Lecture 3 Linear Regression I

ECE 625: Data Analysis and Knowledge Discovery

Di Niu

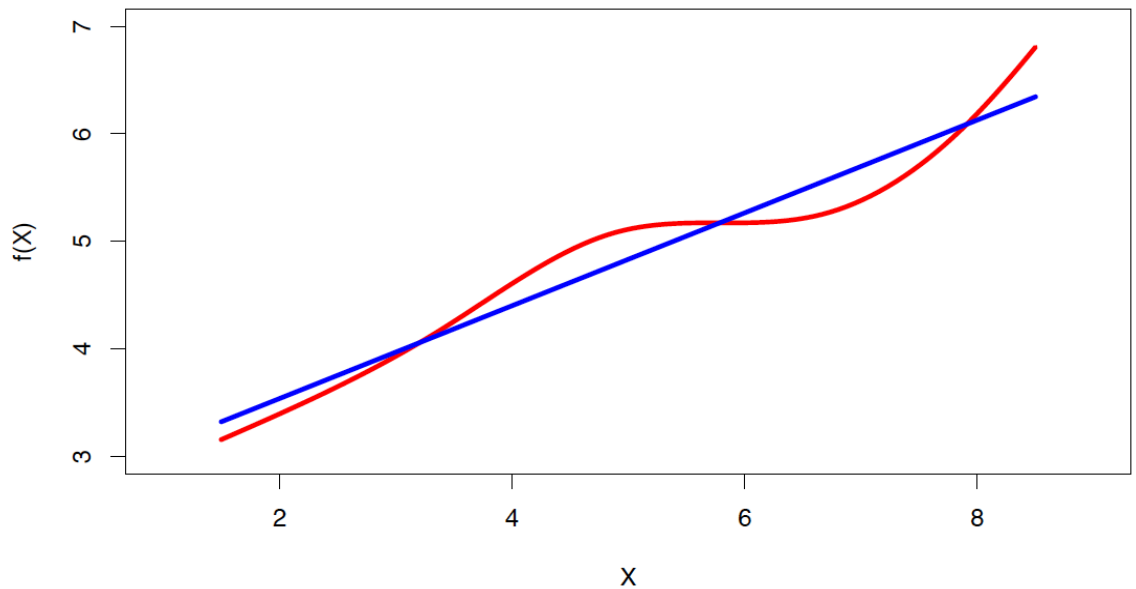
Department of Electrical and Computer Engineering
University of Alberta

January 13, 2021

Outline

Simple Linear Regression

- Linear regression is a simple approach to supervised learning. It assumes that the dependence of Y on $X_1, X_2, \cdots, X_p$ is linear.
- True regression functions are never linear! although it may seem overly simplistic, linear regression is extremely useful both conceptually and practically.



Simple Linear Regression

- Simple Linear Regression Model (SLR) has the form of

$$Y = \beta_0 + \beta_1 X + \varepsilon,$$

where β_0 and β_1 are two unknown parameters (coefficients), called intercept and slope, respectively, and ε is the error term.

- Given the estimates $\hat{\beta}_0$ and $\hat{\beta}_1$, the estimated regression line is

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

- For $X = x$, we predict Y by $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$, where the hat symbol denotes an estimated value.

Estimate the parameters

sample

n training samples

- Let (y_i, x_i) be the i-th observation in the training set and $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$, we call $e_i = y_i - \hat{y}_i$ the ith residual.
- To estimate the parameters, the Least Squares approach minimizes the residual sums of squares (RSS):

batch 1 of n = 10 samples
x1,...,x10
y1,...,y10

$$\text{RSS} = \sum_i e_i^2 = \sum_i \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i, \right)^2.$$

\hat y_i

batch 2 of n = 10 samples
x1,...,x10 are the same as batch 1
y1',...,y10'

- Let $\bar{y} = \sum_i y_i/n$ and $\bar{x} = \sum_i x_i/n$. The minimizers are

beta1 x + e
under e ~ N(0,sigma^2)

$$\hat{\beta}_1 = \frac{\sum_i (y_i - \bar{y})(x_i - \bar{x})}{\sum_i (x_i - \bar{x})^2},$$

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

Taking derivatives of RSS wrt beta0, beta1 and setting them to 0 gives us two linear equations

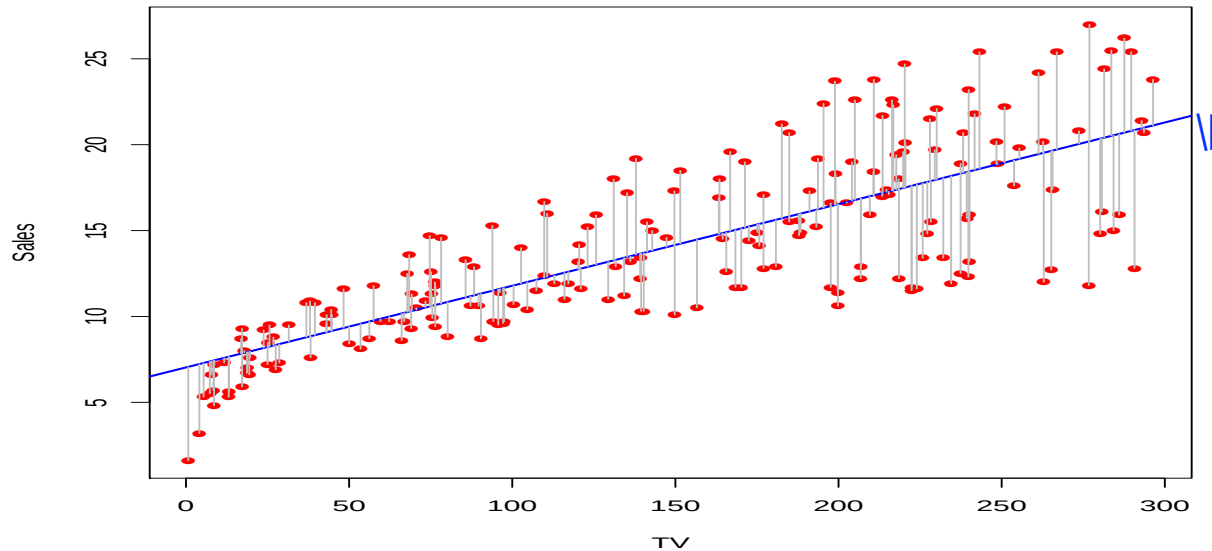
they are just functions of the training data

\hat beta1 is a linear combination of Gaussian RVs y1,...,y10 and thus is a Gaussian RV

Example

Red dots are samples

Blue line is the linear model obtained by minimizing the RSS



\hat y given by blue line

- Advertising data: the least square fit for the regression of `sales` and `TV`. fit: model
- Each grey line segment represents an error, and the fit makes a compromise by minimizing the residual sums of squares (or minimizing training MSE).
- In this case a linear fit captures the essence of the relationship, although it is somewhat deficient in the left of the plot.

E[\hat beta0] = beta0
E[\hat beta1] = beta1

n training samples generated by y = beta0 + beta1 x + e
under e ~ N(0,sigma^2)
Training data are random variables

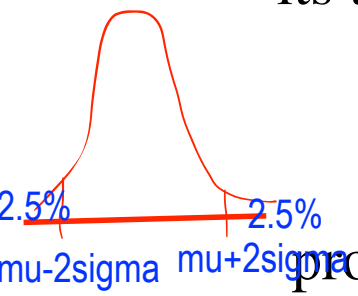
How good are these coefficient estimates?

- $\hat{\beta}_0$ and $\hat{\beta}_1$ are unbiased estimates of β_0 and β_1.
- The standard error of an estimator reflects how it varies around its true value under different training data. We have

standard deviation

under different batches of training data

$$\mathrm{SE}(\hat{\beta}_1) = \sqrt{\frac{\sigma^2}{\sum(x_i - \bar{x})^2}}, \qquad \mathrm{SE}(\hat{\beta}_0) = \sqrt{\sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum(x_i - \bar{x})^2} \right)},$$



provided that the errors ε_i for each observation are uncorrelated and $\mathrm{Var}(\varepsilon_i) = \sigma^2$.

- A 95% confidence interval contains the true unknown value of the parameter with 95% probability, and is given by

\hat beta1 is in [beta1+-1.96 SE(\hat beta1)] with 95% probability

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

beta1 is in [\hat beta1+-1.96 SE(\hat beta1)] with 95% probability

- For the advertising data, the 95% confidence interval for β_1 is $[0.042, 0.053]$, which means, there is *approximately* 95% chance this interval contains the true value of β_1, assuming ε is Gaussian. (In this case, $\hat{\beta}_1$ follows a Gaussian distribution).

Hypothesis testing

- Standard errors can also be used to perform hypothesis tests on the coefficients. The most common hypothesis tests involve testing the null hypothesis:

  H_0: There is no relationship between X and Y

  and the alternative hypothesis:

  H_A: There is *some* relationship between X and Y.
- Mathematically, we are testing

$$H_0 : \beta_1 = 0 \text{ versus } H_A : \beta_1 \neq 0,$$

since if $\beta_1 = 0$ then the model reduces to $Y = \beta_0 + \varepsilon$, and X is not associated with Y.

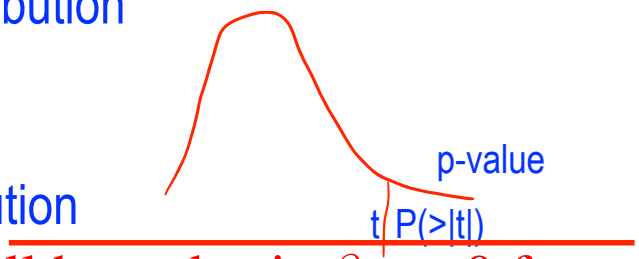
Hypothesis testing

- To test the null hypothesis, we compute a t-statistics,

1. Assume beta1=0, then t statistic follows a t_{n-2} distribution

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

2. Check this specific t value,
see if this t value is a rare event under the above distribution



- This variable follows t_{n-2} under the null hypothesis $\beta_1 = 0$ for an estimated $\hat{\sigma}^2$.
- Using statistical software, it is easy to compute the probability of observing any value equal to $|t|$ or larger. We call this probability the p-value.

a small p-value means that this t is unlikely (rare event), reject null hypothesis

- Results for the advertising data

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 7.032594   0.457843   15.36   <2e-16 ***
TV          0.047537   0.002691   17.67   <2e-16 ***
---
```

Measure of fit

- We compute the Residual Standard Error (which is also an estimate of σ^2) as

$$\text{RSE} = \sqrt{\frac{1}{n-2}\text{RSS}} = \sqrt{\frac{1}{n-2}\sum_i (y_i - \hat{y}_i)^2},$$

  where the residual sum-of-squares is $\text{RSS} = \sum_i (y_i - \hat{y}_i)^2$.
- R-squared or fraction of variance explained is

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

  where $\text{TSS} = \sum_i (y_i - \bar{y})^2$ is the total sum of squares.
- It can be shown that in this simple linear regression setting that $\text{R}^2 = r^2$, where r is the correlation between Y and X:

$$r = \frac{\sum_i (y_i - \bar{y})(x_i - \bar{x})}{\sqrt{\sum_i (y_i - \bar{y})^2}\sqrt{\sum_i (x_i - \bar{x})^2}} = \left(\hat{\beta}_1 \frac{\sqrt{\sum_i (x_i - \bar{x})^2}}{\sqrt{\sum_i (y_i - \bar{y})^2}}\right).$$

R code

```
> TVadDataÂ =Â read.csv('... Advertising.csv')Â Â
> attach(TVadData)
> TVadlm = lm(Sales~TV)
> summary(TVadlm)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 7.032594   0.457843   15.36   <2e-16 ***
TV          0.047537   0.002691   17.67   <2e-16 ***
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1

Residual standard error: 3.259 on 198 degrees of freedom
Multiple R-squared:  0.6119,Adjusted R-squared:  0.6099
F-statistic: 312.1 on 1 and 198 DF,  p-value: < 2.2e-16
```

beta0

beta1

~2 for t value

Summary and Remarks

- Simple linear regression
- Estimation and inference
- Measure of fit R^2
- Read textbook Chapter 3
- Do R lab