Ch 3.1-2 Multi Linear Regression

Lecture 6 - CMSE 381

Rachel Roca and Dr. Munch

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

Mon, Sep 11, 2023

Announcements

Last time:

- 3.1 Linear regression
- Started 3.2 Multiple Linear regression

Announcements:

- Homework #2 Due TONIGHT on Crowdmark
- No office hours tomorrow (office hours were today instead)

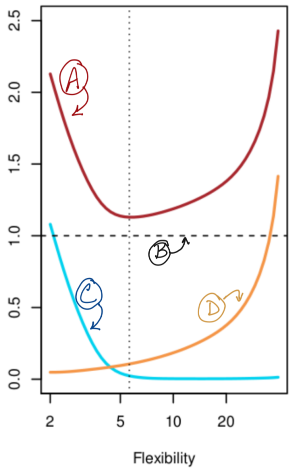
Covered in this lecture

- Review of Quiz
- Multiple linear regression
- Hypothesis test in that case
- R^2 and RSE

Section 1

Quiz Review

Quiz Review: Question 1



- MSE: A
- Bias: C
- Variance of $\hat{f}(x_0)$: D
- Variance of ϵ: B

Quiz Review: Question 2

The *least square coefficients estimates* are the choices for coefficients in our model that minimize *something*. What is it?

- RSS $= \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$
- MSE $= \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$
- RSE $= \sqrt{\frac{1}{n-2}RSS}$
- Difference is the $\frac{1}{n}$ term to take average, does not affect the minimization.
- Square root will also not affect the minimization.

Quiz Review: Question 3

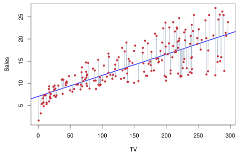
What's the difference between training error and testing error?

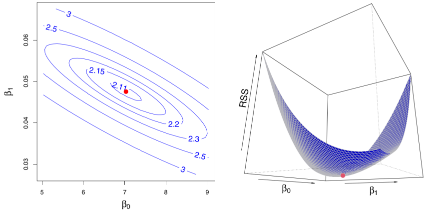
- Training error is the error on the training data, testing error is the error on the testing data.
- Care more about testing error
- Training error can be lower than irreducible error, testing error cannot
- Generally, training error will be lower than testing error
- We want to minimize testing error, not training error

Section 2

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
- $\text{RSS} = \sum_i e_i^2$

- 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:

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

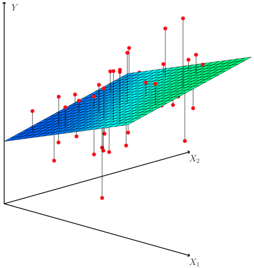
$$TSS = \sum_i (y_i - \overline{y})^2$$

Least Squares Prediction for Multiple Linear Regression

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

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
- β_j is average effect on Y for one unit increase in X_j if all other X_i stay fixed

Coding group work

Let's go back to our Lecture 5 notebook and finish the last part.
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

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

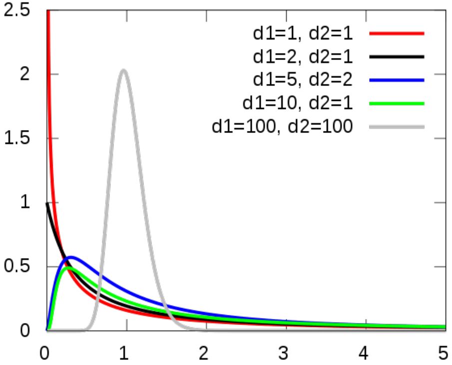


Image from wikipedia, By IkamusumeFan - Own work, CC BY-SA 4.0,

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?

Q3

How well does the model fit the data?

Assessing the accuracy of the module

Almost the same as before

Residual standard error (RSE):

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

R squared:

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

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

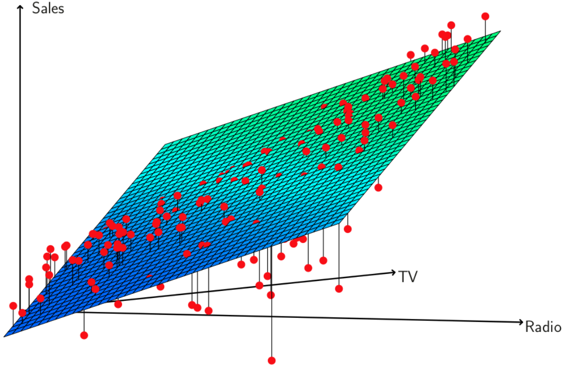
R^2 on Advertising data

- Just TV: $R^2 = 0.61$
- Just TV and radio: $R^2 = 0.89719$
- All three variables: $R^2 = 0.8972$

RSE on Advertising Data

- Just TV: $RSE = 3.26$
- Just TV and radio: $RSE = 1.681$
- All three variables: $RSE = 1.686$

If all else fails, look at the data



Next time

Lec #	Date			Reading
1	Mon	Aug 28	Intro / First day stuff / Python Review Pt 1	1
2	Wed	Aug 30	What is statistical learning?	2.1
	Fri	Sep 1	Assessing Model Accuracy	2.2.1, 2.2.2
3	Mon	Sep 4	No class - Labor day	
4	Wed	Sep 6	Linear Regression	3.1
5	Fri	Sep 8	More Linear Regression	3.1/3.2
6	Mon	Sep 11	Even more linear regression	3.2.2
7	Wed	Sep 13	Probably more linear regression	3.3
8	Fri	Sep 15	Intro to classification, Logisitic Regression	2.2.3, 4.1, 4.2, 4.3
9	Mon	Sep 18	More logistic regression	
10	Wed	Sep 20	Multiple Logistic Regression / Multinomial Logistic Regression	
11	Fri	Sep 22	Overflow/Project day?	
	Mon	Sep 25	***Review***	
	Wed	Sep 27	**Midterm #1**	
	Fri	Sep 29	No class - Dr Munch out of town	
12	Mon	Oct 2	Leave one out CV	5.1.1, 5.1.2
13	Wed	Oct 4	k-fold CV	5.1.3
14	Fri	Oct 6	More k-fold CV,	5.1.4-5
15	Mon	Oct 9	k-fold CV for classification	5.1.5
16	Wed	Oct 11	Resampling methods: Bootstrap	5.2
17	Fri	Oct 13	Subset selection	6.1