Model selection III

Patrick Breheny

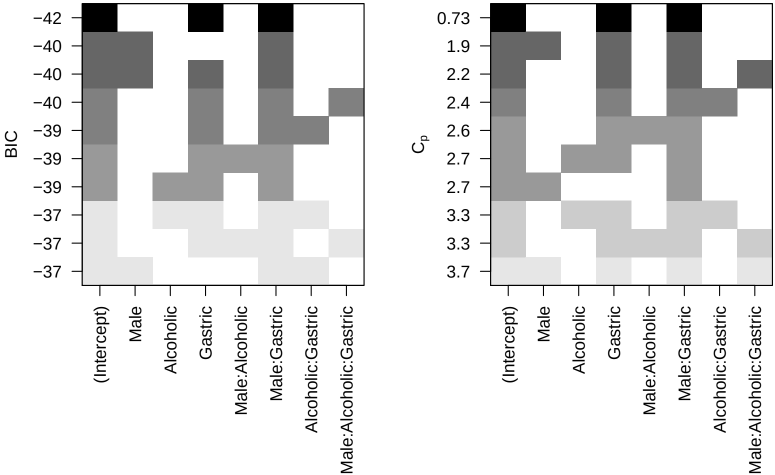
February 24

Introduction

- The picture of model selection that we got from our exploration of the Swiss fertility data set was pretty straightforward
- In today's lecture, we will explore issues of model selection in more complicated scenarios involving interactions and polynomial terms
- We will conclude by making some general remarks about the benefits and pitfalls of automated model selection

Alcohol data

We will begin by investigating our alcohol data set:

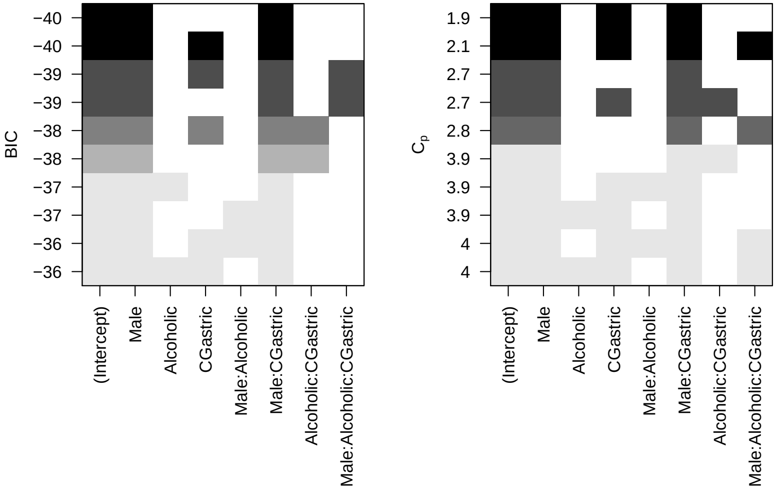


Concerns about the top-ranked model

- Thus, our top-ranked model has an interaction between sex and `Gastric` and a main effect for `Gastric`, but no main effect for `Male`
- Recall, however, that the meaning of main effects is a bit slippery when interactions are present
- Having an explanatory variable in the model to represent the difference between males and females at a `Gastric` level of zero may not be important, but this doesn't mean the variable isn't important at other levels

Alcohol data

For example, if we center `Gastric`:



Concerns about the top-ranked model

- Having our "best model" depend on something as arbitrary as whether or not we center one of the variables does not seem particularly logical
- For this reason, we often prefer our model to have all the main effects that correspond to an interaction, even if they worsen our model selection criterion
- This is especially true in this case, since the model with both main effects is nearly as good as the model with the absolute lowest BIC/C_p

Always include main effects?

- Should you always include main effects when an interaction is present?
- Not necessarily – dominant and recessive inheritance in genetics is a good example of a well-establish scientific phenomenon in which an interaction is present without a main effect
- However, the scientific merits of such a main-effect-free model should be carefully considered, as it is easy to naively propose an absurd model by failing to include main effects

Cotinine study

- A related notion is that of collapsing across the categories of a categorical variable
- For example, I am involved in a study of the correspondence between parents' self-reported description of their child's exposure to second-hand smoke (None/Intermittent/Daily) and a laboratory measurement of cotinine, a biomarker used to measure exposure to tobacco smoke
- One issue that came up in the study was that very few parents responded "Intermittent", leading to high variance in estimates concerning this group
- Perhaps we can combine the intermittent and daily groups

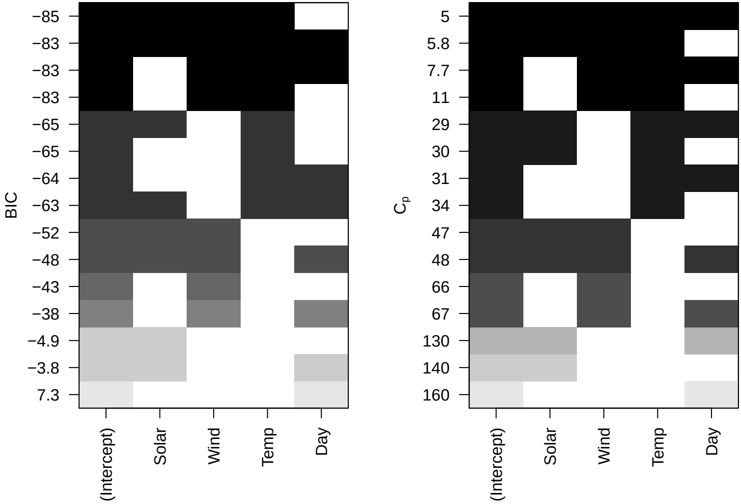
Cotinine study (cont'd)

Combining. . .	PRESS
. . . None and Intermittent	171.07
No combining	175.08
. . . Intermittent and Daily	183.40
. . . All three	221.09
. . . None and Daily	223.79

- These findings indicate that the Intermittent group seems more similar to the None group, and perhaps they should be combined instead
- In the actual study, however, combining these two groups seemed strange from a clinical perspective, and no combining of groups took place

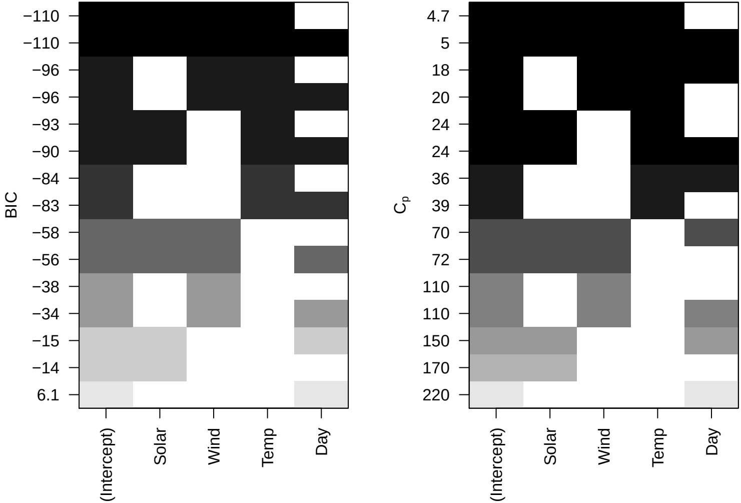
Ozone data

Let us now turn our attention to the ozone data set:



Ozone data

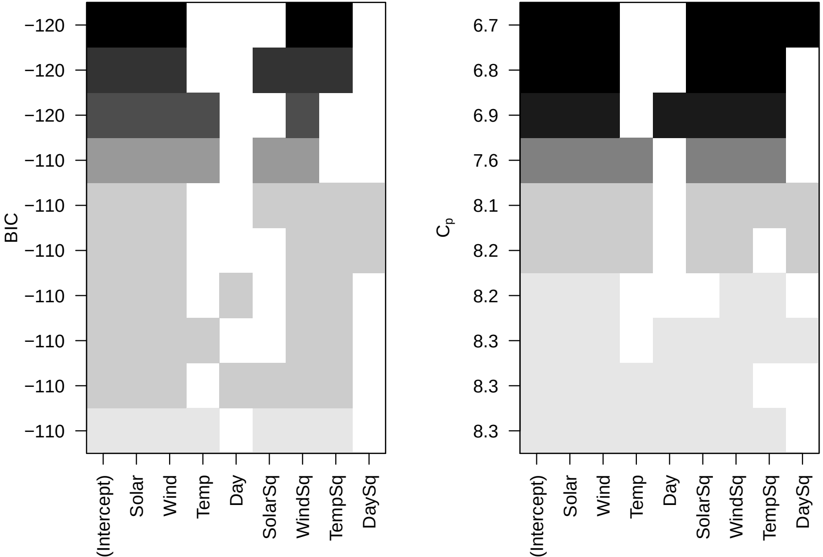
The previous slide was on the original scale; using $\sqrt[5]{\texttt{Ozone}}$,



Quadratic effects?

- Of course, these are not the only possible models we could fit to the ozone data
- In particular, why assume that all the effects are linear?
- Perhaps some of the effects are quadratic
- Let's consider doubling our set of explanatory variables by considering the square of each variable as well

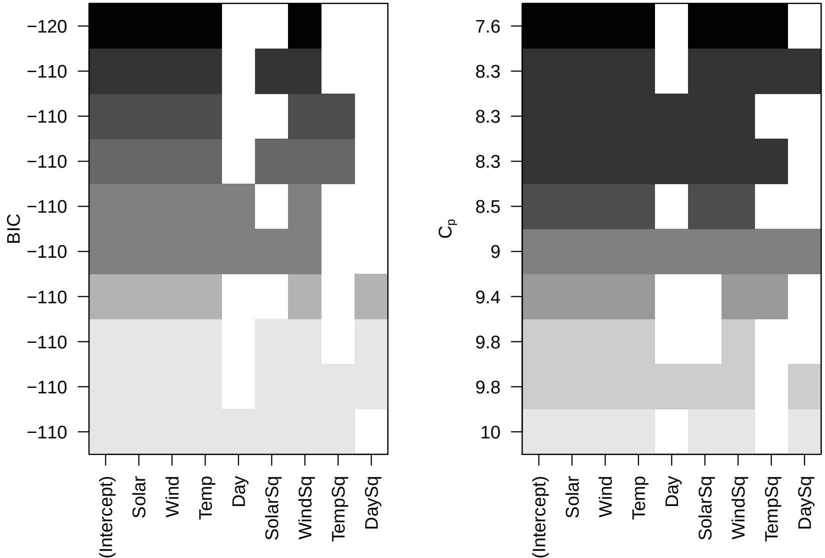
All squared terms



Caution

- Note that we have several terms in the top models with a quadratic effect but no linear effect
- Before you think too deeply about what the scientific reason for all these quadratic effects might be, it needs to be pointed out that quadratic effect models which lack linear effects are not invariant to changes of scale either
- Without a linear term, a quadratic polynomial for X_j assumes that the vertex of its parabolic effect is located at $X_j = 0$; if we re-center X_j, the family of models under consideration is entirely different

All squared terms, standardized predictors



Caveats with polynomial regression

- Like “always include main effect terms when you have an interaction”, “always include lower-order polynomials when you have a higher-order term” is a useful rule of thumb
- Again, there are exceptions, but it is important to realize that if your approach is not invariant to changes of scale, you had better make sure that you’ve given the scale some serious thought!

Caveats with polynomial regression (cont'd)

- A further caveat is that polynomial regression tends to be quite unreliable at the boundaries of the data
- They are even less reliable past the boundaries of the data
- This above problem is referred to as the problem of *extrapolation*; it is generally extremely questionable to assert that whatever trend you have seen in the data will continue past the boundaries of the observed sample

Smooth regression

- There are ways of improving upon polynomial regression by setting up orthogonal terms that represent “pure” linear/quadratic/cubic trends
- Although of course, there is no reason that trends have to be polynomial in the first place
- In general, what we really want is estimate some function f, where

  $$\mathrm{E}(Y) = f(x)$$

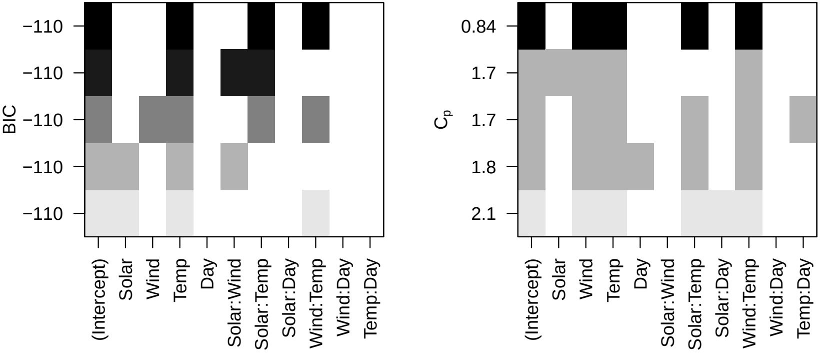
  and f is reasonably smooth
- As you might imagine, this is a complicated topic, and extends beyond the scope of a first-year regression course, but be aware that a number of sophisticated methods have been proposed to address this topic

Naive interactions

- We're still not done, however – we can look at interactions!
- Again, we need to be careful, as

```
regs <- regsubsets(Ozone^.2~.*.,data=ozone)
```

produces



Naive interactions (cont'd)

- So we end up with a model containing a main effect for `Temp`, a `SolarTemp` interaction, and a `WindTemp` interaction, but no main effects for either `Solar` or `Wind`
- What does it mean?
- Well, a `Wind` main effect in this model would be the effect of wind for a temperature of 0 (the minimum value for `Temp` in the data set was 57)
- In other words, this term is meaningless, so it shouldn't come as a big surprise that it wasn't selected by the automatic procedure

More careful handling of interactions

- There are two ways to resolve this issue
- One is to force the main effects into the model:

```
regs <- regsubsets(Ozone^.2~.*.,data=ozone,force.in=1:3)
```

- SAS provides a similar option, `INCLUDE`, to its model statement
- The other is to, once again, standardize the predictors prior to model selection
- In this case, both approaches result in the same model:

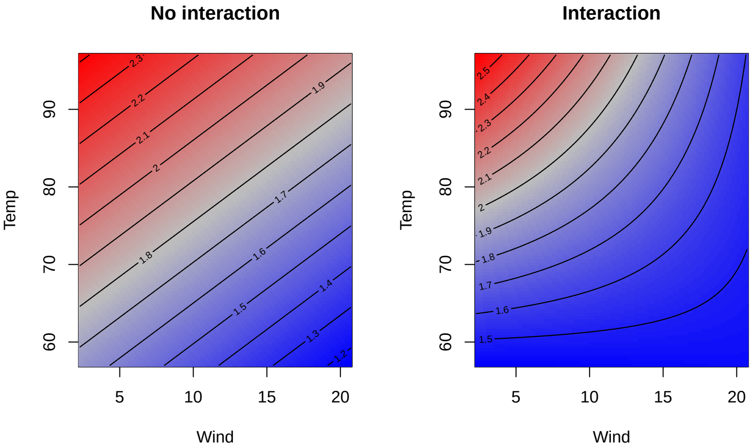
$$\mathrm{E}(\sqrt[5]{\texttt{Ozone}}) = \beta_0 + \beta_1\texttt{Solar} + \beta_2\texttt{Wind} + \beta_3\texttt{Temp} + \beta_4\texttt{WindTemp}$$

Interactions between continuous variables

- What does an interaction between two continuous variables mean?
- It means that the effect of wind depends on temperature, and vice versa
- For example, the effect of increasing wind by 1 mph on a 70 degree day is -0.017; on a 90 degree day, the effect is -0.044
- In other words, the effect of wind on ozone is about 2.5 times greater on warm days than mild days

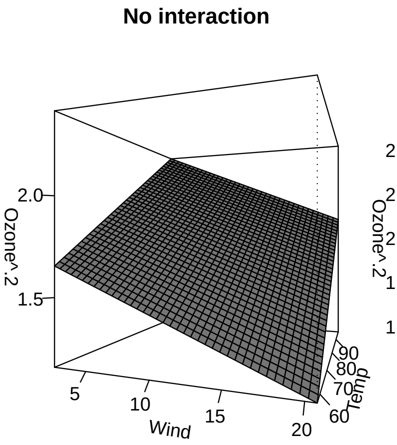
Combined effect of wind and temperature

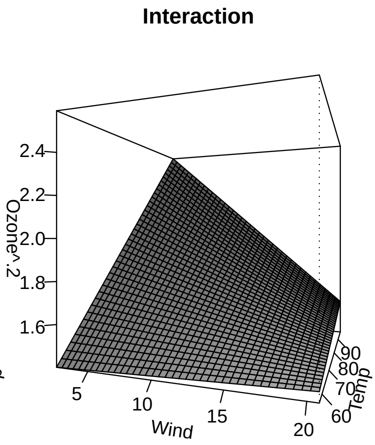
A plot of the combined effect of wind and temperature (while solar radiation remains the same) may make this more clear:



Combined effect of wind and temperature (cont'd)

Perspective plots of the same thing:



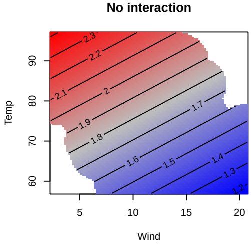


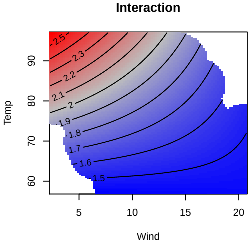
Quadratic effects and interactions

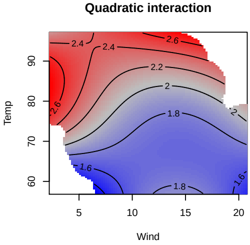
- A final question: given that we have evidence of an interaction between wind and temperature and evidence of nonlinear effects, should we consider a model with both?
- Let's consider one final, rather complicated model:

$$\begin{aligned}\mathrm{E}(\sqrt[5]{\texttt{Ozone}}) =& \beta_0 + \beta_1\texttt{Solar} + \beta_2\texttt{Wind} + \beta_3\texttt{Temp} + \beta_4\texttt{Wind}^2\\ &+ \beta_5\texttt{Temp}^2 + \beta_6\texttt{WindTemp} + \beta_7\texttt{Wind}^2\texttt{Temp}\\ &+ \beta_8\texttt{Temp}^2\texttt{Wind} + \beta_9\texttt{Temp}^2\texttt{Wind}^2\end{aligned}$$

Quadratic effects and interactions (cont'd)







Summary

To summarize our prolonged efforts in modeling ozone concentration:

Model	R^2_{adj}
Original	0.595
$\sqrt[5]{\texttt{Ozone}}$ Transform	0.672
Interaction	0.692
Quadratic terms	0.707
Quadratic interaction	0.751

Pros of automated model selection

- There are certainly advantages to automated model selection, in that it allows the statistician to quickly survey a large number of potential models
- For example, it is doubtful that we would have considered an ozone model with a quadratic interaction between wind and temperature without automated selection (or at least, it would have taken us a very long time to arrive at this model)

Cons of model selection

- However, there are also a great number of cons in automated model selection
- In fact, automated model selection violates every principle of statistical estimation and hypothesis testing:
  - Estimates of R^2 and $\mathrm{R}^2_{\mathrm{adj}}$ are biased high
  - Test statistics no longer follow t/F distributions – all our derivations assumed that the model and hypotheses were prespecified
  - Standard errors are biased low, and confidence intervals falsely narrow
  - p-values are falsely small
  - Regression coefficients are biased away from zero

Simulation

- For example, let's simulate $Y_i, X_{i1}, \ldots, X_{i30}$ all coming from the standard normal distribution (*i.e.*, Y has nothing to do with any of the X's)
- If we run a stepwise variable selection method to find the best model, we end up with

	β	SE	t	p
X_9	0.2145	0.0871	2.46	0.0156
X_{27}	0.2481	0.0971	2.55	0.0122

Is it really “significant”?

- This example was simulated, but there are hundreds and hundreds of published papers with models just like this one
- In reality, models in observational public health studies and the social sciences are rarely prespecified – but are almost always interpreted as if they were
- The result is that a substantial fraction (far higher than .05) of these published findings are false

Final remarks

- Perhaps the biggest negative in automated model selection is that it allows the analyst to not think about the problem
- Automated model selection has its uses, but it is very important to be aware of the heavy cost of over-analysis it carries with it, and to view its output with critical thinking and skepticism

Final remarks (cont'd)

- The role of model selection and overfitting also vary depending on the purpose of the research and the model:

Purpose	Model selection?
Descriptive	Don't worry about it
Prediction	Model selection criteria are useful guides
Causal inference	As little as possible

- An exception to the causal inference remark is that using model selection to identify important confounders does not necessarily undermine the conclusion with respect to the main exposure of interest