CS 477/677 Causal Inference: Homework 1
Probability and Statistics

Ha Bui
hbui13@jhu.edu

1 Analytical (34 Points)

1 (8 points) Consider k features $\mathbf{X}$ and an outcome Y, where $\mu(\mathbf{X};\mathbf{w}) = \mathbb{E}[Y \mid \mathbf{X};\mathbf{w}]$ is parameterized by k parameters $\mathbf{w}$. Define $f(\mathbf{X}, Y; \mathbf{w})$ as the following random function (of size $k \times 1$): $\mathbf{X}^2_{(1:k)} \times \{Y - \mu(\mathbf{X};\mathbf{w})\}$.

1 Consider the sum over n data rows of the j^{th} output of f: $\sum_{i=1}^n (\mathbf{x}_i)^2_j \times \{y_i - \mu(\mathbf{x}_i;\mathbf{w})\}$. Compute the derivative of this object with respect to $\mathbf{w}$. Note that this will be a vector of size k.

Solution:

$$\frac{\nabla f_j}{\nabla \mathbf{w}} = \left[\frac{-\sum_{i=1}^m (\mathbf{x}_i)^2_j \times [\mu(\mathbf{x}_i;\mathbf{w})]'}{\nabla \mathbf{w}_1}, ..., \frac{-\sum_{i=1}^m (\mathbf{x}_i)^2_j \times [\mu(\mathbf{x}_i;\mathbf{w})]'}{\nabla \mathbf{w}_k}\right].$$

2 Give the form for the full derivative of $\sum_{i=1}^n (\mathbf{x}_i)^2 \times \{y_i - \mu(\mathbf{x}_i;\mathbf{w})\}$ with respect to $\mathbf{w}$. Note that this will be a $k \times k$ matrix.

Solution:

$$\frac{\nabla f}{\nabla \mathbf{w}} = \left[\frac{\nabla f_1}{\nabla \mathbf{w}}, ..., \frac{\nabla f_k}{\nabla \mathbf{w}}\right].$$

3 Let $\mathbf{w}_0$ be the true value of $\mathbf{w}$. Will $\mathbb{E}[f(\mathbf{X}, Y; \mathbf{w}_0)] = \sum_{\mathbf{X},Y} f(\mathbf{X}, Y; \mathbf{w}_0) p(\mathbf{X}, Y) = \vec{0}$ (where the expectation is taken with respect to the observed data distribution $p(\mathbf{X}, Y)$) under true values of $\mathbf{w}$? Explain.

Solution: Yes. Because:

$$\begin{aligned}
\mathbb{E}[f(\mathbf{X}, Y; \mathbf{w}_0)] &= \sum_{\mathbf{X},Y} f(\mathbf{X}, Y; \mathbf{w}_0) p(\mathbf{X}, Y) \\
&= \sum_{\mathbf{X},Y} \mathbf{X}^2_{(1:k)} \times \{Y - \mu(\mathbf{X};\mathbf{w_0})\} p(\mathbf{X}, Y) \\
&= \sum_{\mathbf{X},Y} \mathbf{X}^2_{(1:k)} \times \{Y - \mathbb{E}[Y \mid \mathbf{X};\mathbf{w_0}]\} p(\mathbf{X}, Y) \\
&= \sum_{\mathbf{X},Y} \mathbf{X}^2_{(1:k)} \times \{Y - Y\} p(\mathbf{X}, Y) \text{ (because } \mathbf{w}_0 \text{ is the true estimator)} \\
&= \vec{0}.
\end{aligned}$$

4 Fix $f_A(\mathbf{X}, Y; \mathbf{w})$ to be equal to $A(\mathbf{X})\{Y - \mu(\mathbf{X}; \mathbf{w})\}$, where $A(\mathbf{X})$ is any function of $\mathbf{X}$ of the same dimension as $\mathbf{w}$. Will $E[f_A(\mathbf{X}, Y; \mathbf{w})] = \vec{0}$ (where the expectation is taken with respect to the observed data distribution $p(\mathbf{X}, Y)$) under true values of $\mathbf{w}$? Explain.

Solution: Yes. Because:

$$\begin{aligned}
\mathbb{E}[f_A(\mathbf{X}, Y; \mathbf{w})] &= \mathbb{E}_{\mathbf{X}}[\mathbb{E}[f_A(\mathbf{X}, Y; \mathbf{w})] \mid \mathbf{X}] \\
&= \mathbb{E}_{\mathbf{X}}[\mathbb{E}[A(\mathbf{X})\{Y - \mu(\mathbf{X}; \mathbf{w}) \mid \mathbf{X}] \\
&= \mathbb{E}_{\mathbf{X}}[\mathbb{E}[A(\mathbf{X})\{Y - \mathbb{E}[Y \mid \mathbf{X}; \mathbf{w}) \mid \mathbf{X}] \\
&= \mathbb{E}_{\mathbf{X}}[\mathbb{E}[A(\mathbf{X})(Y - Y) \mid \mathbf{X}] \text{ (because } \mathbf{w} \text{ is the true estimator)} \\
&= \vec{0}.
\end{aligned}$$

2 (4 points) Show that the logistic regression model lies in the restricted moments model: $Y = \mu(\mathbf{X}; \beta) + \epsilon$, where $\mathbb{E}[\epsilon \mid \mathbf{X}] = 0$.

Solution:

Due to logistic regression, then:

$$\begin{aligned}
\mathbb{E}(Y \mid \mathbf{X}) &= \sum_y y p(Y = y \mid \mathbf{X}) \\
&= 1 * p(Y = 1 \mid \mathbf{X}) + 0 * p(Y = 0 \mid \mathbf{X}) \\
&= \mu(\mathbf{X}; \beta) = \mathbb{E}(\mu(\mathbf{X}; \beta) \mid \mathbf{X})
\end{aligned} \tag{1}$$

Due to $Y = \mu(\mathbf{X}; \beta) + \epsilon$, then:

$$\begin{aligned}
\mathbb{E}(Y \mid \mathbf{X}) &= \mathbb{E}(\mu(\mathbf{X}; \beta) + \epsilon \mid \mathbf{X}) \\
&= \mathbb{E}(\mu(\mathbf{X}; \beta) \mid \mathbf{X}) + \mathbb{E}(\epsilon \mid \mathbf{X})
\end{aligned} \tag{2}$$

From 1 and 2, we obtain: $\mathbb{E}(\epsilon \mid X) = 0$.

3 (4 points) Assume we conduct a randomized controlled trial where a binary A is randomized to treatment ($A = 1$) or control ($A = 0$) values. Assume the outcome Y is missing completely at random. That is, the observed outcome Y is either the true outcome if $R = 1$ or "?" if $R = 0$. Given data on $p(Y, A, R)$, show that the ACE $\mathbb{E}[Y^{(1)}(A = 1) - Y^{(1)}(A = 0)]$ (with respect to the outcome $Y^{(1)}$ had it been observed for everyone) is identified, and give the identifying formula.

Solution:

Due to MCAR: $Y^{(1)} \perp\!\!\!\perp R = 1$, then:

$$\mathbb{E}[Y^{(1)}] = \mathbb{E}[Y^{(1)} \mid R = 1] = \mathbb{E}[Y \mid R = 1] \text{ (because } Y^{(1)} = Y \text{ if } R = 1)$$

Due to Consistency: $Y = Y(A)$, then:

$$\mathbb{E}[Y|R = 1] = \mathbb{E}[Y(A)|R = 1]$$

Due to Ignoribility: $\{Y(1), Y(0)\} \perp\!\!\!\perp A$, then:

$$\begin{cases}
\mathbb{E}[Y^{(1)}(A = 1)] = \mathbb{E}[Y(A = 1) \mid R = 1] = \mathbb{E}[Y \mid A = 1, R = 1] \\
\mathbb{E}[Y^{(1)}(A = 0)] = \mathbb{E}[Y(A = 0) \mid R = 1] = \mathbb{E}[Y \mid A = 0, R = 1]
\end{cases}$$

So, ACE

$$= \mathbb{E}[Y^{(1)}(A=1) - Y^{(1)}(A=0)]$$
$$= \mathbb{E}[Y^{(1)}(A=1)] - \mathbb{E}[Y^{(1)}(A=0)]$$
$$= \mathbb{E}[Y \mid A=1, R=1] - \mathbb{E}[Y \mid A=0, R=1]$$
$$= \left(\frac{1}{k}\sum_{i=1}^{n-m} Y_i A_i\right) - \left[\frac{1}{n-m-k}\sum_{i=1}^{n-m} Y_i(1-A_i)\right], \text{with m is the number of R=1.}$$

4 (4 points) Assume we have three random variables, A, B, and C. Is is possible for the following statements to both hold? $(A \perp\!\!\!\perp B, C)$ and $(A \not\perp\!\!\!\perp B \mid C)$. If so, give an example real life situation (ie interpretations of the variables A, B, and C) where the above statement should be true. If not, prove it.

Solution: No. Because:

If $(A \perp\!\!\!\perp B, C)$, then:

$$\begin{aligned}
&p(A,B,C) = p(A|B,C)p(B,C)\\
\Leftrightarrow &p(A,B,C) = p(A)p(B,C)\\
\Leftrightarrow &\sum_b p(A,B,C) = \sum_b p(A)p(B,C)\\
\Leftrightarrow &p(A,C) = p(A)\sum_b p(B,C)\\
\Leftrightarrow &p(A,C) = p(A)p(C)\\
\Rightarrow &A \perp\!\!\!\perp C,
\end{aligned} \tag{3}$$

and

$$p(A,B|C) = \frac{p(A,B,C)}{p(C)} = \frac{p(A)p(B|C)p(C)}{p(C)} = p(A)p(B|C). \tag{4}$$

If $(A \not\perp\!\!\!\perp B \mid C)$, then:

$$p(A,B|C) \neq p(A|C)p(B|C) \tag{5}$$

From 4 and 5, we have:

$$p(A) \neq p(A|C) \Rightarrow A \not\perp\!\!\!\perp C \text{ (confliction with 3).}$$

5 (4 points) Given a binary A, assume the usual consistency property for Y, namely: $Y = Y(1)A + Y(0)(1-A)$ does not hold, however there exist known bijective functions g, h such that $Y = g(Y(1))A + h(Y(0))(1-A)$. Is the ACE $\mathbb{E}[Y(1) - Y(0)]$ identified, if ignorability also holds, e.g. $A \perp\!\!\!\perp \{Y(1), Y(0)\}$? If so, derive the identifying formula (show your work). Otherwise explain what goes wrong.

Solution: Yes. Because:

If $Y = g(Y(1))A + h(Y(0))(1-A)$, then:

$$\begin{cases} p(Y|A=1) = p(g(Y(1))A + h(Y(0))(1-A))|A=1) = p(g(Y(1))|A=1) \\ p(Y|A=0) = p(g(Y(1))A + h(Y(0))(1-A)|A=0) = p(h(Y(0))|A=0) \end{cases} \tag{6}$$

Due to Ignoribility: $\{Y(1), Y(0)\} \perp\!\!\!\perp A$, then:

$$\{g(Y(1)), h(Y(0))\} \perp\!\!\!\perp A \tag{7}$$

From 6 and 7, we have:

$$\begin{cases} p(Y|A=1) = p(g(Y(1))|A=1) = p(g(Y(1))) \\ p(Y|A=0) = p(h(Y(0))|A=0) = p(h(Y(0))) \end{cases} \tag{8}$$

From 8 and due to g, h are bijective functions, then:

$$\begin{cases} Y(1) = g^{-1}(Y|A=1) \\ Y(0) = h^{-1}(Y|A=0) \end{cases}$$

ACE

$$\begin{aligned} &= \mathbb{E}[Y(1) - Y(0)] \\ &= \mathbb{E}[Y(1)] - \mathbb{E}[Y(0)] \\ &= \mathbb{E}[g^{-1}(Y|A=1)] - \mathbb{E}[h^{-1}(Y|A=0)] \end{aligned}$$

So, ACE $\mathbb{E}[Y(1) - Y(0)]$ identified (depends only on outcome Y and feature A).

6a (4 points) Define $q(Y, M|A, C)$ as:

$$q(Y, M|A, C) = \sum_{A'} p(Y|A', M, C)p(A'|C)p(M|A, C)$$

Show that $q(Y, M|A, C)$ is a valid probability distribution.

Solution:

We have: $q(Y, M|A, C) = \sum_{A'} p(Y|A', M, C)p(A'|C)p(M|A, C)$, then:

$$q(Y, M|A, C) \geq 0 \text{ (because } p(\cdot) \geq 0) \tag{9}$$

We also have:

$$\begin{aligned} \sum_Y \sum_M q(Y, M|A, C) &= \sum_Y \sum_M \sum_{A'} p(Y|A', M, C)p(A'|C)p(M|A, C) \\ &= \sum_Y \sum_M \sum_{A'} \frac{p(Y, A'|M, C)}{p(A'|M, C)} p(A'|C)p(M|A, C) \\ &= \sum_M p(M|A, C) = 1 \end{aligned} \tag{10}$$

From 9 and 10, $q(Y, M|A, C)$ is a valid probability distribution.

6b (6 points) Let $q(Y, M|A, C)$ be defined as before in 6a. Now, define $q(Y|A, C)$, $q(M|A, C)$, and $q(Y|M, A, C)$ as:

- $q(Y|A, C) = \sum_M q(Y, M|A, C)$
- $q(M|A, C) = \sum_Y q(Y, M|A, C)$
- $q(Y|M, A, C) = \frac{q(Y, M|A, C)}{q(M|A, C)}$

Rewrite $q(Y|A,C)$, $q(M|A,C)$, and $q(Y|M,A,C)$ in terms of the distribution p.

Solution:

$$\begin{aligned}
q(Y|A,C) &= \sum_M \sum_{A'} p(Y|A',M,C)p(A'|C)p(M|A,C) \\
&= \sum_M \sum_{A'} \frac{p(Y,A'|M,C)}{p(A'|M,C)} p(A'|C)p(M|A,C) \\
&= \sum_M \sum_{A'} \frac{p(A'|Y,M,C)p(Y|M,C)p(A'|C)}{p(A'|M,C)} p(M|A,C) \\
&= p(Y|C)
\end{aligned}$$

$$\begin{aligned}
q(M|A,C) &= \sum_Y \sum_{A'} p(Y|A',M,C)p(A'|C)p(M|A,C) \\
&= \sum_Y \sum_{A'} \frac{p(Y,A'|M,C)}{p(A'|M,C)} p(A'|C)p(M|A,C) \\
&= p(M|A,C)
\end{aligned}$$

$$\begin{aligned}
q(Y|M,A,C) &= \frac{\sum_{A'} p(Y|A',M,C)p(A'|C)p(M|A,C)}{p(M|A,C)} \\
&= \frac{p(Y|M,C)p(M|A,C)}{p(M|A,C)} \\
&= p(Y|M,C)
\end{aligned}$$

References

[1] Hill, Jennifer. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 217–240, 2011.

[2] Brooks-Gunn, J., Liaw, F., and Klebanov, P. Effects of Early Intervention on Cognitive Function of Low Birth Weight Preterm Infants. *Journal of Pediatrics*, 350–359, 1991.

[3] Scott, D., and Bauer, C. A Neonatal Health Index for Preterm Infants. *Pediatric Research* , 1989.