Homework 1: Model Building and Model Selection/Fitting

Ziwen Chen
MACSS 19

1. Deviant aggressive behavior

If Theory I about deviant aggressive behavior were correct, an appropriate social policy that reduces such behavior should focus on building a comprehensive reward-punishment system. In this system, deviant behaviors should be punished by longer-term prison sentences or even the death penalty. On the other hand, those who follow the social rules should get proper material or virtual rewards. For example, people with no criminal records could have access to better mortgage or insurance plan. In doing so, people in society would learn to avoid deviant behaviors and get rewards.

If Theory II makes more sense, potential social policies that could lead to a decrease of deviant aggressive behavior should improve the relationship between authority and non-authority figures. For example, a better supportive institution that allows people to relieve their stress or frustrations via professional mental consulting service should help ease the tension and reduce deviant behaviors. Another possible solution is to build a more equal, more harmonious relationship with parents, bosses and public officials. Policies should encourage communication instead of obedience when people have conflicts with authority figures. When children report having friction with their parents, school or local community should help to provide space for equal and calm conversation. In the workplace, a respected third party (i.e., union) could have the power to interfere and fight for the proper right of employees. Social policy should also restrict the power of public officials by legislation. When public officials use their titles for private gain, the law should ensure that they will take accountability for their actions. By building a more salubrious relationship with authority and provide an effective stress alleviation method, there should be less deviant aggressive behaviors.

If Theory III provides the correct answer, to lessen deviant behavior, there should be more social policies that effectively eliminate social discriminations and ensure more equality. Assume that certain minority groups in society suffer from systematic racial or economic discrimination and therefore have a higher tendency to deviate. In such circumstances, social policy should get them out of the difficult situation by a series of affirmative actions. These minority groups should enjoy adequate welfare concerning education, rehabilitation, and financial support. As a result, oppressed groups would gradually gain equal rights as other people in society and thus commit less deviant aggressive behaviors.

If Theory IV reveals the actual mechanism of deviant aggressive behavior, social policies that dealing with deviant behaviors should aim at clearing up deviant subcultures. Stricter restrictions and regulations should be applied to cultural and media products to reduce exposure to deviant values. For example, the government should be more sensitive to the release of cult movies, music or books which depict violence, blood or any other aggressive behaviors. By using a harsh rating standard to restrict the audience, or cutting off those deviant scenes before publishing, the government could control the transmission of deviant subcultures. Apart from publishable culture products, content posted on the internet should also get examined to filter out unhealthy details. Trolling comments, disturbing videos or suspicious organizations on social media platforms should be flagged and removed in time. With less deviant subculture on the internet and other culture product, deviant aggressive behavior should decrease.

2. Waiting until the last minutes
(a) why the observation might be true and write down your explanations
There are several explanations that could support the existence of "waiting until last minute" behavior. One simplest explanation is that people avoid doing certain work because they dislike it. Most students hate writing papers because such work requires tremendous efforts and unpleasant revision procedures. Likewise, most professors hate grading exams because they regard it as a boring task that would occupy their time in research. Therefore, people avoid doing the works they hate until they are forced to do so.

Waiting until the deadline could also be because of the stress and anxiety brought by the planned work. In this case, people may be overemphasizing the work needed to be done, so that they end up under tremendous pressure from the high expectations. Students often believe that the paper could determine their final grades, therefore they give themselves huge pressures to write a perfect paper. Professors may also want to grade each student fairly and comprehensively which could, in turn, stress them. However, such stress could usually lead to extremely low efficiency. Only until the very end when people realize that it is impossible to submit a perfect work will they get rid of the stress and complete the work.

Another possible explanation of the procrastination behavior is to improve efficiency. A lot of people tend to have a higher level of efficiency and concentration when they are under the pressure of the deadline. It is possible that some people may deliberately arrange their plans and do every work just before the deadline to reduce the total time spent. Students may choose to write their papers several days before the submission day so they could have more time on other things. Similarly, professors want to squeeze more time for their research works so they choose to grade the exams until the very end.

(b) explanatory model

The explanatory model is based on the first explanation in (a). The basic hypothesis of the model is that people tend to procrastinate when they hate the content of the work.

The endogenous variable in this model is "delay"— which means the degree to which people tend to procrastinate the work. The factor "delay" could be measured by both self-reported surveys and observed behaviors. The self-reported delay could be based on questions like "From 1-5, to what extent do you feel you delay on work?" Observed "delay" could be measured in an experiment by calculating the gap between submission time and deadline and compare it with population mean. Based on self-reported and observed data, the "delay" indicator could be built.

The major exogenous variable of this model is the degree people dislike their works. This could be measured by survey questions like" From 1(hate) to 5(love), to what extent do you like you the work?". In this model, other exogenous factors could be demographic factors like race, gender, age, income, etc. They are applied to control the potential effect that certain demographic groups may have a special feeling about a specific type of works.

Generally, this model hypotheses a reverse relation between the degree of procrastination and the degree of love people feel about the work. The more people hate the work, the longer they delay.

(c) alternative model

This alternative model builds on the third explanation in (a). That is, the greater the efficiency people will have when chasing the deadline, the higher people value their efficiency, the more likely people tend to wait until the last minutes.

Same as (b), the endogenous variable in this model is "delay"— which means the degree to which people tend to procrastinate the work. The factor "delay" could be measured by both self-reported surveys and observed behaviors. The self-reported delay could be based on questions like "From 1-5, to what extent do you feel you delay on work?" Observed "delay" could be measured in an experiment by calculating the gap between submission time and deadline and normalize it with population mean. Based on self-reported and observed data, the "delay" indicator is built.

There are two exogenous variables in this model: the increase of efficiency, and the degree to which people value efficiency. Both factors could be measured using a self-report survey. For the increase in efficiency, it could be measured by questions like" From 1(lower) to 5(higher), to what degree do you feel about your DDL efficiency compared to your normal efficiency?". And for the value of efficiency, survey question could be like" From 1(don't value) to 5(highly value), to what degree do you value work efficiency?". Other exogenous factors could be

demographic factors like race, gender, age, income, etc. They are applied to control the potential effect that certain demographic groups may have a special feeling about a specific type of works.

Generally, this model will include three major explaining factors: the degree of increase in efficiency, the degree to which people value efficiency, and the interaction term between the two (efficiency*value). The model hypotheses that the higher people value efficiency and the more increases in efficiency when working until deadlines, the more likely people will delay their works until the deadline.

(d) prediction
Model (b) hypotheses that people tend to work until the last minute when they hate the work that they need to complete. If this model is correct, one possible prediction which could be derived from this model is that people will only procrastinate certain types of work while they could be highly motivated by other jobs. For example, social science students may start immediately in their paper reading assignments while choose to finish their statistics homework until the last minutes. This is because social science students generally hate statistics or mathematics while finding social theory fairly inspiring and interesting.

 Another possible prediction build on the model (b) is that mundane jobs will have more "procrastinated" workers than jobs that rely on creativity and intelligence. It is intuitive that most people will agree that being an artist, writer or researcher is more interesting than being a janitor, waiter or factory worker. That said, in general, people should hate more about previous jobs and love more about the latter jobs. Therefore, if the model (b) is correct, we should predict that the previous jobs should have more "wait until last minute" behaviors than the latter jobs.

 Different from model (b), model (c) believes that "waiting until last minutes" behavior is a rational choice that aims at maximizing the efficiency. The possible prediction would be that people who could effectively increase their efficiency near deadline and people who think highly of such benefit should be more likely to procrastinate. Since such qualities (i.e., efficiency and value) are usually independent of the specific work content, people who procrastinate one work is highly possible to procrastinate another work. In other words, constant procrastination behaviors should be observed from a fixed group of people. If one wants to predict whether an individual will procrastinate a future job, he/she should look at the individual's past working record.

Based on the model (c), we could also infer that last minute's works should have the same or even higher level of competency than works submitted several days ahead of the deadline. Again, people who tend to work until the last minutes make this rational choice because they realize that this strategy could allow them to spend less time in work of the same quality.

Therefore, students who submit their papers 10 minutes before the deadline should have an equal or even higher possibility to get an A than students who finish their papers one week ago.

3. Selecting and Fitting model

1. **Flexible vs. inflexible statistical model**
   a) In the situation of a large sample size and small predictor numbers, we should expect that the performance of a flexible model should be **better** than an inflexible model. This is because with a large sample size and limited predictors, the risk of overfitting the flexible model is reduced. Since the flexible model also has a smaller bias than the inflexible model, it is better to use a flexible model in this situation.
   b) When the number of predictors is very large while the sample size is small, inflexible models should have better performance than flexible models. This is because such kind of dataset will have a larger variance in general and inflexible model should be used to minimize the risk of overfitting. Since the dataset cannot contain information about too complex regularity, the benefit of using a flexible model to reduce bias is also very limited.
   c) The flexible model should be a better choice for fitting the dataset when the relationship between predictor and response is highly non-linear. This is because flexible models are more capable of capturing non-linear, complex relationships than inflexible models like linear regression.
   d) An inflexible model will generally perform better than a flexible model in this case. With a high variance of the error term, a flexible model will have a larger possibility to capture the noise hence result in a high variance in the model.

2. **Bias-variance**
The following explanations are based on this figure:
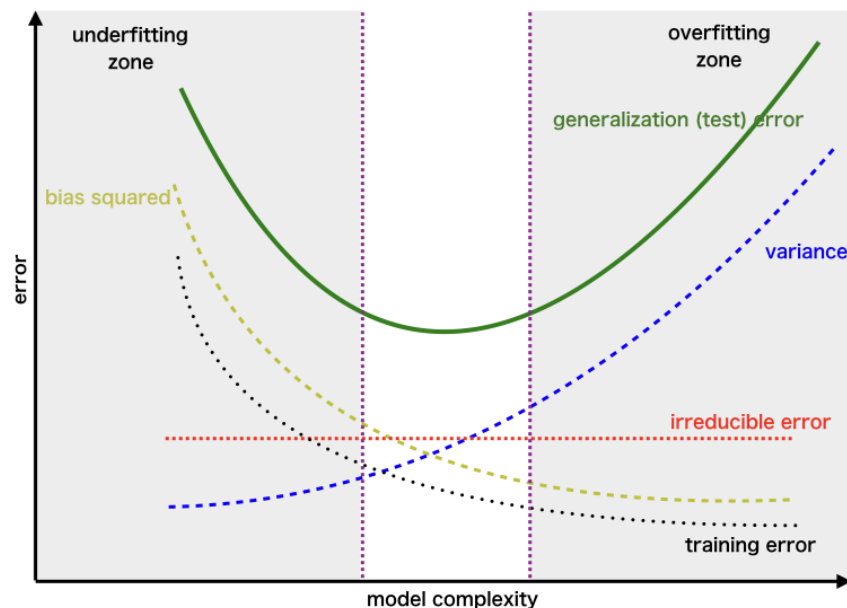
(1) Bias curve

Bias means the degree to which the current model fails to capture the patterns in the training dataset. Moving from a less flexible statistical learning method towards a more flexible statistical learning method (left to right), the model gradually learns more about the training dataset. That said, the more flexible the statistical learning model, the more complexity it captures, the less likely to underfit the training data, and therefore the smaller the bias error.

(2) Variance curve

Variance measures the risk of overfitting of the current model. With a more flexible model and more complexity captured, the possibility of the model fit the noise in the training dataset and lose the generalization power also increase.

(3) Training error

Training error means the difference between the predicted value and actual training data. With a more flexible model to better fit the training data, training error also decreases just like the bias curve. However, since training error also contains other noise which cannot be reduced by modifying model procedure (i.e., the constant irreducible error), the curve of training error is always higher than the bias curve.

(4) Test error

Test error is calculated by running the trained model with the testing dataset. Test error starts to reduce when the trained model learns general patterns from the training dataset (i.e., reduce bias). However, when the model gets too complexed and overfit the data, the variance increases and thus the test error increases again. The tradeoff of bias and variance results in the parabola shape of the test error curve. Moreover, test error is jointly decided by the sum of bias, variance and irreducible error. Therefore, the value of test error is higher than all of the three separate error curves.

(5) Irreducible error

Irreducible error is the noise in the dataset which cannot be minimized by the model selecting procedure. Therefore, the value or irreducible error always holds the same from the inflexible model to the flexible model.