Model building: Case study

Patrick Breheny

April 2

Introduction

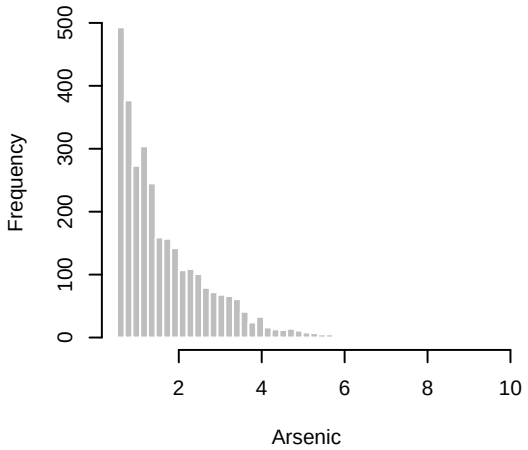
- Today we continue looking at the well-switching data from the previous lecture, but carefully consider the model-building process
- There are no explicit rules to follow when building a model – other than, perhaps, don't build a model by blindly following rules – and different people could look at this same data and build different models, but hopefully this journey will be productive and instructive

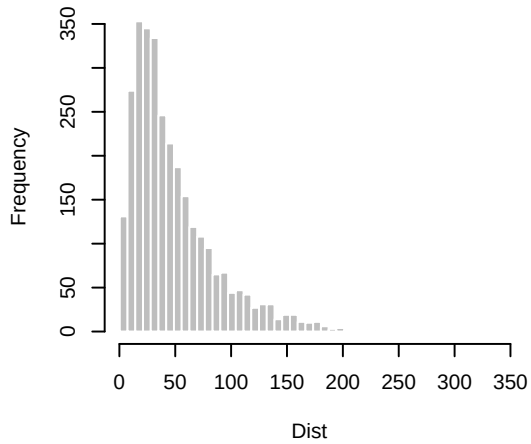
Some general guidelines

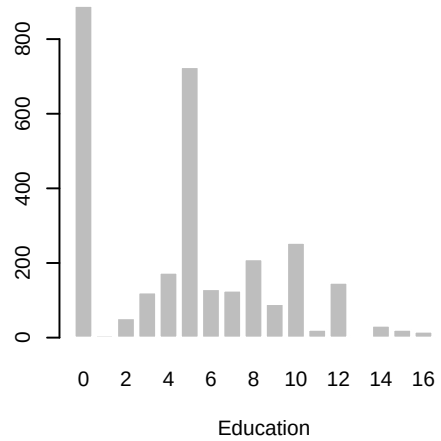
- Generally, I use AIC to judge the quality of a GLM's fit to the data, weighted roughly according to the scale discussed in the previous lecture
- The other primary consideration I use is whether a variable and the sign of its coefficient makes sense (*i.e.*, whether, before seeing the data, I thought it would have been important and act in the appropriate direction)
- So for example, if a variable leaves the AIC essentially unchanged, whether to include it or not depends on external considerations: if it makes sense and has a believable direction, leave it in; if not, we might as well remove it

Descriptive statistics

One should always start by looking at descriptive statistics and getting a sense of each variable:



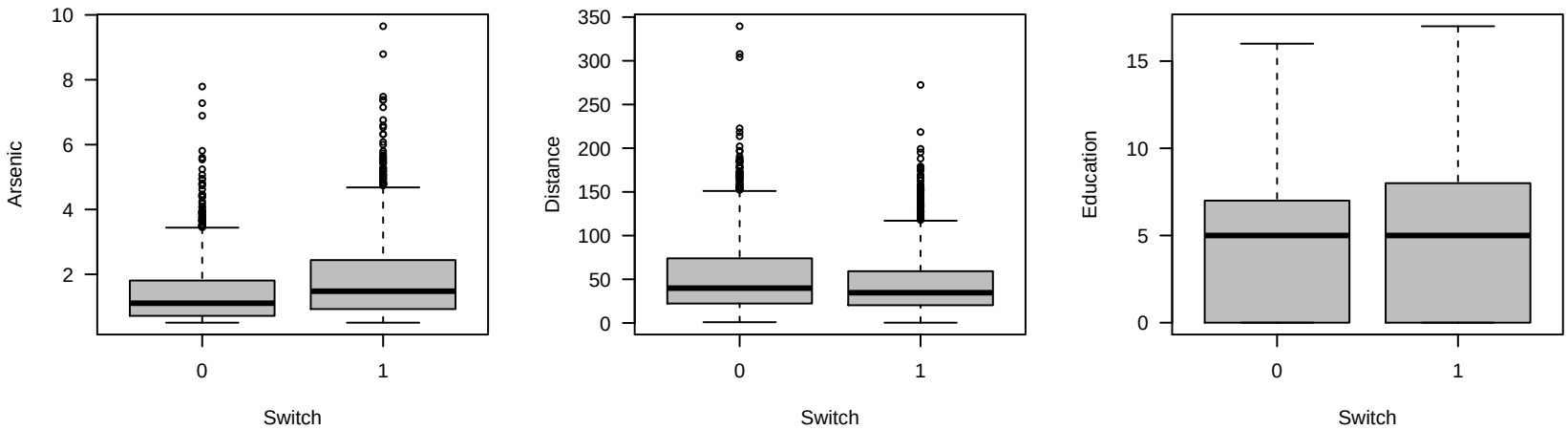




Switch: 58%; Community: 42%

Descriptive statistics

Next, it's useful to compare each variable to the outcome:



Community=Yes: 55% switched; Community=No: 59% switched

Additive model

- A reasonable place to start is the additive model:

$$\text{logit}(\pi_i) = \beta_0 + \beta_1 \texttt{Arsenic} + \beta_2 \texttt{Dist} + \beta_3 \texttt{Educ} + \beta_4 \texttt{Community}$$

- As always, however, the regression coefficients have little meaning unless we consider the scale of each variable
- For example, distance is measured in meters, so all β_2 tells us is how much the log-odds of well-switching change if we compare a well 90 meters away vs. a well 91 meters away; obviously this will be an inconsequential change
- On the other hand, if distance were measured in kilometers, β_2 would be enormous

Standardization

- A reasonable rule is to set changes to be roughly two standard deviations (the reason for the factor of 2 is so that they may be compared to 0-1 variables, which have standard deviations of $\sqrt{p(1-p)} \approx 0.5$)
- This amounts to a change in distance of $\approx$ 100 meters, a change in arsenic levels of $\approx 2\mu g/L$, and a change in education of 8 years
- While we're at it, it would also be helpful to subtract off the mean of each of these variables; otherwise, the intercept, β_0, is essentially meaningless, as it corresponds to an arsenic level of 0

Coefficients

Coefficients and standard errors for standardized variables:

	β	SE
Intercept	0.39	0.05
Arsenic	0.93	0.08
Community	-0.12	0.08
Distance	-0.90	0.10
Education	0.34	0.08

Coefficients

When creating a tables or graphs, it is always a good idea to consider a meaningful order; the following table, with variables sorted in terms of importance, is superior to the previous one:

	β	SE
Arsenic	0.93	0.08
Distance	-0.90	0.10
Education	0.34	0.08
Community	-0.12	0.08

Adjusted versus unadjusted odds ratios

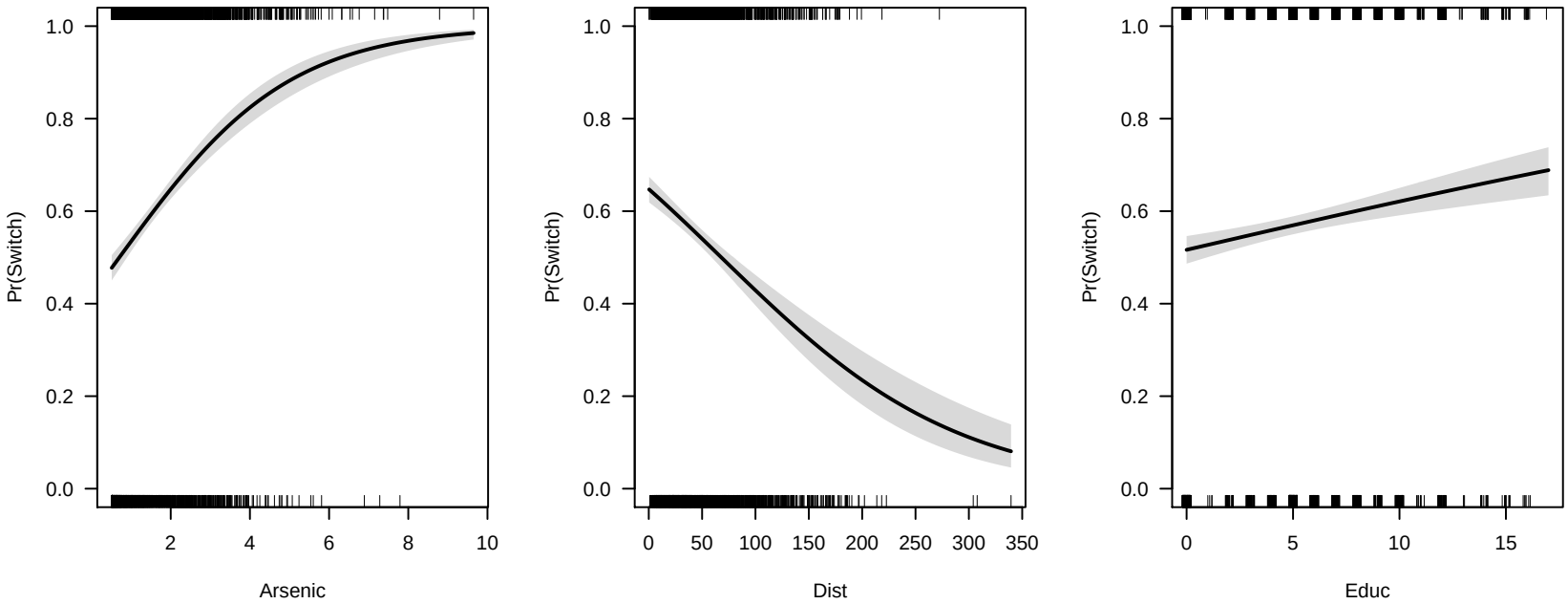
It is also often useful to compare adjusted to unadjusted odds ratios:

	Unadjusted OR	Adjusted OR
Arsenic	2.13	2.54
Distance	0.54	0.41
Education	1.36	1.40
Community	0.86	0.88

Do we really need `Community`?

- From here on, I'll discard Community from the model: there is no statistical justification for it (not significant, doesn't improve AIC), and I find its negative sign difficult to explain/interpret
- In short, it seems to add nothing but clutter to out model
- At this point, it seems reasonable to look at the model using the kind of effect plots you get from `visreg` or `PLOTS=EFFECT` in `PROC LOGISTIC`

What our model looks like



Investigating assumptions

- The main assumptions that our model at this point is making are:
  - There are no other (unmeasured) covariates that affect the probability of switching
  - The effect of each variable is linear
  - No interactions
- The first assumption is almost certainly not true, but there's nothing we can do about it after the data has been collected
- The second and third assumptions, however, we can investigate

Transformations to consider

- Given the highly skewed distributions for distance and arsenic, a log transformation would seem reasonable
- For education, the bar plot we looked at earlier would suggest the following education categories:
  - None: `Educ`$=0$
  - Low: `Educ` $\in$ [1-5]
  - Medium: `Educ` $\in$ [6-10]
  - High: `Educ` ≥ 11
- Unlike linear regression, it is difficult (at least in my opinion) to sense an appropriate transformation in logistic regression by looking at residuals, but we can try fitting these models and seeing what happens with the AIC

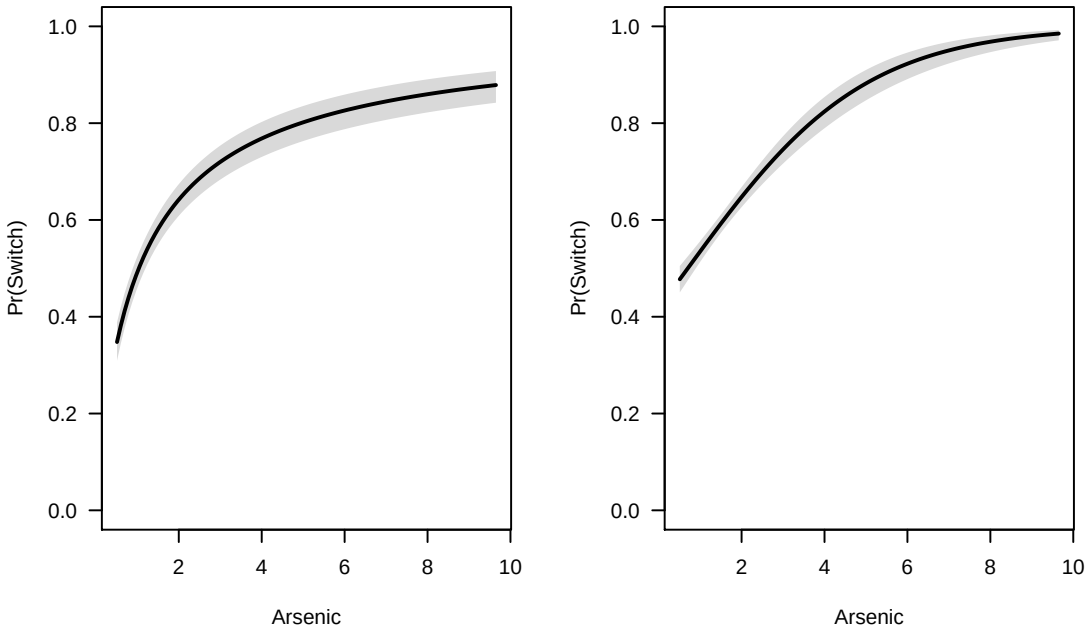
Transformations

- The results:

Transformation	AIC
None	3918
log(Arsenic)	3886
log(Distance)	3951
Ed. Categories	3910

- This is fairly convincing evidence that we should adopt the log(Arsenic) and education category transformations, but stick with a linear effect of distance (this model has an AIC of 3878)

What our new model looks like



Remarks

- Note the effect of the log transformation: compared with the earlier model, the probability of switching changes more rapidly for low arsenic levels, but is flatter for high arsenic levels
- This implies that people react more strongly to the difference between arsenic levels of 1 and 2 than they do between arsenic levels of 5 and 6 (which seems perfectly reasonable)
- For education, our new model indicates that there is little difference between individuals with none or little education, but that higher levels of education are associated with increased probability of switching (again, this seems reasonable)

Interactions

- Finally, let's consider interactions
- In this example, where we only have three variables, might as well just consider all three two-way interactions
- If we had more terms, we would have to narrow our focus; typically this would involve:
  - Concentrating on interactions involving the most important main effects
  - Concentrating on interactions involving the primary treatment or exposure (if there is one)

Interaction results

- The results:

Interaction	AIC
None	3878
Arsenic $\times$ Distance	3879
Education $\times$ Distance	3860
Arsenic $\times$ Education	3878

- So it seems that we have pretty strong evidence that the effect of distance is not the same for all levels of education; the other interactions don't seem to add much

Coefficients for interaction model

Arsenic and distance have been centered and a one-unit change in each corresponds approximately to a 2-SD difference

	β	SE
Intercept	0.26	0.07
log(Arsenic)	0.90	0.07
Distance	-1.17	0.20
Ed: 1-5	-0.12	0.10
Ed: 6-10	0.25	0.10
Ed: 11+	0.71	0.17
DxE: 1-5	-0.25	0.26
DxE: 6-10	0.63	0.27
DxE: 11+	1.54	0.48

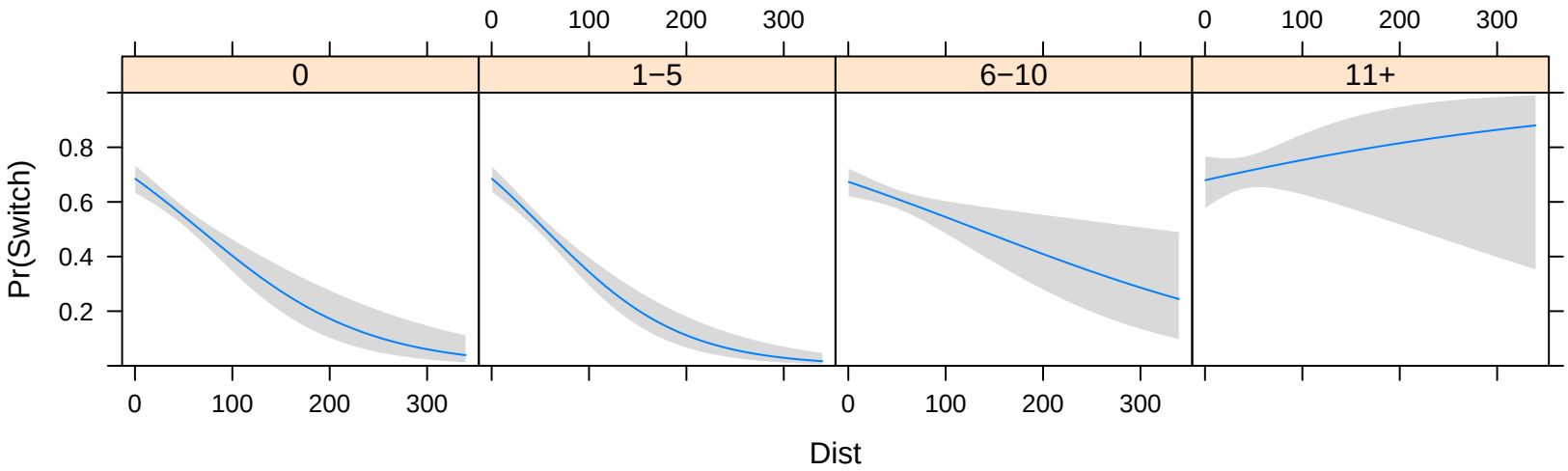
Remarks

- Thus, distance has a large effect for low-education households, a moderate effect for medium-education households, and seemingly no effect for high-education households
- Presumably this is due not so much to education per se, but because education serves as a marker for socioeconomic status
- Thus, while individuals with low education are presumably poorer and have to walk to the next well, better-educated individuals are more likely to be wealthier and able to afford other, less onerous means of transportation

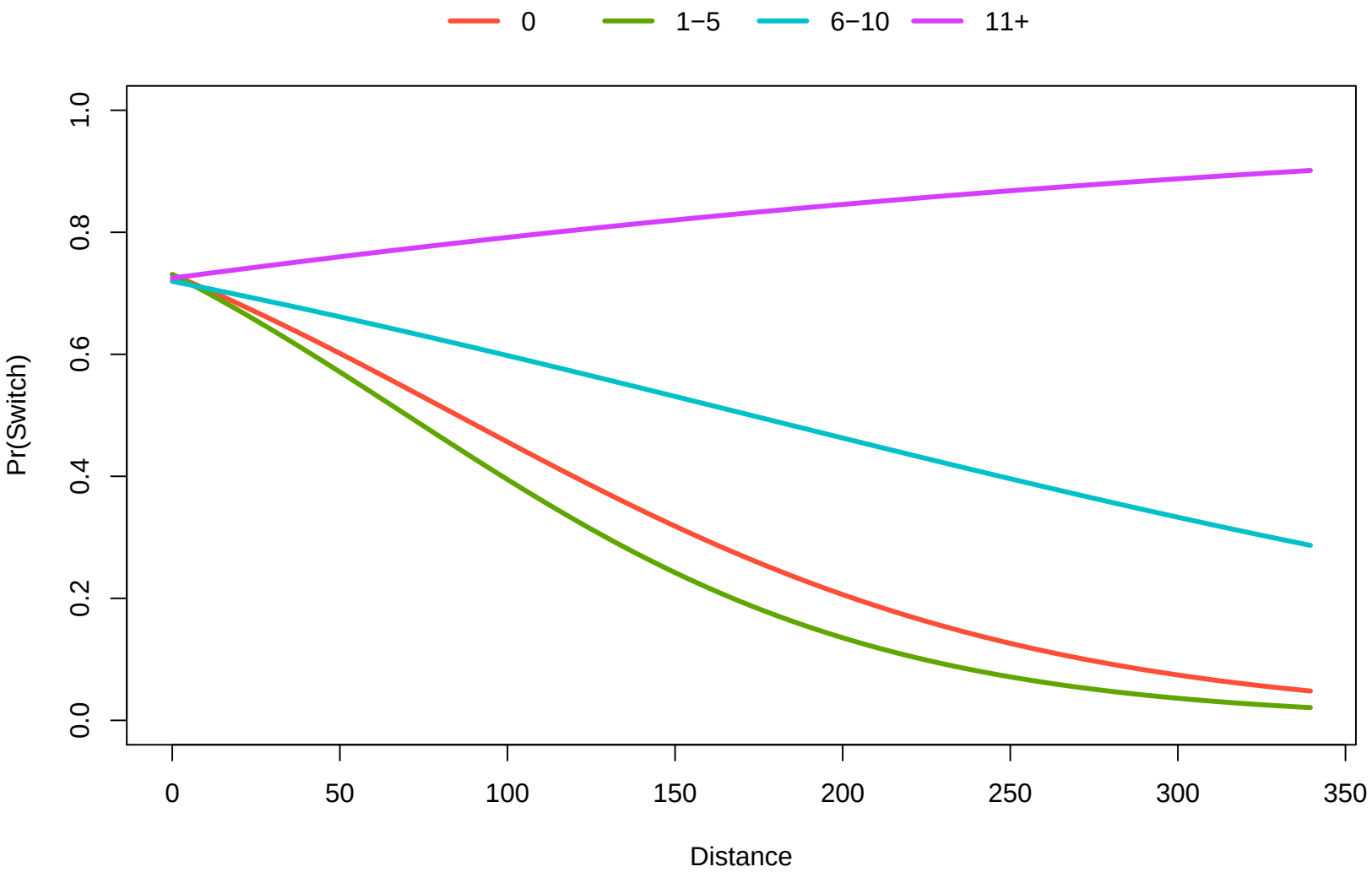
Communicating models with interactions

- When communicating the results of models with interactions, reporting regression coefficients directly is rarely a good idea – only a small fraction of your audience will usually be able to discern their meaning
- Instead, it is better to report separate odds ratios for each level of the interacting variable, or to provide plots indicating the relationship at each level

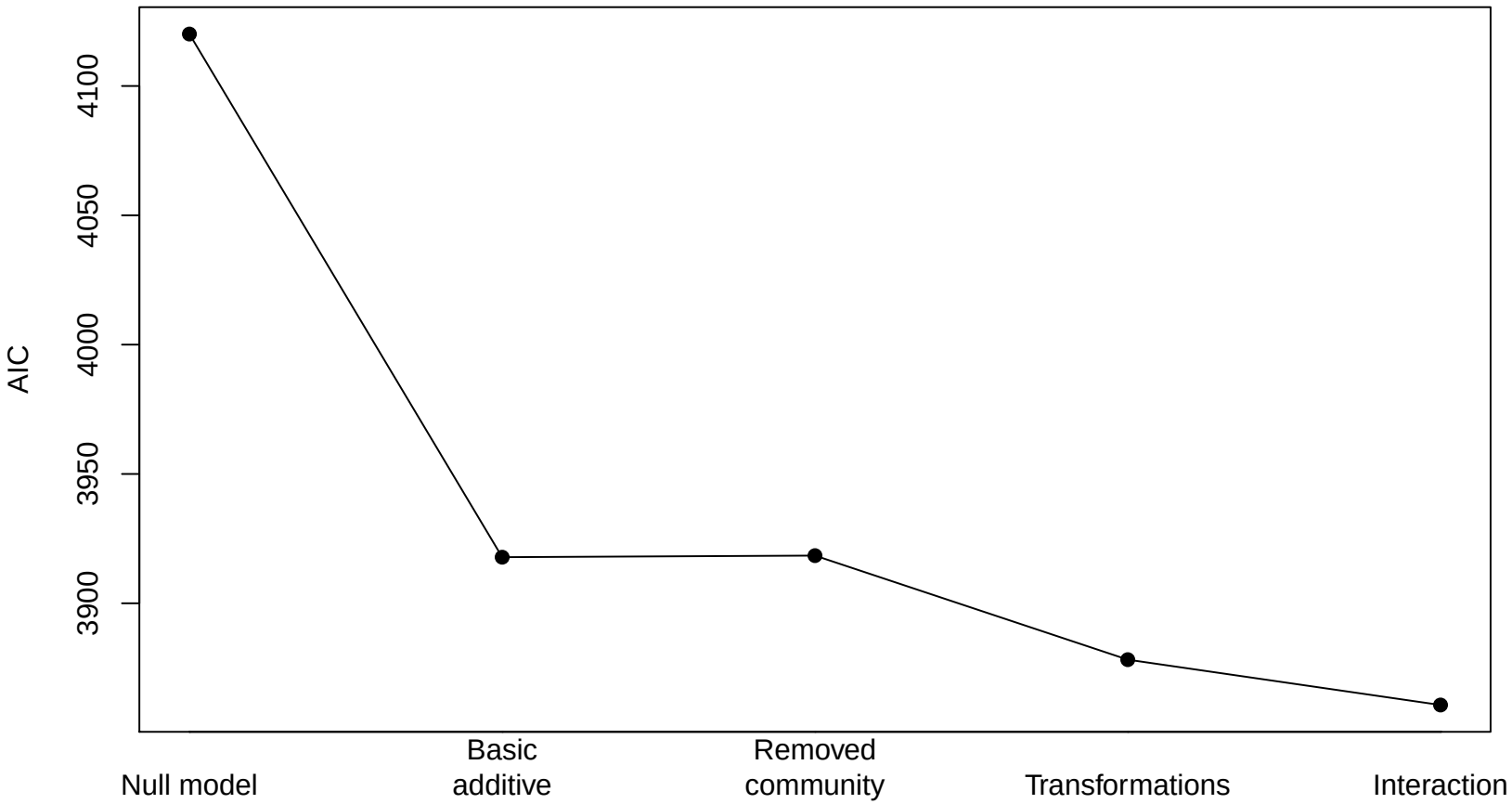
Plot #1



Plot #2


0
1−5
6−10
11+
Pr(Switch)
0.0
0.2
0.4
0.6
0.8
1.0
0
50
100
150
200
250
300
350
Distance

The March of Progress: AIC



The March of Progress: BIC