$ $[1 \; 1 \; 1 \; 1 \; \ldots \; 1]$

$[e_1 \; e_2 \; e_3 \; \ldots \; e_n]$

$col_1 \perp e$ $\quad \mathbb{1} \cdot e = 0$ $\quad \sum e_i = 0$

- In the analytical solution $\mathbb{X}^T e = 0$ and so if our model has a linear intercept term $(x_0 = 1)$ then $1^T e = 0$, meaning that in the optimal model the residuals sum to 0 (mean of residuals is also 0, think about why)
- At least one solution **always** exists, a unique solution exists only if $\mathbb{X}^T\mathbb{X}$ is invertible $\equiv$ full rank
  - if it is not invertible there will an infinite number of solutions
  - $\mathbb{X}^T\mathbb{X}$ is invertible if and only if all columns of $\mathbb{X}$ are linearly independent which is the same as saying that $\mathbb{X}$ is full column rank (same as $\mathbb{X}^T\mathbb{X}$ is full rank–row and column)
- $\mathbb{X}$ will not have full column rank
  - if some features are linear combinations of other features
  - if number columns is greater than number of rows

$X^T X$

1
2
3