Nonparametric Statistics - Exam Like Lab Ex I

Noé Debrois, Person Code 10949145, ID 242751

2024-02-01

Exercise 1

Dr. Simoni, a data scientist, just had his second kid (a boy, 58 cm tall) and he is interested to know how tall his child will be at 25 years of age, given his height at birth. To do so, he has collected the height at birth of 100 males, alongside the height they reached when they were 25. He has collected these data in the file boyheight.rda, height.25 is the height [cm] at 25 years of age, while height.b is the length of the newborn [cm]. Assume that : $height_{25} = f(height_b) + \varepsilon$.

Question 1.1

Build a degree 1 regression spline model, with breaks at the 25th and 75th percentile of $height_b$, to predict the height at 25 from the height at birth. Provide a plot of the regression line, compute the pointwise prediction (round it at the second decimal digit) for the height at 25 of Dr. Simoni newborn child, and calculate, using a bootstrap approach on the residuals, the bias, variance and Mean Squared Error (MSE) of such prediction.

Synthetic description of assumptions, methods, and algorithms

Regression spline model (degree = 1) :

At the contrary of smoothing kernel, you use splines when you want to keep the windows fixed but require appropriate continuity at knots. A spline is a polynomial of order m (= d+1) or degree m-1 (= d) over each subinterval identified by two consecutive knots. The polynomials link with continuity of order m-2 at the knots. Here we use the B-spline basis (computationally efficient ; knots can be placed along the percentiles of $height_b$)

Bootstrap :

Assumptions : S1 = (X1,. . . ,Xn)~p The primary task of bootstrapping is estimating from a random sample S1, the distribution of a statistic, it is to say an estimator theta_hat of an unknown parameter theta. The principle is to sample a new vector having the same size as S1 from S1 itself with replacement ! From this idea, we can provide more information about the quality of the estimator.

Results and brief discussion
Build the model with knots as specified :

```
##
## Call:
## lm(formula = height.25 ~ bs(height.b, knots = Q25_Q75_knots,
##     degree = 1))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.2055 -1.8814  0.0007  2.1218  6.7987
##
```

```
## Coefficients:
##                                                 Estimate Std. Error t value
## (Intercept)                                      155.830      1.701   91.60
## bs(height.b, knots = Q25_Q75_knots, degree = 1)1   22.520      1.997   11.28
## bs(height.b, knots = Q25_Q75_knots, degree = 1)2   32.416      1.738   18.65
## bs(height.b, knots = Q25_Q75_knots, degree = 1)3   37.115      2.430   15.27
##                                                 Pr(>|t|)
## (Intercept)                                       <2e-16 ***
## bs(height.b, knots = Q25_Q75_knots, degree = 1)1   <2e-16 ***
## bs(height.b, knots = Q25_Q75_knots, degree = 1)2   <2e-16 ***
## bs(height.b, knots = Q25_Q75_knots, degree = 1)3   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.984 on 96 degrees of freedom
## Multiple R-squared:  0.8592, Adjusted R-squared:  0.8548
## F-statistic: 195.3 on 3 and 96 DF,  p-value: < 2.2e-16
```

How to interpret summary(model) ?
This table is a return of the degree 1 regression spline model, with breaks at the 25th and 75th percentile of $height_b$, to predict the height at 25 from the height at birth. Here is how to interpret it :

Call: indique la formule du modèle.

Residuals:

- Min: la valeur minimale des résidus (différence entre les valeurs observées et les valeurs prédites par le modèle).
- 1Q: le premier quartile des résidus.
- Median: la médiane des résidus.
- 3Q: le troisième quartile des résidus.
- Max: la valeur maximale des résidus.

Coefficients:

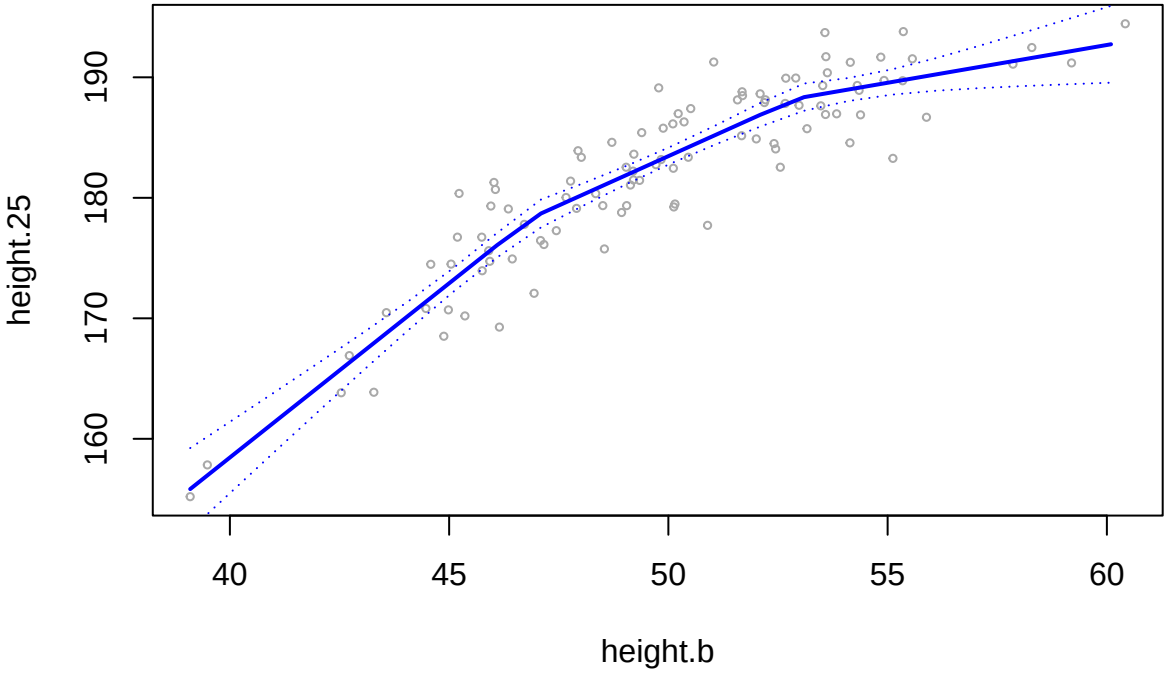
- Intercept : représente l'ordonnée à l'origine du modèle, c'est-à-dire la valeur attendue de la variable de réponse lorsque toutes les variables explicatives sont égales à zéro.
- bs(height.b, knots = Q25_Q75_knots, degree = 1) 1 à 3 : ce sont les coefficients associés aux termes de base spline degré 1 pour height.b. Chaque terme représente une partie différente de la fonction spline.
- Estimate: les valeurs estimées des coefficients.
- Std. Error: les erreurs standard associées à chaque coefficient.
- t value: la statistique de test t pour tester l'hypothèse nulle que le coefficient est égal à zéro.
- Pr(>|t|): la p value associée à la statistique de test t. Elle indique la probabilité d'observer une statistique de test aussi extrême que celle observée, sous l'hypothèse nulle.

En dessous du tableau :

- Signif. codes: indique le niveau de signification des coefficients. Les étoiles (*) sont souvent utilisées pour indiquer les niveaux de significativité, où plus d'étoiles indiquent une plus grande significativité.
- Residual standard error: C'est l'estimation de l'écart-type des résidus. Il mesure la dispersion des résidus autour de la ligne de régression.
- Multiple R-squared / Adjusted R-squared: Ces mesures indiquent la proportion de la variance de la variable de réponse expliquée par le modèle. Adjusted R-squared prend en compte le nombre de variables dans le modèle.
- F-statistic / p-value: Le F-statistic teste l'hypothèse nulle selon laquelle tous les coefficients de régression sont égaux à zéro (aucun effet global). La p-value associée indique la probabilité d'observer un tel F-statistic sous l'hypothèse nulle.

Dans l'ensemble, ce modèle semble expliquer une proportion significative de la variance de la variable de réponse, et tous les termes spline degré 1 pour height.b sont statistiquement significatifs.

Build the standard-error bands and plot the datapoints, the regression line and the se bands :



Compute the pointwise prediction of the height at 25 for Simoni's boy :

```
## [1] 191.43
```

Bootstrap approach to compute the bias, the variance and the MSE of the previous pointwise prediction :

The variance of the pointwise prediction :

```
## [1] 1.142986
```

The bias of the pointwise prediction :

```
## [1] -0.04475737
```

The MSE of the pointwise prediction :

```
## [1] 1.144989
```

Question 1.2

Dr.Simoni is not particularly satisfied with the predictions of such a simple model, and would like you to do more. So build a prediction model for the height at 25 based on a smoothing spline of order 4 (select the lambda parameter via Leave-One-Out CV). Report the optimal lambda value (2 decimal digits), provide a plot of the regression line, alongside the point-wise prediction of the height at 25 of Dr. Simoni's kid, and calculate using a bootstrap approach on the residuals the bias, variance and MSE of such prediction (fix the lambda value to the one obtained via Leave-One-Out CV).

Synthetic description of assumptions, methods, and algorithms
A Spline is a Piecewise Polynomial regression constrained to be continuous at its knots. A Natural Spline is a Spline such that:

- Aimed at reducing variance at boundaries of the domain.
- There is a Polynomial of degree k in each "interior" interval of the partition induced by the knots.
- There is a Polynomial of degree $\frac{k-1}{2}$ on the boundary intervals.

- Continuous derivatives at the knots.
- The spline basis is orthonormal.

A Smoothing Spline is a Natural Spline with a curvature penalty in the objective function.

The goal of cross-validation is to test the model's ability to predict new data that was not used in estimating it, in order to flag problems like overfitting or selection bias and to give an insight on how the model will generalize to an independent dataset. Leave-one-out cross-validation (LOOCV) is a method which learns and tests on all possible ways to divide the original sample into a training and a validation set. Cross-validation computes the statistic on the left-out sample.

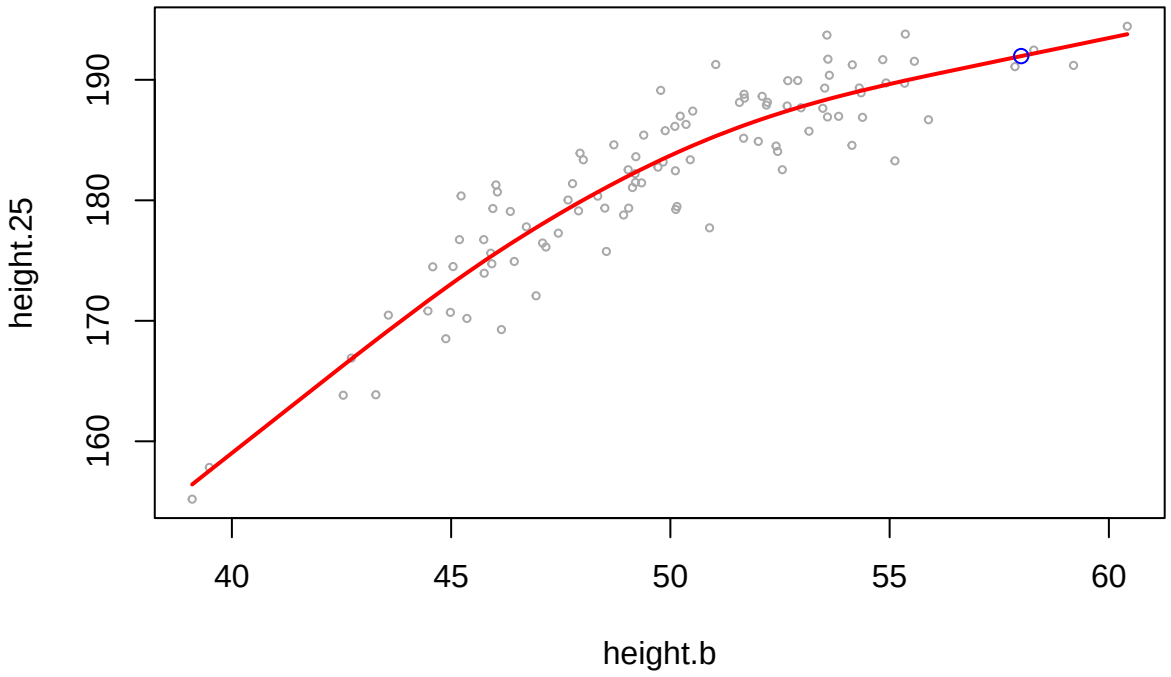
Results and brief discussion

Order 4 so degree 3 : we can use smooth.spline (which "fits a cubic smoothing spline to the supplied data") :

```
## [1] 0.01
```

Above, the lambda found by CV. Here, the pointwise prediction :

```
## [1] 191.97
```



Bootstrap approach to compute the bias, the variance and the MSE of the previous pointwise prediction :

Variance of the pointwise prediction :

```
## [1] 0.8543512
```

Bias of the pointwise prediction :

```
## [1] 0.3625338
```

MSE of the pointwise prediction :

```
## [1] 0.985782
```