Model selection

Prof. Maria Tackett

Click here for PDF of slides

Topics

- Identifying modeling objectives
- Model selection processes
  - Forward selection
  - Backward selection

Which variables should be in the model?

- This is a very hard question that is the subject of a lot of statistical research
- There are many different opinions about how to answer this question
- This lecture will mostly focus on how to approach variable selection
  - We will introduce some specific methods, but there are many others out there

Which variables should you include?

- It depends on the goal of your analysis
- Though a variable selection procedure will select one set of variables for the model, that set is usually one of several equally good sets
- It is best to start with a well-defined purpose and question to help guide the variable selection

Prediction

- **Goal:** to calculate the most precise prediction of the response variable
- Interpreting coefficients is **not** important
- Choose only the variables that are strong predictors of the response variable
  - Excluding irrelevant variables can help reduce widths of the prediction intervals

One variable's effect

- **Goal:** Understand one variable's effect on the response after adjusting for other factors
- Only interpret the coefficient of the variable that is the focus of the study
    - Interpreting the coefficients of the other variables is **not** important
- Any variables not selected for the final model have still been adjusted for, since they had a chance to be in the model

Explanation

- **Goal:** Identify variables that are important in explaining variation in the response
- Interpret any variables of interest
- Include all variables you think are related to the response, even if they are not statistically significant
    - This improves the interpretation of the coefficients of interest
- Interpret the coefficients with caution, especially if there are problems with multicollinearity in the model

Example: SAT Averages by State

- This data set contains the average SAT score (out of 1600) and other variables that may be associated with SAT performance for each of the 50 U.S. states. The data is based on test takers for the 1982 exam.
- Response variable:
  - **`SAT`**: average total SAT score

Data comes from **`case1201`** data set in the **`Sleuth3`** package

SAT Averages: Predictors

- **`Takers`**: percentage of high school seniors who took exam
- **`Income`**: median income of families of test-takers ($ hundreds)
- **`Years`**: average number of years test-takers had formal education in social sciences, natural sciences, and humanities
- **`Public`**: percentage of test-takers who attended public high schools
- **`Expend`**: total state expenditure on high schools ($ hundreds per student)
- **`Rank`**: median percentile rank of test-takers within their high school classes

Suppose you are on a legislative watchdog committee, and you want to determine the impact of state expenditures on state SAT scores. You decide to build a regression model for this purpose. What is the primary modeling objective?

Understand one variable's effect

Suppose you are on a committee tasked with improving the average SAT scores for your state. You have already determined that the number of test takers is an important variable, so you decide to include it in the regression model. Now you want to know what other variables significantly impact the average SAT score after accounting for the number of test takers. What is the primary modeling objective?

Explanation

Model selection criteria

$$Adj.\,R^2 = 1 - \frac{SS_{Error}/(n-p-1)}{SS_{Total}/(n-1)}$$

$$AIC = n\log(SS_{\text{Error}}) - n\log(n) + 2(p+1)$$

$$BIC = n\log(SS_{\text{Error}}) - n\log(n) + log(n) \times (p+1)$$

Selection Process: Backward Selection

- Start with model that includes all variables of interest
- Drop variables one at a time that are deemed irrelevant based on some criterion. Common criterion include
  - Drop variable that results in the model with the highest Adj. R^2

    or
  - Drop variable that results in the model with the lowest value of AIC or BIC
- Stop when no more variables can be removed from the model based on the criterion

Selection Process: Forward Selection

- Start with the intercept-only model (i.e. model with no predictors)
- Include variables one at a time based on some criterion. Common criterion include
  - Add variable that results in the model with highest Adj. R^2 ***or***
  - Add variable that results in the model with the lowest value of AIC or BIC
- Stop when no more variables can be added to the model based on the criterion

Forward selection example

```
sat_scores <- Sleuth3::case1201 %>%
  select(-State)
```

```
int_only_model <- lm(SAT ~ 1, data = sat_scores)
```

```
full_model <- lm(SAT ~ ., data = sat_scores)
```

Step 1

```
add1(int_only_model, full_model, data = sat_scores)
```

```
## Single term additions
##
## Model:
## SAT ~ 1
##         Df Sum of Sq    RSS    AIC
## <none>               246011 427.06
## Takers   1    181024  64987 362.50
## Income   1     84038 161973 408.16
## Years    1     26948 219063 423.25
## Public   1      1589 244422 428.73
## Expend   1       973 245038 428.86
## Rank     1    190471  55539 354.64
```

Add **Rank** to the model.

Step 2

```
current_model <- lm(SAT ~ Rank, data = sat_scores)
```

```
add1(current_model, full_model, data = sat_scores)
```

```
## Single term additions
##
## Model:
## SAT ~ Rank
##         Df Sum of Sq   RSS    AIC
## <none>               55539 354.64
## Takers   1    1761.8 53778 355.03
## Income   1    4601.1 50938 352.32
## Years    1   17913.6 37626 337.17
## Public   1    3847.7 51692 353.05
## Expend   1    7671.0 47868 349.21
```

Add `Years` to the model

Step 3

```
current_model <- lm(SAT ~ Rank + Years, data = sat_scores)
```

```
add1(current_model, full_model, data = sat_scores)
```

```
## Single term additions
##
## Model:
## SAT ~ Rank + Years
##         Df Sum of Sq   RSS    AIC
## <none>               37626 337.17
## Takers  1     778.7 36847 338.13
## Income  1    2782.4 34843 335.33
## Public  1      37.0 37589 339.12
## Expend  1    5917.6 31708 330.62
```

Add **Expend** to the model.

Step 4

```
current_model <- lm(SAT ~ Rank + Years + Expend, data = sat_scores)
```

```
add1(current_model, full_model, data = sat_scores)
```

```
## Single term additions
##
## Model:
## SAT ~ Rank + Years + Expend
##         Df Sum of Sq   RSS    AIC
## <none>               31708 330.62
## Takers   1   1368.28 30340 330.41
## Income   1    848.47 30860 331.26
## Public   1   1462.46 30246 330.25
```

Add **`Public`** to the model.

Step 5

```
current_model <- lm(SAT ~ Rank + Years + Expend + Public, data = sat_scores)
```

```
add1(current_model, full_model, data = sat_scores)
```

```
## Single term additions
##
## Model:
## SAT ~ Rank + Years + Expend + Public
##         Df Sum of Sq   RSS    AIC
## <none>               30246 330.25
## Takers   1    401.32 29844 331.59
## Income   1     70.95 30175 332.14
```

🛑 Stop. We won't add any other variables to the model

Final model

term	estimate	std.error	statistic	p.value
(Intercept)	-204.598	117.687	-1.738	0.089
Rank	10.003	0.603	16.581	0.000
Years	21.890	6.037	3.626	0.001
Expend	2.242	0.678	3.305	0.002
Public	-0.664	0.450	-1.475	0.147

Try backward selection using the **`drop1`** function in R.

Recap

- Identifying modeling objectives
- Model selection processes
  - Forward selection
  - Backward selection