**One of the main challenges in causal inference from observational data is “selection bias.” What do we mean by selection bias? (Hint: Think about the main difference between experimental data and observational data). Then, illustrate the problem of selection bias using a concrete example. Finally, describe briefly one method of causal inference that addresses (i) selection on observables, and (ii) selection on unobservables.**

One of the most important steps to take in causal inference studies when dealing with observational data is to account for selection bias. Selection bias appears in causal inference studies when the two groups that are being compared within the study have underlying differences that make them not truly comparable. This lack of comparability can bias study findings around average causal effect, for example, so it is critical to ensure that selection bias is accounted for and eliminated when undergoing these types of causal inference analyses. As noted, selection bias arises due to a lack of comparability between the groups within a study, and this lack of comparability can be from both “observable” and “unobservable” information about the group and the individual subjects within. At a high-level, causal inference studies have a key challenge where both treatments cannot simultaneously be evaluated on one individual or group. With this type of counterfactual being unobservable, causal inference from observational data instead must rely on these comparisons between groups.

To provide a more concrete example, the Week 1 lecture from SIADS 630 Causal Inference explains very clearly a causal inference study and how selection bias can play a substantial role in biasing the key findings and results. This particular example sought to evaluate two groups, insured and uninsured populations, and assess health outcomes to determine whether insured individuals would end up having improved future outcomes. In this example which evaluated both husbands and wives within households with and without insurance, the researchers found that both husbands and wives reported having better health when compared to those without health insurance. From this, one might think and conclude that having health insurance does, in fact, lead to better health outcomes. However, this finding has been influenced by selection bias and should not be considered reliable as it is not a ceteris paribus comparison. When the insured and uninsured groups were compared demographically, it was shown that the insured group typically was more educated and more likely to be employed, which are direct observable variables that are correlated with improved health status and future health outcomes. Therefore, selection bias was present in the study, the comparison was not ceteris paribus and therefore this causal inference had biased final findings. Additionally, a common note to discuss around selection bias is that when these large differences exist between two groups for observable variables, it is also likely that there will be differences in unobservable variables, as well. For this particular example, it’s possible that insured individuals may also have been more active lifestyles than those uninsured, which would be an unobservable difference that would again also impact future health outcomes. Fortunately, this selection bias problem can be handled by methods of causal inference that address selection on observables and selection on unobservables. Understandably, if the only difference between groups are observable differences, this would be a straightforward fix (with one example discussed below) whereas addressing both observed and unobserved is more challenging. In fact, and again as noted in the SIADS 630 lectures, one of the main challenges in causal inference is developing methods to reduce and/or eliminate selection bias within the two comparison groups to ensure that findings are not biased and that findings do hold value.

One method of causal inference that addresses selection on observables is matching, with one specific example being propensity score matching. This is a very effective approach that helps lead to a more comparable balance of observed covariates between the two groups that does make the comparisons more apples-to-apples especially when compared to causal inference studies that do not take these types of steps. At a high-level, propensity score matching, as noted in the lecture in Week 2 of SIADS 630 Causal Inference, can handle the “curse of dimensionality” problem and allow for matching even in cases where there is high-dimensionality in regards to observed covariates. This method will evaluate the observed covariates of individuals within the groups and match based on a propensity score, which is equivalent to the probability to get the treatment of the study. Matching is therefore one of the methods that addresses selection on observables.

One method of causal inference that addresses selection on unobservables is instrumental variables. Again, when there are covariates/confounders that have a relationship to both the treatment and outcome variables, and these covariates are unobservable, leveraging an instrumental variable is an effective strategy. In order for a variable to meet the instrumental variable requirements, the variable must cause a change in the treatment, must not influence the covariates/confounders and therefore be “as good as randomly assigned”, and, lastly, cannot influence the outcome (only the treatment). When these requirements are met, an instrumental variable can then be used to estimate the causal effect of the treatment without concerns around the impact of covariates on the finding. Instrumental variables are, therefore, one oof the methods that addresses selection on unobservables.

Text

Description automatically generated with medium confidence

Differences-in-Differences would be an effective causal inference method that would allow for the causal impact of mobile phone adoption by fishermen and retailers on price dispersion to be assessed. As can be seen above, Regions 1 and 2 had cell towers built in a staggered fashion and there is, therefore, overlap where Region 1 had mobile phone adoption while Region 2 did not. One potential strategy for this study would be to assess the difference in price dispersion in Regions 1 and 2 during this cross-section of time where the cell tower deployment was staggered. Based on the plots above, it does appear that Region 1’s cell towers completed construction in approximately week 23 of the survey, whereas Region 2’s construction completed approximately week 99 of the survey. Therefore, there’s approximately 80 weeks of survey data that can be leveraged to assist in this causal impact analysis around the adoption of mobile phones on price dispersion.

To set up this differences-in-differences approach, the process would closely mirror an example that was detailed in SIADS 630 Causal Inference week 4 lectures where a sample study wanted to assess a price change and whether it would lead to a revenue increase for a particular product. For this case study, the price of the product was changed in some locations and not others, which provided an effective control to account for any covariates that could impact revenue and sales (supply issues, poor economic conditions, etc.). In the case of the mobile phone adoption study, the two time periods being used would be weeks 0-22 where both regions had no mobile phones and then the proceeding ~80 week period where Region 1 had mobile phone adoption and Region 2 did not. This will allow Region 2 to serve as an effective control to the treatment Region 1. We, therefore, will be leveraging longitudinal data for this causal impact study and assessing price dispersion throughout points in time over the course of ~100 weeks. An important note to make is that the study is NOT set up to evaluate just the before and after price changes for both Region 1 and Region 2. If the treatment were to just be the adoption of mobile phones in both Region 1 and 2, the control would be both Regions 1 and 2 prior to their adoption. This fails, however, to provide an appropriate control to the treatment and whether any change in price dispersion is truly caused by this adoption or whether there were additional covariates biasing results. Therefore, the study will be set up in such a way to focus on the cross-section of data where Region 1 underwent mobile phone adoption while Region 2 was still under cell tower development. To again spell out the setup of this differences-in-differences method, the two groups will be Region 1 and Region 2. The two time periods will be weeks 0 – 22 which was prior to any mobile phone adoption by either region (t = 0) and weeks ~23 – 100 which was during Region 1 adoption with no Region 2 adoption (t = 1). Therefore, Region 1 receives the treatment at t = 1 while Region 2 does not, which is again to provide an appropriate control for this method. Specifically for this study, differences-in-differences will then assess the change in **average** price dispersion for both Regions 1 and 2 during times t = 0 and 1 with the following formula with DD corresponding to differences-in-differences and P corresponding to the average price dispersion seen for the time periods 0 and 1 and the treatment (T) or control (C) groups:

DD = (PT1 – PT0) – (PC1 – PC0)

This method would be an effective way to estimate the causal impact of mobile phone adoption by fishermen and retailers on price dispersion. However, there is one critical assumption that must accompany a differences-in-differences estimate and that is the parallel trends assumption. While untestable, the parallel trends assumption states that, without any treatment, both groups would have seen the same trend leading to their final outcome. Therefore, for this use case around mobile phone adoption, the parallel trends assumption would state that Region 1, during weeks 23 – 100, would have had the same trend exhibited by Region 2 around price dispersion if the treatment (mobile phone adoption) was not applied. This is a key assumption to hold for this method as it allows any deviations from the control group’s trend to be attributed directly to the treatment itself and not due to any observable or unobservable covariates at play. Lastly, in addressing part C of the above question, through both a review of the survey plots above and considering a differences-in-differences style analysis, price dispersion appears to decrease with the introduction of mobile phones, which is in line with expectations given that customers would have a greater ability to identify higher than expected prices with this technology. Through a differences-in-differences approach, we have greater assurance that the causal impact of mobile phone adoption by fishermen and retailers on price dispersion is not biased by any covariates and is due to the specific treatment itself.

**You’re working as a data scientist at a company and your manager is interested in using the social network information of a user, e.g., who their connections are, what their interests are etc., in predicting what content to recommend to users. She has heard of Convolutional Neural Networks (CNNs) and their success stories and is interested in figuring out if and how we can use CNNs for this task? Is such an application of CNNs feasible? How would you design such a system?**

* Input data in 2d format takes advantage of spatial nature on input. Images are 2d spatial data in image. Apply spatial filters that extract some features from this spatial data. High level features, low level features. Deep learning can extract different parts of the data
* **Cannot apply**. Network data would be difficult to project to 2D.
* Random Decision Tree would be simpler and easier to interpret
* CNN best for image detection. Social network like twitter feed. Use CNN to process if people tweeted images to filter out the images and turn into topics
* Combine with a topic model.
* Social network is spatial so maybe yes? Predict who would be in there network. Stephanie didn’t like this
* Matt thinks it’s a bad example. Describe what CNN and why you wouldn’t use it and then use something else (Decision Tree or Topic model)
* Turn recommendation problem into a classification problem

**Principal Component Analysis (PCA) looks for a lower dimensional subspace onto which the input data is projected so that the l2 reconstruction loss (i.e. least-squares) is minimized. Suppose you are now given an ND data matrix X and a target vector y of length N. Explain how PCA can be used to come up with a line that “best describes the data”. Define what “best” means here. How does this line compare with the one that results from ordinary least-squares linear regression and when would you choose one versus the other?   
Draw or plot an example for D=2 to facilitate your discussion.**

* Notebook 1 Unsupervised learning class
* Assuming N x D with vector Y and defined what PCA is
* Helps figure out which input variables matter. Which variables interact with variable of interest
* Could use SVD that produces same result as PCA – one of the lectures that uses principal multiplied by the vector
* The first principal component that PC says all the features that are correlated to the variable of interest
* The first PC the correlated features. 2nd shows the other features that are not correlated
* D = 2. Two features that are two principal components. Means two features and two PCs
* Defining best: SVD line is similar to the line of OLS regression
* PCA or SVD when high dimensional data that have to be reduced down. OLS is more simple but can be applied when high dimensionality is reduced down
* Sourced a diagram from unsupervised learning reading. PC 1 and 2 and the vectors
* Normalize – deduct the mean
* Multiply N x D to the transpose and then get the covariance matrix
* Best: covariance is the maximum
* OLS vs PCA. OLS involves outcomes. PCA involves features
* **Line is a series of points. PCA doesn’t involve target vector. Reduce to N x 1 when it’s best is when covariance is maximized. OLS regress target y. Best in the sense that minimize the square distance. OLS and PCA are different!!**
* PC1 on y, PC2 on x. Two lines that are orthogonal vectors

**Explain the concept of regularization in machine learning, supporting your explanation with at least four specific examples of different approaches to regularization taken from across the various machine learning courses you have taken in the MADS curriculum. Describe the reasons why one regularization method may be preferred over another in what scenario(s) and vice versa. Illustrate one example scenario where regularization is critical to the success of the machine learning method, and discuss whether there are any scenarios where regularization may actually hurt performance.**

There are a couple of regularization methods that can be leveraged in supervised learning tasks to aid in performance. Regularization is the introduction of some penalty with the goal of helping to reduce or control the model’s complexity. By reducing complexity, regularization is an effective tool in avoiding model overfitting and can also help improve cases where a model may be underfit without regularization. The first example of this concept would be L1 regularization, which can be seen in Lasso Regression as one example. L1 regularization looks to minimize the sum of the absolute values of model coefficients and therefore works by preserving influential variables and setting less important model variables to zero. L1 regularization uses an alpha parameter to control this effect, with larger alpha values leading to greater levels of regularization. With this type of effect, Lasso (L1) regularization should be used for models where only a handful of variables have a medium to large impact and the rest are small or negligible. The next example of regularization is L2 (Ridge). L2 regularization looks to minimize the sum of squares for the variable coefficients and therefore penalizes models with high coefficient variation. Similarly to L1, an alpha parameter is leveraged, with a larger alpha corresponding to greater regularization and a model that is likely underfit if the alpha becomes too large. Due to Ridge regularization’s penalty, and unlike Lasso L1 regularization, R2 regularization should be used for models where there are a large number of variables that have small to medium impact on the output prediction.

There are also regularization methods that can be leveraged when training neural networks, and these are dropout, early stopping, and data augmentation. For neural networks, similarly to above, model over and underfitting can be a problem during the training phase and regularization again provides a way to control complexity and allow for a generalizable model that is not overfit to its training data. The first example, dropout, removes a certain percentage of neurons from a hidden layer to ensure that a network does not become too dependent on any particular node when evaluating the input data and subsequent output. The second approach, early stopping, stops model training prior to any convergence to avoid the risk of overfitting. Lastly, data augmentation details the process of rotating or manipulating in some other way input as a way of strengthening the training data being used and again avoiding overfitting. While dropout and early stopping regularization methods can be used across neural network training, data augmentation would be best only for image classification type problems and likely should be limited to convolutional neural network (CNN) training only. The above details five regularization techniques across supervised learning and neural network training and the scenarios best served for their usage.

In terms of a concrete example where regularization is critical to the success of the training stage for a model, take data augmentation. For example, there is a task where the goal of a model is to identify whether an image contains a cat. In this type of image classification problem, the best approach typically involves leveraging a convolutional neural network (CNN). However, when training this CNN it’s discovered that the training data available is very limited and there are only a handful of images with cats to leverage for this work. Due to a lack of sufficient training data, data augmentation can be leveraged as a form of regularization that can help avoid an underfit model, as this example would create without this technique. To ensure adequate training, the limited images can be rotated, mirrored, and/or jittered to create a sufficient amount of examples of cat and cat-like images to ensure that the model is not underfit and can successfully identify a cat on the testing images. Therefore, a situation where training data is limited when training a CNN model would require regularization to ensure success.

Lastly, in terms of scenarios where regularization can actually hurt performance, there are cases where this is possible. Regularization is an effective tool for controlling model complexity, but if a parameter is dialed up too excessively high (like the alpha parameter in L1 and L2 regularization) this will likely impact coefficients to the point where the model is worthless. A complex model that is over-regularized would be bad and likely lead to underfitting. For Lasso regularization, for example, if influential variables are set to 0, this would lead to poor model performance due to an underfit. Similarly a lack of regularization (too low parameters) could result in overfitting not being accounted for, which again would lead to poor model performance when testing data is run through the model.

**Imagine that you are a data scientist working for a healthcare company, and your task is to develop a machine learning model to predict which patients are at risk of developing a certain disease based on their medical records. First describe and justify at least three specific steps you might need to take to preprocess the data. State one particular technical issue related to the data that you need to be vigilant about and how you may avoid it. Then explain how you would go about selecting an appropriate machine learning algorithm for this task (and give a specific example), considering factors such as the size and nature of the data, the complexity of the problem, and any relevant business objectives. Describe and justify the specific evaluation metrics you would use to evaluate the performance of your model and to compare different models. Finally, briefly summarize two ethical considerations that you might encounter while developing and implementing this machine learning model in a real-world setting, and how you might address them.**

To ensure the successful development of a machine learning model, there are a few preprocessing steps that can be taken. Based on my own experience in the healthcare space, one should assume that there will be missing data and steps should be taken to account for this. Data can be considered either missing completely at random (an unlikely assumption), missing at random conditionally, or missing not at random. Regardless, there are a few strategies that can be deployed to address missing values. The first is to undergo a complete case analysis and only include rows that have no missing features. If this approach is taken, it would be important to ensure that the number of rows missing data is small, so as to still have a significant amount of data to train and test on. If this type of approach is not preferred and there are human man-hours available, one could also deploy data wrangling and cleaning tools to help assist. If this is not an option, some learning models can handle missing values naturally, like decision trees, or can account for missing data if a missing value weight is included in the model. Lastly, a very powerful approach for overcoming missing data is to simply impute a value. One can leverage univariate imputation and add missing values for a feature based on the values present for that feature. This type of imputation can rely on the feature’s mean, median, most frequent value or even just impute some constant as needed. For a more robust imputation, multivariate imputation can be leveraged which imputes a missing value for a feature based on **other** features within the record. Therefore, handling missing values would be a critical step for preprocessing and ensuring that the data is robust enough for effective training and testing of the model. As another preprocessing step, one should consider normalization of the data as needed. In the healthcare space, some data may be measured on a 0 to 1 scale or a 0 to 100 scale, with blood pressure readings, glucose readings, HbA1C readings being some examples of commonly available data that would exist on different scales. To ensure that these features would be weighted equally and not based on scale, normalization would be critical as a preprocessing step to ensure adequate model performance. Beyond this, normalization helps combat outliers that may or may not exist in the data by putting the result on a scale that diminishes outlier effect. Lastly, in the healthcare space, the available data likely has class imbalance. For example, certain diseases may be very rare across society and information surrounding these illnesses may not be readily available or in short supply. Additionally, access to care may lead to disparities in terms of race and ethnicity presence within the data. To account for class imbalance and ensure that there is equal representation in the data to avoid bias, one could resample the data or ensure that an adequate number of records corresponding to a minority class are included and accounted for when the model begins training.

In terms of a technical issue related to the data that should be carefully accounted for, data leakage should be avoided. Again speaking from experience in the healthcare space, there are a few common situations where data can leak between training and testing resulting in overfitting. There is often a patient identifier, or even just a patient’s name, date of birth, etc. which, if available in both training and testing sets, could heavily impact the model’s performance. If the model focuses on name and other demographic fields, this would clearly be an overfit model that would not generalize well when presented with new data. As a result, and to ensure a robust model that avoids the overfitting issue, one should be careful around fields or information that could exist across the training and test sets and likely eliminate these features from consideration in the model.

In evaluating what type of model to use, a supervised machine learning model will likely be best in this scenario. The model would be trained on the available medical records and can be used to predict whether or not a patient is at risk of developing a certain disease (binary output). Knowing that healthcare and medical record data will likely be of considerable size, a logistic regression model would be a great option in predicting a binary outcome. Linear models like logistic regression are easy to train, but these simple types of models still provide fast predictions, which would be very valuable when perhaps working in a healthcare space which would require timely intervention of care. Additionally, knowing that healthcare and medical record data can become quite large, logistic regression is a great choice due to its ability to scale to particularly large datasets. As noted above, as well, not only will this type of data be large but its nature is to commonly have missing data and features. If the dataset was sparse even following our preprocessing steps, the model would still work well in this type of scenario. With the business objective in this case being the identification of at risk patients, and accounting for uncertainty around data size and complexity, the best approach to start with in this case of binary classification would be logistic regression, a supervised machine learning model. If we were to find following initial training and testing that the data is not well linearly separated, a more complex non-linear approach that again would work well for high dimensional data sets would be Support Vector Machines **with a non-linear kernel**. However, the model of choice to start in this example would be logistic regression.

While there are a few evaluation metrics that can be leveraged generally for machine learning models, the metrics that I will focus on will be precision and recall. The appropriate evaluation metric to use would depend on the objective of the model. For example, if in this case it is critical to catch as many at risk patients as possible at the expense of perhaps a larger number of false positives (casting a wider net is perhaps appropriate as a precautionary measure), recall would be the best metric to use to assess this model as it evaluates the number of true positives compared to all actual positives in the data. However, if instead researchers were more focused on ensuring that patients that were flagged as positive truly are positive (to perhaps avoid health scares and avoid undue stress placed upon patients) then precision would then be the best metric to use. Finally, If there was a situation where we’d want to have a better balance between precision and recall, then leveraging the F1 score would be the best approach. The F1 score ranges between 0 and 1 and will be higher when both precision and recall are higher. However, in this particular situation, the best evaluation metric for the model itself I believe would be recall to ensure that as many true positive cases as possible are captured, especially if the disease being researched is life-threatening. This metric could then be used to find the “best” model among the various choices detailed above.

Lastly, there are a couple of ethical considerations to note especially when dealing with healthcare data, which could contain PHI and PII, which, in the wrong hands, would be a huge concern. Therefore, the data must be handled to ensure patient privacy. Typically, data use agreements should be signed providing the necessary consent to use the information and then, once consent is granted, the proper guardrails should be put in place to make sure there are no data breaches and that the information is secure at all times. Additionally, while sex and race could at times be a predictor and feature in models, it’ll be important to account for class imbalance and minority representation in the data so as to not develop and train a model with biases in place. Leveraging a biased model may help identify majority patients but if the model performs poorly when assessing minority groups, this project would ethically fail and should be reconsidered. Questioning how the data and, subsequently the model, will be used is a critical step in ensuring that ethical considerations are accounted for early in the process.