The language of modeling

Author: Nicholas G Reich

Today's topics

- Introduction to modeling
- Defining components of models
- Defining model terms

Example: predicting respiratory disease severity ("lung" dataset)

Reading: MDSR: Appendix E and Chapter 9

Watch the first five minutes of Hadley Wickham's 2016 talk on the Tidyverse

" ... every model has to make assumptions, and a model by its very nature cannot question those assumptions...

models can never fundamentally surprise you because they cannot question their own assumptions."

Statistical modeling

The process of using data to describe the relationship between outcomes and predictors is called modeling.

- Models are models, not reality.
- "All models are wrong, but some are useful."
- Introduce structure to our model that balances realism with "goodness of fit".

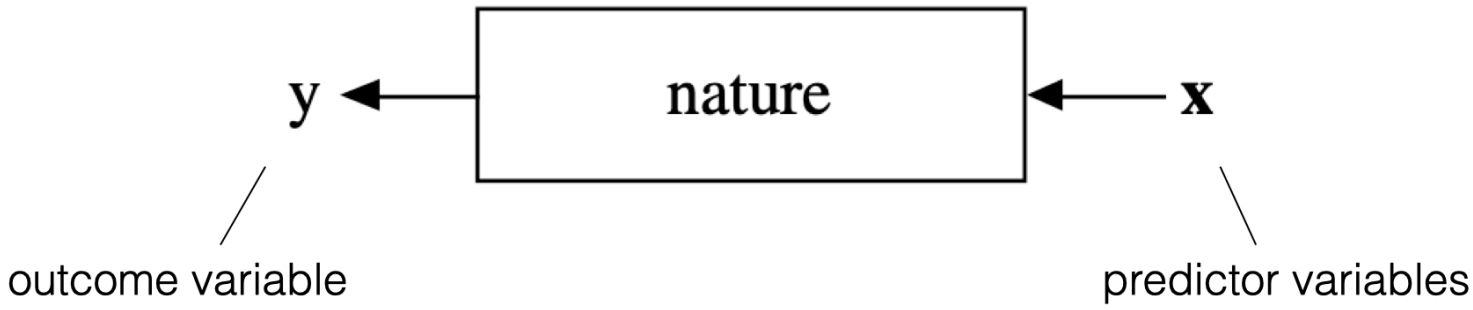
What are models?

Statistical Science
2001, Vol. 16, No. 3, 199–231

Statistical Modeling: The Two Cultures

Leo Breiman

Data arise thanks to the black box of nature.



e.g. covid cases next week in Hampshire county

covid cases this week, behavior, vaccination, cases in all past weeks, ...

What are models?

Statistical Science
2001, Vol. 16, No. 3, 199–231

Statistical Modeling: The Two Cultures

Leo Breiman

"To extract some information about how nature is associating the response variables to the input variables."

One goal: **infer** something about nature from data.

We want to learn something about the "true" state of nature, but we will never be able to observe what the black box relationships are between all the $\mathbf{x}$ and y.

How do population structure, human behavior, biological features of a pathogen, etc... interact to cause an outbreak?

What are models?

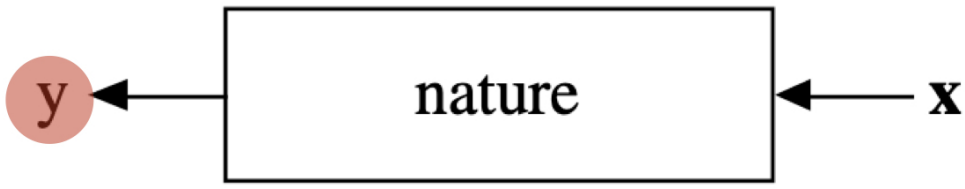
Statistical Science
2001, Vol. 16, No. 3, 199–231

Statistical Modeling: The Two Cultures

Leo Breiman

"To be able to predict what the responses are going to be to future input variables."

Another goal: **predict** new data.

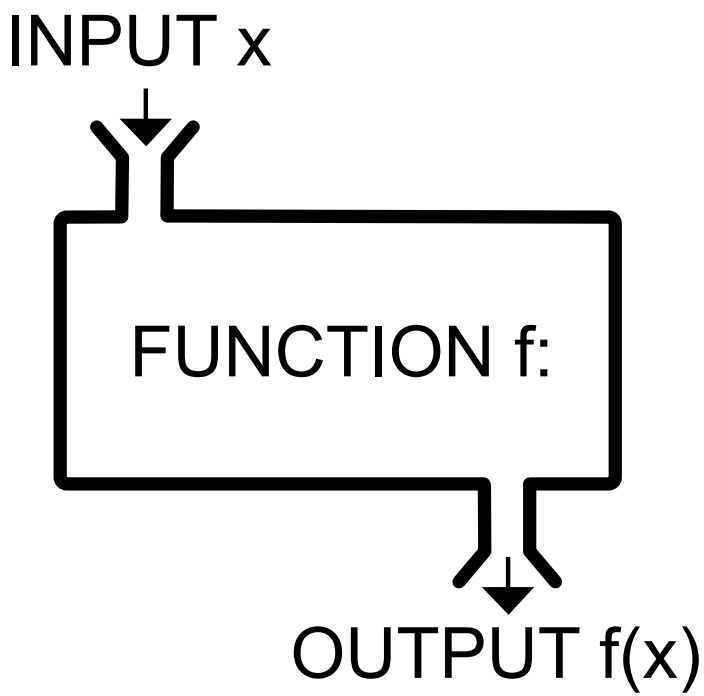


In prediction, we might be less concerned learning about nature, and more with what the the outcome y will be. If we are careful, we can pick problems and settings where we can (eventually) know the truth about what y will be given some $\mathbf{x}$.

How many cases will be observed next week?

Models are functions

Definition: "a **function** is a relation between a set of inputs and a set of permissible outputs with the property that each input is related to exactly one output".[1]



In statistical models, inputs are explanatory variables and outputs are "typical" or "expected" values of response variables.

[1] Wikipedia, https://en.wikipedia.org/wiki/Function_(mathematics)

Models are functions: response variable

Definition: "a **function** is a relation between a set of inputs and a set of permissible outputs with the property that each input is related to exactly one output".[2]

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. We might also call this the **outcome variable**.

[2] Wikipedia, https://en.wikipedia.org/wiki/Function_(mathematics)

Lung data example

99 observations on patients who have sought treatment for the relief of respiratory disease symptoms.
The variables are:

- `disease` measure of disease severity (larger values indicates more serious condition).
- `education` highest grade completed
- `crowding` measure of crowding of living quarters (larger values indicate more crowding)
- `airqual` measure of air quality at place of residence (larger number indicates poorer quality)
- `nutrition` nutritional status (larger number indicates better nutrition)
- `smoking` smoking status (1 if smoker, 0 if non-smoker)

What is the natural response variable here? Which variable are we trying to understand or explain?

Lung data example: looking at the data

What variables will explain variation in disease severity?

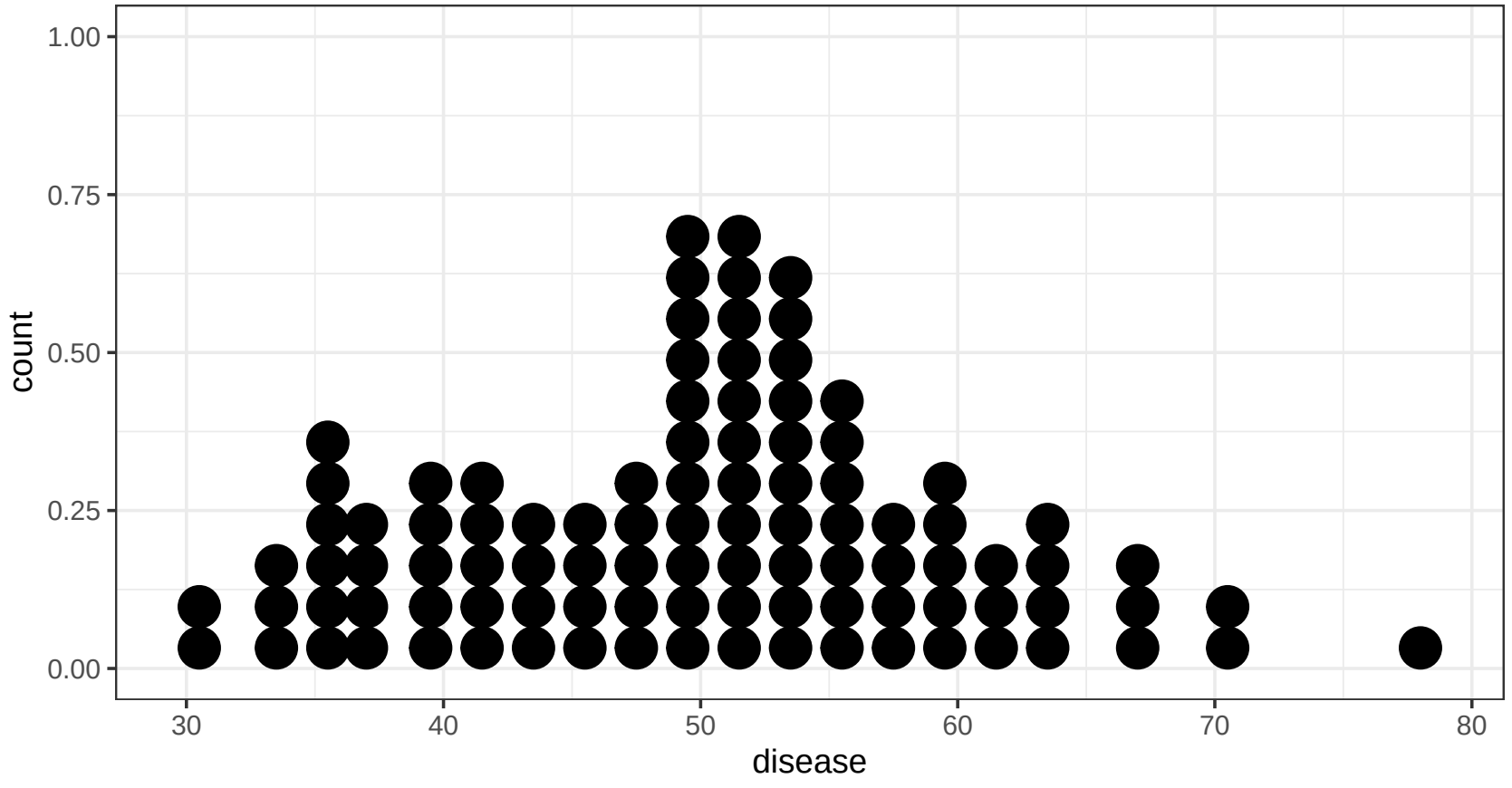
```
dat <- read.table("../../data/lungc.txt", header=TRUE) %>%
  mutate(smoking = factor(smoking,
                          levels=c(0,1),
                          labels=c("non-smoker", "smoker")))
head(dat)
```

```
##   disease education crowding airqual nutrition    smoking
## 1      67         7        33      61        18     smoker
## 2      47        11        21      43       109 non-smoker
## 3      53        10        24      54         0     smoker
## 4      56         9        32      43        97     smoker
## 5      48         8        22      62       131 non-smoker
## 6      64        10        28      76        79     smoker
```

Lung data example: looking at variability in the response

What is the variation in disease severity?

```
ggplot(dat, aes(x=disease)) + geom_dotplot()
```



Models are functions: explanatory variables

We might write generally

$$y = f(x)$$

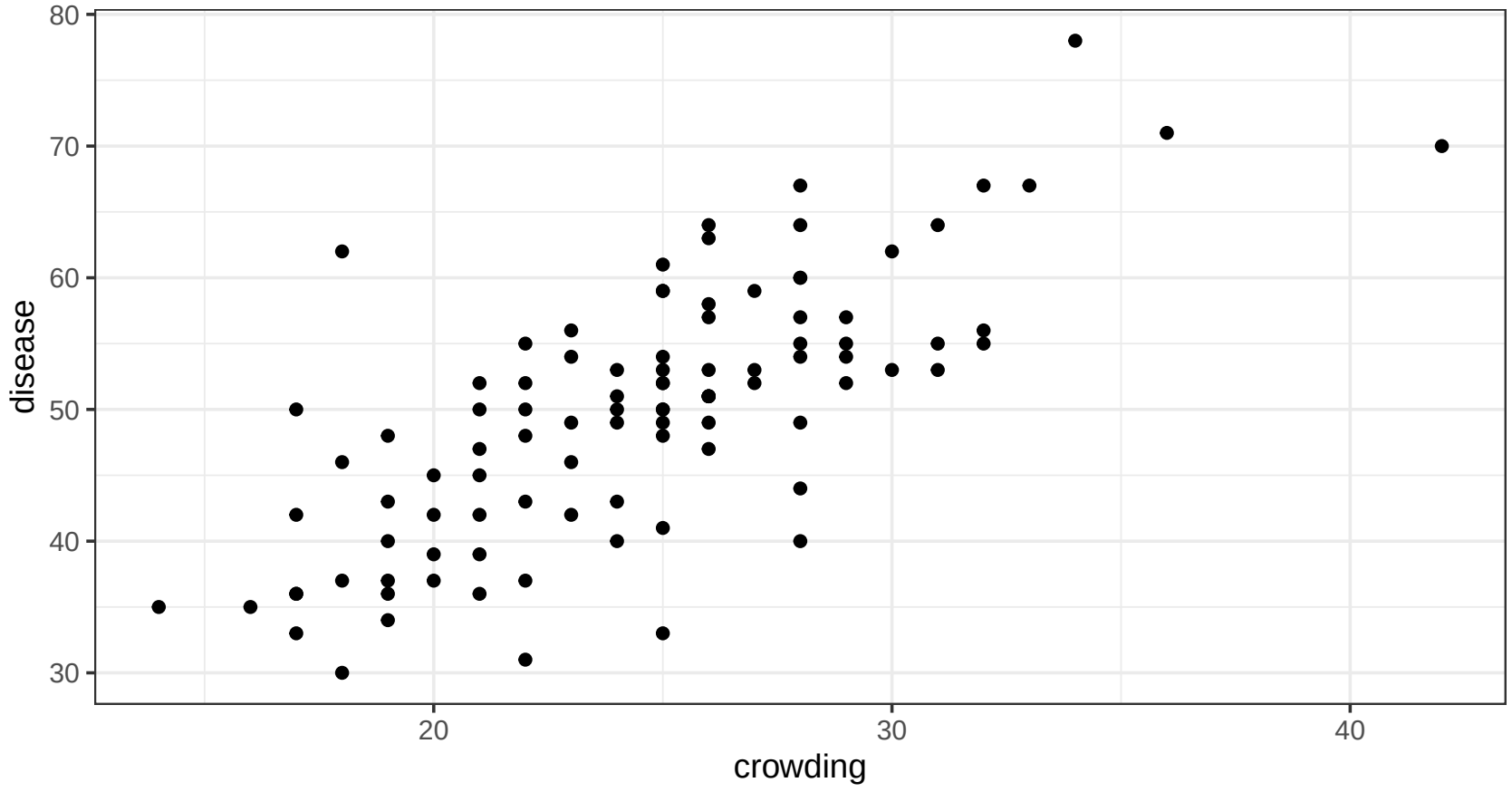
where x could be a single variable or multiple variables.

- **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.

Lung data example: explaining variability in the response

Does crowding of living quarters explain some of the variation in disease severity?

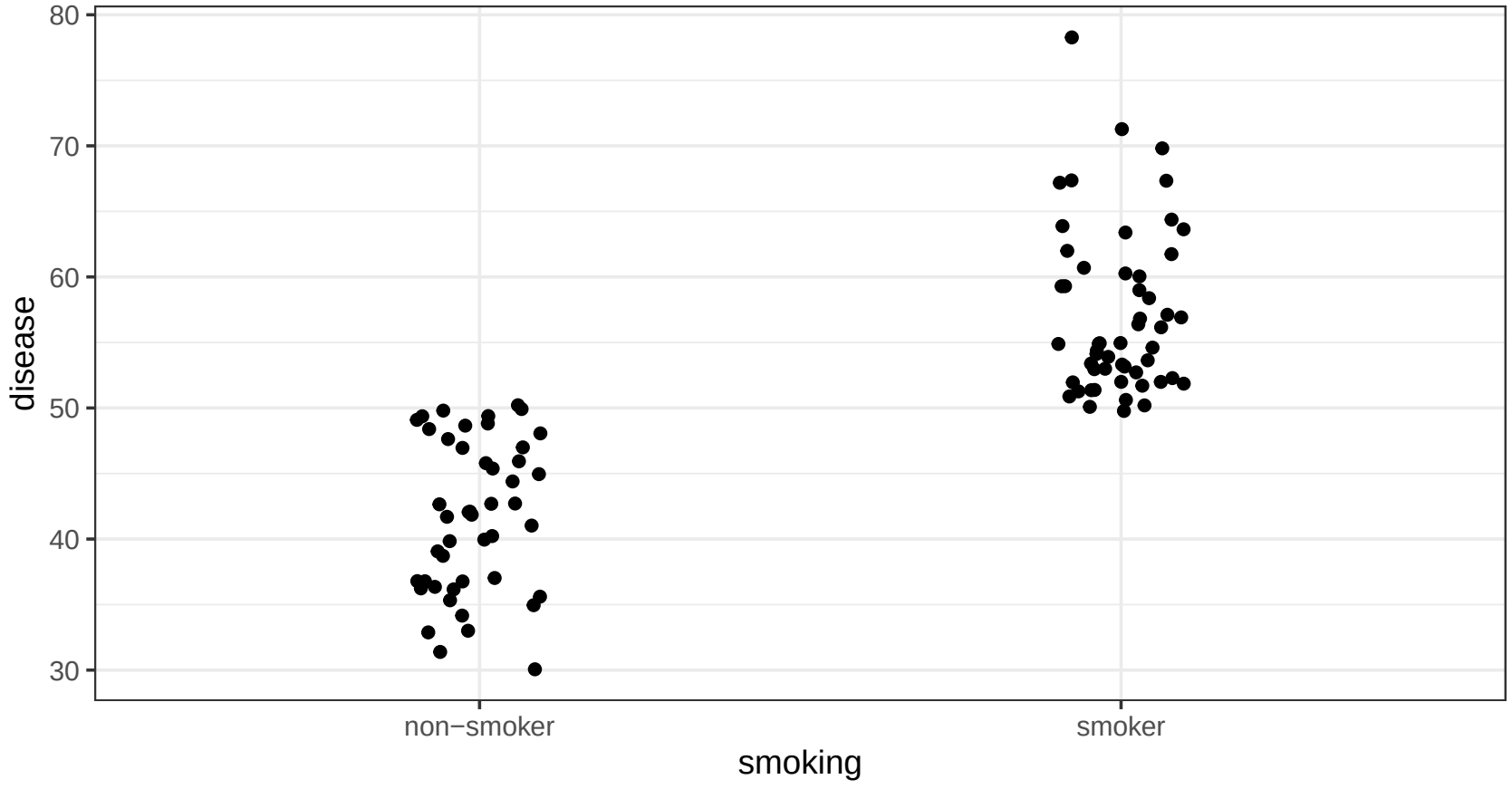
```
ggplot(dat, aes(crowding, disease)) +
  geom_point()
```



Lung Data Example: explaining variability in the response

Does smoking status explain some of the variation in disease severity?

```
ggplot(dat, aes(smoking, disease)) + geom_jitter(width=.1)
```



Modeling recap

We might write generally

$$y = f(x)$$

where x could be a single variable or multiple variables.

What will the "structure" of the model look like?

- Most models we talk about will be a form of **linear models**, e.g.

$$y = f(x) = \beta_0 + \beta_1 \cdot x$$

.

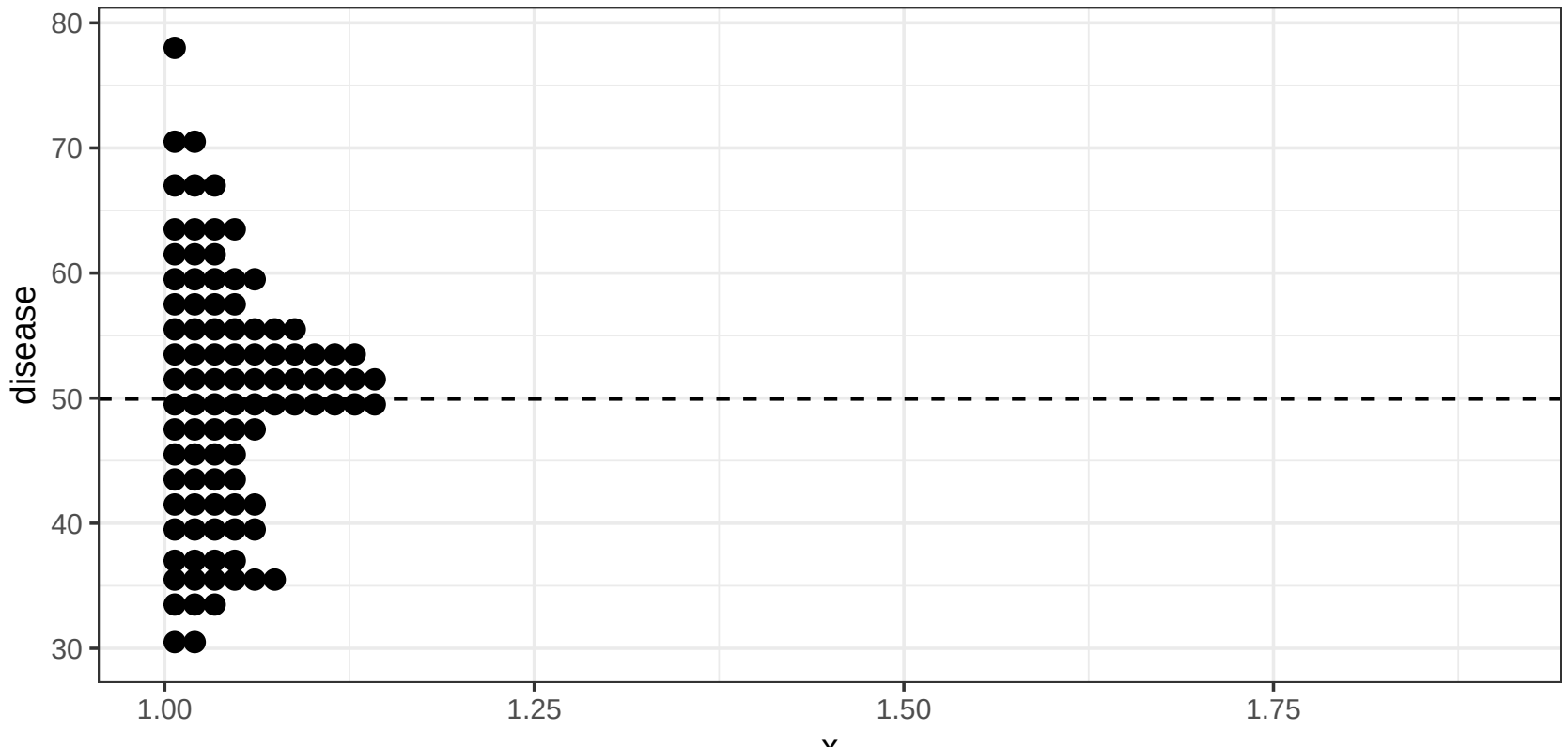
- You must make a choice about **model terms**. What does the right hand side of the above equation look like?

Model terms: the intercept

The intercept is a "baseline" that is included in nearly every model. What would your guess of disease severity be in the absence of any other information?

$$y = \beta_0$$

```
ggplot(dat, aes(x=1, y=disease)) + geom_dotplot(binaxis="y") +
  geom_hline(yintercept = mean(dat$disease), linetype=2)
```

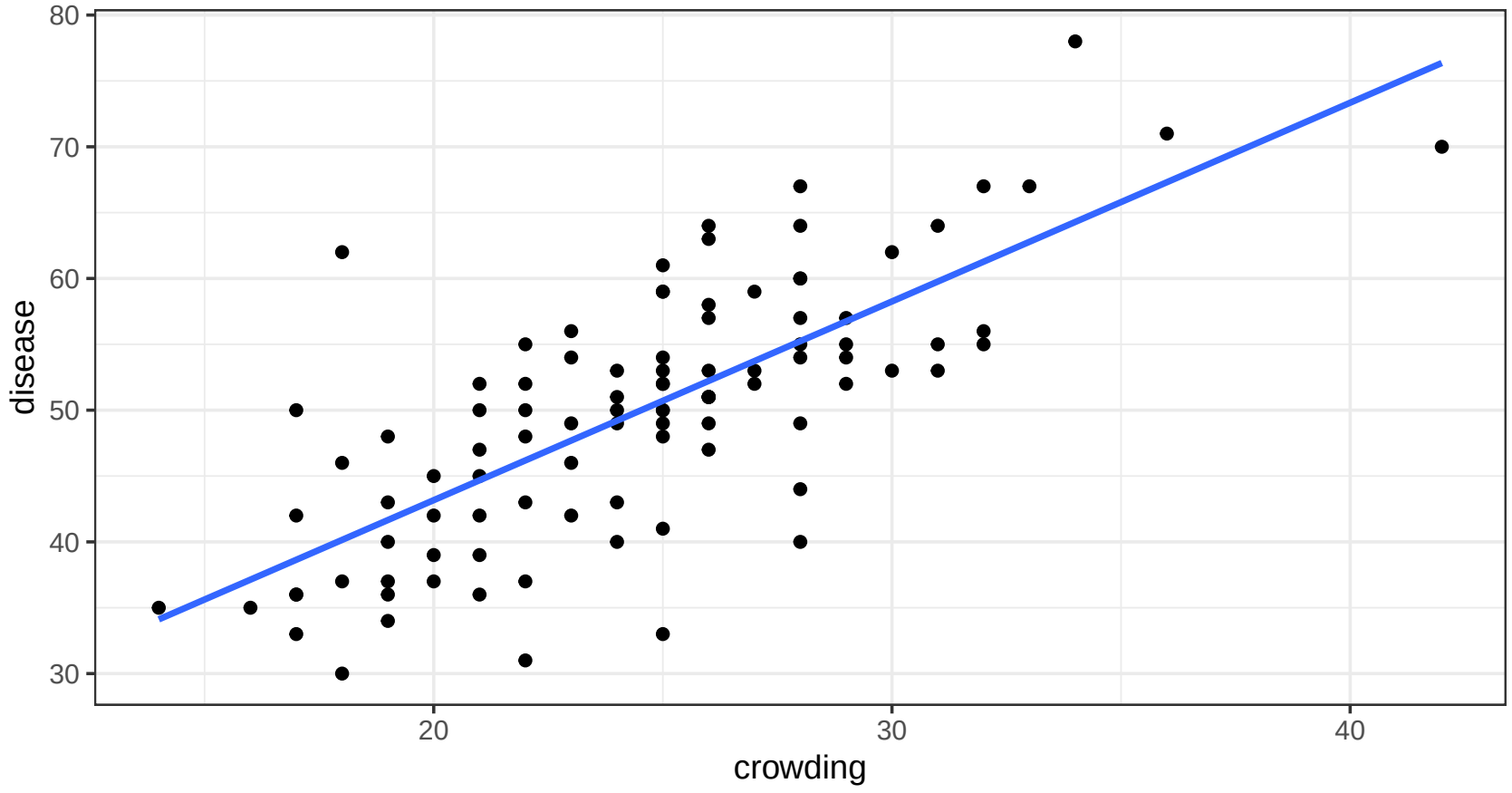


Model terms: main terms

Main terms model the effect of explanatory variables directly.

$$y = \beta_0 + \beta_1 \cdot \textit{crowding}$$

```
ggplot(dat, aes(crowding, disease)) + geom_point() +
  geom_smooth(method="lm", se=FALSE)
```

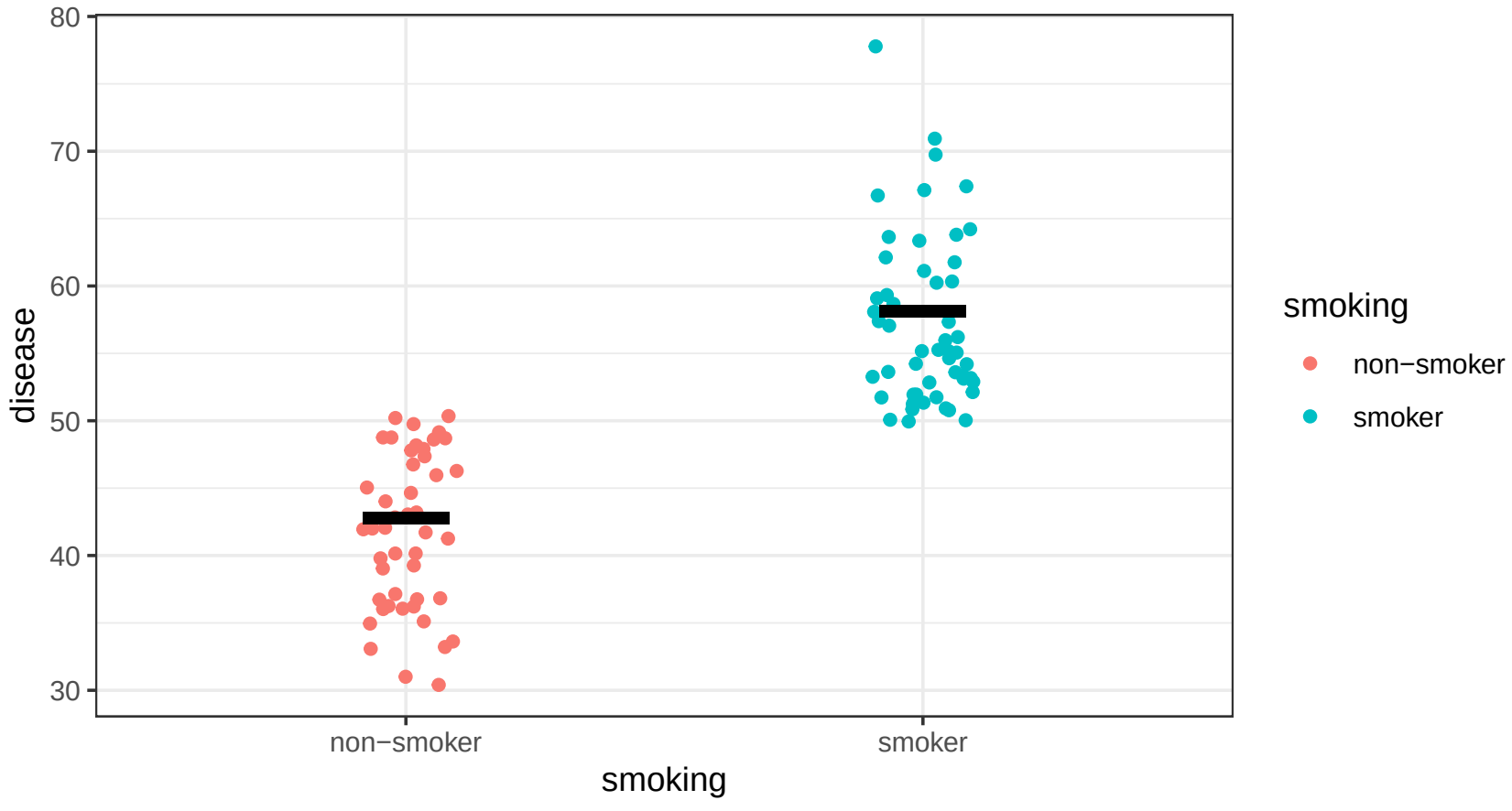


Model terms: main terms

Main terms model the effect of explanatory variables directly.

$$y = \beta_0 + \beta_2 \cdot \textit{smoking}$$

```
ggplot(dat, aes(x=smoking, y=disease, color=smoking)) + geom_jitter(wid
  stat_summary(fun=mean, geom="point", shape="-", size=20, color="black
```

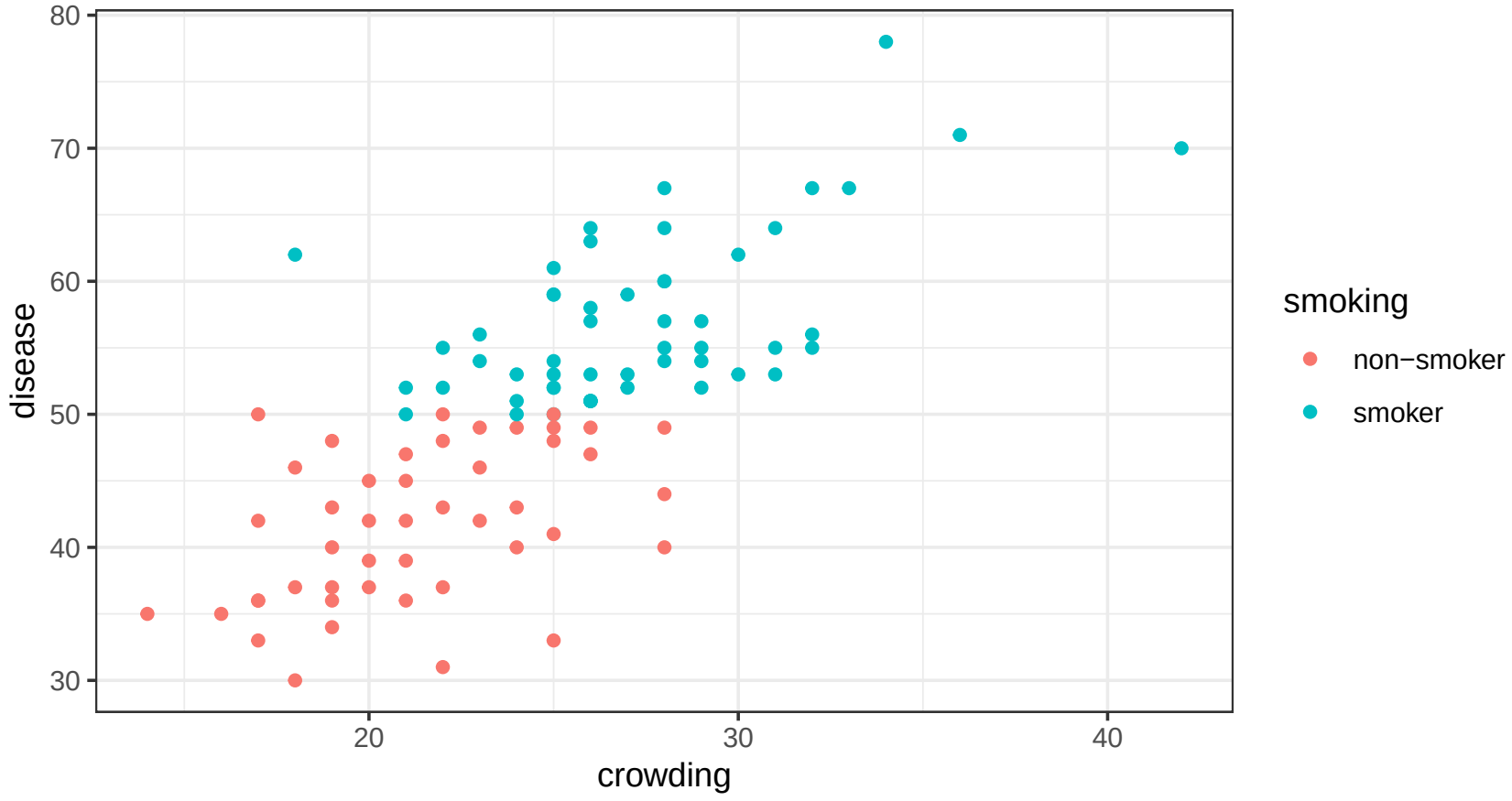


Model terms: two main terms

Main terms model the effect of explanatory variables directly.

$$y = \beta_0 + \beta_1 \mathit{crowding} + \beta_2 \cdot \mathit{smoking}$$

```
ggplot(dat, aes(x=crowding, y=disease, color=smoking)) +
  geom_point()
```

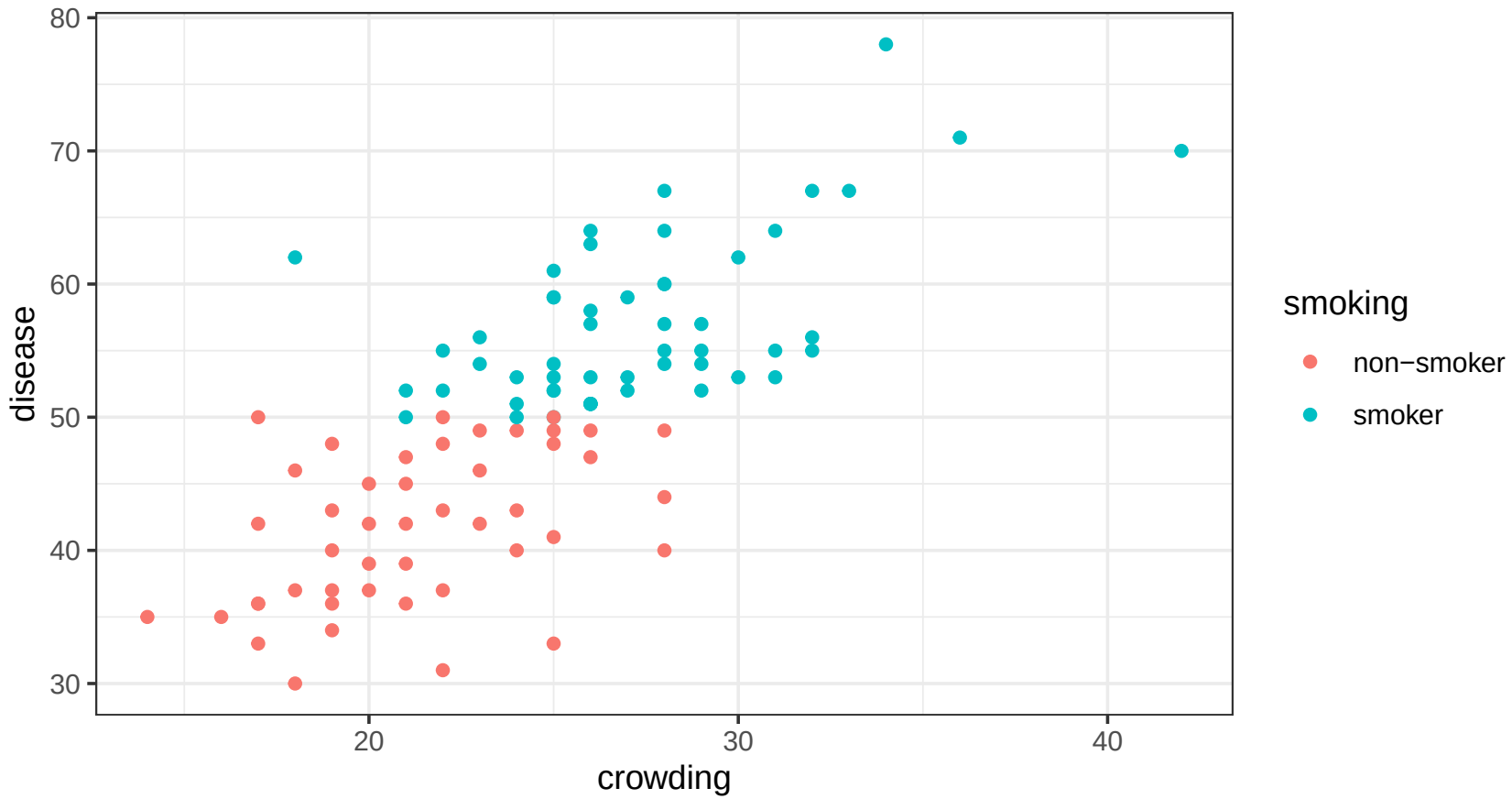


Model terms: interaction terms

Interaction terms allow for different explanatory variables to modulate the relationship of each other to the response variable.

$$y = \beta_0 + \beta_1 \cdot \textit{crowding} + \beta_2 \cdot \textit{smoking} + \beta_3 \cdot \textit{crowding} \cdot \textit{smoking}$$

```
ggplot(dat, aes(x=crowding, y=disease, color=smoking)) +
  geom_point()
```



Model terms: recap

- **The intercept** is a "baseline" that is included in nearly every model. What would your guess of disease severity be in the absence of any other information?
- **Main terms** model the effect of explanatory variables directly.
- **Interaction terms** allow for different explanatory variables to modulate the relationship of each other to the response variable.
- **Smooth terms** and **transformation terms**: to come soon!