Homework 6

Solution

2 April 2017

1 Introduction

In this report, we examine under what circumstances, and to what extent, non-violent movements can achieve large-scale political change in the face of opposition from an existing state. We are especially charged with learning when non-violent movements are more likely to succeed than violent ones, and how large these differences in the probability of success are under different circumstances.

Three factors of particular concern are the role of state violence, the role of interventions from other states, and the role of democracy. Critics of non-violence as a political tactic have long charged that it is only effective when dealing with "nice" governments — ones which will not meet moral force with lethal violence, and which are democratic enough to be swayed by public opinion.[1] The effectiveness of foreign interventions, either on behalf of existing governments or of opposition movements, is meanwhile a continual source of debate within countries considering such interventions, and of course for those on their receiving end.

This report thus examines how well the success of anti-government movements can be predicted on the basis of their non-violence, whether the government engages in violent repression of the movement, the level of democracy of the government, and the presence or absence of foreign aid to the two sides in the conflict. We also consider a number of control variables: the duration of the movement, the year in which it peaked, whether the government was the target of international sanctions, and whether or not significant elements of the government's security forces defected to the movement. Sanctions and defections have obvious relevance to success; the ability of the movement to sustain itself for a long time may also be relevant; and year can serve as a proxy for political conditions extending beyond the borders of any one country, but perhaps affecting how well governments can resist internal pressures, or how willing opposition groups are to challenge state power.

1.1 Preliminary Examination of the Data

The units of analysis in our data set are particular political movements in particular countries. The data file contains 323 such movements, in 136 countries — this indicates that some countries have seen multiple movements, but, on examination, many countries have only one recorded movement.

The summary statistics of the variables reveal that there is little missing data (2 for `defect` and 54 for `democracy`). 32.8% of the movements were non-violent. The oldest movement is from from 1902, the latest from 2006, with most in the last half of the period (Figure 1), and one quarter of them all since 1993.

(This may be a bias in the sample, or it may mean such movements are becoming more frequent, and our analysis is more important than ever.) Otherwise, the only variable whose distribution turned out to be relevant to the analysis was `duration`, which is highly skewed to the right:

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       2     360     730    1928    2190   21170
```

Since it was part of our terms of reference to use logistic-additive models, we need to re-code partial successes as either full successes or complete failures. Since it's rare for any political force to ever be *completely* successful, we will re-code partial successes as successes. 51.4% of the movements then count as successful.

[1] See, for example, George Orwell's 1949 essay "Reflections on Gandhi".

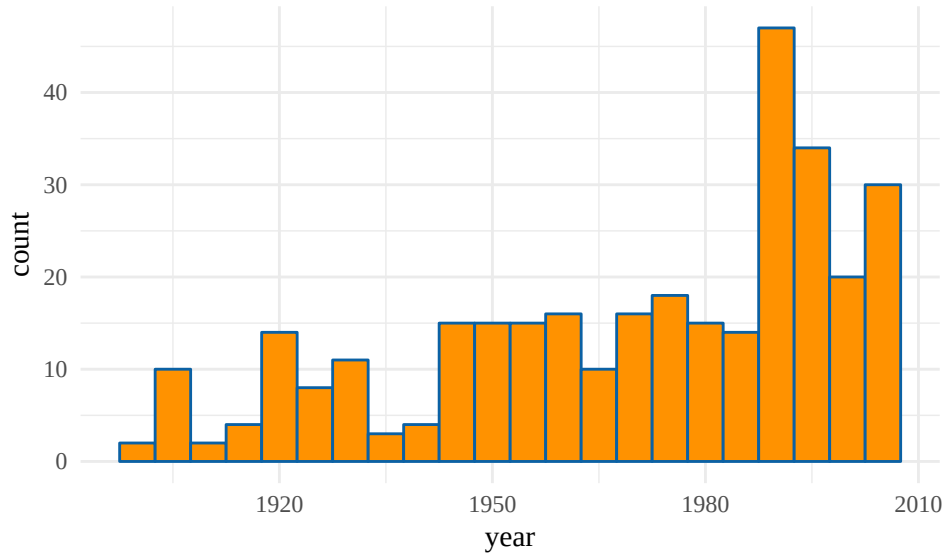

Figure 1: Figure 1: Histogram of peak years

2 The Model

2.1 Formulation

To address the scientific questions, we turn to a generalized additive model, which relates the probability of movement success to the factors of interest. Specifically, following our terms of reference, we use a **logistic** model, meaning that our model will directly predict the log odds of success, $\log p/(1-p)$, and the probability of success only after a transformation. Moreover, again following the terms of reference, we use an **additive** model — each term in the model makes a separate, additive contribution to the log odds (i.e., it multiplies the odds of success). Mathematically, writing x_j for the different predictor variables, we start with a model of the form

$$\log \frac{p}{1-p} = \alpha + \sum_j f_j(x_j)$$

where the f_j are smooth, possibly linear, functions of the predictors. (If x_j is binary, f_j is always linear!) We may also include terms in the sum which are functions of pairs of variables, to capture interactive effects or contrasts between conditions.

Our initial formulation of which terms to put in the model is guided by the terms of reference, and the analytical questions posed to us. We are told that all three continuous predictors — the year, the duration, and the level of democracy — should be included. We also include all of the categorical predictors. The analytical questions specifically ask about the interaction of non-violence (on the part of the movement) with violent repression (on the part of the government), so we include that interaction term. The last of the analytical questions also ask about the interactions between non-violence and democracy. We therefore include both a direct effect for democracy, and a partial response function interacting democracy with non-violence. Categorical predictors may not be smoothed, so they enter into the model as parametric terms; all the continuous variables will be smoothed, because we have no theoretical reason to favor any parametric form for them.

The specification we are led to, then, may be put in R form:

```
mdl.base <- gam(target ~ nonviol*viol.repress + sanctions + aid + support
                + defect + s(year) + s(duration)
                + s(democracy) + s(democracy,by=nonviol),
                data=navc, family="binomial", na.action=na.exclude)
```

The `nonviol*viol.repress` expression in the formula will give terms for the non-violence of the movement, for the presence of violent repression, and for a non-violent movement confronting violent repression; it is that last which tries to captures the difference in response to repression between non-violent and violent movements. Similarly, this model includes a direct effect of democracy, applying to all movements, and an additional smooth function of democracy that *only* gets added for non-violent movements. The last function allows the contrast between violent and non-violent movements to change with the level of democracy.

However, `duration` was extremely skewed; this can cause problems for smoothing when there are not that many observations. We thus also consider a model which is the same as the previous one but smoothes `log10(duration)` instead of `duration`. Using log base 10 rather than natural log helps us keep things interpretable (as most people grasp 10^3 better than $e^{6.9}$). To decide which version is better, we use cross-validation, and in particular the cross-validated deviance score which is computed by the `gam` function as part of its model-fitting. The cross-validated deviance without logging `duration` is 0.133, while that of the model with transformation is 0.127. Since taking the log of `duration` before smoothing it not only unskews, it leads to better predictions, we discard the first model in favor of the logged one.

2.2 Analysis, Including Estimates and Uncertainties

Table 1 gives the model's parametric terms. Here, and throughout, all confidence bands were calculated by resampling of cases.

Table 1: Table 1: Point estimates and 95% confidence intervals for the model's parametric terms based on the bootstrap replications.

	lo	est	hi
(Intercept)	-4.81	-0.36	3.24
nonviol	-521.21	3.12	214.51
viol.repress	-2.71	-0.65	1.75
sanctions	-2.19	-0.05	1.52
aid	-1.09	0.10	1.17
support	-1.07	0.72	2.19
defect	-0.16	0.91	2.43
nonviol:viol.repress	-6.29	-0.57	109.67

The coefficient on `nonviol` is positive, meaning that non-violent movements are more likely to be successful than violent ones are; indeed, the positive coefficient of 3.1 means that a non-violent movement's odds of success are 23 times those of an otherwise-similar but violent movement. Violent repression on the other hand lowers the odds of success. The fact that the interaction term `nonviol:viol.repress` is negative means that violent repression hurts non-violent movements more than violent ones — though a non-violent movement confronting violence still has odds 13 times better than those of a violent movement in the same situation.

The coefficient on `aid` implies that when the government receives aid from other states specifically to help it deal with the movement, the movement's odds of success go *up* by about 10%. This does not necessarily mean that such aid is counter-productive; other countries may only send aid to governments which are already in trouble, dealing with movements which would have been even more likely to win otherwise; the coefficient would combine the real effects of aid with this "selection effect"[2]. The coefficient on support from foreign governments to the anti-government movement is also positive[3], indicating an increase in the odds by a factor of 2. The coefficient on sanctions is negative[4], indicating that sanctions on the government predict a

[2]People who were in a hospital a year ago are more likely to be dead now than those who weren't, but not, for the most part, because their doctors killed them.

[3]Which could also be a selection effect: why waste it on a movement bound to fail?

[4]Perhaps the reverse of the selection-effect story for aid to the government?

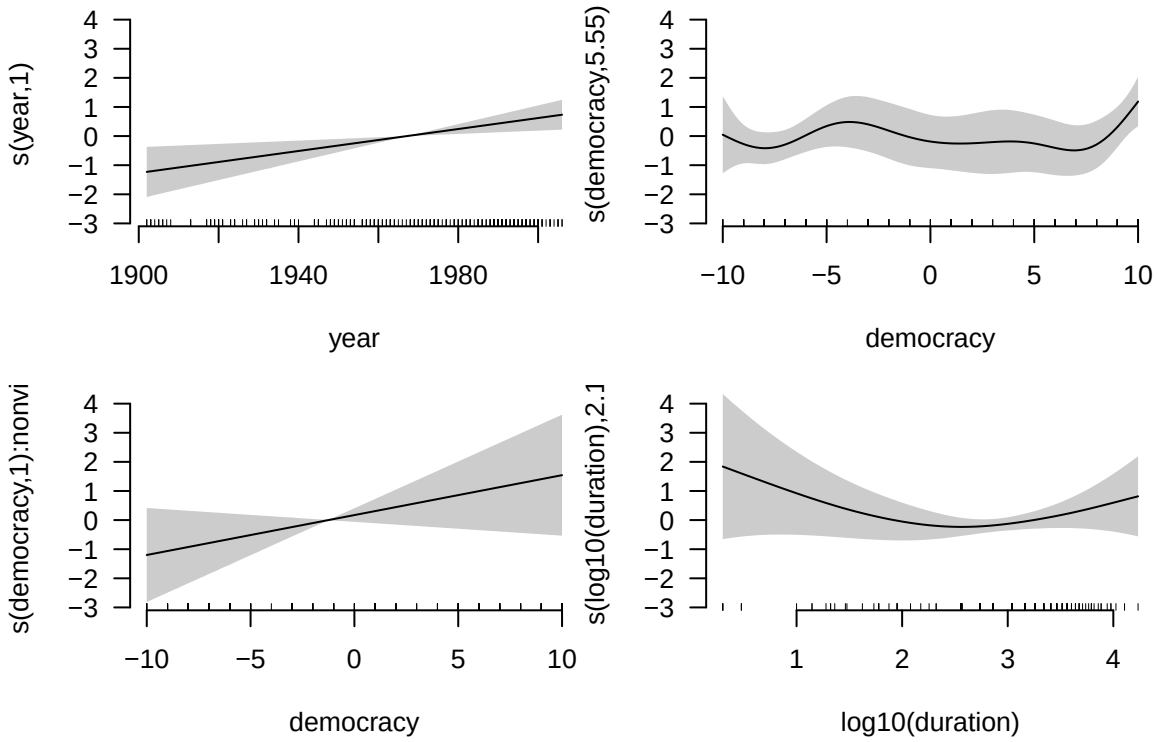

Figure 2: Figure 2: The smooth terms in the GAM, plotted without error bars, but with a common vertical range. The vertical axis is in units of log-odds of success (i.e., on the logit scale), not probability.

lower probability of success for the movement. Finally, and unsurprisingly[5], defection of the security forces increases the movement's chances.

All of the above refers to the point estimates. The other two columns of Table 1 show the 95% confidence limits, which in every single case include 0; some of them (like the coefficients on nonviolence) are remarkably wide. This does not appear to be a bug, but rather a genuine reflection of massive instability in all the coefficients, even while the model as a whole manages to predict quite well.

Figure 2 shows the partitial response functions for our model. The partial response to year increases monotonically over time, tracking the break-up of empires from global military conflicts[6]. The partial response to duration says that very short and very enduring movements are both more likely to succeed than those which last about a year. The high success rate of short movements may be another selection effect — a movement can win quickly, but, once it's mobilized, losing takes times.

Finally, we consider democracy and its interaction with non-violence. Figure 2 shows the two relevant partial response functions: the main effect of democracy, applying to all movements, and the interaction of democracy and non-violence. The former shows a somewhat complicated pattern — movements are relatively more likely to succeed against the most anti-democratic regimes (hereditary monarchies), and then again even more against those which are only moderately anti-democratic (≈ -4: South Korea a few decades ago, Latin American military dictatorships etc.). The odds of success then fall as the country becomes more democratic, only to shoot up in the most democratic countries. The interaction curve, added on for non-violent movements only, is a straight line through the origin, saying that the extra contribution of non-violence is increasingly positive as the country becomes more democratic, and increasingly negative as an anti-democratic country becomes even less democratic. To get the over-all pattern of success for non-violent movements as a function of democracy, we add the two curves (Figure 3); this sum shows that over-all non-violent movements are

[5]But perhaps another selection effect: deserting an already-losing side.

[6]Cf. Fred Halliday, *Revolution and World Politics: The Rise and Fall of the Sixth Great Power* (Durham, North Carolina: Duke University Press, 1999).

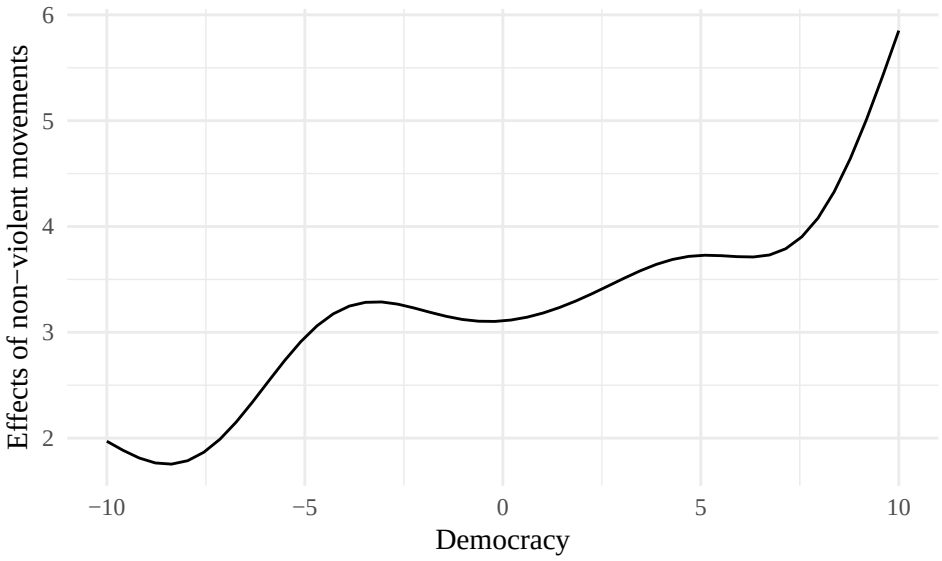

Figure 3: Figure 3: The sum of the two democracy curves

increasingly likely to succeed as the country becomes more democratic, but without much difference over a broad range of values in the middle.

As with the parametric coefficients, however, when we look at the above confidence bands, every single one of them includes zero everywhere at the 95% level (Figure 2).[7] Those for the interaction of democracy and nonviolence, and for duration, have the same huge width we saw for the coefficients related to non-violence. That all the confidence bands include 0 does not, of course, mean that every term should be removed from the model.

2.3 Model Checking

All of the above conclusions presume our model is a good description of the process that generated the data. Thus, our model must be checked before our conclusions can have any credibility. We check the model two ways: can it predict the outcome? are its probabilities calibrated? are the residuals patternless?

Since the outcome variable is a binary category, we can look at how well the model predicts it. As the model gives a *probability* of success, we threshold that probability, predicting success just when $p \geq 0.5$. We can check the error rate of these classifications both in-sample and, through cross-validation, out of sample.

As a baseline, the in-sample classification should have no more error than always predicting the more common class, which is wrong 49 percent of the time. The actual in-sample error rate for our model is 24. Under five-fold cross-validation, both do somewhat worse: the constant prediction has an out-of-sample error rate of 25%, while the GAM is wrong only 20% of the time. We clearly have non-trivial predictive power, even though the model was not built to classify.

2.3.1 Calibration

Figure 4 checks whether the model's probabilities are calibrated, using probability bins 10 percentage points wide. Not only does success get more common as the predicted probability rises, those probabilities are reasonably well-calibrated, with deviations easily explicable by chance.

[7] Note that *really* we should bootstrap the bands around the smooths as well. I did this in the appendix, because it's a bit painful. But do have a look.

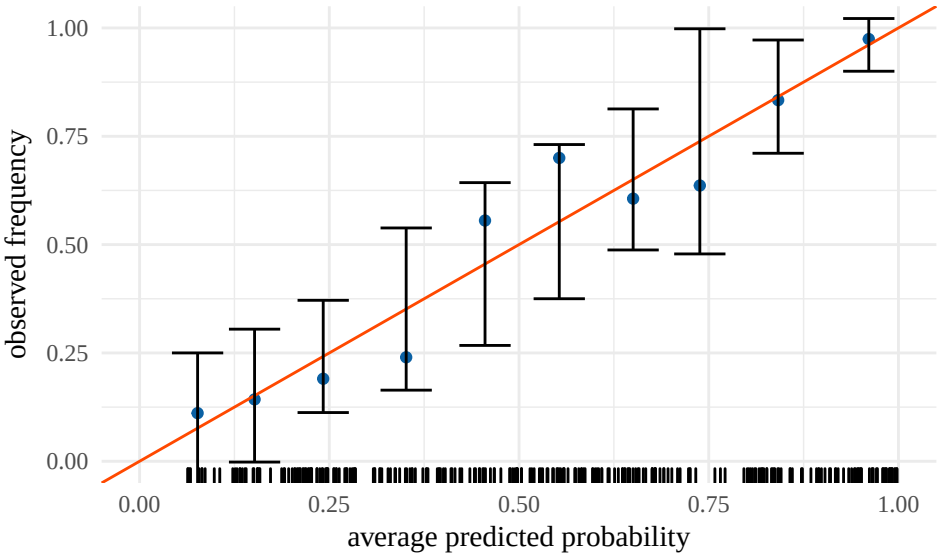

Figure 4: Figure 4: Calibration of probabilities

2.3.2 Residuals

Figure 5 plots the standardized or "Pearson" residuals, $\frac{y_i - p_i}{\sqrt{p_i(1-p_i)}}$ against the predicted probabilities p_i and the three continuous predictors. Each plot also has its own smoothing spline. Ideally, these would be exactly flat; to gauge departures from flatness, we simulated new responses from the fitted model, found their Pearson residuals, and added (faint) splines for them. These indicate little cause for concern.

3 Results and conclusions

Our model lets us answer all our analytical questions, without pre-judging any by excluding relevant interactions. The model can predict out of sample, both in terms of likelihood and of classification, and, while the conditional variance of the residuals is not perfect (last figure), the model otherwise passes our usual checks of calibration, residuals, etc. The point estimates tell a reasonable story: non-violent movements are more successful than violent ones; this advantage grows as the government they confront becomes more democratic, and shrinks in the face of violent repression, though not enough to make violence more likely to succeed. Interventions to aid either side of the dispute predicts a higher probability of success for the anti-government movement, though perhaps for different reasons. The estimates for controls (especially for the year) also make sense.

Sadly for this story, the statistical uncertainty on all of the estimates is huge, swamping every single term (**??** and Figure 6). Further inquiry might try to expand the data set (e.g., it includes no movements from the United States), or find a simpler model which predicts (almost) as well and allows more precise estimates. In the meanwhile, this model has some power to guess the outcomes of confrontations between political movements and governments, but it can't be precise about *why* movements succeed or fail.

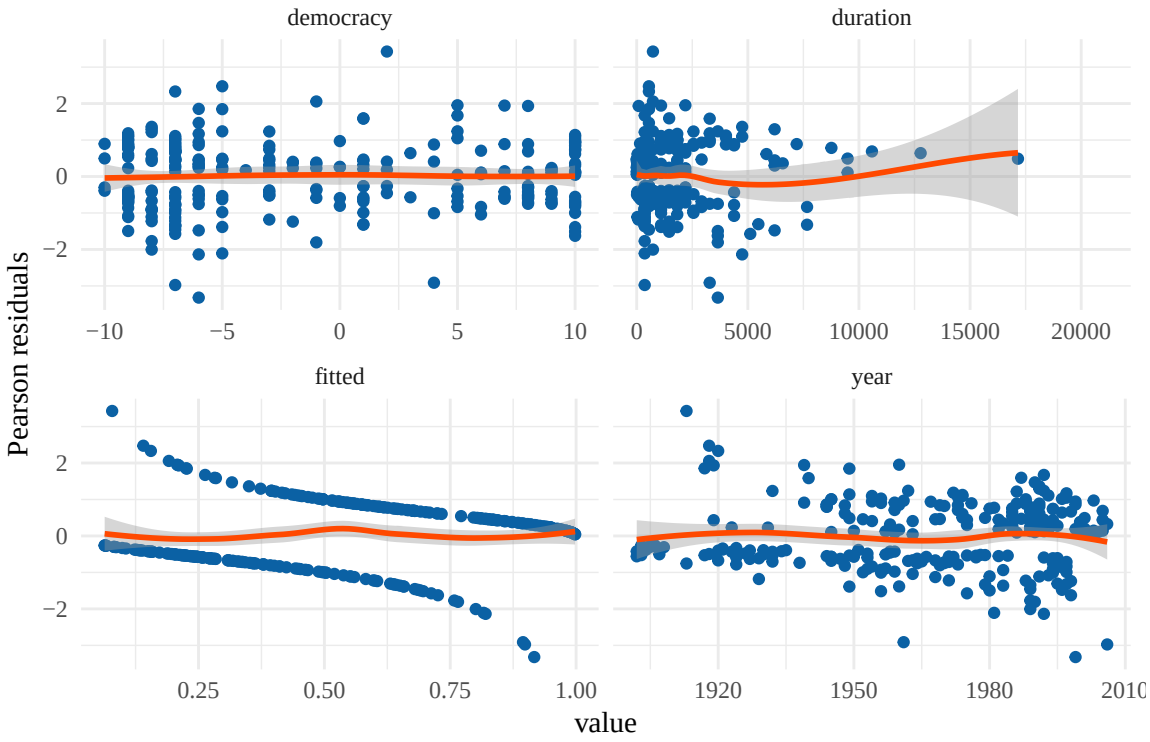

Figure 5: Figure 5: Pearson residuals

4 Alternatives and Notes

4.1 Data Source

This assignment was a re-analysis of the data from

> Maria J. Stephan and Erica Chenowth, "Why Civil Resistance Works; The Strategic Logic of Nonviolent Conflict", *International Security* **33** (2008): 7–44, doi:10.1162/isec.2008.33.1.7

later expanded into a book

> Erica Chenoweth and Maria J. Stephan, *Why Civil Resistance Works; The Strategic Logic of Nonviolent Conflict* (New York: Columbia University Press, 2011)

The paper, and the data set, are available from Prof. Chenoweth's website. For this exercise, we used the version of the data set which accompanied the original paper; expanded and corrected versions are available from Prof. Chenoweth's site, and you should really consult them if you are interested in following this up.

The authors used a multinomial logistic model, but with **only** linear terms in the model. Perhaps for that reason they didn't use `year` directly as a covariate, but rather an indicator for whether the peak year of the movement fell during the Cold War.

Their paper and book contains an account of **why**, under a broad range of circumstances, non-violent civil resistance should be more effective than violent rebellion, supported by cogent theoretical reasoning and detailed case studies. Whether the data allows these theories to be tested with any sort of quantitative precision is another matter.

The analysis which I am presenting here is taken (often without many alterations) from Cosma Shalizi's solutions to his assignment 2 years ago.

4.2 Recoding partial successes

As far as grading goes, having **any** sensible reason to prefer one re-coding over another is enough; it would also be acceptable to try both re-codings and see which one, in some sense, worked better. If you examine the R embedded in the solutions, you'll see that I created new columns for both a loose standard of success (partial success is still success), and a strict one (partial success is total failure). I then added yet a third column (`target`), which copied the loose standard — but all later code refers to that third column, so it would be easy to switch.

An additional, albeit minor, reason to recode partial successes with successes rather than with failures is that this makes the data very nearly balanced between the two classes (166 full-or-partial successes to 157 failures), and an even balance between classes can lead to more stable estimates of classifiers.

An alternative to doing any sort of recoding would be to use a model which can accommodate a three-level categorical response. For logistic regression with a linear predictor, this can be done with the `multinom` function in the `MASS` package, but for additive models, we'd have to use the `VGAM` package, or something like it. `VGAM` is vastly less user-friendly than `mgcv`, and interactions of continuous and discrete variables in particular require hand-coding — see the source code for these solutions for a `VGAM` fit.

4.3 Categorical interactions

Rather than using `nonviol*viol.repress`, and getting main effects and an interaction, we could have made the first expression on the right-hand side of the formula `factor(nonviol)*factor(viol.repress)`. This would have estimated an effect for every combination of the two factors, which would convey the same information, but need us to do some subtraction to see the difference between applying violent repression to a violent movement, and applying it to a nonviolent movement.

Similarly, if instead of `s(democracy) + s(democracy,by=nonviol)` we had written `s(democracy,by=factor(nonviol))`, R would have estimated a different smooth curve for `democracy` at each level of `nonviol`, which would have conveyed the same information. (See `help(gam.models)` for more on the `by` option to smoothers.)

In both cases, the model formulation in the main report was chosen to highlight the contrastive quantities or functions that we were asked to estimate.

4.4 Transforming predictors before smoothing

If a variable is very (say) right-skewed and we don't have all that many measurements of it, we'll have lots of closely-spaced observations at small values, and a few widely-spaced observations at large values. Trying to find **one** smoothing bandwidth which works well for both parts of the data would be hard. In principle, with enough observations this doesn't matter, and splines are somewhat more robust to this effect than are kernels, but it can still be a difficulty. Transforming the predictor before smoothing can help reveal structure which would otherwise be smothered.

4.5 Using the built-in cross-validation in `gam()`

Alternatively, we could use the `cv.gam` function from the earlier lecture. There will be some differences, since what `gam` computes is really the "generalized" cross-validation score, a fast approximation to leave-one-out CV that avoids having to refit the model n times (see Section 3.4.3 of the notes). The circulated code, on the other hand, is for k-fold CV, defaulting to $k = 5$.

4.6 Bootstrapping the smooths

Really, we shouldn't trust the confidence bands that `gam` plots for us. So we could bootstrap those as well. This is a little bit painful however, so you aren't expected to do it. I've done it below though. This is the main reason that I kept around the entire fitted model for each bootstrap replication: I can re-use the fits here and just calculate summaries rather than repeating the entire exercise.

4.7 Choice of bootstrap

Because resampling of residuals works poorly when the response variable is binary, we basically have a choice of simulating from the fitted model, or resampling cases. So as not to rest too strongly on the correctness of the model, I picked resampling of cases. The confidence intervals derived from simulating the model are somewhat narrower than those from resampling, but not drastically so — compare Tables 1 and 2, and the source code for the alternative.

Table 2: Table 2: Point estimates and 95% confidence intervals for the model's parametric terms based on the model-based bootstrap replications.

	lo	est	hi
(Intercept)	-4.809	-0.362	3.235
nonviol	-521.208	3.123	214.507
viol.repress	-2.709	-0.650	1.748
sanctions	-2.189	-0.045	1.518
aid	-1.093	0.098	1.166
support	-1.071	0.716	2.192
defect	-0.160	0.913	2.425
nonviol:viol.repress	-6.293	-0.574	109.674

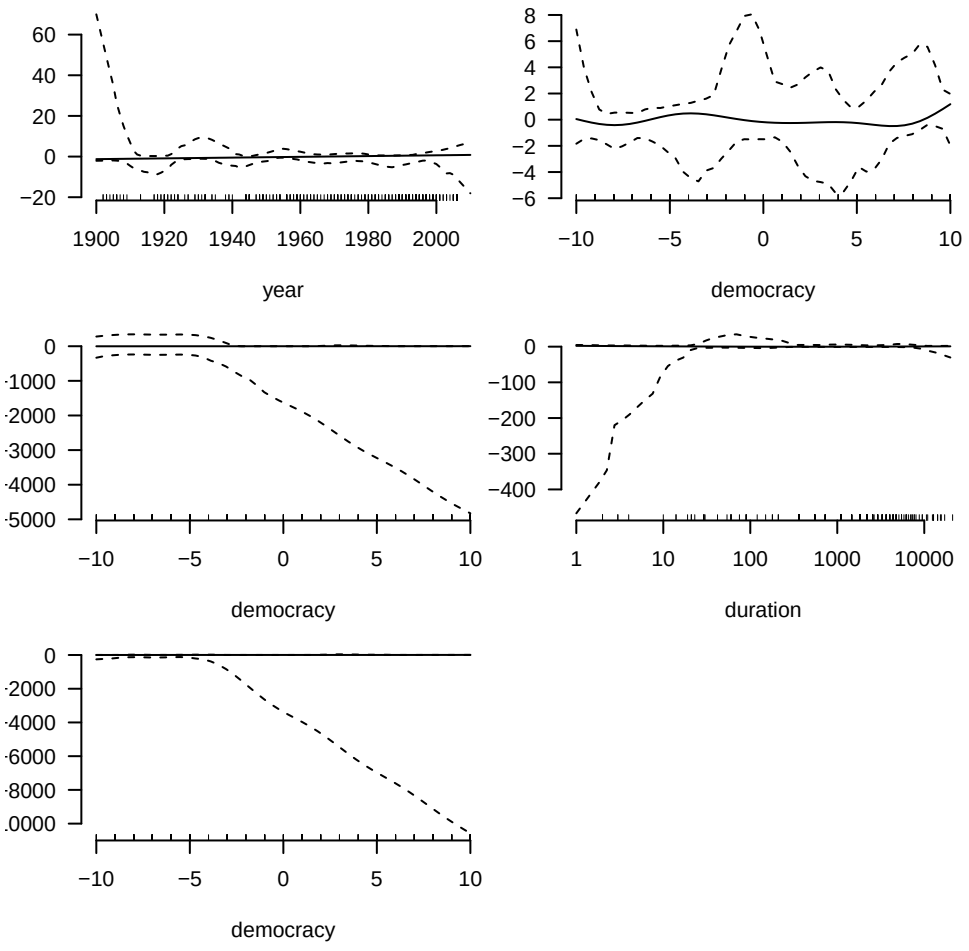

Figure 6: Figure A1: Bootstrap confidence bands. The bottom figure is for the sum of democracy terms.

4.8 Details of bootstrapping

While the resampling scheme is the same for both the parametric and non-parametric parts of the model, and we want confidence intervals for both, we need slightly different pieces of code for the coefficients on indicators and for the partial response functions. When dealing with the coefficients, we treat them like any other set of parameters, e.g., the coefficients in a linear regression. We can extract them from a fitted model with `summary()$p.coef`, or by taking the appropriately-named entries in the vector returned by `coefficients`. On the other hand, the partial response functions are **functions**, curves, and we need to get confidence bands for them the way we got confidence bands for spline and kernel regressions, using `predict`.

Because fitting a GAM is comparatively slow, it would be better not to have to do a completely separate set of resamplings and re-estimations for every term in the model. The code thus modifies the example code from the notes and homework solutions to let different terms in the model share the same set of bootstrap samples. This is, of course, beyond what you were expected to do.

4.9 Model checking: Classification and calibration

The techniques, and the solution code, are lifted from chapter 11 and 12 lectures.

It would have been ever better to check calibration in a cross-validatory way, but that would go beyond what was required here.

4.10 Model checking: Residuals

With a binary response, the raw residuals must always be either $-p_i$ or $1 - p_i$, and the Pearson residuals are similarly constrained. The highly patterned look of the top-left plots in Figure 5 is thus inevitable, however strange it may appear after working with ordinary regressions.

Running a spline through the squared residuals goes back to the discussion of heteroskedasticity, and variance function estimation, in Chapter 7 of the notes.

4.11 Model checking: Comparing to another model

The assignment called for using a GAM. In the notes and homework, we looked at using a GAM to test whether a GLM is well-specified; we can't really turn that around. Even if we had fit a GLM here, and then rejected it, that would just have said "the GAM isn't as bad as the GLM", not "the GAM is a good model". We could try embedding the GAM in a more general model, such as a kernel regression, or conditional kernel density estimate, or even just a GAM with a lot more interactions, and tested it along similar lines, however.

4.12 Model checking: What to do with a bad model?

If any of the checks had detected severe problems with the model, the ideal course would have been to look into **why** the model was bad, and come up with a new one which fixed those problems. The **report** could have been written around the fixed-up model, with at most a brief mention of discarded predecessors. If we couldn't find a tolerable model, then we'd have to say why even our best model wasn't acceptable; some of the ways we'd tried and failed to fix it; and how the model's problems might impair our ability to use its estimates to answer the analytical questions.

4.13 Variable selection

The assignment did **not** call for variable selection or an extensive search over models. It was nonetheless legitimate to do such a search, provided one did so intelligently. Deleting variables from a model because their coefficients or partial response functions are insignificant is **not** a reliable method (as you learned in 431). With the model fit here, even though it has substantial predictive ability, **not one single term** is significant at the 5% level. Cross-validation would work much better.

It should hardly need repeating that "this variable's coefficient or response function was not statistically significant" does **not** mean the same thing as "this variable does not matter for predicting the response". A very wide confidence interval which overlaps with 0 means we **don't know** whether or how much the variable matters. It is only when the confidence interval contains zero **and** is very small that we can be pretty sure the variable is unimportant. (For that matter, a tiny confidence interval which does not include zero, but is centered around a minute value, is also good reason to think that a variable doesn't matter very much.)