Analysis of Covariance

Cathy Maugis-Rabusseau

INSA Toulouse / IMT
GMM 116
cathy.maugis@insa-toulouse.fr

2021-2022

Outline

Context and notation

- ANCOVA= Analysis of covariance
- We want to explain a **quantitative** response variable Y using **qualitative** and **quantitative** variables together
- Here we only consider one covariate z and one factor T with I levels
- n_i = number of observations for the i-th level of T, $n = \sum_{i=1}^{I} n_i$.
- Y_{ij} = value of the response Y for $j = 1, \ldots, n_i$, $i = 1, \ldots, I$
- z_{ij} = value of the covariate z for $j = 1, \ldots, n_i$, $i = 1, \ldots, I$

Example

We want to find if temperature and oxygenation conditions influence the evolution of oyster weight. We have $n = 20$ bags of 10 oysters. We place, during a month, these 20 bags randomly in $I = 5$ different locations of a channel cooling of a power station at the rate of $n_i = 4$ bags per location. These locations are differentiated by their temperature and oxygenation.

For each bag, we have

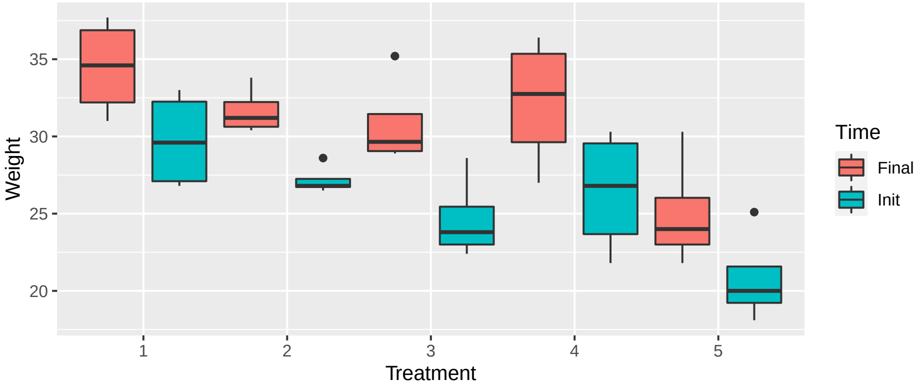
- its weight after the experiment (*Final weight*) = the response Y
- its weight before the experiment (*Init weight*) = the explanatory variable z
- the location (*Treatment* - 1 to 5) = the qualitative variable T

Example

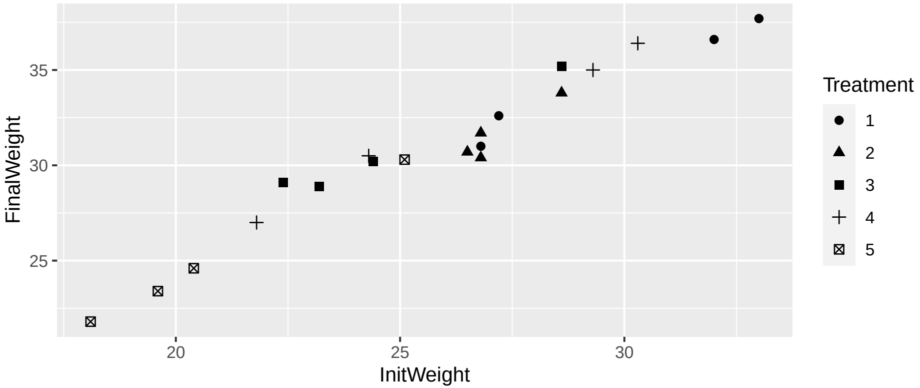
```
print(oyster)
```

```
   InitWeight FinalWeight Treatment
1        27.2        32.6         1
2        32.0        36.6         1
3        33.0        37.7         1
4        26.8        31.0         1
5        28.6        33.8         2
6        26.8        31.7         2
7        26.5        30.7         2
8        26.8        30.4         2
9        28.6        35.2         3
10       22.4        29.1         3
11       23.2        28.9         3
12       24.4        30.2         3
13       29.3        35.0         4
14       21.8        27.0         4
15       30.3        36.4         4
16       24.3        30.5         4
17       20.4        24.6         5
18       19.6        23.4         5
19       25.1        30.3         5
20       18.1        21.8         5
```

Example



Example



Outline

2. **Modelings**

Regular model

- Model:

$$(MR) : \begin{cases} Y_{ij} = a_i + b_i\ z_{ij} + \varepsilon_{ij}, & \forall i = 1, \cdots, I, \forall j = 1, \cdots, n_i \\ \varepsilon_{ij} \text{ i.i.d } \sim \mathcal{N}(0, \sigma^2) \end{cases}$$

$\Leftrightarrow$ Estimating a linear regression of Y on z for each level i of the factor T.

$$\underbrace{\begin{pmatrix} Y_{(1)} \\ \vdots \\ \vdots \\ Y_{(I)} \end{pmatrix}}_{Y} = \underbrace{\begin{pmatrix} X_{(1)} & & & \\ & X_{(2)} & & \\ & & \ddots & \\ & & & X_{(I)} \end{pmatrix}}_{X} \underbrace{\begin{pmatrix} a_1 \\ b_1 \\ \vdots \\ a_I \\ b_I \end{pmatrix}}_{\theta} + \underbrace{\begin{pmatrix} \varepsilon_{(1)} \\ \vdots \\ \vdots \\ \varepsilon_{(I)} \end{pmatrix}}_{\varepsilon}$$

with $Y_{(i)} = (Y_{i1}, \ldots, Y_{in_i})'$, $X_{(i)} = \left(\mathbb{1}_{n_i}, z_{(i)}\right)$.

Singular model

$$(MS): \begin{cases} Y_{ij} = (\mu + \alpha_i) + (\beta + \gamma_i) z_{ij} + \varepsilon_{ij}, \quad \forall i = 1, \cdots, I, \forall j = 1, \cdots, n_i. \\ \varepsilon_{ij} \text{ i.i.d } \sim \mathcal{N}(0, \sigma^2) \end{cases}$$

- In this parametrization,
  - interaction effect between the covariate z and the factor T: γ_i
  - differential effect of the factor T on Y: α_i
  - differential effect of the covariate z on Y: β
- $2I + 2$ parameters $\Rightarrow$ 2 constraints are required to model identifiability

Outline

Estimation in regular model (MR)

In a regular model, $\hat{\theta} = (X'X)^{-1}X'Y$.

Since $X = \text{diag}(X_{(1)}, \ldots, X_{(I)})$, we have

$$(X'X)^{-1} = \text{diag}((X'_{(1)}X_{(1)})^{-1}, \ldots, (X'_{(I)}X_{(I)})^{-1})$$

and

$$X'Y = \text{diag}(X'_{(1)}Y_{(1)}, \ldots, X'_{(I)}Y_{(I)})$$

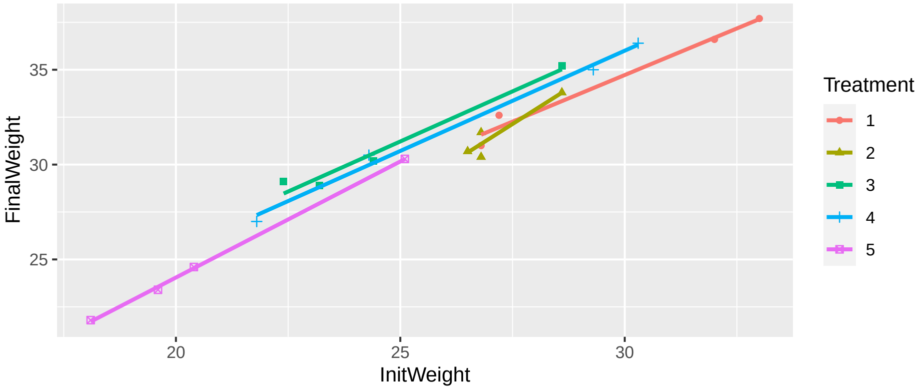
Thus

$$\hat{\theta} = \begin{pmatrix} (X'_{(1)}X_{(1)})^{-1}X'_{(1)}Y_{(1)} \\ \vdots \\ (X'_{(I)}X_{(I)})^{-1}X'_{(I)}Y_{(I)} \end{pmatrix}$$

Using results in simple linear regression, we deduce

$$\begin{cases} \widehat{b}_i = \text{cov}(Y_{(i)}, z_{(i)})/\text{var}(z_{(i)}) \\ \widehat{a}_i = \bar{Y}_{(i)} - \bar{z}_{(i)}\widehat{b}_i \end{cases}$$

Example



Estimation in singular model (MS)

- Identifiability constraints: by default in R $\alpha_1 = \gamma_1 = 0$
- Using the link between the parameters in (MR) and (MS), we can easily deduce

$$\begin{cases} \widehat{\mu} = \widehat{a_1} \\ \\ \widehat{\alpha_i} = \widehat{a}_i - \widehat{a_1} \\ \\ \widehat{\beta} = \widehat{b_1} \\ \\ \widehat{\gamma}_i = \widehat{b}_i - \widehat{b_1} \end{cases}$$

Example

```
complet<-lm(FinalWeight~InitWeight * Treatment,data=oyster)
summary(complet)
```

```
Call:
lm(formula = FinalWeight ~ InitWeight * Treatment, data = oyster)

Residuals:
     Min       1Q   Median       3Q      Max
-0.68699 -0.28193  0.02184  0.10425  0.63075

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)            5.24126    2.86473   1.830   0.0972 .
InitWeight             0.98265    0.09588  10.249 1.27e-06 ***
Treatment2           -14.39058    9.15971  -1.571   0.1472
Treatment3            -0.42330    3.97747  -0.106   0.9174
Treatment4            -0.94550    3.50725  -0.270   0.7930
Treatment5            -5.67309    3.57150  -1.588   0.1433
InitWeight:Treatment2  0.51871    0.33406   1.553   0.1515
InitWeight:Treatment3  0.07342    0.14699   0.499   0.6282
InitWeight:Treatment4  0.07428    0.12229   0.607   0.5571
InitWeight:Treatment5  0.24124    0.13980   1.726   0.1151
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5324 on 10 degrees of freedom
Multiple R-squared:  0.9921,    Adjusted R-squared:  0.985
F-statistic: 139.5 on 9 and 10 DF,  p-value: 2.572e-09
```

Example

```
import statsmodels.api as sm
from statsmodels.formula.api import ols
oysterpy=r.oyster;
completpy = ols('FinalWeight ~ InitWeight * Treatment', data=oysterpy).fit();
completpy.summary()
```

```
<class 'statsmodels.iolib.summary.Summary'>
"""
                            OLS Regression Results
==============================================================================
Dep. Variable:            FinalWeight   R-squared:                       0.992
Model:                            OLS   Adj. R-squared:                  0.985
Method:                 Least Squares   F-statistic:                     139.5
Date:                Jeu, 14 oct 2021   Prob (F-statistic):           2.57e-09
Time:                        21:57:22   Log-Likelihood:                -8.8384
No. Observations:                  20   AIC:                             37.68
Df Residuals:                      10   BIC:                             47.63
Df Model:                           9
Covariance Type:            nonrobust
===============================================================================================
                                coef    std err          t      P>|t|      [0.025      0.975]
-----------------------------------------------------------------------------------------------
Intercept                     5.2413      2.865      1.830      0.097      -1.142      11.624
Treatment[T.2]              -14.3906      9.160     -1.571      0.147     -34.800       6.019
Treatment[T.3]               -0.4233      3.977     -0.106      0.917      -9.286       8.439
Treatment[T.4]               -0.9455      3.507     -0.270      0.793      -8.760       6.869
Treatment[T.5]               -5.6731      3.572     -1.588      0.143     -13.631       2.285
InitWeight                    0.9826      0.096     10.249      0.000       0.769       1.196
InitWeight:Treatment[T.2]     0.5187      0.334      1.553      0.152      -0.226       1.263
InitWeight:Treatment[T.3]     0.0734      0.147      0.499      0.628      -0.254       0.401
InitWeight:Treatment[T.4]     0.0743      0.122      0.607      0.557      -0.198       0.347
InitWeight:Treatment[T.5]     0.2412      0.140      1.726      0.115      -0.070       0.553
------------------------------------------------------------------------------
```

Outline

Absence of any effect

- We want to compare the "null model"

$$(M0) : Y_{ij} = \mu + \varepsilon_{ij}, \forall i = 1, \cdots, I, \forall j = 1, \cdots, n_i$$

against the full model (MS)

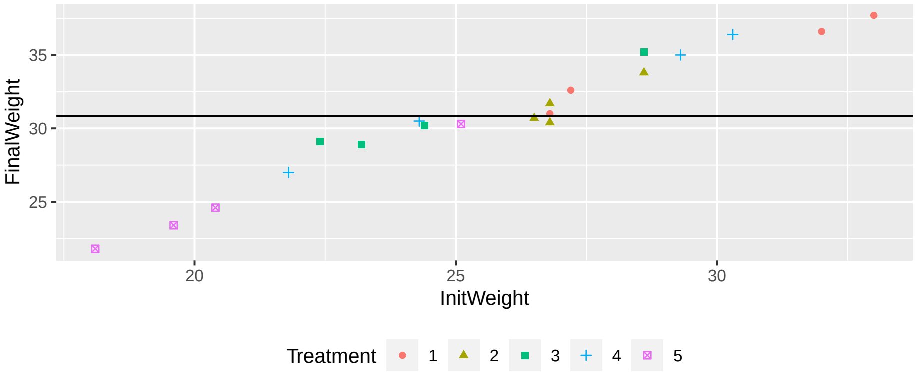
$$(MS) : \; Y_{ij} = (\mu + \alpha_i) + (\beta + \gamma_i) z_{ij} + \varepsilon_{ij}, \forall i = 1, \cdots, I, \forall j = 1, \cdots, n_i.$$

- Fisher's test statistics:

$$F = \frac{SSE/(2I-1)}{SSR/n-2I} \underset{\mathcal{H}_0}{\sim} \mathcal{F}(2I-1, n-2I)$$

with $SSR = \|Y - \widehat{Y}\|^2$ and $SSE = \|\widehat{Y} - \bar{Y}\mathbb{1}_n\|^2$

Example



Example

- With R:

```
M0<-lm(FinalWeight~1,data=oyster)
anova(M0,complet)
```

```
Analysis of Variance Table

Model 1: FinalWeight ~ 1
Model 2: FinalWeight ~ InitWeight * Treatment
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1     19 358.67
2     10   2.83  9    355.84 139.51 2.572e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- With Python:

```
from statsmodels.stats.anova import anova_lm
M0py = ols('FinalWeight~1', data=oysterpy).fit()
anova_lm(M0py,completpy)
```

```
   df_resid         ssr  df_diff     ss_diff           F        Pr(>F)
0      19.0  358.669500      0.0         NaN         NaN           NaN
1      10.0    2.834009      9.0  355.835491  139.510053  2.572066e-09
```

Test of non-interaction between factor and covariate

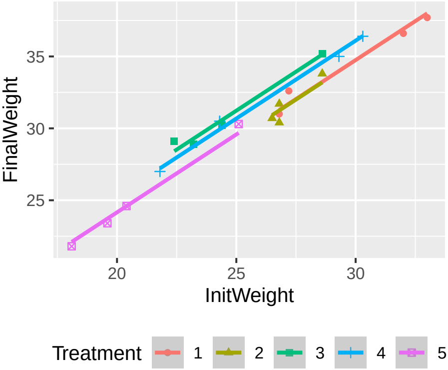
- We want to test the null hypothesis:

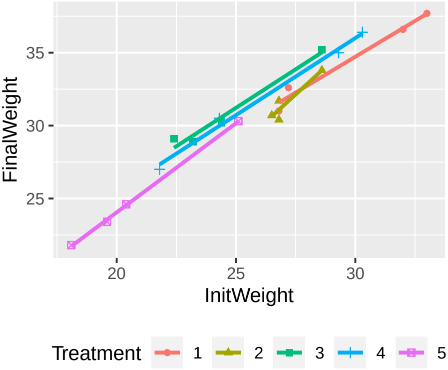
$$\mathcal{H}_0^{(SI)} : b_1 = b_2 = \cdots = b_I \Longleftrightarrow \gamma_1 = \gamma_2 = \cdots = \gamma_I = 0$$

- Fisher's test to compare
  - the full model
    (MS) : $Y_{ij} = (\mu + \alpha_i) + (\beta + \gamma_i) z_{ij} + \varepsilon_{ij}, \forall i = 1, \cdots, I, \forall j = 1, \cdots, n_i$
    (MR) : $Y_{ij} = a_i + b_i z_{ij} + \varepsilon_{ij}, \forall i = 1, \cdots, I, \forall j = 1, \cdots, n_i$
  - the sub-model with non-interaction
    $(MSnonI)$: $Y_{ij} = \mu + \alpha_i + \beta z_{ij} + \varepsilon_{ij}, \forall i = 1, \cdots, I, \forall j = 1, \cdots, n_i.$
    $(MRnonI)$: $Y_{ij} = a_i + b z_{ij} + \varepsilon_{ij}, \forall i = 1, \cdots, I, \forall j = 1, \cdots, n_i.$
- Test statistics:

$$F = \frac{SSR_{nonI} - SSR/(I-1)}{SSR/(n-2I)} \underset{\mathcal{H}_0}{\sim} \mathcal{F}(I-1, n-2I)$$

Test of non-interaction between factor and covariate





Model with non-interaction

```r
nonI<-lm(FinalWeight~InitWeight+Treatment)
summary(nonI)
```

```
Call:
lm(formula = FinalWeight ~ InitWeight + Treatment)

Residuals:
    Min      1Q  Median      3Q     Max 
-0.8438 -0.3154 -0.2171  0.4863  0.8871 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  2.25040    1.44308   1.559 0.141205    
InitWeight   1.08318    0.04762  22.746 1.87e-12 ***
Treatment2  -0.03581    0.40723  -0.088 0.931169    
Treatment3   1.89922    0.45802   4.147 0.000988 ***
Treatment4   1.35157    0.41937   3.223 0.006135 ** 
Treatment5   0.24446    0.57658   0.424 0.678022    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5492 on 14 degrees of freedom
Multiple R-squared:  0.9882,	Adjusted R-squared:  0.984 
F-statistic:   235 on 5 and 14 DF,  p-value: 5.493e-13
```

Example

- With R:

```
anova(nonI,complet)
```

```
Analysis of Variance Table

Model 1: FinalWeight ~ InitWeight + Treatment
Model 2: FinalWeight ~ InitWeight * Treatment
  Res.Df     RSS Df Sum of Sq      F Pr(>F)
1     14 4.2223
2     10 2.8340  4    1.3883 1.2247 0.3602
```

- With Python:

```
nonIpy = ols('FinalWeight ~ InitWeight + Treatment', data=oysterpy).fit()
from statsmodels.stats.anova import anova_lm
anova_lm(nonIpy,completpy)
```

```
   df_resid       ssr  df_diff   ss_diff         F    Pr(>F)
0      14.0  4.222323      0.0       NaN       NaN       NaN
1      10.0  2.834009      4.0  1.388314  1.224691  0.360175
```

ANCOVA with non-interaction

- If the model with non-interaction between the factor and the covariate is retained
  - Singular model:

$$(MSnonI) : Y_{ij} = \mu + \alpha_i + \beta\ z_{ij} + \varepsilon_{ij}, \forall i = 1, \cdots, I, \forall j = 1, \cdots, n_i.$$

  - Regular model:

$$(MRnonI) : Y_{ij} = a_i + b\ z_{ij} + \varepsilon_{ij}, \forall i = 1, \cdots, I, \forall j = 1, \cdots, n_i.$$

- We may test the effect of the factor or the effect of the covariate on the response.

Effect of the covariate z on Y

- Fisher's test to compare
  - the model with non-interaction

$$(MSnonI): \; Y_{ij} = \mu + \alpha_i + \beta z_{ij} + \varepsilon_{ij}, \forall i = 1, \cdots, I, \forall j = 1, \cdots, n_i$$

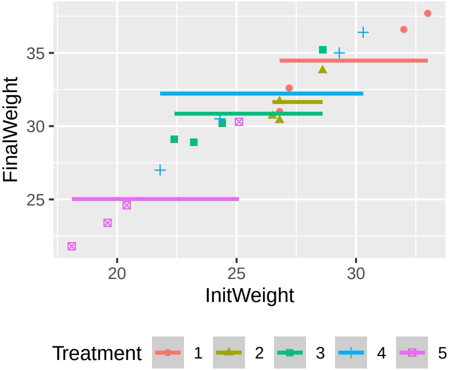
  - the one-way ANOVA

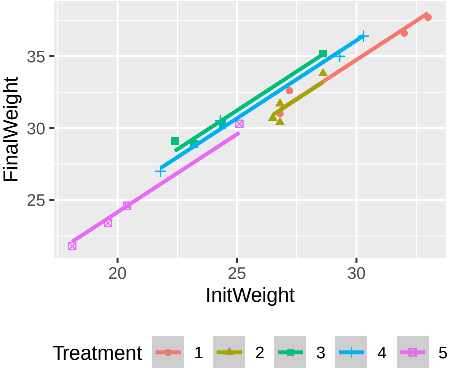
$$(MT): \; Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}, \forall i = 1, \cdots, I, \forall j = 1, \cdots, n_i.$$

- Test statistics:

$$F = \frac{SSR_T - SSR_{nonI}/1}{SSR_{nonI}/(n-(I+1))} \underset{\mathcal{H}_0}{\sim} \mathcal{F}(1, n-(I+1))$$

Effect of the covariate z on Y





Example

```
MT<-lm(FinalWeight~Treatment)
anova(MT,nonI)
```

```
Analysis of Variance Table

Model 1: FinalWeight ~ Treatment
Model 2: FinalWeight ~ InitWeight + Treatment
  Res.Df     RSS Df Sum of Sq      F    Pr(>F)
1     15 160.263
2     14   4.222  1    156.04 517.38 1.867e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
MTpy = ols('FinalWeight ~  Treatment', data=oysterpy).fit()
anova_lm(MTpy,nonIpy)
```

```
   df_resid         ssr  df_diff     ss_diff           F        Pr(>F)
0      15.0  160.262500      0.0         NaN         NaN           NaN
1      14.0    4.222323      1.0  156.040177  517.383995  1.867369e-12
```

Effect of the factor T on Y

- Fisher's test to compare
  - the model with non-interaction

$$(MSnonI): \; Y_{ij} = \mu + \alpha_i + \beta \, z_{ij} + \varepsilon_{ij}, \forall i = 1, \cdots, I, \forall j = 1, \cdots, n_i$$

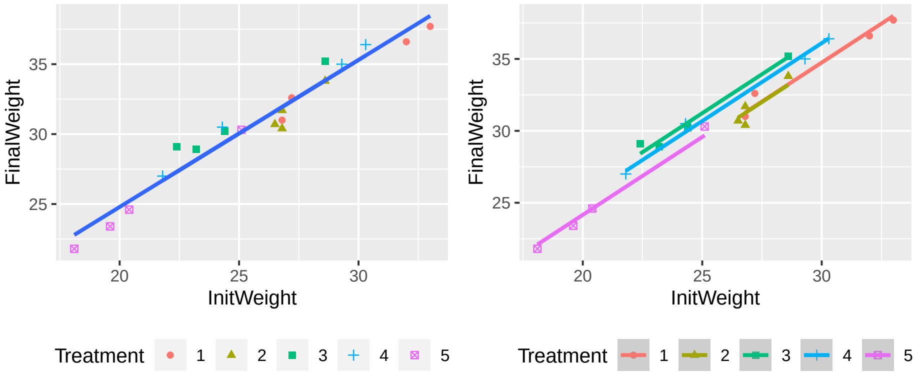
  - the linear regression

$$(Mz): \; Y_{ij} = \mu + \beta \, z_{ij} + \varepsilon_{ij}, \forall i = 1, \cdots, I, \forall j = 1, \cdots, n_i$$

- Test statistics:

$$F = \frac{SSR_z - SSR_{nonI}/(I-1)}{SSR_{nonI}/(n-(I+1))} \underset{\mathcal{H}_0}{\sim} \mathcal{F}(I-1, n-(I+1))$$

Effect of the factor T on Y



Example

- With R:

```
Mz<-lm(FinalWeight~InitWeight)
anova(Mz,nonI)
```

```
Analysis of Variance Table

Model 1: FinalWeight ~ InitWeight
Model 2: FinalWeight ~ InitWeight + Treatment
  Res.Df     RSS Df Sum of Sq      F    Pr(>F)
1     18 16.3117
2     14  4.2223  4    12.089 10.021 0.0004819 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- With Python:

```
Mzpy = ols('FinalWeight ~  InitWeight', data=oysterpy).fit()
anova_lm(Mzpy,nonIpy)
```

```
   df_resid        ssr  df_diff    ss_diff          F    Pr(>F)
0      18.0  16.311683      0.0        NaN        NaN       NaN
1      14.0   4.222323      4.0  12.089359  10.021203  0.000482
```

Summary

Summary

- Know how to write an ANCOVA model (individually and matricially), regular and singular
- Know how to distinguish a regular model from a singular model
- Know how to estimate the parameters of the ANCOVA model in the regular case and in the singular case (by adapting to the chosen constraint(s))
- Know how to construct a confidence interval for a parameter of the ANCOVA model
- Know how to construct a test to test the effect of the factor, the interaction effect, ... and know how to organize these tests
- Know how to associate a graphic representation with a sub-model of ANCOVA