Some Comprehension About Statistical Learning*

Juanling Liu (15320171151901)†

March 17, 2019

1 Statistical Learning and Causal Effect Learning

1.1 Relationships among variables

Statisticians have developed various techniques to help differentiate between causation, a variable directly related to phenomena, and association, a variable who's changes occur concurrently with the phenomena, and could be causal or non-causal. In class we have learned several causal diagrams.There are three structural reasons why two variables may be associated: (1) One causes the other; (2) They share common causes; (3) The analysis is conditioned on their common effects. Correlation implies causation, ie E [y|x = a]=E [y|do(x = a)].

Noncausal association occur in two ways, For example, (1) the disease may cause the exposure(rather than the exposure causing the disease); (2) the disease and exposure are both associated with a third factor , X , known or unknown.[1]

*I thank Muse for valuable comments. All errors are, of course, my own.

†Email: 8542308291@qq.com

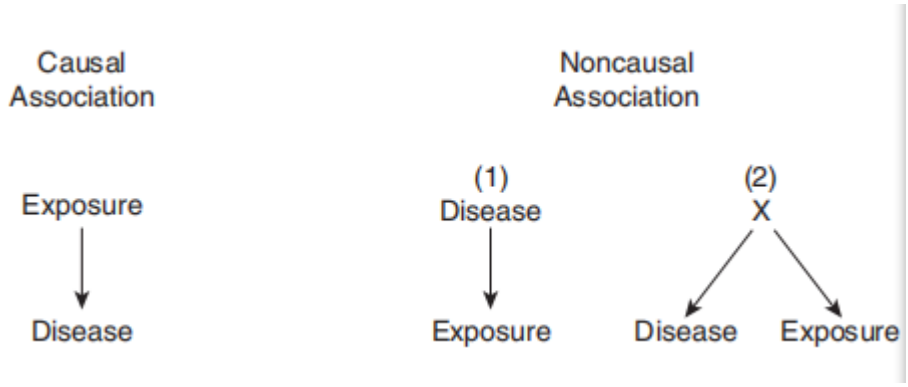

Figure 1: causal or non-causal

1.2 Statistical Learning

In econometrics,we estimate f not only for prediction,but also for inference.In many situations, a set of inputs X are readily available, but the output Y cannot be easily obtained,so we use a function f to predict Y,and we get $\hat{Y} = f\hat{(x)}$.We want to understand the relationship between X and Y , or more specifically, to understand how Y changes as a function of X_1,..., X_p. Since we need to know its exact form, $\hat{f}$ cannot be treated as a black box. What the target function looks like depending on what you want to minimize.Common choice is to minimize the expected squared-error loss: E [$(y - f(x))^2$] , so f (x) = E [y|x] is the target function that we want to learn.

There are two types of Learning. One is Supervised Learning,we wish to fit a model that relates the response to the predictors, with the aim of accurately predicting the response for future observations (prediction) or better understanding the relationship between the response and the predictors (inference). For example, linear regression and logistic regression, GAM, boosting, and support vector machines. The other is Unsupervised Learning, we observe a vector of measurements X_i but no associated responseY_i. The situation is referred to as unsupervised because we lack a response variable that can supervise our analysis. For example, clustering Analysis, PCA.[2]

1.3 Some trade-off

In the process of learning f, there are some trade-off. Typically as the flexibility of $\hat{f}$ increases, its variance increases, and its bias decreases. So choosing the flexibility

based on average test error amounts to a bias-variance trade-off

$$E(y - f(\hat{x}))^2 = Var(f(\hat{x})) + [Bias(f(\hat{x}))]^2 + Var(\epsilon) \tag{1}$$

Now, given these observations, and the hypothesis space H, we minimize the risk over all possible functions in the hypothesis space to find the best fit function g(x)

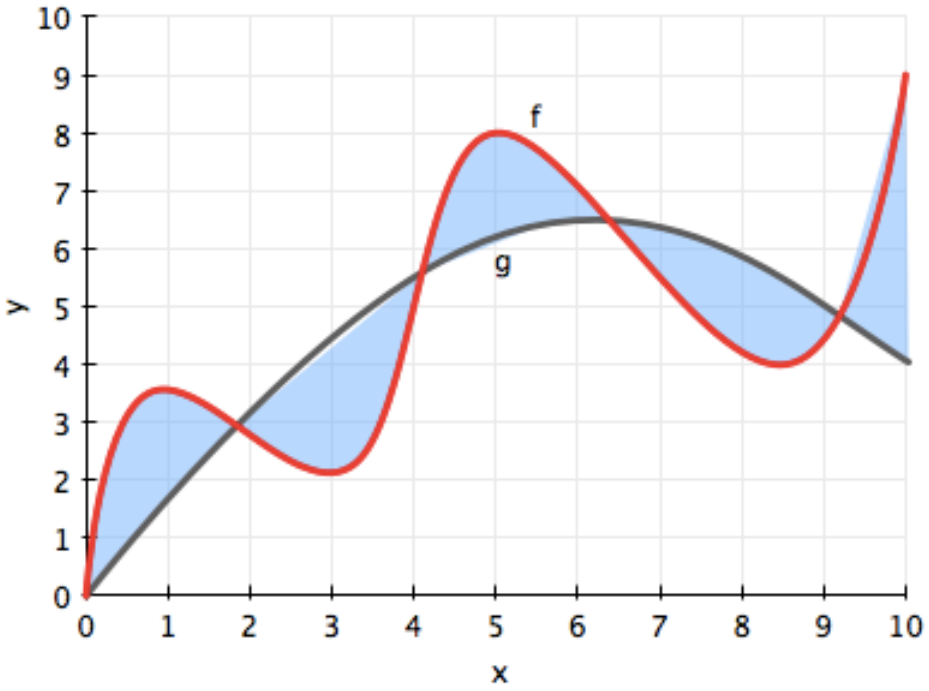

Figure 2: Bias

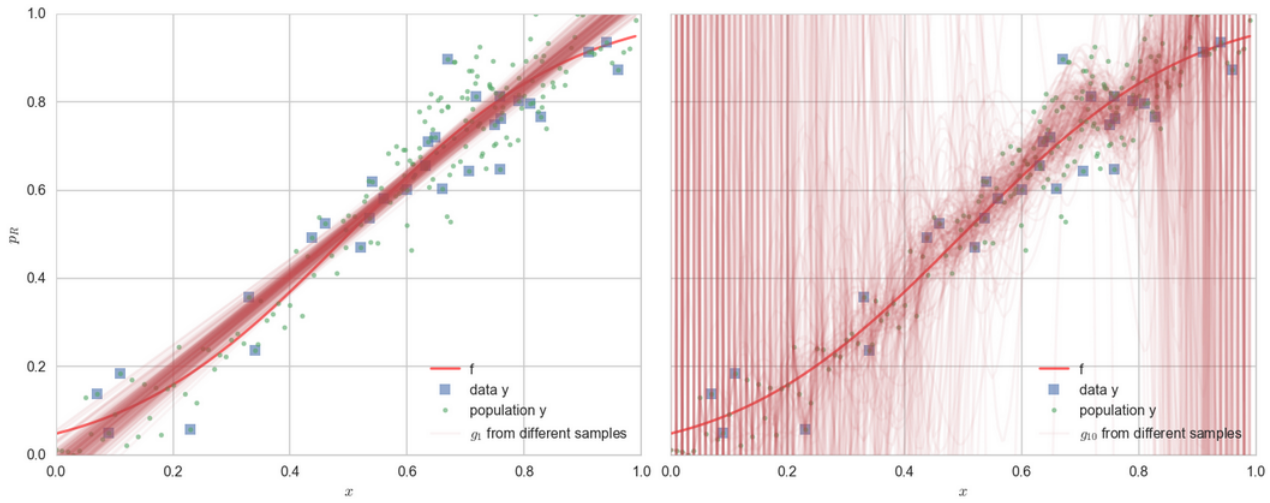

Figure 3: Variance

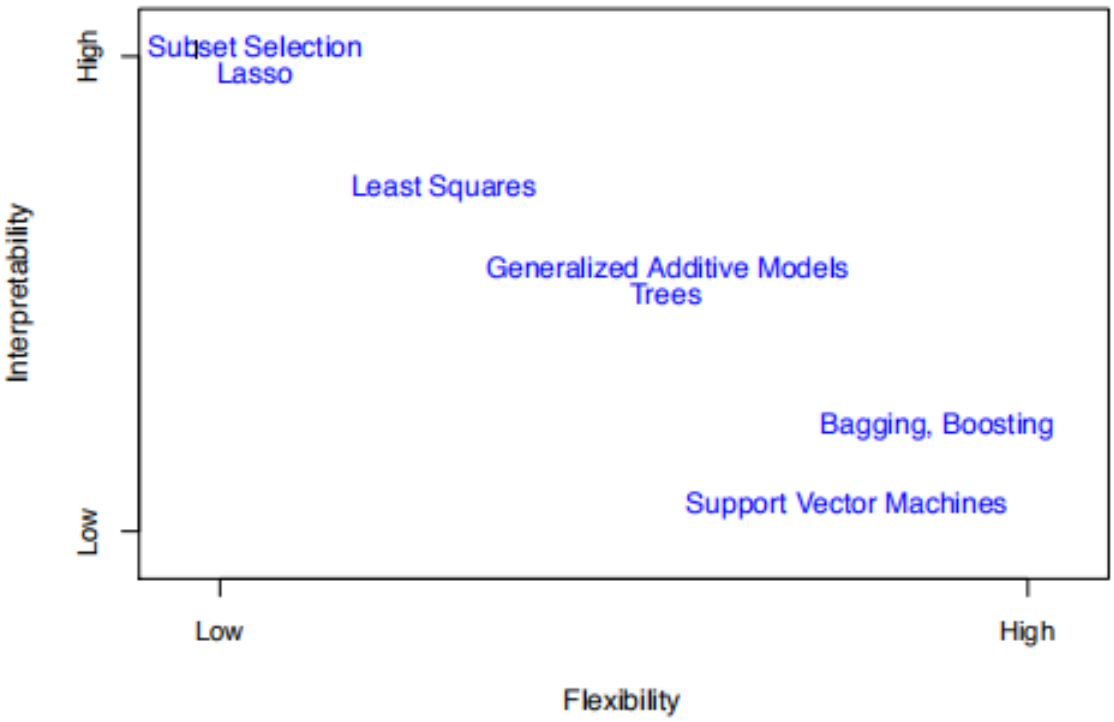

Figure 4: some trade-off

2 Counterfactual Analysis

2.1 Difinition

In the econometrics literature, causal models based on economic theory are referred to as structural models. Because structural estimation learns an entire structural model, once we have learned a model with variables $\{x_1, \ldots, x_n\}$, we can use it to generate data from the distribution p $(x_1, \ldots, x_n|$ do $(x_j = a))$ for any hypothetical manipulation do $(x_j = a)$. This is called counterfactual simulation.

Counterfactual analysis has been considered a potential approach for the assessment of policy impact. It has been extensively used to investigate the impact of various interventions, such as changes in taxation, welfare reform programs, education initiatives, and criminal rehabilitation. The counterfactual approach is suitable for answering fundamental questions, such as what would have happened if there had been no intervention or a different policy regime? For the counterfactual analysis, a non-observable case (referred to as 'counterfactual') is designed to compare with the actual one, shedding light on important factors that explain the impact of a policy .[3]

In reality, the task of constructing a non-observable case has difficulties. Traditional counterfactual approaches divide the entire study area into intervention and control groups (counterfactual). Furthermore, the counterfactual approach is often based on statistical techniques to identify causality between policies and their outcomes. The counterfactual simulation model of causal attribution applies to any domain where people are able to simulate what would have happened in the relevant counterfactual world. It predicts that people's causal judgments are a function of their subjective degree of belief that the causal event made a difference to whether or not the outcome would occur. People are predicted to compare the actual outcome with their belief about what the counterfactual outcome would have been. [4]

2.2 Examples

For example, a paper want to assess the impact of farmland preservation policies in China based on a counterfactual analysis approach by taking Huangmei County, one of the major grain-producing areas in China, as a case study. And find that farmland preservation policies have been successful in that County. Without strict policies, the arable land loss would likely have been exacerbated and urban expansion increasingly dispersed. [5]

Another example, a study aims to estimate the contributions of education unjust

or just to the income inequality by adopting their counterfactual simulation strategy[6]. Assuming that sources of income inequality include circumstances C_i, education E_i, demographic D_i and other unobserved factors, let w_i be income, $\Phi(w_i)$ be the distribution of income among a certain population.

$$\begin{aligned} ln(w_i) &= \lambda + \alpha C_i + \beta E_i + \gamma D_i + u_i \\ E_i &= HC_i + v_i \end{aligned} \tag{2}$$

$$ln(w_i) = \lambda + (\alpha + \beta H)C_i + \gamma D_i + \beta v_i + u_i \tag{3}$$

In Eq.(3),α denotes the direct effect of circumstance variables on income, andβH denotes the indirect effect of circumstances through education on income.The main strategy is to calculate the reduction in inequality, which will attain assuming the variables of circumstances have no effect on incomes, then the reduction can be taken as a measure of the contribution of inequality. Counterfactual estimation of the contribution of circumstances :

attain the predictive coefficients:

$$ln(w_i) = \hat{\lambda} + \hat{\Psi} C_i + \hat{\gamma} D_i + \varepsilon_i \tag{4}$$

calculating the predicted income after the residual is removed:

$$\hat{w_i} = exp[\hat{\lambda} + \hat{\Psi} C_i + \hat{\gamma} D_i] \tag{5}$$

Then, equalize the value of circumstance variable C_i with its mean for removing its impact on inequality, to get the counterfactual predicted value of income, denoted by$\tilde{w_i}^c$, namely:

$$\tilde{w}_i^c = exp[\hat{\lambda} + \hat{\Psi}\overline{C} + \hat{\gamma}D_i] \tag{6}$$

Next, calculate the index of inequality with $\hat{w}_i$ and $\tilde{w}_i{}^c$, denoted by I($\hat{w}_i$) and I($\tilde{w}_i{}^c$) respectively. We are then able to estimate the contribution of the circumstance variable C_i to the inequality of income by calculating the difference between I($\hat{w}_i$)and I($\tilde{w}_i{}^c$), namely:

$$\Phi^c = I(\hat{w}_i) - I(\tilde{w}_i^c), \tag{7}$$

With the prediction coefficients of the above equations, we can adopt the counterfactual simulation strategy to calculate the contribution of circumstances, efforts and decomposed education to the inequality of income.

3 Acknowledgement

Part of this notes is adapted from the following sources:

[1] https://www.studymode.com/essays/Statistics-Association-And-Causation-124293.html

[2] http://kuangnanfang.com/?id=25

[3][4]T.Gerstenberg, et al.From counterfactual simulation to causal judgment[J]. 2014.

[5]He J,et al. A counterfactual scenario simulation approach for assessing the impact of?farmland preservation policies on urban sprawl and food security in a major grain-producing area of China[J]. 2013

[6]JinyanZhou,et al. Contributions of education to inequality of opportunity in income: A counterfactual estimation with data from China[J]. 2019