

Linear Regression

Practical Machine Learning (with R)

UC Berkeley

Fall 2016

Agenda

- ⇒ Administrative
 - Assignment (due 10/19 midnight)
- ⇒ Review
- ⇒ New Topics
 - Linear Regression
 - Model Formula
 - Model Process



QUESTIONS

Amount of time spent on:

- Reading?
 - Did you work along with the example?
- Assignment?

What have you found:

- to be difficult?
- surprising?
- General questions?



EXPECTATIONS

- ➔ Pulled changes from class Git Hub repository as of last Wednesday.
- ➔ Completed Reading
- ➔ Completed/Attempted exercises
 - `git pull/add/commit/push`

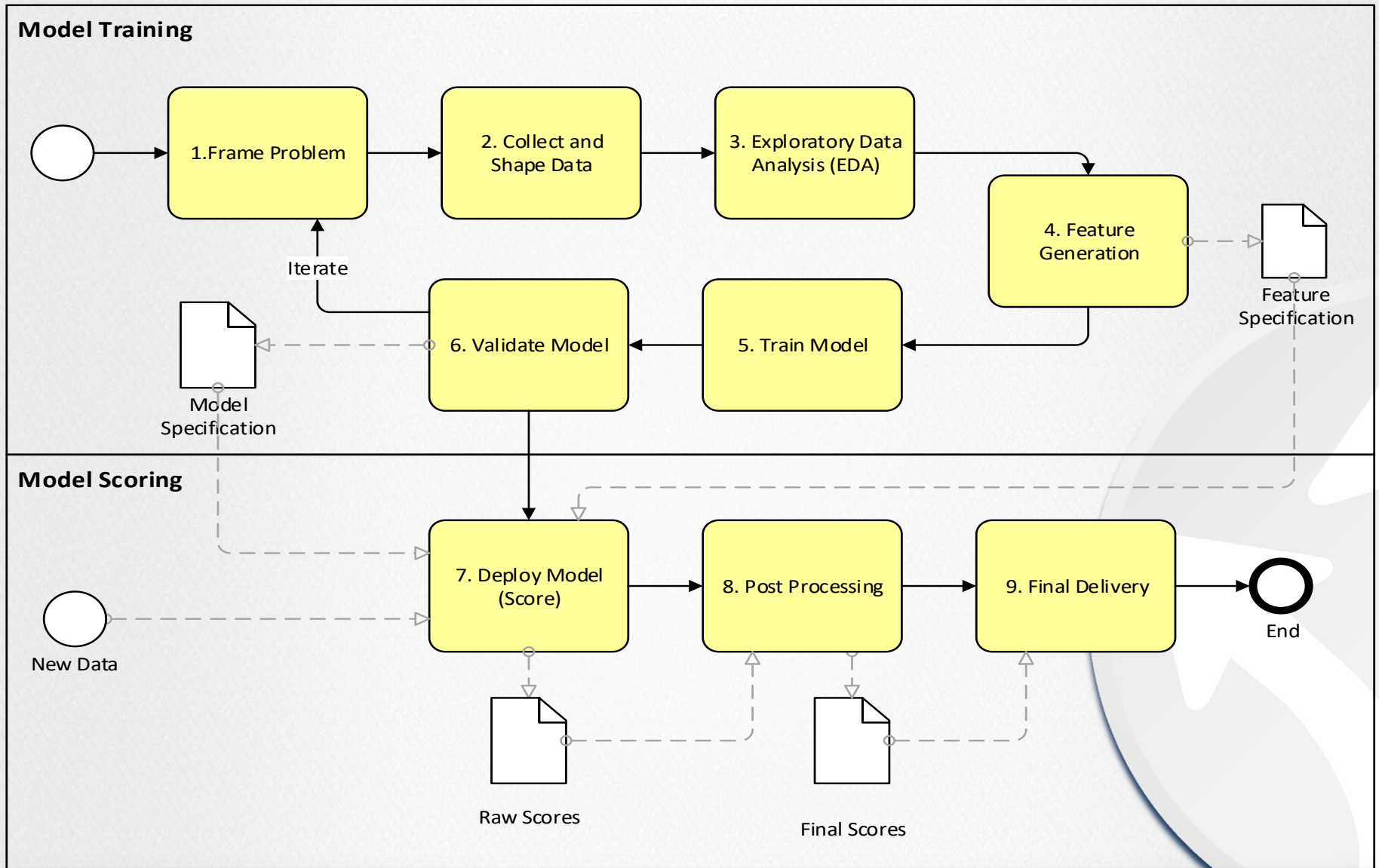


EXPECTATIONS

- ⇒ You have tried
 - ***dplyr/tidyr*** and/or
 - ***data.table***
- ⇒ You know what `%>%` does *and love it*
- ⇒ Understand basic table operations:
 - mutate
 - join
 - ...




Expectations: Process



REVIEW



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:

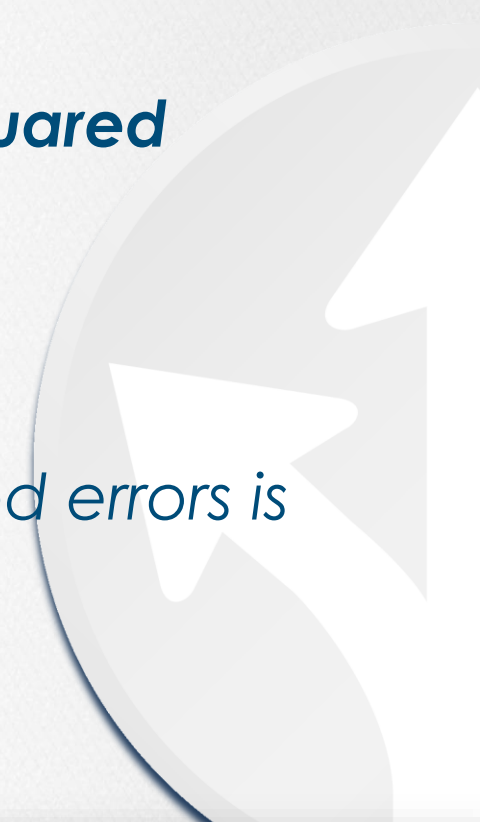


ERRATA: LINEAR REGRESSION ERRORS

- Two different types of errors measured
 - For ***fitting*** models
 - For ***comparing*** models
- Minimize square error loss (SSE) ***sum of squared errors***

$$\operatorname{argmin}_{\beta} \left(\sum (\hat{y} - y)^2 \right)$$

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



READING REVIEW



⇒ Questions on the Reading?



EXAMPLE PREDICTING MEDICAL EXPENSES

Questions:

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

Steps:

- ➔ Step 1: Collect/Load Data
- ➔ Step 2: Explore and Prep the Data
- ➔ Step 3: Train Model
- ➔ Step 4: Evaluating Model Performance
- ➔ Step 5: Improve the 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

1

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 *

2

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

3

HOMEWORK SOLUTION



(MULTIPLE) LINEAR REGRESSION



SIMPLE LINEAR REGRESSION

Naïve Model

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

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 !!!



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)

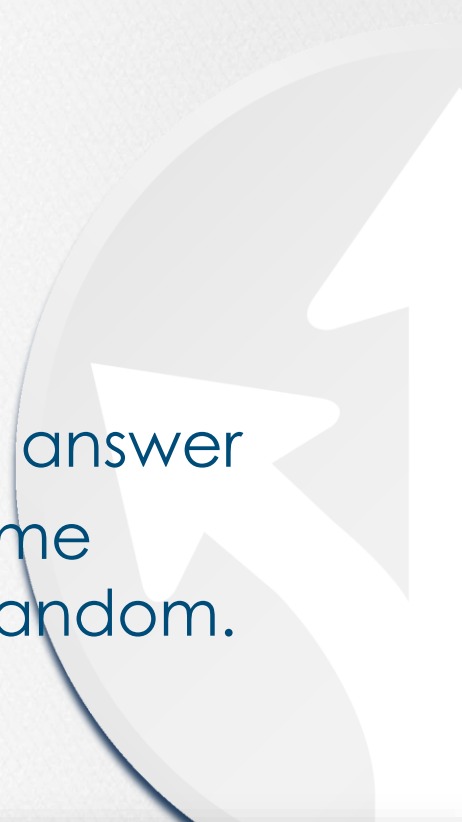
- Start at intercept
- Multiple term by value of the variable
- Move those number of units of y.

LINEAR REGRESSION (INTUITION)

- ⇒ Data is generated by an unknown stochastic process that the model creates the data, i.e. x 's

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

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



LINEAR REGRESSION NUMBER LINE



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)

$$\Pr (> | t |)$$

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	<ul style="list-style-type: none">• Transform data• Express in Model Formula
Insignificant Predictors Left in the Model	<ul style="list-style-type: none">• Use model variant that does feature selection• Use Recursive Feature Elimination (RFE) routines
Sensitive to inputs: Outliers give out-sized influence on model fit	<ul style="list-style-type: none">• Remove outliers• Transform Predictors• Use Robust Regression
Highly correlated predictors yield non-sensical models	<ul style="list-style-type: none">• Use Regularization• RFE
Comparatively not sensitive	<ul style="list-style-type: none">• ???

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)`
- ⇒ Special functions/operators:
 `., /, :, *, I`



BEGIN ASSIGNMENT IN CLASS



APPENDIX

