Model Selection in Regression

Variable Selection

Variable Selection

Why might we not want to include all the variables available to us?

- Overfitting: Using many extra variables gives the model more flexibility; it might be to tailored to the training data.
  - Recall: Polynomials in week 1

- Interpretability: We'd like to know which variables "matter most" to the response, and have accurate coefficient estimates.
  - What if two variables measure the same information?
  - What if the variables are *linearly dependent*?

Data

Recall: 62 unique words describing Cannabis strains. New Response variable: `Rating`

```
cann <- read_csv("https://www.dropbox.com/s/s2a1uoiegitupjc/cannabis_full.csv?dl=1")

cann <- cann %>%
  select(-Type, -Strain, -Effects, -Flavor, -Dry, -Mouth)

head(cann)
```

```
## # A tibble: 6 x 63
##   Rating Creative Energetic Tingly Euphoric Relaxed Aroused Happy Uplifted
##    <dbl>    <dbl>     <dbl>  <dbl>    <dbl>   <dbl>   <dbl> <dbl>    <dbl>
## 1    4          1         1      1        1       1       0     0        0
## 2    4.7        1         1      0        0       1       1     1        0
## 3    4.4        1         1      0        0       1       0     1        1
## 4    4.2        1         0      1        0       1       0     0        1
## 5    4.6        0         0      0        1       1       0     1        1
## 6    0          0         0      0        0       0       0     0        0
## # ... with 54 more variables: Hungry <dbl>, Talkative <dbl>, Giggly <dbl>,
## #   Focused <dbl>, Sleepy <dbl>, Earthy <dbl>, Sweet <dbl>, Citrus <dbl>,
## #   Flowery <dbl>, Violet <dbl>, Diesel <dbl>, Spicy/Herbal <dbl>, Sage <dbl>,
## #   Woody <dbl>, Apricot <dbl>, Grapefruit <dbl>, Orange <dbl>, Pungent <dbl>,
## #   Grape <dbl>, Pine <dbl>, Skunk <dbl>, Berry <dbl>, Pepper <dbl>,
## #   Menthol <dbl>, Blue <dbl>, Cheese <dbl>, Chemical <dbl>, Mango <dbl>,
```

Option 1: Best Subset Selection

Let's try every possible subset of variables and pick the best one.

What do we mean by best?

Penalized metrics:

- BIC
- AIC
- Mallow's Cp
- Adjusted R-squared

Cross-Validation???

Option 1: Best Subset Selection

The problem:

```
Rating ~ Creative
Rating ~ Creative + Energetic
Rating ~ Creative + Energetic + Tingly
Rating ~ Creative + Tingly
...
```

62 variables = 4.6 quintillion models

(Plus cross-validation????)


That's a yikes from me dawg

Option 1: Best Subset Selection

If you have only a few variables, go for it.

In realistic settings, it's not practical.

Use the `leaps` package.

Best Subset Selection with `leaps`

Best model of each size, based on R-squared:

```
library(leaps)
models <- regsubsets(Rating ~ Creative + Energetic + Tingly,
                     data = cann, method = "exhaustive")

summary(models)
```

```
## Subset selection object
## Call: regsubsets.formula(Rating ~ Creative + Energetic + Tingly, data = cann,
##     method = "exhaustive")
## 3 Variables  (and intercept)
##           Forced in Forced out
## Creative      FALSE      FALSE
## Energetic     FALSE      FALSE
## Tingly        FALSE      FALSE
## 1 subsets of each size up to 3
## Selection Algorithm: exhaustive
##          Creative Energetic Tingly
## 1  ( 1 ) "*"      " "       " "
## 2  ( 1 ) "*"      "*"       " "
## 3  ( 1 ) "*"      "*"       "*"
```

Why can we compare same-size models via R-squared, not penalized metrics or cross-validation?

Best Subset Selection with `leaps`

Now compare same-size models:

```
summary(models)$adjr2  # bigger is better
```

```
## [1] 0.01006460 0.01400814 0.01633805
```

```
summary(models)$cp     # smaller is better
```

```
## [1] 16.694102  8.454895  4.000000
```

```
summary(models)$bic    # more negative is better
```

```
## [1]  -8.840652 -11.303084 -10.016876
```

Option 2: Backwards Selection

Start with all candidate variables in the model.

Drop the *worst* variable. (p-vals or R-squared)

Check if dropping it helped. (penalized metric or cross-validation)

Stop when dropping is no longer good.

Backwards selection with `leaps`

```
models <- regsubsets(Rating ~ .,
                     data = cann, method = "forward",
                     nvmax = 61)
```

```
## Reordering variables and trying again:
```

```
summary(models)
```

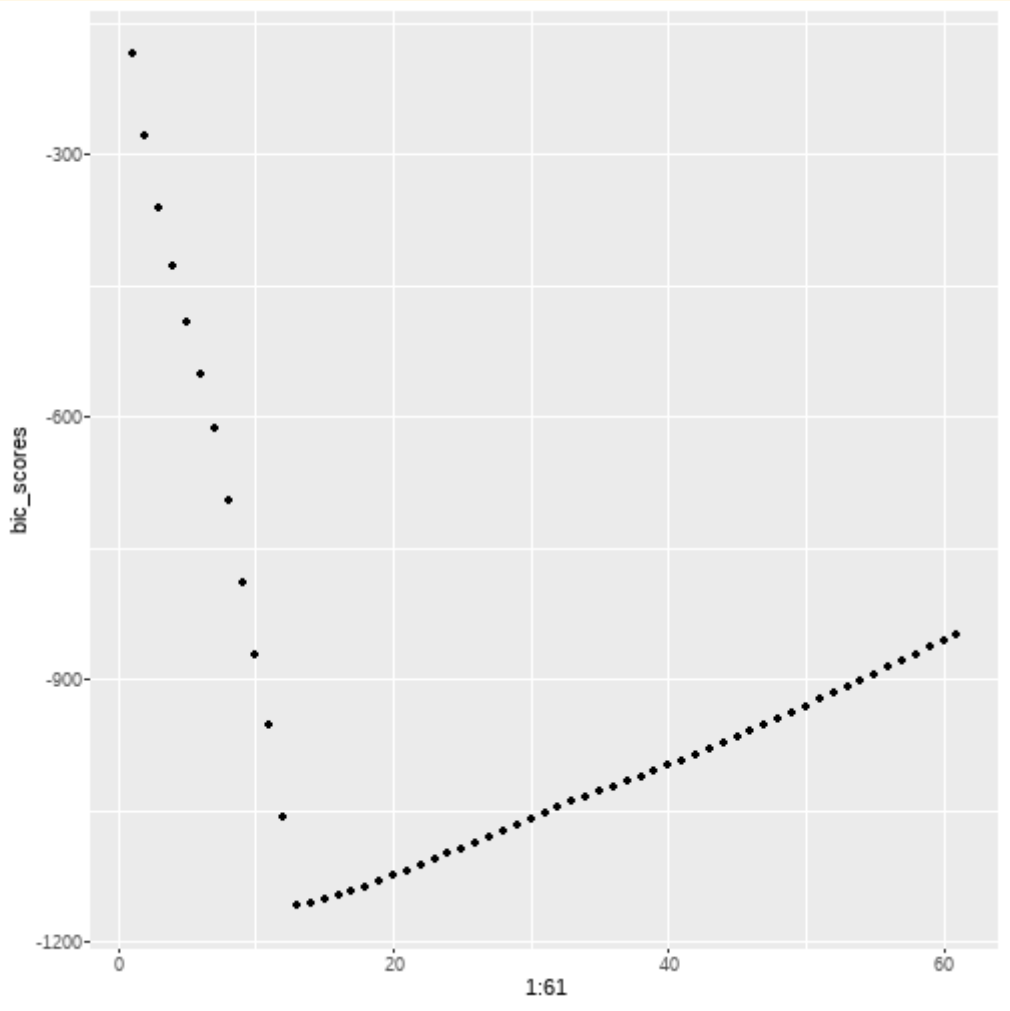
```
## Subset selection object
## Call: regsubsets.formula(Rating ~ ., data = cann, method = "forward",
##     nvmax = 61)
## 62 Variables  (and intercept)
##               Forced in Forced out
## Creative          FALSE      FALSE
## Energetic         FALSE      FALSE
## Tingly            FALSE      FALSE
## Euphoric          FALSE      FALSE
## Relaxed           FALSE      FALSE
## Aroused           FALSE      FALSE
## Happy             FALSE      FALSE
## Uplifted          FALSE      FALSE
## Hungry            FALSE      FALSE
## Talkative         FALSE      FALSE
```

Backwards selection with `leaps`

```
bic_scores <- summary(models)$bic
bic_scores
```

```
##  [1]  -186.0198  -279.0113  -360.9948  -429.0414  -491.9711  -552.6699
##  [7]  -614.3485  -696.0323  -789.0044  -872.2932  -952.9270 -1057.7093
## [13] -1159.2405 -1155.8796 -1151.6375 -1147.1778 -1142.0944 -1136.8903
## [19] -1131.0153 -1124.9173 -1118.6714 -1112.3934 -1106.1547 -1100.0338
## [25] -1093.6837 -1087.3238 -1080.7822 -1073.9061 -1066.9540 -1060.2642
## [31] -1053.5735 -1047.0155 -1040.4737 -1034.2006 -1028.1425 -1022.7423
## [37] -1016.8925 -1010.8436 -1004.8660  -998.7483  -992.8478  -986.6821
## [43]  -980.4454  -973.7643  -967.0669  -959.9159  -952.8325  -945.6575
## [49]  -938.3867  -931.0882  -923.7410  -916.3736  -909.0185  -901.6542
## [55]  -894.1702  -886.6639  -879.0910  -871.4425  -863.7226  -855.9846
## [61]  -848.2431
```

Backwards selection with `leaps`



Backwards selection with `leaps`

```
which.min(bic_scores)
```

```
## [1] 13
```

```
summary(models)$outmat[13,]
```

```
##       Creative      Energetic         Tingly       Euphoric        Relaxed
##            "*"            "*"            "*"            "*"            "*"
##        Aroused          Happy       Uplifted         Hungry      Talkative
##            "*"            "*"            "*"            "*"            "*"
##         Giggly        Focused         Sleepy         Earthy          Sweet
##            "*"            "*"            "*"            " "            " "
##         Citrus        Flowery         Violet         Diesel `Spicy/Herbal`
##            " "            " "            " "            " "            " "
##           Sage          Woody        Apricot     Grapefruit         Orange
##            " "            " "            " "            " "            " "
##        Pungent          Grape           Pine          Skunk          Berry
##            " "            " "            " "            " "            " "
##         Pepper       Menthol           Blue         Cheese       Chemical
##            " "            " "            " "            " "            " "
##          Mango          Lemon          Peach        Vanilla          Nutty
##            " "            " "            " "            " "            " "
##       Chestnut            Tea        Tobacco       Tropical     Strawberry
```

Option 1: Best Subset Selection

Start with one variable that you think is best.

Add the *next best variable*.

Test whether it was worth adding.

Keep going until it's not worth adding any more variables.

Try it!

Open Activity-Variable-Selection

Determine the best model via backwards selection. Fit that model to the data and report results.

Determine the best model via forwards selection. Fit that model to the data and report results.