Econ 5023: Statistics for Decision Making

Univariate Statistics (V): Continuous Variables and Mean (Part II)

Le Wang

Itinerary:

1. Why Mean? Why Median? Why anything?

Our Question continued:

How should we obtain the best forecast for a continuous variabe?

Candidates: Mean, Median, or any quantiles

Question: What is the **best** forecast?

If any measure has to work, it has to minimize the errors.

In the presence of heterogeneity (many possible values), how to define errors?

“Total” (or Overall) Errors

What property would this total error have?

Thought Process of Evaluating Errors

Thought Process of Evaluating Errors

1. We should sum them up

Thought Process of Evaluating Errors

1. We should sum them up $(-100, 0, 100)$

Thought Process of Evaluating Errors

1. We should sum them up
2. Reflect the magnitude of errors!

Thought Process of Evaluating Errors

1. We should sum them up
2. Reflect the magnitude of errors!

Next Question: What kind of function cares only about the magnitudes?

Some choices

Some choices

1. Absolute Function: $|x_i - \text{forecast}|$

Some choices

1. Absolute Function: $|x_i - \text{forecast}|$
2. Squared Function: $(x_i - \text{forecast})^2$

Let's look at a couple of examples first

Example 1:

```r
x <- c(1,2,3,4,10)
x

## [1]  1  2  3  4 10

mean(x)

## [1] 4

median(x)

## [1] 3
```

	x	error_1	error_2	error_3	error_4	error_10
1	1.00	0.00	-1.00	-2.00	-3.00	-9.00
2	2.00	1.00	0.00	-1.00	-2.00	-8.00
3	3.00	2.00	1.00	0.00	-1.00	-7.00
4	4.00	3.00	2.00	1.00	0.00	-6.00
5	10.00	9.00	8.00	7.00	6.00	0.00
6		15.00	10.00	5.00	0.00	-30.00

Table: Error Table

	x	error_1	error_2	error_3	error_4	error_10
1	1.00	0.00	-1.00	-2.00	-3.00	-9.00
2	2.00	1.00	0.00	-1.00	-2.00	-8.00
3	3.00	2.00	1.00	0.00	-1.00	-7.00
4	4.00	3.00	2.00	1.00	0.00	-6.00
5	10.00	9.00	8.00	7.00	6.00	0.00
6		15.00	10.00	5.00	0.00	-30.00

Table: Error Table

	x	error_1	error_2	error_3	error_4	error_10
1	1.00	0.00	1.00	2.00	3.00	9.00
2	2.00	1.00	0.00	1.00	2.00	8.00
3	3.00	2.00	1.00	0.00	1.00	7.00
4	4.00	3.00	2.00	1.00	0.00	6.00
5	10.00	9.00	8.00	7.00	6.00	0.00
6		15.00	12.00	11.00	12.00	30.00

Table: Absolute Error Table

	x	error_1	error_2	error_3	error_4	error_10
1	1.00	0.00	-1.00	-2.00	-3.00	-9.00
2	2.00	1.00	0.00	-1.00	-2.00	-8.00
3	3.00	2.00	1.00	0.00	-1.00	-7.00
4	4.00	3.00	2.00	1.00	0.00	-6.00
5	10.00	9.00	8.00	7.00	6.00	0.00
6		15.00	10.00	5.00	0.00	-30.00

Table: Error Table

	x	error_1	error_2	error_3	error_4	error_10
1	1.00	0.00	1.00	4.00	9.00	81.00
2	2.00	1.00	0.00	1.00	4.00	64.00
3	3.00	4.00	1.00	0.00	1.00	49.00
4	4.00	9.00	4.00	1.00	0.00	36.00
5	10.00	81.00	64.00	49.00	36.00	0.00
6		95.00	70.00	55.00	50.00	230.00

Table: Squared Error Table

Example 2:

```
x <- c(6,10,12,20,52)
x

## [1]  6 10 12 20 52

mean(x)

## [1] 20

median(x)

## [1] 12
```

	x	error_6	error_10	error_12	error_20	error_52
1	6.00	0.00	-4.00	-6.00	-14.00	-46.00
2	10.00	4.00	0.00	-2.00	-10.00	-42.00
3	12.00	6.00	2.00	0.00	-8.00	-40.00
4	20.00	14.00	10.00	8.00	0.00	-32.00
5	52.00	46.00	42.00	40.00	32.00	0.00

Table: Error Table

	x	error_6	error_10	error_12	error_20	error_52
1	6.00	0.00	-4.00	-6.00	-14.00	-46.00
2	10.00	4.00	0.00	-2.00	-10.00	-42.00
3	12.00	6.00	2.00	0.00	-8.00	-40.00
4	20.00	14.00	10.00	8.00	0.00	-32.00
5	52.00	46.00	42.00	40.00	32.00	0.00

Table: Error Table

	x	error_6	error_10	error_12	error_20	error_52
1	6.00	0.00	4.00	6.00	14.00	46.00
2	10.00	4.00	0.00	2.00	10.00	42.00
3	12.00	6.00	2.00	0.00	8.00	40.00
4	20.00	14.00	10.00	8.00	0.00	32.00
5	52.00	46.00	42.00	40.00	32.00	0.00
6		70.00	58.00	56.00	64.00	160.00

Table: Absolute Error Table

	x	error_6	error_10	error_12	error_20	error_52
1	6.00	0.00	-4.00	-6.00	-14.00	-46.00
2	10.00	4.00	0.00	-2.00	-10.00	-42.00
3	12.00	6.00	2.00	0.00	-8.00	-40.00
4	20.00	14.00	10.00	8.00	0.00	-32.00
5	52.00	46.00	42.00	40.00	32.00	0.00

Table: Error Table

	x	error_6	error_10	error_12	error_20	error_52
1	6.00	0.00	16.00	36.00	196.00	2116.00
2	10.00	16.00	0.00	4.00	100.00	1764.00
3	12.00	36.00	4.00	0.00	64.00	1600.00
4	20.00	196.00	100.00	64.00	0.00	1024.00
5	52.00	2116.00	1764.00	1600.00	1024.00	0.00
6		2364.00	1884.00	1704.00	1384.00	6504.00

Table: Squared Error Table

Minimizing the "Total" Error Function:

Mean minimizes $\sum(y_i - a)^2$

Median minimizes $\sum|y_i - a|$

Minimization Problem

$$\begin{aligned}\sum (y_i - a)^2 &= (y_1 - a)^2 + (y_2 - a)^2 + (y_3 - a)^2 \\ &\quad + \dots \\ &\quad + (y_N - a)^2\end{aligned}$$

Mathematically, how do you find an a to minimize this function?

$$
\begin{aligned}
\sum (y_i - a)^2 &= (y_1 - a)^2 + (y_2 - a)^2 + (y_3 - a)^2 \\
&\quad + \dots \\
&\quad + (y_N - a)^2 \\
\longrightarrow \frac{\partial \sum (y_i - a)^2}{\partial a} &= 2(y_1 - a) \times (-1) + 2(y_2 - a) \times (-1) \\
&\quad + 2(y_3 - a) \times (-1) + \dots \\
&\quad + 2(y_N - a) \times (-1) \\
&= \sum_{i=1}^{N} (-2)(y_i - a) \\
&= 0
\end{aligned}
$$

Minimization Problem

$$
\begin{aligned}
\sum_{i=1}^{N}(-2)(y_i - a) &= 0 \\
\sum_{i=1}^{N}((y_i - a) &= 0 \\
\sum_{i=1}^{N} y_i - \sum_{i=1}^{N} a &= 0 \\
\sum_{i=1}^{N} y_i &= \sum_{i=1}^{N} a \\
\sum_{i=1}^{N} y_i &= N \cdot a \\
\frac{\sum_{i=1}^{N} y_i}{N} &= a
\end{aligned}
$$

Minimization of an Absolute Value function

This is more mathematically involved since the objective function is continuous, but not differnetiable. And the gradient or derivative function is not even continuous!

Question: What would this function give you?

$$\text{Function} = \left\{ \begin{array}{ll} .25, & \text{If } (y_i - a) \geq 0 \\ .75, & \text{If } (y_i - a) < 0 \end{array} \right\}$$

Question: What would this function give you?

$$\text{Function} = \left\{ \begin{array}{ll} .25, & \text{If } (y_i - a) \geq 0 \\ .75, & \text{If } (y_i - a) < 0 \end{array} \right\}$$

25^{th} percentile!

Question: What would this function give you?

$$\text{Function} = \left\{ \begin{array}{ll} \tau, & \text{If } (y_i - a) \geq 0 \\ 1 - \tau, & \text{If } (y_i - a) < 0 \end{array} \right\}$$

Question: What would this function give you?

$$\text{Function} = \left\{ \begin{array}{ll} \tau, & \text{If } (y_i - a) \geq 0 \\ 1 - \tau, & \text{If } (y_i - a) < 0 \end{array} \right\}$$

τ^{th} percentile!

Model: An Alternative Approach to Consider these Results

$$y_i = a \cdot x_i + \epsilon_i$$

	y	x
1	1.00	1.00
2	2.00	1.00
3	3.00	1.00
4	4.00	1.00
5	10.00	1.00

Table: Hypothetical Data

Model: An Alternative Approach to Consider these Results

$$
\begin{aligned}
y_i &= a \cdot 1 + \epsilon_i \\
&= a + \epsilon_i
\end{aligned}
$$

Model: An Alternative Approach to Consider these Results

$$\begin{aligned} y_i &= a \cdot 1 + \epsilon_i \\ &= a + \epsilon_i \end{aligned}$$

Finding a that would minimize the “total” error!

[**Sum of Squared Errors**]: $a =$ mean!

[**Sum of Absolute Errors**]: $a =$ median!

Model: An Alternative Approach to Consider these Results

$$\begin{aligned} y_i &= a \cdot 1 + \epsilon_i \\ &= a + \epsilon_i \end{aligned}$$

Finding a that would minimize the "total" error!

[**Sum of Squared Errors**]: $a =$ mean!

[**Sum of Absolute Errors**]: $a =$ median!

Question: Why?!!

No Useful Information! Your best forecast is again **mean** if your goal is to minimize the sum of squared errors.

Model: An Alternative Approach to Consider these Results

$$y_i = a \cdot x_i + \epsilon_i$$

	y	x
1	1.00	2.00
2	2.00	2.00
3	3.00	2.00
4	4.00	2.00
5	10.00	2.00

Table: Hypothetical Data