DATA 100 HW 07

Bryan Ngo

2021-10-09

Properties of Simple Linear Regression

$$\hat{\theta}_0 = \bar{y} - \hat{\theta}_1 \bar{x} \tag{1}$$

$$\hat{\theta}_1 = r\frac{\sigma_y}{\sigma_x} \tag{2}$$

$$\hat{y} = \bar{y} + r\frac{\sigma_y}{\sigma_x}(x - \bar{x}) \tag{3}$$

1

1.a

Theorem 1. *The sum of residuals* $\sum_{i=1}^{n} e_i = 0$.

Proof.

$$\sum_{i=1}^{n} e_i = \sum_{i=1}^{n}(y_i - \hat{y}_i) = \sum_{i=1}^{n}\left(y_i - \left(\bar{y} + r\frac{\sigma_y}{\sigma_x}(x_i - \bar{x})\right)\right) \tag{4}$$

$$= \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} y_i + r\frac{\sigma_y}{\sigma_x}\left(\sum_{i=1}^{n} x_i - \sum_{i=1}^{n} x_i\right) = 0 \tag{5}$$

where we use the fact that the $\frac{1}{n}$ term in the definition of the mean cancels out the summation factor of n. □

1.b

Theorem 2. *The true mean of a sample is equal to the predicted mean of a linear regression of that sample.*

Proof. Using the previous result,

$$\sum_{i=1}^{n}(y_i - \hat{y}_i) = 0 \tag{6}$$

$$\implies \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i) = 0 \tag{7}$$

$$\implies \underbrace{\frac{1}{n}\sum_{i=1}^{n} y_i}_{\bar{y}} - \underbrace{\frac{1}{n}\sum_{i=1}^{n}\hat{y}_i}_{\bar{\hat{y}}} = 0 \tag{8}$$

$$\implies \bar{y} = \bar{\hat{y}} \tag{9}$$

□

1.c

Theorem 3. *The point $(\bar{x}, \bar{y})$ lies on the simple linear regression line.*

Proof. Plugging $\bar{x}$ into the simple linear regression,

$$\hat{y} = \bar{y} + r\frac{\sigma_y}{\sigma_x}(\bar{x} - \bar{x}) = \bar{y} \tag{10}$$

□

Geometric Perspective of Least Squares

$$\hat{\mathbb{Y}} = \mathbb{X}\begin{bmatrix}\hat{\theta}_0 \\ \hat{\theta}_1\end{bmatrix} = \hat{\theta}_0\mathbb{1} + \hat{\theta}_1\boldsymbol{x} \tag{11}$$

2

2.a

Theorem 4. *The sum of the elements of the residual vector $\sum_{i=1}^{n} e_i$, where $\boldsymbol{e} = \begin{bmatrix}e_1 & e_2 & \cdots & e_n\end{bmatrix}^\top$.*

Proof. Note that $\sum_{i=1}^{n} e^i = \mathbb{1}^\top \boldsymbol{e}$. By definition, $\boldsymbol{e}$ is orthogonal to $\mathbb{X}$, as that arises out of minimizing the norm of error in predictions as noted in the geometric perspective. This means that $\boldsymbol{e}$ is *also* orthogonal to any linear combination of the columns of $\mathbb{X}$, being $\text{span}\{\mathbb{X}\} = a\mathbb{1} + b\boldsymbol{x}$, where $a, b \in \mathbb{R}$. If we set $a = 1$ and $b = 0$, we get $\mathbb{1}^\top \boldsymbol{e} = 0$. □

2.b

We can use the same explanation as above, but now we can set $a = 0$ and $b = 1$, meaning that $\boldsymbol{x}^\top \boldsymbol{e} = 0$, so they are orthogonal.

2.c

The predicted response vector $\hat{\mathbb{Y}}$ is nothing more than a linear combination of $\mathbb{1}$ and $\boldsymbol{x}$, with $a = \hat{\theta}_0$ and $b = \hat{\theta}_1$. This means that $\hat{\mathbb{Y}} \in \text{span}\{\mathbb{X}\}$. By the above logic in **2.a**, we have $\hat{\mathbb{Y}}^\top \boldsymbol{e} = 0$, and they are orthogonal.

Properties of a Linear Model with No Constant Term

$$\hat{y} = \gamma x \tag{12}$$

$$R(\gamma) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \gamma x_i)^2 \tag{13}$$

3

Theorem 5. *The minimum of $R(\gamma)$ with respect to γ is*

$$\hat{\gamma} = \frac{\sum x_i y_i}{\sum x_i^2} \tag{14}$$

Proof.

$$\frac{d}{d\gamma}R(\gamma) = \frac{2}{n}\sum_{i=1}^{n} -x_i(y_i - \gamma x_i) \tag{15}$$

$$= -\frac{2}{n}\sum_{i=1}^{n} x_i y_i + \frac{2}{n}\gamma\sum_{i=1}^{n} x_i^2 = 0 \tag{16}$$

$$\implies \gamma\sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i \tag{17}$$

$$\hat{\gamma} = \frac{\sum x_i y_i}{\sum x_i^2} \tag{18}$$

□

4

$$\hat{\mathbb{Y}} = \hat{\gamma}\boldsymbol{x} \tag{19}$$

4.a

This is false, consider the data points $\{(1,2),(2,5),(3,6)\}$. We can find $\hat{\gamma} = \frac{15}{7}$ and $\hat{\mathbb{Y}} = \frac{1}{7}\begin{bmatrix}15 & 30 & 45\end{bmatrix}^\top$. Our residual is $\boldsymbol{e} = \frac{1}{7}\begin{bmatrix}-1 & 5 & -3\end{bmatrix}$. Thus, the sum of the residuals is $\frac{1}{7} \neq 0$.

4.b

Since by definition in the geometric perspective established in lecture, $\boldsymbol{e}$ is orthogonal to $\mathbb{X}$ and any vector in span$\{\mathbb{X}\}$ Thus, $\boldsymbol{e}$ is orthogonal to any linear combination $a\boldsymbol{x}$. Setting the coefficient $a = 1$, we have $\boldsymbol{x}^\top\boldsymbol{e} = 0$, so they are orthogonal.

4.c

Since $\boldsymbol{x}^\top\boldsymbol{e} = 0$, and $\hat{\mathbb{Y}} = \hat{\gamma}\boldsymbol{x}$,

$$\hat{\mathbb{Y}}^\top\boldsymbol{e} = (\hat{\gamma}\boldsymbol{x})^\top\boldsymbol{e} = \hat{\gamma}\boldsymbol{x}^\top\boldsymbol{e} = 0 \tag{20}$$

4.d

This is false, consider the data points $\{(1,2),(2,5),(3,6)\}$. We have $\bar{x} = 2$ and $\bar{y} = \frac{13}{3}$. However, $\hat{\mathbb{Y}} = \frac{30}{7} \neq \frac{13}{3}$.

MSE "Minimizer"

5

5.a

The MSE loss function can be viewed as a sum of **quadratic** terms, each of which can be treated as a function of $\boldsymbol{\gamma}$.

5.b

$$g_i'(\gamma) = -x_i \frac{2}{n}(y_i - \gamma x_i) \tag{21}$$

$$g_i''(\gamma) = \frac{2}{n} x_i^2 \geqslant 0 \tag{22}$$

Since the x_i term is squared, the function will always be nonnegative.

5.c

Since the function is convex, the graph of our function will always "curve" upwards. This implies that the point at which $\frac{d}{dx} g(x) = 0$ will be a minimum.

5.d

5.d.i

Theorem 6. *Given a function $g(x)$ that satisfies the condition*

$$g(cx_1 + (1-c)x_2) \leqslant cg(x_1) + (1-c)g(x_2) \tag{23}$$

for points $(x_1, g(x_1))$, $(x_2, g(x_2))$, and $c \in [0,1]$, the sum of two convex functions $g(x) + h(x)$ is also convex.

Proof. Let $f(x) = g(x) + h(x)$. Adding the inequalities,

$$\begin{aligned} f(cx_1 + (1-c)x_2) &= g(cx_1 + (1-c)x_2) + h(cx_1 + (1-c)x_2) && (24)\\ &\leqslant cg(x_1) + (1-c)g(x_2) + ch(x_1) + (1-c)h(x_2) && (25)\\ &= c(g(x_1) + h(x_1)) + (1-c)(g(x_2) + h(x_2)) && (26)\\ &= cf(x_1) + (1-c)f(x_2) && (27) \end{aligned}$$

□

5.d.ii

Since the sum of two convex functions is also convex, one or both of the convex functions can also be a sum of convex functions, and recursively *ad infinitum*. This can be proved with induction.

5.e

We have also proved that all convex functions' critical points are minimums. We have previously proved that the MSE loss function is convex through the second derivative test. By deduction, this means that the MSE loss function's critical point(s) are minimums.