Ch 3.1-2: (Multi)-Linear Regression

Lecture 5 - CMSE 381

Prof. Elizabeth Munch

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

Mon, Sep 12, 2022

Announcements

Last time:

- Started 3.1 - Single linear regression

Announcements:

- Office Hours: Tues - Fri
- Homework #2 Due Weds, Sep 14
  - Upload solutions to CROWDMARK
  - Look for an email with the link to do this
  - Upload file for each question separately
- Quizzes
  - Approximately once a week
  - Look for email from crowdmark for feedback
  - Grades will be updated on D2L

Covered in this lecture

- Confidence interval, hypothesis test, and p-value for coefficient estimates
- Residual standard error (RSE)
- R squared
- Setup for multiple linear regression

Section 1

Last time

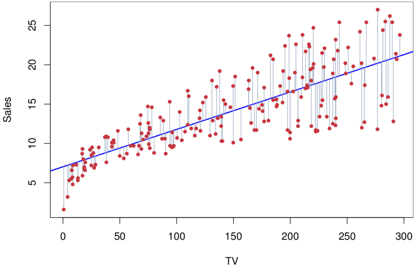
Setup

- Predict Y on a single predictor variable X

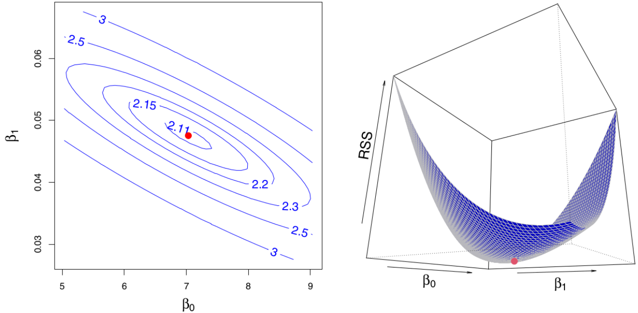
$$Y \approx \beta_0 + \beta_1 X$$

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

- Given $(x_1, y_1), \cdots, (x_n, y_n)$
- Let $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ be prediction for Y on ith value of X.
- $e_i = y_i - \hat{y}_i$ is the ith residual



Least squares criterion: RSS



Residual sum of squares RSS is

$$
\begin{aligned}
RSS &= e_1^2 + \cdots + e_n^2 \\
&= \sum_i (y_i - \hat\beta_0 - \hat\beta_1 x_i)^2
\end{aligned}
$$

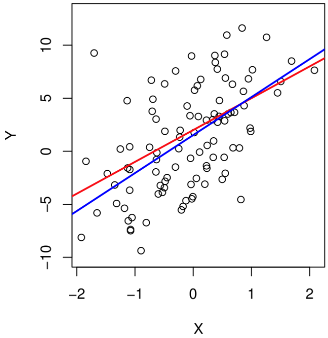
Least squares criterion

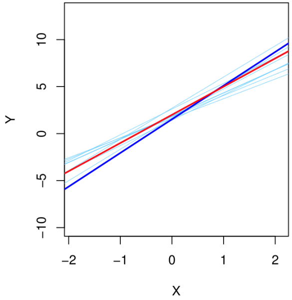
Find β_0 and β_1 that minimize the RSS.

$$\hat\beta_1 = \frac{\sum_{i=1}^n (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^n (x_i - \overline{x})^2}$$

$$\hat\beta_0 = \overline{y} - \hat\beta_1 \overline{x}$$

Linear regression is unbiased





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

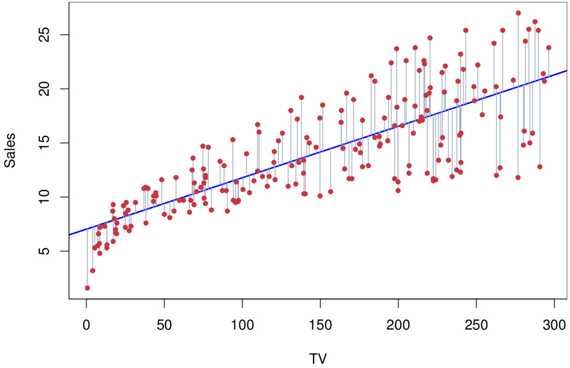
$$\hat{\beta}_1 \pm 2 \cdot \mathrm{SE}(\hat{\beta}_1)$$

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 :: $[6.130, 7.935]$

Section 2

New stuff on evaluating models

Hypothesis testing

H_0: There is no relationship between X and Y

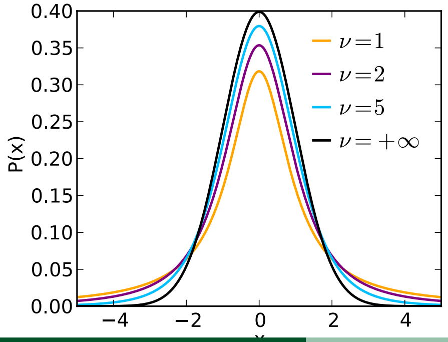
H_1: There is some relationship between X and Y

Test statistic and p-value

Test statistic:

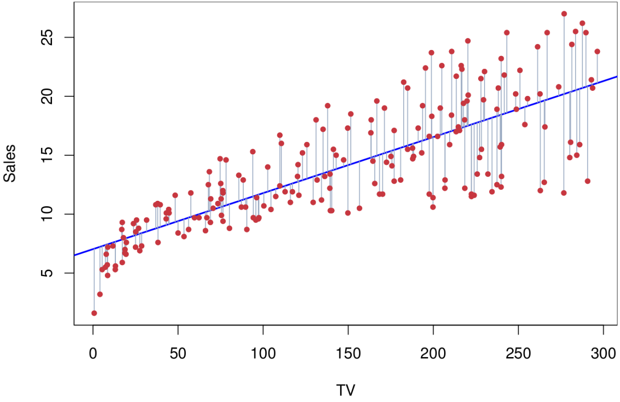
$$t = \frac{\hat{\beta}_1 - 0}{\mathrm{SE}(\hat{\beta}_1)}$$

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):

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

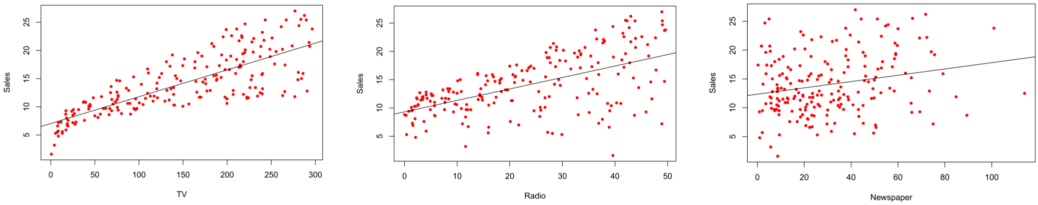
R squared:

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

where total sum of squares is

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

Advertising example



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

Coding group work

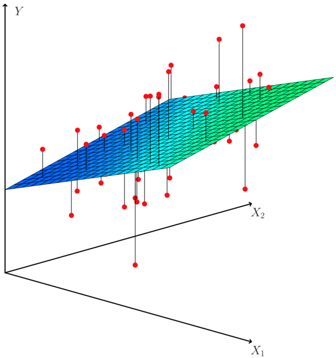
Run the section titled "Assessing Coefficient Estimate Accuracy"

Section 3

Multiple Linear Regression

Setup

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots \beta_p x_p + \varepsilon$$



Estimation and Prediction

Given estimates $\hat\beta_0, \hat\beta_1, \hat\beta_2, \cdots, \hat\beta_p$, prediction is

$$\hat y = \hat\beta_0 + \hat\beta_1 x_1 + \cdots + \hat\beta_p x_p$$

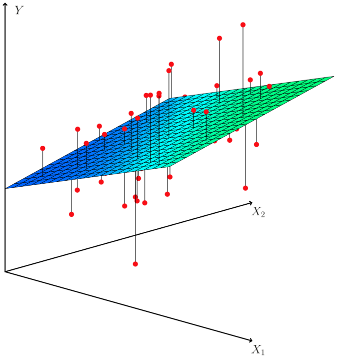
Minimize the sum of squares

$$\begin{aligned} RSS &= \sum_i (y_i - \hat y_i)^2 \\ &= \sum (y_i - \hat\beta_0 - \hat\beta_1 x_1 - \cdots - \hat\beta_p x_p) \end{aligned}$$

Coefficients are closed form but UGLY

Advertising data set example

$$\texttt{Sales} = \beta_0 + \beta_1 \cdot \texttt{TV} + \beta_2 \cdot \texttt{radio} + \beta_3 \cdot \texttt{newspaper}$$



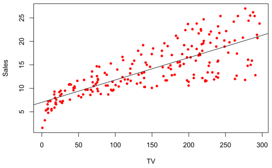
	Coefficient
Intercept	2.939
TV	0.046
radio	0.189
newspaper	−0.001

Interpretation of coefficients

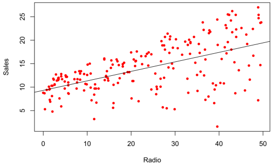
$$\texttt{Sales} = \beta_0 + \beta_1 \cdot \texttt{TV} + \beta_2 \cdot \texttt{radio} + \beta_3 \cdot \texttt{newspaper}$$

	Coefficient
Intercept	2.939
TV	0.046
radio	0.189
newspaper	−0.001

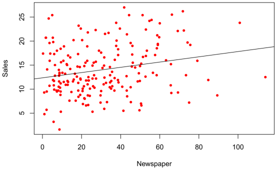
Single regression vs multi-regression



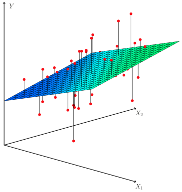
	Coefficient
Intercept	7.0325
TV	0.0475



	Coefficient
Intercept	9.312
radio	0.203



	Coefficient
Intercept	12.351
newspaper	0.055



	Coefficient
Intercept	2.939
TV	0.046
radio	0.189
newspaper	−0.001

Correlation matrix

	TV	radio	newspaper	sales
TV	1.0000	0.0548	0.0567	0.7822
radio		1.0000	0.3541	0.5762
newspaper			1.0000	0.2283
sales				1.0000

Coding group work

Run the section titled "Multiple Linear Regression"

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	