Paul Le Tran
plt377
06/04/22
ECO388E

Problem Set 5 (Problem Set 2 for spring 2022)

1ia). We are estimating $E[wage|education, gender]$ due to uncertainty about the original wording of the problem. We first perform the estimation choosing optimal bandwidth using Silverman's rule-of-thumb. This gives us a bandwidth of $h = 0.3837$. Figure 1 displays the results.

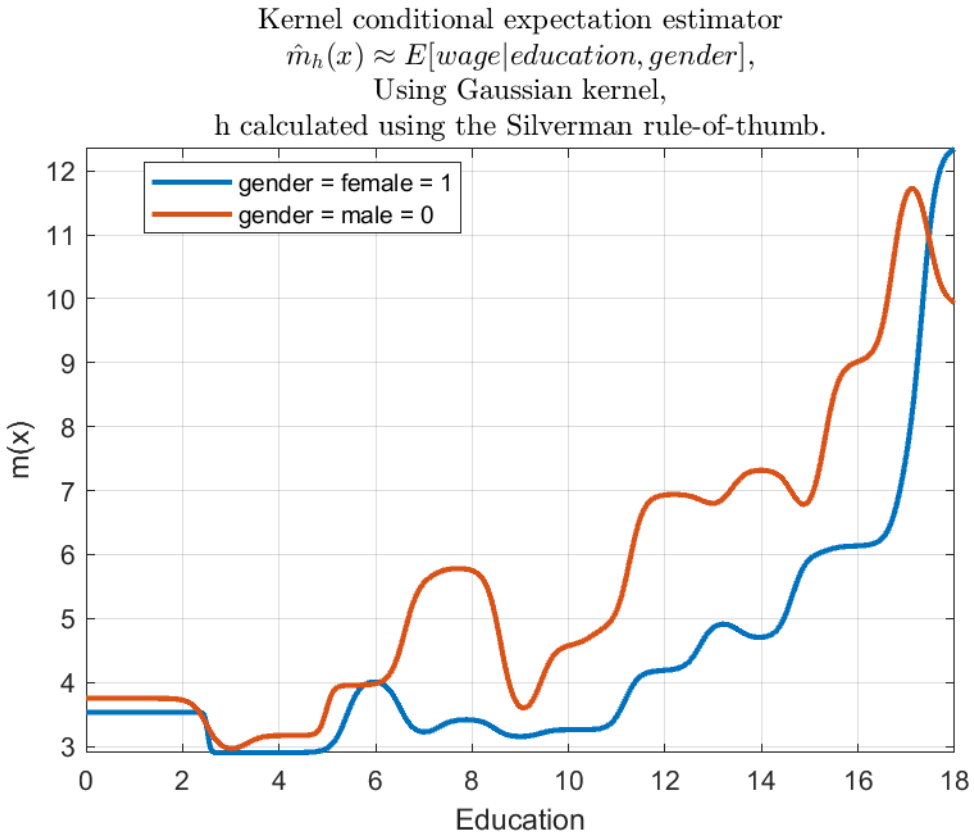

Figure 1: Kernel estimation of conditional expectation $\widehat{m_h}(x) \approx E[wage|education, gender]$. Bandwidth h is chosen by Silverman's rule-of-thumb.

1ib). Alternatively, we perform the kernel conditional expectation estimation using optimal bandwidth calculated with cross-validation. Specifically, we choose to split the sample of education into two equally-sized subsets, S_1 and S_2. The observations in both subsets were randomly assigned. Using the Gaussian kernel, we then created the kernel density equation with S_1 as the training set, and performed maximum likelihood estimation (MLE) for the testing subset, S_2, to find optimal bandwidth $h_{S_2} = 0.1850$. Similarly, we performed the same exercise with S_2 as the training set and S_1 as the testing subset, giving $h_{S_1} = 0.7538$.

For the actual kernel conditional expectation estimation, we took the average of the two bandwidths ($h = 0.4694$) as the chosen bandwidth for the estimation. Figure 2 displays the results.

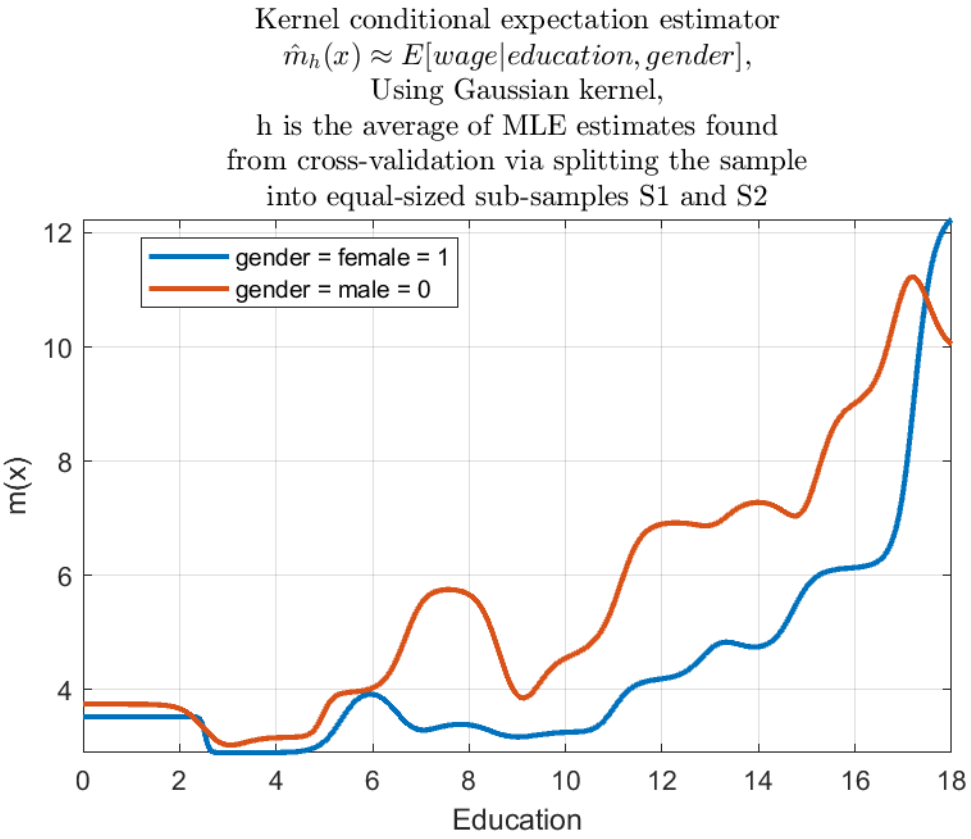

Figure 2: Kernel estimation of conditional expectation $\widehat{m_h}(x) \approx E[wage|education, gender]$. Estimated using the Gaussian kernel. Bandwidth h is the average of the MLE estimates found from cross-validation via splitting the sample into equally sized sub-samples S_1 and S_2.

1iia). We are estimating $E[wage|\text{experience}, gender]$ due to uncertainty about the original wording of the problem. We first perform the estimation choosing optimal bandwidth using Silverman's rule-of-thumb. This gives us a bandwidth of $h = 3.4889$. Figure 3 displays the results.

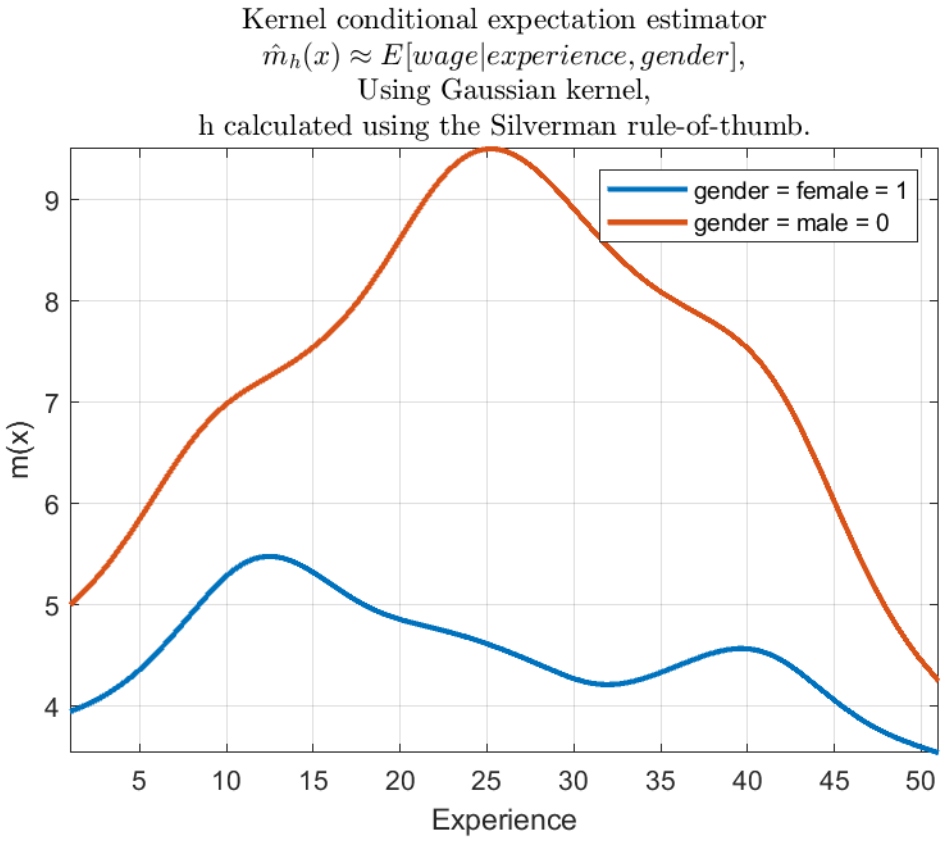

Figure 3: Kernel estimation of conditional expectation $\widehat{m_h}(x) \approx E[wage|\text{experience}, gender]$. Bandwidth h is chosen by Silverman's rule-of-thumb.

1iib). Alternatively, we perform the kernel conditional expectation estimation using optimal bandwidth calculated with cross-validation. Specifically, we choose to split the sample of education into two equally-sized subsets, S_1 and S_2. The observations in both subsets were randomly assigned. Using the Gaussian kernel, we then created the kernel density equation with S_1 as the training set, and performed maximum likelihood estimation (MLE) for the testing subset, S_2, to find optimal bandwidth $h_{S_2} = 1.4143$. Similarly, we performed the same exercise with S_2 as the training set and S_1 as the testing subset, giving $h_{S_1} = 0.1223$.

For the actual kernel conditional expectation estimation, we took the average of the two bandwidths ($h = 0.7688$) as the chosen bandwidth for the estimation. Figure 4 displays the results.

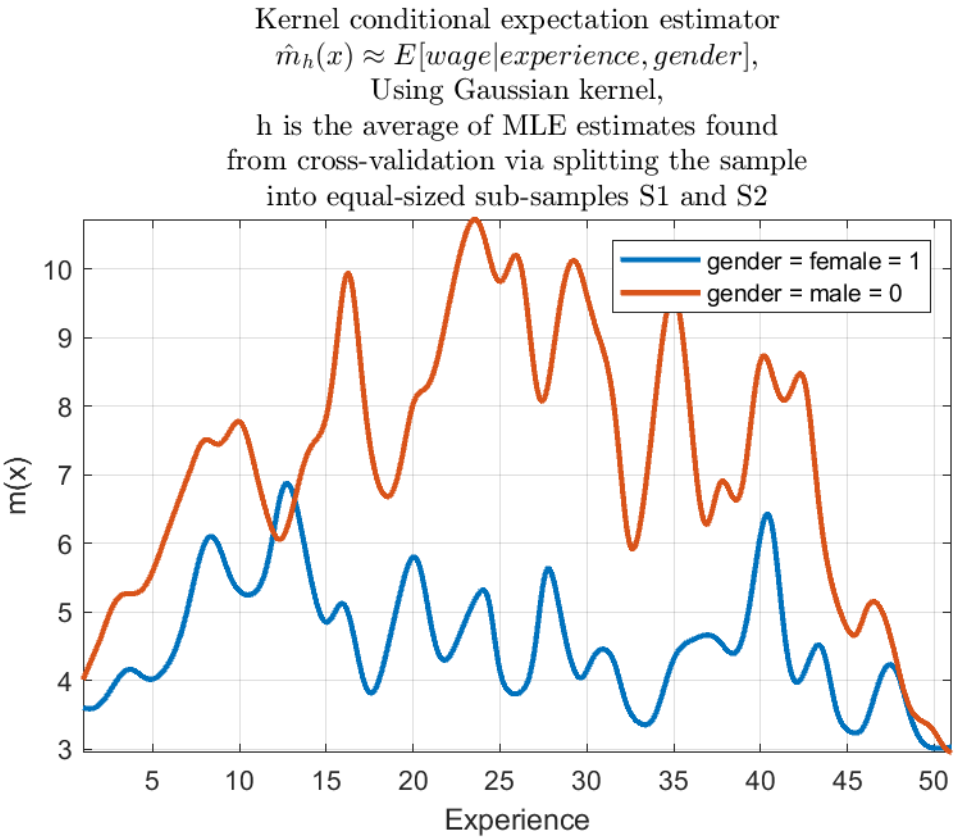

Figure 4: Kernel estimation of conditional expectation $\widehat{m_h}(x) \approx E[wage|\text{experience}, gender]$. Estimated using the Gaussian kernel. Bandwidth h is the average of the MLE estimates found from cross-validation via splitting the sample into equally sized sub-samples S_1 and S_2.

1iiia). We are estimating $E[wage|\text{tenure}, \text{married}]$ due to uncertainty about the original wording of the problem. We first perform the estimation choosing optimal bandwidth using Silverman's rule-of-thumb. This gives us a bandwidth of $h = 1.3429$. Figure 5 displays the results.

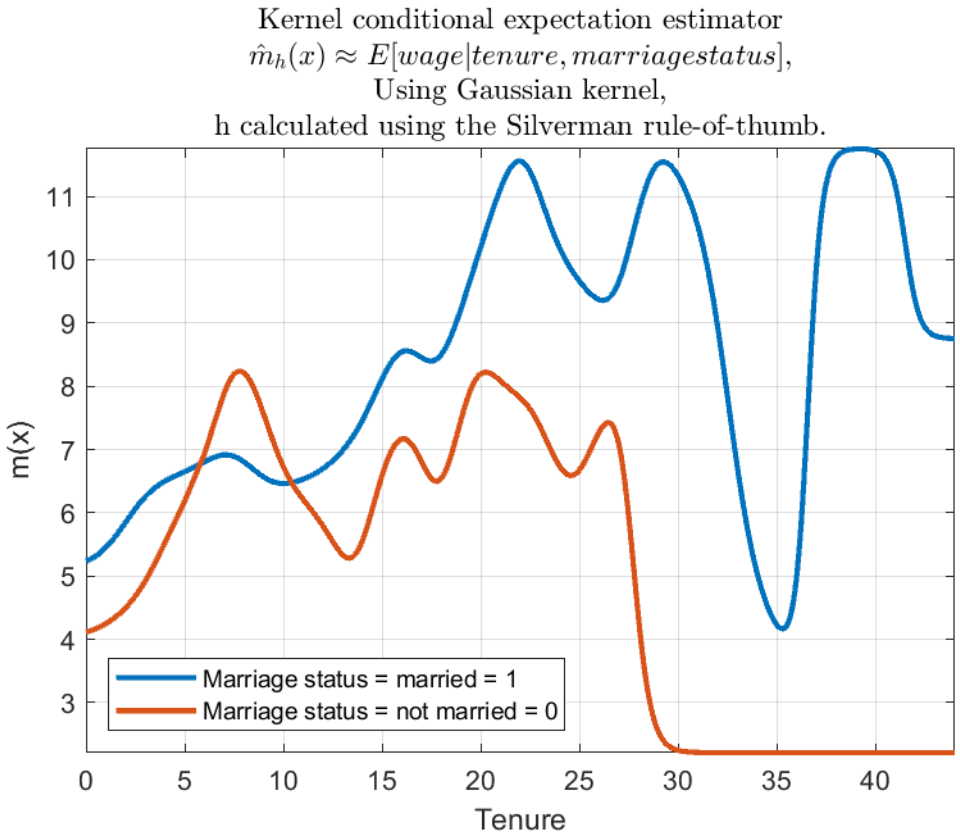

Figure 5: Kernel estimation of conditional expectation $\widehat{m_h}(x) \approx E[wage|\text{tenure}, \text{married}]$. Bandwidth h is chosen by Silverman's rule-of-thumb.

1iiib). Alternatively, we perform the kernel conditional expectation estimation using optimal bandwidth calculated with cross-validation. Specifically, we choose to split the sample of education into two equally-sized subsets, S_1 and S_2. The observations in both subsets were randomly assigned. Using the Gaussian kernel, we then created the kernel density equation with S_1 as the training set, and performed maximum likelihood estimation (MLE) for the testing subset, S_2, to find optimal bandwidth $h_{S_2} = 0.8510$. Similarly, we performed the same exercise with S_2 as the training set and S_1 as the testing subset, giving $h_{S_1} = 0.8834$.

For the actual kernel conditional expectation estimation, we took the average of the two bandwidths ($h = 0.8672$) as the chosen bandwidth for the estimation. Figure 6 displays the results.

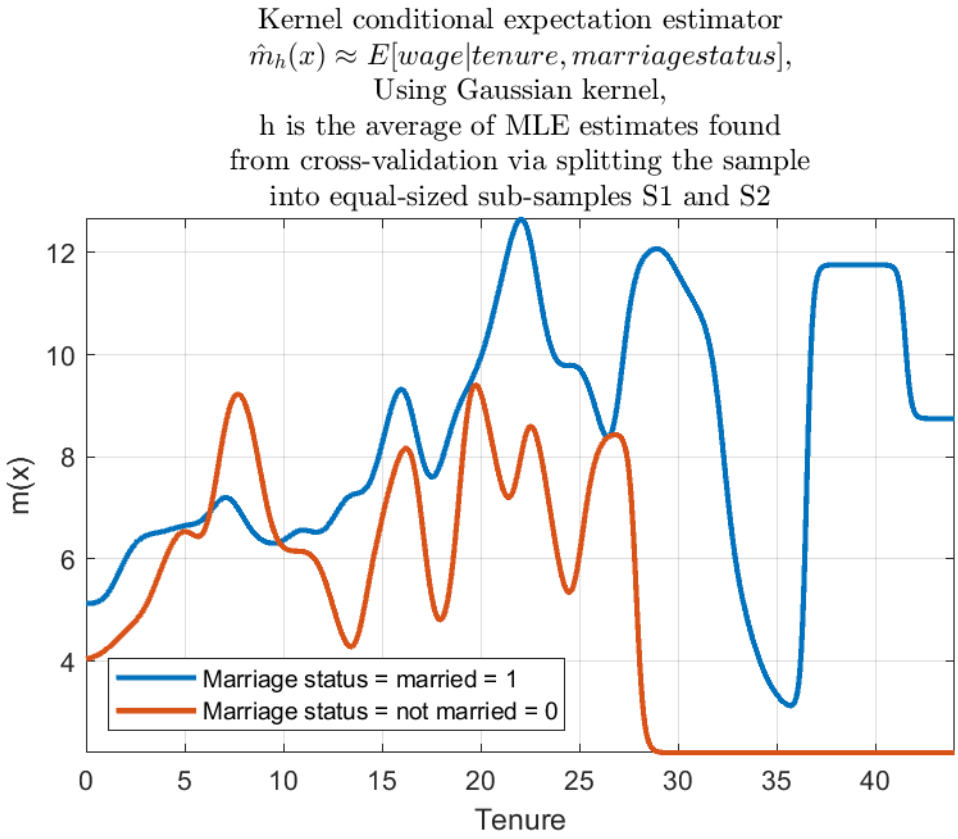

Figure 6: Kernel estimation of conditional expectation $\widehat{m_h}(x) \approx E[wage|\text{tenure}, \text{married}]$. Estimated using the Gaussian kernel. Bandwidth h is the average of the MLE estimates found from cross-validation via splitting the sample into equally sized sub-samples S_1 and S_2.

2i). We are performing the k-nearest-neighbour-search (KNN-search) approach to estimate the conditional expectation $E[wage|education, gender]$. We are choosing $K = \lambda\sqrt{n}$ by performing cross-validation to find λ. Specifically, we split the dataset into training and testing subsets, where the former is 80% and the latter is 20% of the original dataset, respectively. For each education and gender value in the testing subset, we then perform the KNN-search approach using the wages associated with the found neighbours in the training subset. We finally calculate the mean squared error, which we use as our objective function for minimisation. The $\lambda = 2.42$ is finally chosen through this minimisation problem. Figure 7 displays the KNN-search approach estimation results after performing cross-validation.

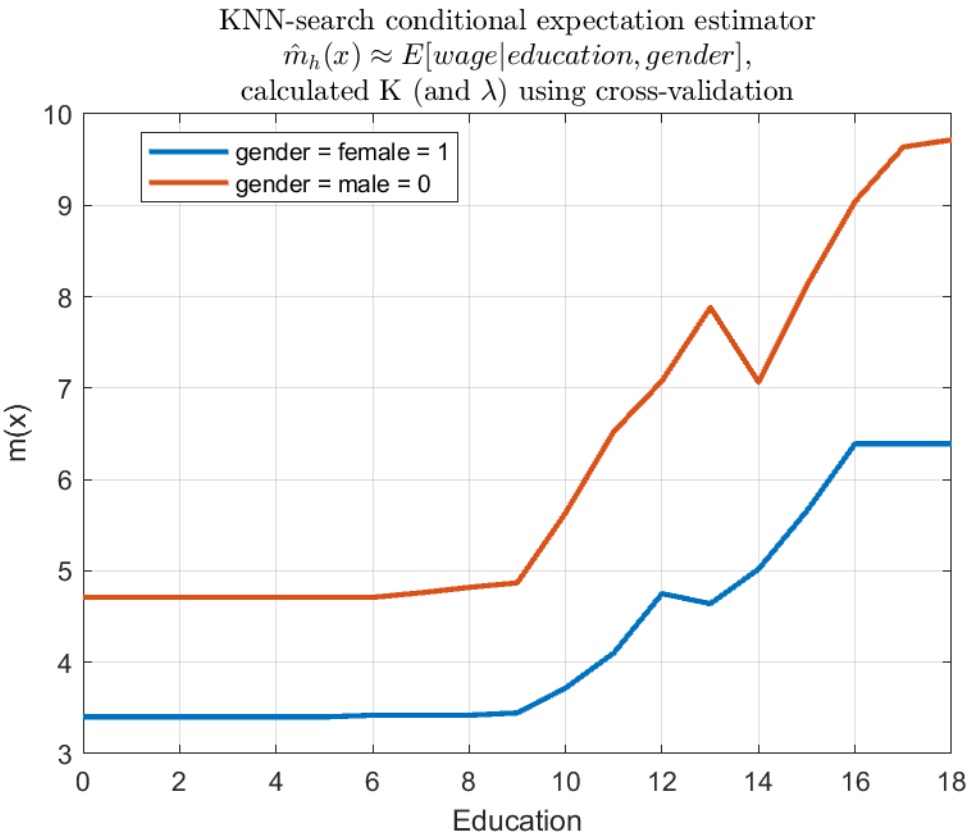

Figure 7: KNN-search approach estimation of conditional expectation $\widehat{m_h}(x) \approx E[wage|\text{education}, gender]$. Optimal λ was chosen using cross-validation.

2ii). We are performing the k-nearest-neighbour-search (KNN-search) approach to estimate the conditional expectation $E[wage|\text{experience}, gender]$. We are choosing $K = \lambda\sqrt{n}$ by performing cross-validation to find λ. Specifically, we split the dataset into training and testing subsets, where the former is 80% and the latter is 20% of the original dataset, respectively. For each education and gender value in the testing subset, we then perform the KNN-search approach using the wages associated with the found neighbours in the training subset. We finally calculate the mean squared error, which we use as our objective function for minimisation. The $\lambda = 2.365$ is finally chosen through this minimisation problem. Figure 8 displays the KNN-search approach estimation results after performing cross-validation.

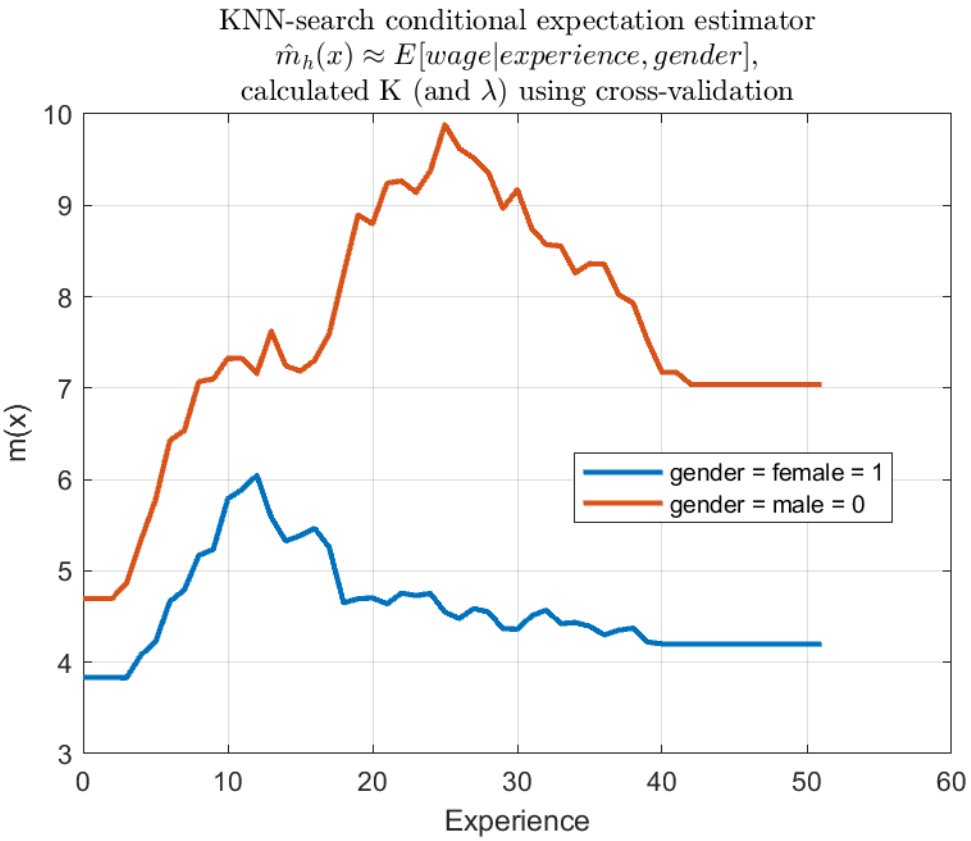

Figure 8: KNN-search approach estimation of conditional expectation $\widehat{m_h}(x) \approx E[wage|\text{experience}, gender]$. Optimal λ was chosen using cross-validation.

2iii). We are performing the k-nearest-neighbour-search (KNN-search) approach to estimate the conditional expectation $E[wage|\text{tenure}, \text{married}]$. We are choosing $K = \lambda\sqrt{n}$ by performing cross-validation to find λ. Specifically, we split the dataset into training and testing subsets, where the former is 80% and the latter is 20% of the original dataset, respectively. For each education and gender value in the testing subset, we then perform the KNN-search approach using the wages associated with the found neighbours in the training subset. We finally calculate the mean squared error, which we use as our objective function for minimisation. The $\lambda = 0.4465$ is finally chosen through this minimisation problem. Figure 9 displays the KNN-search approach estimation results after performing cross-validation.

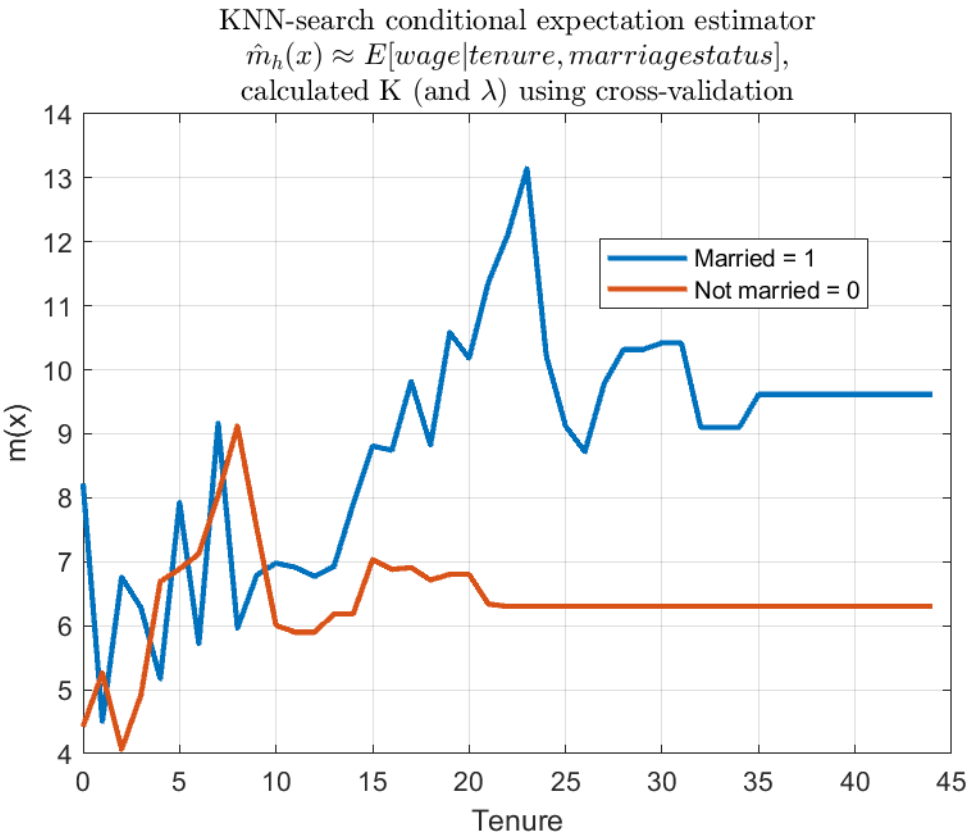

Figure 9: KNN-search approach estimation of conditional expectation $\widehat{m_h}(x) \approx E[wage|\text{tenure}, \text{married}]$. Optimal λ was chosen using cross-validation.