Tutorial 10

Biological Data Analysis

Spring 2023

Outline

- Multiple groups
- ANOVA
- Regression using dummy variables
- Non-linear regression
- Regression with interactions between variables

F test for difference between two groups

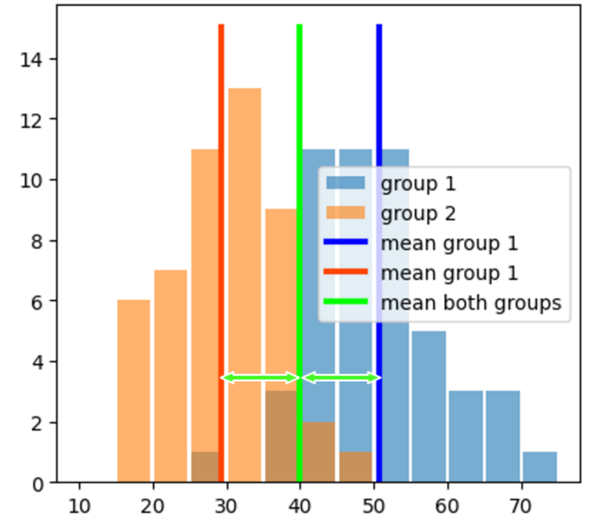
$$SS_{\bar{x}} = N1\left(\bar{x}_1 - \bar{x}\right)^2 + N2\left(\bar{x}_2 - \bar{x}\right)^2$$

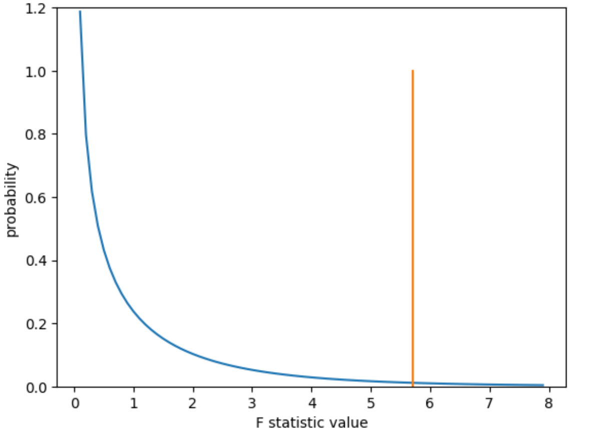
$$MS_{\bar{x}} = \frac{SS_{\bar{x}}}{K-1}$$

$$SS_{\mathrm{res}} = \sum\left(x_{1,i} - \bar{x}_1\right)^2 + \sum\left(x_{2,i} - \bar{x}_2\right)^2$$

$$MS_{res} = \frac{SS_{res}}{N1 + N2 - K}$$

$$\frac{MS_{\bar{x}}}{MS_{\mathrm{res}}} \sim F_{2-1, N1+N2-2}$$





One-way ANOVA

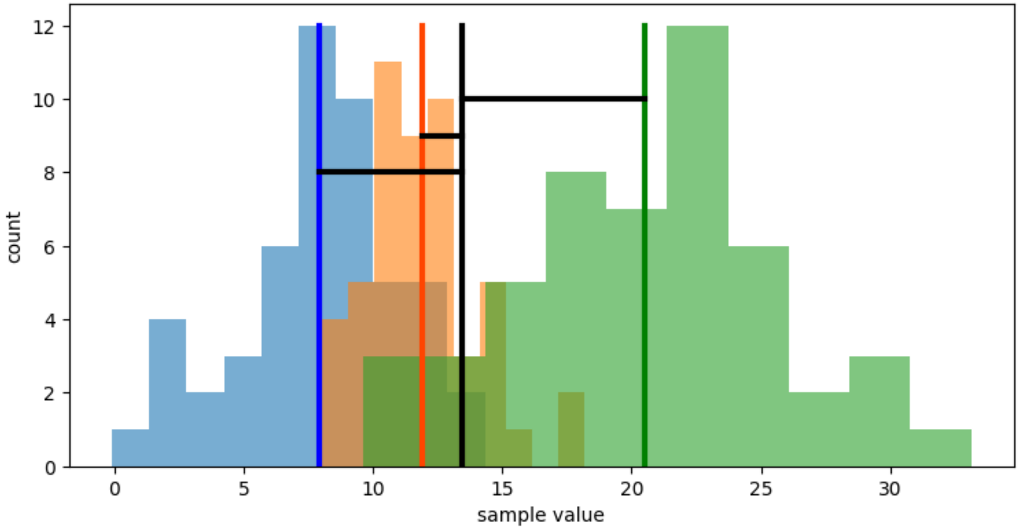
We examine the variance of the groups' mean values (distance between the group mean and the mean value of all groups) and the variance within the groups (distance between the group values and group means)

$$SS_{bet} = \sum_{i=1}^{k} \sum_{j=1}^{N_i} (\overline{x}_i - \overline{x})^2 = \sum_{i=1}^{k} N_i (\overline{x}_i - \overline{x})^2$$

$$SS_{wit} = \sum_{i=1}^{k} \sum_{j=1}^{N_i} (x_{ij} - \overline{x}_i)^2$$

$$MS_{bet} = \frac{SS_{bet}}{K-1}$$

$$MS_{wit} = \frac{SS_{wit}}{N-K}$$



ANOVA

$$SS_{bet} = \sum_{i=1}^{k}\sum_{j=1}^{N_i}(\bar{x}_i - \bar{x})^2 = \sum_{i=1}^{k} N_i(\bar{x}_i - \bar{x})^2$$

$$SS_{wit} = \sum_{i=1}^{k}\sum_{j=1}^{N_i}(x_{ij} - \bar{x}_i)^2$$

$$MS_{bet} = \frac{SS_{bet}}{K-1}$$

$$MS_{wit} = \frac{SS_{wit}}{N-K}$$

$$\frac{MS_{bet}}{MS_{wit}} \sim F_{K-1,N-K}$$

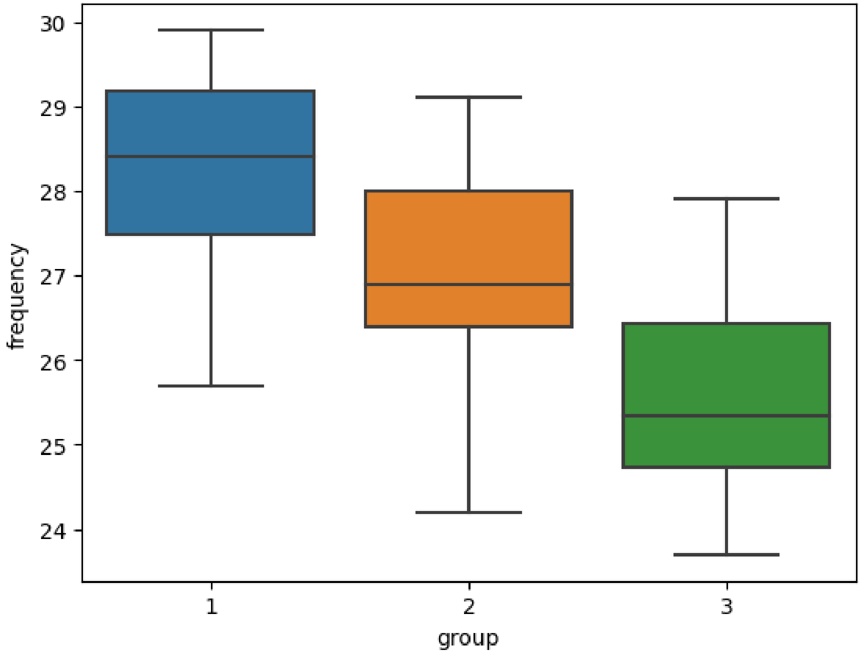
$$F_{stat} = \frac{MS_{bet}}{MS_{wit}}$$

The ratio MS_{bet}/MS_{wit} distributes $F(v_{bet},v_{wit})$ according to the Null Model

Example 1

Data about the eye color and light sensitivity (maximal detectable flickering light frequency)

group	frequency
1	28.3
1	29.4
1	28.5
1	29.9
1	27.2
1	25.7
2	29.1
2	26.9
2	28.0
2	24.2
2	26.4
3	24.5
3	25.7
3	24.8
3	26.3
3	25.0
3	23.7
3	27.9
3	26.8



Example 1 ANOVA

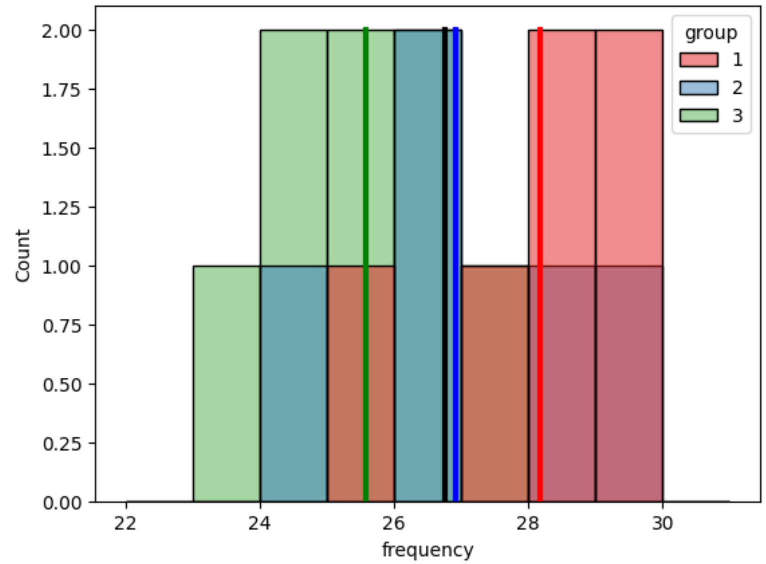
$$SS_{bet} = \sum_{i=1}^{k}\sum_{j=1}^{N_i}(\overline{x}_i - \overline{x})^2 = \sum_{i=1}^{k} N_i(\overline{x}_i - \overline{x})^2$$

$$SS_{wit} = \sum_{i=1}^{k}\sum_{j=1}^{N_i}(x_{ij} - \overline{x}_i)^2$$

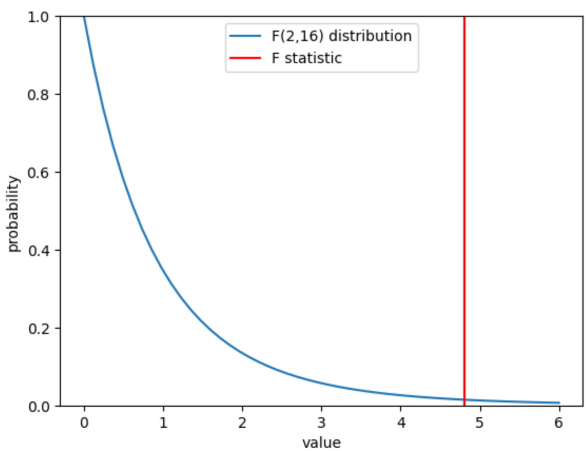
$$MS_{bet} = \frac{SS_{bet}}{K-1}$$

$$MS_{wit} = \frac{SS_{wit}}{N-K}$$

$$\frac{MS_{bet}}{MS_{wit}} \sim F_{K-1,N-K}$$



F statistic = 4.8023
p-value = 0.0232



Example 1 ANOVA

Effect size One-way ANOVA

$$\eta^2 = \frac{SS_{bet}}{SS_{tot}} = \frac{SS_{bet}}{SS_{bet} + SS_{wit}}$$

Small effect – 0.01

Median effect - 0.06

Large effect – 0.14

```
effect size η2 = 0.3751
```

Multivariate regression

$$\hat{y} = b_0 + x_1 b_1 + x_2 b_2 + r; \quad \mathrm{r \sim N(0, s}_{yx}\mathrm{)}$$

$$\mathbf{b} = \begin{bmatrix} b_0 \\ b_1 \\ b_2 \end{bmatrix} \qquad \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}; \quad \mathbf{X} = \begin{bmatrix} 1 & x_{1,1} & x_{1,2} & \cdots & x_{1,m} \\ 1 & x_{2,1} & x_{2,2} & \cdots & x_{2,m} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n,1} & x_{n,2} & \cdots & x_{n,m} \end{bmatrix}; \quad \hat{\mathbf{y}} = \begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_n \end{bmatrix} = \mathbf{Xb}$$

$$\mathbf{b} = \left(\mathbf{X}^T \mathbf{X} \right)^{-1} \mathbf{X}^T \mathbf{y}$$

$$s_{y|x}^2 = MS_{\mathrm{res}} = \frac{SS_{\mathrm{res}}}{\nu_{\mathrm{res}}} = \frac{\left(\mathbf{y} - \mathbf{Xb} \right)^T \left(\mathbf{y} - \mathbf{Xb} \right)}{n - (m+1)}$$

Multivariate regression dummy variables

$$\hat{y} = b_1 + b_2 D_2 + b_3 D_3$$

$$D_2 = \begin{cases} 1 \text{ if group2} \\ 0 \text{ otherswise} \end{cases}$$

$$D_3 = \begin{cases} 1 \text{ if group 3} \\ 0 \text{ otherswise} \end{cases}$$

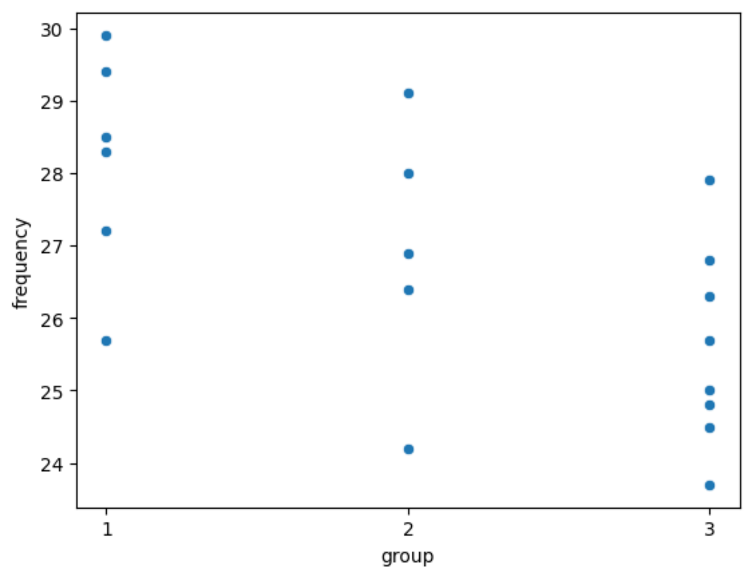
$$\hat{y}_i = \begin{cases} b_1 \text{ for group1} \\ b_1 + b_2 \text{ for group2} \\ b_1 + b_3 \text{ for group3} \end{cases}$$

$$\mathbf{b} = \left(\mathbf{X}^T \mathbf{X}\right)^{-1} \mathbf{X}^T \mathbf{y}$$

Example 1

group	frequency
1	28.3
1	29.4
1	28.5
1	29.9
1	27.2
1	25.7
2	29.1
2	26.9
2	28.0
2	24.2
2	26.4
3	24.5
3	25.7
3	24.8
3	26.3
3	25.0
3	23.7
3	27.9
3	26.8

Relationship between the eye color and light sensitivity



Example 1 – Dummy variables

$$\hat{y} = b_1 + b_2 D_2 + b_3 D_3$$

y

```
28.3
29.4
28.5
29.9
27.2
25.7
29.1
26.9
28.0
24.2
26.4
24.5
25.7
24.8
26.3
25.0
23.7
27.9
26.8
```

X

```
[[1. 0. 0.]
 [1. 0. 0.]
 [1. 0. 0.]
 [1. 0. 0.]
 [1. 0. 0.]
 [1. 0. 0.]
 [1. 1. 0.]
 [1. 1. 0.]
 [1. 1. 0.]
 [1. 1. 0.]
 [1. 1. 0.]
 [1. 0. 1.]
 [1. 0. 1.]
 [1. 0. 1.]
 [1. 0. 1.]
 [1. 0. 1.]
 [1. 0. 1.]
 [1. 0. 1.]
 [1. 0. 1.]]
```

Example 1 – Dummy variables

$$\hat{y} = b_1 + b_2 D_2 + b_3 D_3 + r \; ; \quad \mathrm{r{\sim}N(0,s}_{yx}\mathrm{)}$$

$$\mathbf{b} = \left(\mathbf{X}^T \mathbf{X}\right)^{-1} \mathbf{X}^T \mathbf{y}$$

$$s_{y|x}^2 = MS_{\mathrm{res}} = \frac{SS_{\mathrm{res}}}{\nu_{\mathrm{res}}} = \frac{\left(\mathbf{y} - \mathbf{Xb}\right)^T \left(\mathbf{y} - \mathbf{Xb}\right)}{n - 3}$$

```
b0 = 28.1667 , b1 = -1.2467 , b2 = -2.5792

Syx2 = 2.3944
```

Example 1 – Dummy variables

$$\hat{y} = b_1 + b_2 D_2 + b_3 D_3 + r \; ; \quad \mathrm{r{\sim}N(0,s}_{yx})$$

b_1 - mean for group1

$b_1 + b_2$ - mean for group2

$b_1 + b_3$ - mean for group3

Means:	group 1	group 2	group 3
	28.1667	26.92	25.5875

Example 1 – Dummy variables

Model significance and effect size

$$\frac{MS_{reg}}{MS_{res}} \to F_{K-1,N-K}$$

```
model significance:
Fstat = 4.8023
p-value = 0.0232
```

$$R^2 = \frac{SS_{reg}}{SS_{tot}} \qquad f^2 = \frac{SS_{reg}}{SS_{res}}$$

```
model effect size:
R2 = 0.3751
Cohens f = 0.6003
```

Example 1 – Dummy variables

Coefficient significance

$$b_i - t_{N-K}{}^{\alpha/2} s_{b_i} < \beta_i < b_i + t_{N-K}{}^{\alpha/2} s_{b_i}$$

```
b0 95% CI = [26.8275 , 29.5058]
b1 95% CI = [-3.233 , 0.7397]
b2 95% CI = [-4.3507 , -0.8076]
```

$$\frac{b_i}{s_{b_i}} \rightarrow t_{N-K}$$

```
b0 p-value = 0.0
b1 p-value = 0.101
b2 p-value = 0.0035
```

Example 1 – Dummy variables

Group comparison using t-test

$$\frac{b_1 - b_2}{s_{b_1 - b_2}} \rightarrow t_{vres}$$

$$\mathrm{var}(b_1 - b_2) = \mathrm{var}\left(b_1\right) + \mathrm{var}\left(b_2\right) - 2\,\mathrm{cov}\left(b_1, b_2\right)$$

$$sd\,(b_1 - b_2) = \sqrt{\mathrm{var}(b_1 - b_2)}$$

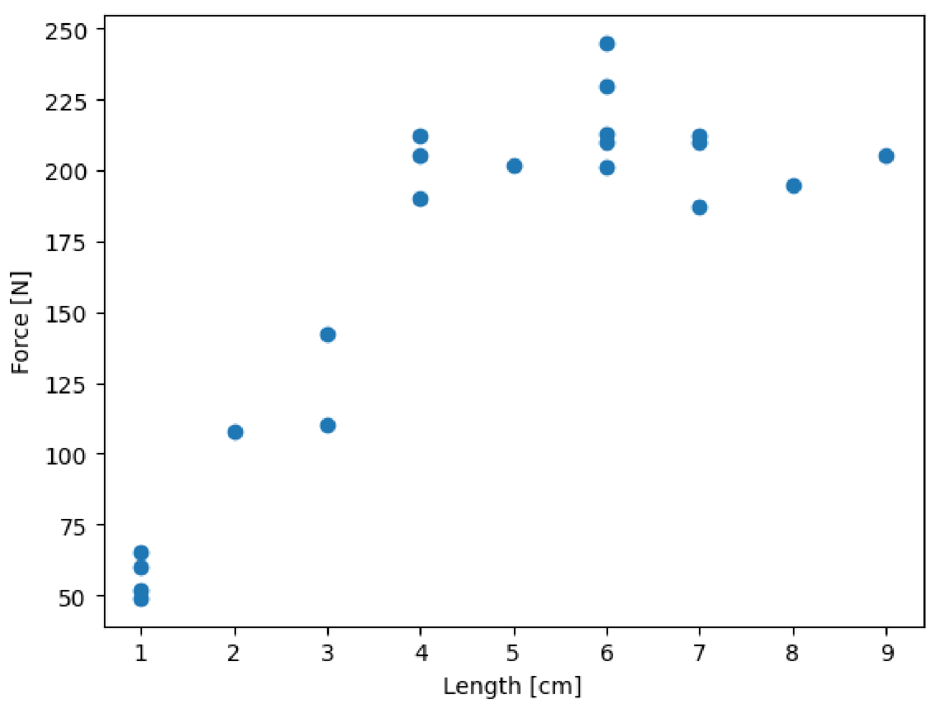
$$t_{b_1 - b_2} = \frac{b_1 - b_2}{sd\,(b_1 - b_2)} \sim T_{vres}$$

```
groups comparison:
groups 1-2 p-value = 0.101
groups 1-3 p-value = 0.0035
groups 2-3 p-value = 0.0752
```

Example 2
Non-linear regression

A group of researches checked the connection between the muscle length and the generated force.

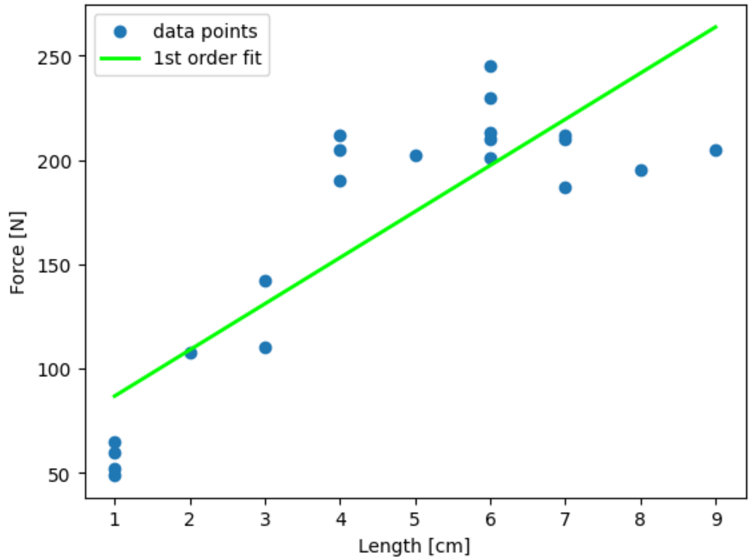
Force	Length
49	1
52	1
60	1
65	1
108	2
110	3
142	3
190	4
205	4
212	4
202	5
201	6
210	6
213	6
230	6
245	6
187	7
210	7
212	7
195	8
205	9



Try linear model

$$y_n = \hat{y}(x_n) + r_n \qquad \hat{y}(x) = b_0 + b_1 x \qquad r_n \sim N\left(0, s^2_{y|x_n}\right)$$

```
b0 = 64.72 , b1 = 22.1019 , Syx2 = 1288.5535
```



Model significance and effect size

```
model p-value = 0.0
R2= 0.7214
```

Try 2nd order model

$$\hat{y}(x) = b_0 + b_1 x + b_2 x^2$$

```
 49      [[ 1.  1.  1.]
 52       [ 1.  1.  1.]
 60       [ 1.  1.  1.]
 65       [ 1.  1.  1.]
108       [ 1.  2.  4.]
110       [ 1.  3.  9.]
142       [ 1.  3.  9.]
190       [ 1.  4. 16.]
205       [ 1.  4. 16.]
212       [ 1.  4. 16.]
202       [ 1.  5. 25.]
201       [ 1.  6. 36.]
210       [ 1.  6. 36.]
213       [ 1.  6. 36.]
230       [ 1.  6. 36.]
245       [ 1.  6. 36.]
187       [ 1.  7. 49.]
210       [ 1.  7. 49.]
212       [ 1.  7. 49.]
195       [ 1.  8. 64.]
205       [ 1.  9. 81.]]
```

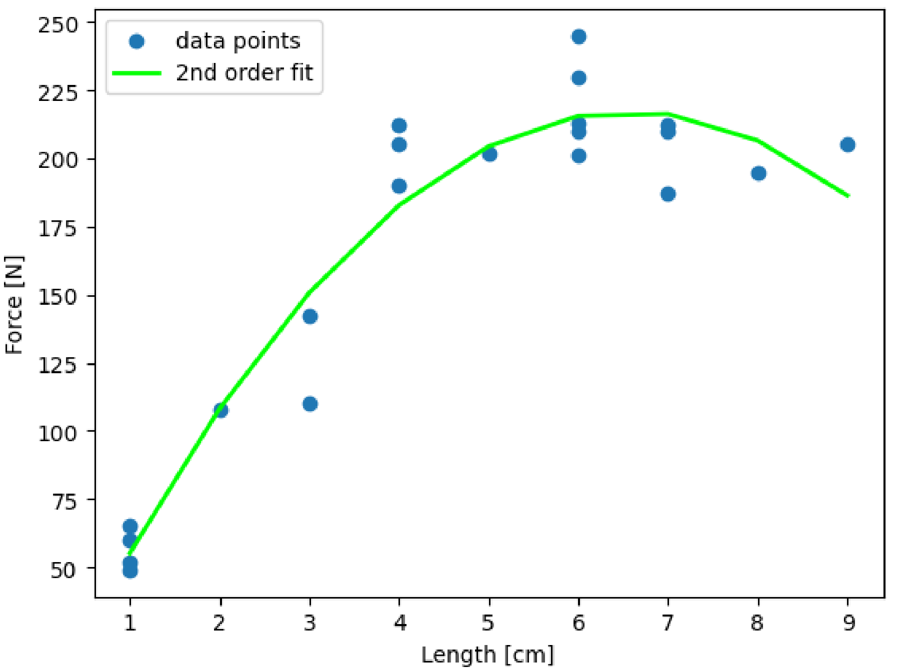
y **X**

Try 2nd order model

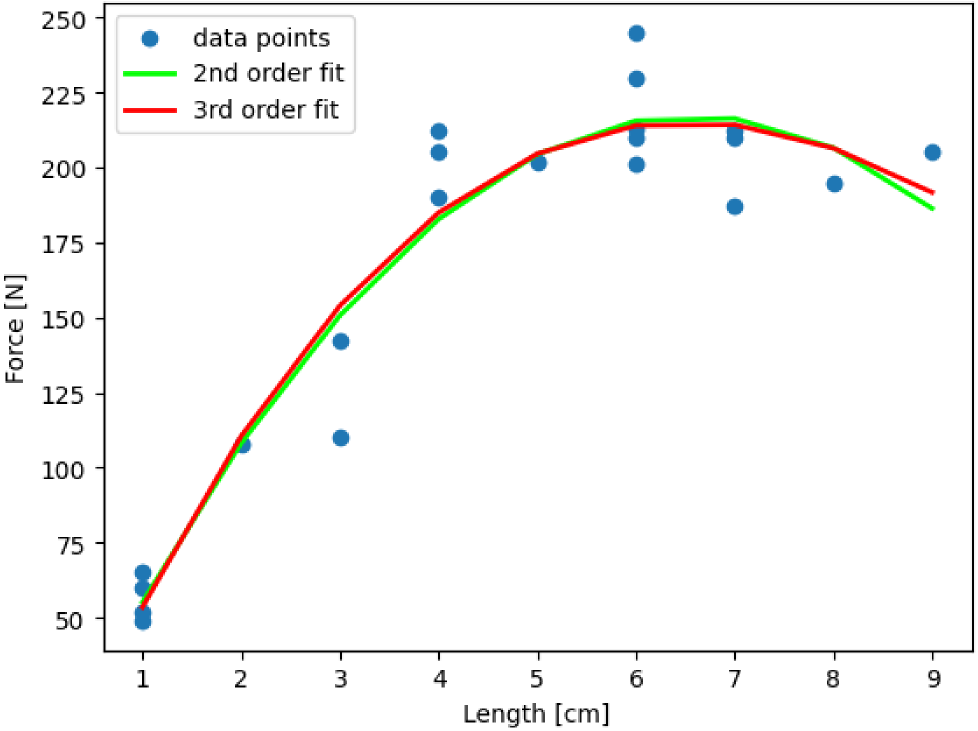
$$\hat{y}(x) = b_0 + b_1 x + b_2 x^2 \qquad r_n \sim N\left(0, s^2_{y|x_n}\right)$$

```
b0 = -8.3022 , b1 = 68.6895 , b2 = -5.2286
Syx2 = 334.8128
```

```
model p-value = 0.0
R2= 0.9276
```



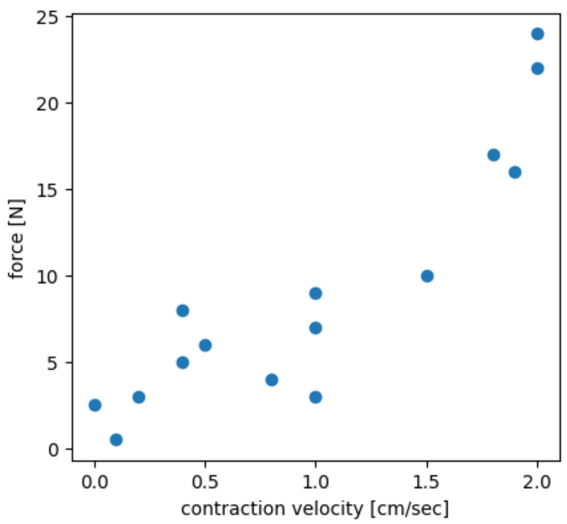
Try 3rd order model

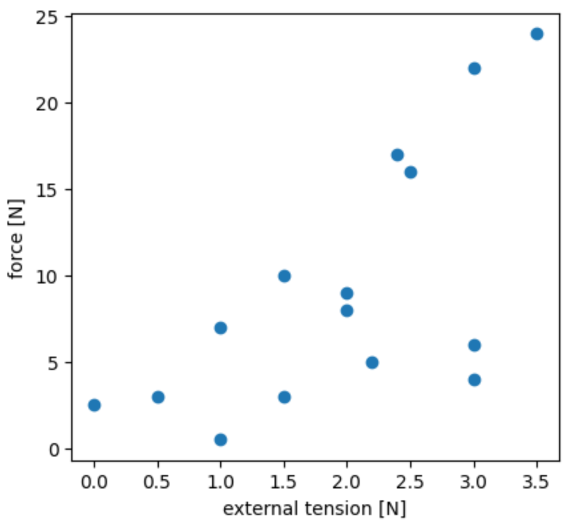


Example 3
Regression with Interaction

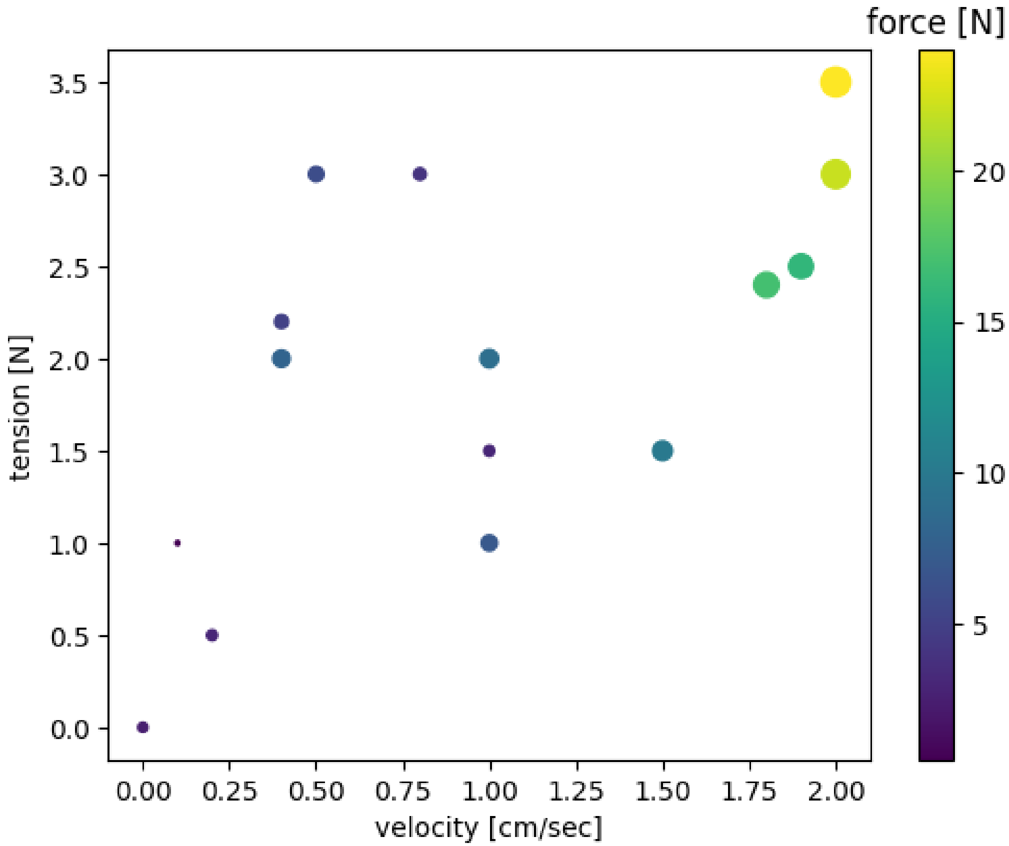
The same group continued the research and checked the connection between the force generated by the muscle, the external force applied to the muscle and its contraction speed

force	velocity	tension
2.5	0.0	0.0
0.5	0.1	1.0
3.0	0.2	0.5
5.0	0.4	2.2
6.0	0.5	3.0
8.0	0.4	2.0
4.0	0.8	3.0
3.0	1.0	1.5
7.0	1.0	1.0
9.0	1.0	2.0
10.0	1.5	1.5
17.0	1.8	2.4
16.0	1.9	2.5
22.0	2.0	3.0
24.0	2.0	3.5





Example 3



force [N]
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
tension [N]
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
2.00
velocity [cm/sec]
20
15
10
5

Example 3
Try multivariate regression

$$\hat{y}(x) = b_0 + b_1 x + b_2 x_2 \qquad r_n \sim N\left(0, s_{y|x_n}^2\right)$$

```
b0 = -1.1362 , b1 = 8.1191 , b2 = 1.2201 , Syx2 = 11.0151
```

Model significance and effect size

```
p_value = 3.015e-05 , R2 = 0.8236
```

Example 3
Try model with interactions

$$\hat{y}(x) = b_0 + b_1 x + b_2 x_2 + b_3 x_1 x_2 \qquad r_n \sim N\left(0, s^2_{y|x_n}\right)$$

b0 = 2.6838 , b1 = 0.6621 , b2 = -0.8938 , b3 = 3.2684
Syx2 = 5.7903

Model significance and effect size

p_value = 3.52e-06 , R2 = 0.915

Example 3
do we need interactions?

Addition test

Full model

$$\hat{y}(x) = b_0 + b_1 x + b_2 x_2 + b_3 x_1 x_2$$

Partial model

$$\hat{y}(x) = b_0 + b_1 x + b_2 x_2$$

`p-value of interaction contribution = 0.0055`