UWEO StatR301 – Spring 2013: Homework 3

Due: Sunday May 12, 11:59 PM (grace period 4 days)

Assaf Oron, *assaf@uw.edu*

Reading related to this assignment: Lectures 4-5; Labs 3-4; Dobson and Barnett Chapter 11; Hastie et al. Sections 5.1-5.5, 9.1; material about GAM uploaded to Lecture 5 folder.

Instructions:

- Please submit online in the class dropbox. Either ***.pdf is accepted as the main submission (with code pasted verbatim), or *.rmd.** .
- **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. My first ggplot plot.

Create a "spaghetti" plot of the `ChickWeight` data, analogous to the first one shown on Lecture 4 (slide 3) – but using `qplot` or another function from the `ggplot2` package. **If you use that function's default, there might be one annoying "feature" to the plot. What is it? Try to suppress it.**

Also, do the plot both in the original and log-transformed scales (for chick weight).

2. Mixed-Effects

Staying with the `ChickWeight` data: in Lecture 4 I explored mixed models, with **the main inference question being the effect of diet upon chick weight gain.** The last model attempted (slides 22-24) used random slope; while it was better than others, the residual patterns were still not quite satisfactory. **Try to fix that.**

The top suspect in my view is that weight-gain is nonlinear, while "my" models assumed that it is.

You have two options – choose one: fixing the residual structure using tools from the `lme4` package (easier and more straightforward), or using a mix of GAM and mixed-effects, via `gam, gamm` or `gamm4` (the former from the `mgcv` package shown in Lecture 5; the latter from the `gamm4` package which wasn't shown).

The goal is to reach a model examining the Diet:Time interaction vs. weight (linear or nonlinear), and whose residuals are not correlated within chick, and don't show a trend vs. time.

Whether or not you succeed, please document and submit you attempts. As said above, credit is for sufficient effort and level of skill commensurate with our stage in the class sequence.

3. GAM for Prediction

In Lecture 5, I explored some GAM specifications for the `autos` data. In the end, a fairly straightforward model was used to predict MPG on the test dataset, with RMSE that seemed better than anything achieved by models used in StatR201 HW4.

But the observed vs. predicted scatter (Lecture 5, slide 18) showed a gross outlier: a car with 46 MPG, predicted to have only around 35 MPG.

a. Identify the particular car. ***(the dataset is available in the Lecture 3 folder)***

b. This calls for a more robust measure. Why? Instead of a conceptual/theory answer, just re-run the model **exactly as given in the notes**, and calculate the test-set RMSE with and without that single car.

c. Usually, **a single point out of 100 observations, should not have such a large impact on our overall prediction metric.** This suggests using a more robust measure. Here are two options: the mean absolute error, and the 3rd quartile (75th percentile) of absolute errors. **Calculate both metrics for the GAM model used in class – again, with and without the single 46-MPG car. Are they really more robust?**

d*. The notes also present a "legacy" model from StatR201 HW4, and one created on-the-fly using stepwise model selection (both ordinary linear models). Repeat (c) for them as well. Does the GAM model still beat them on these metrics?

e. Let's define a semi-robust prediction metric that is the average of the two defined in (c) (so, just a simple average of the mean abs. error and the 3rd quartile of abs. errors). Calculate the metric for the GAM model used in class. **Then, using the GAM modeling and selection tools (or exploring others available on the function's numerous help pages) – try to find a GAM that beats this score. WITHOUT outright cheating (i.e., NOT by using the test data to train the model, or by predicting from a ton of models and post-hoc choosing the best one).**

As in question 2, even if you don't succeed just document your efforts, etc.