Ch 3.1: Linear Regression

Lecture 4 - CMSE 381

Prof. Elizabeth Munch

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

Fri, Sep 9, 2022

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

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

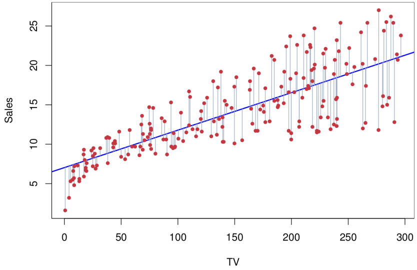
- "$\approx$" .... "is approximately modeled as"

Example

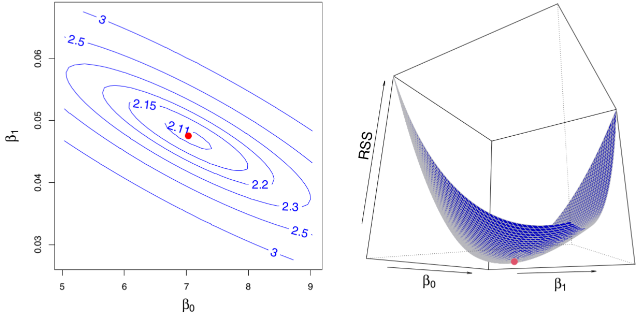
	TV	Radio	Newspaper	Sales
1	230.1	37.8	69.2	22.1
2	44.5	39.3	45.1	10.4
3	17.2	45.9	69.3	9.3
4	151.5	41.3	58.5	18.5
5	180.8	10.8	58.4	12.9
6	8.7	48.9	75	7.2
7	57.5	32.8	23.5	11.8
8	120.2	19.6	11.6	13.2
9	8.6	2.1	1	4.8
10	199.8	2.6	21.2	10.6
11	66.1	5.8	24.2	8.6

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 $E(\hat{\mu} = \mu^*)$

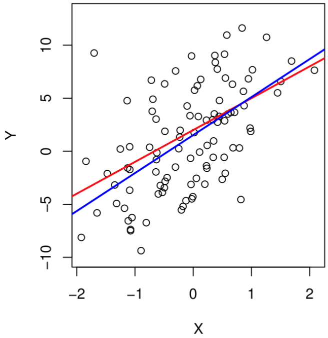
- Sample mean is unbiased for population mean:

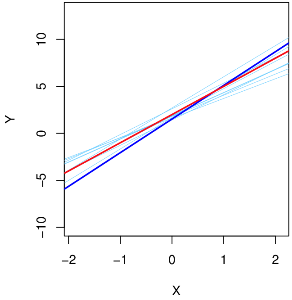
$$E(\hat{\mu}) = E\left(\frac{1}{n}\sum_i X_i\right) = \mu$$

- 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





Coding group work

Run the section titled "Simulating data"

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

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

Confidence Interval

The 95% confidence interval for β_1 approximately takes the form

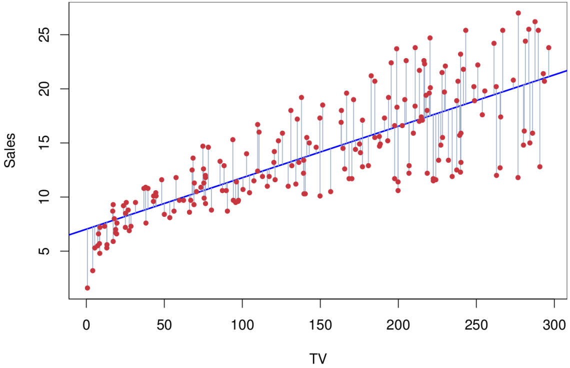
$$\hat{\beta}_1 \pm 2 \cdot \mathrm{SE}(\hat{\beta}_1)$$

Interpretation:
There is approximately a 95% chance that the interval

$$\left[\hat{\beta}_1 - 2 \cdot \mathrm{SE}(\hat{\beta}_1), \hat{\beta}_1 + 2 \cdot \mathrm{SE}(\hat{\beta}_1)\right]$$

will contain β_1 where we repeatedly approximate $\hat{\beta}_1$ using repeated samples.

CI in Advertising data



For the advertising data set, the 95% CIs are:

- β_1 :: $[0.042, 0.053]$
- β_0 :: $[6.130, 7.935]$

Hypothesis testing

H_0: There is no relationship between X and Y

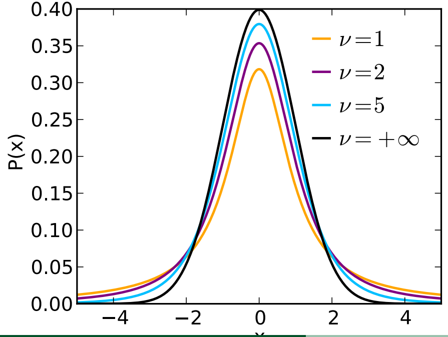
H_1: There is some relationship between X and Y

Test statistic and p-value

Test statistic:

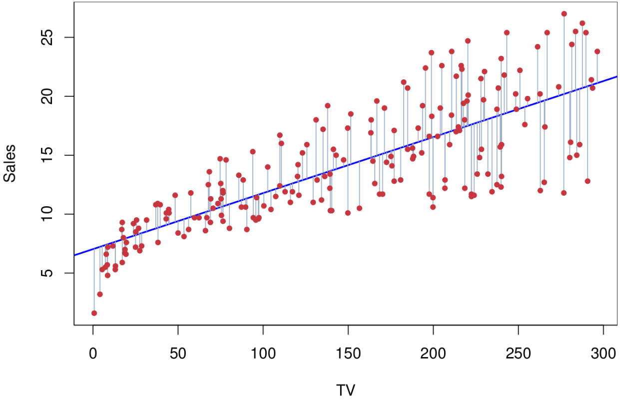
$$t = \frac{\hat{\beta}_1 - 0}{\mathrm{SE}(\hat{\beta}_1)}$$

t-distribution with $n-2$ degrees of freedom



Advertising example

	Coefficient	Std. error	t-statistic	p-value
`Intercept`	7.0325	0.4578	15.36	< 0.0001
`TV`	0.0475	0.0027	17.67	< 0.0001



Assessing the accuracy of the module: RSE

Residual standard error (RSE):

$$
\begin{aligned}
RSE &= \sqrt{\frac{1}{n-2} RSS} \\
&= \sqrt{\frac{1}{n-2} \sum_i (y_i - \hat{y}_i)^2}
\end{aligned}
$$

Assessing the accuracy of the module: R^2

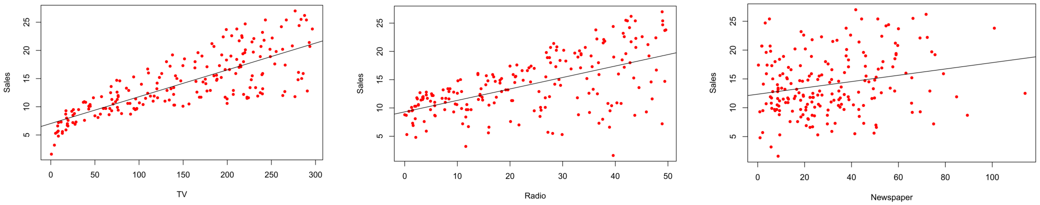
R squared:

$$R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS}$$

where total sum of squares is

$$TSS = \sum_i (y_i - \overline{y})^2$$

Advertising example



$R^2 = 0.61$ $R^2 = 0.33$ $R^2 = 0.05$

Coding group work

Run the section titled "Assessing Coefficient Estimate Accuracy"

Next time

Lec #		Date	Topic	Reading	Homeworks
1	W	Aug 31	Intro / First day stuff / Python Review Pt 1	1	
2	F	Sep 2	What is statistical learning? / Python Review Pt 2	2.1	
	M	Sep 5	No class - Labor day		
3	W	Sep 7	Assessing Model Accuracy	2.2	HW #1 Due
4	F	Sep 9	Linear Regression	3.1	
5	M	Sep 12	More Linear Regression	3.2	
6	W	Sep 14	Even more linear regression	3.3	HW #2 Due
7	F	Sep 16	Probably more linear regression		
8	M	Sep 19	Intro to classification, Logisitic Regression	4.1, 4.2, 4.3	
9	W	Sep 21	More logistic regression		HW #3 Due
10	F	Sep 23	Review		
11	M	Sep 26	***Midterm #1***		
12	W	Sep 28	[No class, Dr Munch out of town]		
13	F	Sep 30	[No class, Dr Munch out of town]		
14	M	Oct 3	Leave one out CV	5.1.1, 5.1.2	
15	W	Oct 5	k-fold CV	5.1.3	
16	F	Oct 7	More k-fold CV	5.1.4	
17	M	Oct 10	CV for classification	5.1.5	HW #4 Due
18	W	Oct 12	Resampling methods: Bootstrap	5.2	
19	F	Oct 14	Subset selection	6.1	
20	M	Oct 17	Shrinkage: Ridge	6.2.1	HW #5 Due
21	W	Oct 19	Shrinkage: Lasso	6.2.2	
22	F	Oct 21	Dimension Reduction	6.3	

Announcements

- We had a quiz last time!
- Homework 2
  - NEW: Upload to crowdmark
  - Due Weds, Sep 14
  - Need to upload individual file for EACH QUESTION