# Lecture 3 Linear Regression I

## STAT 441/505: Applied Statistical Methods in Data Mining

Linglong Kong

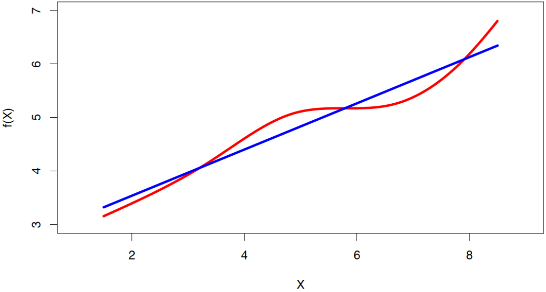
Department of Mathematical and Statistical Sciences
University of Alberta

Winter, 2016

# 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.



UNIVERSITY OF ALBERTA
EDMONTON·ALBERTA·CANADA

# Simple Linear Regression

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

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

  where $\beta_0$ and $\beta_1$ are two unknown parameters (coefficients), called intercept and *slope*, respectively, and $\varepsilon$ is the error term.
- Given the estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ , the estimated regression line is

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

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

UNIVERSITY OF ALBERTA
EDMONTON·ALBERTA·CANADA

# Estimate the parameters

- Let $(y_i, x_i)$ be the $i$-th observation and $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$, we call $e_i = y_i - \hat{y}_i$ the $i$th residual.
- To estimate the parameters, we minimized the residual sums of squares (RSS),

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

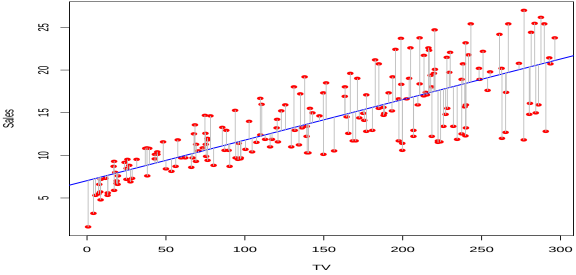
- Denote $\bar{y} = \sum_i y_i/n$ and $\bar{x} = \sum_i x_i/n$. The minimized values are

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

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

UNIVERSITY OF ALBERTA
EDMONTON·ALBERTA·CANADA

# Example




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

UNIVERSITY OF ALBERTA
EDMONTON·ALBERTA·CANADA

# Assess the coefficient estimates

- The standard error of an estimator reflects how it varies under repeated sampling.

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

where $\sigma^2 = \mathrm{Var}(\varepsilon)$.

- 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.
- It has the form

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

- For the advertising data, the 95% confidence interval for $\beta_1$ is $[0.042, 0.053]$, which means, there is approximately 95% chance this interval contains the true value of $\beta_1$ (under a scenario where we got repeated samples like the present sample).

UNIVERSITY OF ALBERTA
EDMONTON·ALBERTA·CANADA

# Hypothesis testing

- Standard errors can also be used to perform hypothesis tests on the coefficients. The most common hypothesis test involves testing the null hypothesis of
  $H_0$: There is no relationship between $X$ and $Y$ versus the alternative hypothesis
  $H_A$: There is some relationship between $X$ and $Y$.
- Mathematically, we test

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

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

UNIVERSITY OF ALBERTA
EDMONTON·ALBERTA·CANADA

# Hypothesis testing

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

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

- This statistics follows $t_{n-2}$ under the null hypothesis $\beta_1 = 0$.
- 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.
- 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 ***
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1
```

UNIVERSITY OF ALBERTA
EDMONTON·ALBERTA·CANADA

# Measure of fit

- We compute the Residual Standard Error

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

where the residual sum-of-squares is $\mathrm{RSS} = \sum_i (y_i - \hat{y}_i)^2$.

- R-squared or fraction of variance explained is

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

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

- It can be shown that in this simple linear regression setting that $\mathrm{R}^2 = r^2$, 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

```
> TVadData = read.csv('... Advertising.csv')
> 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
```

# Summary and Remark

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