UNIVERSITÉ DE GENÈVE

Introduction to Regression

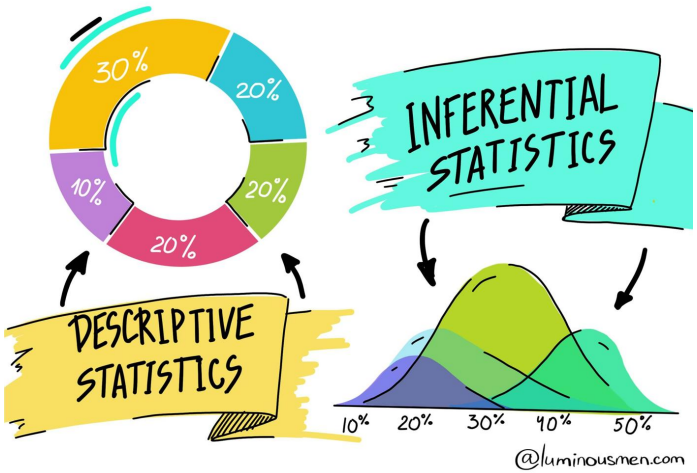
Big Data Foundations

Dominique-L. Couturier, Cancer Research UK, University of Cambridge,
Stéphane Guerrier, Data Analytics Lab, University of Geneva,
Maria-Pia Victoria-Feser, Research Center for Statistics, University of Geneva,
Yuming Zhang, Data Analytics Lab, University of Geneva

Introduction

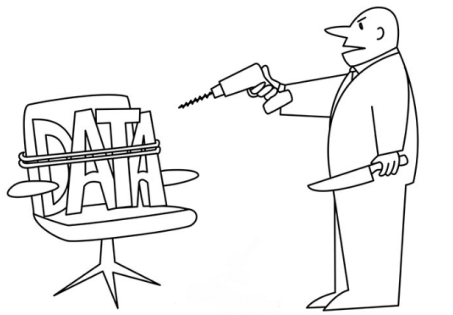
What is statistics?

Statistics is a science that uses mathematics and computer science to deal with the collection, analysis, interpretation, and presentation of masses of numerical data. Informally, it is the science of learning from data.



Source: luminousmen

Statistics is a crucial part of our life. However, statistical methods are often consciously (or not) misused. This can lead to contradictory studies and conclusions (as seen during the current COVID-19 pandemic).



"If you don't reveal some insights soon, I'm going to be forced to slice, dice, and drill!"

Source: Atoz Markets

How can statistics be useful?

Statistics can be used (among others) to

1. Visualize data (e.g. propagation of COVID-19 in different countries).
2. Understand and interpret data (e.g. main causes of cancer).
3. Assess the validity of a hypothesis (e.g. is a drug working?).
4. Make predictions (e.g. predicting unemployment or risk indices).

Learning more about statistics allows to

1. Better understand arguments based on data.
2. Be able to apply critical thinking about statistics used as evidence.
3. Understand how statistical associations are used to evaluate claims (hypotheses) and assess causal connections.

Understanding and knowing how to interpret statistical analyses is therefore becoming an increasingly vital skill.

How to test a (scientific) hypothesis?

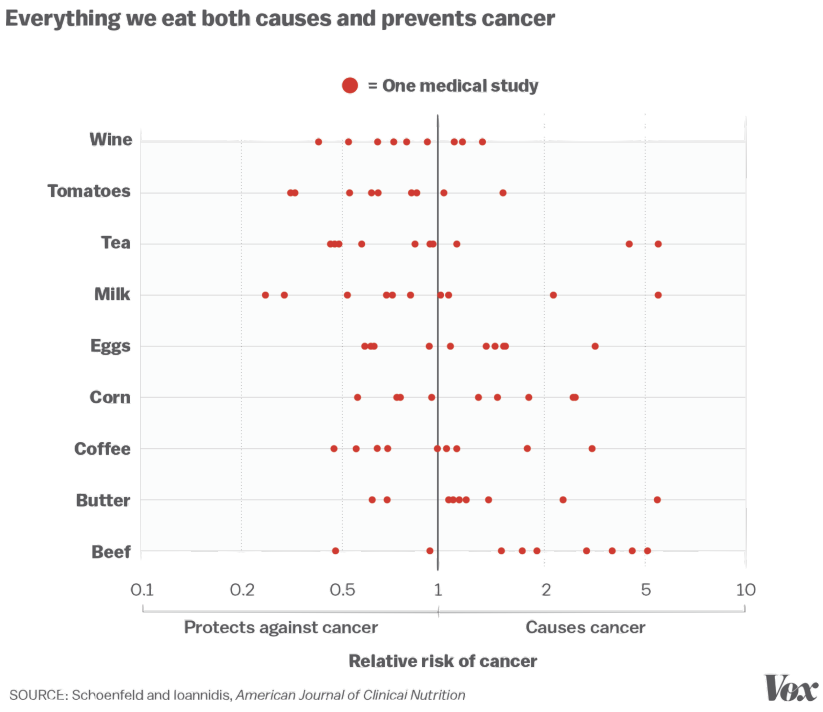
"In god we trust, all others must bring data." 👋

- To assess the validity of a (scientific) hypothesis, the scientific community (generally) agrees on a specific procedure.
- These hypotheses can be nearly anything, such as:
  1. Coffee consumption increases blood pressure.
  2. Republican politicians are bad/good for the American Economy.
  3. A glass of red wine is as good as an hour at the gym.
- This procedure involves the design of an experiment and then the collection of data to compute a metric, called **p-value**, which evaluates the adequacy between the data and your original hypothesis.
- There is generally a specific threshold (typically 5%), and if the p-value falls below this threshold we can claim that we have statistically significant result(s) validating our hypothesis.

👋 From W. Edwards Deming

Statistics vs Truth 🤥

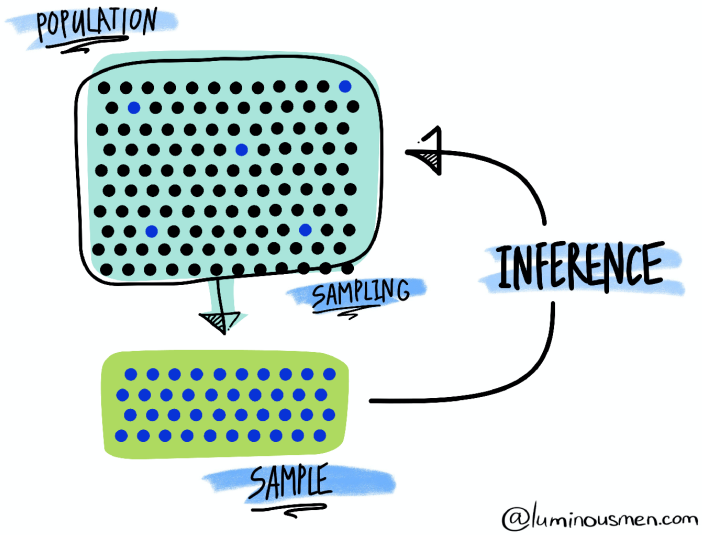
- Statistically significant results are not necessarily the truth, as there isn't a threshold (e.g. 5%) that separates real results from the false ones.
- This procedure simply provides us with one piece of a puzzle that should be considered in the context of other evidence.



👋 Read the original article: "*This is why you shouldn't believe that exciting new medical study*" here.

How does it work?

- Statistical methods are based on several fundamental concepts, the most central of which is to consider the information available (in the form of data) resulting from a random process.
- As such, the data represent a **random sample** of a totally or conceptually accessible **population**.
- Then, statistical inference allows to infer the properties of a population based on the observed sample. This includes deriving estimates and testing hypotheses.



Source: luminousmen

Hypothesis testing

- In general (scientific) hypotheses can be translated into a set of (non-overlapping idealized) statistical hypotheses:

$$H_0 : \theta \in \Theta_0 \text{ and } H_a : \theta \notin \Theta_0.$$

- In a hypothesis test, the statement being tested is called the **null hypothesis** H_0. A hypothesis test is designed to assess the strength of the evidence against the null hypothesis.
- The **alternative hypothesis** H_a is the statement we hope or suspect to be true instead of H_0.
- Each hypothesis excludes the other, so that one can exclude one in favor of the other using the data.
- To be "testable", hypothesis are built on quantities summarizing the question of interest.
- Very often, the summary quantities concern a mean, as for example the mean effect of a drug on patients.

Hypothesis testing

- Example: a drug represses the progression of cancer

$$H_0 : \mu_{\text{drug}} = \mu_{\text{control}} \quad \text{and} \quad H_a : \mu_{\text{drug}} < \mu_{\text{control}}.$$

Hypothesis testing

- To assess the validity of the alternative hypothesis H_a, one can only rely on random quantities that are called test statistics.
- Test statistics are random because for each different sample one could draw from the same population, the value of the test statistic is different.
- Statistical theory allows to calibrate the null probability distribution function of the test statistic under the null hypothesis H_0.
- This null probability distribution function allows to determine if what is observed with the sample at hand, i.e. T_{obs}, is sufficiently probable or not under the null hypothesis.
- The metric that is used is the **p-value**, which is compared to an a priori chosen **significance level** α, usually $\alpha = 5\%$.
- Hence, when rejecting H_0, for a significance level of α means that there are $\alpha\%$ chances that the alternative hypothesis H_a is NOT correct.

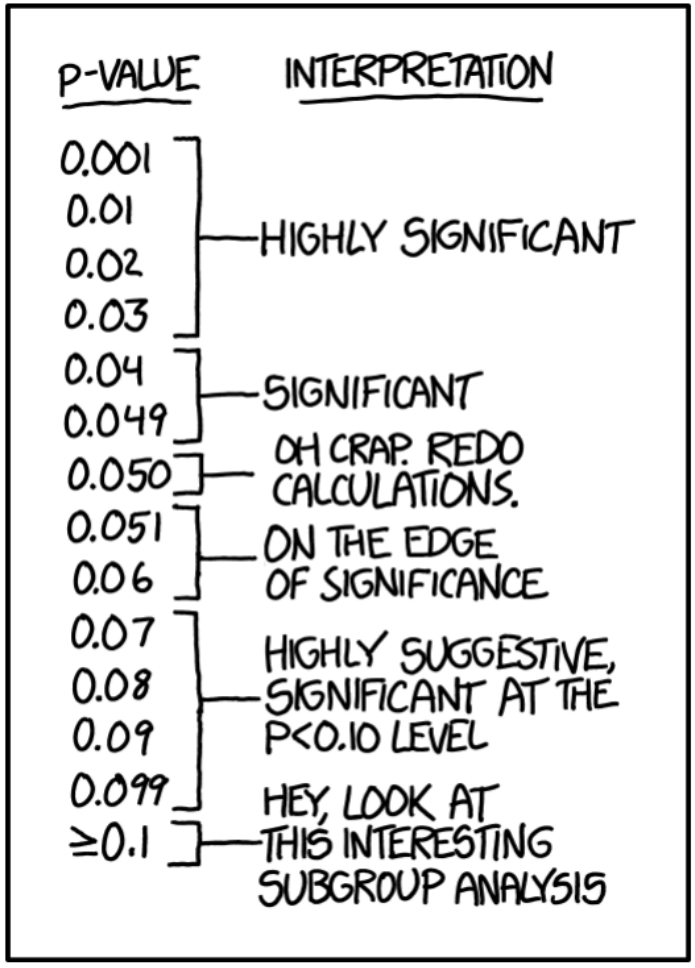
What are p-values?

- The **p-value** can be thought as the probability of observing a test statistic that is at least as extreme as actually observed, assuming that H_0 is true.
- Informally, a p-value can be understood as a measure of plausibility of the null hypothesis given the data. Small p-value indicates strong evidence against H_0.
- When the p-value is small enough (i.e. smaller than the significance level α), one says that the test based on the null and alternative hypotheses is significant or that the null hypothesis is rejected in favor of the alternative. This is generally what we want because it "verifies" our (research) hypothesis.
- When the p-value is not small enough, with the available data, we cannot reject the null hypothesis so nothing can be concluded. 🤔
- The obtained p-value summarizes somehow the incompatibility between the data and the model constructed under the set of assumptions.

"Absence of evidence is not evidence of absence." 👋

👋 From the British Medical Journal.

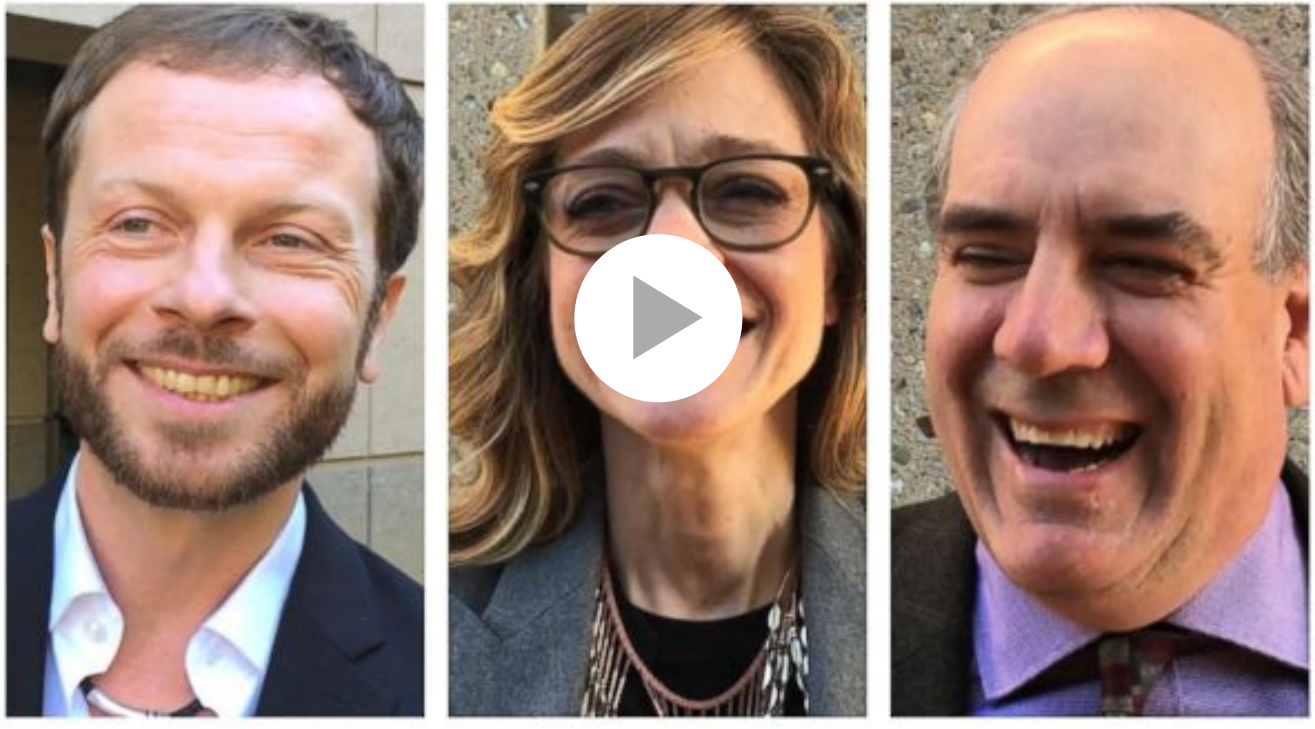
How to understand p-values?



👋 If you want to know more have a look here.

P-values may be controversial

P-values have been misused many times because understanding what they mean is not intuitive!



👋 If you want to know more have a look here.

Two-sample Location Tests

Two-sample location tests

In practice, we often encounter problems where our goal is to compare the means (or locations) of two samples. For example,

1. A scientist is interested in comparing the vaccine efficacy of the Pfizer-BioNTech and the Moderna vaccine.
2. A bank wants to know which of two proposed plans will most increase the use of its credit cards.
3. A psychologist wants to compare male and female college students' impression on a selected webpage.

We will discuss three two-sample location tests:

1. Two independent sample Student's t-test
2. Two independent sample Welch's t-test
3. Two independent sample Mann-Whitney-Wilcoxon test

Two independent sample Student's t-test

Hypotheses for comparing the means of two groups **A** and **B**:

$$H_0 : \mu_A - \mu_B = \mu_0 \quad \text{and} \quad H_a : \mu_A - \mu_B \left[> \text{ or } < \text{ or } \neq \right] \mu_0.$$

Usually one chooses $\mu_0 = 0$ meaning that under H_0 there is no difference between the two samples means.

Test statistic's distribution under H_0:

$$T = \frac{(\overline{X}_A - \overline{X}_B) - \mu_0}{s_p \sqrt{n_A^{-1} + n_B^{-1}}} \underset{H_0}{\sim} \text{Student}(n_A + n_B - 2).$$

- R function:

```
t.test(x = ..., y = ..., alternative = ..., var.equal = TRUE).
```

Two independent sample Student's t-test

One could also formulate the same test by considering a linear model for each observation $X_{iA}, i = 1, \dots n_A$ and $X_{iB}, i = 1, \dots, n_B$, or in short $X_{i(g)}$, where $g = A, B$, $i = 1, \dots, n_g$.

Indeed, each observation in one of the groups can be considered as a realization of a random variable with mean μ_g, and hence

$$X_{i(g)} = \mu_g + \varepsilon_{i(g)},$$

where, for example, $\varepsilon_{i(g)} \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$.

Under H_0 (equal means), then we can write

$$X_{i(g)} = \mu_g + \varepsilon_{i(g)} = \mu + \delta_g + \varepsilon_{i(g)}, \quad \mu_A \leq \mu \leq \mu_B$$

Then the null hypothesis becomes $H_0 : n_A\delta_A + n_B\delta_B = 0$.

Discussion - Student's t-test

- This test relies on the following assumptions:
  - the observations in each group are spread around their respective means following a normal distribution $N(\mu_g, \sigma^2)$,
  - the spread around the mean is the same in each group, i.e. σ^2.
- These assumptions can hence be violated for example when
  - there are outliers in one or both groups,
  - the sample distribution in each group is skewed,
  - the spread (variance) of the distribution is different in the two groups.
- In practice, the assumption of equal variance is hard to verify so **we recommend to avoid this test in practice**.
- In that case, an alternative test is the Welch's t-test (see later.)
- If outliers appear to be present, an alternative test is the Mann-Whitney-Wilcoxon test (see later).

Two independent sample Welch's t-test

The hypotheses are the same as for the Student t-test:

$$H_0 : \mu_A - \mu_B = \mu_0 \quad \text{and} \quad H_a : \mu_A - \mu_B \left[> \text{ or } < \text{ or } \neq \right] \mu_0,$$

with generally $\mu_0 = 0$.

Test statistic's distribution under H_0:

$$T = \frac{(\overline{X}_A - \overline{X}_B) - \mu_0}{\sqrt{s_A^2/n_A + s_B^2/n_B}} \underset{H_0}{\sim} \text{Student}(df).$$

which is different from the one used in the Student t-test (see later).

- R function:

```
t.test(x = ..., y = ..., alternative = ...).
```

Two independent sample Welch's t-test

The corresponding linear model remains

$$X_{i(g)} = \mu_g + \varepsilon_{i(g)} = \mu + \delta_g + \varepsilon_{i(g)}, \quad \mu_A \leq \mu \leq \mu_B,$$

with associated null hypothesis $H_0 : n_A\delta_A + n_B\delta_B = 0$.

The hypotheses, on the other hand, change to $\varepsilon_{i(g)} \overset{iid}{\sim} \mathcal{N}(0, \sigma_g^2)$, i.e. the variance is allowed to be different in each group.

- Notes:
  - The Welch's test strongly relies on the assumed absence of outliers.
  - The sample distribution should be at least approximately normal with no strong skewness to ensure the reliability of the test.
  - Even if the equality of variance is true, there is not real loss in using the Welch's t-test compared to the Student's t-test.

Mann-Whitney-Wilcoxon test

With the Mann-Whitney-Wilcoxon test, one assesses the difference between a more general location measure than the population mean for group **A** and **B**.

Let this location measure (or center) be denoted by respectively θ_A and θ_B.

The hypotheses are

$$H_0 : \theta_A - \theta_B = \theta_0 \quad \text{and} \quad H_a : \theta_A - \theta_B \left[> \text{ or } < \text{ or } \neq \right] \theta_0,$$

with in most cases $\theta_0 = 0$.

Test statistic is based on ranks:

$$Z = \frac{\sum_{i=1}^{n_B} R_{i(g)} - [n_B(n_A + n_B + 1)/2]}{\sqrt{n_A n_B (n_A + n_B + 1)/12}},$$

where $R_{i(g)}$ denotes the global rank of the i-th observation of group g.

Mann-Whitney-Wilcoxon test

- R function:

`wilcox.test(x = ..., y = ..., alternative = ...)`.

- Notes:
  - The distribution of the test statistic under the null is elaborated and can be obtained by different methods (e.g. exact, asymptotic normal, ...). The details are beyond the scope of this class.
  - Since the test statistic is based on the global ranks of the data, i.e. one uses values form 1 to $(n_A + n_B)$ instead of the observed values, outliers can at most get the minimal or maximal rank, independently of the observed value.
  - Also, since the ranks are equally spaced, the possible skweness of the distribution within groups has no effect of the test statistic.
  - Hence, the Mann-Whitney-Wilcoxon test is said to be "robust" to outliers and skweness.

Mann-Whitney-Wilcoxon test

As for the previous tests, one can rewrite the the hypotheses relative to the following linear model

$$X_{i(g)} = \theta_g + \varepsilon_{i(g)} = \theta + \delta_g + \varepsilon_{i(g)}, \quad \theta_A \leq \theta \leq \theta_B,$$

where $\varepsilon_{i(g)} \overset{iid}{\sim} (0, \sigma^2)$.

Then the null hypothesis becomes $H_0 : n_A\delta_A + n_B\delta_B = 0$.

Hence, the assumption does not require a normal distribution... However, if the distributions are symmetric, we have that $\theta_A = \mu_A$ and $\theta_B = \mu_B$.

Compared to the t-tests or Welch's t-test, the Mann-Whitney-Wilcoxon test is less powerful if their requirements (Gaussian and possibly same variances) are met. **Less powerful means that the probability of rejecting H_0 when H_a is true, is smaller.**

Example: Comparing diets A and B

1. Define hypotheses: $H_0 : \mu_A = \mu_B$ and $H_a : \mu_A \neq \mu_B$.
2. Define the significance level α: we consider $\alpha = 5\%$.
3. Get the data and visualize them with e.g. boxplots, for checking which test is most suitable.
4. Compute the p-value and compare it to α.
5. Conclusion: reject or fail to reject the null hypothesis at the significance level of 5%.

Location Tests with Multiple Samples

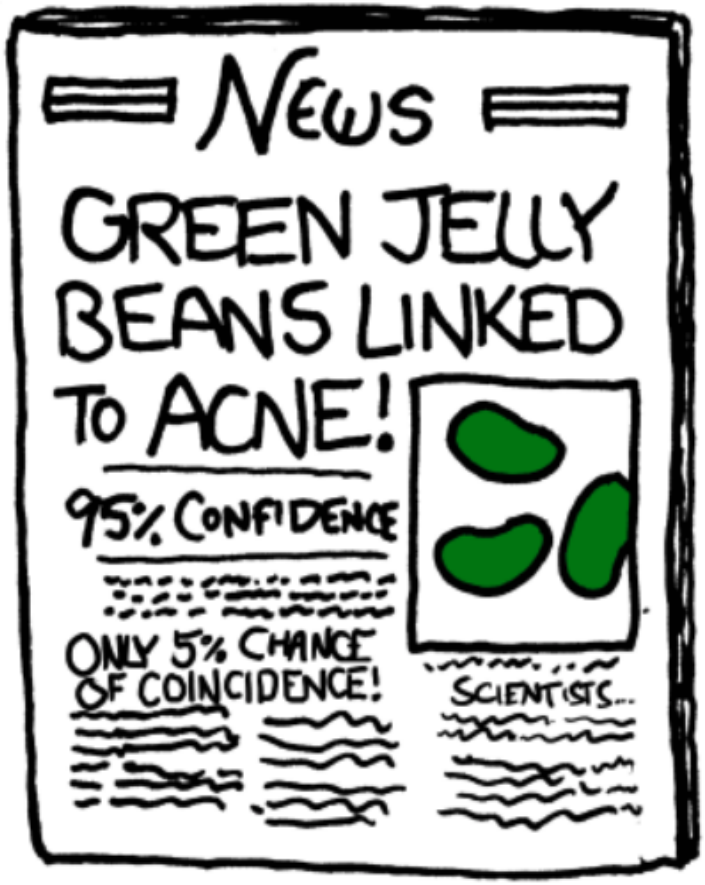
Problems with multiple samples

In practice, we often even encounter situations where we need to compare the means of more than 2 groups. For example, we want to compare the weight loss efficacy of several diets, say diets **A**, **B**, **C**. The aim could, for example, be to evaluate the following hypothesis: $0 < \mu_A = \mu_B < \mu_C$. One intuitive (but **inappropriate**) approach could be:

1. Show that μ_C is greater than μ_A and μ_B (i.e. Test 1: H_0: $\mu_A = \mu_C$, H_a: $\mu_A < \mu_C$; Test 2: H_0: $\mu_B = \mu_C$, H_a: $\mu_B < \mu_C$). Here we hope to reject H_0 in both cases.
2. Show that μ_A and μ_B are greater than 0 (i.e. Test 3: H_0: $\mu_A = 0$, H_a: $\mu_A > 0$; Test 4: H_0: $\mu_B = 0$, H_a: $\mu_B > 0$). Here we also hope to reject H_0 in both cases.
3. Compare μ_A and μ_B (i.e. Test 5: H_0: $\mu_A = \mu_B$, H_a: $\mu_A \neq \mu_B$). Here we hope not to reject H_0. ⚠ This does not imply that $\mu_A = \mu_B$ is true but at least the result would not contradict our theory.

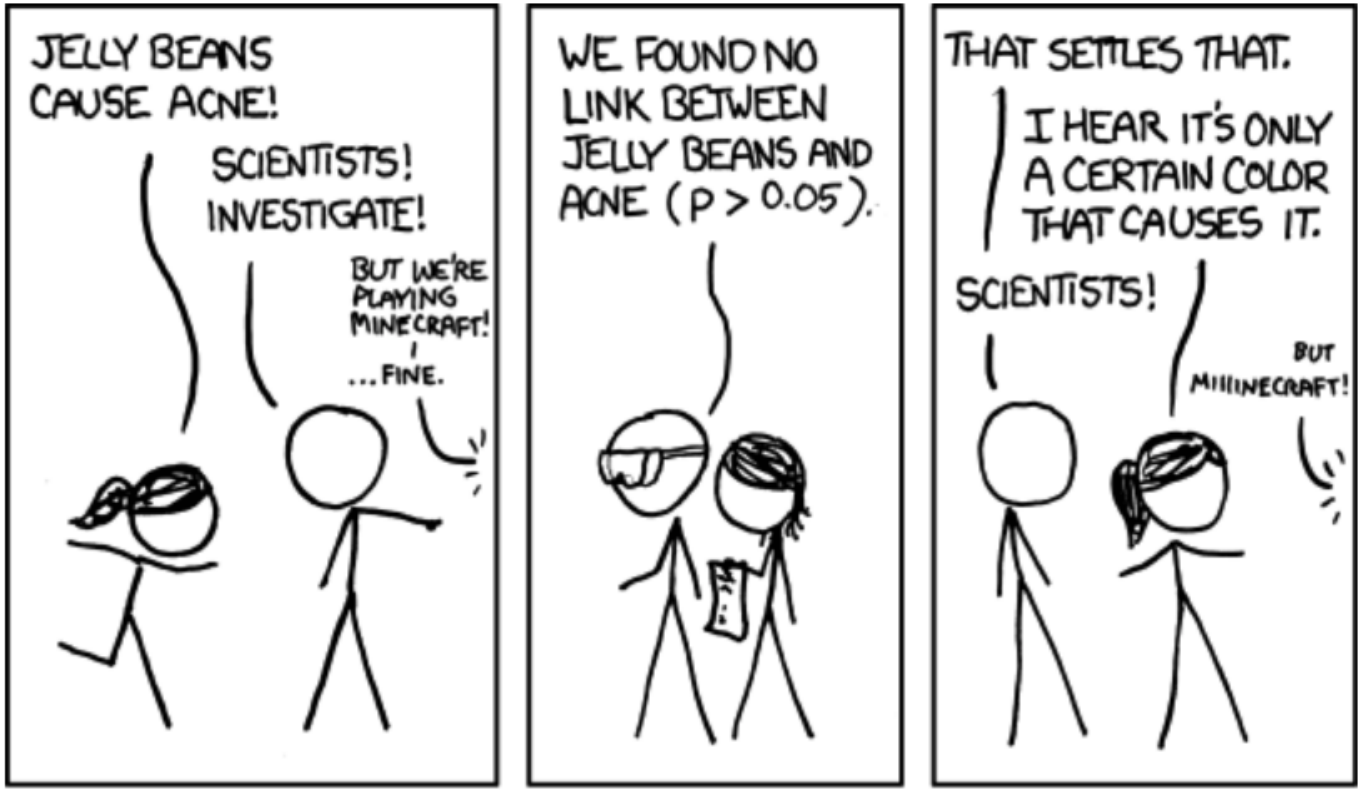
Is there a problem in doing many tests?

Are jelly beans causing acne? Maybe... but why only green ones? 😐



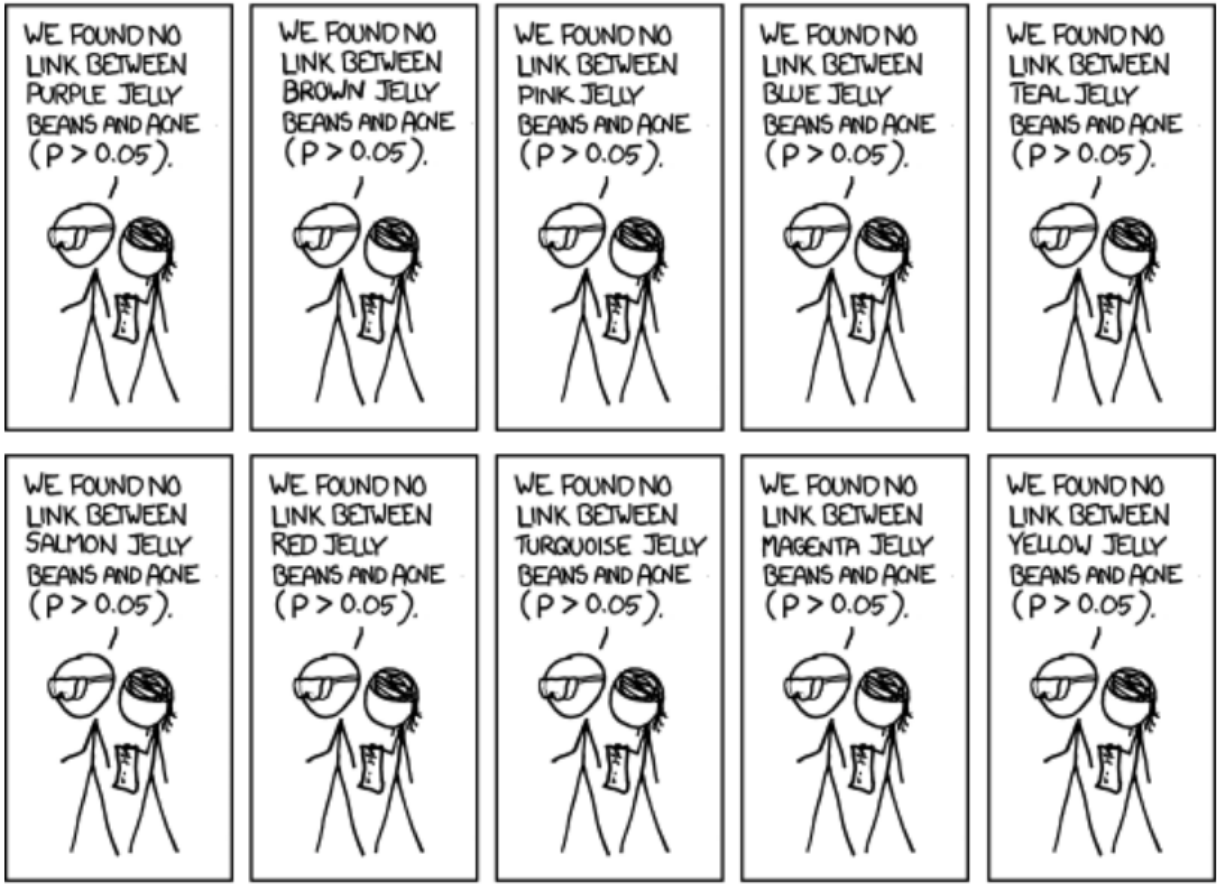
Source: xkcd

Are jelly beans causing acne?



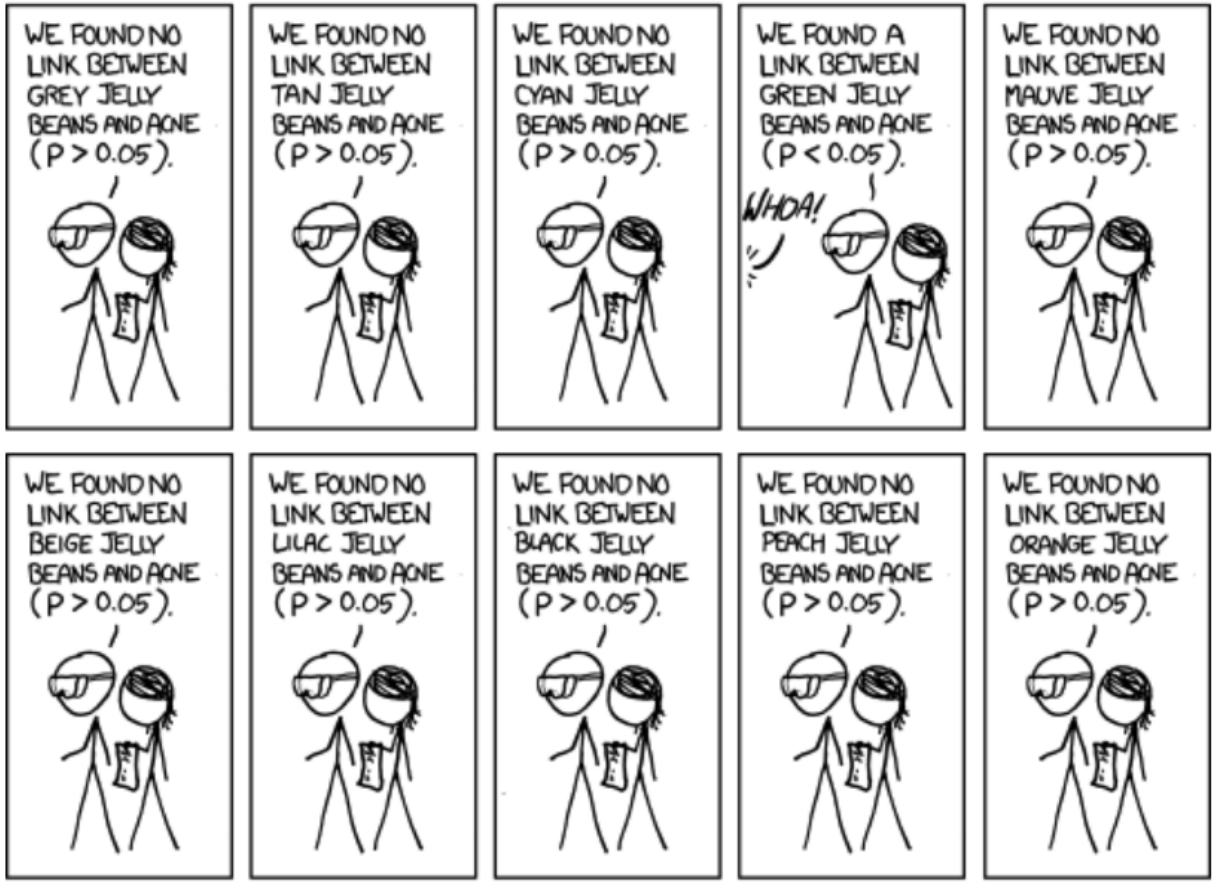
Source: xkcd

Maybe a specific color?



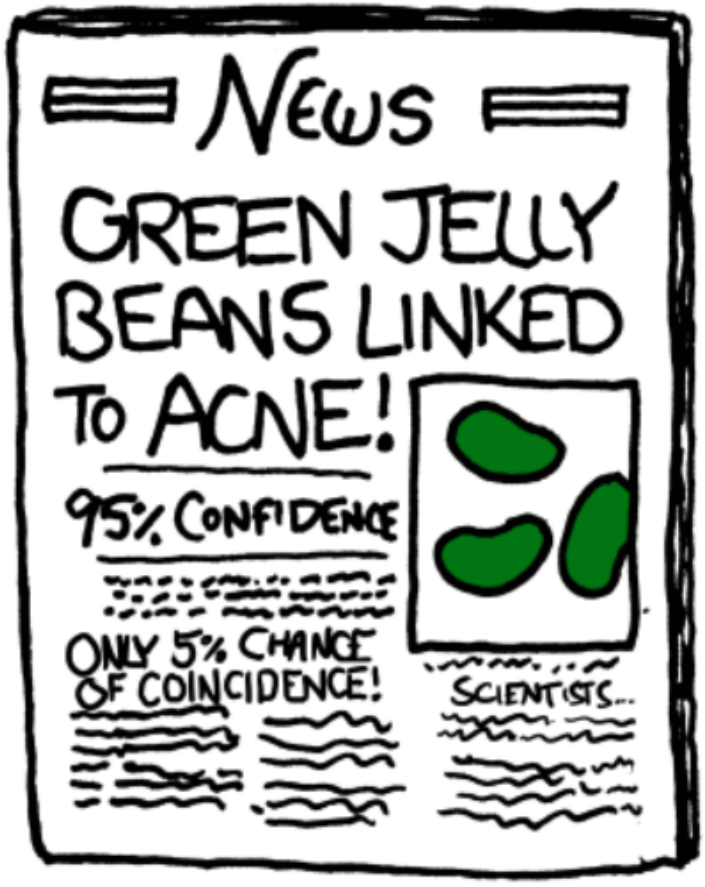
Source: xkcd

Maybe a specific color?



Source: xkcd

And finally...



Source: xkcd

👋 If you want to know more about this comics have a look here.

Multiple testing can be dangerous!

- Remember that a p-value is random as its value depends on the data.
- If multiple hypotheses are tested, the chance of observing a rare event increases, and therefore, the chance to incorrectly reject a null hypothesis (i.e. making a Type I error) increases.
- For example, if we consider k (independent) tests (whose null hypotheses are all correct), we have

$$\begin{aligned}\alpha_k &= \Pr(\text{reject } H_0 \text{ at least once}) \\ &= 1 - \Pr(\text{not reject } H_0 \text{ test 1}) \times \ldots \times \Pr(\text{not reject } H_0 \text{ test k}) \\ &= 1 - (1-\alpha) \times \ldots \times (1-\alpha) = 1 - (1-\alpha)^k\end{aligned}$$

- Therefore, α_k increases rapidly with k (e.g. $\alpha_1 = 0.05$, $\alpha_2 \approx 0.098$, $\alpha_{10} \approx 0.4013$, $\alpha_{100} \approx 0.9941$).
- Hence performing multiple tests, with the same or different data, is dangerous ⚠️ (but very common! 😟) as it can leads to significant results, when actually there are none!

Possible solutions

Suppose that we are interested in making k tests and that we want the probability of rejecting the null at least once (assuming the null hypotheses to be correct for all tests) α_k to be equal to α (typically 5%). Instead of using α for the individual tests we will use α_c (i.e. a corrected α). Then, for k (potentially dependent) tests we have

$$\begin{aligned}\alpha_k = \alpha &= \Pr(\text{reject } H_0 \text{ at least once}) \\ &= \Pr(\text{reject } H_0 \text{ test 1} \cup \ldots \cup \text{reject } H_0 \text{ test k}) \\ &\leq \sum_{i=1}^{k} \Pr(\text{reject } H_0 \text{ test i}) = \alpha_c \times k.\end{aligned}$$

Solving for α_c we obtain: $\alpha_c = \alpha/k$, which is called Bonferroni correction. By making use of the Boole's inequality, this approach does not require any assumptions about dependence among the tests or about how many of the null hypotheses are true.

Possible solutions

The Bonferroni correction can be conservative if there are a large number of tests, as it comes at the cost of reducing the power of the individual tests (e.g. if $\alpha = 5\%$ and $k = 20$, we get $\alpha_c = 0.05/20 = 0.25\%$). There exists a (slightly) "tighter" bound for α_k, which is given by

$$\alpha_k = \Pr(\text{reject } H_0 \text{ at least once}) \leq 1 - (1 - \alpha_c)^k.$$

Solving for α_c we obtain: $\alpha_c = 1 - (1 - \alpha)^{1/k}$, which is called Dunn–Šidák correction. This correction is (slightly) less stringent than the Bonferroni correction (since $1 - (1 - \alpha)^{1/k} > \alpha/k$ for $k \geq 2$).

There exist many other alternative methods for multiple testing corrections. It is important to mention that when k is large (say > 100) the Bonferroni and Dunn–Šidák corrections become inapplicable and methods based on the idea of False Discovery Rate should be preferred. However, these recent methods are beyond the scope of this class.

Multiple-sample location tests

To compare several means of different populations, a standard approach is to start our analysis by using the multiple-sample location tests. More precisely, we proceed our analysis with the following steps:

- Step 1: We first perform the multiple-sample location tests, where the null hypothesis states that all the locations are the same. If we cannot reject the null hypothesis, we stop our analysis here. Otherwise, we move on to Step 2.
- Step 2: We compare the groups mutually (using α_c) with two-sample location tests in order to verify our hypothesis.

We will discuss three multiple-sample location tests:

1. Fisher's one-way ANalysis Of VAriance (ANOVA)
2. Welch's one-way ANOVA
3. Kruskal-Wallis test

Fisher's one-way ANOVA

This test considers the following assumed model for G groups

$$X_{i(g)} = \mu_g + \varepsilon_{i(g)} = \mu + \delta_g + \varepsilon_{i(g)},$$

where μ is the overall mean, $g = 1, \ldots, G$, $i = 1, \ldots, n_g$, $\varepsilon_{i(g)} \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$ so that $H_0 : \sum n_g \delta_g = 0$.

This corresponds to:

$$H_0 : \mu_1 = \mu_2 = \ldots = \mu_G \quad \text{and} \quad H_a : \mu_i \neq \mu_j \text{ for at least one pair of } (i, j).$$

Test statistic's distribution under H_0:

$F = \frac{N s^2_{\overline{X}}}{s^2_p} \underset{H_0}{\sim} \text{Fisher}(G-1, N-G)$, where $s^2_{\overline{X}} = \frac{1}{G-1} \sum_{g=1}^{G} \frac{n_g}{N} (\overline{X}_g - \overline{X})^2$,
$s^2_p = \frac{1}{N-G} \sum_{g=1}^{G} (n_g - 1) s^2_g$, $N = \sum_{g=1}^{G} n_g$, and $\overline{X} = \frac{1}{N} \sum_{g=1}^{G} n_g \overline{X}_g$.

Discussion - Fisher's one-way ANOVA

R function:

`aov(response ~ groups, data = mydata)`.

where `response` is the column in `mydata` of all the $X_{i(g)}$ and `groups` is the column in `mydata` that identifies g (the groups).

- Notes:
  - The Fisher's test strongly relies on the assumed absence of outliers. If outliers appear to be present the Kruskal-Wallis test (see later) is (probably) a better option.
  - For moderate and small sample sizes, the sample distribution should be at least approximately normal with no strong skewness and equality of variance within the groups.
  - In practice, the assumption of equal variance is hard to verify so **we recommend to avoid testing for this in practice**.

Welch's one-way ANOVA

This test considers the same assumed model for G groups

$$X_{i(g)} = \mu_g + \varepsilon_{i(g)} = \mu + \delta_g + \varepsilon_{i(g)},$$

but with $\varepsilon_{i(g)} \overset{iid}{\sim} \mathcal{N}(0, \sigma_g^2)$ and with the same $H_0 : \sum n_g \delta_g = 0$.

This corresponds to:

$$H_0 : \mu_1 = \mu_2 = \ldots = \mu_G \quad \text{and} \quad H_a : \mu_i \neq \mu_j \text{ for at least one pair of } (i, j).$$

Test statistic's distribution under H_0:

$$F^* = \frac{s^{*2}_{\overline{X}}}{1 + \frac{2(G-2)}{3\Delta}} \underset{H_0}{\sim} \text{Fisher}(G-1, \Delta),$$

where $s^{*2}_{\overline{X}} = \frac{1}{G-1} \sum_{g=1}^{G} w_g (\overline{X}_g - \overline{X}^*)^2$, $\Delta = [\frac{3}{G^2-1} \sum_{g=1}^{G} \frac{1}{n_g} (1 - \frac{w_g}{\sum_{g=1}^{G} w_g})]^{-1}$, $w_g = \frac{n_g}{s_g^2}$, and $\overline{X}^* = \sum_{g=1}^{G} \frac{w_g \overline{X}_g}{\sum_{g=1}^{G} w_g}$.

Discussion - Welch's one-way ANOVA

- R function:

```
oneway.test(response ~ groups, data = mydata).
```

- Notes:
  - This test strongly relies on the assumed absence of outliers. If outliers appear to be present the Kruskal-Wallis test (see later) is (probably) a better option.
  - For moderate and small sample sizes, the sample distribution should be at least approximately normal with no strong skewness to ensure the reliability of the test.
  - This test does not require the variances of the groups to be equal. If the variances of all the groups are the same (which is rather unlikely in practice), the Welch's one-way ANOVA losses a little bit of power compared to the Fisher's one-way ANOVA.

Kruskal-Wallis test

This test considers the same assumed model for G groups

$$X_{i(g)} = \theta_g + \varepsilon_{i(g)} = \theta + \delta_g + \varepsilon_{i(g)},$$

but with $\varepsilon_{i(g)} \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$ and $H_0 : \sum n_g \delta_g = 0$.

The $\theta_g, g = 1, \ldots, G$ are location measures, not necessarily means.

This corresponds to:

$$H_0 : \theta_1 = \theta_2 = \ldots = \theta_G \quad \text{and} \quad H_a : \theta_i \neq \theta_j \text{ for at least one pair of } (i, j).$$

Test statistic's distribution under H_0: $K = \frac{\frac{12}{N(N+1)} \sum_{g=1}^{G} \frac{\overline{R}_g}{ng} - 3(N-1)}{1 - \frac{\sum_{v=1}^{V} t_v^3 - t_v}{N^3 - N}} \underset{H_0}{\sim} \mathcal{X}(G-1)$, where

$\overline{R}_g = \frac{1}{n_g} \sum_{i=1}^{n_g} R_{i(g)}$ with $R_{i(g)}$ denoting the global rank of the i^{th} observation of group g, V is the number of different values/levels in X and t_v denotes the number of times a given value/level occurred in X.

Discussion - Kruskal-Wallis test

- R function:

`kruskal.test(response ~ groups, data = mydata)`.

- Notes:
  - This test is "robust" in the sense that it is not overly affected by skewness and outliers.
  - For the Kruskal-Wallis test to be comparable to the one-way ANOVAs (i.e. testing for the mean) we need to assume: (1) The distributions are symmetric, (2) the variances are the same. Then, we have $\theta_i = \mu_i, i = 1, \ldots, G$.
  - Compared to the one-way ANOVA, the Kruskal-Wallis test is less powerful if their requirements (Gaussian and possibly same variances) are met.

Example: Comparing diets A, B and C

To compare diets A, B, and C, we first test if there's any difference among these 3 diets.

1. Define hypotheses:

$$H_0 : \mu_A = \mu_B = \mu_C \quad \text{and} \quad H_a : H_0 \text{ is false, or}$$

$$H_0 : \theta_A = \theta_B = \theta_C \quad \text{and} \quad H_a : H_0 \text{ is false.}$$

2. Define α: we consider $\alpha = 5\%$.
3. Get the data and visualize them with e.g. boxplots, for checking which test is most suitable.
4. Compute the p-value and compare it to α.
5. Conclusion: reject or fail to reject the null hypothesis at the significance level of 5%.

Example: Comparing diets A, B and C

1. Define hypotheses:

$$\begin{aligned} H_0 : \mu_A = \mu_B \quad &\text{and} \quad H_a : \mu_A < \mu_B. \\ H_0 : \mu_A = \mu_C \quad &\text{and} \quad H_a : \mu_A < \mu_C. \\ H_0 : \mu_B = \mu_C \quad &\text{and} \quad H_a : \mu_B < \mu_C. \end{aligned}$$

2. Define significance level: we consider $\alpha/3 = 5\%/3 \approx 1.67\%$ for each pair of hypotheses.
3. Compute the p-values and compare them to $\alpha/3$.
4. Conclusion: reject or fail to reject the null hypotheses at the "familywise" significance level of 5%.

Linear Regression

Motivating Example: Reading Ability

Problem | Data | Graph | Test | R Code

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?

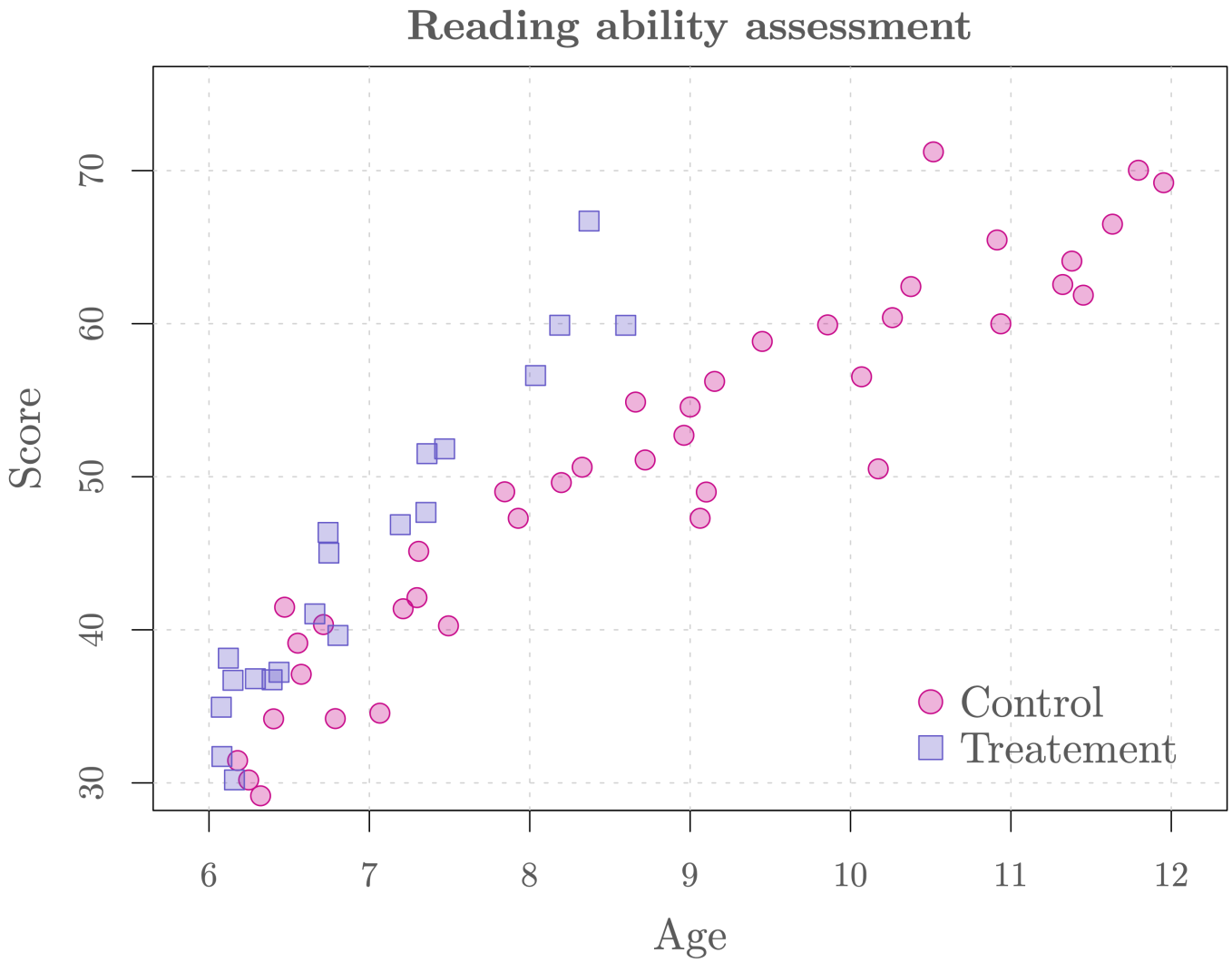
Is our analysis comprehensive?

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:

```
treatment_age = dat$age[dat$group == "Treatment"]
control_age   = dat$age[dat$group == "Control"]
```

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

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

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

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. $\varepsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$, at least when the sample size is small to medium.

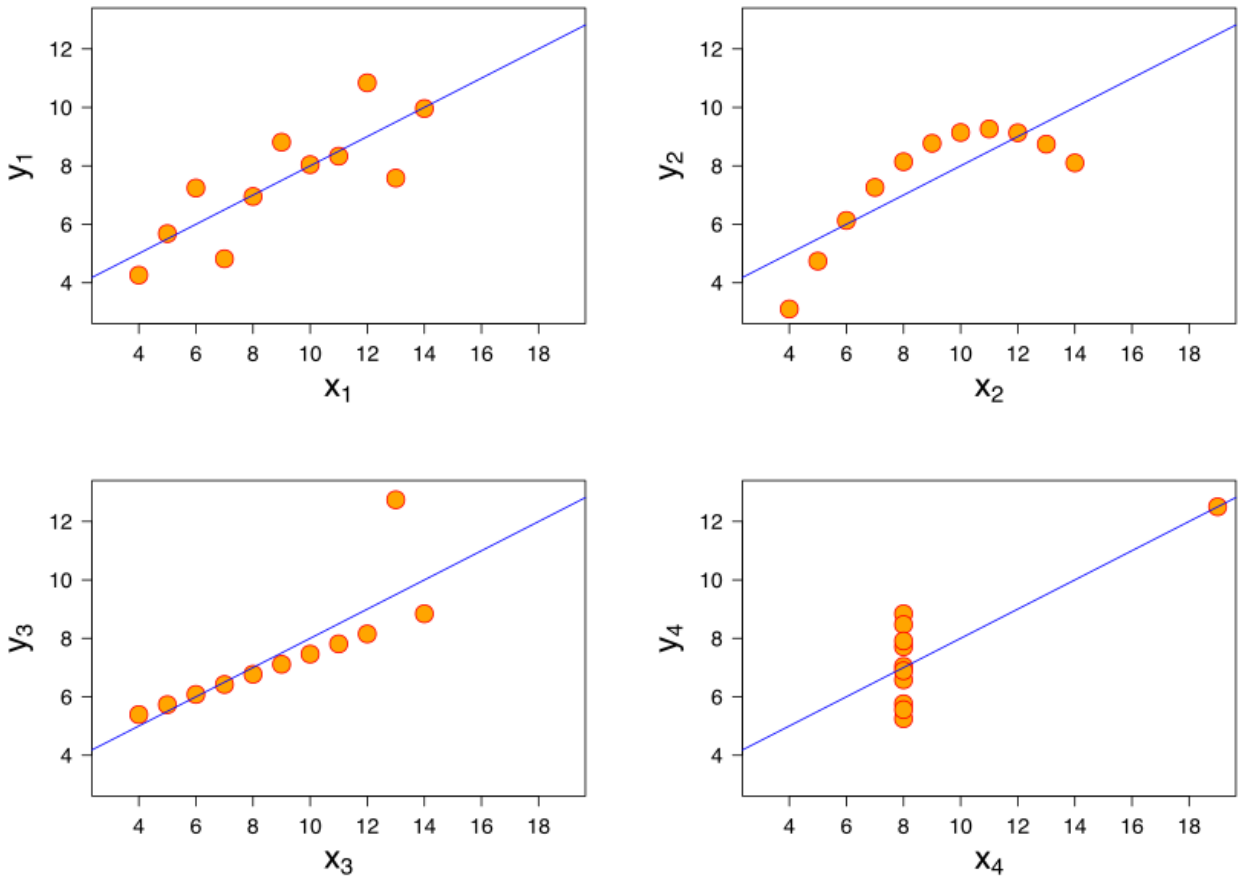
⚠ In practice, it is important to assess if these assumptions are plausible.

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



👋 Source: Wikipedia.

Example: Reading ability assessment

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

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

- 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

R Code | Output

R function: `lm(y ~ x1 + ... + xp, data = mydata)`.

Here is the code for our example:

```
# Import data (if you haven't already)
dat = read.csv("data/reading.csv")

# Fit linear regression model
mod1 = lm(score ~ group + age, data = dat)
summary(mod1)
```

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

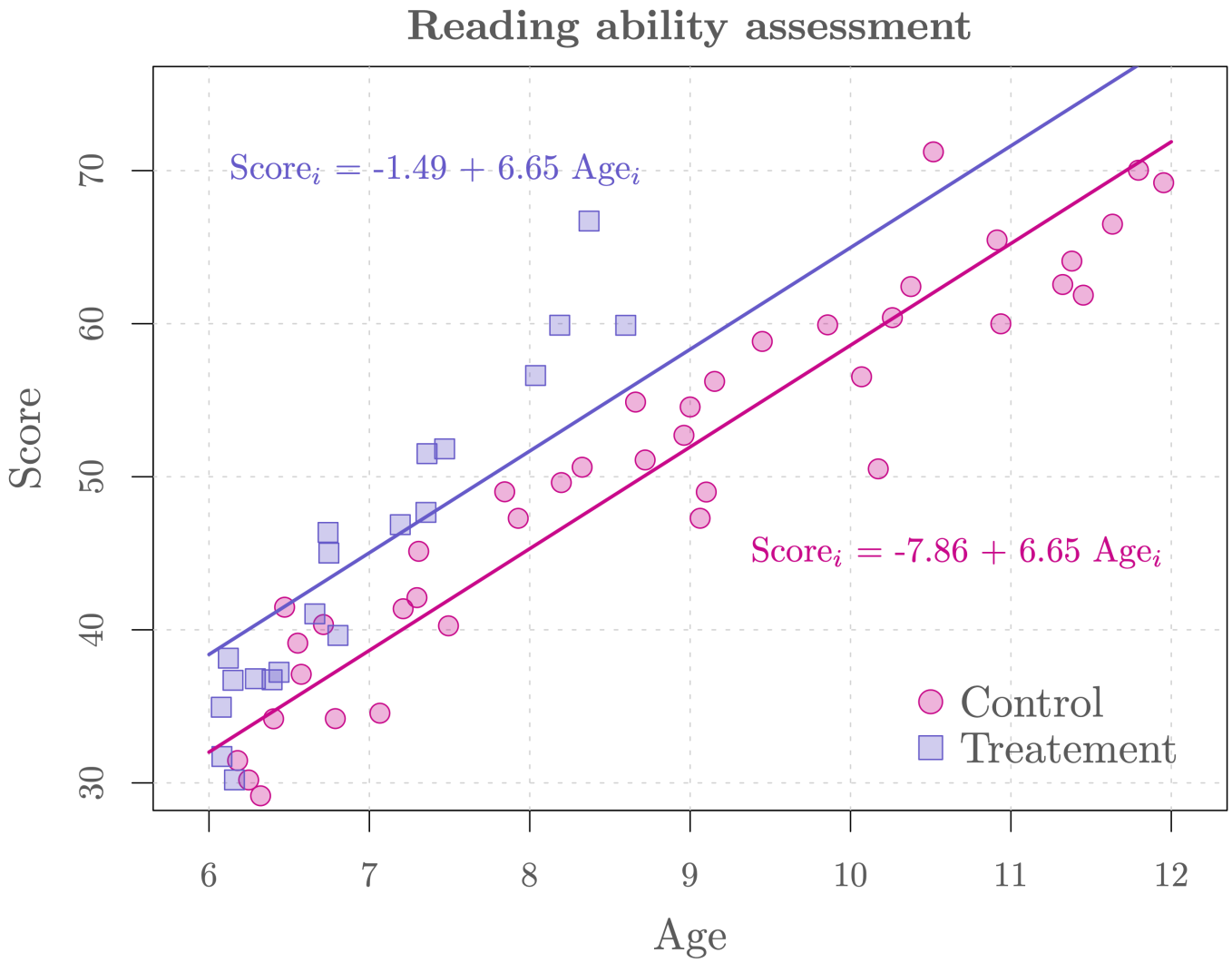
In the linear regression output, the coefficient p-value (which we denote as p below) corresponds to a two-sided test. We can use this result to compute the p-value of a one-sided test using the following relations:

	$H_a : \beta_j > 0$	$H_a : \beta_j < 0$
$\hat{\beta}_j > 0$	$p/2$	$1 - p/2$
$\hat{\beta}_j < 0$	$1 - p/2$	$p/2$

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? 🤔

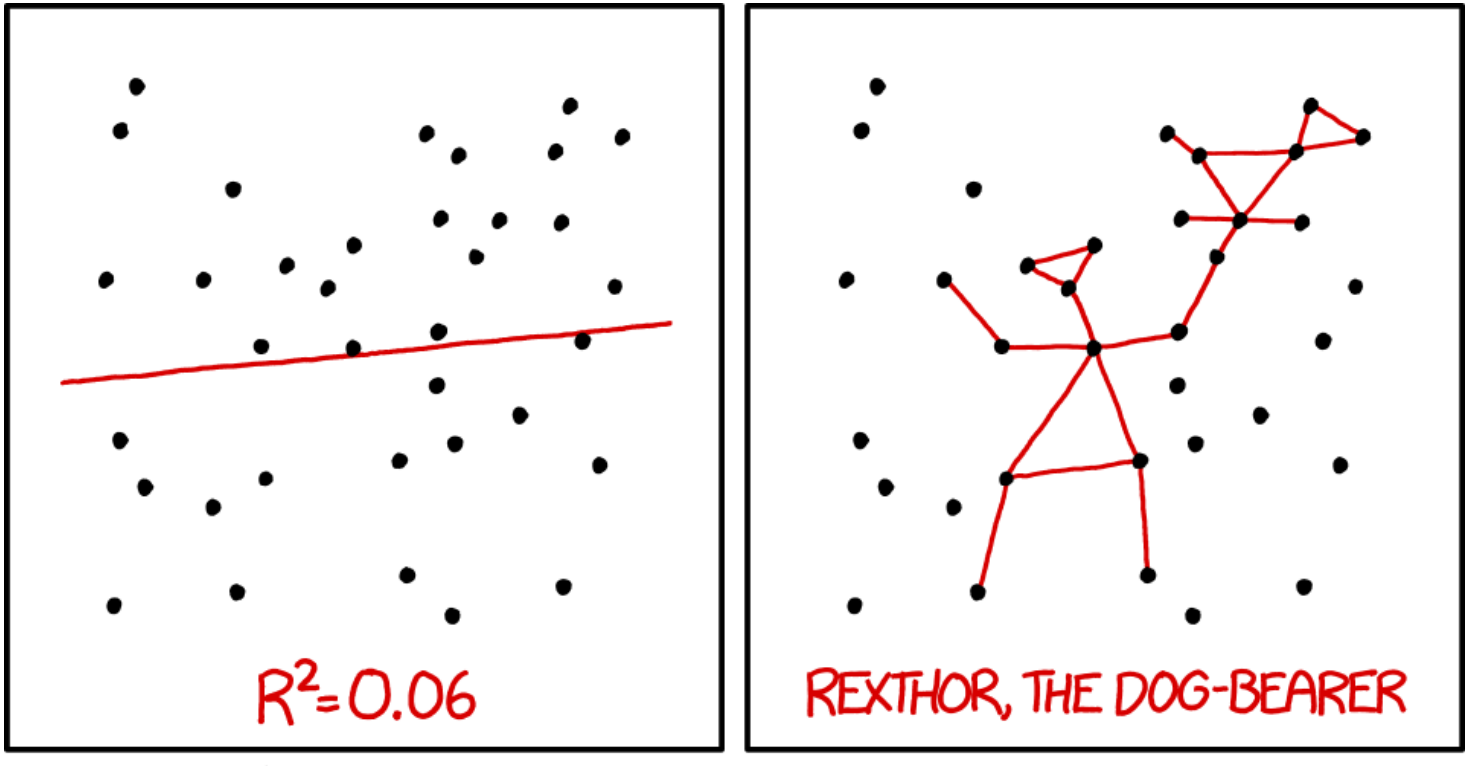
How good is our model? 🤔



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.

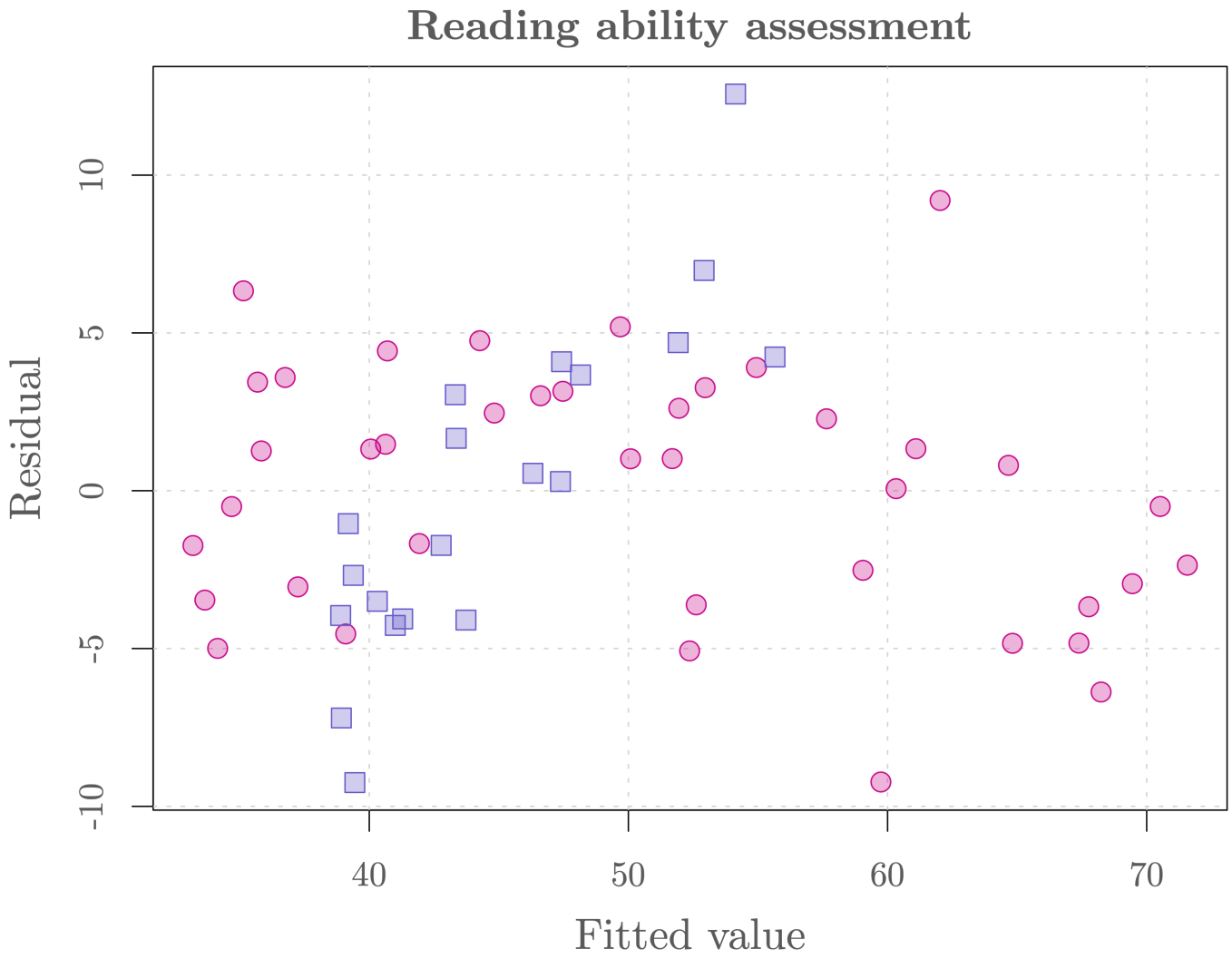
Rexthor, the Dog-Bearer!



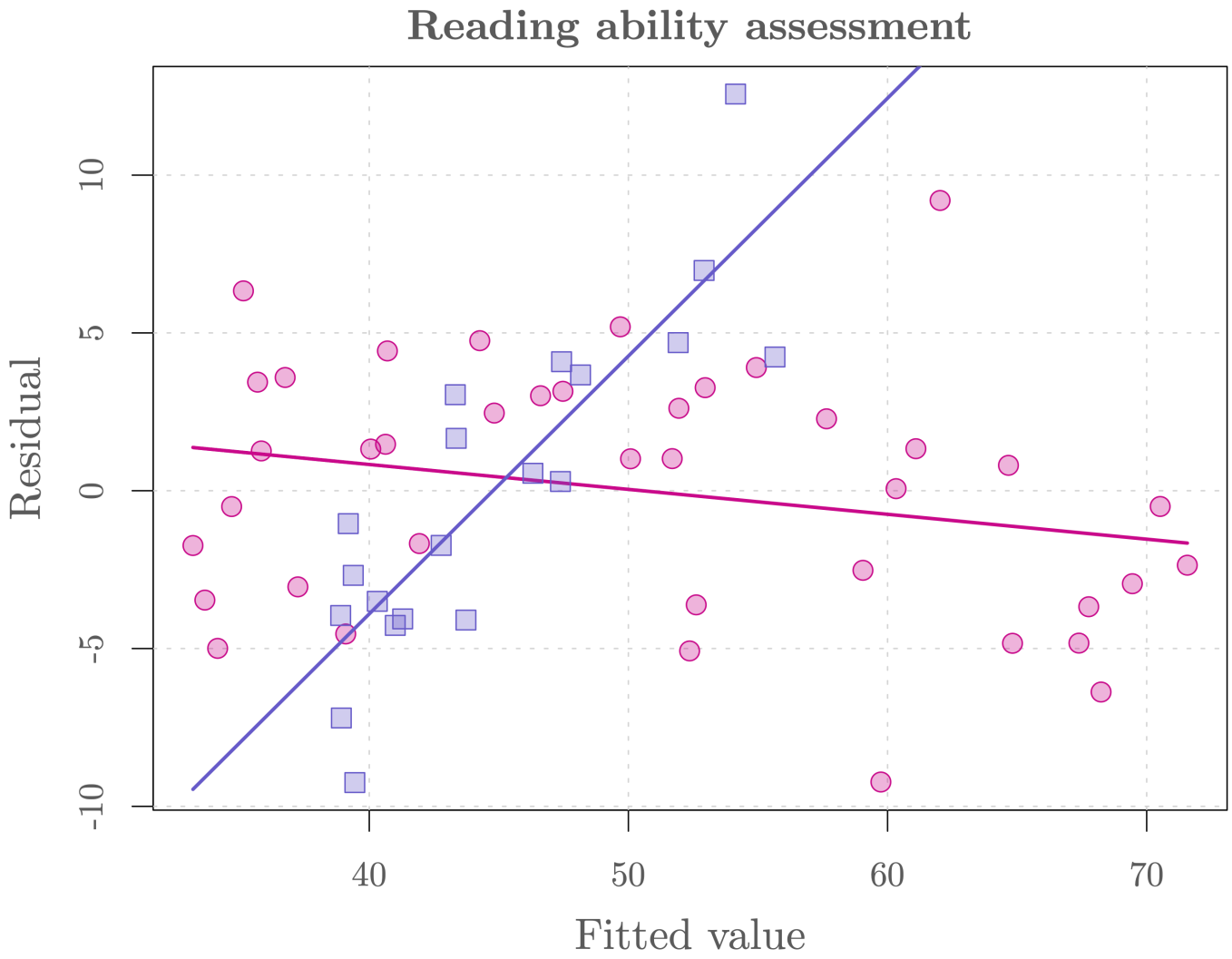
I DON'T TRUST LINEAR REGRESSIONS WHEN IT'S HARDER TO GUESS THE DIRECTION OF THE CORRELATION FROM THE SCATTER PLOT THAN TO FIND NEW CONSTELLATIONS ON IT.

👋 If you want to know more have a look here.

Model diagnostic

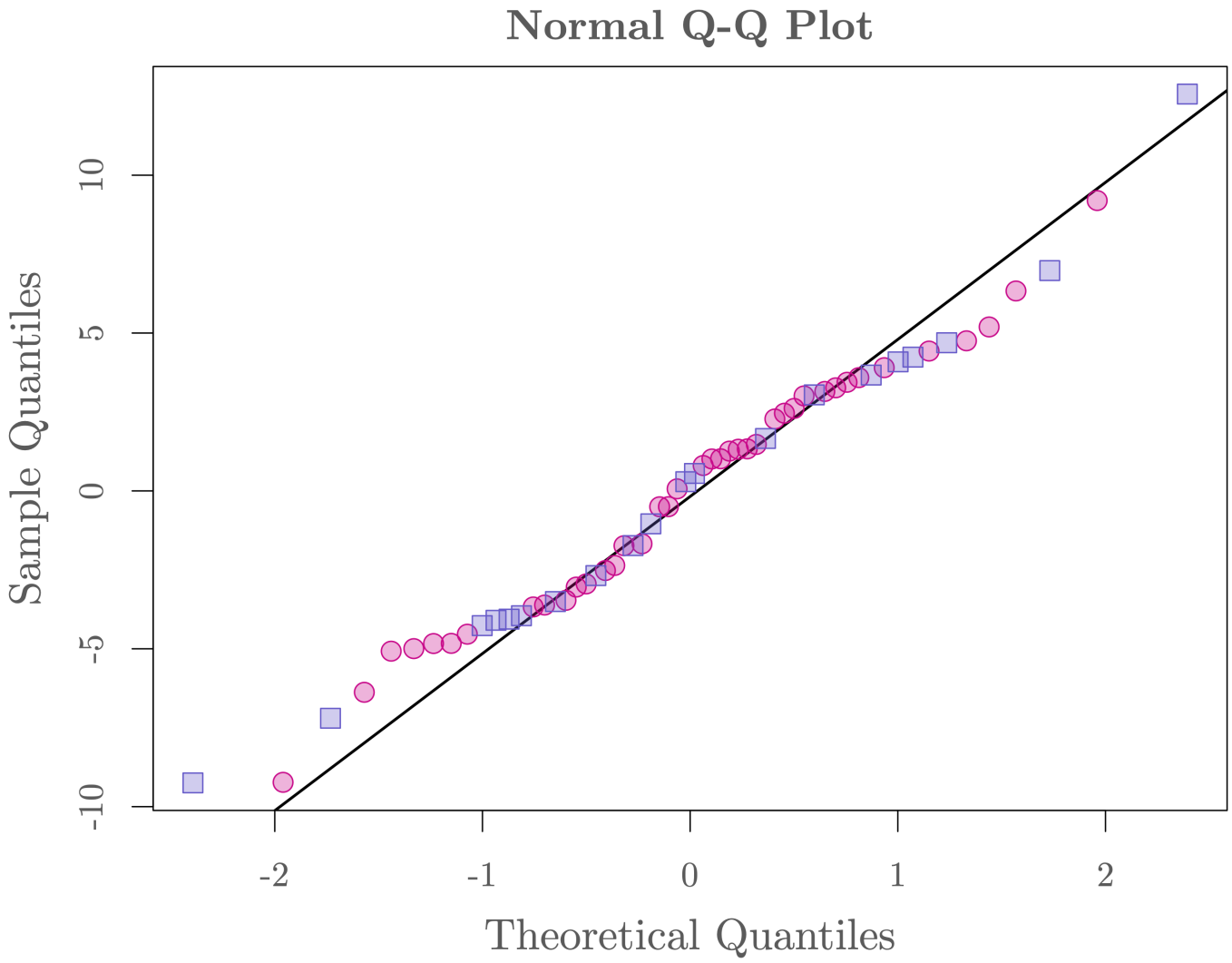


Model diagnostic ⚠️



Model diagnostic ⚠️

Model diagnostic



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:

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

- 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

R Code | Output

Here is the code to fit our second model:

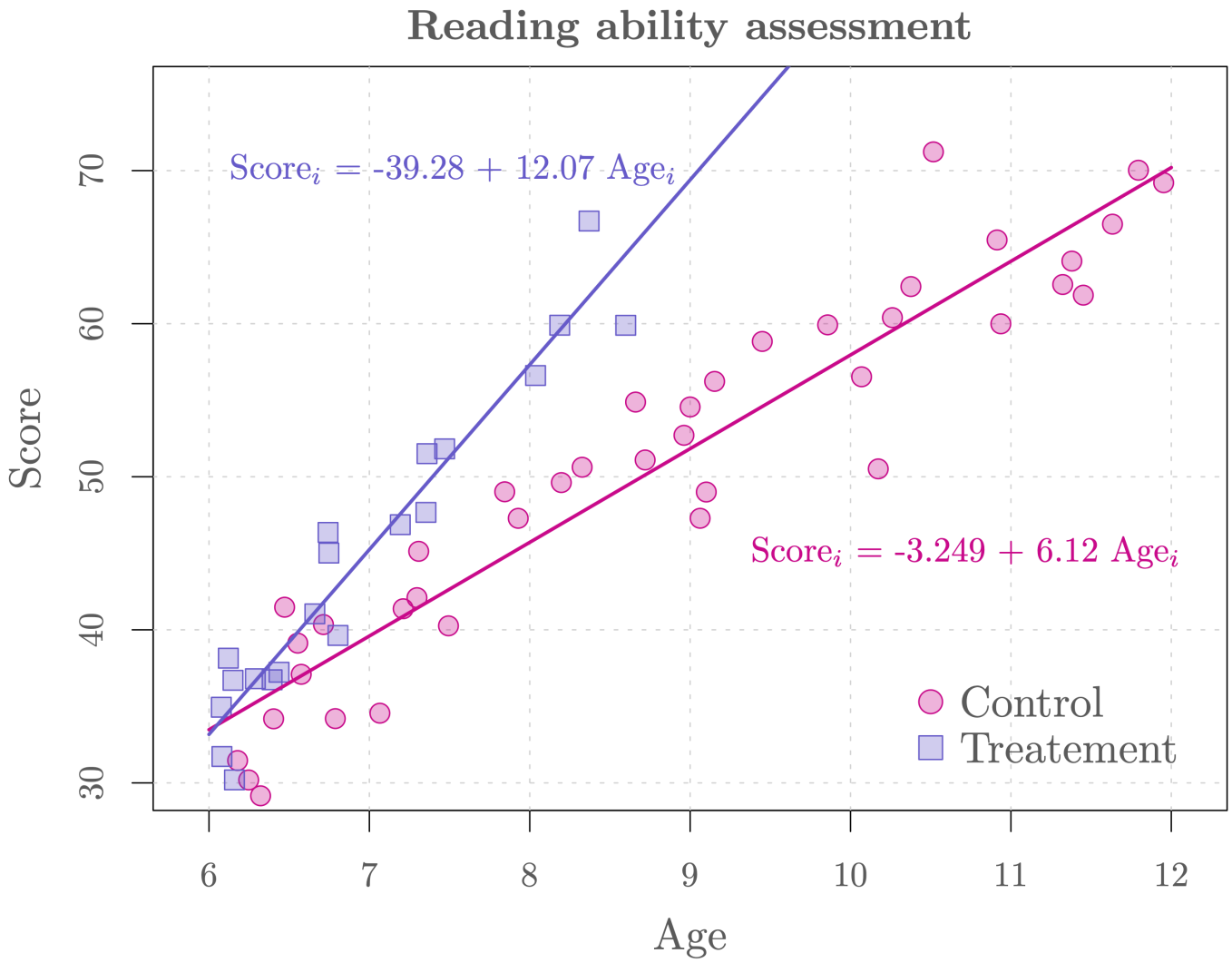
```
# Import data (if you haven't already)
dat = read.csv("data/reading.csv")

# Fit linear regression model
mod2 = lm(score ~ group*age, data = dat)
summary(mod2)
```

Interpretation of coefficients

We can obtain the estimated coefficients. Specifically,

- $\hat{\beta}_0 = -\mathbf{3.2489}$ represents the estimated baseline average score of the DRP test at birth (but *again* what does it mean? 🙄)
- $\hat{\beta}_1 = -\mathbf{36.0307}$ 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? 🧐).
- $\hat{\beta}_2 = \mathbf{6.1207}$ 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.
- $\hat{\beta}_3 = \mathbf{5.9539}$ 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.

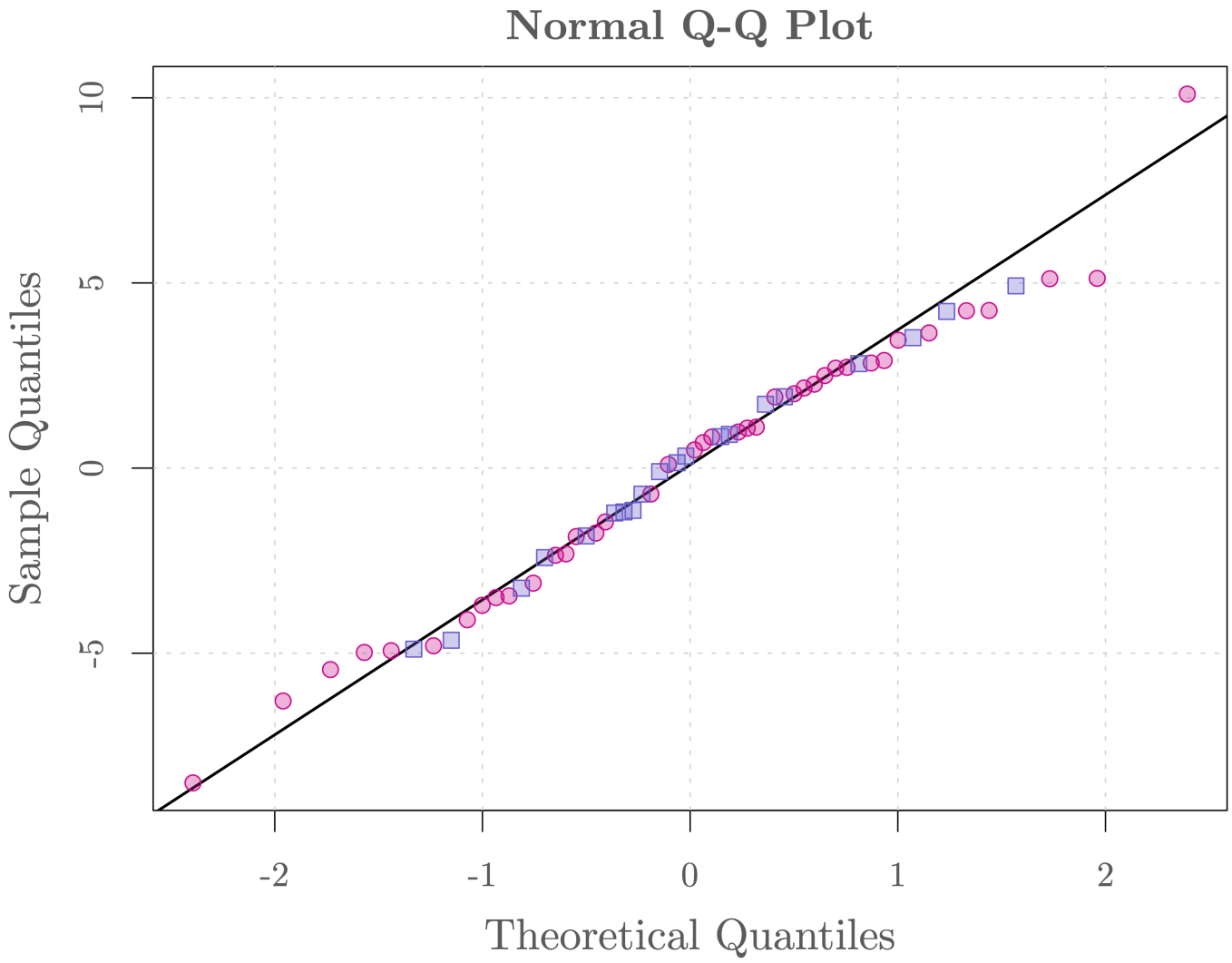
Model fit


Reading ability assessment
Score
Age
30
40
50
60
70
6
7
8
9
10
11
12
Score$_i$ = -39.28 + 12.07 Age$_i$
Score$_i$ = -3.249 + 6.12 Age$_i$
Control
Treatement

Model diagnostic

Model diagnostic

Model diagnostic



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

```
AIC(mod1)     # First model (no interaction)
```

```
#> [1] 354.2688
```

```
AIC(mod2)     # Second model (with interaction)
```

```
#> [1] 328.9095
```

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.

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

- 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

R Code | Output | AIC

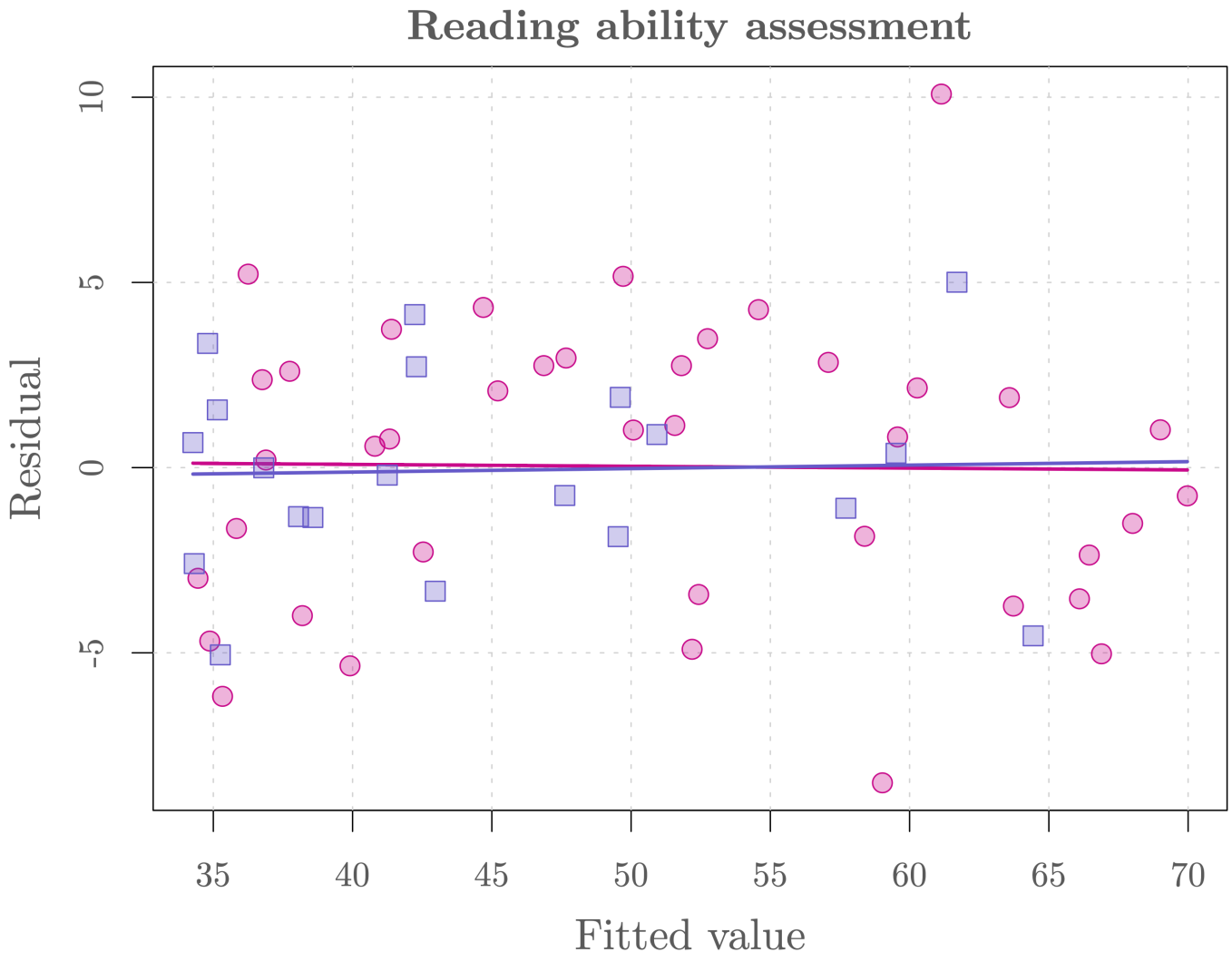
Here is the code to fit our third model:

```
# Import data (if you haven't already)
dat = read.csv("data/reading.csv")
dat$age_minus_6 = dat$age - 6

# Fit linear regression model
mod3 = lm(score ~ age_minus_6 + group:age_minus_6, data = dat)
summary(mod3)
```

Model fit

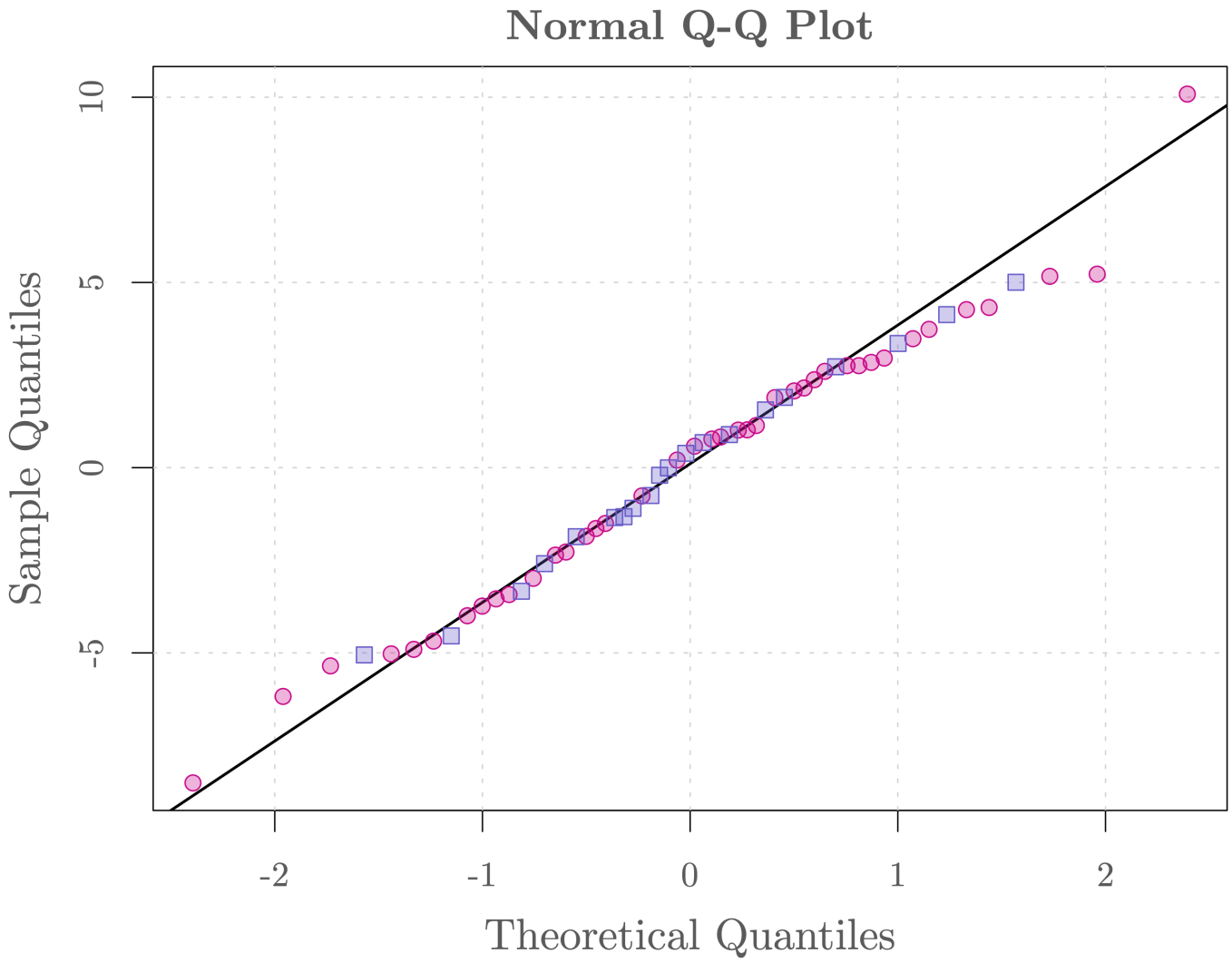

Reading ability assessment
Score
Age
30
40
50
60
70
6
7
8
9
10
11
12
$\text{Score}_i = 33.35 + 11.96\ (\text{Age}_i - 6)$
$\text{Score}_i = 33.35 + 6.15\ (\text{Age}_i - 6)$
Control
Treatement

Model diagnostic



Model diagnostic

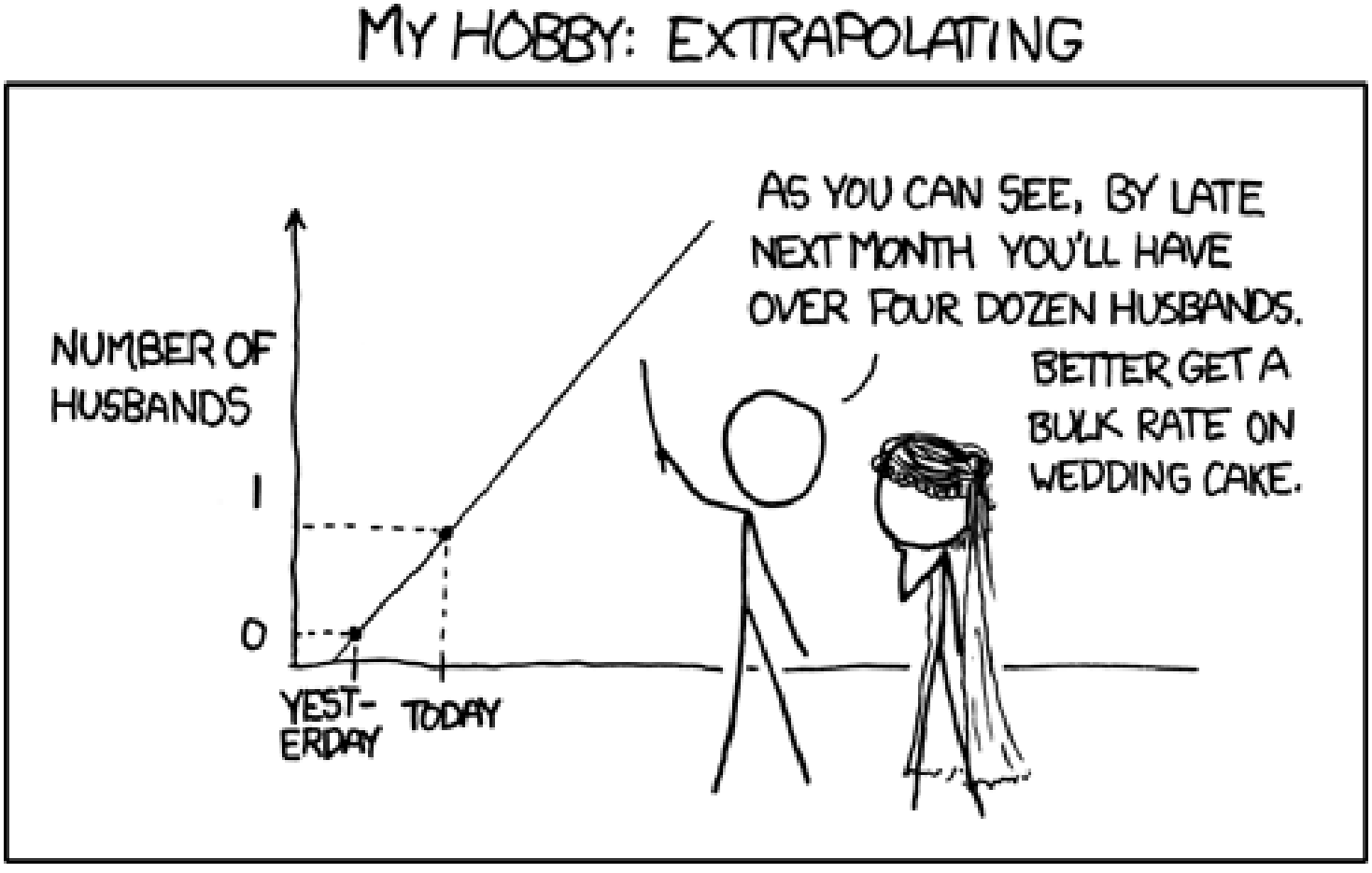
Model diagnostic



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? 😯

Extrapolating



👋 If you want to know more have a look here.

Exercise: Crime Rate

Problem | Import

A police governor is interested in studying the effects of various variables by the US states on the **crime rate**. To study this issue, 47 crime rate data were collected, together with the following variables: `Youth` (number of males aged 18-24 per 1k), `Southern` (southern state, 1=yes, 0=no), `Education` (education time), `Expenditure` (expenditure on police), `LabourForce` (number of young men employed per 1k), `Males` (number of males per 1k females), `StateSize` (state size in hundred thousands), `HighYouthUnemploy` (1=yes, 0=no), `Wage` (median weekly wage), and `BelowWage` (number of families below half wage per 1k). Based on this data, can you find a suitable model to predict the crime rate?

Exercise: Crime Rate

Full Back Back For For Comp Check

We start by fitting an initial model with all covariates included (without interactions):

```
fit.full = lm(CrimeRate ~ ., data = crime)
summary(fit.full)
```

We can see that some variables appear not significant, implying that we may be able to find a smaller model with less variables.