R Stats Bootcamp

2.12 - ANOVA

Megan Lewis

2025-05-01

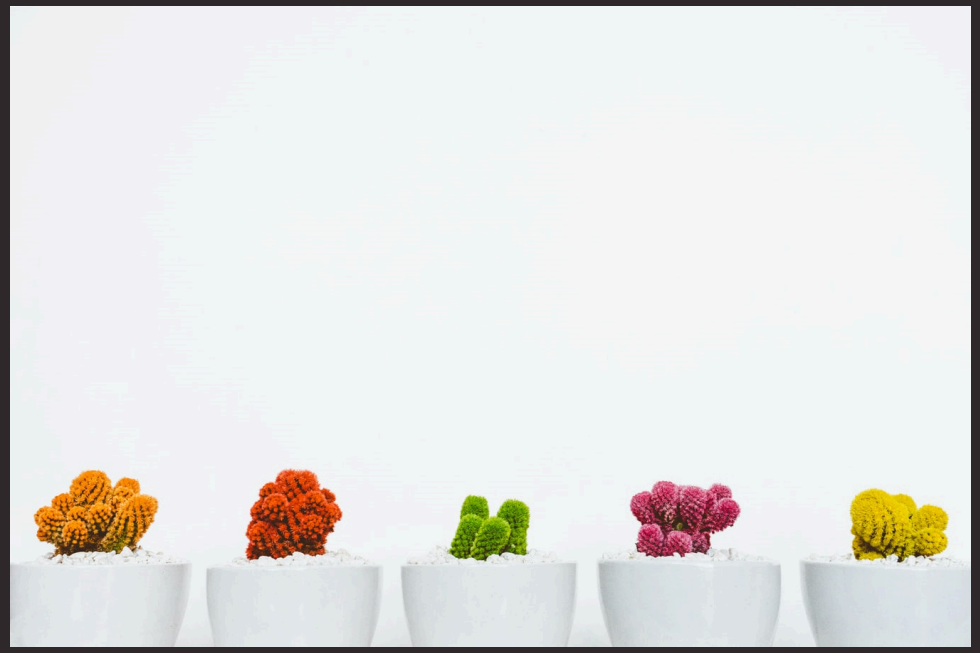
R stats bootcamp - Module 2

Schedule:

- ~~Session 7: Explore data~~
- ~~Session 8: Distributions~~
- ~~Session 9: Correlation~~
- ~~Session 10: Regression~~
- ~~Session 11: T-test~~
- Session 12: ANOVA

R Stats Bootcamp

Arrange the arithmetic to compare more than two groups



Session 12 objectives:

- The question of 1-way ANOVA
- Data and assumptions
- Graphing
- Tests and alternatives
- Practice exercises

*AN*alysis *O*f *VA*riance

“The analysis of variance is not a mathematical theorem, but rather a convenient method of arranging the arithmetic.” - R. A. Fisher (via Wishart 1934. Sppl. J. Roy. Soc. 1(1):26-61.)

*AN*alysis *O*f *VA*riance

- Revolutionizing the idea of objectivity in using data to produce evidence
- Invented by R. A. Fisher while ar Rothamsted
- Intention that ANOVA would be useful tool for analysis of agricultural experiments
- Remains foundation of best practice of statistics

The question of one-way ANOVA

- Several reasons to use ANOVA
- Usual scenario:
  - one numeric continuous dependent variable
  - a factor with two or more levels (often with a control)
    - When only two levels, the 1-way ANOVA is conceptually equiv. to the t-test

The question of one-way ANOVA

- Example:
  - Classic field trial
    - Crop pest damage measured (numeric continuous variable)
    - Factor with three levels:
      - Control (No pesticide)
      - Organic pesticide
      - Chemical pesticide

The question of one-way ANOVA

- Basic question:
  - Is there an overall difference in numeric independent variable (pest damage) amongst the factor levels (treatment)

The question of one-way ANOVA

- Several kinds of question are also possible to answer:
  - Overall difference test of means between factor levels
  - Comparision of difference of each factor level with control or other reference factor level
  - posthoc tests of difference between specific factor levels (i.e., pairwise tests)
  - Examination of the “sources of variation” obsevred in dependent variable (i.e., what prop. of total var can be accounted for by the factor)

The question of one-way ANOVA

- The test statistic for ANOVA = F-ratio
  - Proportion of variance in the dependent variable between groups, relative to that within the categories

Data and Assumptions

- R demo:
  - Explore these in R
  - Experiment in animal genetics
    - Weight of male chickens (8-week old weight in grams)
    - Factor variable = Sire identity
      - Five levels: A, B, C, D, E

Wide Data vs. Long Data

- Data in wide or long format
  - Wide format: storing data in five vectors (one per factor level)
    - Not the usual way to store data nowadays
  - Long format: numeric data in single vector with sire vector
    - More preferred layout, conforms more to “Tidy Data” principles

Off to R!

ANOVA assumptions

- Similar to assumptions for regression
  - Both are a specific kind of linear model
- Most important to consider are:
  - Gaussian residuals
  - Homoscedastcity
  - Equality of variance
  - Independent observations

Off to R!

- Checking assumptions in R
  - Gaussian residuals: Test graphically and using NHST
  - Homoscedastcity: Test graphically with residuals vs fitted plot
  - Equality of variance: Test graphically with plot of residual vs factor and NHST for equal variance
  - Independent observations: Not testing formally, will assume with example data
    - Would likely know based off of your experimental design

Graphing

- Classic graphing: Boxplot
  - Including display of central tendency of data separately for each factor level
  - For continuous variables, boxplots show this perfectly
  - With count variables, barplots are sometimes used
    - Height set to mean
    - Error bar included

Off to R!

ANOVA : F test statistic and alternatives

- Basic application on ANOVA in R uses the `aov()` fucntion
  - BUT there are a lot of alternative ways to perform the EXACT same test in R
- ANOVA is essentially a subset of the Gaussian linear model
  - Which is a subset o the General Linear Model …

Performing the ANOVA

- Perform a one-way ANOVA
  - Look at basic output for *overall effect of sire*
  - Look at how to examine *contrasts* (i.e., diff between control or reference level)
  - Look at *post hoc* testing (i.e., pairwise comparisons)
  - Examine what happens in ANOVA in a little more detail…

Off to R!

Contrasts … aov() versus lm()

```
aov(formula = Weight ~ factor(Sire), data = new.long)

                Df    Sum Sq   Mean Sq  F value  Pr(>F)
factor(Sire)    4     17426    4356     1.872    0.137
Residuals       35    81442    2327
```

##

Contrasts … aov() versus lm()

Contrasts . . . aov() versus lm()

```
aov(formula = Weight ~ factor(Sire), data = new.long)



                Df     Sum Sq    Mean Sq   F value   Pr(>F)
factor(Sire)    4      17426     4356      1.872     0.137
Residuals       35     81442     2327
```

Contrasts . . . aov() versus lm()

```
aov(formula = Weight ~ factor(Sire), data = new.long)
```

Main effect
Residual error

Degrees of freedom

Test statistic

P-value

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
factor(Sire)	4	17426	4356	1.872	0.137
Residuals	35	81442	2327		

Contrasts . . . aov() versus lm()

```
aov(formula = Weight ~ factor(Sire), data = new.long)
```

```
                Df
factor(Sire)     4
Residuals       35
```

= number of factor levels (5) minus 1

= total number of observations (40) minus number of factor levels (5)

Contrasts . . . aov() versus lm()

```
Call:
lm(formula = Weight ~ Sire, data = new.long)

Residuals:
Min        1Q        Median     3Q        Max
-93.625    -29.312    -2.875     33.906     104.750

Coefficients:
             Estimate       Std. Error     t value    Pr(>|t|)
(Intercept)     696.63        17.05          40.846    <2e-16 ***
SireA            18.38        24.12           0.762     0.451
SireB           -31.50        24.12          -1.306     0.200
SireD           -39.13        24.12          -1.622     0.114
SireE           -13.38        24.12          -0.555     0.583

---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ‘ 1

Residual standard error: 48.24 on 35 degrees of freedom
Multiple R-squared: 0.1763,       Adjusted R-squared: 0.08211
F-statistic: 1.872 on 4 and 35 DF,    p-value: 0.1373
```

Contrasts . . . aov() versus lm()

```
Call:
lm(formula = Weight ~ Sire, data = new.long)

Residuals:
Min        1Q        Median     3Q        Max
-93.625    -29.312    -2.875     33.906    104.750

Coefficients:
             Estimate      Std. Error    t value    Pr(>|t|)
(Intercept)   696.63        17.05         40.846    <2e-16 ***
SireA          18.38        24.12          0.762     0.451
SireB         -31.50        24.12         -1.306     0.200
SireD         -39.13        24.12         -1.622     0.114
SireE         -13.38        24.12         -0.555     0.583

---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ‘ 1

Residual standard error: 48.24 on 35 degrees of freedom
Multiple R-squared: 0.1763,      Adjusted R-squared: 0.08211
F-statistic: 1.872 on 4 and 35 DF,    p-value: 0.1373
```

Contrasts . . . aov() versus lm()

5 rows

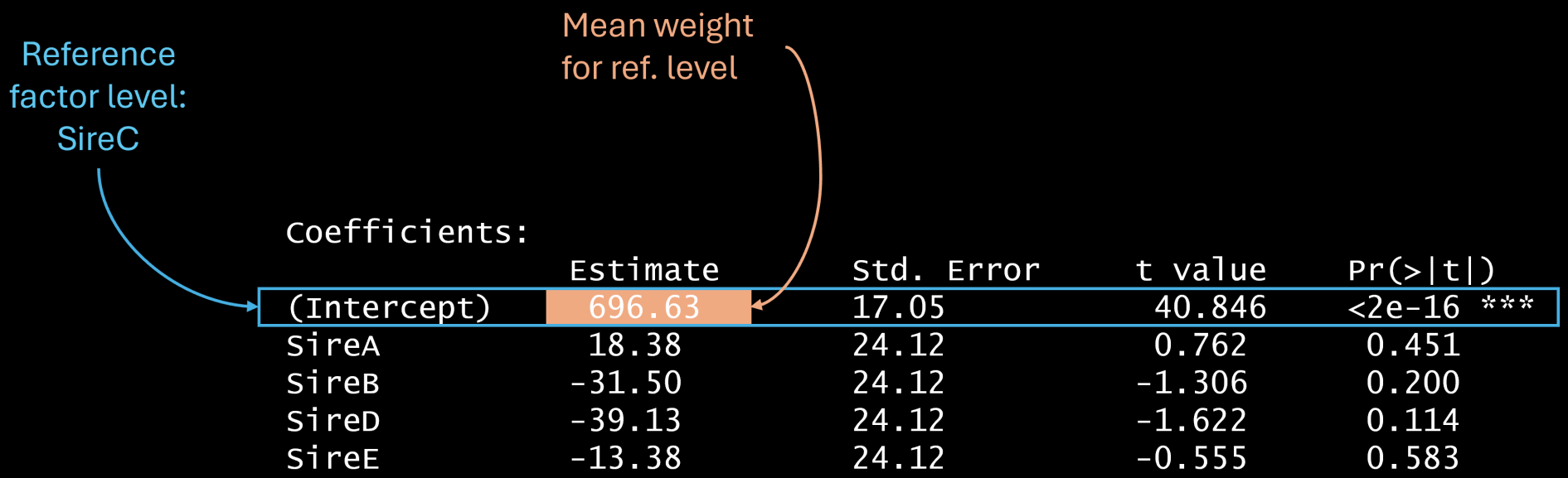
```
Coefficients:
             Estimate      Std. Error    t value    Pr(>|t|)
(Intercept)   696.63       17.05          40.846    <2e-16 ***
SireA          18.38       24.12           0.762     0.451
SireB         -31.50       24.12          -1.306     0.200
SireD         -39.13       24.12          -1.622     0.114
SireE         -13.38       24.12          -0.555     0.583
```

Contrasts … aov() versus lm()

Reference factor level: SireC

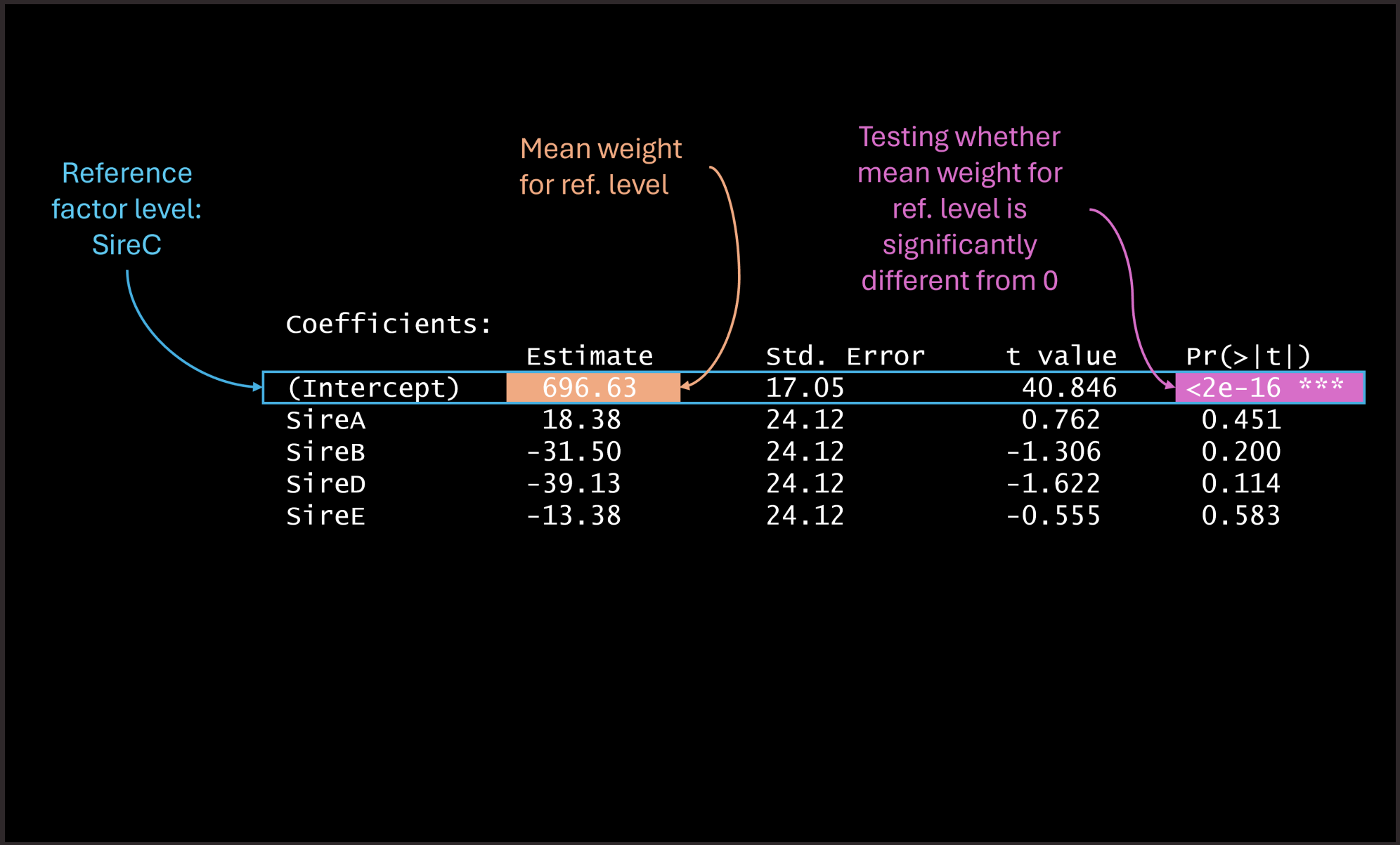
```
Coefficients:
             Estimate    Std. Error    t value    Pr(>|t|)
(Intercept)   696.63       17.05        40.846    <2e-16 ***
SireA          18.38       24.12         0.762     0.451
SireB         -31.50       24.12        -1.306     0.200
SireD         -39.13       24.12        -1.622     0.114
SireE         -13.38       24.12        -0.555     0.583
```

Contrasts … aov() versus lm()

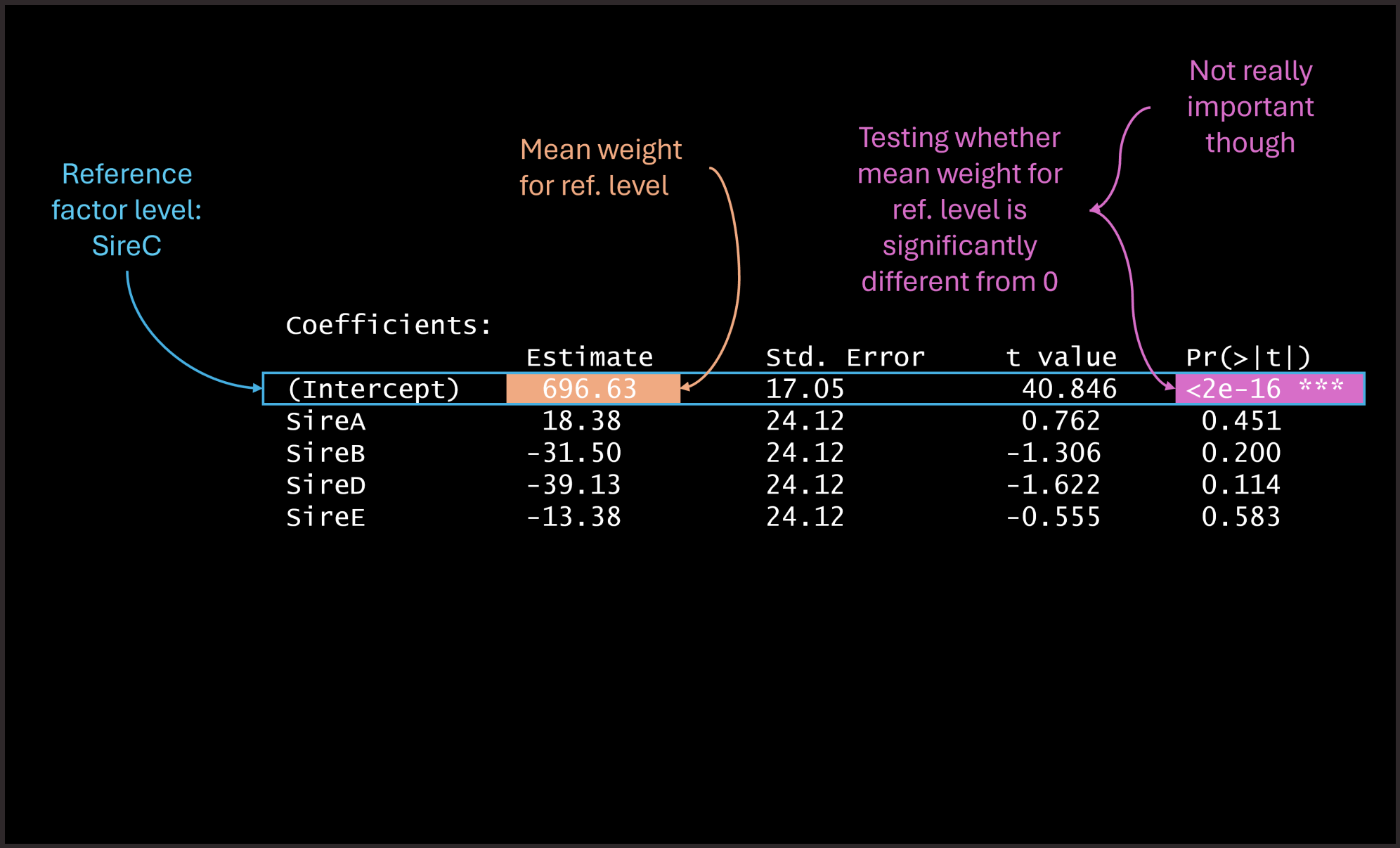



```
Coefficients:
             Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)   696.63      17.05      40.846  <2e-16 ***
SireA          18.38      24.12       0.762   0.451
SireB         -31.50      24.12      -1.306   0.200
SireD         -39.13      24.12      -1.622   0.114
SireE         -13.38      24.12      -0.555   0.583
```

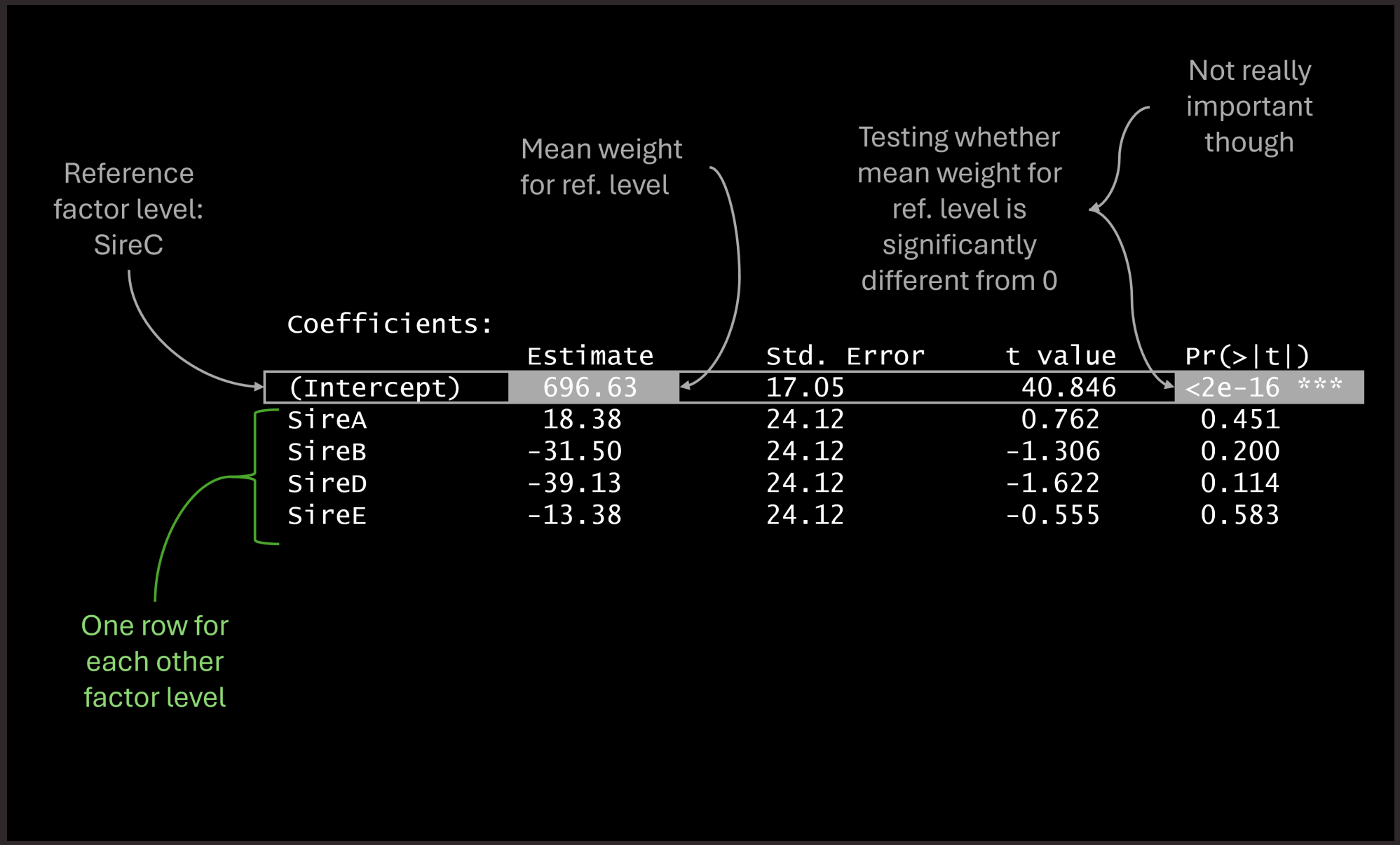
Contrasts . . . aov() versus lm()



Contrasts … aov() versus lm()



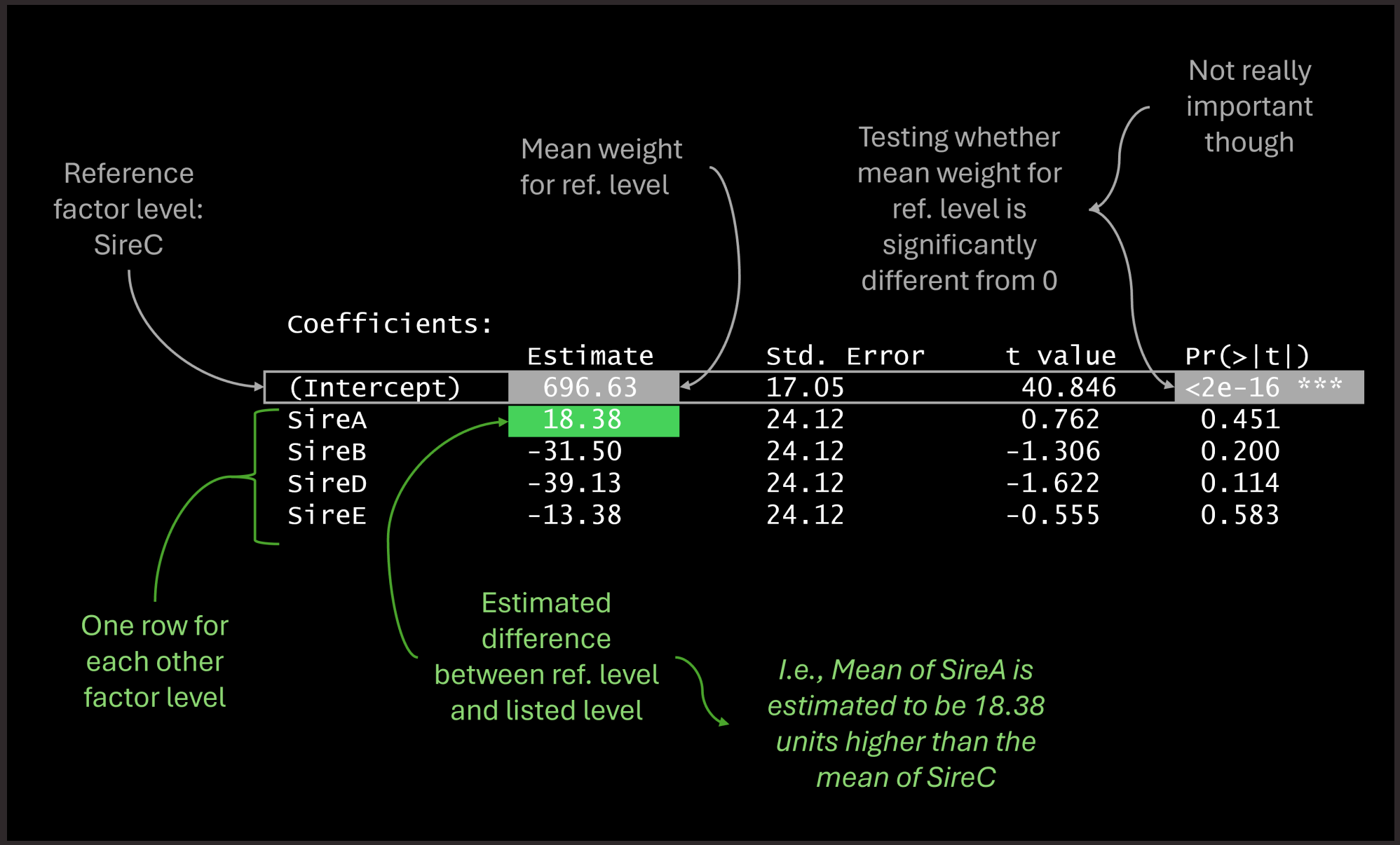
Contrasts … aov() versus lm()



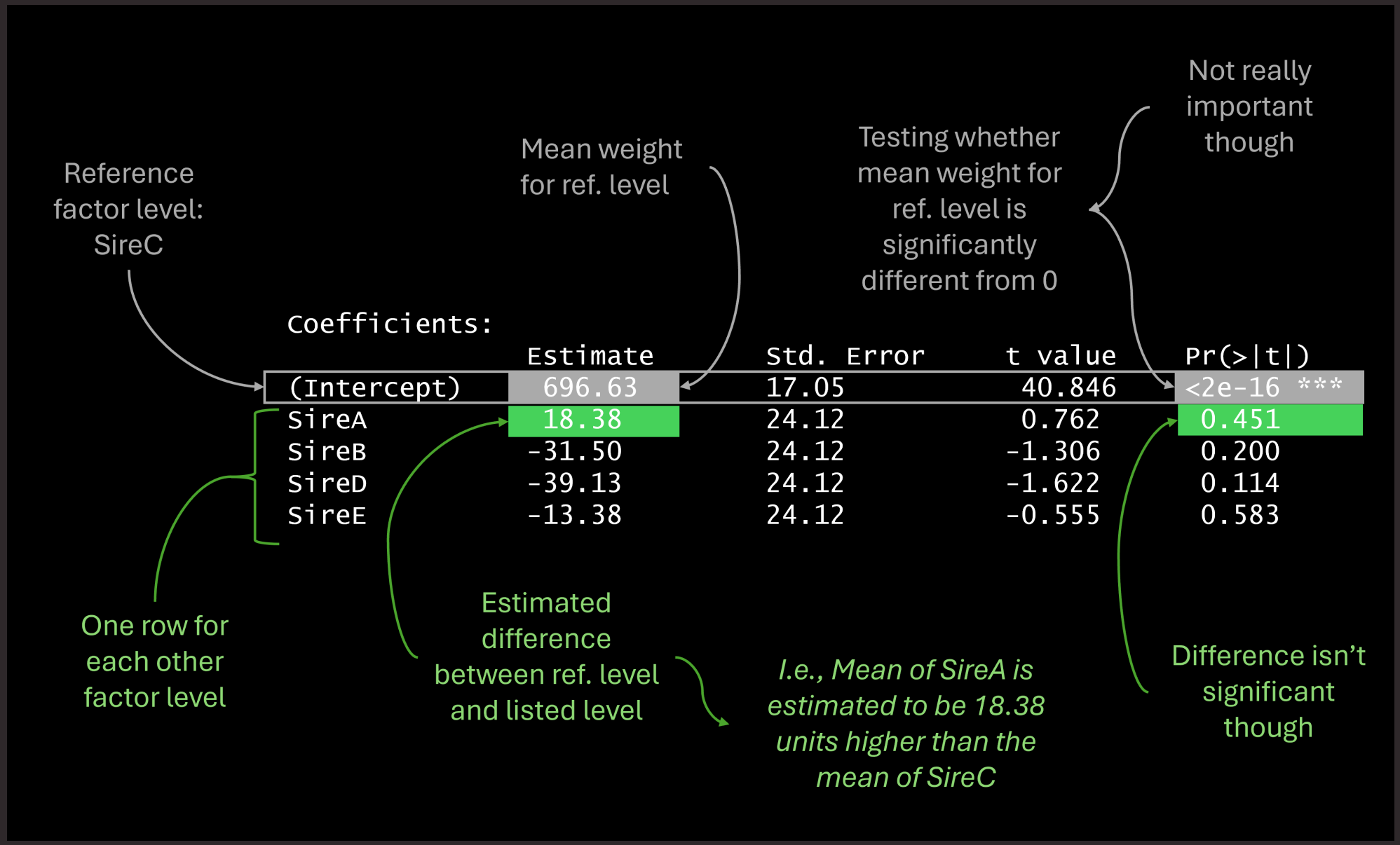

```
Coefficients:
            Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)   696.63       17.05   40.846  <2e-16 ***
SireA          18.38       24.12    0.762     0.451
SireB         -31.50       24.12   -1.306     0.200
SireD         -39.13       24.12   -1.622     0.114
SireE         -13.38       24.12   -0.555     0.583
```

Contrasts … aov() versus lm()

Contrasts … aov() versus lm()



Contrasts … aov() versus lm()



Post Hocs

- Overall question of “is there a difference?” can be enhanced by asking whether there are specific differences
- The phrase **post hoc** implies these questions are an ‘afterthought’
  - Thus, consideration should be given to whether these specific questions *NEED* to be asked

Type I errors

- The alpha (α) value which we compare our P-value can be interpreted as the (maximum) probability we are willing to accept of being wrong if we conclude there is significance in our data
  - i.e., We are are **happy** to accept the *chance* of our significant finding being wrong 5% of the time when alpha = 0.05.

Type I errors

- When multiple tests are made on the same data, it increases the chance of discovering a false positive, purely by chance alone, to above 5%
- As such there are methods used to avoid this by adjusting the p-value o keep the overall likelihood of a false positive error at 5%
- Examples:
  - The Bonferroni adjustment
  - Tukey HSD test

Bonferroni Adjustment

- Baseline, conservative adjustment to the alpha value to avoid false positive errors
- Typically applied in one-way ANOVA situation where specific, pairwise comparisons between means are required and aren’t covered by the overall tests or contrasts
- E.g., In our chicken data, we know there is no difference between offspring weights between Sire C and all other Sires - C:A, C:B, C:D, C:E
  - But don’t know about other comparisons i.e. A:B, A:D, A:E etc…

Tukey HSD

- Ideal for one-way ANOVA
- Less conservative than Bonferroni

Off to R!

Alternatives to one-way ANOVA

- When assumptions can not be met, some options:
  - Attempt to transform data (e.g., `sqrt()`, `log()`) to 'coerce' data to conform to assumptions
  - Use an alternative test for which assumptions are not violated
    - i.e., a non-parametric test like Kruskall Wallis

Off to R!

Non-parametric tests

- Advantages:
  - Very easy to use
  - Very easy to interpret
- Disadvantages
  - Less statistical power
  - i.e., less likely than parametric cousins to detect a significant effect even when one exists

ANOVA calculation details

ANOVA equations

$$(1)\ Grand\ Mean = GM = \sum_{i=1}^{k}\sum_{j=1}^{n} Y_{ij} \times \frac{1}{n_{total}}$$

$$(2)\ Sum\ of\ Squares\ Between = SS_{between} = \sum_{i=1}^{k} n_i(\bar{Y}_i - GM)^2$$

$$(3)\ Sum\ of\ Squares\ Within = SS_{within} = \sum_{i=1}^{k}\sum_{j=1}^{n} (Y_{ij} - \bar{Y}_i)^2$$

$$(4)\ Sum\ of\ Sqaures\ Total\ = SS_{between} + SS_{within}$$

ANOVA variables

- k = Number of factor levels
- n_i = number of samples for group i
- n_{total} = total number of samples for all groups
- Yij = observation j in group i
- $\bar{Y}_i$ = mean of group i

ANOVA sources of variation

Source of Variation	df	Sum Sq	Mean Sq	F
Between groups	$k - 1$	$Eq\ (2)$	$\frac{Eq\ (2)}{k-1}$	$\frac{MeanSq_{between}}{MeanSq_{within}}$
Within groups	$n_{total} - k$	$Eq\ (3)$	$\frac{Eq\ (3)}{n_{total}-k}$	
Total	$n_{total} - 1$	$EQ\ (4)$		

ANOVA ‘by hand’ in R

Practice Exercises

Practice excercise data

For these exercises, run the code below to recreate the data object pest. There are 40 rows and 2 variables with 2 variables: damage and treatment.

See bootcamp page

Think of this data as the result of an experiment looking at the effectiveness of pesticide treatment on leaf damage. Let us imagine that this experiment measured leaf damage (variable "damage" measured in mm squared) and that the plants were treated with one of 4 `treatment` levels: - `Control` - `x.half` - `x.full` - `organic`

Practice excercise data

The experiment is of course designed to look at an overall effect the various treatments may have to reduce leaf damage relative to the control. In addition, it is of interest to examine the effect of the organic treatment compared to that of the x.half to the x.full.

The experiment ran using 40 potted plants spaced 1m from each other in a greenhouse setting. Each treatment was randomly assigned to 10 plants. Onto each plant was placed 5 red lily beetle (Lilioceris lilii) pairs.

Practice exercise 1

- Make a good, appropriate graph representing the overall experiment.
- Show your code.
- Describe any trends in the data that are apparent from the graph, as well as an initial assessment of principle assumptions of 1-way ANOVA based only on your single graph.

Practice exercise 2

- Test the assumption of Gaussian residuals for 1-way ANOVA using any graphs or NHST approach that you deem appropriate.
- Show your code and briefly describe your EDA findings and conclusion as to whether these data adhere to the Gaussian assumption.

Practice exercise 3

- Test the assumption of homoscedasticity of residuals for 1-way ANOVA using any graphs or NHST approach that you deem appropriate.
- Show your code and briefly describe your EDA findings and conclusion as to whether these data adhere to the homoscedasticity assumption.

Practice exercise 4

- Perform either a 1-way ANOVA or an appropriate alternative based on your findings in the previous answers.
- Show your code, state your results in the technical style and briefly interpret your findings.

Practice exercise 5

- Perform an appropriate set of post hoc tests to compare pairwise mean differences in these data.
- Focus on the post hoc questions of interest:
  - Is the organic pesticide effective?
  - Does dose matter in the non-organic treatments?

Practice exercise 6

- Write a plausible practice question involving any aspect of data handling, graphing or analysis for the 1-way ANOVA framework for the iris data (`data(iris); help(iris)`).