ECON 883-6 Problem Set 1

Otavio Braga, Ricardo Mirando, Alexander Whitefield

November 2022

1

In this question we assume that x is scalar. We also assume that the marginal and joint densities of X, Y have well defined PDFs.

Consider an arbitrary h. First we note that,

$$\begin{aligned}
Q(\tau) = \mathbb{E}\left[(Y-\tau)^2 K\left(\frac{X-x}{h}\right)\right] &= \mathbb{E}\left[(Y^2+2\tau Y-\tau^2)K\left(\frac{X-x}{h}\right)\right] \\
&\propto_\tau \mathbb{E}\left[(-2\tau Y+\tau^2)K\left(\frac{X-x}{h}\right)\right] \\
&= -2\tau\,\mathbb{E}\left[YK\left(\frac{X-x}{h}\right)\right] + \tau^2\,\mathbb{E}\left[K\left(\frac{X-x}{h}\right)\right]
\end{aligned}$$

Therefore,

$$\arg\min_\tau Q(\tau) = \arg\min_\tau -2\tau\,\mathbb{E}\left[YK\left(\frac{X-x}{h}\right)\right] + \tau^2\,\mathbb{E}\left[K\left(\frac{X-x}{h}\right)\right]$$

Differentiating with respect to τ gives the following first order condition.

$$Q'_\tau = -2\,\mathbb{E}\left[YK\left(\frac{X-x}{h}\right)\right] + 2\tau\,\mathbb{E}\left[K\left(\frac{X-x}{h}\right)\right] = 0$$

Therefore,

$$\tau^* = \frac{\mathbb{E}\left[YK\left(\frac{X-x}{h}\right)\right]}{\mathbb{E}\left[K\left(\frac{X-x}{h}\right)\right]} \tag{1}$$

To ensure that this is a global minimum, we look at the second order derivative.

$$Q''_\tau = 2\,\mathbb{E}\left[K\left(\frac{X-x}{h}\right)\right] < 0$$

So the function is globally increasing implying τ^* is the global optimum.

We would like to show that

$$\lim_{h\to 0}\arg\min_\tau Q(\tau^*) = m(x) \tag{2}$$

First we consider $\mathbb{E}\left[K\left(\frac{X-x}{h}\right)\right]$

$$
\begin{aligned}
\mathbb{E}\left[K\left(\frac{X-x}{h}\right)\right] &= \int K\left(\frac{\tilde{x}-x}{h}\right) f_X(\tilde{x})d\tilde{x} \\
&= \int K(\psi) f_X(n\psi + x)hd\psi \\
&= h\int K(\psi) f_X(h\psi + x)d\psi \\
&= h\int K(\psi)[f_X(x) + f_X'(x)h\psi + f_X^{''}(x')h^2\psi^2]d\psi \\
&= h\left[\int K(\psi)[f_X(x)]d\psi + \int K(\psi)[f_X'(x)h\psi]d\psi + \int K(\psi)[f_X^{''}(x')h^2\psi^2]d\psi\right] \\
&= h\left[f_X(x)\int K(\psi)d\psi + f_X'(x)h\int K(\psi)\psi d\psi + h^2 f_X^{''}(x')\int K(\psi)\psi^2 d\psi\right]
\end{aligned}
$$

We use a change of variables with $\psi = \frac{\tilde{x}-x}{h}$ in line 4, and take a Taylor expansion around x in line 6. We assume that f_X is two time continuously differentiable. If we also assume that,

- $\int K(\psi)d\psi = 1$
- $\int K(\psi)\psi d\psi = 0$ (this would be implied by symmetry)
- $\int K(\psi)\psi^2 d\psi = \sigma_x^2 < \infty$

Then we have

$$
\mathbb{E}\left[K\left(\frac{X-x}{h}\right)\right] = h\left[f_X(x) + h^2 f_X^{''}(x')\sigma_x\right]
$$

Next we consider $\mathbb{E}\left[YK\left(\frac{X-x}{h}\right)\right]$.

$$
\begin{aligned}
\mathbb{E}\left[YK\left(\frac{X-x}{h}\right)\right] &= \int\int \tilde{y}K\left(\frac{\tilde{x}-x}{h}\right) f_{X,Y}(\tilde{x},\tilde{y})d\tilde{x}d\tilde{y} \\
&= \int K\left(\frac{\tilde{x}-x}{h}\right) f_X(\tilde{x})\int \tilde{y} f_{Y|X=x}(\tilde{y})d\tilde{y}d\tilde{x} \\
&= \int K\left(\frac{\tilde{x}-x}{h}\right) f_X(\tilde{x})\int \tilde{y} f_{Y|X=x}(\tilde{y})d\tilde{y}d\tilde{x} \\
&= \int K\left(\frac{\tilde{x}-x}{h}\right) f_X(\tilde{x})m(\tilde{x})d\tilde{x} \\
&= \int K(\psi) f_X(h\psi + x)m(h\psi + x)hd\psi \\
&= h\int K(\psi)[f_X(x) + f_X'(x)h\psi + f_X^{''}(x')h^2\psi^2][m(x) + m'(x)h\psi + m''(x'')h^2\psi^2]d\psi
\end{aligned}
$$

In line 4 we use $m(x) = E[Y|X = \tilde{x}]$. In line 5 we use a change of variables, and in line 6 we can use a Taylor expansion, while noting that $m(x)$ and f_X are both (two times) continuously differentiable. If we further assume that,

- $\int K(\psi)^3 d\psi = \sigma_\psi^3$ where $|\sigma_\psi^3| \leq \infty$
- $\int K(\psi)^4 d\psi = \sigma_\psi^4 \leq \infty$

Then the expression simplifies to

$$
\begin{aligned}
\mathbb{E}\left[YK\left(\frac{X-x}{h}\right)\right] = h\Bigg(\bigg[& f_X(x)m(x)\bigg] \\
& + \sigma_\psi^2 h^2\bigg[m(x)f_X''(x') + f_X(x)m''(x'') + f_X'(x)m'(x)\bigg] \\
& + \sigma_\psi^3 h^3\bigg[f_X'(x)m''(x'') + m'(x)f_X''(x')\bigg] \\
& + \sigma_\psi^4 h^4\bigg[f_X''(x')m''(x'')\bigg]\Bigg)
\end{aligned}
$$

Therefore, using (1) and (2), we have

$$
\lim_{h\to 0} \arg\min_\tau Q(\tau^*) = \frac{h((f_X(x)m(x)) + O(h^2) + O(h^3) + O(h^4))}{h(f_X(x) + O(h^2) + O(h^3) + O(h^4))} = m(x)
$$

2

2.1

Suppose $p(x) \approx k$ is in some neighbourhood $[x - \delta, x + \delta]$. The likelihood of a set of N observations Y_N in this neighbourhood is roughly given by

$$\begin{aligned}
\mathbb{P}[Y = Y_n | p(x)] &= \prod_{i=1}^{N} k_i^y \cdot (1-k)^{1-y_i} \\
&\propto \log\left(\prod_{i=1}^{N} k_i^y \cdot (1-k)^{1-y_i}\right) \\
&= \sum_{i=1}^{n} (y_i \ln k + (1-y_i)\ln(1-k))
\end{aligned}$$

As $p(x) \approx k$, the final line is 'approximately' the log likelihood of the data given k. We could pretend this is the likelihood function, and maximise it to obtain k^* that we could interpret as being close to $p(x)$ in the region $[x - \delta, x + \delta]$. Expression (2) given below is similar to the process just described. However, rather than use an indicator function kernel, it uses an arbitrary kernel.

$$\max_{\tau} Q(\tau) = \max_{\tau} \sum_{i=1}^{n} (y_i \ln \tau + (1-y_i)\ln(1-\tau)) K\left(\frac{x_i - x}{h}\right) \tag{3}$$

Given x, it can be interpreted as maximising the likelihood of the observed data, weighting observations close to x_i higher.

2.2

We write the object as follows

$$\max_{\tau} Q(\tau) = \max_{\tau} \sum_{i=1}^{n} (y_i \ln \tau + (1-y_i)\ln(1-\tau)) K\left(\frac{x_i - x}{h}\right)$$

Differentiating with respect to τ gives the following first order condition:

$$\sum_{i=1}^{n} \left(\frac{y_i}{\tau} - \frac{(1-y_i)}{(1-\tau)}\right) K\left(\frac{x_i - x}{h}\right) = 0$$

Rearranging gives

$$
\begin{aligned}
\sum_{i=1}^{n} \frac{y_i}{\tau} K\left(\frac{x_i - x}{h}\right) &= \sum_{i=1}^{n} \frac{(1-y_i)}{(1-\tau)} K\left(\frac{x_i - x}{h}\right) \\
(1-\tau)\sum_{i=1}^{n} y_i K\left(\frac{x_i - x}{h}\right) &= \tau \sum_{i=1}^{n} (1-y_i) K\left(\frac{x_i - x}{h}\right) \\
\sum_{i=1}^{n} y_i K\left(\frac{x_i - x}{h}\right) &= \tau \sum_{i=1}^{n} (1-y_i) K\left(\frac{x_i - x}{h}\right) + \tau \sum_{i=1}^{n} y_i K\left(\frac{x_i - x}{h}\right) \\
\sum_{i=1}^{n} y_i K\left(\frac{x_i - x}{h}\right) &= \tau \left[\sum_{i=1}^{n} (1-y_i) K\left(\frac{x_i - x}{h}\right) + \sum_{i=1}^{n} y_i K\left(\frac{x_i - x}{h}\right)\right] \\
\sum_{i=1}^{n} y_i K\left(\frac{x_i - x}{h}\right) &= \tau \left[\sum_{i=1}^{n} (1-y_i+y_i) K\left(\frac{x_i - x}{h}\right)\right] \\
\sum_{i=1}^{n} y_i K\left(\frac{x_i - x}{h}\right) &= \tau \left[\sum_{i=1}^{n} K\left(\frac{x_i - x}{h}\right)\right]
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
\tau^* &= \frac{\sum_{i=1}^{n} y_i K\left(\frac{x_i - x}{h}\right)}{\sum_{i=1}^{n} K\left(\frac{x_i - x}{h}\right)} \\
&= \sum_{i=1}^{n} l_i y_i
\end{aligned}
$$

where $l_i = \frac{K\left(\frac{x_i - x}{h}\right)}{\sum_{i=1}^{n} K\left(\frac{x_i - x}{h}\right)}$. Therefore, we can write $\hat{p}(x)$ is a kernel regression estimator.

3

3.1

$$
\begin{aligned}
S_n(\tau) &= \frac{1}{nh} \sum_{i=1}^{n} (y_i - m - (x_i - x)\beta)^2 K\left(\frac{x - x_i}{h}\right) \\
&= \frac{1}{nh} \sum_{i=1}^{n} \left((y_i - m - (x_i - x)\beta))\sqrt{K\left(\frac{x - x_i}{h}\right)}\right)^2
\end{aligned}
$$

Define $\tilde{x}_i = (x_i - x)\sqrt{K\left(\frac{x_i - x}{h}\right)}$, $\tilde{y_i} = y_i\sqrt{K\left(\frac{x_i - x}{h}\right)}$ and $\tilde{k}_i = \sqrt{K\left(\frac{x_i - x}{h}\right)}$

Then, fixing x, the problem becomes equivalent to estimating a linear model without constant using OLS:

$$
\min_{m,\beta} \sum_{i=1}^{n} (\tilde{y}_i - m\tilde{k}_i - \tilde{x}_i\beta)^2
$$

The first order conditions are:

$$\sum_{i=1}^{n}(\tilde{y}_i - m\tilde{k} - \beta\tilde{x}_i)\tilde{k}_i = 0$$
$$\sum_{i=1}^{n}(\tilde{y}_i - m\tilde{k} - \beta\tilde{x}_i)\tilde{x}_i = 0$$

Solving for β and m yields:

$$\beta = \frac{\sum_{i=1}^{n}(\tilde{y}_i - m\tilde{k}_i)}{\sum_{i=1}^{n}\tilde{x}_i^2}$$
$$m = \frac{\sum_{i=1}^{n}(\tilde{y}_i - \beta\tilde{x}_i)}{\sum_{i=1}^{n}\tilde{k}_i^2}$$

3.2

Substituting β into the equation for m:

$$m = \Big(\sum_{i=1}^{n}\tilde{k}_i^2\Big)^{-1}\Big[\sum_{i=1}^{n}\tilde{y}_i - \Big(\frac{\sum_{i=1}^{n}(\tilde{y}_i - m\tilde{k}_i)}{\sum_{i=1}^{n}\tilde{x}_i^2}\Big)\sum_{i=1}^{n}\tilde{x}_i\Big]$$

$$\implies m\Big[\frac{\sum_{i=1}^{n}\tilde{k}_i^2\sum_{i=1}^{n}\tilde{x}_i^2 - \sum_{i=1}^{n}\tilde{k}_i\sum_{i=1}^{n}\tilde{x}_i}{\sum_{i=1}^{n}\tilde{x}_i^2}\Big] = \frac{\sum_{i=1}^{n}\tilde{y}_i\sum_{i=1}^{n}\tilde{x}_i^2 - \sum_{i=1}^{n}\tilde{y}_i\sum_{i=1}^{n}\tilde{x}_i}{\sum_{i=1}^{n}\tilde{x}_i^2}$$

$$\implies m = \frac{\sum_{i=1}^{n}\tilde{y}_i\sum_{i=1}^{n}\tilde{x}_i^2 - \sum_{i=1}^{n}\tilde{y}_i\sum_{i=1}^{n}\tilde{x}_i}{\sum_{i=1}^{n}\tilde{k}_i^2\sum_{i=1}^{n}\tilde{x}_i^2 - \sum_{i=1}^{n}\tilde{k}_i\sum_{i=1}^{n}\tilde{x}_i}$$

Substituting x_i and y_i the last equation becomes:

$$\begin{aligned} m &= \frac{\sum_{i=1}^{n} y_i\sqrt{K\Big(\frac{x_i-x}{h}\Big)}\Big[\sum_{i=1}^{n}(x_i-x)^2K\Big(\frac{x_i-x}{h}\Big) - \sum_{i=1}^{n}(x_i-x)\sqrt{K\Big(\frac{x_i-x}{h}\Big)}\Big]}{\sum_{i=1}^{n}K\Big(\frac{x_i-x}{h}\Big)\sum_{i=1}^{n}(x_i-x)^2K\Big(\frac{x_i-x}{h}\Big) - \sum_{i=1}^{n}(x_i-x)\sum_{i=1}^{n}\sqrt{K\Big(\frac{x_i-x}{h}\Big)}} \qquad (4) \\ &= \sum_{i=1}^{n} w_i(x)y_i \qquad (5) \end{aligned}$$

The solution obtained is a weighted average of the y_i's where the weights are:

$$w_i(x) = \frac{\sqrt{K\Big(\frac{x_i-x}{h}\Big)}\Big[\sum_{i=1}^{n}(x_i-x)^2K\Big(\frac{x_i-x}{h}\Big) - \sum_{i=1}^{n}(x_i-x)\sqrt{K\Big(\frac{x_i-x}{h}\Big)}\Big]}{\sum_{i=1}^{n}K\Big(\frac{x_i-x}{h}\Big)\sum_{i=1}^{n}(x_i-x)^2K\Big(\frac{x_i-x}{h}\Big) - \sum_{i=1}^{n}(x_i-x)\sum_{i=1}^{n}\sqrt{K\Big(\frac{x_i-x}{h}\Big)}} \qquad (6)$$

3.3

Using the definitions for $\tilde{y}_i$, $\tilde{y}$, $\tilde{k}$ in the first part of the problem, for any given x it is sufficient to calculate $\tilde{k}_i, \tilde{y}_i$ and $\tilde{k}_i$ for all $i \in \{1, 2, ..., n\}$. To compute $\hat{m}(x)$ and $\hat{\beta}(x)$ we would fit a linear model $\tilde{y}_i = m\tilde{k}_i - \beta\tilde{x}_i$ using software for OLS estimation (note that the linear model should be estimated without intercept).

4

We implement the locally linear estimator using the *loess* function in R. We set set the degree parameter to 1. We implement a NW kernal estimator by defining the function: *nw_kernal_estimator* as shown below.

```
nw_kernal_estimator <- function(h,x,data) {
  #h: bandwidth
  #x: test datapoint
  #data: training dataset
  Kx <- sapply(data$accel, function(i) dnorm(((x - i) / h)))
  l <- Kx/sum(Kx) #weights
  r_hat <- data$times %*%  l
  return(r_hat[1,1])
}
```

We implement leave one out cross validation by defining the function: *cv_score*. The R code for the function is presented below. The function takes in as input a list of bandwidths (*h_list*), and the type of non-parametric estimator that should be used). The inner loop of the function loops through each observation, and for each, computes a non-parametric estimate for that observations x value by fitting a non-parametric model to the entire dataset excluding that observation. The outer loop iterates across bandwidths, and computes the mean squared loss across all of the predictions in the inner loop.

```
cv_score <- function(h_list,est_type = 'local_polynomial') {
  res_list_outer <- rep(-1,length(h_list))
  for (j in 1:length(h_list)) {
    res_list_inner <- rep(-1,dt[,.N])
    for (i in 1:dt[,.N]) {
      dt_temp <- dt[!i]
      if (est_type == 'local_polynomial'){
        mod.temp <- loess(times ~ accel , span=h_list[j], degree=1, data=dt_temp)
        y_hat <- predict(mod.temp, newdata = dt[i,accel])
      } else if (est_type == 'nw_kernal_estimator') {
        y_hat <- nw_kernal_estimator(h=h_list[j],x=dt[i,accel],dt_temp)
      } else {
        stop("error, unknown est_type")
      }

      res_list_inner[i] <- (y_hat - dt[i,times])^2
    }
    res_list_outer[j] <- sum(res_list_inner,na.rm=T)
  }
  return(res_list_outer)
}
```

For each estimator, we then pick the bandwidth associated the lowest mean square. The results for cross validation are plotted in Figure 1. For locally linear regression and the NW kernel estimator, the selected bandwidths are and 0.4 and 3.8 respectively.

The results are presented in Figure 2. The locally linear estimate of the conditional mean function looks reasonable, while the NW kernal estimator seems to overfit the data.

References

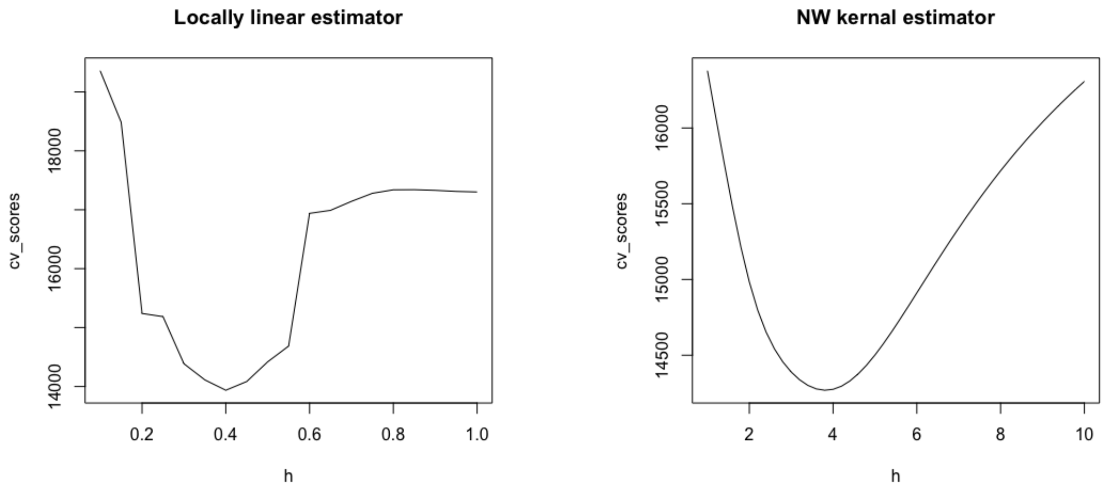

Figure 1: Bandwidth selection using leave one out cross validation.

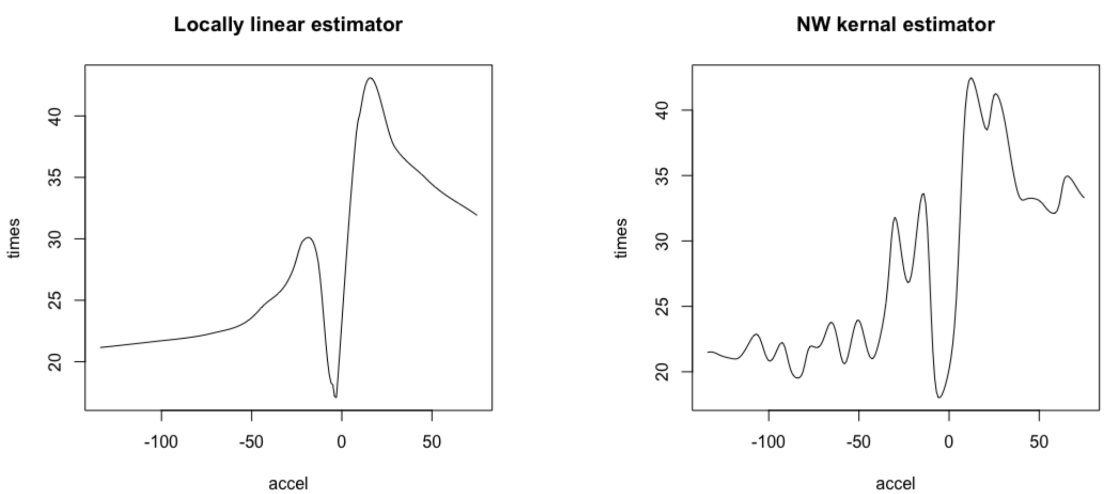

Figure 2: Non-parametric estimators of the conditional mean function