Homework 6: "The sound of gunfire, off in the distance"

36-402, Spring 2016

Due at 11:59 pm on Thursday, 25 February 2016

AGENDA: Explicitly, logistic models, generalized additive models, and checking regression specifications. Implicitly, the perils of science by p-value.

Our data this week, `http://www.stat.cmu.edu/~cshalizi/uADA/16/hw/06/ch.csv`, comes from a study of the causes of civil wars. Every row of the data represents a combination of a country and of a five year interval — the first row is `Afghanistan, 1960`, really meaning Afghanistan, 1960–1965. The variables are:

- The country name;
- The year;
- An indicator for whether a civil war *began* during that period — the code of `NA` means an on-going civil war, while `0` denotes continuing peace;
- Exports, really a measure of how dependent the country's economy is on *commodity* exports;
- Secondary school enrollment rate for males, as a percentage[1];
- Annual growth rate of GDP;
- An index of the geographic concentration of the country's population (which would be 1 if the entire population lives in one city, and `0` if it evenly spread across the territory);
- The number of months since the country's last war or the end of World War II, whichever is more recent[2];
- The natural logarithm of the country's population;
- An index of social "fractionalization", which tries to measure how much the country is divided along ethnic and/or religious lines;

[1] I have been unable to find an explanation anywhere of why this rate is greater than 100 for some data points.

[2] This appears to count only civil and not foreign wars.

- An index of ethnic dominance, which tries to measure how much one ethnic group runs affairs in the country.

Some of these variables are NA for some countries.

1. (10) Fit logistic regression for the start of civil war on all other variables except country and year; include a quadratic term for exports. Report the coefficients and their standard errors, together with R's p-values. Which ones does R say are significant at the 5% level?

2. *Interpretation* All parts of this question refer to the logistic regression model you just fit.

   (a) (5) What is the model's predicted probability for a civil war in India in the period beginning 1975? What probability would it predict for a country just like India in 1975, except that its male secondary school enrollment rate was 30 points higher? What probability would it predict for a country just like India in 1975, except that the ratio of commodity exports to GDP was 0.1 higher?

   (b) (5) What is the model's predicted probability for a civil war in Nigeria in the period beginning 1965? What probability would it predict for a country just like Nigeria in 1965, except that its male secondary school enrollment rate was 30 points higher? What probability would it predict for a country just like Nigeria in 1965, except that the ratio of commodity exports to GDP was 0.1 higher?

   (c) (5) In parts (a) and (b), you changed the same predictor variables by the same amounts. If you did your calculations properly, the changes in predicted probabilities are not equal. Explain why not. (The reasons may or may not be the same for the two variables.)

3. *Confusion* Logistic regression predicts a probability of civil war for each country and period. Suppose we want to make a definite prediction of civil war or not, that is, to **classify** each data point. The probability of mis-classification is minimized by predicting war if the probability is ≥ 0.5, and peace otherwise.

   (a) (5) Build a 2×2 "confusion matrix" (a.k.a. "classification table" or "contigency table") which counts: the number of outbreaks of civil war correctly predicted by the logistic regression; the number of civil wars not predicted by the model; the number of false predictions of civil wars; and the number of correctly predicted absences of civil wars. (Note that some entries in the table may be zero.) Make sure the rows and columns of the table are clearly labeled.

   (b) (3) What fraction of the logistic regression's predictions are correct? (Note that this is if anything too kind to the model, since it's an in-sample evaluation.)

(c) (2) Consider a foolish (?) pundit who always predicts "no war". What fraction of the pundit's predictions are correct on the whole data set? What fraction are correct on data points where the logistic regression model also makes a prediction?

4. *Calibration* (10) Divide the data points into groups where the predicted probability of a civil war is 0–10%, those where it is 10–20%, etc. Calculate the actual proportion of civil wars for each group of data points. Give a plot where the horizontal axis is the predicted probability, and the vertical is the actual frequency. Does the plot go up the 45-degree diagonal? Should it, if the model is right? If it does not, do observed frequencies at least increase as the predicted probability goes up, so that civil war really is more common when the model says it has higher probability? (Again, this is if anything too kind to the logistic regression, because it's an in-sample comparison.)

5. (10) Fit a GAM with the same variables to the same data: smooth all the continuous predictor variables; do *not* include an explicit quadratic term for exports. (The ethnic-dominance variable is binary, and should be included in the model with `as.factor`.) Provide plots of the partial response functions. Which ones are at least roughly linear, and which are not?

6. (10) Calculate the confusion matrix for the GAM. What fraction of its predictions are accurate? How does that compare both to the logistic regression and the peace-always pundit?

7. (10) Repeat the calibration checking plot for the GAM. Are its probabilities closer to tracking actual frequencies, or further, than those of the logistic regression?

8. (15) Test whether the logistic regression is properly specified, using the GAM as the alternative model. (Follow the procedure in the notes.) What is the p-value? Explain, based on this test and any other results you have reported, which model you prefer.

RUBRIC (10): The text is laid out cleanly, with clear divisions between problems and sub-problems. The writing itself is well-organized, free of grammatical and other mechanical errors, and easy to follow. Questions which ask for a plot or table are answered with both the figure itself and the command (or commands) use to make the plot. Plots are carefully labeled, with informative and legible titles, axis labels, and (if called for) sub-titles and legends; they are placed near the text of the corresponding problem. All quantitative and mathematical claims are supported by appropriate derivations, included in the text, or calculations in code. Numerical results are reported to appropriate precision. Code is properly integrated with a tool like R Markdown or knitr, and both the knitted file and the source file are submitted. The code is indented, commented, and uses meaningful names. All code is relevant to the text; there are no dangling or useless commands. All parts of all problems are answered with actual coherent sentences, and never with raw computer code or its output.

EXTRA CREDIT (15): Start with the model which predicts a constant probability of civil war for all countries and years. Evaluate its log-likelihood out of sample through five-fold cross-validation. Now consider all one-variable GAMs, using all available predictor variables except country and year. Which one variable has the highest cross-validation log-likelihood, and is it higher than the trivial, intercept-only model? Consider all two-variable GAMs which extend the one-variable GAM you just picked: report their cross-validated log-likelihoods. Are the two variables you picked the two variables with the smallest p-values in the logistic regression? Should they be?