Residuals, outliers, and influence

Author: Nicholas G Reich

Today's topics

- Model terms recap
- Model values and residuals
- Smooth terms

Models are functions (recap)

We might write generally

$$y = f(x)$$

where x could be a single variable or multiple variables.

- **The response variable** is y the variable whose behavior or variation you are trying to understand.
- **The explanatory variables** x are the variable(s) that you want to use to explain the variation in the response variable.

Statistical models are functions with randomness

- The equation

$$y = f(x)$$

  describes a **deterministic** function. There is no randomness.
- A statistical model uses a deterministic kernel, e.g. $f(x)$, but layers in **stochastic**, or random, variation. Remember, our goal is to explain and understand the variation in y.
- A statistical model will always have an additional layer that accounts for the random variation. In linear models with continuous outcomes, this can be expressed as a **residual** term:

$$y = f(x) + \epsilon$$

Statistical models are functions with randomness

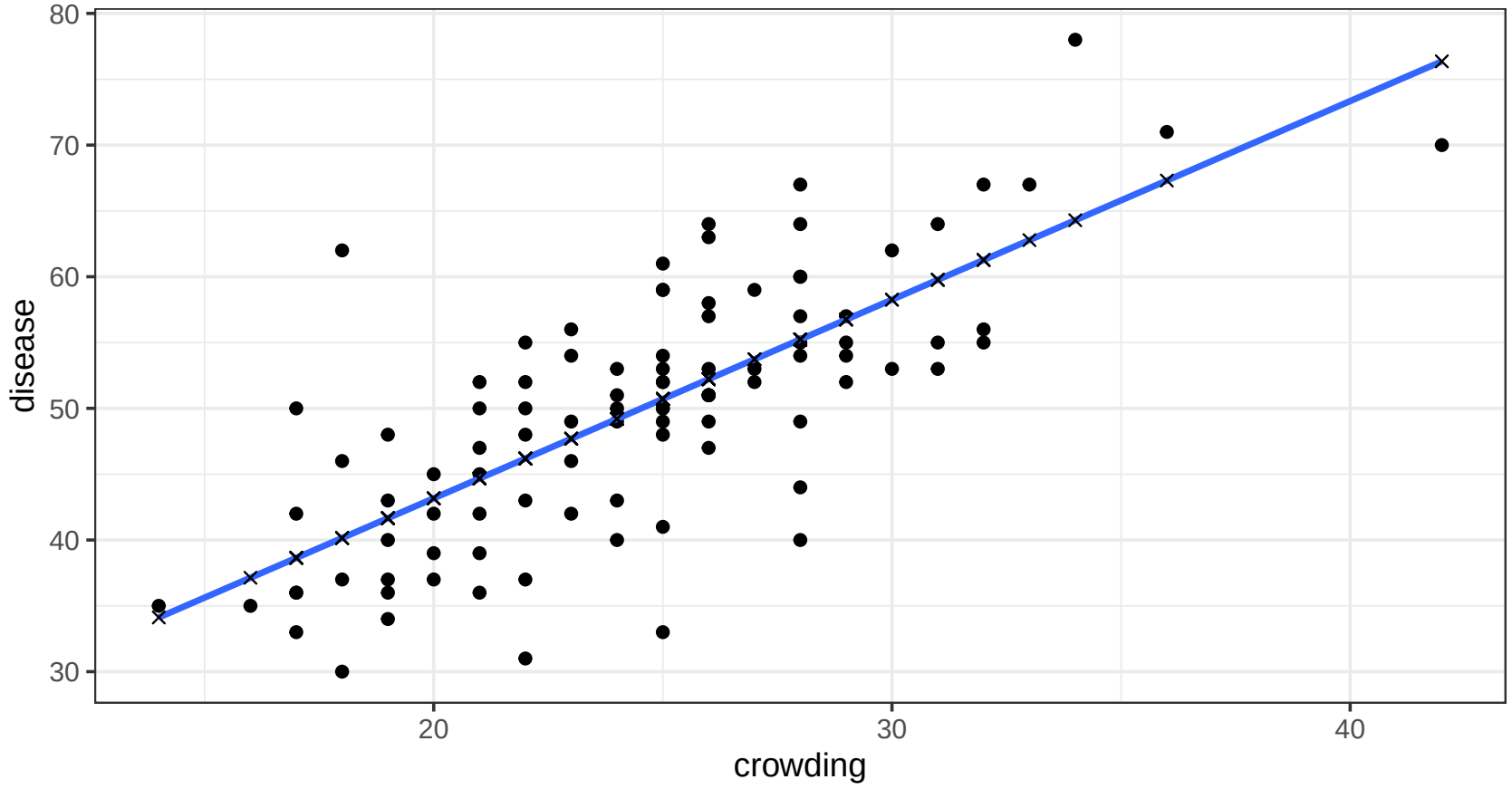
- Let's say we have N observations in a dataset. The observed values of the response variables are represented by $y_1, y_2, ..., y_N$.
- The observed values of the explanatory variables are $x_1, x_2, ..., x_N$.
- For a given observation with index i, the **model value** $\hat{y}_i$ is the output of the deterministic function part of your model:

$$\hat{y}_i = f(x_i)$$

Statistical models are functions with randomness

For a given observation with index i, the **model value** $\hat{y}_i$ is the output of the deterministic function part of your model:

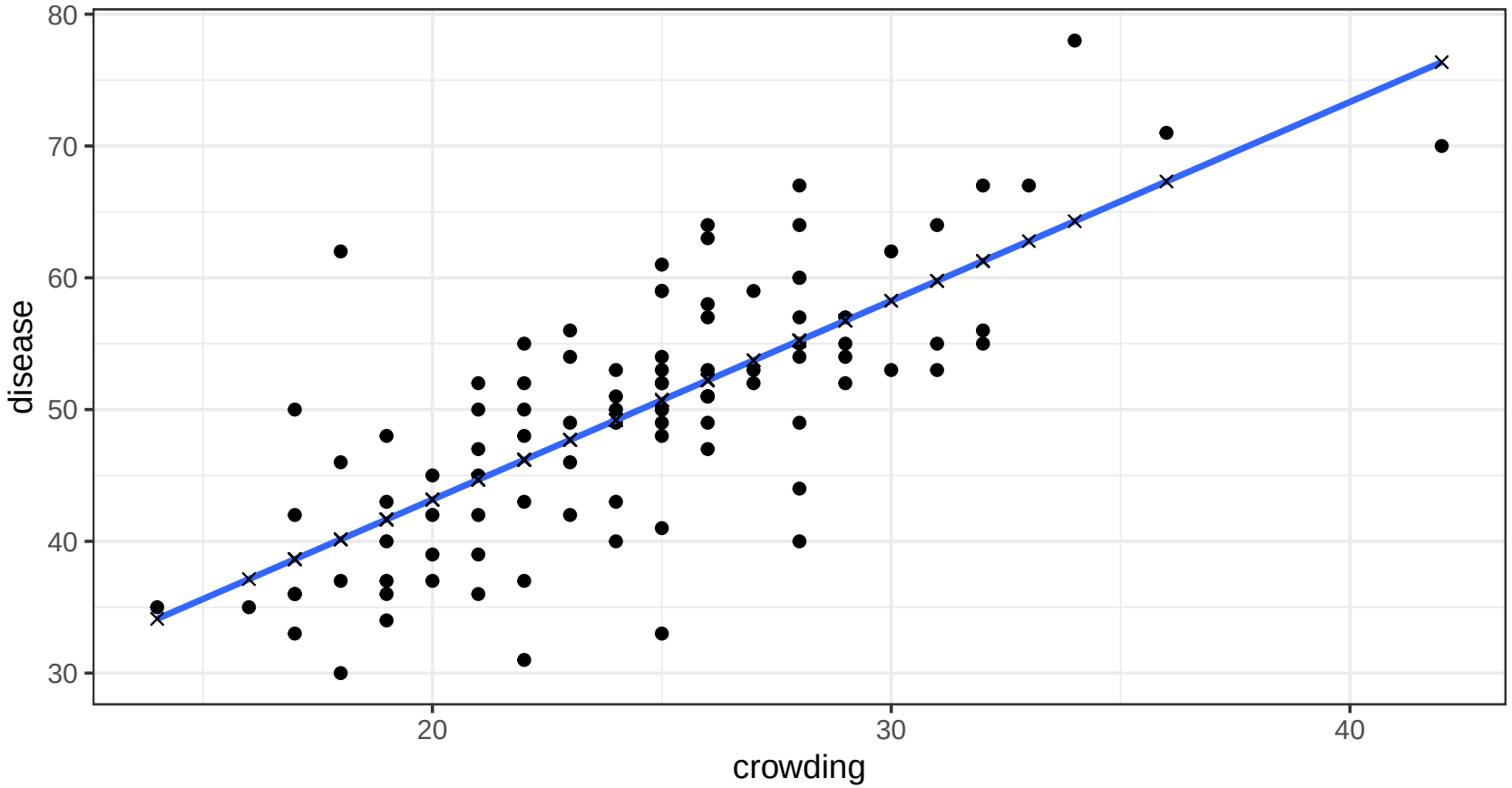
$$\hat{y}_i = f(x_i) = 13.0 + 1.5 * \text{crowding}_i$$



Statistical models are functions with randomness

For a given observation with index i, the original observation y_i can be reconstructed as the deterministic function part of your model plus the residual (ϵ_i), or the part of the variation that your model cannot explain:

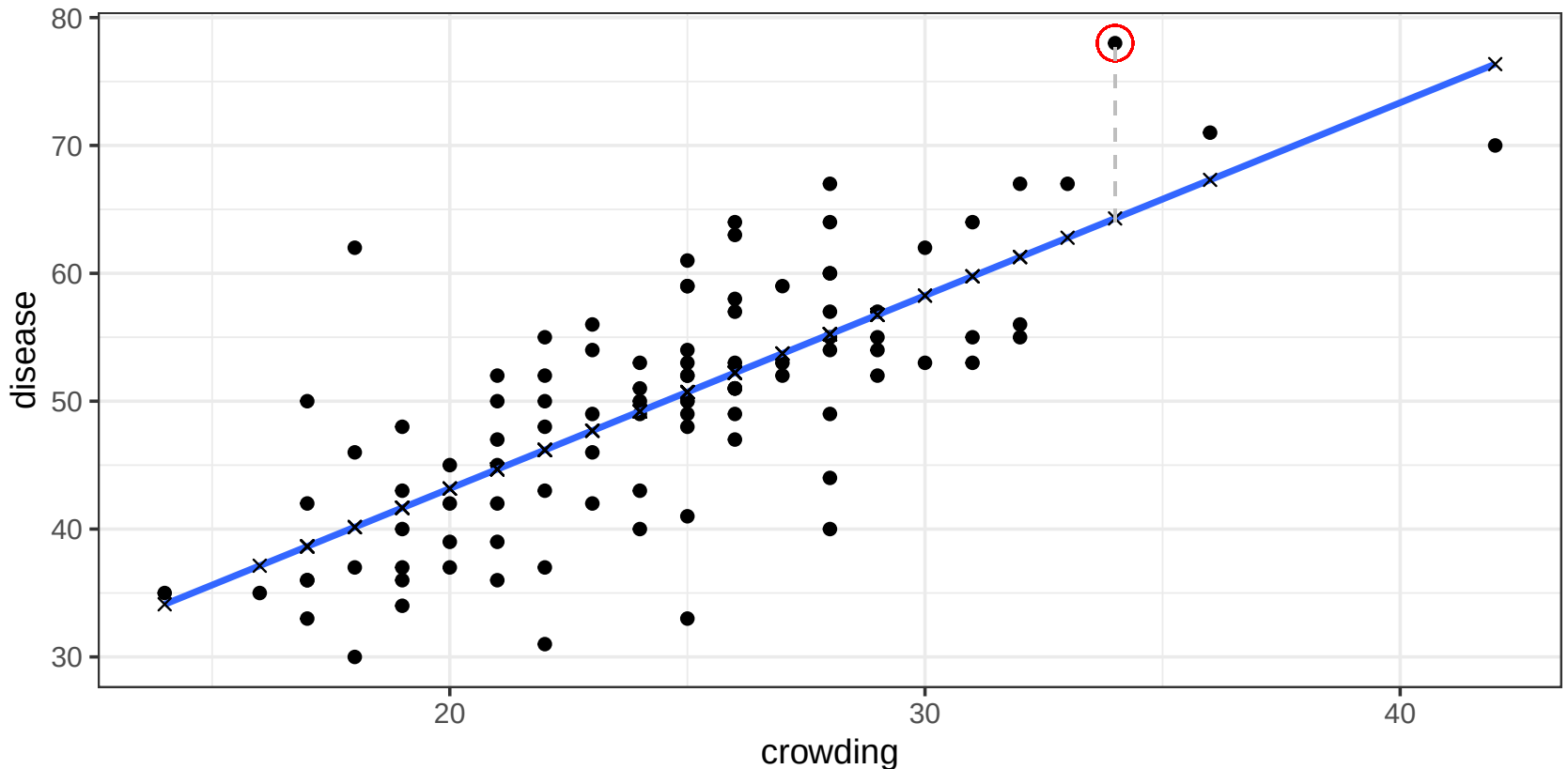
$$y_i = f(x_i) + \epsilon_i = 13.0 + 1.5 * \text{crowding}_i + \epsilon_i$$



Statistical models are functions with randomness

Let's look at one specific model value and residual. One observation in this dataset has disease $= 78$ and crowding $= 34$.

$$
\begin{aligned}
y_i &= f(x_i) + \epsilon_i \\
78 &= 13.0 + 1.5 * 34 + \epsilon_i \\
\implies 14 &= \epsilon_i
\end{aligned}
$$

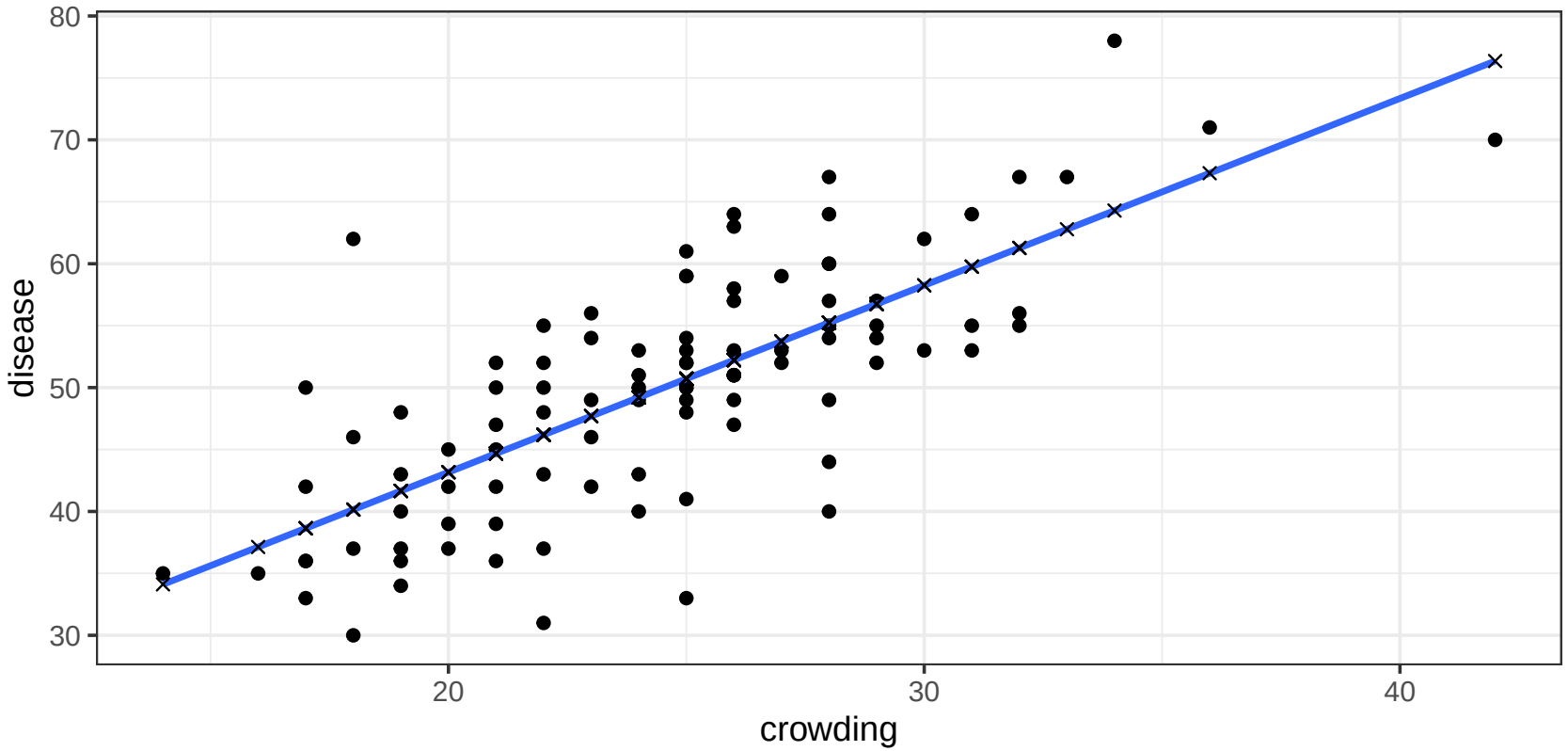


Assessing variation in response explained by model

- The mean squared error (MSE) is a commonly used measure of model accuracy:

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

- The MSE will be small if the observed and predicted responses are close to one another.



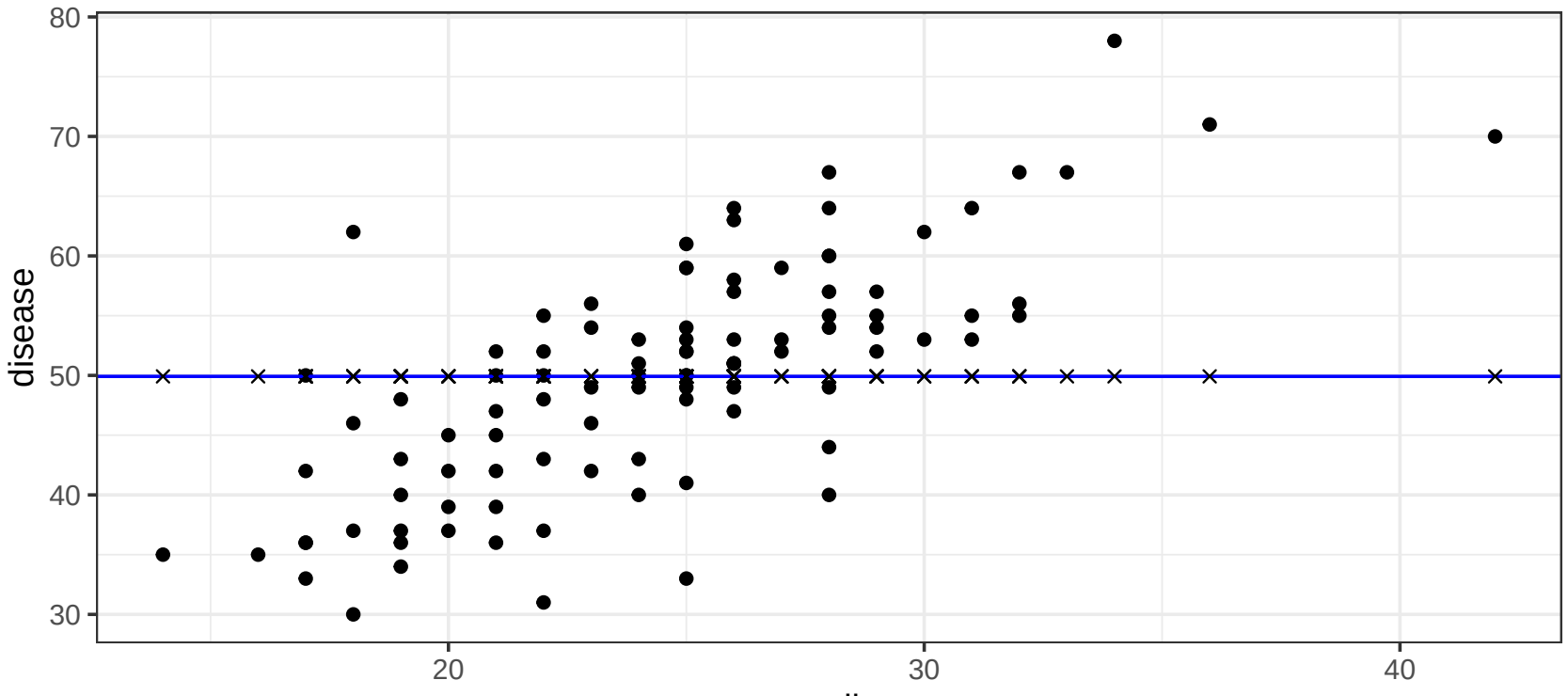
Connection to "sample variance"

From Stats 101: the sample variance of y is commonly defined as

$$Var(y) = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2 .$$

Note similarity with MSE for an intercept-only model:

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



Connecting residuals to outliers

- Outliers are points that lie away from the cloud of points.
- Outliers that lie horizontally away from the center of the cloud are called high leverage points.
- High leverage points that actually influence the slope of the regression line are called influential points.
- In order to determine if a point is influential, visualize the regression line with and without the point. Does the slope of the line change considerably? If so, then the point is influential. If not, then it's not an influential point.

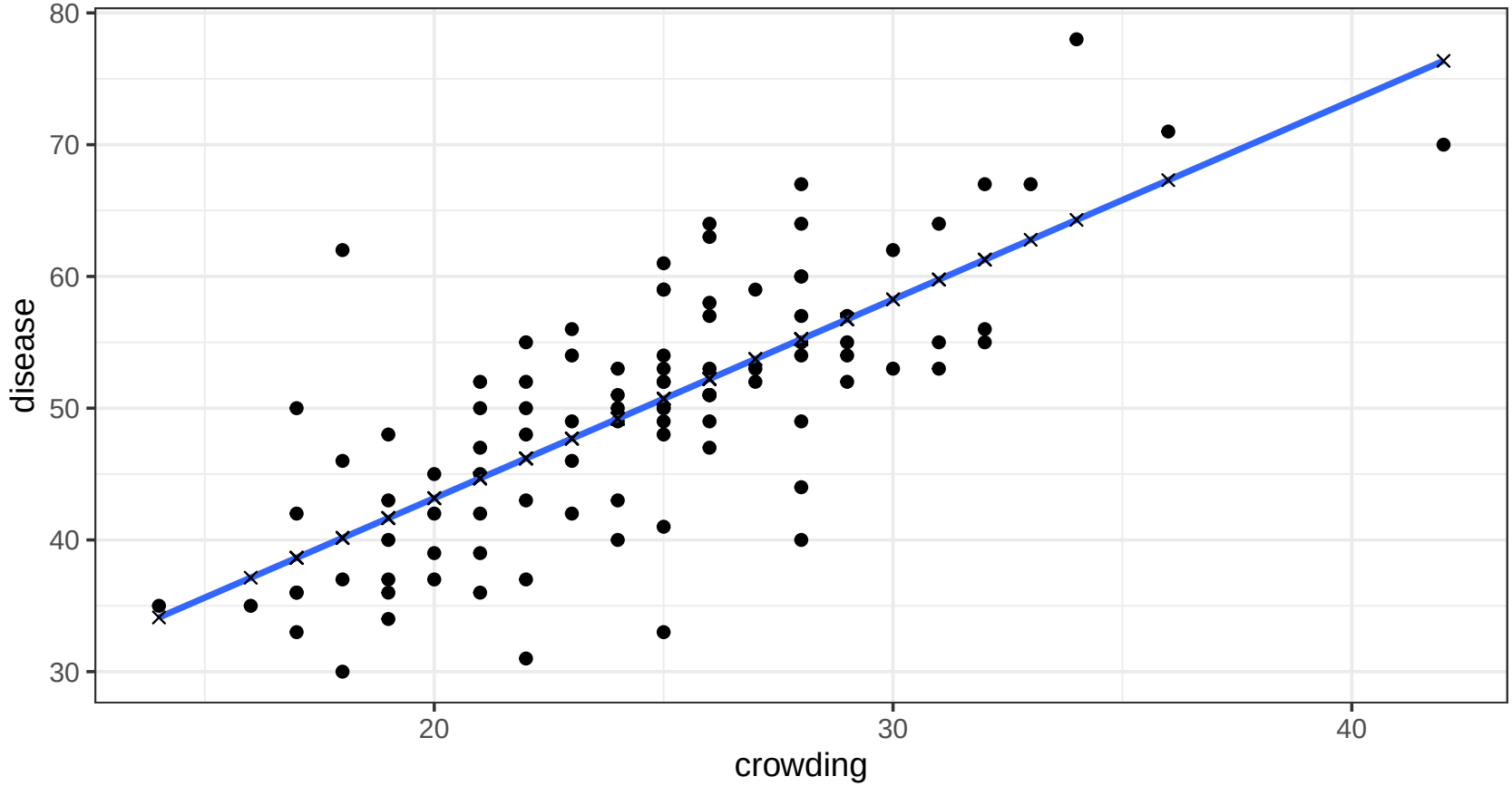
Influence

Intuitively, "influence" is a combination of outlying-ness and leverage. More specifically, we can measure the "deletion influence" of each observation: quantify how much $\hat{\boldsymbol{\beta}}$ changes if an observation is left out.

- Mathematically: $|\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(-i)}|$
- Cook's distance is a value we can calculate for each observation in our dataset that measures this deletion influence. (It uses some nice tricks of linear algebra without having to refit the regression iteratively without each point.)

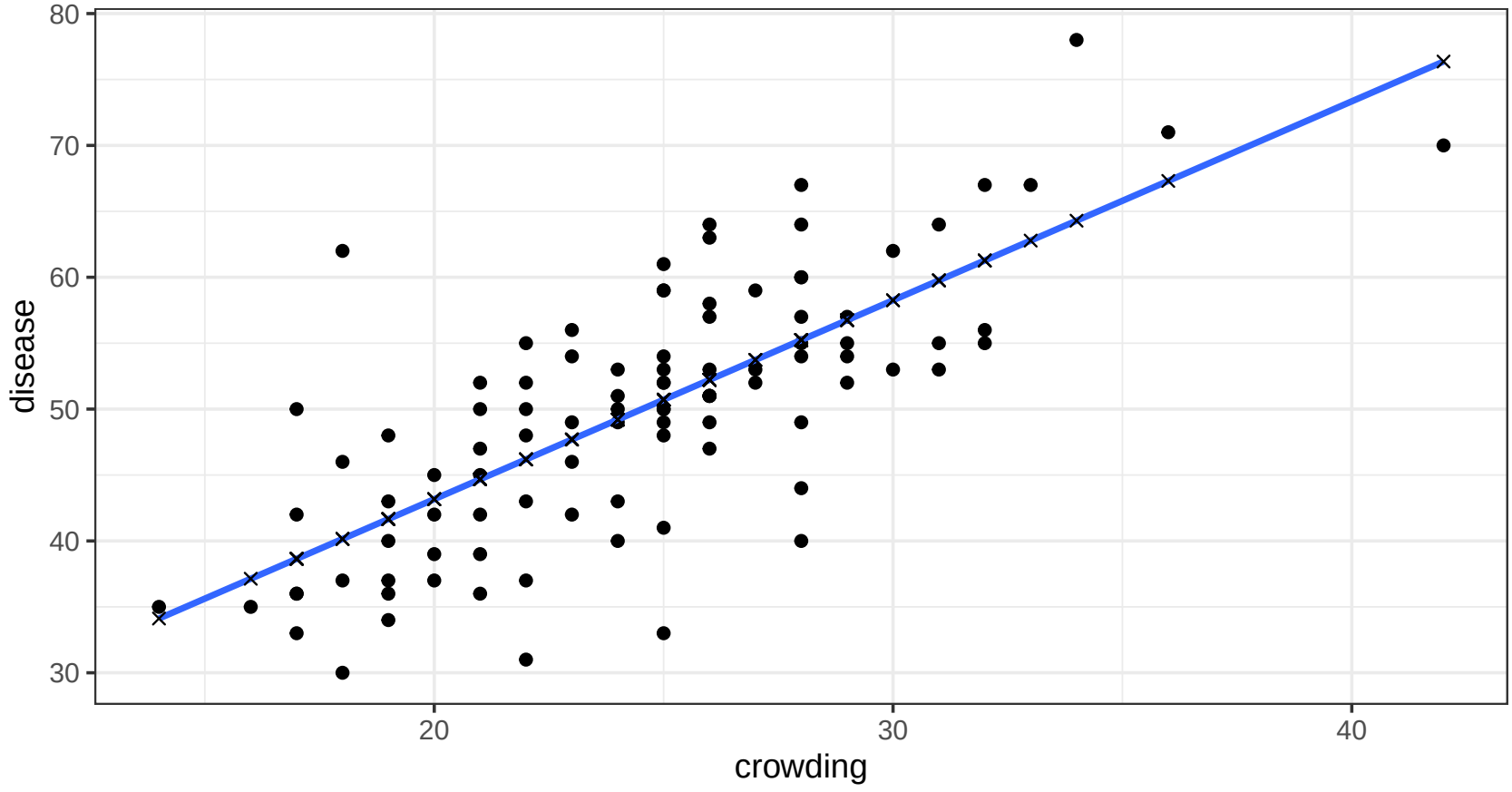
Geometrical thinking about influence

Imagine the regression line as a see-saw, being pulled by springs attached to each point. Which points do you think have the most influence?



Geometrical thinking about influence

- Points near $\bar{x}$ (the average value of x) do not influence the slope but might influence the position of the line.
- Only points far away from $\bar{x}$ can influence the slope.



Influential points

```
m1 <- lm(disease ~ crowding, data=dat)
dat$yhat <- predict(m1)
dat$resid <- resid(m1)
dat$cooksd <- cooks.distance(m1)
dat %>% select(disease, crowding, yhat, resid, cooksd) %>%
  arrange(desc(cooksd)) %>% head()

## # A tibble: 6 x 5
##   disease crowding  yhat  resid cooksd
##     <int>    <int> <dbl>  <dbl>  <dbl>
## 1      62       18  40.2  21.8  0.167
## 2      78       34  64.3  13.7  0.121
## 3      70       42  76.4  -6.36 0.0937
## 4      50       17  38.6  11.4  0.0556
## 5      40       28  55.2 -15.2  0.0430
## 6      33       25  50.7 -17.7  0.0375
```

Connecting back to data stories

It's really important to keep an eye out for points that might be unduly influential in telling the story!