Linear Regression

Practical Machine Learning (with R)

UC Berkeley

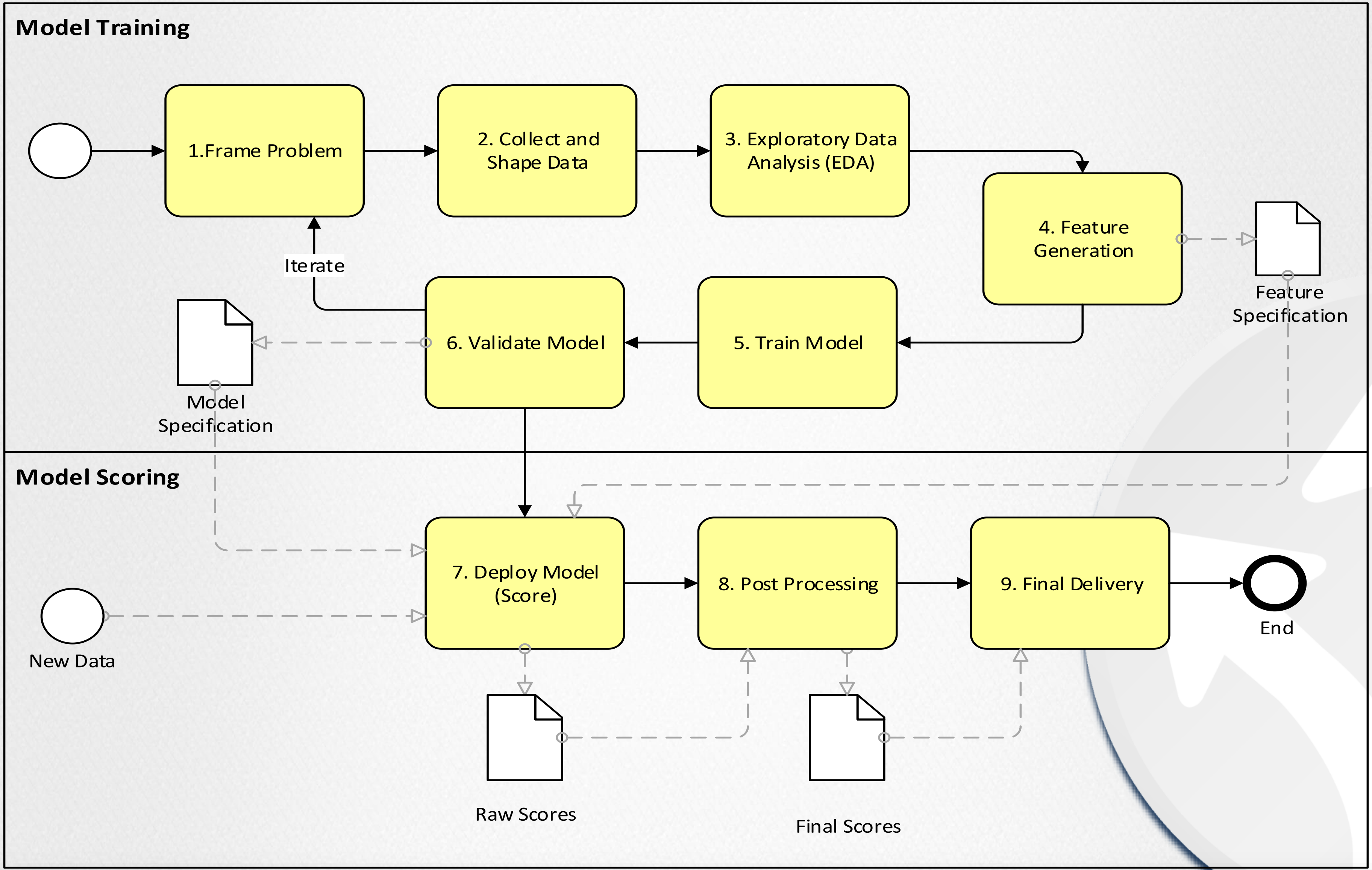
Expectations

- You know what `%>%` does ***and love it***

- You have worked with:
  - ***dplyr/tidyr*** and/or
  - ***data.table***

- Understand how to perform operations:
  - Single table: `select, filter, transform/mutate`
  - Multiple table: `join/merge`
  - …

Expectations: Process



Concepts

- Difference between
  - supervised and unsupervised models
  - semi-supervised
  - reinforcement / adaptive learning

- Difference between classification and regression

- Three components for ML algorithms …

3 Requirement for Algorithm

- A method for evaluating how well the algorithm performs (**ERRORS**)
- A restricted class of function (**MODEL**)
- A process for proceeding through the restricted class of functions to identify the functions (**SEARCH/OPTIMIZATION**)

Simple Linear Regression

Errors:

Model:

Search Optimization:

Strengths / Weaknesses (Limitations)

HOMEWORK SOLUTION

Clarification: Linear Regression Errors

- Two different types of errors measured
  - For ***fitting*** models
  - For **comparing** models

- Minimize square error loss (SSE) ***sum of squared errrors***

$$argmin_{\boldsymbol{\beta}}\left(\sum(\hat{y}-y)^2\right)$$

- *choose **β** such that the sum of squared errors is minimized.*

READING REVIEW APM

Example Predicting Medical Expenses

Questions:

- What is the goal?
- What tools were used?
- What are some alternatives?

Steps:

- Step 1: Understand Background
- Step 2: Set-up environment …
- Step 2: Collect/Load Data
- Step 3: Fix Data

- Step 2: Explore and Transform the Data
- Step 3: Train Model
- Step 4: Evaluating Model Performance

Step 1: Understand Problem

… And what you are delivering

Before modeling you should know: Who, When, What AND How

Who

- Client (org. and individual)
- User
- Is effected

When

- When is the solution required
- How often will this need to be revisited

What

- Goal(s)
  - Strategic: Impact(s) of goal
  - Operational:
- How will solution be judged?
- Data
  - Available Data
  - Ideal (additional) Data
- Existing solution/performance

How

- Will model be accessed (deployment)
- Goal mapped to solution
  - Method: algorithm(s) and evaluation

Step 2: Set-Up Environment (Code)

Depends on the deliverable

Types of deliverables:

Complexity

- Simple Answer (e.g. communicated via email)
- Report : [ProjectTemplate]
- R Package : [devtools]
- Model Package: caret, etc.
- Application : Shiny, OpenCPU/Deployer/Plumber

- More complicated solutions often require multiple deliverables

Step 2: Collect and Read Data

Ways to “read” data From

Files System

- `base::readLines`
- `utils::read.*`
- `readr::read_*`
- `foreign::read.*`

Web

- `utils::download.file`
- `httr::GET`

Database

- `DBI`
- `RODBC`

... Many Others

(Spark, Hadoop, SAP/Hana, Mongo, etc. ...)

Step 3: Fix Data (Outside → Inside)

- Change data to tbl or DT
- Standardize names (`lettercase`)
- Remove non-predictor/disallowed variables
- Coerce Types (`as.*`)

Step 4: Initial Tests

Identify response

- Does Y align with goal(s)
- Fix erroneous response value(s) as needed

Fit **Naïve** Model

simple estimate most often does not use preictors (X's)

- Measure Naïve Model

Fit **Simple** Model (one table model)

model of one or a few predictors often from same table or data source

- Measure: Simple Model
- Eval. perf. diff between simple and naïve models

Step 5: EDA

Explore Y

- Consider/try normalizing transformation
  - Match performance criteria
- Refit model

Explore Best-known X's

- Consider transformations
- Refit Model
- Does model meet performance criteria

FIT MODEL

lm.summary

```
Call:
lm(formula = expenses ~ ., data = insurance)

Residuals:
     Min       1Q   Median       3Q      Max
-11302.7  -2850.9   -979.6   1383.9  29981.7

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     -11941.6      987.8 -12.089  < 2e-16 ***
age                256.8       11.9  21.586  < 2e-16 ***
sexmale           -131.3      332.9  -0.395 0.693255
bmi                339.3       28.6  11.864  < 2e-16 ***
children           475.7      137.8   3.452 0.000574 ***
smokeryes        23847.5      413.1  57.723  < 2e-16 ***
regionnorthwest   -352.8      476.3  -0.741 0.458976
regionsoutheast  -1035.6      478.7  -2.163 0.030685 *
regionsouthwest   -959.3      477.9  -2.007 0.044921 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6062 on 1329 degrees of freedom
Multiple R-squared:  0.7509,  Adjusted R-squared:  0.7494
F-statistic: 500.9 on 8 and 1329 DF,  p-value: < 2.2e-16
```

1

2

3

LINEAR REGRESSION (INTUITION)

- Which is the more important variable?

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    51.3541     0.4593 111.814  < 2e-16 ***
EngDispl       -3.7454     0.2507 -14.941  < 2e-16 ***
NumCyl         -0.5880     0.1722  -3.414 0.000664 ***
```

- Coefficients … multiply then sum
- Number Line (in units of the response)
  - Start at intercept
  - Multiple term by value of the variable
  - Move those number of units of y.

Model Intuition

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ \ldots$$

But: $\hat{y} \neq y$

Data is generated by an unknown *stochastic* process:

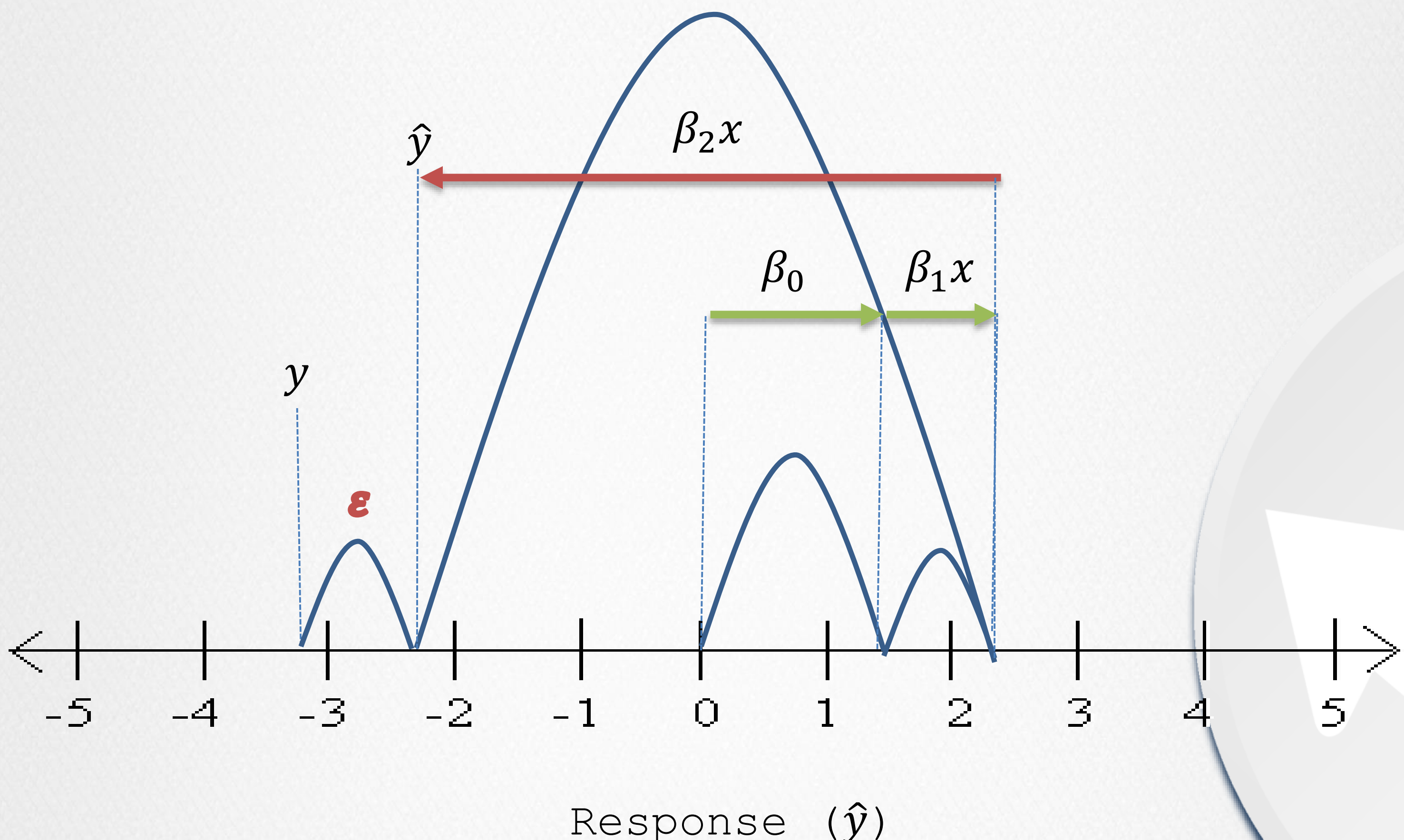
$$y = \hat{y} + \boldsymbol{\varepsilon}$$

Random Term

- Deterministic : always produces the same answer
- Stochastic: non-deterministic, contains some element of randomness, but not entirely random.

Linear Regression Number Line

$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2$



Linear Regression

- train a linear regression model
- Interpret linear regression model
  - "stars" (significance), `Estimate`, `Std. Error`, `R-squared`, `Pr(>|t|)`

```
Call:
lm(formula = FE ~ EngDispl, data = cars2010)

Residuals:
    Min      1Q  Median      3Q     Max
-14.486  -3.192  -0.365   2.671  27.215

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  50.5632     0.3985  126.89   <2e-16 ***
EngDispl     -4.5209     0.1065  -42.46   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.624 on 1105 degrees of freedom
Multiple R-squared:   0.62,	Adjusted R-squared:  0.6196
F-statistic:  1803 on 1 and 1105 DF,  p-value: < 2.2e-16
```

Linear Regression (Predictor Significance)

`Pr(>|t|) (p-value)`

Probability that there is NO relationship between the predictor and the response

- Is expressed as a probability.
- Lower is "better" i.e. more significant

Think of it (loosely) as the probability of the coefficient being wrong. It's an estimate after-all.

Indication of Bad Model Fit

These are signs of a bad model fit:

- No significant coefficients / predictors
- Many insignificant predictors
- Coefficients … too large or too small
- Low R-squared
- Skewed or non-zero centered residuals

Linear Regression Limitations

Limitation	Solution
Linear Response Does not fit higher order functions or interactions	• Transform data • Express in Model Formula
Insignificant Predictors Left in the Model	• Use model variant that does feature selection • Use Recursive Feature Elimination (RFE) routines
Sensitive to inputs: Outliers give out-sized influence on model fit	• Remove outliers • Transform Predictors • Use Robust Regression
Highly correlated predictors yield non-sensical models	• Use Regularization • RFE
Comparatively not sensitive	• ???

Transformations

- Centering and Scaling: `scale`*
- Resolve skewness: `log, sqrt, inv`
- Resolve outliers: spatial sign, `PCA`

Some algorithms require scaling

Some are insensitive

Time consuming

Somewhat of an art

- Genetic algorithms (GA)

Add complexity

Contribute to loss of interpretability

Model Formula

- DSL for expressing relationships between responses and predictors

- Specified by: `~`

```
response ~ predictor(s)
```

- Formula Components as interpreted by `lm`

```
Values: .
Operators: +, :, *
Functions: I
```

BEGIN ASSIGNMENT IN CLASS

APPENDIX

(MULTIPLE) LINEAR REGRESSION

Simple Linear Regression

Naïve Model

$$\hat{y} = mean(y)$$

Simple linear model:

$$\hat{y} = \beta_0 + \beta_1 x_1$$

LINEAR REGRESSION MODEL

→ Abstract to multiple dimensions

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ \dots$$

$$\hat{y} = \beta_0 + \sum_{i=1}^{p} \beta_i x_i$$

Mathy-r !!!