Semiparametric regression - Homework 7

Klaudia Weigel

1 Exercise 1

We consider the nonparametric regression model

$$y_i = f(x_i) + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma_\epsilon^2), \quad 1 \leq i \leq n$$

and the hypothesis testing problem

$$H_0 : f \text{ is linear} \quad \text{versus} \quad H_1 : f \text{ is a smooth nonlinear function.}$$

1.1 (a)

We will use the `RLRsim` library.

```
library(RLRsim)
```

1.2 (b)

We will first generate a dataset from the H_0 hypothesis, with $f(x) = x, \sigma_\epsilon = 1$ and $n = 200$.

```
set.seed(1)
x <- seq(0, 1, length = 200)
y <- x + rnorm(200)
```

1.3 (c)

In this point we will perform the following statistical tests:

1. F-test using the `gam()` default penalized spline fit,
2. F-test using an ordinary least squares fit with 2 linear and 3 spline basis functions,
3. exact restricted likelihood ratio test.

```
library(mgcv)
fitLine <- gam(y ~ x)
fitDfltPenSpl <- gam(y ~ s(x))
anova(fitLine, fitDfltPenSpl, test = "F")$"Pr(>F)"[2]
```

```
## [1] 7.514685e-09
```

```
fitOLSspl <- gam(y ~ s(x, k=5, sp=0))
anova(fitLine, fitOLSspl, test="F")$"Pr(>F)"[2]
```

```
## [1] 0.6872285
```

```
fitGAMM <- gamm(y~s(x), method="REML")
exactRLRT(fitGAMM$lme)[2]
```

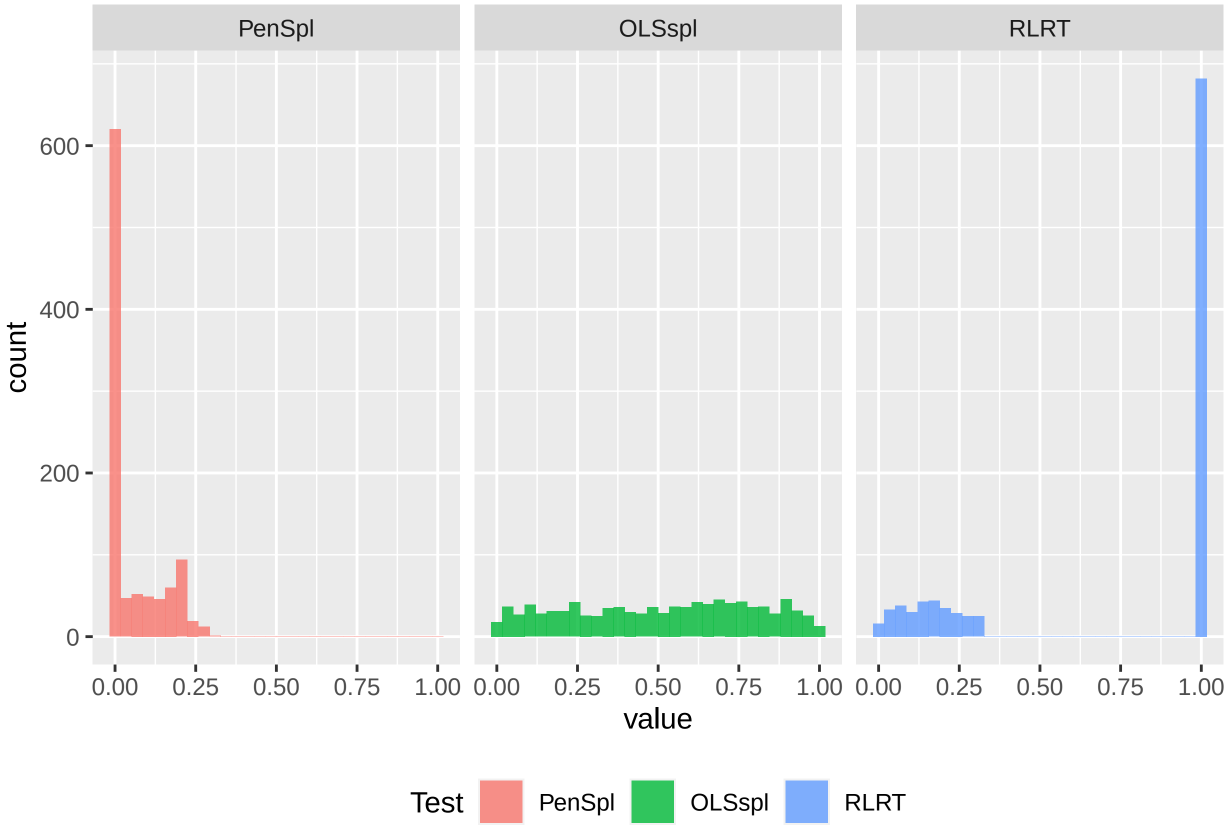
```
## $p.value
## [1] 1
```

Since our data is generated from the null hypothesis we should obtain large p-values and the tests should not reject the null hypothesis. This is not the case in the first test where the p-value is very close to zero.

1.4 (d)

We will replicate the experiment from point (c) 1000 times and plot histograms of the p-values.

```r
# one dataset in each column
rep <- 1000; Y <- replicate(rep, x + rnorm(200))

get_pvalues <- function(y) {
  fitLine <- gam(y ~ x)
  fitDfltPenSpl <- gam(y ~ s(x))
  p1 <- anova(fitLine, fitDfltPenSpl, test = "F")$"Pr(>F)"[2]

  fitOLSspl <- gam(y ~ s(x, k=5, sp=0))
  p2 <- anova(fitLine, fitOLSspl, test="F")$"Pr(>F)"[2]

  fitGAMM <- gamm(y~s(x), method="REML")
  p3 <- as.numeric(exactRLRT(fitGAMM$lme)[2])

  return(c(p1,p2,p3))
}

library(dplyr); library(reshape2); library(ggplot2)
pvals <- data.frame(t(apply(Y, 2, get_pvalues)))
colnames(pvals) <- c("PenSpl", "OLSspl", "RLRT")
pvals %>%
  melt() %>%
  ggplot(aes(x=value, fill=variable)) +
    geom_histogram(alpha=0.8) +
    facet_wrap(~variable) +
    theme(legend.position = 'bottom') +
    labs(fill = "Test")
```



We see from the above histograms that the test based on the default penalized spline most frequently rejects the null hypothesis, even though it is true. The distribution for the p-values of the F-test based on the OLS fit is the most uniform. In the case of the RLRT majority of the p-values are close to one.

1.5 (e)

For the significance level at 0.05 we will see what proportion of tests reject the null hypothesis.

```r
res <- apply(pvals, 2, function(p) sum(p < 0.05))/rep
res
```

```
## PenSpl OLSspl   RLRT
##  0.665  0.051  0.047
```

We see that a significant number of F-tests using default penalized spline fit reject the null hypothesis. Most of the F-tests using OLS fit accept the null hypothesis, similarly with RLRT. The last two tests both have rejection probabilities that are close to the chosen significance level.