STATISTICS 452/652: Statistical Learning and Prediction

September 22, 2020

Lecture 8: Step Functions

(Reading: ISLR 7.1–7.2)

1 Goals of lecture

- In Lecture 4 we saw that we can expand the notion of "variables" to include categorical variables, transformations, cross-products, and other engineered features
- This expands the number of possible surface shapes we can fit to data
- They are still limited, in that
  - each new variable type represents one specific shape
  - we have to choose the shapes that go into any model
- We can't [yet] create models than have one shape in one region and different shapes on other regions
  - Let's work toward that!!!

2 Indicator Variables and Step Functions

How do we create functions that can have different shapes in different regions of X?

- Start with 1 variable problem so we can see how things work
- Recall the definition of an INDICATOR VARIABLE (or "dummy variable") as a way to make numeric variables out of categories
  - For a categorical variable with Q levels, $X_q,\ q = 1, \ldots, Q$ is 1 if the observed value of X is the qth level, and 0 if not
- We now more generally define an INDICATOR FUNCTION, $I(a)$:

  - "a" here is just a placeholder representing any "test" or "event"
  - Then

$$I(a) = \begin{matrix} 1 \text{ if } a \text{ is true} \\ 0 \text{ if } a \text{ is false} \end{matrix}$$

- For example, we can represent all of the indicator variables we learned earlier as indicator functions:
  - $X_q = I(X = q)$
- We can define indicator functions ("indicators") more generally than for categorical variables, though
  - Roll a die and score if the result, X, is an even number. We only care about even vs. odd.
    - $I(X \text{ is even})$ represents this
  - Is someone legally allowed to drink alcohol in BC? Let X=their age.
    - $I(X \geq 19)$ represents this
  - The last example is exactly how we can create functions that define different regions of X

Defining regions within X

- Basically, to define regions in X you
  - choose K CUTPOINTS where you want to break X into $K+1$ intervals
    - Denote cutpoints by $c_1 < c_2 < \ldots < c_K$
  - define $K+1$ indicators based on these cutpoints
    - $C_0(X) = I(X < c_1)$
    - $C_1(X) = I(c_1 \leq X < c_2)$
    - $C_2(X) = I(c_2 \leq X < c_3)$
    - etc
    - $C_{K-1}(X) = I(c_{K-1} \leq X < c_K)$
    - $C_K(X) = I(c_K \leq X)$
- Note that
  1. The difference between $<$ and $\leq$ decides which interval the exact value of the cutpoint goes into
     - Intervals are often expressed in parentheses or (square) brackets
       - Parentheses mean that the interval is "open", meaning that the endpoint is *not* in the interval
       - Brackets mean that the interval is "closed", meaning that the endpoint IS included in the interval

        * For example, with integers 1 to 10,
            - [1,10] includes all 10 numbers
            - (1,10) excludes 1 and 10 but includes everything in between, equivalent to [2,9]
            - (1,10] or [1,10) each include only one endpoint
    2. It doesn't matter whether the cutpoints are included into the lower or upper end of an interval, as long as every cutpoint goes into exactly one interval.

- For example, suppose X is age, and you want to define age groups as < 18, 18–25, 25–35, 35–45, 45–55, 55–65, and > 65
    - We are making $K + 1 = 7$ intervals based on $K = 6$ cutpoints
    - Cutpoints are $c_1 = 18$, $c_2 = 25$, $c_3 = 35$, $c_4 = 45$, $c_5 = 55$, $c_6 = 65$
    - We have to decide, e.g., whether someone who turned 25 today is 18–25 or 25–35, but this is an open choice and there is no rule that says how it must go.
    - So our indicators to define these regions might be
        * $C_0(X) = I(X < 18)$
        * $C_1(X) = I(18 \leq X < 25)$
        * $C_2(X) = I(25 \leq X < 35)$
        * $C_3(X) = I(35 \leq X < 45)$
        * $C_4(X) = I(45 \leq X < 55)$
        * $C_5(X) = I(55 \leq X < 65)$
        * $C_6(X) = I(65 \leq X)$
    - I have chosen to put the "$\leq$" on the lower end, since certain laws or rules might change *on* the 18th and 65th birthdays.

- Essentially the process of creating regions turns numerical variable into a categorical one with $K + 1$ levels
    - I suggested earlier that we should avoid this, but we will soon see a special reason for doing it.

3 Step Functions

- A STEP FUNCTION is just a structure built on indicator functions
    - The step function takes a different value in each region
    - *Within* a region, the value is constant
    - Called a STEP FUNCTION because at the region boundaries you suddenly step from one value of the function to another

- We can create prediction functions that are step functions by doing regression on region indicators

  - Each region has its own mean.

- Recall that a regression model on a categorical explanatory variable requires that we drop one indicator
  - Dropping the indicator for the first region first leads to the model

$$f(X) = \beta_0 + \beta_1 C_1(X) + \beta_2 C_2(X) + \ldots + \beta_K C_k(X) \quad (1)$$

∗ **Exercise: what does each of these parameters measure?**

Example: Age regions in the Prostate Data (L8 - Step functions.R) Prostate Cancer is a disease that affects men in their later years, so let's make 4 groups with cutpoints 50, 60, and 70 groupings. (With a larger data set I could make more regions if I wanted to, but in general we will get poor estimates if regions have very few observations.) In each case, I want the interval to be closed on the left, so that ages are 0-49.99, 50-59.99, etc. (although I know that ages are just integers here).

The `cut()` function does just what we want, listing the cutpoints with additional numbers cover the end intervals:

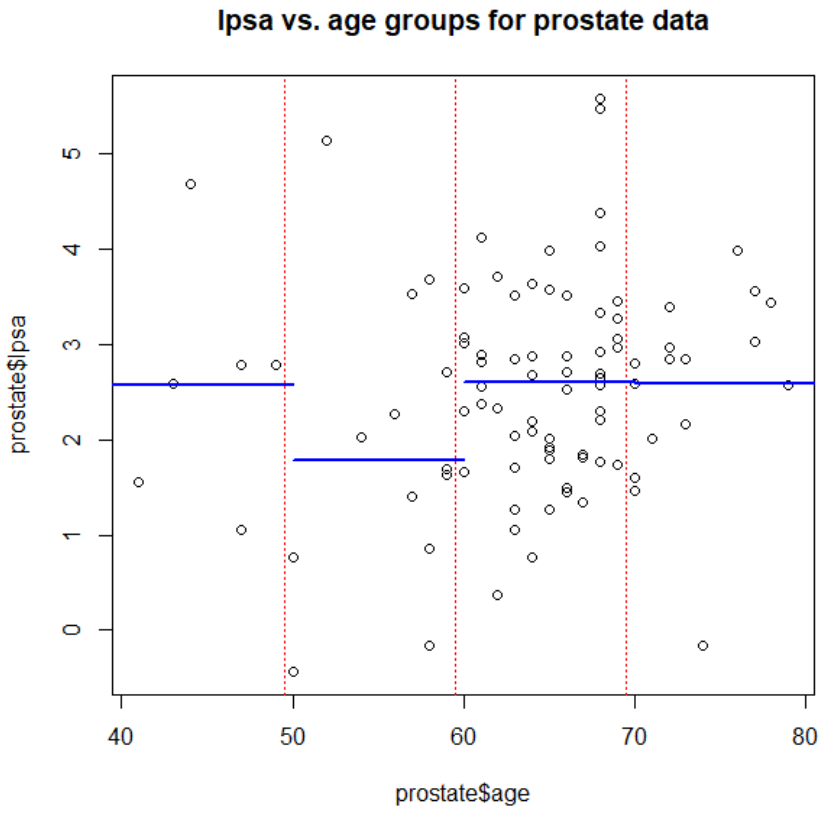
```
> regions = cut(x=prostate$age, breaks=c(0, 50, 60, 70, 100),
+                 right=FALSE)
> regions[1:10]
 [1] [50,60)  [50,60)  [70,100) [50,60)  [60,70)  [50,60)
 [7] [60,70)  [50,60)  [0,50)   [60,70)
Levels: [0,50) [50,60) [60,70) [70,100)
> head(data.frame(prostate$age,regions))
  prostate.age  regions
1           50  [50,60)
2           58  [50,60)
3           74 [70,100)
4           58  [50,60)
5           62  [60,70)
6           50  [50,60)
```

I performed the regression using `regions` as a factor in `lm()`. The results are in Figure 1. The plot of `lpsa` against `age` doesn't show much pattern, and the four group means are not hugely different from one another. Certainly, there is no apparent general increase or decrease in `lpsa` as age groups represent older patients.

3.1 Using the distribution of X to define regions

- It is very common to use percentiles or other statistics of X to define cutpoints
  - Often use median to compare top 50% to bottom 50%

Figure 1: Example data set from simulation. True relationship with X_1 (labelled as "X" in plot) is in red. Points from one sample of $n = 50$ are in blue.



  - Medical studies often break into 10 groups to define "deciles of risk"

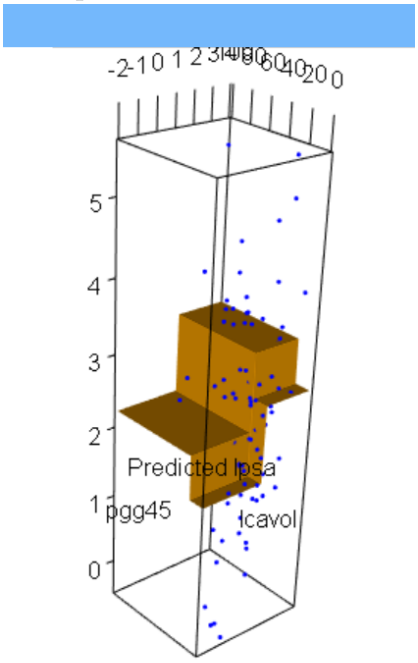
- This presents no problem:
  - Define in advance what aspects of the distribution of X you want to base cutpoints on
  - Use observed data to find the actual values

3.2 More than one variable

- Making regions based on more than one variable is possible
  - e.g., might have age and income groupings
    - Age: as above, cutpoints 18, 25, 35, 45, 55, 65
    - Income ($1,000): 20, 50, 75, 100, 150, 200, 500
  - We might naturally want to use the same income regions in each age group
- **Most commonly, we make the regions separately in each variable, and then combine them**
  - K_j cutpoints variable j
  - Can label them something like $c_{11}, c_{12}, \ldots, c_{1K_1}$ and $c_{21}, c_{22}, \ldots, c_{2K_2}$, etc.
  - This means that each group in one variable uses the same cutpoints in the other variables
  - Indicators for a combined region are cross-products of indicator functions from respective variables
    - e.g. Indicator for combining first interval for X_1 and 4th interval for X_2 is $I(X_1 < c_{11}; c_{23} \leq X_2 < c_{24}) = I(X_1 < c_{11})I(c_{23} \leq X_2 < c_{24})$
    - =1 if BOTH conditions are met, =0 otherwise
- Regression on the cross-products creates a separate mean in each combined region
  - Still a step function, but with (hyper-)rectangular regions
- As with a single variable, can use percentiles of any or all X_j to define cutpoints

Example: Multivariate regions in the Prostate Data (L8 - Step functions.R) I use this example just to show what a step function looks like in 3-D. We continue from the earlier examples with the Prostate data where we modeled on both `lcavol` and `pgg45`. For no particular reason other than it is something that people sometimes do, I split each variable at the median value to create a "low" and "high" group within each variable. Then I ran `lm()` on the model with both group variables in it. The fitted surface is shown in Figure 2. (For some reason, the `persp3d()` function in `rgl` made the bottom axes tiny for this plot! I could not see how to fix it.)

Figure 2: Example data set from simulation. True relationship with X_1 (labelled as "X" in plot) is in red. Points from one sample of $n = 50$ are in blue.



We see that the median values are not necessarily in the middle of the range, particularly for `pgg45`, which is very skew. The surface has created 4 separate means, according to the high and low regions of each variable. For both variables, the higher region also has higher mean. Also, the sizes of the steps, or jumps, taken between the low and high regions of `lcavol` are the same for both regions of `pgg45`, because I did not include an interaction in the model. (In this particular data set, if I add an interaction, the parameter estimate is nearly zero, so you wouldn't see its effects anyway.)

Optimal cutpoint selection

- All of the methods for selecting cutpoints use either prior knowledge or the distribution of X
  - They do not use Y in the selection of regions
- May not necessarily align well with biggest changes in mean of Y
- Natural question: is there an way to select cutpoints so that means are "optimally" modeled?
  - The answer is yes, but we won't do it now... :0
  - See "Regression trees"

4 What to take away from this

- We have created a whole new shape for regression functions
- It is a strange shape, because we rarely expect a true structure to jump suddenly from one value to another
- So step functions will not often be thought of as excellent prediction models off the shelf
- However, they play a *huge* role in developing other "machines" that are among the best predictors
- Also, they play a surprisingly important role in creating very flexible smooth regression models that can adapt to data automatically

5 Exercises

Concepts

1. Refer to Equation (1):
   (a) In terms of X and Y, what does the parameter β_0 measure?
   (b) In terms of X and Y, what does the parameter β_K measure?

Application

Refer to the Air Quality data described previously, and the analyses we have done with `Ozone` as the response variable, and the five explanatory variables (including the two engineered features).

1. Find and **report the median value for wind speed and temperature**
2. Use this median value to create high and low regions on both variables. Show values for `Temp`, `Wind`, and the two high-low region factors for these variables.
3. Fit a linear regression with the two region variables.
   (a) **Report the results from `summary()`.**
   (b) Do the two variables have statistically significant influence on the mean ozone level at the 5% Type 1 error rate? **Report their p-values and your conclusion. (No hypotheses needed.)**
   (c) Make a 3-D plot of the surface. **Report a screenshot from some angle that shows the whole surface and describe how the surface changes with each variable (use one short sentence each).**
4. Add the interaction of the two region variables to the model
   (a) **Report the results from `summary()`.**
   (b) Does the interaction have statistically significant influence on the mean ozone level at the 5% Type 1 error rate? **Report the p-values and your conclusion. (No hypotheses needed.)**
   (c) Make a 3-D plot of the surface. **Report a screenshot from some angle that shows the whole surface and describe how the interaction affects the surface (use one sentence).**