Ch 3.1: Linear Regression

Lecture 4 - CMSE 381

Prof. Elizabeth Munch

Michigan State University
::
Dept of Computational Mathematics, Science & Engineering

Weds Sep 6, 2023

Announcements

Last time:

- 2.2 Assessing Model Accuracy

Announcements:

- Office Hours

Covered in this lecture

- Least squares coefficient estimates for linear regression
- Residual sum of squares (RSS)
- Confidence interval, hypothesis test, and p-value for coefficient estimates
- Residual standard error (RSE)
- R squared

Section 1

Simple Linear Regression

Setup

- Predict Y on a single predictor variable X

$$Y \approx \beta_0 + \beta_1 X$$

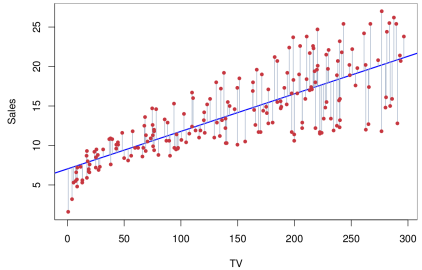
- "$\approx$" .... "is approximately modeled as"

Example

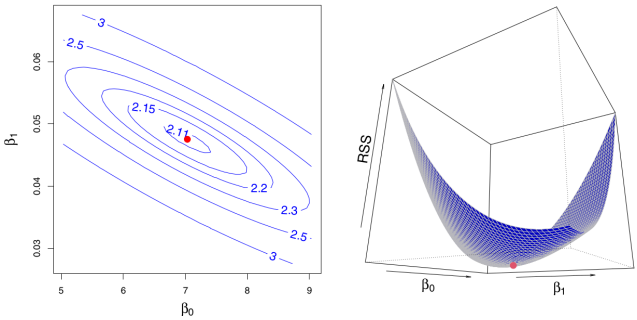
		TV	Radio	Newspaper	Sales
1	1	230.1	37.8	69.2	22.1
2	2	44.5	39.3	45.1	10.4
3	3	17.2	45.9	69.3	9.3
4	4	151.5	41.3	58.5	18.5
5	5	180.8	10.8	58.4	12.9
6	6	8.7	48.9	75	7.2
7	7	57.5	32.8	23.5	11.8
8	8	120.2	19.6	11.6	13.2
9	9	8.6	2.1	1	4.8
10	10	199.8	2.6	21.2	10.6
11	11	66.1	5.8	24.2	8.6

Least squares criterion: Setup

How do we estimate the coefficients?



Least squares criterion: RSS



Residual sum of squares RSS is

$$\begin{aligned} RSS &= e_1^2 + \cdots + e_n^2 \\ &= \sum_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 \end{aligned}$$

$$\text{sales} \approx \beta_0 + \beta_1 \text{TV}$$

Least squares criterion

Find β_0 and β_1 that minimize the RSS.

Least squares coefficient estimates

$$\min_{\beta_0,\beta_1} \sum_i (\hat{\beta}_0 + \hat{\beta}_1 x_i)^2$$

$$\frac{\partial RSS}{\partial \beta_0} = -2 \sum_i (y_i - \beta_0 - \beta_1 x_i) = 0$$

$$\frac{\partial RSS}{\partial \beta_1} = -2 \sum_i x_i (y_i - \beta_0 - \beta_1 x_i) = 0$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^n (x_i - \overline{x})^2}$$

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}$$

Coding group work

Section 2

Assessing Coefficient Estimate Accuracy

Bias in estimation

Analogy with mean

- Assume a true value μ^*
- An estimate from training data $\hat{\mu}$
- The estimate is unbiased if $E(\hat{\mu} = \mu^*)$

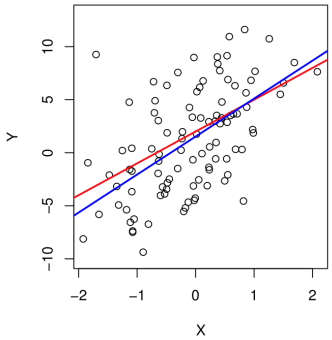
- Sample mean is unbiased for population mean:

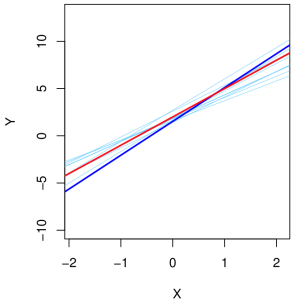
$$E(\hat{\mu}) = E\left(\frac{1}{n}\sum_i X_i\right) = \mu$$

- Standard variance estimate is biased

$$E(\hat{\sigma}^2) = E\left[\frac{1}{n}\sum_i (X_i - \overline{X})^2\right] \neq \sigma^2$$

Linear regression is unbiased





Variance in estimation

Continuing analogy with mean

- True value μ^*
- Estimate from training data $\hat{\mu}$
- Variance of sample mean
  $\mathrm{Var}(\hat{\mu}) = \mathrm{SE}(\hat{\mu})^2 = \frac{\sigma^2}{n}$

Variance of linear regression estimates

- Variance of linear regression estimates:

$$\mathrm{SE}(\hat{\beta}_0) = \sigma^2 \left[\frac{1}{n} + \frac{\overline{x}^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \right]$$

$$\mathrm{SE}(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

  where $\sigma^2 = \mathrm{Var}(\varepsilon)$

- Residual standard error is an estimate of σ

$$RSE = \sqrt{RSS/(n-2)}$$

Coding group work

Run the section titled “Simulating data”

Next time

Lec #	Date			Reading
1	Mon	Aug 28	Intro / First day stuff / Python Review Pt 1	1
2	Wed	Aug 30	What is statistical learning?	2.1
	Fri	Sep 1	Assessing Model Accuracy	2.2.1, 2.2.2
3	Mon	Sep 4	No class - Labor day	
4	Wed	Sep 6	Linear Regression	3.1
5	Fri	Sep 8	More Linear Regression	3.1/3.2
6	Mon	Sep 11	Even more linear regression	3.2.2
7	Wed	Sep 13	Probably more linear regression	3.3
8	Fri	Sep 15	Intro to classification, Logisitic Regression	2.2.3, 4.1, 4.2, 4.3
9	Mon	Sep 18	More logistic regression	
10	Wed	Sep 20	Multiple Logistic Regression / Multinomial Logistic Regression	
11	Fri	Sep 22	Overflow/Project day?	
	Mon	Sep 25	***Review***	
	Wed	Sep 27	**Midterm #1**	
	Fri	Sep 29	No class - Dr Munch out of town	
12	Mon	Oct 2	Leave one out CV	5.1.1, 5.1.2
13	Wed	Oct 4	k-fold CV	5.1.3
14	Fri	Oct 6	More k-fold CV,	5.1.4-5
15	Mon	Oct 9	k-fold CV for classification	5.1.5
16	Wed	Oct 11	Resampling methods: Bootstrap	5.2
17	Fri	Oct 13	Subset selection	6.1

Announcements

- We had a quiz last time!
- Homework 2
  - Due Mon, Sep 11
  - Need to upload individual file for EACH QUESTION