UWEO StatR201 – Winter 2013: Homework 4

Due: Monday March 4, 8 AM Pacific (grace period until March 7 lecture, with notification)

Assaf Oron, assaf@uw.edu

<u>Reading related to this assignment</u>: Lectures 6-7; Hastie et al., Sections 3.3-3.6, 7.1-7.7, 7.10,

- Please submit online in the class dropbox. Please submit either ***.pdf is accepted as the main submission (with code pasted verbatim), or *.rmd.** (tip: to save time when using knitr, use `'cache=TRUE'` in the chunk header for any code chunks that run big simulations – this will prevent them from re-running each time you recompile the code).
- **Starred (*) questions and question-parts are not required.** You may submit them if you choose, or do any part of them without submitting.
- **Grading is determined chiefly by effort, not by correctness. If your submission shows evidence of independent, honest effort commensurate with the amount of homework assigned – you will receive full credit.**

1. Complete the exploratory analysis of the automotive fuel-efficiency training dataset, such as:

- Examine and triage outliers and unusual values (e.g., odd-cylinder engines)
- Covariate transformation and discretization
- Nonlinear effect, e.g., examine splines for the `year` covariate, by plotting it vs. the residuals from a model that only includes vehicle weight (as done in Lecture 5 for the Boston dataset). **If you use spline terms, it's better to specify fixed knots than d.f.**
- Mine the vehicle-name information for potentially useful covariates
- Examine potential interactions via multi-way plotting
- Collinearity/VIF
- etc.

Note 1: you can use HW1 key for Question 4 (Boston dataset exploratory analysis) for a rough guidelines and ideas. Additional material is in the lecture notes.

Note 2: For all following questions, use the inverted, gas-consumption (rather than efficiency) variable `gp100m` (gallons per 100 miles) as our response. All models should be trained to predict this response. Needless to say, the `mpg` variable must be ignored.

Note 3: At the end of question 1, you need to have a script (or function) that takes the dataset and does all the covariate additions/manipulations you've decided upon. You will need that when you get the test dataset.

2. Implement a penalized-likelihood model selection tool via `step` or a similar function. Enable two-way interactions. Choose **only one** of the available directions ('forward', 'backward' or 'both'), and run it once using AIC and once using BIC. Also, run the tool once on a covariate set that was VIF-filtered

(i.e., if there are any covariates with VIF>5, remove the one with largest VIF – then re-calculate VIFs, and so forth until all VIFs are <5). Record the 4 resulting models.

3. Implement a constrained-subset cross-validation selection routine on your chosen collection of covariates and interactions, as demonstrated in Lecture 6-7. Record both the "top" model, and the best among the most parsimonious ones within 1 SE of the best CV RMSE.

4. Similarly, implement the Lasso and elastic nets, as shown in Lecture 7. Choose two values for α. One of them should be α=1 ("pure Lasso"), and the other no greater than, say, 0.7.

Then choose **one** of the three: "top" model, the one within 1 SE, or the one exactly between them. For this one model, store **both** the lasso-shrunk parameter estimates, and only the covariates for use via plain least-square estimates (i.e., using Lasso only as a model selection tool).

5*. Similarly, implement PCA **or** PLS. Do it once via "naive" CV, and once via "proper" CV.

6. When the test dataset is released, score all the models created/chosen in questions 2-5, for their RMSE, correlation-R-squared, and bias.

Warning: take care to use the training dataset's values and parameters wherever relevant. For example, if you scale the covariates, you need to scale the test-set covariates using the training-set mean and s.d.

Draw "Observed vs. Predicted" plots for all models/methods you used, with diagonal lines representing the +/- 1 gallon/100miles errors – as done in Lecture 7. Heuristically rank the methods according to their overall performance.

***** Starred option: take a dataset you are already familiar with – so that Question #1 is needless or easy for you (could be the Boston one, or the one you did your StatR101 project on), and repeat the Question #2-6 drill on it.**