Model Selection

Dr. Maria Tackett

10.16.19

Click for PDF of slides

Announcements

- Team Feedback #2 **due Friday, Oct 18 at 11:59p**
- Project Proposal **due Wed, Oct 30 at 11:59p**
- Datathon **November 2 - 3**

Spring 2020 STA Courses

- STA 240: Probability for Statistics
- STA 250: Statistics (pre-req: multivariable calculus, STA 230/STA 240)
- STA 322: Study Design
- STA 323: Computing (pre-req: STA 230/STA 240, co-req: STA 250)
- STA 360: Bayesian and Modern Statistics (pre-req: multivariable calculus, linear algebra, STA 230/STA 240)

Today's Agenda

- Understanding Variance Inflation Factor (VIF)
- Model Selection

Understanding VIF

Detecting Multicollinearity (VIF)

- Variance Inflation Factor (VIF): Measure of multicollinearity

$$VIF(\hat{\beta}_j) = \frac{1}{1 - R^2_{X_j|X_{-j}}}$$

where $R^2_{X_j|X_{-j}}$ is the proportion of variation X that is explained by the linear combination of the other explanatory variables in the model.

- Typically $VIF > 10$ indicates concerning multicollinearity
- Use the `vif()` function in the `rms` package to calculate VIF

Example: Diamonds Data

- Recall the `diamonds` data in the ggplot2 package.
- Suppose we fit a model using `carat`, `color` and `clarity` to predict the price of a diamond.
- We will use the log-transformed price for this model.

```
price_mod <- lm(log(price) ~ carat + color + clarity, data = diamo
kable(tidy(price_mod), format = "html", digits = 3)
```

term	estimate	std.error	statistic	p.value
(Intercept)	5.404	0.014	394.919	0
carat	2.193	0.004	625.200	0
colorE	-0.057	0.005	-10.521	0
colorF	-0.054	0.005	-9.885	0
colorG	-0.129	0.005	-24.318	0
colorH	-0.263	0.006	-46.477	0
colorI	-0.419	0.006	-65.915	0
colorJ	-0.582	0.008	-74.291	0
clarityIF	1.052	0.015	69.790	0
claritySI1	0.742	0.013	57.477	0
claritySI2	0.557	0.013	42.884	0
clarityVS1	0.903	0.013	68.572	0
clarityVS2	0.838	0.013	64.708	0
clarityVVS1	0.965	0.014	69.304	0

VIF for diamonds model

```
library(rms)
tidy(vif(price_mod))
```

```
## # A tibble: 14 x 2
##    names           x
##    <chr>       <dbl>
##  1 carat        1.30
##  2 colorE       2.01
##  3 colorF       2.01
##  4 colorG       2.19
##  5 colorH       1.95
##  6 colorI       1.71
##  7 colorJ       1.42
##  8 clarityIF    3.42
##  9 claritySI1  14.3
## 10 claritySI2  11.2
## 11 clarityVS1  10.4
## 12 clarityVS2  13.8
## 13 clarityVVS1  5.74
## 14 clarityVVS2  7.32
```

- Which variables are highly correlated?
- How do you know?
- Why are these variables highly correlated?

You can explore the data in the **Model Selection** project on RStudio Cloud.

Model 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.
- **Data:** `case1201` data set in the `Sleuth3` package
- Response variable:
  - **SAT**: average total SAT score

SAT Averages: Explanatory Variables

- **`State`**: U.S. State
- **`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

Model Selection Practice

- Copy the **Model Selection** project on RStudio Cloud
- Complete Part I of the exercise with your lab group

Practice: What's the primary objective?

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?
  - One variable's effect
  - Prediction
  - Explanation
- What strategy would you use to select variables for the model?

Practice: What's the primary objective?

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?
  - One variable's effect
  - Prediction
  - Explanation
- What strategy would you use to select variables for the model?

Model Selection Criterion

- Akaike's Information Criterion (AIC):

$$AIC = n\log(RSS) - n\log(n) + 2(p+1)$$

- Schwarz's Bayesian Information Criterion (BIC):

$$BIC = n\log(RSS) - n\log(n) + log(n) \times (p+1)$$

See the supplemental note on AIC & BIC for derivations.

AIC & BIC

$$AIC = n\log(RSS) - n\log(n) + 2(p+1)$$
$$BIC = n\log(RSS) - n\log(n) + \log(n) \times (p+1)$$

- First Term: Decreases as p increases
- Second Term: Fixed for a given sample size n
- Third Term: Increases as p increases

Using AIC & BIC

$$AIC = n\log(RSS) - n\log(n) + 2(p+1)$$
$$BIC = n\log(RSS) - n\log(n) + \log(n) \times (p+1)$$

- Choose model with smallest AIC or BIC
- If $n \geq 8$, the penalty for BIC is larger than that of AIC, so BIC tends to favor *more parsimonious* models (i.e. models with fewer terms)

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 with highest p-value over some threshold (e.g. 0.05, 0.1)
  - Drop variable that leads to smallest value of AIC or BIC
- Stop when no more variables can be removed from the model based on the criterion

Forward Selection

- Start with the intercept-only model
- Include variables one at a time based on some criterion. Common criterion include
  - Add variable with smallest p-value under some threshold (e.g. 0.05, 0.1)
  - Add variable that leads to the smallest value of AIC or BIC
- Stop when no more variables can be added to the model based on the criterion

Stepwise Selection (Hybrid)

- Start with intercept-only model
- Conduct one forward step to potentially add a variable to the model based on some criterion
- Conduct one backward step to potentially remove a variable from the model based on some criterion
- Stop when no other variables can be added or removed from the model

Caution!

- Different automated selection methods may choose different models
- You may miss key transformations or interaction effects that are not selected by the automated procedure
- You may find models that have no scientific use, if automation rather than science is used to select model
- Standard errors for the coefficients are difficult to interpret, since there is additional variability from the model selection procedure that should also be accounted for

Model Selection in R

- Use **`step`** function for forward, backward, and stepwise selection using AIC as the selection criteria
- Use **`regsubsets`** function in the **leaps** package for forward, backward, and stepwise selection using BIC or Adj. R^2 as the selection criteria

`step` function (AIC)

```
null_model <- lm(Y ~ 1, data = my_data)
full_model <- lm(Y ~ ., data = my_data)
```

- Forward selection

```
regfit_forward <- step(null_model, scope = formula(full_model),
                       direction = "forward")
```

- Backward selection

```
regfit_backward <- step(full_model, direction = "backward")
```

- Stepwise (hybrid) selection

```
regfit_hybrid <- step(null_model, scope = formula(full_model),
                      direction = "both")
```

`regsubsets` function (BIC, Adj. R^2)

- Forward selection

```
regfit_forward <- regsubsets(Y ~ ., data = my_data,
                             method="forward")
```

- Backward selection

```
regfit_backward <- regsubsets(Y ~ ., data = my_data,
                              method="backward")
```

- Choose the best model:
  - Code shown for forward selection; use similar code for backward selection

```
sel_summary <- summary(regfit_forward)
coef(regfit_forward, which.max(sel_summary$adjr2)) # Adj R-sq
coef(regfit_forward, which.min(sel_summary$bic)) # BIC
```

Model Selection Practice

- Complete Part II of the application exercise in RStudio Cloud.