Inference 2: Fitting models to data: Variable selection

Last updated: October 20, 2017

Multiple regression

Multiple regression = more than one input variables.
Model:

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k.$$

$X_1, \ldots, X_k$: input variables. Can be either numerical or indicator.
Examples

- We saw: one categorical with $k + 1$ categories recoded as k binaries
- teacher: salary vs degree + fulltime + years
- marioKart: totalPrice vs everything else

Multiple regression: diagnostics

Model: $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + \epsilon_i$.
Assumptions on the ϵ_i's:

- Independence
- Normally distributed
- Constant variance

Examples: teacher, marioKart

The model is not a good fit! What to do?

In teacher: salary vs degree + year, we saw that the independence assumption is violated, though adjusted R^2 is 87%.

Is this a good model?

The model is not a good fit! What to do?

In teacher: salary vs degree + year, we saw that the independence assumption is violated, though adjusted R^2 is 87%.

Is this a good model?

- High R^2: good prediction of salary based on degree and year
- Assumptions violated: the actual relationship is NOT linear!

The model is not a good fit! What to do?

In teacher: salary vs degree + year, we saw that the independence assumption is violated, though adjusted R^2 is 87%.

Is this a good model?

- High R^2: good prediction of salary based on degree and year
- Assumptions violated: the actual relationship is NOT linear!

What should we do?

The model is not a good fit! What to do?

In teacher: salary vs degree + year, we saw that the independence assumption is violated, though adjusted R^2 is 87%.

Is this a good model?

- High R^2: good prediction of salary based on degree and year
- Assumptions violated: the actual relationship is NOT linear!

What should we do? Variable selection.

Variable selection: bigger $\neq$ better

Variable selection $=$ discard useless X's, put in useful X's.
A variable is useful if it does ONE or more of the following:

- improve fit
- improve prediction
- improve interpretability

Variable selection: bigger $\neq$ better

Variable selection $=$ discard useless X's, put in useful X's.
A variable is useful if it does ONE or more of the following:

- improve fit
- improve prediction
- improve interpretability

Why not always include all variables?

eg: we ran: `teacher`: total vs degree + year
why not fit `teacher`: total vs degree + year + fulltime + fica + retirement?

Variable selection: bigger $\neq$ better

Variable selection = discard useless X's, put in useful X's.
A variable is useful if it does ONE or more of the following:

- improve fit
- improve prediction
- improve interpretability

Why not always include all variables?

eg: we ran: `teacher`: total vs degree + year
why not fit `teacher`: total vs degree + year + fulltime + fica + retirement?

How to select variables?

How to select variables?

Good tools

- pairwise plots of all variables
- scatterplot: predicted vs observed
- adjusted R^2: R^2 with a penalty on number of variables in model
- p-values of the coefficients

Remedies

- Exclude useless inputs
- transform inputs (eg: $salary^2$)

Example: teacher, marioKart.
Rule of thumb: **Simple = best**

How to select variables?

Variable selection strategies:

- Backward: put all variables in, then gradually remove the useless ones
- Forward: put in one variable at a time
- Stepwise: backward + forward at the same time

Criterion: adjusted R^2, **AIC**, BIC etc

How to select variables?

Variable selection strategies:

- Backward: put all variables in, then gradually remove the useless ones
- Forward: put in one variable at a time
- Stepwise: backward + forward at the same time

Criterion: adjusted R^2, **AIC**, BIC etc
Good news: R can do this automatically for you!
Bad news: R is not that smart. YOU need to check if the results make sense.