Fitting and interpreting model coefficients

Author: Nicholas G Reich

Today's topics

- Model terms: recap
- Fitting and interpreting models

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.

Formulas for Statistical Models (Linear Regression)

In general, linear models can be thought of as having these components

$$\text{y} \quad = \quad \text{intercept} + \text{terms} + \text{error}$$

With a single predictor variable, this is simply a line (plus error):

$$y_i \quad = \quad \beta_0 + \beta_1 \cdot x_i + \epsilon_i$$

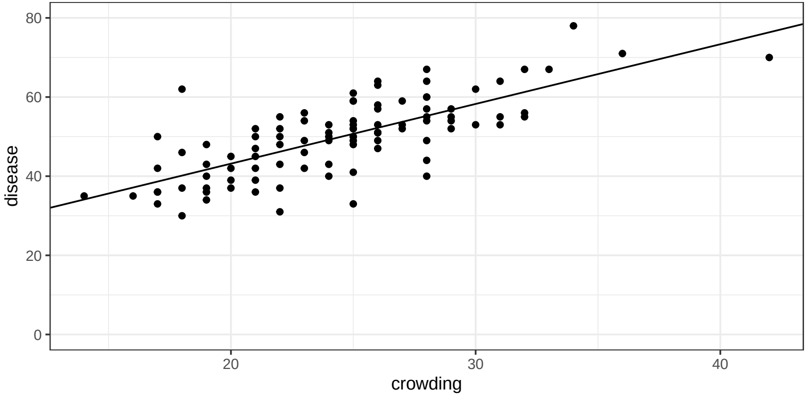
However, there can be multiple variables and different types of "terms" in this equation

- intercept
- main effects
- interaction terms
- smooth terms

Main effects model terms

$$\text{equation:} \quad \widehat{disease}_i = \beta_0 + \beta_1 \cdot crowding_i$$

```r
m1 <- lm(disease ~ crowding, data=dat)
```



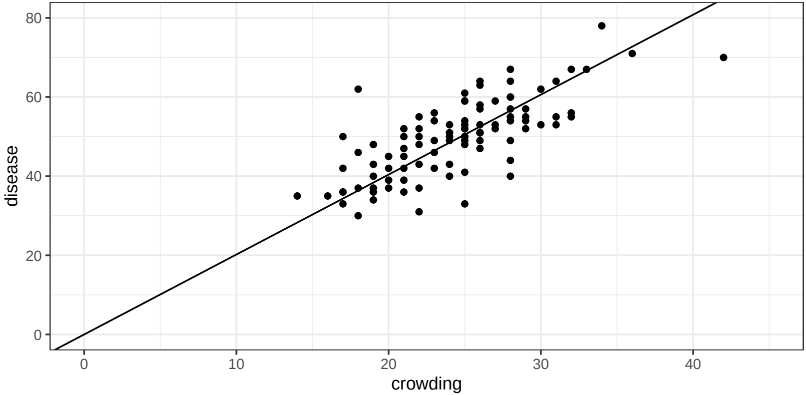
Dust off your algebra: what is an intercept?

$$y = \beta_1 \cdot x \quad \text{vs.} \quad y = \beta_0 + \beta_1 \cdot x$$

Main effects with no intercept: bad idea

$$\text{equation:}\quad \widehat{disease}_i = \beta_1 \cdot crowding_i$$

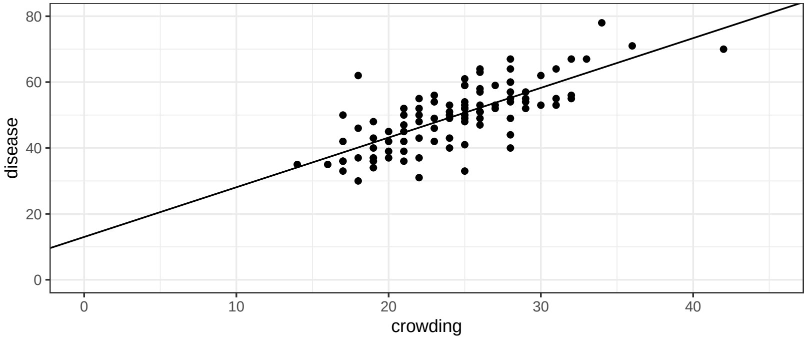
```
m1_no_intcpt <- lm(disease ~ crowding - 1, data=dat)
```



Main effects model terms

$$\text{equation:} \quad \widehat{disease}_i = \beta_0 + \beta_1 \cdot crowding_i$$

```
m1 <- lm(disease ~ crowding, data=dat)
coef(m1)

## (Intercept)    crowding
##   12.991536    1.508806
```



Main effects model terms: interpretation

$$\text{equation:} \quad \widehat{disease}_i = \beta_0 + \beta_1 \cdot crowding_i$$

```
m1 <- lm(disease ~ crowding, data=dat)
coef(m1)

## (Intercept)    crowding
##   12.991536    1.508806
```

Main effects model terms: interpretation

$$\text{equation:} \quad \widehat{disease}_i = \beta_0 + \beta_1 \cdot crowding_i$$

```
m1 <- lm(disease ~ crowding, data=dat)
coef(m1)

## (Intercept)    crowding
##   12.991536    1.508806
```

β_0 is the expected value of *disease* when $crowding = 0$.

Main effects model terms: interpretation

$$\text{equation:}\quad \widehat{disease}_i = \beta_0 + \beta_1 \cdot crowding*_i$$

```
dat$crowding_ctr <- dat$crowding - mean(dat$crowding)
m1a <- lm(disease ~ crowding_ctr, data=dat)
coef(m1a)

##  (Intercept) crowding_ctr
##    49.919192     1.508806
```

β_0 is the expected value of *disease* when $crowding_{ctr} = 0$, in other words, when crowding is the average value.

Main effects model terms: interpretation

$$\text{equation:} \quad \widehat{disease}_i = \beta_0 + \beta_1 \cdot crowding_i$$

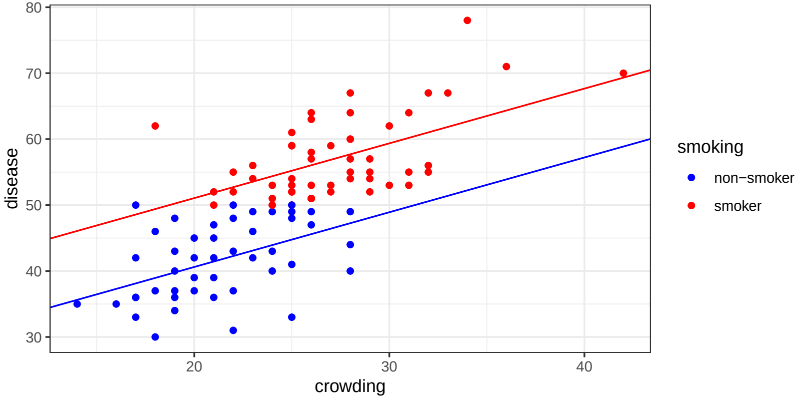
```r
m1 <- lm(disease ~ crowding, data=dat)
coef(m1)

## (Intercept)    crowding
##   12.991536    1.508806
```

β_1 is the expected change in disease for a 1 unit increase of crowding.

2 main effects: 1 continous, 1 categorical

$$\text{equation:} \quad \widehat{disease}_i = \beta_0 + \beta_1 \cdot crowding_i + \beta_2 \cdot smoker$$

```
m2 <- lm(disease ~ crowding + smoking, data=dat)
```



2 main effects: 1 continous, 1 categorical

equation: $\widehat{disease}_i = \beta_0 + \beta_1 \cdot crowding_i + \beta_2 \cdot smoker$

```
m2 <- lm(disease ~ crowding + smoking, data=dat)
coef(m2)

##   (Intercept)      crowding smokingsmoker
##    24.0079027     0.8302413    10.4442068
```

2 main effects: 1 continous, 1 categorical

equation: $\widehat{\mathit{disease}}_i = \beta_0 + \beta_1 \cdot \mathit{crowding}_i + \beta_2 \cdot \mathit{smoker}$

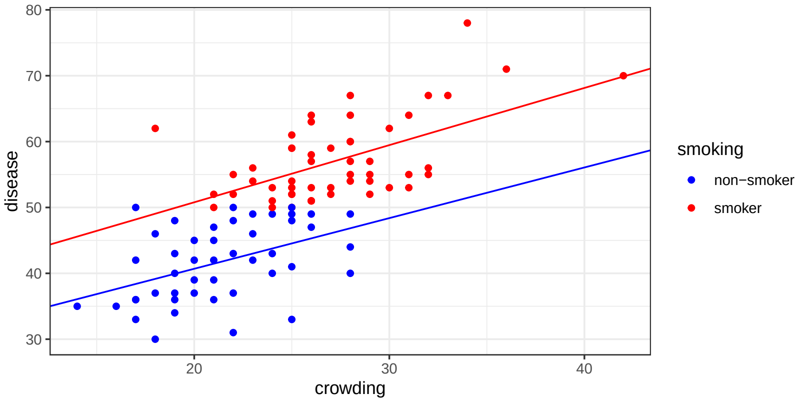
```
m2 <- lm(disease ~ crowding + smoking, data=dat)
coef(m2)

##   (Intercept)      crowding smokingsmoker
##    24.0079027     0.8302413    10.4442068
```

β_0 is the expected value of disease when both crowding and smoker are zero.
β_1 is the expected change in disease for a one-unit change in crowding, holding smoking status constant.
β_2 is the expected difference in disease between smokers and non-smokers, holding crowding constant.

Interaction model terms

equation: $\widehat{disease}_i = \beta_0 + \beta_1 \cdot crowd_i + \beta_2 \cdot smoke_i + \beta_3 \cdot crowd_i \cdot smoke_i$

```
m3 <- coef(lm(disease ~ crowding*smoking, data=dat))
```



Interaction vs. confounding

Definition of interaction

Interaction occurs when the relationship between two variables depends on the value of a third variable. E.g. you could hypothesize that the true relationship between nutritional intake and disease severity may be different for smokers and non-smokers.

Definition of confounding

Confounding occurs when the measurable association between two variables is distorted by the presence of another variable. Confounding can lead to biased estimates of a true relationship between variables.

- It is important to include confounding variables. Not doing so may bias your results.
- Unmodeled interactions do not lead to "biased" estimates in the same way that confounding does, but it can lead to a richer and more detailed description of the data at hand.

How to include interaction in a MLR

Model A: $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \epsilon_i$
Model B: $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{1i} \cdot x_{2i} + \epsilon_i$

Key points

- “easily” conceptualized with 1 continuous, 1 categorical variable
- models possible with other variable combinations, but interpretation/visualization harder
- two variable interactions are considered “first-order” interactions
- still a **linear** model, but no longer a strictly **additive** model

How to interpret an interaction model

For now, assume x_1 is continuous, x_2 is 0/1 binary.
Model A: $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \epsilon_i$
Model B: $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{1i} \cdot x_{2i} + \epsilon_i$

How to interpret an interaction model

For now, assume x_1 is continuous, x_2 is 0/1 binary.
Model A: $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \epsilon_i$
Model B: $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{1i} \cdot x_{2i} + \epsilon_i$

β_3 is the change in the slope of the line that describes the relationship of $y \sim x_1$ comparing the groups defined by $x_2 = 0$ and $x_2 = 1$.
$\beta_1 + \beta_3$ is the expected change in y for a one-unit increase in x_1 in the group $x_2 = 1$.
$\beta_0 + \beta_2$ is the expected value of y in the group $x_2 = 1$ when $x_1 = 0$.

Fitting models in R: syntax summary

Quick recap of key syntax for linear models

- For linear models, use `lm()`.
- Equations look like `y ~ x1 + x2`.
- Plus signs (+) indicate main effect terms.
- Multiplication signs (∗) indicate main effect AND interaction terms.

Group work