Question 1

(1)

Plot 1a shows a local constant regression, whereas plot 1b shows a local linear regression. The local constant regression tends to be more biased around the boundary points, which can follow from a somewhat oddly upward bending regression line in the boundary regions. Instead, local linear regression performs better in the boundary regions, which follows from plot 1b where the regression line seems to follow the data a bit better.

(2)

Plot 2a uses $\lambda = 5 \cdot 10^{-7}$, plot 2b uses $\lambda = 5 \cdot 10^{-5}$, plot 2c uses $\lambda = 5 \cdot 10^{-2}$. The larger the smoothing parameter, the less spikey and the smoother the curve looks like. This is the case for plot 2c, which has the largest smoothing parameter. Plot 2a shows a less smooth and less rough line more precisely following the data points, due to the smaller smoothing parameter. Plot 2b is in between.

(3)

Plot 3a shows 3-nearest neighbours, plot 3b shows 20-nearest neighbours, plot 3c shows 50-nearest neighbours. A low tuning parameter k yields low bias and high variance which closely follows the data points (with more spikes), whereas a high tuning parameter k yields high bias and low variance since it is more rough and smoother. Hence, flexibility of the line is reduced as k increases. The most flexible and spikey line is found in plot 3a, such that this one corresponds to the lowest k of 3. The least flexible and smoothest line is displayed in plot 3c, such that this one corresponds to the highest k of 50. Plot 3b is in between.

(4)

Plot 4a shows a natural cubic spline, plot 4b shows a quadratic cubic spline, plot 4c shows a cubic spline. Quadratic splines are characterized by the parabola, quadratic nature between each pair of knots (or the minimum and maximum), its continuous first derivative, and five basis functions with corresponding coefficients. It is less smooth than cubic splines since less basis functions are used. Out of all plots, plot 4b seems to be the least smooth. Cubic splines are characterized by its continuous second derivative, and six basis functions with corresponding coefficients, where it extends the basis functions of the quadratic splines. A natural cubic spline moreover has a linear shape in the outer regions (beyond the most left and most right know), indicating plot 4a shows the natural cubic spline. Then, plot 4c shows the cubic spline which is the smoothest.

Question 2

(1)

We fit a Nadaraya-Watson estimator with a Gaussian kernel and bandwidth $h = 2$ to estimate $f(x)$. Figure 1 displays the fitted regression line with the data points. The expected income of an employee who has worked 10 years in the company is calculated as 57478.02.

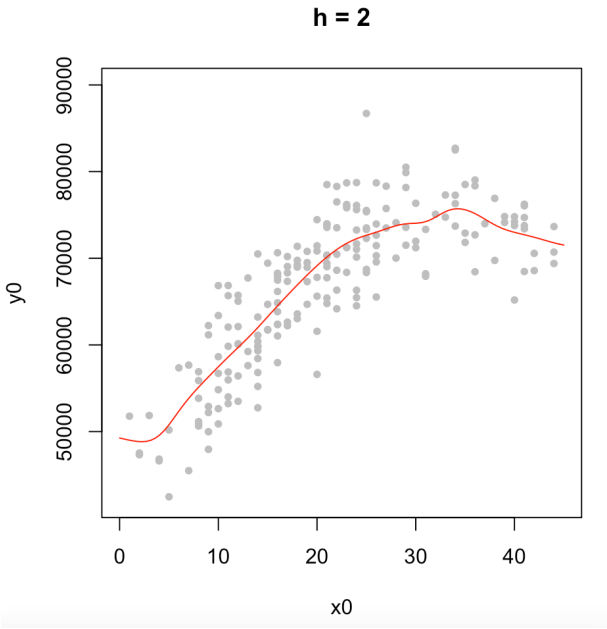

Figure 1: Fitted Nadaraya-Watson regression line with a Gaussian kernel and bandwidth $h = 2$ with the data points.

(2)

We fit a Nadaraya-Watson estimator with a uniform kernel and bandwidth $h = 5$ to estimate $f(x)$. Figure 2 displays the fitted regression line with the data points. The expected income of an employee who has worked 10 years in the company is calculated as 57895.98.

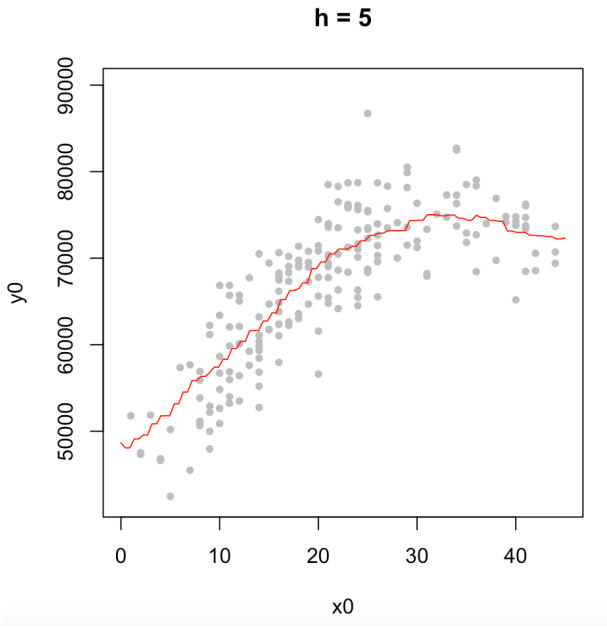

Figure 2: Fitted Nadaraya-Watson regression line with a uniform kernel and bandwidth $h = 5$ with the data points.

(3)

The models in part (1) and (2) are compared using a leave-one-out cross validation. The Mean Squared Error of the N-W estimator with a Gaussian kernel and bandwidth $h = 5$ (19350379) outperforms the N-W estimator with a uniform kernel and bandwidth $h = 2$ (19678381).

(4)

We fit a local linear regression estimator with a Gaussian kernel and bandwidth $h = 2$ to estimate $f(x)$. Figure 3 displays the fitted regression line with the data points. The expected income of an employee who has worked 10 years in the company is calculated as 56962.48.

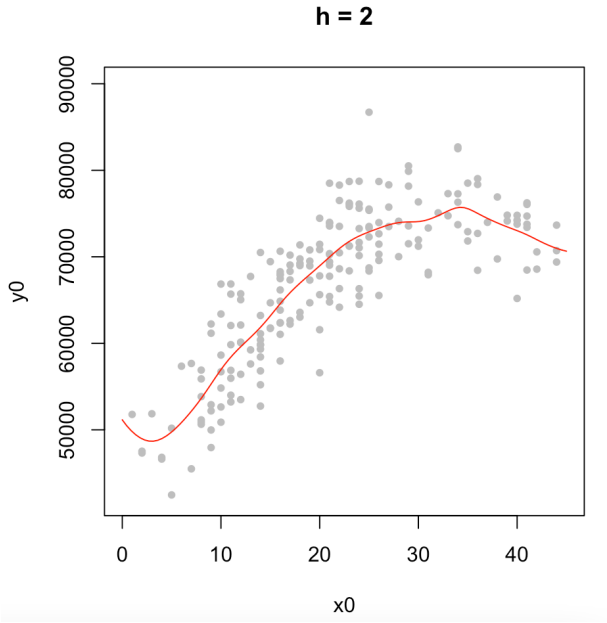

Figure 3: Fitted local linear regression line with a Gaussian kernel and bandwidth $h = 2$ with the data points.

(5)

We fit a k-nearest neighbours estimator with $k = 10$ to estimate $f(x)$. Figure 4 displays the fitted regression line with the data points. The expected income of an employee who has worked 10 years in the company is calculated as 57376.24.

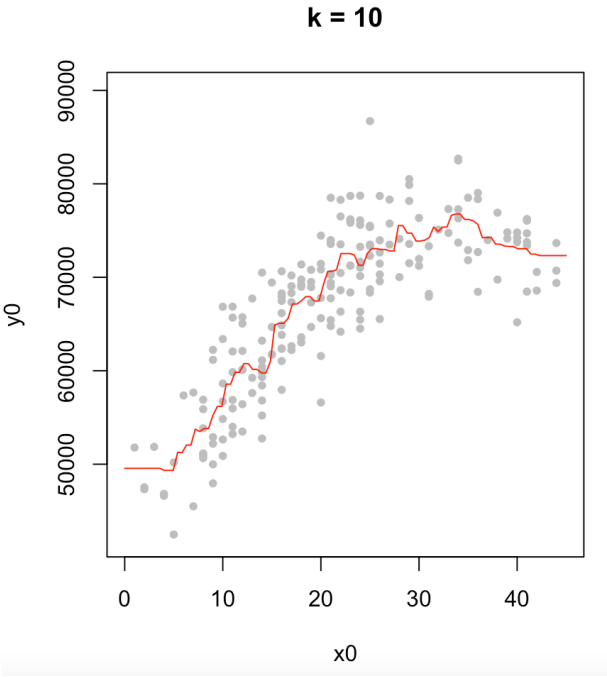

Figure 4: Fitted k-nearest neighbours estimator with $k = 10$ and with the data points.

(6)

We select the optimal k_{opt} for the k-nearest neighbours estimator to estimate $f(x)$ using 10-fold cross validation. Figure 5 displays the plotted CV values against k, indicating that out of the range $k = 2, ..., 100$, $k_{opt} = 10$ is again optimal. Figure 4 already displays this fitted regression line with the data points.

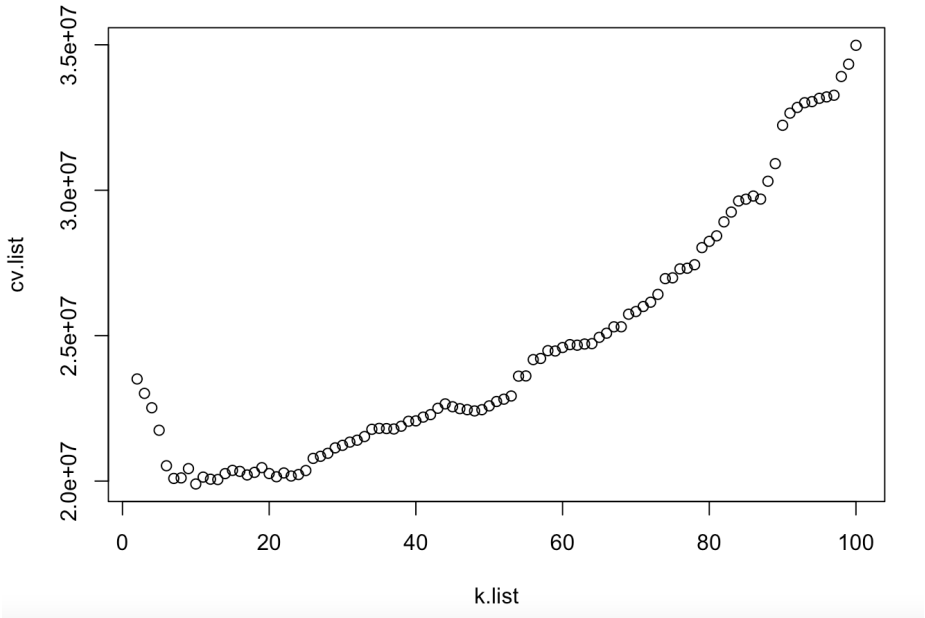

Figure 5: CV values of 10-fold cross validation plotted against k.

Question 3

(1)

A visualization of the data is displayed in Figure 6. With x_i and y_i on the axes for $i = 1, ..., n$, crosses denote data points for which the corresponding value $g_i = 0$, whereas circles denote data points for which $g_i = 1$.

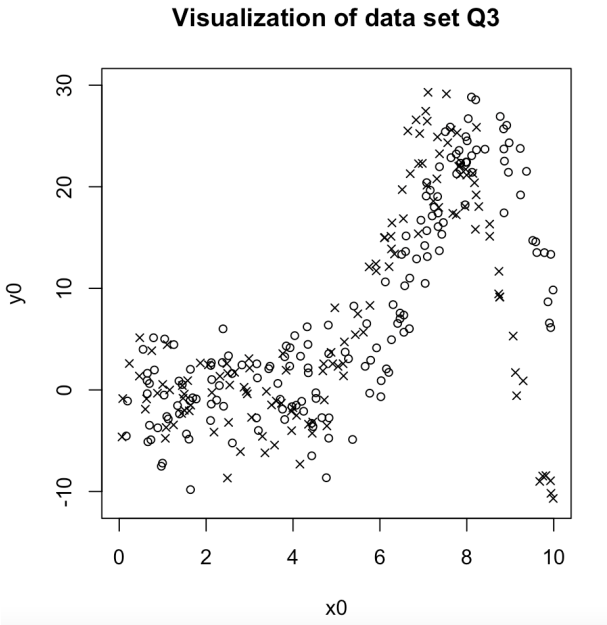

Figure 6: Visualization of the data set, where crosses (circles) denote data points (x_i, y_i) for which $g_i = 0$ ($g_i = 1$).

(2)

Let $x_i^{(0)}$ and $y_i^{(0)}$ denote the values of x and y for which $g_i = 0$. The best fitting linear spline model to estimate the conditional expectation $f_0(x) = E[Y|G = 0, X = x]$ (with two knots at $x = 5$ and $x = 7$) is found as follows. The basis functions of the linear space are $h1 = 1$, $h2 = x^{(0)}$, $h3 = (x^{(0)} - 5)_+$ (such that $h3 = 0$ if $x^{(0)} < 5$), and $h4 = (x^{(0)} - 7)_+$ (such that $h4 = 0$ if $x^{(0)} < 7$). The estimated coefficients are: -0.360 (for $h1$), -0.162 (for $h2$), 9.678 (for $h3$), -9.175 (for $h4$). The fitted regression line overlapped with the observed data points are plotted in Figure 7, where knots are indicated at $x = 5$ and $x = 7$.

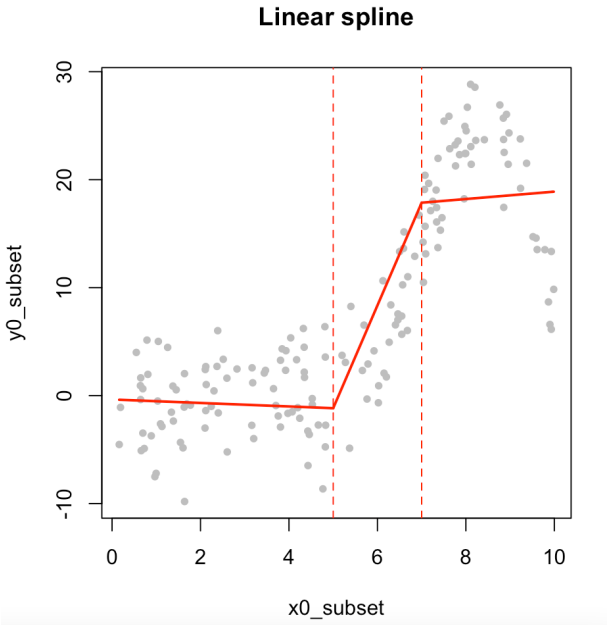

Figure 7: Fitted linear spline model regression line on data points (x_i, y_i) for which $g_i = 0$ overlapped with the actual observed data points.

(3)

We still consider the data points $x_i^{(0)}$ and $y_i^{(0)}$ for which x and y have value $g_i = 0$. The best fitting cubic spline model to estimate the conditional expectation $f_0(x) = E[Y|G = 0, X = x]$ (with one knot at $x = 5$) is found as follows. The basis functions of the linear space are $h1 = 1$, $h2 = x^{(0)}$, $h3 = x^{(0)2}$, $h4 = x^{(0)3}$, $h5 = (x^{(0)} - 5)_+^3$ (such that $h5 = 0$ if $x^{(0)} < 5$). The estimated coefficients are: -7.836 (for $h1$), 11.652 (for $h2$), -4.620 (for $h3$), 0.520 (for $h4$), and -1.258 (for $h5$). The fitted regression line overlapped with the observed data points are plotted in Figure 8, where a knot is indicated at $x = 5$.

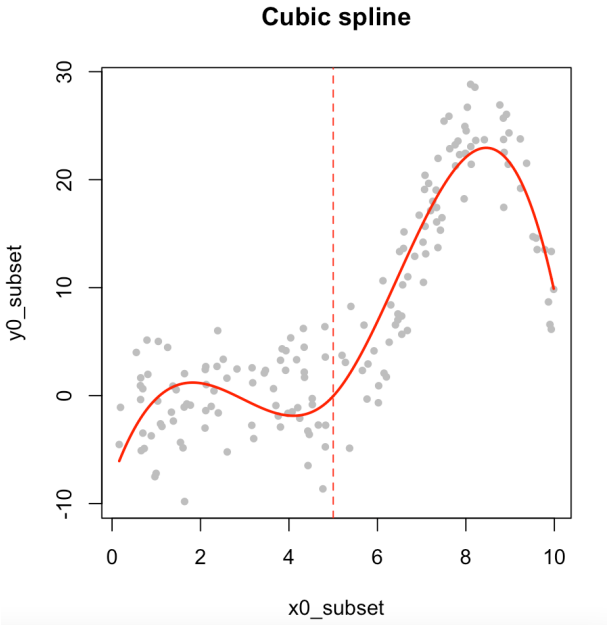

Figure 8: Fitted cubic spline model regression line on data points (x_i, y_i) for which $g_i = 0$ overlapped with the actual observed data points.

(4)

We still consider the data points $x_i^{(0)}$ and $y_i^{(0)}$ for which x and y have value $g_i = 0$. Three smoothing splines are fitted to estimate the conditional expectation $f_0(x) = E[Y|G = 0, X = x]$ with smoothing parameter $\lambda = 10^{-6}$, $\lambda = 10^{-4}$, and $\lambda = 10^{-2}$. The fitted regression line overlapped with the observed data points are plotted in Figures 9, 10, and 11.

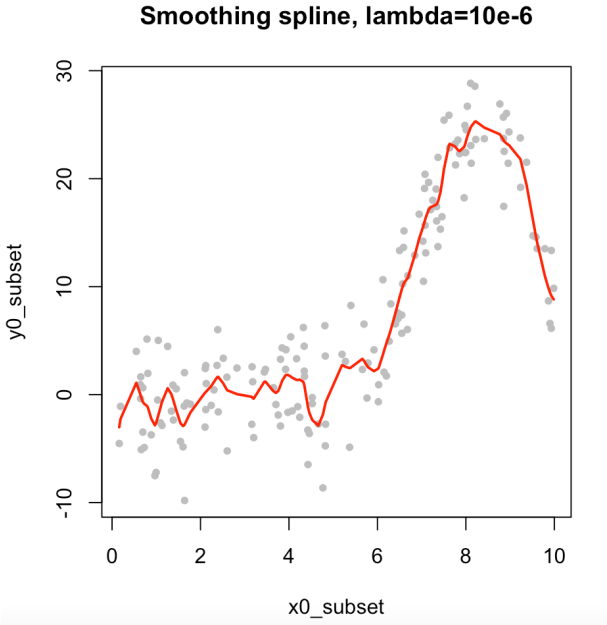

Figure 9: Fitted smoothing spline model regression line (smoothing parameter $\lambda = 10^{-6}$ on data points (x_i, y_i) for which $g_i = 0$ overlapped with the actual observed data points.

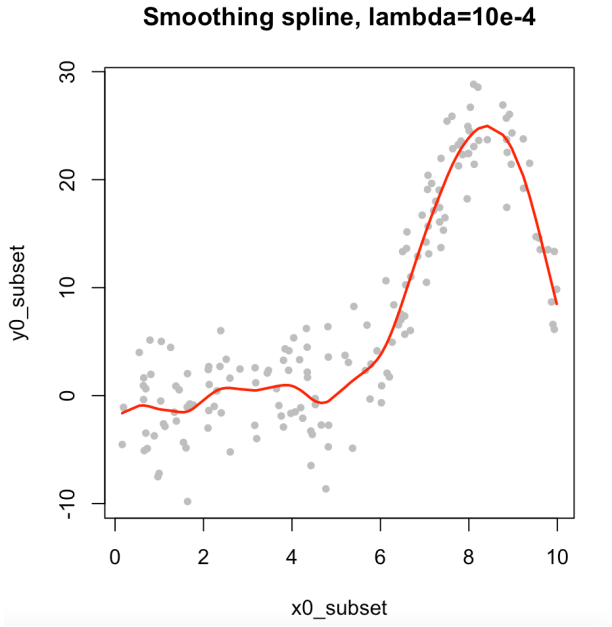

Figure 10: Fitted smoothing spline model regression line (smoothing parameter $\lambda = 10^{-4}$ on data points (x_i, y_i) for which $g_i = 0$ overlapped with the actual observed data points.

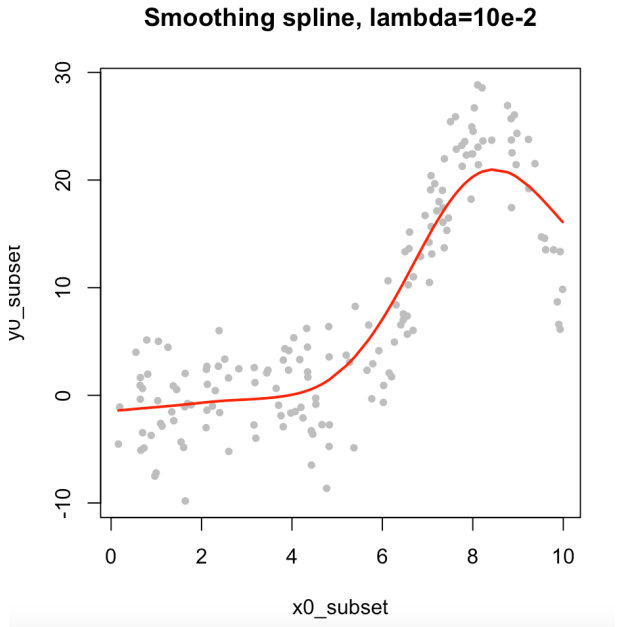

Figure 11: Fitted smoothing spline model regression line (smoothing parameter $\lambda = 10^{-2}$ on data points (x_i, y_i) for which $g_i = 0$ overlapped with the actual observed data points.

(5)

Now we consider the whole data set. Since Figure 6 implies a similar conditional expectation $f_0(x) = f_1(x)$ (where $f_i(x) = E[Y|G = i, X = x]$) for $x5$ and a different conditional expectation between the two for $x > 5$, we fit a cubic spline to f_0 and f_1 each, such that $f_0(x) = f_1(x)$ for $x5$.

That is, the data points are fitted using the basis functions $h1 = 1$, $h2 = x$, $h3 = x^2$, $h4 = x^3$,

$h5 = (x-5)^3 * (1-g)$ (for data points such that $g = 0$, and $h5 = 0$ if $x < 5$), and $h6 = (x-5)^3 * g$) (for data points such that $g = 1$, and $h6 = 0$ if $x < 5$). The estimated coefficients are: -3.605 (for $h1$), 6.048 (for $h2$), -2.906 (for $h3$), 0.382 (for $h4$), -1.128 (for $h5$), and -1.307 (for $h6$). The fitted regression lines overlapped with the observed data points are plotted in Figure 12, where a knot is indicated at $x = 5$. The red line indicates values with $x > 5$ such that $g = 0$, whereas the blue line indicates values with $x > 5$ such that $g = 1$.

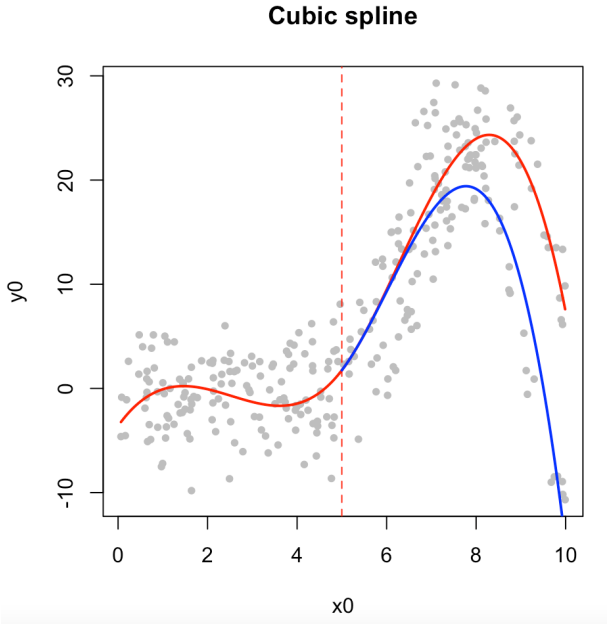

Figure 12: Fitted cubic spline model regression line on data points (x_i, y_i) for which $x < 5$ irrespective of g, and for which $x > 5$ with either $g = 0$ (red) or $g = 1$ (blue).

Question 4 Part I.

(1)

A visualization of the data is displayed in Figure 13. With x_{i1} and x_{i2} on the axes for $i = 1, ..., n$, crosses denote data points for which the corresponding value $y_i = 0$, whereas circles denote data points for which $y_i = 1$.

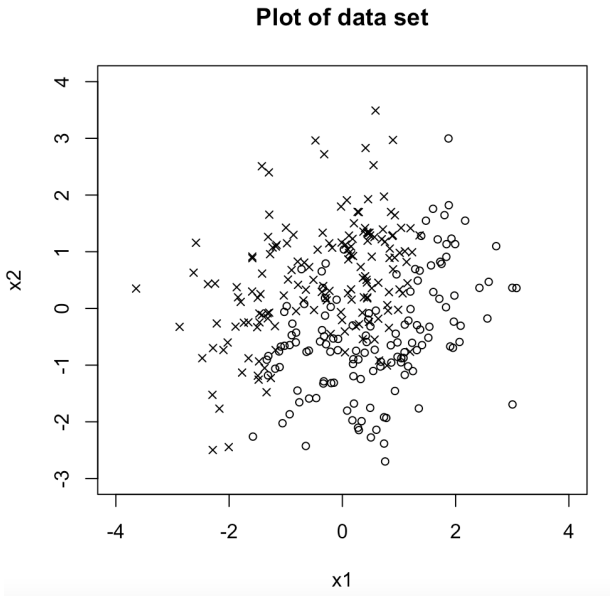

Figure 13: Visualization of the data set, where crosses (circles) denote data points (x_{i1}, x_{i2}) for which $y_i = 0$ ($y_i = 1$).

(2)

The mean μ_0 and covariance matrix Σ_0 for observations (x_{i1}, x_{i2}) which follow a $N(\mu_0, \Sigma_0)$ distribution when $y_i = 0$ are estimated using Maximum Likelihood as follows:

$$\hat{\mu}_0 = \begin{bmatrix} 0.581 \\ -0.482 \end{bmatrix}$$

$$\hat{\Sigma}_0 = \begin{bmatrix} 1.175 & 0.471 \\ 0.471 & 1.033 \end{bmatrix}$$

The mean μ_1 and covariance matrix Σ_1 for observations (x_{i1}, x_{i2}) which follow a $N(\mu_1, \Sigma_1)$ distribution when $y_i = 1$ are estimated using Maximum Likelihood as follows:

$$\hat{\mu}_1 = \begin{bmatrix} -0.376 \\ 0.508 \end{bmatrix}$$

$$\hat{\Sigma}_1 = \begin{bmatrix} 1.193 & 0.438 \\ 0.438 & 1.007 \end{bmatrix}$$

(3)

The predictions $\hat{y}_i$ for all the data points (x_{i1}, x_{i2}), $i = 1, ..., n$, are plotted in Figure 14. Crosses denote data points for which $y_i = 0$, whereas circles denote data points with $y_i = 1$. Moreover, blue data points denote predictions $\hat{y}_i = 0$, whereas red data points denote predictions $\hat{y}_i = 1$. Hence, blue circles and red crosses are misclassified data points according to the prediction.

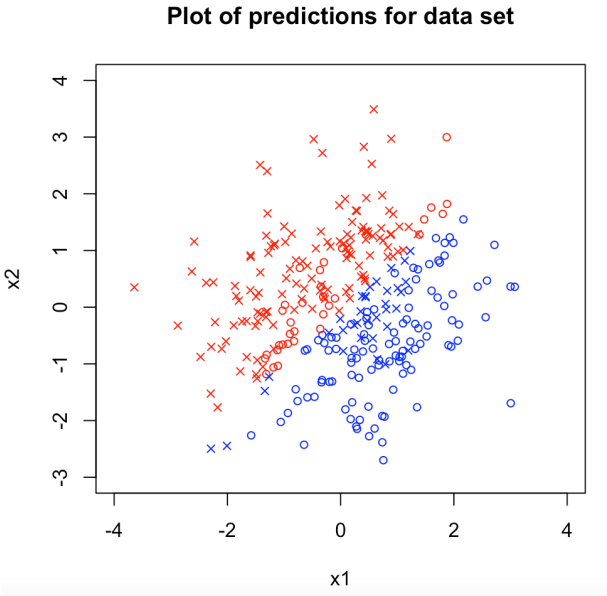

Figure 14: Predictions $\hat{y}_i$ for (x_{i1}, x_{i2}), $i = 1, ..., n$, using a parametric approach.

(4)

The amount of misclassified data points equals 66.

Question 4 Part II.

(1)

The estimated values are: $\hat{f}_0(1,1) = 0.035$ and $\hat{f}_1(0,1) = 0.124$.

(2)

The predictions $\hat{y}_i$ for all the data points (x_{i1}, x_{i2}), $i = 1, ..., n$, are plotted in Figure 15. Crosses denote data points for which $y_i = 0$, whereas circles denote data points with $y_i = 1$. Moreover, blue data points denote predictions $\hat{y}_i = 0$, whereas red data points denote predictions $\hat{y}_i = 1$. Hence, blue circles and red crosses are misclassified data points according to the prediction.

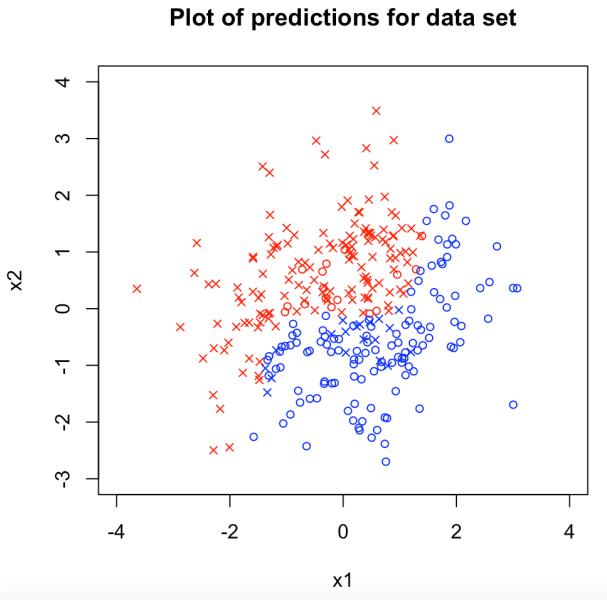

Figure 15: Predictions $\hat{y}_i$ for (x_{i1}, x_{i2}), $i = 1, ..., n$, using a non-parametric approach.

(3)

The amount of misclassified data points equals 35.