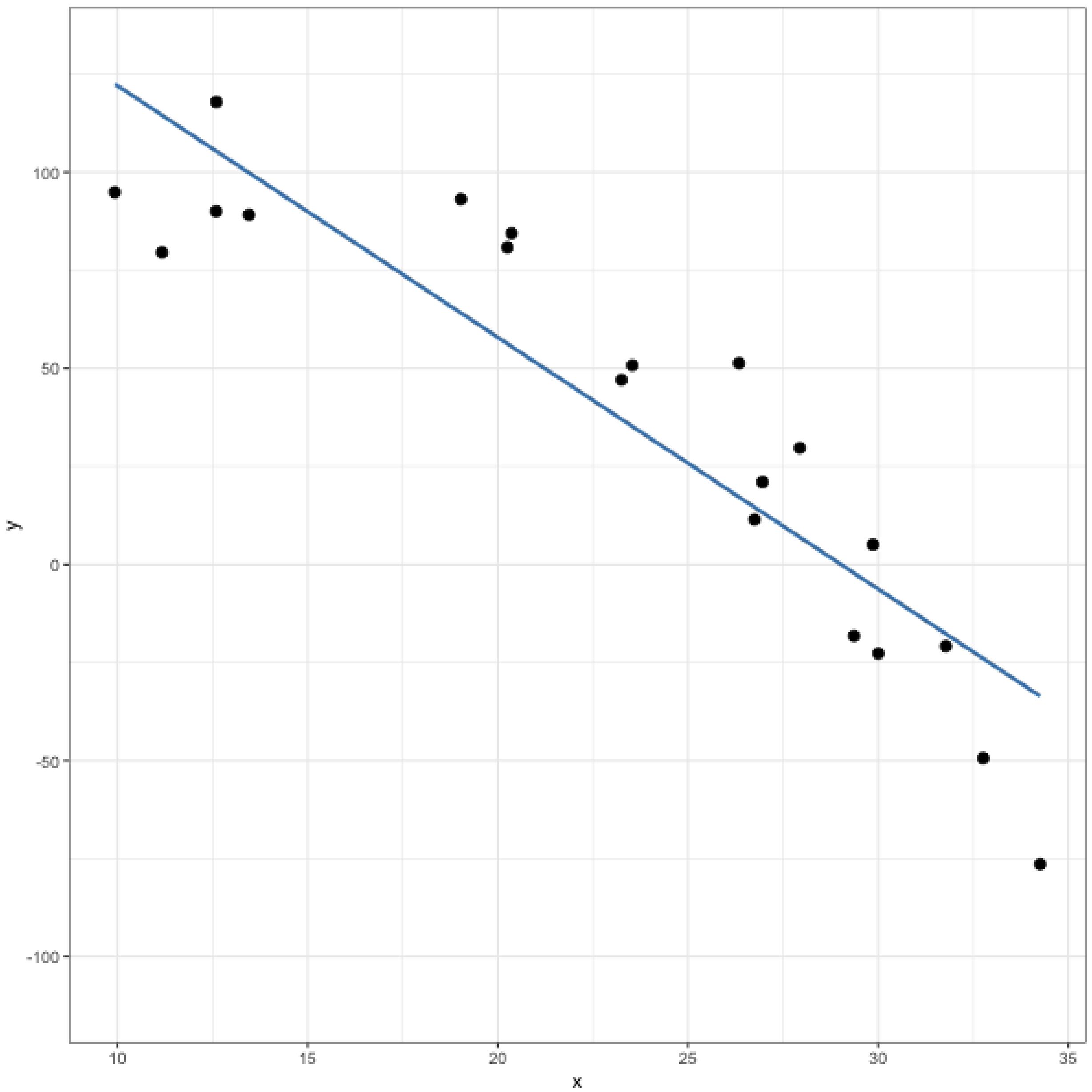

100
50
0
-50
-100
y
10
15
20
25
30
35
x

Generating from Generative Models

Many statistical models are **generative**: they represent probability distributions from which we're able to take random samples. That is, once you have the model (true or estimated) you are able to generate as much data as you like.

Simulation from true model (ex: polynomial regression)

True mean function:

$$\begin{aligned} f(x) &= \beta_0 + \beta_1 x + \beta_2 x^2 \\ &= 23 + 4x - 3.2x^2 \end{aligned}$$

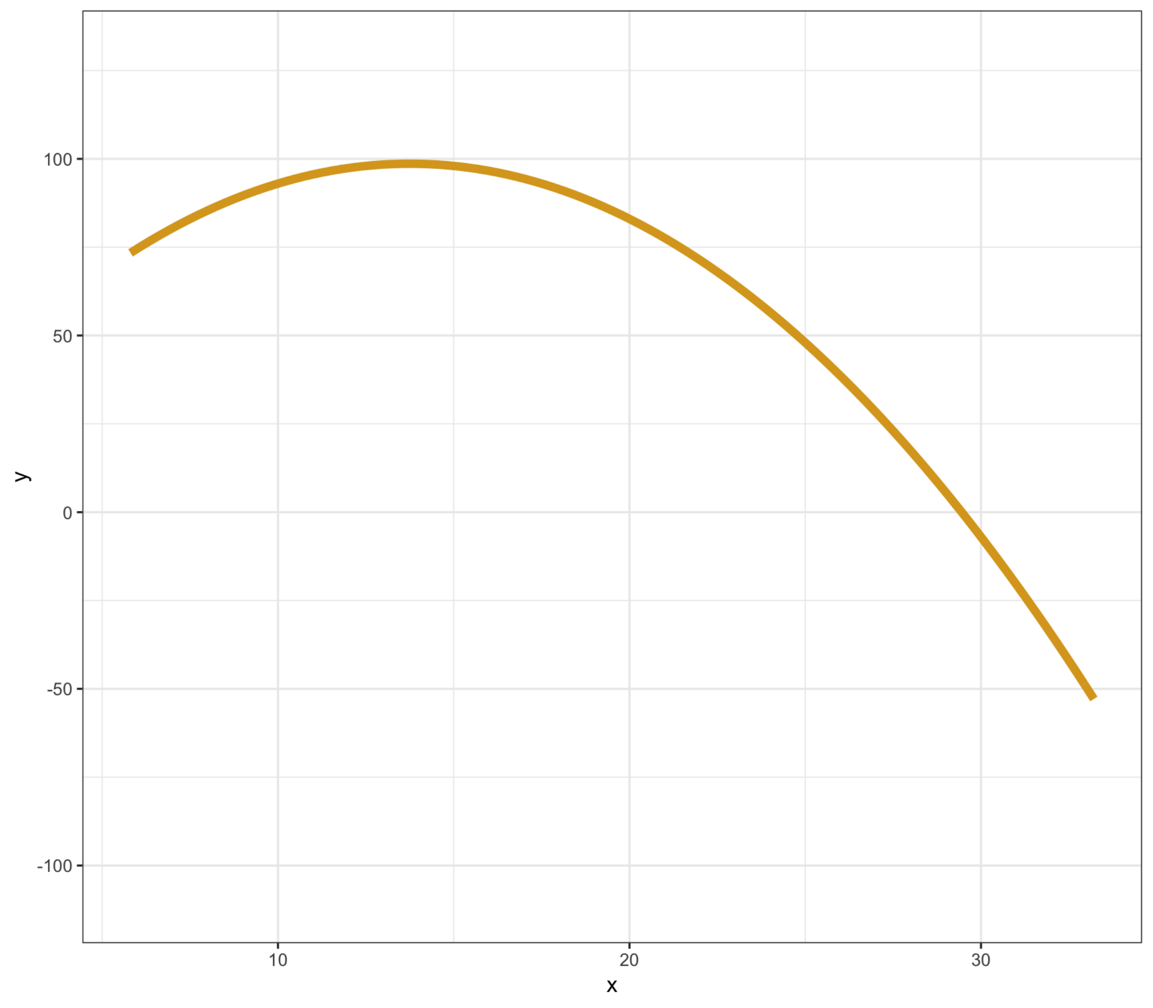
True data generating function:

$$y = f(x) + \epsilon; \quad \epsilon \sim N(0, 11)$$

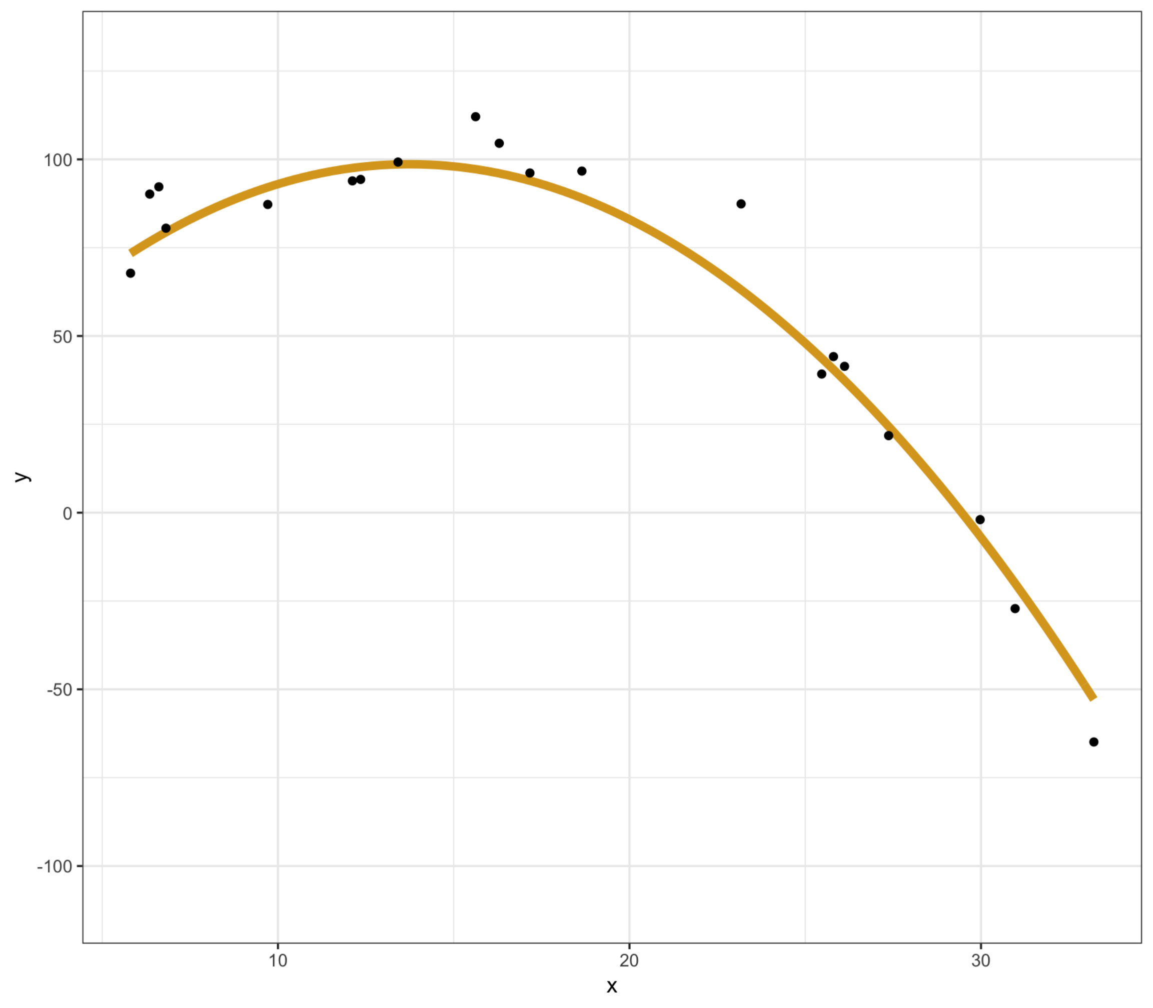
$$Y|X \sim N(\mu = f(x),\, \sigma = 11)$$

```r
n <- 20
x <- runif(n, min = 5, max = 35)
f <- function(x) {
  23 + 11 * x - .4 * x^2
}
y <- rnorm(length(x), mean = f(x), sd = 11)
```

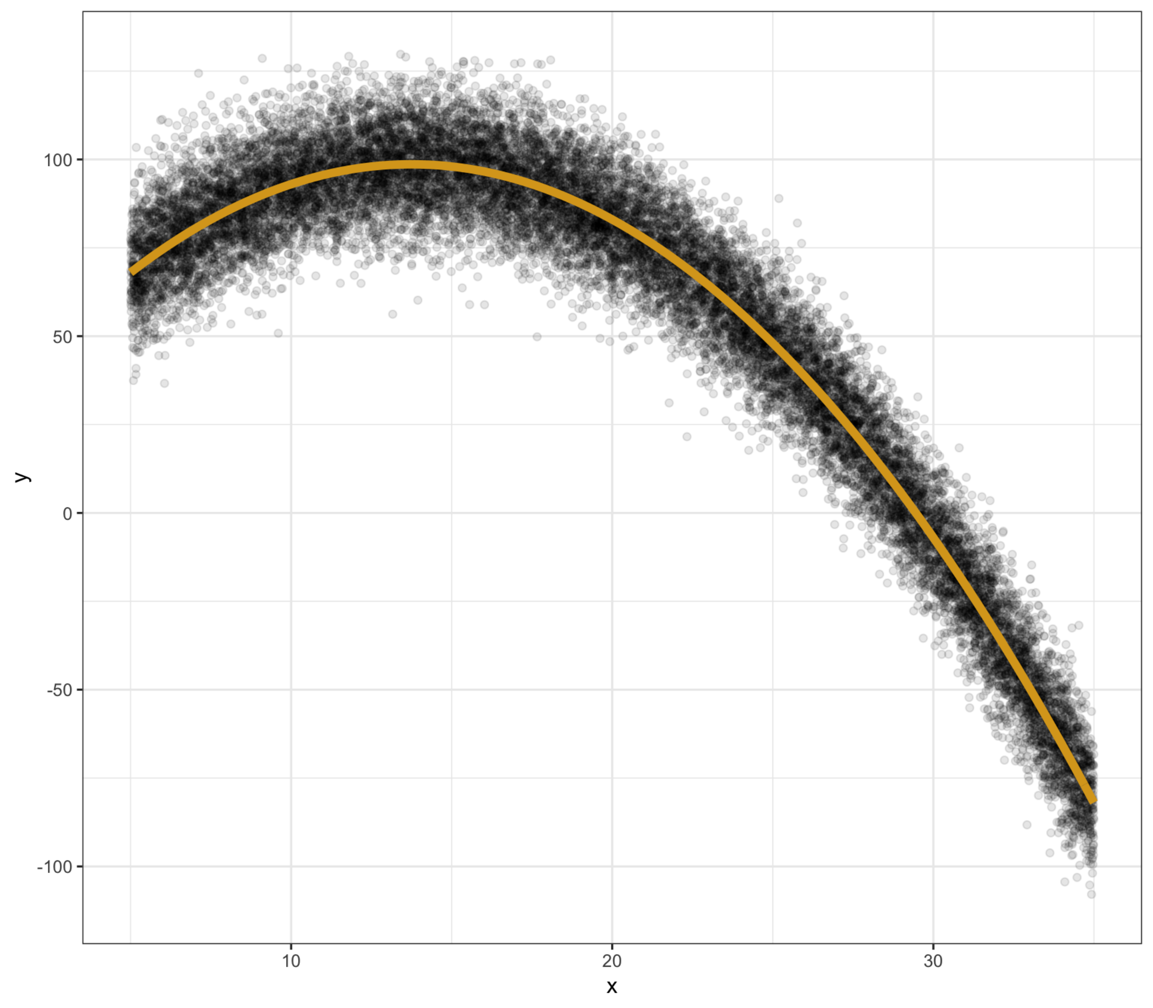
$$f(x) = \beta_0 + \beta_1 x + \beta_2 x^2$$



$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon;\ n = 20$$



$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon; \; n = 20000$$



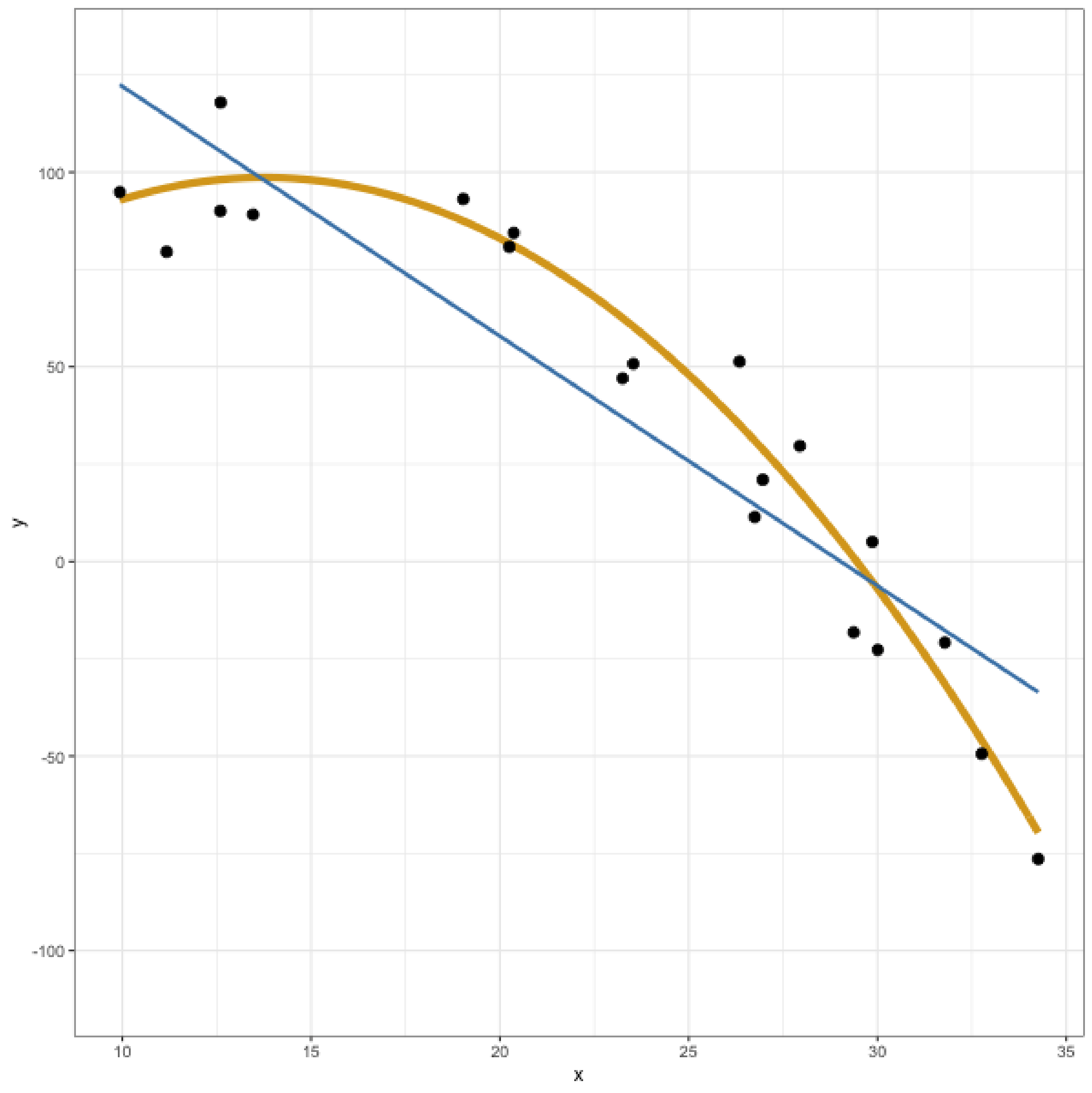
Visualizing Bias and Variance

Procedure

1. Assume true generative model
2. Generate data set of size n
3. Estimate $\hat{f}$
4. Repeat 2 and 3 many times to get a sense of the variation in $\hat{f}$

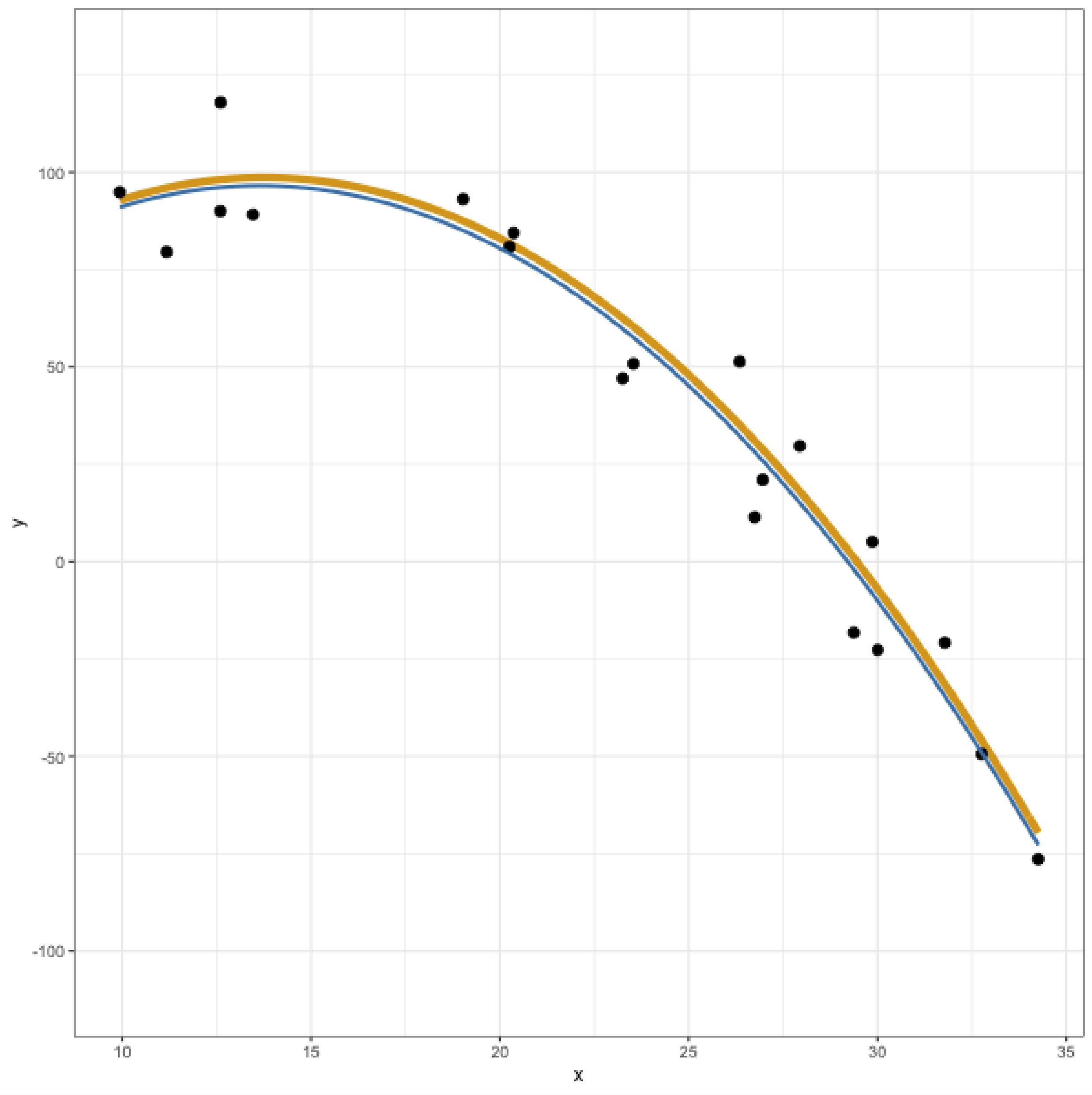
Estimating $\hat{f}$

Let's naively assume a *linear form*, work with data sets of size 20, and fit $\hat{f}$ by least squares.


100
50
0
-50
-100
y
10
15
20
25
30
35
x

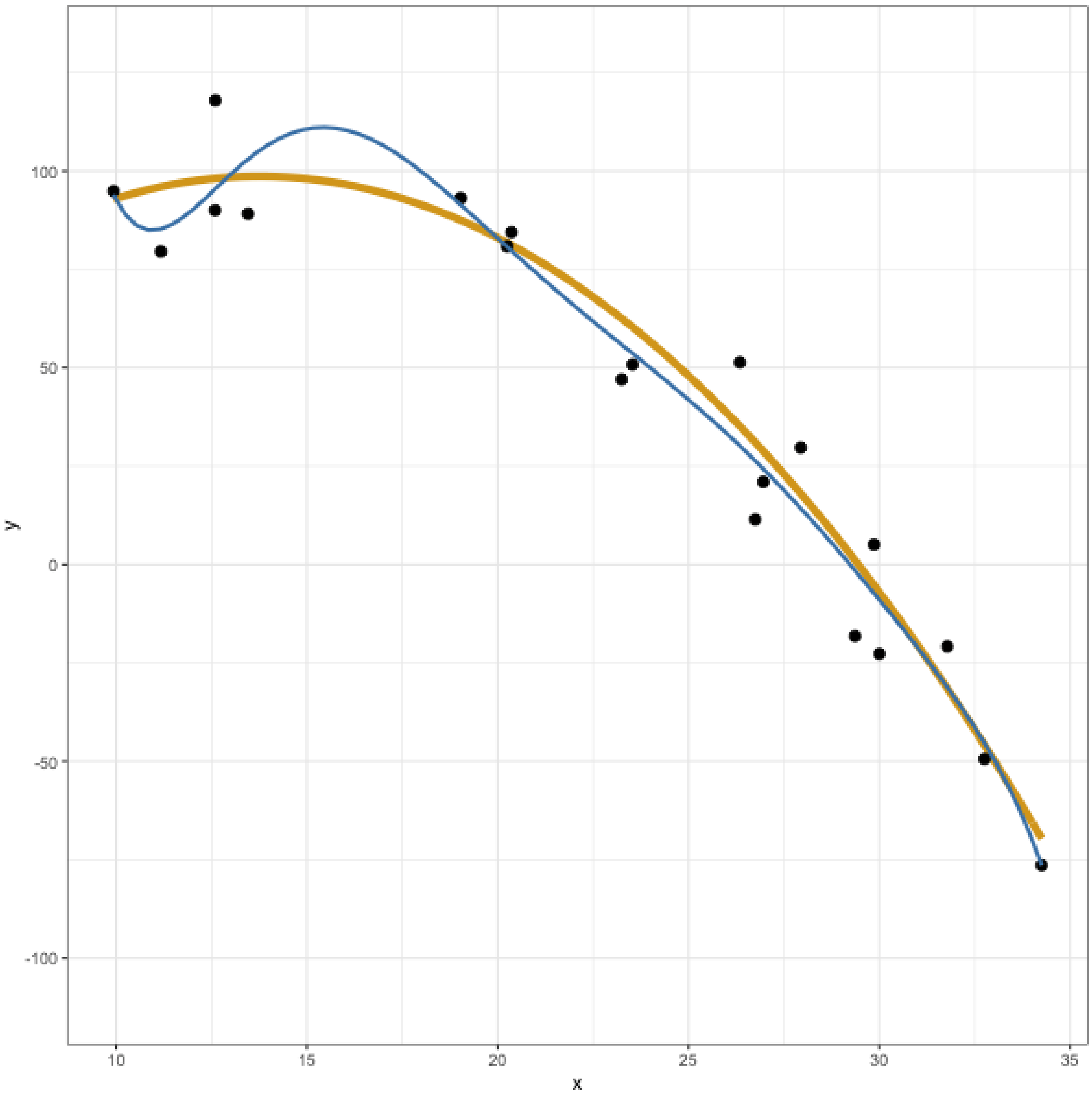
Estimating $\hat{f}$, take two

Next, let's presciently assume a quadratic form...


100
50
0
-50
-100
y
10
15
20
25
30
35
x

Estimating $\hat{f}$, take three (or seven?)

Finally, let's get ridiculous and assume a septic form...


100
50
0
-50
-100
y
10
15
20
25
30
35
x