厦门大学《微观计量》课程试卷

主考教师：茅家铭　　试卷类型:(A 卷)

Microeconometrics

Final Examination

Suggested Solutions (62 Points)

Multiple Choices (2 points each)

1. For binary discrete choice problems, let $y \in \{A, b\}$, and U denote the corresponding utility. For the model

$$U_{iA} = \alpha_A + x_i'\beta_A + e_{iA}$$
$$U_{iB} = \alpha_B + x_i'\beta_B + e_{iB}$$

   Which statement is true? (*Haihua Xie*)

   (a) $\alpha_A, \alpha_B, \beta_A, \beta_B$ can all be identified.

   (b) α_A and α_B cannot be separately identified. β_A and β_B can be separately identified.

   (c) β_A and β_B cannot be separately identified. α_A and α_B can be separately identified.

   (d) **Neither α_A and α_B, or β_A and β_B, can be separately identified.**

2. Suppose you are doing forward stepwise selection to select the variables, using AIC as your criterion. The results of step 3 and step 4 are as follows. Would you continue to do step 5 and what's the best model among all models you have fitted from step 1 to step 4? (*Jingyan Jiang*)

Step 3	**AIC**	**Step 4**	**AIC**
Y ~ A+E+D	3590.43	B	4017.19
		C	3595.37
		F	3596.63
		G	3592.82

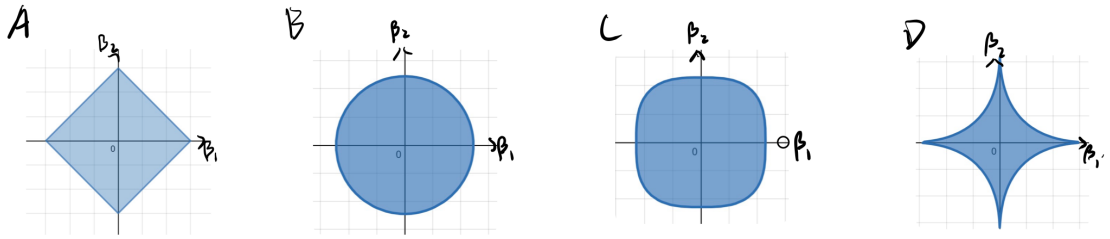
   (a) Continue; $Y \sim A + E + D + B$

   (b) Continue; $Y \sim A + E + D + G$

   (c) Stop; $Y \sim A + E + D + G$

   (d) **Stop;** $Y \sim A + E + D$

3. Regularization can be expressed as constrained optimization (Ivanov form). Which of the following represents the lasso constraint set in $\mathbb{R}^2$? (*Siyuan Wang*)

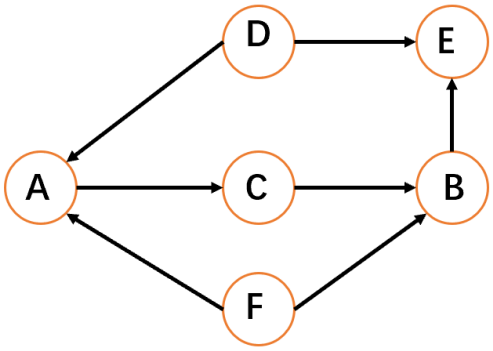


`Ans: A`

4. Which model is likely to perform best when predicting nonlinear relationships with unknown structure and modeling interactions with a requirement for high interpretability? (*Penghuan Huang*)

   (a) Linear regression
   (b) Logistic regression
   (c) Lasso regression
   (d) **Random forest**

5. Given the following causal graph, suppose we are interested in investigating the causal effect of A on B, in which of the following cases we will not be able to obtain a nonparametric identification, but will be able to parametrically identify the causal effect? (*Haihua Xie*)

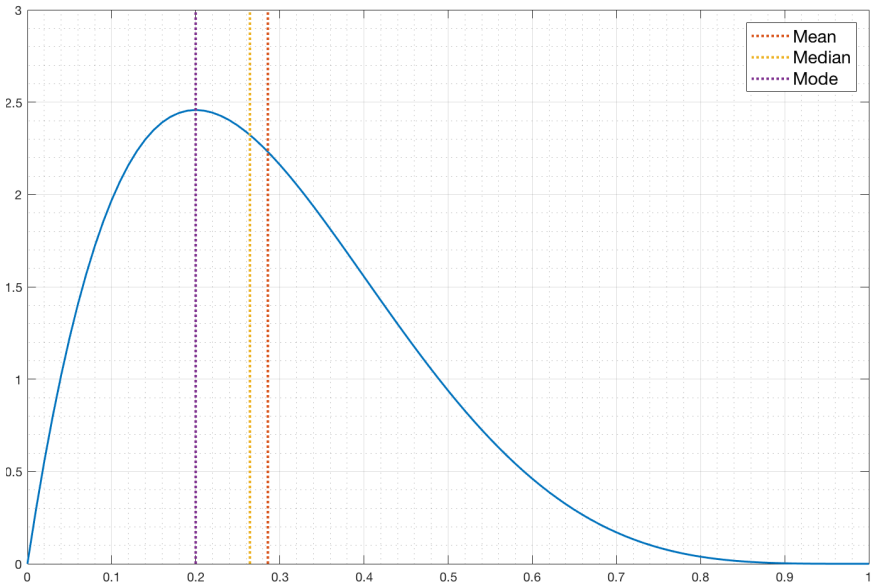


   (a) We observe $\{A, B, C\}$.
   (b) **We observe** $\{A, B, D\}$**.**
   (c) We observe $\{A, B, E\}$.
   (d) We observe $\{A, B, F\}$.
   (e) **We observe** $\{A, B, D, E\}$**.**
   (f) We observe $\{A, B, E, F\}$.

Problems

Problem 1 (8 points)

Given variables $\{x, y\}$ where y is a continuous variable, suppose we know the true conditional distribution $p(y|x)$ and want to predict the value of y at $x = a$. Let our prediction be $\widehat{y}$. (*Simrit Rattan*)



$p(y|x = a)$

1. Given different loss functions $\ell(y, \widehat{y})$, we will obtain different $\widehat{y}$. Write down the mathematical expression of $\widehat{y}$ as a function of $\ell(y, \widehat{y})$.

$$\widehat{y} = \arg\min_{c} E\left[\ell(y, c) | x = a\right]$$

2. For each question below, write down the name as well as the mathematical expression for the loss function.

   (a) What $\ell(y, \widehat{y})$ will result in $\widehat{y} = E(y|x)$?

   `Squared-error loss:` $\ell(y, \widehat{y}) = (y - \widehat{y})^2$

   (b) What $\ell(y, \widehat{y})$ will result in $\widehat{y} = \text{Median}(y|x)$?

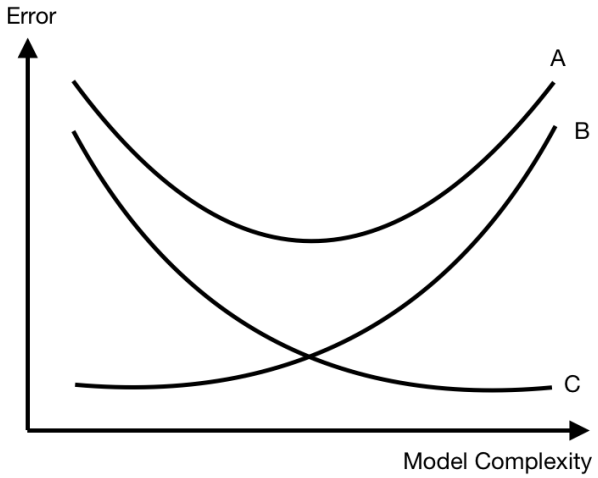
   `Absolute-error loss:` $\ell(y, \widehat{y}) = |y - \widehat{y}|$

   (c) What $\ell(y, \widehat{y})$ will result in $\widehat{y} = \text{Mode}(y|x)$?

   `Zero-one loss:` $\ell(y, \widehat{y}) = \mathcal{I}(y \neq \widehat{y})$

Problem 2 (4 points)

Use the graph to explain the bias-variance-tradeoff and indicate what the lines A, B and C are. Write down the formula that shows the Bias-Variance-Tradeoff. (*Elena Riccarda Ziege*)



Solution. A = Total Error, B = Variance, C = Bias^2, Formula: $E[y - \widehat{y}] = Var(f(x)) + bias(f(x))]^2 + Var(e)$. As a general rule, as model flexibility increases, bias(f) will decrease and Var(f) will increase. More flexible models tend to have higher variance because they have the capacity to follow the data more closely. Thus changing any of the data points may cause the estimate $\widehat{f}$ to change considerably. The bias decreases because a more complex model fits the true model better. As the model flexibility increases, the bias tends to initially decrease faster than the variance increases. Then at some point increasing flexibility has little impact on the bias but starts to significantly increase the variance.

Problem 3 (4 points)

1. Write down a K class multinomial logistic model.

$$\Pr(y = j|x) = \frac{\exp(x'\beta_j)}{\sum_{\ell=1}^{J} \exp(x'\beta_\ell)}$$

   , with one class – say class m – serving as reference class, for which $\beta_m = 0$.

2. Write down the expression for decision boundary separating class j and class k.

$$\log \frac{\Pr(y = j|x)}{\Pr(y = k|x)} = x'(\beta_j - \beta_k) = 0$$

Problem 4 (4 points)

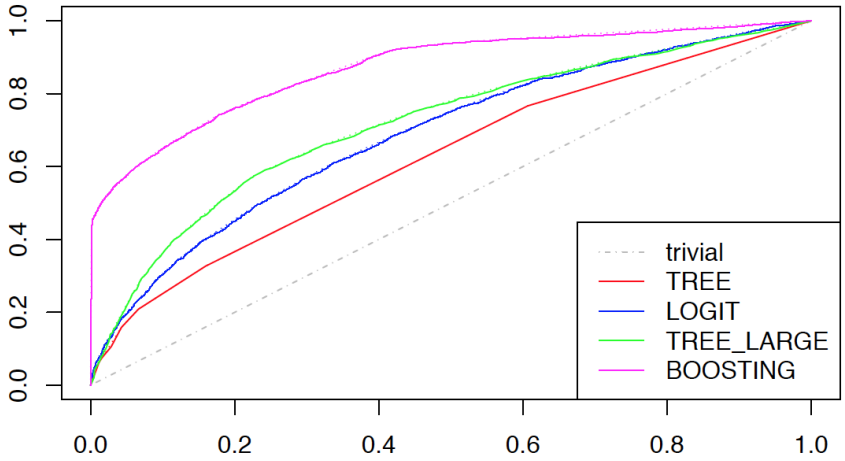
Suppose you are using a discriminative probabilistic classification model (e.g., logistic regression) to decide whether a patient needs to undergo further examinations for heart disease based on the data of his/her regular health check-up. Let $y = 1$ denote *yes* and $y = 0$ denote no. The classification result using $\widehat{p}(y = 1|x) = 0.5$ as the threshold is as follows. If you want to improve your prediction (to better help the doctors), should you increase or decrease the cutoff threshold for predicting $y = 1$ and why? (*Jingyan Jiang*)

		True	
		0 (No)	1 (Yes)
Predicted	0 (No)	923	55
	1(Yes)	10	43

Solution. I should decrease the cutoff threshold. Because from the perspective of the doctors that are trying to identify individuals with high risks of heart disease, the FNR is what's important. The FNR of the result above is about 56% >50%. So, I should turn down the cutoff threshold to reduce the FNR.

Problem 5 (4 points)

What is the AUC and what does it tell us? The following graph shows the ROC curve of the different models: Classification tree (TREE), Logistic regression (LOGIT), a large classification tree (TREE LARGE) and boosting (BOOSTING). State what two measures are plotted against each other to construct the ROC curve. Describe how an ideal ROC curve looks like. According to the plot, which models are preferred? (*Elena Riccarda Ziege*)



Solution. The AUC is the area under the ROC curve. The larger the AUC the better the classifier. The ROC curve plots the sensitivity against 1-specificity. (alternative solution: It plots 1-False Negative Rate against the False Positive Rate.) An ideal ROC curve hugs the top left corner. According to the plot, the best model is the boosting, followed by the large classification tree. The logistic regression is the third best and the small tree is the worst model.

Problem 6 (2 points)

Piet Mondrian is a famous Dutch artist who got the idea of abstract painting from drawing trees (Fig 1). Let's take a look at one of his representative abstract paintings (Fig 2) and manage to draw a decision tree based on it. Which of the following trees can be used to describe it? (*Chang Liu*)

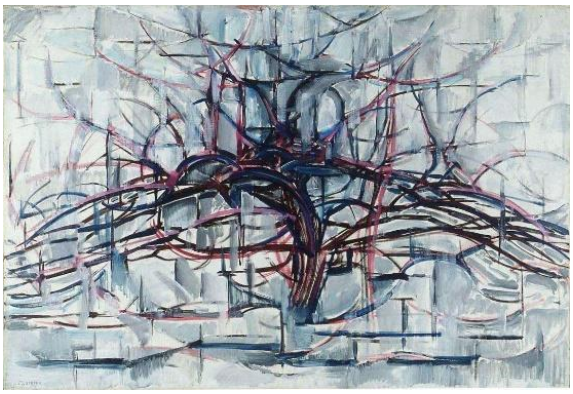


Fig 1 The Gray Tree

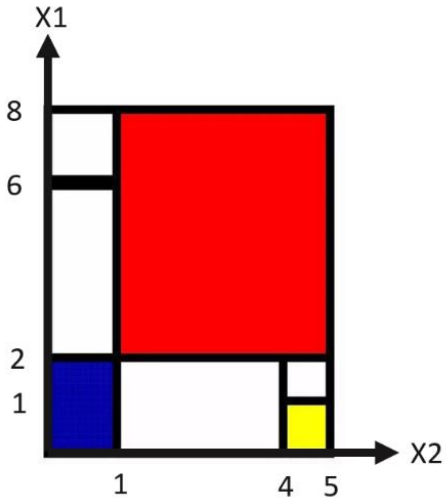
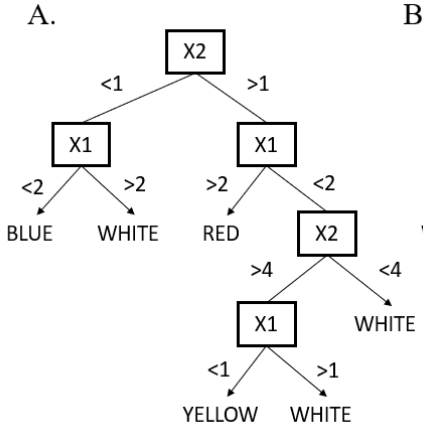
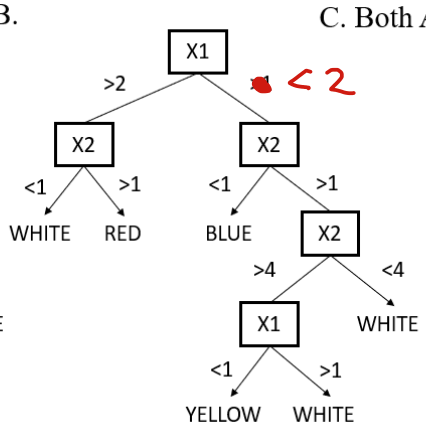


Fig 2 Decision Tree

A.



B.



C. Both A and B

D. Neither A nor B

Solution. C

Problem 7 (6 points)

Given a sample $\{(x_{1i}, x_{2i}, y_i)\}_{i=1}^{6} = \{(4,1,1), (6,6,0), (9,5,1), (1,2,0), (7,3,1), (5,4,0)\}$, we try to build a decision tree to classify y based on $x = (x_1, x_2)$. We use cross-entropy as the node impurity measure. For region R_m, the cross-entropy is defined as:

$$Q_m = -\sum_{j=0}^{1} \widehat{p}_j^m \log \widehat{p}_j^m$$

, where $\widehat{p}_j^m$ is the proportion of observations in R_m that belong to class j. (*Yu Keren*)

1. What is the cross-entropy at the root of the tree?

$$-\frac{1}{2}\log\frac{1}{2} - \frac{1}{2}\log\frac{1}{2} = \log 2$$

2. What is the optimal first split? Write your answer in a form like $x_1 \geq 4$ or $x_2 \geq 3$. Hint: you should be able to eyeball the best split without calculating the entropies.

   $x_1 \geq 7$

3. For each of the two leaves after the first split, what is the rule for the second split?

   $x_2 \geq 2$

Problem 8 (4 points)

Briefly state the advantages and the disadvantages of the decision tree model. (*too many came up with this question, so no one gets the credit, hahaha – this is called "Tragedy of the commons".)*

Solution. Pros:

- Highly interpretable
- Automatically detecting nonlinear relationships
- Automatically modeling interactions

  Cons:

- High variance; relatively poor predictive performance
- Difficulty in capturing simple relationships

Problem 9 (2 points)

Describe why it is difficult for bagging to improve the performance of linear regression models. (*Yifei Fang*)

Solution. A bagged linear model is still a linear model and the OLS solution is already BLUE (best linear unbiased estimator). (In contrast, a bagged decision tree is no longer a tree. Bagging decision trees enlarges the hypothesis set, while lowering the variance of the tree estimator).

Problem 10 (4 points)

As the critic Neil Postman points out in his book *Amusing ourselves to death*, "all public discourse is increasingly emerging as entertainment and becoming a cultural spirit." In contrast to RAI, Italy's state monopoly, Mediaset, a privately owned television station founded by Silvio Berlusconi, broadcasts entertainment programmes to the area that the tower's signal could reach. (The coverage of the area can be seen as a quasi-natural experiment). Did exposure to Mediaset's programming influence voters' support for Berlusconi?

	Area Mediaset's signal can reach (broadcast entertainment programmes about Berlusconi)	Area Mediaset's signal cannot reach (entertainment programmes about Berlusconi were not able to wacth)
Low-educated	531/894 voted for Berlusconi (60%)	2342/5328 voted for Berlusconi (44%)
High-educated	1792/4835 voted for Berlusconi (37%)	265/769 voted for Berlusconi (34%)
Combined data	2323/5729 voted for Berlusconi (41%)	2607/6097 voted for Berlusconi (43%)

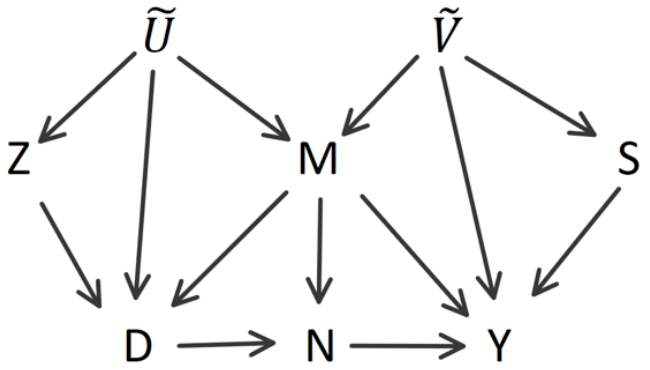
Real event but simulated data

In the table above, the combined data shows that watching Berlusconi's entertainment programmes was associated with lower voter support. But the pattern is reversed once we look at each education group separately. What is this phenomenon called? How to interpret the pattern? (*Jiabao Song*)

Solution. Simpson's paradox. One possible interpretation of the observed data is that less educated people are more likely to identify with populist parties whose language styles are straightforward and hence more likely to be influenced by Mediaset programming. However, in the area where Mediaset broadcast, there were more high-educated than low-educated people, leading to the observed aggregation reversal. Whether this interpretation is the right one depends on our causal assumptions on the data-generating mechanisms.

Problem 11 (6 points)

In the graph, we are interested in the causal effect of D on Y. $\widetilde{U}$ and $\widetilde{V}$ are not observed. (*Giacomo Sanna*)



1. Find all the open back-door paths from D to Y.

   $D \leftarrow Z \leftarrow \widetilde{U} \rightarrow M \rightarrow Y$, $D \leftarrow Z \leftarrow \widetilde{U} \rightarrow M \rightarrow N \rightarrow Y$, $D \leftarrow \widetilde{U} \rightarrow M \rightarrow Y$, $D \leftarrow \widetilde{U} \rightarrow M \rightarrow N \rightarrow Y$, $D \leftarrow M \rightarrow Y$, $D \leftarrow M \rightarrow N \rightarrow Y$, $D \leftarrow M \leftarrow \widetilde{V} \rightarrow Y$, $D \leftarrow M \leftarrow \widetilde{V} \rightarrow S \rightarrow Y$

2. Can we identify the causal effect of D on Y using the back-door criterion? If so, (1) write down the set of variables that would satisfy the back-door criterion. (2) write down the formula that expresses $E[Y|\text{do}(D=a)]$ in terms of statistical (conditional) distributions that you can estimate from the observed data.

   `No observed sets of variables satisfy the back-door criterion.`

3. Can we identify the causal effect of D on Y using the front-door criterion? If so, write down the formula that expresses $E[Y|\text{do}(D=a)]$ in terms of statistical (conditional) distributions that you can estimate from the observed data.

   `Yes. Assuming all variables are discrete, we have:`

$$E[Y|\text{do}(D=a)] = \sum_n \left\{ p(N=n|D=a) \times \left[\sum_d p(Y|N=n, D=d)\, p(D=d) \right] \right\}$$

Problem 12 (4 points)

Suppose $\{z_1\}$, $\{z_1, z_2\}$ both satisfy the back-door criterion for identifying the causal effect of x on y. Then we have:

$$E[y|\text{do}(x)] = \int E[y|x, z_1]\, p(z_1)\, dz_1 \tag{10}$$

$$= \int\int E[y|x, z_1, z_2]\, p(z_1, z_2)\, dz_1 dz_2 \tag{11}$$

Given infinite data, (10) and (11) are equivalent ways of computing $E[y|\text{do}(x)]$.

In finite sample, however, they are *not* equivalent.

Now, ideally I should type these equations up rather than pasting them as an image, but the student who designed this problem pasted them as an image, so I am pasting them as an image here by taking the image from her submission, which she took from my lecture slides ... so what is this problem about? Explain why and how equations (10) and (11) are not equivalent in finite samples and the trade-offs involved in choosing the best set of variables to control for using the back-door strategy. (*Congying Yuan*)

Solution. Given *nested* models $\mathcal{H}_1 = \{h(x, z_1)\} \subset \mathcal{H}_2 = \{h(x, z_1, z_2)\}$, if the integration in (10) and (11) can be done perfectly – if we know the true $p(z_1, z_2)$ – then choosing between $\mathcal{H}_1$ and $\mathcal{H}_2$ is a statistical variable selection problem: the best model generates the smallest prediction error. In practice, adding more control variables means we have to integrate over higher dimensions to obtain $E[y|\text{do}(x)]$. Doing so increases the variance of the estimator in finite sample.