# BIOST 527 - HOMEWORK 1

KATIE WOOD

## Problem 1

1. The probability that a single $X_i$ falls in $B_j$ equals the area under $f$ that is within the bounds of $B_j$:
$$p_j = \int_{B_j} f(x)dx$$
Then the event "$X_i$ is observed in $B_j$" is a Bernoulli trial with success probability $p_j$. We know that $n$ repeated Bernoulli trials give rise to a Binomial distribution. Since $X_1, ..., X_n$ are i.i.d. $F$, we observe $n$ Bernoulli trials with success probability $p_j$. Thus, $v_j$, the number of observations in $B_j$, is a random variable distributed according to $v_j \sim \text{Binom}(p_j, n)$.
2. With $x_0 \in B_j$ being fixed,
$$\begin{aligned}
Var[\hat{f}_n(x_0)] &= Var\left[\sum_{k=1}^{n} \frac{\hat{p}_k}{h}\mathbb{I}\{x_0 \in B_k\}\right] \\
&= Var\left[\frac{\hat{p}_j}{h}\right] \\
&= Var\left[\frac{v_j}{nh}\right] \\
&= \frac{1}{n^2h^2}Var[v_j] \\
&= \frac{1}{n^2h^2}np_j(1-p_j) \qquad \text{Binomial variance} \\
&\leq \frac{1}{n^2h^2}np_j \\
&= \frac{1}{nh^2}\int_{B_j} f(x)dx \\
&\leq \frac{1}{nh^2}\int_{B_j} f_{max}dx \\
&= \frac{1}{nh^2}f_{max}\int_{B_j} dx \\
&= \frac{1}{nh^2}f_{max}h \\
&= \frac{1}{nh}f_{max} \\
&= \frac{C_2}{nh}
\end{aligned}$$

3. (a) For $x_0 \in B_j$, we verify

$$\begin{aligned}
\mathbb{E}[\hat{f}_n(x_0)] &= \mathbb{E}\left[\sum_{k=1}^{n} \frac{\hat{p}_k}{h}\mathbb{I}\{x_0 \in B_k\}\right] \\
&= \mathbb{E}\left[\frac{\hat{p}_j}{h}\right] \\
&= \mathbb{E}\left[\frac{v_j}{nh}\right] \\
&= \frac{1}{nh}\mathbb{E}[v_j] \\
&= \frac{1}{nh}np_j \qquad \text{Binomial mean} \\
&= \frac{p_j}{h}
\end{aligned}$$

(b) By the Mean Value Theorem for Integrals, $\exists c \in B_j$ such that

$$p_j = \int_{B_j} f(x)dx = hf(c)$$

Thus,

$$\begin{aligned}
|\mathbb{E}[\hat{f}_n(x_0)] - f(x_0)| &= \left|\frac{p_j}{h} - f(x_0)\right| \\
&= \left|\frac{hf(c)}{h} - f(x_0)\right| \\
&= |f(c) - f(x_0)| \\
&\leq |f(x_0) - f(c)|
\end{aligned}$$

(c) Then by the Fundamental Theorem of Calculus,

$$\begin{aligned}
|\mathbb{E}[\hat{f}_n(x_0)] - f(x_0)| &\leq |f(x_0) - f(c)| \\
&= \left|\int_c^{x_0} f'(x)dx\right| \\
&\leq \int_c^{x_0} |f'(x)|dx \\
&\leq \int_{B_j} |f'(x)|dx \\
&\leq \int_{B_j} |f'_{max}|dx \\
&= |f'_{max}| \int_{B_j} dx \\
&= |f'_{max}|h \\
&= C_1 h
\end{aligned}$$

4. Therefore,

$$
\begin{aligned}
MSE(\hat{f}_n(x_0), f(x_0)) &= \mathbb{E}\left[(\hat{f}_n(x_0) - f(x_0))^2\right] \\
&= \mathbb{E}\left[(\hat{f}_n(x_0) - \mathbb{E}[\hat{f}_n(x_0)] + \mathbb{E}[\hat{f}_n(x_0)] - f(x_0))^2\right] \\
&= \mathbb{E}\left[(\hat{f}_n(x_0) - \mathbb{E}[\hat{f}_n(x_0)])^2\right] + \mathbb{E}\left[(\mathbb{E}[\hat{f}_n(x_0)] - f(x_0))^2\right] \\
&\qquad \cancel{+ \mathbb{E}\left[\hat{f}_n(x_0) - \mathbb{E}[\hat{f}_n(x_0)]\right] \mathbb{E}\left[\mathbb{E}[\hat{f}_n(x_0)] - f(x_0)\right]} \\
&= Var[\hat{f}_n(x_0)] + (\mathbb{E}[\hat{f}_n(x_0)] - f(x_0))^2 \\
&\leq \frac{C_2}{nh} + (C_1 h)^2 \\
&= C_1 h^2 + \frac{C_2}{nh} \qquad \text{renaming } C_1 = C_1^2
\end{aligned}
$$

5. We seek to minimize $MSE(\hat{f}_n(x_0), f(x_0))$ with respect to $h$, so we take the $h$-partial derivative and set it equal to 0:

$$
\begin{aligned}
\frac{\partial}{\partial h} MSE(\hat{f}_n(x_0), f(x_0)) = 2C_1 h_{opt} - \frac{C_2}{n h_{opt}^2} &= 0 \\
\Rightarrow 2C_1 h_{opt} &= \frac{C_2}{n h_{opt}^2} \\
h_{opt}^3 &= \frac{C_2}{2C_1} \frac{1}{n} \\
h_{opt} &= \left( \frac{C_2}{2C_1} \frac{1}{n} \right)^{1/3}
\end{aligned}
$$

Then using the result from part 4, we have

$$
\begin{aligned}
\min_h MSE(\hat{f}_n(x_0), f(x_0)) &\leq C_1 \left( \frac{C_2}{2C_1} \frac{1}{n} \right)^{2/3} + \frac{C_2}{n} \left( \frac{C_2}{2C_1} \frac{1}{n} \right)^{-1/3} \\
&= C_1^* n^{-2/3} + C_2^* n^{-1+1/3} \\
&= C_1^* n^{-2/3} + C_2^* n^{-2/3} \\
&= C^* n^{-2/3}
\end{aligned}
$$

6. The parametric rate of convergence is known to be $n^{-1}$. Since $\frac{1}{n} \ll \frac{1}{n^{2/3}}$ as $n \to \infty$, the histogram rate of convergence is slower than the parametric rate. Thus, we are able to avoid making any structural assumptions on the density function $f$ at the cost of a slower rate of convergence for the MSE.

---

# Problem 2

1. Looking at Figure 1, the scalar value $\mathbb{E}[Y|X = x]$ may be unsatisfactory because for values of the predictor variable (bmi) greater than 30, the conditional density $f_{Y|X}(y|x)$ appears to be bimodal. Taking the expectation $\mathbb{E}[Y|X = x]$ presumably results in some intermediate value of the response variable (charges), which obscures the fact that for $bmi \geq 30$ there seem to be two separate groups: one

with low charges (lower than $\mathbb{E}[Y|X=x]$) and a second, smaller group with high charges (much higher than $\mathbb{E}[Y|X=x]$).

2. The density function $f_{Y|X}(y|x)$ represents the distribution of health insurance charges for a given bmi. We could think of taking 2D slices for every value of $x$ (bmi), such that within each slice, $f_{Y|X}(y|x)$ is a curve indicating the distribution of outcomes $y$ (charges). We could then imagine compiling all of these 2D slices and associated density curves together to create a surface, whose height above the $x, y$-plane would represent the probability mass at every point $(x, y)$.

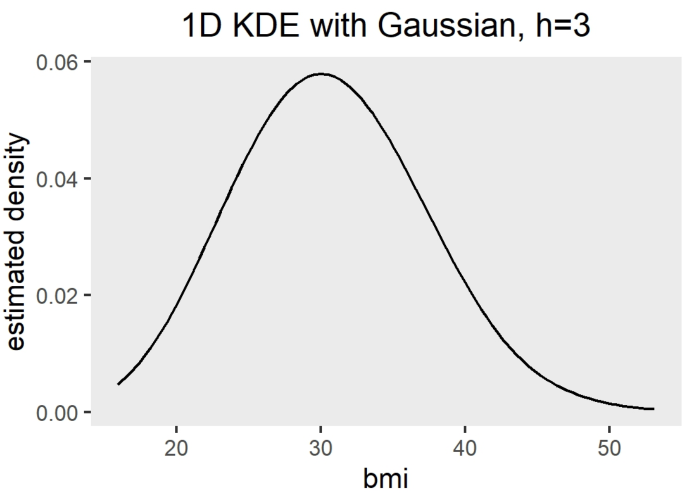
   Estimating $f_{Y|X}(y|x)$ is a challenging problem because as stated before, the conditional distribution of $y$ appears to be bimodal for $x \geq 30$, but unimodal for $x < 30$. We could make additional assumptions, e.g. that there are actually three distinct groups hidden within the data: one with $bmi < 30$, one with $bmi \geq 30$ and high charges, and one with $bmi \geq 30$ and low charges. Then perhaps we could approximate each group separately by a Normal distribution. But we have no evidence for this assumption; we are asked to find a distribution $f_{Y|X}(y|x)$ to explain the whole dataset. The change in behavior from unimodal to bimodal as $x$ surpasses 30 is not a classic behavior of any known family of distributions. Thus we look to KDE to provide a more flexible approach.

3. Given a kernel $K : \mathbb{R} \to \mathbb{R}$ where $K$ is symmetric and $\int K(x)dx = 1$, the explicit forms of the KDE estimators are

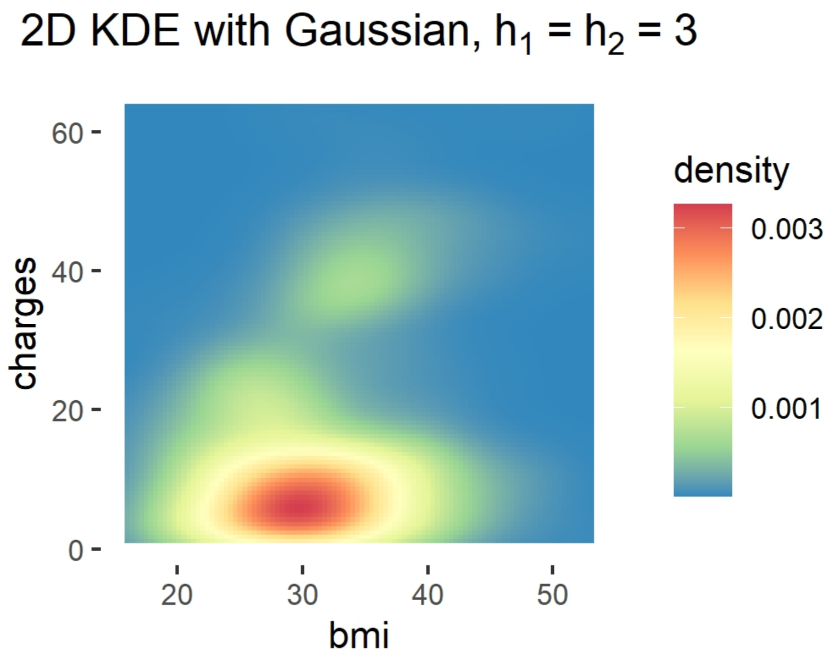
$$\hat{f}_{Y,X}^{h_1,h_2}(y,x) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h_1} K\left(\frac{X_i - x}{h_1}\right) \frac{1}{h_2} K\left(\frac{Y_i - y}{h_2}\right)$$

$$\hat{f}_X^{h_3}(x) = \frac{1}{nh_3}\sum_{i=1}^{n} K\left(\frac{X_i - x}{h_3}\right)$$

   where $h_1$, $h_2$, $h_3$ are the respective bandwidths.

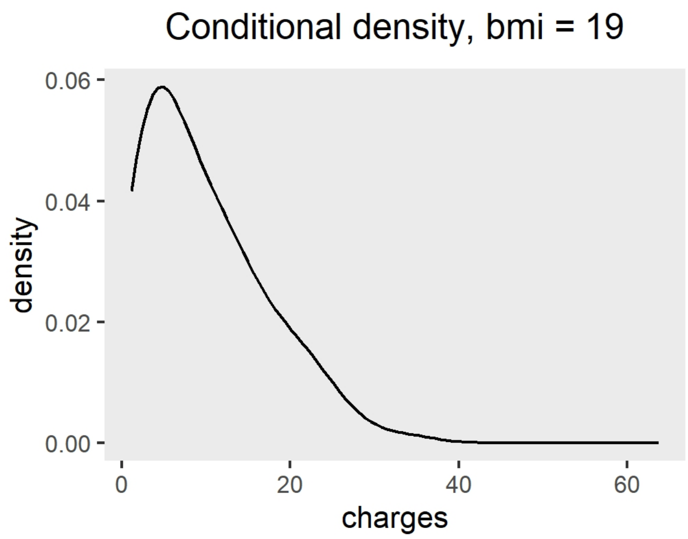
4. 

5.

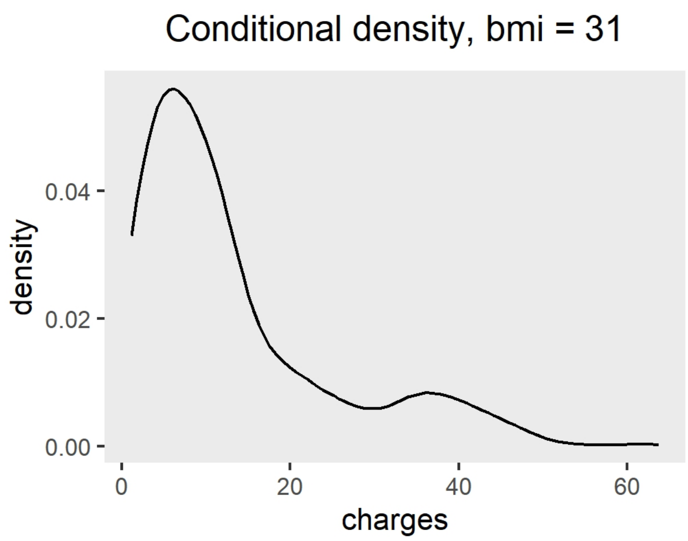


2D KDE with Gaussian, $h_1 = h_2 = 3$
charges
60
40
20
0
20
30
40
50
bmi
density
0.003
0.002
0.001

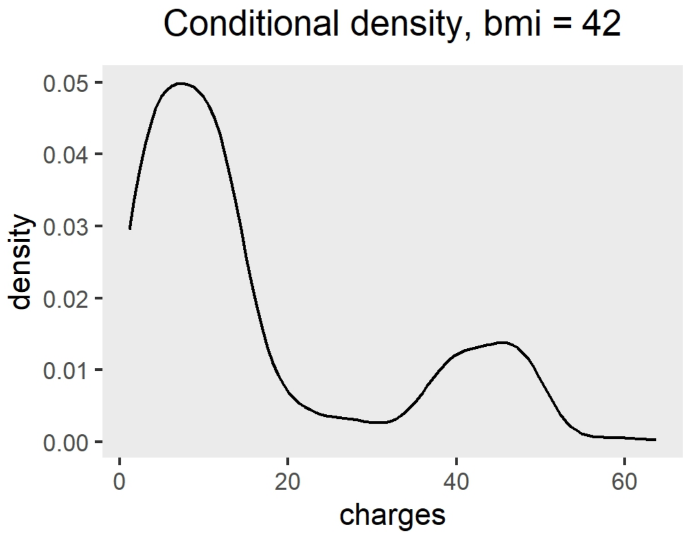

6.



Conditional density, bmi = 19
density
0.06
0.04
0.02
0.00
0
20
40
60
charges

Conditional density, bmi = 31
density
0.04
0.02
0.00
0
20
40
60
charges




Above, we have plotted the estimated conditional densities for each of three values of $x$ (a small, a medium, and a large value). Consistent with our earlier discussion, for the small value of $x$ (bmi $= 19$), we observe a unimodal distribution, while for the large value of $x$ (bmi $= 42$), we observe a distribution that is clearly bimodal. At the medium value of $x$ (bmi $= 31$), we are able to observe something like a transition between unimodal and bimodal, with a small second peak beginning to form near charges $= 40$.

7. Below, we overlay the original data plot with the 0.1-, 0.5-, and 0.9-quantiles of the conditional density. We notice that the 0.9-quantile lies significantly farther above the 0.5-quantile than the 0.1-quantile lies below, especially when bmi $\geq 30$. This would seem to indicate that charges for the most cost-burdened patients are disproportionately high.