
Data Analytics

Yong Zheng

Illinois Institute of Technology
Chicago, IL, 60616, USA



School of Applied Technology
ILLINOIS INSTITUTE OF TECHNOLOGY

Review: Steps in Multiple Linear Regression



Multiple Linear Regression

Important Steps in Multiple Linear Regression

- Data Splits – build a model based on train set, and evaluate it based on the test set; hold-out or N-fold cross validation
- Determine x and y, examine their linear relationships
- Build a multiple linear regression model by parameter estimates → build diff models by using feature selection
- Goodness of fit test
- Residual analysis – the last step to tell your model is qualified
- Interpret the performance of the training process
- Evaluations and predictions – evaluate it based on test set

In-Class Practice

- Use the Case1_Student Grades_Regular.csv in case study 1
- Use finalGrade as dependent variable, and other numerical variables as the independent variables
- Use feature selection to build multiple models
- Use 75% as training, 25% as testing. Finally evaluate the models by using RMSE
- Identify the best model and explain it.



Data Splits for Evaluations

1). Hold-out Evaluation



If your data is large enough

Color	Weight (lbs)	Stripes	Tiger?
Orange	300	no	no
White	50	yes	no
Orange	490	yes	yes
White	510	yes	yes
Orange	490	no	no
White	450	no	no
Orange	40	no	no
Orange	200	yes	no
White	500	yes	yes
Green	560	yes	no
Orange	500	yes	?
White	50	yes	?

Training Data Set

Validation Data Set

Unseen data set

Example

```
mydata=read.table("clerical.txt",header=T)
mydata=mydata[sample(nrow(mydata)),]
select.data = sample (1:nrow(mydata), 0.8*nrow(mydata))
train.data = mydata[select.data,]
test.data = mydata[-select.data,]
```



Do not forget to shuffle the data



We use hold-out evaluation
For example. 80% as training





Data Splits for Evaluations

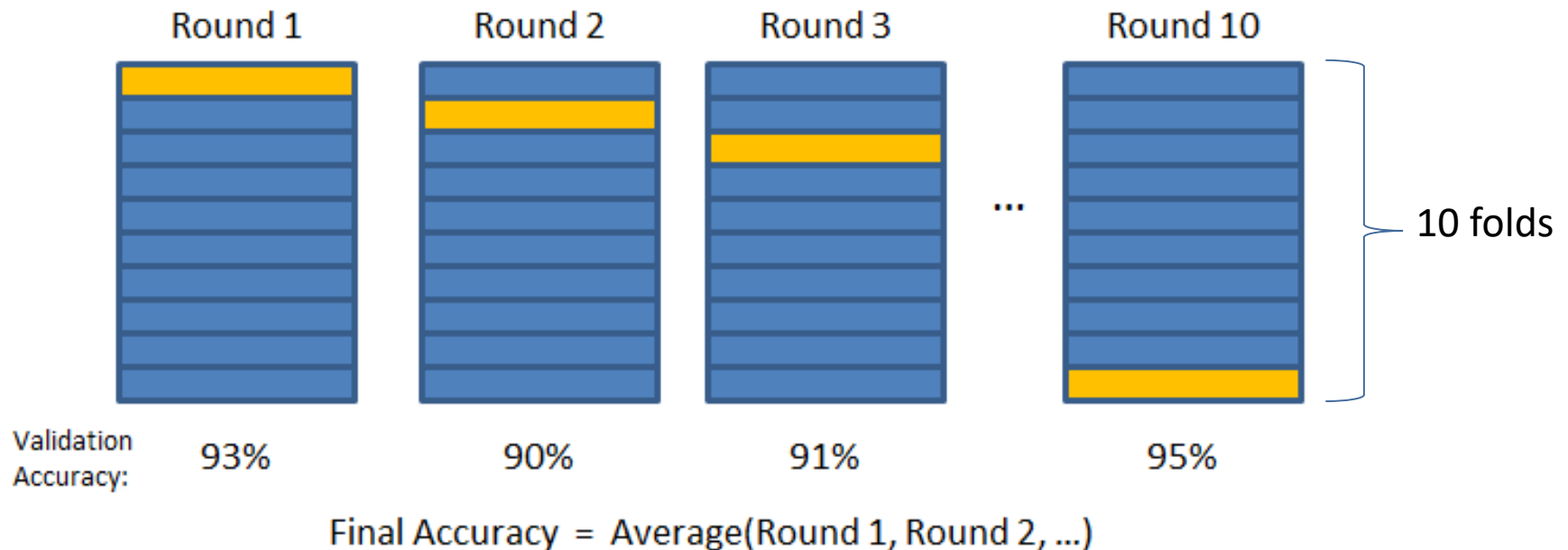
2). N-folds Cross Evaluation



If your data is relatively small

 Validation Set
 Training Set

Usually we choose N as 5 or 10



N-fold Cross Validation

- To use N-fold Cross validation, you need to build models first, and evaluate the model in a way of cross validation
- In linear regression, you need to use feature selections to figure out a list of possible models
- N-fold cross validation can be applied at the last step → model evaluations



Multiple Linear Regression

Important Steps in Multiple Linear Regression

- ~~Data Splits – build a model based on train set, and evaluate it based on the test set~~
- Determine linear relationship between y and x variables
- Build a multiple linear regression model by parameter estimates
- Goodness of fit test
- Residual analysis – the last step to tell your model is qualified
- Interpret the performance of the training process
- Evaluations and predictions – evaluate it based on test set



N-fold Cross validation

Example

We use the clerical example again

You can use the full data to examine

- Linear relationship between y and x
- Use feature selections to build models
- Examine whether they are qualified models or not



Example

Then we have different models

M3 = Backward selection by manually drop x based on p-value

M4 = Backward and Forward selection by step() based on AIC

M5 = Best Subset selection by Cp with least number of x variables

M3: `hours~cert+acc+change+check`

M4: `hours~cert+acc+change+check+misc`

M5: `hours~acc+check`

Example

Run 5-fold cross validation
`cv.glm()` in the package `boot`

```
> m3=glm(hours~cert+acc+change+check)
> m4=glm(hours~cert+acc+change+check+misc)
> m5=glm(hours~acc+check)
>
> mse3=cv.glm(mydata,m3,K=5)$delta
> mse4=cv.glm(mydata,m4,K=5)$delta
> mse5=cv.glm(mydata,m5,K=5)$delta
>
> mse3
[1] 137.1981 134.5955
> mse4
[1] 132.9957 129.9830
> mse5
[1] 168.4418 166.0293
```

You should build
models based on
`glm()` function

Raw MSE value

Adjusted MSE value

$$\text{MSE} = \text{RMSE}^2$$

lm() vs glm() in R

Lm function is used to build linear regression models and it has strict requirements about the linear relationship and the residuals

Glm function refers to Generalized function for linear regression models, it accepts a parameter “family” which is a description of the error distribution and link function to be used in the model.