Model building

STA9750

Fall 2018

Logistics

- HW2 up on the website, ***due 10/25***
- ***Midterm assigned Oct 25, Due Oct 31st @ 11:59***
- It covers everything covered in HW1 and HW2
- It will look like a HW assignment (5 exercises)
- It's a take-home exam, which means that you can consult textbooks, notes, and online sources
- Please, don't work in groups! [or ask a friend]

Today

- Review of model selection / model building
- Multiple comparisons
  - ANOVA tells us if there is a difference when we have more than 2 means, but it doesn't tell us what it is
  - There are methods that tell us where the differences are

Model building

General problem: Variable selection

- You have an outcome y and predictors $x_1, x_2, \ldots, x_p$
- Do put all p predictors in the model?
- Some reasons we might not want to include all of them
  - In the application, the client might be interested in knowing which variables seem to be "active" ("predictive")
  - If you don't need some of them, you might be able to get rid of them and get more precise estimates and predictions [*there are some caveats here*]

Two classes of approaches

- All subsets
  - Fit *all possible models* (with all the possible subsets of predictors in and out of the model)
  - *Rank/score* the model according to some criterion
    - Almost infinitely many possibilities, no single criterion is uniformly better than the rest
- Search strategies
  - Look for *good models*, without exploring all the subsets
  - Sometimes you just have to do this because the model space is too big, and you can't go through all subsets…

All subsets

- You go through all subsets, find a “score”… A score like what?
- We saw some last time
  - Adjusted R^2
  - BIC
  - C_p

$$R^2_{\text{adjusted}} = 1 - (1 - R^2)\left(\frac{n-1}{n-p-1}\right)$$

- Unfortunately, R^2 can't get worse as you add in more variables [the residual sum of squares can't get worse after adding a variable… Worst case scenario, the coefficient of that variable is set to 0, and we're done]
- Fortunately, somebody found out a way to penalize the so that there isn't a *bias* towards bigger models
- If all predictors are garbage: $E[R^2] = p/(n-1)$
  - BAD! It increases as we put in bogus predictors
  - Adjusted R^2 is modified so that $E[R^2_{adj}] = 0$ if all predictors are bad

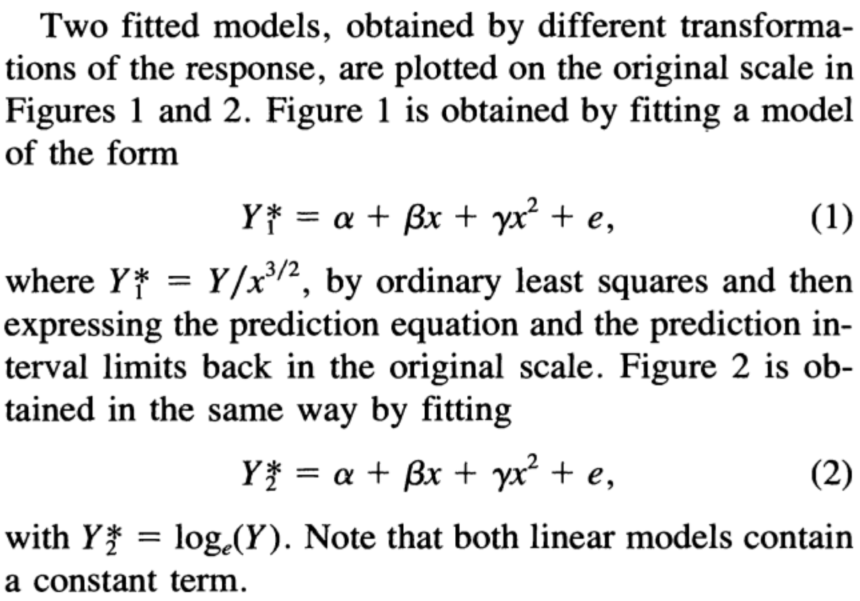
BIC and C_p

- BIC: smaller is better
  - Again, it looks at the tradeoff between smaller residual sum of squares (RSS) and the fact that bigger models (tend to) have smaller residual sum of squares
  - So, it has a term that increases in RSS and some penalty on model "complexity" (p * log n)
- C_p: Pick smallest model whose C_p is roughly p
  - Idea: Same tradeoff between small RSS and penalizing big models
  - Can be derived by thinking how E(RSS) should behave if the model is "correct"

Searching for *good* models

- Sometimes you can't go through all models
- Some strategies for finding *good models*
  - Forward selection: start with no variables, and keep on adding variables one at a time until it doesn't pay off (according to some criterion)
  - Backward selection: start with all of the variables, and keep on dropping variables until it doesn't pay off (according to some criterion)
  - Stepwise selection: start with no variables, and keep on adding variables one at a time until it doesn't pay off. If a variable that seemed useful at some previous step isn't useful anymore, you drop it
- You can use p-values as the criterion to include/exclude variables
- You can use other criteria, such as BIC, etc.

Don't compare model scores if you transformed *y*!


Two fitted models, obtained by different transformations of the response, are plotted on the original scale in Figures 1 and 2. Figure 1 is obtained by fitting a model of the form

$$Y_1^* = \alpha + \beta x + \gamma x^2 + e, \qquad (1)$$

where $Y_1^* = Y/x^{3/2}$, by ordinary least squares and then expressing the prediction equation and the prediction interval limits back in the original scale. Figure 2 is obtained in the same way by fitting

$$Y_2^* = \alpha + \beta x + \gamma x^2 + e, \qquad (2)$$

with $Y_2^* = \log_e(Y)$. Note that both linear models contain a constant term.

Source:

Transformations and *R* ^{2}

Alastair Scott & Chris Wild

Don't compare model scores if you transformed *y*!

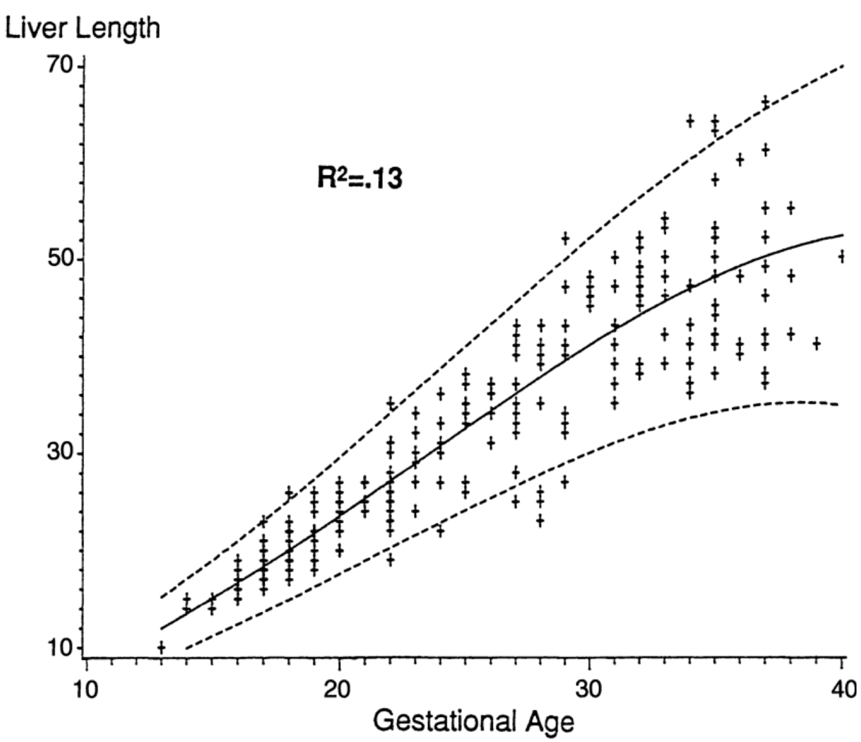


Figure 1. Fitted Model Based on $Y_1^* = Y/x^{3/2}$.

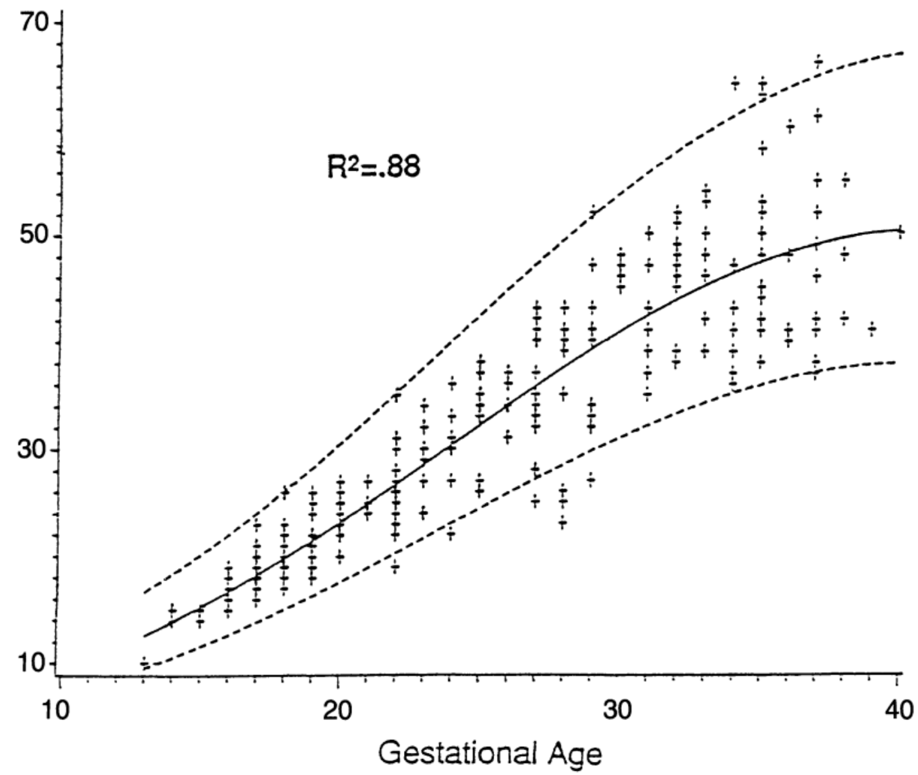


Figure 2. Fitted Model Based on $Y_2^* = \log Y$.

Source:

Transformations and $\boldsymbol{R^2}$

Alastair Scott & Chris Wild