Ch 3.2: Multiple Linear Regression

Lecture 6 - CMSE 381

Prof. Elizabeth Munch

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

Weds, Sep 11, 2024

Announcements

Last time:

- 3.1 Linear regression

Announcements:

- Homework #2 Due Sunday on Crowdmark

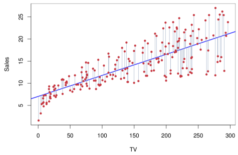
Covered in this lecture

- Multiple linear regression
- Hypothesis test in that case
- Forward and Backward Selection

Section 1

Review from last time

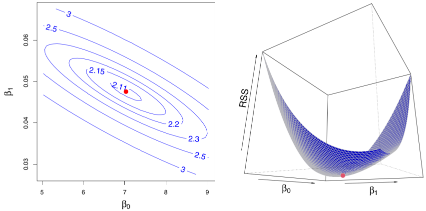
Linear Regression with One Variable



- Predict Y on a single variable X

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

- Find good guesses for $\hat{\beta}_0$, $\hat{\beta}_1$.
- $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$
- $e_i = y_i - \hat{y}_i$ is the ith residual
- RSS $= \sum_i e_i^2$



- RSS is minimized at *least squares coefficient estimates*

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

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

Evaluating the model

- Linear regression is unbiased
- 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 $\sigma^2 = \mathrm{Var}(\varepsilon)$

- Estimate σ: $\hat\sigma^2 = \frac{RSS}{n-2}$

- The 95% confidence interval for β_1 approximately takes the form

$$\hat\beta_1 \pm 2 \cdot \mathrm{SE}(\hat\beta_1)$$

- Hypothesis test:

  H_0: $\beta_1 = 0$
  H_a: $\beta_1 \neq 0$
  - Test statistic $t = \frac{\hat\beta_1 - 0}{\mathrm{SE}(\hat\beta_1)}$

Assessing the accuracy of the model

Residual standard error (RSE):

$$RSE = \sqrt{\frac{1}{n-2} RSS}$$

R squared:

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

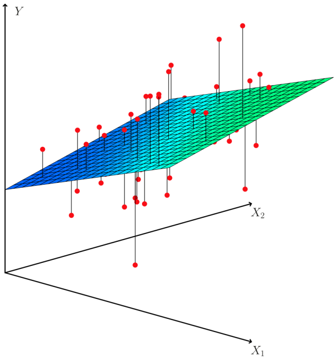
$$TSS = \sum_i (y_i - \overline{y})^2$$

Section 2

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

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_p x_p$$

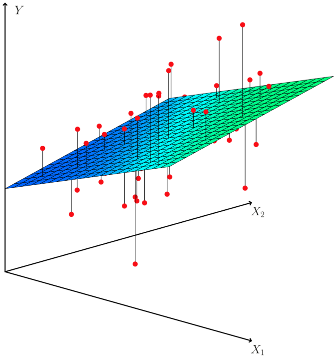
Minimize the sum of squares

$$\begin{aligned} RSS &= \sum_i (y_i - \hat{y}_i)^2 \\ &= \sum (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_1 - \cdots - \hat{\beta}_p x_p) \end{aligned}$$

Coefficients are closed form but UGLY

Advertising data set example

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



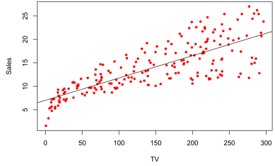
	Coefficient
Intercept	2.939
TV	0.046
radio	0.189
newspaper	−0.001

Interpretation of coefficients

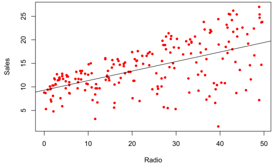
$$\texttt{Sales} = \beta_0 + \beta_1 \cdot \texttt{TV} + \beta_2 \cdot \texttt{radio} + \beta_3 \cdot \texttt{newspaper}$$

	Coefficient
Intercept	2.939
TV	0.046
radio	0.189
newspaper	−0.001

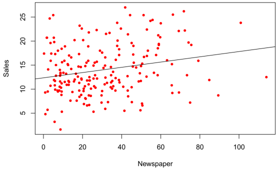
Single regression vs multi-regression



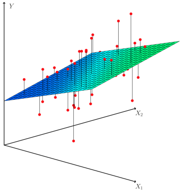
	Coefficient
Intercept	7.0325
TV	0.0475



	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"

Section 3

Ch 3.2.2: Questions to ask of your regression

Question 1

Is at least one of the predictors $X_1, \cdots, X_p$ useful in predicting the response?

Q1: Hypothesis test

$H_0: \beta_1 = \beta_2 = \cdots = \beta_p = 0$

$H_a:$ At least one β_j is non-zero

F-statistic:

$$F = \frac{(TSS - RSS)/p}{RSS/(n-p-1)} \sim F_{p,n-p-1}$$

The F-statistic for the hypothesis test

$$F = \frac{(TSS - RSS)/p}{RSS/(n-p-1)} \sim F_{p,n-p-1}$$

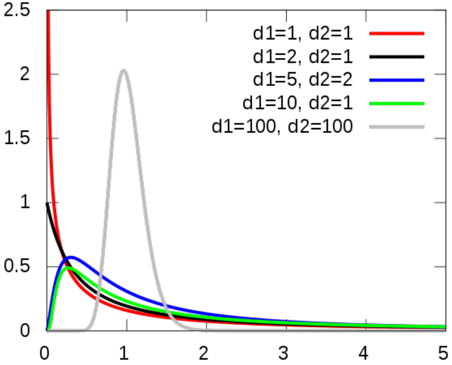


Image from wikipedia, By IkamusumeFan - Own work, CC BY-SA 4.0,

Do Q1 section in jupyter notebook

Q2

Do all the predictors help to explain Y, or is only a subset of the predictors useful?

Q2: A first idea

Great, you know at least one variable is important, so which is it?....

Do Q2 section in jupyter notebook

Why is this a bad idea?

Next time

Lec #	Date			Reading	HW
1	Mon	8/26	Intro / First day stuff / Python Review Pt 1	1	
2	Wed	8/28	What is statistical learning?	2.1	
	Fri	8/30	Class Cancelled (Dr Munch out of town)		
	Mon	9/2	No class - Labor day		
3	Wed	9/4	Assessing Model Accuracy	2.2.1, 2.2.2	
4	Fri	9/6	Linear Regression	3.1	HW #1 Due Sun 9/8
5	Mon	9/9	More Linear Regression	3.1/3.2	
6	Wed	9/11	Even more linear regression	3.2.2	
7	Fri	9/13	Probably more linear regression	3.3	Hw #2 Due Dun 9/15
8	Mon	9/16	Linear regression coding module		
9	Wed	9/18	Intro to classification, Bayes classifier, KNN classifier	2.2.3	
10	Fri	9/20	Logistic Regression	4.1, 4.2, 4.3.1-3	Hw #3 Due Sun 9/22
11	Mon	9/23	Multiple Logistic Regression / Multinomial Logistic Regression /Project day	4.3.4-5	
	Wed	9/25	***Review***		