Modeling in Practice

Model Validation & Missing Data

Dr. Maria Tackett

11.18.19

Click for PDF of slides

Announcements

- Project Regression Analysis **due Wed, Nov 20 11:59p**
- Looking ahead:
  - Exam 02: Mon, Nov 25 in class
  - Exam review on Nov 20

Model Validation

Model Validation

- **Goal:** Want to find set of variables that give the lowest test (not training) error
    - Want a model that is generalizable, i.e. can be used to make predictions for new observations
- If we have a large data set, we can achieve this goal by randomly splitting the data into training and test (validation) sets
- Use the training set to fit a model, then use the fitted model to predict the responses for the observations in the test set
- Assess the error when applied to the test set and choose the model with the lowest error

Error

- **Quantitative response**: use Mean Square Error

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

- **Categorical response**: use misclassification rate

$$\text{Misclassification Rate} = \frac{1}{n}\sum_{i=1}^{n} Err_i = \frac{1}{n}\sum_{i=1}^{n} I(y_i \neq \hat{y}_i)$$

Training and test set

- There is no set split for the training and test sets. Common splits are
  - 50% training; 50% test
  - 80% training; 20% test
- Assigning observations to training and test sets:
  - Random assignment
  - Pick a certain time point to make split, if data is collected over time. Generally use earlier data in training and later data in test.
  - Use other relevant characteristic to make split

Cautions

- Be sure the training set is large enough to build a reliable model
    - The number of observations should be at least 10 times larger than the number of predictors
- Standard errors for model coefficients fit using training data are larger than standard errors if entire dataset was used
    - If the training set is reasonably large, then the difference in standard errors is small
- The test error is highly variable depending on which observations are in the test set

k-fold Cross Validation

- There are numerous validation methods that address the variability in testing error; we will focus on **k-fold cross validation**
  - More in-depth discussion of model validation in STA 325
- **k-fold Cross Validation**
  - Randomly split the data into k folds (typically 5 or 10)
  - Use $k - 1$ folds to fit a model (this is the training data)
  - Assess how well model predicts on remaining fold (this is the test data)
  - Repeat k using a different fold as the test set each time
- Calculate estimated testing error by average the k different error rates
- Once the variables for the final model have been selected, use the entire dataset to estimate coefficients for final model

5-fold Cross Validation in R

- Split data into 5 folds. Don't forget to set a seed!

```
library(modelr)
set.seed(04012019)
mydata_cv <- crossv_kfold(my.data, 5)
```

- Fit model on each training set

```
models <- map(mydata_cv$train, ~ lm(Y ~ X1 + ... + XP, data = .))
```

- Calculate MSE on each test set

```
# map2_dbl in purrr package that's loaded with tidyverse
test_mse <- map2_dbl(models, mydata_cv$test, mse)
```

Example: Advertising

We want to use spending on TV, radio, and newspaper advertising ($thousands) to predict total sales ($millions). The data contains the advertising and sales for 200 markets.

```
glimpse(advertising)
```

```
## Observations: 200
## Variables: 4
## $ tv        <dbl> 230.1, 44.5, 17.2, 151.5, 180.8, 8.7, 57.5, 120.2, 8.6
## $ radio     <dbl> 37.8, 39.3, 45.9, 41.3, 10.8, 48.9, 32.8, 19.6, 2.1, 2
## $ newspaper <dbl> 69.2, 45.1, 69.3, 58.5, 58.4, 75.0, 23.5, 11.6, 1.0, 2
## $ sales     <dbl> 22.1, 10.4, 9.3, 18.5, 12.9, 7.2, 11.8, 13.2, 4.8, 10.
```

We'll start by looking at the 5-fold cross validation results for the model using the predictors `radio` and `newspaper`

Adveritising: Split into 5 folds

```
set.seed(11182019)
ad_cv <- crossv_kfold(advertising, 5)
ad_cv
```

```
## # A tibble: 5 x 3
##   train        test         .id
##   <named list> <named list> <chr>
## 1 <resample>   <resample>   1
## 2 <resample>   <resample>   2
## 3 <resample>   <resample>   3
## 4 <resample>   <resample>   4
## 5 <resample>   <resample>   5
```

Advertising: Fit models

- Fit model on each training set

```
models <- map(ad_cv$train,
              ~ lm(sales ~ radio + newspaper, data = .))
models
```

```
## $`1`
##
## Call:
## lm(formula = sales ~ radio + newspaper, data = .)
##
## Coefficients:
## (Intercept)        radio    newspaper
##     9.07396      0.18314      0.01907
##
##
## $`2`
##
## Call:
## lm(formula = sales ~ radio + newspaper, data = .)
##
## Coefficients:
## (Intercept)        radio    newspaper
```

Advertising: Test error

```
test_mse <- map2_dbl(models, ad_cv$test, mse)
test_mse
```

```
##        1        2        3        4        5
## 16.20811 15.72228 14.17436 24.07450 23.80238
```

```
(error_model1 <- mean(test_mse))
```

```
## [1] 18.79633
```

Advertising

We will look at the cross-validation results for the model that includes `radio`, `newspaper`, and `tv` as predictors

```
models <- map(ad_cv$train,
              ~ lm(sales ~ radio + newspaper + tv, data = .))
```

```
test_mse <- map2_dbl(models, ad_cv$test, mse)
test_mse
```

```
##        1        2        3        4        5
## 2.907184 2.286179 1.999073 2.857006 4.381569
```

```
(error_model2 <- mean(test_mse))
```

```
## [1] 2.886202
```

Comparing Models

- The estimated testing error for
  - Model 1: `radio` and `newspaper`: 18.796325
  - Model 2: `radio`, `newspaper`, `tv`: 2.8862023
- Model 2 performs better than Model 1 when predicting the sales for new markets.

Missingness

What missing data looks like

```
##       age  bmi  hyp chl
## 1   20-39   NA <NA>  NA
## 2   40-59 22.7   no 187
## 3   20-39   NA   no 187
## 4   60-99   NA <NA>  NA
## 5   20-39 20.4   no 113
## 6   60-99   NA <NA> 184
## 7   20-39 22.5   no 118
## 8   20-39 30.1   no 187
## 9   40-59 22.0   no 238
## 10  40-59   NA <NA>  NA
```

Why is missing data an issue?

Do you have missingness in your data for the final project?

Why is missing data an issue when doing an analysis?

Dealing with missingness

- Deal with missing ness before doing any analysis
    - This is one of the many reasons exploratory data analysis is an important first step!
- Some things to consider if you find missing values:
    - Why are the values missing?
    - Is there a pattern of missingness? If so, what is it?
    - What is the proportion of missing values?
- The answers to these questions will help you determine how to deal with the missing data

Why are the values missing?

- Missing Completely at Random (MCAR): Missingness does not depend on the observed data or missing data, i.e. the probability of missing is the same for each observation
  - Example: People used a die to decide whether to share their income on a survey
- Missing at Random (MAR): Missingness depends on other observed variables but is random after conditioning on those variables, i.e. the probability that a variable is missing only depends on available information
  - Example: People with a college degree are less likely to share income than people without college degree
- Missing Not at Random (MNAR): Missingness dependents on the variable itself
  - Example: People with higher incomes are less likely to share them on a survey

How to deal with missing?

1. Only use observations with no missingness (complete-case analysis)
2. Only use variables with no missingness
3. Impute the missing values

Complete-case analysis

Use only complete observations in the analysis, i.e. those that have a value for each variable

What are potential disadvantages of dealing with missing data this way?

Complete-case analysis

- This may be OK if there are very few observations with missing values
- R does this automatically in its regression functions

Potential problems:

- Could result in a model being built on very few observations
  - This is especially true if there are many variables included in the model
  - Standard errors of model coefficients increase since you're losing information from the partially complete data
- If the observations with missingness differ systematically from the complete observations, then resulting analysis could be biased
  - This is especially true if the missingness is not random

Single Imputation

Single Imputation: Replace each missing value with a single number/category

- Mean imputation
- Use information from related observations
- Indicator variable for missingness
- Logical rule

Mean Imputation

- Replace missing values of a variable with the mean calculated from the observed data
- **Advantage:** Easy and straightforward method
- **Disadvantages:**
  - Can distort the distribution of the variable
  - Standard deviation underestimated
  - Results in inaccurate regression coefficients; relationships between variables become distorted

Related observations

- Replace the missing values using information from another observation that is "similar" to the one with missingness
- The "similar" observation can come from within the same dataset (hot deck) or from an external dataset (cold deck)
- Examples:
  - Hot Deck: Mother's income can be used to fill in missing values for father's income
  - Cold Deck: Use respondents from 2009 NHANES survey to fill in missing values for the 2011 NHANES survey
- **Disadvantage:** Could expand effects of measurement error

Indicator variable: categorical predictor

- Make "missing" an additional category for the variable
  - Use this updated variable in the regression model; "missing" becomes a term for the model

What can you conclude if the term for missing is significant in the model?

Indicator variable: quantitative predictor

- Impute the missing in the original variable using the mean (or some other method) and create a new indicator variable for the missingness
- Can lead to inaccurate estimates of the coefficients of other variables, since the slope is forced to be the same for the groups with and without missingness
- Reduce some of this bias by including interactions between the missing indicator and the other predictors

Logical Rule

- Can use some logical rule to impute missing values
- Example: The Social Indicators Survey includes a question on the "number of months worked in the previous year" which was answered by all 1501 respondents. Of the people who didn't answer the question about total earnings in the previous year, 10 reported working 0 months during the previous year.

For these 10 respondents, what is a logical value to use to impute their earnings?

How would you impute the earnings for the other respondents who didn't share their earnings?

Missing Data Exercise

- Copy the `Missing Data` project in Rstudio Cloud.

Acknowledgements

These slides draw material from

- Missing Data
- Handling Missing Data: An Introduction
- *Data Analysis Using Regression and Multilevel/Hierarchical Models*, "Chapter 25: Missing-data Imputation"