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

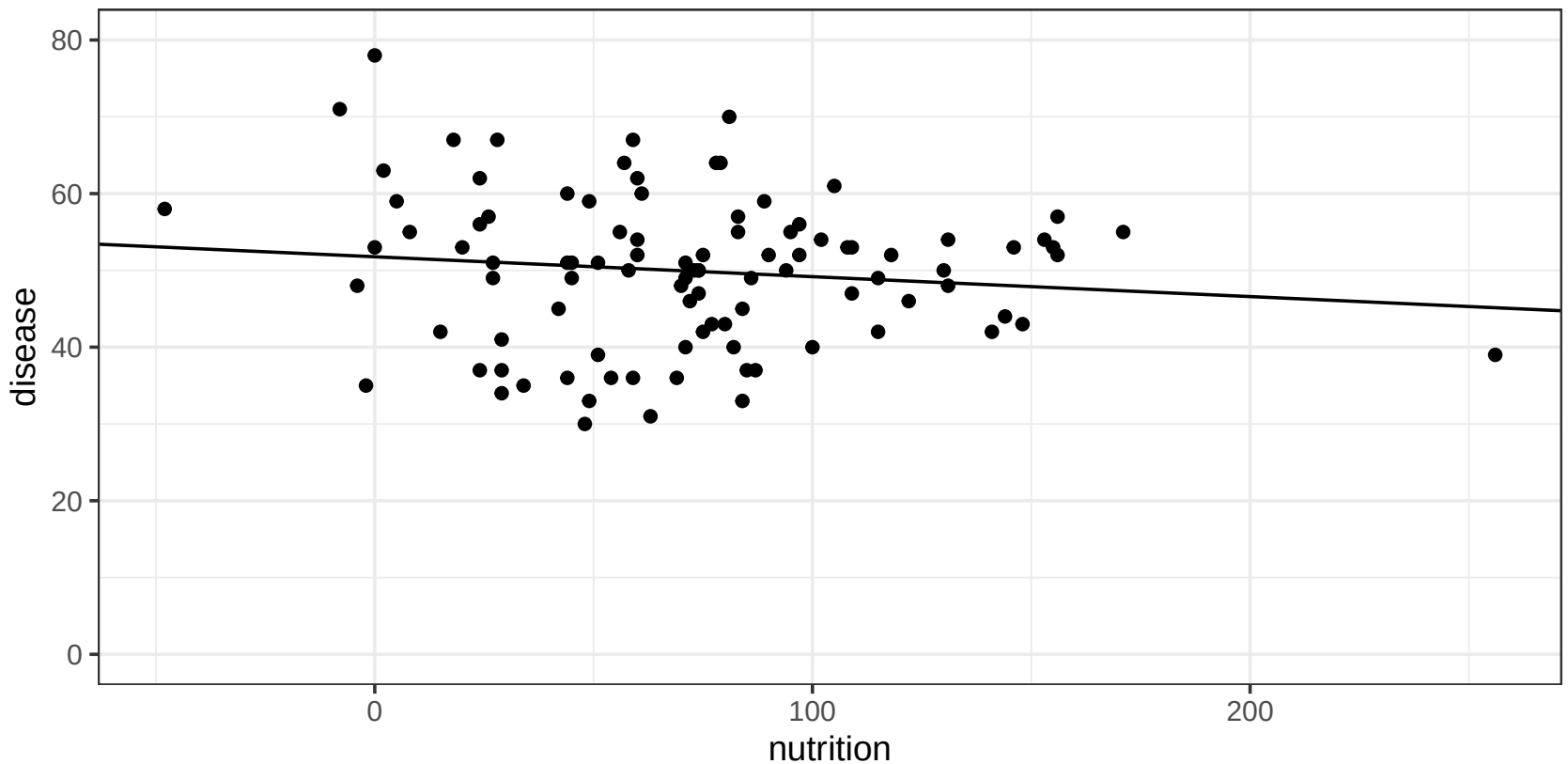
- intercept
- main effects
- interaction terms
- smooth terms

Main effects model terms

equation: $\widehat{disease}_i = \beta_0 + \beta_1 \cdot nutrition_i$

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

```
ggplot(dat, aes(x=nutrition, y=disease)) +
  geom_point() + ylim(c(0,80)) +
  geom_abline(intercept = coef(m1)[1], slope = coef(m1)[2])
```



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 model terms

equation: $\widehat{disease}_i = \beta_0 + \beta_1 \cdot nutrition_i$

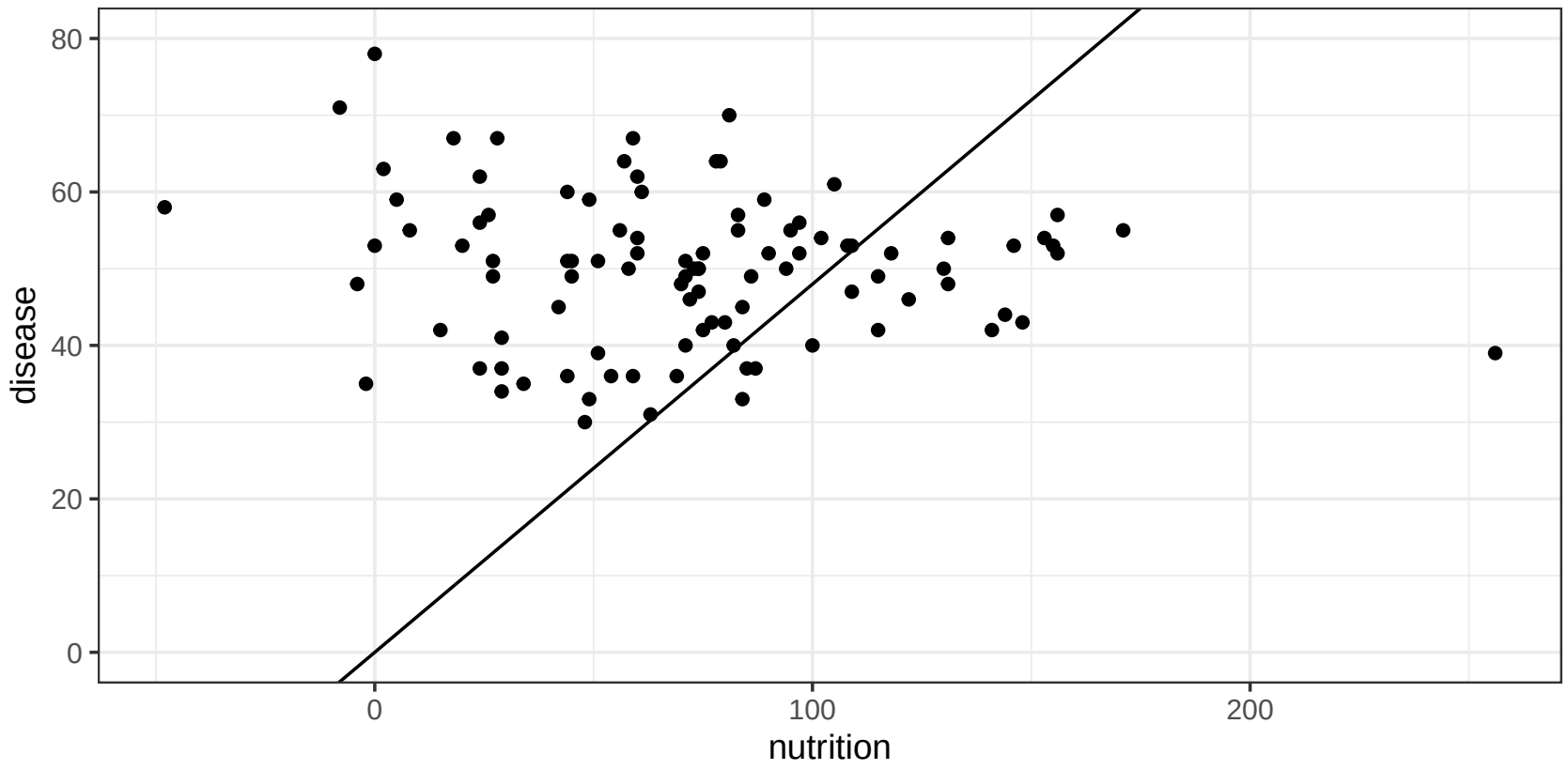
```
m1 <- lm(disease ~ nutrition, data=dat)
summary(m1)

##
## Call:
## lm(formula = disease ~ nutrition, data = dat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -20.5306  -6.4197   0.3916   5.6545  26.2250
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 51.77502    1.78403  29.021   <2e-16 ***
## nutrition   -0.02592    0.02084  -1.244    0.216
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.735 on 97 degrees of freedom
## Multiple R-squared:  0.0157,Adjusted R-squared:  0.005556
## F-statistic: 1.548 on 1 and 97 DF,  p-value: 0.2165
```

Main effects w/no intercept

equation: $\widehat{disease}_i = \beta_1 \cdot nutrition_i$

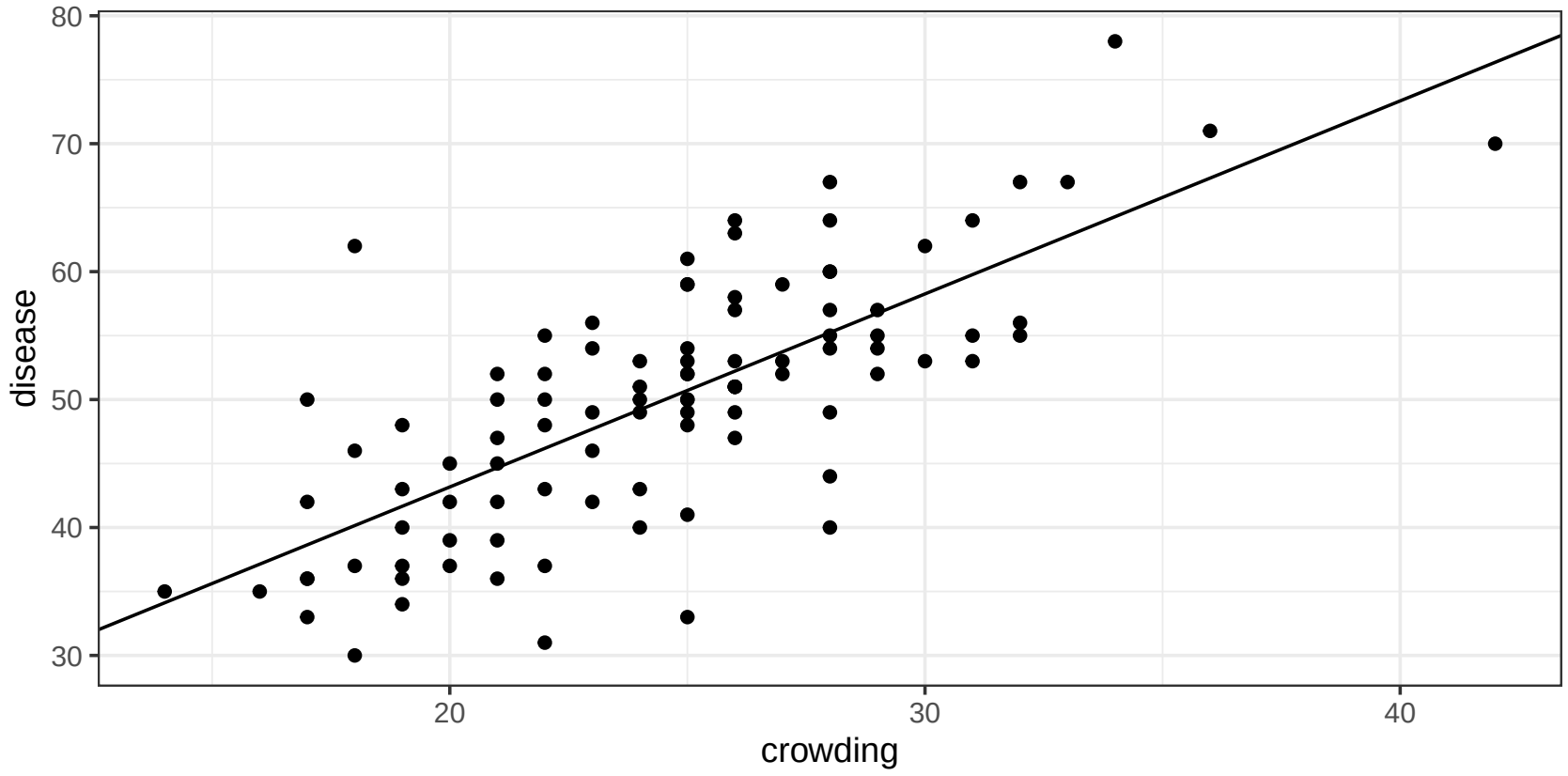
```
m1_no_intcpt <- lm(disease ~ nutrition - 1, data=dat)
ggplot(dat, aes(x=nutrition, y=disease)) +
  geom_point() + ylim(c(0,80)) +
  geom_abline(intercept = 0, slope = coef(m1_no_intcpt)[1])
```



Main effects model terms (crowding)

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

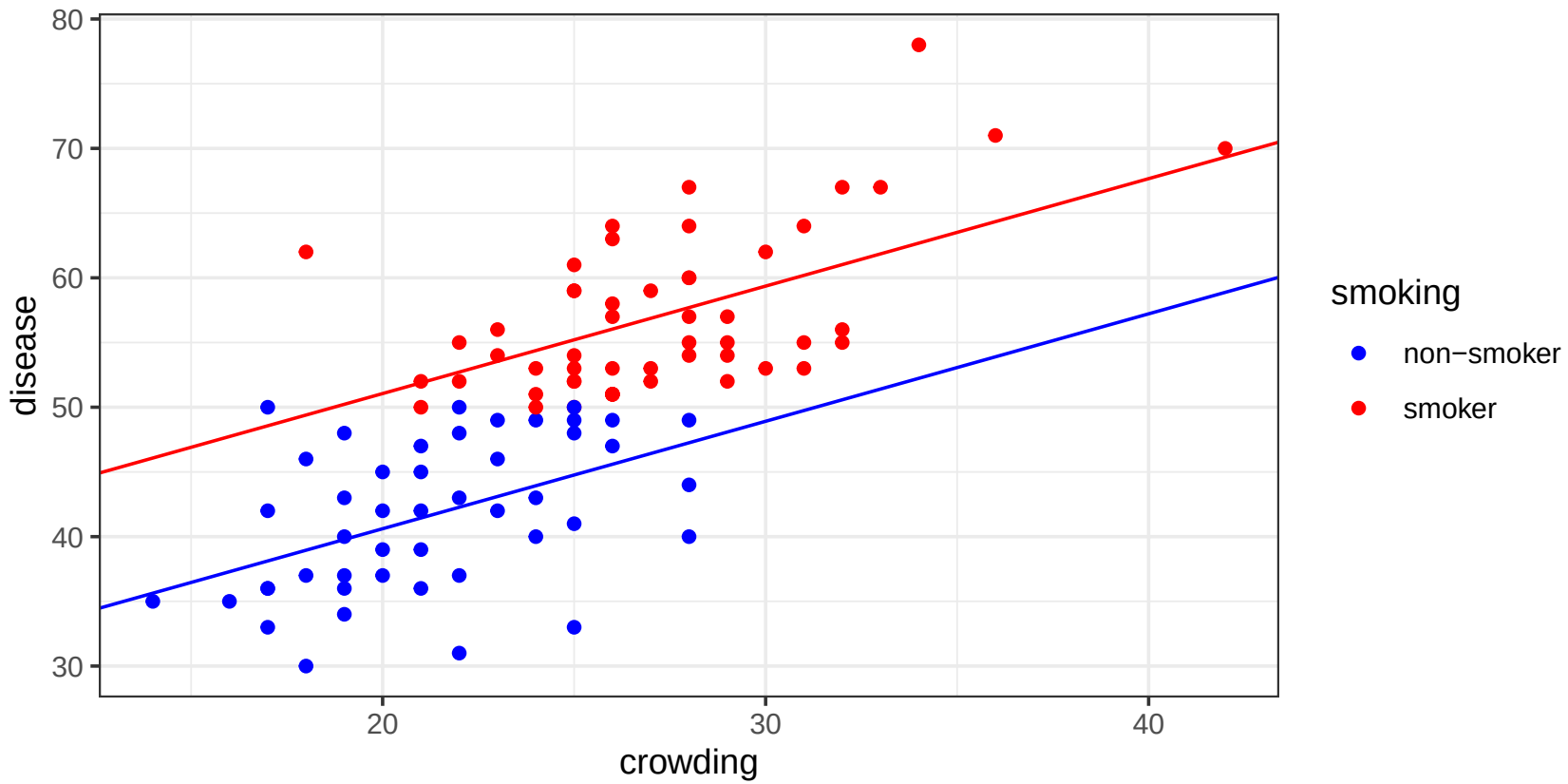
```
m1 <- lm(disease ~ crowding, data=dat)
ggplot(dat, aes(x=crowding, y=disease)) + geom_point() +
  geom_abline(intercept = coef(m1)[1], slope = coef(m1)[2])
```



2 main effects: 1 continous, 1 categorical

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

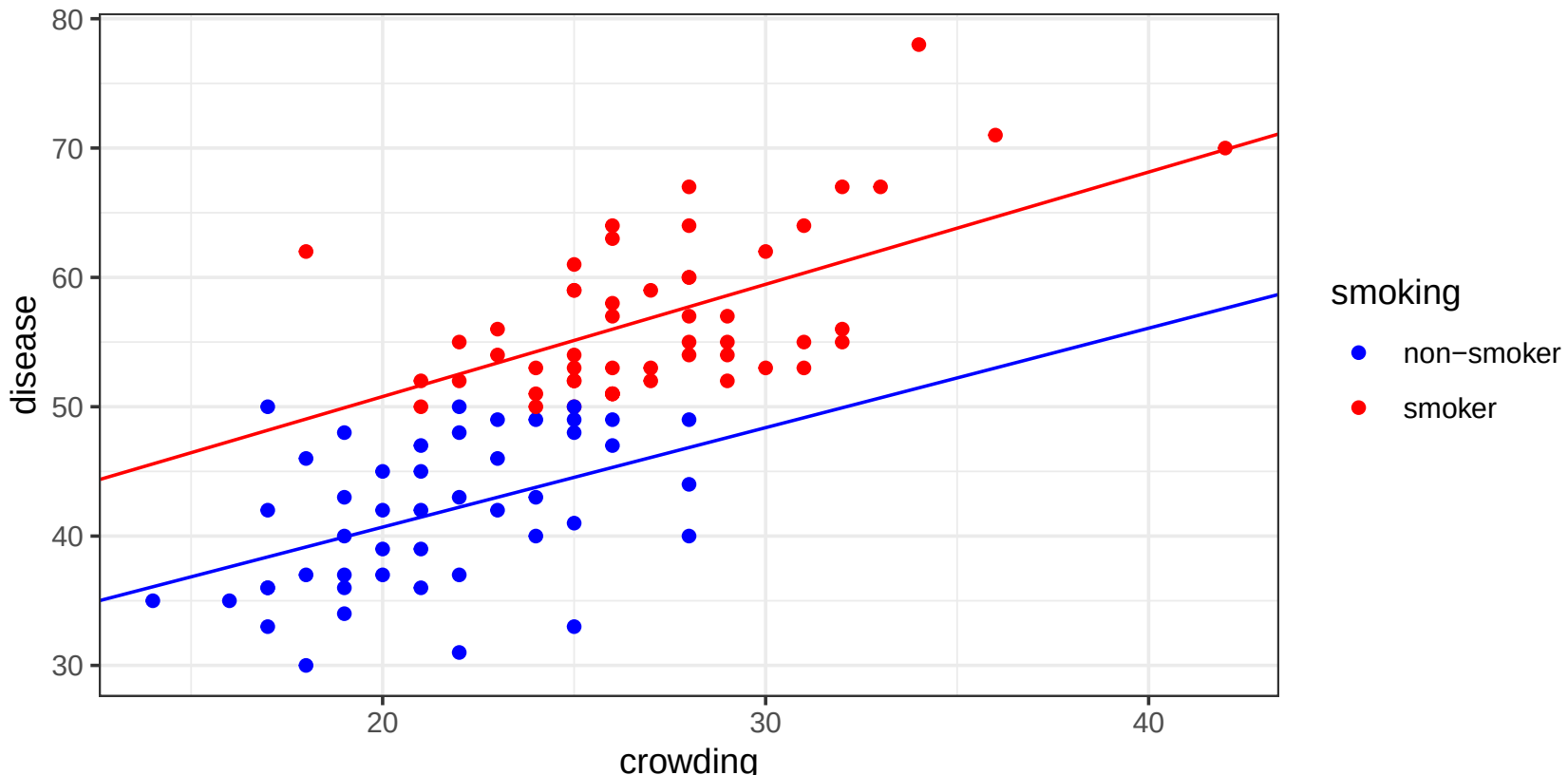
```
m2 <- lm(disease ~ crowding + smoking, data=dat)
ggplot(dat, aes(x=crowding, y=disease, color=smoking)) + geom_point() +
  scale_color_manual(values = c("blue", "red")) +
  geom_abline(intercept = coef(m2)[1], slope = coef(m2)[2], color="blue") +
  geom_abline(intercept = coef(m2)[1]+coef(m2)[3], slope = coef(m2)[2], color="
```



Interaction model terms

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

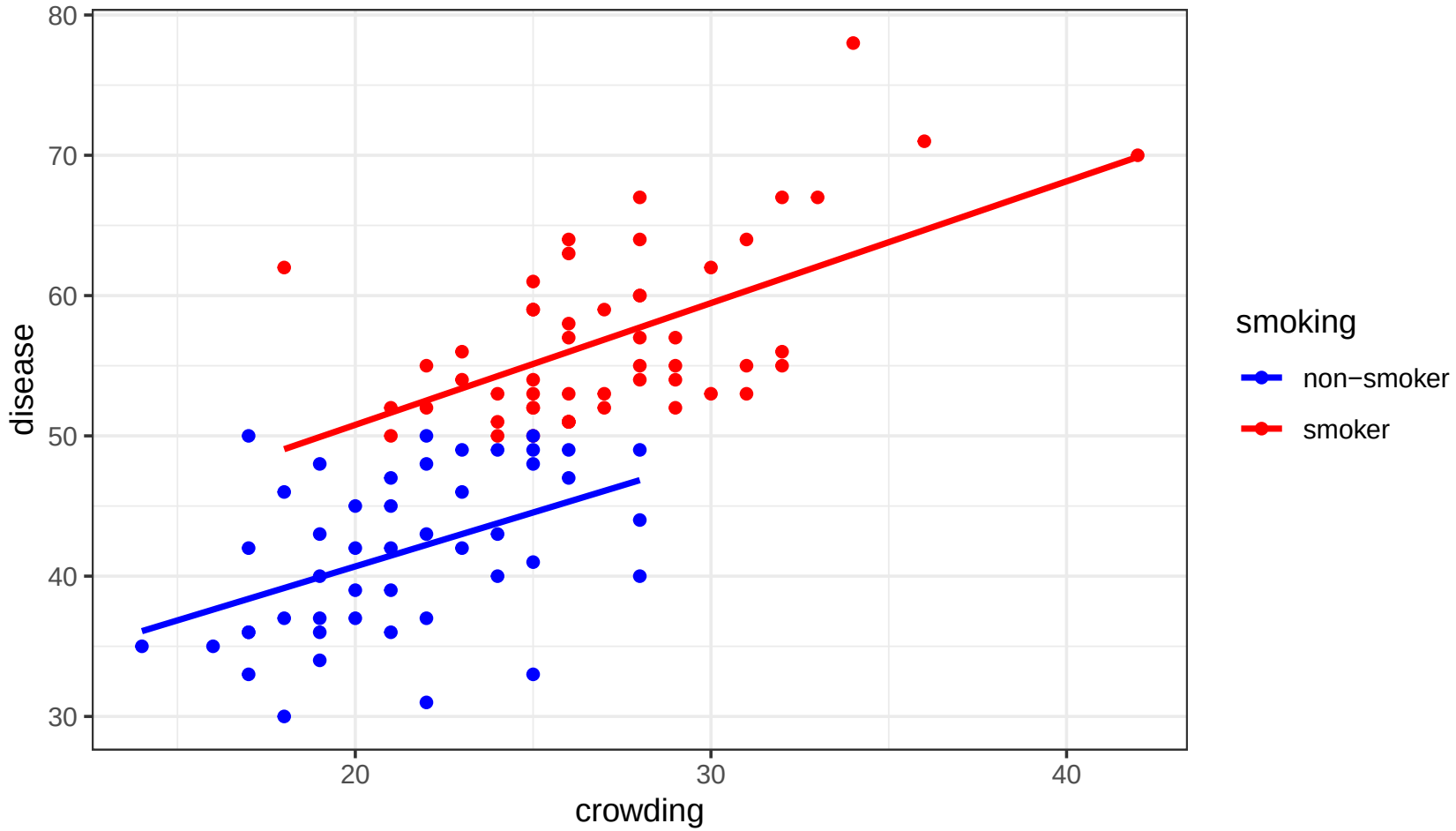
```
m3 <- coef(lm(disease ~ crowding*smoking, data=dat))
ggplot(dat, aes(x=crowding, y=disease, color=smoking)) +  geom_point() +
  scale_color_manual(values = c("blue", "red")) +
  geom_abline(intercept = m3[1], slope = m3[2], color="blue") +
  geom_abline(intercept = m3[1]+m3[3], slope = m3[2]+m3[4], color="red")
```



Interaction via geom_smooth()

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

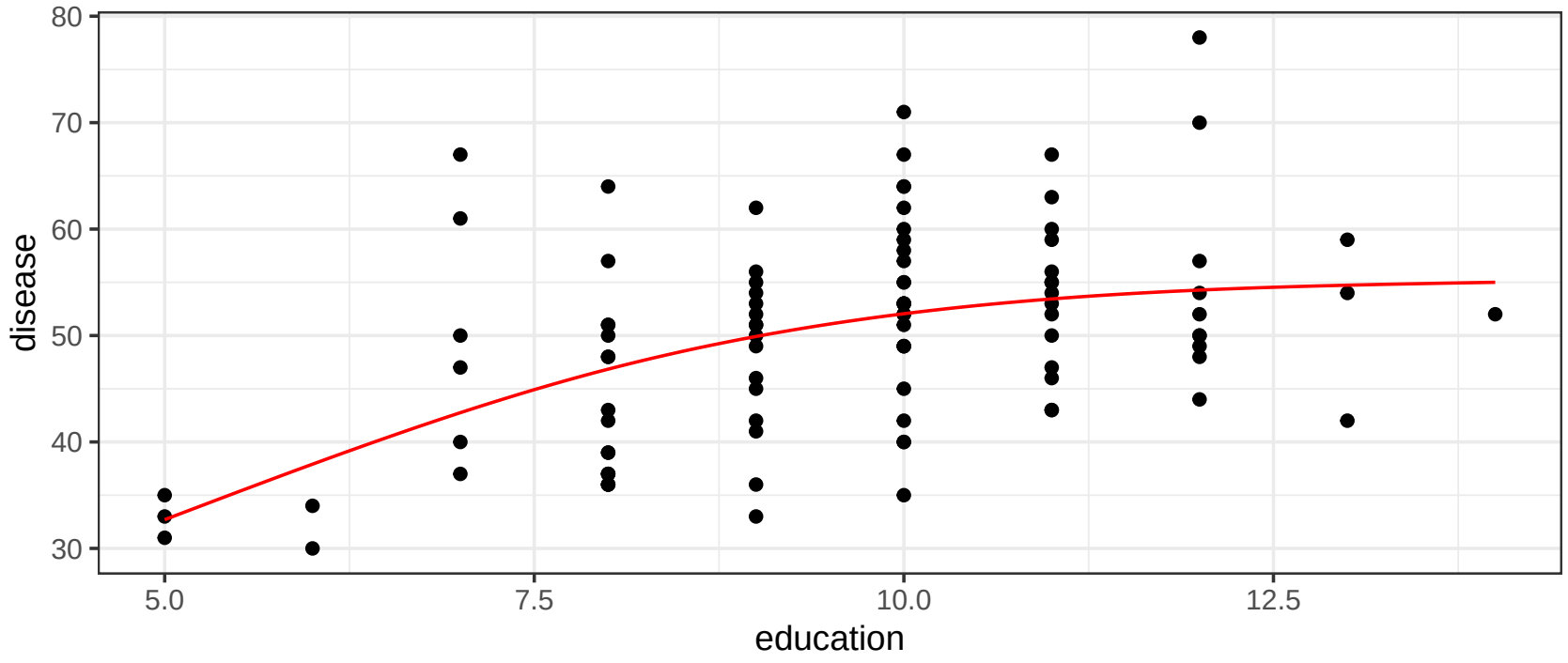
```
ggplot(dat, aes(x=crowding, y=disease, color=smoking)) +  geom_point() +
  geom_smooth(method="lm", se=FALSE) + scale_color_manual(values = c("blue", "r
```



Smooth model terms

equation: $\widehat{disease} = \beta_0 + s(education)$

```
library(splines)
m4 <- lm(disease ~ ns(education, knots = 8), data=dat)
x.new <- seq(min(dat$education), max(dat$education), by=.1)
yhat.m4 <- predict(m4, newdata = data.frame(education=x.new))
ggplot() + geom_point(data=dat, aes(x=education, y=disease)) +
    geom_path(aes(x=x.new, y=yhat.m4), color="red")
```



Fitting models in R

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.

Reading model output

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 <- lm(disease ~ crowding*smoking, data=dat)
round(summary(m3)$coef, 2)

##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 25.31       4.90    5.17     0.00
## crowding                     0.77       0.23    3.39     0.00
## smokingsmoker                8.11       6.93    1.17     0.24
## crowding:smokingsmoker       0.10       0.29    0.34     0.73
```

Breakout rooms

Work on, and share progress on your Lab 3 with other group-mates.