Biostatistics 201a - Lecture 25 - 11/23/11

Contents

- Model Selection Criteria
- Variance Stabilizing + Normalizing Tranformations
- Problems w/ automatic selection procedures (deferred to Lec. 26)

pages = 8

11/23/11
Lecture 25

More on Model Selection

How you do model selection depends on your goal:

- Best predictive model: then you ~~don't~~ want any non-sig variables in your model (leads to overfitting, multic and their associated problems)
- If your goal is understanding relationships among a theoretical set of variables and their possible confounds you may want to be sure that key variables stay in regardless of significance

We need in order to do the first kind of model selection some criteria for what makes a good (or "best") model:

① Percentage of Variability Explained - Higher is Better

- Could use R^2 but as we've seen R^2 always goes up when we add more variables even if they're not useful so this isn't very reliable

Better

- Use R^2_{adj} which penalizes you for adding variables that don't help much
- Use change in R^2 associated with adding a new variable (a la R^2 difference test / partial F test)

② RMSE - "average" error we make using our model ON THE POINTS USED TO CREATE THE MODEL

Lower is better.

③ Predictive Error for New Data

- Training / test data: Instead of using all the data to fit model, you can just

use part of it (half, 2/3rds, ...) and then use that model to make predictions for the points you left out and look at their root mean squared error. Great if you have data to spare but often you don't.

- PRESS / Cross-Validation: "Poor man's train/test setup. Fit the model without one of the data points and then use it to make a prediction for that point. Do this for each point in turn. That way we predict for each point using a model that it wasn't involved in fitting:

$e_{i(-i)} = \hat{Y}_{i(-i)} - Y_i$ ← observed value

fitted value when ith point left out

$PRESS = \sum_{i=1}^{n} e_{i(-i)}^2$ — Pick model with the lowest score

(4) Information Criteria:
- Mallow's Cp statistic
- Akaike Information Criterion (AIC)
- Bayesian Information Criterion (BIC)
⋮

- Based on maximum likelihood ideas ("best models" are the ones most likely to have produced the current data but with a penalty for the complexity of the model (number of variables) and accounting for sample size
- Rank models based on these scores with smaller generally being better → looking for the smallest/simplest model that doesn't result in worse fit.

(5) Want a model that meets your regression assumptions and deals successfully with outliers

What selection are you going to make for transformations?

- If your normality and constant variance assumptions are ok, but residuals have a curved pattern, you transform X.

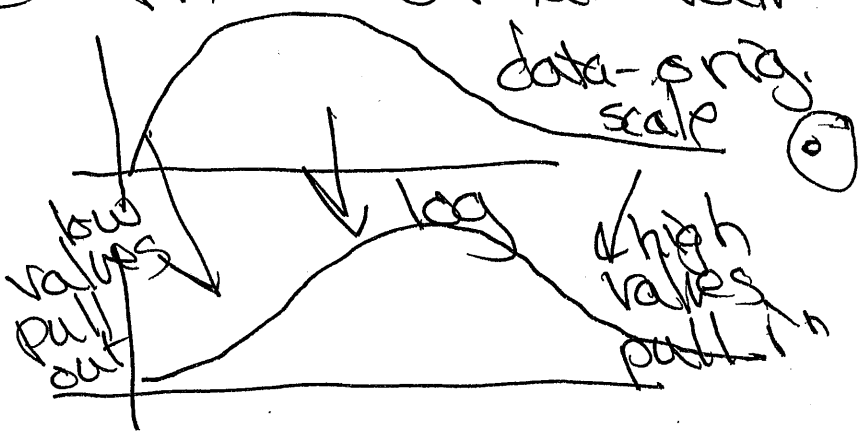
← looks like parabola so I'd use X, X^2

- If you have a problem with normality, constant variance or a large outlier a transform. of Y may be a good idea.

"Variance stabilizing transformations"
"Normalizing transformations"

Example: Log Y: shrinks large values more than small values

$\log_{10} 100 = 2$
$\log_{10} 10 = 1$
$\log_{10} 1 = 0$
$\log_{10} 1/10 = -1$



Also helps to make outliers much less extreme and therefore less influential

- Other transformations that act like this include $\sqrt{Y}$ or $1/y$
- Whole sequences of such transformations have been studied
  - Box-Cox Transformations
  - Tukey Ladder of Powers
- In general transforming Y changes the scale of the problem and this changes the weights / influence of the points in the model

---

How do we use these criteria to pick our model?

* We can't just fit the model with everything in it, throw out the variables that look insignificant and check whether our scores improve

When I remove one variable that affects the interpretation of everything else. This is what led to the stepwise procedures from last time where we either

(a) Backwards: Start w/ all variables in; at each step remove the variable that as a result most improves our fit criterion (p-value least sig; one that has made R^2_{adj} go down, etc.)

(b) Forwards: Start w/ nothing and at each step add variable that most improves the fit statistics.