

Simple linear regression

Applied Data Science using R, Sessions 11 & 12

Prof. Dr. Claudius Gräbner-Radkowitz

Europa-University Flensburg, Department of Pluralist Economics

www.claudius-graebner.com | [@ClaudiusGraebner](https://twitter.com/ClaudiusGraebner) | claudius@claudius-graebner.com

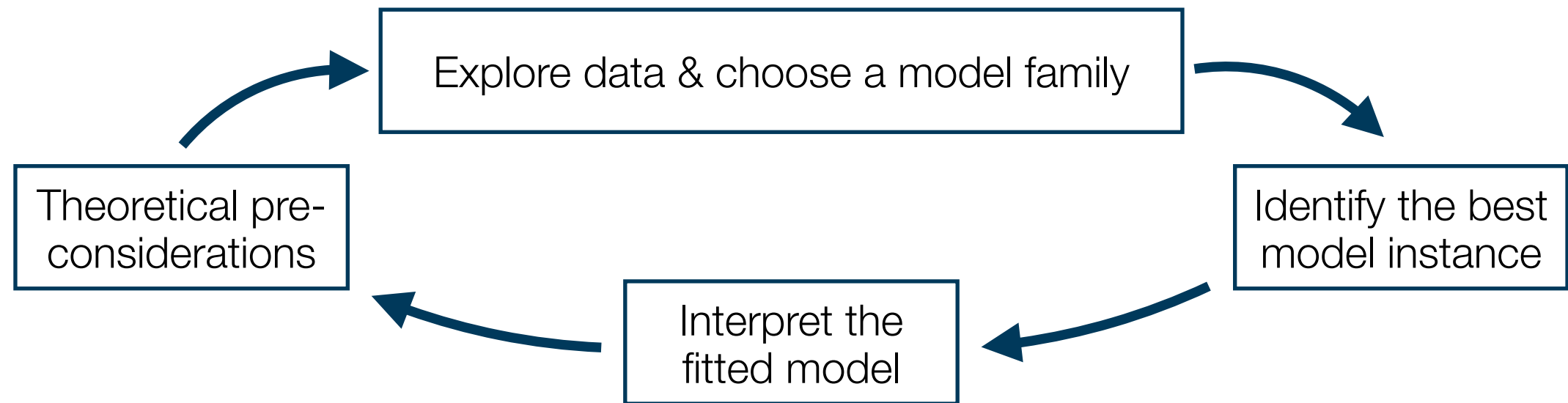
Goals for today

- I. Understand how the four steps of modelling data are operationalised within simple linear regression framework
- II. Understand the concept of ordinary least squares
- III. Learn how to conduct a simple lineare regression in R

Simple linear regression

Introduction

- In the previous session we learned about the four steps of modelling:



- In this session, we will go through these four steps for the modelling technique of **simple linear regression**
 - It its multiple variant, it is among the most widespread modelling techniques
 - It belongs to the class of supervised machine learning
 - While it can be used for exploratory purposes, its main strength lies in explanatory analysis

Modelling data - general workflow

1. Theoretical pre-considerations

- During the theoretical pre-considerations you think about the goal of your modelling exercise
 - What is your subject of interest?
 - Do you want to engage in an exploratory or explanatory analysis?
 - If the latter, what are your main hypothesis? If the former, what is the goal of exploration?
 - What is the data you need and how was it collected?
- **Example:**
 - We are interested in what drives beer consumption
 - We first want to explore the survey data we obtained to derive hypotheses, which we then want to test

Modelling data - general workflow

2. Data exploration and choice of family

- Based on our theoretical considerations we need to obtain data
- Then we need to inspect the data and think about how it could be modelled
- Assume we have a data set with survey results on beer consumption
 - First need to take a `glimpse` at the data set:

```
> glimpse(beer_data)
```

```
Rows: 30
```

```
Columns: 5
```

```
$ consumption <dbl> 81.7, 56.9, 64.1, 65.4, 6...  
$ price       <dbl> 1.78, 2.27, 2.21, 2.15, 2...  
$ price_liquor <dbl> 6.95, 7.32, 6.96, 7.18, 7...  
$ price_other  <dbl> 1.11, 0.67, 0.83, 0.75, 1...  
$ income      <dbl> 25088, 26561, 25510, 2715...
```

- We have 30 observations of five variables, all of which are numeric
 - We should also have a look at common descriptive statistics

Modelling data - general workflow

2. Data exploration and choice of family

- The function `skimr::skim()` provides a nice statistical summary
 - We can complement this via some easy visualisations* (`geom_jitter()` and `geom_violin()`)

Data Summary

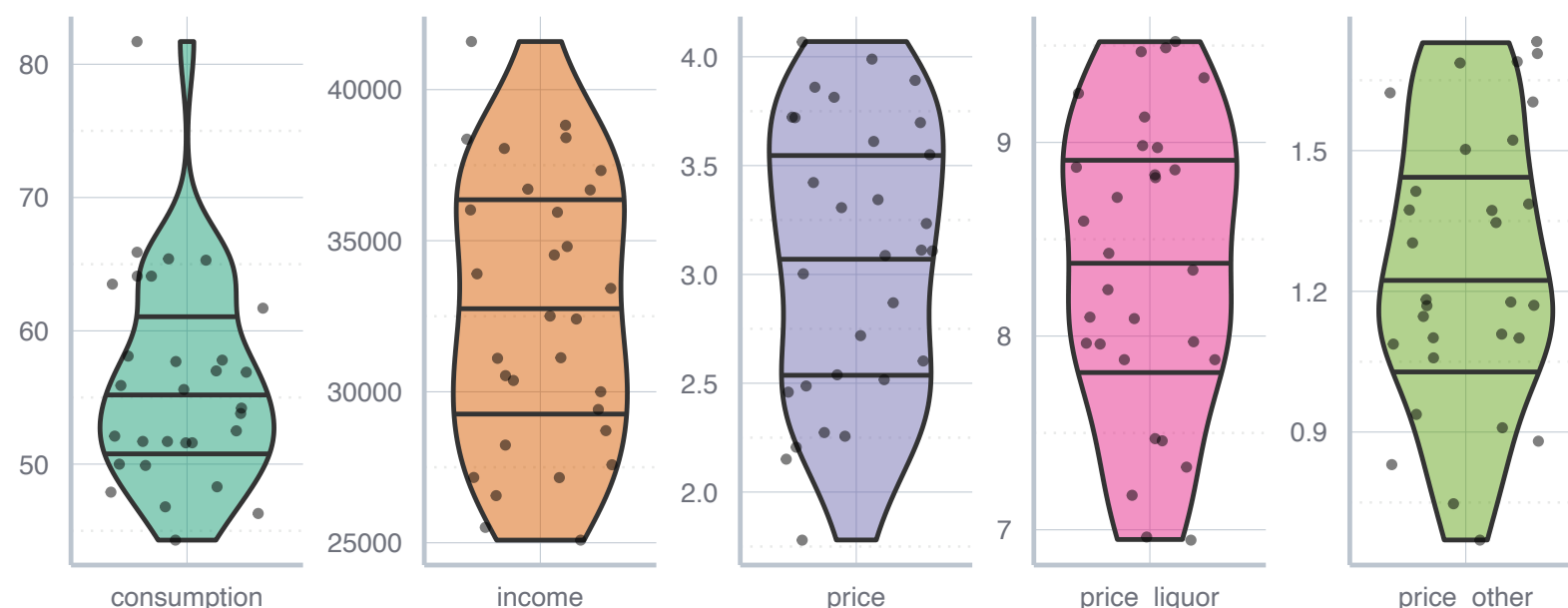
	Values
Name	beer_data
Number of rows	30
Number of columns	5

Column type frequency:	
numeric	5

Group variables	None

Variable type: numeric

	skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
1	consumption	0	1	56.1	7.86	44.3	51.6	54.9	60.8	81.7	
2	price	0	1	3.08	0.642	1.78	2.53	3.11	3.68	4.07	
3	price_liquor	0	1	8.37	0.770	6.95	7.9	8.38	8.94	9.52	
4	price_other	0	1	1.25	0.298	0.67	1.09	1.18	1.48	1.73	
5	income	0	1	32602.	4542.	25088	28888	32457	36516.	41593	



It seems feasible and interesting to look at the relationship between **consumption**, **price** and **income**

Modelling data - general workflow

2. Data exploration and choice of family

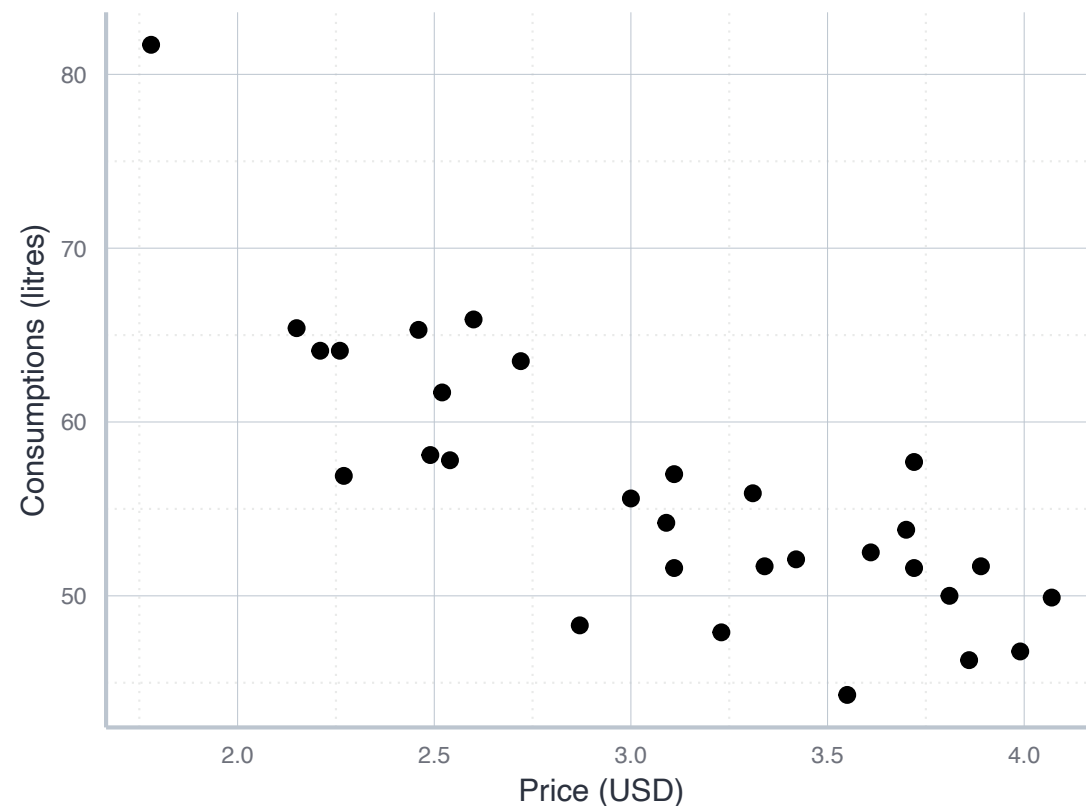
- It seems feasible and interesting to look at the relationship between **consumption**, **price** and **income**
 - Economic theory would suggest a close relationship between them
 - Consumption and price do correlate with each other:
 - ```
cor(
 x = beer_data$consumption,
 y = beer_data$price,
 method = "pearson"
)
```
- **x** and **y** give the vectors, **method** the kind of correlation coefficient
  - If you do not remember the different kind of correlation coefficients, please review
- Beware: correlation only means association or co-movement, it **does not imply causation**! We should look at the relationship in more detail!



# Modelling data - general workflow

## 2. Data exploration and choice of family

- To get more information and choose the right model family, it is always a good idea to **visualise** the data
  - Since both variables are numeric, we choose a scatter plot



- There seems to be a strong and **linear** relationship
- This suggests to choose the **family of linear models**
- It has the general form:

$$y = a + b \cdot x$$

# Modelling data - general workflow

## 2. Data exploration and choice of family

- The family of linear models has the general form  $y = a + b \cdot x$
- In the context of economic modelling, we use the following notation:

The diagram shows the linear model equation  $y = \beta_0 + \beta_1 x_1 + \epsilon$  with the following components and annotations:

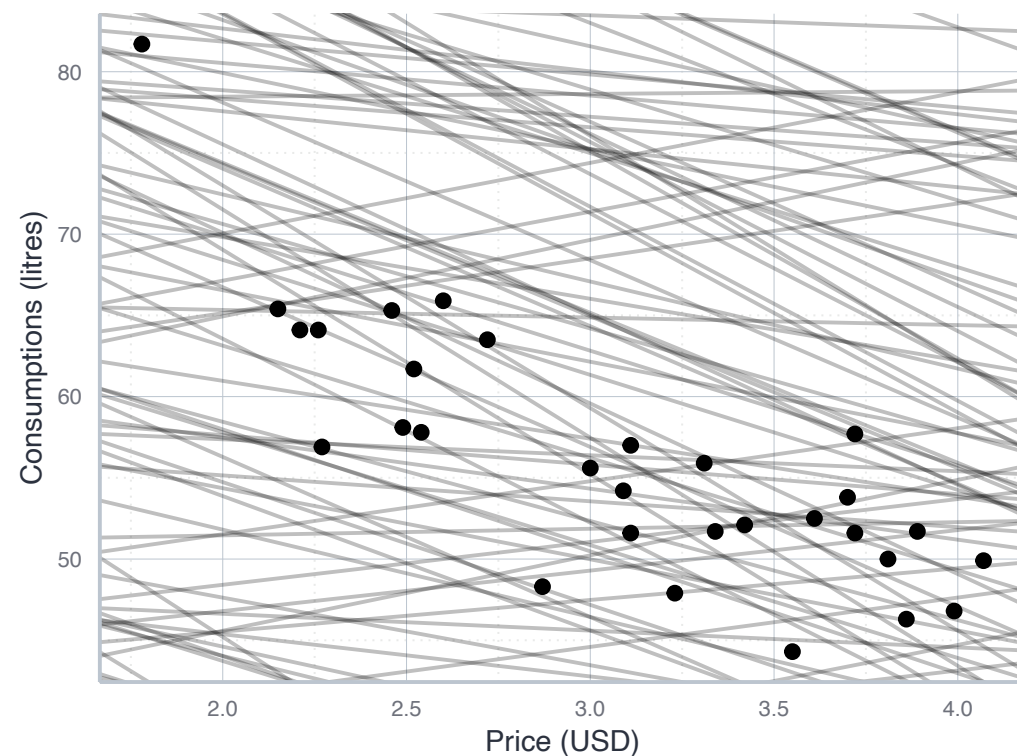
- Dependent variable** (Synonyms: response variable, regressand, explained variable, outcome variable): Points to the variable  $y$  in a light blue box.
- Parameters to be estimated**: Two orange curved arrows point from this text to the parameters  $\beta_0$  and  $\beta_1$ , which are in light orange boxes.
- Independent variable** (Synonyms: predictor, regressor, explanatory variable, input variable): A red arrow points from this text to the variable  $x_1$  in a light red box.
- Error term**: A green arrow points from this text to the error term  $\epsilon$  in a light green box.

- The **error term** absorbs all effects on  $y$  not covered by  $x \rightarrow$  unobservable & probabilistic
- Everything on the left side of the  $=$  is called the left-hand-side (**LHS**)
- Everything on the right side of the  $=$  is called the right-hand-side (**RHS**)

# Modelling data - general workflow

## 3. Fitting a model

- So far we have chosen a family of models:  $y = \beta_0 + \beta_1 \cdot x$ 
  - It posits a linear relationship between the **dependent variable**  $y$  and the **independent variable**  $x \rightarrow$  can be represented by a straight line
  - It has two parameters for which we need to choose particular values:  $\beta_0$  and  $\beta_1$
- Depending on the values for  $\beta_0$  and  $\beta_1$ , these relationships can look very differently:

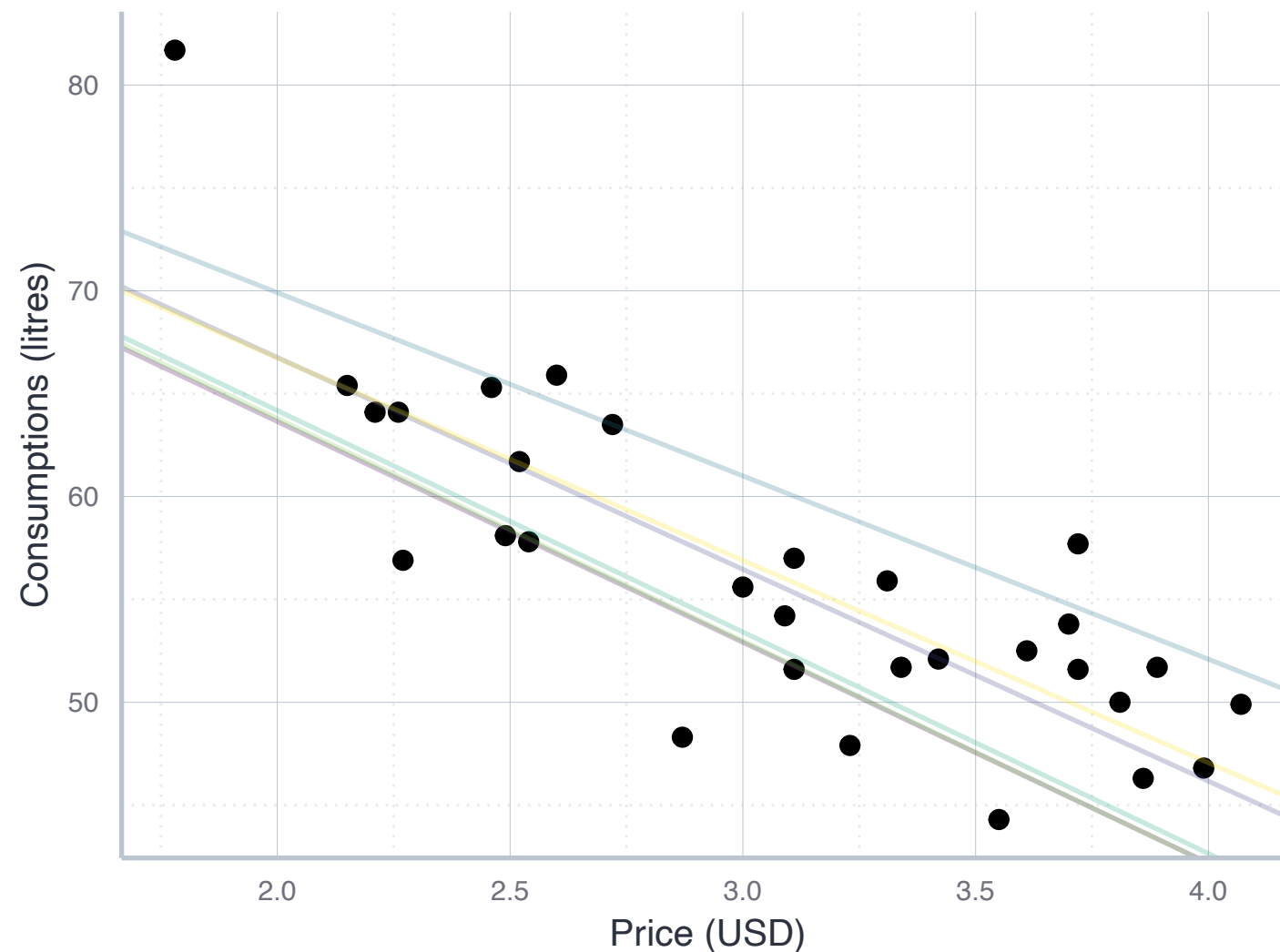


- We see that some of these different members of the linear family are clearly off the mark
- The job of fitting a model means to choose the member of the family that fits the data best  $\rightarrow$  criterion needed!

# Modelling data - general workflow

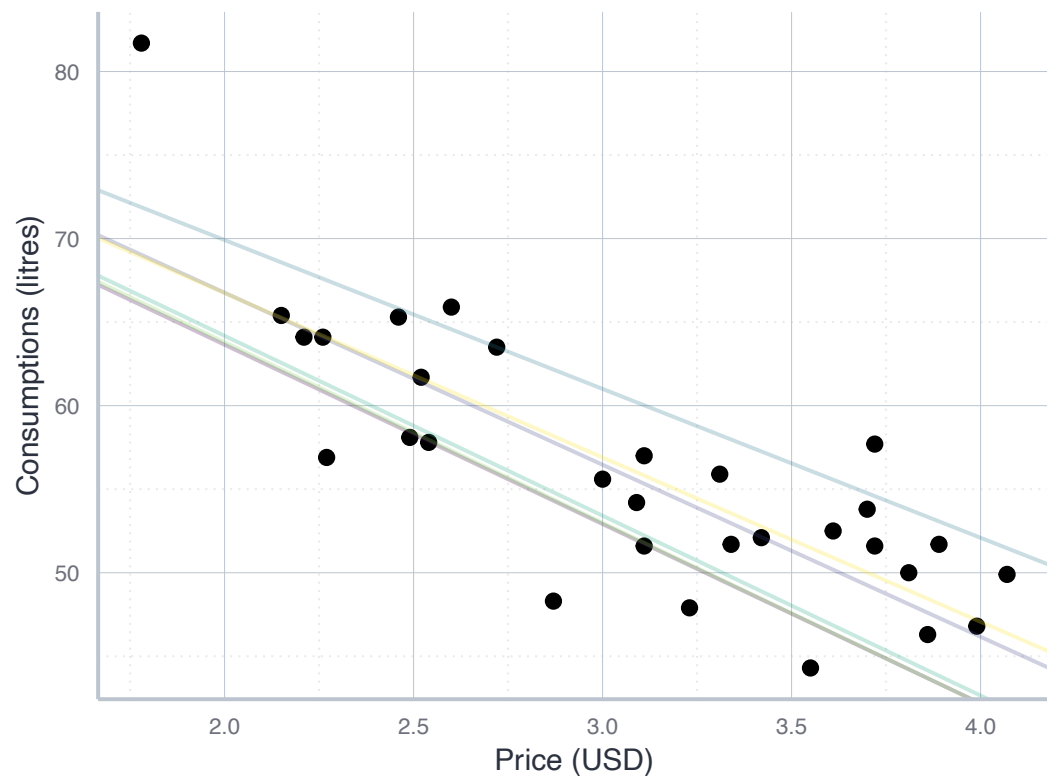
## 3. Fitting a model

- Fitting a model means to choose the 'best' member of a model family
  - How would you, for instance, evaluate the following models?



# Modelling data - general workflow

## 3. Fitting a model

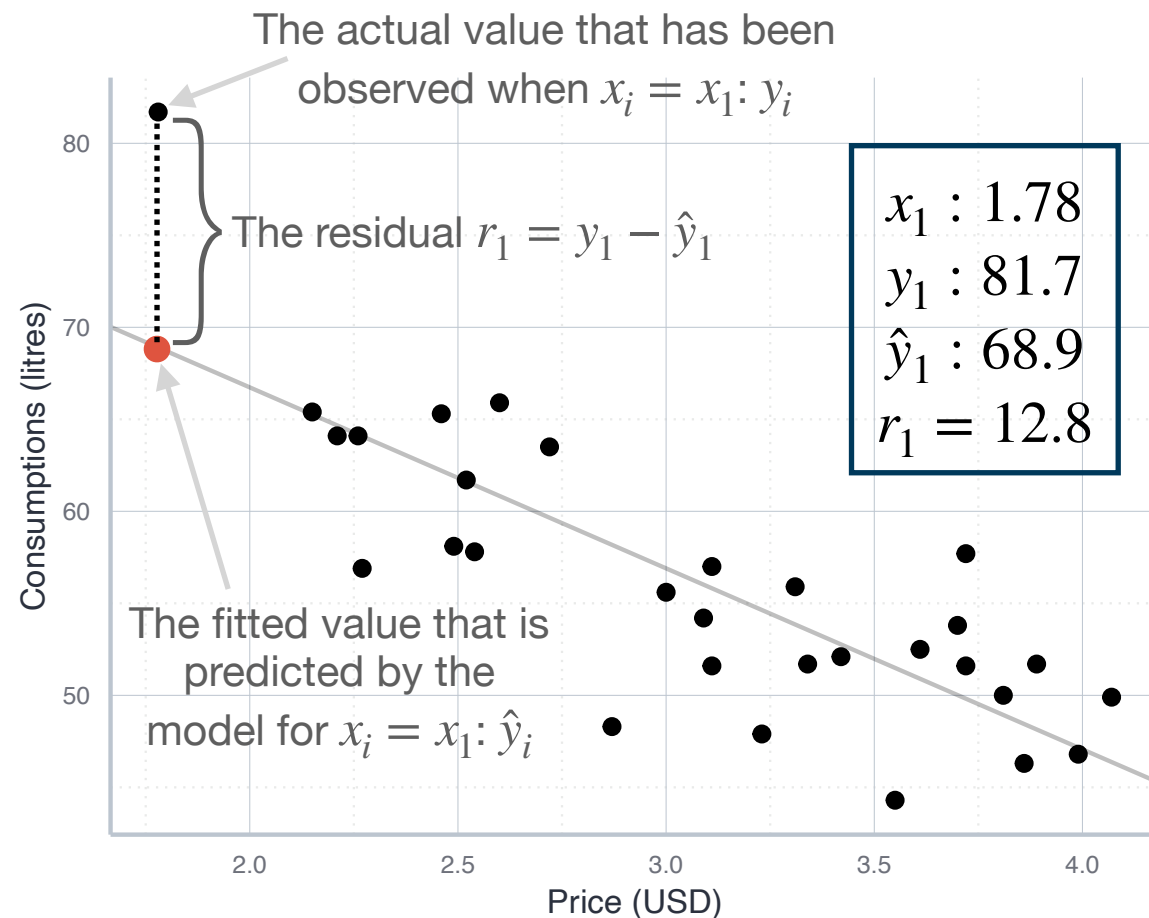


- Each of the model is a particular realisation of the general form  $y = \beta_0 + \beta_1 x$
- If we talk about a particular model instance, where values for  $\beta_0$  and  $\beta_1$  were chosen, we write  $\hat{\beta}_0$  and  $\hat{\beta}_1$

- Such model gives a prediction for each value of  $x$ 
  - We call this prediction a **fitted value** and denote it by  $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$
- A good model would give fitted values  $\hat{y}$  that are close to the true values  $y$ 
  - Thus, a reasonable cost function would consider the difference between true and fitted values: the **residuals**

# Modelling data - general workflow

## 3. Fitting a model



- A good model has fitted values that are close to the actual values
- To get the best model out of a family we should choose the parameters such that the residuals are small
- Since we do not prioritise particular observations, we consider all residuals

- Thus, we can get a measure for the ability of the model to represent the true values by summing up all residuals?
  - We need to square the residuals first → otherwise positive and negative residuals would cancel each other out
  - The sum of squared residuals is called the **RSS**: residual sum of squares

# Modelling data - general workflow

## 3. Fitting a model

- The general approach in machine learning is to choose parameters by first defining a **cost function**, and then to minimise it
- A cost function maps the chosen parameters into a cost measure
  - Here we could use the RSS as a cost measure
  - More widespread is, however, the **Root Mean Squared Error** (RMSE):

$$RSS = \sum_{i=1}^N (y_i - \hat{y}_i)^2$$

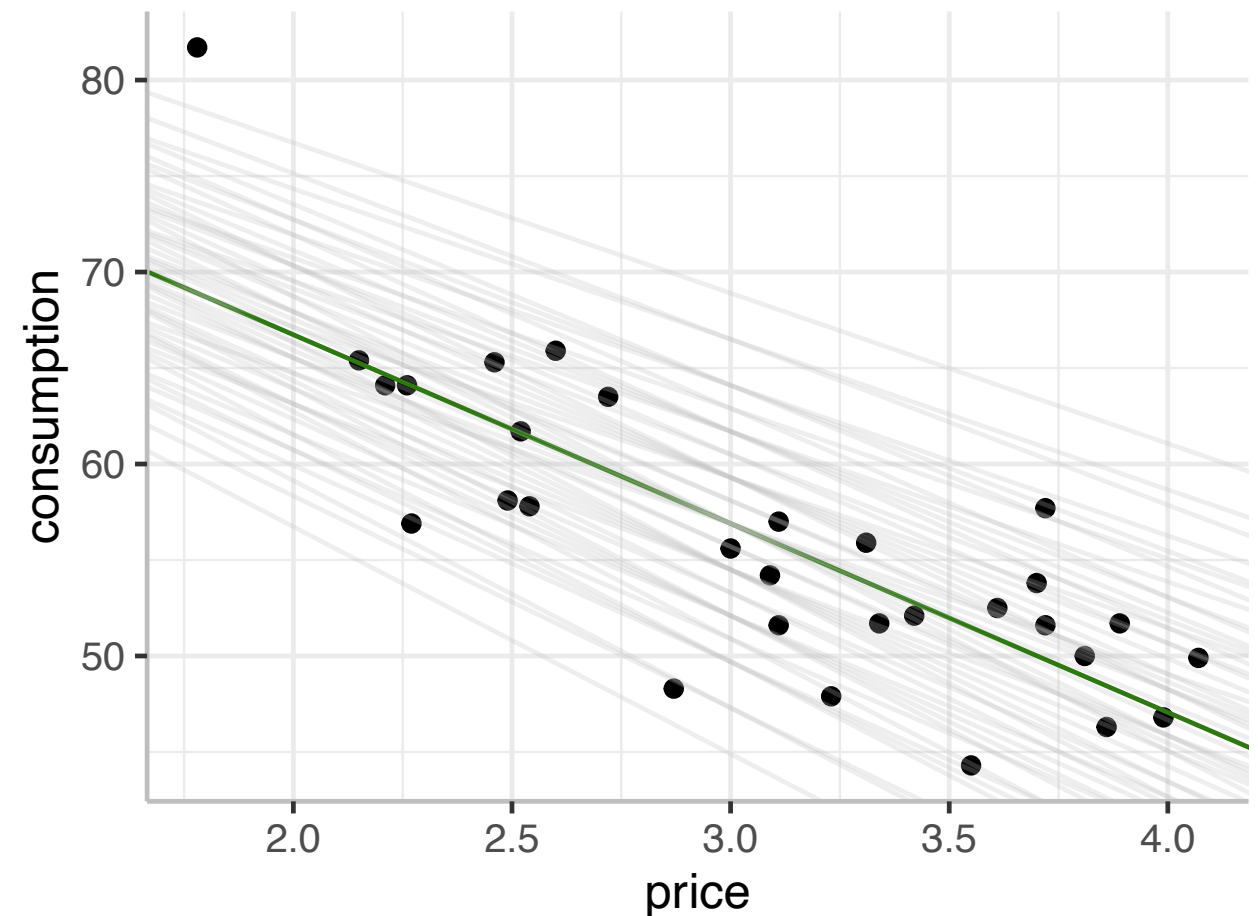
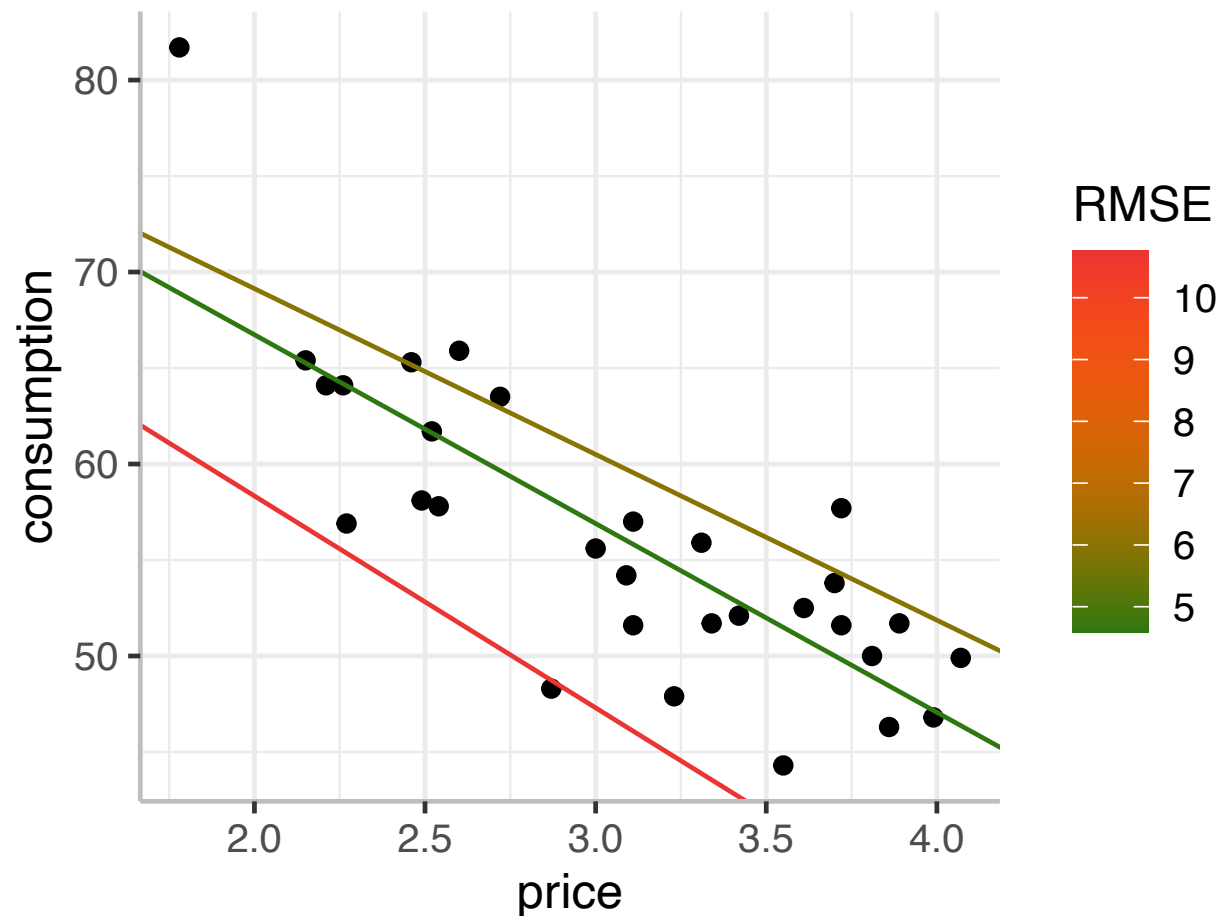
$$MSE = \frac{\sum_{i=1}^N (y_i - \hat{y}_i)^2}{N}$$

$$RMSE = \sqrt{MSE} = \sqrt{\frac{\sum_{i=1}^N (y_i - \hat{y}_i)^2}{N}}$$

# Modelling data - general workflow

## 3. Fitting a model

- Fitting a model means to choose the 'best' member of a model family
  - To evaluate these models we look at their RMSE → the best fit is given by the model with the smallest RMSE → the minimisation problem of **ordinary least squares** (OLS)

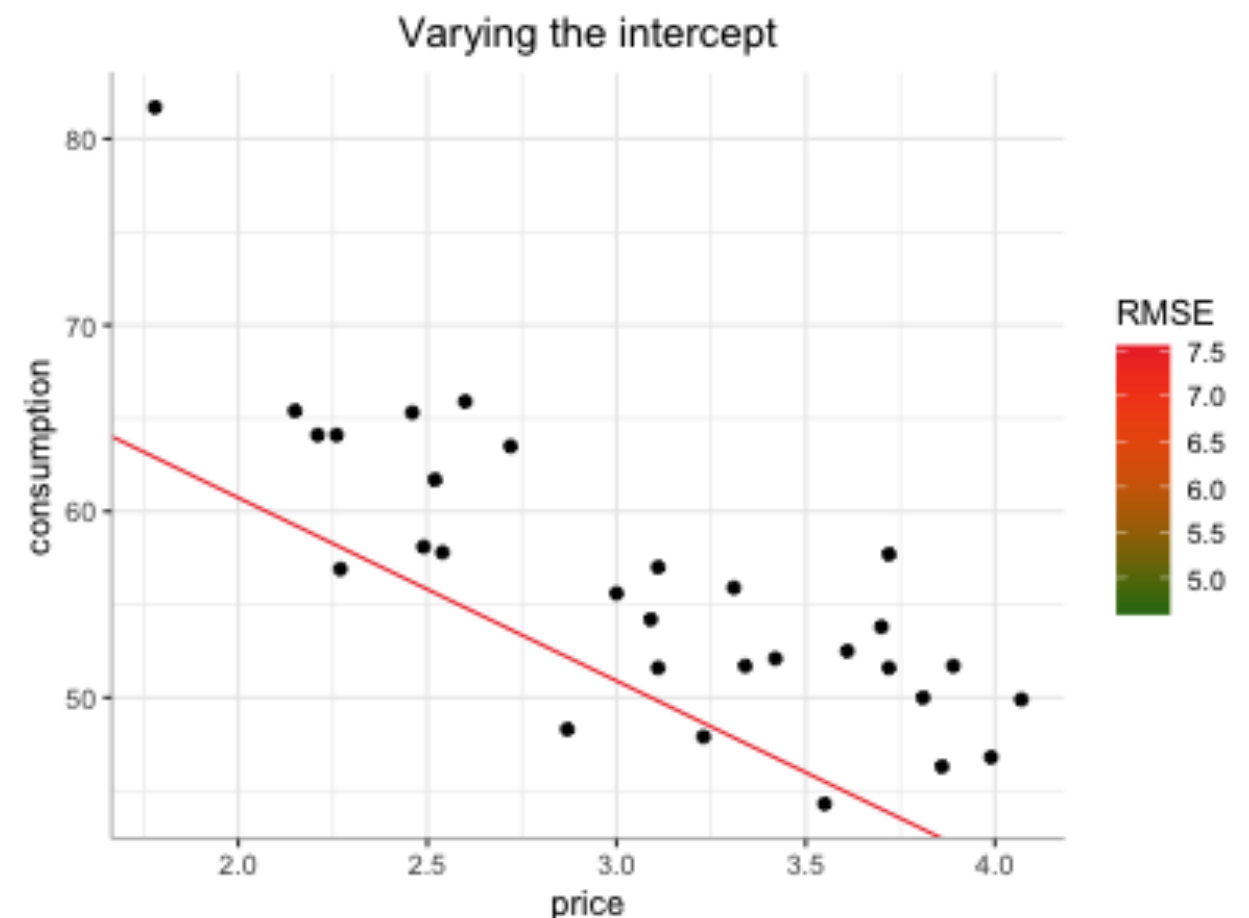
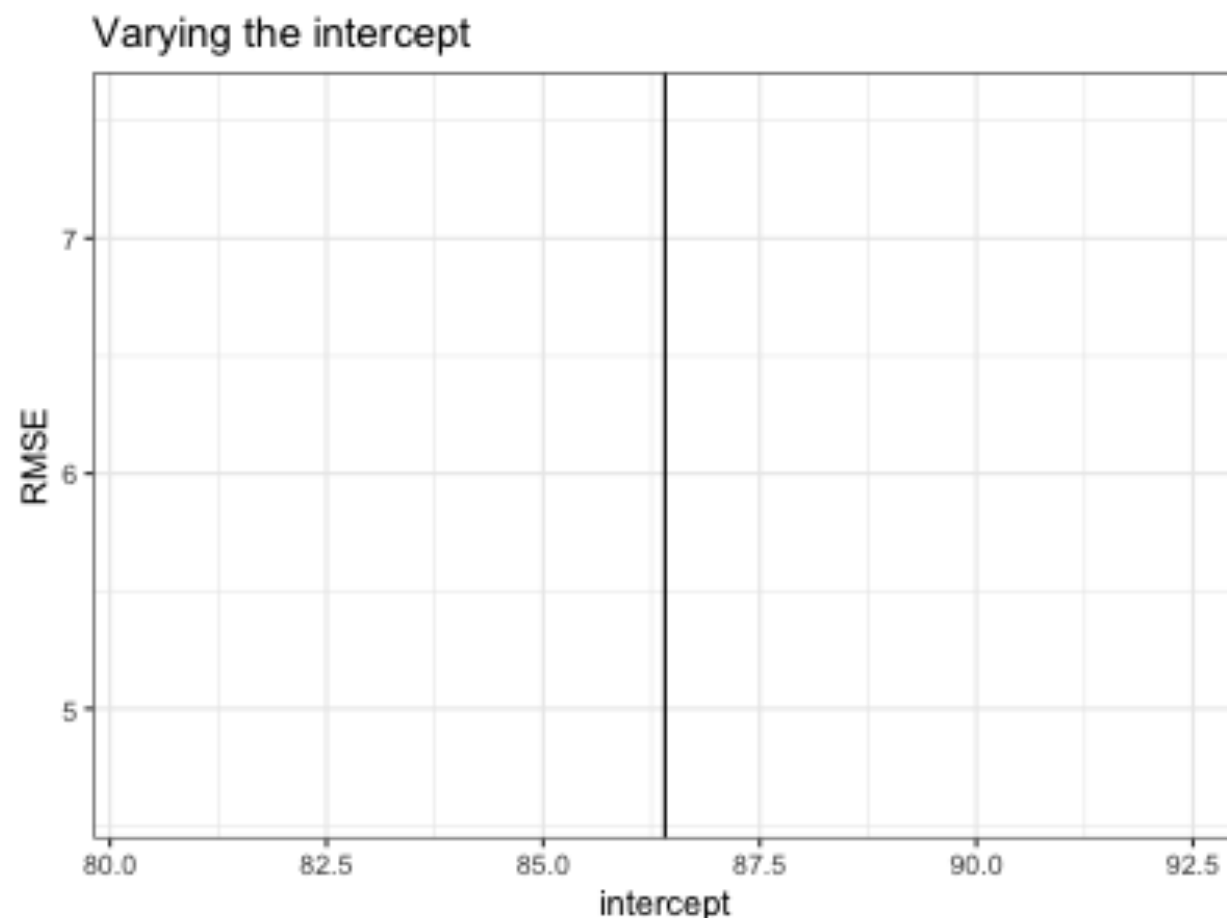




# Modelling data - general workflow

## 3. Fitting a model

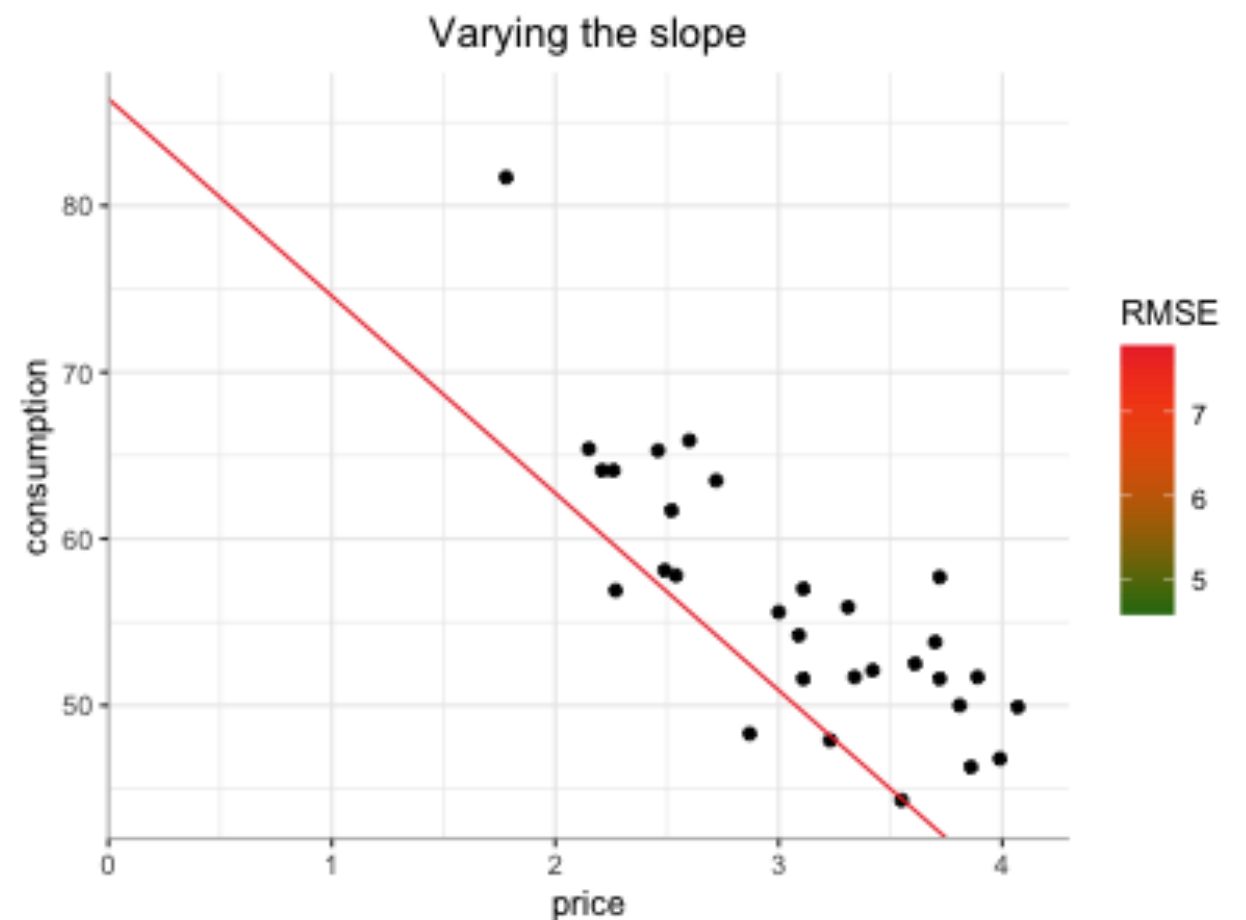
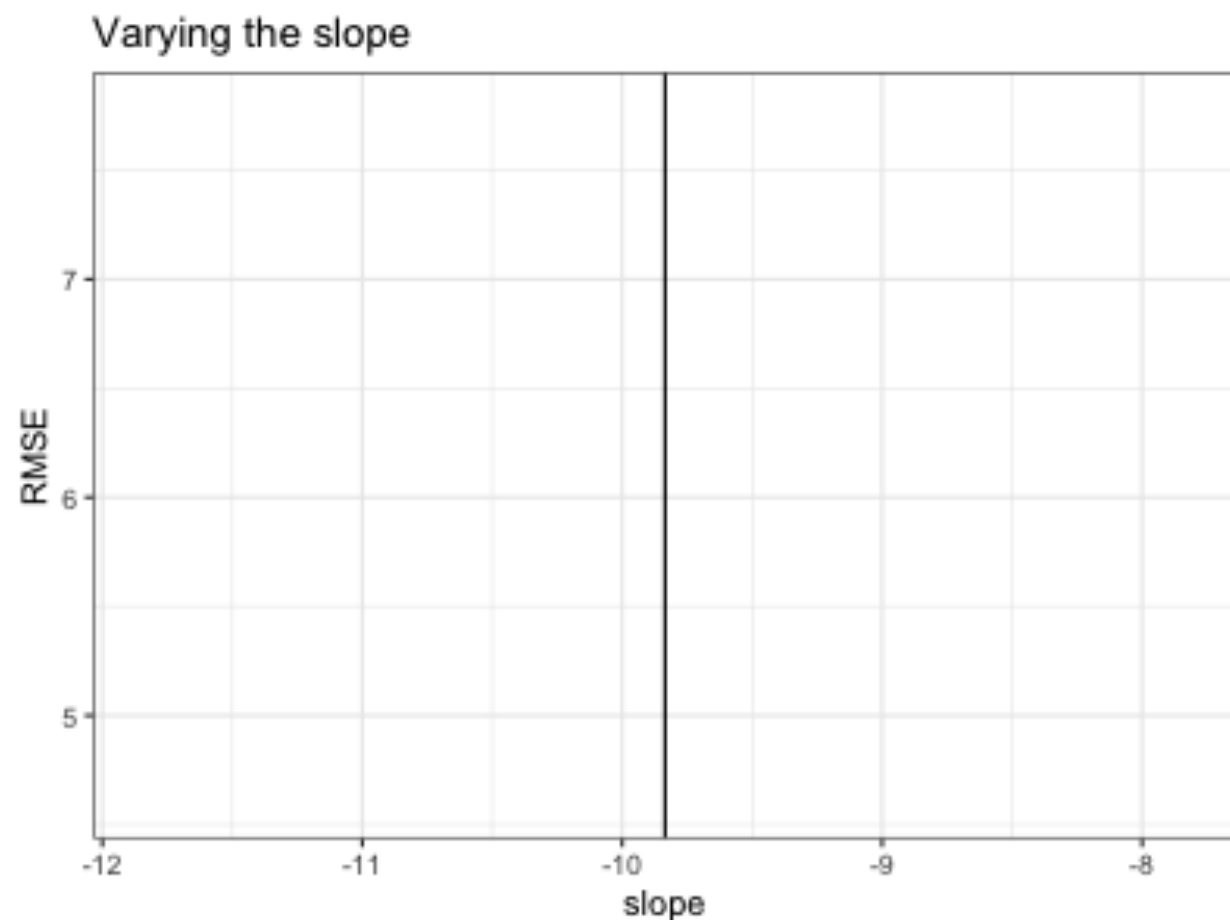
- Fitting a model means to choose the 'best' member of a model family
  - To evaluate these models we look at their RMSE → the best fit is given by the model with the smallest RMSE → the minimisation problem of **ordinary least squares** (OLS)



# Modelling data - general workflow

## 3. Fitting a model

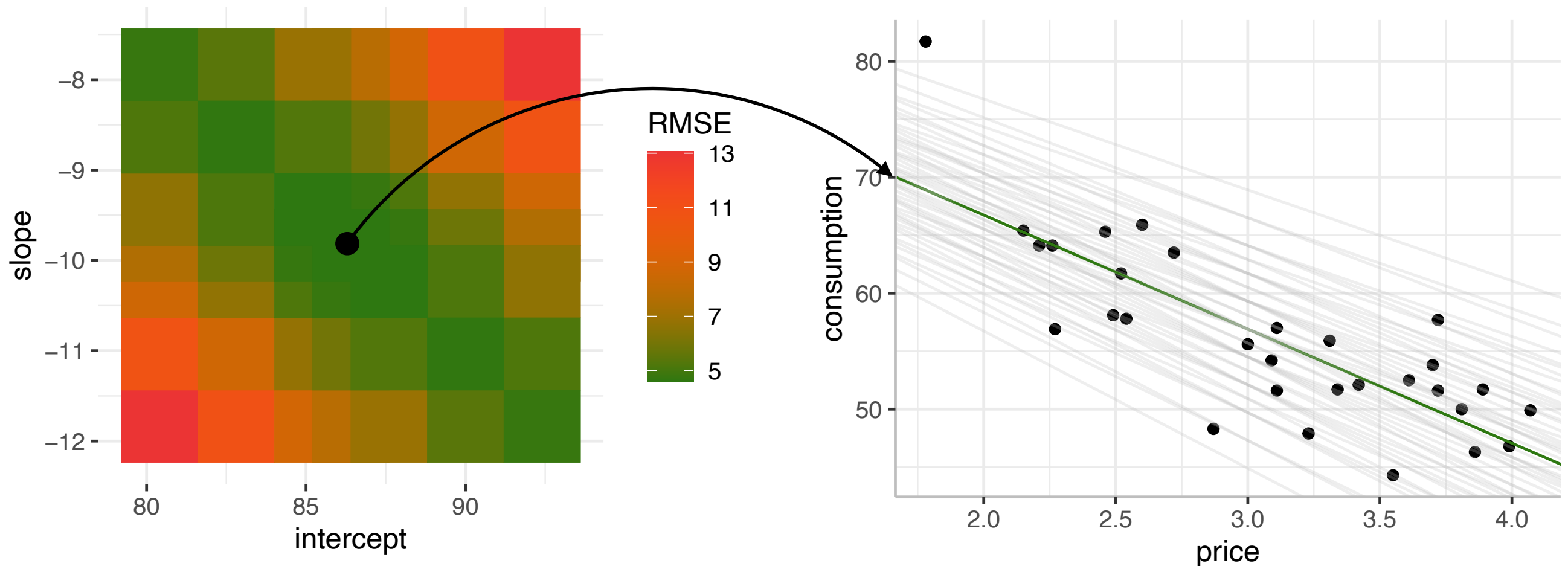
- Fitting a model means to choose the 'best' member of a model family
  - To evaluate these models we look at their RMSE → the best fit is given by the model with the smallest RMSE → the minimisation problem of **ordinary least squares** (OLS)



# Modelling data - general workflow

## 3. Fitting a model

- Fitting a model means to choose the 'best' member of a model family
  - To evaluate these models we look at their RMSE → the best fit is given by the model with the smallest RMSE → the minimisation problem of **ordinary least squares** (OLS)



**Note:** For the linear case, the best model can actually be computed using a formula!

# Modelling data - general workflow

## 3. Fitting a model

- If the family of linear models is adequate for the modelling purpose at hand we can use the function `lm()` to find the model with the smallest RMSE:

```
lm(formula = consumption~price, data = beer_data_red)
```

The regression formula with the dependent variable on the LHS, and the independent variable on the RHS of the `~`

The data set used; the variables in the formula must correspond to the variables in the data set

```
> head(beer_data_red, 2)
A tibble: 2 × 2
 consumption price
 <dbl> <dbl>
1 81.7 1.78
2 56.9 2.27
```

- The immediate output of `lm()` is already quite informative:

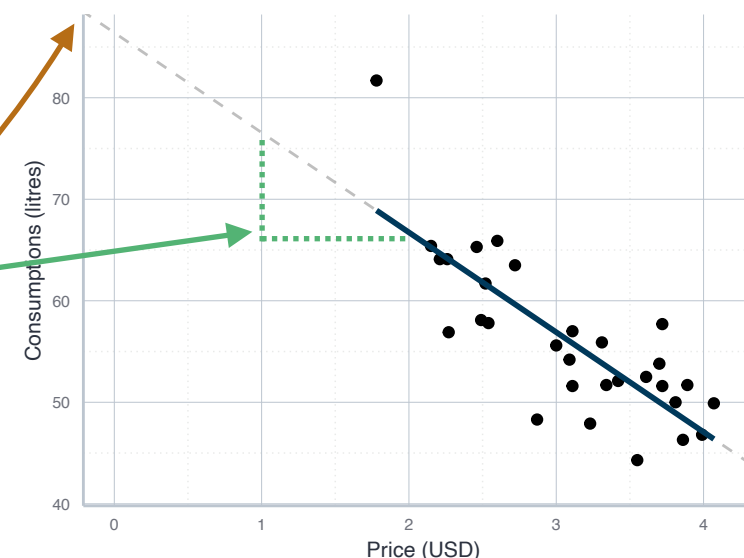
Call:

```
lm(formula = consumption ~ price, data = beer_data_red)
```

Coefficients:

(Intercept)  
86.406

price  
-9.835



# Modelling data - general workflow

## 4. Evaluate and interpret the model

- Usually we want to have more information about our regression result than the function `lm()` provides
  - The classical option is to call `summary()` on the resulting object
- A neat alternative is `moderndive::get_regression_table()`

```
> linmod_c_price <- lm(
+ formula = consumption~price, data = beer_data_red)
> moderndive::get_regression_table(linmod_c_price)
```

```
A tibble: 2 × 7
```

|   | term      | estimate | std_error | statistic | p_value | lower_ci | upper_ci |
|---|-----------|----------|-----------|-----------|---------|----------|----------|
|   | <chr>     | <dbl>    | <dbl>     | <dbl>     | <dbl>   | <dbl>    | <dbl>    |
| 1 | intercept | 86.4     | 4.32      | 20.0      | 0       | 77.5     | 95.3     |
| 2 | price     | -9.84    | 1.38      | -7.15     | 0       | -12.7    | -7.02    |

Subject to later  
sessions!

# Modelling data - general workflow

## 4. Evaluate and interpret the model

```
> linmod_c_price <- lm(
+ formula = consumption~price, data = beer_data_red)
> moderndive::get_regression_table(linmod_c_price)
```

# A tibble: 2 × 7

|   | term      | estimate | std_error | statistic | p_value | lower_ci | upper_ci |
|---|-----------|----------|-----------|-----------|---------|----------|----------|
|   | <chr>     | <dbl>    | <dbl>     | <dbl>     | <dbl>   | <dbl>    | <dbl>    |
| 1 | intercept | 86.4     | 4.32      | 20.0      | 0       | 77.5     | 95.3     |
| 2 | price     | -9.84    | 1.38      | -7.15     | 0       | -12.7    | -7.02    |

Subject to later  
sessions!

- The intercept is often practically irrelevant: hypothetical consumption when *price* = 0
- The coefficient of price (or any explanatory variable) is more important:

For every increase of 1 unit in **price**, there is an **associated decrease** of, **on average**, 9.84 units of consumption.

- Our model is only about association, **not about causation**
- Our model does not say anything about particular comparisons, but the **average over all possible cases**

# Your turn!

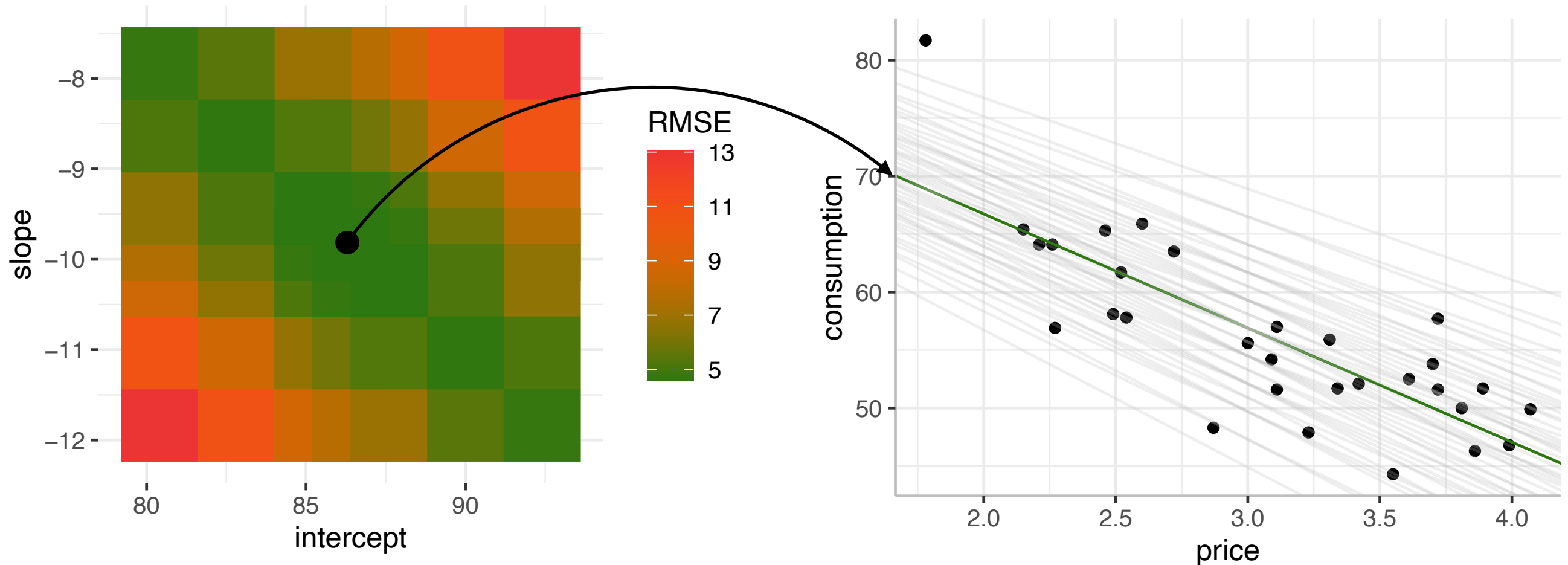
- Consider the data set `DataScienceExercises::beer`, but focus on the relationship between `consumption` and `income`
- Go through all the relevant steps for conducting a regression:
  1. Theoretical pre-considerations
  2. Data exploration and choice of a model family
  3. Fit the model
  4. Evaluate and interpret your model
- Keep in mind that we have used the following functions:
  - `dplyr::glimpse()`, `skimr::skim()`, `lm()` and `moderndive::get_regression_table()`
- *Note: To add a regression line to a ggplot you may use `geom_smooth(method="lm", se=FALSE)`*

# Ordinary Least Squares (OLS) estimation



# Estimating a model using OLS

- Above we argued that estimating a linear model means to identify the model instance with the smallest RMSE
  - Now we look at how this is being done in practice → the OLS method



# Estimating a model using OLS

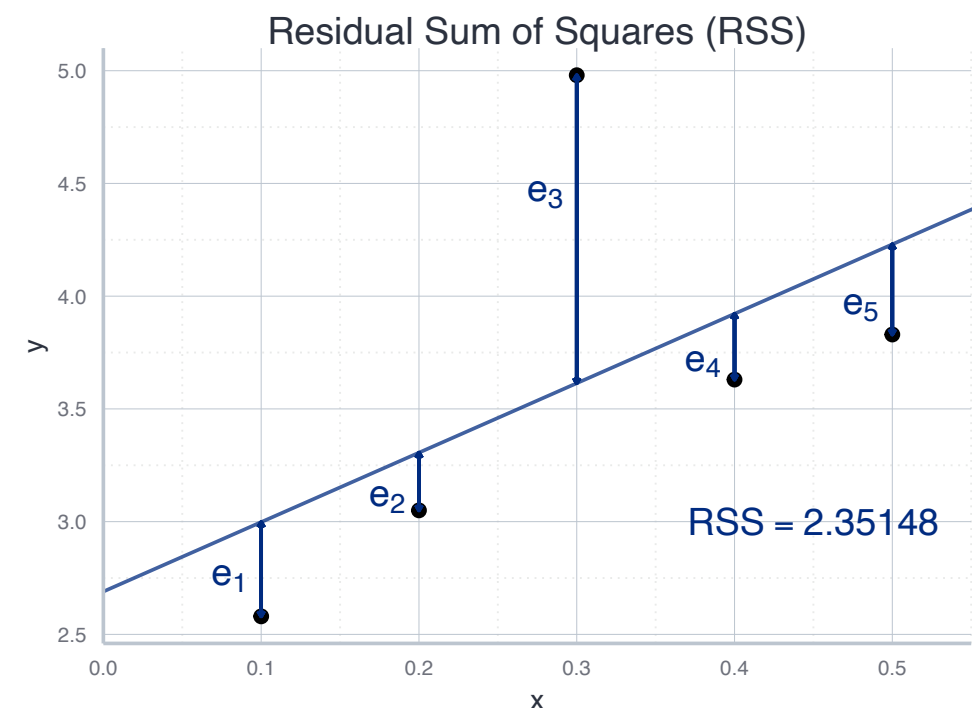
## The general idea

- In principle we could minimise the loss function numerically
  - But this is very inefficient and dangerous
- For the linear case, the best model can be derived analytically
  - This also allows us to derive some further properties of the model
- The idea is to choose  $\beta_0$  and  $\beta_1$  such that the RSS gets minimised

$$RSS = \sum_{i=1}^n e_i^2$$

- Put mathematically:

$$\hat{\beta}_0, \hat{\beta}_1 = \operatorname{argmin}_{\beta_0, \beta_1} \sum_{i=1}^n (y_i - \hat{y}_i)^2$$



# Estimating a model using OLS

## Deriving the OLS estimator

$$\hat{\beta}_0, \hat{\beta}_1 = \operatorname{argmin}_{\beta_0, \beta_1} \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

- Since  $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 \cdot x_i$  this equals have:

$$\hat{\beta}_0, \hat{\beta}_1 = \operatorname{argmin}_{\beta_0, \beta_1} \sum_{i=1}^n (y_i - \hat{\beta}_0 + \hat{\beta}_1 \cdot x_i)^2$$

- With a little bit of algebra we can rearrange this expression to:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \quad \text{and} \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

- All the variables are included in our data  $\rightarrow \hat{\beta}_0$  and  $\hat{\beta}_1$  are identified

# Estimating a model using OLS

## Exercise: computing the OLS estimator manually

- Let us compute the estimated values  $\hat{\beta}_0$  and  $\hat{\beta}_1$  for our example data set by hand

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

```
> data_set
```

```
A tibble: 5 × 2
```

|   | x     | y     |
|---|-------|-------|
|   | <dbl> | <dbl> |
| 1 | 0.1   | 2.58  |
| 2 | 0.2   | 3.05  |
| 3 | 0.3   | 4.98  |
| 4 | 0.4   | 3.63  |
| 5 | 0.5   | 3.83  |

- $\bar{x} = 0.3$
- $\bar{y} = 3.614$
- $\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = (0.1 - 0.3)(2.58 - 3.614) + \dots = 0.308$
- $\sum_{i=1}^n (x_i - \bar{x})^2 = (0.1 - 0.3)^2 + (0.2 - 0.3)^2 + \dots = 0.1$
- $\hat{\beta}_1 = \frac{0.308}{0.1} = 3.08$
- $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 3.614 - 3.08 \cdot 0.3 = 2.69$

# Estimating a model using OLS

## Exercise: computing the OLS estimator manually

```
> data_set
```

```
A tibble: 5 × 2
```

|   | x     | y     |
|---|-------|-------|
|   | <dbl> | <dbl> |
| 1 | 0.1   | 2.58  |
| 2 | 0.2   | 3.05  |
| 3 | 0.3   | 4.98  |
| 4 | 0.4   | 3.63  |
| 5 | 0.5   | 3.83  |

- $\bar{x} = 0.3$

- $\bar{y} = 3.614$

- $\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = (0.1 - 0.3)(2.58 - 3.614) + \dots = 0.308$

- $\sum_{i=1}^n (x_i - \bar{x})^2 = (0.1 - 0.3)^2 + (0.2 - 0.3)^2 + \dots = 0.1$

- $\hat{\beta}_1 = \frac{0.308}{0.1} = 3.08$

- $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 3.614 - 3.08 \cdot 0.3 = 2.69$

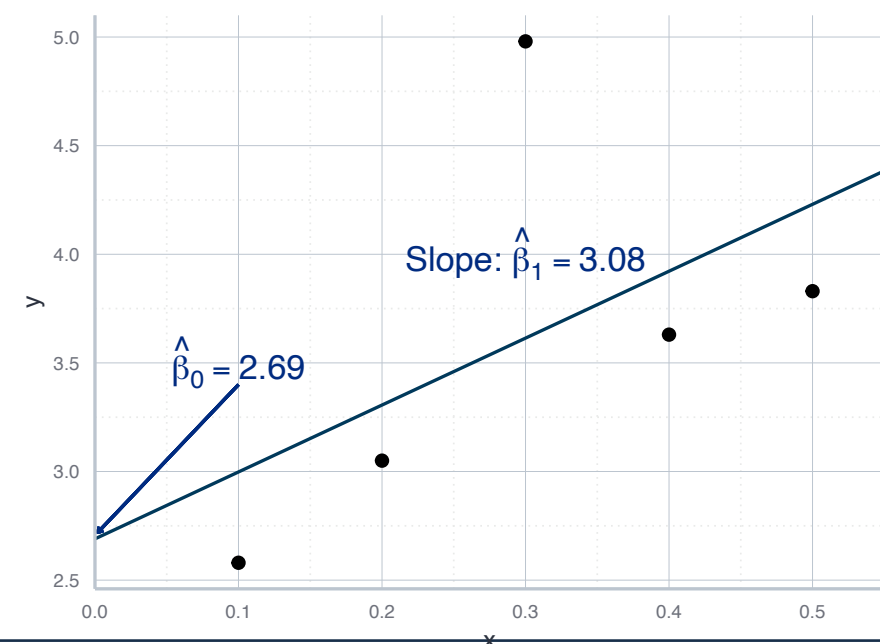
- Let us now verify our result by computing  $\hat{\beta}_0$  and  $\hat{\beta}_1$  using `lm()`:

Call:

```
lm(formula = y ~ x, data = data_set)
```

Coefficients:  
(Intercept)  
2.69

x  
3.08



# Estimating a model using OLS

## Final remarks on the OLS method

- $\beta_i$  and  $\hat{\beta}_i$  are different: the former is the **true value**, the latter the **estimate**
  - This distinction refers to the fundamental distinction between a **population** and a **sample**
  - We will discuss this in more detail after our session on sampling
- In this context we also need to distinguish an **estimator** and the **estimate**
  - An estimator is a way to compute the estimate: it's a formula or an algorithm
  - The estimate is the result of this procedure: for each sample, it corresponds to a single number

# Estimating a model using OLS

## Final remarks on the OLS method

- The OLS estimation method has some great mathematical properties
  - E.g., if you can only obtain a sample of the population of interest, the estimates obtained via OLS are **unbiased** and **efficient**
- These properties hinge, however, on some **assumptions**, e.g. a linear relationship between  $y$  and  $x$ 
  - In practice you always need to test whether your assumptions are met
  - Otherwise there is no way to tell whether the estimates obtained via OLS are not terribly misleading → see session on **regression diagnostics**

# Model evaluation

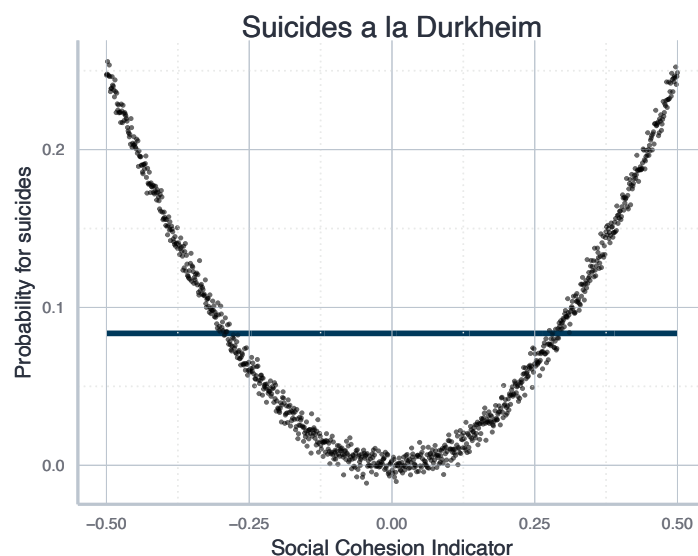


# Evaluating models - assumptions

- We identified the best model by minimising the RMSE → the method of ordinary least squares (OLS)
  - Identifying the model this way is based on a number of assumptions
- Part of any model evaluations should be the test of whether these assumptions were satisfied in the case at hand
  - We will have a specific session about how to do this
- **Example:** one central assumption of the simple OLS regression is that the relationship between the two variables is **linear**
- What would happen if this assumption was not met?

# Evaluating models - assumptions

- The French sociologist Emile Durkheim distinguished two types of suicides:
  - Moral confusing and a lack of social embeddednes in modern societies
  - Neglect of individual desires in archaic societies
- This could be summarised in a u-shaped relationship between social cohesion and the likelihood of suicides



- This is not a linear relationship, and fitting a linear model would lead to very misleading results
  - Here the estimate for  $\beta_1$  would be zero  $\rightarrow$  suggests no systematic relationship
- Its always important to visualise the data and then choose the right family

# Evaluating models - explanatory power

- We will learn more about the underlying assumptions and how to test for them in a later session
- At this point we want to focus on one additional measure for the goodness of fit of a model: its  $R^2$ 
  - The  $R^2$  measures how much variation in the explained variable can be explained by the variation of the explanatory variable
  - Lets look at an artificial example:

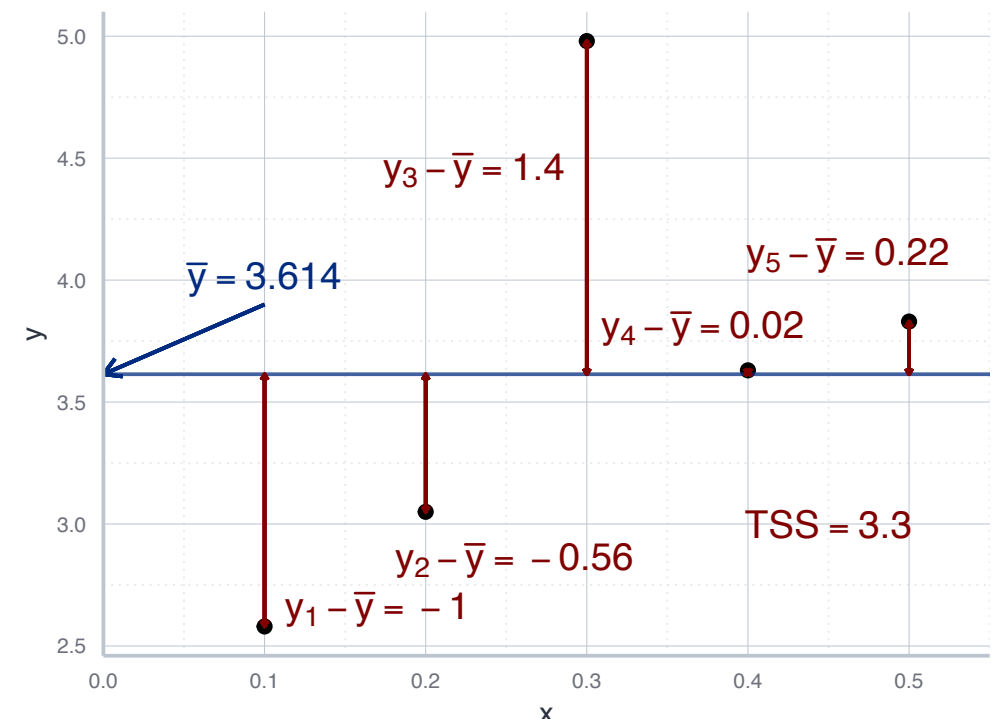
datensatz

```
#> x y
#> 1 0.1 2.58
#> 2 0.2 3.05
#> 3 0.3 4.98
#> 4 0.4 3.63
#> 5 0.5 3.83
```

- How to measure the total variation in the explained variable?

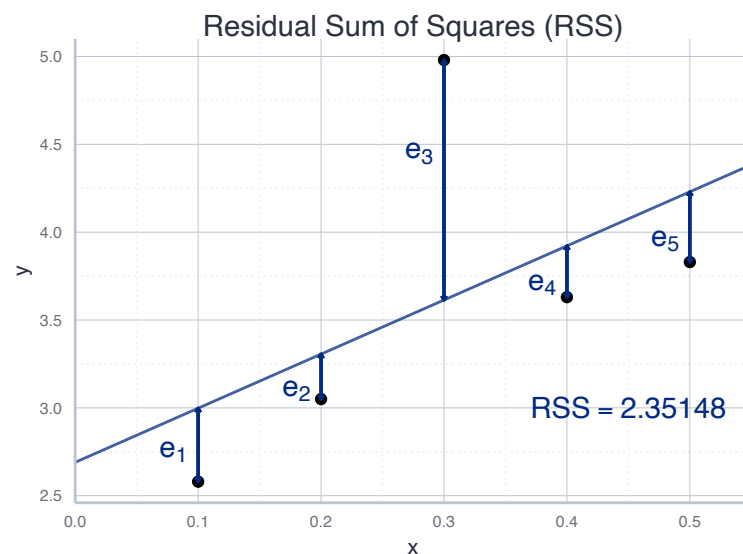
- Deviations from its mean value:  
total sum of squares:

- $$TSS = \sum_{i=1}^n (Y_i - \bar{Y})^2$$



# Evaluating models - explanatory power

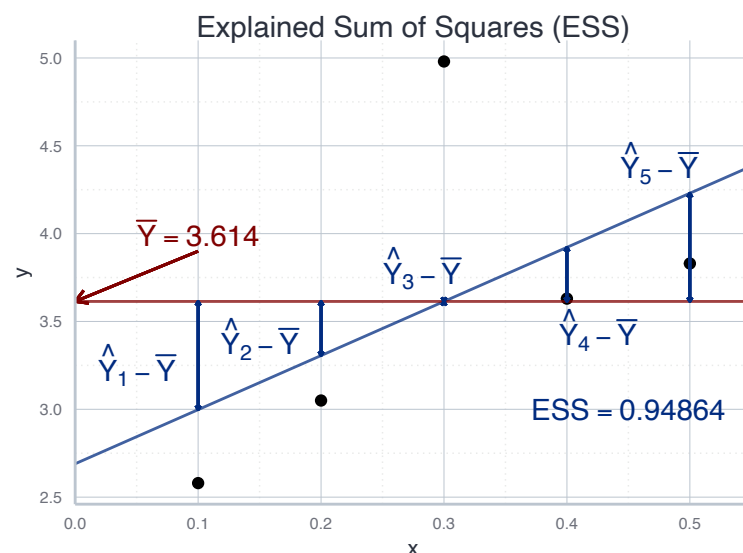
- TSS as the total variation in the outcome variable:  $TSS = \sum_{i=1}^n (Y_i - \bar{Y})^2$
- We separate the total variation into two parts:



- **Explained sum of squares** (ESS): the variation explained by our model
- **Residual sum of squares** (RSS): the variation left unexplained
- RSS: the sum of squared residuals:

$$RSS = \sum_{i=1}^n e_i^2$$

- Residuals  $e$ : observable counterpart to the error term  $\epsilon$
- ESS: squared deviations between the fitted values and  $\bar{y}$ :



$$ESS = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2$$

# Evaluating models - explanatory power

- We separate the total variation into two parts:

$$TSS = ESS + RSS$$

- The  $R^2$  is defined as the share of explained variation:

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS}$$

- In general, a higher  $R^2$  comes with higher explanatory power
- A very high  $R^2$ , however, should also make you suspicious
- But in general, its a good indication for the usefulness of your model

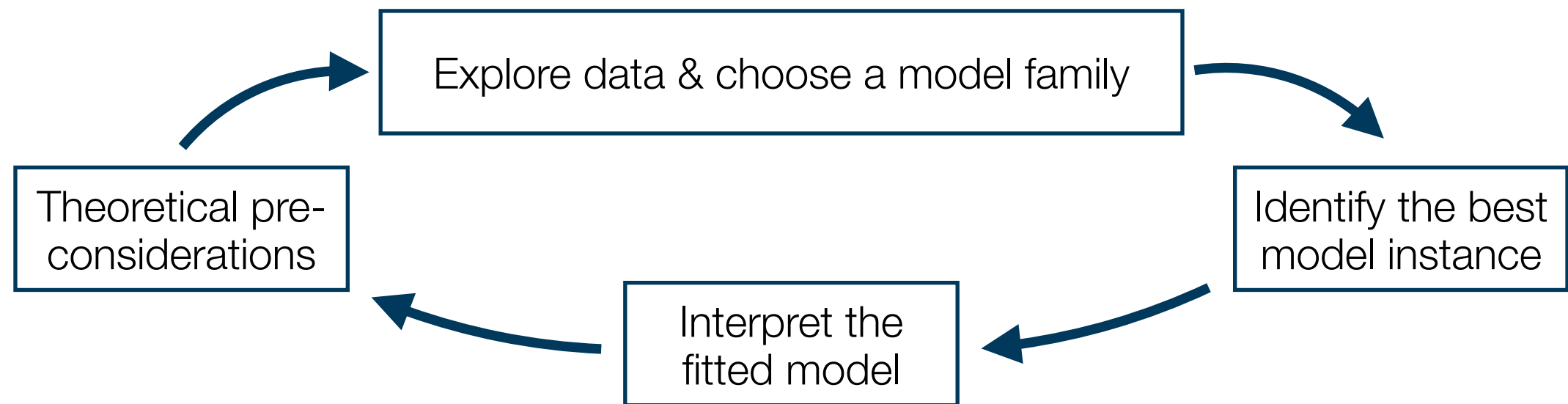
# Exercise: computing $R^2$

- Consider again our example of beer consumption and the linear model you fitted before (i.e. on beer consumption and income).
  - Now compute the  $R^2$  of your model by hand.
- Remember:
  - $TSS = \sum_{i=1}^n (Y_i - \bar{Y})^2$
  - $RSS = \sum_{i=1}^n e_i^2$
  - $ESS = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2$
  - Any `lm`-object has the elements `residuals` and `fitted.values`, through which you can obtain the respective vectors
- How can you interpret your  $R^2$ ?
- Bonus: compare it to the  $R^2$  of the model including price instead of income. How would you interpret this?

# Summary & outlook

# Summary and outlook

- We applied the general **workflow** of empirical modelling in the context of simple linear regression:



- The idea is to use the **family of linear models** with **two variables**
- Thus, SLR is used to study the association of two numerical variables
- The idea is to fit a regression line that minimises the squared differences between the actual and fitted values → method of **ordinary least squares**



# Summary and outlook

- Using SLR makes sense if you are interested in a **linear relationship** between numerical variables
  - Thus, prior theoretical considerations and descriptive exploration of your data is necessary
- SLR is built upon the **family of linear models**, which in the context of economic applications is specified as  $y = \beta_0 + \beta_1 x_1 + \epsilon$ 
  - In this context we introduced the concepts of the *LHS* and *RHS* of a regression equation, as well as the terms *parameters*, *dependent & independent variables*, and the *error term*
- We defined the best model instance of the family of linear models as the one that has the smallest **RMSE** for the data at hand
  - To find the particular model, we used the method of **OLS**

# Summary and outlook

- OLS produces concrete **estimates**  $\hat{\beta}_0$  and  $\hat{\beta}_1$  by minimising the RMSE for the data at hand
  - Once estimated, we can use our model to create predictions: the **fitted values**  
 $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$
- The deviations from the fitted and actual values are called **residuals** → sample equivalent to the theoretical error term
- Once estimated, we can interpret the estimated values of our model
  - The model has **no causal interpretation** → its about associations
- The OLS method is built upon **assumptions**, which we need to check for each application
- There are other tools to assess our estimated model, such as its  $R^2$

# Summary and outlook

- Next time we will extend the approach of simple linear regression and learn about **multiple** linear regression
  - We study not the relationship between two, but between many variables
  - This will allow us to isolate the relationship between two variables from the confounding effects of other variables
  - After this, we consider the process of taking samples from bigger populations theoretically, and then learn how to assess the quality of our regression models

## Tasks until next time:

1. Fill in the **quick feedback survey** on Moodle
2. Read the **tutorials** posted on the course page
3. Do the **exercises** provided on the course page and **discuss problems** and difficulties via the Moodle forum