Introduction to regression analysis Lecture 6a – 2021-06-16

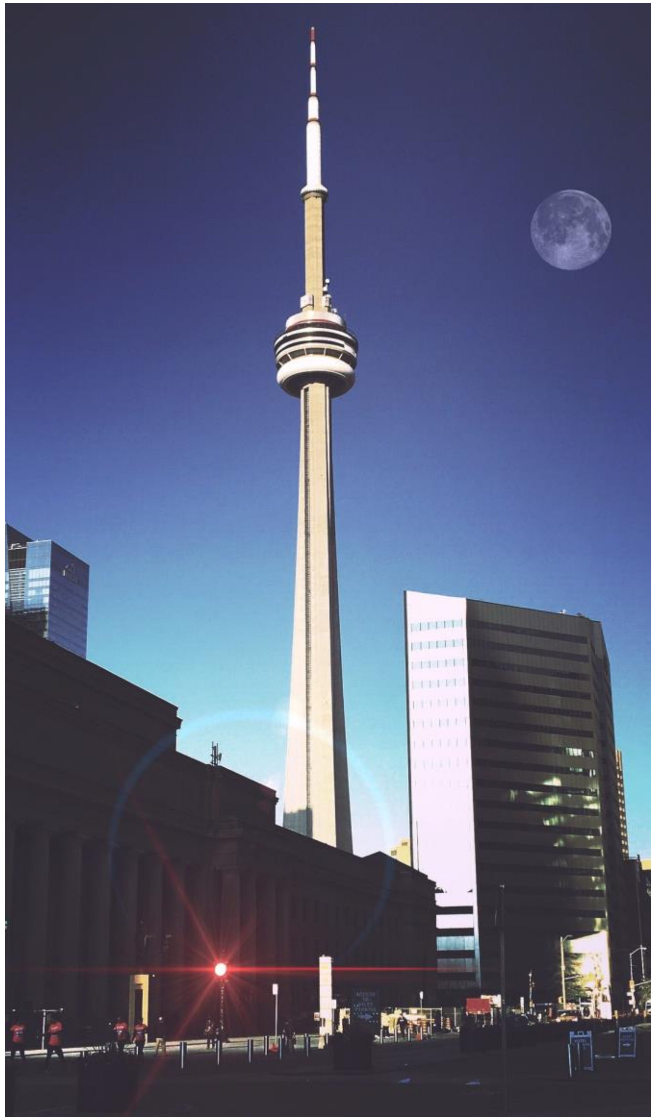
MAT A35 – Summer 2021 – UTSC

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

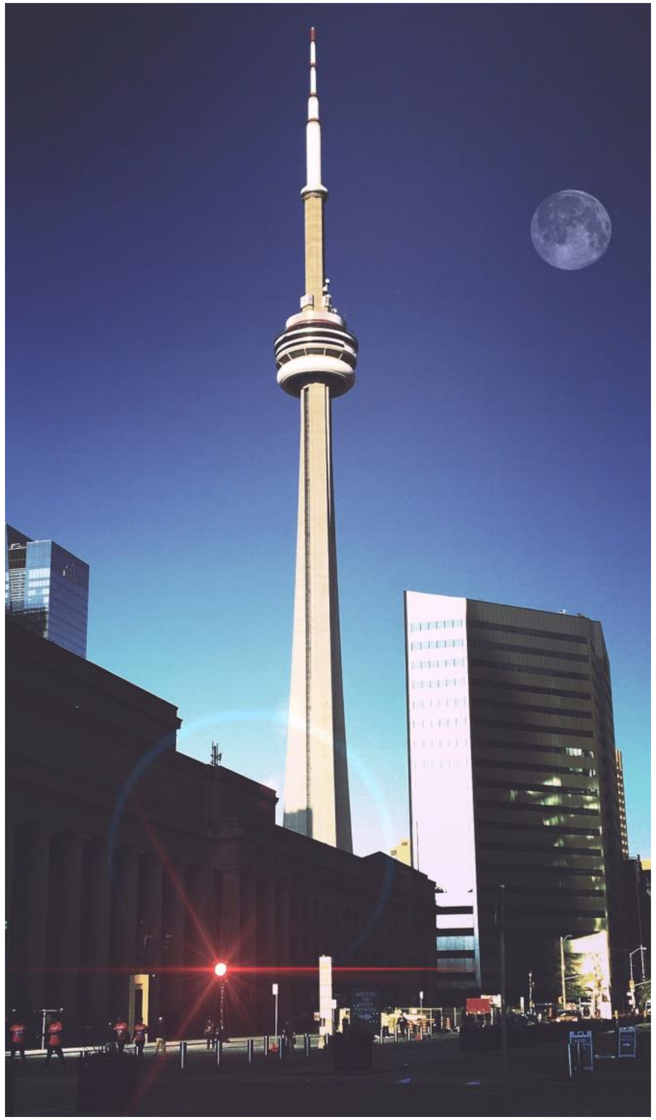
- On May 15, you measured that a weeping willow measures 420 inches tall.
- On June 15, you measured that the same weeping willow is 424 inches tall.
- How tall is the weeping willow on July 15?

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.
  21,785
- 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.

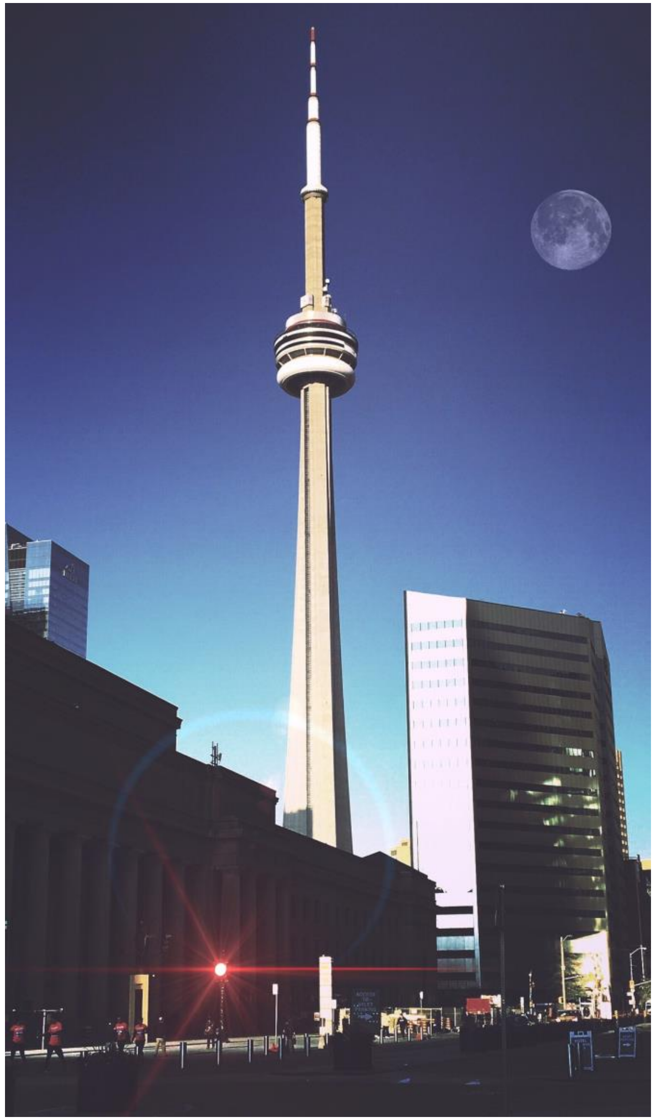
One-parameter model

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

$h(\text{May}) = 21,786 \Rightarrow b = 21,786$

$h(\text{June}) = 21,785 \Rightarrow b = 21,785$

$21,786 \neq 21,785$



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

May = 5 June = 6

$h(5) = 420 = 5m + b$

$h(6) = 424 = 6m + b$

$\Rightarrow m = 4 \quad b = 400$

Prediction: $h(\text{July}) = h(7) = 428$ inches

Three data points

- On April 15, you measured a height of 417 inches tall.
- On May 15, you measured 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

$h(t) = b$

- 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	???

Error $h(t)=20{,}000$	Mean Error: $h(t)=21785$	Median Error: $h(t)=21{,}786$
1,779	-6	-7
1,787	2	1
1,788	3	2
1,786	1	0
1,786	1	0
1,785	0	-1

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	Abs error	Squared error	Error from median estimator $h(t) =$ 21786	Abs error	Squared error
-6	6	36	-7	7	49
2	2	4	1	1	1
3	3	9	2	2	4
1	1	1	0	0	0
1	1	1	0	0	0
0	0	0	1	1	1
Average:	$\frac{13}{6} = 2.1\overline{6}$	$\frac{51}{6} = 8.5$		$\frac{11}{6} = 1.8\overline{3}$	$\frac{55}{6} = 9.1\overline{6}$

"Best" estimators depend on error metric

- Mean absolute error
  - Given data points $h_1, h_2, \ldots, h_n$ 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 $h_1, h_2, \ldots, h_n$ and a guessed height b,
$$MSE(b) = \frac{1}{n}\sum_{i=1}^{n}(h_i - b)^2$$
  - Optimal guess is the mean = $\frac{1}{n}\sum_{i=1}^{n} h_i$

Root Mean Squared Error

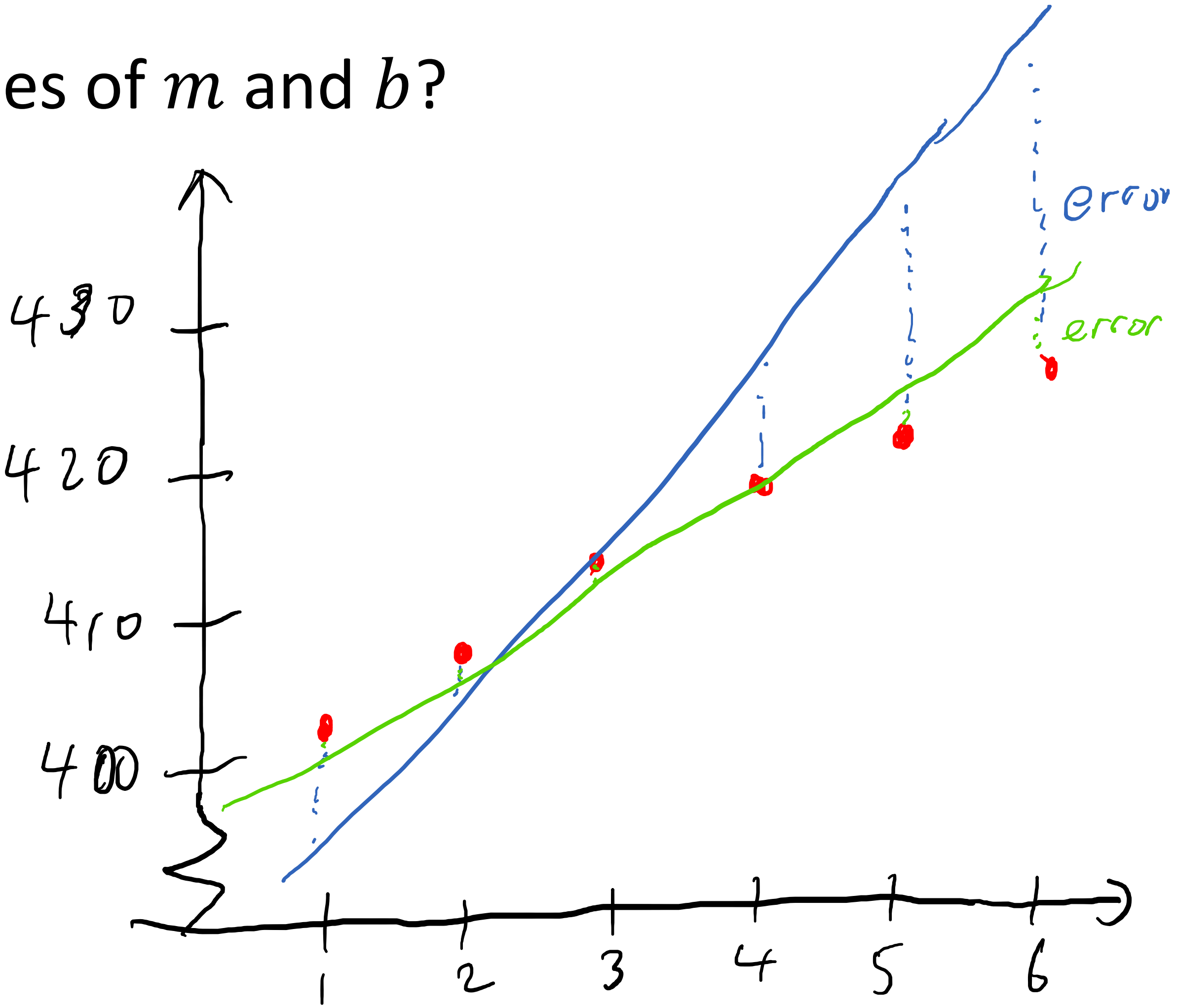
RMSE

$\sqrt{MSE}$

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)

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

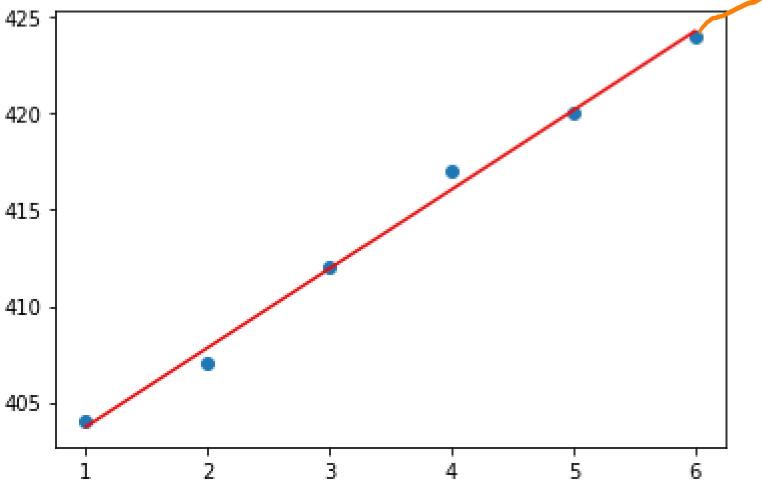
  - Computing mean absolute error is hard because absolute value is not differentiable. (See Linear Programming)
- Mean squared error
  - Given data points $h_1, h_2, \dots, h_n$ at times $t_1, \dots, t_n$ and parameters (m, b)

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

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

Best-fit line for willow tree

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



7

$f(7) = 428.37$

July: 428.37 inches

predicted height

Derivation for simple example

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

$$= \frac{1}{6}\Big[(404 - m - b)^2 + (407 - 2m - b)^2 + (412 - 3m - b)^2 + (417 - 4m - b)^2 + (420 - 5m - b)^2 + (424 - 6m - b)^2\Big]$$

$$\frac{\partial S}{\partial m} = \frac{1}{6}\Big[-2(404 - m - b) - 4(407 - 2m - b) - 6(412 - 3m - b) - 8(417 - 4m - b) - 10(420 - 5m - b) - 12(424 - 6m - b)\Big] = 0$$

$$\frac{\partial S}{\partial b} = \frac{1}{6}\Big[-2(404 - m - b) - 2(407 - 2m - b) - 2(412 - 3m - b) - 2(417 - 4m - b) - 2(420 - 5m - b) - 2(424 - 6m - b)\Big] = 0$$

Month t_1	Height of willow tree h_1
1 – January	404
2 – February t_2	407 h_2
3 – March t_3	412 h_3
4 – April $\vdots$	417 $\vdots$
5 – May	420
6 – June	424

$$\frac{\partial}{\partial m}\left[(404 - m - b)^2\right]$$

$$= 2(404 - m - b)(-1)$$

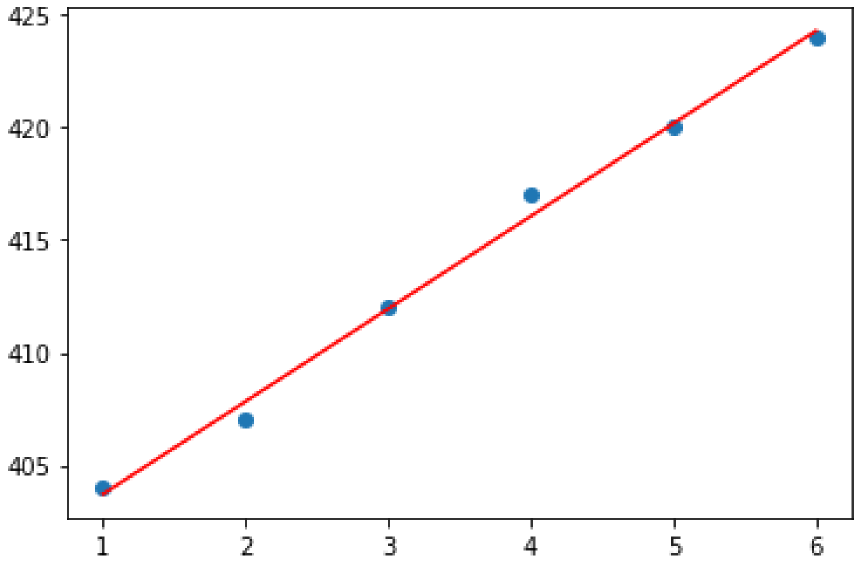
because $\frac{\partial}{\partial m}(-m) = -1$ (chain rule)

$$\frac{\partial}{\partial m}(m+1)^2 = 2(m+1)$$

$$\frac{\partial}{\partial m}(1-m)^2 = 2(1-m)\cdot(-1)$$

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:
  - $\begin{bmatrix} \frac{\partial^2 s}{\partial m^2} & \frac{\partial^2 s}{\partial b \partial m} \\ \frac{\partial^2 s}{\partial m \partial s} & \frac{\partial^2 s}{\partial b^2} \end{bmatrix}$ 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

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

- 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 $y_1, y_2, \ldots, y_n$ at values $x_1, \ldots, x_n$. 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 $S(m, s) = \frac{1}{n}\sum_{i=1}^{n}\big(y_i - (mx_i + b)\big)^2$.