Tools and Techniques for Machine Learning
Homework 2: Regression imputation, covariate shift, and control variates

Instructions: Your answers to the questions below, including plots and mathematical work, should be submitted as a single PDF file. It's preferred that you write your answers using software that typesets mathematics (e.g. LaTeX, LyX, or Jupyter), though if you need to you may scan handwritten work. For submission, you can also export your Jupyter notebook and merge that PDF with your PDF for the written solutions into one file. **Don't forget to complete the Jupyter notebook as well, for the programming part of this assignment**.

General hint on Adam's Law

A couple times in this assignment we'll need a variant on the basic Adam's Law. Adam's Law is that $\mathbb{E}\left[\mathbb{E}\left[Y \mid X\right]\right] = \mathbb{E}Y$. The variant we'll need is that Adam's Law still holds when everything is conditioned on a particular event. For example, $\mathbb{E}\left[\mathbb{E}\left[Y \mid X, Z > a\right] \mid Z > a\right] = \mathbb{E}\left[Y \mid Z > a\right]$. We could see this by defining (X', Y') to have joint distribution that's equal to the conditional distribution of $(X, Y) \mid Z > a$. Then

$$\begin{aligned} & \mathbb{E}\left[\mathbb{E}\left[Y \mid X, Z > a\right] \mid Z > a\right] \\ = \quad & \mathbb{E}\left[\mathbb{E}\left[Y' \mid X'\right]\right] = \mathbb{E}\left[Y'\right] = \mathbb{E}\left[Y \mid Z > a\right]. \end{aligned}$$

Another approach would be to define the random variable $W = \mathbb{1}\left[Z > a\right]$. Then

$$\mathbb{E}\left[\mathbb{E}\left[Y \mid X, W\right] \mid W\right] = \mathbb{E}\left[Y \mid W\right],$$

by the generalized form of Adam's Law. This implies that

$$\mathbb{E}\left[\mathbb{E}\left[Y \mid X, W = 1\right] \mid W = 1\right] = \mathbb{E}\left[Y \mid W = 1\right],$$

1 Complete case mean is unbiased for MCAR (when it's defined)

Let $R_i \in \{0, 1\}$ be the response indicator, $Y_i \in \mathbb{R}$ the response. Consider the MCAR setting, in which $R_i \perp\!\!\!\perp Y_i$, and suppose $(R, Y), (R_1, Y_1), \ldots, (R_n, Y_n)$ are i.i.d. We observe data $\mathcal{D} = ((R_1, R_1 Y_1), \ldots, (R_n, R_n Y_n))$. The complete case estimator is defined as

$$\hat{\mu}_{\text{cc}} = \hat{\mu}_{\text{cc}}(\mathcal{D}) = \frac{\sum_{i=1}^{n} R_i Y_i}{\sum_{i=1}^{n} R_i}.$$

Note that if $R_1 = \cdots = R_n = 0$, then $\hat{\mu}_{\text{cc}} = \frac{0}{0}$, which is undefined. Since $\hat{\mu}_{\text{cc}}$ is undefined with nonzero probability, it doesn't have an expectation or bias. In this problem, we consider whether $\hat{\mu}_{\text{cc}}$ is unbiased after ruling out the case where it's undefined.

1. Show that $\mathbb{E}\left[\hat{\mu}_{\text{cc}} \mid \sum_{i=1}^n R_i > 0\right] = \mathbb{E}Y$. (Hint: Show that $\mathbb{E}\left[\hat{\mu}_{\text{cc}} \mid R_1, \ldots, R_n, \sum_{i=1}^n R_i > 0\right] = \mathbb{E}Y$.)

2 Regression imputation with $\mathbb{E}\left[Y \mid X = x\right]$

Consider the MAR setting. Let $\hat{f}(x)$ be a regression function fit to the complete cases. Then the regression imputation estimator for $\mathbb{E}Y$ that we defined in class is given by

$$\hat{\mu}_{\hat{f}} := \frac{1}{n} \sum_{i=1}^n \left[R_i Y_i + (1 - R_i)\, \hat{f}(X_i)\right].$$

There is an alternative form of regression imputation where we apply $\hat{f}(x)$ to all the X_i's, not just the incomplete cases. This estimator is given by

$$\hat{\mu}_{\hat{f}\text{-full}} := \frac{1}{n} \sum_{i=1}^n \hat{f}(X_i).$$

In this problem, we will verify that if we use $\mathbb{E}\left[Y \mid X = x\right]$ for our regression imputation, then both of these regression imputation estimators are unbiased. This will give us some hope that, under the appropriate technical conditions, if our model is well-specified, then each method of regression imputation is consistent.

1. If $f(x) = \mathbb{E}\left[Y \mid X = x\right]$,show that $\mathbb{E}\left[\hat{\mu}_{f\text{-full}}\right] = \mathbb{E}Y$.
2. If $f(x) = \mathbb{E}\left[Y \mid X = x\right]$,show that $\mathbb{E}\left[\hat{\mu}_{f}\right] = \mathbb{E}Y$. (Hint: See the slide on "Adam's Law / Law of iterated expectation" for inspiration, and you'll also need to use the MAR assumption that $Y_i \perp\!\!\!\perp R_i \mid X_i$.)
3. If we expand out the two forms of the imputation estimator for a particular set of observed data, we might get something like

$$\begin{aligned} \hat{\mu}_{\hat{f}} &= \frac{1}{n}\left(Y_1 + Y_2 + \hat{f}(X_3) + Y_4 + \hat{f}(X_5) + \cdots + Y_n\right) \\ \hat{\mu}_{\hat{f}\text{-full}} &= \frac{1}{n}\left(\hat{f}(X_1) + \hat{f}(X_2) + \hat{f}(X_3) + \hat{f}(X_4) + \hat{f}(X_5) + \cdots + \hat{f}(X_n)\right). \end{aligned}$$

   Written in this way, it's easy to see that $\hat{\mu}_{\hat{f}\text{-full}}$ and $\hat{\mu}_{\hat{f}}$ differ only in how they handle the complete cases. We generally do not expect to have $\hat{f}(X_i) = Y_i$ for all the compete cases – that would indicate overfitting of our imputation function. Nevertheless, you might be surprised to learn that $\hat{\mu}_{\hat{f}} = \hat{\mu}_{\hat{f}\text{-full}}$ in some common scenarios. Show that $\hat{\mu}_{\hat{f}\text{-full}} = \hat{\mu}_{\hat{f}}$ if we fit the regression function $\hat{f}(x)$ to the complete cases using a linear model with intercept:

$$\hat{f} = \underset{\{f : f(x) = a + w^T x\}}{\arg\min} \sum_{i=1}^n R_i (f(X_i) - Y_i)^2.$$

3 A family of simple AIPW estimators

(Continuing the "IPW estimator is not equivariant" problem (1.3) in Homework #1.)

Suppose $\mathcal{D}$ represents the dataset $(X_1, R_1, R_1Y_1), \ldots, (X_n, R_n, R_nY_n)$ from a MAR setting. For any $a \in \mathbb{R}$, we'll write $\mathcal{D} - a$ for the dataset $(X_1, R_1, R_1(Y_1 - a)), \ldots, (X_n, R_n, R_n(Y_n - a))$, which is the same as $\mathcal{D}$, but with each Y value shifted by a. Recall the following estimators:

$$\begin{aligned}\hat{\mu}_{\text{ipw}} = \hat{\mu}_{\text{ipw}}(\mathcal{D}) \quad &:= \quad \frac{1}{n}\sum_{i=1}^{n}\frac{R_iY_i}{\pi(X_i)} \\ \hat{\mu}_{\text{ipw},a} = \hat{\mu}_{\text{ipw},a}(\mathcal{D}) \quad &:= \quad \hat{\mu}_{\text{ipw}}(\mathcal{D} - a) + a,\end{aligned}$$

for any $a \in \mathbb{R}$. In the last homework, we showed that

$$\hat{\mu}_{\text{ipw}}(\mathcal{D} - a) = \hat{\mu}_{\text{ipw}}(\mathcal{D}) - \frac{a}{n}\sum_{i=1}^{n}\frac{R_i}{\pi(X_i)}$$

and that $\mathbb{E}\hat{\mu}_{\text{ipw},a}(\mathcal{D}) = \mathbb{E}Y$.

1. We can view $\hat{\mu}_{\text{ipw},a}$ as an augmented IPW (AIPW) estimator -- that is, as a control-variate adjusted IPW estimator. With this view, what is the control variate and what is its expectation?

2. Given what we learned about control variates, how would you choose $a \in \mathbb{R}$? (There are many reasonable answers to this question, and I don't believe there is a single best answer without additional assumptions. That said, the section on "Optimal scaling to improve variance" in the control variates module may be a source of some ideas.)

4 Election forecasting

Suppose we want to forecast the outcome of an election with two candidates. We have a budget to call n people and ask who they'll vote for. Each individual i is described by the following random variables:

$X_i \in \mathcal{X}$	covariates describing individual i
$T_i \in \{0,1\}$	indicator for whether i will vote in the election ("turnout indicator")
$R_i \in \{0,1\}$	indicator for whether i will respond to a survey question if called
$Y_i \in \{0,1\}$	indicator for which candidate an individual will vote for, if they vote

We'll assume the existence of an "eligible voter generating distribution[1]", and we'll refer to it as P. To carry out the survey, n individuals are sampled from P. For individuals who respond (i.e. for whom $R = 1$), we will assume they reveal their true value of Y. We'll write the full data corresponding to this scenario as

$$(X, R, Y, T), (X_1, R_1, Y_1, T_1), \ldots, (X_n, R_n, Y_n, T_n),$$

[1] In reality, there is a fixed set of potential voters. We're taking the "eligible voter generating distribution" approach to align more with the framework of our class. For large elections, the list of all potential voters is so much larger than the size of the survey sample that this is a very reasonable approximation.

sampled i.i.d. from P. However, since we only observe Y when $R = 1$, and we don't observe T at all, we'll write the observed data as

$$(X, R, RY), (X_1, R_1, R_1Y_1), \ldots, (X_n, R_n, R_nY_n).$$

We'll make the following assumptions:

1. R, Y, and T are mutually independent given X. (In particular, this implies $Y \perp\!\!\!\perp R \mid X$ and $Y \perp\!\!\!\perp T \mid X$.)

2. We have access to a function $\pi_t(x) = \mathbb{P}(T = 1 \mid X = x)$ that gives the "turnout probability", i.e. the probability that an individual will go vote, given their covariates[2].

3. We have access to a function $\pi_r(x) = \mathbb{P}(R = 1 \mid X = x)$ that gives the "response probability." This can function be estimated using the observed data using, for example, logistic regression. But we'll also assume that this part of the problem has already been solved and we know $\pi_r(x)$.

4. Every voter has at least some chance of responding to a survey. To put this in mathematical terms: $\pi_t(x) > 0 \implies \pi_r(x) > 0 \quad \forall x \in \mathcal{X}$.

To forecast the election, we want to estimate $\mathbb{P}(Y = 1 \mid T = 1)$, i.e. the rate of voting for candidate 1 among individuals who actually go vote. In this problem, we'll use a variant of regression imputation that accounts for the covariate shift between the survey respondent distribution and the voter distribution.

4.1 Fitting the regression

If we fit a model to the survey responses (i.e. the complete cases, i.e. the (X_i, Y_i) pairs corresponding to $R_i = 1$) in the usual way (say empirical risk minimization over some space of functions), we'll end up with a function $\hat{f}(x)$ that has low risk with respect to the distribution $p(x, y \mid R = 1)$. In other words, $\hat{f}(x)$ will perform well for survey responders, but what we really need is for $\hat{f}(x)$ to perform well for voters, i.e. to have low risk w.r.t. the distribution $p(x, y \mid T = 1)$.

1. If we're fitting $\hat{f}(x)$ to data from $p(x, y \mid R = 1)$ (without importance weighting), then we expect $\hat{f}(x) \approx \mathbb{E}[Y \mid X = x, R = 1]$. And if we could fit $\hat{f}(x)$ to data from $p(x, y \mid T = 1)$ then we would have $\hat{f}(x) \approx \mathbb{E}[Y \mid X = x, T = 1]$. Naturally, you think that we'll want to try importance weighting to use data from $p(x, y \mid R = 1)$ to estimate $\mathbb{E}[Y \mid X = x, T = 1]$, which is what we'd get with data from $p(x, y \mid T = 1)$. But wait! A colleague reminds you that we've assumed $Y \perp\!\!\!\perp T \mid X$ and $Y \perp\!\!\!\perp R \mid X$, which implies $\mathbb{E}[Y \mid X = x, R = 1] = \mathbb{E}[Y \mid X = x, T = 1] = \mathbb{E}[Y \mid X = x]$. And so, your colleague claims that importance weighting doesn't make a difference: we're estimating $\mathbb{E}[Y \mid X = x]$ no matter which data we're fitting on. Describe a circumstance when this claim is reasonable and a circumstance when it is not reasonable. (Hint: model misspecification)

2. Give an appropriate importance-weighted empirical risk estimate for $f(x)$ in terms of a loss function $\ell(f(X), Y)$ and the observed data described above. We'll only use it for learning $\hat{f}$, so don't worry about scale factors.

[2] There are organizations and companies that produce this type of thing. It's not a straightforward statistics or machine learning problem, since it's not clear there are any high quality labels to fit a model to. But we'll assume that somebody else has already solved this problem for us.

4.2 Using our regression to forecast the election

The goal of this section is come up with an estimator for $\mathbb{P}(Y = 1 \mid T = 1)$. As noted in the introduction, this will be our forecast of the election outcome.

1. Let $f(x) = \mathbb{P}(Y = 1 \mid X = x, T = 1) = \mathbb{E}[Y \mid X = x, T = 1]$. Show that

$$\mathbb{P}(Y = 1 \mid T = 1) = \frac{\mathbb{E}[\pi_t(X) f(X)]}{\mathbb{E}[\pi_t(X)]}.$$

   You can follow your own path, or use the steps in the subproblems below.

   (a) Show that $\mathbb{P}(Y = 1 \mid T = 1) = \mathbb{E}[f(X) \mid T = 1]$.

   (b) Show that $\mathbb{E}[T f(X)] = \mathbb{P}(T = 1)\, \mathbb{E}[f(X) \mid T = 1]$. (Hint Remember that $T \in \{0, 1\}$.)

   (c) Use the previous two results to show that $\mathbb{P}(Y = 1 \mid T = 1) = \mathbb{E}[\pi_t(X) f(X)] / \mathbb{P}(T = 1)$. (Hint: $\pi_t(X) = \mathbb{E}[T \mid X]$.)

   (d) Conclude the proof of this section by showing that $\mathbb{P}(T = 1) = \mathbb{E}[\pi_t(X)]$.

2. Propose an estimator for $\mathbb{P}(Y = 1 \mid T = 1)$ that uses an estimated regression function $\hat{f}(x)$ (such as the one developed in Section 4.1) as a plug-in estimate for $f(x)$, together with $\pi_t(x)$, $\pi_r(x)$, and a new large sample[3] $X_1, \ldots, X_N$ of covariates from P. Your estimator should converge to $\frac{\mathbb{E}[\pi_t(X)\hat{f}(X)]}{\mathbb{E}[\pi_t(X)]}$ as $N \to \infty$, though proving this is optional.

[3] In the election context, getting samples of just covariates X is generally cheap compared to getting samples of (X, Y) pairs.