Advanced Statistical Modeling

Non-parametric models - Comparisson of regression functions

Haoran Mo, Alexandra Yamaui

December 17th, 2017

In this exercise, we are going to compare different nonparametric regression functions graphically and formally in a hypothesis test performing analysis of covariance with the different populations.

In this opportunity, we are going to use the Hirsutism dataset, which contains information about female patients who suffer from this condition. Hirsutism is the excessive hairiness on women in those parts of the body where terminal hair does not normally appear or is minimal. However, Hirsutism is a symptom rather than a disease and may be a sign of other medical conditions. The amount and location of the hair are measured by a Ferriman-Gallwey score in 9 body areas. In this way, hair growth is rated from 0 (no growth of terminal hair) to 4 (extensive hair growth) in each of the nine locations. A patient's score may, therefore, range from a minimum score of 0 to a maximum score of 36.

A clinical trial was conducted to evaluate the effectiveness of an antiandrogen combined with an oral contraceptive in reducing hirsutism for 12 consecutive months. Patients were split into 4 treatment levels: level 0 (only contraceptive), 1, 2, and 3 of the antiandrogen in the study (always in combination with the contraceptive).

This dataset contains artificial values of measures corresponding to some patients in this study. The variables are the following:

- Treatment, with values 0, 1, 2 or 3.
- FGm0, it indicates the baseline hirsutism level at the randomization moment (the beginning of the clinical trial). Only women with baseline FG values grater than 15 where recruited.
- FGm3, FG value at 3 months.
- FGm6, FG value at 6 months.
- FGm12, FG value at 12 months, the end of the trial.
- SysPres, baseline systolic blood pressure.
- DiaPres, baseline diastolic blood pressure.
- weight, baseline weight.
- height, baseline height.

With this information we are going to make hypothesis testing to compare nonparametric regression functions over the data.

Considering the following form of the I regression functions to compare:

$$y_{ij} = m_i(x_{ij}) + \varepsilon_{ij}, j = 1, .., n_i, i = 1, ..., I.$$

we want to test the hypothesis:

$$H0 : m_i(x) = m(x), i = 1, .., I \text{ for all } x,$$

$$H1 : \text{not all the regression functions are equal.}$$

$m_i(x)$ being the regression function using data from subpopulation $i, i = 1, ..., I$

For the test hypothesis we are going to use the *ancova* function from the *sm* package, which is a developed version of code originally written by Stuart Young.

Point 7

First, we are going to compare the regression curves of $FGm12$ as a function of $FGm0$ in the four groups defined by $Treatment$. To do this, we will use the bandwidth values $h1$ and $h2$ obtained from the $h.select$ function using cross-validation and the AICc criterion, respectively.

```
attach(hirs)
(h1 <- h.select(FGm0,FGm12,method="cv", group = Treatment))
```

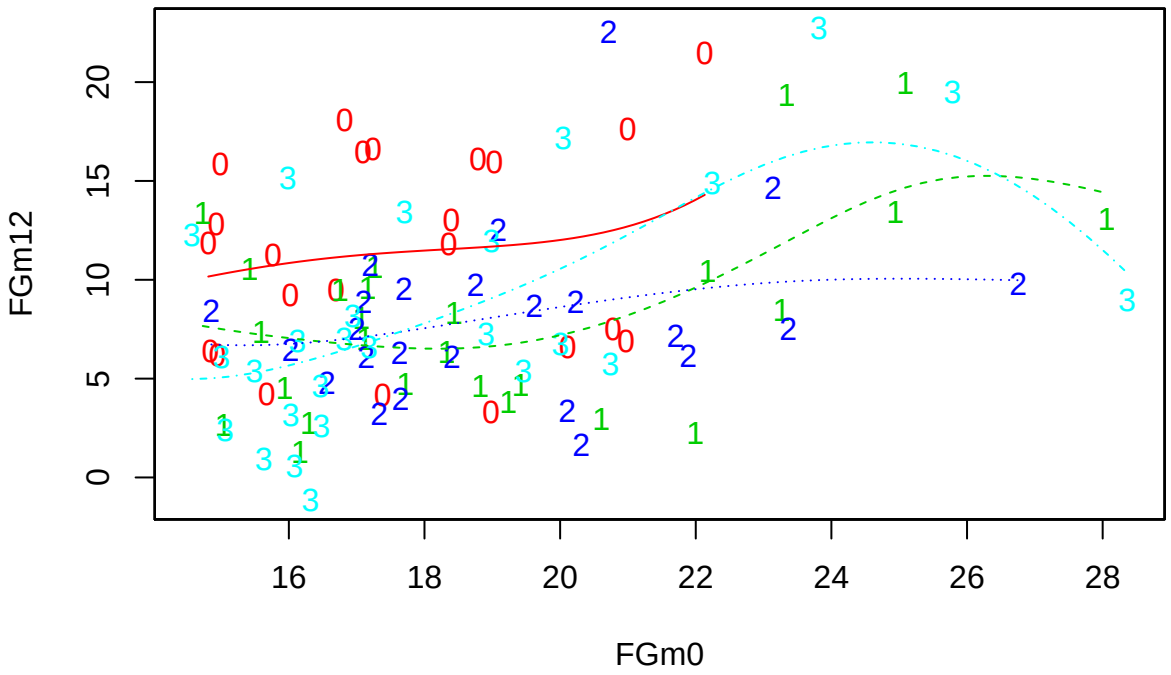
```
## [1] 2.504583
```

```
(h2 <- h.select(FGm0,FGm12,method="aicc", group = Treatment))
```

```
## [1] 3.259888
```

```
hvec = seq(min(h1,h2)/3,3*max(h1,h2), length=20)
s1 = sm.ancova(FGm0,FGm12,g=Treatment,h = h1,model = 'equal')
```

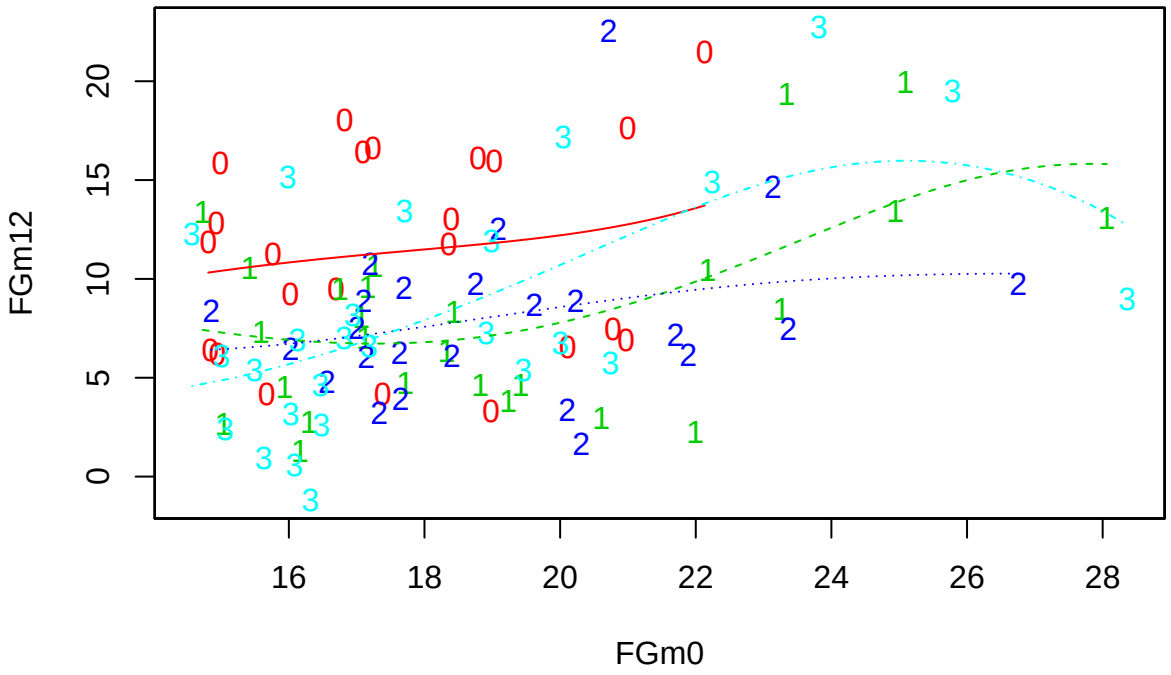
```
## Test of equality :  h =  2.50458    p-value =  0.0475
```




```
## Band available only to compare two groups.
```

```
sm.ancova(FGm0,FGm12,g=Treatment,h = h2,model = 'equal')
```

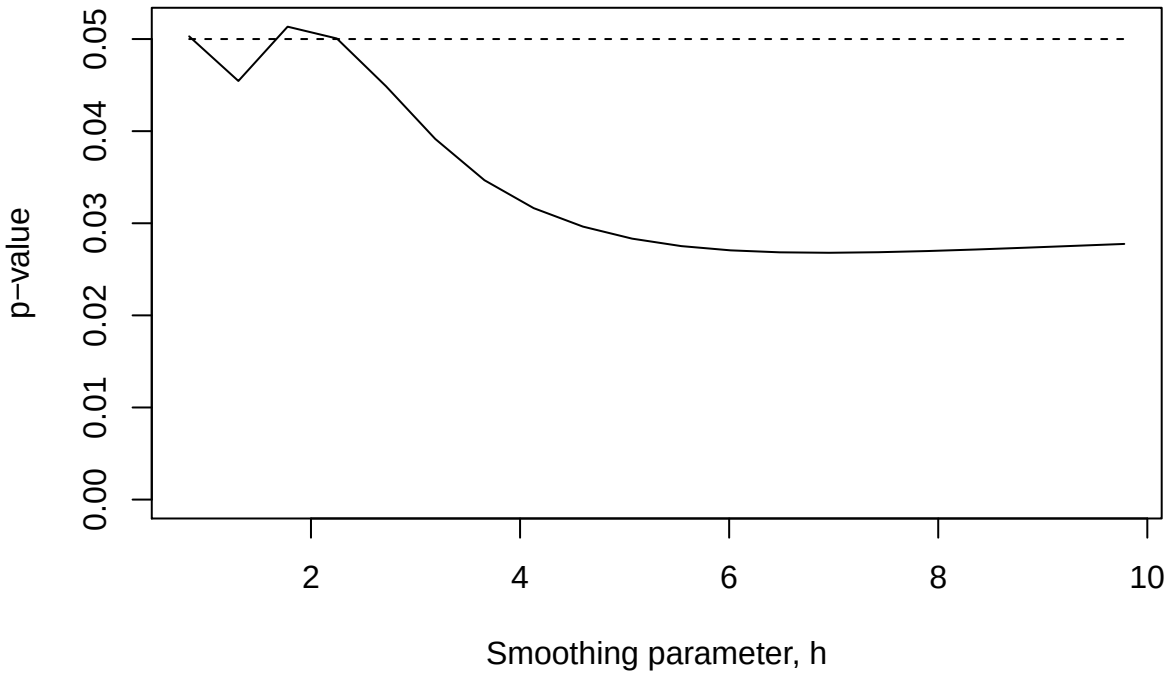
```
## Test of equality :  h =  3.25989    p-value =  0.0384
```




```
## Band available only to compare two groups.
```

```
sig.trace(sm.ancova(FGm0, FGm12,Treatment, model="equal",display="none"),
          h=hvec)
```

```
## Test of equality :  h =  0.834861    p-value =  0.0503
## Test of equality :  h =  1.30564    p-value =  0.0454
## Test of equality :  h =  1.77642    p-value =  0.0513
## Test of equality :  h =  2.2472    p-value =  0.0501
## Test of equality :  h =  2.71798    p-value =  0.0449
## Test of equality :  h =  3.18876    p-value =  0.0392
## Test of equality :  h =  3.65954    p-value =  0.0347
## Test of equality :  h =  4.13031    p-value =  0.0316
## Test of equality :  h =  4.60109    p-value =  0.0296
## Test of equality :  h =  5.07187    p-value =  0.0283
## Test of equality :  h =  5.54265    p-value =  0.0275
## Test of equality :  h =  6.01343    p-value =  0.0271
## Test of equality :  h =  6.48421    p-value =  0.0268
## Test of equality :  h =  6.95499    p-value =  0.0268
## Test of equality :  h =  7.42577    p-value =  0.0269
## Test of equality :  h =  7.89655    p-value =  0.027
## Test of equality :  h =  8.36733    p-value =  0.0272
## Test of equality :  h =  8.83811    p-value =  0.0273
## Test of equality :  h =  9.30888    p-value =  0.0276
## Test of equality :  h =  9.77966    p-value =  0.0278
```



From the plot of the first ancova using $h1 = 2.5045831$ we can see that the 4 curves behave differently, however, the group 1 and 3 have a slightly similar curves. If we check the p-value we see that this is close but lower than 0.05, which means the rejection of null hypothesis and indicating that the regression functions are not equal.

From the second ancova using $h2 = 3.2598879$ we can see the curves behave more or less the same as before with a p-value still lower than 0.05 and therefore, we reject again the null hypothesis.

Additionally, from the signifcance trace we can see that while the bandwith parameter gets bigger the p-values decrease, having a little bump near to value 2.

Point 8

Now, we are going to test if the regression function $FGm12 \sim FGm0$ can be considered **equal** or **parallel** in the two subpopulations defined according to $Treatment = 0$ or not.

```
data8 <- hirs
value <- c(1,0,0,0)
index <- c(0,1,2,3)
data8$Tr0 <- value[match(Treatment,index)]
detach(hirs)
attach(data8)
(h1 <- h.select(FGm0,FGm12,method="cv", group = Tr0))
```

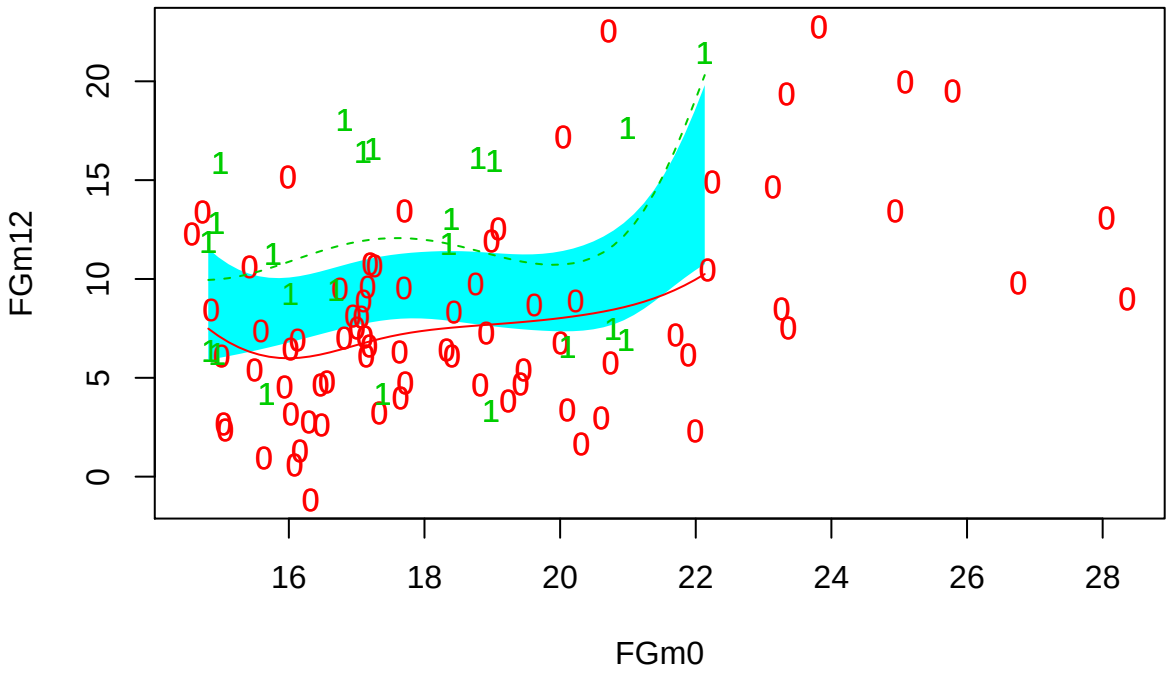
```
## [1] 1.121807
```

```
(h2 <- h.select(FGm0,FGm12,method="aicc", group = Tr0))
```

```
## [1] 2.464908
```

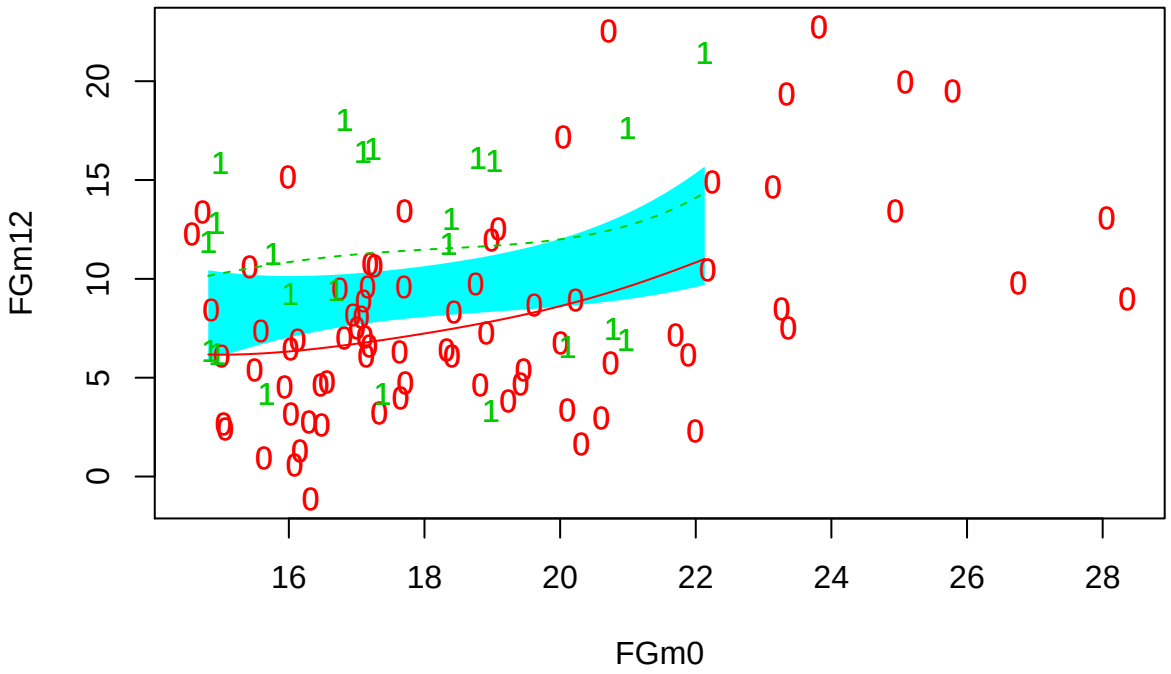
```
hvec = seq(min(h1,h2)/3,3*max(h1,h2), length=20)

sm.ancova(FGm0,FGm12,g=Tr0,h = h1,model = 'equal')
```

```
## Test of equality :  h =  1.12181    p-value =  0.0034
```




```
sm.ancova(FGm0,FGm12,g=Tr0,h = h2,model = 'equal')
```

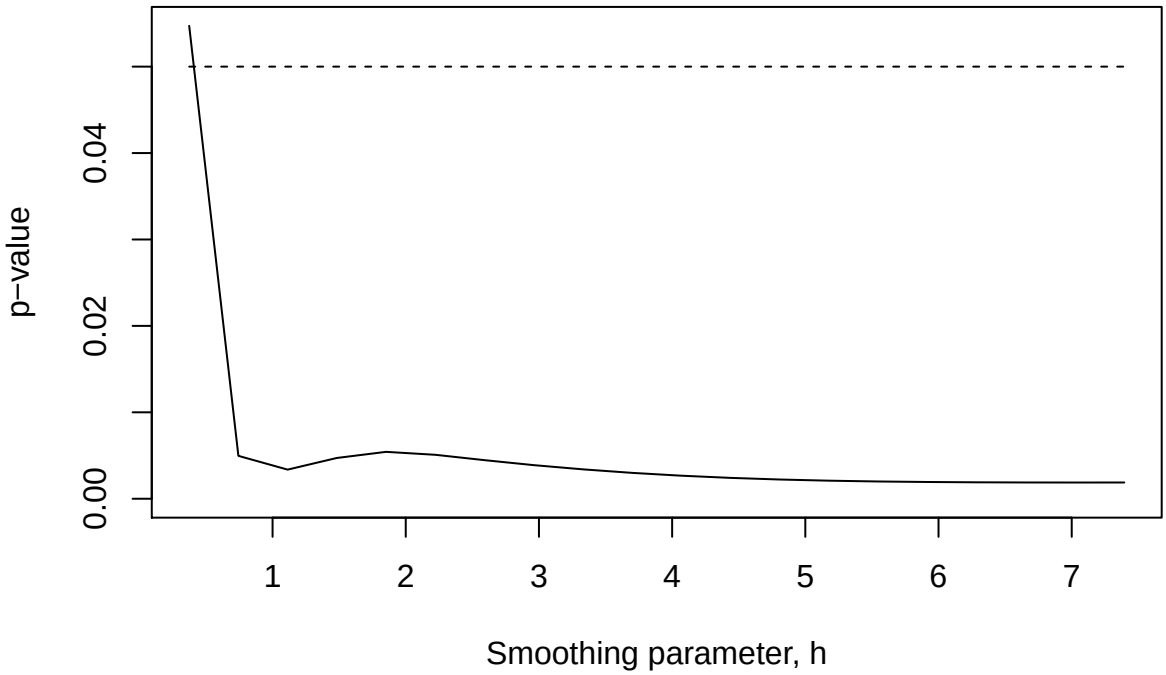
```
## Test of equality :  h =  2.46491    p-value =  0.0047
```




```
sig.trace(sm.ancova(FGm0, FGm12,g=Tr0, model="equal",display="none"),
          h=hvec)
```

```
## Test of equality :  h =  0.373936    p-value =  0.0547
## Test of equality :  h =  0.743451    p-value =  0.005
## Test of equality :  h =  1.11297    p-value =  0.0034
```

```
## Test of equality :  h =  1.48248    p-value =  0.0047
## Test of equality :  h =  1.852    p-value =  0.0054
## Test of equality :  h =  2.22151    p-value =  0.0051
## Test of equality :  h =  2.59103    p-value =  0.0045
## Test of equality :  h =  2.96054    p-value =  0.0039
## Test of equality :  h =  3.33006    p-value =  0.0034
## Test of equality :  h =  3.69957    p-value =  0.003
## Test of equality :  h =  4.06909    p-value =  0.0027
## Test of equality :  h =  4.4386    p-value =  0.0024
## Test of equality :  h =  4.80812    p-value =  0.0022
## Test of equality :  h =  5.17763    p-value =  0.0021
## Test of equality :  h =  5.54715    p-value =  0.002
## Test of equality :  h =  5.91666    p-value =  0.0019
## Test of equality :  h =  6.28618    p-value =  0.0019
## Test of equality :  h =  6.65569    p-value =  0.0019
## Test of equality :  h =  7.02521    p-value =  0.0019
## Test of equality :  h =  7.39473    p-value =  0.0019
```



From the equality ancova test plot (using $h1$) we can see that both curves are mostly outside of the reference bands and they behave in different ways. Indeed, if we check the p-value we can see that it is far lower than 0.05, making us reject the null hypothesis for equality. The same occurs using $h2$ bandwidth, with a p-value (0.0047) lower than 0.05.

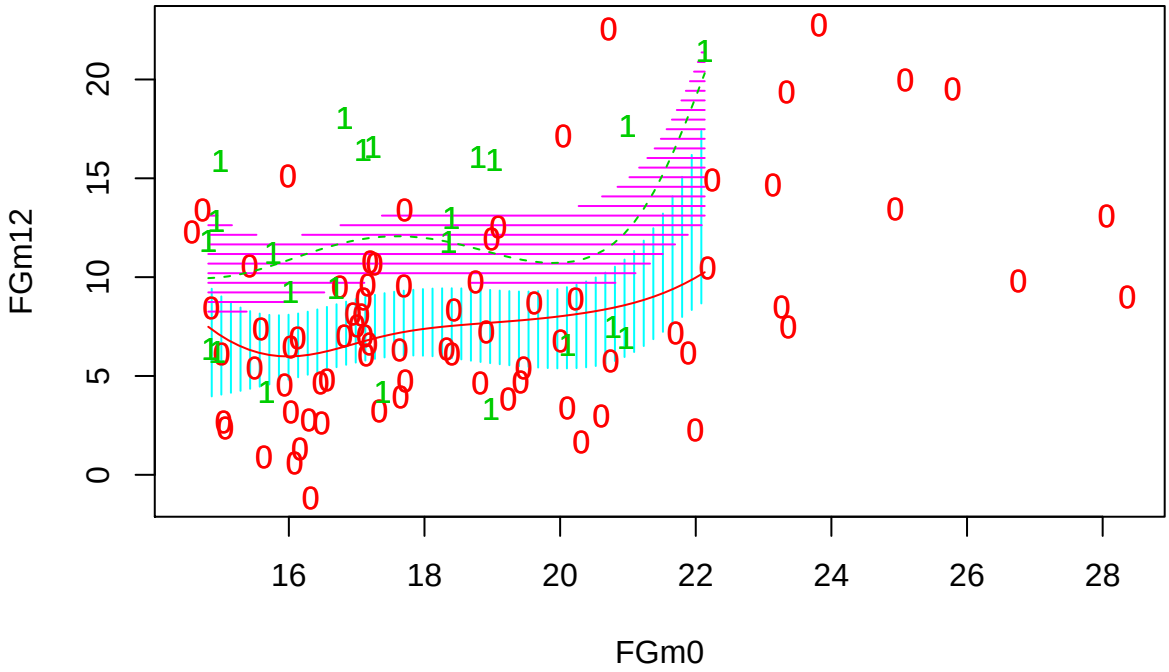
Looking at the significance trace plot we can see that the p-values are all lower than 0.05, existing a light increase around value 2.

On the other hand, when we perform the ancova test for parallel regression functions with $h1$ (below) we can see that now both curves are inside the reference bands and the p-value (0.5152) is significant. Hence, we cannot reject the null hypothesis of these regression functions being parallel. In the same way, when we perform the ancova with $h2$, again both curves are inside the bands and the p-value is greater than 0.05, meaning that both curves are parallel.

The significance trace shows that almost all the p-values are higher than 0.05, existing an increase around values 2 and 3.

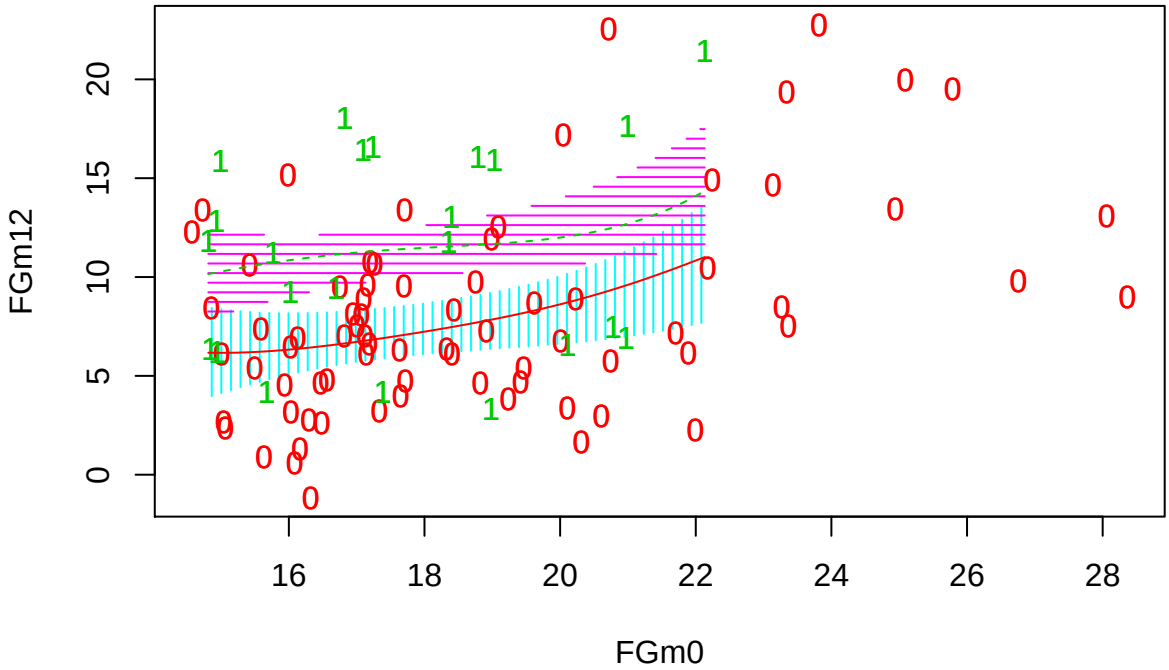
```
sm.ancova(FGm0,FGm12,g=Tr0,h = h1,model = 'parallel')
```

```
## Test of parallelism :  h =  1.12181    p-value =  0.5152
```




```
sm.ancova(FGm0,FGm12,g=Tr0,h = h2,model = 'parallel')
```

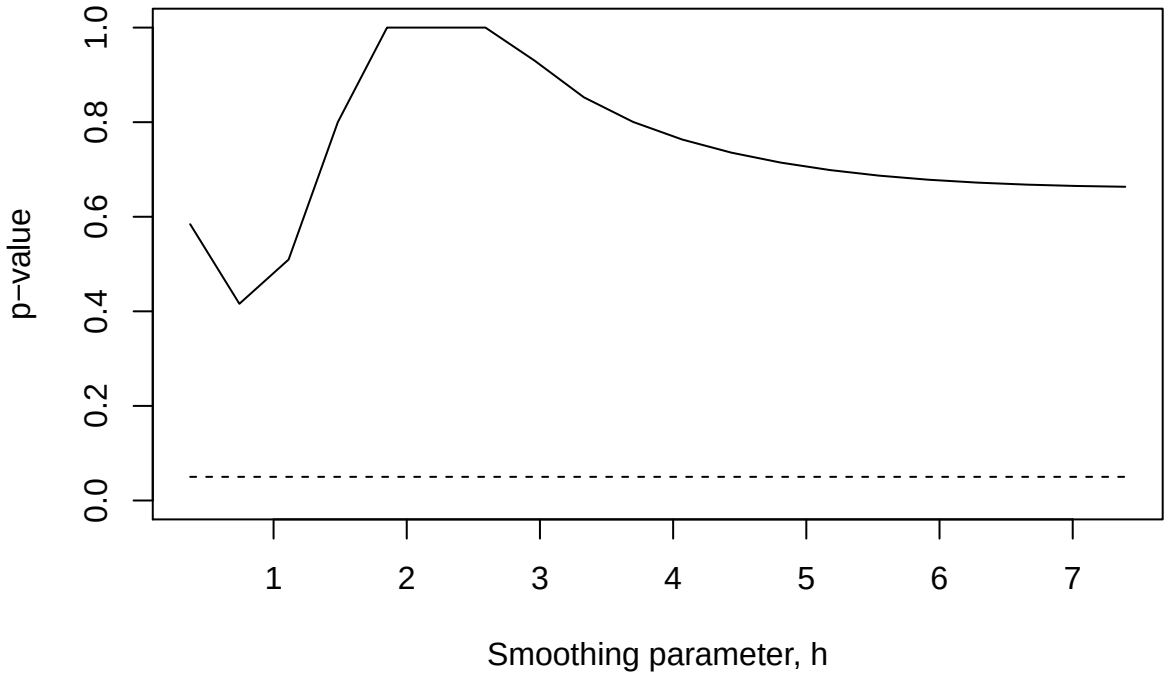
```
## Test of parallelism :  h =  2.46491    p-value =  1
```




```
sig.trace(sm.ancova(FGm0, FGm12,g=Tr0, model="parallel",display="none"),
          h=hvec)
```

```
## Test of parallelism :  h =  0.373936    p-value =  0.5842
```

```
## Test of parallelism :  h =  0.743451    p-value =  0.416
## Test of parallelism :  h =  1.11297    p-value =  0.5095
## Test of parallelism :  h =  1.48248    p-value =  0.8001
## Test of parallelism :  h =  1.852    p-value =  1
## Test of parallelism :  h =  2.22151    p-value =  1
## Test of parallelism :  h =  2.59103    p-value =  1
## Test of parallelism :  h =  2.96054    p-value =  0.93
## Test of parallelism :  h =  3.33006    p-value =  0.8526
## Test of parallelism :  h =  3.69957    p-value =  0.8005
## Test of parallelism :  h =  4.06909    p-value =  0.7632
## Test of parallelism :  h =  4.4386    p-value =  0.7355
## Test of parallelism :  h =  4.80812    p-value =  0.7145
## Test of parallelism :  h =  5.17763    p-value =  0.6987
## Test of parallelism :  h =  5.54715    p-value =  0.6869
## Test of parallelism :  h =  5.91666    p-value =  0.6783
## Test of parallelism :  h =  6.28618    p-value =  0.6722
## Test of parallelism :  h =  6.65569    p-value =  0.6679
## Test of parallelism :  h =  7.02521    p-value =  0.6651
## Test of parallelism :  h =  7.39473    p-value =  0.6634
```




```
detach(data8)
```

Point 9

Then we are going to test if the regression function $FGm12 \sim FGm0$ can be considered equal or parallel using only patients with treatments 1 and 3.

```
attach(hirs)
data9 <- hirs[which(Treatment==1 | Treatment ==3),]

detach(hirs)
```

```
attach(data9)

(h1 <- h.select(FGm0,FGm12,method="cv", group = Treatment))
```

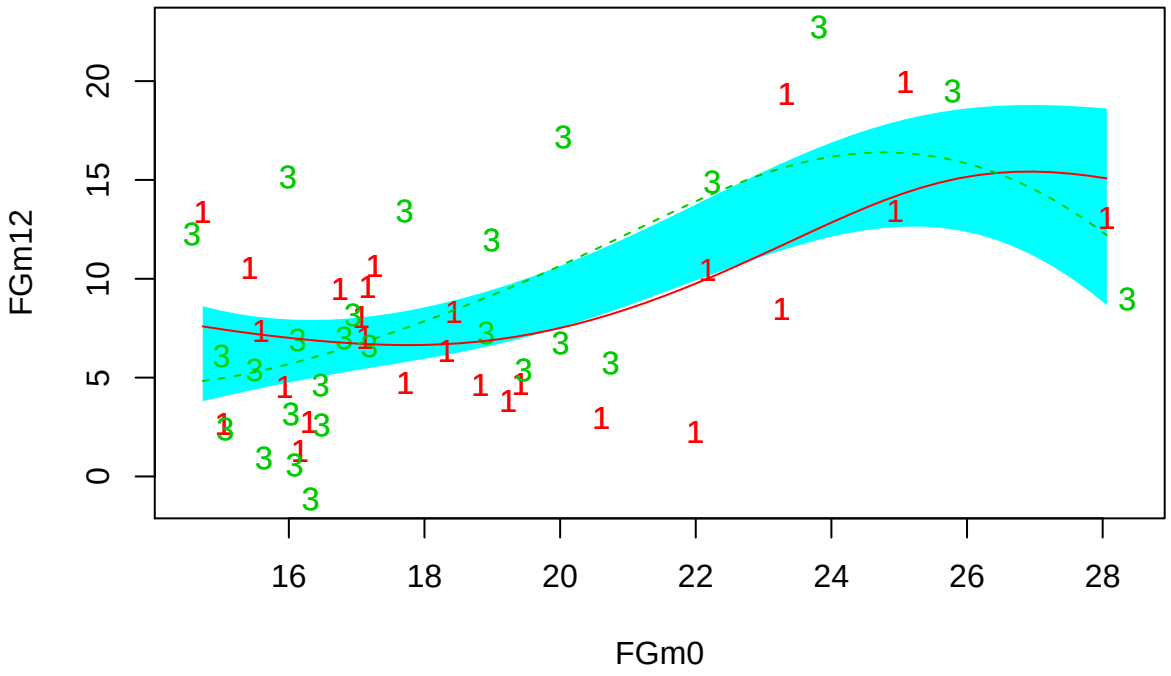
```
## [1] 2.848915
```

```
(h2 <- h.select(FGm0,FGm12,method="aicc", group = Treatment))
```

```
## [1] 2.123256
```

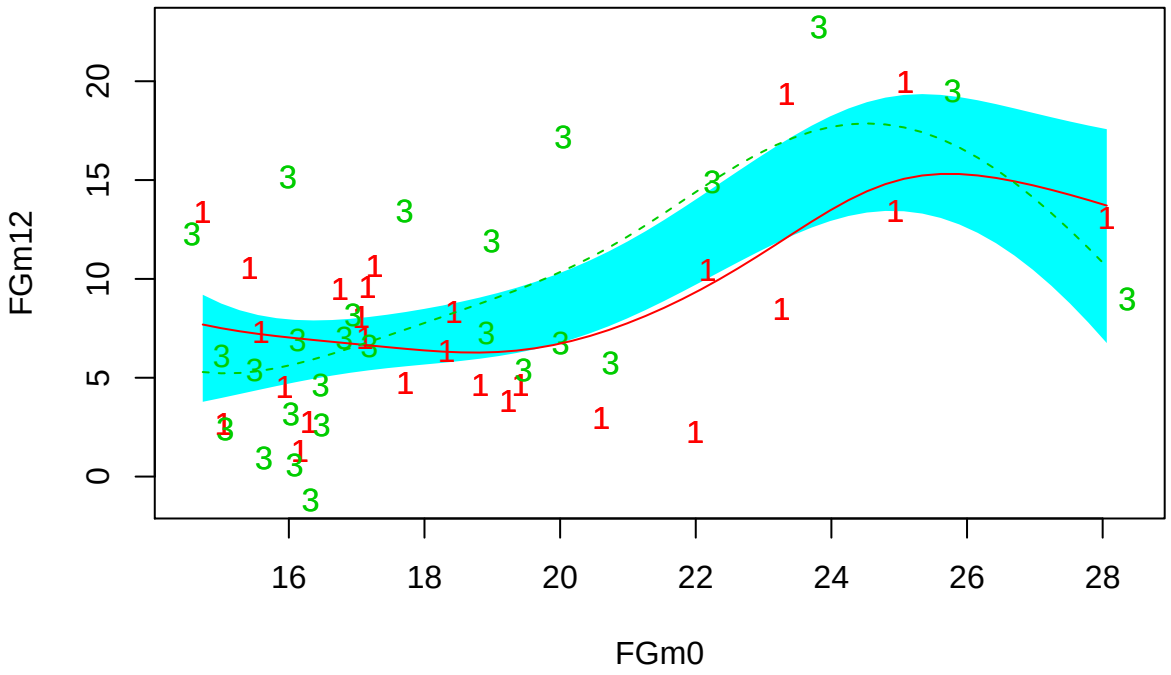
```
hvec = seq(min(h1,h2)/3,3*max(h1,h2), length=20)

sm.ancova(FGm0,FGm12,g=Treatment,h = h1,model = 'equal')
```

```
## Test of equality :  h =  2.84892    p-value =  0.2974
```



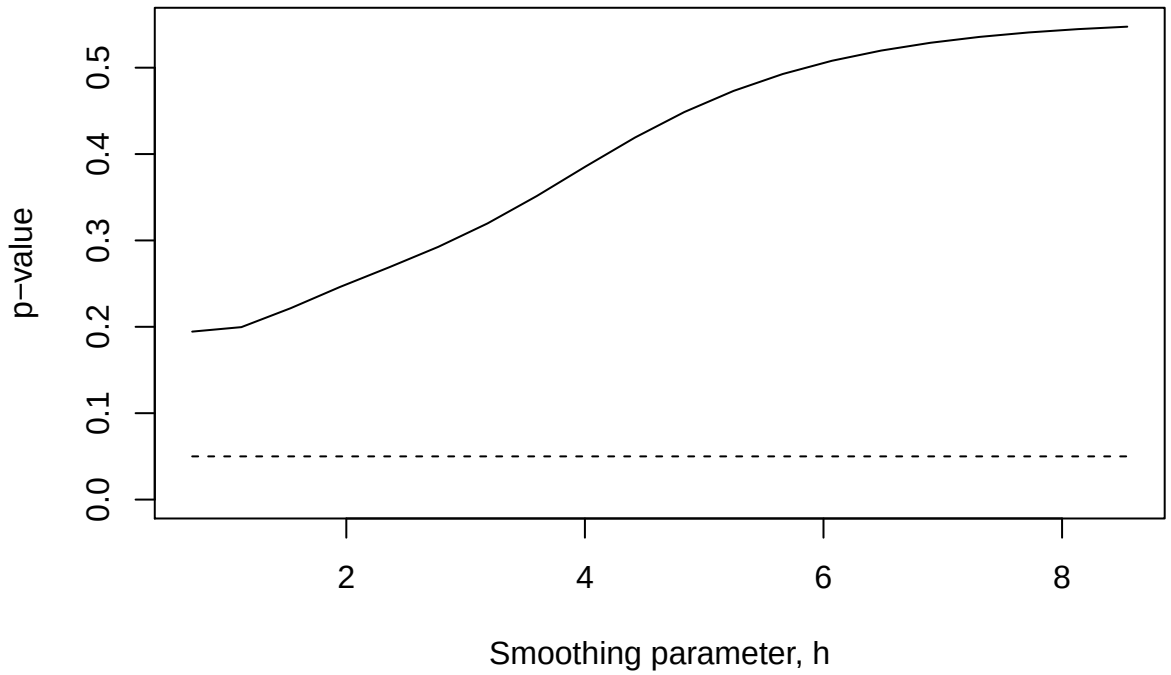

```
sm.ancova(FGm0,FGm12,g=Treatment,h = h2,model = 'equal')
```

```
## Test of equality :  h =  2.12326    p-value =  0.256
```



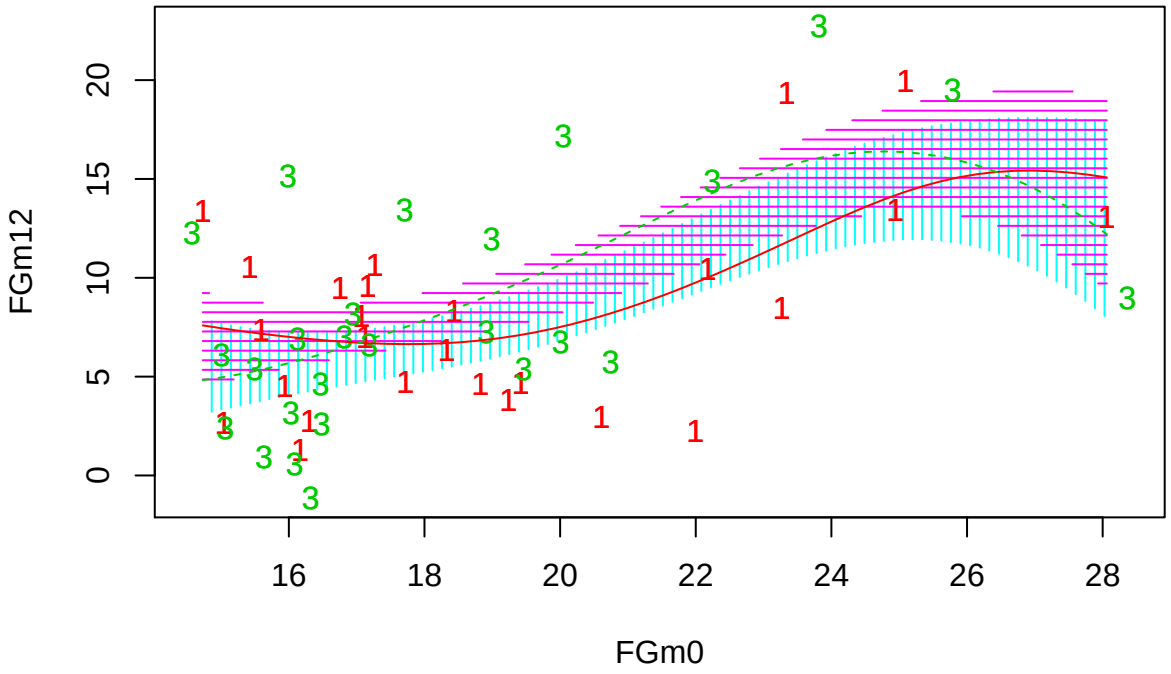

```
sig.trace(sm.ancova(FGm0, FGm12,g=Treatment, model="equal",display="none"),
          h=hvec)
```

```
## Test of equality :  h =  0.707752    p-value =  0.1945
## Test of equality :  h =  1.12033    p-value =  0.1996
## Test of equality :  h =  1.53291    p-value =  0.2215
## Test of equality :  h =  1.94549    p-value =  0.2461
## Test of equality :  h =  2.35807    p-value =  0.2688
## Test of equality :  h =  2.77065    p-value =  0.2926
## Test of equality :  h =  3.18322    p-value =  0.3197
## Test of equality :  h =  3.5958    p-value =  0.3513
## Test of equality :  h =  4.00838    p-value =  0.3858
## Test of equality :  h =  4.42096    p-value =  0.4192
## Test of equality :  h =  4.83354    p-value =  0.4486
## Test of equality :  h =  5.24612    p-value =  0.4731
## Test of equality :  h =  5.6587    p-value =  0.4927
## Test of equality :  h =  6.07127    p-value =  0.508
## Test of equality :  h =  6.48385    p-value =  0.5198
## Test of equality :  h =  6.89643    p-value =  0.5288
## Test of equality :  h =  7.30901    p-value =  0.5357
## Test of equality :  h =  7.72159    p-value =  0.5408
## Test of equality :  h =  8.13417    p-value =  0.5446
## Test of equality :  h =  8.54675    p-value =  0.5475
```



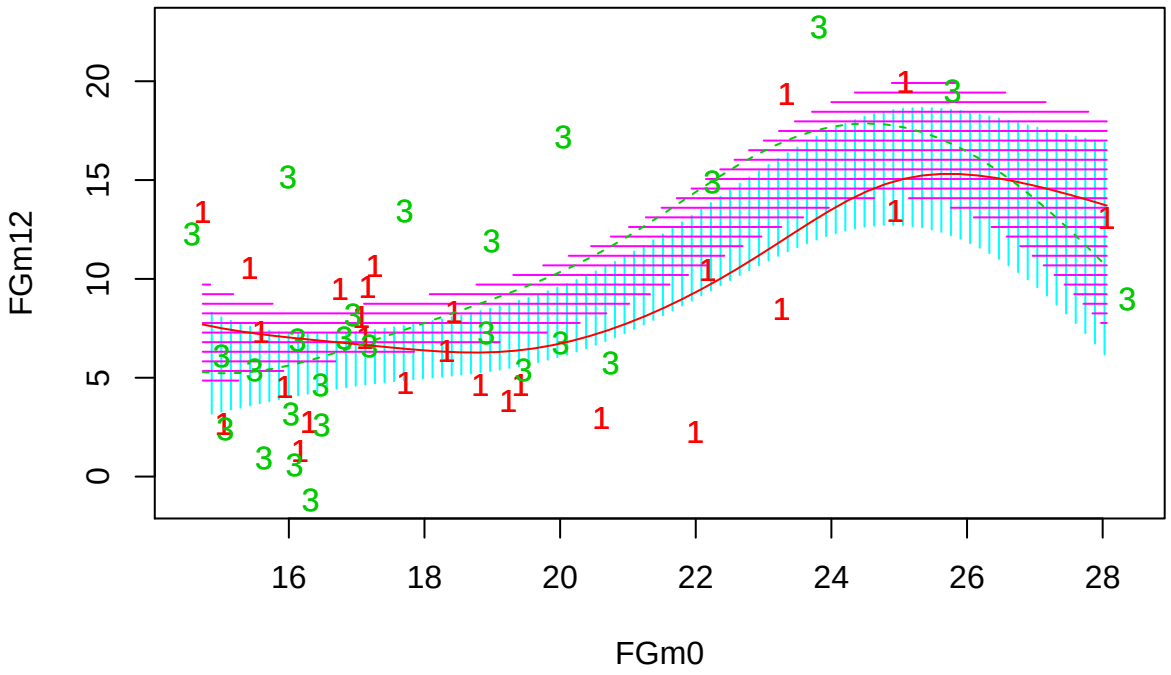

```
sm.ancova(FGm0,FGm12,g=Treatment,h = h1,model = 'parallel')
```

```
## Test of parallelism :  h =  2.84892    p-value =  0.1554
```



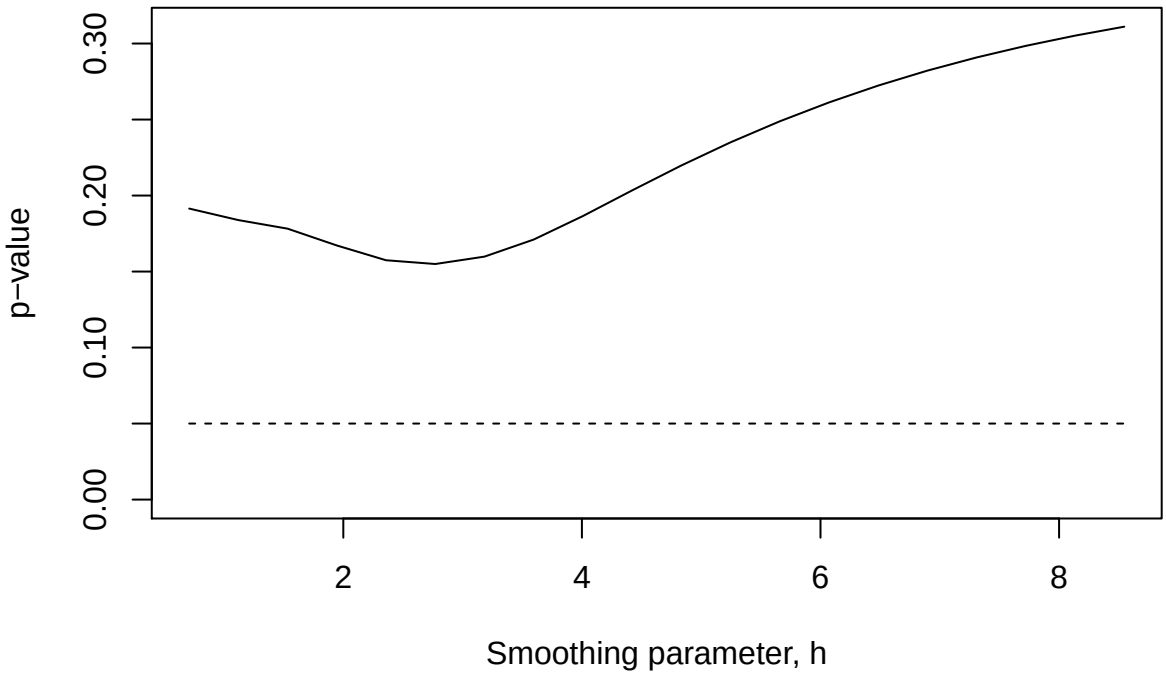

```
sm.ancova(FGm0,FGm12,g=Treatment,h = h2,model = 'parallel')
```

```
## Test of parallelism :  h =  2.12326    p-value =  0.1623
```




```
sig.trace(sm.ancova(FGm0, FGm12,g=Treatment, model="parallel",display="none"),
          h=hvec)
```

```
## Test of parallelism :  h =  0.707752    p-value =  0.1914
## Test of parallelism :  h =  1.12033    p-value =  0.1839
## Test of parallelism :  h =  1.53291    p-value =  0.1782
## Test of parallelism :  h =  1.94549    p-value =  0.1672
## Test of parallelism :  h =  2.35807    p-value =  0.1574
## Test of parallelism :  h =  2.77065    p-value =  0.155
## Test of parallelism :  h =  3.18322    p-value =  0.1598
## Test of parallelism :  h =  3.5958    p-value =  0.1711
## Test of parallelism :  h =  4.00838    p-value =  0.1865
## Test of parallelism :  h =  4.42096    p-value =  0.2034
## Test of parallelism :  h =  4.83354    p-value =  0.2198
## Test of parallelism :  h =  5.24612    p-value =  0.235
## Test of parallelism :  h =  5.6587    p-value =  0.2488
## Test of parallelism :  h =  6.07127    p-value =  0.2613
## Test of parallelism :  h =  6.48385    p-value =  0.2723
## Test of parallelism :  h =  6.89643    p-value =  0.2822
## Test of parallelism :  h =  7.30901    p-value =  0.2908
## Test of parallelism :  h =  7.72159    p-value =  0.2985
## Test of parallelism :  h =  8.13417    p-value =  0.3052
## Test of parallelism :  h =  8.54675    p-value =  0.3111
```




```
detach(data9)
```

Performing an ancova test using bandwidth $h1$ we can see graphically that both curves have a similar behaviour and are inside the reference bands. Actually, if we check the p-value we can see that it is higher than 0.05, which indicates that we cannot reject the null hypothesis of equality between the 2 populations. A similar scenario occurs using $h2$, the p-value is signficant and the curves are inside the bands, with which we cannot reject the null hypothesis.

We can see in the significance trace plot that the p-values increase while the bandwidth parameter increases.

As is expected, performing a parallel hypothesis ancova over the two populations we cannot reject the null hypothesis for none of the 2 values of bandwidth. In this case, the significance trace shows that after the value 4 of bandwidth the p-value increases in a linear way.

Point 10

At last, we are going to test the linearity for the regression function $FGm12 \sim FGm0$ using only the patients with treatments 1, 2 or 3.

```
attach(hirs)
data10 <- hirs[which(Treatment==1 | Treatment ==2 | Treatment ==3),]
detach(hirs)
attach(data10)

(h1 <- h.select(FGm0,FGm12,method="cv", group = Treatment))
```

```
## [1] 3.546784
```

```
(h2 <- h.select(FGm0,FGm12,method="aicc", group = Treatment))
```

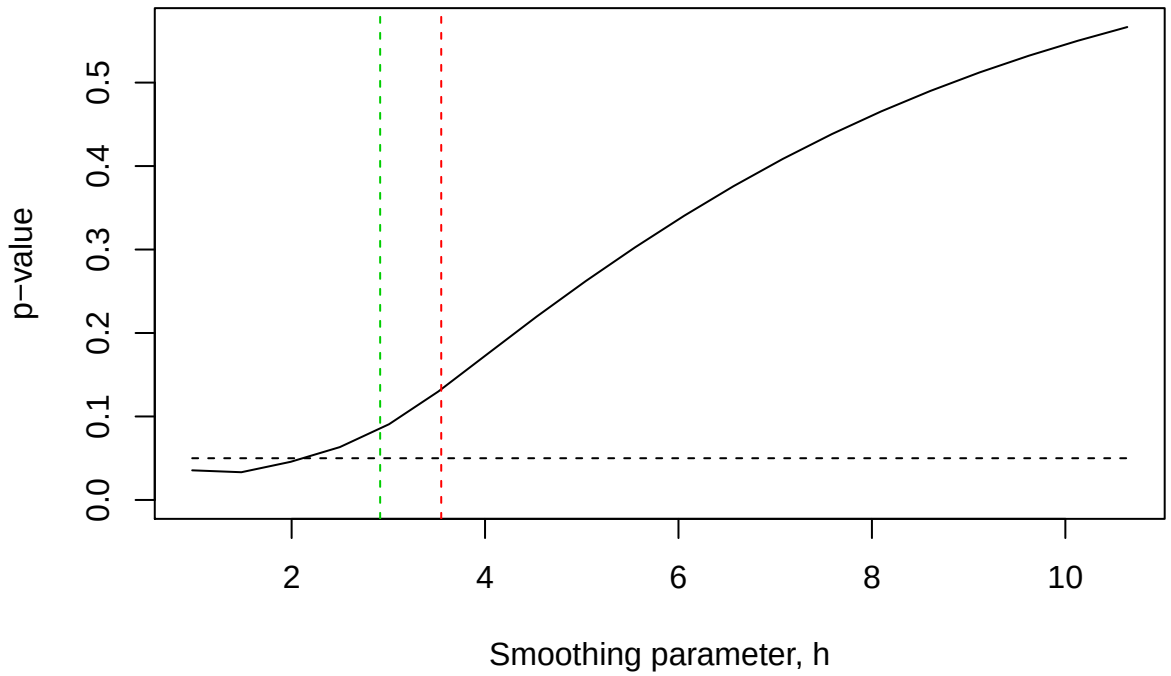
```
## [1] 2.915549
```

```
hvec = seq(min(h1,h2)/3,3*max(h1,h2), length=20)
```

```
sig.trace(sm.regression(FGm0, FGm12, model="linear",display="none"),
          h=hvec)
```

```
## Test of linear model:  significance =  0.035
## Test of linear model:  significance =  0.033
## Test of linear model:  significance =  0.046
## Test of linear model:  significance =  0.063
## Test of linear model:  significance =  0.091
## Test of linear model:  significance =  0.13
## Test of linear model:  significance =  0.175
## Test of linear model:  significance =  0.22
## Test of linear model:  significance =  0.262
## Test of linear model:  significance =  0.303
## Test of linear model:  significance =  0.341
## Test of linear model:  significance =  0.376
## Test of linear model:  significance =  0.409
## Test of linear model:  significance =  0.438
## Test of linear model:  significance =  0.465
## Test of linear model:  significance =  0.49
## Test of linear model:  significance =  0.512
## Test of linear model:  significance =  0.532
## Test of linear model:  significance =  0.55
## Test of linear model:  significance =  0.567
```

```
abline(v=h1,col=2,lty=2)
abline(v=h2,col=3,lty=2)
```



From the plot we can see that for both values of bandwith (h1=blue, h2=red) the p-value is greater than 0.05, therefore, we cannot reject the null hypothesis of linearity.