note @25 stop following 37 views

Actions

Notes on HW04 and a new due date

Hi all,

All three of the problems for HW04 turned out to be more involved than I expected when I made the assignment, and in addition we haven't really talked about the tools you need to do problem #2 (primarily, forward and backward selection).

SOoooo, first things first:

I'm changing the due date for HW04 to Weds at 1159, instead of Mon at 1159. This will also delay when I publish the take-home midterm for you to do.

Now, on to the individual problems…

Problem #1:

There's nothing that weird about this problem, except that you should re-generate all the output using the `cars04.csv` file, rather that using the pre-done output in Sheather. The reason for this is that there are some errors in Sheather's output (I suspect he didn't use the same transformations that he claimed to use in the problem. Please use the transformations he claimed to use, when you re-do the fits, graphs, etc.

Note that for part (f) all you have to do is say what you would do to accommodate the analyst's boss; you don't have to actually do anything in R. On the other hand, if you do try to figure out how to do it in R, I'd love to see it. I will put my way of dealing with it in the solutions.

Problem #2:

Part 2(a):

The method I showed in class worked to get all-subsets model selection using R^2_{adj} or BIC. However it will not work for AIC or CAIC (Sheather calls this AIC_c), because the `subsets()` function doesn't know how to do AIC or CAIC.

You can still use the output from a command like

```
all.subsets <- regsubsets(log(PrizeMoney) ~ ., data=golf.red)
```

but you have to do some things by hand. You will want to start by making

```
tmp <- summary(all.subsets)
```

If you ask for `names(tmp)` you will get

```
## [1] "which"  "rsq"    "rss"    "adjr2"  "cp"     "bic"    "outmat" "obj"
```

run code snippet

- The first element, `which`, is a matrix with as many rows as there are subset sizes, and columns labelled with the names of the variables in your data frame. In each row there are `TRUE` and `FALSE` values. `TRUE` means you should include that variable in the model for that row, `FALSE` means you should not.

- The next several elements are more or less self-explanatory: R^2, RSS, R^2_{adj}, Mallows' C_p, and BIC. We can safely ignore the last two elements, `outmat` and `obj`.

If you make a data frame like this

```
attach(tmp)
results <- data.frame(which,rss,adjr2,bic)
detach()
```

and then print `results` out, you can easily scan the `adjr2` and `bic` columns to see which model (which row) maximizes R^2_{adj} and which model minimizes BIC. Your "by hand" answers should agree with the output from

```
subsets(all.subsets,statistic="adjr2")
subsets(all.subsets,statistic="bic")
```

run code snippet

To calculate AIC and CAIC, we need to look back at lecture 08, slide 7, at the bottom: We can write the log-likelihood as

$$(\text{log-likelihood}) = c_1(n) - c_2(n)\log(RSS) = -\tfrac{n}{2}\log(2\pi) - \tfrac{n}{2}\log(RSS)$$

which means we can calculate AIC by hand as

$$AIC = -2(\text{log-liklihood}) \; + 2 * (p+2) \qquad = [n * log(2pi)] + n * log(RSS) + 2 * (p+2)$$

and since the `[n*log(2pi)]` term will cancel when we subtract AIC's we can ignore it and just write

$$AIC = n * log(RSS) + 2 * (p+2)$$

Similarly,

$$CAIC = AIC + 2 * (p+2) * (p+3)/(n-p-1)]$$
$$= n * log(RSS) + 2 * (p+2) + 2 * (p+2) * (p+3)/(n-p-1)$$

and

$$BIC = n * log(RSS) + log(n) * (p+2)$$

Here, n is the number of observations (rows) in the golf data frame, p is the number of parameters (which you could get from `row.names(tmp$which)` if you want) in the model (row of `tmp$which`) you are evaluating, and you can get RSS from the corresponding element of `tmp$rss`.

In this way, you can calculate AIC, CAIC, and recalculate BIC from scratch, all by hand, and you can find the best AIC, CAIC (and BIC again if you like) models.

Note that different authors will use somewhat different definitions of AIC and BIC. The differences are usually just in what is done with the constants $c_1(n)$ and $c_2(n)$, so the value of the criterion changes, but the model that minimizes the criterion does not change. If you compare `tmp$bic` with your "by-hand" BIC values you will see this unimportant difference in the values (you may even find that the `tmp$bic` are negative while the by-hand BIC's are positive, but if you make a scatterplot you will see they line up almost perfectly on a straight line.

Another way to tackle this would be to fit the model indicated by the `TRUE` and `FALSE` values in each row of `tmp$which` and then apply the functions `AIC()` and `BIC()` from `library(MASS)` to that model. Again, the exact AIC and BIC values might not be the same as you get using the other methods above, but the model that minimizes each version of AIC should be the same across all the different ways to get AIC, and the model that minimizes each version of BIC should be the same across all the different versions of BIC, etc.

Part 2(b):

This part asks you to do backwards selection: start with the largest model, then remove the variable that does the least damage to RSS, then from this model remove the next variable that does the least damage to RSS, and so forth, until you get down to one variable remaining. You should end up with as many models as you have predictors in the model. Fortunately, `regsubsets()` can do this for you:

```
backward <- regsubsets(log(PrizeMoney) ~ ., data=golf.red, method ="backward")
```

Then you would want to set

```
tmp <- summary(backward)
```

and proceed as in part (a) above to compare the models with R^2_{adj}, BIC, AIC and CAIC.

Part 2(c):

Now you should do forward selection: start with just the intercept model and add the variable that improves RSS the most. Then add the next variable to that model that improves RSS the most, etc., and keep going until you arrive at the full model. Again, `regsubsets()` can do this for you:

```
forward <- regsubsets(log(PrizeMoney) ~ ., data=golf.red, method ="forward")
```

Then you would want to set

```
tmp <- summary(forward)
```

and proceed as in part (a) above to compare the models with R^2_{adj}, BIC, AIC and CAIC.

Parts 2(d), 2(e) and 2(f):

I don't think these require any special hints.

Problem 3:

Part 3(a):

This is long but it doesn't require any special hints.

Part 3(b):

I suggest taking out the variables that you are not going to use anyway in Part 3(c) right here. That way you have less variables to worry about.

Also, there are two groups of variables that have very special relationships with `btystdave`; be on the lookout for them, and try to decide something sensible to do with them once you find them.

Part 3(c):

This seems pretty straightforward.

Part 3(d):

Remember to justify each variable you remove with some combination (or subset) of the t-statistic, the vif value, and your high level of knowledge about how universities, classes, and course and teacher evaluations work. At some point you may also wish to remove `profevauation` (depending on what model you end up with) since it seems to mask the effects of other variables.

After you have removed all the variables you are going to remove, refit the remaining model and see if it has better interpretations than the original model.

Part 3(e):

I don't think any special hints are needed.

I hope all of this helps!

-BJ

hw4

Edit good note | 1 Updated 4 days ago by Brian Junker

followup discussions, *for lingering questions and comments*

Start a new followup discussion

Compose a new followup discussion