Nonparametric Statistics

Noé Debrois, Person Code 10949145, ID 242751

2024-01-12

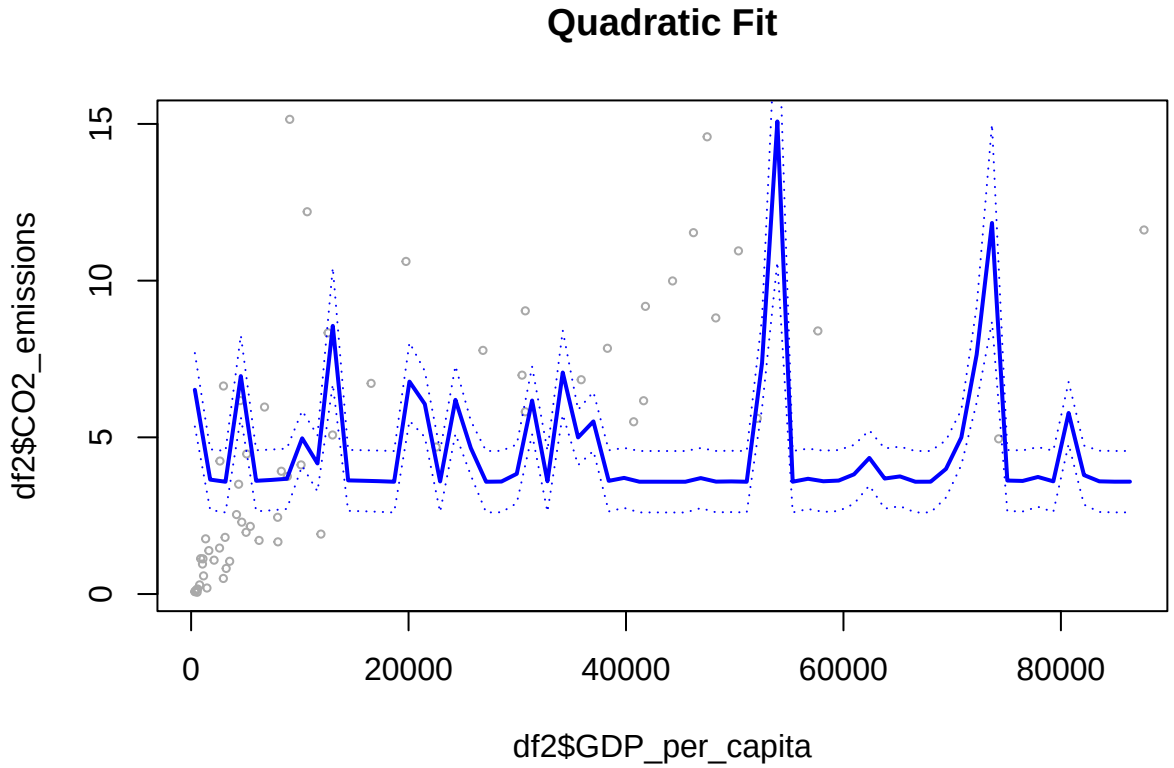
Exercise 2

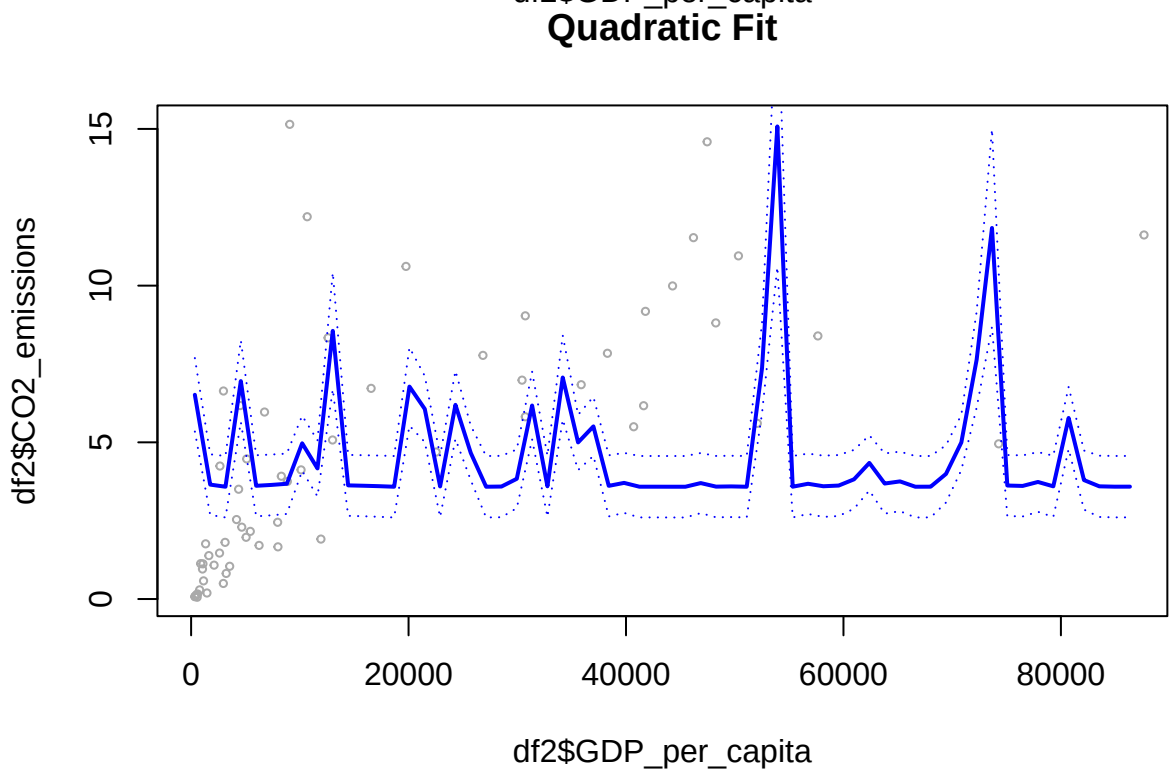
Question 2.1

Synthetic description of assumptions, methods, and algorithms
We apply a raw polynomial fit here twice. On the first, we see thanks to the pvalues that all the coefficient are *** significant.

Results and brief discussion

```
##
## Call:
## lm(formula = df2$CO2_emissions ~ I(df2$GDP_per_capita^2))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.8873 -2.4968 -0.8194  2.0520 11.4408
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)              3.583e+00  4.902e-01   7.308 7.45e-10 ***
## I(df2$GDP_per_capita^2) 1.497e-09  3.156e-10   4.742 1.35e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.425 on 60 degrees of freedom
## Multiple R-squared:  0.2726, Adjusted R-squared:  0.2605
## F-statistic: 22.48 on 1 and 60 DF,  p-value: 1.348e-05
```





Question 2.2

Synthetic description of assumptions, methods, and algorithms

Assumptions : X1,. . . ,Xn~p We want to predict the location of Xn+1 following the same distribution p. In particular, we want to build a confidence region for this new point. We choose a Non-conformity

measure. Then, from this non conformity measure, we can now rank the observations of an augmented vector (x1,. . . ,xn,xtest) and so determine the p-value corresponding to xtest : pval=R_{n+1}/(n+1). From that we can deduce prediction bands or prediction region with confidence 1-alpha by taking C(X1,. . . ,Xn)={xn+1 in Rˆp st r_{n+1}<=ceil((1-alpha)*(n+1))}

Results and brief discussion
123

```
##
## Call:
## lm(formula = df2_1$CO2_emissions ~ I(df2_1$GDP_per_capita))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.0973 -1.9567 -0.9054  1.3710 11.5448
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)              2.4265277  0.4925469   4.926 6.91e-06 ***
## I(df2_1$GDP_per_capita) 0.0001296  0.0000184   7.040 2.14e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.972 on 60 degrees of freedom
## Multiple R-squared:  0.4524, Adjusted R-squared:  0.4433
## F-statistic: 49.56 on 1 and 60 DF,  p-value: 2.136e-09
```

Question 2.3

Synthetic description of assumptions, methods, and algorithms
By fitting the GAM and looking at the p-values, we see that only the intercept and the second coefficient from the bs are statistically significant (others have higher than 0.05 pvalues). The coefficients represent the estimated effects of the basis splines for the variable GDP. The asterisks in the p-values indicate statistical significance. The pvalues < 0.05 suggest that the related coefficients contribute significantly to the model. The others suggest that these terms may not be contributing significantly to the model.

Results and brief discussion
123

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## Gini ~ bs(GDP_per_capita, degree = 2, knots = median(3000))
##
## Parametric coefficients:
##                                                    Estimate Std. Error
## (Intercept)                                        -0.41489    0.12685
## bs(GDP_per_capita, degree = 2, knots = median(3000))1  0.07025    0.16043
## bs(GDP_per_capita, degree = 2, knots = median(3000))2 -0.56672    0.24313
## bs(GDP_per_capita, degree = 2, knots = median(3000))3 -0.52120    0.30256
##                                                    t value Pr(>|t|)
## (Intercept)                                         -3.271  0.00181 **
## bs(GDP_per_capita, degree = 2, knots = median(3000))1   0.438  0.66312
## bs(GDP_per_capita, degree = 2, knots = median(3000))2  -2.331  0.02326 *
```

```
## bs(GDP_per_capita, degree = 2, knots = median(3000))3  -1.723  0.09028 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##
## R-sq.(adj) =  0.213   Deviance explained = 25.2%
## GCV = 0.11618  Scale est. = 0.10868   n = 62
```

Question 2.4

Synthetic description of assumptions, methods, and algorithms
Bootstrap Assumptions : S1 = (X1,. . . ,Xn)~ iid 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
123