A tale of two variables

INTRODUCTION TO REGRESSION IN R

Richie Cotton
Learning Solutions Architect at DataCamp

Swedish motor insurance data

- Each row represents one geographic region in Sweden.
- There are 63 rows.

n_claims	total_payment_sek
108	392.5
19	46.2
13	15.7
124	422.2
40	119.4
...	...

1 http://college.hmco.com/mathematics/brase/understandable_statistics/7e/students/datasets/slr/frames/frame.l

Descriptive statistics

```r
library(dplyr)
swedish_motor_insurance %>%
  summarize_all(mean)
```

```
# A tibble: 1 x 2
  n_claims total_payment_sek
     <dbl>             <dbl>
1     22.9              98.2
```

```r
swedish_motor_insurance %>%
  summarize(
    correlation = cor(n_claims, total_payment_sek)
  )
```

```
# A tibble: 1 x 1
  correlation
        <dbl>
1       0.881
```

What is regression?

- Statistical models to explore the relationship a response variable and some explanatory variables.
- Given values of explanatory variables, you can predict the values of the response variable.

n_claims	total_payment_sek
108	392.5
19	46.2
13	15.7
124	422.2
40	119.4
200	**???**

Jargon

Response variable (a.k.a. dependent variable)

The variable that you want to predict.

Explanatory variables (a.k.a. independent variables)

The variables that explain how the response variable will change.

Linear regression and logistic regression

Linear regression

- The response variable is numeric.

Logistic regression

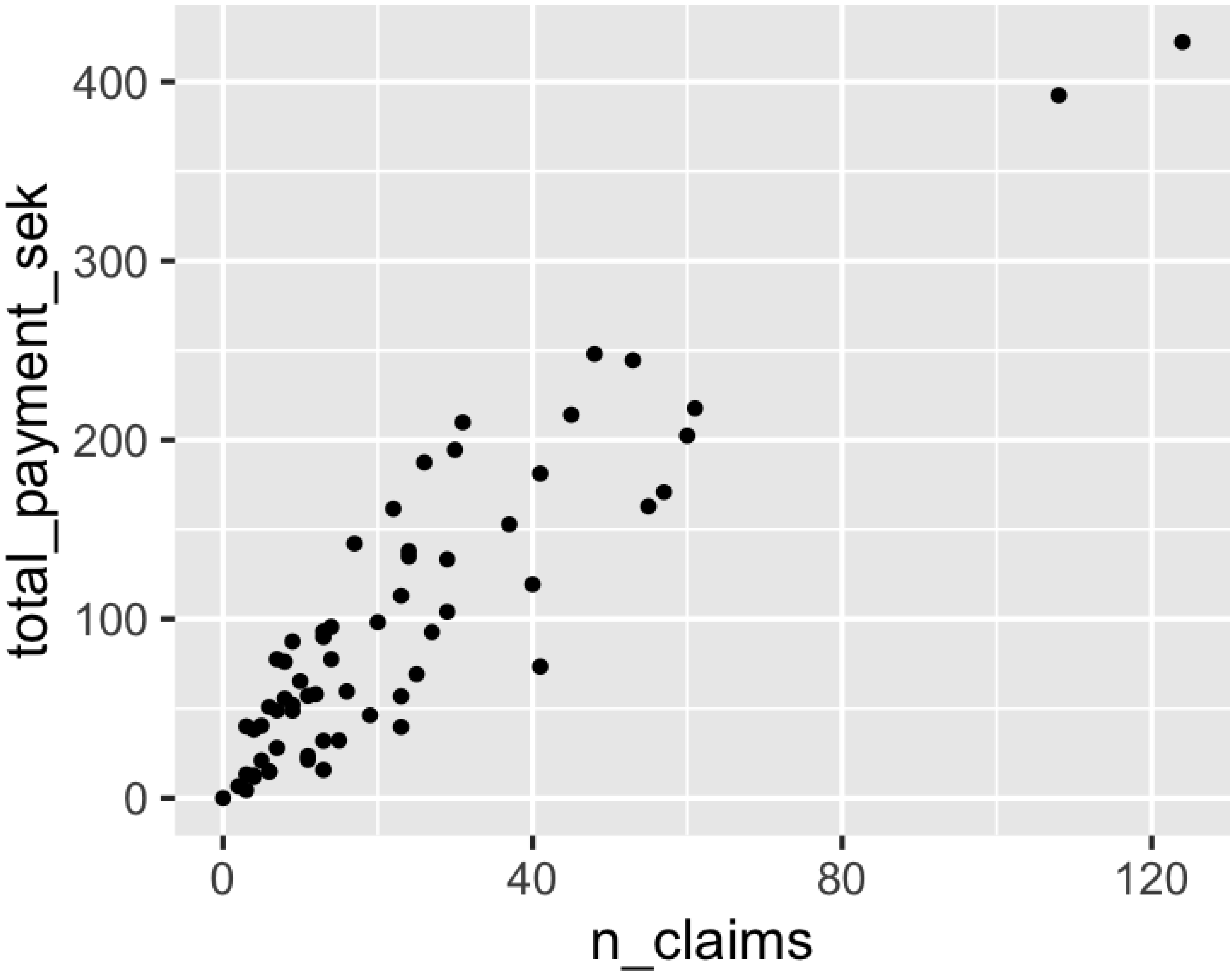
- The response variable is logical.

Simple linear/logistic regression

- There is only one explanatory variable.

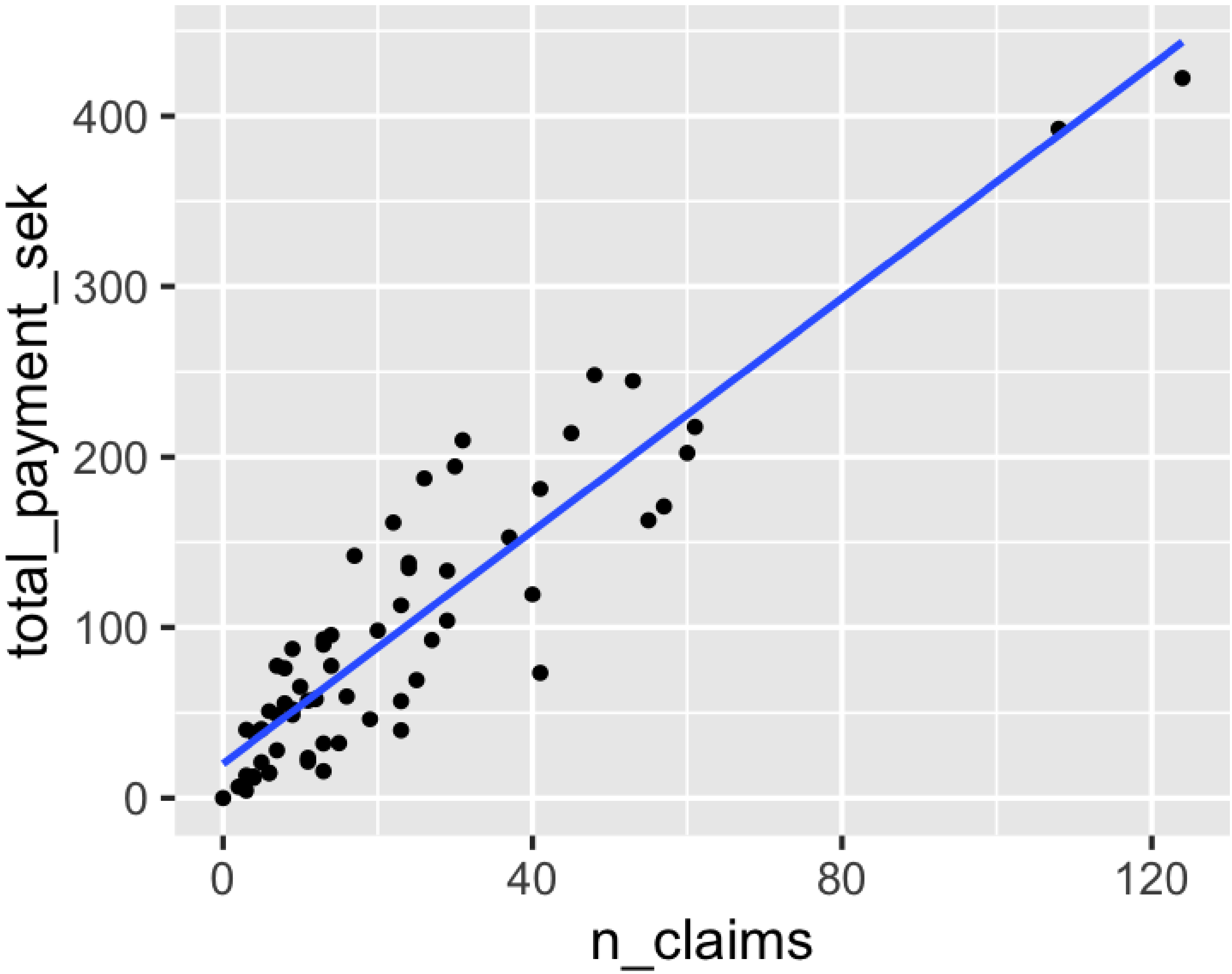
Visualizing pairs of variables

```r
library(ggplot2)

ggplot(
  swedish_motor_insurance,
  aes(n_claims, total_payment_sek)
) +
  geom_point()
```



Adding a linear trend line

```r
library(ggplot2)

ggplot(
  swedish_motor_insurance,
  aes(n_claims, total_payment_sek)
) +
  geom_point() +
  geom_smooth(
    method = "lm",
    se = FALSE
  )
```



Course flow

Chapter 1

Visualizing and fitting linear regression models.

Chapter 2

Making predictions from linear regression models and understanding model coefficients.

Chapter 3

Assessing the quality of the linear regression model.

Chapter 4

Same again, but with logistic regression models

Let's practice!

Fitting a linear regression

INTRODUCTION TO REGRESSION IN R

Richie Cotton
Learning Solutions Architect at DataCamp

Straight lines are defined by two things

Intercept

The y value at the point when x is zero.

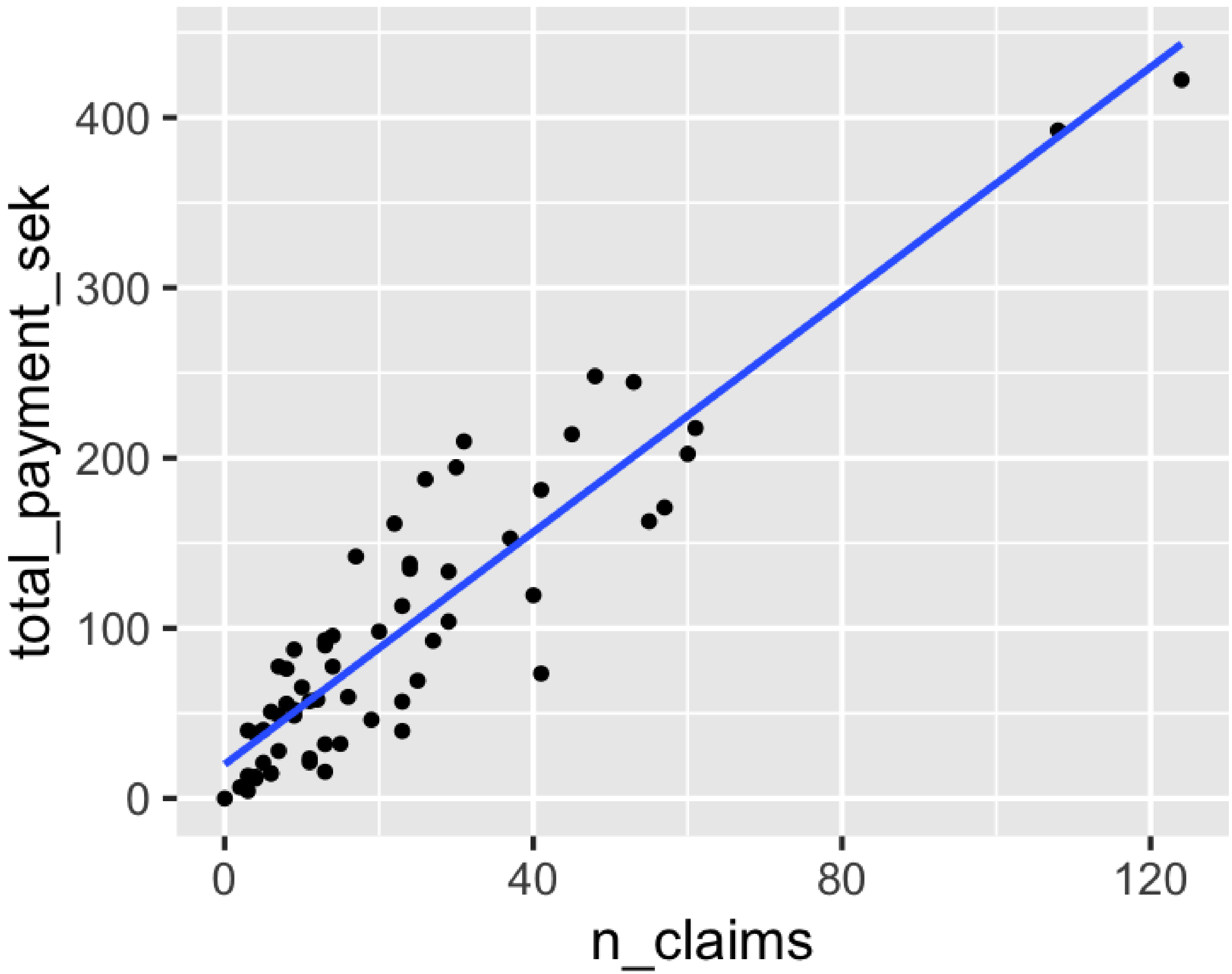
Slope

The amount the y value increases if you increase x by one.

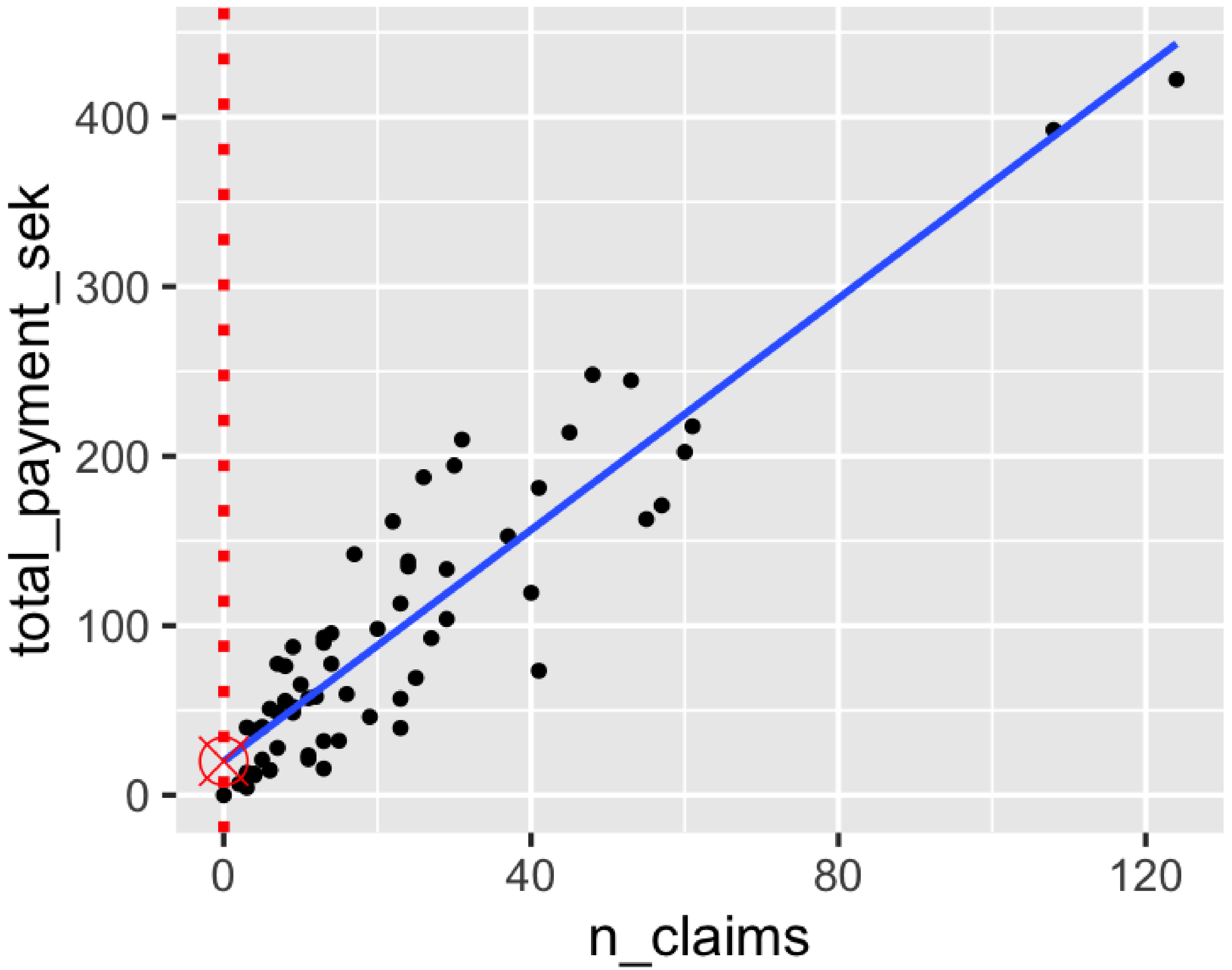
Equation

$y = intercept + slope * x$

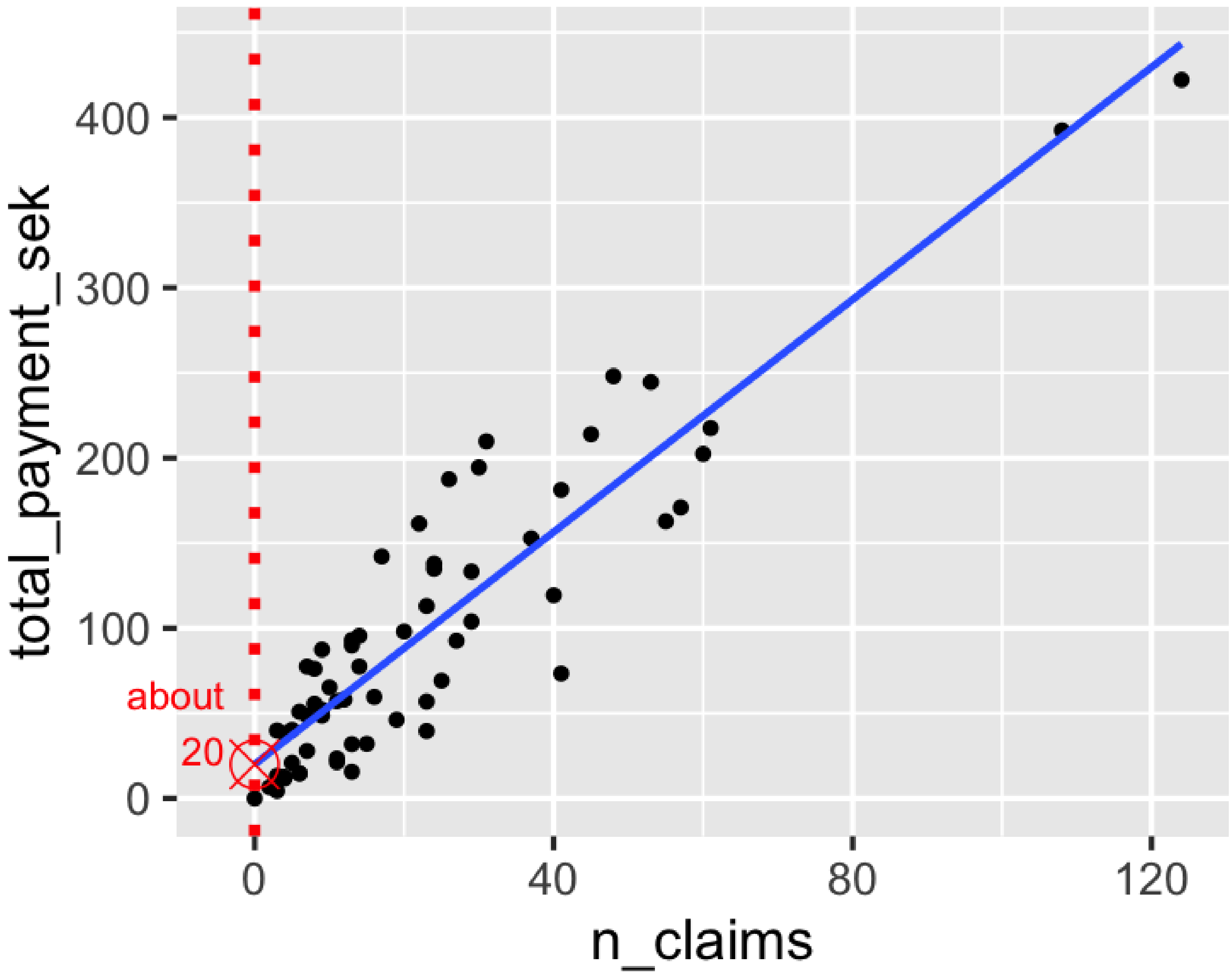
Estimating the intercept



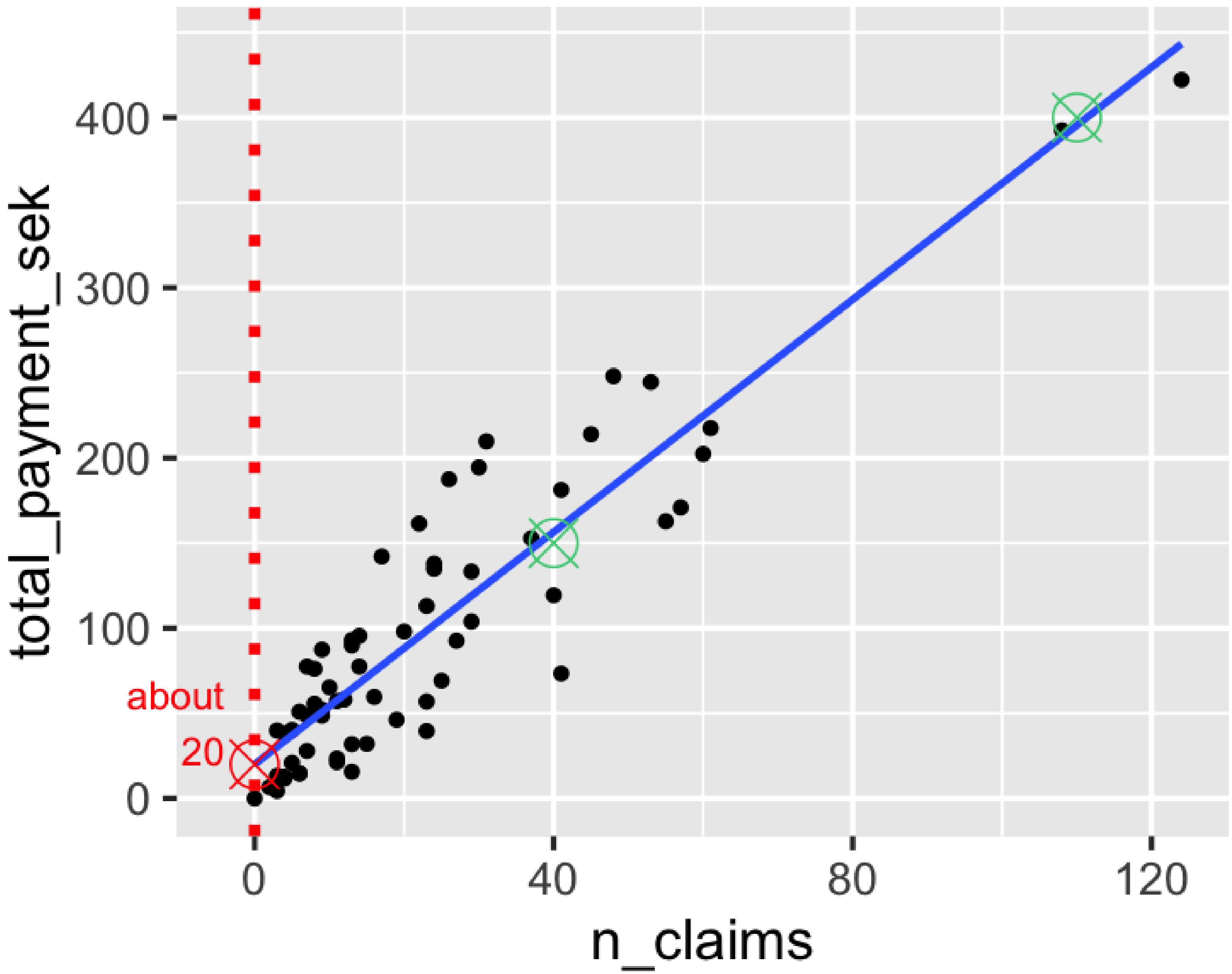
Estimating the intercept



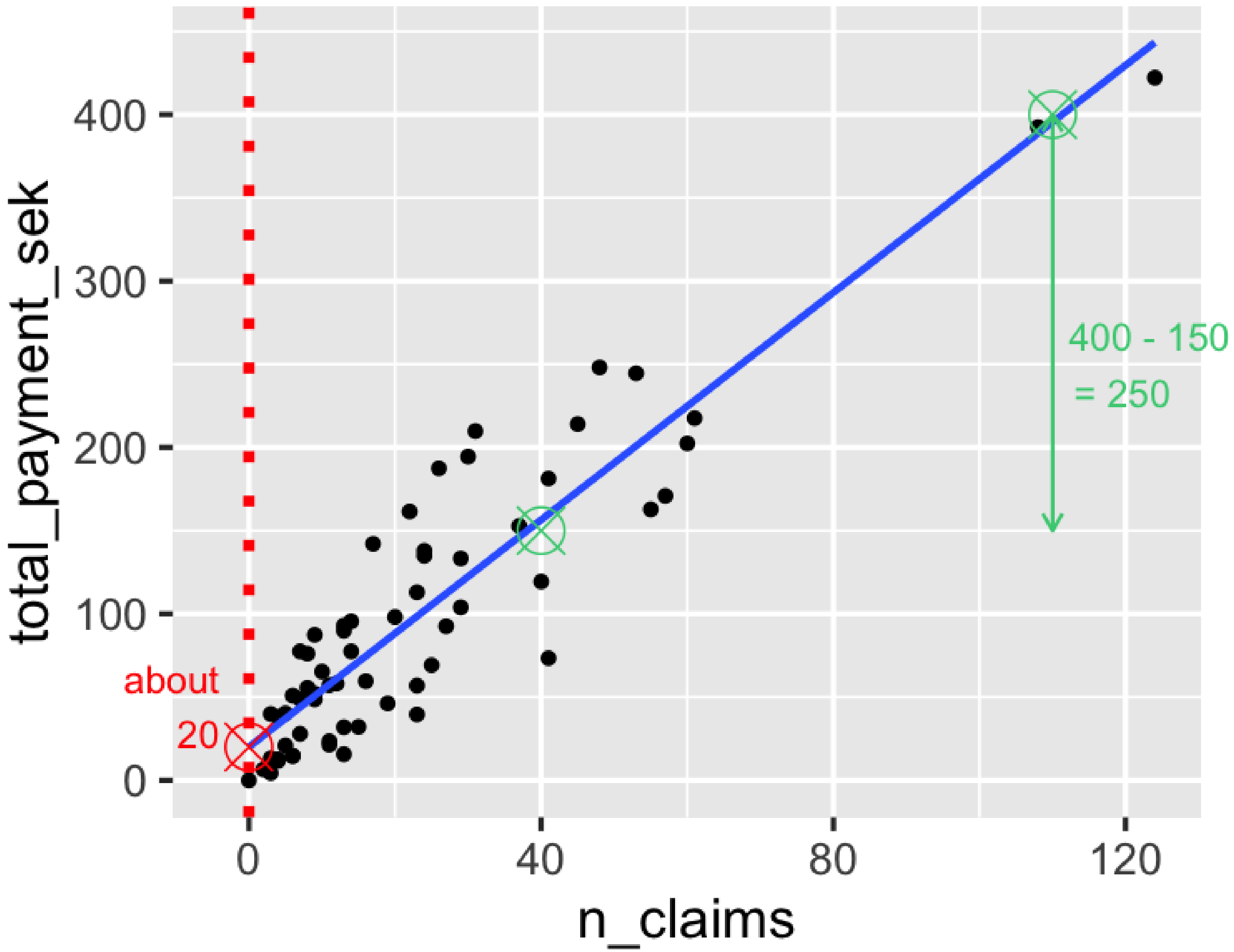
Estimating the intercept



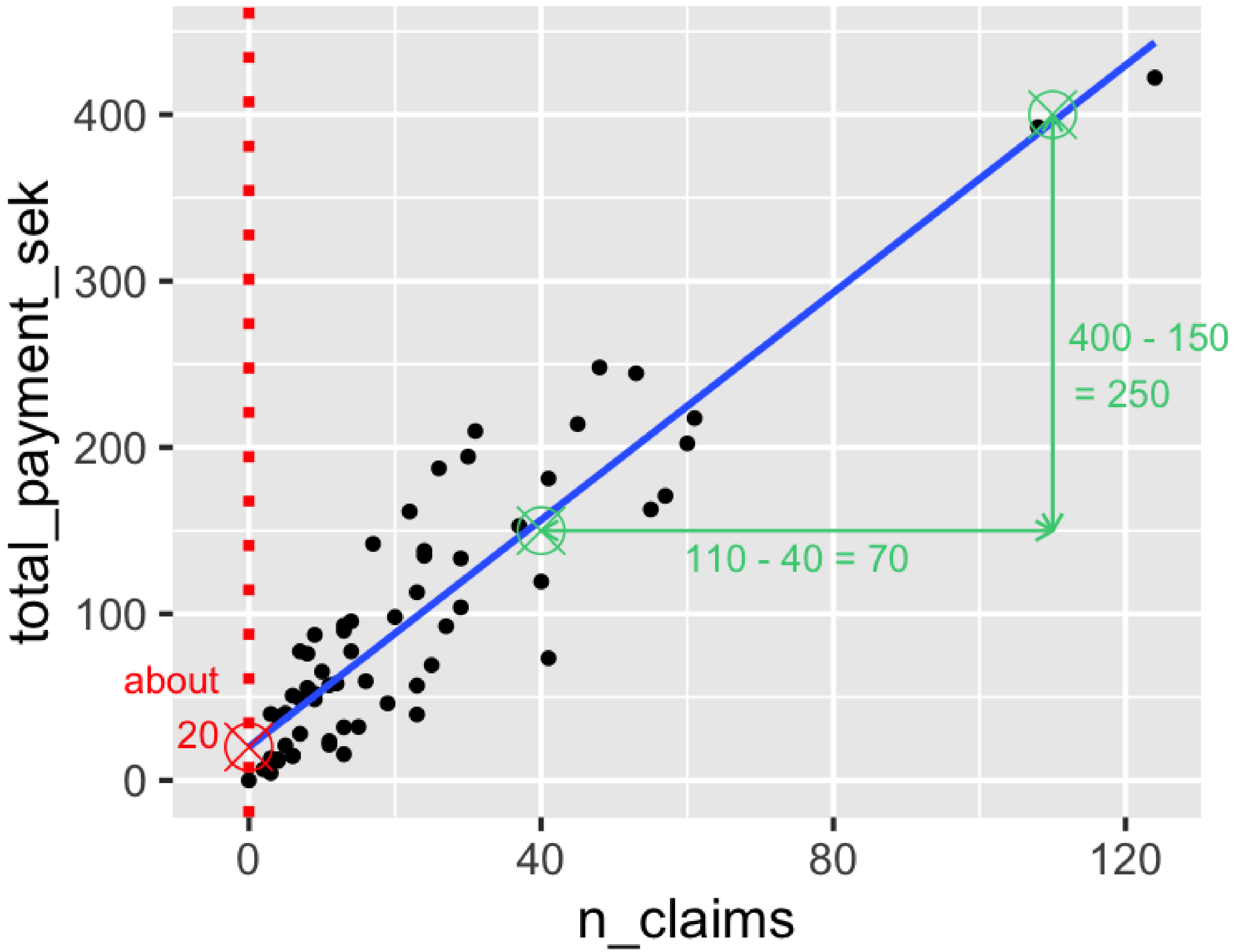
Estimating the slope



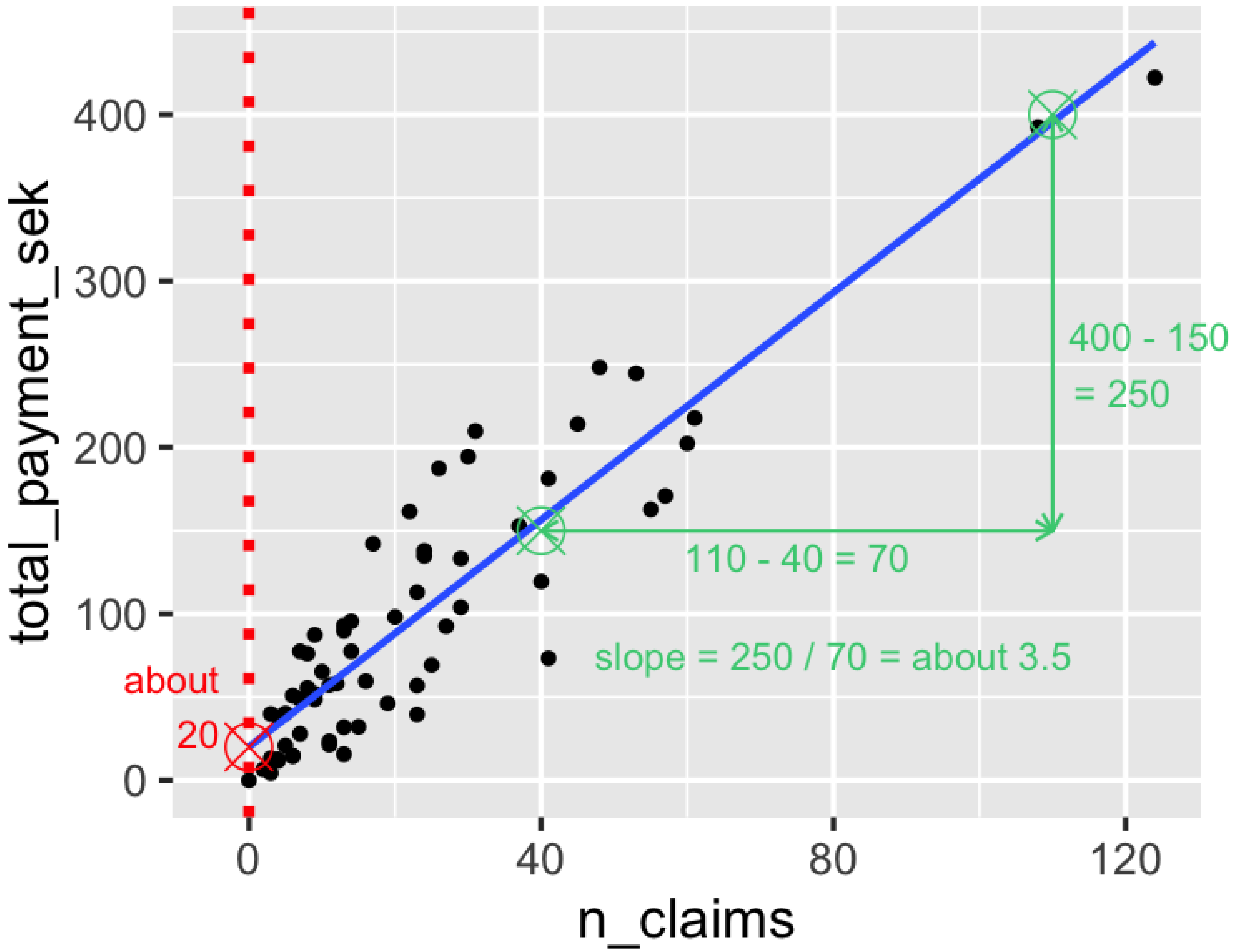
Estimating the slope



Estimating the slope



Estimating the slope



Running a model

```
lm(total_payment_sek ~ n_claims, data = swedish_motor_insurance)
```

```
Call:
lm(formula = total_payment_sek ~ n_claims, data = swedish_motor_insurance)

Coefficients:
(Intercept)     n_claims
     19.994        3.414
```

Interpreting the model coefficients

```
Call:
lm(formula = total_payment_sek ~ n_claims, data = swedish_motor_insurance)


Coefficients:
(Intercept)     n_claims
     19.994        3.414
```

Equation

$$total\_payment\_sek = 19.994 + 3.414 * n\_claims$$

Let's practice!

INTRODUCTION TO REGRESSION IN R

Categorical explanatory variables

INTRODUCTION TO REGRESSION IN R

Richie Cotton
Learning Solutions Architect at DataCamp

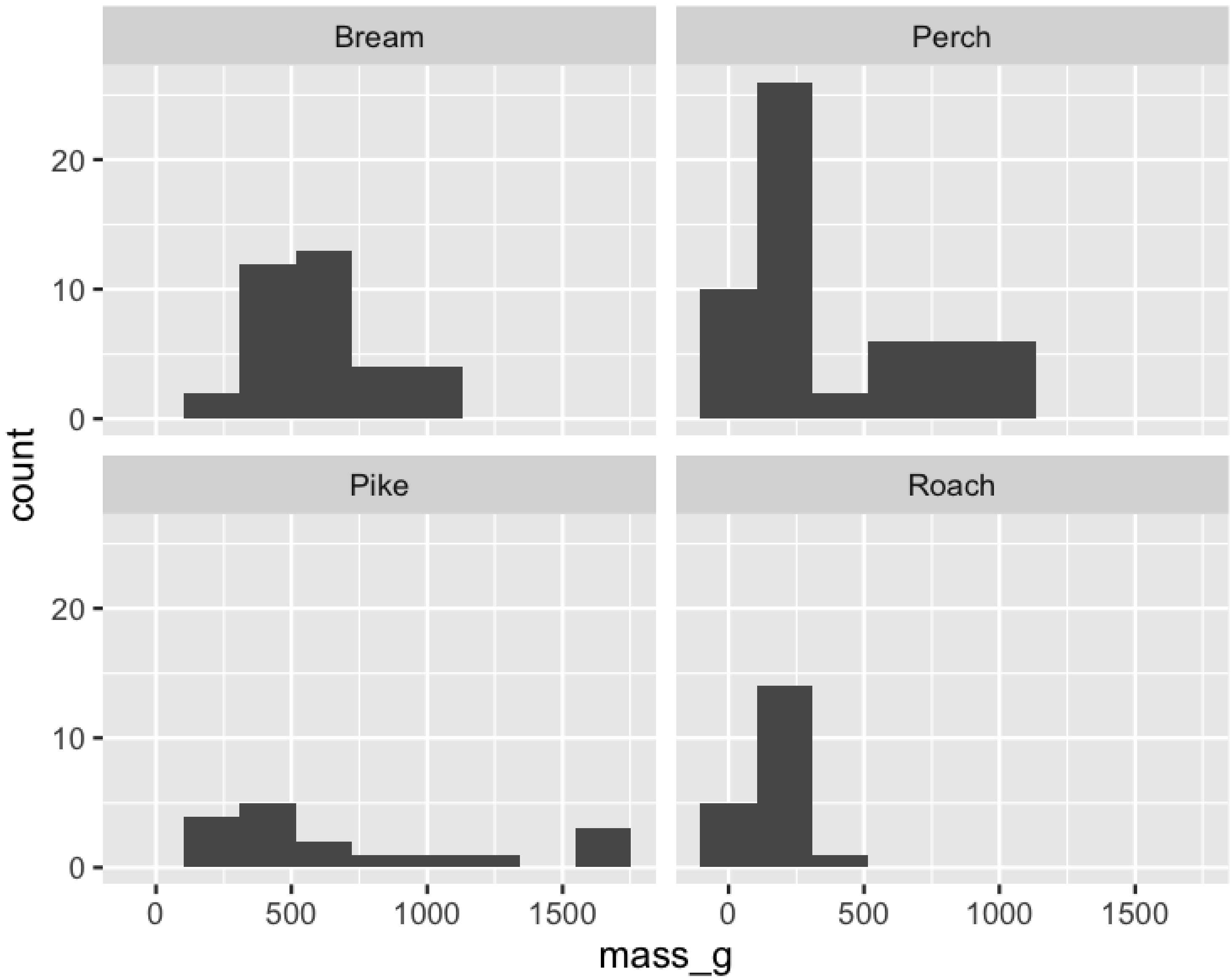
Fish dataset

- Each row represents one fish.
- There are 128 rows in the dataset.
- There are 4 species of fish.

species	mass_g
Bream	242.0
Perch	5.9
Pike	200.0
Roach	40.0
...	...

Visualizing 1 numeric and 1 categorical variable

```r
library(ggplot2)

ggplot(fish, aes(mass_g)) +
  geom_histogram(bins = 9) +
  facet_wrap(vars(species))
```



Summary statistics: mean mass by species

```
fish %>%
  group_by(species) %>%
  summarize(mean_mass_g = mean(mass_g))
```

```
# A tibble: 4 x 2
  species mean_mass_g
  <chr>         <dbl>
1 Bream          618.
2 Perch          382.
3 Pike           719.
4 Roach          152.
```

Linear regression

```
lm(mass_g ~ species, data = fish)
```

```
Call:
lm(formula = mass_g ~ species, data = fish)

Coefficients:
 (Intercept)  speciesPerch   speciesPike  speciesRoach
       617.8        -235.6         100.9        -465.8
```

No intercept

```
lm(mass_g ~ species + 0, data = fish)
```

```
Call:
lm(formula = mass_g ~ species + 0, data = fish)

Coefficients:
speciesBream  speciesPerch   speciesPike  speciesRoach
       617.8         382.2         718.7         152.0
```

Let's practice!

INTRODUCTION TO REGRESSION IN R