STA314, Lecture 4

Sept 27, 2019

Stanislav Volgushev

Multiple linear regression: qualitative predictors

Motivation: so far we have considered numeric predictors. But sometimes predictors are qualitative (not numeric). Example: the `Credit` data set contains information data on student status (yes/no), ethnicity (African American, Asian, Caucasian) etc.

It does not make sense to treat student status or ethnicity as numeric variable. To include those variables in linear models, they are converted to *dummies*.

Example 1: `student` has two possible values. One possibility to convert this into dummy

$$x_i = \begin{cases} 1 & \textit{if} \quad \text{i'th person student} \\ 0 & \textit{if} \quad \text{i'th person not student} \end{cases}$$

Then

$$b_1 + b_2 x_i = \begin{cases} b_1 + b_2 & \textit{if} \quad \text{i'th person student} \\ b_1 & \textit{if} \quad \text{i'th person not student} \end{cases}$$

Interpretation:

- b_1 mean for non-students
- $b_1 + b_2$ mean for students
- b_2 is 'effect of being student'.

Multiple linear regression: qualitative predictors with $k > 2$ levels

Example 2: the credit data set contains a variable called `ethnicity` which takes values *African American, Asian, Caucasian*. Clearly no numerical ordering.

Idea: create $k - 1$ dummies. For example

$$x_{i,1} = \begin{cases} 0 & \textit{if} \quad \text{i'th person not African American} \\ 1 & \textit{if} \quad \text{i'th person African American} \end{cases}$$

$$x_{i,2} = \begin{cases} 0 & \textit{if} \quad \text{i'th person not Asian} \\ 1 & \textit{if} \quad \text{i'th person Asian} \end{cases}$$

Then

$$b_1 + b_2 x_{i,1} + b_3 x_{i,3} = \begin{cases} b_1 & \textit{if} \quad \text{i'th person Caucasian} \\ b_1 + b_2 & \textit{if} \quad \text{i'th person African American} \\ b_1 + b_3 & \textit{if} \quad \text{i'th person Asian} \end{cases}$$

Running linear regression in **R** using the `lm` function will automatically convert qualitative predictors into dummies, *if qualitative predictors are not coded as numeric in advance*.

Variable selection

Motivation

- With Advertisement data, we determined the important predictors by 'trial and error'. This worked because there were only 3 predictors.
- The `Credit` data set has (after introducing dummies) 11 predictors. R output below indicates that not all are important.
- We will now look at systematic approaches to finding the important predictors.

```
Call:
lm(formula = Balance ~ ., data = Credit)

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)        -479.20787   35.77394 -13.395  < 2e-16 ***
Income               -7.80310    0.23423 -33.314  < 2e-16 ***
Limit                 0.19091    0.03278   5.824 1.21e-08 ***
Rating                1.13653    0.49089   2.315   0.0211 *
Cards                17.72448    4.34103   4.083 5.40e-05 ***
Age                  -0.61391    0.29399  -2.088   0.0374 *
Education            -1.09886    1.59795  -0.688   0.4921
GenderFemale        -10.65325    9.91400  -1.075   0.2832
StudentYes          425.74736   16.72258  25.459  < 2e-16 ***
MarriedYes           -8.53390   10.36287  -0.824   0.4107
EthnicityAsian       16.80418   14.11906   1.190   0.2347
EthnicityCaucasian   10.10703   12.20992   0.828   0.4083
```

Selecting relevant predictors

If we have 'many' candidate predictors, how do we select 'the right ones'? Reasons for selecting fewer predictors:

Interpretation: if we can use fewer predictors to describe the response this helps to understand relationship between response and predictor and explain it to others. In the future, less data need to be collected and stored.

Improved prediction: in previous lectures we have seen that including many irrelevant predictors can lead to bad prediction performance.

Best subset selection

Assume we have p candidate predictors. Best subset selection proceeds as follows

---

1. Denote by $\mathcal{M}_0$ a model with no predictors and just the intercept.
2. For $k = 1, 2, ..., p$
   - 2.1 For each set $\{j_1, ..., j_k\} \subset \{1, ..., p\}$ with exactly k elements compute RSS of linear model with predictors $x_{1,j_1}, ..., x_{1,j_k}$.
   - 2.2 Among models from step 2.1 select the one with smallest RSS. This is model $\mathcal{M}_k$.
3. From steps 1 and 2, we have *candidate models* $\mathcal{M}_0, \mathcal{M}_1, ..., \mathcal{M}_p$ with $0, 1, ..., p$ predictors. Select a model among those candidates, details later in this lecture.

- Motivation for step 2: models with same number of predictors have same flexibility.

How many models do we need to fit for this approach? *See notes on blackboard for details and an example.*

Potential issue: computational complexity when p is large.

- Doing cross-validation or the validation set approach in this setting right is tricky but very important. Comments on that later.

Forward stepwise selection

1. Denote by $\mathcal{M}_0$ a model with no predictors and just the intercept.
2. For $k = 1, 2, ..., p$, starting with $\mathcal{M}_0$ do this iteratively
   - 2.1 Consider all possible $p + 1 - k$ predictors which are not in $\mathcal{M}_{k-1}$, this gives candidate models $\mathcal{M}_{k,1}, ..., \mathcal{M}_{k,p+1-k}$.
   - 2.2 Among models $\mathcal{M}_{k,1}, ..., \mathcal{M}_{k,p+1-k}$ select the one with smallest RSS, call it $\mathcal{M}_k$.
3. From steps 1 and 2, we have candidate models $\mathcal{M}_0, \mathcal{M}_1, ..., \mathcal{M}_p$ with $0, 1, ..., p$ predictors. Select a model among those candidates, details later in this lecture.

How many models do we need to fit for this approach? *See notes on blackboard for details and an example.*

- Less computation than best subset selection (difference can be substantial).
- Price for reducing computation: this is not guaranteed to not find the best model with given number of predictors.
- Doing cross-validation or the validation set approach in this setting right is tricky...

Backward stepwise selection

1. Denote by $\mathcal{M}_p$ a model which includes all predictors.
2. For $k = p-1, p-2, ..., 1, 0$, starting with $\mathcal{M}_p$ do this iteratively
   - 2.1 Consider all possible $k+1$ predictors which are in $\mathcal{M}_{k+1}$ and remove one at a time, this gives candidate models $\mathcal{M}_{k,1}, ..., \mathcal{M}_{k,k+1}$.
   - 2.2 Among models $\mathcal{M}_{k,1}, ..., \mathcal{M}_{k,k+1}$ select the one with smallest RSS, call it $\mathcal{M}_k$.
3. From steps 1 and 2, we have candidate models $\mathcal{M}_0, \mathcal{M}_1, ..., \mathcal{M}_p$ with $0, 1, ..., p$ predictors. Select a model among those candidates, details later in this lecture.

How many models do we need to fit for this approach? *See notes on blackboard for details and an example.*

- Similar amount of computation as forward stepwise.
- As forward stepwise, this is not guaranteed to not find the best model with given number of predictors.
- Doing cross-validation or the validation set approach in this setting right is tricky...

Analytic corrections for training error I

- So far we discussed how to use cross-validation to obtain estimators of the test error. We have also seen (in earlier lectures) that for linear models R^2 is not helpful for model selection (recall why?).
- For linear models, there are other popular approaches which are based on analytic corrections. Those can be faster than cross-validation and it is important to know their meaning.

Mathematical motivation: let predictors $X_i = (1, x_{i,1}, ..., x_{i,d})^\top$ be $d+1$-dimensional.

- Consider n data (X_i, y_i) with $y_i = X_i^\top b + \varepsilon_i$, ε_i i.i.d. with $\mathbb{E}[\varepsilon_i] = 0, \mathbb{E}[\varepsilon_i^2] = \sigma^2$ and $\mathbf{X}\mathbf{X}^\top$ of full rank.
- Let $\hat{y}_i = X_i^\top \hat{b}$
- Consider $y_i^{new} = X_i^\top b + \varepsilon_i^{new}$ with ε_i^{new} i.i.d. $\mathbb{E}[\varepsilon_i^{new}] = 0, \mathbb{E}[(\varepsilon_i^{new})^2] = \sigma^2$ independent of $\varepsilon_1, ..., \varepsilon_n$.

One can prove (see derivation in class)

$$\sum_i \mathbb{E}[(y_i^{new} - \hat{y}_i)^2] - \sum_i \mathbb{E}[(y_i - \hat{y}_i)^2] = 2(d+1)\sigma^2.$$

Interpretation: on average training error underestimates test error by $2(d+1)\sigma^2$. This motivates the AIC on next slide.

Analytic corrections for training error II

Most popular approaches for least squares linear regression

$$\text{Mallow's } C_p: \quad C_p := \frac{1}{n}(RSS + 2(d+1)\hat{\sigma}^2)$$

$$\text{Akaike information criterion}: \quad AIC := \frac{1}{n\hat{\sigma}^2}(RSS + 2(d+1)\hat{\sigma}^2)$$

$$\text{Bayesian information criterion}: \quad BIC := \frac{1}{n\hat{\sigma}^2}(RSS + (d+1)(\log n)\hat{\sigma}^2)$$

$$\text{Adjusted } R^2: \quad aR^2 := 1 - \frac{RSS/(n-d-1)}{TSS/(n-1)}$$

- In each case $\hat{\sigma}^2$ denotes an estimator of the error variance $Var(\varepsilon_1)$. Usually variance estimator from 'full model' (i.e. including all predictors).
- Note: in textbook C_p, AIC, BIC are given with d instead of $d+1$. This does not matter for model selection discussed next.
- C_p, AIC, BIC introduce additional 'penalty' term. Including additional predictors will decrease RSS but increase penalty term.
- Adjusted R^2 smaller than 'usual' R^2. No real theoretical justification, so should not be used in practice.

R examples part 1: file `STA314-Lecture04-AIC-BIC-CreditAnalysis`

Comments

- Historically, AIC and BIC were derived for general *maximum likelihood* methods. Least squares regression is just one special example.
- Depending on book, software etc. AIC and BIC have different forms. All forms lead to the same model selection.
- For approaches that follow, AIC and C_p select exactly the same model (why is this true mathematically?). Deeper reason: maximum likelihood for Gaussian errors is the least squares estimator that we considered in lectures.

Some facts and statistical folklore: AIC vs BIC

- Fact: for $n \geq 8$ BIC corresponds to larger penalties compared to AIC (why?). Thus BIC tends to prefer smaller models compared to AIC.
- Fact: AIC does usually not lead to 'consistent model selection', i.e. it includes noise variables with high probability. BIC can perform consistent model selection (this can be formalized in a mathematical framework).
- Folklore (often true): AIC leads to models with better prediction performance. This is especially true when there are some 'weak' predictors which BIC can miss.

AIC vs BIC and consistent model selection I

Example illustrating why BIC can perform consistent model selection and AIC in general won't:

- true model $y_i = \varepsilon_i$, $\varepsilon_i \sim \mathcal{N}(0,1)$. x_i fixed (non-random) one-dimensional predictors unrelated to y_i.
- Use AIC and BIC to select between model with no predictor and model with x_i as predictors.
- denote by $RSS(\mathcal{M}_0)$ residual sum of squares for model with just intercept and by $RSS(\mathcal{M}_1)$ residual sum of squares for model including x_i.
- Assume that we plug in the true value for $\sigma^2 = 1$ in AIC and BIC.

For simplicity assume $\overline{x} = n^{-1}\sum_i x_i = 0$. Then

$$\hat{b}_1 = \bar{y}, \quad \hat{b}_2 = \frac{\overline{xy}}{\overline{x^2}} = \frac{n^{-1}\sum_{i=1}^n x_i y_i}{n^{-1}\sum_{i=1}^n x_i^2}$$

This implies (see derivation in class) $RSS(\mathcal{M}_0) - RSS(\mathcal{M}_1) \sim \chi^2_1$ i.e. $RSS(\mathcal{M}_0) - RSS(\mathcal{M}_1)$ follows a chi-squared distribution with one degree of freedom.