Analytics Cycle in R

ITEC 621 W2

Best predictive model:

High accuracy (low error) on *new data*

Cross-validation:

The process of splitting up data in order to validate potential models against a set of held-out data. Tests how well a model will generalize to new data.

Training set

The data seen when estimating a model.

Test set

The data held-out during estimation.

Model Error

Difference between true data values and a model's predicted values: $y - \hat{y}$

Mean Squared Error (MSE): $\frac{1}{N} \sum_{n=1}^{N} (y_n - \hat{y}_n)^2$

MSE makes each errors positive before summing, which is important!

Goal

Find model with smallest MSE on training set. If model generalizes, then test MSE will be small too; otherwise model is

overfit.

Ordinary Least Squares

$$y = X \cdot \beta + \epsilon$$

y: Response variable (dependent variable)

X: Design matrix (independent variables)

β: Parameters (model)

ϵ or $y - X \cdot \beta$: Error

Generate random data

```
# Generate some random data
nsamples = 100
set.seed(1)
x1 = rnorm(nsamples, mean=1, sd=1)
x2 = rnorm(nsamples, mean=5, sd=5)
epsilon = rnorm(nsamples, mean=0, sd=0.3)
ones = matrix(1,nsamples,1)
X = cbind(ones, x1, x2)

# Here the true model for y is
# y = 1.6 + 3*x1 - 2*x2 + epsilon
b = c(1.6, 3, -2)
y = X %*% b + epsilon

# Prepping data frame
my_df = data.frame(cbind(y, x1, x2))
names(my_df) = c('y', 'x1', 'x2')
head(my_df)
```

```
            y        x1         x2
1  -0.9528741 0.3735462  1.8981666
2  -4.7635668 1.1836433  5.2105794
3   1.6783072 0.1643714  0.4453918
4  -2.2937177 2.5952808  5.7901439
5   1.4487991 1.3295078  1.7270768
6 -24.7849794 0.1795316 13.8364363
```

Ordinary Least Squares

```
my_model = lm(y ~ 1 + x1 + x2, data=my_df)
summary(my_model)
```

```
Call:
lm(formula = y ~ 1 + x1 + x2, data = my_df)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.88308 -0.13094  0.00061  0.19107  0.79182 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  1.617313   0.059072   27.38   <2e-16 ***
x1           3.006333   0.035026   85.83   <2e-16 ***
x2          -2.003208   0.006569 -304.96   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.313 on 97 degrees of freedom
Multiple R-squared:  0.999, Adjusted R-squared:  0.999 
F-statistic: 5.021e+04 on 2 and 97 DF,  p-value: < 2.2e-16
```

Calculating mean squared error

```
mean((y - X %*% my_model$coefficients)**2)
```

```
[1] 0.09504482
```

```
mean((my_model$residuals)**2)
```

```
[1] 0.09504482
```