**Discussion Questions:**

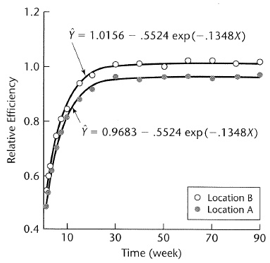
1. In the expression for the likelihood function, which rows have a pi, and which have a 1-pi?

and

2. Intuitively, why does it make sense to maximize the likelihood function when fitting the model?

Maximizing the likelihood function makes sense because we are finding the model that is the most likely to fit the data (Maximize the likelihood that you observed what you observed)

3. If facilities A and B had different asymptotic efficiencies as in Fig. 13.5, how would you modify the model?



I would add a dummy variable

4. If facilities A and B had different exponential rates, how would you modify the model?

I would capture it using an interaction term

5. If the objective was to determine if the two facilities had different asymptotic efficiencies, how could you do this?

Look at the p-value for the interaction term (Hypothesis test)

6. Are the formulae for t-tests, standard errors, etc. in a linear regression still valid? If not, how would you calculate and use the analogous quantities in nonlinear regression?

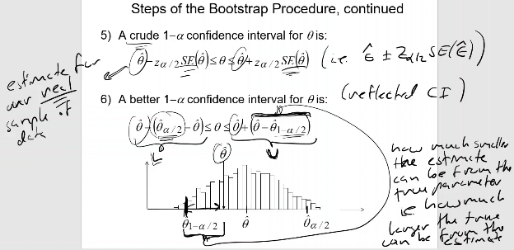
Answer

7. What is the difference between the two CIs (crude versus reflected) on the previous slide?

Crude: use

Reflected (Quantls): use or

The reflected CI was sifted left a little on the prev.



8. In general, when would the two confidence intervals differ?

When the distribution of is highly non-normal

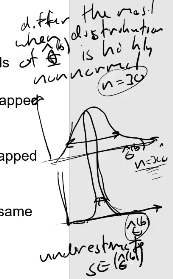
9. What are the effects of increasing B (Number of bootstrapped samples) on the bootstrapped histogram of a parameter estimate? Would the histogram become tighter?

A smoother/better estimate of the true parameter

10. What are the effects of increasing n on the bootstrapped histogram of a parameter estimate? Would the histogram become tighter?

The standard error would be less. The distribution would be very tight. Would underestimate the standard error. That’s why in bootstrapping, you should always use n where n is the size of your dataset

11. Why must n for each bootstrapped sample be the same as n for the real sample?



12. In boot.ci, type = "norm" gives our crude CI based on the SE and the normal percentiles, but translated by subtracting out the estimated Bias (taken to be the bootstrap average minus the original parameter estimate); type = “basic” interval gives the better CI obtained by reflecting the percentiles.

R calls it norm and basic whereas Apley calls it crude and reflected

13. How can we determine if there is statistically significant evidence that the asymptotic relative efficiencies of the two manufacturing facilities differ?

Perform hypothesis test. Does the confidence interval for contain 0?

14. What is a 95% CI on the asymptotic relative efficiency of the older facility (x1 = 0) and the newer facility (x1 = 1)?

Older: 95% CI on

Newer: 95% CI on

15. **In general, given the covariance matrix S of a random vector Z, the variance of the linear combination aTZ is Var(aTZ) = aTSa**

Ans

16. Which estimated coefficients (for l = 0 or for l = 0.2) make more sense?

0.2, Lambda always needs to be greater than 0 or else over-fitting will occur

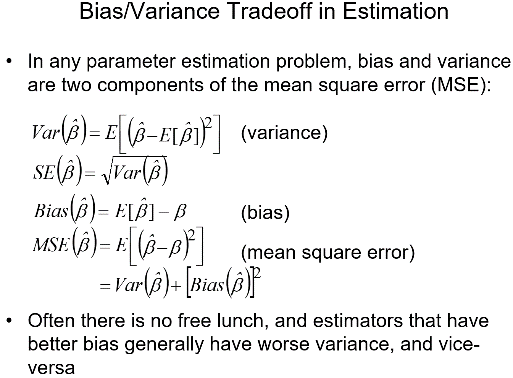
17. n-fold CV indicates that l = 0.001 is slightly better than l = 0.137 or l = 0.667, but this is probably not correct

18. In general, averaging across multiple replicates of K-fold CV is better than using n-fold CV

19. Regarding the bias/variance tradeoff, as increases:

increases

decreases



20. Why does the LASSO penalty tend to completely drop some terms from the model, as opposed to keeping the terms in the model but shrinking their coefficients (like ridge regression)?

When the value of lambda is large, the LASSO penalty will drive many of the coefficients towards zero, effectively dropping those terms from the model. Vice Versa for small lambda. LASSO penalty helps to reduce overfitting and improve the interpretability of the model by selecting a subset of the most important variables and setting the coefficients of the less important variables to zero.

21. What is the potential advantage of shrinking coefficients AND dropping some terms via LASSO, versus just shrinking coefficients via ridge regression?

Feature selection, model generalization by reducing the number of coefficients, and reducing overfitting, handles multicollinearity, sparse models, when coefficients are set to zero relationships are more easily

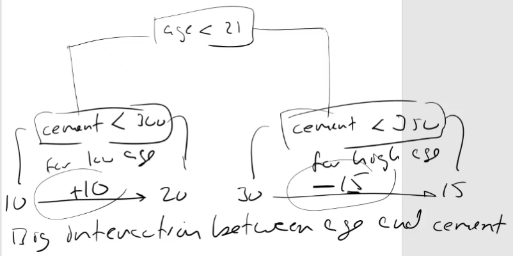
22. How can we get r^2 when R doesn’t provide it?

23. Provide an interpretation of the model and which predictor variables are most important

Which variable appears highest on the tree/the greatest number of times in the tree

24. Do there appear to be any interactions between Age and any other predictor?

Below or else ALE plots



25. To grow the initial tree larger (it is better to overgrow the initial tree, then prune it back) you can decrease minbucket and/or cp

26. What is the variable importance measure for trees?

Number assigned by R for variable importance. Branch length is proportional to reduction in SSE. If we want a numerical value for the branch length, we get that from the printout of the tree

27. If you want a summary measure of the predictive quality of the tree model, what would it be?

Use CV(misclass) or CV(deviance)

28. When fitting a tree model, never use misclass to split. Instead, use deviance (information) or gini

29. How can you tell what impurity measure rpart used to fit the model? R help command lol (gini default)

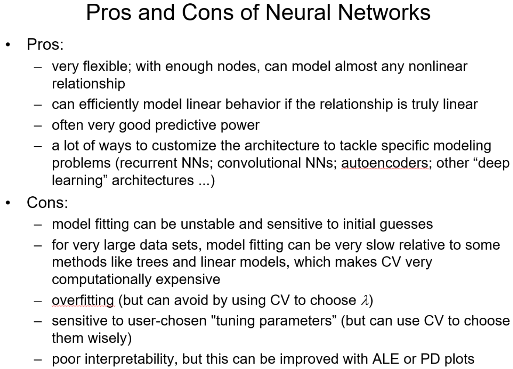
30. You choose the best size the same way – choosing the size with the lowest CV error measure (the best M was about 15—20)

31. If you want a summary measure of the predictive quality for the 6-class tree model, what would it be? Same as binary case

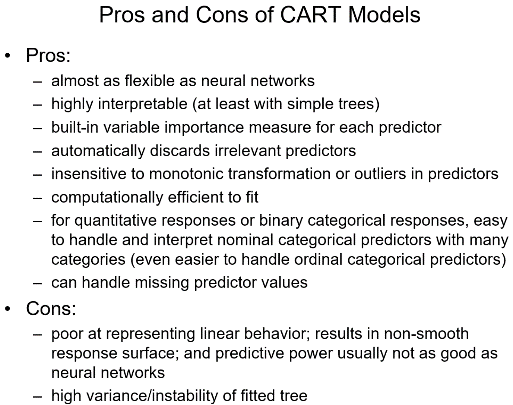
32. As always, you should deliberately overgrow the tree and then prune it back. How do know if you have overgrown the tree enough?

**Models:**

**Neural Net**

****

**Tree**

****

**Formulas:**

**AIC** (smaller is better | df(LogLik) = p+1)

**Extras:**

