Linear Regression Models
P8111

Lecture 15

Jeff Goldsmith
March 22, 2016

THE DEPARTMENT OF BIOSTATISTICS
Columbia University
MAILMAN SCHOOL OF PUBLIC HEALTH

Today's Lecture

- Welcome back!!
- Model selection vs. model checking
- Stepwise model selection
- Criterion-based approaches

Model selection vs. model checking

In a model of the form

$$y|\boldsymbol{x} = f(\boldsymbol{x}) + \epsilon$$

model selection focuses on how you construct $f(\cdot)$; model checking asks whether the ϵ match the assumed form.

Model selection

Things to keep in mind

- Why am I building a model?
- Is this my primary or secondary analysis?
- What predictors will I allow?
- What forms for $f(x)$?

Motivation

Why am I building a model?

- Estimate associations between x and y
- Test significance of association between x and y
- Predict future y for new x

These goals will generally not result in the same final model.

Primary vs secondary

Is this my primary or secondary analysis?

- Seriously – have you (or anyone else) analyzed this data before?
- Primary analyses are often very constrained or have the goal of confirming a hypothesis
- Secondary analyses are often less constrained; may be examining hunches or generating new hypotheses

Both are valid, but have different implications for multiple comparisons

Model structure

What predictors will I allow? What forms for $f(x)$?

- All variables? All continuous variables? Binary versions of continuous variables? Known significant variables?
- Linear models? Non-linearity? Interactions?

Some of this you know ahead of time, some you discover as you go

Model selection is hard

- If we're asking which is the "true" model, we're gonna have a bad time
- In practice, issues with sample size, collinearity, and available predictors are real problems
- It is often possible to differentiate between better models and less-good models, though

Estimating associations

- We may not care about whether an association is significant in our data; we're just looking for associations
- Some covariates should be included regardless of significance – models have to be convincing in the scientific context
- This can affect the baseline model, or at least the class of models one considers

Basic idea for model selection

- Specify a class of models
- Define a criterion to summarize the fit of each model in the class
- Select the model that optimizes the criterion you're using

Again, we're focusing on $f(x)$ in the model specification. Once you've selected a model, you should subject it to regression diagnostics – which might change or augment the class of models you specify or alter your criterion.

Classes of models

Some examples of classes of models:

- Linear models including all subsets of $x_1, ..., x_p$
- Linear models including all subsets of $x_1, ..., x_p$ and their first order interactions
- All functions $f(x_1)$ such that $f''(x_1)$ is continuous
- Additive models of the form $f(\boldsymbol{x}) = f_1(x_1) + f_2(x_2) + f_3(x_3)...$ where $f_k''(x_k)$ is continuous

Popular criteria

- Akaike Information Criterion
- Bayes Information Criterion
- F- or t-tests
- Prediction RSS (PRESS) or CV

Sequential methods (Forward Stepwise)

- Start with “baseline” (usually intercept-only) model
- For every possible model that adds one term, evaluate the criterion you’ve settled on
- Choose the one with the best “score” (lowest AIC, smallest p-value)
- For every possible model that adds one term to the current model, evaluate your criterion
- Repeat until either adding a new term doesn’t improve the model or all variables are included

Sequential methods (Backward Stepwise)

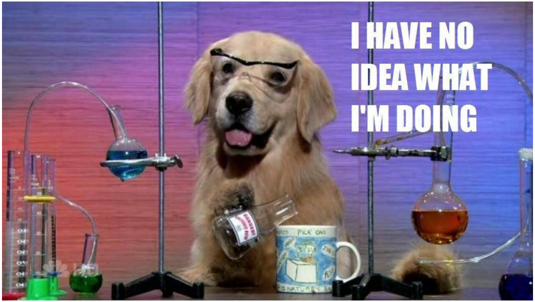
- Start with every term in the model
- Consider all models with one predictor removed
- Remove the term that leads to the biggest score improvement
- Repeat until removing additional terms doesn't improve your model

Sequential methods

- There are many potential models – usually exhausting the model space is difficult or infeasible
- Stepwise methods don't consider all possibilities
- Stepwise methods work well for *F*- and *t*-tests, which require nested models
- Other criteria don't require nested models (which can be nice) but don't ascertain significance (which can be a downer)

Sequential methods

Sequential methods are basically an admission that you had no idea what you were doing with the data



AIC

AIC ("An Information Criterion") measures goodness-of-fit through RSS (equivalently, log likelihood) and penalizes model size:

$$AIC = n \log(RSS/n) + 2p$$

- Small AIC's are better, but scores are not directly interpretable
- Penalty on model size tries to induce *parsimony*

BIC

BIC ("Bayes Information Criterion") similarly measures goodness-of-fit through RSS (equivalently, log likelihood) and penalizes model size:

$$BIC = n \log(RSS/n) + p \log(n)$$

- Small BIC's are better, but scores are not directly interpretable
- AIC and BIC measure goodness-of-fit through RSS, but use different penalties for model size. They won't always give the same answer

Adjusted R^2

- Recall:
$$R^2 = 1 - \frac{RSS}{TSS}$$
- Definition of adjusted R^2:
$$\begin{aligned} R_a^2 &= 1 - \frac{RSS/(n-p-1)}{TSS/(n-1)} = 1 - \frac{\hat{\sigma}^2_{model}}{\hat{\sigma}^2_{null}} \\ &= 1 - \frac{n-1}{n-p-1}(1-R^2) \end{aligned}$$
- Minimizing the standard error of prediction means minimizing $\hat{\sigma}^2_{model}$ which in turn means maximizing R_a^2
- Adding a predictor will not necessarily increase R_a^2 unless it has some predictive value

PRESS

Prediction residual sum of squares is the most clearly focused on prediction

$$PRESS = \sum (y_i - \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}_{(-i)})^2$$

Looks computationally intensive, but for linear regression models this is equivalent to

$$PRESS = \sum \left(\frac{r_i}{1 - h_{ii}} \right)^2$$

PRESS is leave-one-out cross validation; other forms of cross validation are equally vaid

Life expectancy example

- Response: life expectancy
- Predictors: population, capital income, illiteracy rate, murder rate, percentage of high-school graduates, number of days with minimum temperature < 32, land area
- Data for 50 US states
- Time span: 1970-1975

Example

```
> data(state)
> statedata = data.frame(state.x77,row.names=state.abb)
> g = lm(Life.Exp ~., data=statedata)
> summary(g)
...

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.094e+01  1.748e+00  40.586  < 2e-16 ***
Population   5.180e-05  2.919e-05   1.775   0.0832 .
Income      -2.180e-05  2.444e-04  -0.089   0.9293
Illiteracy   3.382e-02  3.663e-01   0.092   0.9269
Murder      -3.011e-01  4.662e-02  -6.459 8.68e-08 ***
HS.Grad      4.893e-02  2.332e-02   2.098   0.0420 *
Frost       -5.735e-03  3.143e-03  -1.825   0.0752 .
Area        -7.383e-08  1.668e-06  -0.044   0.9649

...
> AIC(g)
[1] 121.7092
```

Example

```
> g = lm(Life.Exp ~ . - Area, data=statedata)
> summary(g)
...

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.099e+01  1.387e+00  51.165  < 2e-16 ***
Population   5.188e-05  2.879e-05   1.802   0.0785 .
Income      -2.444e-05  2.343e-04  -0.104   0.9174
Illiteracy   2.846e-02  3.416e-01   0.083   0.9340
Murder      -3.018e-01  4.334e-02  -6.963 1.45e-08 ***
HS.Grad      4.847e-02  2.067e-02   2.345   0.0237 *
Frost       -5.776e-03  2.970e-03  -1.945   0.0584 .

...
> AIC(g)
[1] 119.7116
```

Example

```
> g = lm(Life.Exp ~ . - (Area + Illiteracy), data=statedata)
> summary(g)
...

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.107e+01  1.029e+00  69.067  < 2e-16 ***
Population   5.115e-05  2.709e-05   1.888   0.0657 .
Income      -2.477e-05  2.316e-04  -0.107   0.9153
Murder      -3.000e-01  3.704e-02  -8.099 2.91e-10 ***
HS.Grad      4.776e-02  1.859e-02   2.569   0.0137 *
Frost       -5.910e-03  2.468e-03  -2.395   0.0210 *

...
> AIC(g)
[1] 117.7196
```

Example

```
> g = lm(Life.Exp ~ . - (Area + Illiteracy + Income), data=statedata)
> summary(g)
...

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.103e+01  9.529e-01  74.542  < 2e-16 ***
Population   5.014e-05  2.512e-05   1.996  0.05201 .
Murder      -3.001e-01  3.661e-02  -8.199 1.77e-10 ***
HS.Grad      4.658e-02  1.483e-02   3.142  0.00297 **
Frost       -5.943e-03  2.421e-03  -2.455  0.01802 *

...
> AIC(g)
[1] 115.7326
```

Example

```
> g = lm(Life.Exp ~ . - (Area + Illiteracy + Income + Population), data=statedata)
> summary(g)
...

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 71.036379   0.983262  72.246  < 2e-16 ***
Murder      -0.283065   0.036731  -7.706 8.04e-10 ***
HS.Grad      0.049949   0.015201   3.286  0.00195 **
Frost       -0.006912   0.002447  -2.824  0.00699 **

...
> AIC(g)
[1] 117.9743
```

So now what?

- It's common to treat the final model as if it were the only model ever considered – to base all interpretation on this model and to assume the inference is accurate
- This doesn't really reflect the true model building procedure, and can misrepresent what actually happened
- Inference is difficult in this case; it's hard to write down a statistical framework for the entire procedure
- Predictions can be made from the final model, but uncertainty around predictions will be understated
- P-values, CIs, etc will be incorrect

What to do?

- Remember the bootstrap?
- We can resample subjects with replacement, and repeat the entire process
- Produce predicted values $\hat{y}_i^b$ for $\boldsymbol{x} = \{\boldsymbol{x}_i\}_{i=1}^I$ based on the final bootstrap model
- Base inference for predictions on the distribution of $\{\hat{y}_i^b\}_{b=1}^B$

Downside – only gives inference for predicted values, not for the parameter estimates. Bootstrap models might not be the same as the final model (which is kind of the point).

Shrinkage/penalization

As a preview of things to come -

- There are other strategies for model/variable selection or tuning
- *Penalized regression* adds an explicit penalty to the least squares criterion
- That penalty can keep regression coefficients from being too large, or can shrink coefficients to zero
- We'll worry more about this next time

Variable selection in polynomial models

A quick note about polynomials. If you fit a model of the form

$$y_i = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon_i$$

and find the quadratic term is significant but the linear term is not...

- You should still keep the linear term in the model
- Otherwise, your model is sensitive to centering – shifting x will change your model

Today's big ideas

- Model selection

- Suggested reading: Ch 10