

Introduction to regression analysis

Lecture 6a – 2021-06-16

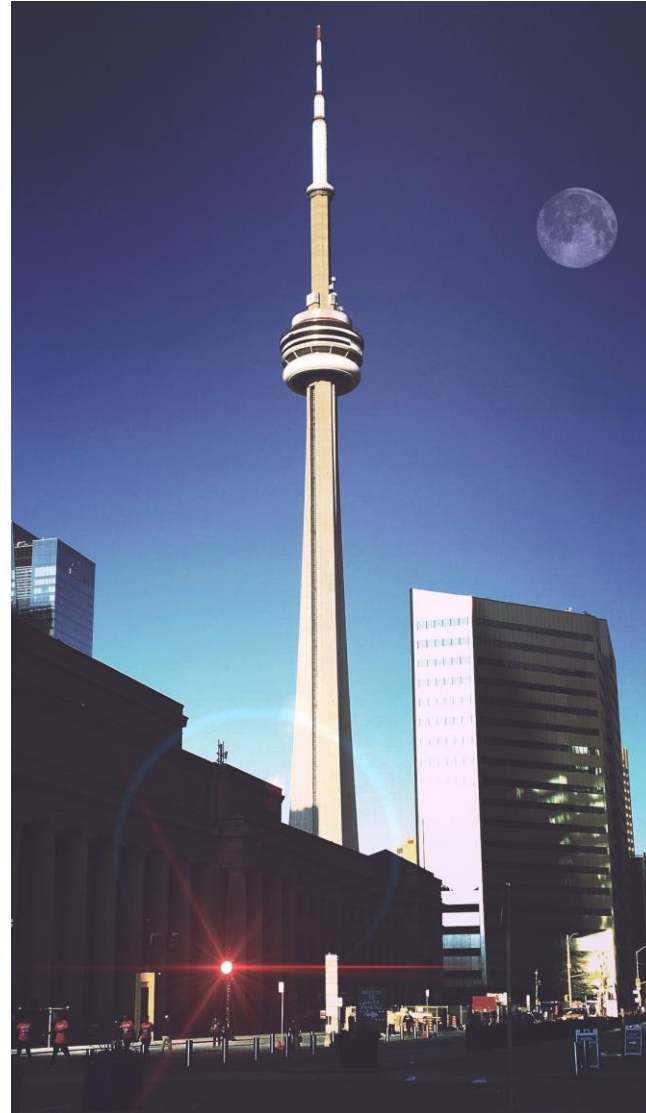
MAT A35 – Summer 2021 – UTSC

Prof. Yun William Yu

Height of the CN tower

- On June 15, you measure that the CN tower is 21,785 inches tall.
- How tall will the CN tower be on July 15?

A: 10892.5 inches
B: 21785 inches
C: 43570 inches
D: ???
E: None of the above



Growth of a willow tree

- On June 15, you measure that a weeping willow measures 424 inches tall.
- How tall will the tree be on July 15?

A: 212 inches
B: 424 inches
C: 848 inches
D: ???
E: None of the above



Two data points

- On May 15, you measured that a weeping willow measures 420 inches tall.
- On June 15, you measured that the same weeping willow is 424 inches tall.
- How tall is the weeping willow on July 15?

- A: 420 inches
- B: 424 inches
- C: 428 inches
- D: ???
- E: None of the above



Two data points

- On May 15, you measure that the CN tower is 21,786 inches tall.
- On June 15, you measure that the CN tower is 2,1785 inches tall.
- How tall will the CN tower be on July 15?

- A: 21,784 inches
- B: 21,785 inches
- C: 21,786 inches
- D: ???
- E: None of the above



Model assumptions

- Model assumption: the CN tower should stay a roughly constant height, subject to experimental errors.

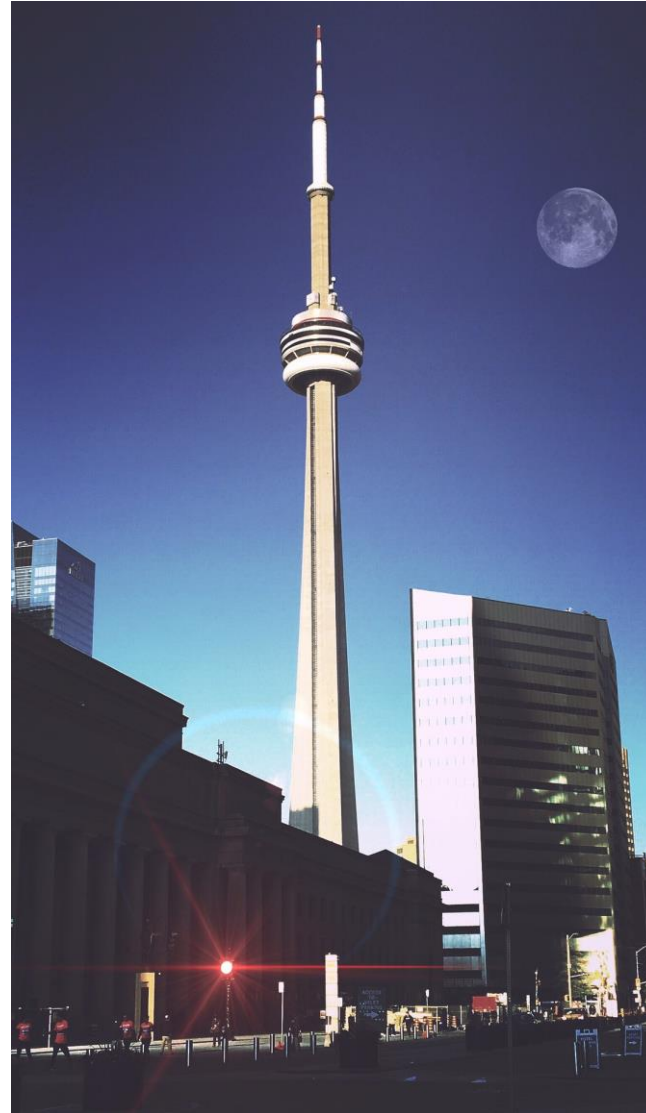


- Model assumption: a willow tree grows roughly linearly, subject to experimental errors.



One-parameter model

- Model assumption: the CN tower should stay a constant height, subject to experimental errors.
- $h(t) = b$, where b is a constant.



Two-parameter model

- Model assumption: a willow tree grows roughly linearly, subject to experimental errors.
- $h(t) = mt + b$, where m and b are constants, and t is time in months



Three data points

- On April 15, you measured a height of 417 inches tall.
- On May 15, you measured a a height of 420 inches tall.
- On June 15, you measured that the same weeping willow is 424 inches tall.
- How tall is the weeping willow on July 15?

- A: 424 inches
- B: 427 inches
- C: 428 inches
- D: ???
- E: None of the above



The “best”-fit model



- A model is good if it predicts future data accurately.
- Since the model cannot see into the future, the model is built to accurately explain (“fit” to) existing data.

| Date | Height of CN tower |
|----------|--------------------|
| January | 21,779 |
| February | 21,787 |
| March | 21,788 |
| April | 21,786 |
| May | 21,786 |
| June | 21,785 |
| July | ??? |

Errors in both directions matter

- We want to minimize average errors, but pos/neg errors are both bad.
- Can use either absolute value or squaring before summing errors.

| Error from mean estimator $h(t) =$ 21785 |
|--|
| -6 |
| 2 |
| 3 |
| 1 |
| 1 |
| 0 |

| Error from median estimator $h(t) =$ 21786 |
|--|
| -7 |
| 1 |
| 2 |
| 0 |
| 0 |
| 1 |

“Best” estimators depend on error metric

- Mean absolute error

- Given data points h_1, h_2, \dots, h_n and a guessed height b ,

$$MAE(b) = \frac{1}{n} \sum_{i=1}^n |h_i - b|$$

- Optimal guess is the median (the middle element if odd, or the sum of the two middle elements divided by two if even)

- Mean squared error

- Given data points h_1, h_2, \dots, h_n and a guessed height b ,

$$MSE(b) = \frac{1}{n} \sum_{i=1}^n (h_i - b)^2$$

- Optimal guess is the mean $= \frac{1}{n} \sum_{i=1}^n h_i$

Two-parameter model fitting

- $h(t) = mt + b$, where m and b are constants, and t is time in months
- What are the optimal values of m and b ?



| | Month | Height of willow tree |
|---|----------|-----------------------|
| 1 | January | 404 |
| 2 | February | 407 |
| 3 | March | 412 |
| 4 | April | 417 |
| 5 | May | 420 |
| 6 | June | 424 |
| 7 | July | ??? |

Error of linear model: $f(t) = mt + b$

- Mean absolute error

- Given data points h_1, h_2, \dots, h_n at times t_1, \dots, t_n and parameters (m, b)

$$MAE(m, b) = \frac{1}{n} \sum_{i=1}^n |h_i - f(t_i)| = \frac{1}{n} \sum_{i=1}^n |h_i - (mt_i + b)|$$

- Computing mean absolute error is hard because absolute value is not differentiable. (See Linear Programming)

- Mean squared error

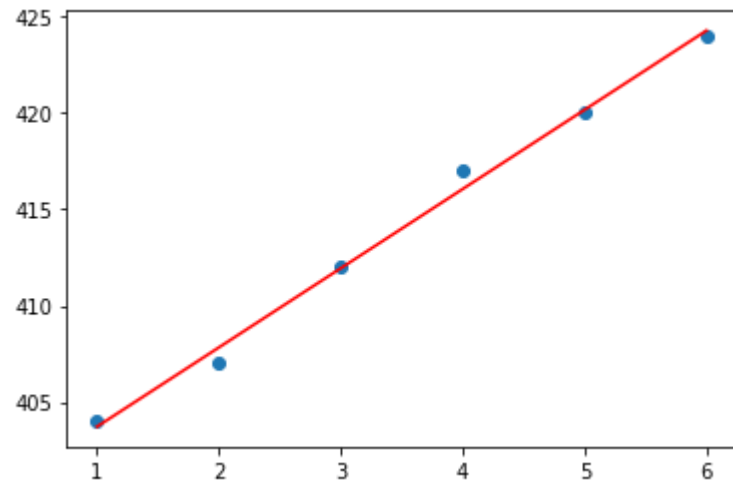
- Given data points h_1, h_2, \dots, h_n at times t_1, \dots, t_n and parameters (m, b)

$$MSE(m, b) = \frac{1}{n} \sum_{i=1}^n (h_i - f(t_i))^2 = \frac{1}{n} \sum_{i=1}^n (h_i - (mt_i + b))^2$$

- We can find the minimum of this function using tools from calculus.

Best-fit line for willow tree

- $f(t) = 4.11t + 399.6$



Derivation for simple example

$$S(m, b) = \frac{1}{n} \sum_{i=1}^n (h_i - (mt_i + b))^2$$



| Month | Height of willow tree |
|--------------|-----------------------|
| 1 – January | 404 |
| 2 – February | 407 |
| 3 – March | 412 |
| 4 – April | 417 |
| 5 – May | 420 |
| 6 – June | 424 |

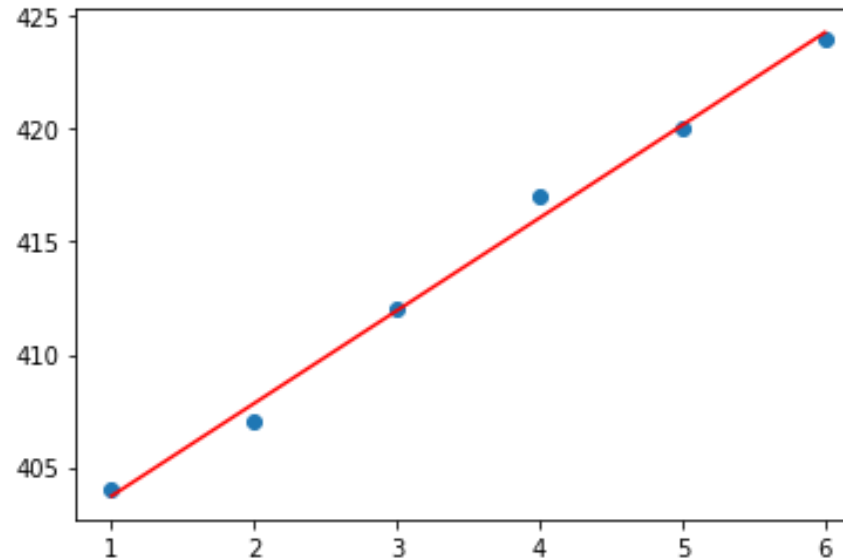
Find critical points & check the Hessian

- Set $\frac{\partial S}{\partial m} = 0$ and $\frac{\partial S}{\partial b} = 0$.
 - End up with $m \approx 4.11$ and $b = 399.6$
- Then need to check that the eigenvalues of the Hessian are both positive:

- $$\begin{bmatrix} \frac{\partial^2 S}{\partial m^2} & \frac{\partial^2 S}{\partial b \partial m} \\ \frac{\partial^2 S}{\partial m \partial b} & \frac{\partial^2 S}{\partial b^2} \end{bmatrix}$$
 has positive

eigenvalues at (4.11, 399.6)

- Therefore, the model $f(x) = 4.11x + 399.6$ is the best-fit line



Linear model error

- Given measurements y_1, y_2, \dots, y_n at values x_1, \dots, x_n , a linear model is a function $f(x) = mx + b$, with parameters m and b where $y_i \approx f(x_i)$ with some error.
 - Ex: the x-axis coordinates might be time, and the y-axis might be height of a tree as a function of time.
- The Mean Squared Error of the model is given by

$$MSE(m, b) = \frac{1}{n} \sum_{i=1}^n (y_i - f(x_i))^2 = \frac{1}{n} \sum_{i=1}^n (y_i - (mx_i + b))^2$$

- We want to find the model parameters that give the minimum mean squared error, so consider the function $S(m, b) = MSE(m, b)$. We want to find the minimum of the function $S(m, b)$.

Theorem (linear models)

- Suppose we are given measurements y_1, y_2, \dots, y_n at values x_1, \dots, x_n . Let $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$ and $\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$ be the respective averages.
- Then the linear model $f(x) = mx + b$ that minimizes the mean squared error is given by:

$$m = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

$$b = \bar{y} - m\bar{x}$$

- Proof involves using the partial derivatives to find the minimum of the function $S(m, s) = \frac{1}{n} \sum_{i=1}^n (y_i - (mx_i + b))^2$.