

EFSA-Project

Statistical Learning Theory

UNIPV

A.Y. 2020/2021

Professor: Giuseppe De Nicolao

Team members:

- Andrea Vergine
- Domenico Ragusa

Goals



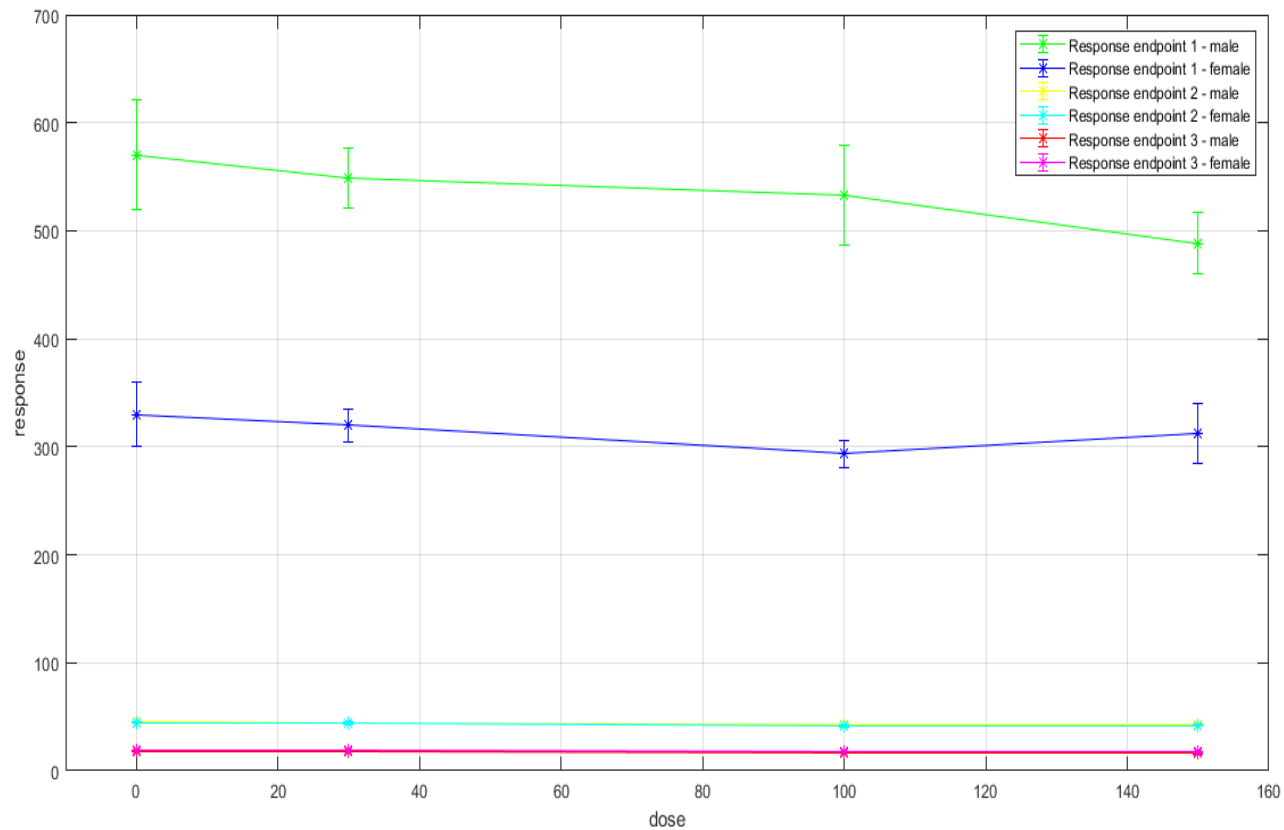
1. Plot dose-response data for each pair endpoint-gender (6 plots) with error bars reflecting error size on response measurements.
2. Use subset selection to estimate separate models for the 3 endpoints using gender as categorical variable.
3. Use subset selection to estimate a unique model using gender and endpoint as categorical variables.
4. In points 2. and 3., heteroscedasticity of the data should be taken into account.

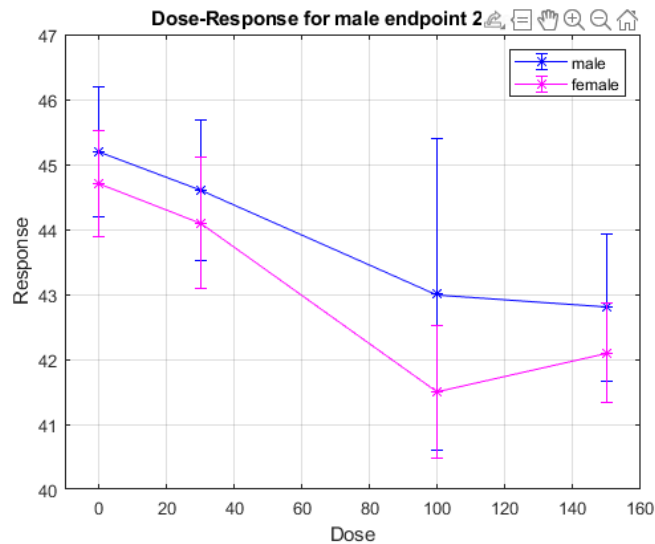
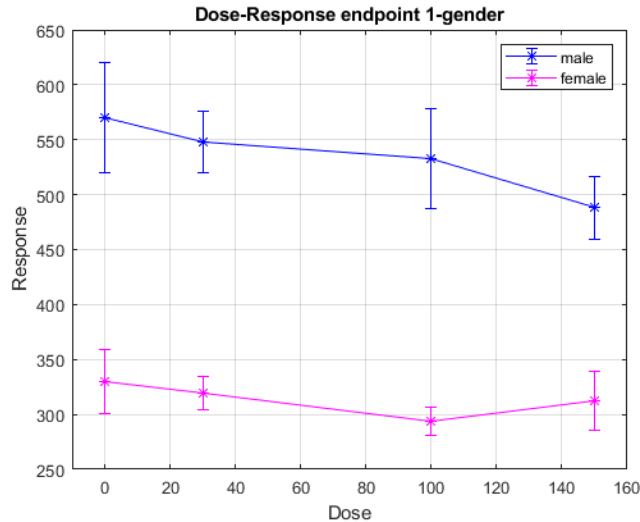
Data table



response	number of animals	SD	SE	variance	dose	sex (0=M, 1=F)	endpoint
570.4	9	75.6	25.200	635.040	0	0	1
548.5	10	44.3	14.009	196.249	30	0	1
533	10	72.3	22.863	522.729	100	0	1
488.1	9	42.5	14.167	200.694	150	0	1
329.8	10	46.8	14.799	219.024	0	1	1
319.7	10	24.3	7.684	59.049	30	1	1
293.3	10	20.1	6.356	40.401	100	1	1
312.4	10	43.1	13.629	185.761	150	1	1
45.2	9	1.5	0.500	0.250	0	0	2
44.6	10	1.7	0.538	0.289	30	0	2
43	10	3.8	1.202	1.444	100	0	2
42.8	9	1.7	0.567	0.321	150	0	2
44.7	10	1.3	0.411	0.169	0	1	2
44.1	10	1.6	0.506	0.256	30	1	2
41.5	10	1.6	0.506	0.256	100	1	2
42.1	10	1.2	0.379	0.144	150	1	2
18.3	9	0.5	0.167	0.028	0	0	3
17.9	10	0.5	0.158	0.025	30	0	3
16.5	10	0.6	0.190	0.036	100	0	3
16.3	9	0.8	0.267	0.071	150	0	3
19.1	10	0.5	0.158	0.025	0	1	3
19	10	0.4	0.126	0.016	30	1	3
17.9	10	0.5	0.158	0.025	100	1	3
17.3	10	0.6	0.190	0.036	150	1	3

Data plot

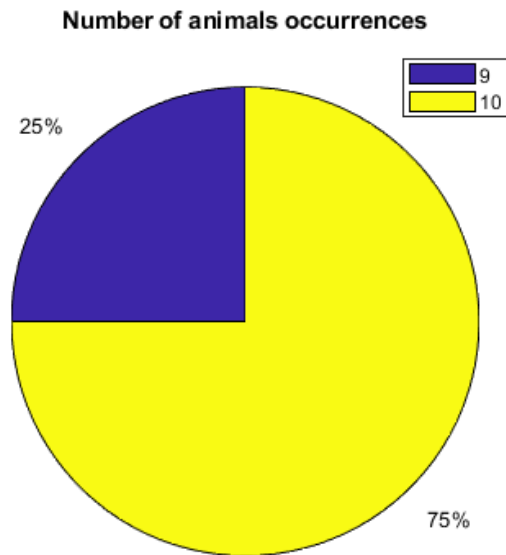
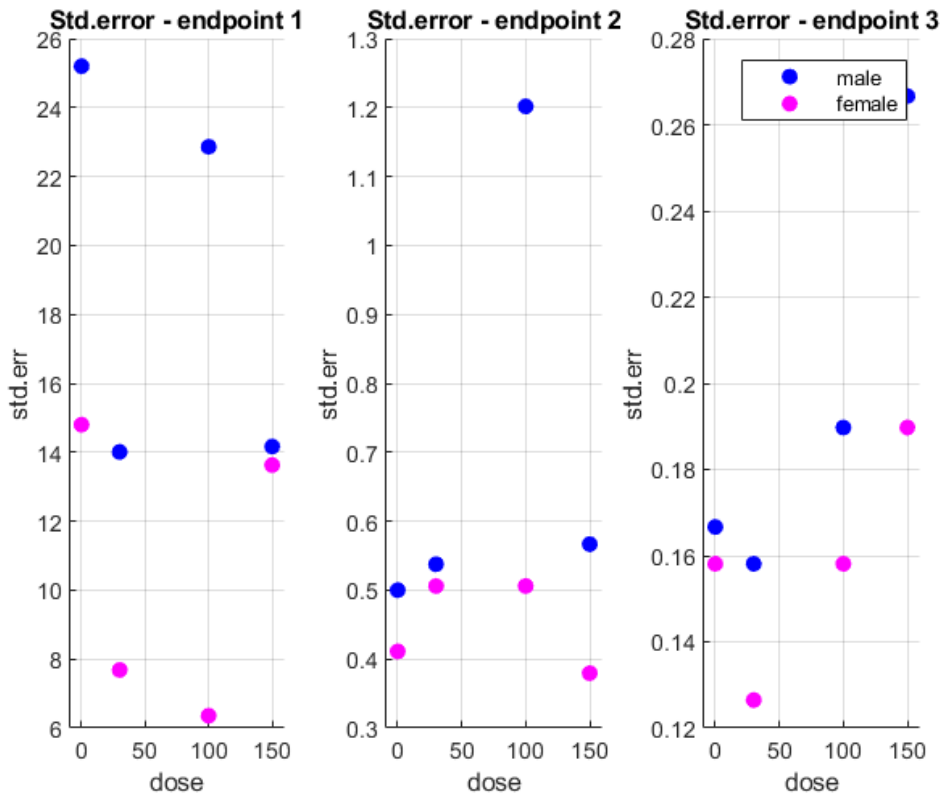




Dose - response data plot




Looking at the data in advance



How to work

Alternatives:

- Linear models -> 
- Fourier series -> No period, few data
- Polynomial models -> few data (not beyond 2° degree)

Approaches:

- Forward stepwise selection
- Backward stepwise selection
- Forward-Backward stepwise selection (stepwiselm)

Take into account heteroscedasticity:

$$w_{ii} = \frac{1}{std.error^2}, \quad std.error^2 = \frac{\sigma_i^2}{n}$$



Forward stepwise selection

1. M_0 : null model (no predictors).
 2. For $k = 0, \dots, p - 1$:
 - 2.1 Consider all $p - k$ models with one additional predictor.
 - 2.2 Choose the best, call it M_{k+1} .
Best: smallest RSS or highest R^2 .
 3. Select the best model using objective tests.
-

Backward Stepwise Selection

1. M_p : full model (all p predictors).
 2. For $k = p, p-1, \dots, 1$:
 - 2.1 Consider all k .
 - 2.2 Choose the best among these k models and call it M_{k1} . (smallest RSS or highest R^2).
 3. Select the best model using objective tests.
-

Function: stepwiselm

stepwiselm

Perform stepwise regression

Syntax

```
mdl = stepwiselm(tbl)
mdl = stepwiselm(X,y)
mdl = stepwiselm(__,modelspec)
mdl = stepwiselm(__,Name,Value)
```

Description

`mdl = stepwiselm(tbl)` creates a linear model for the variables in the table or dataset array `tbl` using stepwise regression to add or remove predictors, starting from a constant model. `stepwiselm` uses the last variable of `tbl` as the response variable. `stepwiselm` uses forward and backward stepwise regression to determine a final model. At each step, the function searches for terms to add the model to or remove from the model, based on the value of the 'Criterion' argument.

`mdl = stepwiselm(X,y)` creates a linear model of the responses `y` to the predictor variables in the data matrix `X`.

`mdl = stepwiselm(__,modelspec)` specifies the starting model `modelspec` using any of the input argument combinations in previous syntaxes.

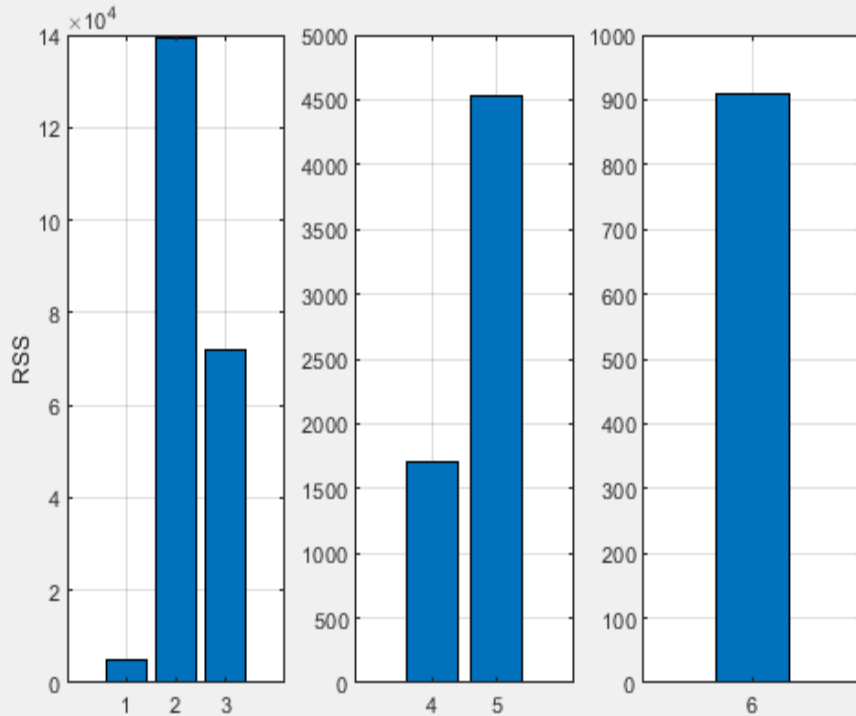
`mdl = stepwiselm(__,Name,Value)` specifies additional options using one or more name-value pair arguments. For example, you can specify the categorical variables, the smallest or largest set of terms to use in the model, the maximum number of steps to take, or the criterion that `stepwiselm` uses to add or remove terms.

<http://it.mathworks.com/help/stats/stepwiselm.html>

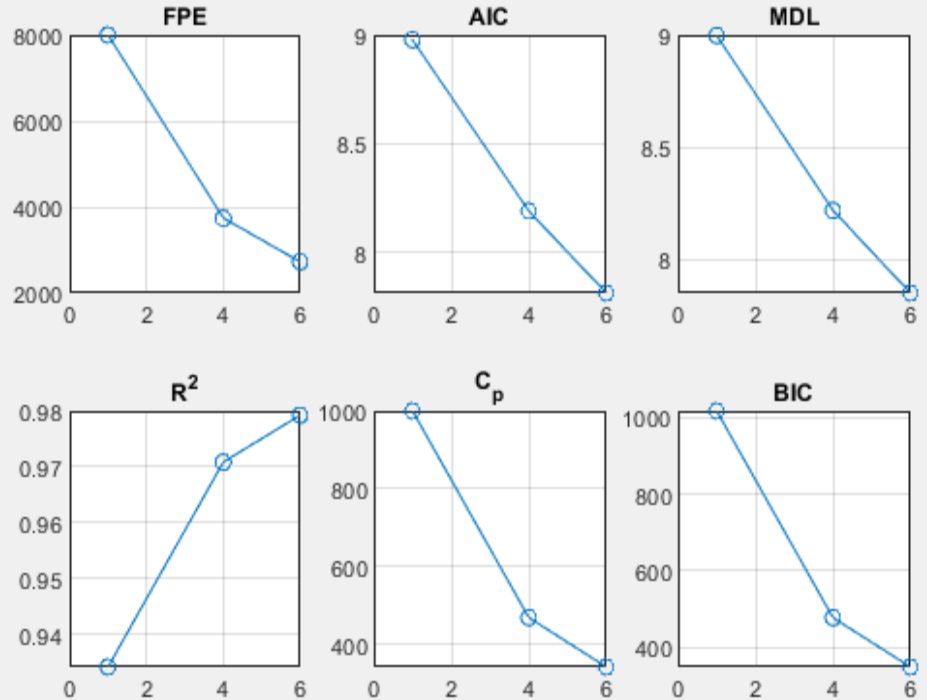
Forward (Backward) Selection – endpoint 1

$$response = 567.3801 - 243.6787 * sex - 0.5059 * dose + 0.2808 * sex * dose$$

RSS for each iteration

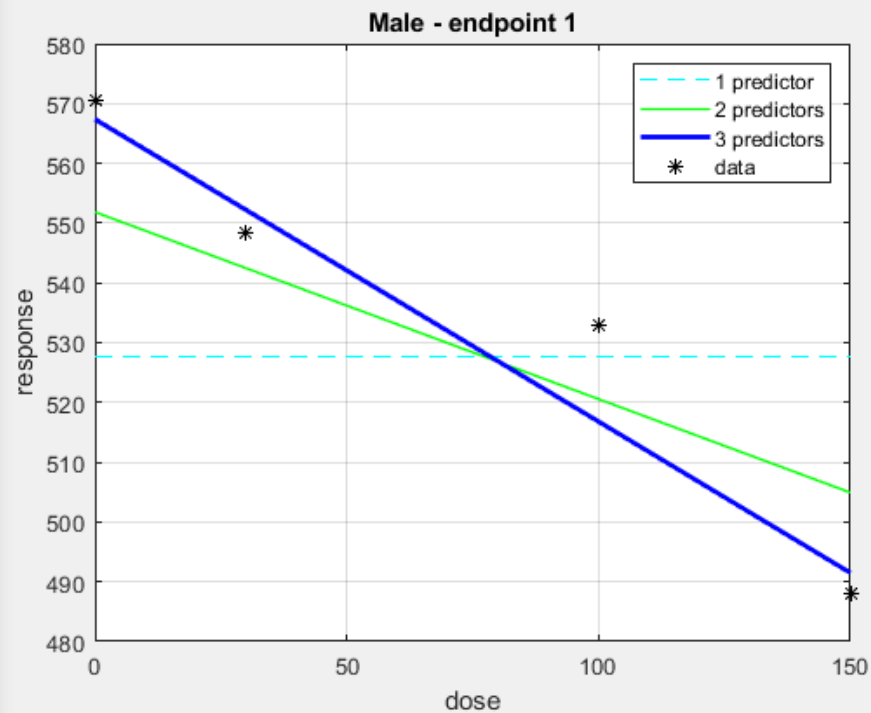
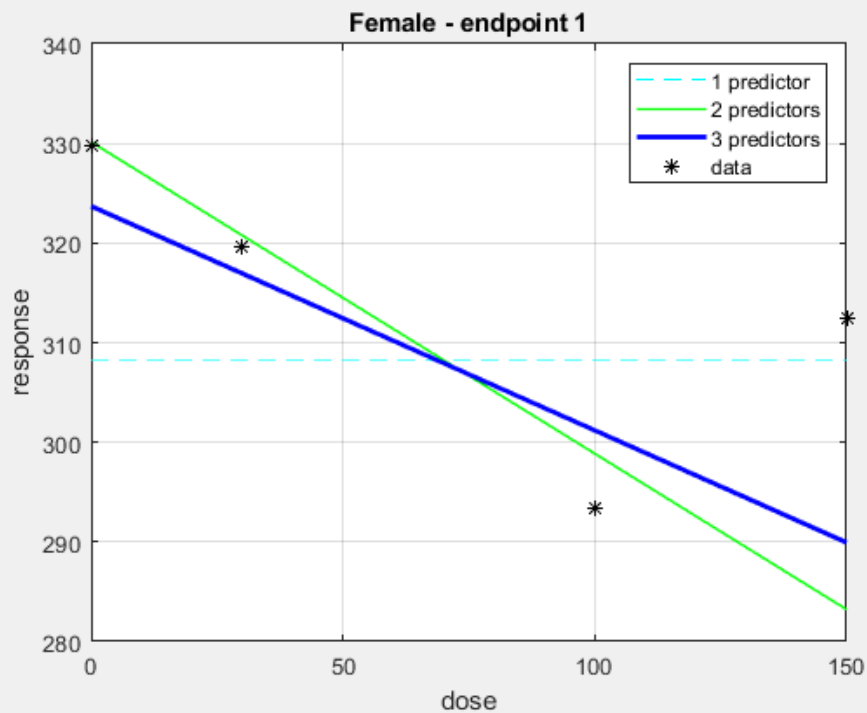


Objective tests evolution



Forward (Backward) Selection – endpoint 1

theta	std.error
567.3801	16.4684625
-243.679	19.0457353
-0.50592	0.16458333
0.280831	0.19849252



StepwiseLM (forward-backward) – endpoint 1

```
mdl_1 =  
Linear regression model:  
  response_1 ~ 1 + dose_1 + gender_female_1  
  
Estimated Coefficients:  


|                   | Estimate | SE     | tStat   | pValue     |
|-------------------|----------|--------|---------|------------|
| (Intercept)       | 551.84   | 13.447 | 41.039  | 1.6201e-07 |
| dose_1            | -0.31284 | 0.1008 | -3.1037 | 0.026742   |
| gender_female_1_1 | -221.7   | 12.073 | -18.363 | 8.8074e-06 |

  
Number of observations: 8, Error degrees of freedom: 5  
Root Mean Squared Error: 1.25  
R-squared: 0.986, Adjusted R-Squared: 0.98  
F-statistic vs. constant model: 171, p-value = 2.5e-05
```

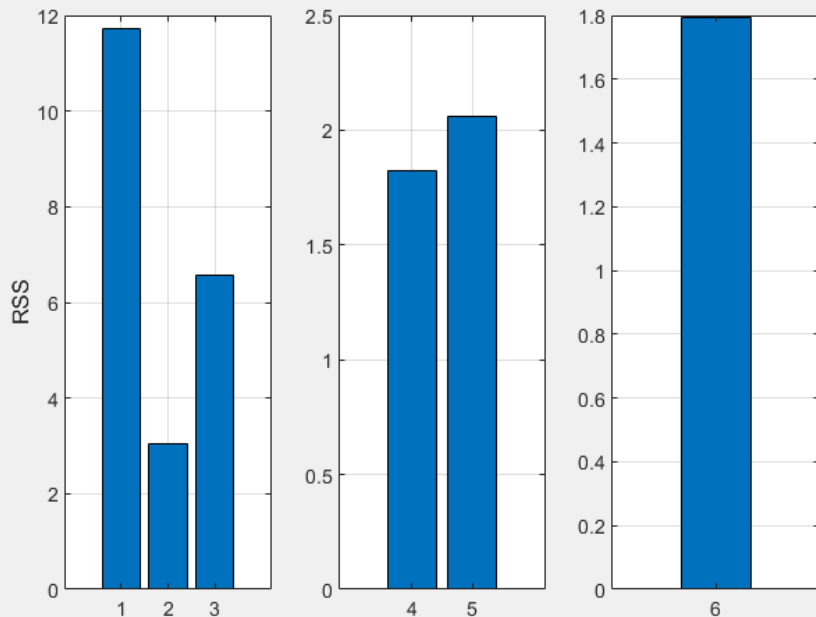
Confidence intervals

517.2787	586.4103
-0.5720	-0.0537
-252.7355	-190.6660

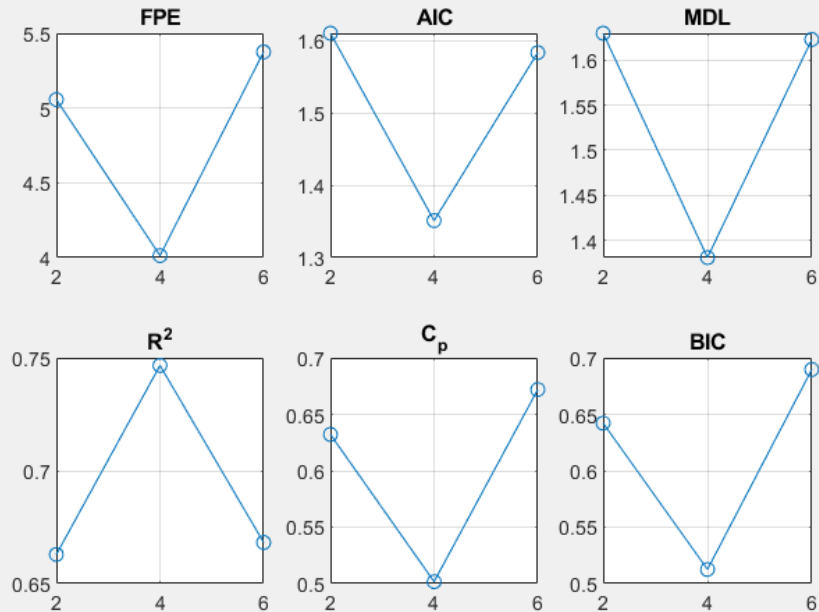
Forward (Backward) Selection – endpoint 2

$$\text{response} = 45.2371 - 0.0180 * \text{dose} - 0.7831 * \text{sex}$$

RSS for each iteration



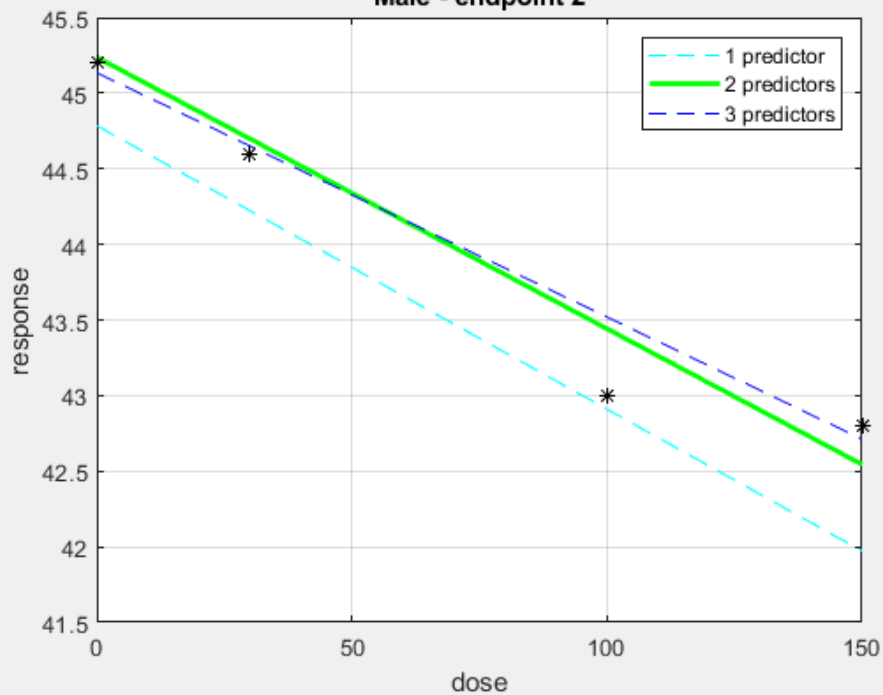
Objective tests evolution



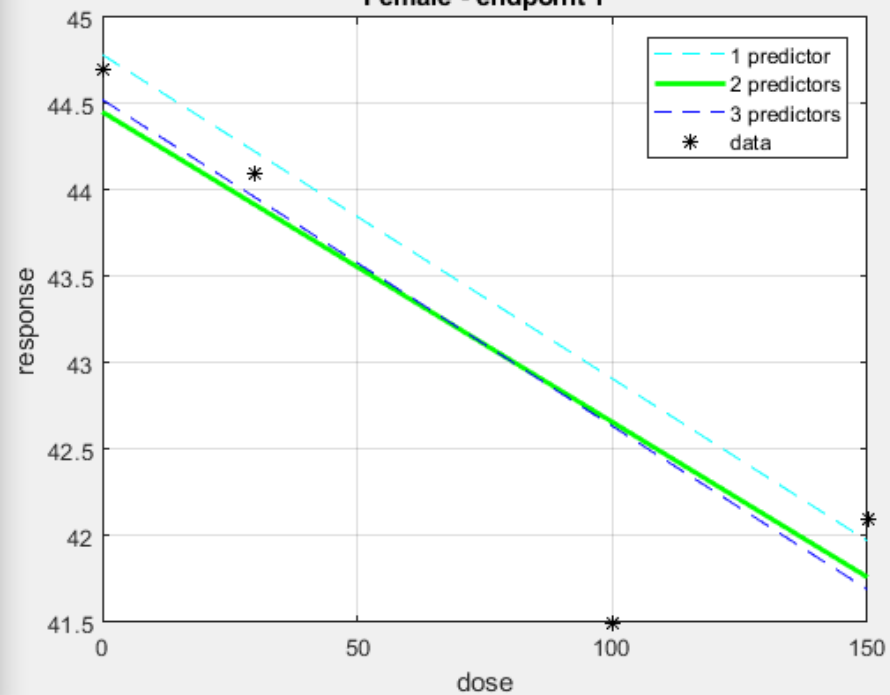
Forward (Backward) Selection – endpoint 2

theta	std.error
45.23707	0.39747548
-0.01795	0.00330656
-0.78308	0.43948087

Male - endpoint 2



Female - endpoint 1



StepwiseLM (forward-backward) – endpoint 2

```
mdl_2 =  
Linear regression model:  
  response_2 ~ 1 + dose_2  
  
Estimated Coefficients:  
              Estimate      SE      tStat      pValue  
-----  
(Intercept)    44.785    0.35731    125.34    1.7391e-11  
dose_2         -0.018754  0.0038235    -4.9048    0.0026978  
  
Number of observations: 8, Error degrees of freedom: 6  
Root Mean Squared Error: 1.37  
R-squared: 0.8, Adjusted R-Squared: 0.767  
F-statistic vs. constant model: 24.1, p-value = 0.0027
```

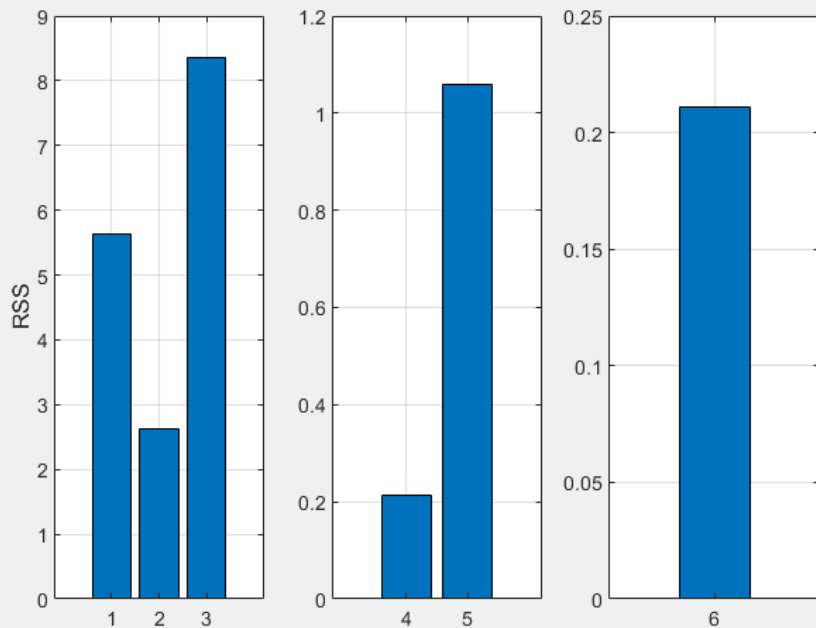
Confidence intervals

43.991	45.6596
-0.0281	-0.0094

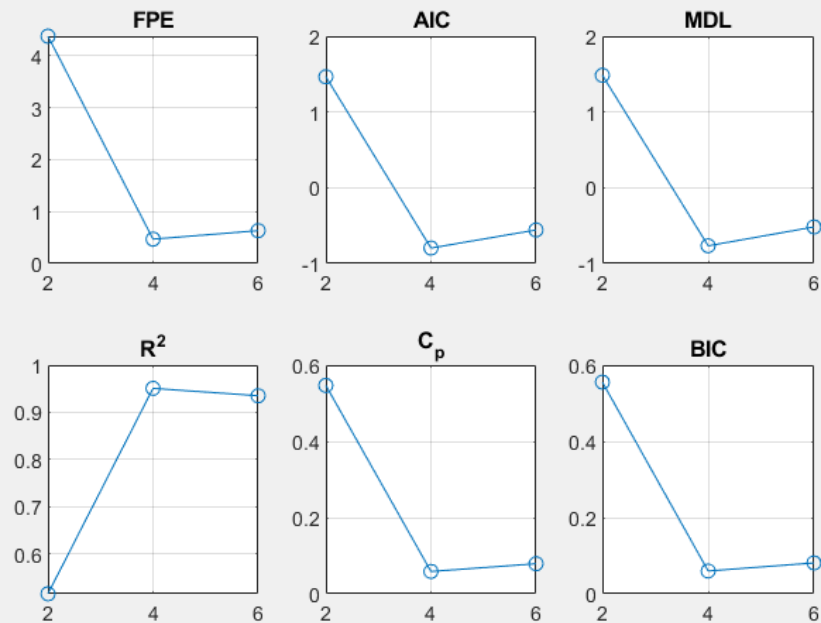
Forward (Backward) Selection – endpoint 3

$$response = 18.2186 - 0.0139 * sex + 1.0881 * dose$$

RSS for each iteration

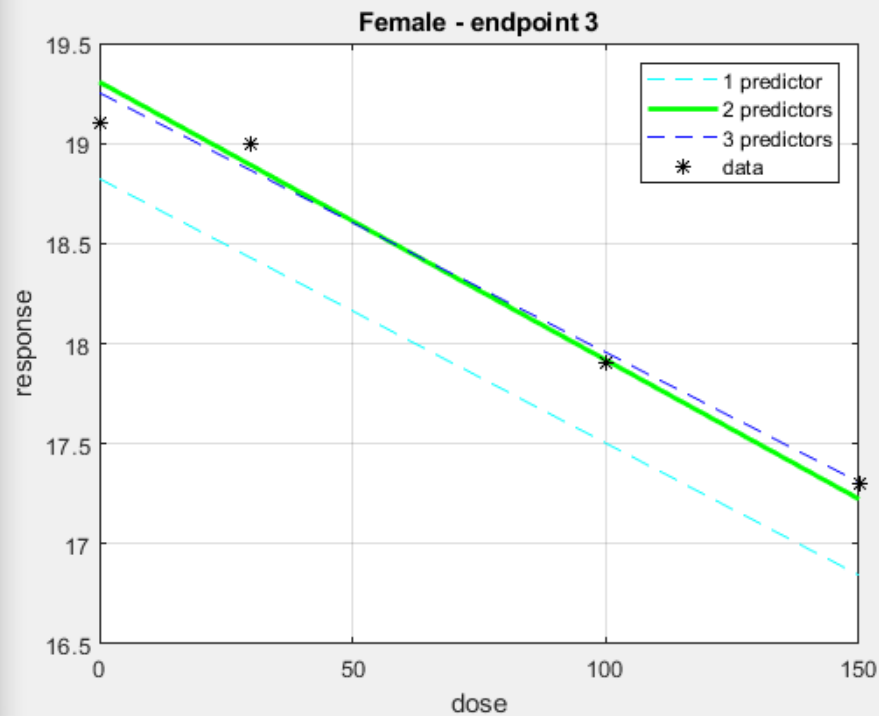
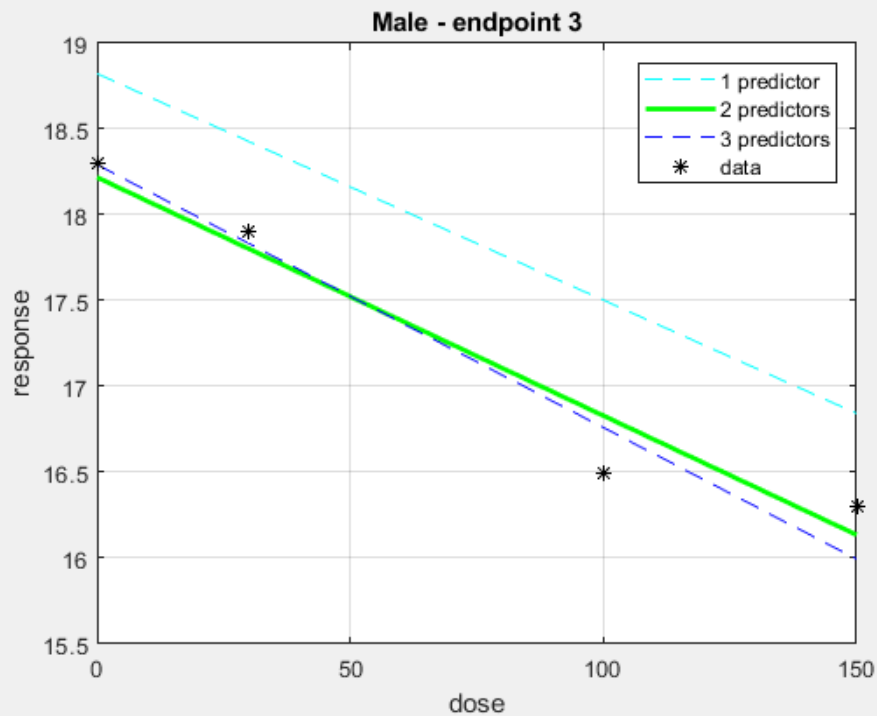


Objective tests evolution



Forward (Backward) Selection – endpoint 3

theta	std.error
18.21858	0.1254263
-0.01389	0.00128698
1.088115	0.13866997



StepwiseLM (forward-backward) – endpoint 3

```
mdl_3 =  
Linear regression model:  
  response_3 ~ 1 + dose_3 + gender_female_3  
  
Estimated Coefficients:  
  
              Estimate      SE      tStat      pValue  
-----  
(Intercept)      18.219      0.12543     145.25     2.934e-10  
dose_3            -0.013887    0.001287     -10.79     0.00011858  
gender_female_3_1    1.0881      0.13867      7.8468     0.00053962  
  
Number of observations: 8, Error degrees of freedom: 5  
Root Mean Squared Error: 1.15  
R-squared: 0.971, Adjusted R-Squared: 0.959  
F-statistic vs. constant model: 83.6, p-value = 0.000144
```

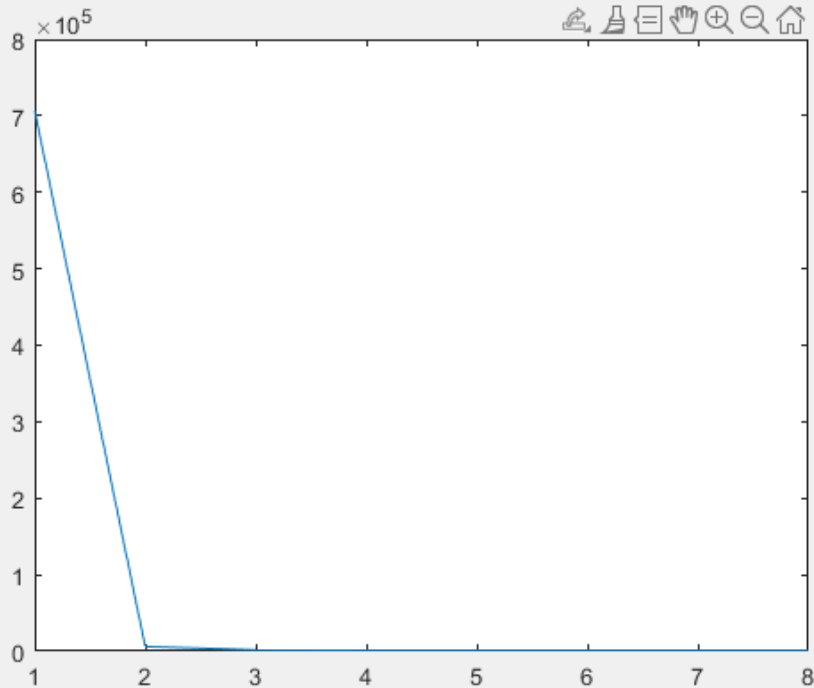
Confidence intervals

17.8962	18.5410
-0.0172	-0.0106
0.7317	1.4446

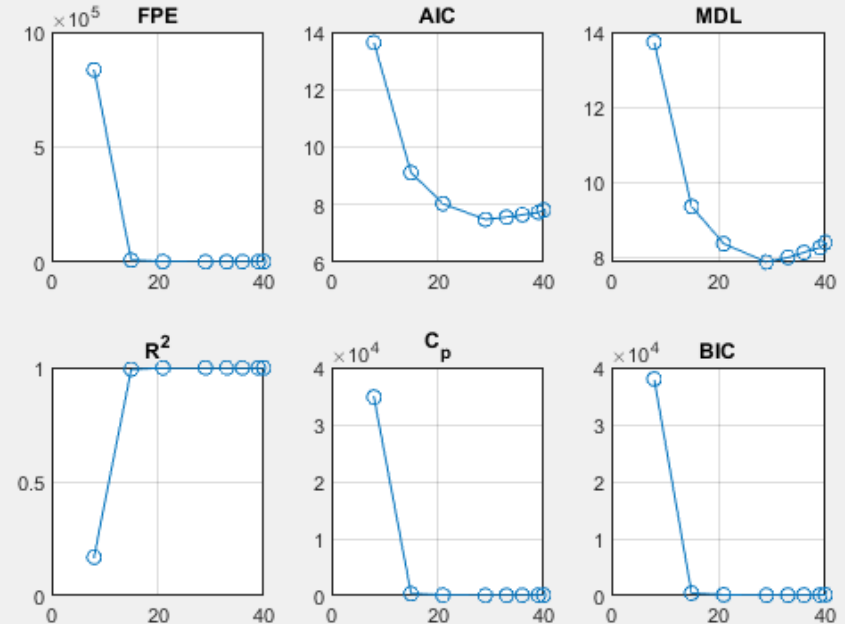
Forward (backward) selection – unique model

response

$$\begin{aligned} = & 18.36327 + 0.909607 * sex - 0.01459 * dose + 549.0169 * endpoint1 + 25.54763 \\ & * endpoint2 - 0.49133 * dose * endpoint1 - 244.588 * sex * endpoint1 + 0.280831 \\ & * sex * endpoint1 * dose \end{aligned}$$

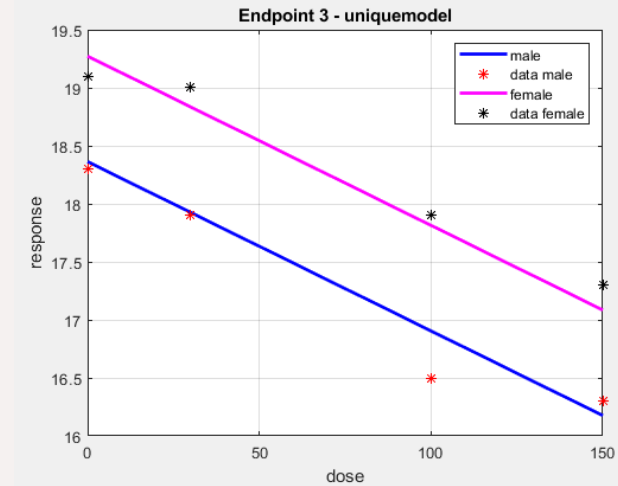
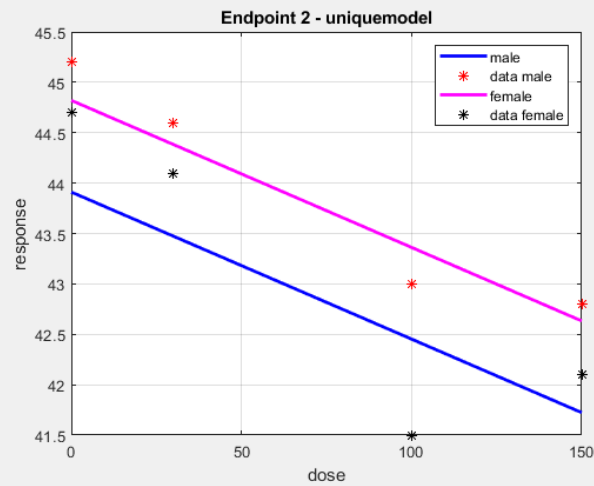
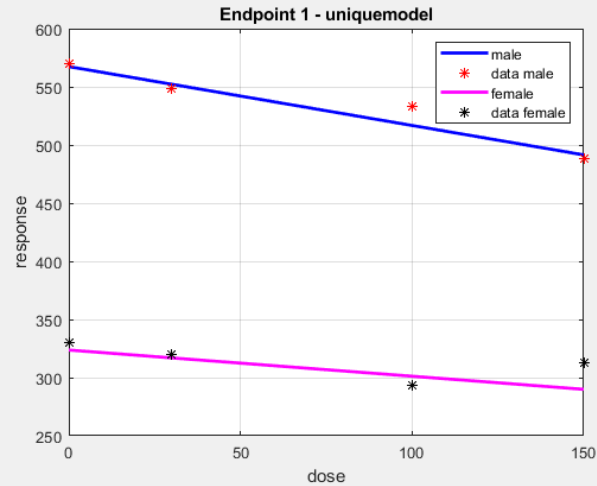


Objective tests evolution



Forward (Backward) selection – unique model

Theta	Std. Error
18.36327	0.1750687
0.909607	0.1925467
-0.01459	0.0017444
549.0169	24.3243
25.54763	0.3149547
-0.49133	0.2430934
-244.588	28.130922
0.280831	0.2931705



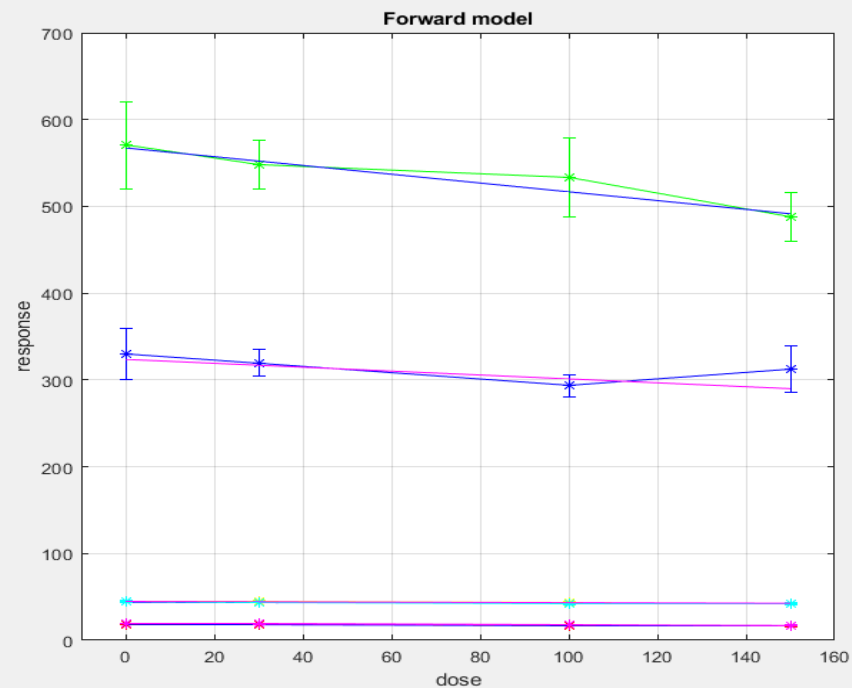
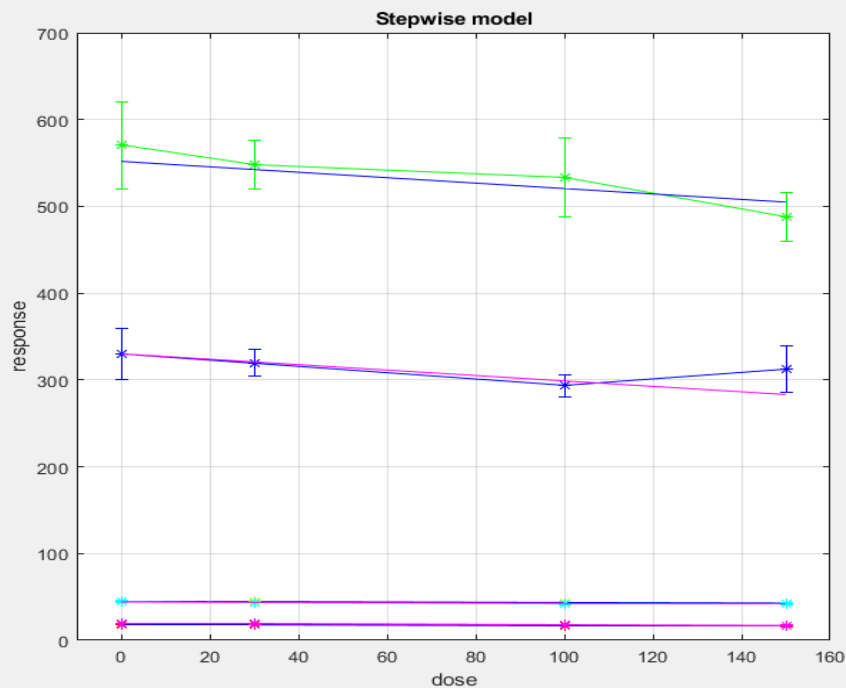
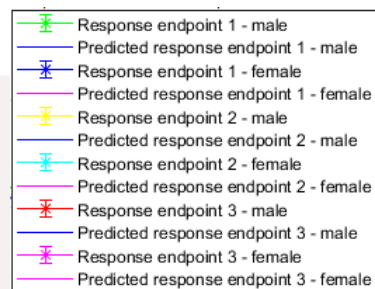
StepwiseLM (forward-backward) – unique model

Confidence intervals

17.9752	18.5190
-0.0171	-0.0118
0.7858	1.3985
506.1281	561.0667
25.9946	27.5863
-0.5043	-0.0925
-247.4567	-198.1291
-2.9325	-0.9453

```
mdl =  
Linear regression model:  
  response ~ 1 + dose*endpoint_1 + female*endpoint_1 + female*endpoint_2  
  
Estimated Coefficients:  
  
              Estimate      SE      tStat      pValue  
-----  
(Intercept)      18.247      0.12825     142.28     2.9732e-26  
dose             -0.014438     0.0012474    -11.575     3.4626e-09  
female_1          1.0922      0.14453      7.5566     1.1526e-06  
endpoint_1_1      533.6       12.958      41.18      1.149e-17  
endpoint_2_1       26.79      0.37542     71.361     1.8213e-21  
dose:endpoint_1_1 -0.29841     0.097136     -3.072      0.0072948  
female_1:endpoint_1_1 -222.79     11.634     -19.149     1.8681e-12  
female_1:endpoint_2_1 -1.9389      0.4687     -4.1368     0.00077449  
  
Number of observations: 24, Error degrees of freedom: 16  
Root Mean Squared Error: 1.2  
R-squared: 0.999, Adjusted R-Squared: 0.999  
F-statistic vs. constant model: 2.62e+03, p-value = 2.96e-23
```

Jump to the conclusions



Thanks for your attention!



Repository: <https://github.com/domenico-rgs/EFSA-Project>