Ch 3.1: Linear Regression

Lecture 3 - CMSE 381

Prof. Rongrong Wang

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

January 17, 2024

Covered in this lecture

- Least squares coefficient estimates for linear regression
- Residual sum of squares (RSS)

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

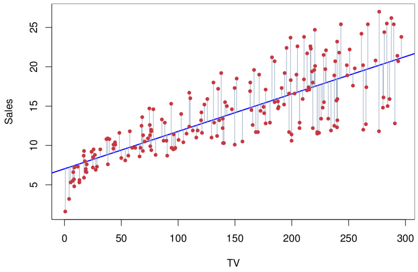
- $$\text{sales} \approx \beta_0 + \beta_1 \text{TV}$$
- β_0 intercept; β_1 slope
- Coefficients or parameters : $\{\beta_0, \beta_1\}$
- Once we have good guesses for $\hat{\beta}_i$, model is

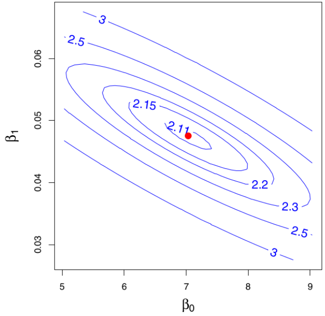
$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

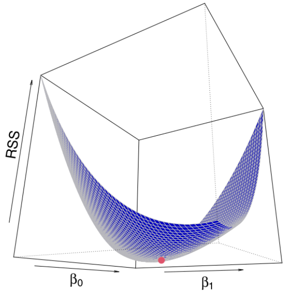
Least squares criterion: Setup

How do we estimate the coefficients?

- 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







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

Minimizing RSS:

$$(\hat{\beta}_0, \hat{\beta}_1) = \arg\min_{\beta_0,\beta_1} \sum_i (y_i - \beta_0 - \beta_1 x_i)^2$$

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

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

$$y_i = f(x_i) + \varepsilon_i$$

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

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

- Closed form!

$$\overline{x} = \frac{\sum x_i}{n} \qquad \overline{y} = \frac{\sum y_i}{n}$$

Group work

Derive the closed form expression of $\hat{\beta}_0$ and $\hat{\beta}_1$ by yourself.

$$\bar{x} = \frac{\sum x_i}{n}$$

$$\begin{cases} \sum (y_i - \beta_0 - \beta_1 x_i) = 0 & ① \\ \sum x_i (y_i - \beta_0 - \beta_1 x_i) = 0 & ② \end{cases}$$

$$\boxed{\sum (x_i - \bar{x}) = 0} = \sum x_i - \sum_{i=1}^{n} \bar{x} = \sum x_i - n\bar{x} = \sum x_i - \sum x_i = 0$$

$$\boxed{\sum (y_i - \bar{y}) = 0}$$

$$① \Rightarrow \frac{\sum y_i}{n} - \frac{\sum \beta_0}{n} - \frac{\sum \beta_1 x_i}{n} = \frac{0}{n} \Rightarrow y - \beta_0 - \beta_1 \bar{x} = 0$$

$$② - \bar{x} \cdot ① : \sum x_i (y_i - \beta_0 - \beta_1 x_i) - \sum \bar{x} (y_i - \beta_0 - \beta_1 x_i) = 0$$

$$\Leftrightarrow \sum (x_i - \bar{x})(y_i - \beta_0 - \beta_1 x_i) = 0$$

$$\Leftrightarrow \sum (x_i - \bar{x}) y_i - \cancel{\beta_0 \sum (x_i - \bar{x})} - \beta_1 \sum (x_i - \bar{x}) x_i = 0$$

$$\Rightarrow \beta_1 = \frac{\sum (x_i - \bar{x}) y_i}{\sum (x_i - \bar{x}) x_i}$$

Section 2

Assessing Coefficient Estimate Accuracy

Bias in estimation

Analogy with mean

Everything here is about determining if linear model is doing a good job

$\frac{\sum X_i}{n}$

- 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:

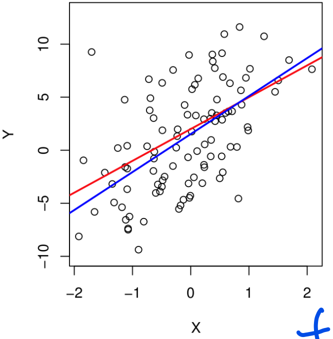
$$E(\hat{\mu}) = E\left(\frac{1}{n}\sum_i X_i\right) = \mu \quad {}^{\prime\prime}\sum X_i$$

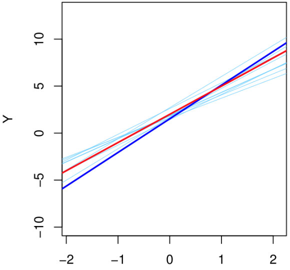
- Standard variance estimate is biased

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

$n-1$

Linear regression is unbiased





$f(x) = \beta_0 + \beta_1 X$

$\mathbb{E}(\hat{\beta}_0) = \beta_0$ $\quad$ $\mathbb{E}(\hat{\beta}_1) = \beta_1$

- 100 data points drawn from $Y = 2 + 3X + \varepsilon$
- ε drawn from normal distribution with mean 0
- Red line is true relationship, blue is least squares estimate
- Repeat this 10 times and plot all the found lines (in variations of blue)
- The resulting models are slightly different but are all around the red true relationship

Variance in estimation

Continuing analogy with mean

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

- Standard error
- The more data you have, the smaller variance, the better the estimate

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

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

Variance of linear regression estimates

- the Standard errors can be used to compute confdence intervals.
- A 95% confidence interval is defined as a range of values such that with 95% probability, the range will contain the true unknown value of the parameter.
- For linear regression, the 95% confidence interval for β_0 β_1 approximately take the form

$$\hat{\beta}_0 \pm 2SE(\hat{\beta}_0), \quad \hat{\beta}_1 \pm 2SE(\hat{\beta}_1)$$

Coding group work

Work on the in-class assignment titled
"LinRegLab"

Next time

Announcements

- Quiz 1 is on Friday !
- Homework 2 is to be released on Friday.
- next time: Linear regression (II)