Advanced Econometrics I (Econ 273) Assignment 1
Lin Fan
November 1, 2017

Problem 1

Assume \((x_i, y_i)\) are i.i.d. where \(x_i \in \mathbb{R}^K\), and for the “true” \(\theta_0\), assume \(\epsilon_i = y_i - \theta_0'x_i\) are continuous random variables with density \(f_\epsilon\). We do not make any assumptions about restricting estimates \(\hat{\theta}\) to some compact subset of \(\mathbb{R}^K\) containing \(\theta_0\). Hence, we use a slightly more complicated approach. Consider the approximate first order condition:

\[
\frac{1}{2n} \sum_{i=1}^{n} x_i \text{sgn}(y_i - \theta'x_i) \overset{a.s.}{=} \frac{1}{n} \sum_{i=1}^{n} x_i (1/2 - \mathbb{1}(y_i - \theta'x_i \leq 0)) = 0
\]

where the almost sure equality holds for every fixed \(\theta\). Suppose that

\[
\hat{\theta} = \arg\min_{\theta} \frac{1}{n} \sum_{i=1}^{n} |y_i - \theta'x_i|.
\]

Then considering the elements of the subgradient of the argmin expression at the minimizer \(\hat{\theta}\), we have

\[
\left\| \frac{1}{n} \sum_{i=1}^{n} x_i \text{sgn}(y_i - \hat{\theta}'x_i) \right\| \leq \left\| \frac{1}{n} \sum_{i=1}^{n} x_i \mathbb{1}(y_i - \hat{\theta}'x_i = 0) \right\| \leq \left( \sum_{i=1}^{n} \mathbb{1}(y_i - \hat{\theta}'x_i = 0) \right) \frac{\max_{i=1}^{n} \|x_i\|}{n} \overset{a.s.}{\leq} K \frac{\max_{i=1}^{n} \|x_i\|}{n} = o_p(1/\sqrt{n})
\]

We may use any norm \(\|\cdot\|\) since all norms on finite-dimensional vector spaces are equivalent. The last equality is due to the fact that for any \(\epsilon > 0\),

\[
P \left( \max_{i=1}^{n} \|x_i\| > \sqrt{n\epsilon} \right) \leq nP \left( \|x_1\| > \sqrt{n\epsilon} \right) = \mathbb{E} \left[ n \mathbb{1} (\|x_1\|^2 > n\epsilon^2) \right] \leq \mathbb{E} \left[ (\|x_1\|/\epsilon)^2 \mathbb{1}(\|x_1\|/\epsilon)^2 > n \right] \downarrow 0
\]

as \(n \to \infty\) by the Dominated Convergence Theorem assuming that \(\mathbb{E} \left[ \|x_1\|^2 \right] < \infty\).
Now we show that
\[
\sup_{\theta \in \mathbb{R}^K} \left| \frac{1}{n} \sum_{i=1}^n x_i \text{sgn}(y_i - \theta' x_i) - \mathbb{E} [x_1 \text{sgn}(y_1 - \theta' x_1)] \right| = o_P(1).
\]

We use some empirical process VC results. By Lemma 2.6.15 of van der Vaart and Wellner (Weak Convergence and Empirical Processes), the class of functions
\[
\{(x, y) \mapsto y - \theta' x : \theta \in \mathbb{R}^K\}
\]
is a VC-class since it is a subcollection of the collection of linear combinations of the \(K+1\) coordinate-wise projections \(\mathbb{R}^{K+1} \rightarrow \mathbb{R}\). Thus, because \(\text{sgn}\) is a monotone function, by the VC-class-preserving transformations in Lemma 2.6.18 of van der Vaart and Wellner, the class of functions
\[
\mathcal{F} = \{(x, y) \mapsto x \text{sgn}(y - \theta' x) : \theta \in \mathbb{R}^K\}
\]
is a VC-class. Note that \((x, y) \mapsto \|x\|_1 (\|\cdot\|_1\) is the vector 1-norm\) is a coordinate-wise envelope function for the collection \(\mathcal{F}\). Therefore, Theorem 2.6.7 of van der Vaart and Wellner gives
\[
\sup_Q N(\epsilon \|\mathcal{F}\|_{L^1(Q)}, \mathcal{F}, L^1(Q)) \leq CV(\mathcal{F})(16e)^{V(\mathcal{F})(1/\epsilon)}^{V(\mathcal{F})-1} < \infty
\]
for every \(\epsilon \in (0, 1)\), where the supremum is taken over all probability measures \(Q\), \(C\) is a constant, \(V(\mathcal{F})\) is the VC-index of the set of subgraphs of collection \(\mathcal{F}\), and \(N\) is covering number. Assuming the \(x_i\) are integrable random vectors, \(\mathcal{F}\) is a so-called \(P\)-Glivenko Cantelli class and we obtain the stronger almost sure convergence coordinate-wise (under the measure \(P\) from which the \((x_i, y_i)\) are drawn):
\[
\sup_{\theta \in \mathbb{R}^K} \left| \frac{1}{n} \sum_{i=1}^n x_i \text{sgn}(y_i - \theta' x_i) - \mathbb{E} [x_1 \text{sgn}(y_1 - \theta' x_1)] \right| \overset{a.s.}{\rightarrow} 0.
\]

It only remains for us to show that for any \(\epsilon > 0\),
\[
\inf_{\theta \in \mathbb{R}^K, \|\theta - \theta_0\| \geq \epsilon} \left\| \mathbb{E} [x_1 \text{sgn}(y_1 - \theta' x_1)] \right\| > 0 = \left\| \mathbb{E} [x_1 \text{sgn}(y_1 - \theta_0' x_1)] \right\|,
\]
or equivalently (by the almost sure equality for the approximate first order condition mentioned at the beginning),
\[
\inf_{\theta \in \mathbb{R}^K, \|\theta - \theta_0\| \geq \epsilon} \left\| \mathbb{E} [x_1 (1/2 - \mathbb{1}(y_1 - \theta' x_1 \leq 0))] \right\| > 0 = \left\| \mathbb{E} [x_1 (1/2 - \mathbb{1}(y_1 - \theta_0' x_1 \leq 0))] \right\|.
\]
Once we show this, we can directly apply Theorem 5.9 of van der Vaart (Asymptotic Statistics) to obtain weak consistency: $\hat{\theta} \Rightarrow \theta_0$. We get the following under the assumption that $\mathbb{P}(\epsilon_1 \leq 0 \mid x_1) = 1/2$.

$$
\mathbb{E} \left[ x_1 (1/2 - \mathbb{I}(y_1 - \theta'_0 x_1 \leq 0)) \right] = \mathbb{E} \left[ x_1 \mathbb{E}[1/2 - \mathbb{I}(y_1 - \theta'_0 x_1 \leq 0) \mid x_1] \right] \\
= \mathbb{E} \left[ x_1 (1/2 - \mathbb{P}(y_1 - \theta'_0 x_1 \leq 0 \mid x_1)) \right] \\
= \mathbb{E} \left[ x_1 (1/2 - \mathbb{P}(\epsilon_1 \leq 0 \mid x_1)) \right] \\
= 0
$$

Now we return to the minimization of the function: $\theta \mapsto \mathbb{E} \left[ |y_1 - \theta' x_1| \right]$. The first order condition is obtainable by differentiation, which we carried out in class.

$$
\frac{\partial}{\partial \theta} \mathbb{E} \left[ |y_1 - \theta' x_1| \right] = \frac{\partial}{\partial \theta} \mathbb{E}_{x_1} \left[ \mathbb{E}[|y_1 - \theta' x_1| \mid x_1] \right] \\
= \frac{\partial}{\partial \theta} \mathbb{E}_{x_1} \left[ \int |y - \theta' x_1| f_\epsilon(y - \theta'_0 x_1 \mid x_1) dy \right] \\
= \frac{\partial}{\partial \theta} \mathbb{E}_{x_1} \left[ \int_{\theta' x_1}^{\infty} (y - \theta' x_1) f_\epsilon(y - \theta'_0 x_1 \mid x_1) dy + \int_{-\infty}^{\theta' x_1} (\theta' x_1 - y) f_\epsilon(y - \theta'_0 x_1 \mid x_1) dy \right] \\
= \mathbb{E}_{x_1} \left[ \frac{\partial}{\partial \theta} \int_{\theta' x_1}^{\infty} (y - \theta' x_1) f_\epsilon(y - \theta'_0 x_1 \mid x_1) dy + \frac{\partial}{\partial \theta} \int_{-\infty}^{\theta' x_1} (\theta' x_1 - y) f_\epsilon(y - \theta'_0 x_1 \mid x_1) dy \right] \\
(\text{assume regularity such that } \mathbb{E}_{x_1} [\cdot] \text{ and } \frac{\partial}{\partial \theta} \text{ may be interchanged, which can typically be justified using a combination of the mean value theorem applied to finite differences, and then the Dominated Convergence Theorem to get limit and integral interchange}) \\
= \mathbb{E}_{x_1} \left[ x_1 (\mathbb{P}(y_1 \leq \theta' x_1 \mid x_1) - \mathbb{P}(y_1 > \theta' x_1 \mid x_1)) \right] \\
= -2 \mathbb{E} \left[ x_1 (1/2 - \mathbb{I}(y_1 - \theta' x_1 \leq 0)) \right]
$$

Thus, the first order condition

$$
\frac{\partial}{\partial \theta} \mathbb{E} \left[ |y_1 - \theta' x_1| \right] = 0
$$

translates into

$$
\mathbb{E} \left[ x_1 (1/2 - \mathbb{I}(y_1 - \theta' x_1 \leq 0)) \right] = 0.
$$

Continuing the differentiation, under regularity conditions analogous to those required earlier concerning interchange of integration and differentiation, we have

$$
\frac{\partial^2}{\partial \theta \partial \theta'} \mathbb{E} \left[ |y_1 - \theta' x_1| \right] \Bigg|_{\theta = \theta_0} = \mathbb{E} \left[ x_1 x_1' f_\epsilon(0 \mid x_1) \right].
$$
Assuming this Hessian is nonsingular, we see that it then must have positive eigenvalues, which implies that it is symmetric, positive-definite. Thus, $\theta_0$ is a strict local minimum of the function $\theta \mapsto \mathbb{E} [ |y_1 - \theta' x_1| ]$. Because $\theta \mapsto \mathbb{E} [ |y_1 - \theta' x_1| ]$ is a convex function, $\theta_0$ is actually the unique global minimizer of the function. This translates into the statement that

$$\| \mathbb{E} [x_1 (1/2 - \mathbb{I} (y_1 - \theta' x_1 \leq 0))] \| > 0 = \| \mathbb{E} [x_1 (1/2 - \mathbb{I} (y_1 - \theta_0' x_1 \leq 0))] \|$$

for any $\theta \neq \theta_0$. Finally, continuity of the function $\theta \mapsto \| \mathbb{E} [x_1 (1/2 - \mathbb{I} (y_1 - \theta' x_1 \leq 0))] \|$ (which may be justified using the Dominated Convergence Theorem) yields, for any $\epsilon > 0$,

$$\inf_{\theta \in \mathbb{R}^K, \| \theta - \theta_0 \| \geq \epsilon} \| \mathbb{E} [x_1 (1/2 - \mathbb{I} (y_1 - \theta' x_1 \leq 0))] \| > 0 = \| \mathbb{E} [x_1 (1/2 - \mathbb{I} (y_1 - \theta_0' x_1 \leq 0))] \| .$$

Our proof is thus complete.
Problem 2

Part a

Assume we have at least 50 moments of the i.i.d. $X_i$ random variables. For $p = 3, 5, 7$, we may apply Taylor’s Theorem with Lagrange remainder to obtain:

$$
\frac{1}{n} \sum_{i=1}^{n} (X_i - \tilde{\mu})^p = \frac{1}{n} \sum_{i=1}^{n} (X_i - \mu)^p + p(\mu - \tilde{\mu}) \frac{1}{n} \sum_{i=1}^{n} (X_i - \mu)^{p-1} + \frac{p(p-1)}{2} (\mu - \tilde{\mu})^2 \frac{1}{n} \sum_{i=1}^{n} (X_i - \tilde{\mu})^{p-2}
$$

where $\tilde{\mu}$ is between $\mu$ and $\hat{\mu}$. We claim that

$$
\frac{1}{n} \sum_{i=1}^{n} (X_i - \tilde{\mu})^{p-2} = O_p(1).
$$

Why? The Strong LLN says that $\hat{\mu} \xrightarrow{a.s.} \mu$. For $n$ sufficiently large, $|\mu - \hat{\mu}| < 1$, and thus

$$
\mathbb{P} \left( \left| \frac{1}{n} \sum_{i=1}^{n} (X_i - \hat{\mu})^{p-2} \right| > M \right) \leq \mathbb{P} \left( \left| \frac{1}{n} \sum_{i=1}^{n} (|X_i - \mu| + 1)^{p-2} \right| > M \right) \leq \frac{1}{M} \mathbb{E} \left[ (|X_1 - \mu| + 1)^{p-2} \right].
$$

This last quantity can be made arbitrarily small by letting $M$ get large. By the classical CLT, $\sqrt{n}(\hat{\mu} - \mu)$ is asymptotically normal and thus,

$$
\frac{p(p-1)}{2} (\mu - \tilde{\mu})^2 \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (X_i - \tilde{\mu})^{p-2} = \frac{p(p-1)}{2} (\mu - \hat{\mu}) \cdot \sqrt{n}(\mu - \tilde{\mu}) \cdot \frac{1}{n} \sum_{i=1}^{n} (X_i - \tilde{\mu})^{p-2} = o_p(1).
$$

By the Strong LLN,

$$
\frac{1}{n} \sum_{i=1}^{n} (X_i - \mu)^{p-1} - \mathbb{E} \left[ (X_1 - \mu)^{p-1} \right] \xrightarrow{a.s.} 0.
$$

Therefore,

$$
p\sqrt{n}(\mu - \tilde{\mu}) \left( \frac{1}{n} \sum_{i=1}^{n} (X_i - \mu)^{p-1} - \mathbb{E} \left[ (X_1 - \mu)^{p-1} \right] \right) = o_p(1).
$$

Putting all of this together, we have

$$
\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left( (X_i - \hat{\mu})^p - \mathbb{E} \left[ (X_i - \mu)^p \right] \right) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left( (X_i - \mu)^p - \mathbb{E} \left[ (X_i - \mu)^p \right] \right) - p\mathbb{E} \left[ (X_1 - \mu)^{p-1} \right] \sqrt{n}(\hat{\mu} - \mu) + o_p(1).
$$
By the classical CLT, we have

\[
\frac{1}{\sqrt{n}} \begin{bmatrix}
\sum_{i=1}^{n} (X_i - \mu) \\
\sum_{i=1}^{n} (X_i - \mu)^3 - \mathbb{E}[(X_i - \mu)^3] \\
\sum_{i=1}^{n} (X_i - \mu)^5 - \mathbb{E}[(X_i - \mu)^5] \\
\sum_{i=1}^{n} (X_i - \mu)^7 - \mathbb{E}[(X_i - \mu)^7]
\end{bmatrix} \Rightarrow \mathcal{N}(0, \Sigma)
\]

where the covariance matrix \( \Sigma \in \mathbb{R}^{4 \times 4} \) is trivially derivable analytically. Consider the function

\[
f(y_1, y_2, y_3, y_4) = \begin{bmatrix}
y_2 - 3\mathbb{E}[(X_1 - \mu)^2] y_1 \\
y_3 - 5\mathbb{E}[(X_1 - \mu)^4] y_1 \\
y_4 - 7\mathbb{E}[(X_1 - \mu)^6] y_1
\end{bmatrix}.
\]

Then

\[
\nabla f(y_1, y_2, y_3, y_4) = \begin{bmatrix}
-3\mathbb{E}[(X_1 - \mu)^2] & -5\mathbb{E}[(X_1 - \mu)^4] & -7\mathbb{E}[(X_1 - \mu)^6] \\
1 & 0 & 0 \\
0 & 1 & 0 \\
0 & 0 & 1
\end{bmatrix} := T.
\]

Then by the delta method,

\[
\frac{1}{\sqrt{n}} \begin{bmatrix}
\sum_{i=1}^{n} (X_i - \hat{\mu})^3 - \mathbb{E}[(X_i - \hat{\mu})^3] \\
\sum_{i=1}^{n} (X_i - \hat{\mu})^5 - \mathbb{E}[(X_i - \hat{\mu})^5] \\
\sum_{i=1}^{n} (X_i - \hat{\mu})^7 - \mathbb{E}[(X_i - \hat{\mu})^7]
\end{bmatrix} \Rightarrow \mathcal{N}(0, T'\Sigma T).
\]

Under the null hypothesis of symmetry about the mean, we have

\[
\frac{1}{\sqrt{n}} \begin{bmatrix}
\sum_{i=1}^{n} (X_i - \hat{\mu})^3 \\
\sum_{i=1}^{n} (X_i - \hat{\mu})^5 \\
\sum_{i=1}^{n} (X_i - \hat{\mu})^7
\end{bmatrix} \Rightarrow \mathcal{N}(0, T'\Sigma T).
\]

A statistic to test for symmetry could be a linear functional of the vector

\[
\frac{1}{\sqrt{n}} \begin{bmatrix}
\sum_{i=1}^{n} (X_i - \hat{\mu})^3 \\
\sum_{i=1}^{n} (X_i - \hat{\mu})^5 \\
\sum_{i=1}^{n} (X_i - \hat{\mu})^7
\end{bmatrix}
\]
and the corresponding limiting distribution is easily obtained via the delta method or the continuous mapping theorem.

**Part b**

We consider the test statistic

\[ \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}(X_i \leq \hat{\mu}) - \frac{1}{2}, \]

Let’s rigorously derive the limiting distribution under the null hypothesis of symmetry about the true mean \( \mu \). We could use the abstract empirical process theory like in Problem 1, but here we will take a more classical approach using weak convergence in \( D[-\infty, \infty] \). Consider the random functions:

\[ \overline{Y}_n(y) = \frac{1}{n} \sum_{i=1}^{n} Y_i(y) := \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}(X_i \leq y) \]

where \( F \) is the distribution function of the \( X_i \) random variables. Then we may write the test statistic as:

\[ \overline{Y}_n(\hat{\mu}) - \frac{1}{2} = (\overline{Y}_n(\hat{\mu}) - F(\hat{\mu})) + (F(\hat{\mu}) - \frac{1}{2}). \]

We claim that

\[ \sqrt{n}(\overline{Y}_n(\mu) - F(\hat{\mu})) - \sqrt{n}(\overline{Y}_n(\mu) - \frac{1}{2}) \Rightarrow 0. \]

The classical result by Donsker for the empirical distribution function yields weak convergence in \( D[-\infty, \infty] \):

\[ \sqrt{n}(\overline{Y}_n(\cdot) - F(\cdot)) \Rightarrow B(F(\cdot)) \]

where \( B \) is a standard Brownian bridge. Thus, for any fixed \( \delta > 0 \), by the continuous mapping theorem, as \( n \to \infty \),

\[ \mathbb{E} \left[ \min \left( \sup_{|u-\mu| \leq \delta} \sqrt{n} \left| (\overline{Y}_n(u) - F(u)) - (\overline{Y}_n(\mu) - \frac{1}{2}) \right| , 1 \right) \right] \]

\[ \to \mathbb{E} \left[ \min \left( \sup_{|u-\mu| \leq \delta} \sqrt{n} |B(F(u)) - B(F(\mu))|, 1 \right) \right]. \]
Then for any \( \epsilon \in (0, 1) \), we have

\[
P \left( \sqrt{n} \left| \bar{Y}_n(\hat{\mu}) - F(\hat{\mu}) - \bar{Y}_n(\mu) - \frac{1}{2} \right| > \epsilon \right) \\
\leq P (|\hat{\mu} - \mu| > \delta) + P \left( \sup_{|u - \mu| \leq \delta} \sqrt{n} \left| \bar{Y}_n(u) - F(u) - \bar{Y}_n(\mu) - \frac{1}{2} \right| > \epsilon \right) \\
\leq P (|\hat{\mu} - \mu| > \delta) + \frac{1}{\epsilon} E \left[ \min \left( \sup_{|u - \mu| \leq \delta} \sqrt{n} \left| \bar{Y}_n(u) - F(u) - \bar{Y}_n(\mu) - \frac{1}{2} \right|, 1 \right) \right] .
\]

Sending \( n \to \infty \) first, followed by \( \delta \downarrow 0 \), and an application of the Bounded Convergence Theorem taking advantage of the almost sure path-continuity of \( B(F(\cdot)) \) (assuming \( F \) is continuous in a neighborhood of \( \mu \)), yields

\[
\limsup_{n \to \infty} P \left( \sqrt{n} \left| \bar{Y}_n(\hat{\mu}) - F(\hat{\mu}) - \bar{Y}_n(\mu) - \frac{1}{2} \right| > \epsilon \right) = 0.
\]

With the claim established, we have

\[
\sqrt{n}(\bar{Y}_n(\hat{\mu}) - \frac{1}{2}) = \sqrt{n}(\bar{Y}_n(\mu) - \frac{1}{2}) + \sqrt{n}(F(\hat{\mu}) - \frac{1}{2}) + o_p(1).
\]

Now we may apply Taylor’s Theorem, and assuming \( F \) is continuously differentiable at \( \mu \) with density \( f \), we get

\[
\sqrt{n}(\bar{Y}_n(\hat{\mu}) - \frac{1}{2}) = \sqrt{n}(\bar{Y}_n(\mu) - \frac{1}{2}) + \sqrt{n} f(\mu)(\hat{\mu} - \mu) + o_p(1) \\
\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left( \mathbb{I}(X_i \leq \mu) + f(\mu)X_i - \frac{1}{2} - f(\mu)\frac{\mu}{2} \right) + o_p(1).
\]

Therefore, by the classical CLT, we have under the null hypothesis,

\[
\sqrt{n}(\bar{Y}_n(\hat{\mu}) - \frac{1}{2}) \Rightarrow \sigma N(0, 1)
\]

where \( \sigma^2 = \text{Var}\left( \mathbb{I}(X_1 \leq \mu) + f(\mu)X_1 \right) \).
Part c

We can use GMM to improve the efficiency of estimating the mean. Assuming the null hypothesis is correct, we might for instance make use of, say, four estimating equations

\[
\frac{1}{n} \begin{bmatrix}
\sum_{i=1}^{n}(X_i - \hat{\mu}) \\
\sum_{i=1}^{n}(X_i - \hat{\mu})^3 \\
\sum_{i=1}^{n}(X_i - \hat{\mu})^5 \\
\sum_{i=1}^{n}(X_i - \hat{\mu})^7
\end{bmatrix} = 0.
\]

The sample mean only makes use of the first estimating equation, and thus, we will usually get more efficiency if we can combine all four equations in a sensible way using GMM. Of course if four estimating equations is not enough to significantly improve efficiency, and higher moments exist, we can use more odd powers for estimating equations. Under the null hypothesis, such vectors involving odd powers should be close to zero. Defining

\[
g(X_i, u) = \begin{bmatrix}
(X_i - u) \\
(X_i - u)^3 \\
(X_i - u)^5 \\
(X_i - u)^7
\end{bmatrix}
\]

we can set the GMM weighting matrix to be

\[
\widehat{W}_n(u) = \left(\frac{1}{n} \sum_{i=1}^{n} g(X_i, u)g(X_i, u)'\right)^{-1}.
\]

Then the GMM estimator is

\[
\hat{\mu} = \arg\min_{u} \left(\frac{1}{n} \sum_{i=1}^{n} g(X_i, u)\right)' \widehat{W}_n(u) \left(\frac{1}{n} \sum_{i=1}^{n} g(X_i, u)\right).
\]

We might also consider using the sample median, perhaps as part of the GMM estimation, to improve efficiency. This could be helpful if the underlying distribution has relatively heavy tails and not many moments exist.
Problem 3

We show consistency of $\tilde{\mu}$.

\begin{align*}
\tilde{\mu} &= \int x\hat{f}(x)dx \\
&= \int x \frac{1}{nh} \sum_{i=1}^{n} K \left( \frac{x - X_i}{h} \right) dx \\
&= \frac{1}{n} \sum_{i=1}^{n} \int (hy + X_i) K(y) dy \\
&= \frac{1}{n} \sum_{i=1}^{n} \left( h \int y K(y) dy + X_i \int K(y) dy \right) \\
&= \frac{1}{n} \sum_{i=1}^{n} \left( h \int y K(y) dy + X_i \right) \\
&= h \int y K(y) dy + \frac{1}{n} \sum_{i=1}^{n} X_i \\
\xrightarrow{a.s.} \mu & \quad \text{by Strong LLN as long as either } h = o(1) \text{ or } K \text{ is symmetric}
\end{align*}

We show consistency of $\tilde{m}_2$.

\begin{align*}
\tilde{m}_2 &= \int x^2\hat{f}(x)dx \\
&= \frac{1}{n} \sum_{i=1}^{n} \int (hy + X_i)^2 K(y) dy \quad \text{by our work above} \\
&= \frac{1}{n} \sum_{i=1}^{n} \int \left( h^2 y^2 + 2hyX_i + X_i^2 \right) K(y) dy \\
&= h^2 \int y^2 K(y) dy + 2h \int yK(y)dy \frac{1}{n} \sum_{i=1}^{n} X_i + \int K(y)dy \frac{1}{n} \sum_{i=1}^{n} X_i^2 \\
&= h^2 \int y^2 K(y) dy + 2h \int yK(y)dy \frac{1}{n} \sum_{i=1}^{n} X_i + \frac{1}{n} \sum_{i=1}^{n} X_i^2 \\
\xrightarrow{a.s.} m_2 & \quad \text{by Strong LLN as long as } h = o(1)
\end{align*}

Assuming we use a symmetric kernel $K$, by our work above, we have

\begin{align*}
\tilde{\mu} = h \int yK(y)dy + \frac{1}{n} \sum_{i=1}^{n} X_i = \frac{1}{n} \sum_{i=1}^{n} X_i.
\end{align*}
Therefore,
\[
\sqrt{n}(\bar{\mu} - \mu) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (X_i - \mu),
\]
which is asymptotically normal by the standard CLT. Also, assuming we use a symmetric kernel \( K \), by our work above, we have
\[
\tilde{m}_2 = h^2 \int y^2 K(y) dy + \frac{1}{n} \sum_{i=1}^{n} X_i^2.
\]
And thus,
\[
\sqrt{n}(\tilde{m}_2 - m_2) = \sqrt{nh^2} \int y^2 K(y) dy + \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (X_i^2 - m_2).
\]
Assuming \( nh^4 = o(1) \), we get \( \sqrt{nh^2} \int y^2 K(y) dy = o(1) \). Hence,
\[
\sqrt{n}(\tilde{m}_2 - m_2) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (X_i^2 - m_2) + o_p(1),
\]
which is asymptotically normal by the standard CLT and Slutsky’s Lemma.
Problem 4

Part a

First, notice that

\[
\mathbb{E} \left[ \hat{f}^{(1)}(x_0) \right] = \mathbb{E} \left[ \frac{1}{nh^2} \sum_{i=1}^{n} K^{(1)} \left( \frac{x_0 - X_i}{h} \right) \right]
\]

\[
= \frac{1}{h^2} \int K^{(1)} \left( \frac{x_0 - y}{h} \right) f(y) dy
\]

\[
= \frac{1}{h} \int K^{(1)}(y) f(x_0 - hy) dy
\]

\[
= \frac{1}{h} f(x_0 - hy) K(y) \bigg|_{y=-\infty}^{\infty} + \int K(y) f^{(1)}(x_0 - hy) dy
\]

\[
= \int K(y) f^{(1)}(x_0 - hy) dy
\]

by integration by parts and the bounded support of \( K \)

Therefore, we have

\[
\text{bias}(\hat{f}^{(1)}(x_0)) = \int K(y) f^{(1)}(x_0 - hy) dy - f^{(1)}(x_0)
\]

\[
= \int K(y) f^{(1)}(x_0 - hy) - f^{(1)}(x_0) dy
\]

\[
= \int K(y) \left[ f^{(2)}(x_0) (-hy) + \frac{1}{2} f^{(3)}(x_0) (-hy)^2 + \cdots + \frac{1}{r!} f^{(r+1)}(x_0) (-hy)^r + O(1)(-hy)^{r+1} \right] dy
\]

where the big \( O(1) \) holds uniformly in \( y \) due to Taylor’s Theorem with Lagrange remainder and the boundedness of \( f^{(r+2)} \)

\[
= \int K(y) \left[ \frac{1}{r!} f^{(r+1)}(x_0)(-hy)^r + O(1)(-hy)^{r+1} \right] dy \quad \text{since } K \text{ is of order } r
\]

\[
= \frac{1}{r!} f^{(r+1)}(x_0) h^r \int K(y) y^r dy + O(h^{r+1})
\]

\[
\approx \frac{1}{r!} f^{(r+1)}(x_0) c_K h^r \quad \text{for small } h > 0
\]
Part b

\[
\text{Var}(\hat{f}^{(1)}(x_0)) = \text{Var} \left( \frac{1}{nh^2} \sum_{i=1}^{n} K^{(1)} \left( \frac{x_0 - X_i}{h} \right) \right)
\]

\[
= \frac{1}{n} \text{Var} \left( \frac{1}{h^2} K^{(1)} \left( \frac{x_0 - X_1}{h} \right) \right)
\]

\[
= \frac{1}{n} \left[ \int \left( \frac{1}{h^2} K^{(1)} \left( \frac{x_0 - y}{h} \right) \right)^2 f(y)dy - \left( \int \frac{1}{h^2} K^{(1)} \left( \frac{x_0 - y}{h} \right) f(y)dy \right)^2 \right]
\]

\[
= \frac{1}{nh^3} \left[ \int (K^{(1)}(y))^2 f(x_0 - hy)dy - \left( \int K^{(1)}(y) f(x_0 - hy)dy \right)^2 \right]
\]

\[
= \frac{1}{nh^3} \left[ \int (K^{(1)}(y))^2 f(x_0 - hy)dy - \left( \int K^{(1)}(y) f(x_0 - hy)dy \right)^2 \right]
\]

\[
= \frac{1}{nh^3} \left[ \int (K^{(1)}(y))^2 dy f(x_0) - hf^{(1)}(x_0) + O(1)\right]
\]

where the big \(O(1)\) holds uniformly in \(y\) due to Taylor’s Theorem with Lagrange remainder and the boundedness of \(f^{(r+2)}\)

\[
\approx \frac{1}{nh^3} d_K f(x_0) \quad \text{for small } h > 0
\]

Part c

We know that \(\text{MSE} = \text{bias}^2 + \text{variance}\), so the optimal value of \(h\) is determined by the equation:

\[
\left( \frac{1}{r!} f^{(r+1)}(x_0) c_K h^r \right)^2 = \frac{1}{nh^3} d_K f(x_0).
\]

So the optimal value is

\[
h = \left( \frac{(r!)^2 d_K}{n c_K^2 f^{(r+1)}(x_0)^2} \right)^{1/(2r+3)} = O(n^{-1/(2r+3)}).
\]

Part d

We would ideally plug in the optimal value of \(h\) into the estimated bias and variance and use the expression

\[
\text{MSE} = \text{bias}^2 + \text{variance}
\]
to estimate MSE. However, the optimal value of $h$ actually involves the unknown density through $f^{(r+1)}(x_0)$ and $f(x_0)$, which in turn depend on $h$ in highly nonlinear ways. A crude option would be to simply plug in $h = n^{-1/(2r+3)}$ to get

$$\hat{\text{MSE}} = \left[ \left( \frac{\hat{f}^{(r+1)}(x_0)c_K}{r!} \right)^2 + \hat{f}(x_0)d_K \right] n^{-2r/(2r+3)}$$

where we have

$$\hat{f}^{(r+1)}(x_0) = n^{-(r+1)/(2r+3)} \sum_{i=1}^{n} K^{(r+1)} \left( \frac{x_0 - X_i}{n^{-1/(2r+3)}} \right)$$

$$\hat{f}(x_0) = n^{-(2r+2)/(2r+3)} \sum_{i=1}^{n} K \left( \frac{x_0 - X_i}{n^{-1/(2r+3)}} \right).$$

We can compute $c_K$ and $d_K$ using numerical quadrature once the kernel $K$ is chosen. More sophisticated techniques such as cross-validation can be used to pick $h$. We explore some of those methods in the next problem.
Problem 5

Parts a and b

Kernel smoothing with the standard Gaussian kernel is used for nonparametric estimation. The bandwidth is obtained using leave-one-out cross-validation, and an optimal value $h = 0.2596$ is used. The first plot below shows the marginal relationship between log family income and labor force participation probability. The second plot below shows the marginal effect on the labor force participation probability of one percent increase in family income, as a function of the log family income. Also shown in the two plots below are the probit model results. For very low and very high family incomes, the probit model predicts a lower probability than that of the nonparametric model, but at intermediate ranges, the nonparametric model oscillates significantly between lower and higher probabilities, while the probit model is monotonic as expected. The difference is not uniform across different log family income levels.
Part c

Kernel smoothing with the standard multivariate Gaussian kernel on $\mathbb{R}^3$ is used for nonparametric estimation. The bandwidth is again obtained using leave-one-out cross-validation, and an optimal value $h = 0.6734$ is used. The two plots below repeat the plots from Parts a and b above. Here, the nonparametric model that is plotted is the standard multivariate Gaussian kernel smoothing model on the regressors SES, log family income, and AFQT, but evaluated at the mean level of the SES and AFQT regressors. The mean levels of SES, log family income, and AFQT respectively are 0.0834, 10.2569, and 0.0503. The same is plotted for a logit model. The marginal effect of one percent increase in family income on the probability of labor force participation at the mean levels of all three regressors is $-1.149 \cdot 10^{-4}$ for the nonparametric model, while it is $-0.001076$ for the logit model. In the first plot below, the logit model is monotonic and similar to the probit model from Parts a and b, as expected. Also, here the nonparametric model involving three regressors is smoother than the nonparametric model involving only a single regressor, like in Parts a and b. Perhaps this has to do with using a standard multivariate Gaussian kernel on $\mathbb{R}^3$ and the bandwidth that is used (which is constrained to be the same in all three coordinate directions). If we relax the constraint and allow each coordinate, i.e., each regressor, to have its own bandwidth, then we might get a more parsimonious model. However, such a model would be more difficult to fit computationally, and so we did not pursue that here.
Part d

A standard Gaussian kernel is used to estimate the density of log family income. The bandwidth is chosen using maximum pseudo-likelihood leave-one-out cross-validation, which was originally proposed by Habbema, Hermans and Van den Broek (1974) as well as Duin (1976). Specifically, we find $h$ to maximize
\[
\frac{1}{n} \sum_{i=1}^{n} \log \left[ \sum_{j \neq i} K \left( \frac{X_j - X_i}{h} \right) \right] - \log[(n - 1)h].
\]

A value $h = 0.2138$ was obtained. An asymptotic pointwise 95% confidence interval was obtained using an approximation for the variance of the kernel density estimator. In particular, we may derive the following under the assumption that the true density $f$ has bounded derivative everywhere.

\[
\text{Var} \left( \hat{f}(x) \right) = \text{Var} \left( \frac{1}{nh} \sum_{i=1}^{n} K \left( \frac{x - X_i}{h} \right) \right)
\]
\[
= \frac{1}{nh^2} \text{Var} \left( K \left( \frac{x - X_1}{h} \right) \right)
\]
\[
= \frac{1}{nh^2} \left[ \int \left( K \left( \frac{x - y}{h} \right) \right)^2 f(y)dy - \left( \int K \left( \frac{x - y}{h} \right) f(y)dy \right)^2 \right]
\]
\[
= \frac{1}{nh^2} \left[ h \int (K(y))^2 f(x - hy)dy - h^2 \left( \int K(y)f(x - hy)dy \right)^2 \right]
\]
\[
= \frac{1}{nh^2} \left[ hf(x) \int (K(y))^2dy + O(h^2) \right]
\]
\[
\approx \frac{1}{nh} f(x) \int (K(y))^2dy
\]

So with the consistency $\hat{f}(x) \Rightarrow f(x)$ at each point, an asymptotic pointwise 95% confidence interval can be taken to be
\[
\hat{f}(x) \pm 1.96 \sqrt{\frac{1}{nh} \hat{f}(x) \int (K(y))^2dy}.
\]

Below is an illustration of the nonparametric pointwise 95% confidence bands.
Part e

Suppose \( X_i \overset{i.i.d.}{\sim} \mathcal{N}(\mu, \sigma^2) \). We first obtain joint asymptotic normality of the sample mean and sample variance:

\[
\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} X_i \quad \text{and} \quad \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \hat{\mu})^2.
\]

We recall that these two quantities are independent by Basu’s theorem, so marginal asymptotic normality will imply joint asymptotic normality. Of course we have

\[
\sqrt{n}(\hat{\mu} - \mu) \Rightarrow \mathcal{N}(0, \sigma^2).
\]

Notice also that

\[
\sqrt{n}(\hat{\sigma}^2 - \sigma^2) = \sqrt{n} \left( \frac{1}{n} \sum_{i=1}^{n} (X_i - \mu)^2 - (\hat{\mu} - \mu)^2 - \sigma^2 \right).
\]

Notice that the following quantity can be neglected:

\[
\sqrt{n}(\hat{\mu} - \mu)^2 = \frac{1}{\sqrt{n}} \left( \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (X_i - \mu) \right)^2 = \frac{1}{\sqrt{n}} O_p(1).
\]
And by the standard CLT,
\[
\sqrt{n} \left( \frac{1}{n} \sum_{i=1}^{n} (X_i - \mu)^2 - \sigma^2 \right) \Rightarrow \mathcal{N}(0, \mathbb{E} [(X_1 - \mu)^4] - \sigma^4).
\]

Therefore, we have joint asymptotic normality:
\[
\sqrt{n} \left[ \hat{\mu} - \mu \atop \hat{\sigma}^2 - \sigma^2 \right] \Rightarrow \mathcal{N}(0, \Sigma),
\]
where with \( \mu_4 = \mathbb{E} [(X_1 - \mu)^4] \),
\[
\Sigma = \begin{bmatrix} \sigma^2 & 0 \\ 0 & \mu_4 - \sigma^4 \end{bmatrix}.
\]

Now notice that for \( f \), the \( \mathcal{N}(\mu, \sigma^2) \) density,
\[
\nabla f(x; \mu, \sigma^2) = f(x; \mu, \sigma^2) \left[ \frac{x - \mu}{\sigma^2}, \frac{1}{2\sigma^2} \left( \frac{(x - \mu)^2}{\sigma^2} - 1 \right) \right].
\]

By the delta method,
\[
\sqrt{n}(f(x; \hat{\mu}, \hat{\sigma}^2) - f(x; \mu, \sigma^2)) \Rightarrow \mathcal{N}(0, \nabla f(x; \mu, \sigma^2) \Sigma(\nabla f(x; \mu, \sigma^2))').
\]

The function
\[
(\mu, \sigma^2, \mu_4) \mapsto \nabla f(x; \mu, \sigma^2) \Sigma(\nabla f(x; \mu, \sigma^2))'
\]
is continuous. Let \( \hat{\mu}_4 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \hat{\mu})^4 \). Since
\[
(\hat{\mu}, \hat{\sigma}^2, \hat{\mu}_4)
\]
consistently estimates \((\mu, \sigma^2, \mu_4)\), by the continuous mapping theorem, the plug-in estimator
\[
f^2(x; \hat{\mu}, \hat{\sigma}^2) \left[ \frac{(x - \hat{\mu})^2}{\hat{\sigma}^2} + \frac{1}{4\hat{\sigma}^4} \left( \frac{(x - \hat{\mu})^2}{\hat{\sigma}^2} - 1 \right)^2 (\hat{\mu}_4 - \hat{\sigma}^4) \right]
\]
consistently estimates
\[
f^2(x; \mu, \sigma^2) \left[ \frac{(x - \mu)^2}{\sigma^2} + \frac{1}{4\sigma^4} \left( \frac{(x - \mu)^2}{\sigma^2} - 1 \right)^2 (\mu_4 - \sigma^4) \right] = \nabla f(x; \mu, \sigma^2) \Sigma(\nabla f(x; \mu, \sigma^2))'.
\]
Thus, an asymptotic pointwise 95% confidence interval can be taken to be
\[
f(x; \hat{\mu}, \hat{\sigma}^2) \pm 1.96 \frac{f(x; \hat{\mu}, \hat{\sigma}^2)}{\sqrt{n}} \sqrt{\frac{(x - \hat{\mu})^2}{\hat{\sigma}^2} + \frac{1}{4\hat{\sigma}^4} \left( \frac{(x - \hat{\mu})^2}{\hat{\sigma}^2} - 1 \right)^2 (\hat{\mu}^4 - \hat{\sigma}^4)}.
\]
In particular, we have the estimated values $\hat{\mu} = 10.2569$, $\hat{\sigma}^2 = 0.7498$ and $\hat{\mu}_4 = 4.6086$. Below is a comparison of the nonparametric and parametric pointwise 95% confidence bands. They are similar for the most part. The confidence bands seem to be of similar width in both cases.

![Graph comparing nonparametric and parametric confidence bands.](image)

**Part f**

Kernel smoothing with the fourth-order Gaussian kernel
\[
K(x) = \frac{1}{2} (3 - x^2) \frac{1}{\sqrt{2\pi}} \exp \left( -x^2 / 2 \right)
\]
is used for nonparametric estimation. The bandwidth is once again obtained using leave-one-out cross-validation, and an optimal value $h = 0.4047$ is used. The plots below compare the fourth-order and second-order Gaussian kernel smoothing. Note in the first plot, the fourth-order kernel predicts a probability that is greater than one near a log family income of 6. This is an issue with using a higher order kernel where negative weights are given to some data points. We can get such artifacts.