Answer Hints

Basic Concept

In machine learing, Sparsity refers to the concept that a significant portion of data or parameters are zero or haing a negligible impact in a dataset or model. Stundents can find reference to this question in the slides of lecture 3, and there two types of sparsity mentioned in the lecture slides:
1. The first type is *exact sparsity*

$$s = \sum_{j=1}^{p} \mathbb{I}\{\beta_j \neq 0\} \text{ is small.}$$

2. The second type is *approximate sparsity*, which means that:

- most coefficients are close to zero.
- only a small number of coefficients are far away from zero.

ECON434 Final Exam, 06/09/2024, Question 2

As long as the student provides valid explanations, full scores of this question will be given. The answer listed below is from one of last year's student's reponse.
The selection bias formula is as follows:

$$\begin{aligned}\text{Selection Bias} = {} & P(D=0)\left(\mathbb{E}[Y_1 \mid D=1] - \mathbb{E}[Y_1 \mid D=0]\right) \\ & + P(D=1)\left(\mathbb{E}[Y_0 \mid D=1] - \mathbb{E}[Y_0 \mid D=0]\right)\end{aligned}$$

where $D = 0$ indicates that the individual is not insured, and $D = 1$ indicates that the individual is insured. The term $\mathbb{E}[Y_1 \mid D = 1] - \mathbb{E}[Y_1 \mid D = 0]$ represents the difference in health outcomes if everyone were insured, between those who choose to be insured and those who do not. Similarly, $\mathbb{E}[Y_0 \mid D = 1] - \mathbb{E}[Y_0 \mid D = 0]$ represents the difference in health outcomes if no one were insured, between those who choose to be insured and those who do not.
Now consider the mechanism that results in the **over-estimation** of the insurance health benefits. Individuals who participate in health insurance may be the type of people who care more about their health and tend to have healthier living habits, such as no use of drugs, no smoking, and no use of alcohol. These living habits would yield better health outcomes even without insurance. In the selection bias formula, this results in:

$$\mathbb{E}[Y_1 \mid D=1] - \mathbb{E}[Y_1 \mid D=0] > 0 \quad \text{and} \quad \mathbb{E}[Y_0 \mid D=1] - \mathbb{E}[Y_0 \mid D=0] > 0$$

Both terms contribute positively to the selection bias, leading to an over-estimation of the insurance health benefit.
Next, consider the second mechanism that results in the **under-estimation** of the insurance health benefits. Individuals who decide to participate in health

insurance may be those with poorer health, anticipating higher healthcare needs. Even with insurance, their observed health outcomes may be worse than those uninsured. In the selection bias formula, this leads to:

$$\mathbb{E}[Y_1 \mid D=1] - \mathbb{E}[Y_1 \mid D=0] < 0 \quad \text{and} \quad \mathbb{E}[Y_0 \mid D=1] - \mathbb{E}[Y_0 \mid D=0] < 0$$

Both terms contribute negatively to the selection bias, resulting in an under-estimation of the insurance health benefit.
Finally, the third mechanism concerns **differential access to healthcare resources**. The availability and quality of healthcare can vary significantly by geographic region. Individuals in areas with better health infrastructure and more healthcare providers tend to have better outcomes regardless of insurance. If individuals in such areas are also more likely to be insured, the effect of insurance might be overestimated, as improved outcomes may be due to better access, not insurance itself. In the selection bias formula, this implies:

$$\mathbb{E}[Y_1 \mid D=1] - \mathbb{E}[Y_1 \mid D=0] > 0 \quad \text{and} \quad \mathbb{E}[Y_0 \mid D=1] - \mathbb{E}[Y_0 \mid D=0] > 0$$

Again, both terms contribute positively, leading to an over-estimation of the insurance health benefit.

Coding Question

Answer to this problem can be found in the slides for lecture 12:

- n_estimators: number of trees in the forest
- max_depth: depth of each tree

ECON434 Final Exam, 06/09/2024, Question 1

Since it applied the OLS regression here, we have the following:

$$E[Y|X=x] = \alpha + \beta x,\ x = -1, 0, 1$$

with,

$$\begin{aligned} E[e] &= 0 \\ cov(e, X) &= 0 \end{aligned}$$

Then, we can deduce:

$$\begin{aligned} E[Y - \alpha - \beta X] &= 0 \\ E[(Y - \alpha - \beta X)X] &= 0 \end{aligned}$$

Then we have the following results:

$$\begin{aligned} \alpha &= E[Y] - \beta E[X] \\ \beta &= \frac{cov(X, Y)}{Var(X)} \end{aligned}$$

According to the description,

$$\begin{aligned} E[XY] &= P(X=1)(1)E[Y|X=1] + P(X=0)(0)E[Y|X=0] + P(X=-1)(-1)E[Y|X=-1] \\ &= \frac{1}{3}E[Y|X=1] - \frac{1}{3}E[Y|X=-1] \\ E[X] &= \sum xP(X=x), \text{for } x=-1,0,1 \end{aligned}$$

Therefore, $cov(X,Y) = E[XY] - E[X]E[Y] = \frac{1}{3}E[Y|X=1] - \frac{1}{3}E[Y|X=-1]$
Now consdier the value of $Var(X)$,

$$Var(X) = E[X^2] - E^2[X] = \frac{2}{3}$$

Therefore,

$$\begin{aligned} \beta &= \frac{cov(X,Y)}{Var(X)} \\ &= \frac{1}{2}(E[Y|X=1] - E[Y|X=-1]) \end{aligned}$$

In conclusion, the probability limit of the slope coefficient of such a regression in terms of the conditional mean function $E[Y|X=x]$, with $x=-1,0,1$ is:

$$\lim_{n\to\infty} \hat{\beta} \xrightarrow{p} \frac{1}{2}(E[Y|X=1] - E[Y|X=-1])$$

ECON434 Final Exam, 06/09/2024, Question 3

(a). To show that θ can be rewritten as the formula provided, we have the following:

$$\theta = \mathbb{E}[Zh(X)^4] = \mathbb{E}[Zh(X)^4 + 4(Y-h(X))h(X)^3p(X)]$$

With the property of expectations:

$$\mathbb{E}[Zh(X)^4 + 4(Y-h(X))h(X)^3p(X)] = \mathbb{E}[Zh(X)^4] + \mathbb{E}[4(Y-h(X))h(X)^3p(X)]$$

Then, we need to show that:

$$\mathbb{E}[4(Y-h(X))h(X)^3p(x)] = 0$$

$$\mathbb{E}[4(Y-h(X))h(X)^3p(x)] = 4\mathbb{E}[Yh(X)^3p(X)] - 4\mathbb{E}[h(X)^4p(X)]$$

By applying the law of iterated expectations on the first term on the right hand side, we have:

$$4\mathbb{E}[Yh(X)^3p(X)] = 4\mathbb{E}[\mathbb{E}[Yh(X)^3p(X) \mid X]] = 4\mathbb{E}[h(X)^3p(X)\mathbb{E}[Y \mid X]]$$

By definition of $h(x) = \mathbb{E}[Y \mid X=x],\ x \in \mathbb{R}$:

$$4\mathbb{E}[h(X)^3p(X)\mathbb{E}[Y \mid X]] = 4\mathbb{E}[h(X)^4p(X)]$$

Therefore,

$$\begin{aligned}
\mathbb{E}[Zh(X)^4 + 4(Y - h(X))h(X)^3 p(X)] &= \mathbb{E}[Zh(X)^4] + \mathbb{E}[4(Y - h(X))h(X)^3 p(X)] \\
&= \mathbb{E}[Zh(X)^4] + 4\mathbb{E}[Yh(X)^3 p(X)] - 4\mathbb{E}[h(X)^4 p(X)] \\
&= \mathbb{E}[Zh(X)^4] + 4\mathbb{E}[h(X)^4 p(X)] - 4\mathbb{E}[h(X)^4 p(X)] = \mathbb{E}[Zh(X)^4] \\
&= \theta
\end{aligned}$$

(b). To prove the estimating equation (2) satisfies the Neyman orthogonality condition, it's equivalent to prove that the estimating equation is insensitive to the changes of $h(x)$ and $p(x)$ (in the course contents, it's to prove the insensitivity of the estimator to the change of two machine learning estimators used).

First, we consider the small change of $h(x) = \mathbb{E}[Y|X = x],\ x \in \mathbb{R}$: Assuming that, with a small change we have $h_{\text{new}}(x) = h(x) + r(\tilde{h}(x) - h(x))$, then for equation (2), we have:

$$\theta_h(r) = \mathbb{E}[Zh_{\text{new}}(X)^4 + 4(Y - h_{\text{new}}(X))h_{\text{new}}(X)^3 p(X)]$$

Take the derivative with respect to r:

$$\begin{aligned}
\frac{\partial \theta_h(r)}{\partial r} &= \mathbb{E}[4Z(h(x) + r(\tilde{h}(x) - h(x)))^3(\tilde{h}(x) - h(x))] \\
&\quad - \mathbb{E}[4(\tilde{h}(x) - h(x))(h(x) + r(\tilde{h}(x) - h(x)))^3 p(x)] \\
&\quad + \mathbb{E}[12(\tilde{h}(x) - h(x))(Y - h(x) - r(\tilde{h}(x) - h(x)))(h(x) + r(\tilde{h}(x) - h(x)))^2 p(x)]
\end{aligned}$$

The third term can be rewritten as:

$$\mathbb{E}[12(\tilde{h}(x) - h(x))(Y - h_{\text{new}}(x))h_{\text{new}}(x)^2 p(x)]$$

From previous question, we know that:

$$\mathbb{E}[Y - h(x)|X] = 0$$

Therefore:

$$\mathbb{E}[12(\tilde{h}(x) - h(x))(Y - h_{\text{new}}(x))h_{\text{new}}(x)^2 p(x)] = 0$$

As for the first two terms:

$$\begin{aligned}
&\mathbb{E}[4Z(h(x) + r(\tilde{h}(x) - h(x)))^3(\tilde{h}(x) - h(x))] \\
&\quad - \mathbb{E}[4(\tilde{h}(x) - h(x))(h(x) + r(\tilde{h}(x) - h(x)))^3 p(x)] \\
&= 4\mathbb{E}[(\tilde{h}(x) - h(x))(h(x) + r(\tilde{h}(x) - h(x)))^3(Z - p(x))]
\end{aligned}$$

By the law of iterated expectations:

$$4\mathbb{E}[(\tilde{h}(x) - h(x))(h(x) + r(\tilde{h}(x) - h(x)))^3(\mathbb{E}[Z|X = x] - p(x))] = 0$$

By definition, $p(x) = \mathbb{E}[Z|X = x]$, so

$$\frac{\partial \theta_h(r)}{\partial r} = 0$$

Now consider the small change of $p(x)$. Assuming that, with a small change we have $p_{\text{new}}(x) = p(x) + r(\tilde{p}(x) - p(x))$, then we have:

$$\theta_p(r) = \mathbb{E}[Zh(x)^4 + 4(Y - h(x))h(x)^3(p(x) + r(\tilde{p}(x) - p(x)))]$$

Then:

$$\frac{\partial \theta_p(r)}{\partial r} = \mathbb{E}[4(Y - h(x))h(x)^3(\tilde{p}(x) - p(x))]$$

By the law of iterated expectations:

$$= 4\mathbb{E}[h(x)^3(\tilde{p}(x) - p(x))\mathbb{E}[Y - h(x)|X = x]]$$

By definition $h(x) = \mathbb{E}[Y|X = x],\ x \in \mathbb{R}$, we have:

$$\mathbb{E}[Y - h(x)|X = x] = 0 \Rightarrow \frac{\partial \theta_p(r)}{\partial r} = 0$$

In conclusion, the estimation equation (2) satisfies the Neyman orthogonality condition with respect to both $h(x)$ and $p(x)$.
(c). Similar to question (b), we assume that with a small change, we have

$$h_{\text{new}}(x) = h(x) + r(\tilde{h}(x) - h(x))$$

Then:

$$\theta(r) = \mathbb{E}[Zh_{\text{new}}(x)^4]$$

$$\frac{\partial \theta(r)}{\partial r} = \mathbb{E}\left[4Z(h(x) + r(\tilde{h}(x) - h(x)))^3(\tilde{h}(x) - h(x))\right]$$

Since r is arbitrarily chosen, Z is a random variable here, and Z and $h(x)$ are not necessarily independent of each other, then

$$\frac{\partial \theta(r)}{\partial r} \neq 0$$

for arbitrary r.
Therefore, the estimating equation (1) does not satisfy the Neyman orthogonality condition.
Note:The answers provided for the last two questions are based on a response from a student in last year's class. They are intended for reference only and do not imply that your solutions must be as detailed as this.