Linear Regression

Sébastien Biass ⓘD

sebastien.biasse@unige.ch

Earth Sciences, University of Geneva 🇨🇭

Stéphane Guerrier ⓘD

Stephane.Guerrier@unige.ch

Earth Sciences & Pharmaceutical Sciences, University of Geneva 🇨🇭

Dominique Couturier ⓘD

Cancer Research UK, University of Cambridge 🇬🇧🇺🇸

Yuming Zhang ⓘD

T.H. Chan School of Public Health, Harvard University

Motivating Example: Reading Ability

Problem

An educator believes that new directed reading activities in the classroom can help elementary school students (6-12 years old) improve their reading ability. She arranged a pilot study where some students (chosen at random) of age 6 start to take part in these activities **(treatment group)**, meanwhile other students continue with the classical curriculum **(control group)**. The educator wishes to evaluate the effectiveness of these activities so all students are given a Degree of Reading Power (DRP) test, which assesses their reading ability.

Can we conclude that these new directed reading activities can help elementary school students improve their reading ability?

Motivating Example: Reading Ability

Data

```python
import pandas as pd

reading = pd.read_csv("https://raw.githubusercontent.com/ELSTE-Master/Data-Science/main/Data/reading.csv")

control = reading.loc[reading["group"] == "Control", "score"]
print(control.head())
```
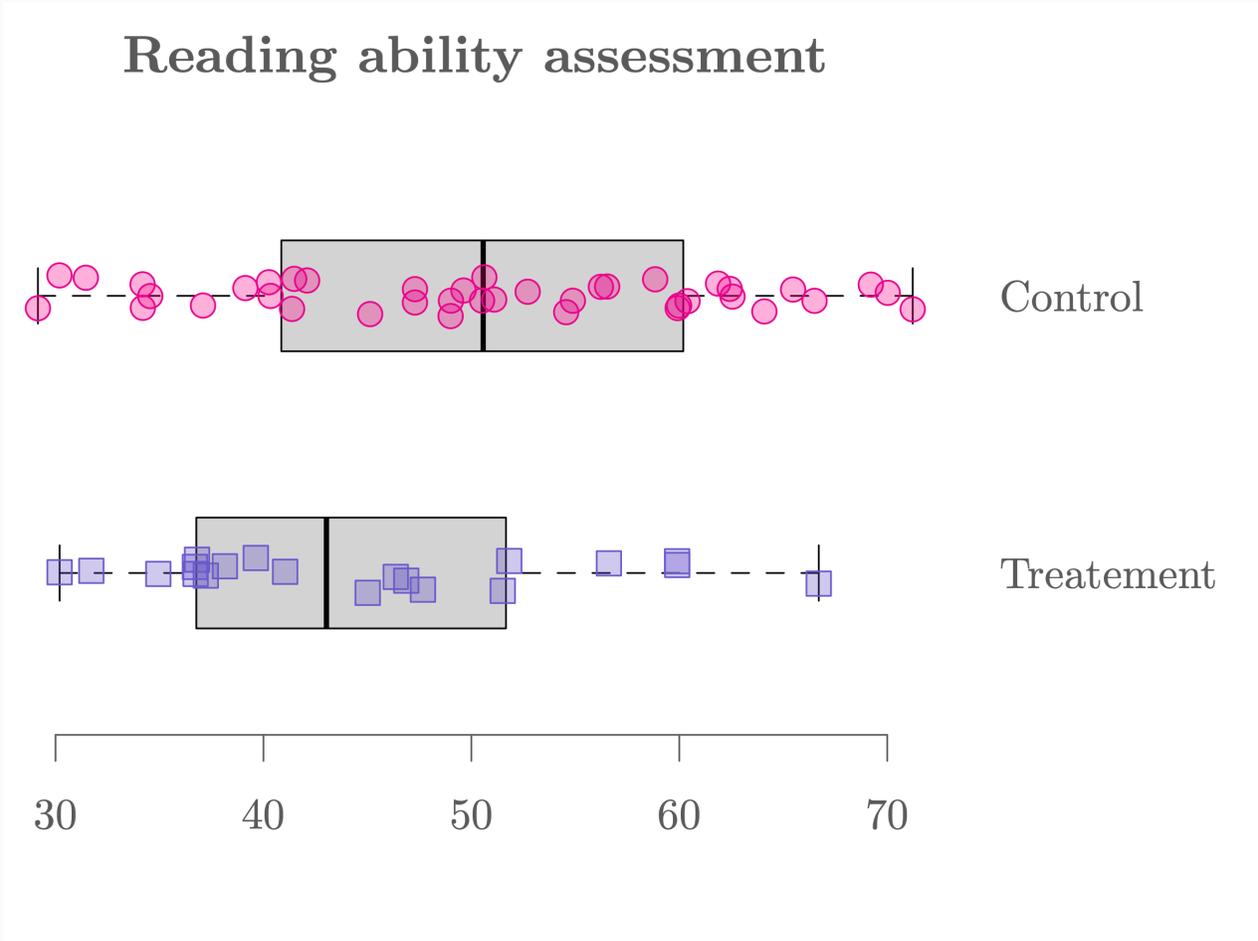
```
0    59.988809
1    60.394867
2    70.023947
3    41.477072
4    29.154968
Name: score, dtype: float64
```

```python
treatment = reading.loc[reading["group"] == "Treatment", "score"]
print(treatment.head())
```

```
40    36.805579
41    36.702008
42    59.893620
43    31.722885
44    41.042415
Name: score, dtype: float64
```

Motivating Example: Reading Ability

Graph

Motivating Example: Reading Ability

Test

1. Define hypotheses: $H_0 : \mu_T = \mu_C$ and $H_a : \mu_T > \mu_C$.

2. Define α: We consider $\alpha = 5\%$.

3. Compute p-value: p-value = **97.12%** (see Python code tab for details).

4. Conclusion: We have p-value > α so we cannot reject the null hypothesis at the significance level of 5%.

Remark: Notice that the p-value for the test with opposite hypotheses is actually $100\% - 97.12\% = 2.88\%$ $< \alpha$ (Why? 🤔). So we conclude, at the significance level of 5%, that the classical curriculum without these new directed reading activities actually improve students' reading ability more compared to the curriculum with these activities. 😬

Motivating Example: Reading Ability

Python *Code*

```python
from scipy import stats
t_test_reading, p_value_reading = stats.ttest_ind(
    treatment, control, equal_var=True, alternative='greater'
)

print(f"t-test: {t_test_reading:.4f}")
```

```
t-test: -1.8553
```

```python
print(f"p-value: {p_value_reading:.4f}")
```

```
p-value: 0.9657
```

Is our analysis comprehensive?

The educator points out that only students of 6-8 years old have participated in the new directed reading activities. In other words, in the sample she collected, the students in the treatment group are only of age 6 to 8, whereas the students in the control group vary from 6 to 12 years old. Is age a potential explanation to the difference we observe among the students' reading ability?

To make sure that the analysis is reliable, she includes the age information of the students, which can be accessed as follows:

```
1  treatment_age = reading.loc[reading["group"] == "Treatment", "age"]
2  control_age   = reading.loc[reading["group"] == "Control", "age"]
```

Should age be taken into account?
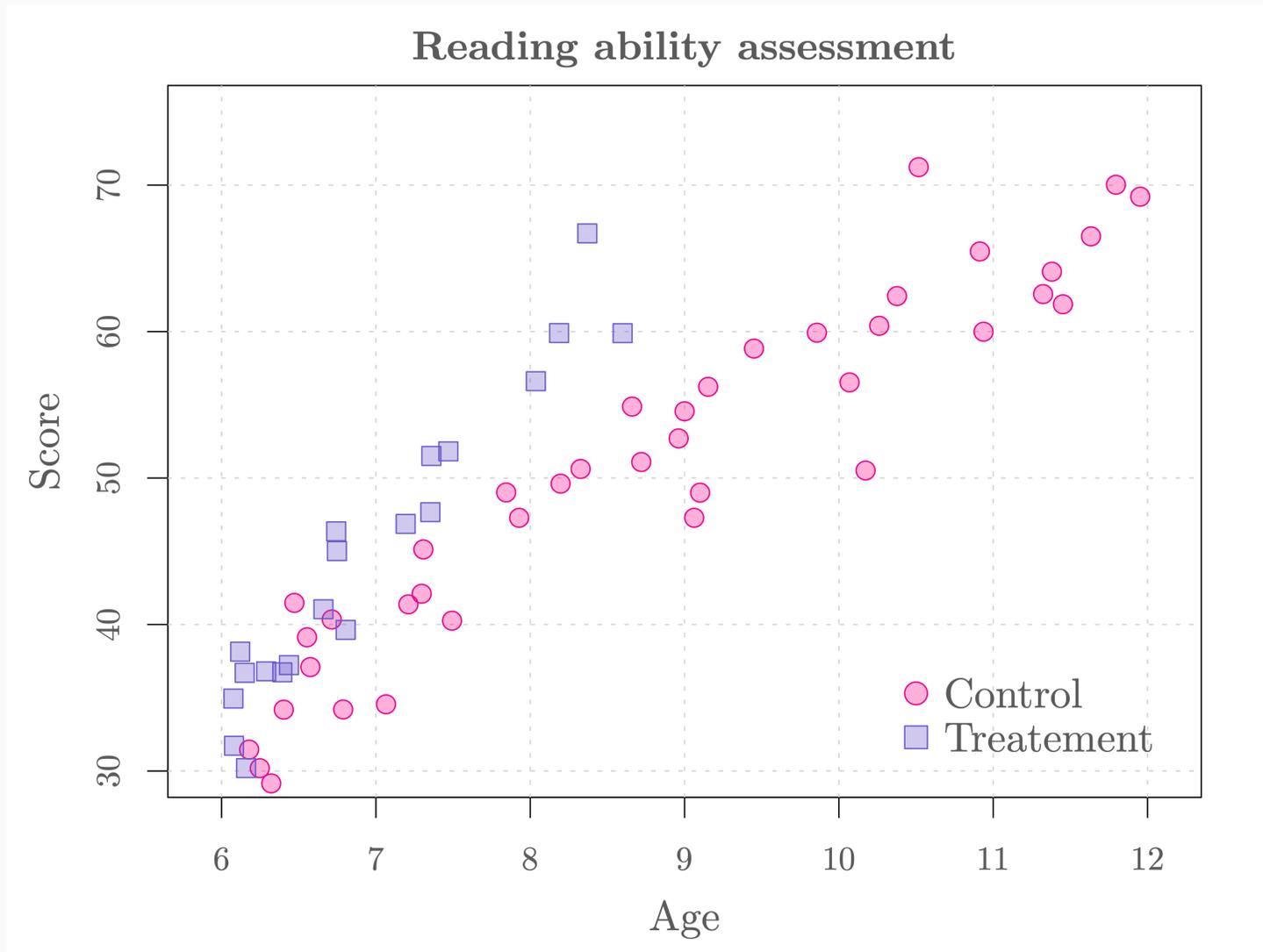
Regression analysis

- Regression analysis corresponds to a set of statistical methods for estimating the relationships between a response variable Y of primary interest (also called the *outcome variable*) and some explanatory variables $X_1, \ldots, X_p$ (also called *covariates*, *regressors*, *features* or *predictors*).

- The relationship between the response variable Y and the covariates is not deterministic and we model the conditional expected value (i.e. $\mathbb{E}[Y|X_1, \ldots, X_p]$).

- Therefore, we consider the following (general) model:

$$Y_i = \mathbb{E}[Y_i|X_{i1}, \ldots, X_{ip}] + \varepsilon_i,$$

where $\mathbb{E}[\varepsilon_i] = 0$ and $i = 1, \ldots, n$.

- Example: $\mathbb{E}[\text{reading abilities}_i| \text{ age}_i, \text{treatment}_i, \ldots]$.

Linear regression

- The most common form of regression analysis is linear regression, in which the conditional expected value $\mathbb{E}[Y|X_1, \ldots, X_p]$ takes the form

$$\mathbb{E}[Y_i|X_{i1}, \ldots, X_{ip}] = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p.$$

and $\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$.

- Our general model can be expressed as

$$Y_i = \mathbb{E}[Y_i|X_{i1}, \ldots, X_{ip}] + \varepsilon_i = \beta_0 + \sum_{j=1}^{p} \beta_j X_{ij} + \varepsilon_i,$$

and therefore,

$$Y_i \sim N\left(\beta_0 + \sum_{j=1}^{p} \beta_j X_{ij}, \sigma^2\right).$$

Linear regression

Therefore, this approach makes two (strong) assumptions:

1. The conditional expected value $\mathbb{E}[Y|X_1, \ldots, X_p]$ is assumed to be a linear function of the covariates.

2. The errors are assumed to be iid Gaussian, i.e. $\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$, at least when the sample size is small to medium.
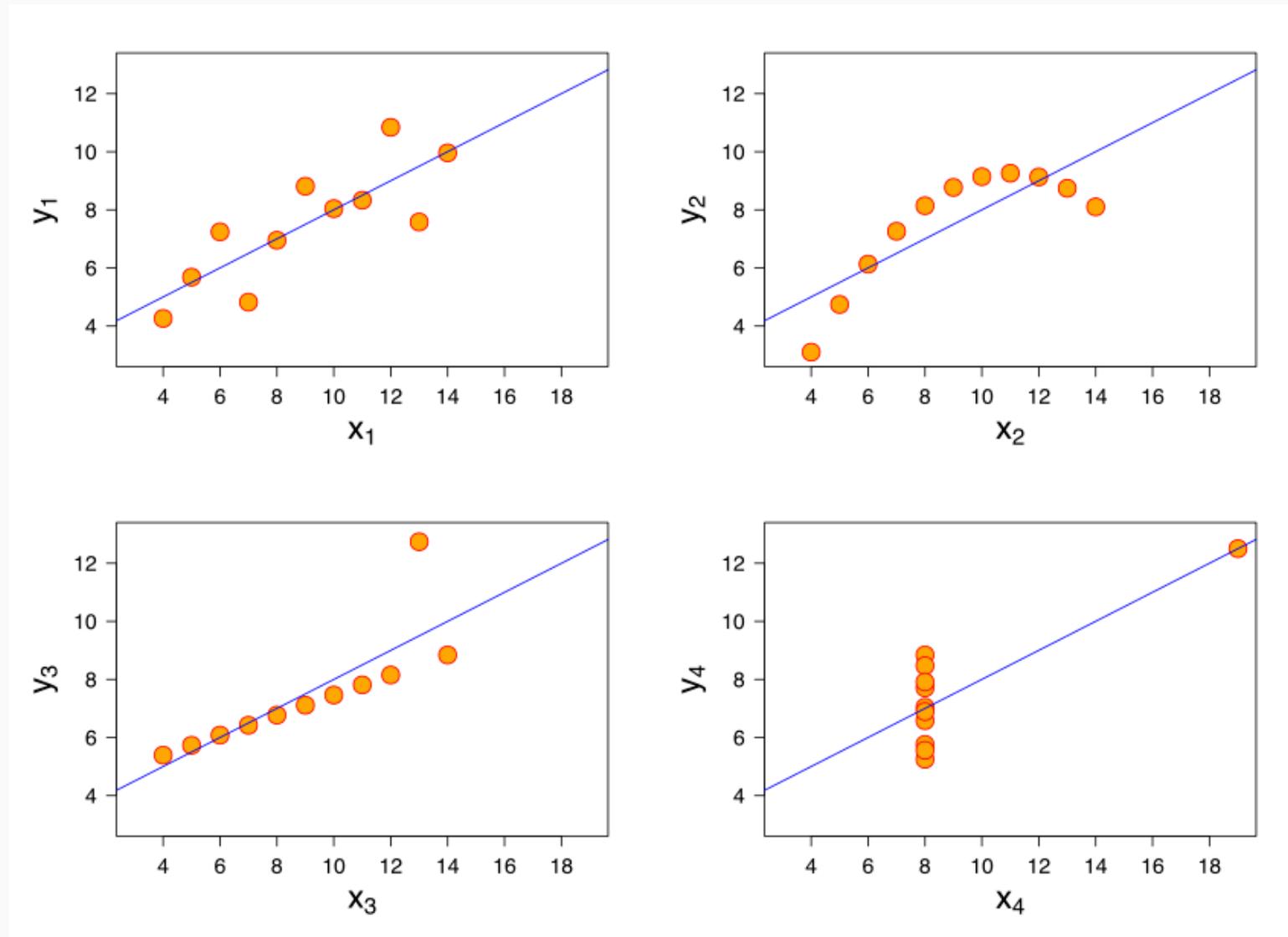
⚠️ In practice, it is important to assess if these assumptions are plausible.

The parameters of the model (i.e. $\beta_0, \beta_1, \ldots, \beta_p$ and σ^2) can be estimated by several methods. The most commonly used is the least squares approach where $\beta_0, \beta_1, \ldots, \beta_p$ are chosen such that

$$\sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} \left(Y_i - \mathbb{E}[Y_i|X_{i1}, \ldots, X_{ip}] \right)^2 = \sum_{i=1}^{n} \left(Y_i - \beta_0 - \sum_{j=1}^{p} \beta_j X_{ij} \right)^2$$

is minimized, which then allows to estimate σ^2 further on.

Anscombe's quartet

Example: Reading ability assessment

In the reading ability example, we can formulate a linear regression model (without interaction) as follows:

$$\text{Score}_i = \beta_0 + \beta_1 \text{Group}_i + \beta_2 \text{Age}_i + \epsilon_i, \quad \epsilon_i \overset{iid}{\sim} N(0, \sigma^2).$$

- Score_i: score of the DRP test of the i-th student.
- Group_i: indicator of participation of the new directed reading activities for the i-th student (i.e. $\text{Group}_i = 1$ if participate and $\text{Group}_i = 0$ if not participate).
- Age_i: age of the i-th student (related to *time since start of treatment*).

With this model the two groups can be compared as the age effect is taken into account. The goal of the educator is now to assess if β_1 is significantly larger than 0.

Example: Reading ability assessment

Python *Code*

Python function:

```python
import statsmodels.formula.api as smf

# Fit linear model using formula syntax
model = smf.ols(formula="y ~ x1 + x2 + ... + xp", data=mydata).fit()

# Show results
print(model.summary())
```

Example: Reading ability assessment

Output

```
                           OLS Regression Results
==============================================================================
Dep. Variable:                  score   R-squared:                       0.859
Model:                            OLS   Adj. R-squared:                  0.854
Method:                 Least Squares   F-statistic:                     173.0
Date:                Mon, 03 Nov 2025   Prob (F-statistic):           6.16e-25
Time:                        17:34:57   Log-Likelihood:                -173.13
No. Observations:                  60   AIC:                             352.3
Df Residuals:                      57   BIC:                             358.6
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                    coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept         -7.8639      3.324     -2.366      0.021     -14.520      -1.208
group[T.Treatment] 6.3771      1.393      4.578      0.000       3.587       9.167
age                6.6457      0.370     17.985      0.000       5.906       7.386
==============================================================================
Omnibus:                        0.633   Durbin-Watson:                   2.068
Prob(Omnibus):                  0.729   Jarque-Bera (JB):                0.386
Skew:                           0.196   Prob(JB):                        0.824
Kurtosis:                       3.013   Cond. No.                         50.7
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified
```

Interpretation of coefficients

We can obtain the estimated coefficients. Specifically,

- $\hat{\beta}_0 = -7.8639$ represents the estimated baseline average score of the DRP test at birth (but what does it mean? 🤨).

- $\hat{\beta}_1 = 6.3771$ means that for a student of the same age, participating in the new directed reading activities is estimated to increase their average score of the DRP test by 6.3771.

- $\hat{\beta}_2 = 6.6457$ means that when a student receives the same treatment (either participate or not in the activities), their average score increases by 6.6457 as they become 1 year older.

Regression coefficients represent the mean change in the response variable for one unit of change in the predictor variable while holding other covariates in the model constant.

Interpretation of coefficient p-values

- We notice that for each coefficient β_j, there is a corresponding p-value associated to the (Wald t-)test of $H_0 : \beta_j = 0$ and $H_a : \beta_j \neq 0$.

- A covariate with a small p-value (typically smaller than 5%) is considered to be a significant (meaningful) addition to the model, as changes in the values of such covariate can lead to changes in the response variable.

- On the other hand, a large p-value (typically larger than 5%) suggests that the corresponding covariate is not (significantly) associated with changes in the response or that we don't have enough evidence (data) to show its effect.

- In this example, the coefficient p-value associated to the `group` covariate is 2.6×10^{-3}%. This suggests that taking into account the effect of `age`, the reading abilities of the students receiving the treatment is significantly **different** from the control group, at the significance level of 5%. But this is not what we want!

Interpretation of coefficient p-values

In the linear regression output, the coefficient p-value (which we denote as p below) corresponds to a two-sided test. We can use this result to compute the p-value of a one-sided test using the following relations:

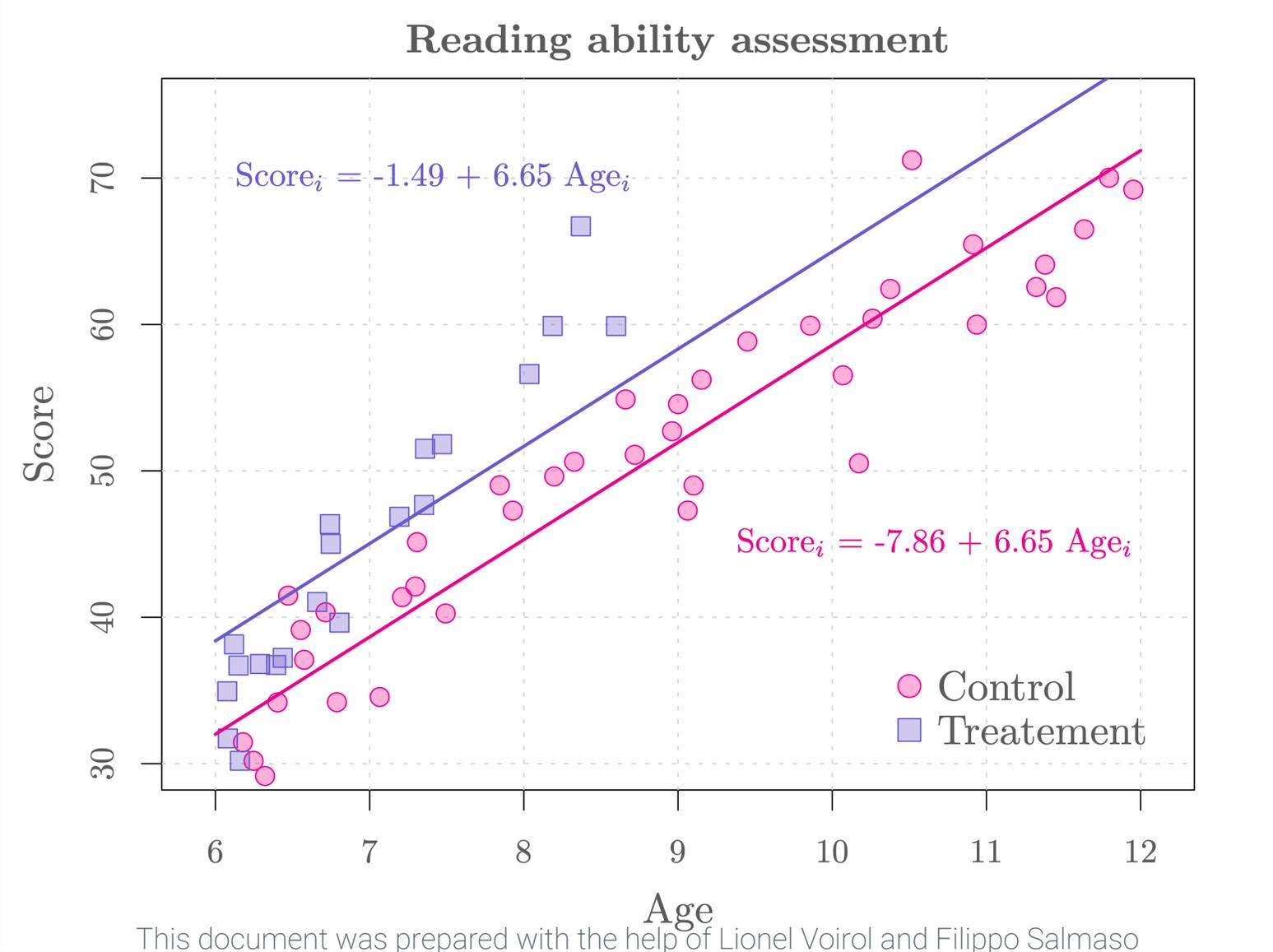|	$H_a : \beta_j > 0$	$H_a : \beta_j < 0$
$\hat{\beta}_j > 0$	✅ Correct tail (prob = $p/2$)	⚠️ Wrong tail (prob = $1 - p/2$)
$\hat{\beta}_j < 0$	⚠️ Wrong tail (prob = $1 - p/2$)	✅ Correct tail (prob = $p/2$)

✅ = rejection region corresponds to the alternative
⚠️ = rejection region is in the opposite direction of the alternative

In our example, $\beta_1 = 6.3771$ and $p = 2.6 \times 10^{-3}\%$. So the p-value of the test with hypotheses $H_0 : \beta_1 = 0$ and $H_a : \beta_1 > 0$ is $2.6 \times 10^{-3}\%/2 \approx 1.3 \times 10^{-3}\% < \alpha$. So we can conclude that these new directed reading activities can significantly improve students' reading ability compared to classical curriculum.

However, is our model plausible? 🤔

How good is our model? 🤔

Reading ability assessment

$$\text{Score}_i = -1.49 + 6.65 \text{ Age}_i$$

$$\text{Score}_i = -7.86 + 6.65 \text{ Age}_i$$

- Control
- Treatement

Score

Age

Could we use the R^2?

- The coefficient of determination, denoted as R^2 and often referred to as R-squared, corresponds to the proportion of the variance in the response variable that is "explained" by the model.

- R^2 gives certain information about the goodness of fit of a model. It measures how well the regression predictions approximate the real data points. An R^2 of 1 indicates that the regression predictions perfectly fit the data.

- However, the value of R^2 is not related to the adequacy of our model to the data.

- ⚠️ Moreover, adding new covariates to the current model always increases R^2, whether the additional covariates are significant or not. Therefore, R^2 alone cannot be used as a meaningful comparison of models with different covariates.

- The adjusted R^2 is a modification of R^2 that aims to limit this issue.

Rexthor, the Dog-Bearer!
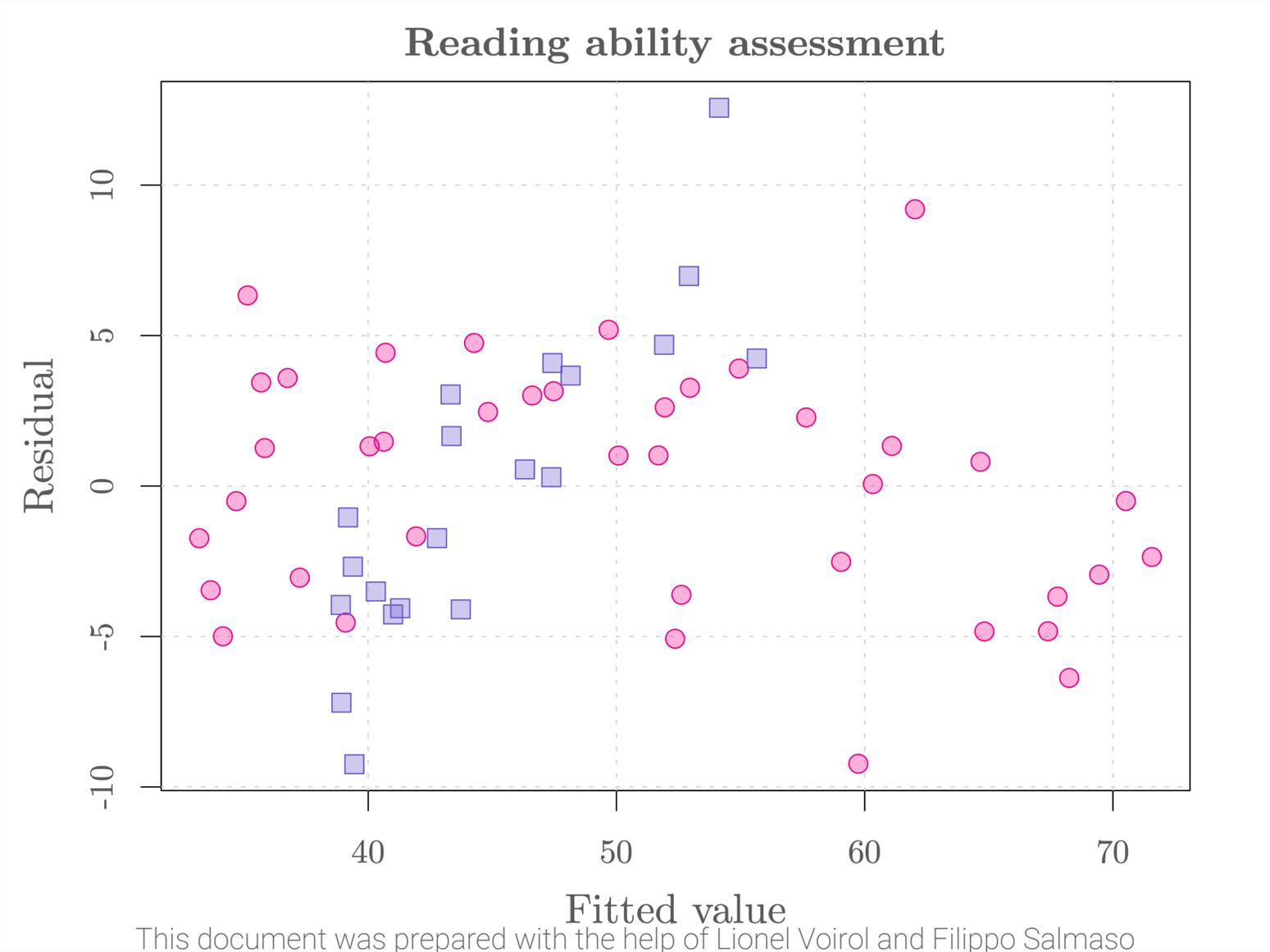

R²=0.06

REXTHOR, THE DOG-BEARER
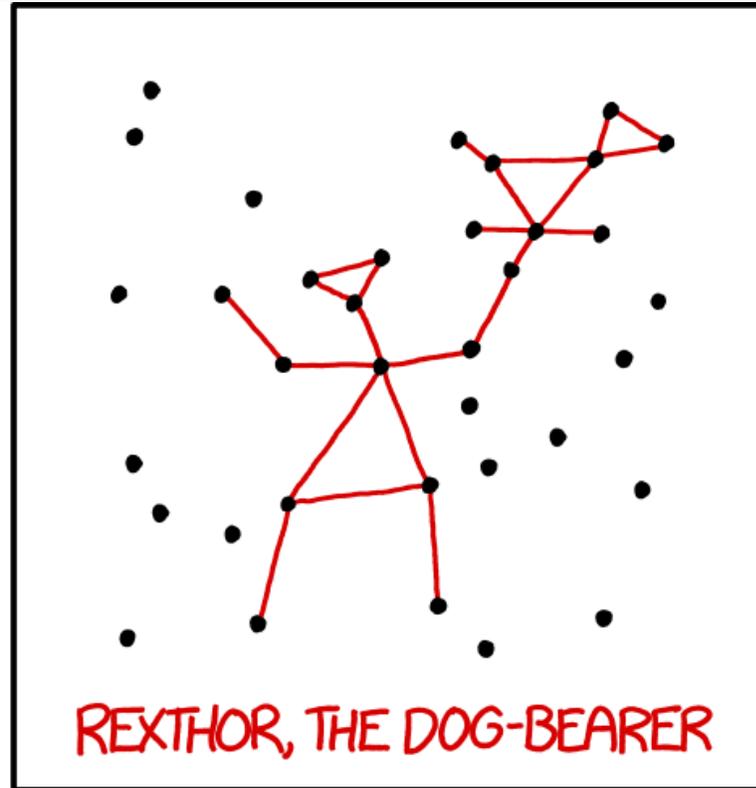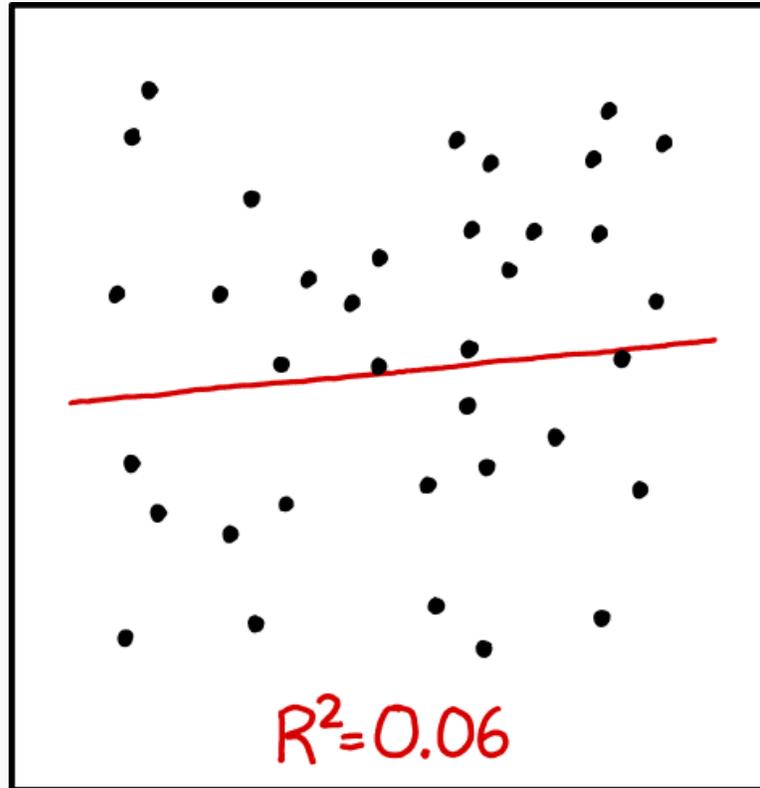
I DON'T TRUST LINEAR REGRESSIONS WHEN IT'S HARDER TO GUESS THE DIRECTION OF THE CORRELATION FROM THE SCATTER PLOT THAN TO FIND NEW CONSTELLATIONS ON IT.
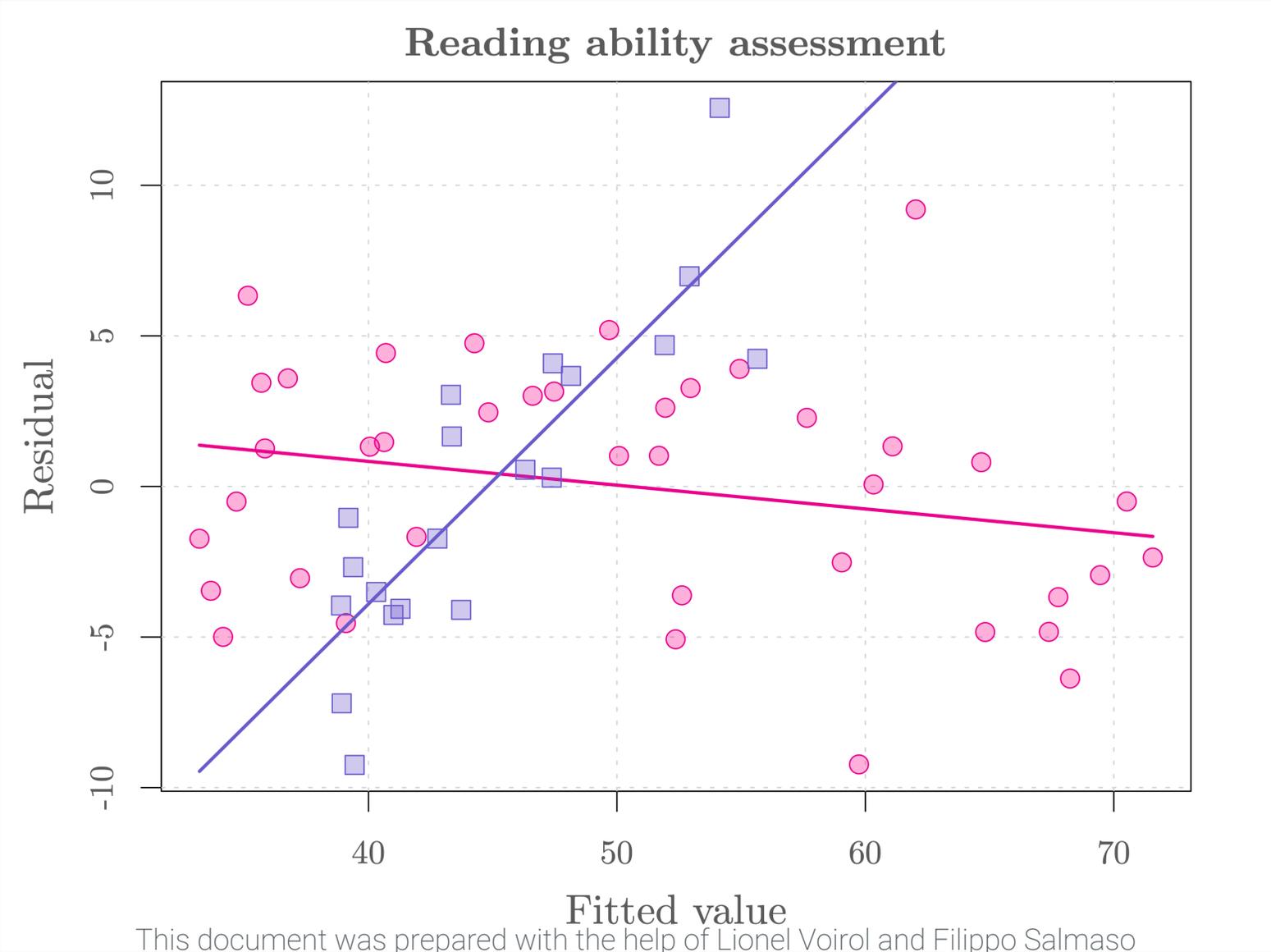
👋 If you want to know more have a look here.

Model diagnostic

Reading ability assessment

Model diagnostic ⚠️

Reading ability assessment

Model diagnostic ⚠️

Reading ability assessment

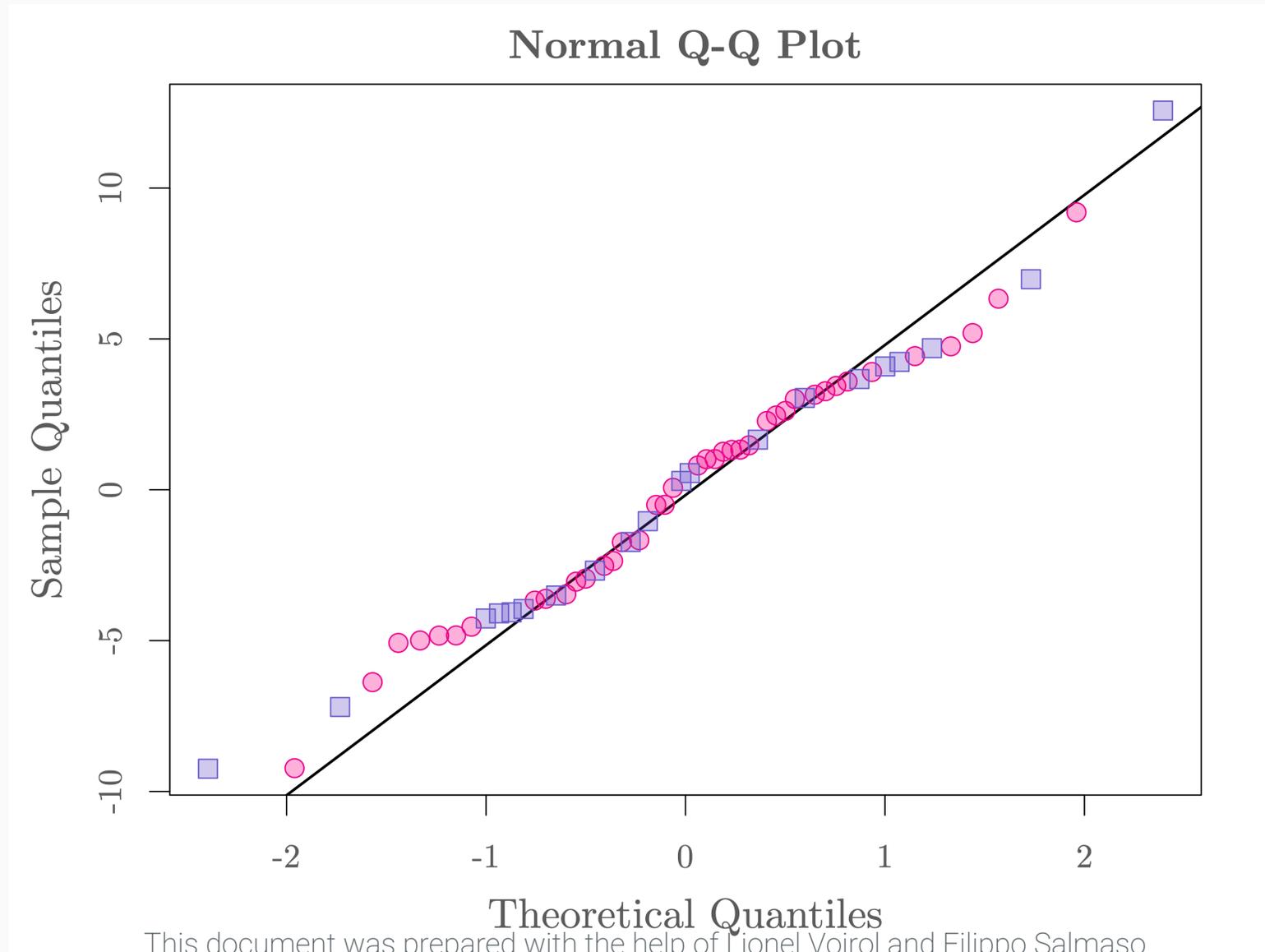# Model diagnostic ⚠️

Normal Q-Q Plot

Let's update our model

Our results suggest that the students of the group participating in these new directed reading activities progress more rapidly (which is actually more reasonable than our initial model 🤔). Therefore, we modify our model by adding an interaction term:

$$\text{Score}_i = \beta_0 + \beta_1 \text{Group}_i + \beta_2 \text{Age}_i + \beta_3 \text{Group}_i \text{Age}_i + \epsilon_i, \quad \epsilon_i \overset{iid}{\sim} N(0, \sigma^2).$$

- Score_i: score of the DRP test of the i-th student.
- Group_i: indicator of participation of the new directed reading activities for the i-th student (i.e. $\text{Group}_i = 1$ if participate and $\text{Group}_i = 0$ if not participate).
- Age_i: age of the i-th student (related to *time since start of treatment*),

The goal of the educator is now to assess if β_1 and/or β_3 are significantly larger than 0.

Example: Reading ability assessment

Python *Code*

Here is the code to fit our second model:

```python
import pandas as pd
import statsmodels.formula.api as smf

# Import data (if you haven't already)
reading = pd.read_csv("https://raw.githubusercontent.com/ELSTE-Master/Data-Science/main/Da

# Fit a linear regression model
mod2 = smf.ols(formula="score ~ group*age", data=reading).fit()

# Show summary
print(mod2.summary())
```

Example: Reading ability assessment

Output

```
                          OLS Regression Results
==============================================================================
Dep. Variable:                  score   R-squared:                       0.910
Model:                            OLS   Adj. R-squared:                  0.906
Method:                 Least Squares   F-statistic:                     189.6
Date:                Mon, 03 Nov 2025   Prob (F-statistic):           2.70e-29
Time:                        17:34:57   Log-Likelihood:                -159.45
No. Observations:                  60   AIC:                             326.9
Df Residuals:                      56   BIC:                             335.3
Df Model:                           3
Covariance Type:            nonrobust
==============================================================================
                         coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept              -3.2489      2.790     -1.164      0.249      -8.839       2.341
group[T.Treatment]    -36.0307      7.539     -4.779      0.000     -51.134     -20.928
age                     6.1207      0.311     19.693      0.000       5.498       6.743
group[T.Treatment]:age  5.9539      1.047      5.688      0.000       3.857       8.051
==============================================================================
Omnibus:                        0.210   Durbin-Watson:                   2.013
Prob(Omnibus):                  0.901   Jarque-Bera (JB):                0.005
Skew:                           0.013   Prob(JB):                        0.998
Kurtosis:                       3.036   Cond. No.                         145.
==============================================================================

Notes:
```
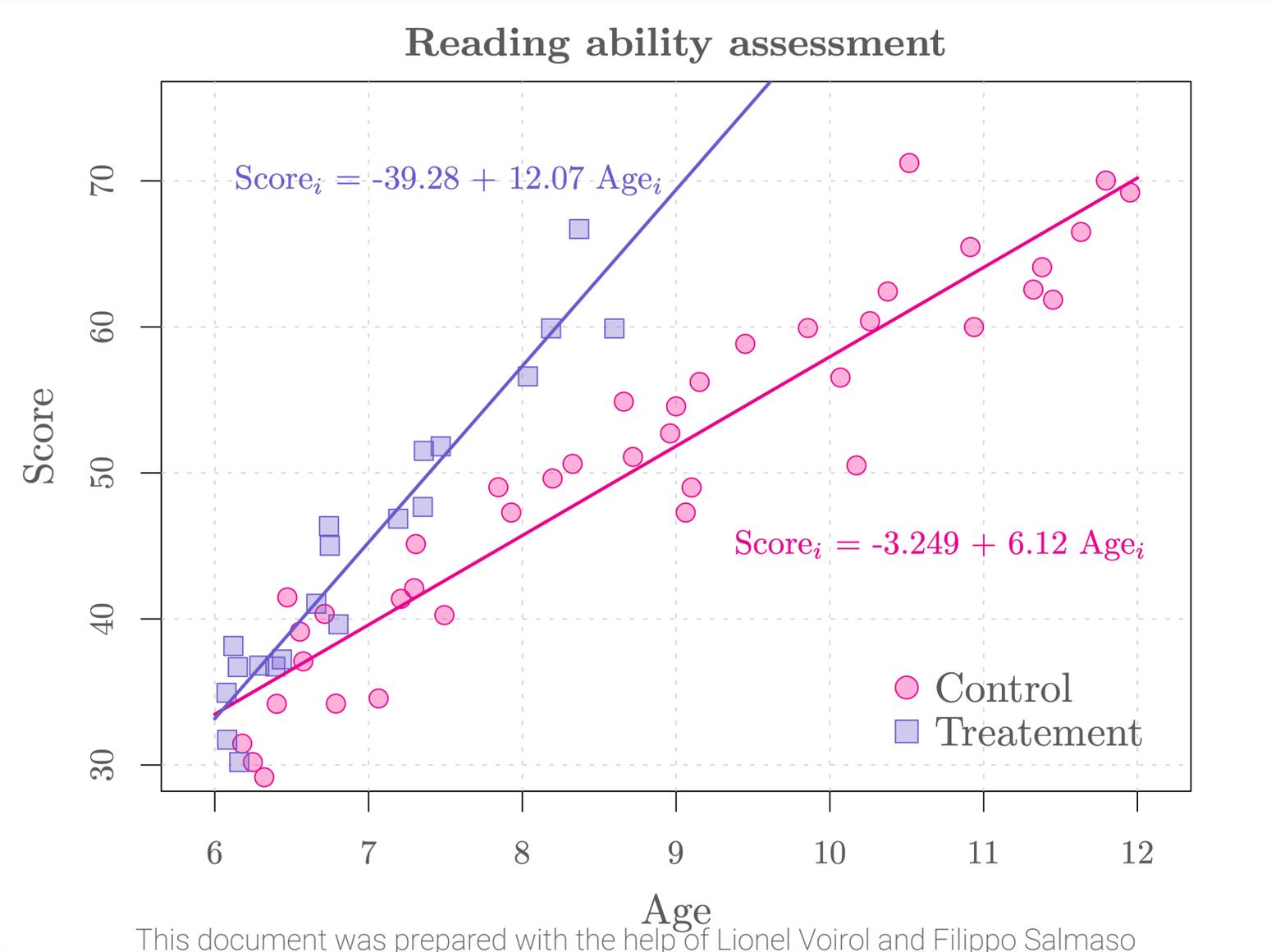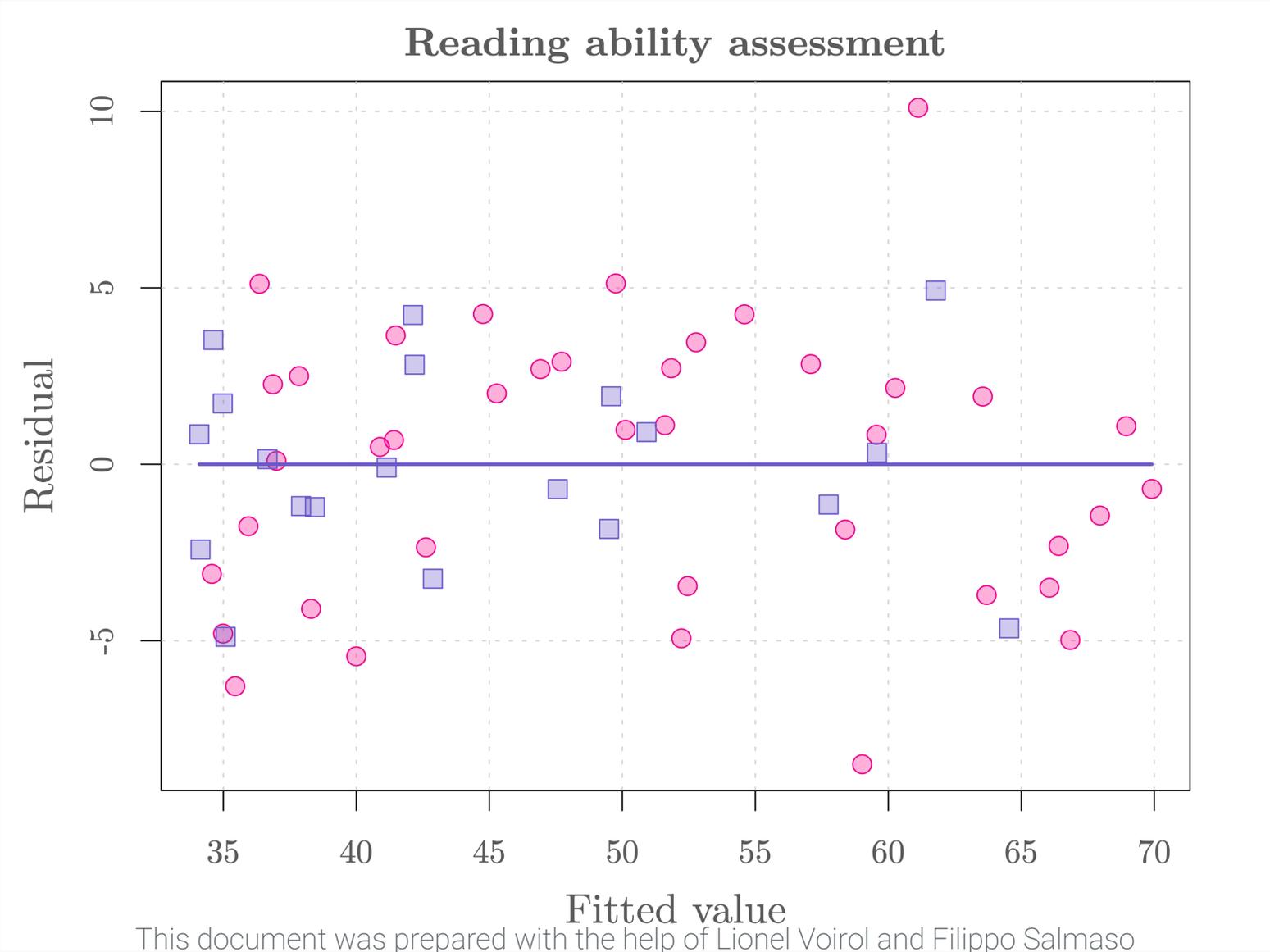
Interpretation of coefficients

We can obtain the estimated coefficients. Specifically,

- $\hat{\beta}_0 = -3.2489$ represents the estimated baseline average score of the DRP test at birth (but *again* what does it mean? 🙄)

- $\hat{\beta}_1 = -36.0307$ means that for a student of the same age, participating in the new directed reading activities is estimated to decrease their average score of the DRP test by 36.0307 (does this make sense? 🧐).

- $\hat{\beta}_2 = 6.1207$ means that for students not participating to the new directed reading activities, their average score increases by 6.1207 as they become 1 year older.

- $\hat{\beta}_3 = 5.9539$ means that the average score of students participating in the new directed reading activities increases by 5.9539 as they become 1 year older compared to the other students. This means that the average score of students participating to the new program increases by 12.0746 (i.e. 6.1207 + 5.9539) as they become 1 year older.

Model fit

Reading ability assessment

$$\text{Score}_i = \text{-39.28} + 12.07\ \text{Age}_i$$

$$\text{Score}_i = \text{-3.249} + 6.12\ \text{Age}_i$$

- Control
- Treatement

Score

Age

Model diagnostic

Reading ability assessment

Model diagnostic

Reading ability assessment
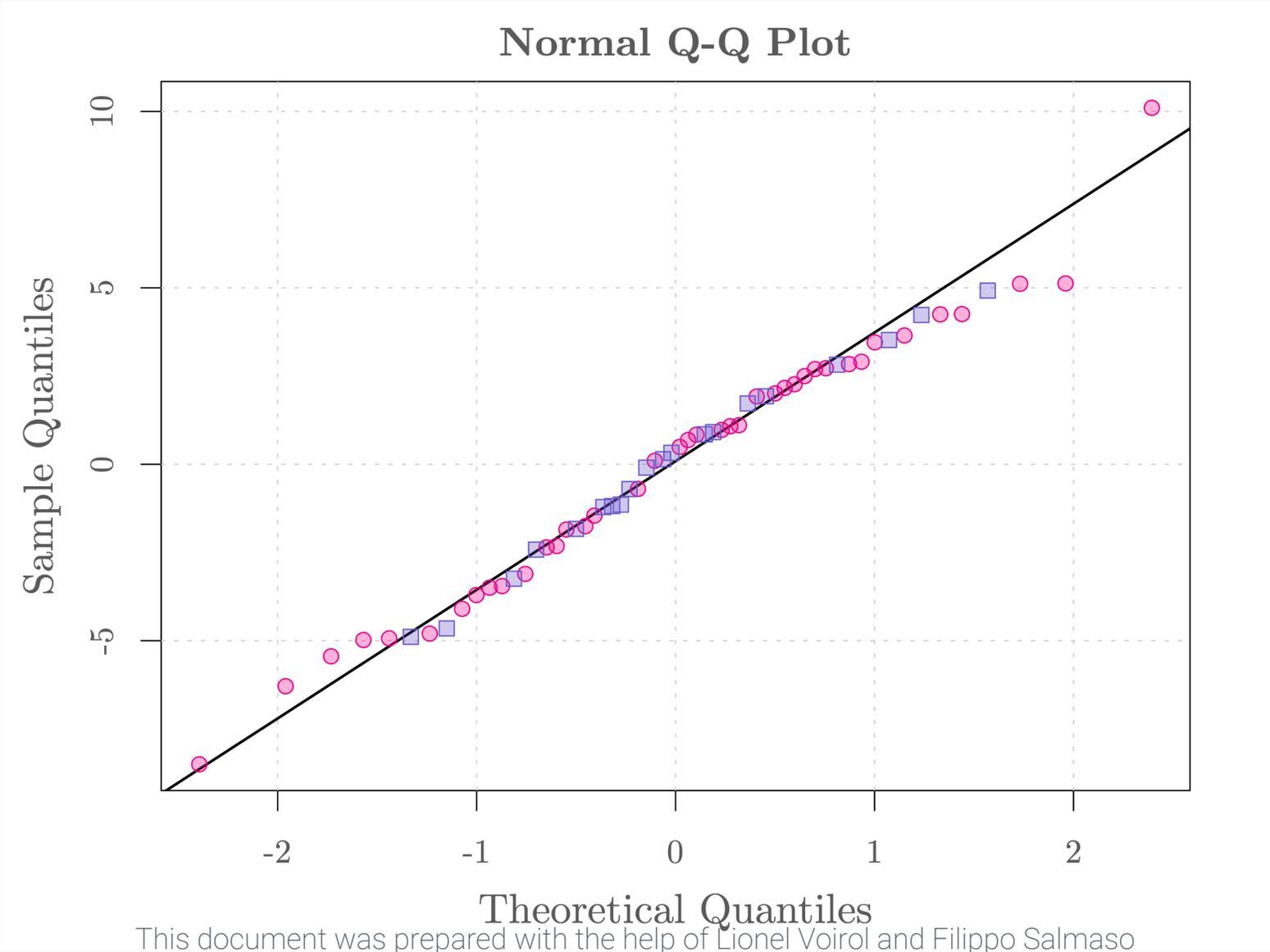
Model diagnostic

Normal Q-Q Plot

Model selection

In general, we prefer a parsimonious approach to modeling in the sense that we only choose a more complex model if the benefits are "substantial"✋. We want our model to be such that:

1. The model fits the data well.
2. Avoid (excessive) overfitting.

Naturally these two objectives are contradictory and there are many ways to select a suitable model (actually this is one of the most active areas of research in modern Statistics). In this class, we will only consider one (simple) approach based on the Akaike Information Criterion (AIC). This criterion corresponds to an estimator of prediction error and thereby relative quality of statistical models for a given set of data.

✋ This point of view is based on Occam's razor (or law of parsimony), the problem-solving principle stipulating that "the simplest explanation is usually the right one".

Model selection

In `Python`, the AIC can be computed for a given model (i.e. use the output of the function `sfm.ols(...)` in `model.aic`.) For example, we can compare the AIC of the previously considered models as follows:

```
1  print(round(mod1.aic, 4))      # First model (no interaction)
```

```
352.2688
```

```
1  print(round(mod2.aic, 4))      # Second model (with interaction)
```

```
326.9095
```

As expected, these results suggest that the second model is more appropriate. But should we further improve it?

Let's update our model (again)

It should be reasonable that the average reading scores of the two groups are the same at the start of the program.

$$\text{Score}_i = \beta_0 + \beta_1(\text{Age}_i - 6) + \beta_2\text{Group}_i(\text{Age}_i - 6) + \epsilon_i, \quad \epsilon_i \overset{iid}{\sim} N(0, \sigma^2).$$

- Score_i: score of the DRP test of the i-th student.
- Group_i: indicator of participation of the new directed reading activities for the i-th student (i.e. $\text{Group}_i = 1$ if participate and $\text{Group}_i = 0$ if not participate).
- $\text{Age}_i - 6$: corresponds to the time since start of treatment of the i-th student.

With this model the two groups can be compared as the age effect is taken into account. The goal of the educator now is (only!) to assess if β_1 is significantly larger than 0.

Example: Reading ability assessment

Python *Code*

Here is the code to fit our third model:

```python
import pandas as pd
import statsmodels.formula.api as smf

# Import data (if you haven't already)
reading = pd.read_csv("https://raw.githubusercontent.com/ELSTE-Master/Data-Science/main/Da
reading["age_minus_six"] = reading["age"] - 6

# Fit a linear regression model
mod3 = smf.ols(formula="score ~ age_minus_six + group:age_minus_six", data=reading).fit()

# Show summary
print(mod3.summary())
```

Example: Reading ability assessment

Output

```
                          OLS Regression Results
==============================================================================
Dep. Variable:                  score   R-squared:                       0.910
Model:                            OLS   Adj. R-squared:                  0.907
Method:                 Least Squares   F-statistic:                     289.2
Date:                Mon, 03 Nov 2025   Prob (F-statistic):           1.42e-30
Time:                        17:34:57   Log-Likelihood:                 -159.47
No. Observations:                  60   AIC:                             324.9
Df Residuals:                      57   BIC:                             331.2
Df Model:                           2
Covariance Type:            nonrobust
===================================================================================================
                                    coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------------------------
Intercept                        33.3504      0.790     42.192      0.000      31.768      34.933
age_minus_six                     6.1522      0.260     23.625      0.000       5.631       6.674
group[T.Treatment]:age_minus_six  5.8103      0.716      8.119      0.000       4.377       7.243
==============================================================================
Omnibus:                        0.192   Durbin-Watson:                   2.013
Prob(Omnibus):                  0.909   Jarque-Bera (JB):                0.008
Skew:                           0.027   Prob(JB):                        0.996
Kurtosis:                       3.020   Cond. No.                        5.84
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified
```

Example: Reading ability assessment

AIC

```
1  print(round(mod1.aic, 4))
```

352.2688

```
1  print(round(mod2.aic, 4))
```

326.9095
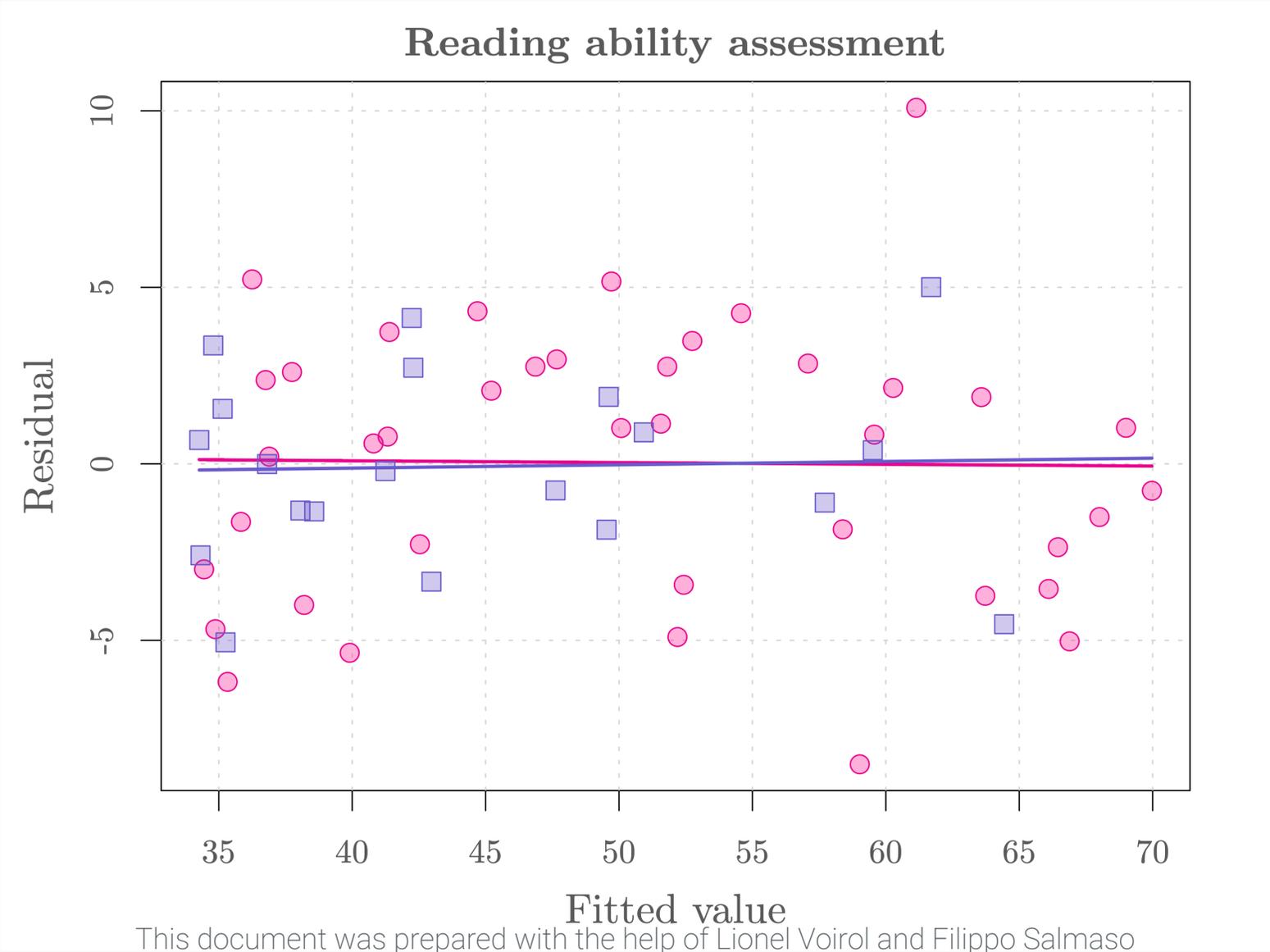
```
1  print(round(mod3.aic, 4))
```
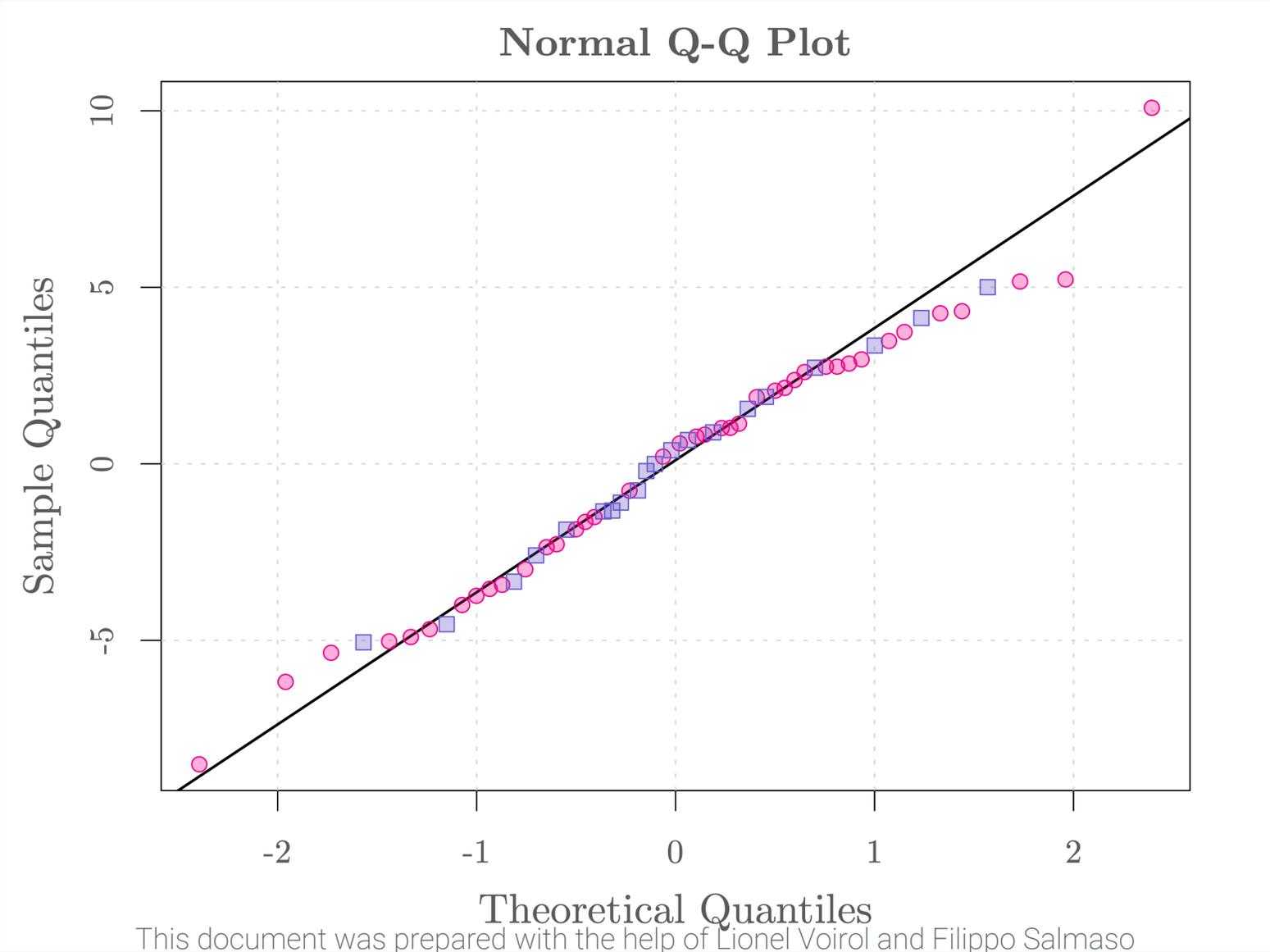
324.9479

Model fit
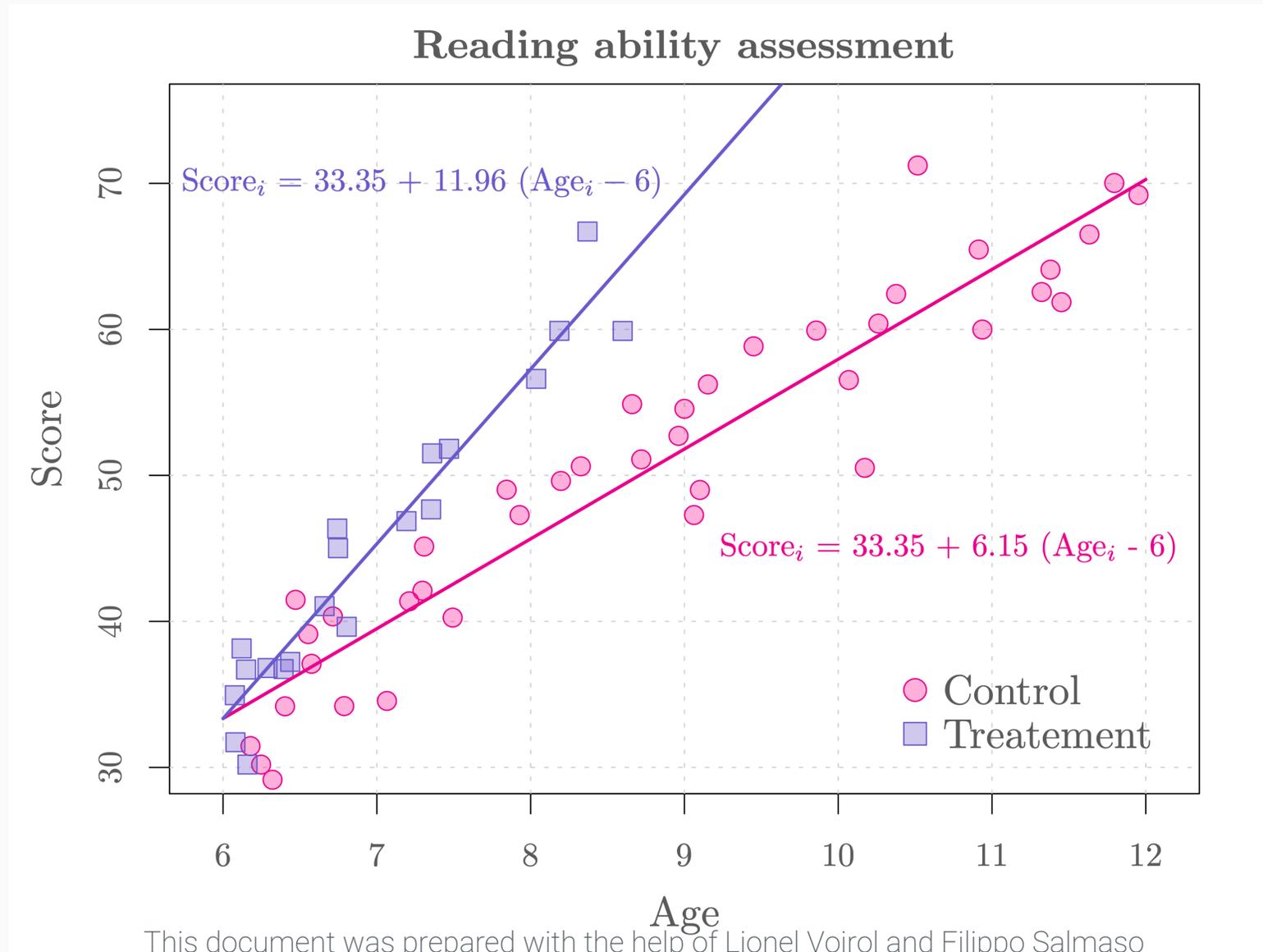
Reading ability assessment

$\text{Score}_i = 33.35 + 11.96 \, (\text{Age}_i - 6)$

$\text{Score}_i = 33.35 + 6.15 \, (\text{Age}_i - 6)$

Score

Age

● Control
■ Treatement

Model diagnostic

Reading ability assessment

Model diagnostic

Reading ability assessment

Model diagnostic

Normal Q-Q Plot

Concluding remarks

- The last model we consider appears to fit the data, avoids overfitting and allows to answer whether the new reading activities are of interest. Indeed, the programs significantly improve the reading performance of the students (e.g. 5.81 more per year compared to control, p-value < 5%).

- Our model assumes a linear relationship between the response and the covariates. However, this may be incorrect.

- Our model only considers independent data (which may not be correct here).

- Finally, linear regression should not be used to extrapolate, i.e. to estimate beyond the original observation range. For example, if we consider a 100 year-old person in this reading ability example, we would predict (using the third model) that the corresponding score of the DRP test would be 1157.59 and 611.45, respectively, with and without these activities. Does it really make sense? 😲