Excercise 1 - Multivariate Analysis

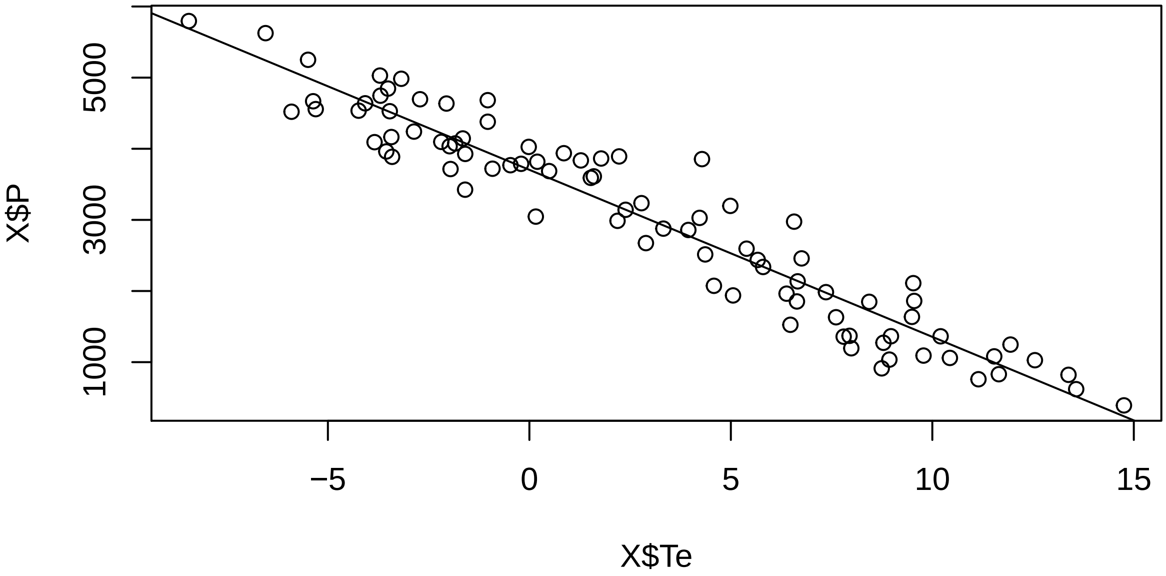
Marco Hernandez Velasco

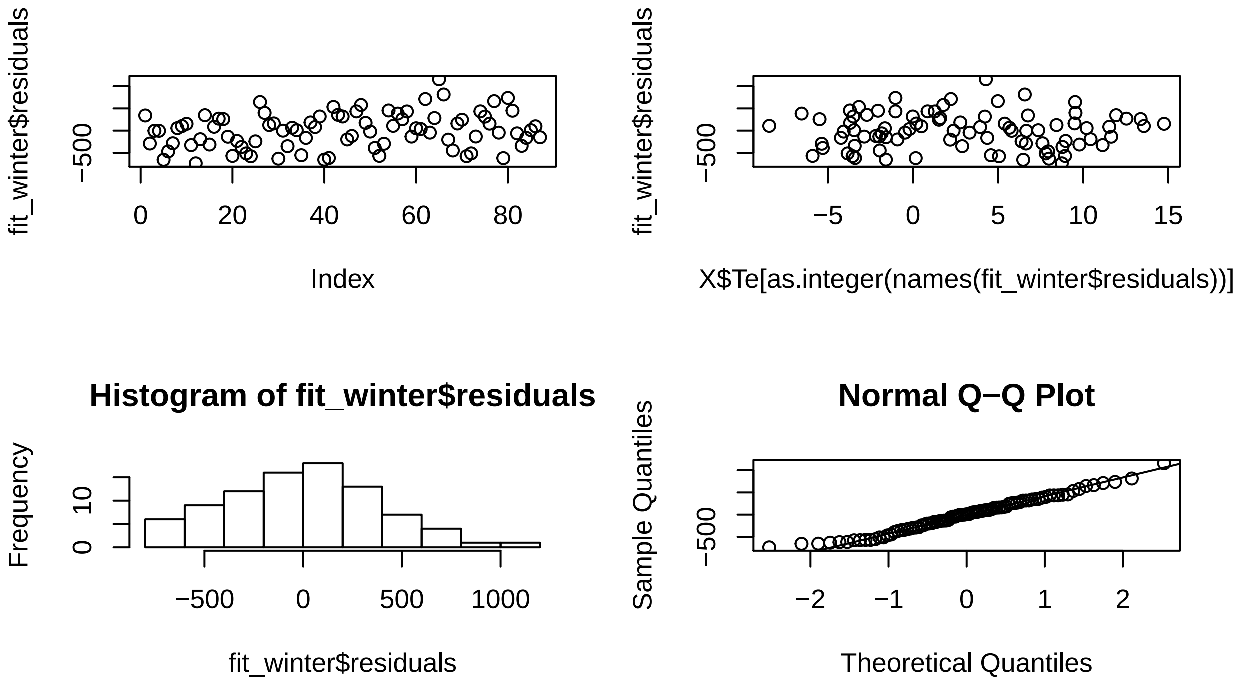
August 2018

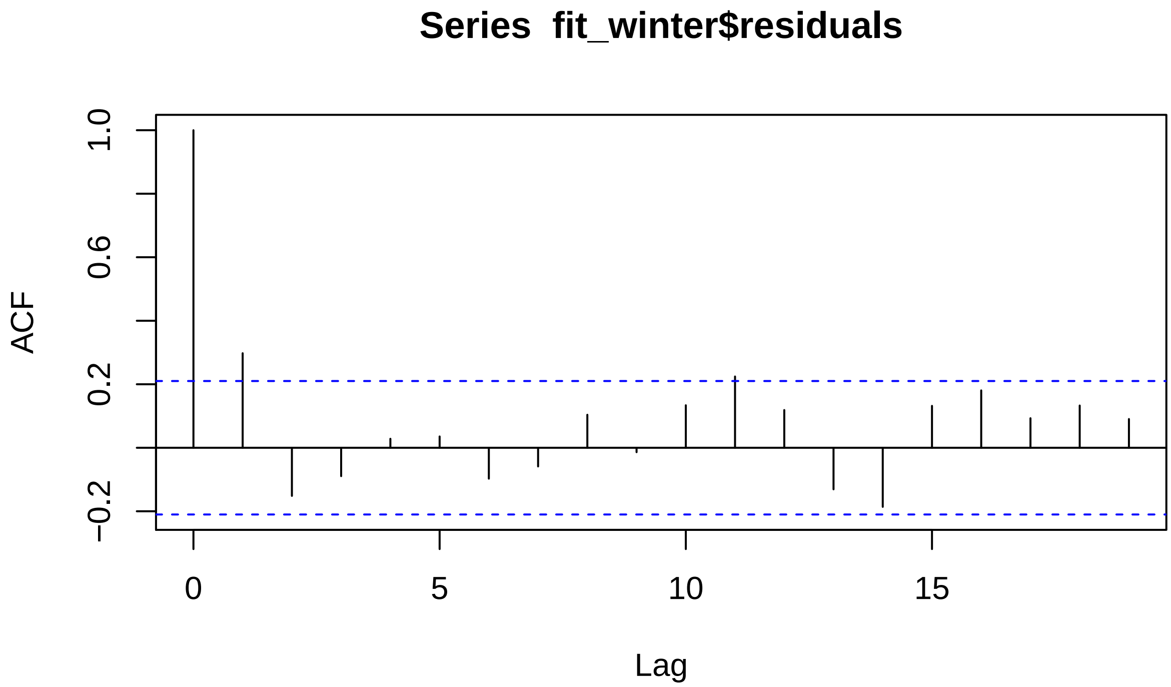
Q1 - Read data and lm in R

Read the data into a dataframe. The data consists of hourly average values

Only winter period



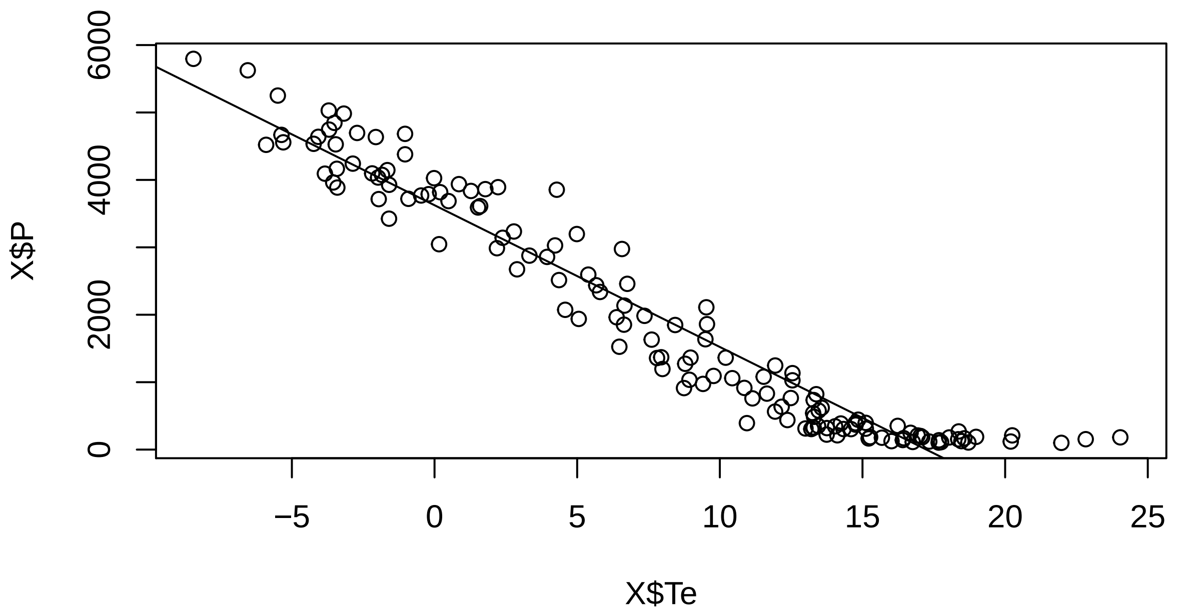


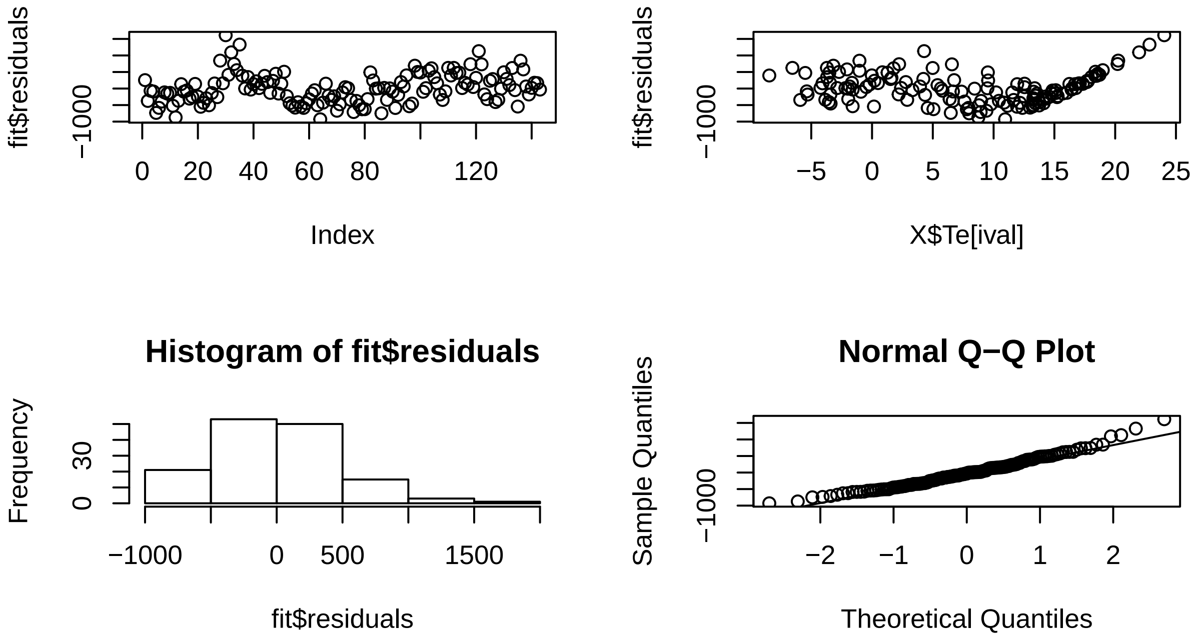


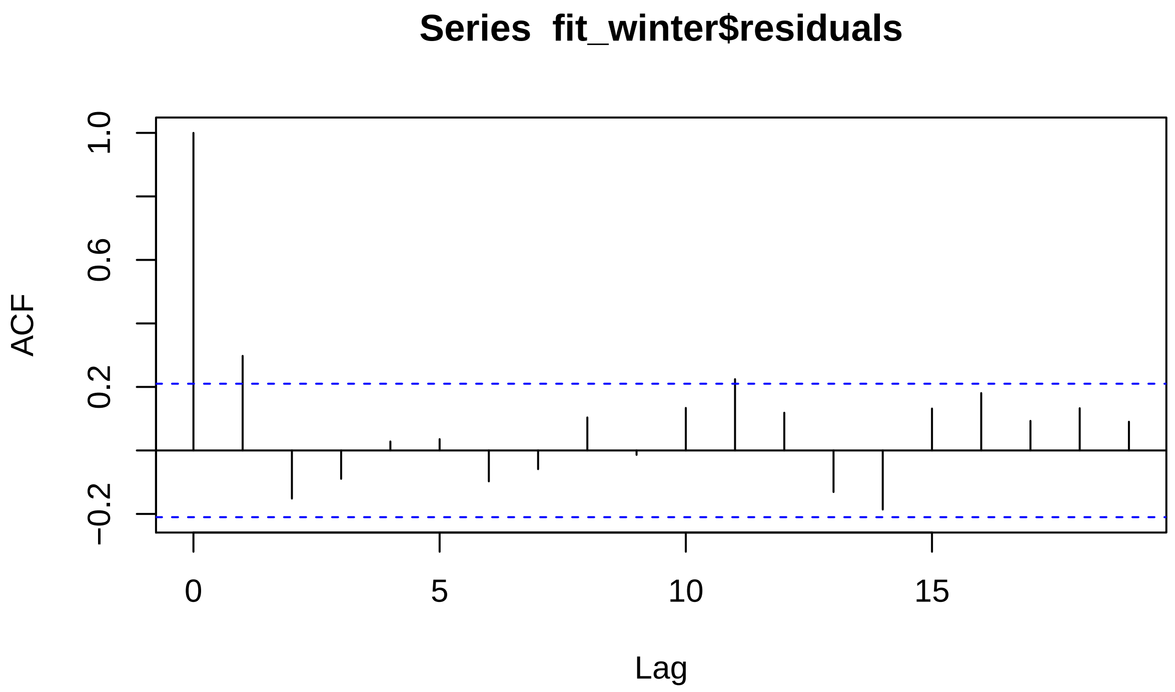
How does a linear regression model fit the data when using the winter period only?

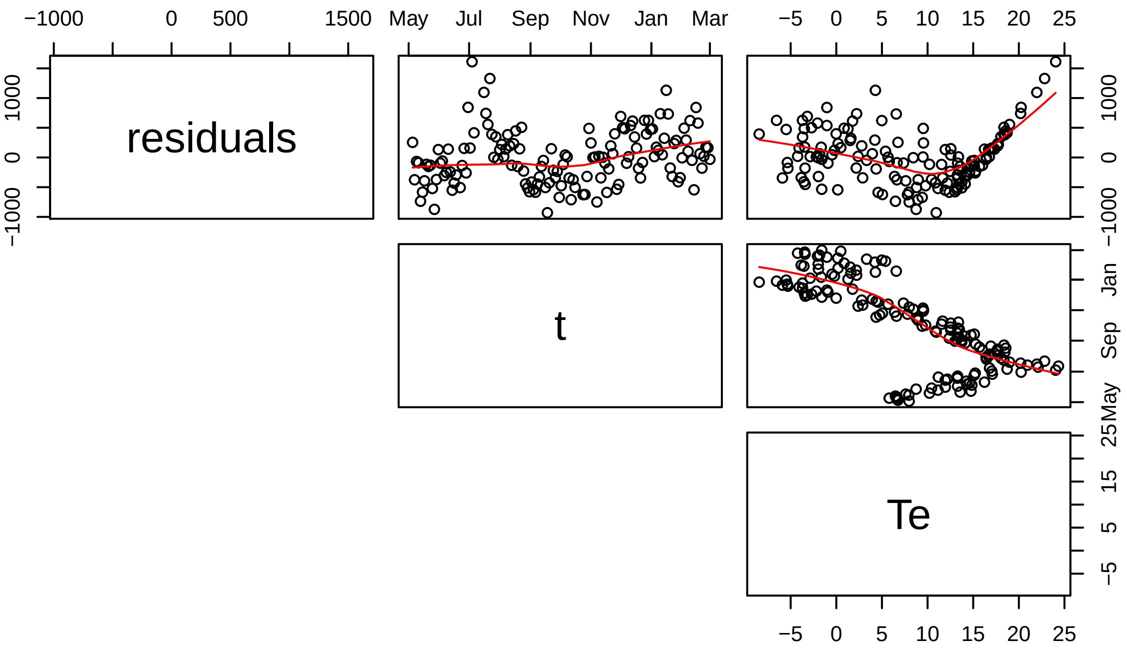
The linear regression fits the data well enough. Looking from the residual plots, the residuals are independent and normally distributed. Also the ACF shows no auto significant correlation in the residuals so the linear *regression is a good fit.*

Both summer and winter





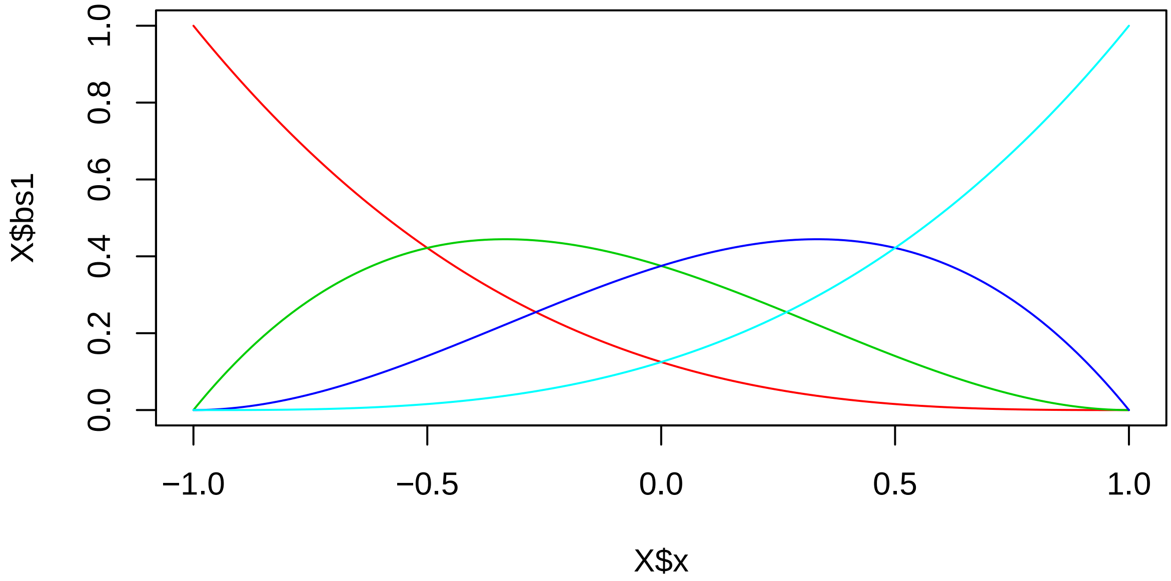




How does a linear regression model fit the data when using all the data? *Including all the data reduces the fit of the model. The residuals are not anymore i.i.d. specially at higher temperatures.*

Q2 - Base splines intro

The aim of this question is to give you an idea of how the base splines behave and how parameters can change them. Play around with the **bs** function.

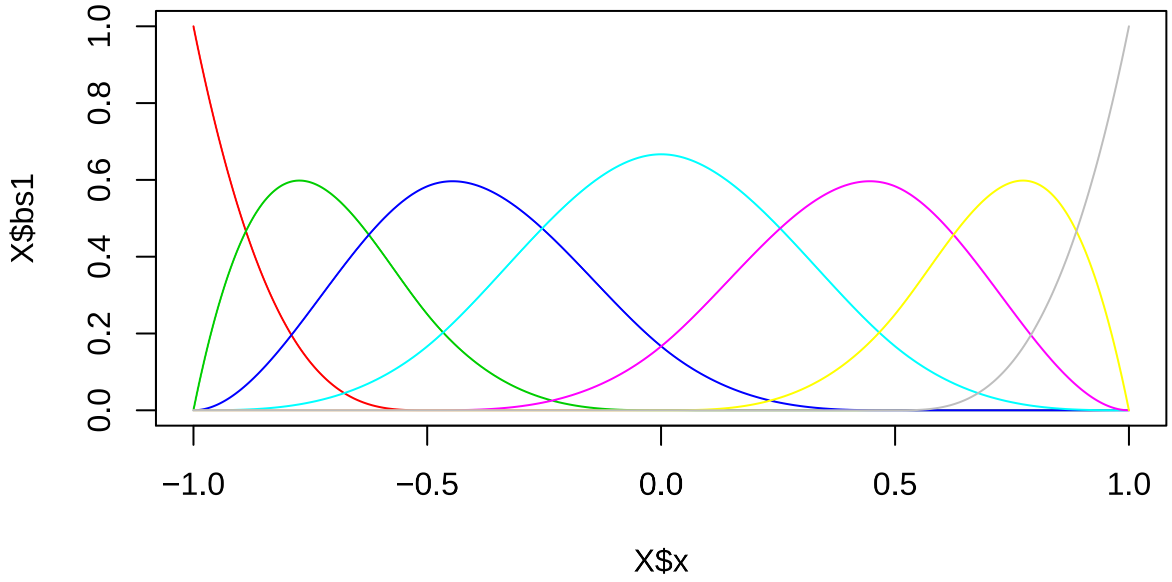


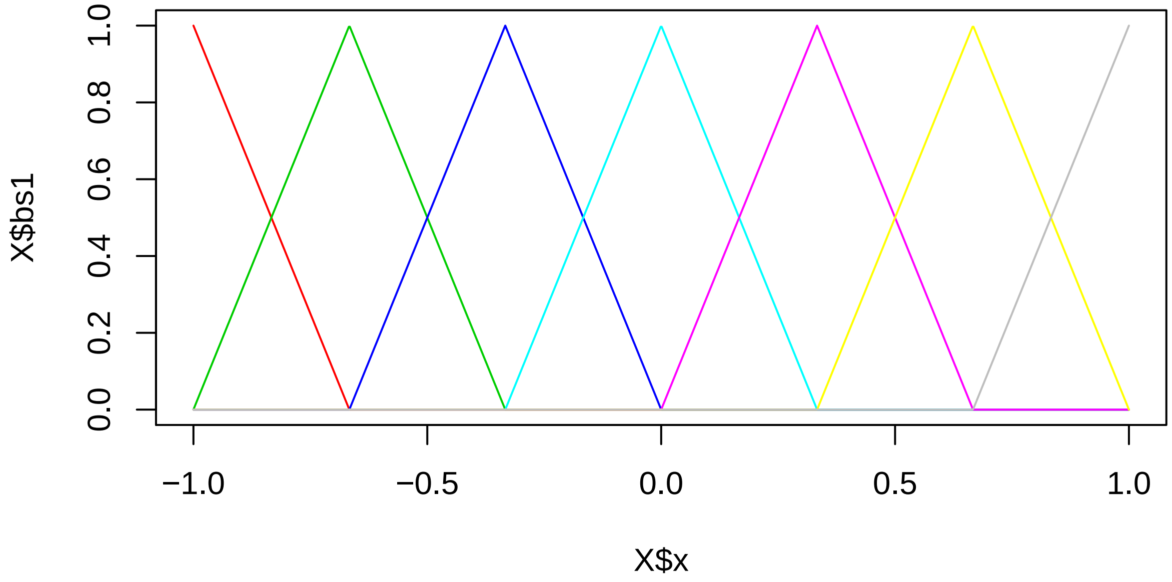
Try to vary the degrees of freedom (df). What happens to the base splines generated?

A higher degree of freedom generates more base splines.

Try to vary the degree (degree). What happens to the base splines generated?

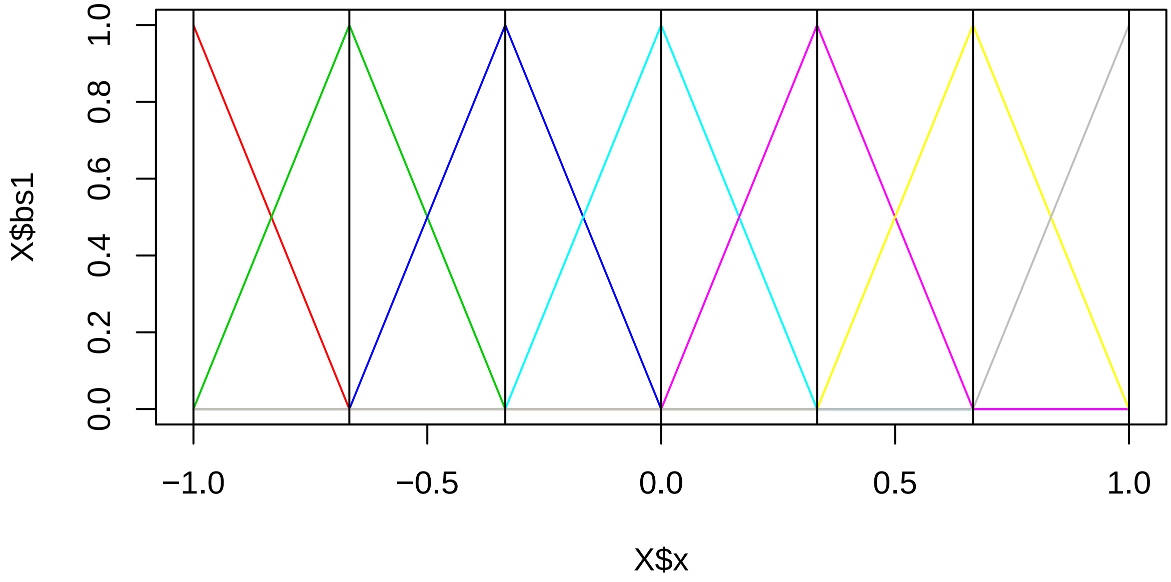
A higher degree changes the "shape" of the splines. Higher degree make polynomial or "curve" lines.

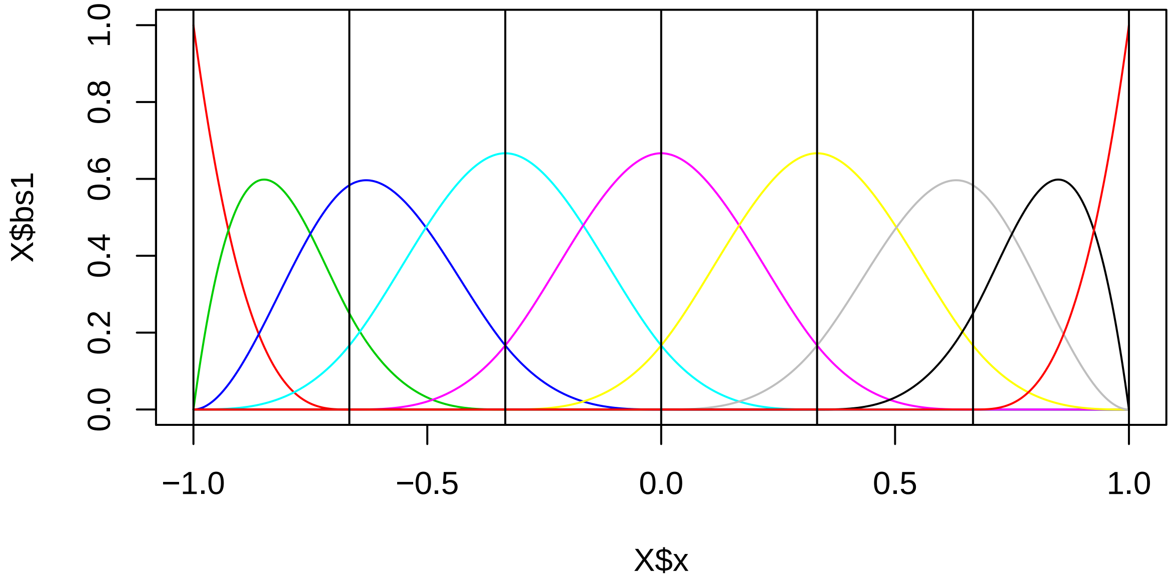




Give the knot points directly (using the knots argument). Give the knots as the quantiles of x, what happens when degree = 1 and what happens when degree = 3?

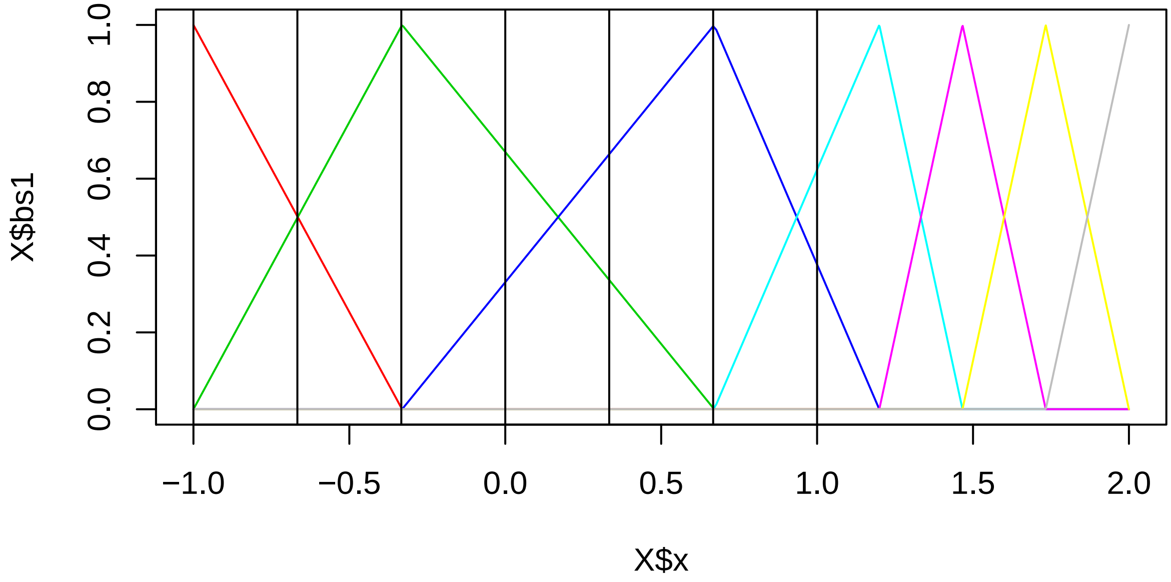
Giving the knots directly we can change manually the breaking points. When defining the quantiles, we assign *the knots depending on the amount of data. The degree changes the shape of the splines.*





Try with some non-equidistant x sequence, such that the quantiles are not equidistant. What happens with the base splines?

The base splines are then divided according to the definitions of the knots, with more splines where there is more data.

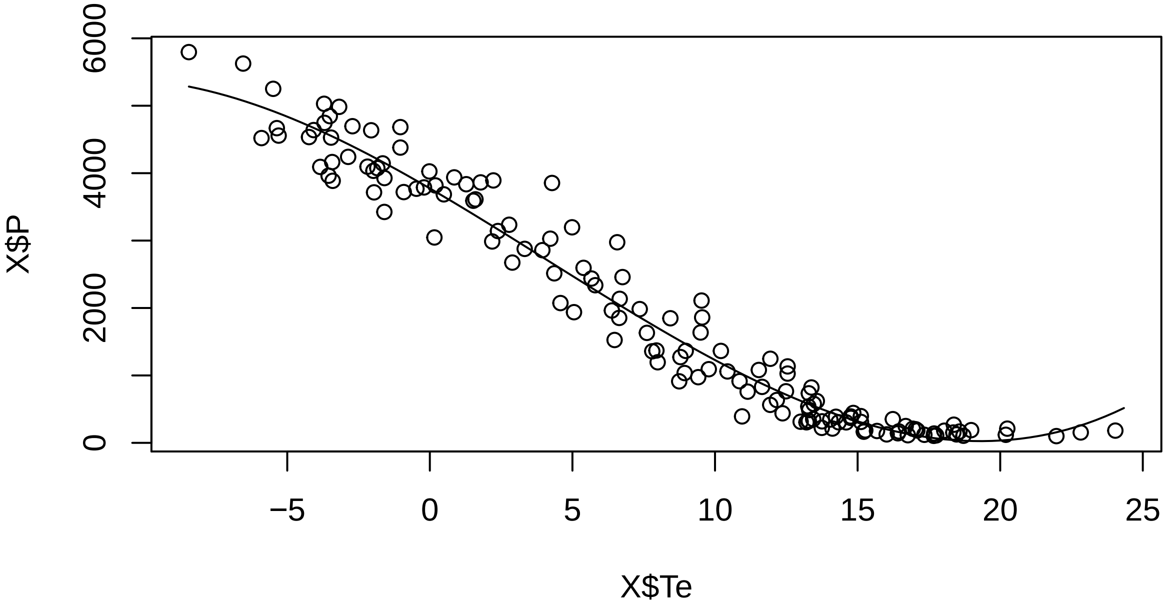


Q3 - Base splines model

Now we want to calculate the base splines as a function of the external temperature. And then use them as input to a linear regression model. In this way, it becomes a non-linear model The characteristics of *f()* depend on how the base splines are generated, so it does not have any direct parameters, therefore such a model is called a non-parametric model.

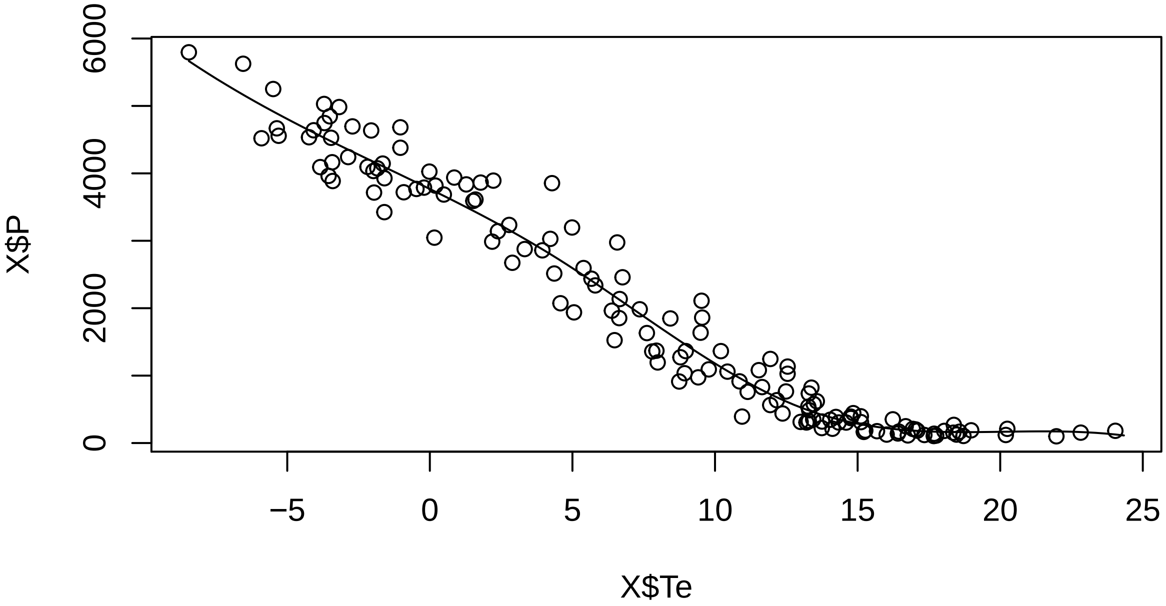
Try to fit a model, is it linear or how is it shaped?

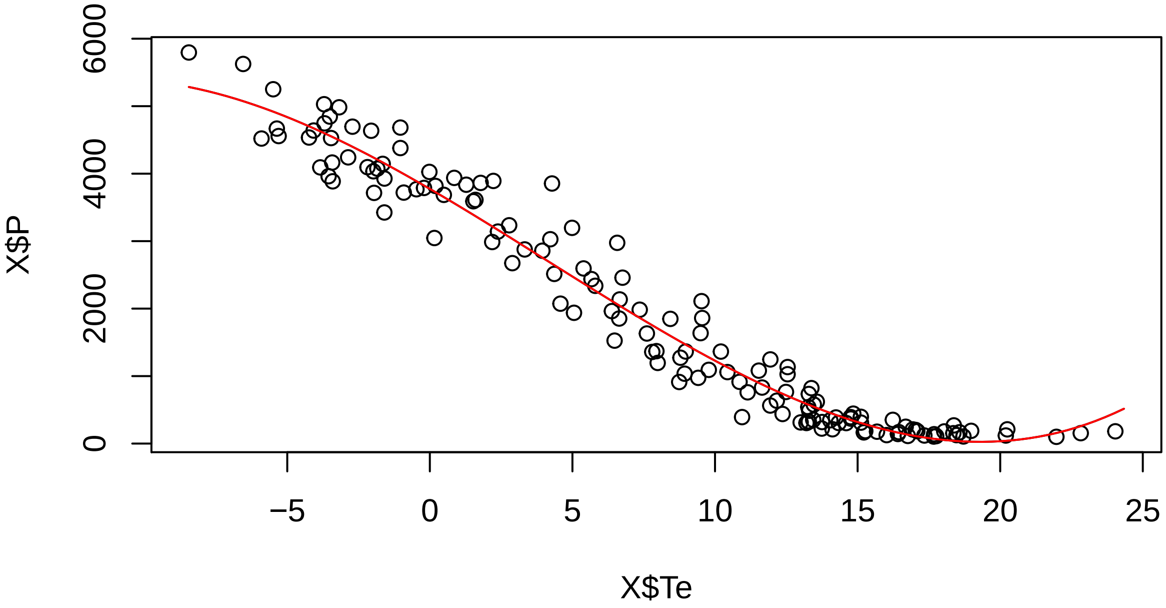
The fitted model is not linear, but a polynomial fit. However there are some fit problems at the extremes.



Try to change the degrees of freedom (df), what happens?

When increasing the degrees of freedom the line fits better in the extreme temperatures. With a df=5 there is a good fit without overfitting.


6000
4000
2000
0
X$P
−5
0
5
10
15
20
25
X$Te


6000
4000
2000
0
X$P
−5
0
5
10
15
20
25
X$Te

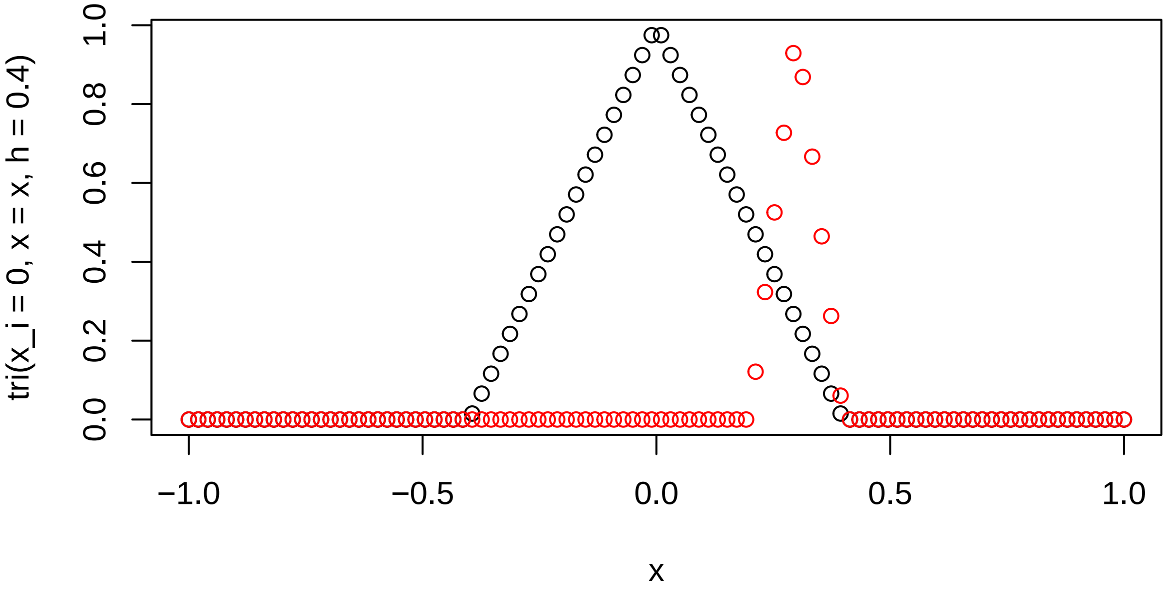
Q4 - Kernel functions

Another way to make non-parametric models is to use locally weighted regression. To do this we need a kernel function.

Try to calculate and plot the triangular kernel and play around with the parameters.

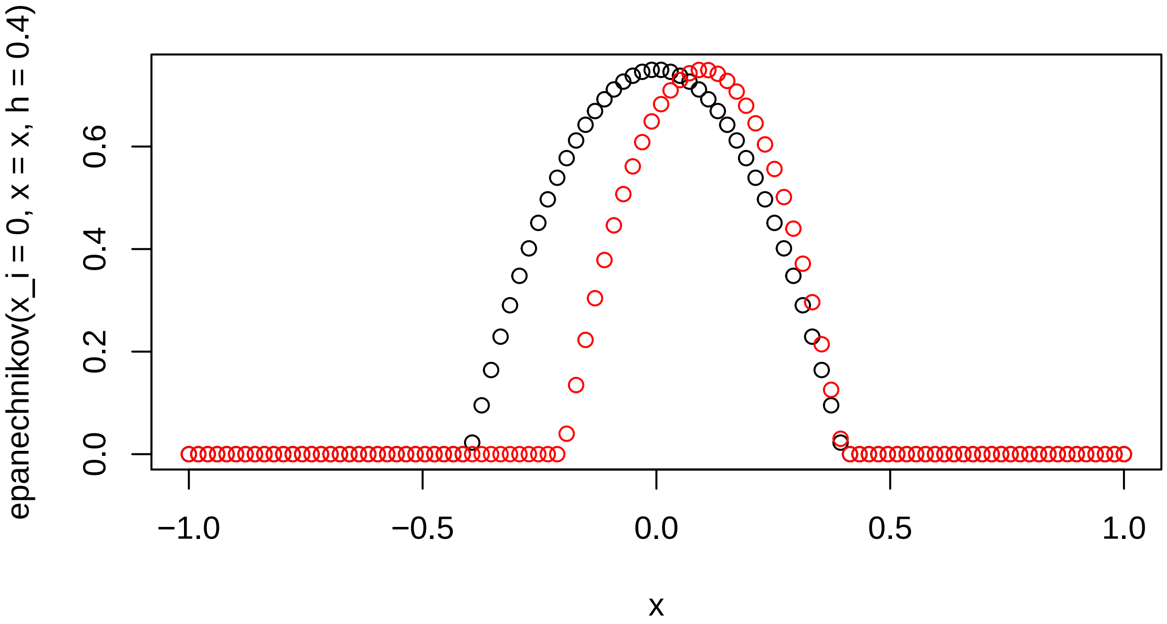
How do they affect the shape of the kernel?

Changing the x_i moves the center or peak of the kernel. Adjusting the h defines the width of the base.



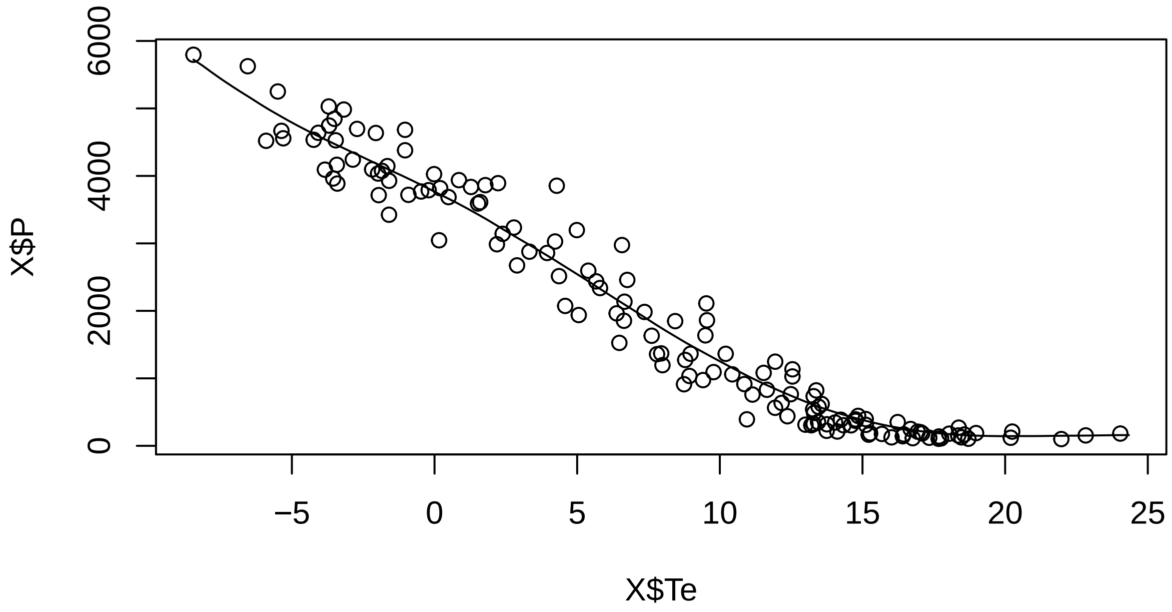
Try to calculate and plot the Epanechnikov kernel and play around with the parameters. How do they affect the shape of the kernel?

Similarly to the triangular kernel, changing the x_i in the Epanechnikov kernel function moves the center or peak of the kernel. Adjusting the h defines the width of the base.


epanechnikov(x_i = 0, x = x, h = 0.4)
0.0
0.2
0.4
0.6
−1.0
−0.5
0.0
0.5
1.0
x

Q5 - Locally weighted model with a kernel function

Fit a locally weighted model for a single point and predict the heat load. Fit for a sequence and make the plot of the function. Try to change the bandwidth h.



How does changing the h change the estimated function between Te and P?
A smaller h reduces the bandwith for the kernel function which takes less data points for the model fit. However, a very small h causes overfitting and if it comes to a part where the distance between points is larger than the h, then the line is fitted to 0.

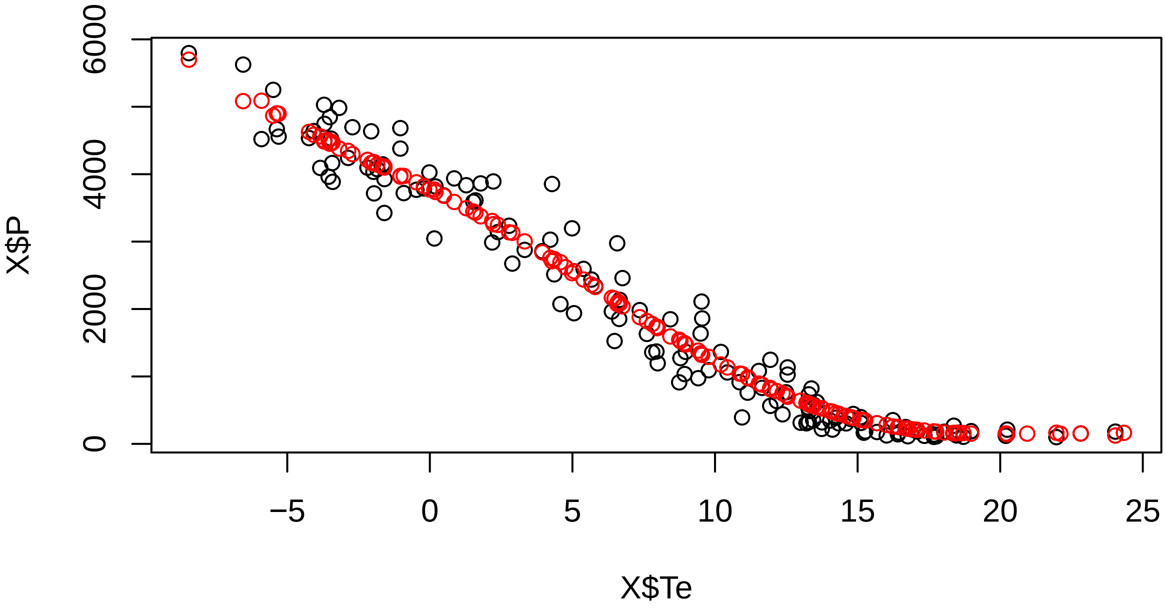
What should the bandwidth h be?
A bandwith between 6 and 7 gives a good fit. To find the best option a cross-validation has to be conducted to find the optimal.

Q6 - Locally weighted model with a kernel function

Now we have a challenge of finding the optimal values for the smoothing parameters, either the bandwidth **h** in the kernel or the degrees of freedom for the base splines. If the model is over-fitted it varies too much (the function is too flexible and bends around too much), and on the other hand if it is under-fitted, then it is not "bending" and adapting enough to the observations.

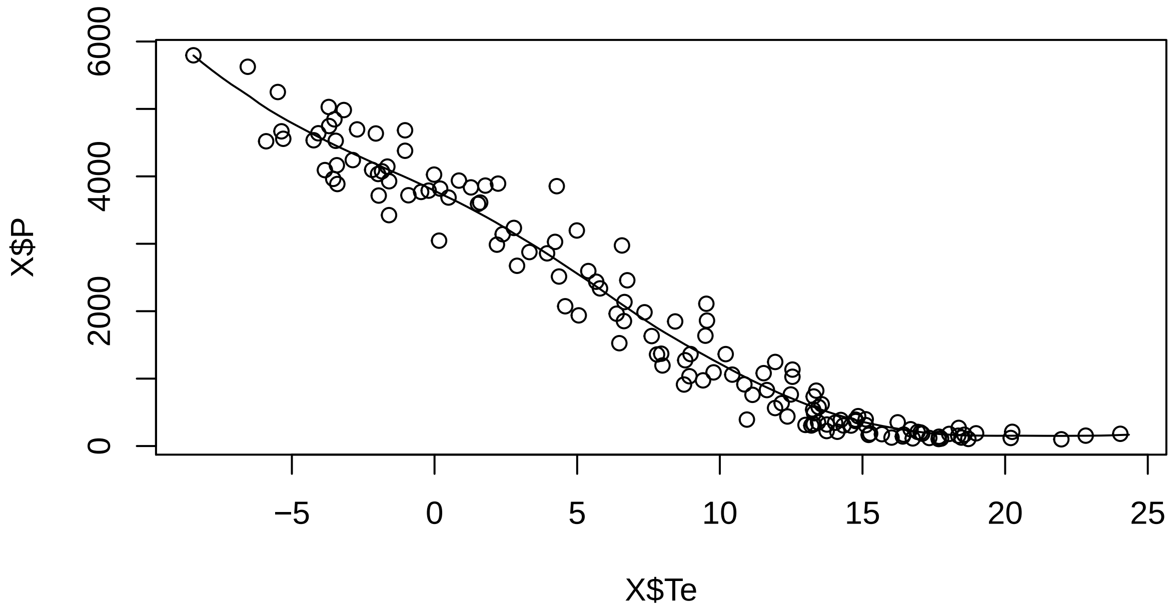
One approach is to do a cross-validation optimization of a score function. In the case of estimating the (conditional) mean value, the score function should almost always be the **Root Mean Squared Error (RMSE).**

Do this for all the observations and then calculate the score function using the predictions. In this way we can find the right balance between under- and over-fitting. Carry out leave-one-out cross validation. Try to change the bandwidth.



What happens to the RMSE score?
Depending on the bandwith h, the RMSE score of the model changes. What should the bandwidth h be set to?
A h = 6 gives the lowest RMSE.

Of course we cannot use our time doing manually optimization, so use an optimizer to optimize the bandwidth.



Does the result look reasonable for the locally weighted model? *Using the optimized h = 5.51, the model fits the data properly.*

Use leave-one-out cross validation for the base spline model.

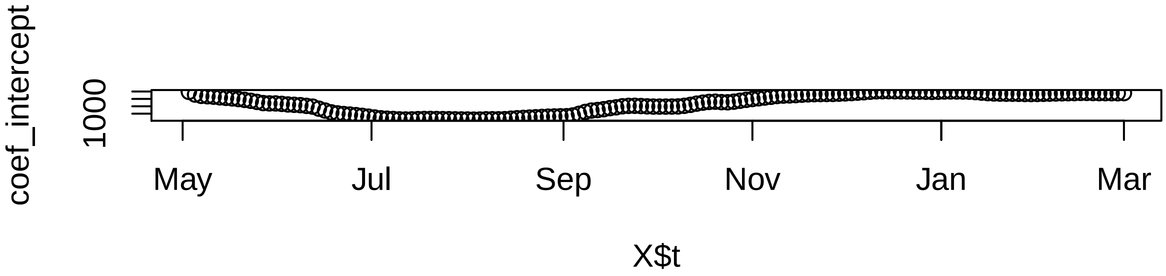
Does the result look reasonable? *yes, the df = 6 gives a good fit to the data without overfitting and low boundary bias. The optimized RMSE is 325 which is also lower than the 329 that the Locally weighted model with h = 5.51 had*

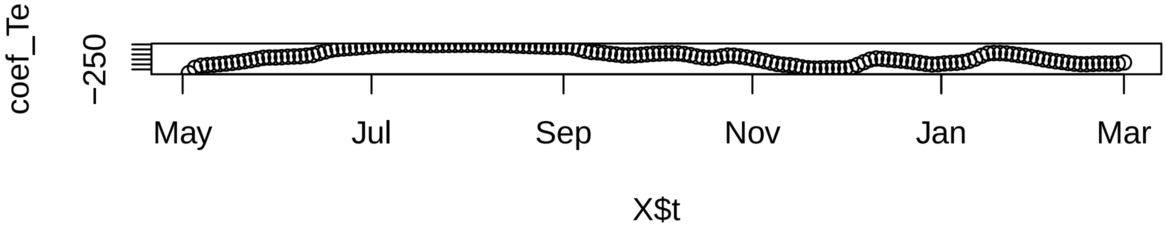
For the base spline models a model selection criteria, such as AIC or BIC can be used. Try that for the base spline model and compare. Do you get the same results?
The results of the AIC and BIC don't come the same result. AIC find a df = 6, while BIC optimizes to df = 5.

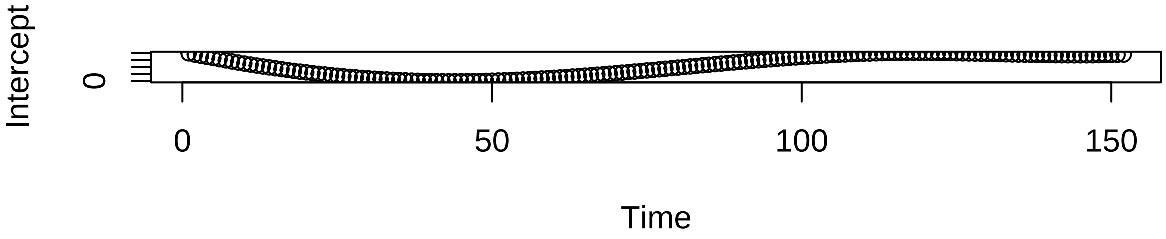
Q7 - Semi- and conditional parametric models

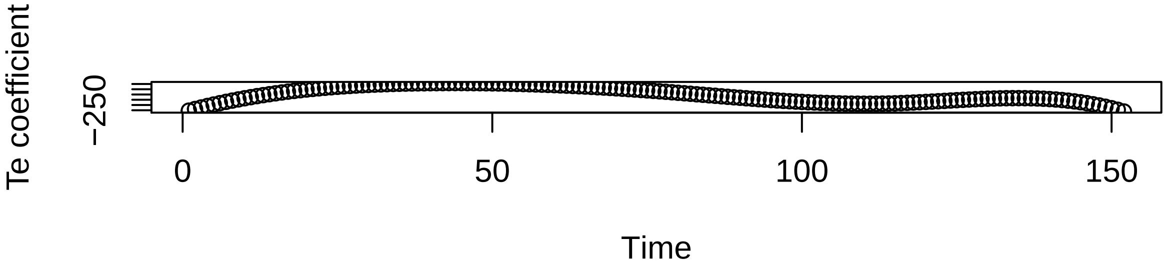
Until now, we have calculated the weights and base splines using the input to the model. What if we used another variable, but still fitted the same model?

Lets try to calculate the weights using the time t. By doing that we actually allow the coefficients in the model to change as a function of time. We would usually in this case add a t to the parameters.









What happens with the coefficients during the summer? *answer*

Can you explain the result in relation to how the heating system of the building isoperating? *answer*

Q8 - Semi- and conditional parametric models

...

References

JJ Allaire, Yihui Xie, Jonathan McPherson, Javier Luraschi, Kevin Ushey, Aron Atkins, Hadley Wickham, Joe Cheng, and Winston Chang. *rmarkdown: Dynamic Documents for R*, 2018. URL https://CRAN.R-project.org/package=rmarkdown. R package version 1.10.

R. Kabacoff. *R in Action: Data Analysis and Graphics with R*. Data, statistics, programming. Manning, 2015. ISBN 9781617291388.

Henrik Madsen. *Time series analysis*. Chapman and Hall/CRC, 2007.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2018. URL https://www.R-project.org/.

Hadley Wickham. *tidyverse: Easily Install and Load the 'Tidyverse'*, 2017. URL https://CRAN.R-project.org/package=tidyverse. R package version 1.2.1.

Yihui Xie. *knitr: A General-Purpose Package for Dynamic Report Generation in R*, 2018. URL https://CRAN.R-project.org/package=knitr. R package version 1.20.