CIS 490 Machine Learning

Lecture 5

Instructor: (Julia) Hua Fang

1

2

Last time

- Review: Probability Theory (III)
  - Central limited theory (II)
  - Multivariate distribution
    - Joint probability
    - Marginal probability
    - Conditional probability

Reminder: Sectional Homework 1, Due Feb 4^{th}.

Adapted from Jeff Howbert, Greg Shakhnarovich

2

Answers to class exercises in LS 4: Multivariate Discrete Distribution

3

Example: Probability Table			**X= Model type**		
		Sedan	**Minivan**	**SUV**	**Sport**
Y= Manufacturer	**European**	**n/a**	**0.1481**	**n/a**	**n/a**
	Asian	**n/a**	**0.1111**	**n/a**	**n/a**
	American	**n/a**	**0.0741**	**n/a**	**n/a**
		Marginal	**0.3333**		
Assuming calculating the probability of selling a high volume of cars					
Joint Probability	p(X = x, Y = y) p (X = minivan, Y= European) = 0.1481				
Marginal Probability	p (X=x) = $\sum_{b= \text{all values of } Y}$ p(X=x,Y=b)				
	p(X = minivan) = 0.0741 + 0.1111 + 0.1481 = 0.3333				
Conditional probability	p(X = x | Y = y) = p(X = x, Y = y) / p(Y = y) p(Y = European | X = minivan) = 0.1481 / (0.0741 + 0.1111 + 0.1481) = 0.4433				

3

4

Outline

- Supervised Learning (I): Two major categories
  - Regression
  - Classification

What is the difference between these two categories?

4

5

Regression: Continuous Outcome (Y)

- Linear Regression

5

6

Linear Regression: Simple vs Multiple linear regression

- Simple linear regression & Multiple linear regression

  You need to understand:
  - Loss function: sum of squared loss (SSL), sum of squared residual (SSR), or residual sum of squares (RSS)
  - Estimation: Least square, the best fit line
  - Accuracy checking/quality of model fit: Residual standard error (RSE), R-squared (R^2), Mean squared error (MSE),

    or use simulation (not required for this course).

6

7

Simple linear regression

7

8

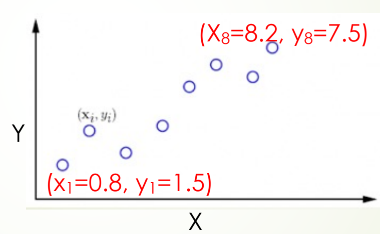
Simple linear regression (1)

- Simple linear regression: only 1 predictor/explanatory variable/attribute (X); outcome is continuous variable (Y)

Faked data:

ID	X	Y
1	0.8	1.5
2	2.1	2.5
3	3.2	2.4
4	4.1	2
5	5.2	5.3
6	6.0	6.4
7	7.2	6.1
8	8.2	7.5

Regress Y on X: Predict Y based on X



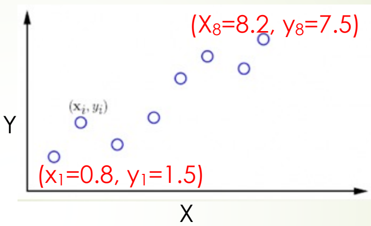
Assuming approximately linear relationship, how would you write this relationship mathematically?
(next slides)

8

9

Simple linear regression (2)

$$Y = f(X) + \epsilon$$



If this unknown function *f* is approximately linear, then

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

How can you find Betas?

So,

$Y = \beta_0 + \beta_1 X + \epsilon$ for Population

β_0 : intercept; when X =0, Y =?

β_1 : Slope, the average increase in Y given one unit increase in X

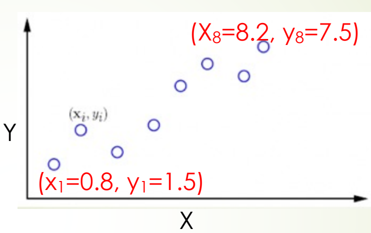
ϵ : a mean-zero random error term

9

10

Simple linear regression (3)

ID	X	Y
1	0.8	1.5
2	2.1	2.5
3	3.2	2.4
4	4.1	2
5	5.2	5.3
6	6.0	6.4
7	7.2	6.1
8	8.2	7.5



Given this sample, how would you find the estimates for Betas, $\hat{\beta}_0$ $\hat{\beta}_1$?

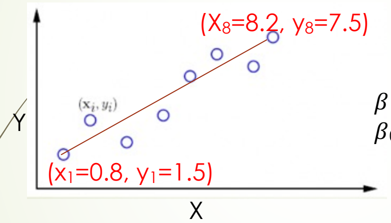
$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

Hint: How do you define a line?

10

Simple linear regression (4)

- A Fit line



β_1= (7.5 -1.5)/(8.2-0.8) = 6/7.4
β_0= sum (1.5 + …+7.5)/8 = avg (Y)

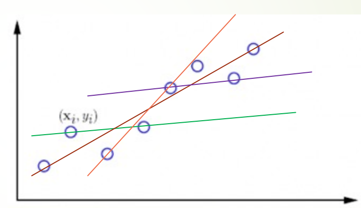
Is this the best fit line?

11

Simple linear regression (5)

- A few Fit lines based on sample data

ID	X	Y
1	0.8	1.5
2	2.1	2.5
3	3.2	2.4
4	4.1	2
5	5.2	5.3
6	6.0	6.4
7	7.2	6.1
8	8.2	7.5



Is one of them the best line?

12

13

Simple linear regression: Loss function

13

14

Simple linear regression-Loss function

To find the best fit line with best estimated $\hat{\beta}_0$ $\hat{\beta}_1$, we need to find the function *f*, that minimizes the sum of squared loss (SSL)

Loss function: defines the penalty for predicting $\hat{y}$ when the true value is y. (ie., penalizing errors in prediction)

- Loss function for regression:

$$L(y, \hat{y}) = (y - \hat{y})^2$$

Recall: sum of squared loss (SSL), sum of squared residual (SSR), or residual sum of squares (RSS), refer to the same thing.

14

15

Simple linear regression—(Ordinary) Least Squares Estimation

15

Simple linear regression: Ordinary) Least Squares Estimation (I)

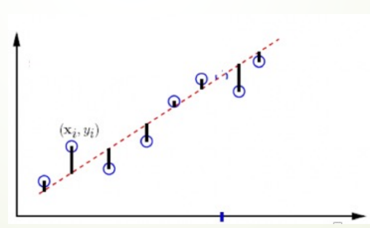
Likewise, in statistics, we call Least Squares (LSQ), or Ordinary Least Squares (OLS) to estimate $\hat{\beta}_0$ $\hat{\beta}_1$, and find the best line

Let $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ be predication for y based on x_i

ID	X	Y
1	0.8	1.5
2	2.1	2.5
3	3.2	2.4
4	4.1	2
5	5.2	5.3
6	6.0	6.4
7	7.2	6.1
8	8.2	7.5

Step 1: Compute residual (ie, loss)

$$e_i = y_i - \hat{y}_i$$



16

Simple linear regression: (Ordinary) Least Squares Estimation (II)

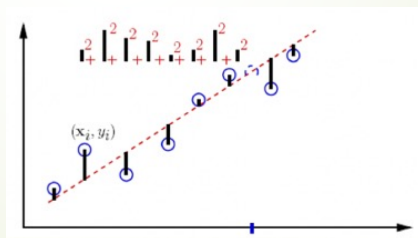
ID	X	Y
1	0.8	1.5
2	2.1	2.5
3	3.2	2.4
4	4.1	2
5	5.2	5.3
6	6.0	6.4
7	7.2	6.1
8	8.2	7.5

Step 2: Compute residual sum of squares (RSS). recall, also called as sum of squared residuals (SSR) or sum of squared loss (SSL).

$$\mathrm{RSS} = e_1^2 + e_2^2 + \cdots + e_n^2$$

Equivalently,

$$\mathrm{RSS} = (y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1)^2 + (y_2 - \hat{\beta}_0 - \hat{\beta}_1 x_2)^2 + \ldots + (y_n - \hat{\beta}_0 - \hat{\beta}_1 x_n)^2$$



17

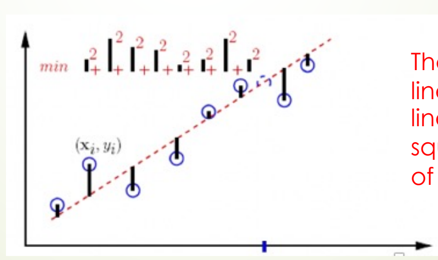
Simple linear regression: (Ordinary) Least Squares Estimation (III)

ID	X	Y
1	0.8	1.5
2	2.1	2.5
3	3.2	2.4
4	4.1	2
5	5.2	5.3
6	6.0	6.4
7	7.2	6.1
8	8.2	7.5

Step 3: use calculus to find the values of $\hat{\beta}_0$ $\hat{\beta}_1$ that give the minimum RSS/SSR/SSL.

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

where $\bar{y} \equiv \frac{1}{n}\sum_{i=1}^{n} y_i$ and $\bar{x} \equiv \frac{1}{n}\sum_{i=1}^{n} x_i$ are the sample means



The dashed red line is the best fit line with least squares estimates of $\hat{\beta}_0$ $\hat{\beta}_1$

18

Suppl.: Simple linear regression:(Ordinary) Least Squares Estimation (III) deviation of betas

SSR, the quantity to minimize.

$$
\begin{aligned}
SSR &= \sum_{i=1}^{n}(y_i - (\beta_0 + \beta_1 x_i))^2 \\
&= \sum_{i=1}^{n}\left(y_i^2 - 2y_i(\beta_0 + \beta_1 x_i) + \beta_0^2 + 2\beta_0\beta_1 x_i + \beta_1^2 x_i^2\right)
\end{aligned}
$$

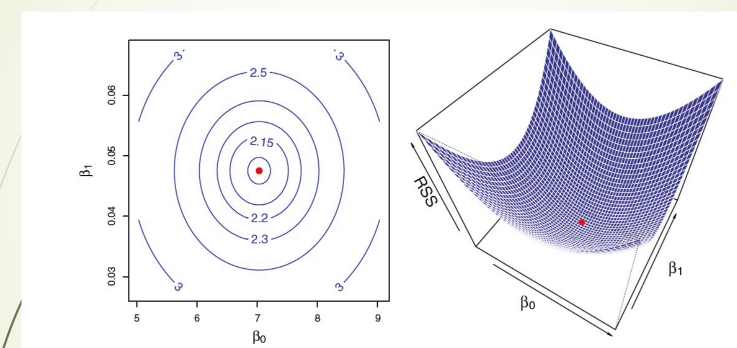
$$
\begin{aligned}
\frac{\partial SSR}{\partial \beta_0} &= \sum_{i=1}^{n}(-2y_i + 2\beta_0 + 2\beta_1 x_i) \\
0 &= \sum_{i=1}^{n}\left(-y_i + \hat\beta_0 + \hat\beta_1 x_i\right) \\
0 &= -n\bar y + n\hat\beta_0 + \hat\beta_1 n\bar x \\
\hat\beta_0 &= \bar y - \hat\beta_1 \bar x \\
\frac{\partial SSR}{\partial \beta_1} &= \sum_{i=1}^{n}\left(-2x_i y_i + 2\beta_0 x_i + 2\beta_1 x_i^2\right) \\
0 &= -\sum_{i=1}^{n} x_i y_i + \hat\beta_0 \sum_{i=1}^{n} x_i + \hat\beta_1 \sum_{i=1}^{n} x_i^2 \\
0 &= -\sum_{i=1}^{n} x_i y_i + (\bar y - \hat\beta_1 \bar x)\sum_{i=1}^{n} x_i + \hat\beta_1 \sum_{i=1}^{n} x_i^2 \\
\hat\beta_1 &= \frac{\sum_{i=1}^{n} x_i(y_i - \bar y)}{\sum_{i=1}^{n} x_i(x_i - \bar x)}
\end{aligned}
$$

See details in Section 9.4 of Chapter 9 Simple Linear Regression at http://www.stat.cmu.edu/~hseltman/309/Book/chapter9.pd

19

20

Example: the least squares estimates $\hat\beta_0$ $\hat\beta_1$ for Advertising data (ISLR_chpt3)



Contour and three-dimensional plots of the RSS on the Advertising data, using sales as the response and TV as the predictor. The red dots correspond to the least squares estimates: $\hat\beta_0$ =7.03. $\hat\beta_1$ =0.0475

Recall: ISLR refers to the book called An introduction to statistical learning with application in R

20

21

Multiple linear regression

22

Multiple linear regression (1)

- Multiple linear regression: Multiple dimensions (i.e., multiple X, called multiple regression)

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon,$$

- Predict the response Y on the basis of a set of values for the predictors $X_1, X_2, \ldots, X_p$, where p = number of predictors

Multiple linear regression (2)

23

- Multiple linear regression:
  - Find the Least squares plane ("the best fit plane", e.g. 2-Dimensions/Predictors) ; hyperplane, if *d*-Dimensions.

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \cdots + \hat{\beta}_p X_p$$

Choose $\beta_0, \beta_1, \ldots, \beta_p$. to minimize the SSR using the same Least squares approach.

$$\begin{aligned} \text{RSS} &= \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \\ &= \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \cdots - \hat{\beta}_p x_{ip})^2 \end{aligned}$$

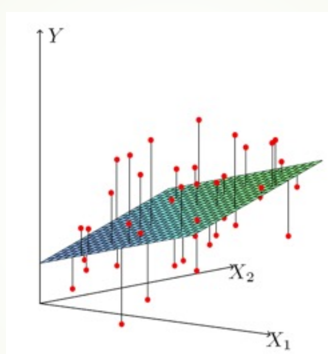
$$\boldsymbol{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

where **X** is the matrix with all ones in the first column (for the intercept)

23

Example: Least squares plane

24



24

25

linear regression: Accuracy Checking/Quality of Model Fit

25

Linear regression Accuracy Checking

Faked example:

Quantity	Value
Residual standard error	1.69
R^2	0.897
F-statistic	570

- **Residual Standard Error (RSE)**: an estimate of the standard deviation of error term ϵ. Roughly speaking, it is the average amount that the response will deviate from the true regression line

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

Simple linear regression: p =1. $\mathrm{RSE} = \sqrt{\frac{1}{n-2}\mathrm{RSS}}$

The RSE is considered a measure of the lack of fit of the model to the data: the smaller the better.

26

Linear regression Accuracy Checking

Faked example:

Quantity	Value
Residual standard error	1.69
R^2	0.897
F-statistic	570

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

where $\mathrm{TSS} = \sum(y_i - \bar{y})^2$ is the *total sum of squares*.

- R^2 : the fraction of variance in Y explained by X, it is the square of the correlation between the response and the fitted linear model. E.g., 89.7% of variance in Y is explained by X in this example.
  R^2 : the larger the better.
  In the **simple** linear regression setting, $R^2 = r^2$ (r: correlation coefficient)

Adjusted R^2: adjusted for the number of predictors in the model: the larger the better.

27

Linear regression Accuracy Checking

Faked example:

Quantity	Value
Residual standard error	1.69
R^2	0.897
F-statistic	570

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

- F-statistics: assess multiple coefficients simultaneously; compare the fitted model with the intercept model; if significant, the fitted model is better.

  e.g. 570 is significantly larger than 1 and at least one of X must be related to the outcome/response.

28

Linear regression Accuracy Checking

- Mean Squared Error:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2$$

That is, MSE = RSS/n

- Root Mean Squared Error (RMSE)= Sqrt (MSE)

Note: MSE and RMSE Generic criteria for regression (Y is a continuous variable)

29

30

Reading assignments:

- Review Lecture 5 Slides, and read
  - Simple linear regression: Section 9.4-9.5 (Must); Section 9.6-9.7 (Encouraged) http://www.stat.cmu.edu/~hseltman/309/Book/chapter9.pdf,
  - ISLR Chapter 3.1 and 3.2:

ISLR: An Introduction To Statistical Learning, (2013) Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani, 2013, ISBN: 9781461471387 (online) and 9781461471370.

Will use ISLR to refer to this book from now on.

30