CENG 222
Statistical Methods for Computer Engineering

Week 12

Chapter 11 Regression

11.1 Least squares estimation

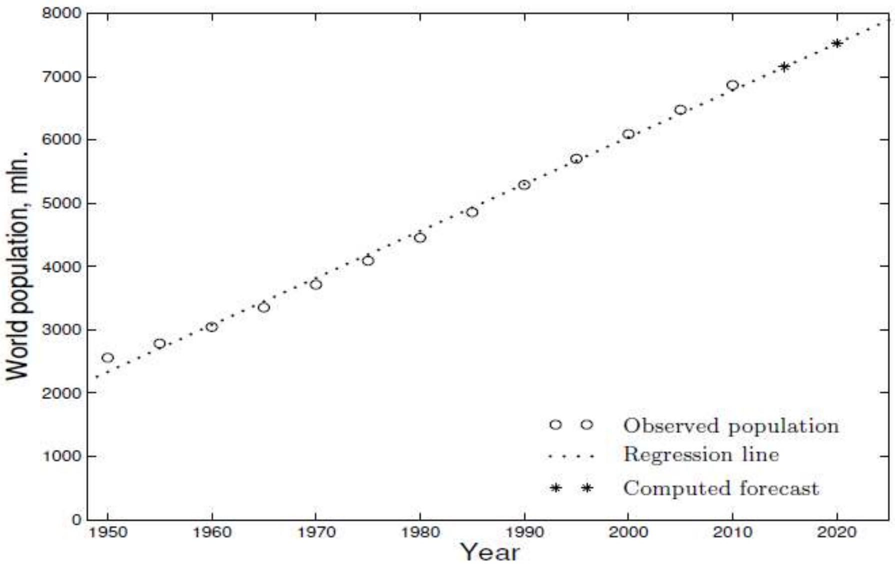
Regression

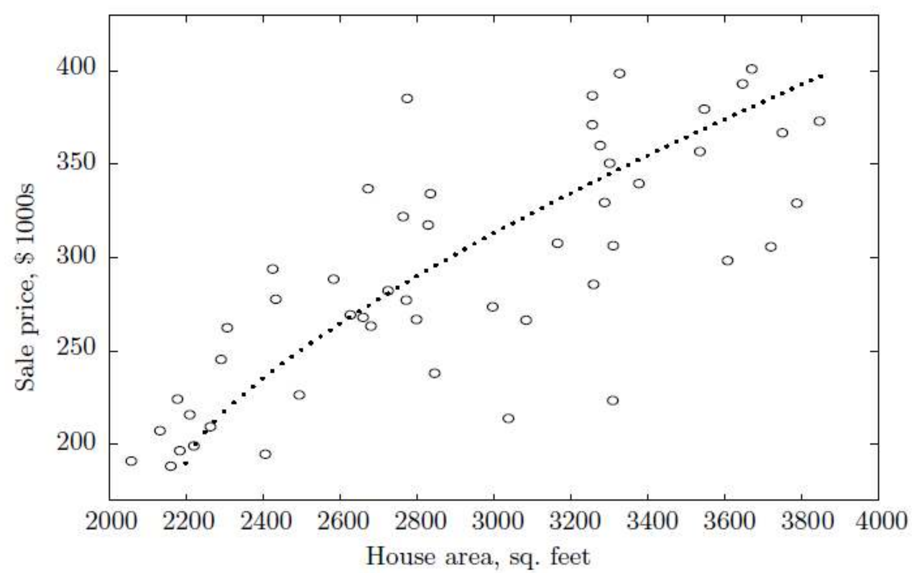
- Analysis of relations between random variables
- Regression of Y on $X^{(1)}, \dots, X^{(k)}$ is the conditional expectation:
  - $\mathbf{E}(Y | X^{(1)} = x^{(1)}, \dots, X^{(k)} = x^{(k)})$
  - Y is called the *response* or *dependent* variable. It is the variable we want to predict
  - $X^{(i)}$s are called the *predictors* or *independent* variables.
- Linear multi-variate regression
  - $Y = \beta_0 + \beta_1 X^{(1)} + \beta_2 X^{(2)} + \cdots + \beta_k X^{(k)}$

Regression

- In this course, we will cover the simplest form:
  - Univariate, linear regression
  - $G(x) = \mathbf{E}(Y|X = x) = \beta_0 + \beta_1 X$
  - Intercept: $\beta_0 = G(0)$
  - Slope: $\beta_1 = G(x + 1) - G(x)$

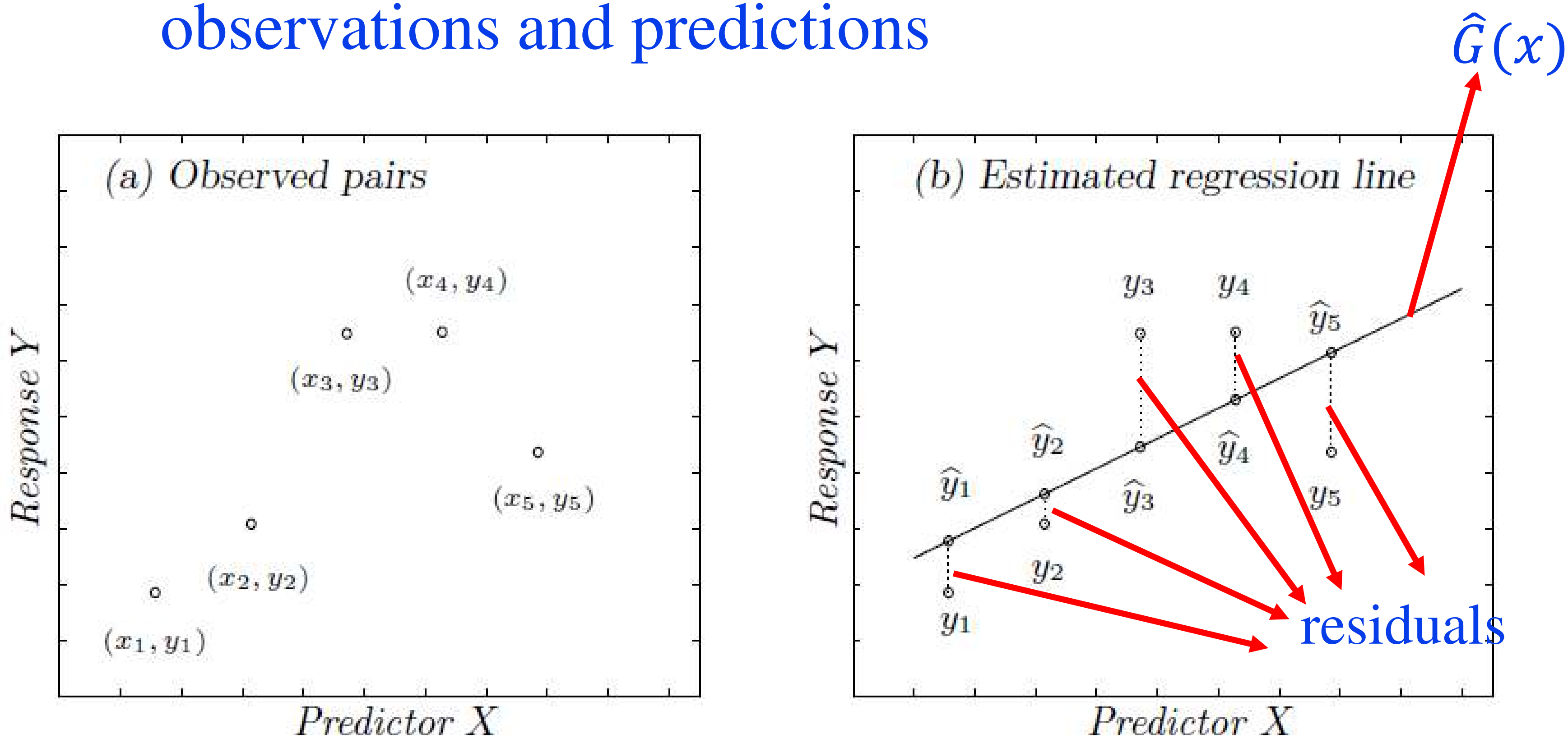
Linear versus Non-Linear Regression





Method of least squares

- Estimate the function $G(x)$ with $\hat{G}(x)$
  - $\hat{G}(x)$: try to minimize the distance between real observations and predictions



Method of least squares

- Find $\hat{G}(x)$ that minimizes the sum of squares of the residuals
  - $e_i = y_i - \hat{y}_i$
  - Minimize $\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$

Method of least squares

- $Q = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}\left(y_i - \hat{G}(x_i)\right)^2$
  $= \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2$
- Take partial derivatives wrt β_0 and β_1 and equate to 0

normal equations

$$\begin{cases} \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i) = 0 \\ \sum_{i=1}^{n} x_i(y_i - \beta_0 - \beta_1 x_i) = 0 \end{cases}$$

Method of least squares

- $\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i\ ) = 0$

$\rightarrow \beta_0 = \frac{\sum y_i - \beta_1 \sum x_i}{n} = \bar{y} - \beta_1 \bar{x}$

- Substitute β_0 in the second normal equation:

$\rightarrow \sum_{i=1}^{n} x_i((y_i - \bar{y}) - \beta_1(x_i - \bar{x})) = 0$

$\rightarrow S_{xy} - \beta_1 S_{xx} = 0$ where

sum of squares

$$S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2$$

sum of cross products

$$S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

Method of least squares steps

1. Compute $\bar{x}$ and $\bar{y}$
2. $S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2$
3. $S_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$
4. $b_1 = \hat{\beta}_1 = S_{xy}/S_{xx}$
5. $b_0 = \hat{\beta}_0 = \bar{y} - b_1\bar{x}$

- Example 11.3 (World Population)

Regression and correlation

- Recall that covariance and correlation coefficient are:
  - $Cov(X, Y) = E\Big(\big(X - E(X)\big)\big(Y - E(Y)\big)\Big)$
  - $\rho = \frac{Cov(X,Y)}{\sigma_x \sigma_y}$
- Sample covariance and sample correlation coefficient can be used to estimate $Cov(X, Y)$ and ρ

Sample covariance and correlation coefficent

- $s_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$
- $r = \frac{s_{xy}}{s_x s_y}$
- s_x and s_y are sample standard deviations
- $s_x = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$ and $s_y = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}}$

$\rightarrow b_1 = \frac{S_{xy}}{S_{xx}} = \frac{s_{xy}}{s_x s_x} = r\left(\frac{s_y}{s_x}\right)$

Why linear
when we can have 0 sum of errors?

- Answer: to avoid overfitting