Multiple linear regression

Applied Data Science using R

Prof. Dr. Claudius Gräbner-Radkowitsch
Europa-University Flensburg, Department of Pluralist Economics
www.claudius-graebner.com | @ClaudiusGraebner | claudius@claudius-graebner.com

Europa-Universität Flensburg

Europa-Universität Flensburg
International Institute of Management and Economic Education
Department of Pluralist Economics

Introduction

- Build upon the example from previous session: what are the determinants of beer consumption?
    - We considered two variables **separately**: income and beer price
- Multiple regression analysis allows us to consider **both variables at once**
    - This changes the interpretation of the obtained estimates
    - They now give the association with the outcome variable, assuming that all other variables are held constant

Goals for today

I. Learn how to implement and interpret multiple linear regression models

II. Learn how to deal with categorial variables within a regression

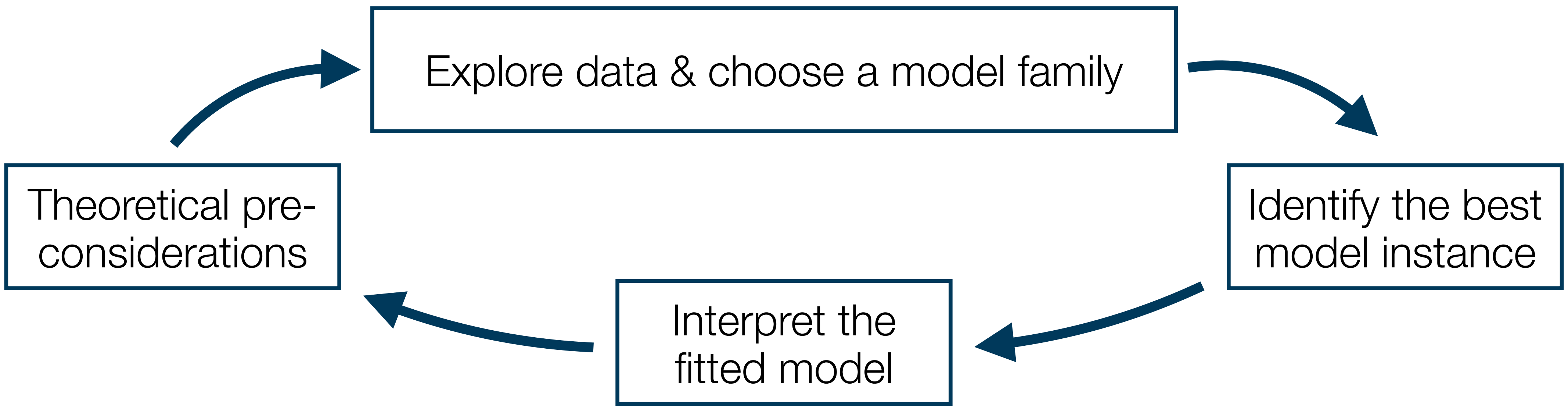
III. Understand the concept of interaction effect and the difference between interaction and parallel slopes models

Multiple linear Regression

Introduction

- In terms of theoretical background and technical implementation, multiple regression analysis is very similar to simple linear regression

- The overall sequence of considerations remains the same:



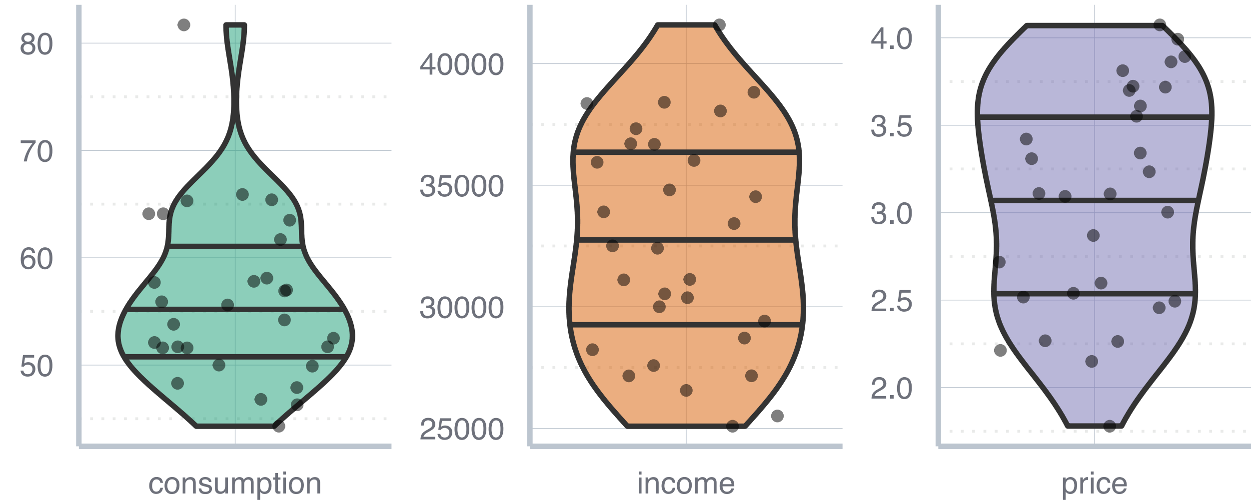
- Let us take this opportunity to recap what we have learned

Data exploration

- We again use the data set `DataScienceExercises::beer`, but only the three variables of interest

```
── Data Summary ────────────────────────
                           Values
Name                       beer_data
Number of rows             30
Number of columns          3
Key                        NULL
________________________
Column type frequency:
  numeric                  3
________________________
Group variables            None

── Variable type: numeric ──────────────────────────────────────────────────────────────────────
  skim_variable n_missing complete_rate     mean       sd       p0      p25      p50      p75
1 consumption           0             1     56.1     7.86     44.3     51.6     54.9     60.8
2 price                 0             1     3.08    0.642     1.78     2.53     3.11     3.68
3 income                0             1  32602.    4542.     25088    28888    32457    36516.
     p100 hist
1    81.7
2    4.07
3 41593
```



- Since `consumption`, `income` and `price` are all numerical, we can basically proceed as in the previous session

Data exploration

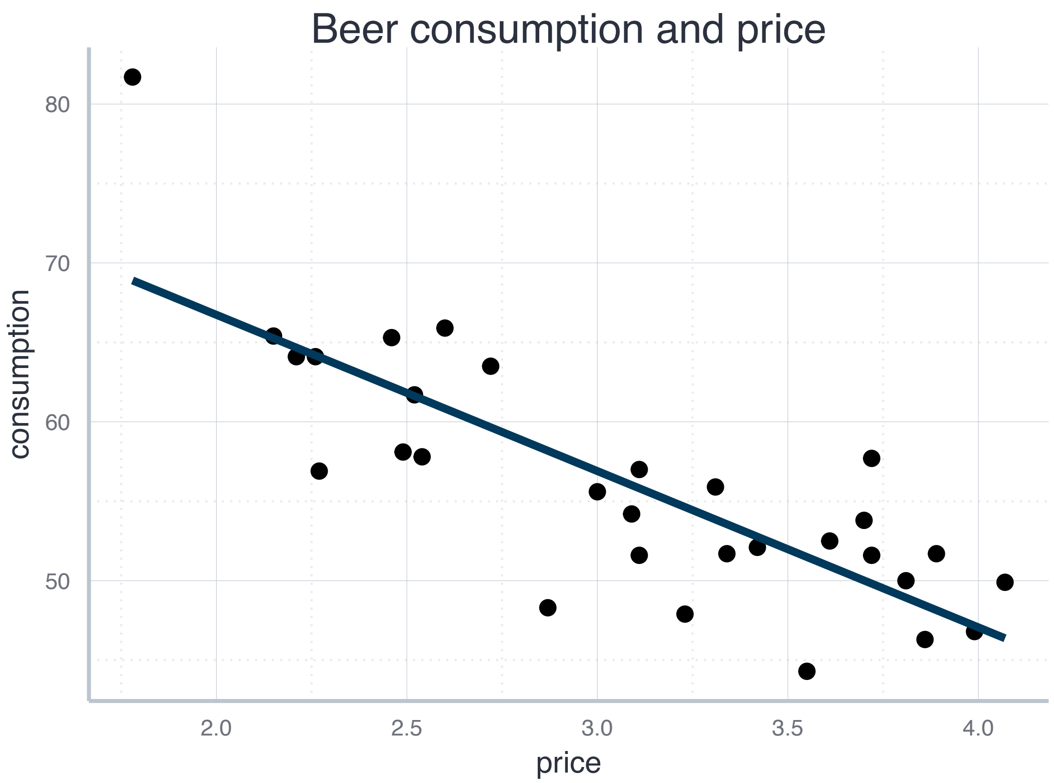
- Our focus on both income and price can be justified **theoretically** via reference to economic theory....
- ...and **empirically** by looking at the correlations:

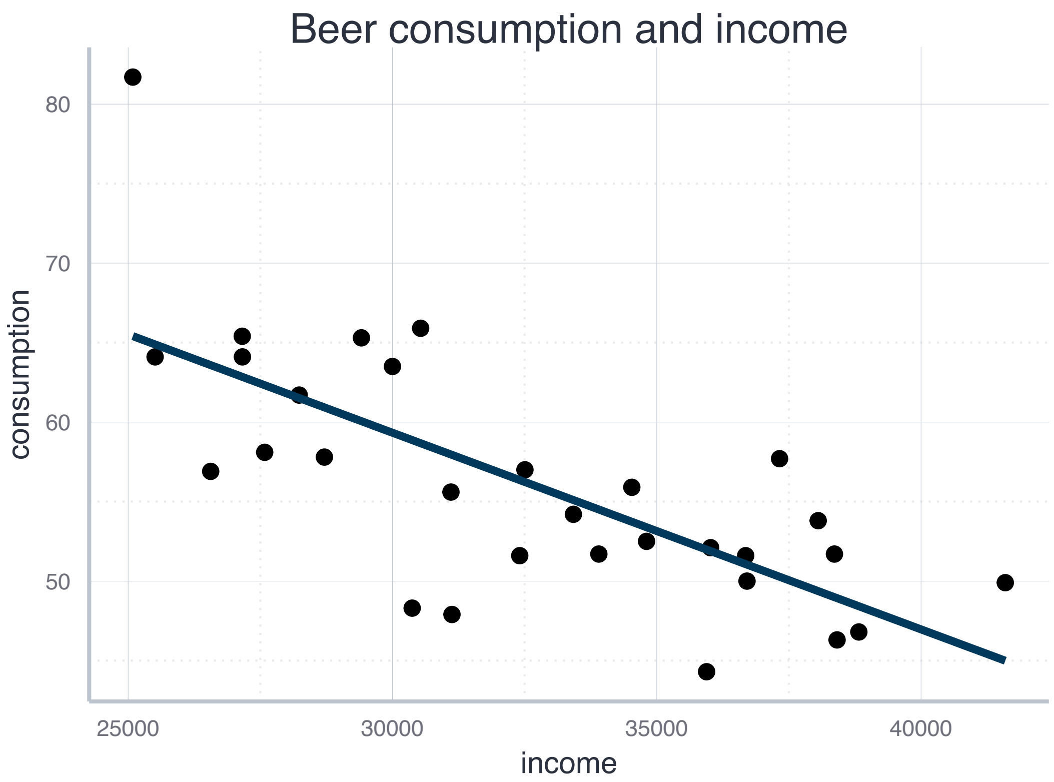
```
> cor(beer_data$consumption, beer_data$price)
[1] -0.8038513
> cor(beer_data$consumption, beer_data$income)
[1] -0.714995
> cor(beer_data)
            consumption      price     income
consumption   1.0000000 -0.8038513 -0.7149950
price        -0.8038513  1.0000000  0.9763155
income       -0.7149950  0.9763155  1.0000000
```

Note: very strong correlations between explanatory variables should be a warning sign! More on this later!

Data exploration

- Our focus on both income and price can be justified theoretically via reference to economic theory....
- ...and empirically by looking at the correlations:





➡ In both cases, a linear models seems to be an adequate choice!

Estimate a multiple regression model

- Writing down our regression model with two explanatory variables is very similar to the case with only one variable:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$$

$$CONS = \beta_0 + \beta_1 PRICE + \beta_2 INCOME + \epsilon$$

- The computation in R is equally similar → here is the general form:

```
lm(y ~ x1 + x2, data=data_used)
```

- **Exercise**: adjust the code to the actual data set `DataScienceExercises::beer` and estimate the model!

Interpret a multiple regression model

```
> cons_model <- lm(consumption ~ price + income, data = beer_data)
> moderndive::get_regression_table(cons_model)
# A tibble: 3 × 7
  term      estimate std_error statistic p_value lower_ci upper_ci
  <chr>        <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
1 intercept   57.2      9.47        6.04   0         37.7     76.6
2 price      -27.7      5.44       -5.08   0        -38.8    -16.5
3 income       0.003    0.001       3.36   0.002     0.001    0.004
```

- In the multiple case, the coefficients must be interpreted in a *ceteris paribus* fashion:

For every increase of 1 unit in `price`, there is an associated decrease of, on average **and ceteris paribus**, 27.7 units of `consumption`.

For every increase of 1 unit in `income`, there is an associated increase of, on average **and ceteris paribus**, 0.003 units of `consumption`.

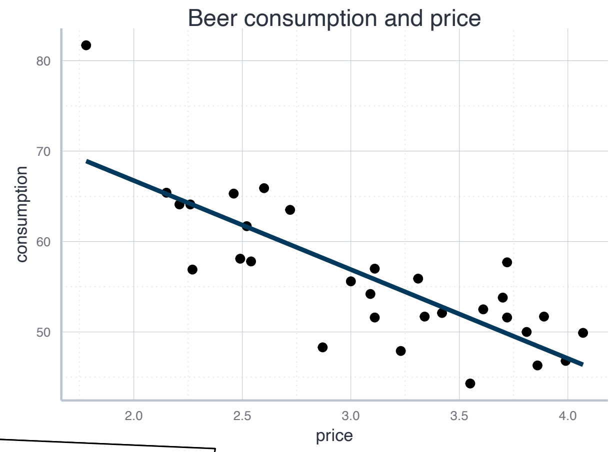
Graphical interpretation

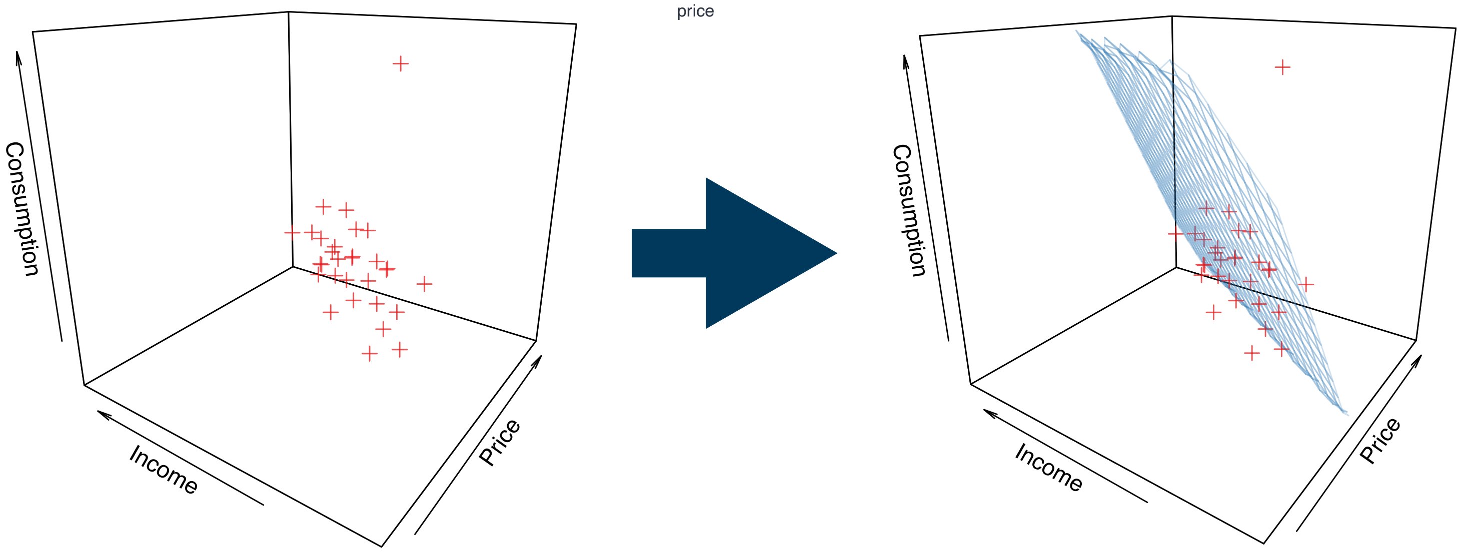
- The more explanatory variables you use, the more difficult it becomes to think about the regression problem graphically

- Simple regression: fit a regression **line**

- Two explanatory variables: fit a regression **plane**

- More than two variables: fit a regression **hyperplane**





Outlook: the choice of variables matters

- **Guess:** how do the estimates for income and price from the simple regression models and the multiple regression model relate to each other?

	Model 1	Model 2	Model 3
(Intercept)	86.406 ***	96.439 ***	57.160 ***
	(4.324)	(7.521)	(9.468)
price	-9.835 ***		-27.653 ***
	(1.375)		(5.438)
income		-0.001 ***	0.003 **
		(0.000)	(0.001)
R^2	0.646	0.511	0.750
Adj. R^2	0.634	0.494	0.732
Num. obs.	30	30	30

- This points to an important concept: **omitted variable bias**
  - When you forget one important variable in your model, all resulting estimates can be misleading → more on this in later sessions

Exercise I

- Get together in groups and use again the data on beer consumption
- But this time use all potential explanatory variables for the RHS:
    - `price`: the price for beer
    - `price_liquor`: the price for other strong alcoholic beverages
    - `price_other`: price of other goods and services
    - `income`: household income
- Before you do the estimation, what would you expect regarding their effect?
- How can you interpret the estimates you obtained? How did the estimates change over different specifications?
- What specification would you prefer? Why?

Exercise

- Before you do the estimation, what would you expect regarding their effect?
- How can you interpret the estimates you obtained? How did the estimates change over different specifications?
- What specification would you prefer? Why?
- *Note: get back to this table at the end of lecture!*

	(1)	(2)	(3)	(4)
(Intercept)	86.406***	57.160***	82.159***	65.738***
	(4.324)	(9.468)	(17.962)	(8.800)
price	−9.835***	−27.653***	−23.743***	−26.426***
	(1.375)	(5.438)	(5.429)	(4.797)
income		2.580**	1.995*	1.726*
		(0.769)	(0.776)	(0.734)
price_liquor			−4.077	
			(3.890)	
price_other			12.924**	12.394**
			(4.164)	(4.141)
Num.Obs.	30	30	30	30
R2	0.646	0.750	0.822	0.814
R2 Adj.	0.634	0.732	0.794	0.793
RMSE	4.60	3.86	3.26	3.33

+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

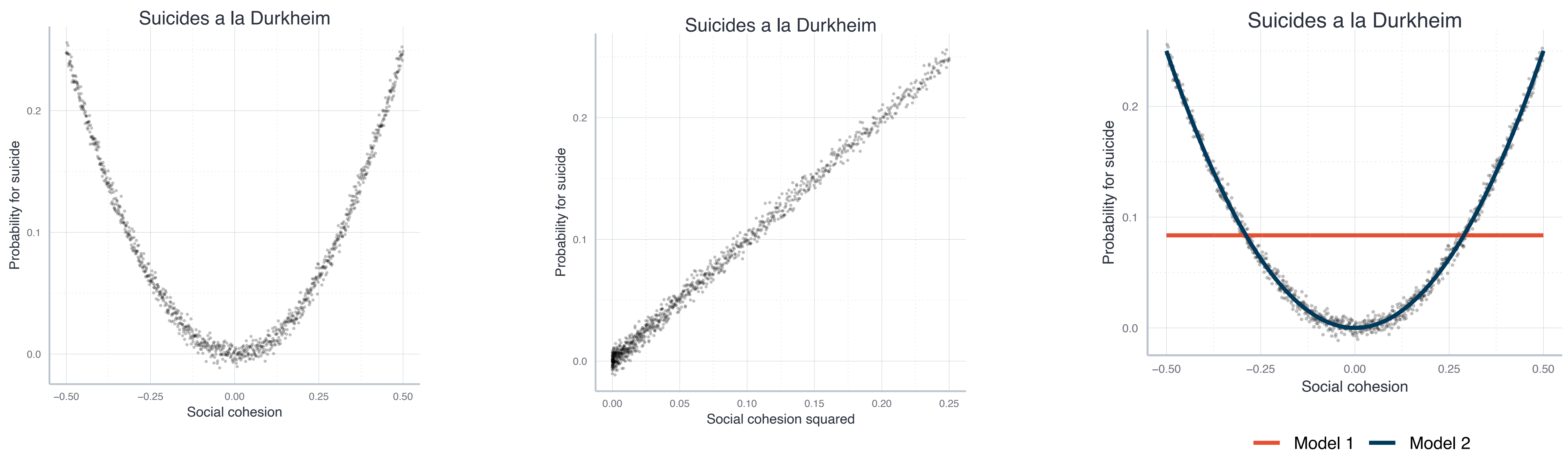
Linear regression and nonlinear relationships

Linear regression and nonlinear relationships

- Linear regression is a parametric approach
    - Focus on linear models → assumes a **linear relationship**

Meaning: **linearity in parameters**

- Fitting a linear model to nonlinear relationships is misleading, except…
    - …we transform the data to make the relationship linear

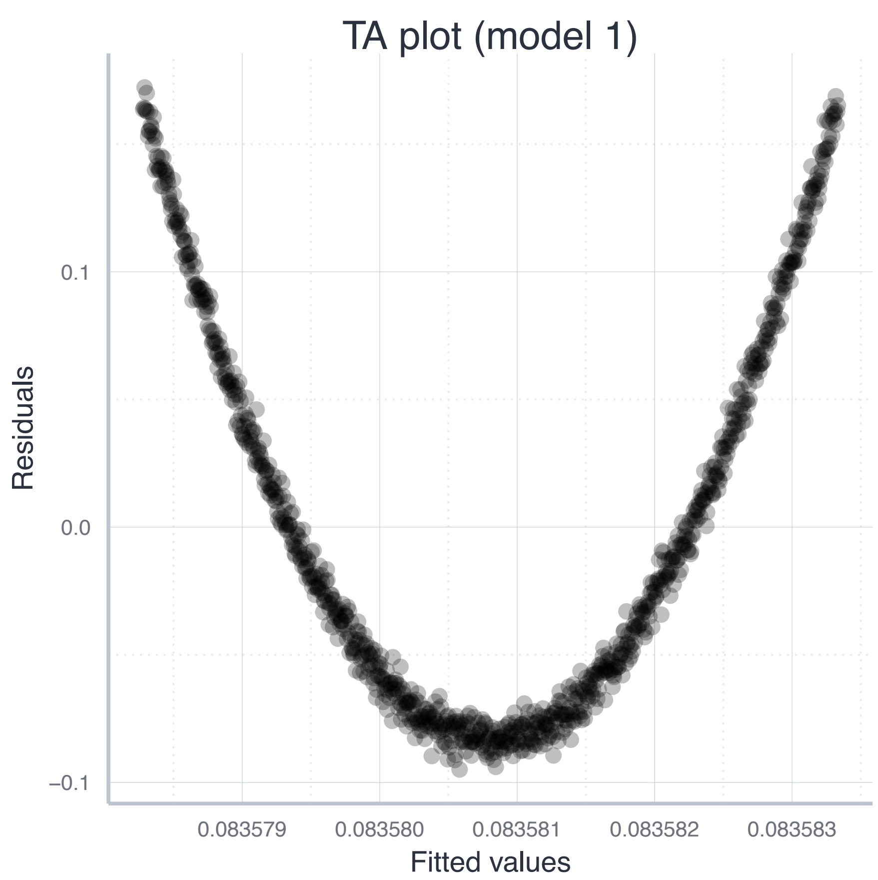


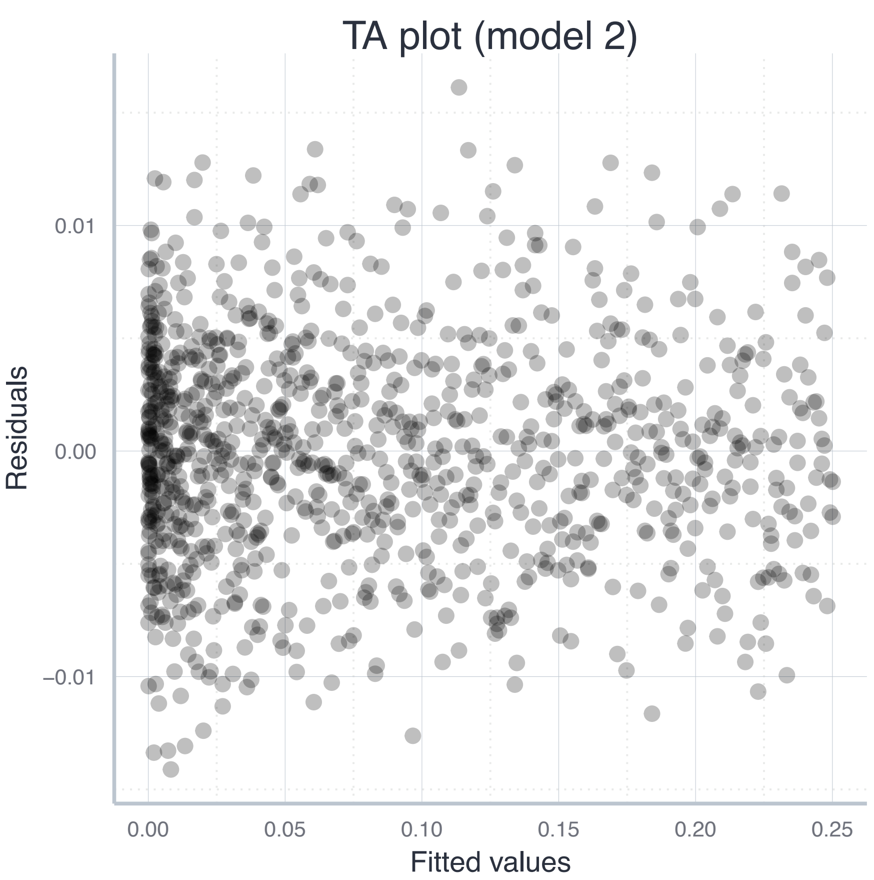
Model 1: $SuicideProb = \beta_0 + \beta_1 COH + \epsilon$

Model 2: $SuicideProb = \beta_0 + \beta_1 COH + \beta_2 COH^2 + \epsilon$

Linear regression and nonlinear relationships
The Tukey-Anscombe Plot

- How to decide whether transformation was successful?
- The residuals should not show any structure → **Tukey-Anscombe Plot**
  - x-Axis: predicted values (`predict()`), y-axis: residuals (`residuals()`):

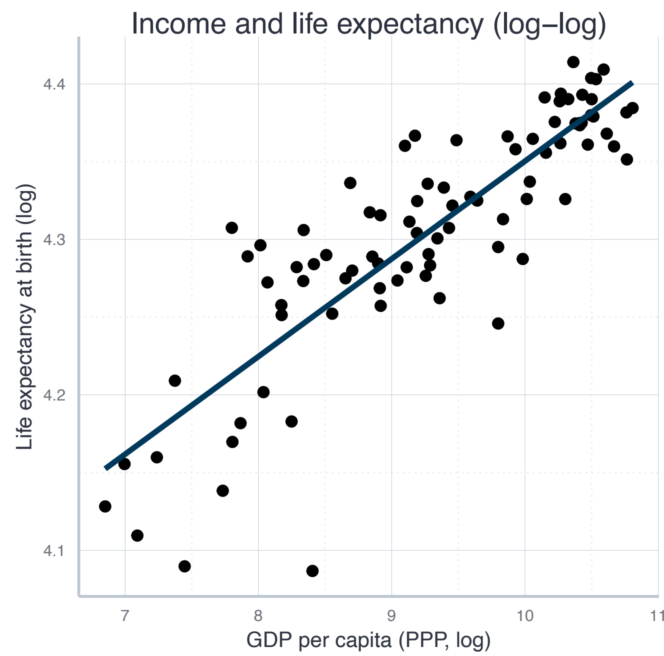


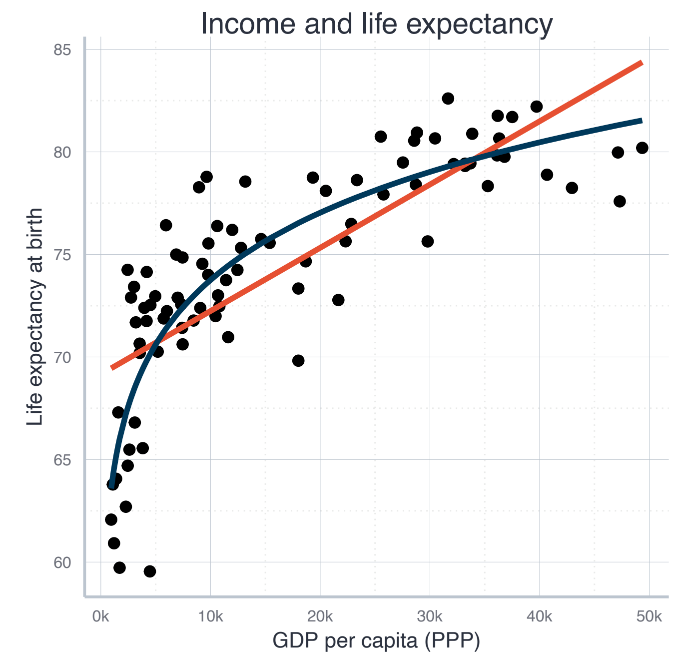


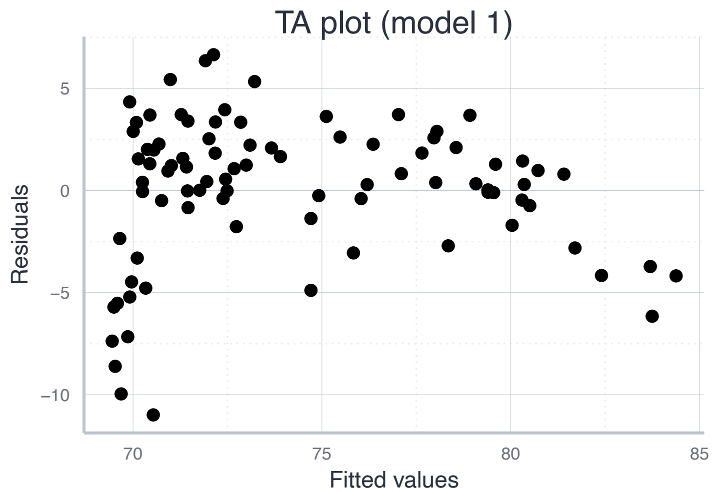
Linear regression and nonlinear relationships

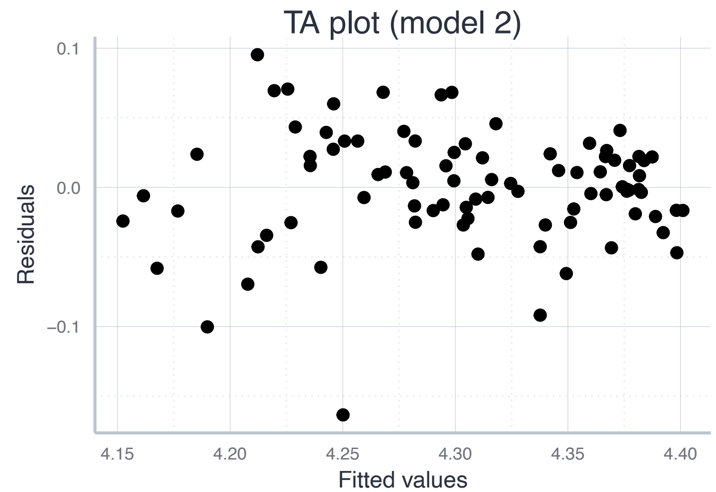
Linearising exponential relationship with logs

- Another very common transformation is taking logs → linearises otherwise exponential relationships:









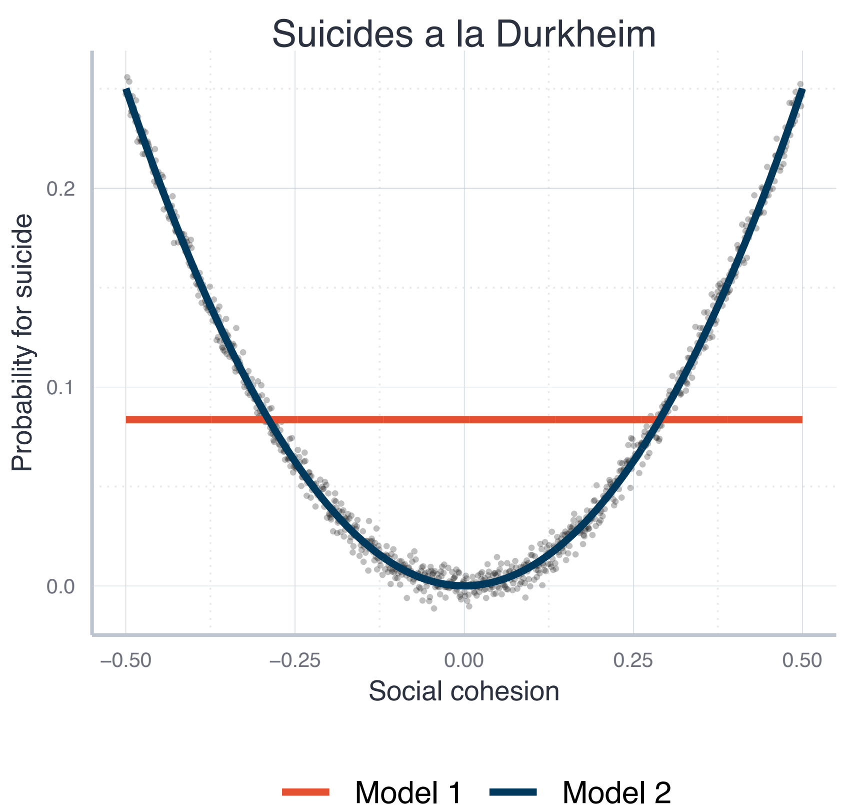
Linear regression and nonlinear relationships
Interpreting models with transformed variables: logs

- Not all relationships can be linearised
  - Sometimes linear regression is just not the right tool!
- Transformation of the variables must be considered in interpretation:

Model	Equation	Interpretation
Level-Level	$y = \beta_0 + \beta_1 x_1$	Change in x by 1 unit comes with change in y by β_1 units
Log-Level	$\ln(y) = \beta_0 + \beta_1 x_1$	Change in x by 1 unit comes with change in y by $100 \cdot \beta_1 \%$
Level-Log	$y = \beta_0 + \beta_1 \ln(x_1)$	Change in x by 1% comes with change in y by $\beta_1 / 100$
Log-Log	$\ln(y) = \beta_0 + \beta_1 \ln(x_1)$	Change in x by 1% comes with change in y by $\beta_1 \%$

Linear regression and nonlinear relationships
Interpreting models with quadratic terms

- Not all relationships can be linearised
  - Sometimes linear regression is just not the right tool!
- Transformation of the variables must be considered in interpretation:



	Model 1	Model 2
(Intercept)	0.084	0.000
Social cohesion	0.000	0.000
I(\`Social cohesion\`^2)		1.000
R2	0.000	0.996

The change in the slope of `Social Cohesion`

Categorial variables: Simple regression

Using categorical variables

- So far we only worked with numerical and **continuous variables**
  - Income, prices, consumption,...
- But there are other types of variables, e.g. **categorial data**
  - Gender, continent of origin, employment status,...
- In the following we want to learn how to consider categorial data as **explanatory variables**
  - If you have categorial variables on the LHS → different estimation methods
- Let us illustrate the procedure using the data on life expectancy, but focus on the role of different continents
  - Data: `DataScienceExercises::gdplifexp2007`
  - Variables of interest: `continent`, `lifeExp`, and `gdpPercap`

Exploratory analysis

```
── Data Summary ────────────────────────
                           Values
Name                       life_exp
Number of rows             142
Number of columns          3
Key                        NULL
_______________________
Column type frequency:
  factor                   1
  numeric                  2
________________________
Group variables            None

── Variable type: factor ──────────────────────────────────────────────────────────────
  skim_variable n_missing complete_rate ordered n_unique top_counts
1 continent             0             1 FALSE          5 Afr: 52, Asi: 33, Eur: 30, Ame: 25

── Variable type: numeric ─────────────────────────────────────────────────────────────
  skim_variable n_missing complete_rate   mean     sd   p0   p25   p50   p75   p100 hist
1 lifeExp               0             1   67.0   12.1 39.6  57.2  71.9  76.4   82.6 ▁▁▂▃▇
2 gdpPercap             0             1 11680. 12860. 278. 1625. 6124. 18009. 49357. ▇▁▁▁▁
```

No problems with missing data, 142 observations in total

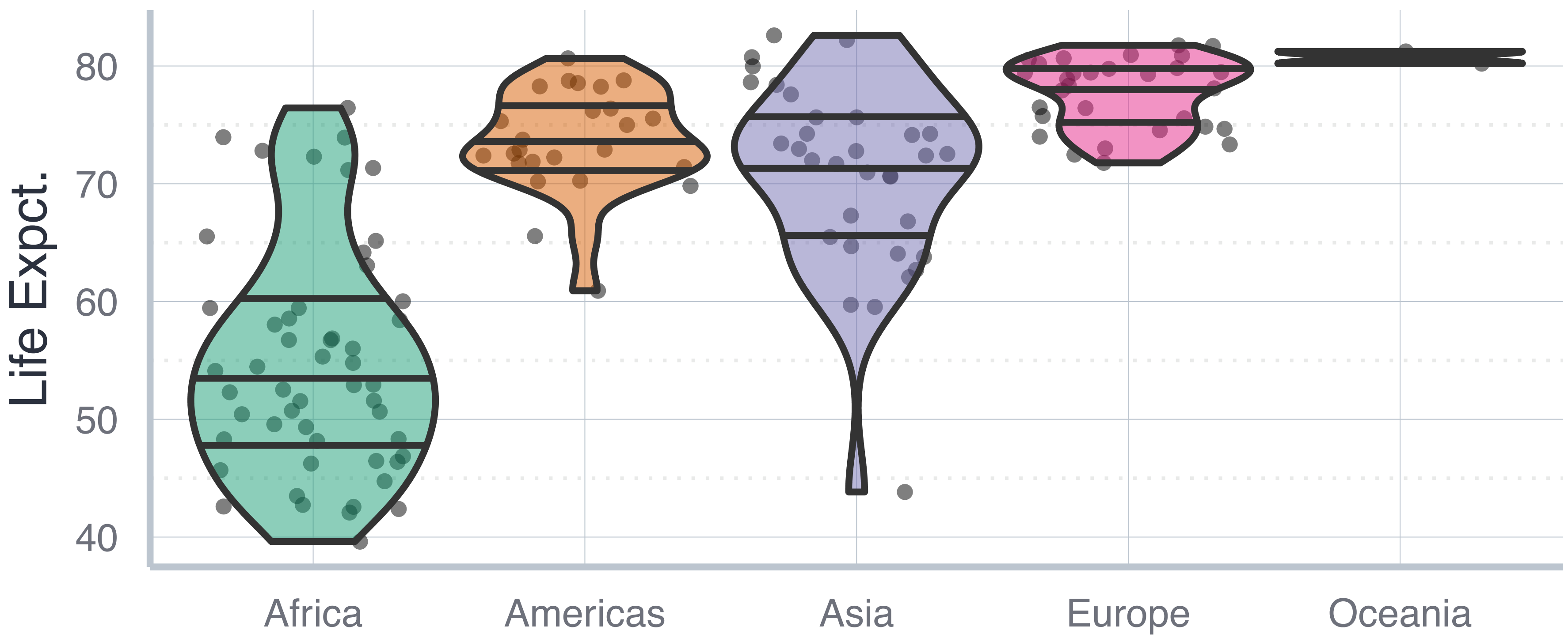
There are five different continents with Africa comprising most countries (52)

Mean and median differ considerable due to skewed distribution of the variables!

- Note: `continent` was saved as `character`, but we transformed it into `factor`

Exploratory analysis

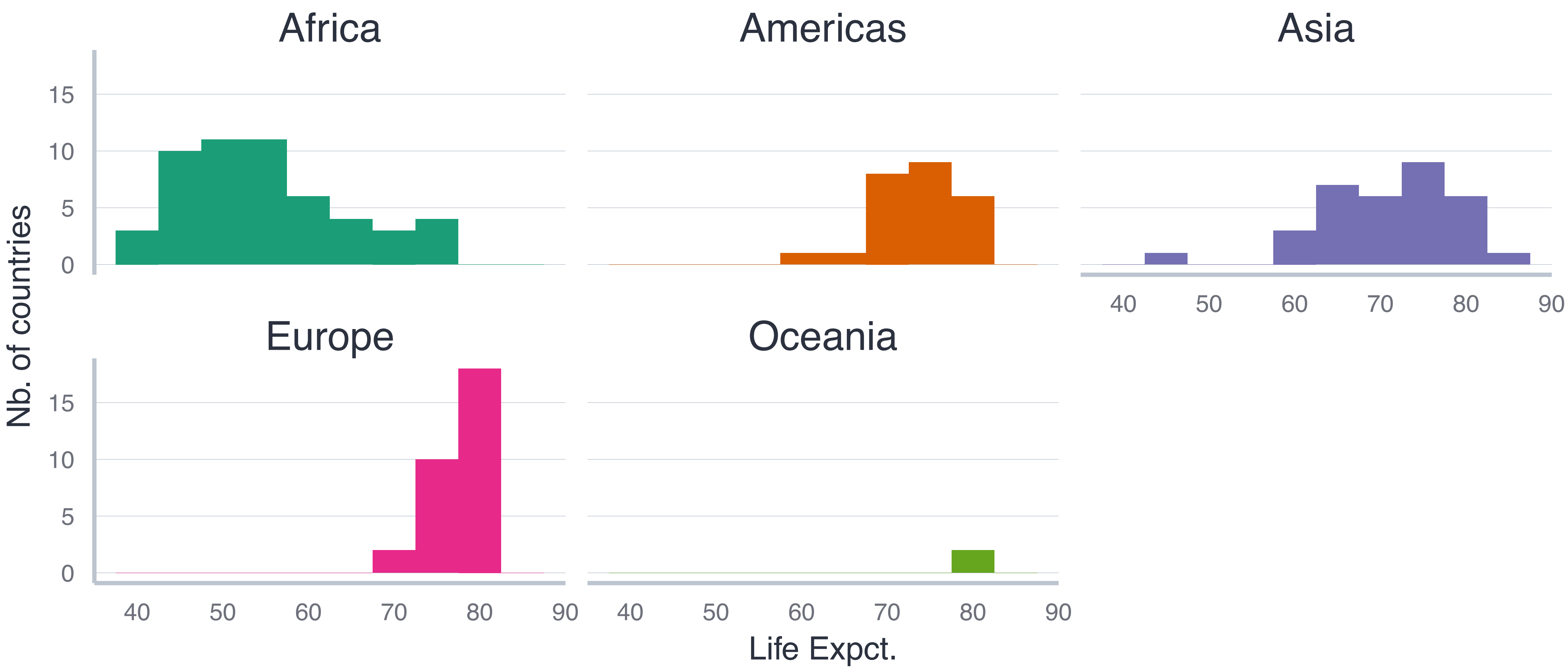
- We see considerable differences also within continents:



- Especially Oceania will be hard to interpret since it comprises only two countries

Exploratory analysis

- To look at the distribution within countries, histograms are also useful:



- For categorial variables, fitting a regression line has a different meaning

Fitting a model with categorical variables

- The notation for a model with a categorical variable on the RHS is similar...
  - ...but the technical implementation is quite different
- We write:

$$lifeExp = \beta_0 + \beta_1 \cdot CONT + \epsilon$$

- We estimate:

$$lifeExp = \beta_0 + \beta_{Am.} \cdot \mathbb{I}_{Am.}CONT + \beta_{As.} \cdot \mathbb{I}_{As.}CONT + \beta_{Eu.} \cdot \mathbb{I}_{Eu.}CONT + \beta_{Oc.} \cdot \mathbb{I}_{Oc.}CONT + \epsilon$$

- $\mathbb{I}_x(X)$ is an **indicator function:** takes the value 1 if $X = x$ and zero otherwise
  - $\mathbb{I}_{Am.}(CONT) = 1$ iff $CONT$ equals $Am.$ (i.e. Americas), and 0 otherwise
  - There are four indicator functions → four continents (plus one as a baseline level)
- The estimates must always be interpreted against a baseline value
  - Here: the first factor level, i.e. Africa

Interpreting a model with categorical variables

$$lifeExp = \beta_0 + \beta_{Am.} \cdot \mathbb{1}_{Am.}CONT + \beta_{As.} \cdot \mathbb{1}_{As.}CONT + \beta_{Eu.} \cdot \mathbb{1}_{Eu.}CONT + \beta_{Oc.} \cdot \mathbb{1}_{Oc.}CONT + \epsilon$$

- Lets consider the results from estimating this formula one by one:
  - Note that the code for the regression remains `lm(lifeExp~continent)`

```
> cont_linmod <- lm(lifeExp~continent, data = life_exp)
> get_regression_table(cont_linmod)
# A tibble: 5 × 7
  term                estimate
  <chr>                  <dbl>
1 intercept               54.8
2 continent: Americas     18.8
3 continent: Asia         15.9
4 continent: Europe       22.8
5 continent: Oceania      25.9
```

```
  continent lifeExp_mean diff_africa
  <fct>            <dbl>       <dbl>
1 Africa            54.8           0
2 Americas          73.6        18.8
3 Asia              70.7        15.9
4 Europe            77.6        22.8
5 Oceania           80.7        25.9
```

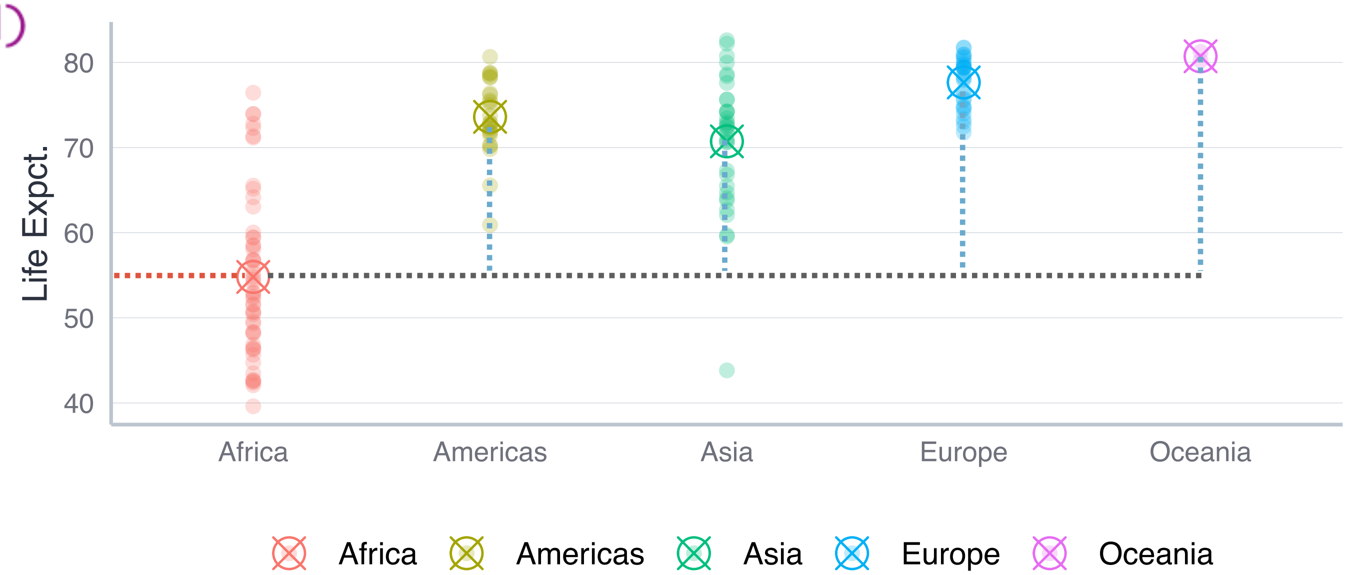
- The intercept corresponds to the mean value of the baseline category
  - The other estimates correspond to the deviation of the group mean from this baseline

Interpreting a model with categorical variables

$$lifeExp = \beta_0 + \beta_{Am.} \cdot \mathbb{1}_{Am.}CONT + \beta_{As.} \cdot \mathbb{1}_{As.}CONT + \beta_{Eu.} \cdot \mathbb{1}_{Eu.}CONT + \beta_{Oc.} \cdot \mathbb{1}_{Oc.}CONT + \epsilon$$

- Lets consider the results from estimating this formula one by one:
  - Note that the code for the regression remains `lm(lifeExp~continent)`

```
> cont_linmod <- lm(lifeExp~continent, data = life_exp)
> get_regression_table(cont_linmod)
# A tibble: 5 × 7
  term                 estimate
  <chr>                   <dbl>
1 intercept                54.8
2 continent: Americas      18.8
3 continent: Asia          15.9
4 continent: Europe        22.8
5 continent: Oceania       25.9
```



- The intercept corresponds to the mean value of the baseline category
  - The other estimates correspond to the deviation of the group mean from this baseline

Quick recap

- The result of the following regression model...

$$SUGAR = \beta_0 + \beta_1 KIND + \epsilon$$

```
lm(`residual sugar` ~ kind, data = wine_data)
```

- ...is as follows:

term	estimate
<chr>	<dbl>
intercept	2.54
kind: white	3.85

- The variables are as follows:
  - `` `residual sugar` ``: the amount of sugar left in the wine
  - `kind`: the kind of wine, red or white

- How would you interpret the estimated coefficients?

Categorical variables: Multiple regression

Introduction

- How to consider both continuous and categorical variables?
- Two cases: an **interaction model**, and a **parallel slope model**
  - Note: both also occur in the case of continuous variables
- Example: data set on the prices of economics journals: `DataScienceExercises::econjournals`
  - Only consider journals that published at least 10 papers and cost under 5000 USD per year: `dplyr::filter(papers>10, sub_price<5000)`
- **Main interest:** what is the impact of the paper length on the subscription price? Are there differences between profit and nonprofit publishers?

The parallel slopes model

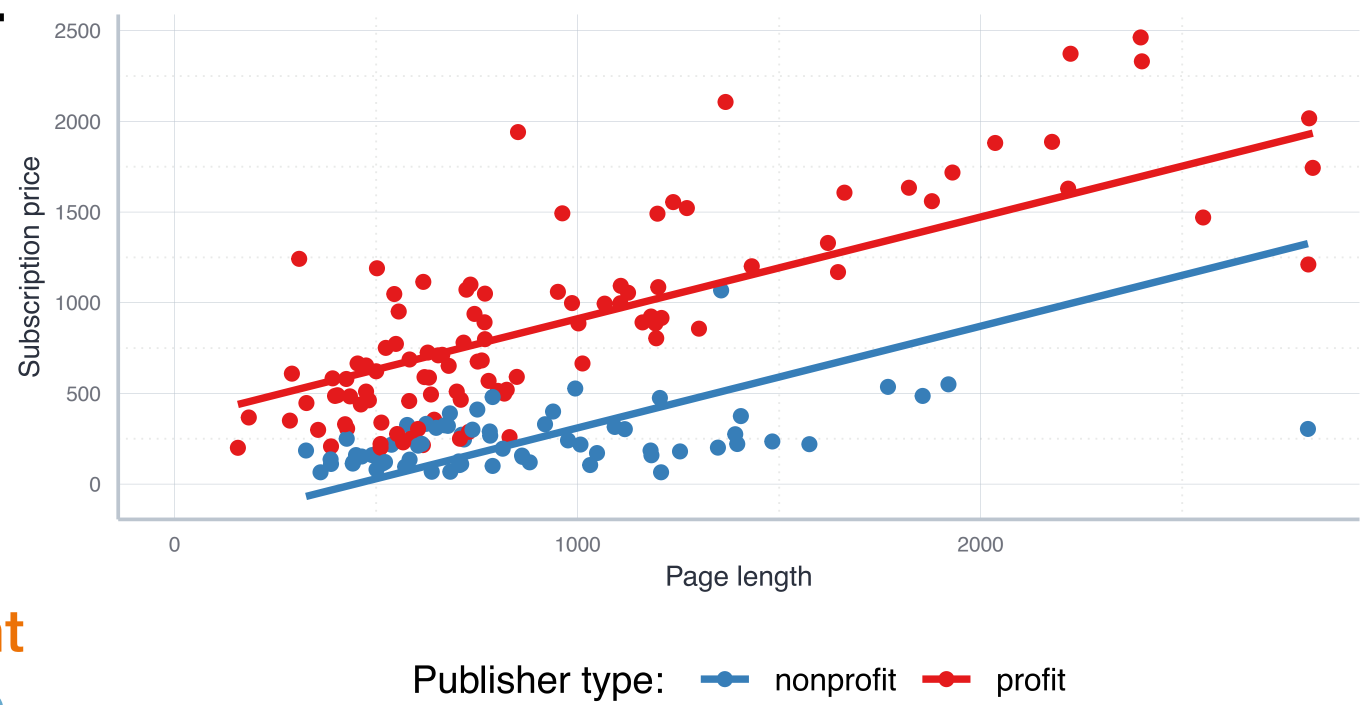
- The variables `pages_py` and `sub_price` are continuous, the variable `publisher_type` is categorical

- What if we simple add both explanatory variables to the RHS?

```
lm(sub_price~pages_py+publisher_type)
```

- Lets look at the resulting estimates:

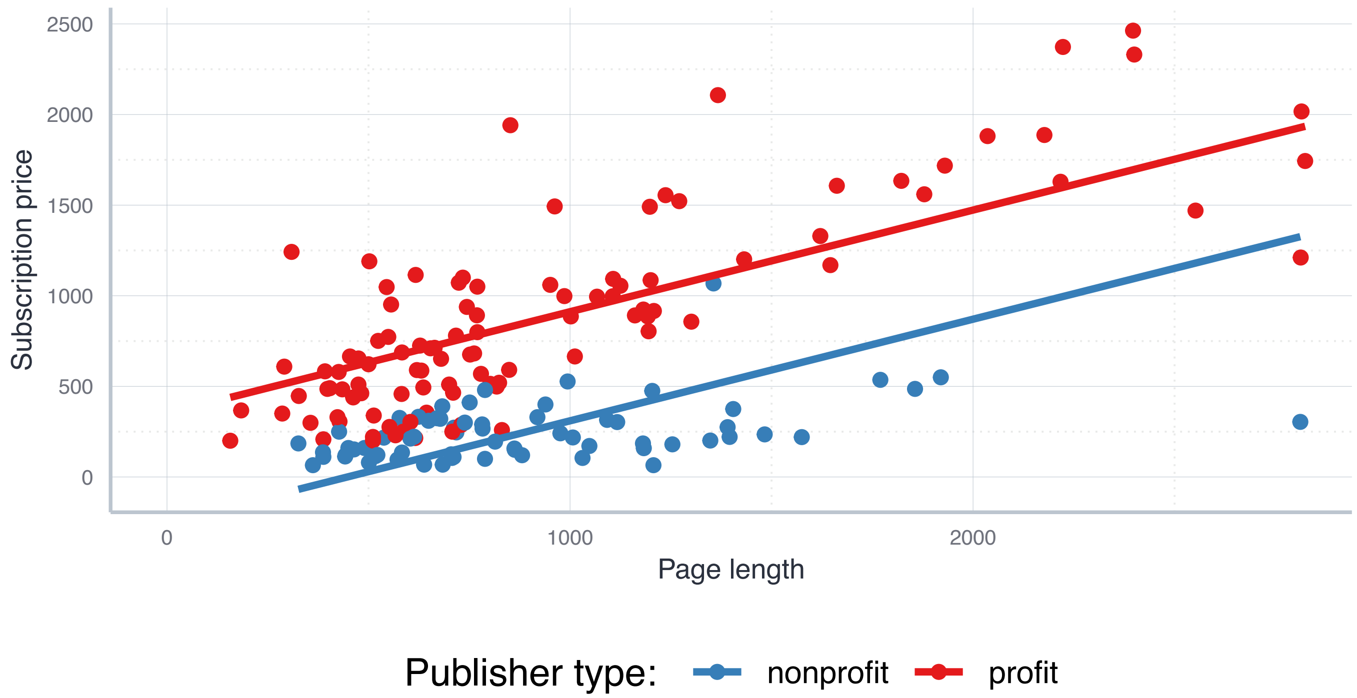
	term	estimate
	<chr>	<dbl>
1	intercept	-251.
2	pages_py	0.561
3	publisher_type: profit	602.



The categorical variables correspond to different intercepts, but each group has the same slope

The parallel slopes model

- Journals from non-profit publishers are cheaper
- An additional page comes with the same increase in journal price
- Visual inspection: relationship might differ across groups 🧐




```
  term                    estimate
  <chr>                      <dbl>
1 intercept                -251.
2 pages_py                    0.561
3 publisher_type: profit   602.
```

To capture the idea that the association between page length and price differs across groups we need an **interaction model**

The interaction model

- The interaction model is more complex: it does not assume that sloper are the same in the different groups → variables interact with each other

- Technically, we just replace the **+** by an ***** in the model formula:

```
lm(sub_price~pages_py*publisher_type)
```

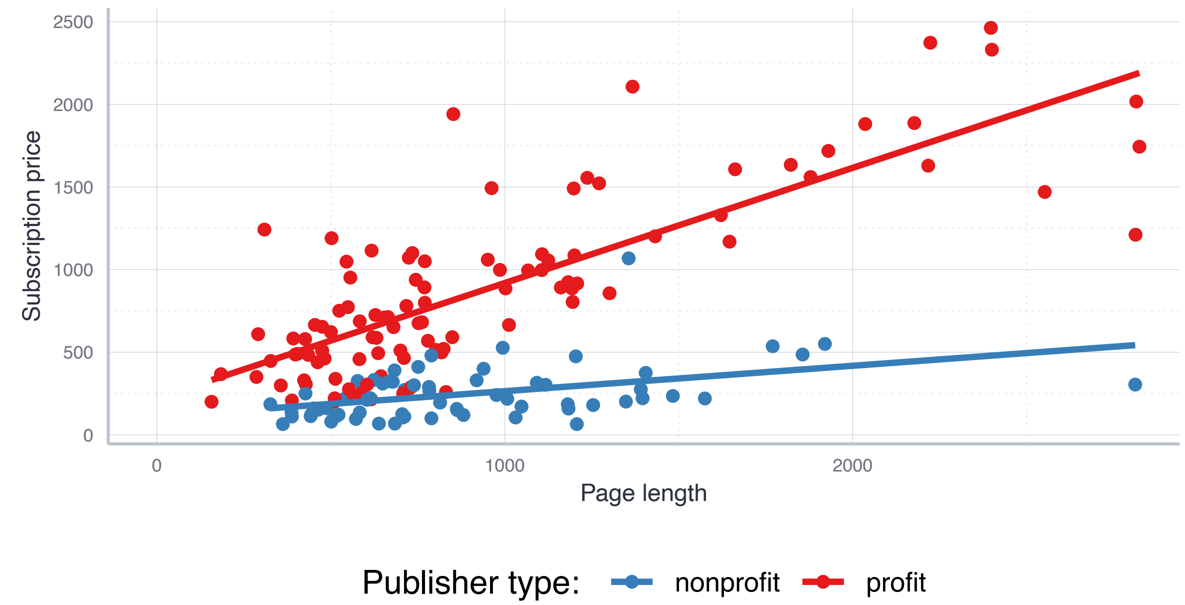
term	estimate
<chr>	<dbl>
intercept	111.
pages_py	0.154
publisher_type: profit	111.
pages_py:publisher_typeprofit	0.543

- There is **one more parameter to estimate** than in the PSM

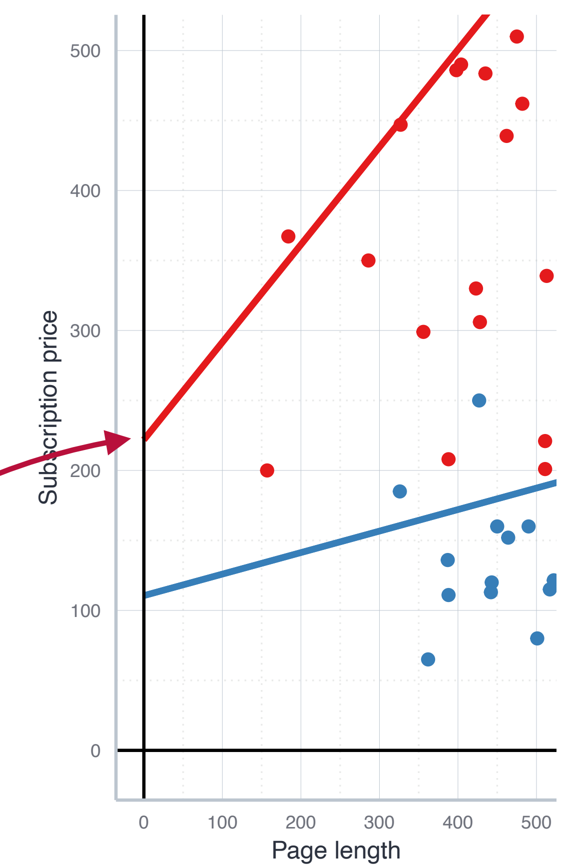
- But the plot suggests that this additional complexity is warranted: for-profit publisher charge more per additional page

The interaction model
Interpretation

term	estimate
<chr>	<dbl>
intercept	111.
pages_py	0.154
publisher_type: profit	111.
pages_py:publisher_typeprofit	0.543

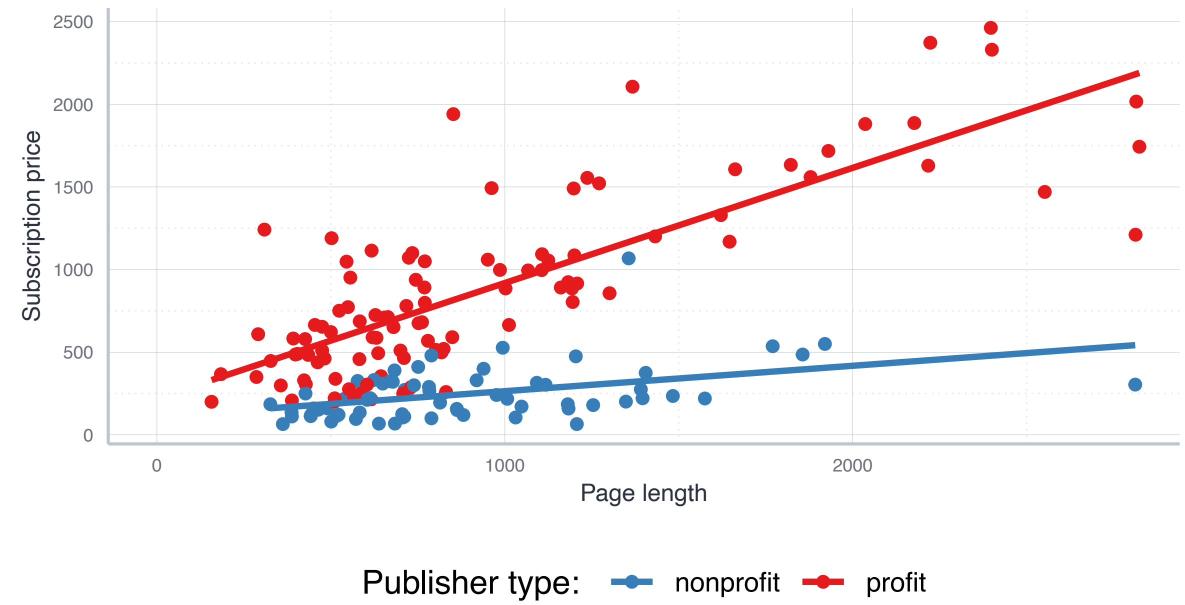


- The estimate **`intercept`** is the intercept only for the reference group → 111
- The estimate **`pages_py`** gives the slope only for the reference group → 0.154
- The estimate **`publisher_type:profit`** gives the difference in the intercept for the profit group
  - `intercept + publisher_type:profit = 222`

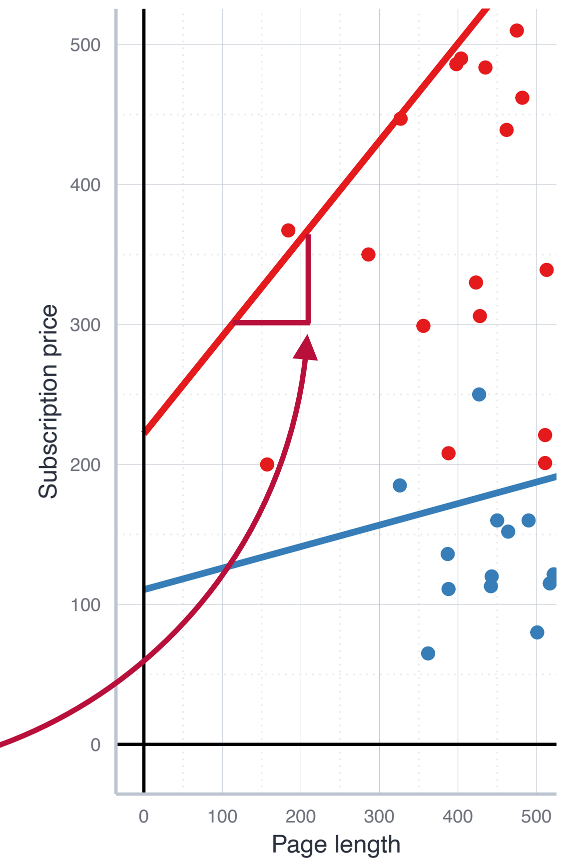


The interaction model
Interpretation

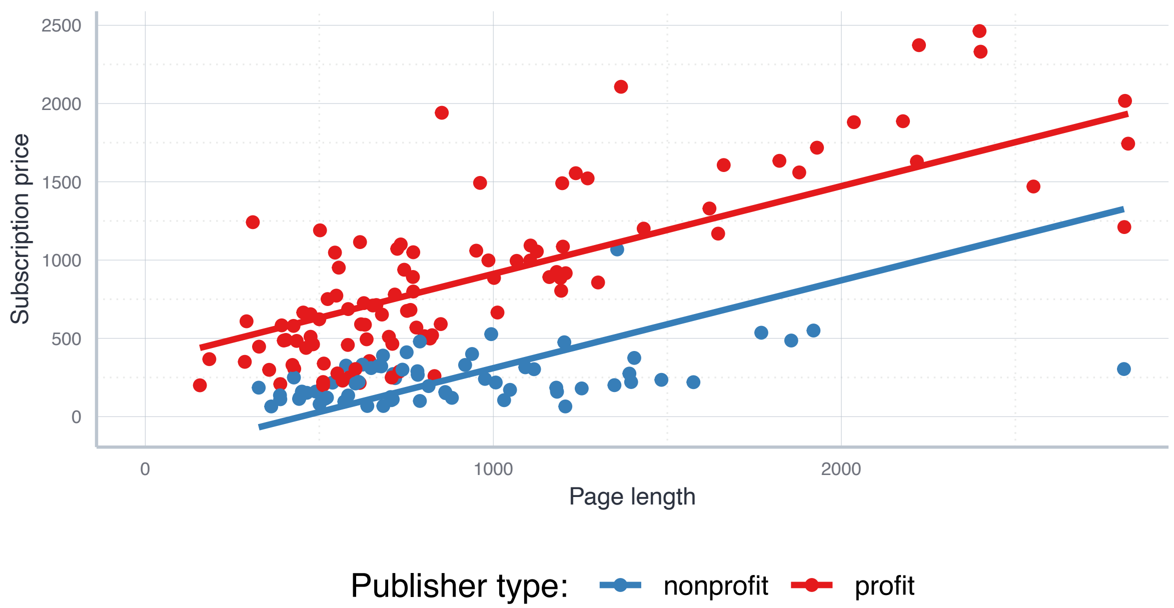
term	estimate
<chr>	<dbl>
intercept	111.
pages_py	0.154
publisher_type: profit	111.
pages_py:publisher_typeprofit	0.543



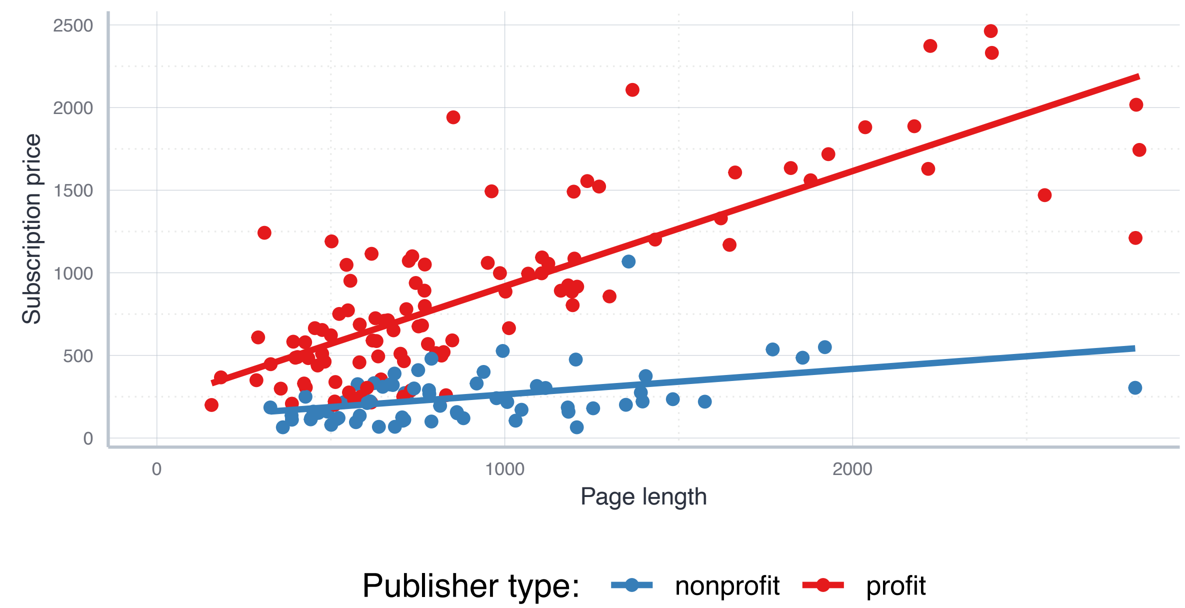
- The estimate **`intercept`** is the intercept only for the reference group → 111
- The estimate **`pages_py`** gives the slope only for the reference group → 0.154
- The estimate **`pages_py:profit`** gives the difference in the slope for the profit group
  - `pages_py+pages_py:publisher_typeprofit=0.697`



The interaction and parallel slopes model



	term	estimate
	<chr>	<dbl>
1	intercept	-251.
2	pages_py	0.561
3	publisher_type: profit	602.

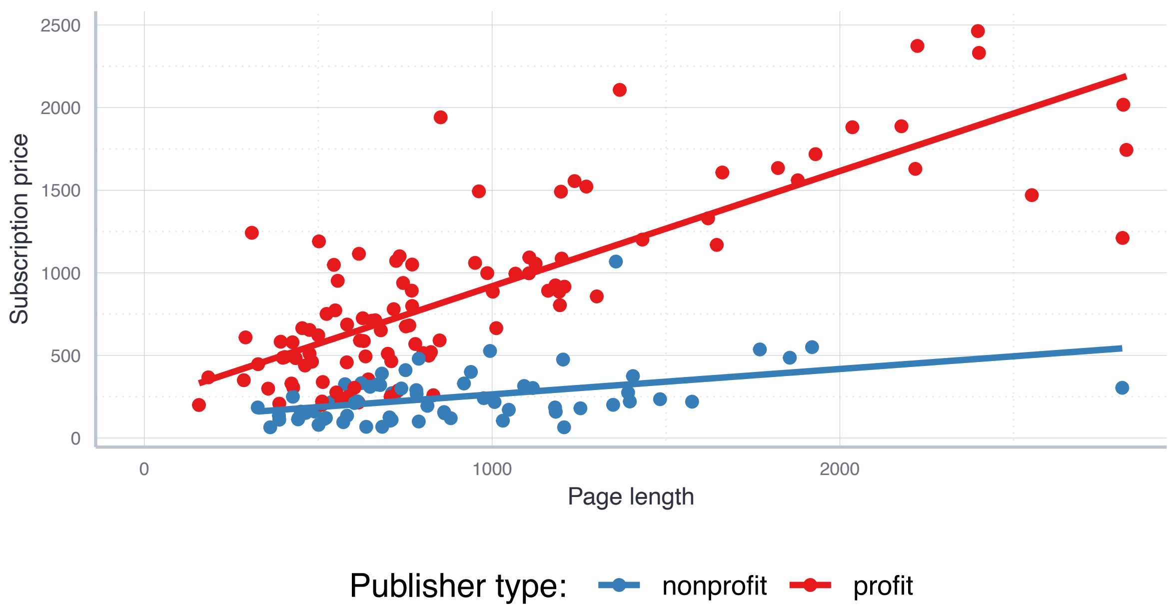


term	estimate
<chr>	<dbl>
intercept	111.
pages_py	0.154
publisher_type: profit	111.
pages_py:publisher_typeprofit	0.543

- As a general rule of thumb: the PSM is better if nothing suggests that slopes differ → then the estimation is more efficient
  - In other cases, its safer to use the interaction mode
  - We learn how to test for the right model in later sessions

Model selection in the multiple variable case

Model selection using visual inspection



- Visual inspecting the estimated model is mandatory and often very insightful: the interaction model is clearly preferable due to different slopes

Model selection using R^2

- You can use R^2 as one argument for model selection, i.e. when you need to decide which models works best for your purpose at hand
- Compare, for instance, the R^2 of the PSM and interaction model we estimated before:
- `summary(journal_linmod_intct)[["r.squared"]]`: 0.75
- `summary(journal_linmod_psm)[["r.squared"]]`: 0.68
- The reference to R^2 confirms our impression that the more complex interaction model is warranted
- But: using R^2 in the multiple regression context can be misleading: adding more variables typically increases the R^2 for purely mathematical reasons

Model selection using R^2

- To see why consider the formal definition of R^2:

$$R^2 = \frac{ESS}{TSS} = \frac{\sum_{i=1}^{N}\left(\hat{y}_i - \bar{y}\right)^2}{\sum_{i=1}^{N}\left(y_i - \bar{y}\right)^2}$$

  - An additional explanatory variable never changes TSS, but mostly increases ESS at least a bit → bias towards 'too complex' models

- There is an alternative, the adjusted R^2, denoted as $\bar{R}^2$:

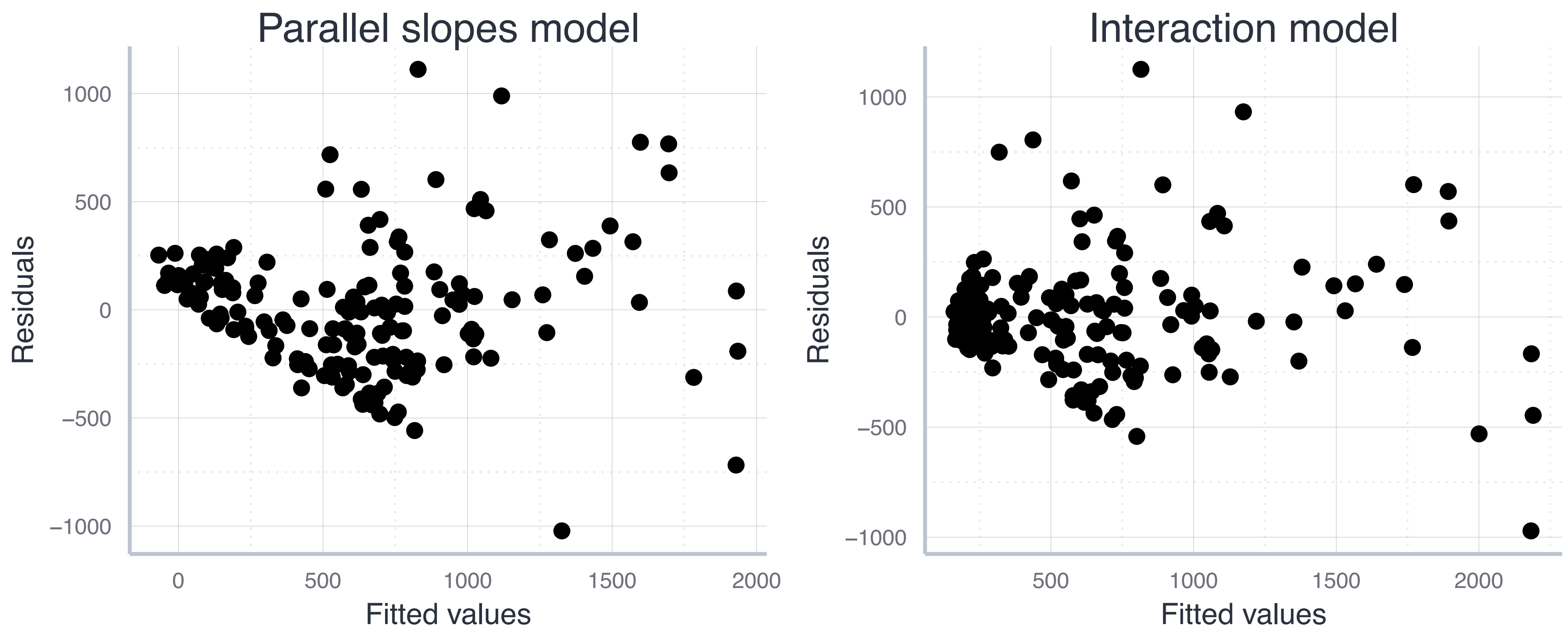
$$\bar{R}^2 = 1 - \frac{\sum_{i=1}^{n} e^2/(N-K-1)}{\sum_{i=1}^{n}(y_i - \bar{y})^2/(N-1)}$$

  - Here, N is the number of observations and K the nb. of estimated parameters

Model selection using R^2

- $\bar{R}^2$ only increases if the additional variables contribute to the explanatory power for substantial reasons
  - **Drawback**: cannot interpret it as the share of explained variation any more
- Both R^2 and $\bar{R}^2$ provide valuable information, but they should be complemented by other diagnostic tools
- In the present PSM vs. IM case, using $\bar{R}^2$ instead of R^2 does not alter the conclusion, but you find plenty other examples in the readings

Model selection based on residual plots



- Residuals more equally distributed in the interaction model
- Strong indication for heteroscedasticity → adjust p values

All arguments suggest superiority of interaction model!

Summary & outlook

Summary

- We extended the simple to the multiple regression model
- This allows us to have more than one variable on the RHS
- The interpretation of the estimates is different:
  - For every increase of 1 unit in the explanatory variable i, there is an associated decrease of, on average **and ceteris paribus**, of $\hat{\beta}_i$ units in the response
  - Ceteris paribus: holding all other variables constant
- This allows us to separate the variation in the response variable according to the different explanatory variables
- Forgetting relevant explanatory variables seems to cause problems since adding a variable changes estimates of all other variables

Summary

- We also learned about how to include **categorical variables** to regressions
- Technically this is easy, but the interpretation becomes a bit trickier
- When both continuous and categorical variables are used, we learner about the difference between **interaction and parallel slope models**
- The latter are simpler, but often the complexity of the former model is warranted
- We saw that selecting models using R^2 requires a bit more caution in the multiple regression context
- Finally, linear regression assumes linearity only in parameters
    - Adequate data transformation until residual distribution is adequate