Statistical Modeling 2
Exercise 3

March 7, 2017

Basic concepts

$$\begin{aligned}
\text{MSE}(\hat{f}, f) &= E\{[f(x) - \hat{f}(x)]^2\} \\
&= E[f(x)^2 + \hat{f}(x)^2 - 2f(x)\hat{f}(x)] \\
&= f(x)^2 + E[\hat{f}(x)^2] - E[\hat{f}(x)]^2 + E[\hat{f}(x)]^2 - 2f(x)E[\hat{f}(x)] \\
&= \{E[\hat{f}(x)^2] - E[\hat{f}(x)]^2\} + \{f(x)^2 + E[\hat{f}(x)]^2 - 2f(x)E[\hat{f}(x)]\} \\
&= V + B^2
\end{aligned}$$

Curve fitting by linear smoothing

A

In the least square estimate, we minimize:

$$L = \sum_{i=1}^{n} (\hat{\beta} x_i - y_i)^2$$

We take the derivative and set to 0:

$$\begin{aligned}
\partial L / \partial \hat{\beta} = \sum_{i=1}^{n} 2(\hat{\beta} x_i - y_i) x_i = 0 \\
\implies \hat{\beta}(\sum_{i=1}^{n} x_i^2) - \sum_{i=1}^{n} y_i x_i = 0 \\
\implies \hat{\beta} = \frac{\sum_{i=1}^{n} y_i x_i}{\sum_{i=1}^{n} x_i^2}
\end{aligned}$$

We have that the prediction is:

$$\begin{aligned}\hat{f}(x^*) &= \hat{\beta}x^* \\ &= \frac{\sum_{i=1}^n x_i x^* y_i}{\sum_{i=1}^n x_i^2} \\ &= \sum_{i=1}^n w(x_i, x^*) y_i\end{aligned}$$

where

$$w(x_i, x^*) = \frac{x_i x^*}{\sum_{i=1}^n x_i^2}$$

This weight function weights examples by the product with predictors, using all the points in the dataset. The K nearest function weights the K nearest neighbors equally and ignore all other points.

B

Code: kernel.r
I use the function sin from -5 to 5 with 100 samples and try three bandwidths: $h = 0.01, h = 0.1, h = 1$.

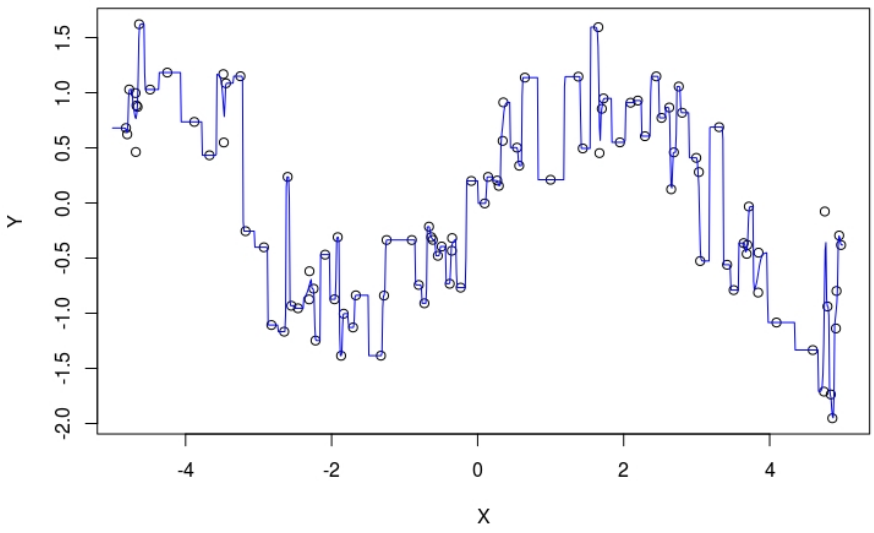

Figure 1: Bandwidth $h = 0.01$

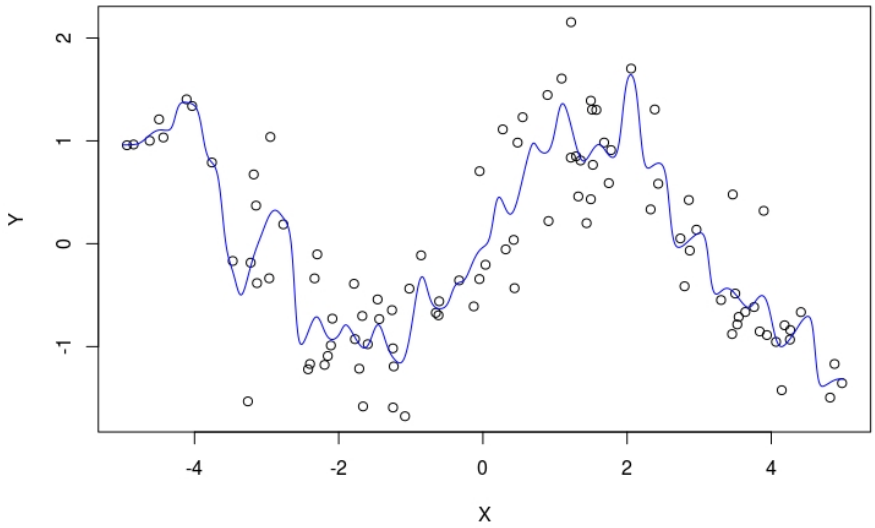

Figure 2: Bandwidth $h = 0.1$

The bandwidth h affects the smoothness of the prediction line. With smaller h, more weight is given to neighbor points and the line fluctuates. With larger h, more weight is given to points that are further away, resulting in a smoother line.

Cross Validation

B

For the smooth function, I use x^2; for the wiggly function, I use $sin(10x)$. 'Not so noisy' has a Normal noise of 0.05, 'noisy' has a Normal noise of 0.25. I generate 100 points in train and 100 points in test. The bandwidth h is varied from 0.01 to 0.20. For the smooth function with less noise, the RMSE is stable across different settings of h. But for other cases, using the right h seems to improve RMSE significantly.

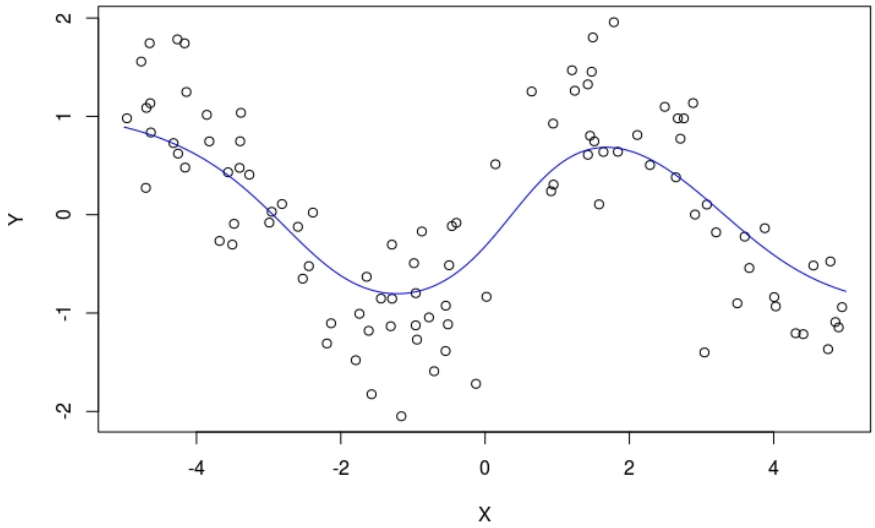

Figure 3: Bandwidth $h = 1$

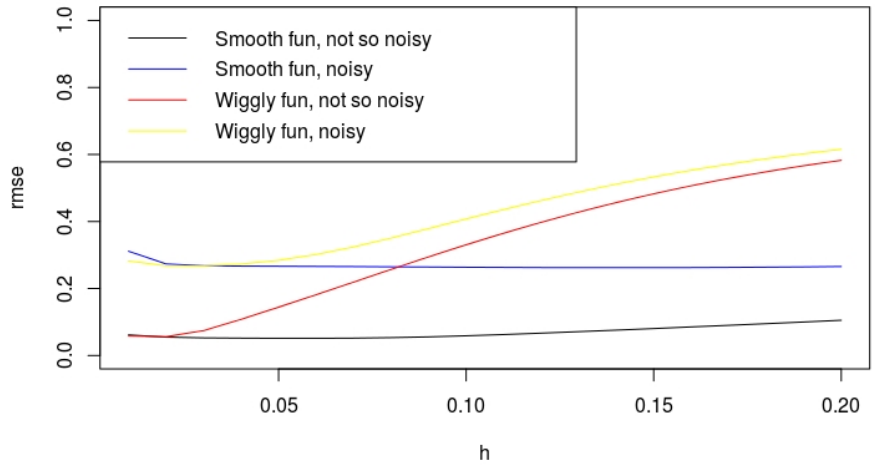

Figure 4: Cross Validation

Local polynomial regression

A

Let R_x be a $n \times D$ matrix whose (i, j) entry is $(x_i - x)^{j-1}$. We have that

$$g_x(x_i; a) = R_{x,i}^T a$$

where a is the column vector of coefficients of the polynomial g and $R_{x,i}$ is the row i of R_x. We want to minimize:

$$\begin{aligned}
&\sum_{i=1}^{n} w_i \{y_i - g_x(x_i; a)\}^2 \\
&= \sum_{i=1}^{n} w_i \{y_i - R_{x,i}^T a\}^2 \\
&= \sum_{i=1}^{n} \frac{1}{h} K\left(\frac{x_i - x}{h}\right) \{y_i - R_{x,i}^T a\}^2 \\
&= (R_x a - y)^T K_x (R_x a - y) \\
&= a^T R_x^T K_x R_x a - a^T R_x^T K_x y - y^T K_x R_x a + y^T y \\
&= F_x
\end{aligned}$$

where K_x is the diagonal matrix whose (i, i) entry is $\frac{1}{h} K(\frac{x_i - x}{h})$. We take the derivative and set to zero:

$$\begin{aligned}
\partial F_x / \partial a = 2R_x^T K_x R_x a - 2R_x^T K_x y = 0 \\
\implies a = (R_x^T K_x R_x)^{-1} R_x^T K_x y
\end{aligned}$$

The estimate at x is $\hat{f}(x) = a_0 = S_x y$, where

$$S_x = e_1 (R_x^T K_x R_x)^{-1} R_x^T K_x$$

B

For $D = 1$, we have:

$$\begin{aligned}
a &= \begin{pmatrix} \sum_i K_{x,i} & \sum_i K_{x,i}(x_i - x) \\ \sum_i K_{x,i}(x_i - x) & \sum_i K_{x,i}(x_i - x)^2 \end{pmatrix}^{-1} \begin{pmatrix} K_{x,1} & \dots & K_{x,n} \\ K_{x,1}(x_1 - x) & \dots & K_{x,n}(x_n - x) \end{pmatrix} y \\
&= \frac{1}{C} \begin{pmatrix} \sum_i K_{x,i}(x_i - x)^2 & -\sum_i K_{x,i}(x_i - x) \\ -\sum_i K_{x,i}(x_i - x) & \sum_i K_{x,i} \end{pmatrix} \begin{pmatrix} \sum_i K_{x,i} y_i \\ \sum_i K_{x,i}(x_i - x) y_i \end{pmatrix} \\
&= \frac{1}{C} \begin{pmatrix} s_2(x) & -s_1(x) \\ -s_1(x) & s_0(x) \end{pmatrix}
\end{aligned}$$

where $C = s_0(x)s_2(x) - s_1(x)^2$. We then have:

$$\begin{aligned} a_0 &= \frac{1}{C}\left[s_2(x)\sum_i K_{x,i}y_i - s_1(x)\sum_i K_{x,i}(x_i - x)y_i\right] \\ &= \frac{1}{C}\left\{\sum_i K_{x,i}[s_2(x) - (x_i - x)s_1(x)]y_i\right\} \end{aligned}$$

and

$$\begin{aligned} C &= \sum_i K_{x,i}s_2(x) - \sum_i K_{x,i}(x_i - x)s_1(x) \\ &= \sum_i K_{x,i}\left\{s_2(x) - (x_i - x)s_1(x)\right\} \end{aligned}$$

Let $w_i = K_{x,i}\left\{s_2(x) - (x_i - x)s_1(x)\right\}$, we have:

$$\hat{f}(x) = \frac{\sum_i w_i y_i}{w_i}$$

C

The mean is

$$E\hat{f}(x) = E[S_x y] = S_x f(X) = \sum_i S_{x,i} f(x_i)$$

and the variance is:

$$\begin{aligned} \mathrm{var}\hat{f}(x) = \mathrm{var}[S_x y] &= \mathrm{var}\left(\sum_i S_{x,i}(f(x_i) + e_i)\right) \\ &= \sum_i S_{x,i}^2 \mathrm{var}(e_i) \\ &= \sum_i S_{x,i}^2 \sigma^2 \end{aligned}$$

D

E

Code: local_linear.r

F

The variance is larger for temperature 20 to 60 and smaller for temperature below 20 or above 60. By taking log on the Daily gasbill, the hemoskedastic assumption is more reasonable.

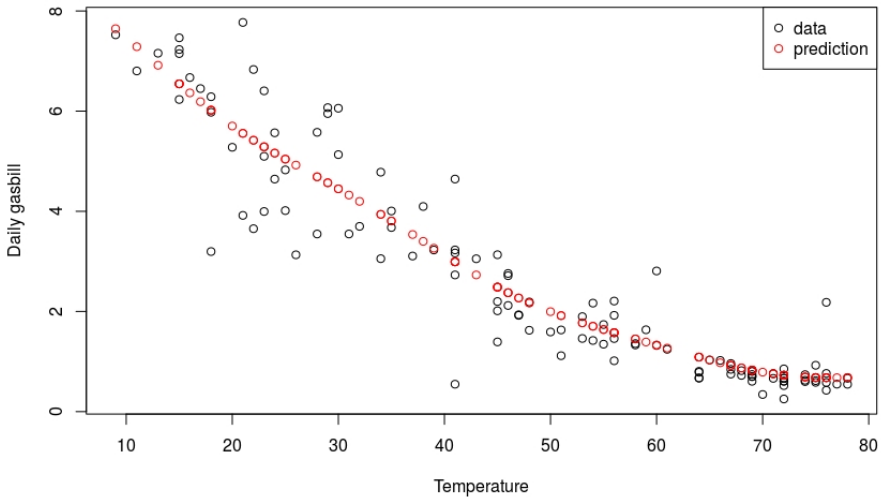

Figure 5: The data and prediction by local linear estimator

G

Gaussian processes

A

With τ_2^2 small, the parameter τ_1^2 controls the magnitude of the function while the parameter b controls the smoothness.

B

We have that the random variables $f(x^*), f(x_1), \ldots, f(x_N)$ are Normal with mean $m(x^*), m(x_1), \ldots, m(x_N)$ and variance

$$\begin{aligned} C &= \begin{pmatrix} C(x^*, x^*) & \ldots & C(x^*, x_N) \\ \ldots & \ldots & \ldots \\ C(x_N, x^*) & \ldots & C(x_N, x_N) \end{pmatrix} \\ &= \begin{pmatrix} C_* & C_{*x} \\ C_{*x}^T & C_{xx} \end{pmatrix} \end{aligned}$$

Now given $f(x_1), \ldots, f(x_N)$, we have that $f(x^*)$ is Normal with mean

$$\mu_{*|1..N} = m(x^*) + C_{*x} C_{xx}^{-1} (f(x_{1..N}) - m(x_{1..N}))$$

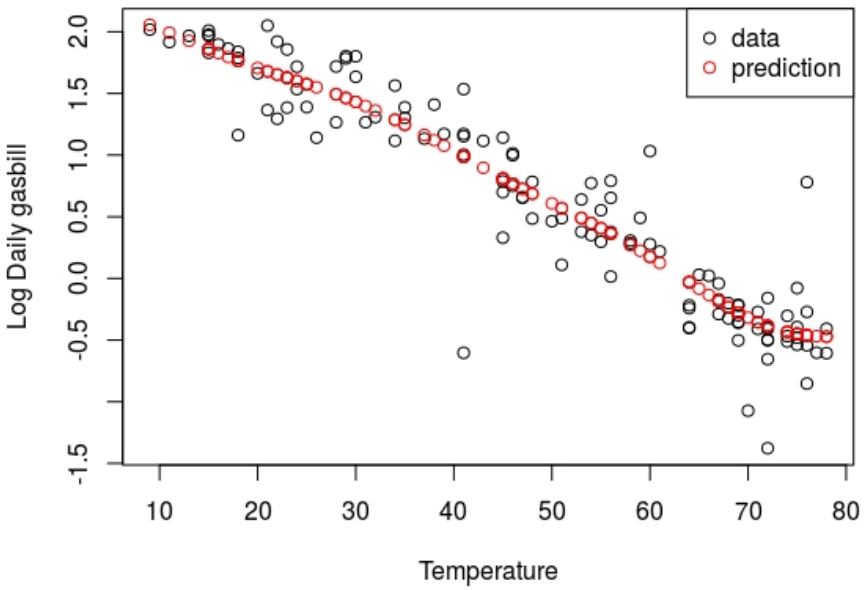

Figure 6: The data and prediction by local linear estimator after a log transformation on the daily gasbill.

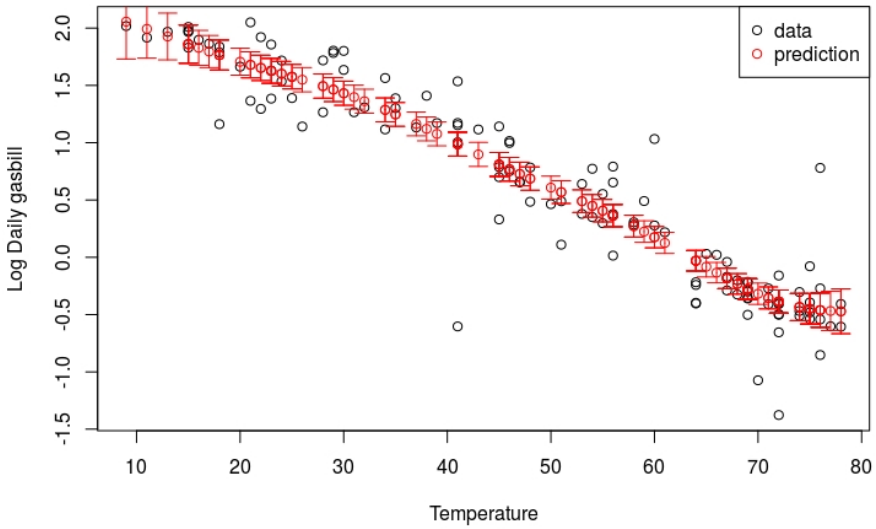

Figure 7: The data and prediction by local linear estimator after a log transformation on the daily gasbill, with the 95 % confidence interval.

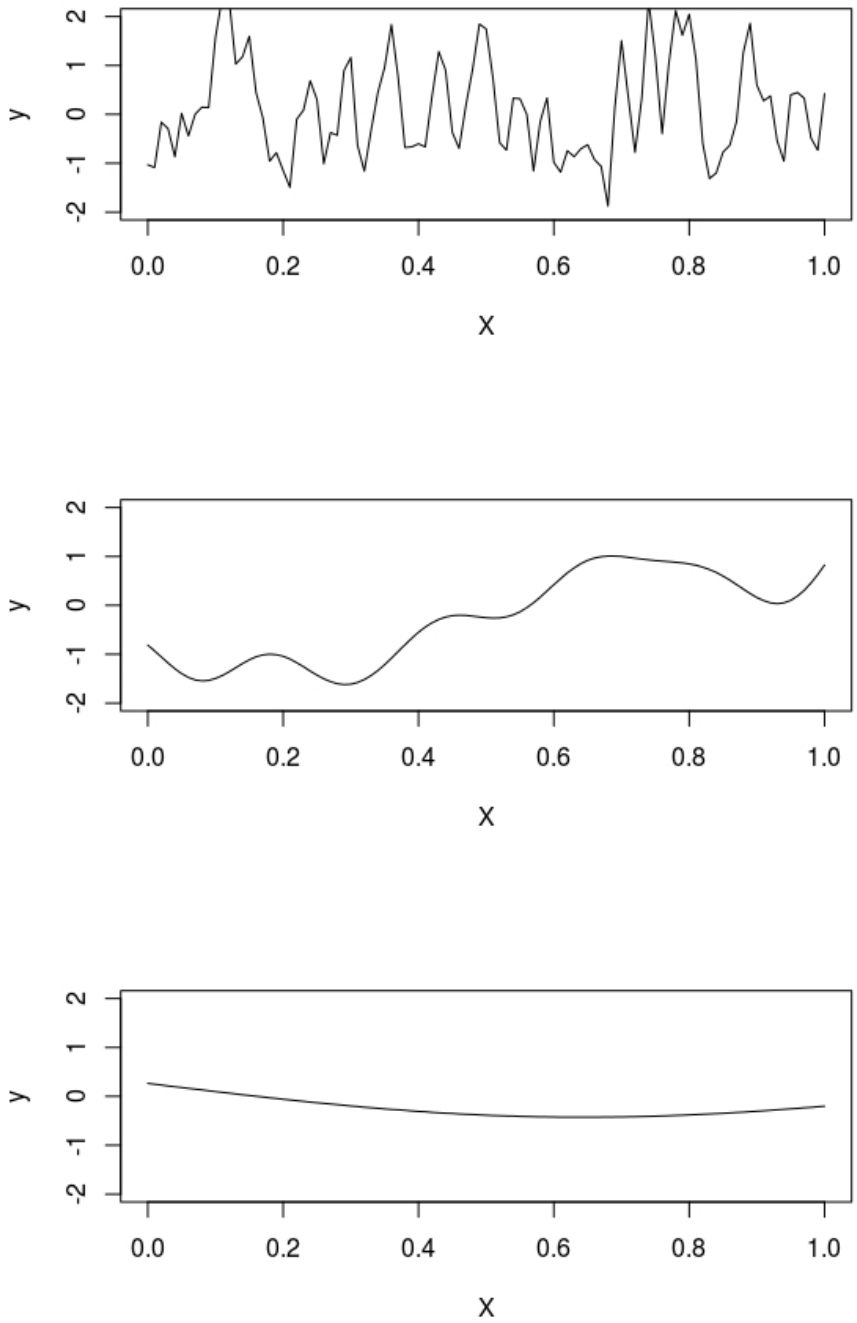

Figure 8: Functions sampled the a Gaussian Process with $\tau_1^2 = 1$ and $\tau_2^2 = 10^{-6}$. Top: $b = 0.01$; Middle: $b = 0.1$; Bottom: $b = 1$. We see that the function is smoother as b increases.

and variance:

$$C_{*|1..N} = C_* - C_{*x} C_{xx}^{-1} C_{*x}^T$$

C

We have:

$$\begin{aligned}
p(y, \theta) &= p(y \mid \theta) p(\theta) \\
&= N(R\theta, \Sigma) N(m, V) \\
&\propto \exp\left(-\frac{1}{2}(y - R\theta)^T \Sigma^{-1} (y - R\theta)\right) \exp\left(-\frac{1}{2}(\theta - m)^T V^{-1} (\theta - m)\right) \\
&= \exp\left[-\frac{1}{2}\left(y^T \Sigma^{-1} y - 2 y^T \Sigma^{-1} R\theta + \theta^T R^T \Sigma^{-1} R\theta + (\theta - m)^T V^{-1} (\theta - m)\right)\right]
\end{aligned}$$

We then see that precision matrix of y and θ is:

$$C = \begin{pmatrix} \Sigma^{-1} & -\Sigma^{-1} R \\ -R^T \Sigma^{-1} & V^{-1} + R^T \Sigma^{-1} R \end{pmatrix}$$

y and θ are Normal with mean Rm, m and precision C.

In nonparametric regression and spatial smoothing

A

Let $f_x = [f(x_1), \ldots, f(x_N)]^T$ and $y = [y_1, \ldots, y_N]^T$.

$$\begin{aligned}
p(f_x, y) &= p(y \mid f_x) p(f_x) \\
&= \prod_{i=1}^{N} p(y_i \mid f(x_i)) p(f_x) \\
&= \prod_{i=1}^{N} N(y_i \mid f(x_i), \sigma^2) N(f_x \mid 0, C) \\
&= N(y \mid f_x, \sigma^2 I) N(f_x \mid 0, C)
\end{aligned}$$

By Lemma 1 in part (C), we have that y and f_x are jointly Normal with mean 0 and precision:

$$\begin{pmatrix} 1/\sigma^2 I & -1/\sigma^2 I \\ -1/\sigma^2 I & C^{-1} + 1/\sigma^2 I \end{pmatrix}$$

Then the posterior of f_x given y is Normal with mean

$$\frac{1}{\sigma^2}(C^{-1} + 1/\sigma^2 I)^{-1} y$$

and precision:

$$C^{-1} + 1/\sigma^2 I$$

B

We have that $f^* = f(x^*)$ and $y = [y_1, \ldots, y_N]^T$ is jointly Normal with mean 0 and variance:

$$\begin{pmatrix} C_{**} & C_{*x} \\ C_{*x}^T & C_x + \sigma^2 I \end{pmatrix}$$

Now given y, we have that f^* is Normal with mean

$$C_{*x}(C_x + \sigma^2 I)^{-1} y$$

and variance:

$$C_{**} - C_{*x}(C_x + \sigma^2 I)^{-1} C_{*x}^T$$

From the expression for the mean, we can recognize a linear smoother with the smoothing weight $C_{*x}(C_x + \sigma^2 I)^{-1}$.

C

D

$$\begin{aligned} p(y \mid b, \tau_1^2, \tau_2^2) &= \int p(y \mid f) p(f \mid b, \tau_1^2, \tau_2^2) df \\ &= \int N(y \mid f, \sigma^2 I) N(f \mid 0, C) df \\ &= N(y \mid 0, C + \sigma^2 I) \end{aligned}$$

The marginal likelihood is:

$$(2\pi)^{-n/2} |C|^{-1/2} \exp\left(\frac{1}{2} y^T (C + \sigma^2 I)^{-1} y\right)$$

E

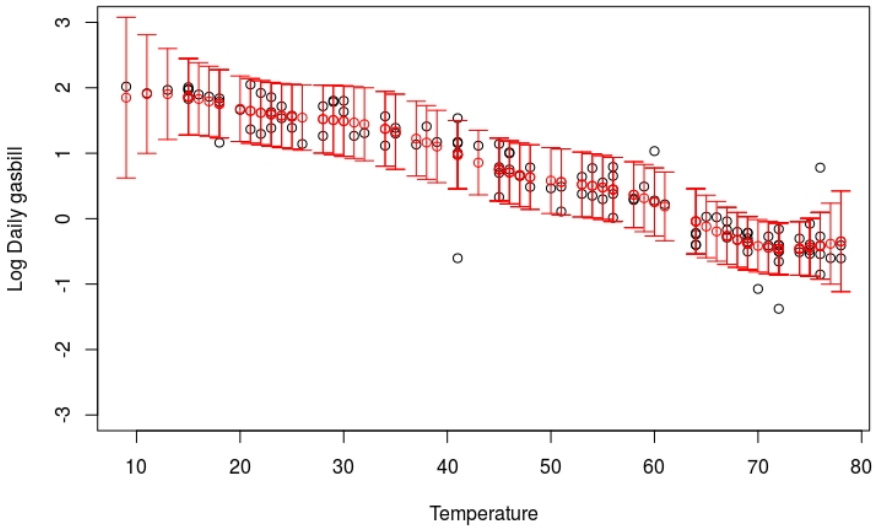

Figure 9: The data and prediction by Gaussian Process after a log transformation on the daily gasbill, with the 95 % confidence interval.

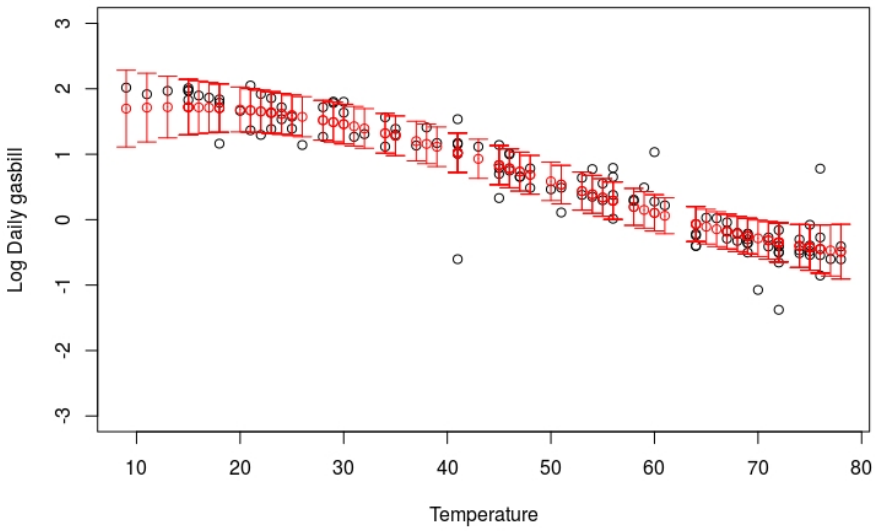

Figure 10: The data and prediction by Gaussian Process after a log transformation on the daily gasbill, with the 95 % confidence interval, after optimizing the parameters by grid search.