797ML Handbook

Steve Linberg

2022-04-03

Contents

Chapter 1

About

This book is being written as part of a final project for 797ML at UMass Amherst, spring 2022. It contains a simple reference and breakdown for a couple of dozen core methods used in machine learning.

The intent is twofold:

1. Serve as a reference for the basics of the material covered in the class, using language and examples that are as simple as possible to explain the core concepts and how to do them;
2. Force myself to learn these techniques better by carrying out the above.

The main purpose of this work is to be *simple*, not to be *comprehensive*. We won't cover every facet of every technique, or every possible permutations of outcomes. The goal is to simply express the broad strokes and core concepts in a way that can be easily remembered, and to serve as a jumping-off point when more information is needed.

1.1 Authoring guidelines

The goal is to have no more than a few short paragraphs for each section, and to keep each explanation of the meanings of outcome variables to one sentence each.

1.2 Resources

A lot of the material from this work is from the class textbook, James et al. [2021]. I also find UNC geneticist Josh Starmer's StatQuest video series on YouTube to be immensely helpful for simple explanations of statistics and machine learning concepts.

Chapter 2

Simple Linear Regression

2.1 TL;DR

What it does Looks to see how well a single predictor variable predicts an outcome, like *how well do years of education predict salary?*

When to do it When you want to see if pretty much the simplest possible model provides enough of an explanation of variance for your purposes

How to do it With the `lm()` function, among other ways

How to assess it Look for a significant p-value for the predictor, and a reasonable R^2

2.2 What it does

Simple linear regression is where it all begins; among the simplest of all of the regression techniques in analysis, which attempts to estimate a slope and an intercept line for a set of observations using a single predictor variable X and an output variable Y. It uses ordinary least squares (OLS) to build its model, looking for the line through the mean of X and Y that has the smallest sum of squares between the predicted and observed values.

The figure above shows a plot of a simple linear regression, attempting to use the variable `TV` to predict `Sales`. The blue line is the line defined by the regression with its Y intercept and slope; the red dots are the actual observations of `Sales` for each measure of the predictor `TV` on the X axis. The thin blue lines are the error in the prediction.

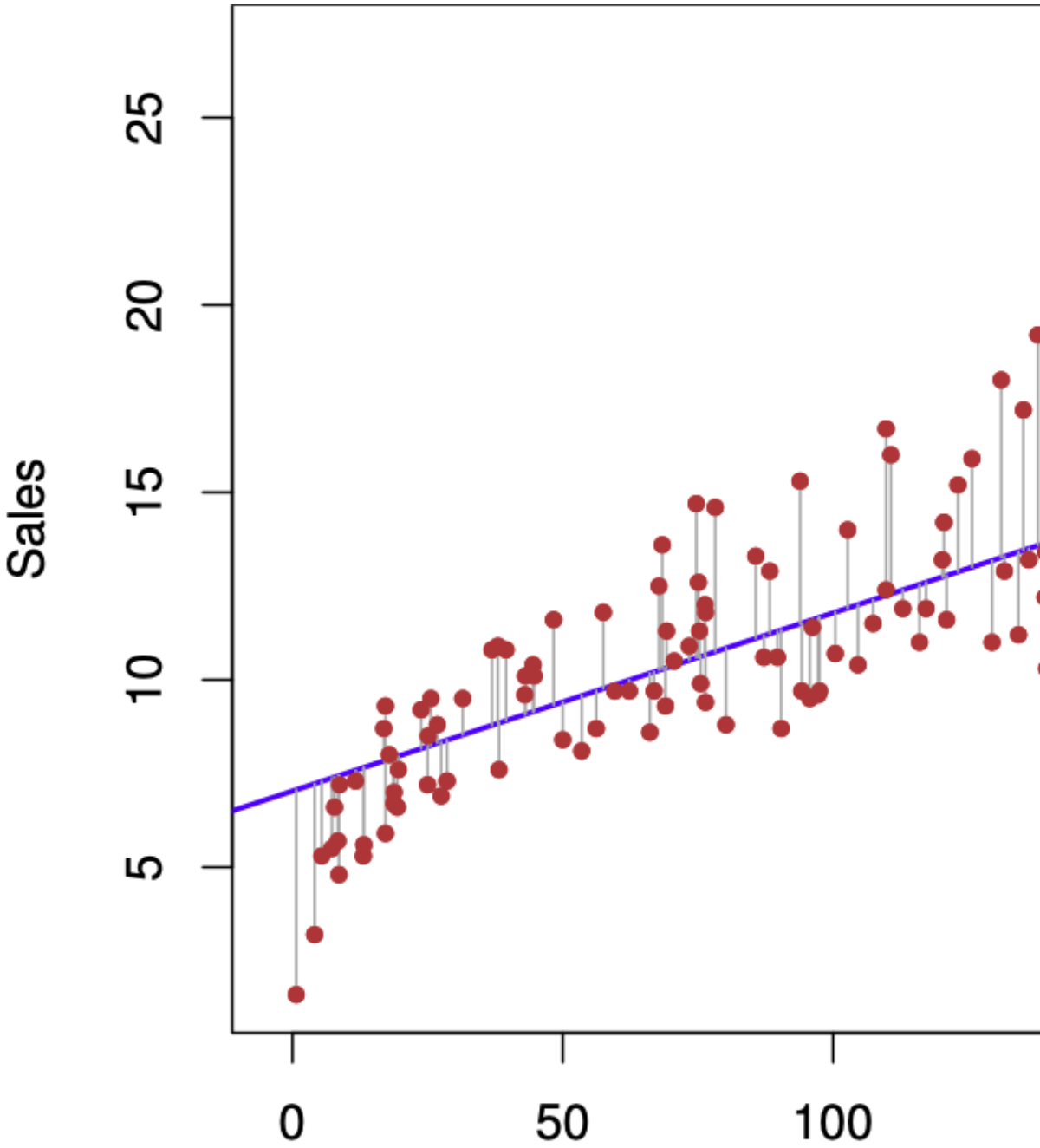


Figure 2.1: Simple linear regression plot (source: James et al. [2021], p. 62)

2.3 When to do it

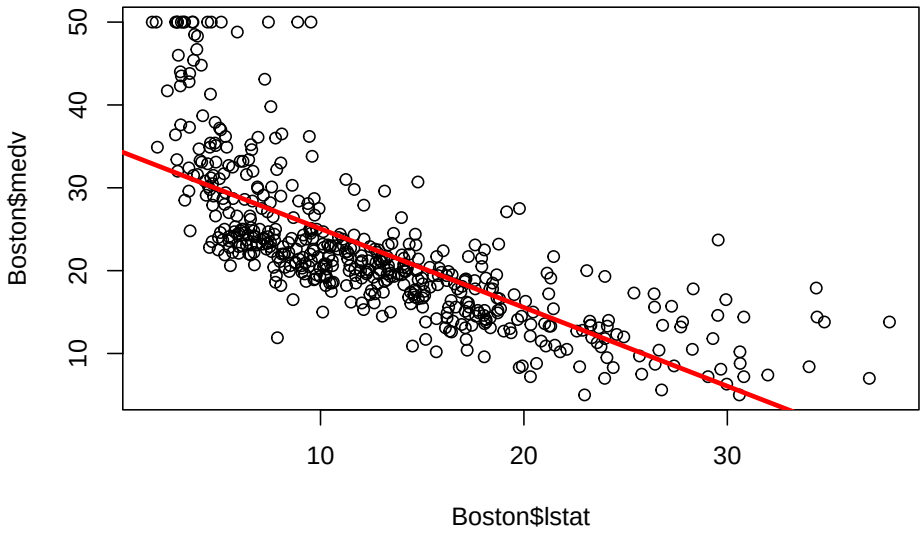
It is a simple first step for looking at data to see if there is an easy single-variable model that does a reasonable job predicting outcomes using one predictor variable. Sometimes, it can be good enough! It has the advantage of being easy to execute, to understand and to communicate, and the value of these factors should not be underestimated. Communicating with non-specialists is an important aspect of a data scientist's job.

Linear regression requires a dataset with a continuous outcome variable; it is easiest and most effective if the predictor variable is also numeric, whether continuous or discrete. It is possible to do linear regression with non-numeric predictors, such as true/false or ordered responses, by converting the predictors to a numeric scale.

2.4 How to do it

This example from James et al. [2021] shows a simple linear regression of **medv** onto **lstat**, attempting to predict median housing prices from percentage of "lower-status" population in the Boston data set from the late 1970s.

```
lm.fit <- lm(medv ~ lstat, data = Boston)
plot(Boston$lstat, Boston$medv)
abline(lm.fit, lwd = 3, col = "red")
```



The results of the regression are stored in the output of `lm()`, and may be viewed with `summary()`:

```
summary(lm.fit)
```

```
##
## Call:
## lm(formula = medv ~ lstat, data = Boston)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -15.168  -3.990  -1.318   2.034  24.500
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 34.55384    0.56263   61.41   <2e-16 ***
## lstat       -0.95005    0.03873  -24.53   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.216 on 504 degrees of freedom
## Multiple R-squared:  0.5441, Adjusted R-squared:  0.5432
## F-statistic: 601.6 on 1 and 504 DF,  p-value: < 2.2e-16
```

2.5 How to interpret the output

The output contains a lot of information, but the key points are:

- `Adjusted R-squared`: 0.5432 is the percentage of the variation in the model that is explained by the fit's prediction, compared to just looking at how much the observations vary from their own average with no predictor variable.
- `p-value`: < 2.2e-16 (basically zero) means that the model is (very) statistically significant, with a near-zero percent chance that the data in this set could have resulted from a random draw from the (unknown) population if there were no relationship between `medv` and `lstat`
- `Estimated intercept` of 34.55 is the Y intercept on the graph, or the estimated value of `medv` when `lstat` is zero
- `Estimated (lstat)`: -0.95 is the estimated coefficient of `lstat`, or the amount that `medv` changes by for each unit change in `lstat`.

The intercept and coefficient are β_0 and β_1, respectively, of the linear regression formula

$$Y = \beta_0 + \beta_1 X$$

Substituting the coefficients and variables, we transform this to:

$$\text{medv} = 34.55 - 0.95 \times \text{lstat}$$

The low p-value shows that there is (almost) definitely a relationship between `medv` and `lstat`, but the R^2 of 0.54 shows that the model only explains slightly more than half of the variation in the data. It's not a bad start, but we would probably want to find a model that explains more of it.

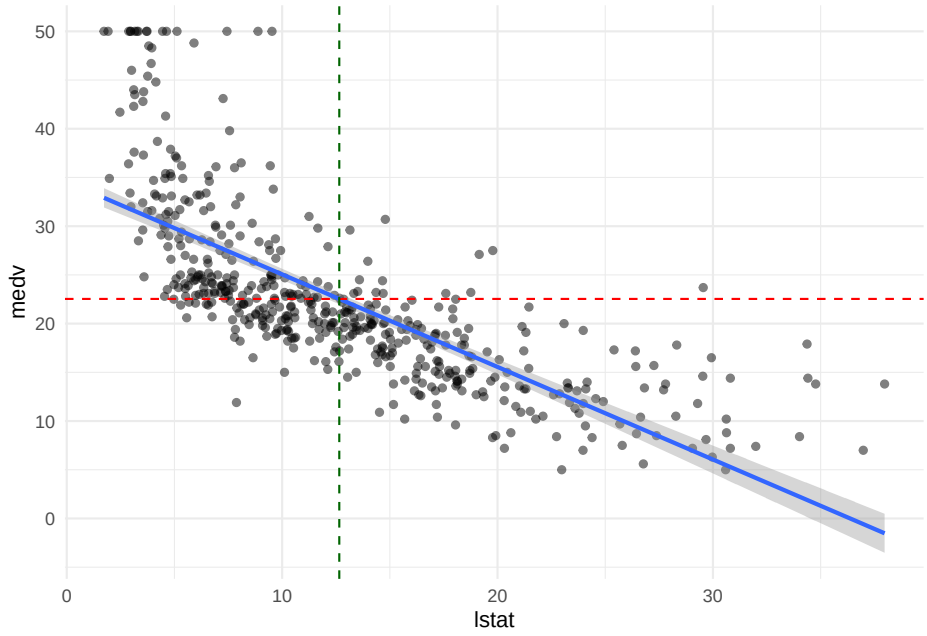
2.6 Where to learn more

- Chapter 3 - 3.1 in James et al. [2021]
- StatQuest: Linear Regression - this goes into good depth on the meaning of the F statistic as well, and how it used to calculate the p value.

2.7 Notes

You can also do a scatterplot in ggplot2 and add a regression line with `geom_smooth()`; it doesn't show the coefficients or other output information, but it gives a quick visual that's a little prettier than the base R plot:

```
ggplot(data = Boston, aes(x = lstat, y = medv)) +
  geom_point(alpha = 0.5) +
  geom_smooth(method = "lm", formula = y ~ x) +
  geom_hline(yintercept = mean(Boston$medv), linetype = "dashed", color = "red") +
  geom_vline(xintercept = mean(Boston$lstat), linetype = "dashed", color = "darkgreen") +
  theme_minimal()
```



The means of `lstat` and `medv` are shown with dashed red and green lines,

showing that the regression line goes through the center of the data.

Chapter 3

Multiple Linear Regression

3.1 TL;DR

What it does Looks to see how well multiple predictor variables predict an outcome, like *how well do years of education and age predict salary?*

When to do it When a simple linear regression doesn't provide a good enough explanation of variance, and you want to see if adding additional variables provides a better one

How to do it With the `lm()` function, utilizing more than one predictor

How to assess it Look for significant p-values for the predictors, and a reasonable adjusted-R^2

3.2 What it does

Multiple linear regression is the first natural extension of simple linear regression. It allows for more than one predictor variable to be specified. It is also possible to combine predictors in interactions, to find out if combinations of predictors have different effects than simply adding them to the model. XXX explain/demo

3.3 When to do it

Use multiple linear regression when a simple linear regression doesn't provide a good enough explanation of the variance you're observing, and you want to see if adding more predictors provides a better fit. Typically, this would be in response to either a low R^2 that leaves a lot of unexplained variance, or even just a visual conclusion drawn from seeing a plot of a linear model with an unsatisfactory regression line.

3.4 How to do it

The `lm()` function, using more than one predictor in the formula.

3.5 How to interpret the output

3.6 Where to learn more

Chapter 4

Logistic Regression

4.1 TL;DR

What it does :

When to do it :

How to do it :

How to assess it :

4.2 What it does

4.3 When to do it

4.4 How to do it

4.5 How to interpret the output

4.6 Where to learn more

Chapter 5

Multiple Logistic Regression

5.1 TL;DR

What it does :

When to do it :

How to do it :

How to assess it :

5.2 What it does

5.3 When to do it

5.4 How to do it

5.5 How to interpret the output

5.6 Where to learn more

Chapter 6

Linear Discriminant Analysis

6.1 TL;DR

What it does :

When to do it :

How to do it :

How to assess it :

6.2 What it does

6.3 When to do it

6.4 How to do it

6.5 How to interpret the output

6.6 Where to learn more

Chapter 7

Quadratic Discriminant Analysis

7.1 TL;DR

What it does :

When to do it :

How to do it :

How to assess it :

7.2 What it does

7.3 When to do it

7.4 How to do it

7.5 How to interpret the output

7.6 Where to learn more

Chapter 8

Naive Bayes

8.1 TL;DR

What it does :

When to do it :

How to do it :

How to assess it :

8.2 What it does

8.3 When to do it

8.4 How to do it

8.5 How to interpret the output

8.6 Where to learn more

Chapter 9

K-Nearest Neighbors

9.1 TL;DR

What it does :

When to do it :

How to do it :

How to assess it :

9.2 What it does

9.3 When to do it

9.4 How to do it

9.5 How to interpret the output

9.6 Where to learn more

Chapter 10

Poisson Regression

10.1 TL;DR

What it does :

When to do it :

How to do it :

How to assess it :

10.2 What it does

10.3 When to do it

10.4 How to do it

10.5 How to interpret the output

10.6 Where to learn more

Chapter 11

Cross-Validation

11.1 TL;DR

What it does :

When to do it :

How to do it :

How to assess it :

11.2 What it does

11.3 When to do it

11.4 How to do it

11.5 How to interpret the output

11.6 Where to learn more

Chapter 12

Bootstrap

12.1 TL;DR

What it does :

When to do it :

How to do it :

How to assess it :

12.2 What it does

12.3 When to do it

12.4 How to do it

12.5 How to interpret the output

12.6 Where to learn more

Chapter 13

Best Subset Selection

13.1 TL;DR

What it does :

When to do it :

How to do it :

How to assess it :

13.2 What it does

13.3 When to do it

13.4 How to do it

13.5 How to interpret the output

13.6 Where to learn more

Chapter 14

Stepwise Selection

14.1 TL;DR

What it does :

When to do it :

How to do it :

How to assess it :

14.2 What it does

14.3 When to do it

14.4 How to do it

14.5 How to interpret the output

14.6 Where to learn more

Chapter 15

Ridge Regression

15.1 TL;DR

What it does :

When to do it :

How to do it :

How to assess it :

15.2 What it does

15.3 When to do it

15.4 How to do it

15.5 How to interpret the output

15.6 Where to learn more

Chapter 16

Lasso

16.1 TL;DR

What it does :

When to do it :

How to do it :

How to assess it :

16.2 What it does

16.3 When to do it

16.4 How to do it

16.5 How to interpret the output

16.6 Where to learn more

Chapter 17

Principal Component Regression

17.1 TL;DR

What it does :

When to do it :

How to do it :

How to assess it :

17.2 What it does

17.3 When to do it

17.4 How to do it

17.5 How to interpret the output

17.6 Where to learn more

Chapter 18

Bagging

18.1 TL;DR

What it does :

When to do it :

How to do it :

How to assess it :

18.2 What it does

18.3 When to do it

18.4 How to do it

18.5 How to interpret the output

18.6 Where to learn more

Chapter 19

Random Forests

19.1 TL;DR

What it does :

When to do it :

How to do it :

How to assess it :

19.2 What it does

19.3 When to do it

19.4 How to do it

19.5 How to interpret the output

19.6 Where to learn more

Chapter 20

Boosting

20.1 TL;DR

What it does :

When to do it :

How to do it :

How to assess it :

20.2 What it does

20.3 When to do it

20.4 How to do it

20.5 How to interpret the output

20.6 Where to learn more

Chapter 21

Bayesian Additive Regression Trees

21.1 TL;DR

What it does :

When to do it :

How to do it :

How to assess it :

21.2 What it does

21.3 When to do it

21.4 How to do it

21.5 How to interpret the output

21.6 Where to learn more

Chapter 22

Support Vector Machines

22.1 TL;DR

What it does :

When to do it :

How to do it :

How to assess it :

22.2 What it does

22.3 When to do it

22.4 How to do it

22.5 How to interpret the output

22.6 Where to learn more

Chapter 23

Principal Component Analysis

23.1 TL;DR

What it does :

When to do it :

How to do it :

How to assess it :

23.2 What it does

23.3 When to do it

23.4 How to do it

23.5 How to interpret the output

23.6 Where to learn more

Chapter 24

K-Means Clustering

24.1 TL;DR

What it does :

When to do it :

How to do it :

How to assess it :

24.2 What it does

24.3 When to do it

24.4 How to do it

24.5 How to interpret the output

24.6 Where to learn more

Chapter 25

Hierarchical Clustering

25.1 TL;DR

What it does :

When to do it :

How to do it :

How to assess it :

25.2 What it does

25.3 When to do it

25.4 How to do it

25.5 How to interpret the output

25.6 Where to learn more

Bibliography

Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An Introduction to Statistical Learning*. Springer, 2nd edition, 2021. URL https://link.springer.com/book/10.1007/978-1-0716-1418-1. ISBN 978-1-0716-1417-4.