Homework 7

Advanced Methods for Data Analysis (36-402)

Due Friday April 5, 2019, at 3:00 PM

See the syllabus for general instructions regarding homework. Note that you should **always show all your work** and submit **exactly two files** (Rmd or R file as *code*, and knitted or scanned/merged pdf or html as *writeup*). Make sure that everything you submit is readable.

Problem 1: Additive Models

Refer back to previous HWs for a description of the abalone data set. (Yes, that data set again.) Once again, we will be looking to predict the "Shucked.weight" of abalones, in a nonparametric fashion, from other predictors.

Most of the tasks in the parts below can be performed using the `gam` package (though the linear models can also be fit with the standard `lm` function, of course). *General tip:* Make sure you utilize the functionality provided by the `gam` package to answer the questions; this will save you a lot of unnecessary programming. Also take a look at the additive lecture example (*R* Demo 8.1).

As in all data analysis problems, you should do exploratory analysis and examine residuals.

Part a

Fit a linear model of `Shucked.weight` on `Diameter`, `Length`, `Height`, `Rings`, and `Sex`. Call this `modelA`.

Part b

Fit an additive model of `Shucked.weight` on `Diameter`, `Length`, `Height`, `Rings`, and `Sex`. For each of `Diameter`, `Length`, `Height`, and `Rings`, use a smoothing spline with 4 degrees of freedom. For `Sex`, use a step function. Call this `modelB`. Plot the estimated regression functions (here there should be five), along with estimates of their (pointwise) standard errors. Describe and interpret the plots.

Part c

For each variable in the model from Part b (`Diameter`, `Length`, `Height`, `Rings`, `Sex`), fit two additional reduced models:

(i) one in which the variable enters linearly, and

(ii) one in which the variable does not enter at all.

For example, for `Diameter`, (i) would be the model

```
modc1=gam(Shucked.weight~Diameter+s(Length,4)+s(Height,4)+s(Rings,4)+Sex,
data=abalone}
```

while (ii) would be the model

```
modc2=gam(Shucked.weight~s(Length,4)+s(Height,4)+s(Rings,4)+Sex,data=abalone}
```

You can compare the three models using

`anova(modc2,modc1,modelB)`

For each variable, perform two F tests:

- for the null hypotheses that model (i) is true versus the alternative that `modelB` is true, and
- for the null hypothesis that model (ii) is true versus the alternative that model (i) is true.

(Note that for `Sex`, `modelB` is the same as (i), so only one additional model needs to be fit and only one F test needs to be performed.) What are the results of the F tests, and what effects would you make linear, or drop from the model, if any?

Part d

Next you will fit competing additive models to compare to `modelA` and `modelB`. The first two comparison models are:

`modelC` : `Shucked.weight` on `Diameter` (smoothing spline, 4 df), `Length` (smoothing spline, 4 df), `Height` (smoothing spline, 4 df), `Rings` (linear), and `Sex` (linear)

`modelD` : `Shucked.weight` on `Diameter` (smoothing spline, 4 df), `Length` (smoothing spline, 4 df), `Height` (linear), `Rings` (linear), and `Sex` (linear)

The final comparison model, `modelE`, will be introduced in part (g). To compare the predictive accuracy of the first four models, use 5-fold cross-validation. Report the cross-validation estimates of prediction error for each of the four models, as well as the standard errors of these estimates. Be sure to save the 5-folds so that you can use them again for `modelE`.

Part e

Compare the cross-validation results for `modelB` and `modelC`. Are the estimated prediction errors of the two models comparable (within one standard error)? Hence, is the action you took to drop a variable in part (c) justified, from the perspective of prediction error?

Part f

Compare the cross-validation results for `modelC` and `modelD`. Again, are their estimated prediction errors close, taking the standard errors into consideration (your solution does not need to include a formal hypothesis test)? What do you conclude about the importance of using nonlinear terms in `modelC` versus linear ones in `modelD`?

Part g

We see that `modelC` generalizes `modelD`, in the sense that it replaces the linear term (for `Height`) by a nonlinear term. Another way to generalize `modelD` is to include linear interaction terms. That is, consider fitting a final model:

`modelE` : `Shucked.weight` on `Diameter` (smoothing spline, 4 df), `Length` (smoothing spline, 4 df), `Height*Sex` (linear), and `Rings*Sex` (linear)

where in the above we have used `a*b` to denote the interaction between two variables `a` and `b` (as per the formula notation used in `gam` or `lm`). Use 5-fold cross-validation again to estimate the prediction error of this model. Compare the estimated prediction errors for `modelD` and `modelE`. Is one better than the other, again taking into account standard errors?

Part h

Now, we consider the complexities of all of the models as measured by their effective degrees of freedom. For additive models, you can think of the effective degrees of freedom of the final fit as simply the sum of the effective degrees of freedom of the individual smoothing operators. As a concrete example, consider the additive model:

- y on x_1 (smoothing spline, 7 degrees of freedom), x_2 (smoothing spline, 4 degrees of freedom), and $x_3 * x_4$ (linear),

where x_1, x_2, x_3 are continuous variables, and x_4 is a factor with 3 levels. This model has 7 degrees of freedom from x_1, 4 degrees of freedom from x_2, 3 degrees of freedom for the three slope parameters in $x_3 * x_4$, plus 3 degrees of freedom for the three intercepts in $x_3 * x_4$. This makes for a total of 17 degrees of freedom.

Rank the five models by complexity as measured by effective degrees of freedom. Assume that being less complex is better than being more complex. Based on complexity, the estimated prediction errors, and the analysis of residuals, which model seems best?