Deconstructing Data Science

David Bamman, UC Berkeley

Info 290
Lecture 4: Regression overview

Jan 26, 2017

Regression

A mapping from input data x (drawn from instance space $\mathcal{X}$) to a point y in $\mathbb{R}$

($\mathbb{R}$ = the set of real numbers)

x = the empire state building
y = 17444.5625"

task	x	y
predicting box office revenue	movie	total box office
predicting stock movements	$TWTR	price at time t+1
predicting vote share	Clinton	47%

Regression

Supervised learning

Given training data in the form of <x, y> pairs, learn $\hat{h}(x)$

Regression

- Can you create (or find) labeled data that marks that value for a bunch of examples? ~~Can *you* make that choice?~~

- Can you create features that might help in distinguishing those classes?

Experiment design

	training	development	testing
size	80%	10%	10%
purpose	training models	model selection	evaluation; never look at it until the very end

Metrics

- Measure difference between the prediction ŷ and the true y

Mean squared error (MSE)

$$\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2$$

Mean absolute error (MAE)

$$\frac{1}{N}\sum_{i=1}^{N}\left|\hat{y}_i - y_i\right|$$

y	ŷ	MAE	MSE
1	2	1	1
1	1.1	0.1	0.01
1	100	99	9801
1	5	4	16
1	-5	6	36
1	10	9	81
1	3	2	4
1	0.9	0.1	0.01
1	1	0	0
		121.2	9939.02

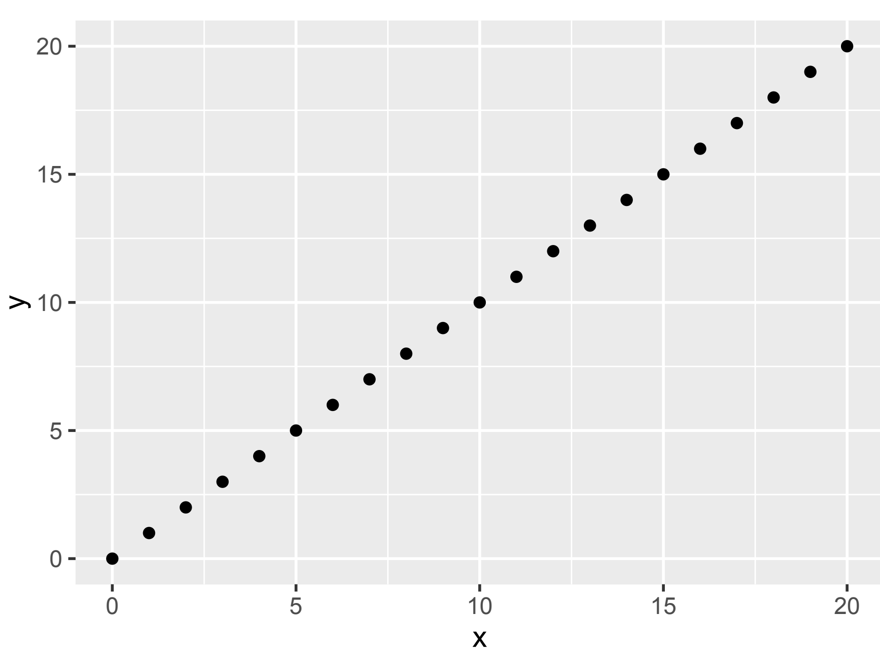
81.7% of total MAE →

← 98.6% of total MSE

MSE error penalizes outliers more than MAE

Linear regression

$$\hat{y} = \sum_{i=1}^{F} x_i \beta_i$$

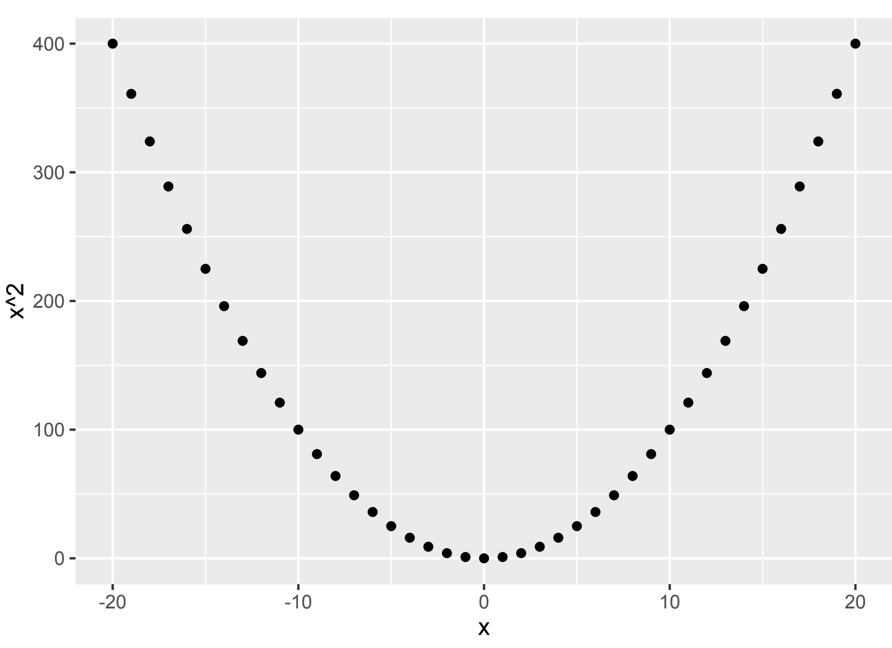


$\beta \in \mathbb{R}^F$

(F-dimensional vector of real numbers)

Polynomial regression

$$\hat{y} = \sum_{i=1}^{F} x_i \beta_{a,i} + \sum_{i=1}^{F} x_i^2 \beta_{b,i}$$

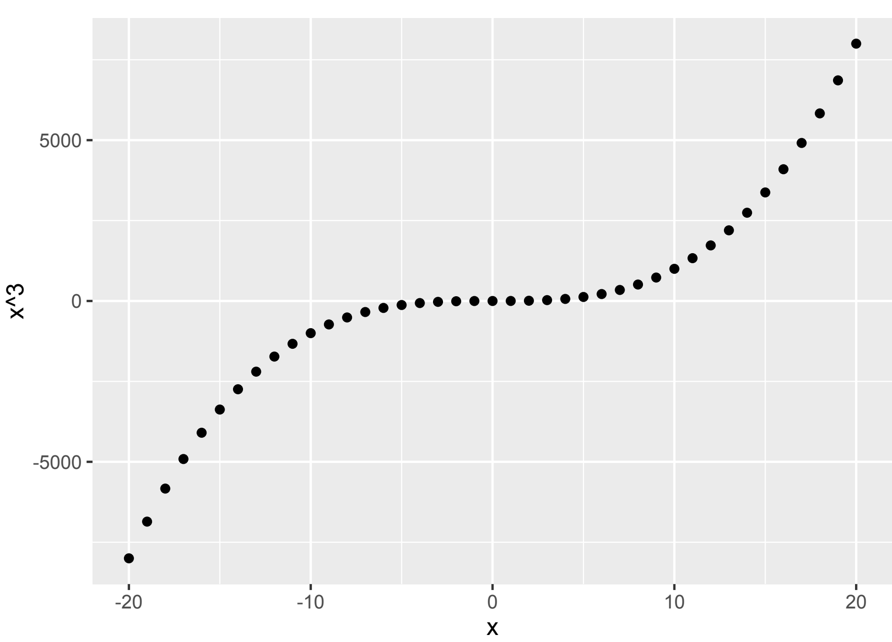


$\beta_a, \beta_b \in \mathbb{R}^F$

(F-dimensional vector of real numbers)

Polynomial regression

$$\hat{y} = \sum_{i=1}^{F} x_i \beta_{a,i} + \sum_{i=1}^{F} x_i^2 \beta_{b,i} + \sum_{i=1}^{F} x_i^3 \beta_{c,i}$$



$\beta_a, \beta_b, \beta_c \in \mathbb{R}^F$

(F-dimensional vector of real numbers)

Nonlinear regression

Number of Parameters

order 1
(linear reg.)

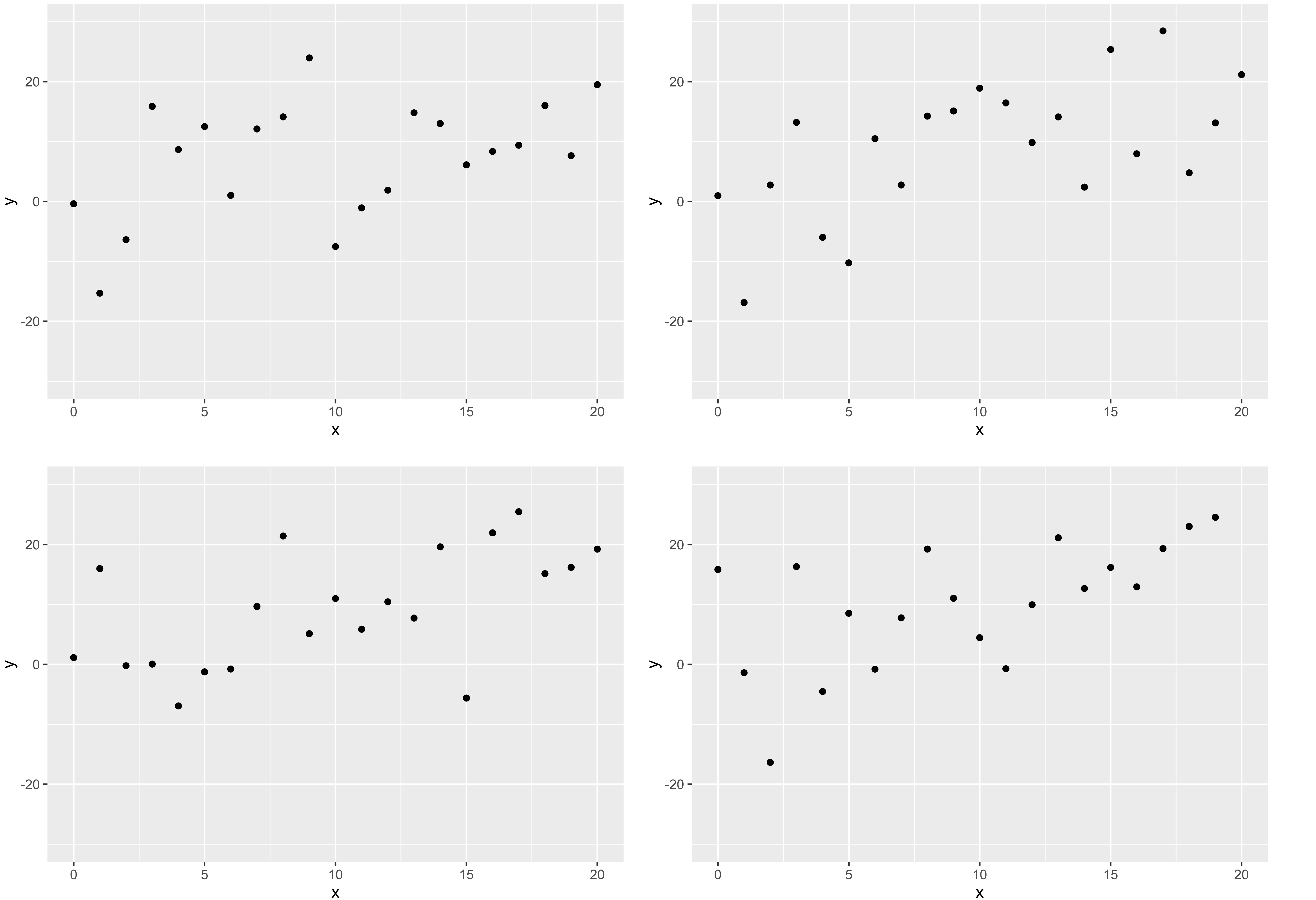
$$\hat{y} = \sum_{i=1}^{F} x_i \beta_{a,i}$$

order 2

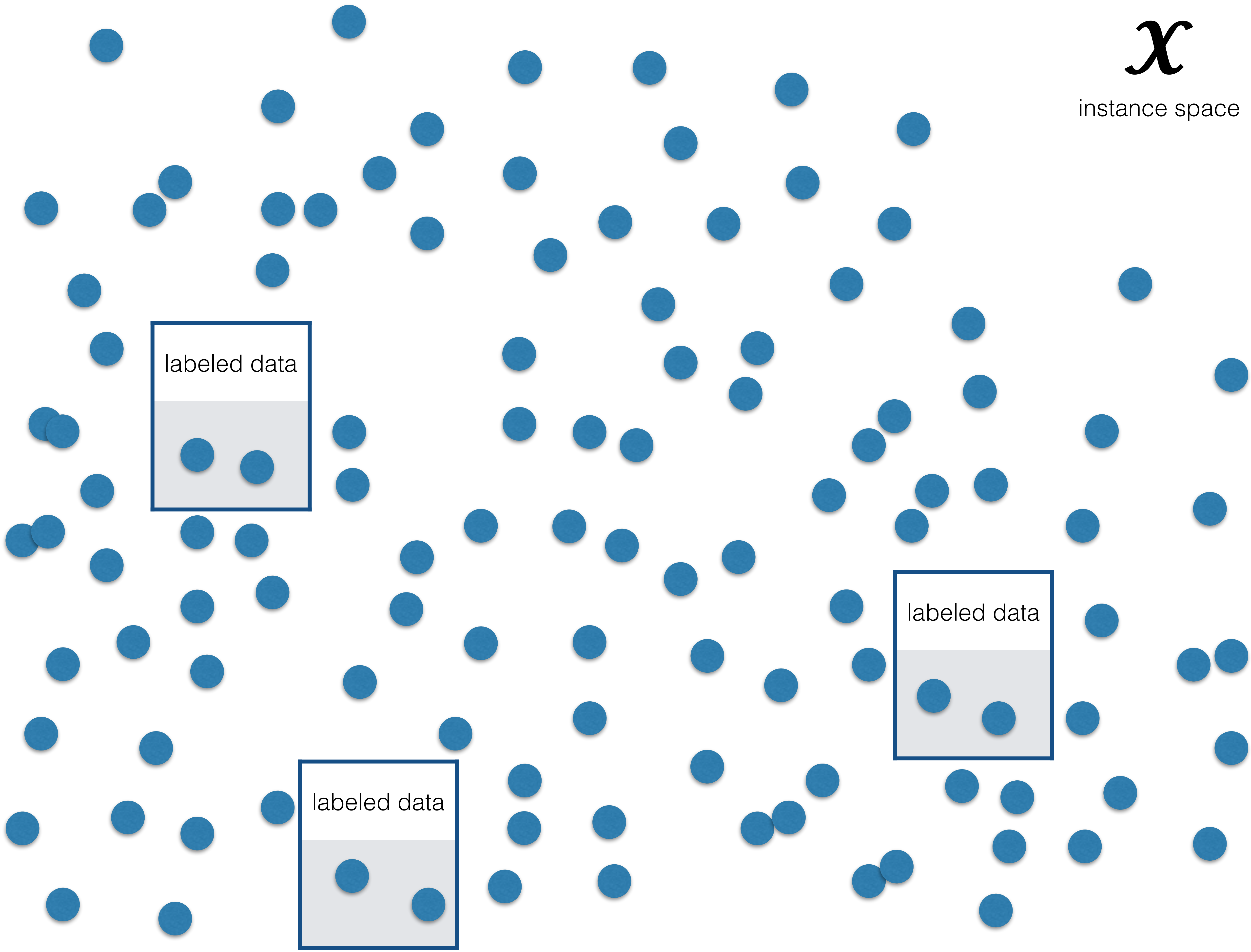
$$\hat{y} = \sum_{i=1}^{F} x_i \beta_{a,i} + \sum_{i=1}^{F} x_i^2 \beta_{b,i}$$

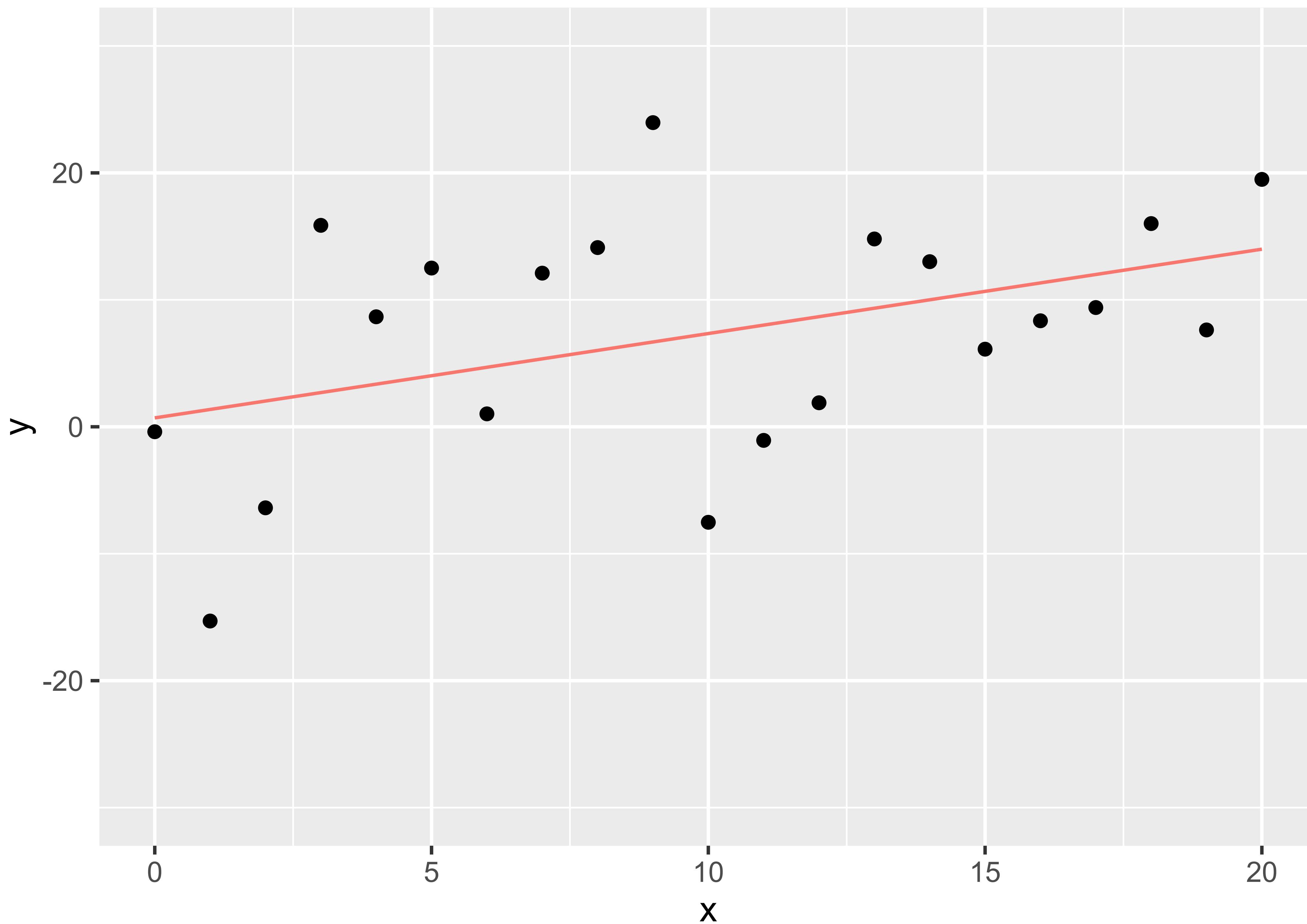
order 3

$$\hat{y} = \sum_{i=1}^{F} x_i \beta_{a,i} + \sum_{i=1}^{F} x_i^2 \beta_{b,i} + \sum_{i=1}^{F} x_i^3 \beta_{c,i}$$

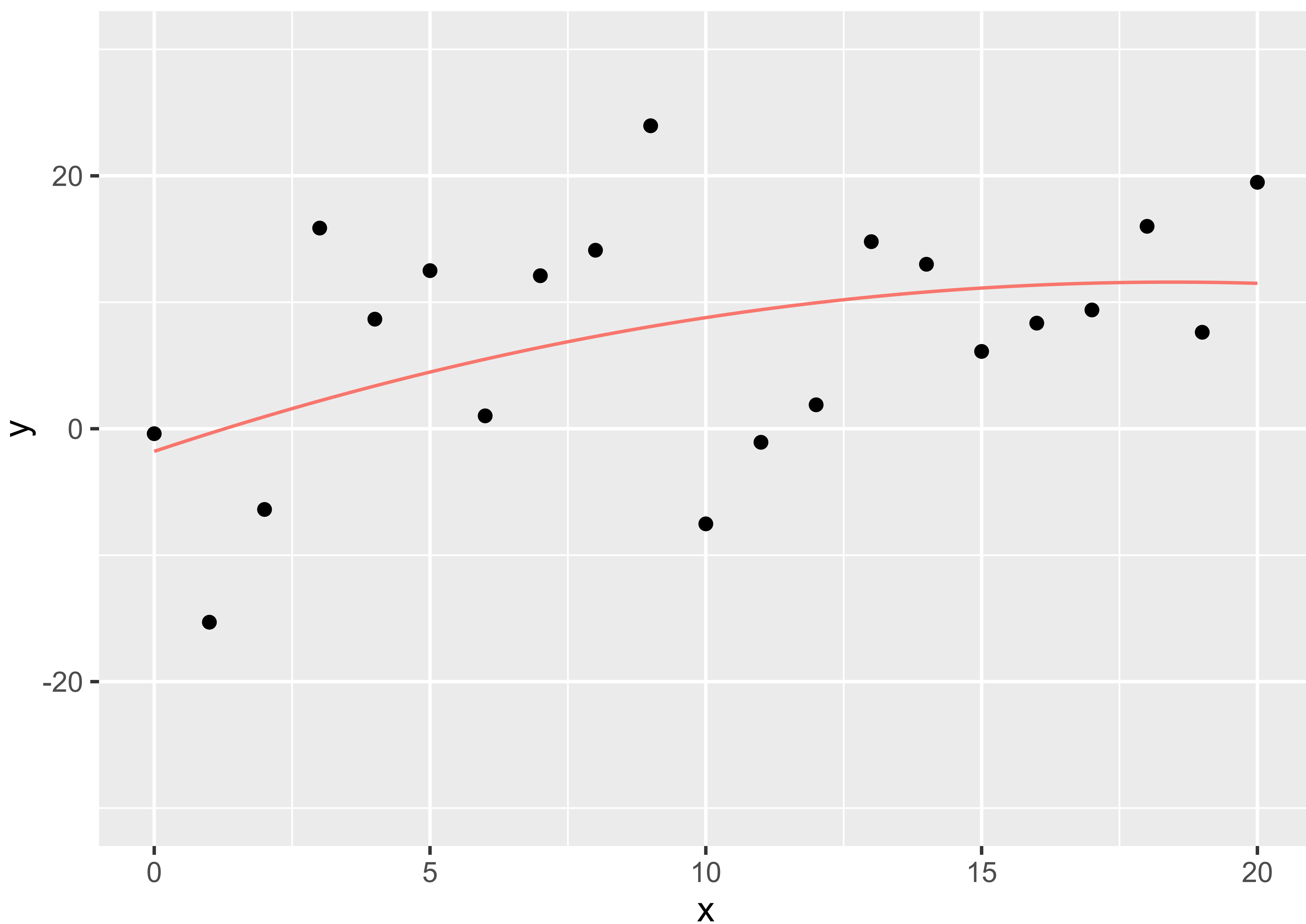


y
x
20
0
-20
0
5
10
15
20

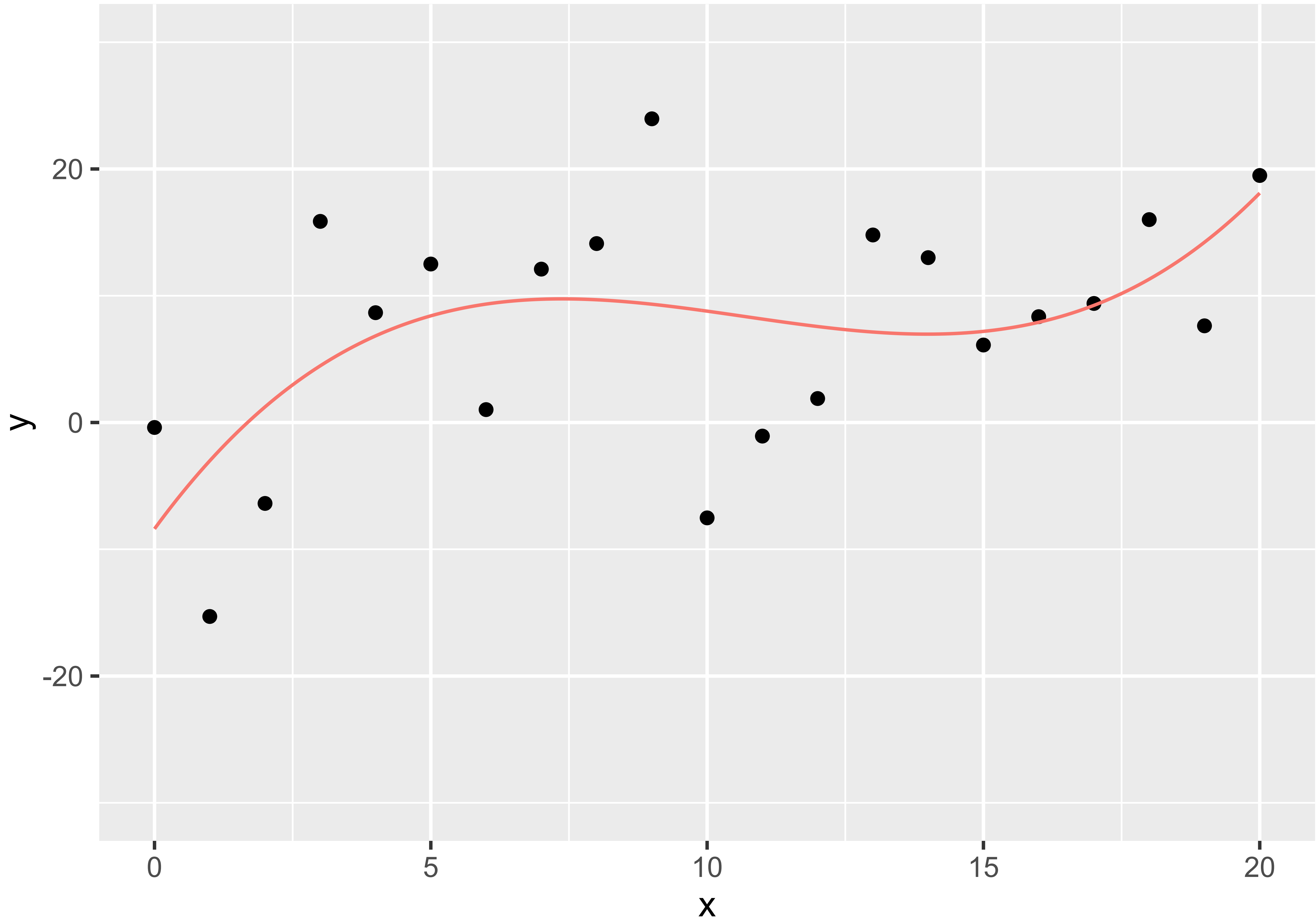

x
instance space
labeled data
labeled data
labeled data



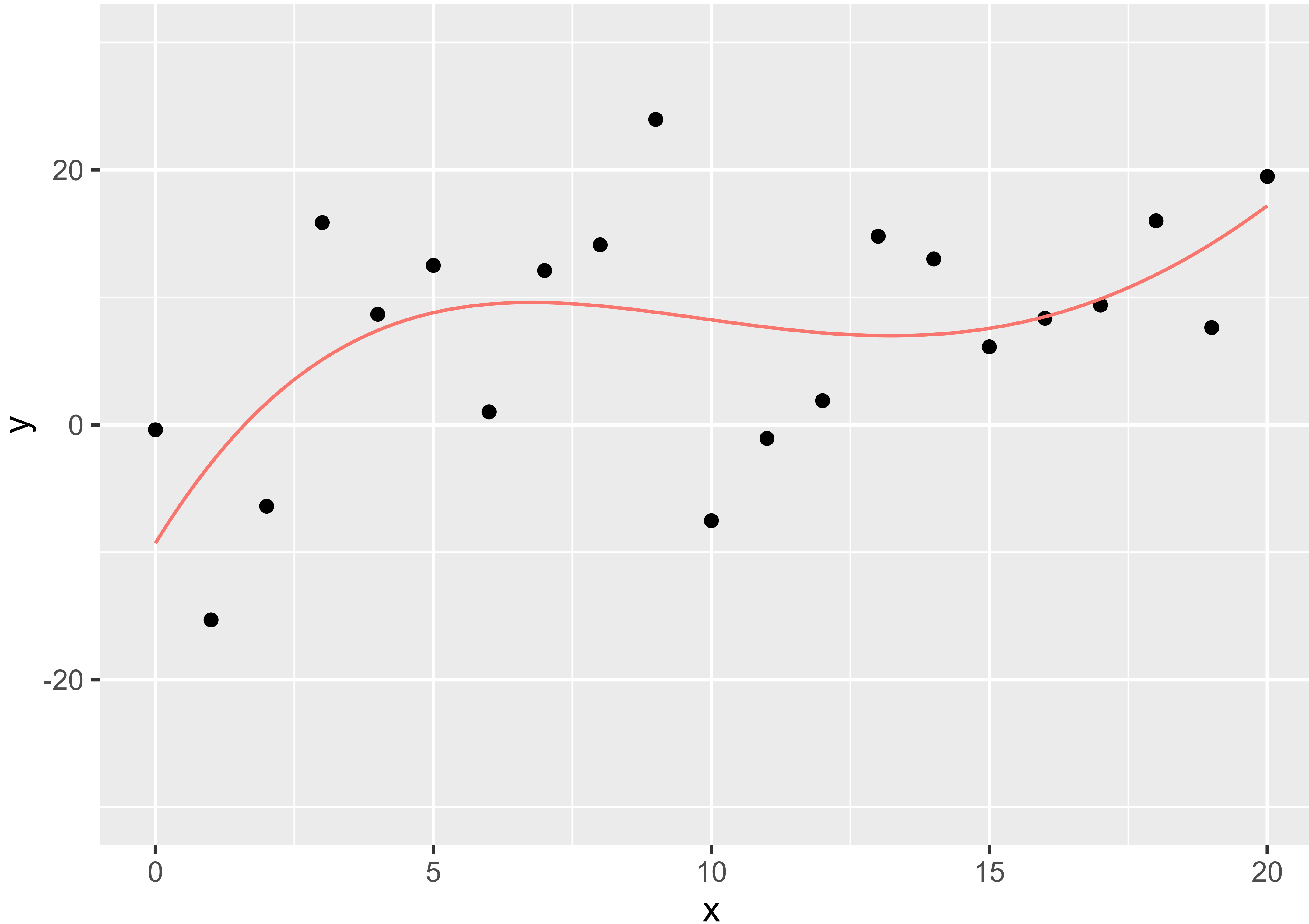
degree 1, training MSE = 73.4



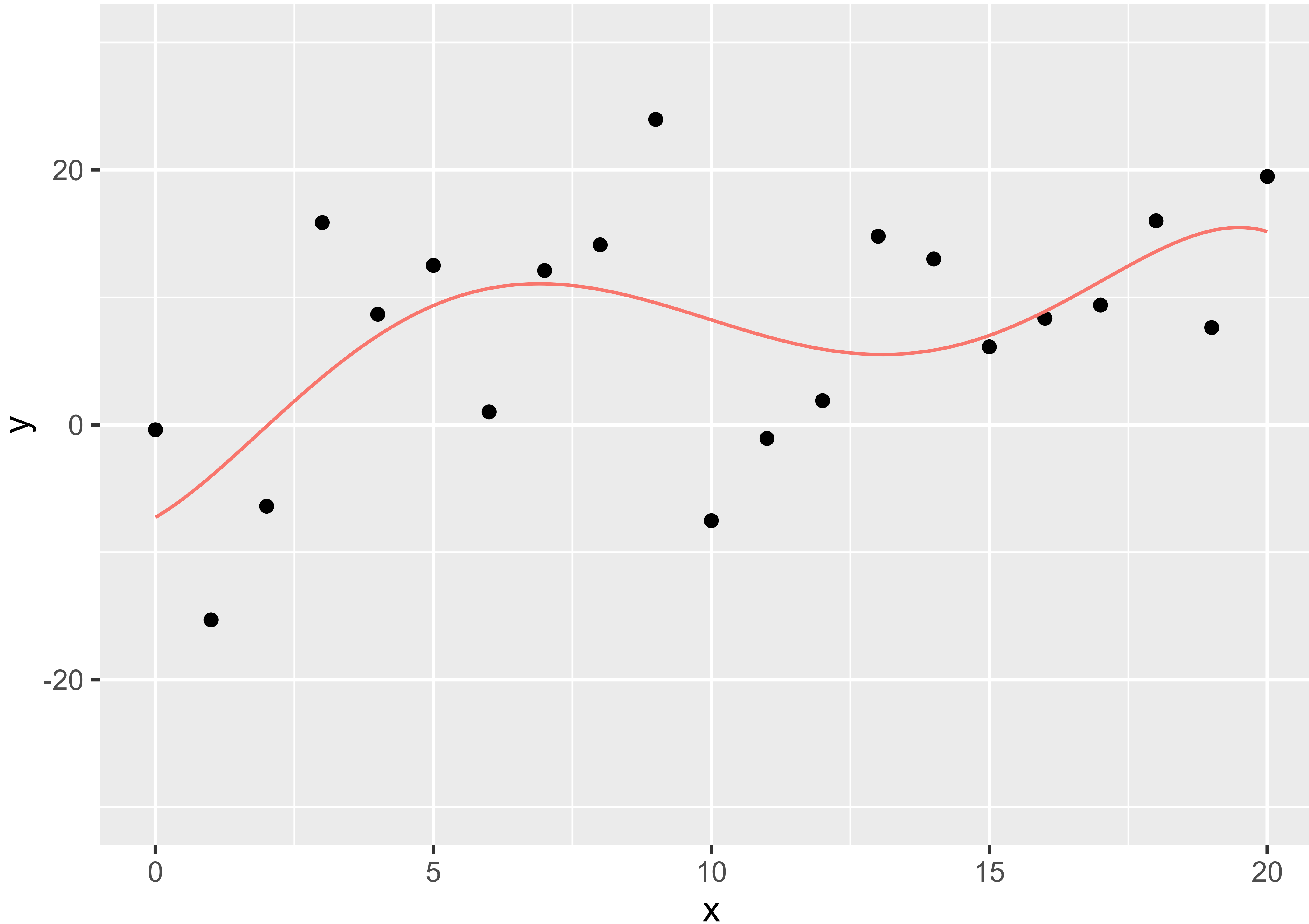
degree 2, training MSE = 71.9



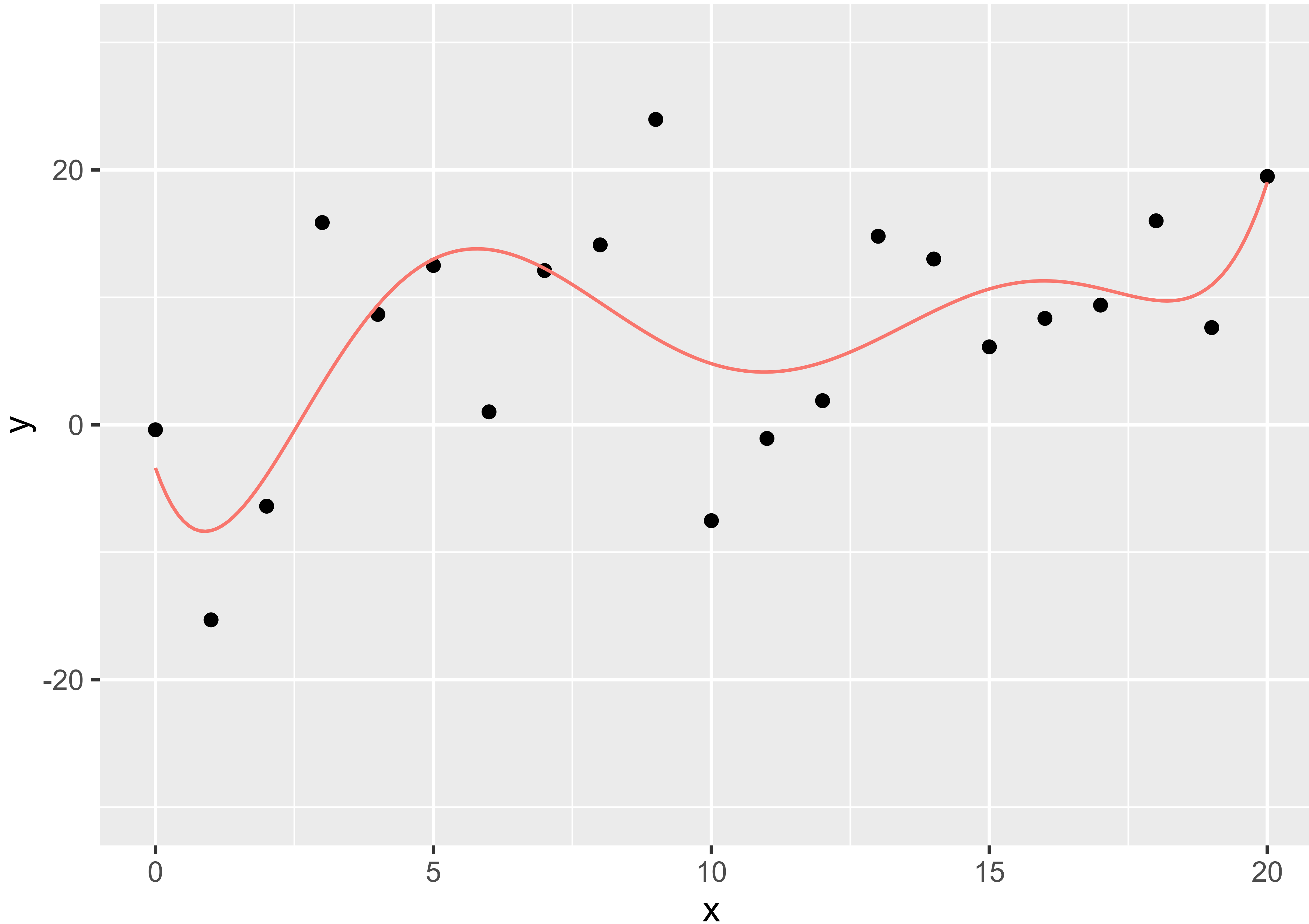
degree 3, training MSE = 60.9



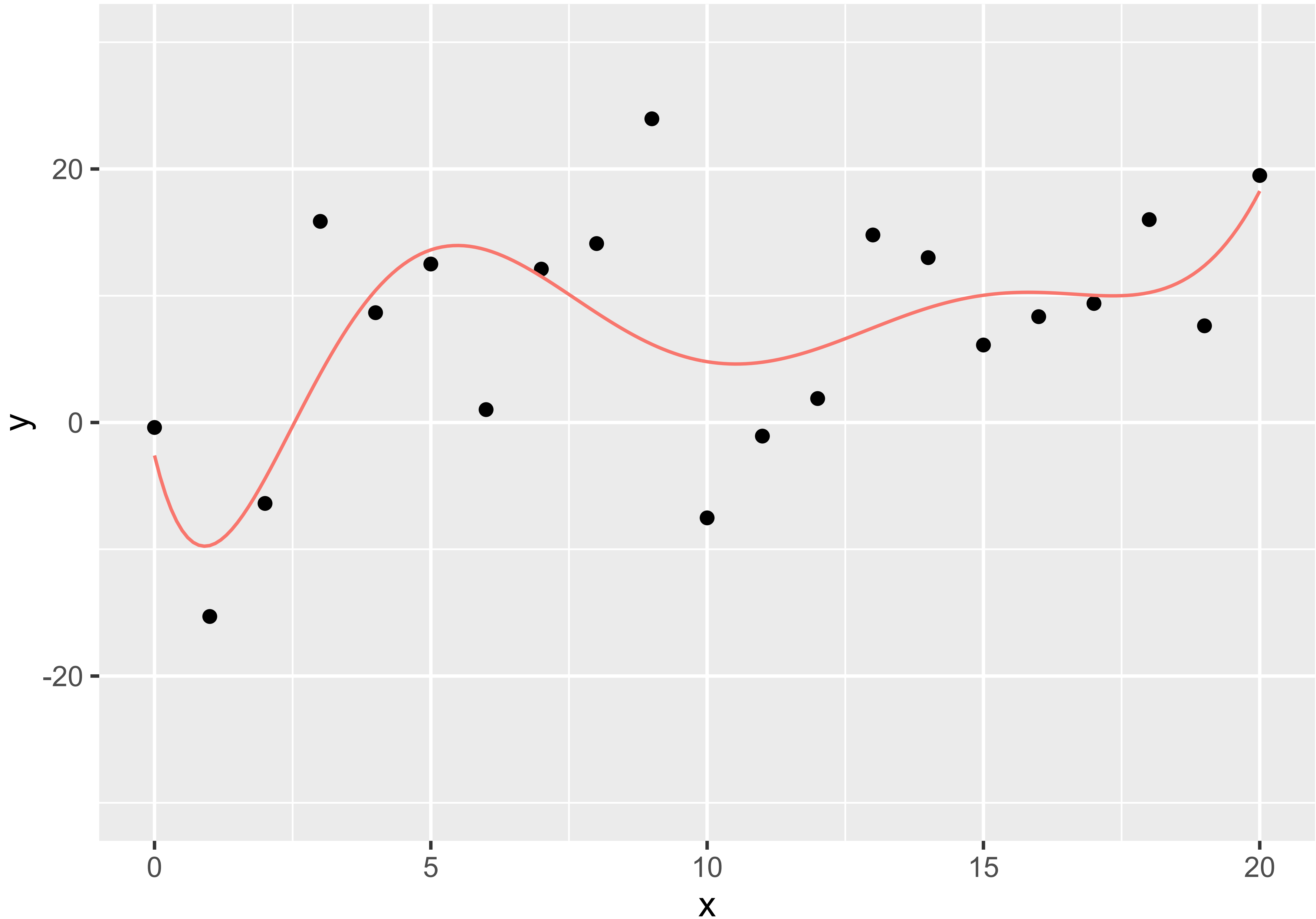
degree 4, training MSE = 60.6



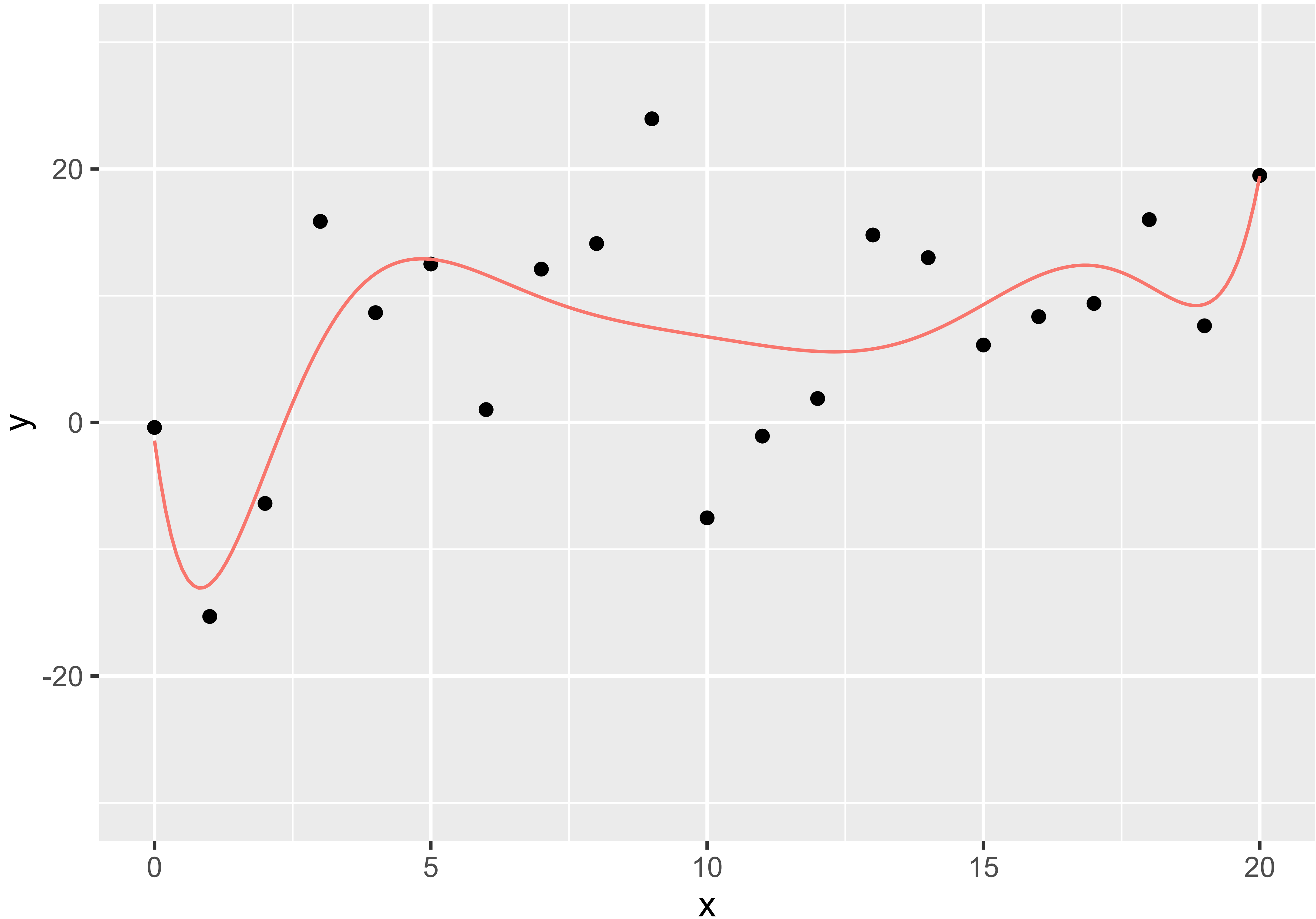
degree 5, training MSE = 59.1



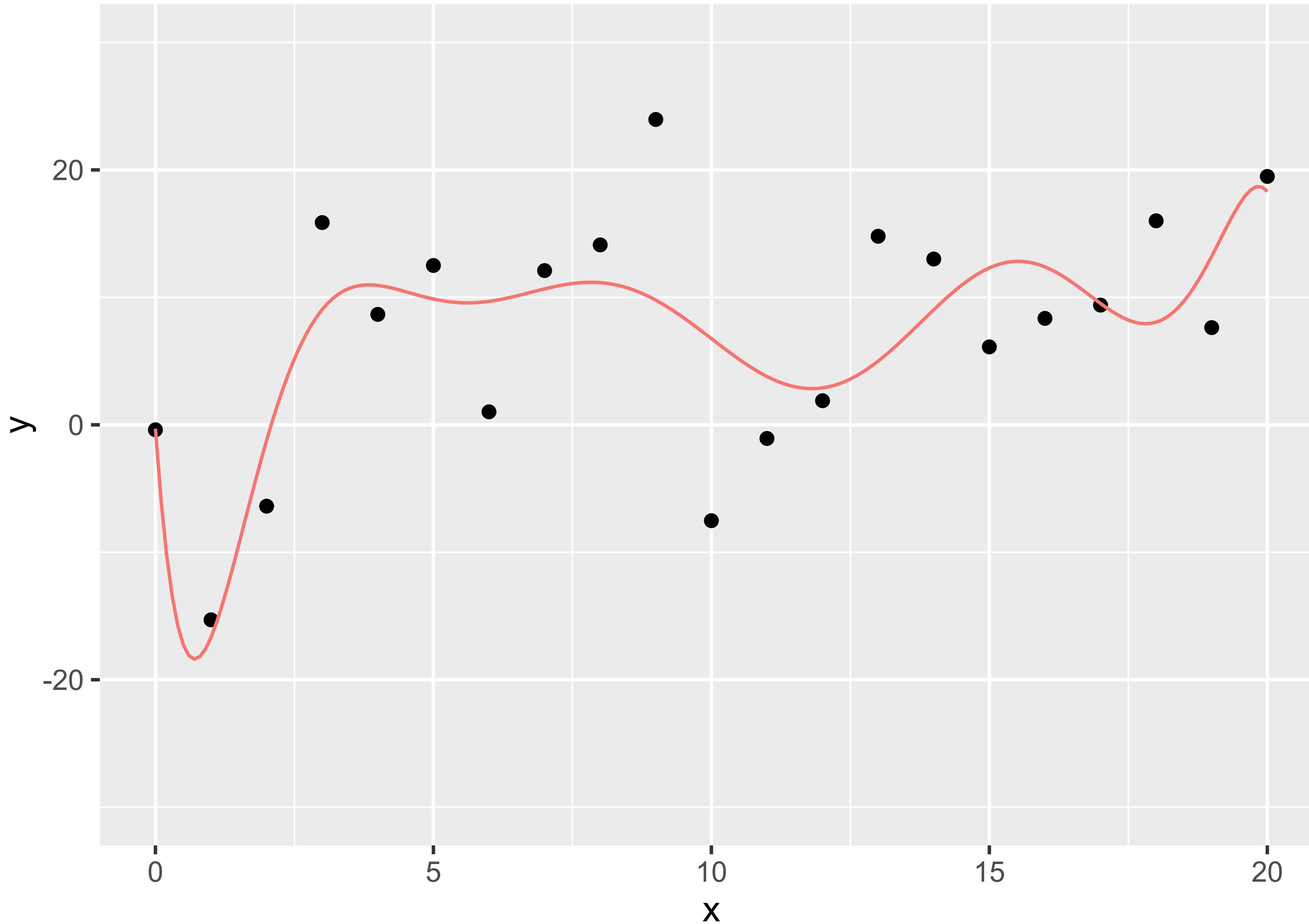
degree 6, training MSE = 50.2



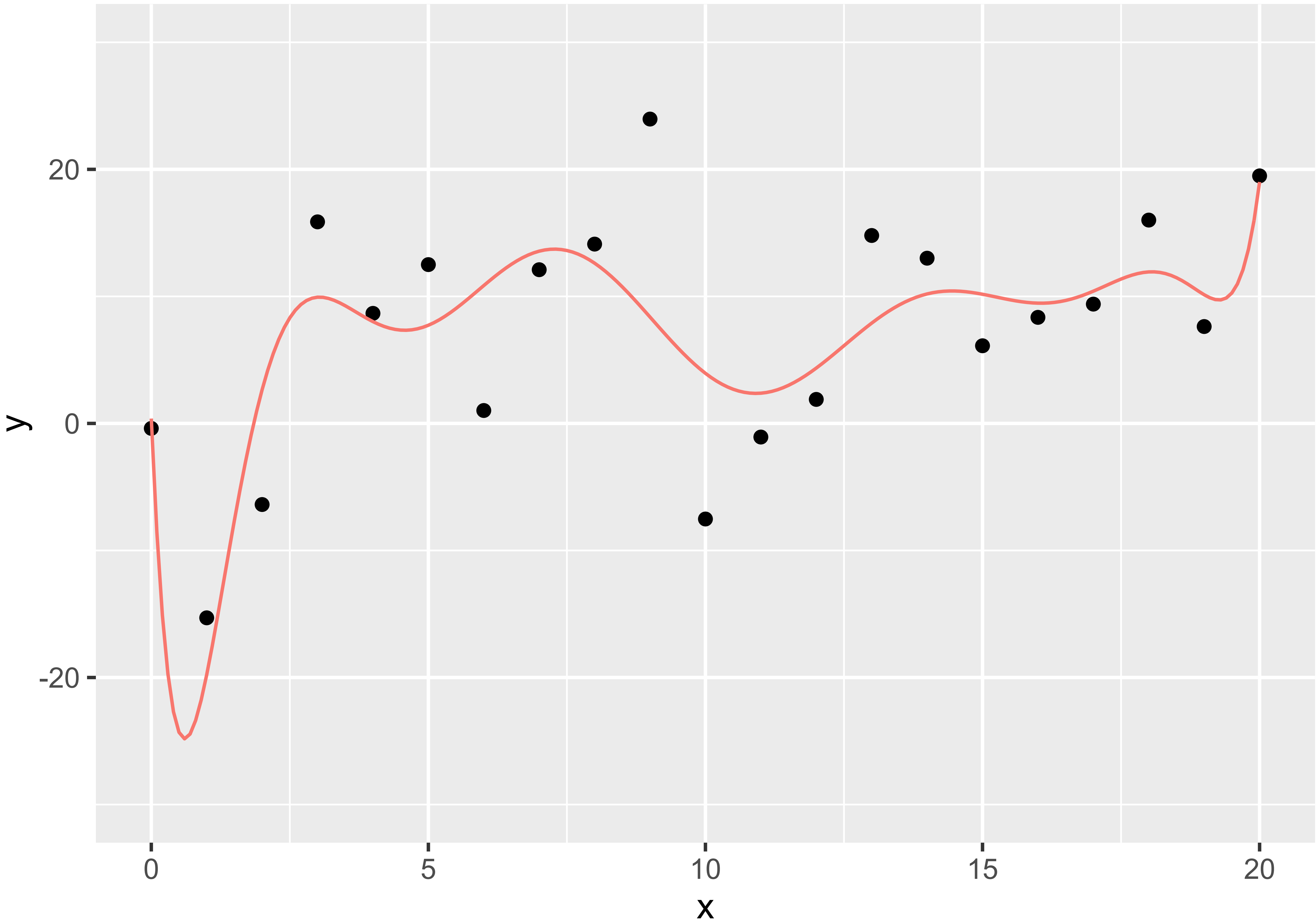
degree 7, training MSE = 49.6



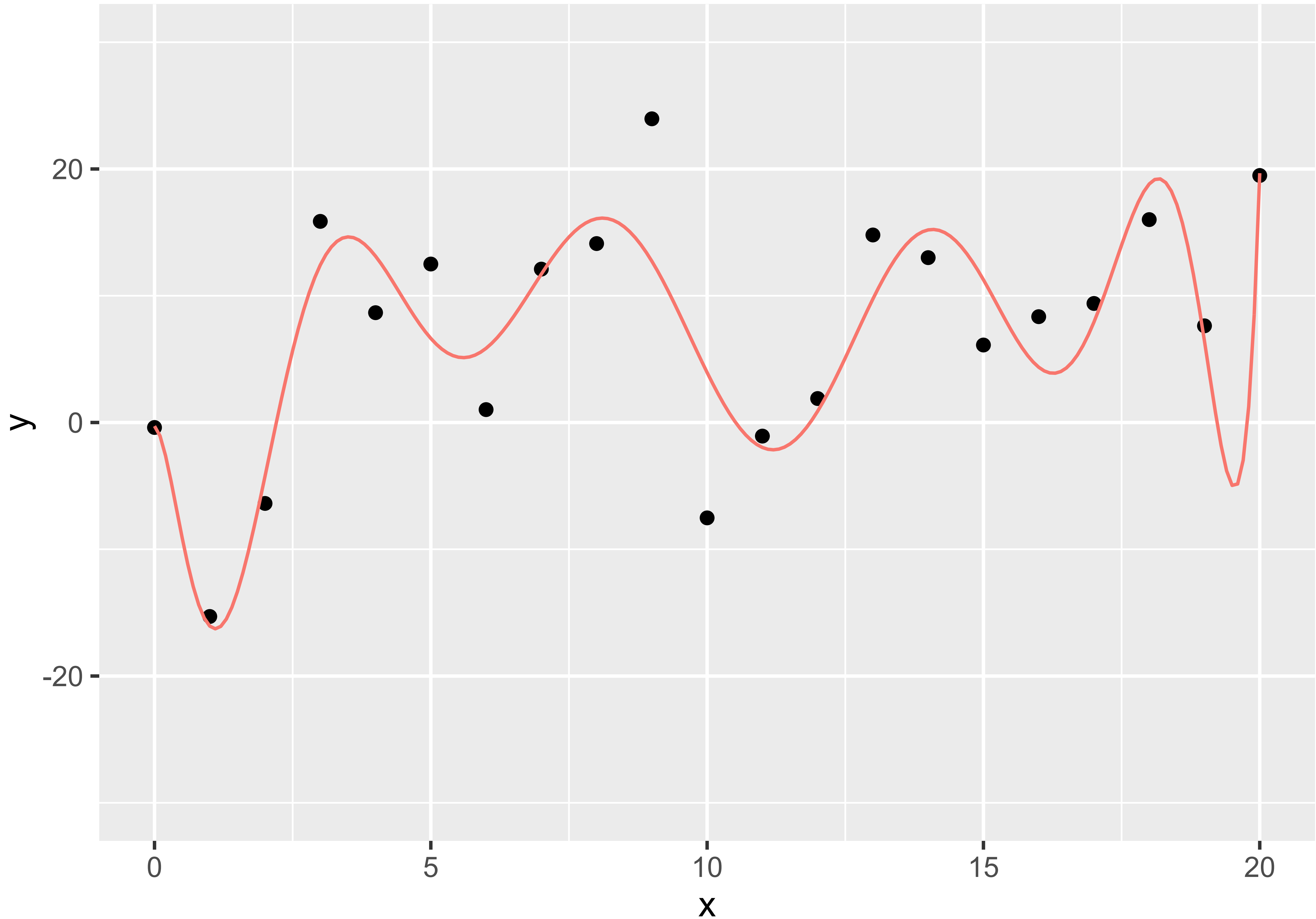
degree 8, training MSE = 46.8



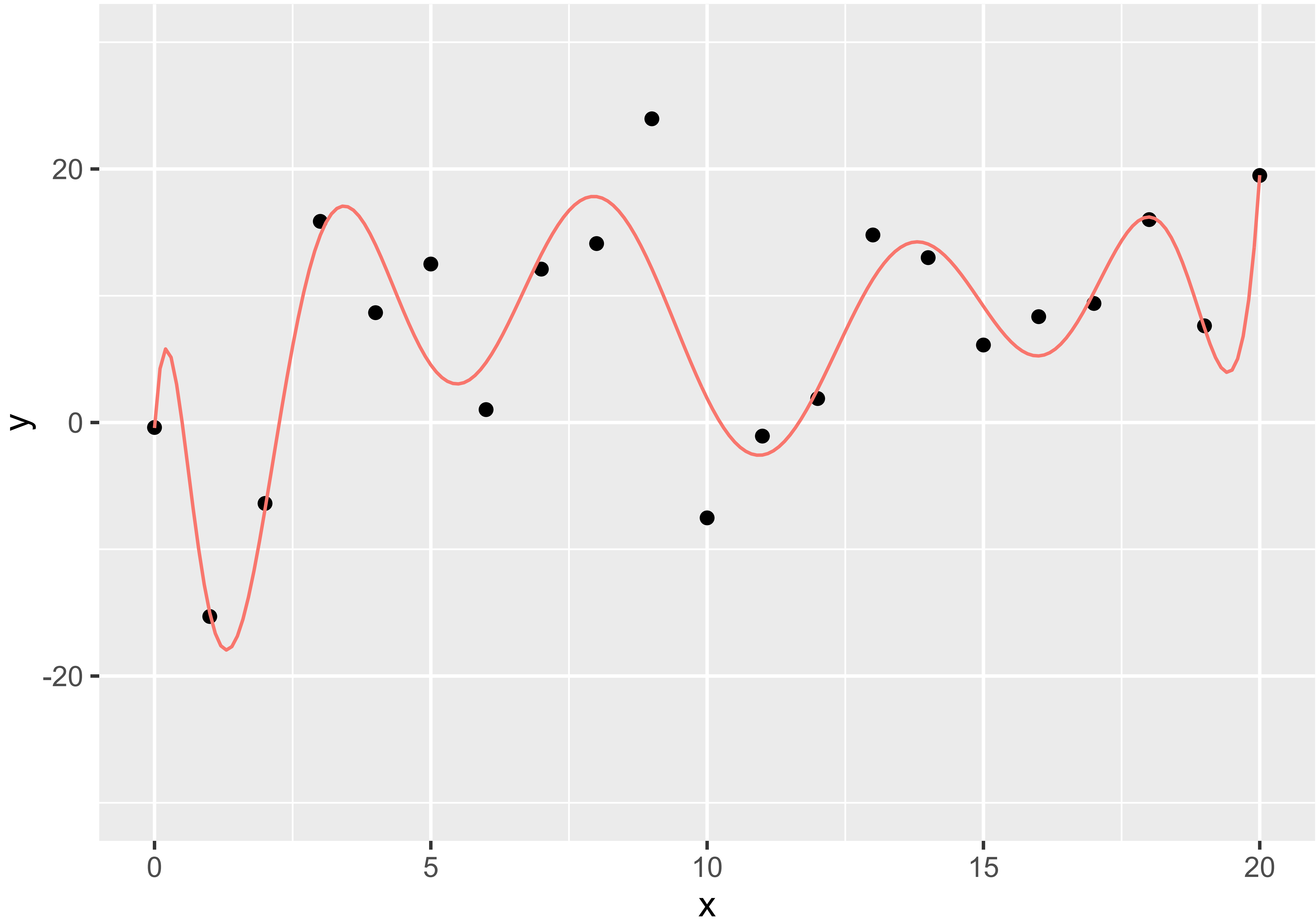
degree 9, training MSE = 41.2



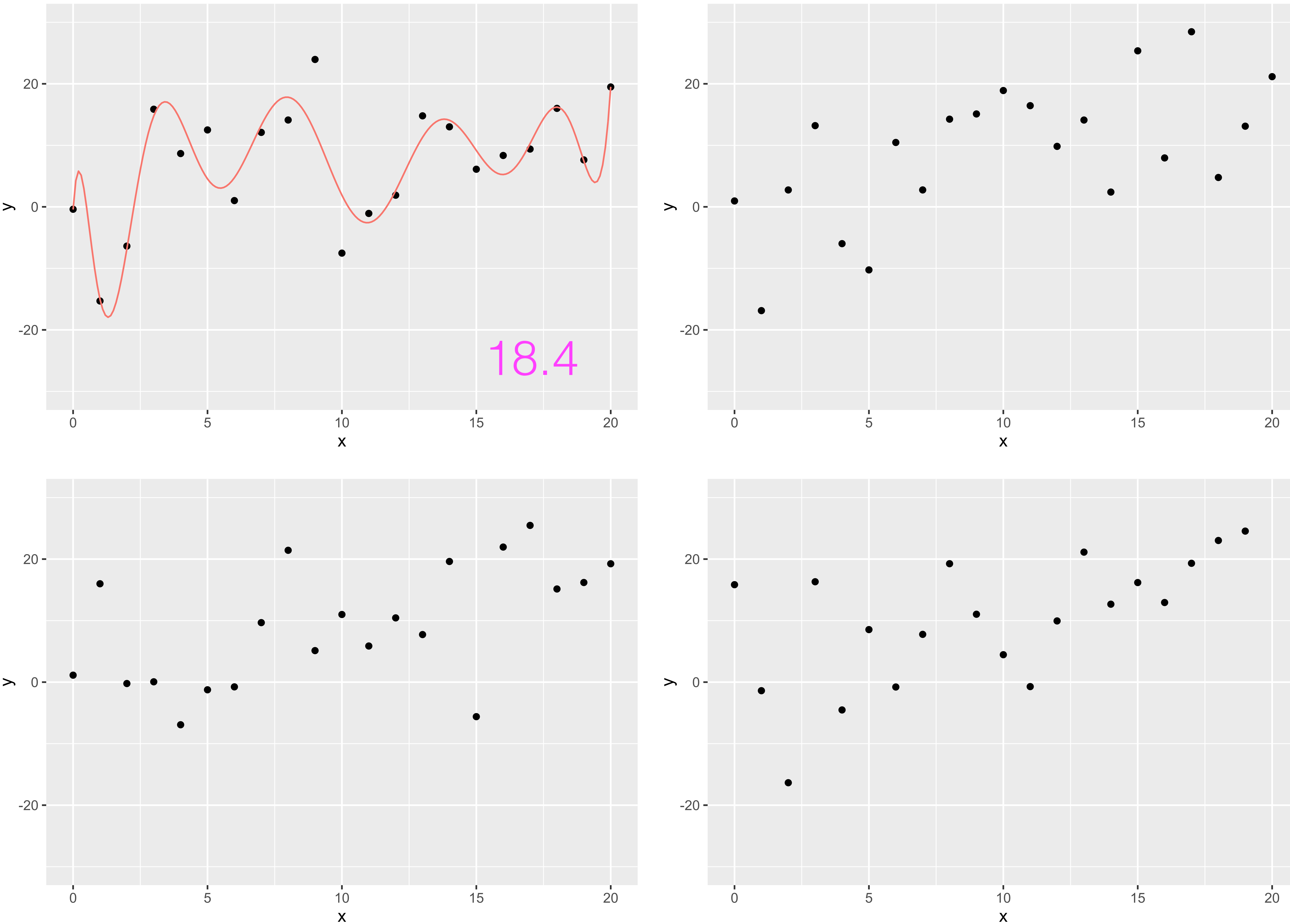
degree 10, training MSE = 35.8



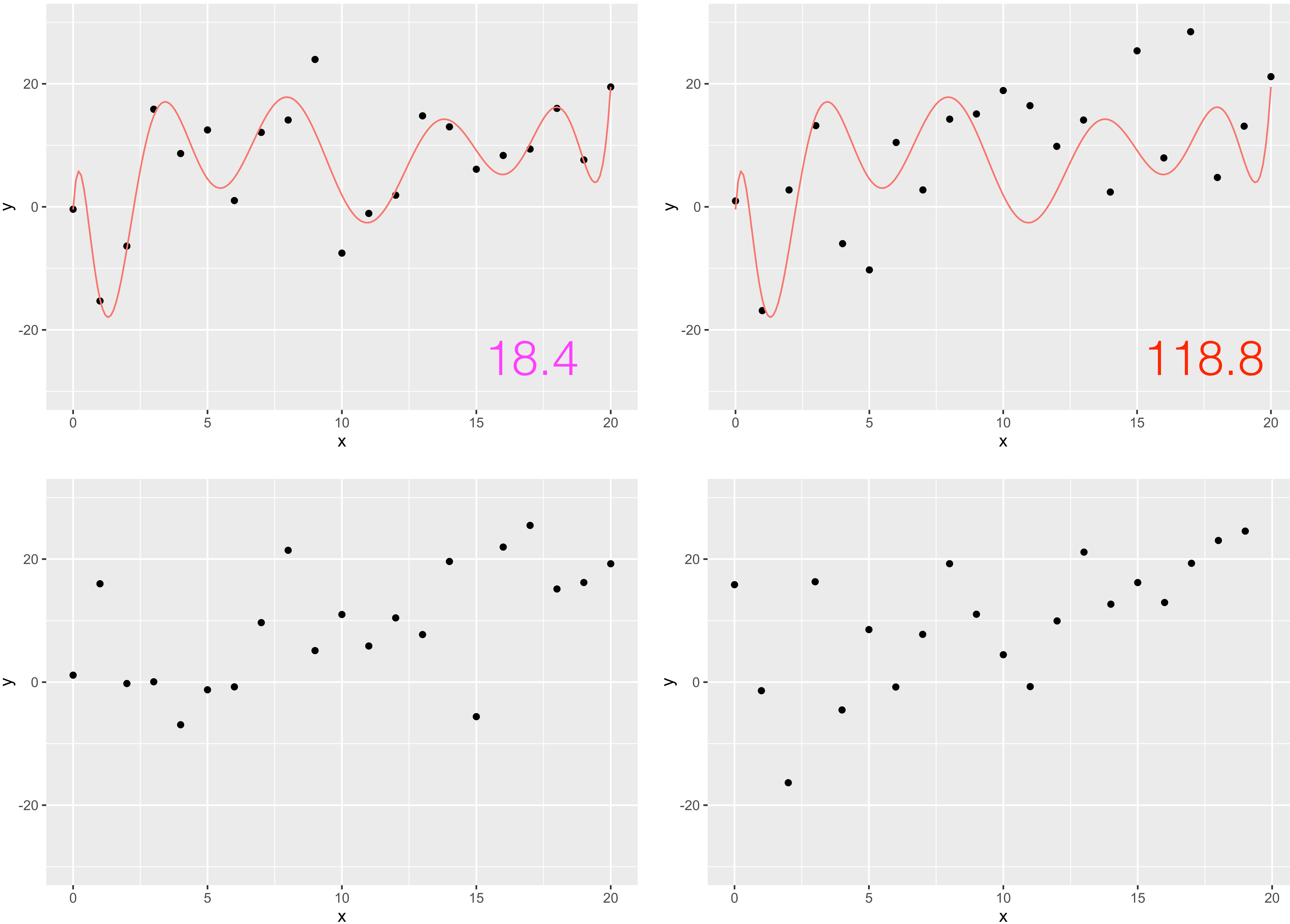
degree 11, training MSE = 21.1

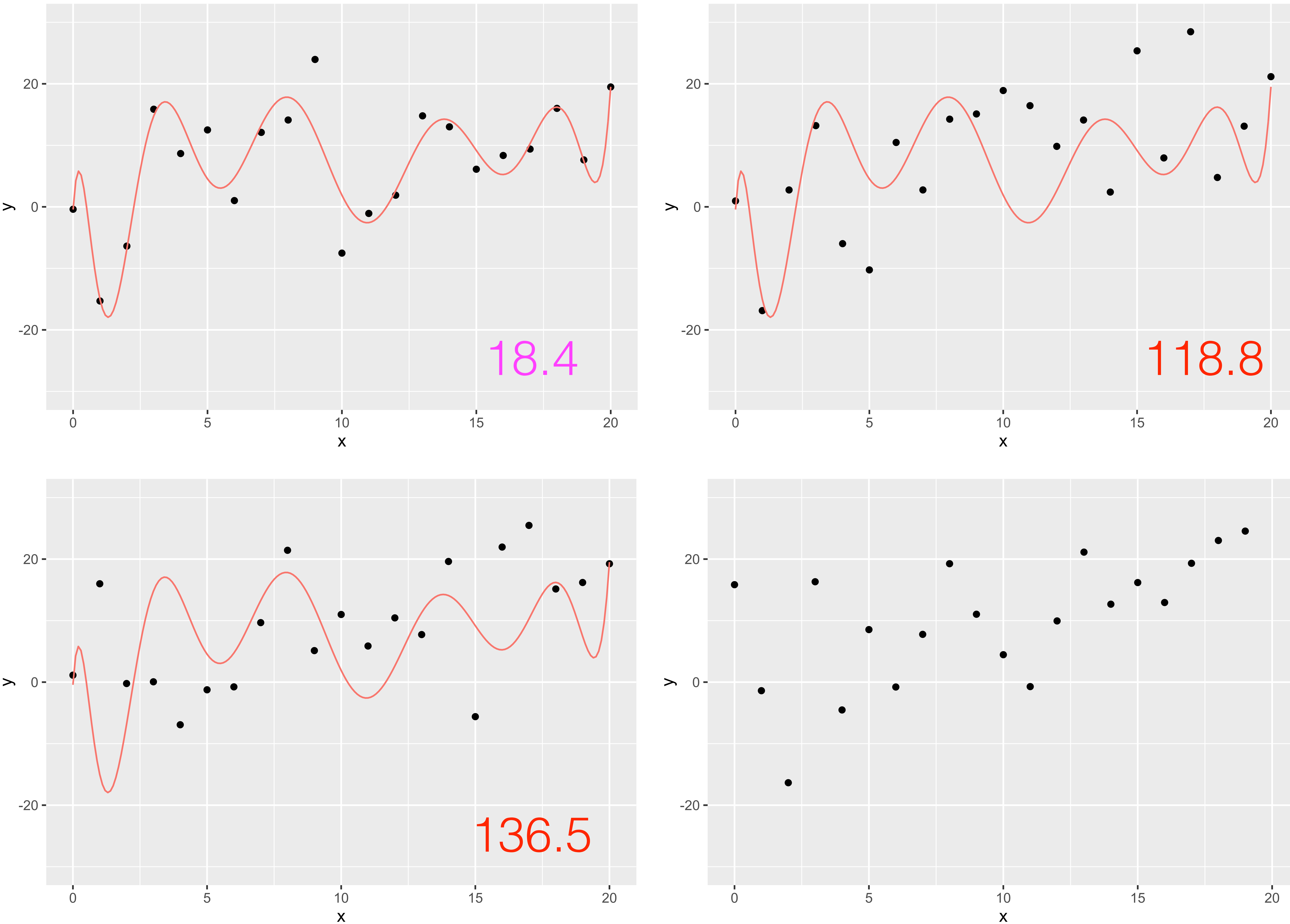


degree 12, training MSE = 18.4

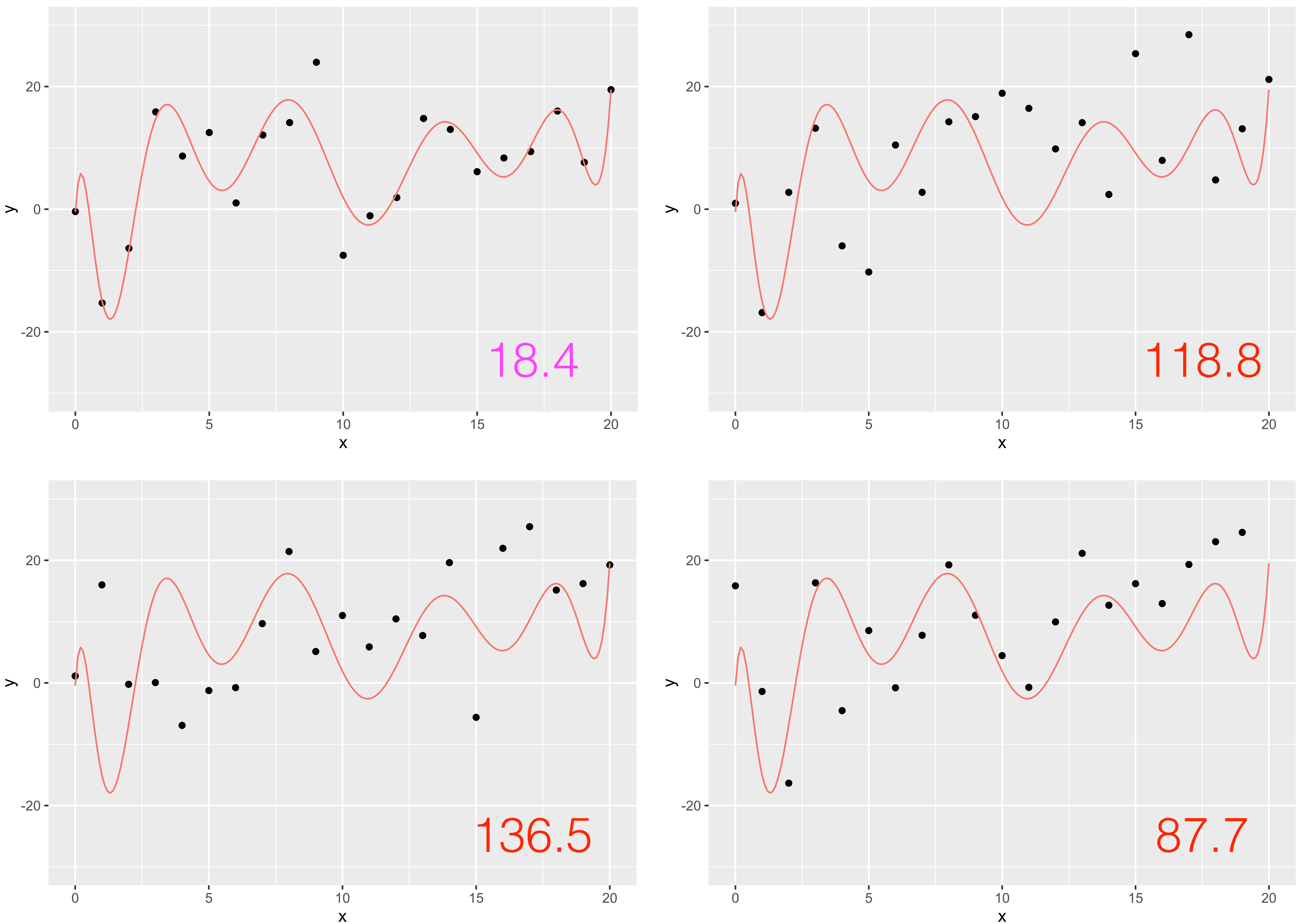


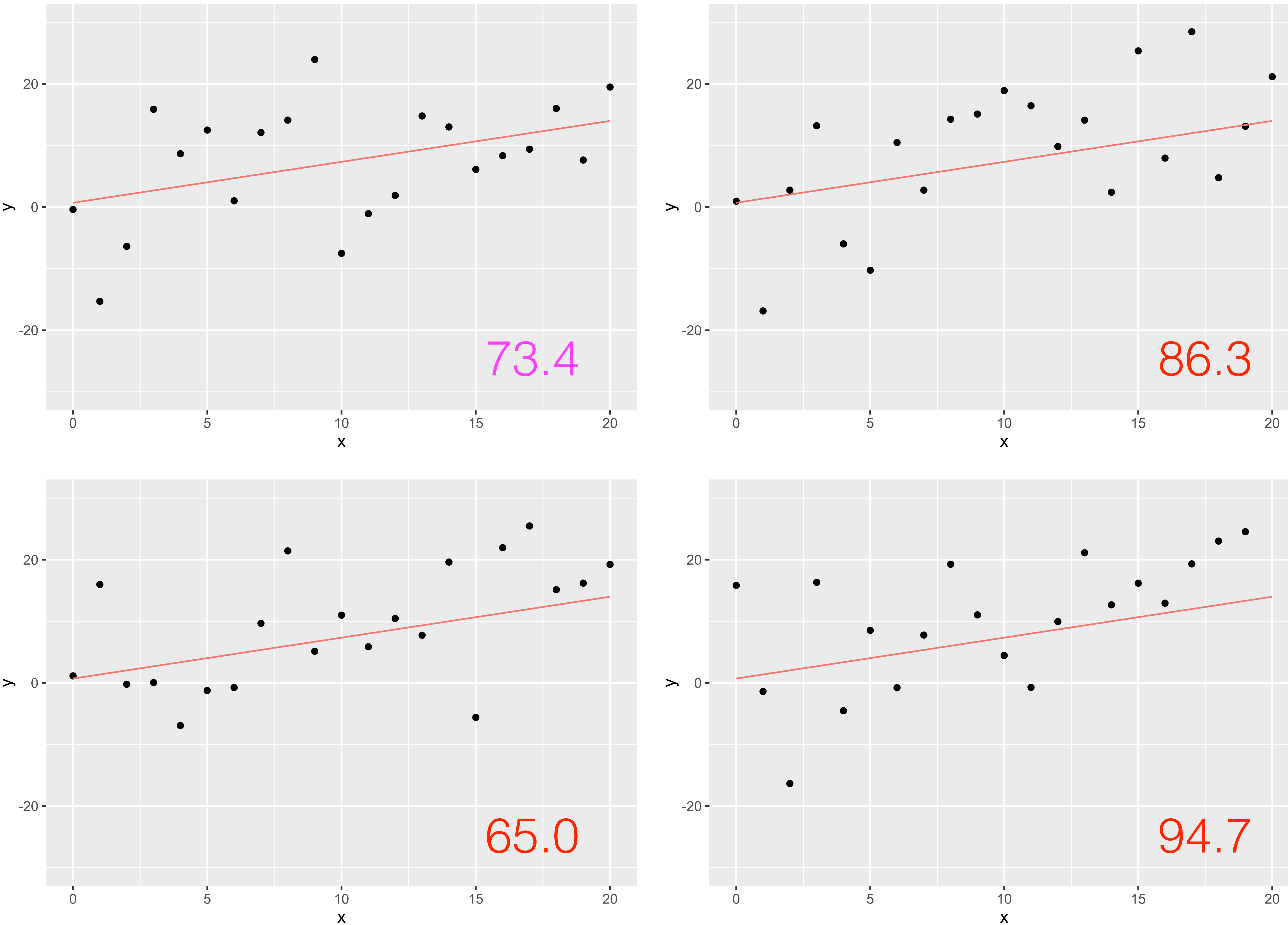
18.4
y
x
20
0
-20
0
5
10
15
20


18.4
118.8
y
x
20
0
-20
0
5
10
15
20



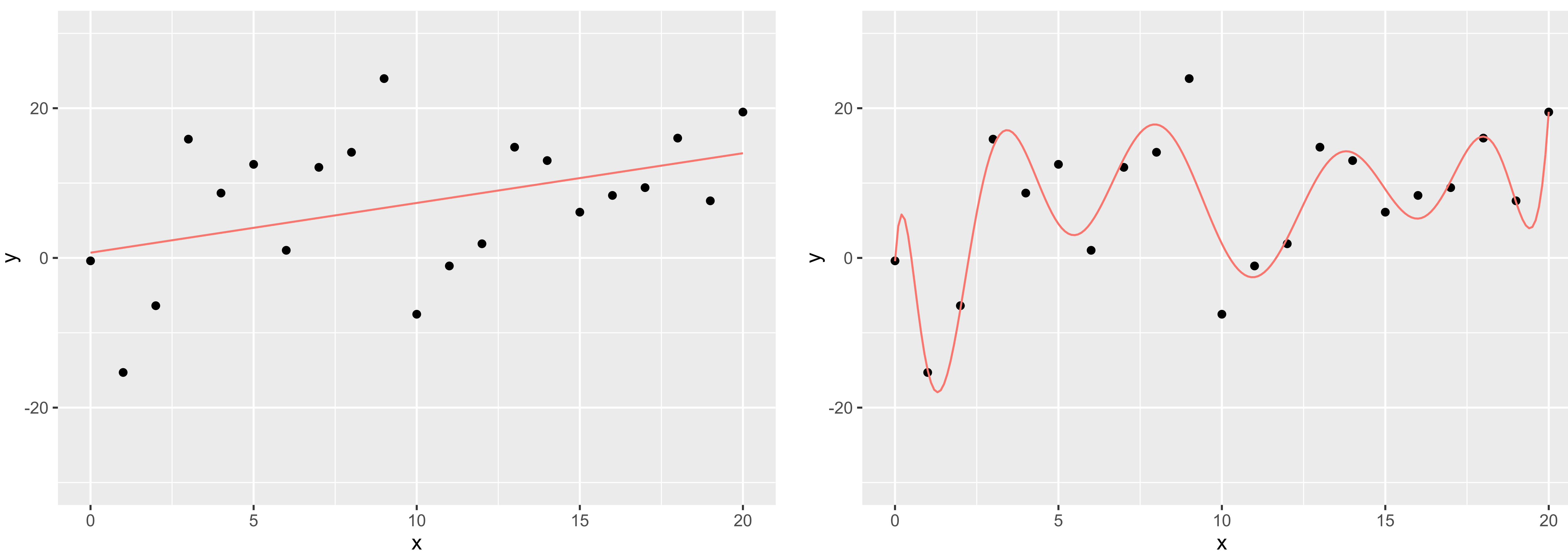
18.4
118.8
136.5
y
x
20
0
-20
0
5
10
15
20


18.4
118.8
136.5
87.7
x
y


73.4
86.3
65.0
94.7
y
x
20
0
-20
0
5
10
15
20

Overfitting

- Memorizing the nuances (and noise) of the training data that prevents generalizing to unseen data



Sources of error

- Bias: Error due to mis-specifying the relationship between input and the output.
  [too few parameters, or the wrong kinds]

- Variance: Error due to sensitivity to random fluctuations in the training data. If you train on different data, do you get radically different predictions?
  [too many parameters]

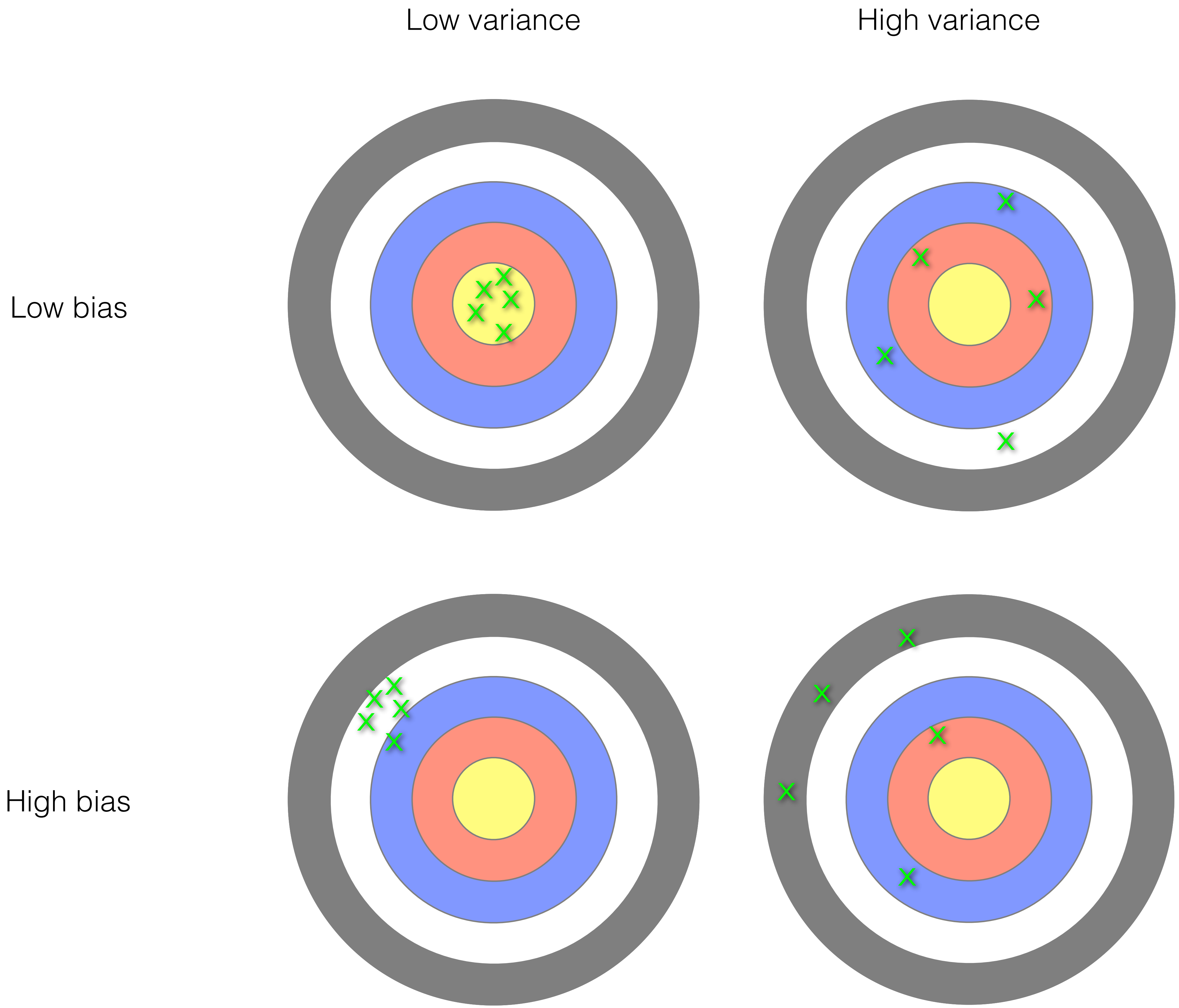
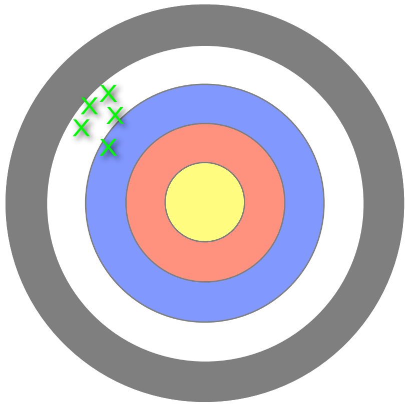
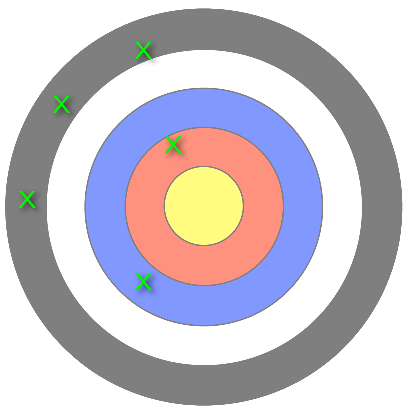


Image from Flach 2012

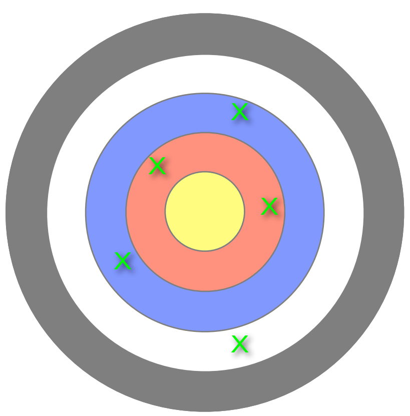
Example: geolocation on Twitter



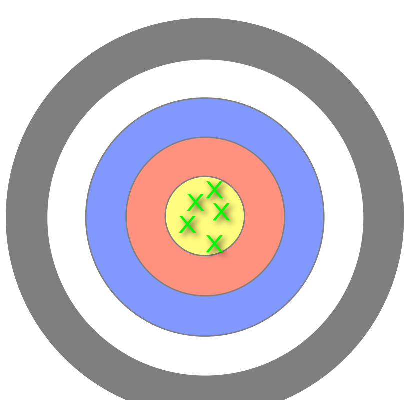
High bias, low variance: Always predict “Berkeley”



High bias, high variance: Predict most frequent city in training data



Low bias, high variance: many features, some of which capture true signal but capture random noise



Low bias, low variance: enough features to capture the true signal

Ordinal regression

- In between classification and regression
  - $\mathcal{Y}$ is categorical (e.g.,☆, ☆☆, ☆☆☆)
  - Elements of $\mathcal{Y}$ *are ordered*
    - ☆ < ☆☆
    - ☆☆ < ☆☆☆
    - ☆ < ☆☆☆

Ordinal regression

task	x	y
predicting star ratings	movie	$\{☆, ☆☆, ☆☆☆\}$

Computational Journalism

- Sarah Cohen, James T. Hamilton, and Fred Turner, “Computational Journalism,” *Communications of the ACM* (2011)
- Sylvain Parasie, “Data-Driven Revelation? Epistemological tensions in investigative journalism in the age of ‘big data,’” *Digital Journalism* (2015)

Computational Journalism

- “Changing how stories are discovered, presented, aggregated, monetized and archived” (Cohen et al. 2012)
- Draws on earlier tradition of computer-assisted reporting and “precision journalism” (Meyer 1972)

Computational Journalism

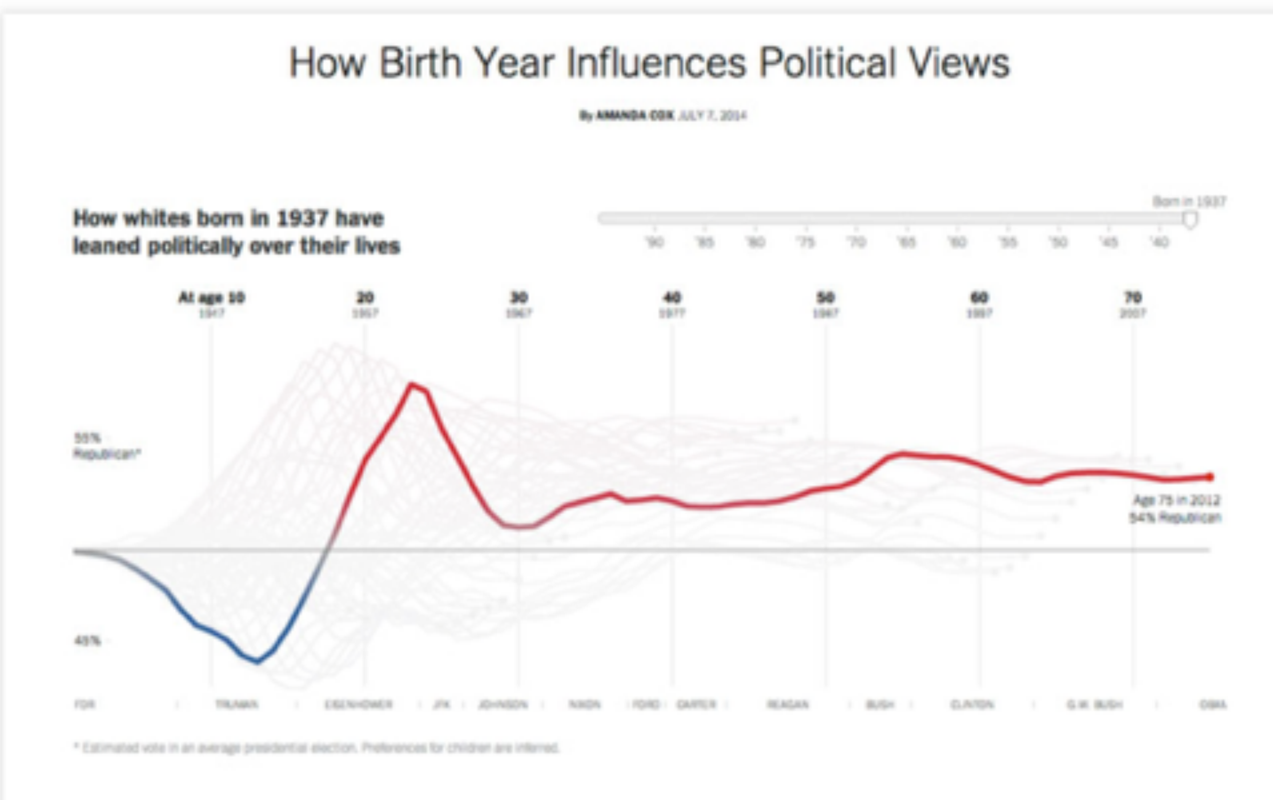
- Database linking, e.g.:
    - voting records to the deceased
    - press releases from different members of congress
    - indictments/settlements from U.S. attorneys
    - documents from SEC, Pentagon, defense contractors to note movement to industry (Cohen 2012)
    - DSA database of safety status of CA public schools + US seismic zones + school list from CA Dept of (Parasie 2015)

Computational Journalism

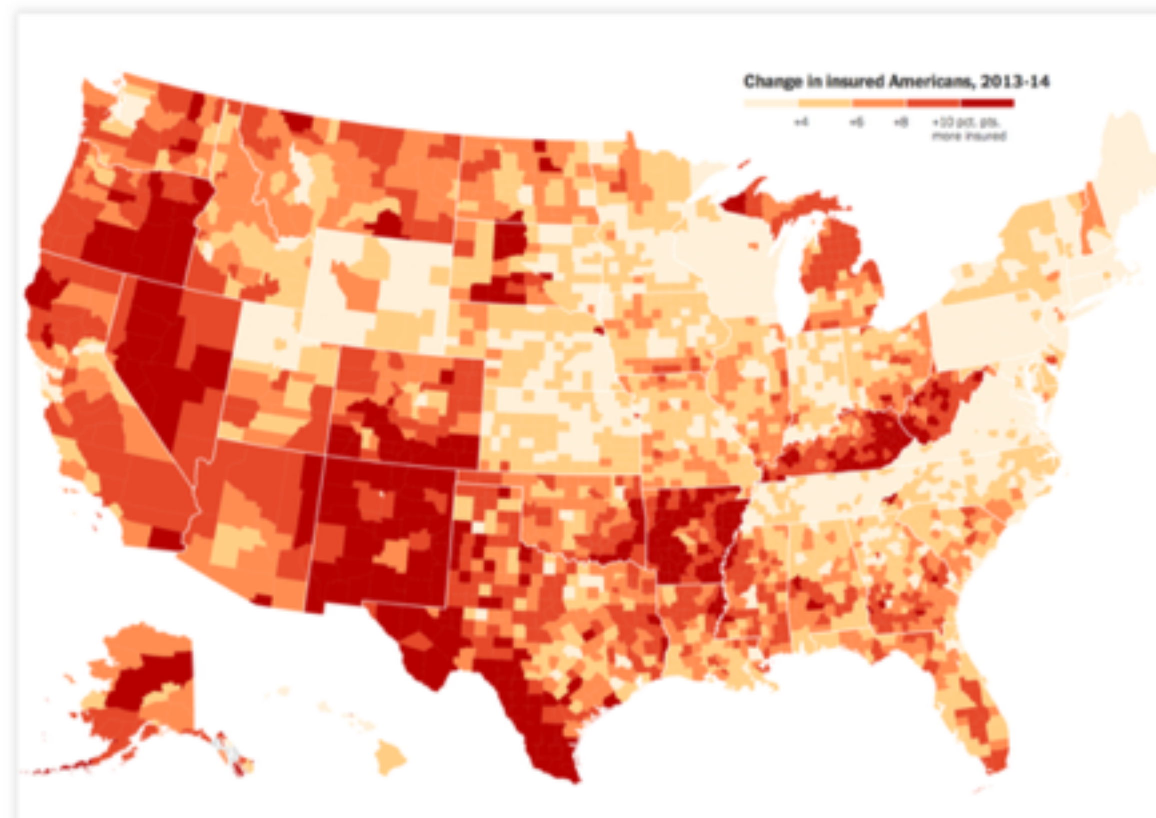
- Information extraction: need to pull out people, places, organizations and their relationship from large (often sudden) dumps of documents.
- Analyzing the relationship between entities

Computational Journalism

- Data-driven stories about large-scale trends



Relationship between birth year and political views
NY Times (July 7, 2014)



Change in insured Americans under the ACA,
NY Times (Oct 29, 2014)

Computational Journalism

- Data-driven lead generation; the outliers in analysis that point to a story

Computational Journalism

- Demands:
  - High precision
  - Fast turnaround
- Needs (Stray 2016):
  - Accurate document analysis
  - Guided search
  - Interactive methods

Project proposal, due 2/16

- Collaborative project (involving up to 3 students), where the methods learned in class will be used to draw inferences about the world and critically assess the quality of those results.

- Proposal (2 pages):

    - outline the work you're going to undertake
    - formulate a hypothesis to be examined
    - motivate its rationale as an interesting question worth asking
    - assess its potential to contribute new knowledge by situating it within related literature in the scientific community. (cite 5 relevant sources)
    - who is the team and what are each of your responsibilities (everyone gets the same grade)