Overfitting and Model Tuning

Overfitting and Model Tuning

author: Son Nguyen

Reading Materials

- Max Kuhn. Chapter 4.

Prediction Problem

Given data of $X = [X_1, X_2, ..., X_d]$ and Y. Find the relation between X and Y.

input variables

label (output variable)

Prediction Problem - Examples

- One Input Variable X

continuous

X	Y
13	4.0
6	3.5
14	3
10	3.9
7	2.7
12	3.8
1	1.5

How are X and Y related?

a regression problem

Prediction Problem - Examples

- Multiple Input Variables

categorical

X_1	X_2	...	X_{35}	Y
1	-1	...	2	Tree
2.1	0	...	6	Not a Tree
3	0	...	8	Tree

How are X and Y related?

a classification problem

Prediction Problem

- If Y is continous, we have a **regression** problem.
- If Y is categorical, we have a **classification** problem.
- If Y is binary, we have a **binary classification** problem.

Prediction Problem - Examples

- This is a regression problem since Y is continuous.

X	Y
13	4.0
6	3.5
14	3
10	3.9
7	2.7
12	3.8
1	1.5

Prediction Problem - Examples

- This is a binary classification Problem since Y is binary.

X_1	X_2	...	X_{35}	Y
1	-1	...	2	Tree
2.1	0	...	6	Not a Tree
3	0	...	8	Tree

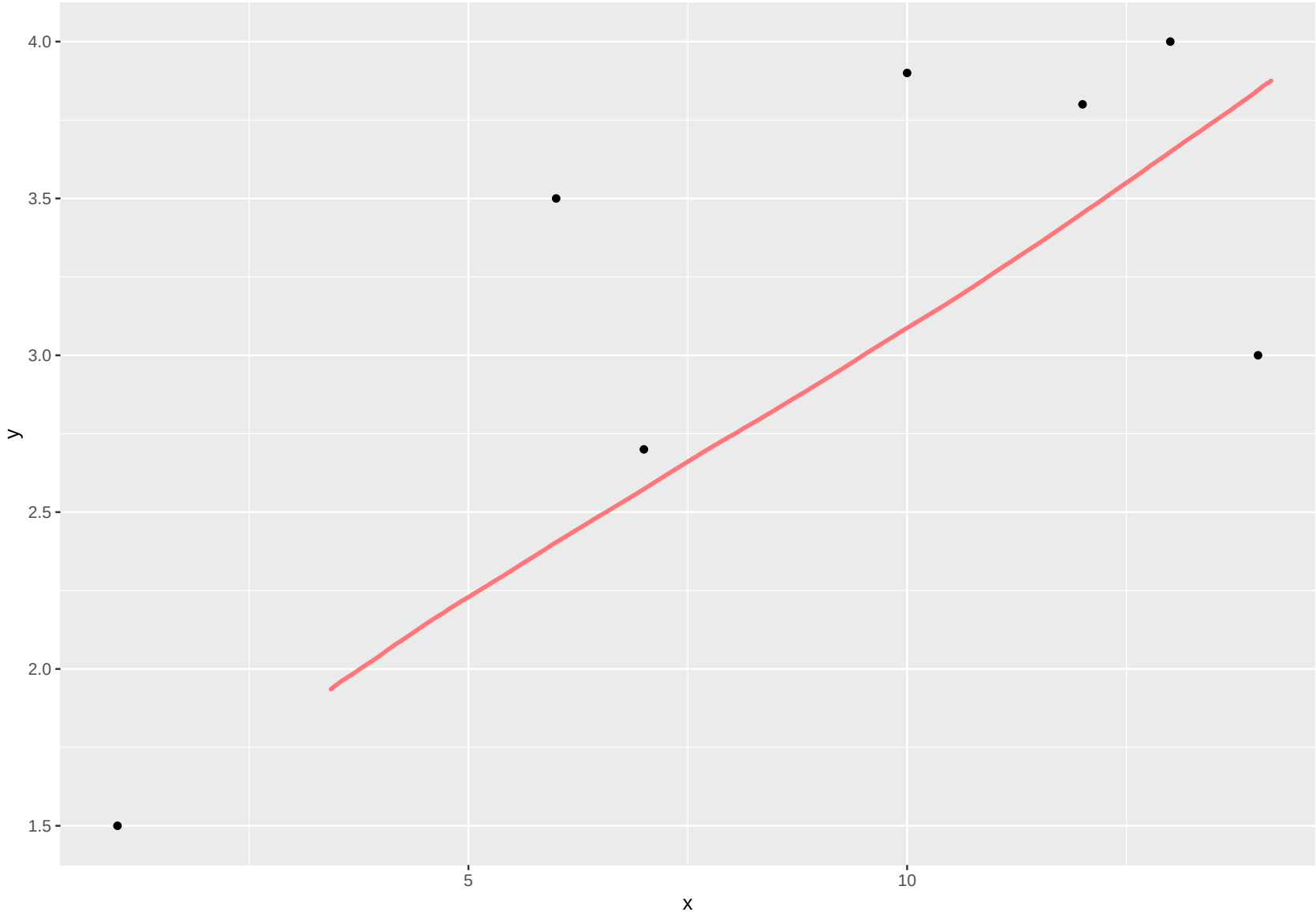
Overfitting

- Consider the data:

X	Y
13	4.0
6	3.5
14	3
10	3.9
7	2.7
12	3.8
1	1.5

- We will fit these data by polynomial model.
- In polynomial model, Y is a polynomial function of X.

Overfitting



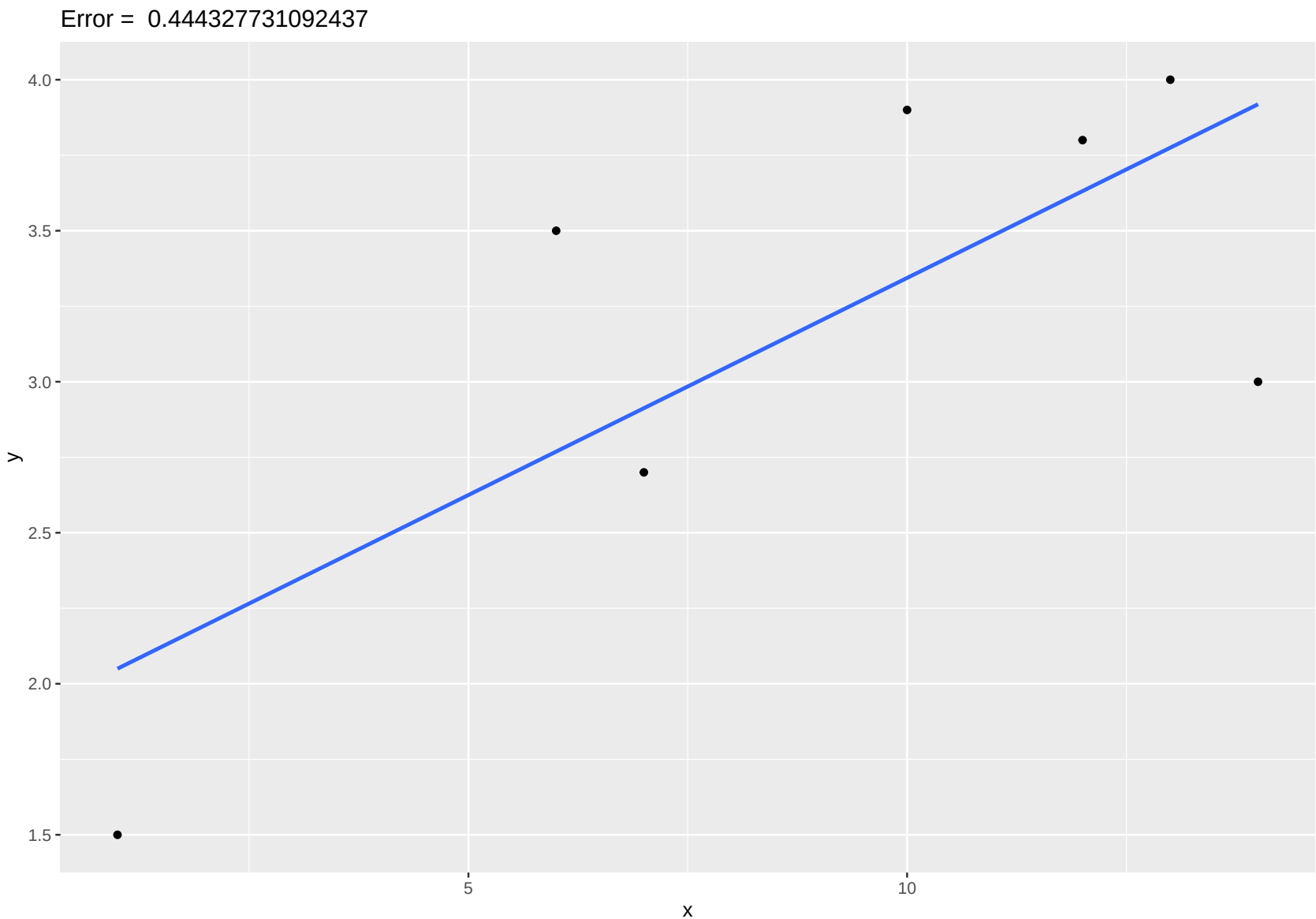
- We will fit these data by **polynomial model**.
- In polynomial model, Y is a polynomial function of X.

Overfitting - Polynomial Model

- In polynomial model, we need to specify the degree of the polynomial, n. Let try a few.
- If $n = 1$, we have a familiar **linear model**.
- **Question: Does increasing n resuls in a better model?**

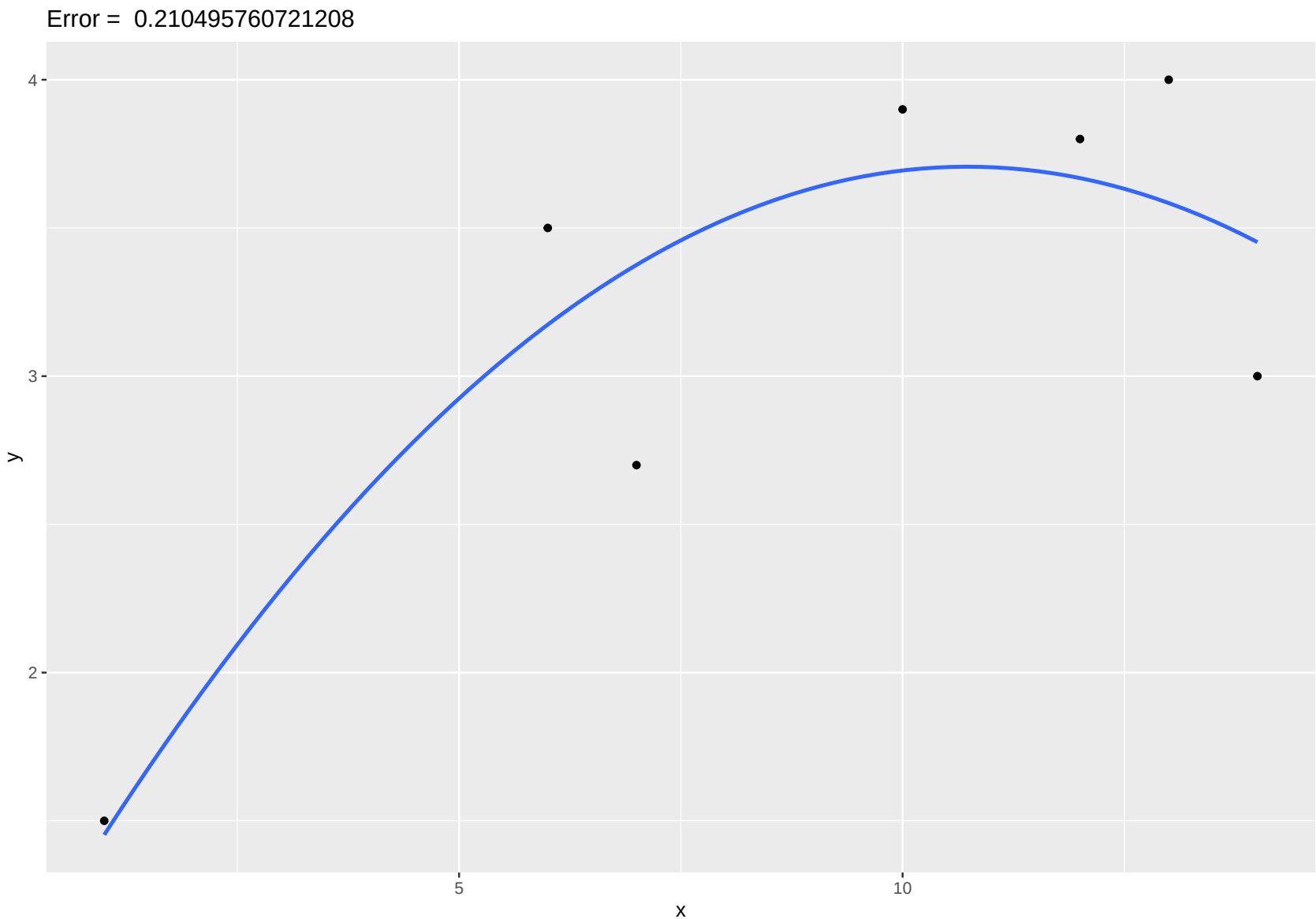
Overfitting - Polynomial Model

- $n = 1$. (linear model)



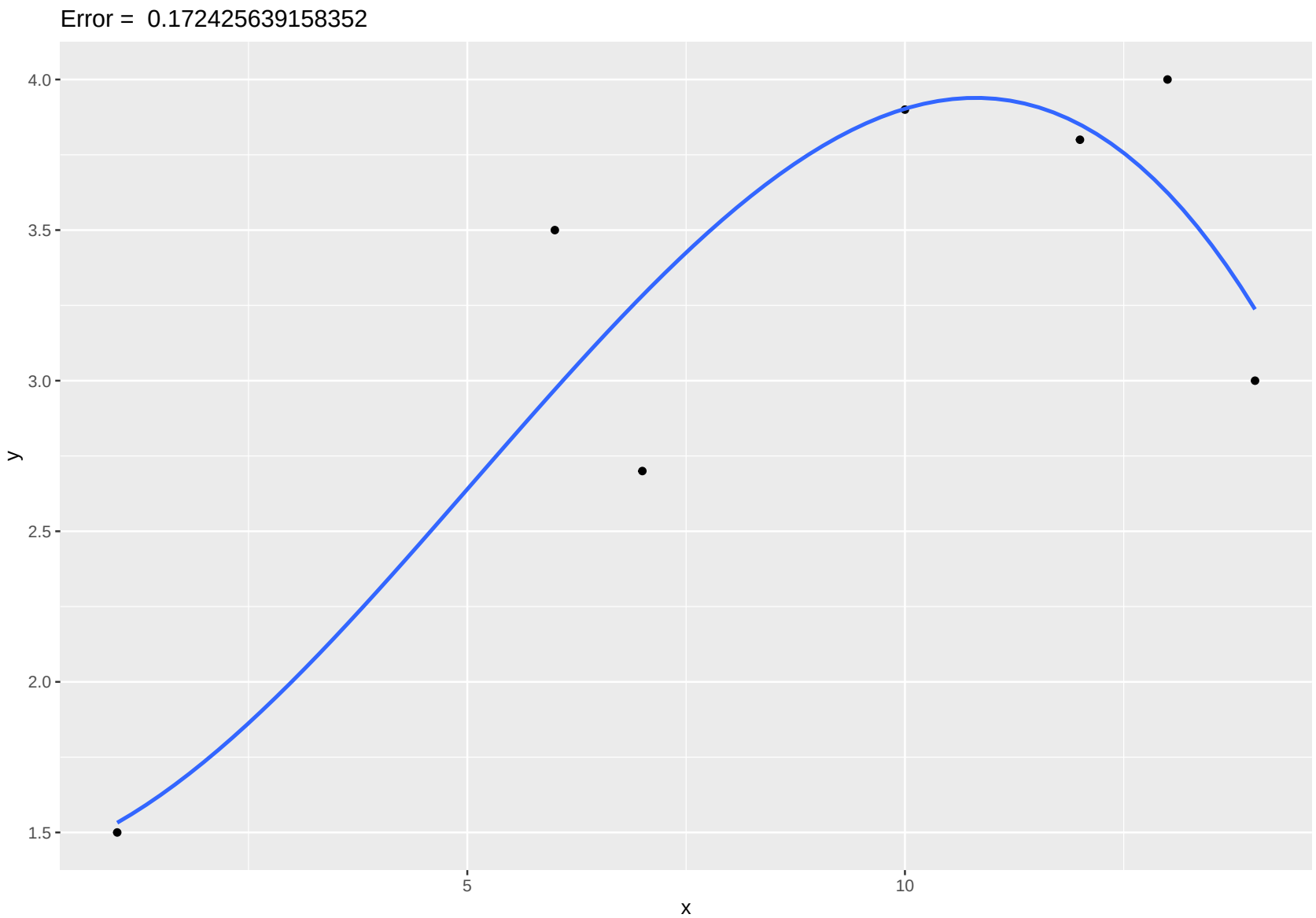
Overfitting - Polynomial Model

- $n = 2$.



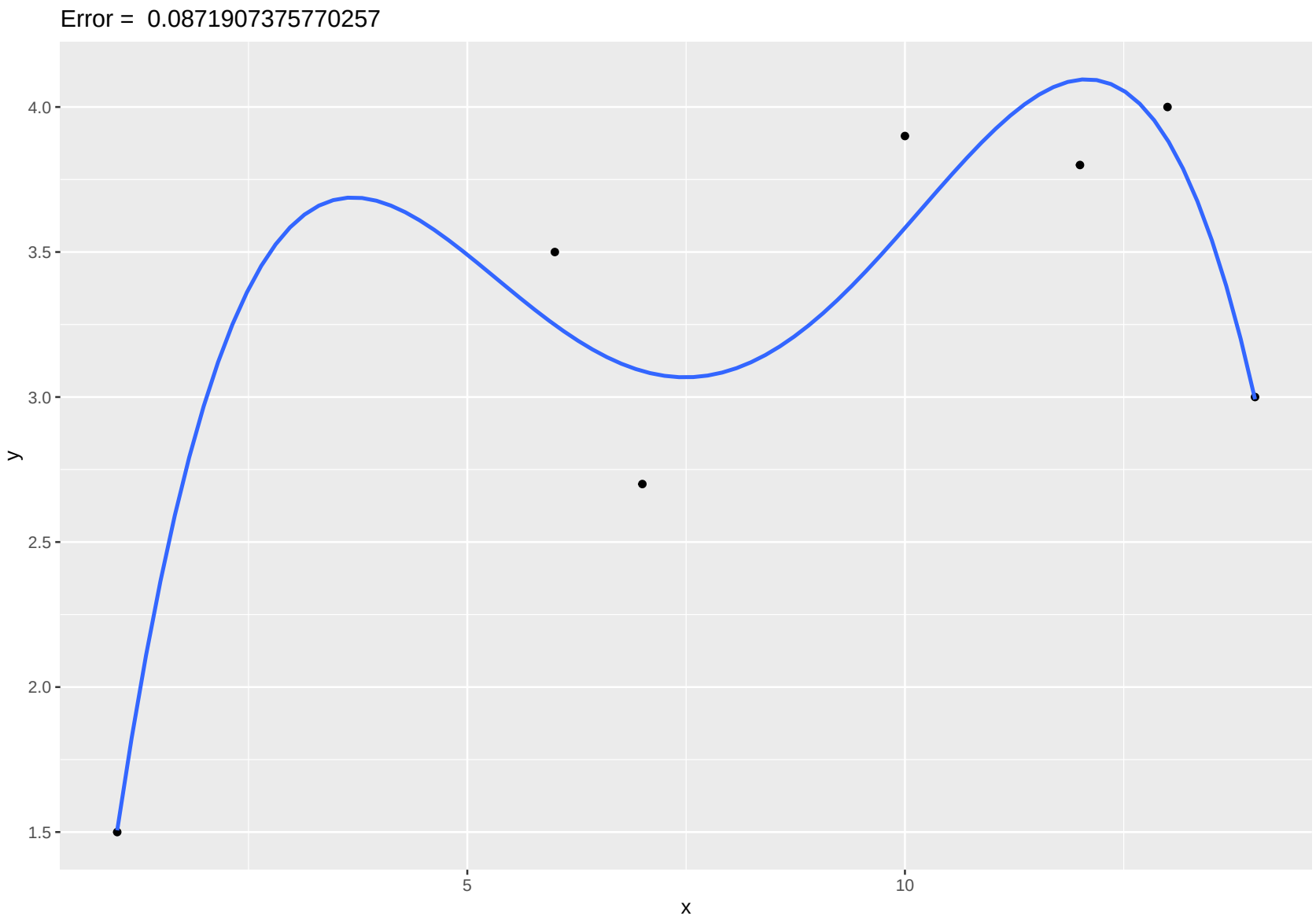
Overfitting - Polynomial Model

- $n = 3$.



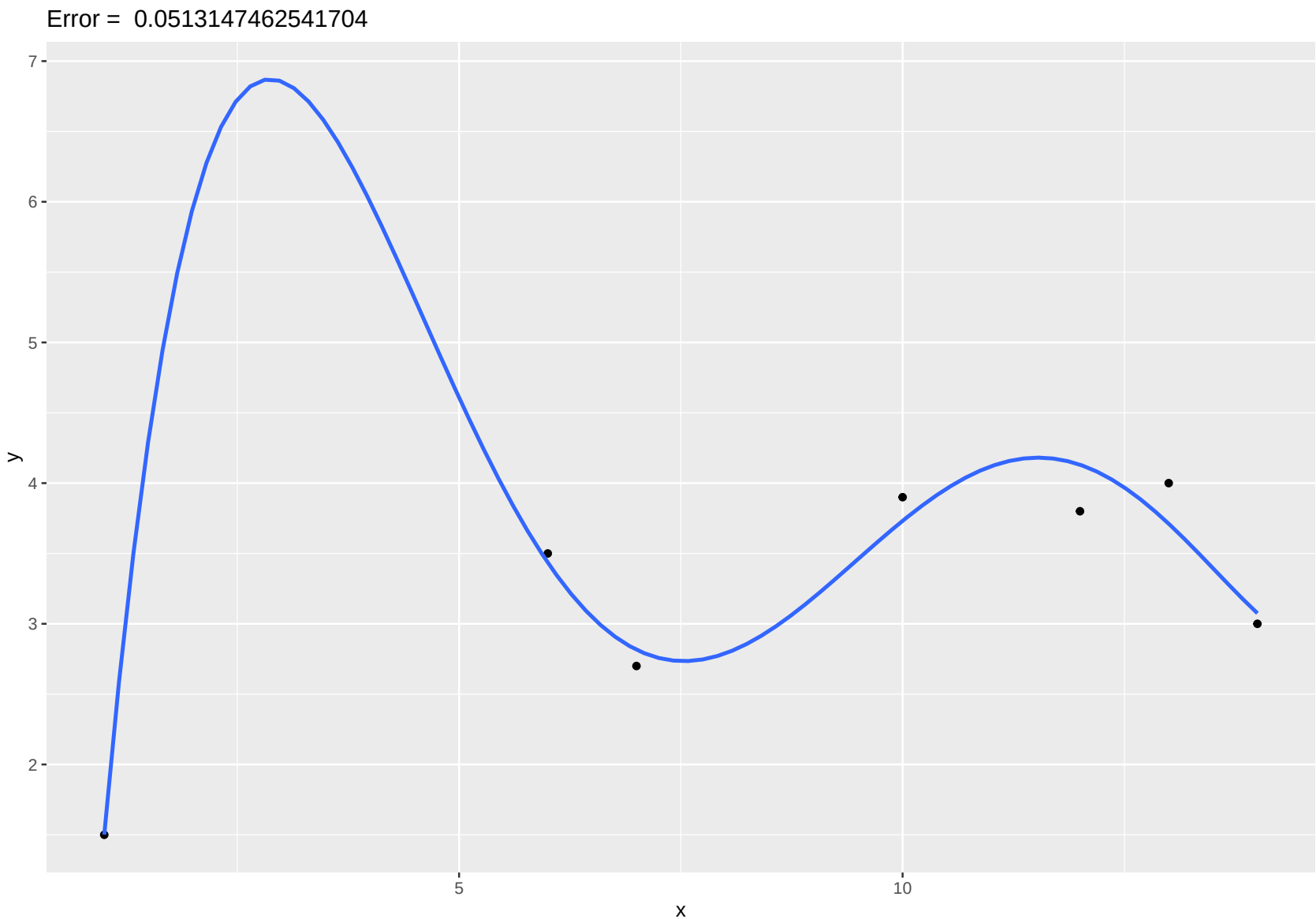
Overfitting - Polynomial Model

- $n = 4$.



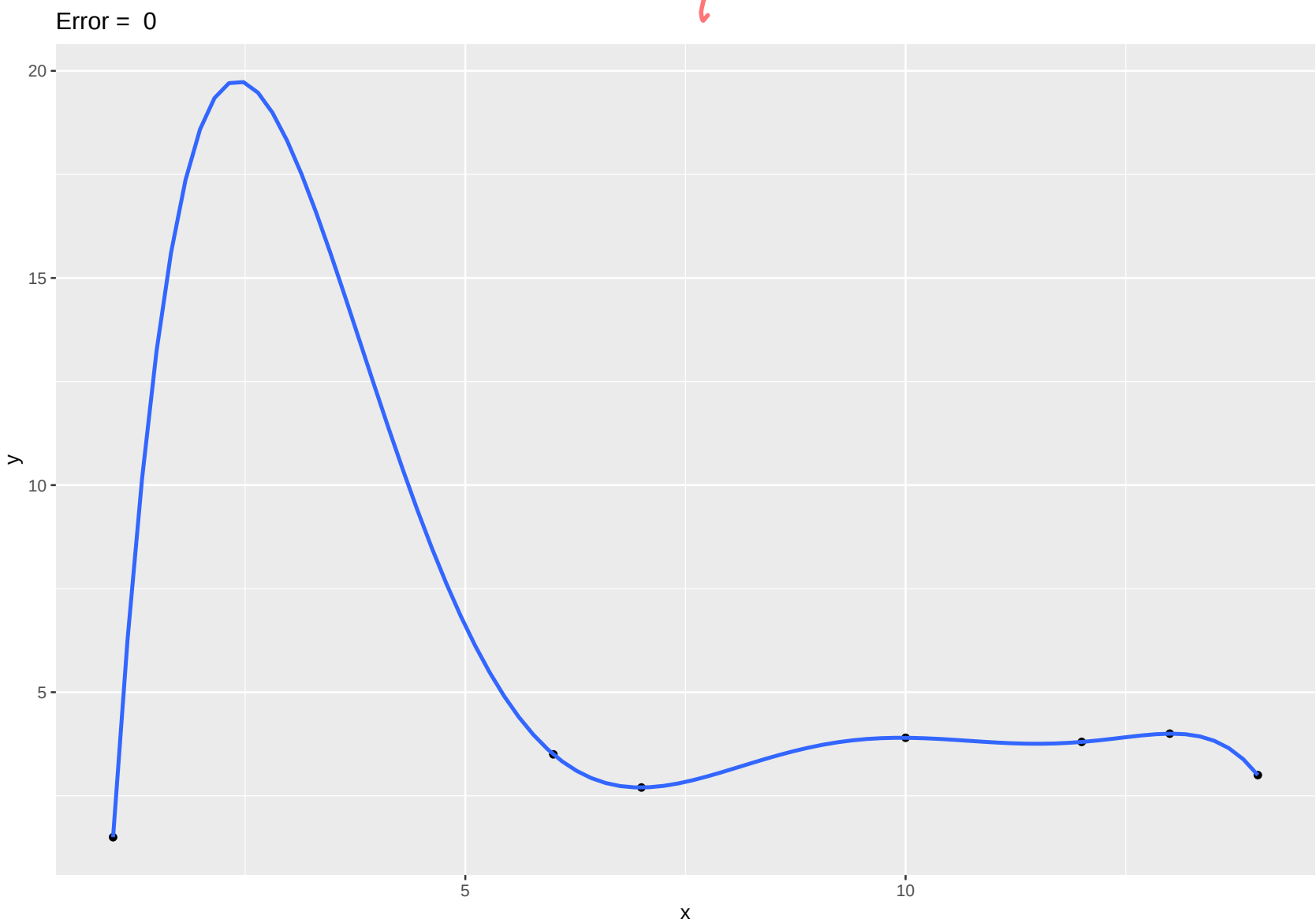
Overfitting - Polynomial Model

- $n = 5$.



Overfitting - Polynomial Model

- $n = 6$.



Overfitting - Polynomial Model

- **Question**: What are the errors when $n > 6$?