The language of modeling

Author: Nicholas G Reich

Today's topics

- The language of models
- Model formulas and coefficients

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

Reading: Kaplan, Chapters 4, 6-10.

Watch the first five minutes of Hadley Wickham's UseR! 2016 talk

" ... 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."

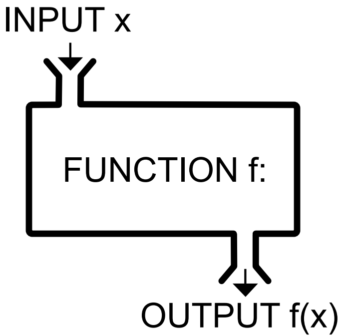
Statistical modeling

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”.

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)

A common modeling tool: regression

- The goal is to learn about the relationship between "explanatory" (or "predictor") variables of interest and a "response" (or "outcome") of interest.
  - Some models focus on prediction.
  - Other models focus on description.
- Regression is an exercise in inferential statistics: we are drawing evidence and conclusions from data about "complex aspects of reality", i.e. "noisy" systems.

Lung data example

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

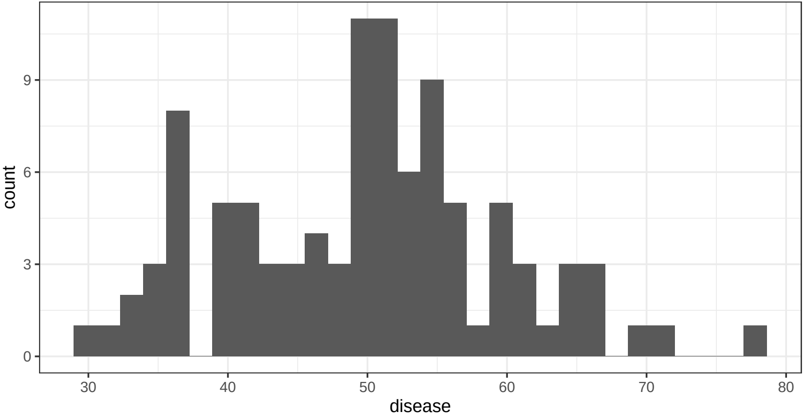
- `disease` measure of disease severity (larger values indicates more serious condition).
- `education` highest grade completed
- `crowding` measure of crowding of living quarters (larger values indicate more crowding)
- `airqual` measure of air quality at place of residence (larger number indicates 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 variability in the response

What variables will explain variation in disease severity?

```
dat <- read.table("../../data/lungc.txt", header=TRUE)
ggplot(dat, aes(x=disease)) +
  geom_histogram()
```



Models are functions: explanatory variables

We might write generally

$$y = f(x)$$

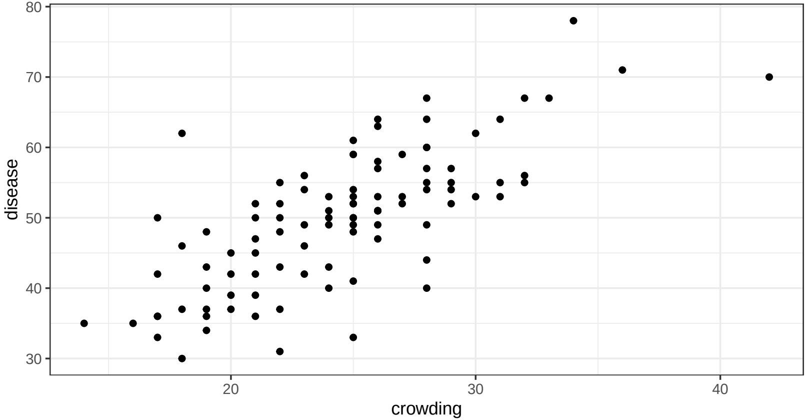
where x could be a single variable or multiple variables.

- **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?

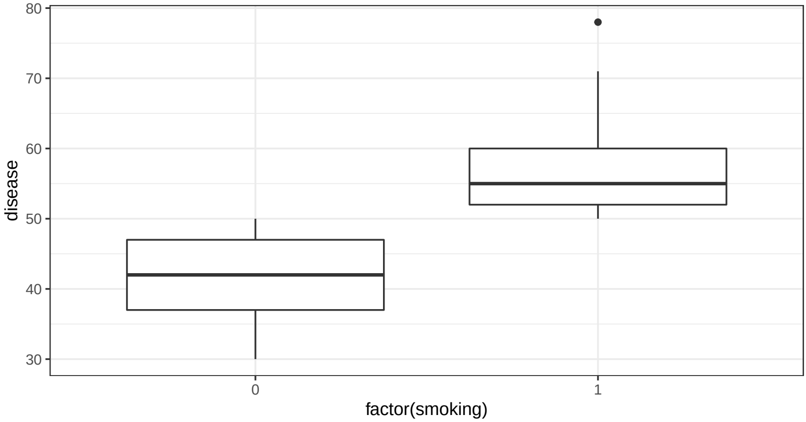
```
ggplot(dat, aes(crowding, disease)) +
  geom_point()
```



Lung Data Example: explaining variability in the response

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

```
ggplot(dat, aes(factor(smoking), disease)) + geom_boxplot()
```



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?

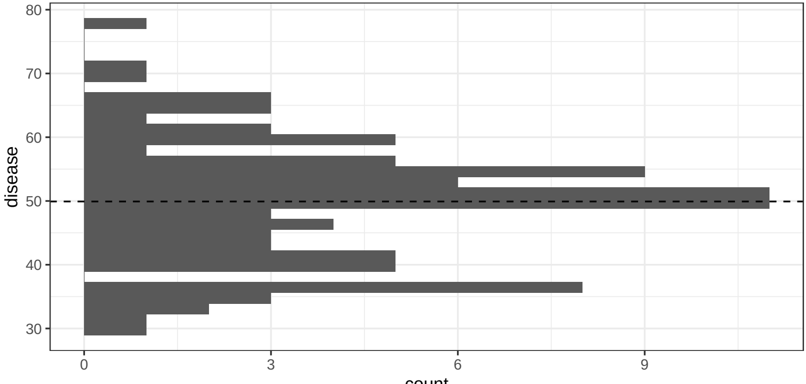
- Most models we talk about will be a form of **linear models**, e.g.
$$y = f(x) = \beta_0 + \beta_1 \cdot x$$
.
- You must make a choice about **model terms**. What does the right hand side of the above equation look like?

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(y=disease)) +
  geom_histogram() + geom_hline(yintercept = mean(dat$disease), linetyp
```

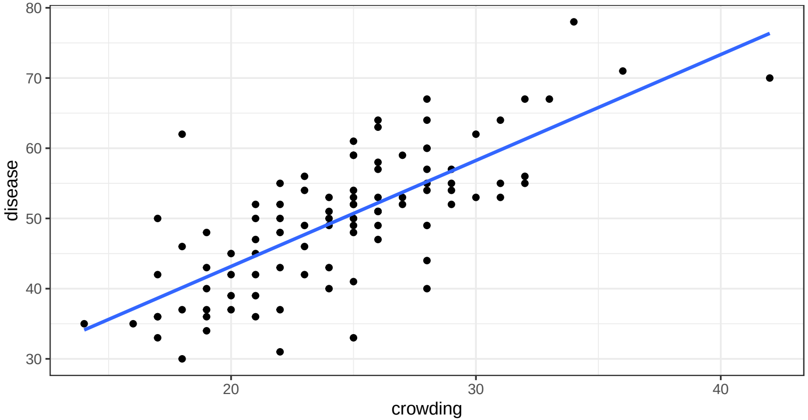


Model terms: main terms

Main terms model the effect of explanatory variables directly.

$$y = \beta_0 + \beta_1 \cdot crowding$$

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

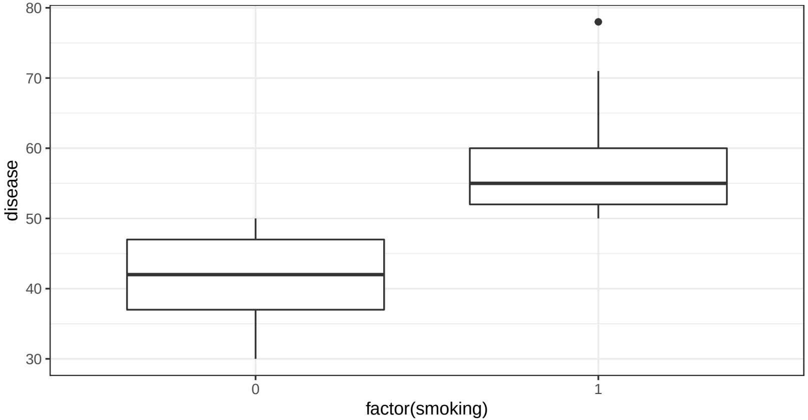


Model terms: main terms

Main terms model the effect of explanatory variables directly.

$$y = \beta_0 + \beta_2 \cdot smoking$$

```
ggplot(dat, aes(x=factor(smoking), y=disease)) +
  geom_boxplot()
```

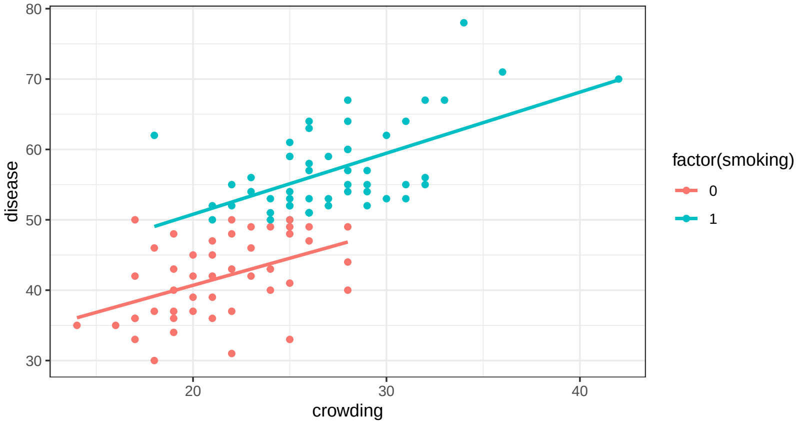


Model terms: interaction terms

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

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

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



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!