

Ch 3.1: More Linear Regression

Lecture 5 - CMSE 381

Prof. Mengsen Zhang

Michigan State University

::

Dept of Computational Mathematics, Science & Engineering

Fri, Sep 5, 2025

Last time:

- Started 3.1 - Simple linear regression (least squares)

Announcements:

- Homework #1 Due Sun, Sep 7
- Homework #2 Due Sun, Sep 14

Covered in this lecture

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

Section 1

Last time

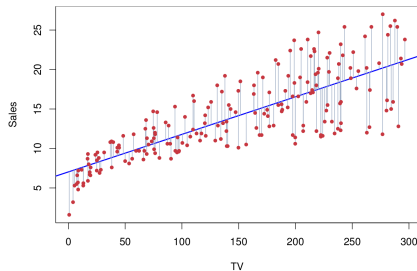
Setup

- Predict Y on a single predictor variable X

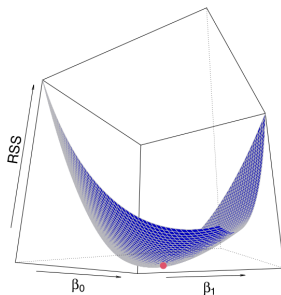
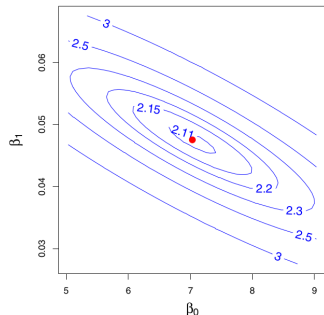
$$Y \approx \beta_0 + \beta_1 X$$

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

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

$$\begin{aligned} \hat{\beta}_1 &= \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \\ \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x} \end{aligned}$$

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

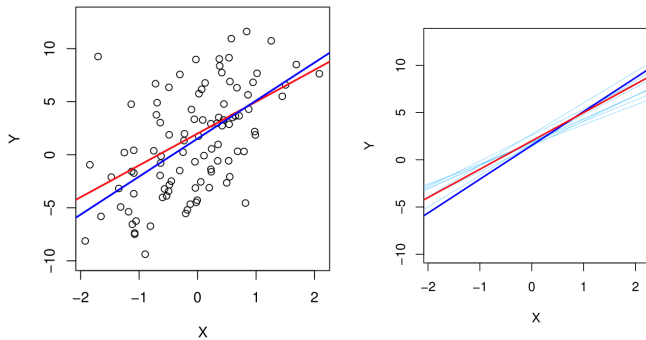
- Sample mean is unbiased for population mean:

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

- Standard variance estimate is biased

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

Linear regression is unbiased



Variance in estimation

Continuing analogy with mean

- True value μ^*
- Estimate from training data $\hat{\mu}$
- Variance of sample mean
$$\text{Var}(\hat{\mu}) = \text{SE}(\hat{\mu})^2 = \frac{\sigma^2}{n}$$
- Test your understanding: [PollEv](#)

Variance of linear regression estimates

- Variance of linear regression estimates:

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

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

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

- Residual standard error is an estimate of σ

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

Coding group work

Run the section titled “Simulating data”

The 95% confidence interval for β_1 approximately takes the form

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

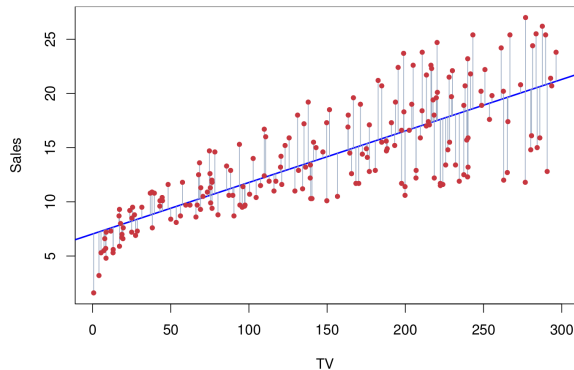
Interpretation:

There is approximately a 95% chance that the interval

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

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:

- $\beta_1 :: [0.042, 0.053]$

- $\beta_0 :: [6.130, 7.935]$

Hypothesis testing

H_0 : There is no relationship between X and Y

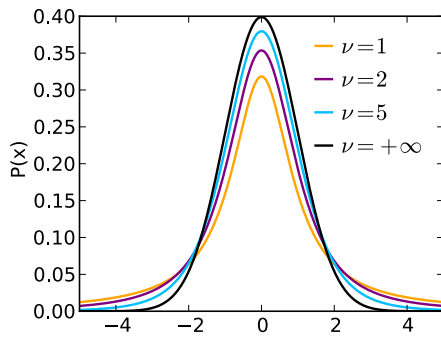
H_1 : There is some relationship between X and Y

Test statistic and p-value

Test statistic:

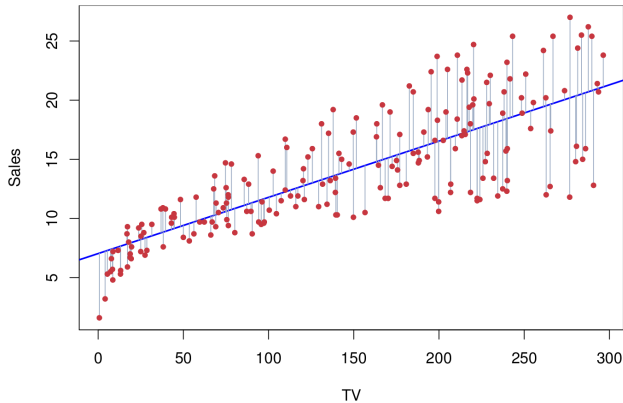
$$t = \frac{\hat{\beta}_1 - 0}{\text{SE}(\hat{\beta}_1)}$$

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



Advertising example

	Coefficient	Std. error	<i>t</i> -statistic	<i>p</i> -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

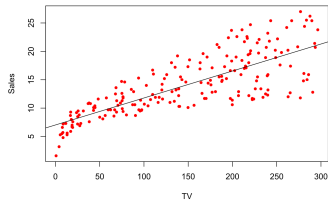
R squared:

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

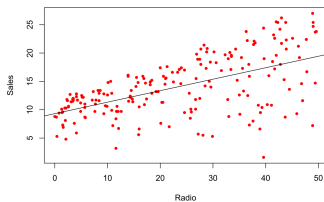
where total sum of squares is

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

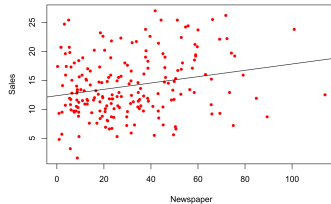
Advertising example



$$R^2 = 0.61$$



$$R^2 = 0.33$$



$$R^2 = 0.05$$

Coding group work

Run the section titled “Assessing
Coefficient Estimate Accuracy”

Next time

CMSE381_F2025_Schedule : Schedule

Lec #	Date	Topic	Reading	HW
1	M 8/25	Intro / Python Review	1	
2	W 8/27	What is statistical learning	2.1	
3	F 8.29	Assessing Model Accuracy	2.2.1, 2.2.2	
	M 9/1	Labor Day - No Class		
4	W 9/3	Linear Regression	3.1	
5	F 9/5	More Linear Regression	3.1	HW #1 Due Sun 9/7
6	M 9/8	Multi-linear Regression	3.2	
7	W 9/10	Probably More Linear Regression	3.3	
8	F 9/12	Last of the Linear Regression		HW #2 Due Sun 9/14
9	M 9/15	Intro to classification, Bayes classifier, KNN classifier	2.2.3	
10	W 9/17	Logistic Regression	4.1, 4.2, 4.3, 4.4	

Announcements

- Homework 1
 - ▶ Due Sun, Sep 7
- Homework 2
 - ▶ Due Sun, Sep 14