# Regression: Interactions and dummy variables

Author: Nicholas G Reich

# Outline

- Interpretation: main effects
- Interpretation: interactions

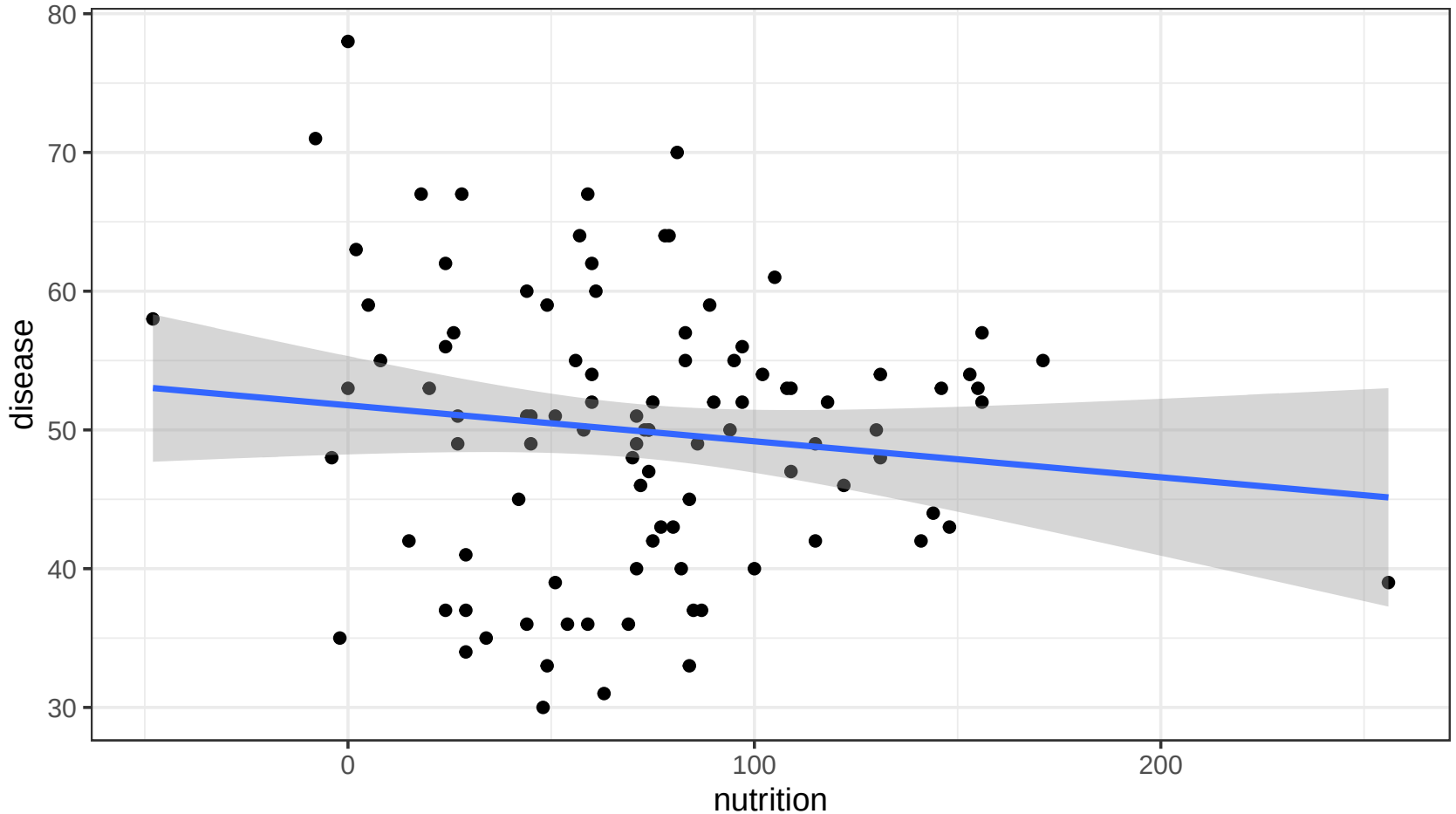
# Lung data example

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).
- `nutrition` nutritional status (larger number indicates better nutrition)
- `smoking` smoking status (1 if smoker, 0 if non-smoker)

# Nutrition's impact on lung disease

```
ggplot(dat, aes(nutrition, disease)) +
    geom_point() + geom_smooth(method="lm")
```

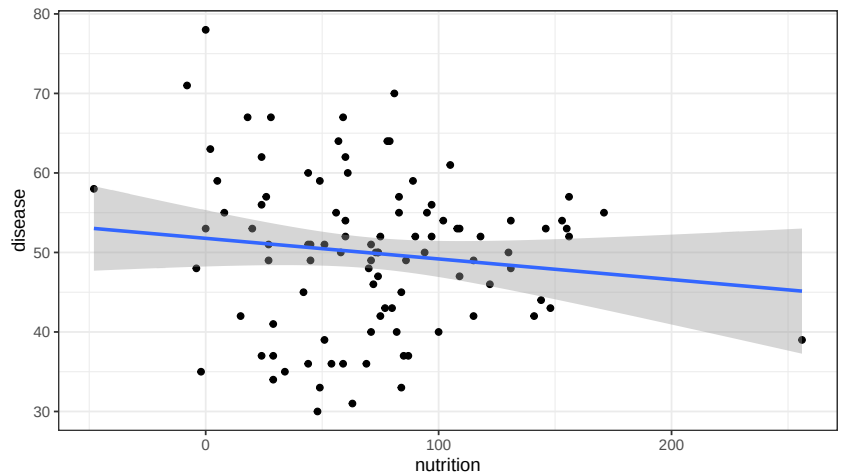


# A main effects model

Using the model $\text{disease}_i = \beta_0 + \beta_1 \text{nutrition}_i + \epsilon_i$, interpret

$\beta_0 =$

$\beta_1 =$



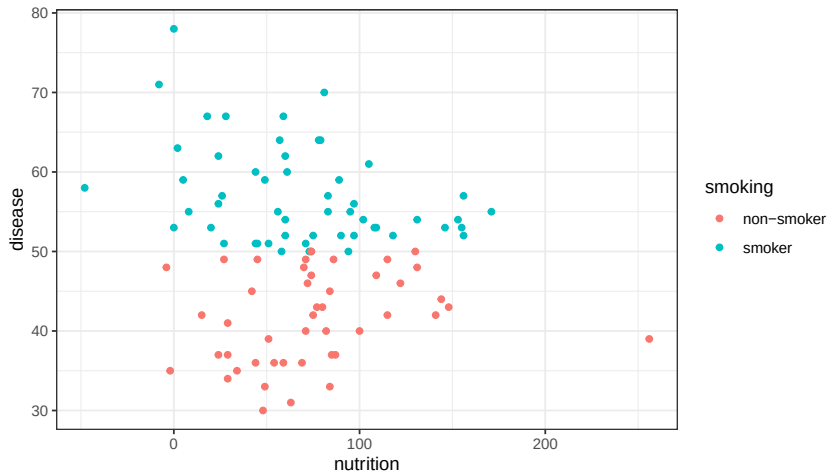
# A main effects model with two variables

Interpret from $\text{disease}_i = \beta_0 + \beta_1 \text{nutrition}_i + \beta_2 \text{smoking}_i + \epsilon_i$

$\beta_0 =$

$\beta_1 =$

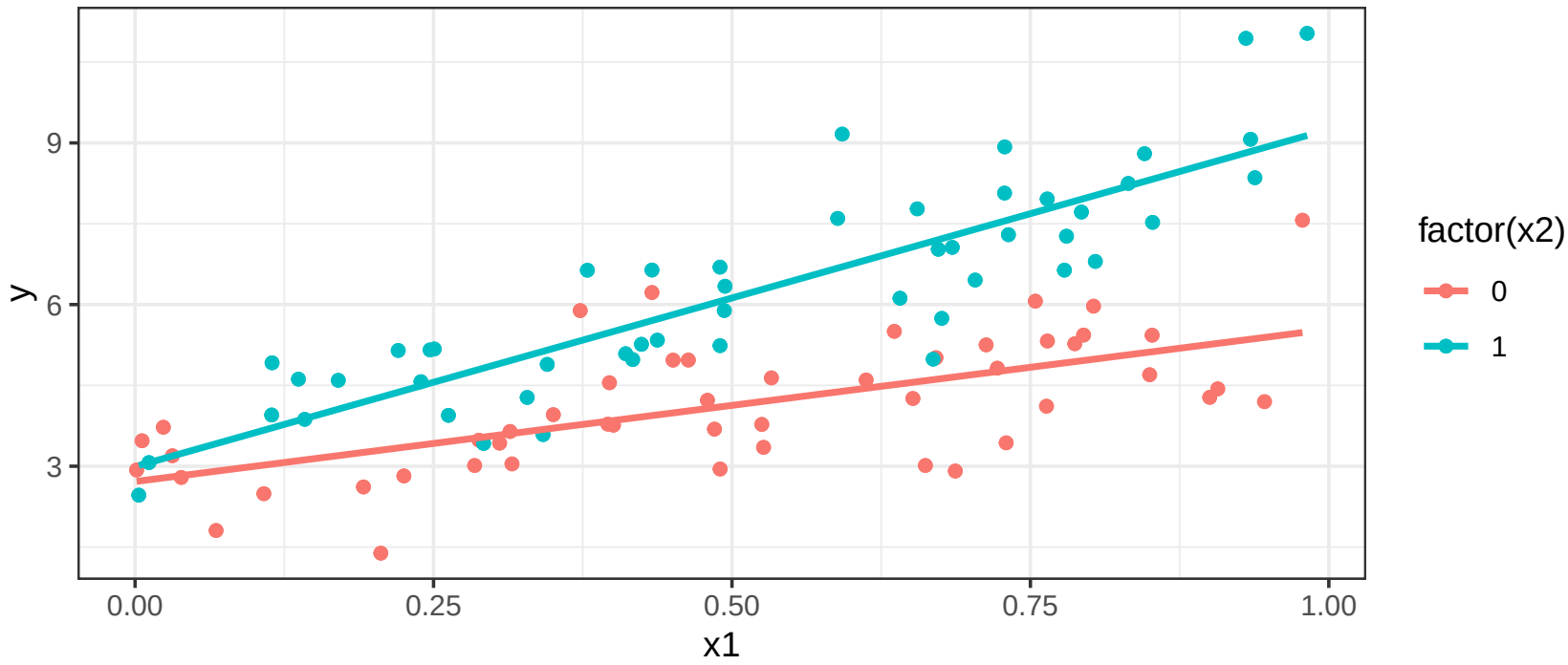
$\beta_2 =$



# What is interaction?

## Definition of interaction

Interaction occurs when the relationship between two variables depends on the value of a third variable.



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

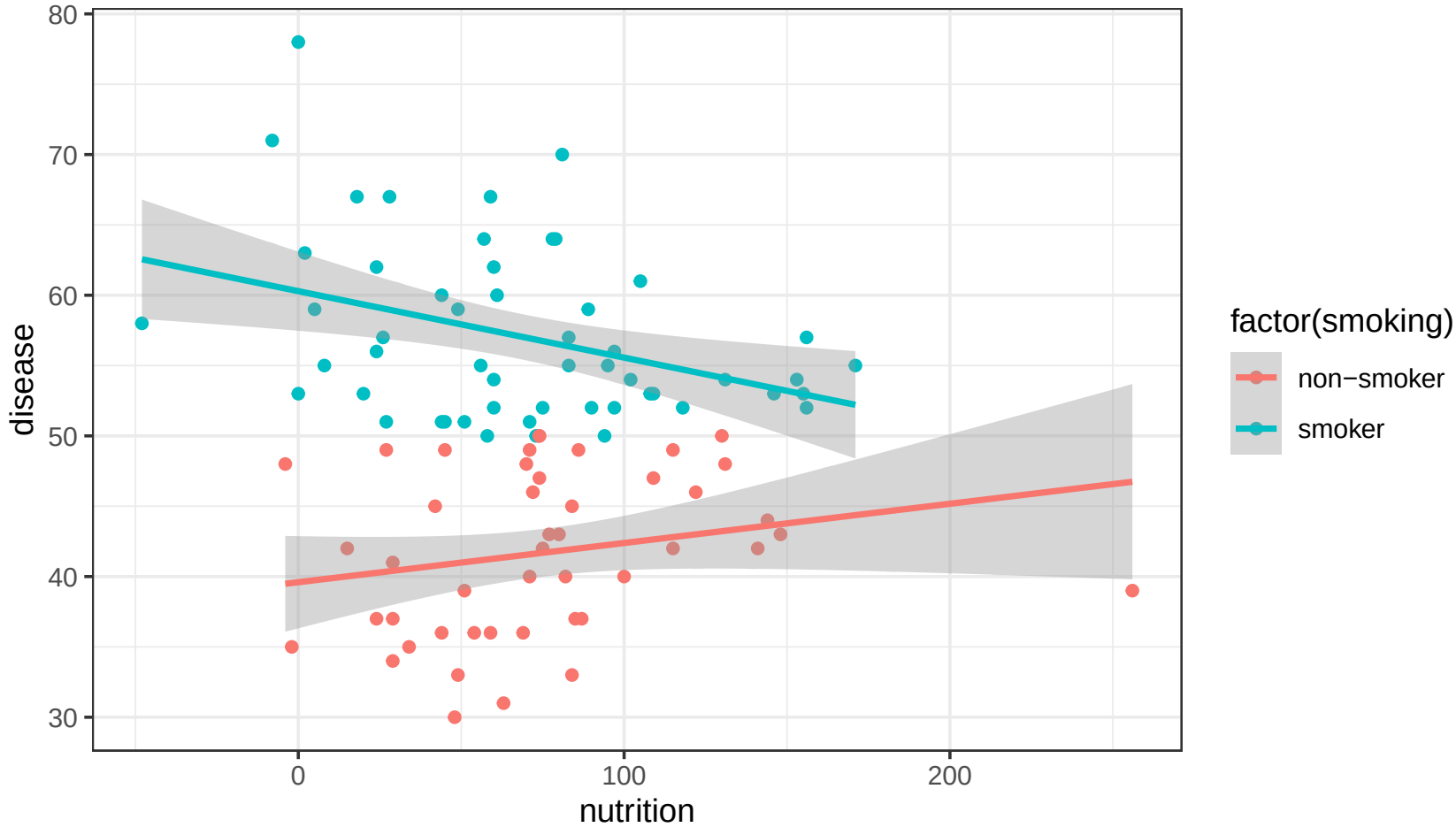
# How to interpret an interaction model

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

$\beta_3$ 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 $x_1 = 0$ .

# Example interaction model with lung data

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



# Example interaction model with lung data

$$dis_i = \beta_0 + \beta_1 nutrition_i + \beta_2 smoking_i + \beta_3 nutrition \cdot smoking_i + \epsilon_i$$

```
mi1 <- lm(disease ~ nutrition + smoking, data=dat)
mi2 <- lm(disease ~ nutrition*smoking, data=dat)
c(summary(mi1)$adj.r.squared, summary(mi2)$adj.r.squared)
```

```
## [1] 0.6190283 0.6483849
```

```
round(summary(mi2)$coef,2)
```

```
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 39.60       1.65   24.05     0.00
## nutrition                    0.03       0.02    1.49     0.14
## smokingsmoker               20.69       2.15    9.61     0.00
## nutrition:smokingsmoker     -0.08       0.03   -3.00     0.00
```

# Example interaction model with lung data

$$dis_i = \beta_0 + \beta_1 nutrition_i + \beta_2 smoking_i + \beta_3 nutrition \cdot smoking_i + \epsilon_i$$

```
mi1 <- lm(disease ~ nutrition + smoking, data=dat)
mi2 <- lm(disease ~ nutrition*smoking, data=dat)
c(summary(mi1)$adj.r.squared, summary(mi2)$adj.r.squared)


## [1] 0.6190283 0.6483849


round(summary(mi2)$coef,2)


##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 39.60       1.65   24.05     0.00
## nutrition                    0.03       0.02    1.49     0.14
## smokingsmoker               20.69       2.15    9.61     0.00
## nutrition:smokingsmoker     -0.08       0.03   -3.00     0.00
```

Among non-smokers, for every 1 unit improvement (increase) in nutrition score, the expected disease severity increases by 0.03 points. OR, among non-smokers, for every 100 unit improvement (increase) in nutrition score, the expected disease severity increases by 3 points. For smokers, for every 100 unit improvement (increase) in nutrition score, the expected disease severity would decrease by 5 points.

# Example interaction model with lung data

$$dis_i = \beta_0 + \beta_1 nut_i + \beta_2 smoking_i + \beta_3 nut_i \cdot smoking_i + \beta_4 crowd_i + \epsilon_i$$

```
mi3 <- lm(disease ~ nutrition*smoking + crowding, data=dat)
round(summary(mi3)$coef,2)
```

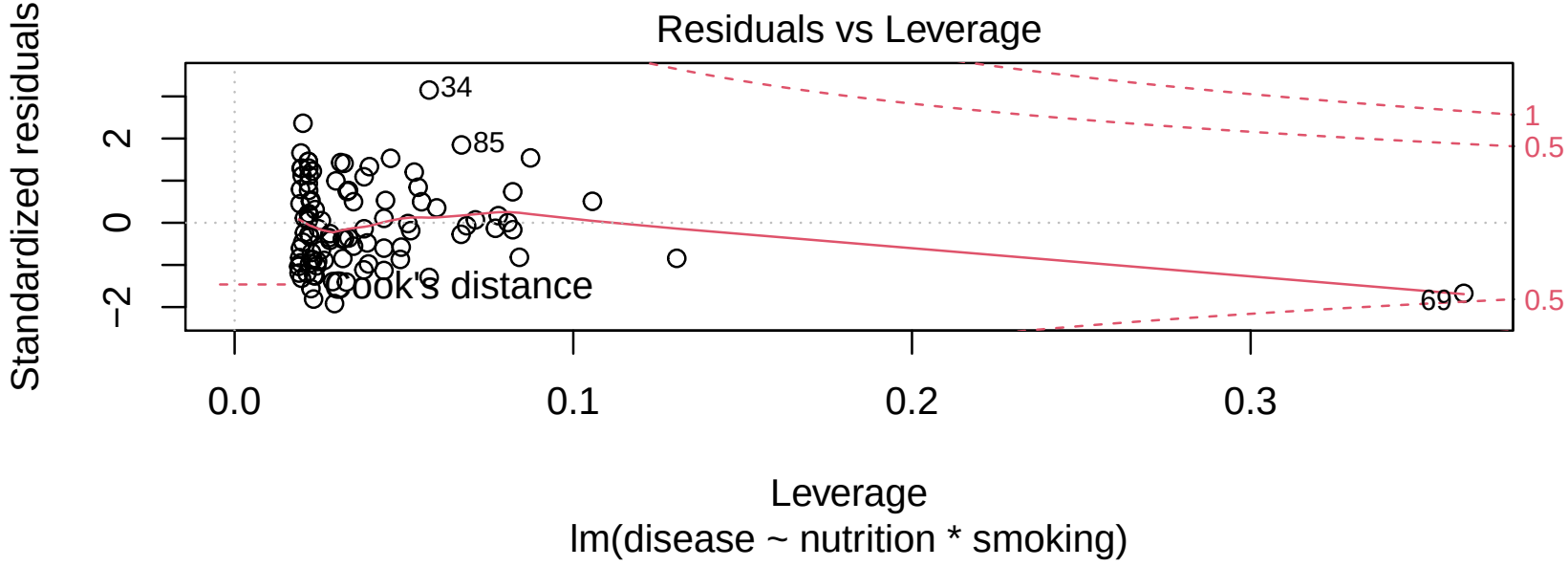
```
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 22.90       3.02    7.59     0.00
## nutrition                    0.02       0.02    0.95     0.34
## smokingsmoker               15.02       2.03    7.38     0.00
## crowding                     0.83       0.13    6.24     0.00
## nutrition:smokingsmoker     -0.07       0.02   -3.07     0.00
```

Adjusting for level of crowding (or, "holding crowding constant"), for every 100 unit improvement (increase) in nutrition score in smokers, the expected disease severity would decrease by 5 points.

# Checking influential points

We note that these results are sensitive to the inclusion of an influential outlying observation which had a much higher value of nutrition than any other observation.

```
plot(mi2, which=5)
```




```
dat[69,]
```

```
##    disease education crowding airqual nutrition    smoking
## 69      39         8       20      54       256 non-smoker
```

# Results sensitivity to outlier

It is important to experiment and note the sensitivity to the outlier, but unless you have REALLY good reason to do so, you should in general NOT remove outlying points from your primary analysis.

```
round(summary(mi2)$coef, 2)
```

```
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  39.60       1.65   24.05     0.00
## nutrition                     0.03       0.02    1.49     0.14
## smokingsmoker                20.69       2.15    9.61     0.00
## nutrition:smokingsmoker      -0.08       0.03   -3.00     0.00
```

```
mi2a <- lm(disease ~ nutrition*smoking, data=dat, subset=-69)
round(summary(mi2a)$coef, 2)
```

```
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  38.13       1.85   20.66     0.00
## nutrition                     0.05       0.02    2.21     0.03
## smokingsmoker                22.15       2.30    9.63     0.00
## nutrition:smokingsmoker      -0.10       0.03   -3.47     0.00
```

# Interaction modeling summary

- Interactions can give you a more detailed story about your data.
- They are 'easier' to interpret/visualize with a binary and continuous variable interaction.
- They are also valid for continuous x continuous variables: as the value of variable $A$ increases, the association between $B$ and $Y$ changes.
- Interaction is sometimes referred to as 'effect modification'.