14.1) Modeling and Estimation

Vitor Kamada

February 2020

Reference

Tables, Graphics, and Figures from

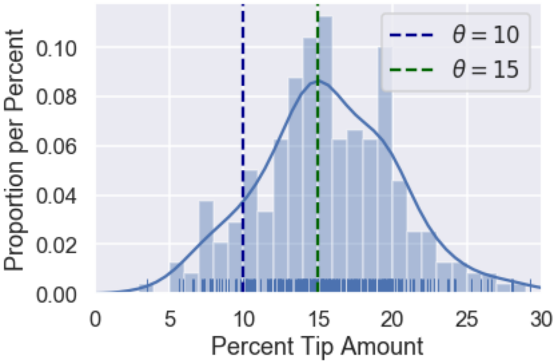
Principles and Techniques of Data Science

Lau et al. (2019): Ch 10 Modeling and Estimation

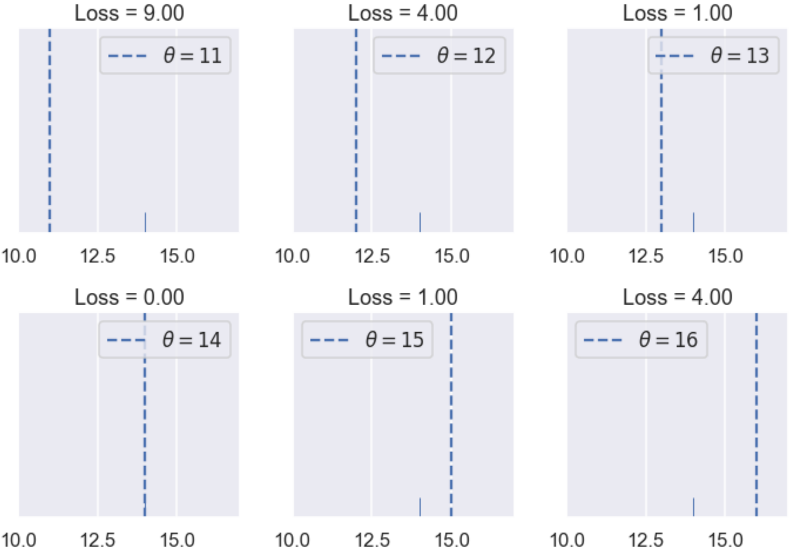
https://www.textbook.ds100.org/ch/12/prob_exp_var.html

```python
tips = sns.load_dataset('tips')
```

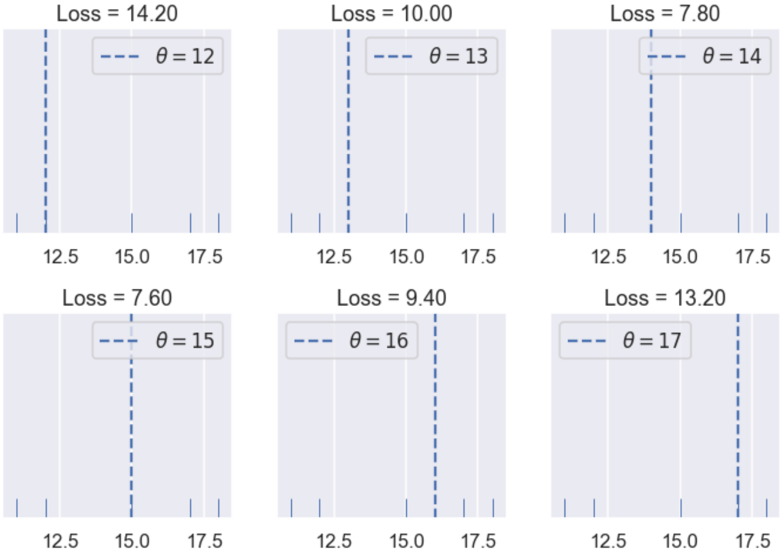
	total_bill	tip	sex	smoker	day	time	size
0	16.99	1.01	Female	No	Sun	Dinner	2
1	10.34	1.66	Male	No	Sun	Dinner	3
2	21.01	3.50	Male	No	Sun	Dinner	3



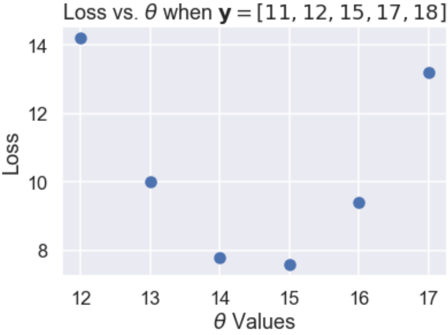
$$L(\theta, \mathbf{y}) = \frac{1}{n} \sum_{i=1}^{n} (y_i - \theta)^2$$



$\mathbf{y} = [11, 12, 15, 17, 18]$



Minimizing the Mean Squared Error (MSE)



$$\hat{\theta} = 14.6$$

$$\begin{aligned}
L(\theta, \mathbf{y}) &= \frac{1}{5}\big((11-\theta)^2 + (12-\theta)^2 + (15-\theta)^2 + (17-\theta)^2 + (18-\theta)^2\big) \\
\frac{\partial}{\partial\theta} L(\theta, \mathbf{y}) &= \frac{1}{5}\big(-2(11-\theta) - 2(12-\theta) - 2(15-\theta) - 2(17-\theta) - 2(18-\theta)\big) \\
&= \frac{1}{5}\big(10 \cdot \theta - 146\big)
\end{aligned}$$

$L(\theta, \mathbf{y}) = \frac{1}{n} \sum_{i=1}^{n} (y_i - \theta)^2$

$$-\frac{2}{n}\sum_{i=1}^{n}(y_i - \theta) = 0$$

$$\sum_{i=1}^{n}(y_i - \theta) = 0$$

$$\sum_{i=1}^{n}\theta = \sum_{i=1}^{n} y_i$$

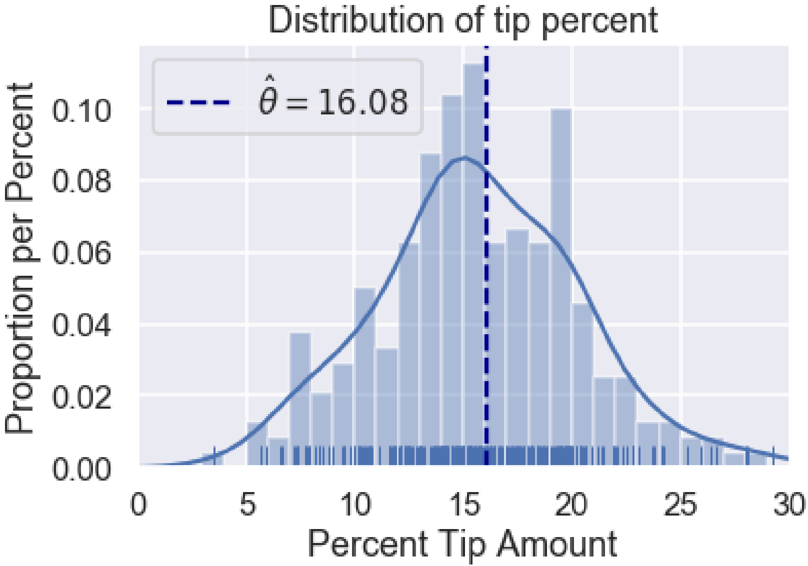
$$n \cdot \theta = y_1 + \ldots + y_n$$

$$\theta = \frac{y_1 + \ldots + y_n}{n}$$

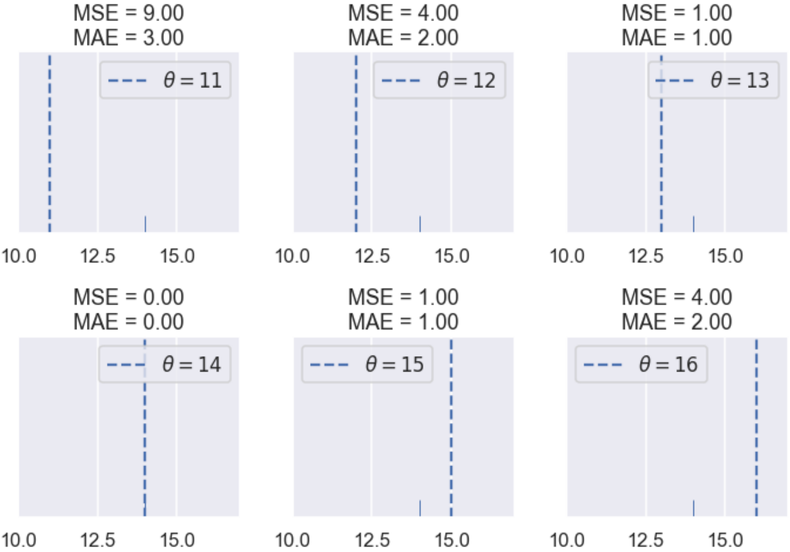
$$\hat{\theta} = \theta = \text{mean}(\mathbf{y})$$

```python
np.mean(tips['pcttip'])
```

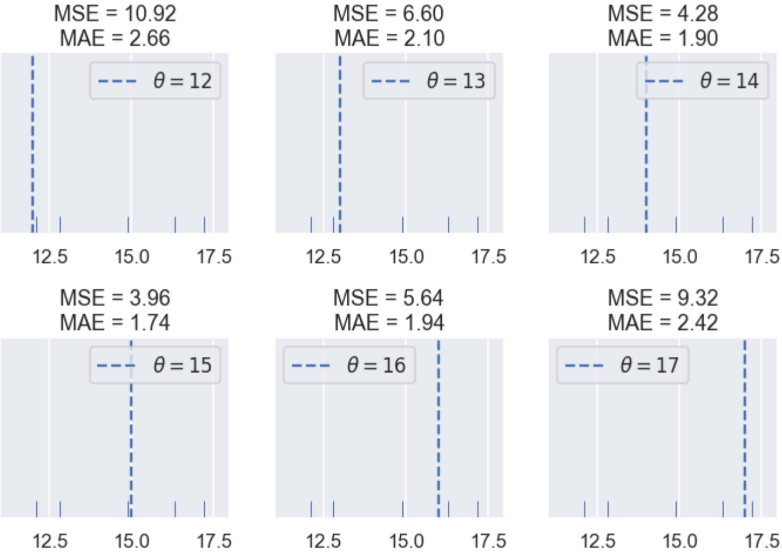
16.08



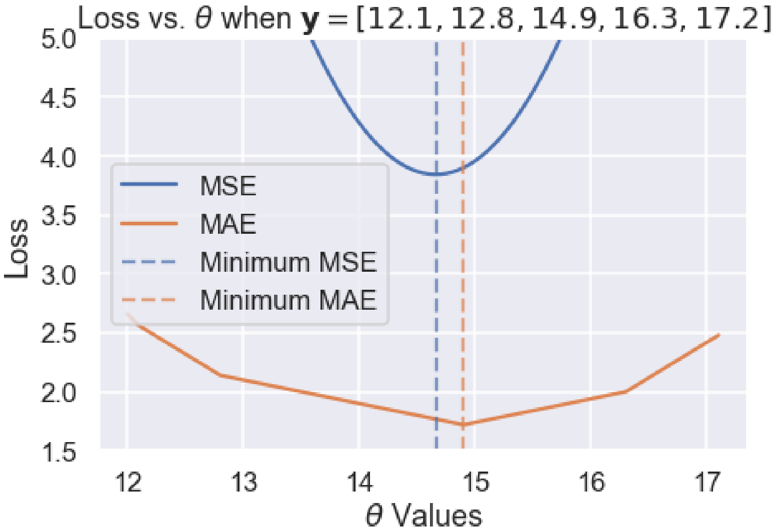
$L(\theta, \mathbf{y}) = \frac{1}{n} \sum_{i=1}^{n} |y_i - \theta|$



$\mathbf{y} = [12.1, 12.8, 14.9, 16.3, 17]$



MSE vs MAE



Loss vs. θ when $\mathbf{y} = [12.1, 12.8, 14.9, 16.3, 17.2]$
5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
Loss
MSE
MAE
Minimum MSE
Minimum MAE
12
13
14
15
16
17
θ Values

$L(\theta, \mathbf{y}) = \frac{1}{n}\sum_{i=1}^{n} |y_i - \theta|$

$$= \frac{1}{n}\left(\sum_{y_i<\theta}|y_i-\theta| + \sum_{y_i=\theta}|y_i-\theta| + \sum_{y_i>\theta}|y_i-\theta|\right)$$

$$\frac{1}{n}\left(\sum_{y_i<\theta}(-1) + \sum_{y_i=\theta}(0) + \sum_{y_i>\theta}(1)\right) = 0$$

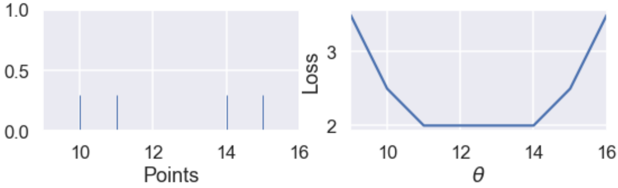
$$\sum_{y_i<\theta}(-1) + \sum_{y_i>\theta}(1) = 0$$

$$-\sum_{y_i<\theta}(1) + \sum_{y_i>\theta}(1) = 0$$

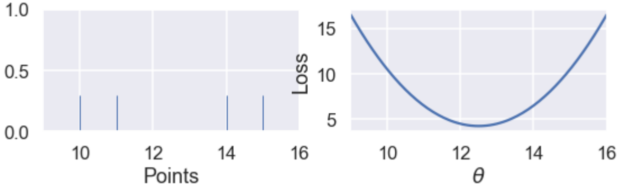
$$\sum_{y_i<\theta}(1) = \sum_{y_i>\theta}(1)$$

$\mathbf{y} = [10, 11, 14, 15]$

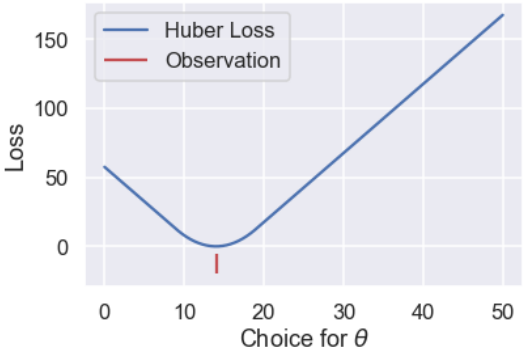
$$\hat{\theta} = median(y)$$



$$\hat{\theta} = mean(y)$$



Huber Loss for $\mathbf{y} = [14]$



$$L_\alpha(\theta, \mathbf{y}) = \frac{1}{n} \sum_{i=1}^{n} \begin{cases} \frac{1}{2}(y_i - \theta)^2 & |y_i - \theta| \leq \alpha \\ \alpha(|y_i - \theta| - \frac{1}{2}\alpha) & \text{otherwise} \end{cases}$$