Model Selection

ST552 Lecture 21

Charlotte Wickham

2019-02-27

Overview

The process of selecting a few of many possible variables to include in a regression model is known as **variable** or **model selection**.

Reasons for preferring smaller models:

- Occam's razor
- Variables unrelated to the response in a model result in more noise in our estimates of interest
- May be cheaper to collect future data for fewer predictors
- May be easier to communicate/explain

Model Selection in regression problems

Model selection doesn't replace thinking hard about a problem.

Do I want a model that explains/predicts the response well? In what way?

- Set down a criteria for a good model.
- Search for models that do well on your criteria.
- You can often learn about the structure of your data, by examining a few of the "good" models.

Do I want to answer a specific question of interest about the value of parameters in the model?

This generally means you are very interested in a particular p-value and/or confidence interval. In general, how to do valid inference after model selection is an unsolved problem.

- Thinking hard about the problem beforehand (before seeing data) should elicit a model. What are important covariates, should terms enter linearly, what terms will interact etc? If you are familiar enough with the field of application you should be able to do this.
- Model selection will not be done at all. There may be a small set of prespecified models for comparison.
- Diagnostics are still important, you want to check your prespecified model is reasonable.

I think of model selection as:

- a tool for finding predicitve models
- a tool for exploratory data analysis

Respecting heirachy

Some models are heirachical in nature, in that, a lower order term should not be dropped without dropping all higher order terms.

- Polynomials: $y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \epsilon_i$ We wouldn't drop x_i^2 without also dropping x_i^3, similarly we wouldn't drop x_i^2 without dropping x_i^2 and x_i^3.
- Interactions: $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1} x_{i2} + \epsilon_i$. We wouldn't drop the main effect, x_{i1}, without also dropping the interaction, $x_{i1} x_{i2}$.
- Categorical variables: generally, consider keeping or dropping all the indicator variables for a single categorical variable as a group.

You might argue this isn't important for predictive models, but it removes dependence of models on the scale of variables, and makes comparing models easier.

Stepwise methods

(Unless best subsets is infeasible, you **should not** use a stepwise method)

Stepwise methods rely on adding or removing a variable one at a time. Each step chooses the best variable to add/remove based on some criterion, often based on a hypothesis test p-value.

For example, we'll use the p-value from the F-test comparing our current model to the candidate model.

Stepwise methods

Backward Elimination Start with full model. Drop the variable that has the highest p-value above some critical level, α_{crit}. Repeat until all variables in the model have p-values below α_{crit}.

Forward Selection Start with only a constant mean in the model. Add the variable that has the lowest p-value below some critical level, α_{crit}. Repeat until no variable can be added with a p-value below α_{crit}.

Stepwise Selection (many variants) Start with forward selection until there are two terms in the model. Then consider a backwards step. Repeat a forwards step and a backwards step until a final model is reached.

α_{crit} does not have to be 0.05.

Example from Faraway

Backward elimination

```
library(faraway)
data(state)
state_data <- data.frame(state.x77)
lmod <- lm(Life.Exp ~ ., data = state_data)
sumary(lmod)
```

```
##                Estimate  Std. Error t value  Pr(>|t|)
## (Intercept)  7.0943e+01  1.7480e+00 40.5859 < 2.2e-16
## Population   5.1800e-05  2.9187e-05  1.7748   0.08318
## Income      -2.1804e-05  2.4443e-04 -0.0892   0.92934
## Illiteracy   3.3820e-02  3.6628e-01  0.0923   0.92687
## Murder      -3.0112e-01  4.6621e-02 -6.4590  8.68e-08
## HS.Grad      4.8929e-02  2.3323e-02  2.0979   0.04197
## Frost       -5.7350e-03  3.1432e-03 -1.8246   0.07519
## Area        -7.3832e-08  1.6682e-06 -0.0443   0.96491
##
## n = 50, p = 8, Residual SE = 0.74478, R-Squared = 0.74
```

```
drop1(lmod, test = "F")  # will work better when factors are involved
```

```r
# one step of backward elimination
lmod <- update(lmod, . ~ . - Area)
sumary(lmod)
```

```
##                Estimate  Std. Error t value  Pr(>|t|)
## (Intercept)  7.0989e+01  1.3875e+00 51.1652 < 2.2e-16
## Population   5.1883e-05  2.8788e-05  1.8023   0.07852
## Income      -2.4440e-05  2.3429e-04 -0.1043   0.91740
## Illiteracy   2.8459e-02  3.4163e-01  0.0833   0.93400
## Murder      -3.0182e-01  4.3344e-02 -6.9634 1.454e-08
## HS.Grad      4.8472e-02  2.0667e-02  2.3454   0.02369
## Frost       -5.7758e-03  2.9702e-03 -1.9446   0.05839
##
## n = 50, p = 7, Residual SE = 0.73608, R-Squared = 0.74
```

Your turn

What would be the next step of backward elimination using $\alpha_{crit} = 0.05$?

Forward selection

```
lmod <- lm(Life.Exp ~ 1, data = state_data)
add1(lmod, ~ Population + Income + Illiteracy + Murder +
    HS.Grad + Frost + Area,
  test = "F")
```

```
## Single term additions
##
## Model:
## Life.Exp ~ 1
##            Df Sum of Sq    RSS     AIC F value    Pr(>F)
## <none>                  88.299  30.435
## Population  1     0.409 87.890  32.203  0.2233   0.63866
## Income      1    10.223 78.076  26.283  6.2847   0.01562 *
## Illiteracy  1    30.578 57.721  11.179 25.4289 6.969e-06 ***
## Murder      1    53.838 34.461 -14.609 74.9887 2.260e-11 ***
## HS.Grad     1    29.931 58.368  11.737 24.6146 9.196e-06 ***
## Frost       1     6.064 82.235  28.878  3.5397   0.06599 .
## Area        1     1.017 87.282  31.856  0.5594   0.45815
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# one step of forwards selection
lmod <- update(lmod, . ~ . + Murder)
add1(lmod, ~ Population + Income + Illiteracy + Murder +
    HS.Grad + Frost + Area,
  test = "F")
```

```
## Single term additions
##
## Model:
## Life.Exp ~ Murder
##            Df Sum of Sq    RSS     AIC F value   Pr(>F)
## <none>                  34.461 -14.609
## Population  1    4.0161 30.445 -18.805  6.1999 0.016369 *
## Income      1    2.4047 32.057 -16.226  3.5257 0.066636 .
## Illiteracy  1    0.2732 34.188 -13.007  0.3756 0.542910
## HS.Grad     1    4.6910 29.770 -19.925  7.4059 0.009088 **
## Frost       1    3.1346 31.327 -17.378  4.7029 0.035205 *
## Area        1    0.4697 33.992 -13.295  0.6494 0.424375
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Your turn

What would be the next step in forward selection process using $\alpha_{crit} = 0.05$?

Limitations of stepwise methods

1. The make a very limited search through all possible models, so they may miss an "optimal" one.
2. p-values will generally overstate the importance of remaining predictors.
3. Inclusion in the model doesn't correspond to important, and exclusion doesn't correspond to unimportant.
4. Tend to pick smaller models than optimal for prediction.

Next time . . . criterion based procedures