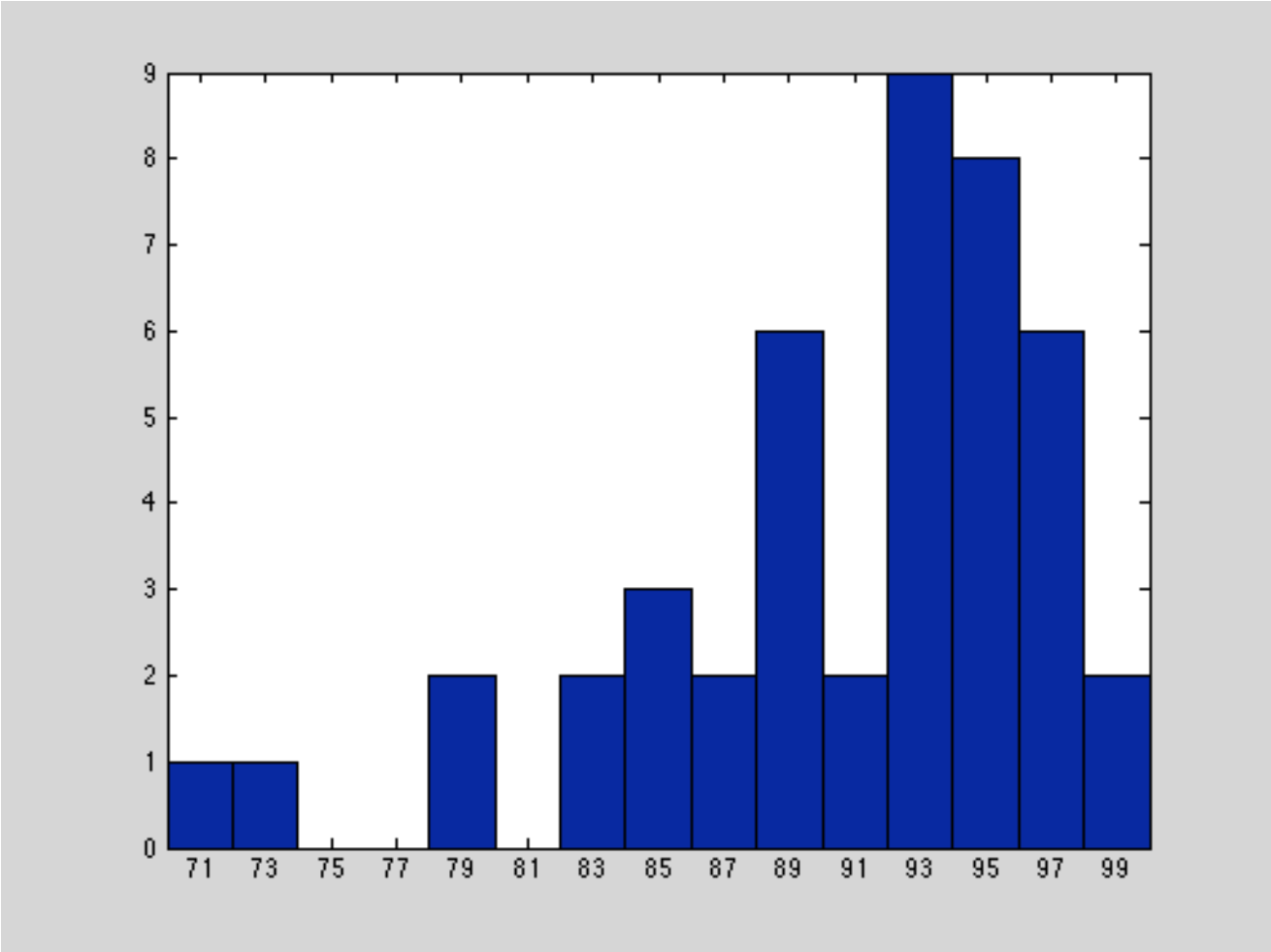
Lecture 11
Simple Linear Regression

Fall 2013

Prof. Yao Xie, yao.xie@isye.gatech.edu
H. Milton Stewart School of Industrial Systems & Engineering
Georgia Tech

Midterm 2

- mean: 91.2
- median: 93.75
- std: 6.5



Meddicorp Sales

Meddicorp Company sells medical supplies to hospitals, clinics, and doctor's offices.

Meddicorp's management considers the effectiveness of a new advertising program.

Management wants to know if the **advertisement** in 1999 is related to **sales**.

Data

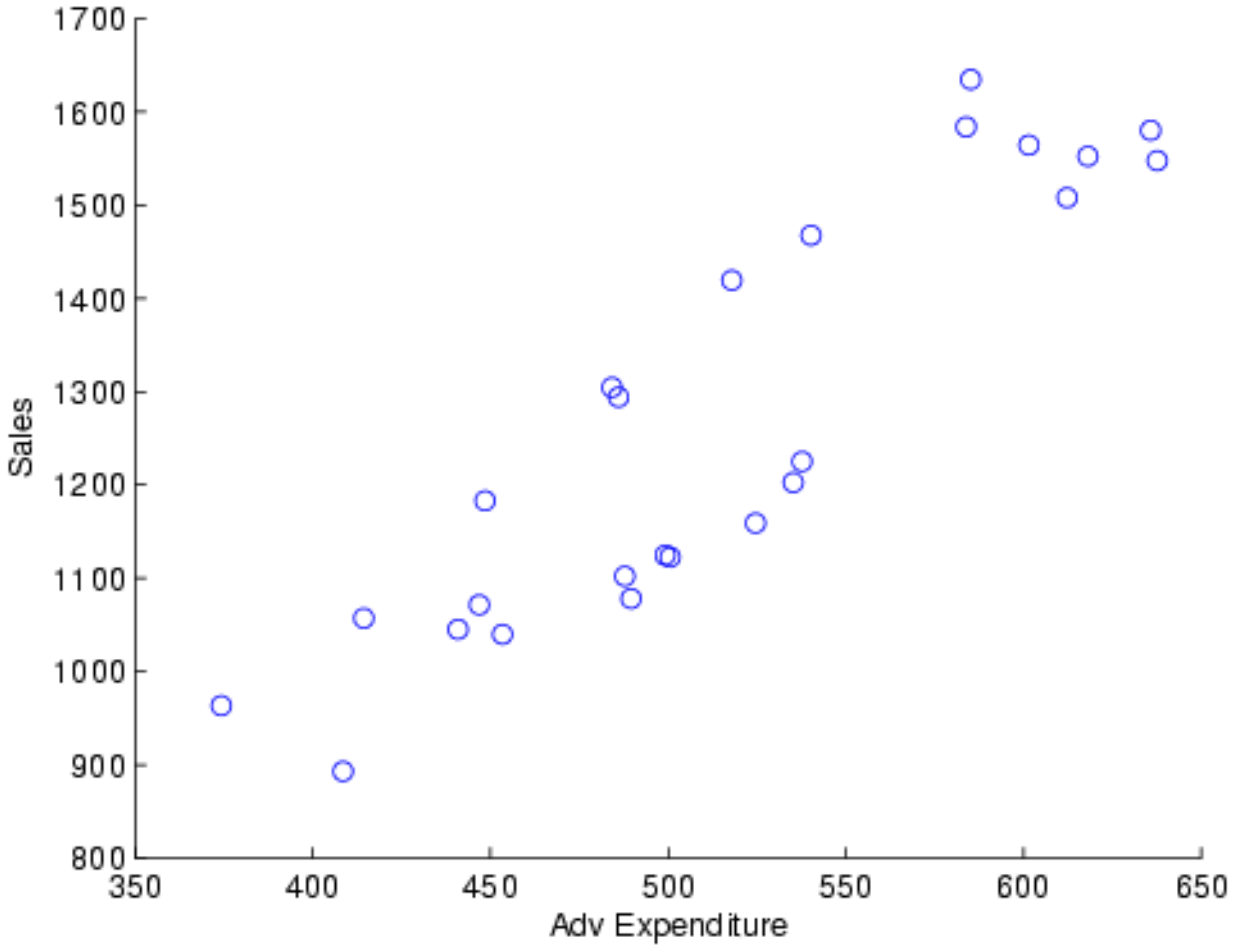
The company observes for 25 offices the yearly sales (in thousands) and the advertisement expenditure for the new program (in hundreds)

	SALES	ADV
1	963.50	374.27
2	893.00	408.50
3	1057.25	414.31
4	1183.25	448.42
5	1419.50	517.88

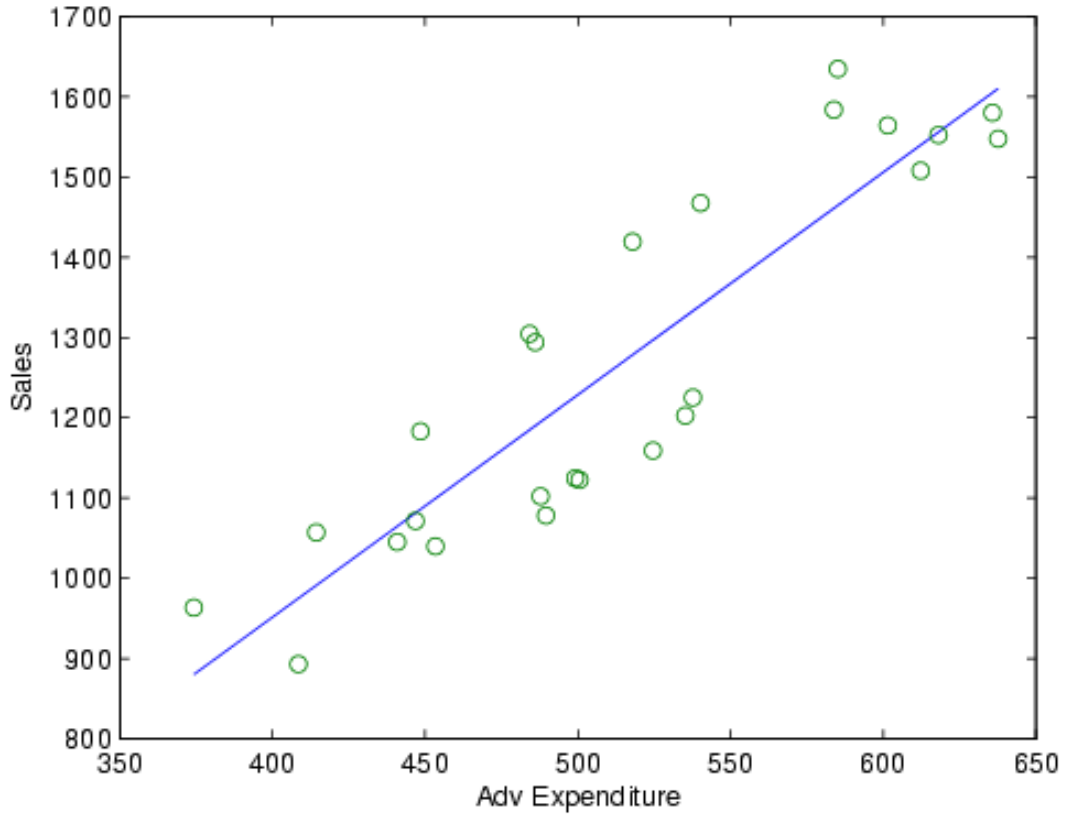
.....

Regression analysis

- Step 1: graphical display of data — scatter plot: sales vs. advertisement cost



- Step 2: find the relationship or association between Sales and Advertisement Cost — Regression



Regression Analysis

- The collection of statistical tools that are used to model and explore relationships between variables that are related in nondeterministic manner is called regression analysis

- Occurs frequently in engineering and science

Scatter Diagram

Many problems in engineering and science involve exploring the relationships between two or more variables.

Regression analysis is a statistical technique that is very useful for these types of problems

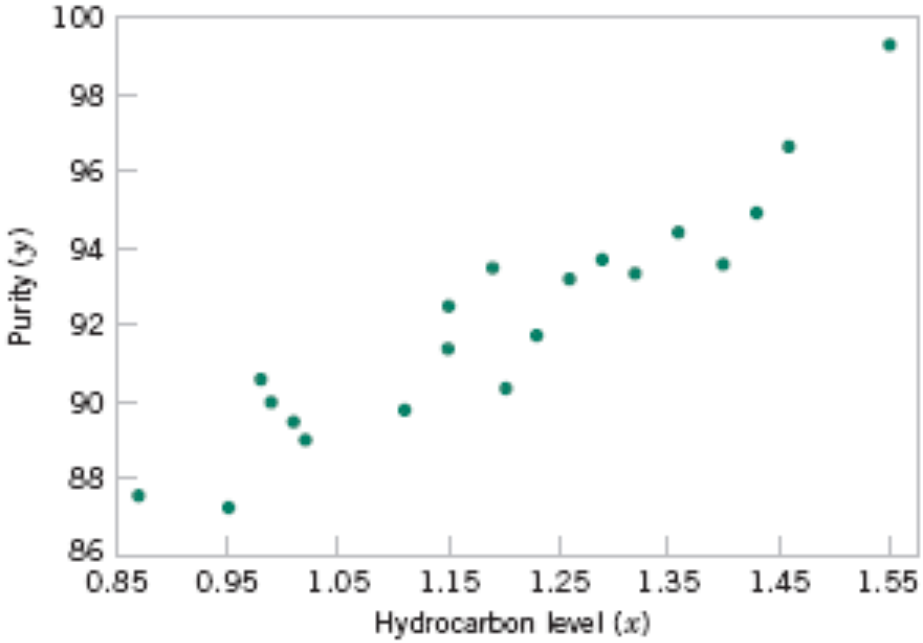


Table 11-1 Oxygen and Hydrocarbon Levels

Observation Number	Hydrocarbon Level x(%)	Purity y(%)
1	0.99	90.01
2	1.02	89.05
3	1.15	91.43
4	1.29	93.74
5	1.46	96.73
6	1.36	94.45
7	0.87	87.59
8	1.23	91.77
9	1.55	99.42
10	1.40	93.65
11	1.19	93.54
12	1.15	92.52
13	0.98	90.56
14	1.01	89.54
15	1.11	89.85
16	1.20	90.39
17	1.26	93.25
18	1.32	93.41
19	1.43	94.98
20	0.95	87.33

$$\hat{\rho} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \times \sum_{i=1}^{n}(y_i - \bar{y})^2}} \qquad -1 \le \hat{\rho} \le 1$$

Basics of Regression

- We observe a **response** or **dependent** variable (Y)

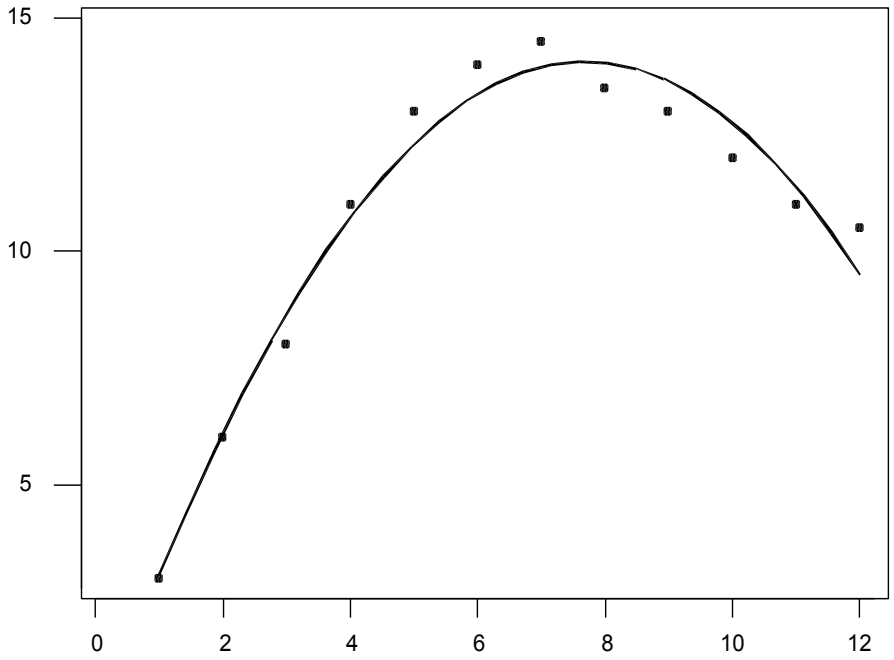
- With each (Y), we also observe **regressors** or **predictors** $\{X_1, \ldots, X_n\}$

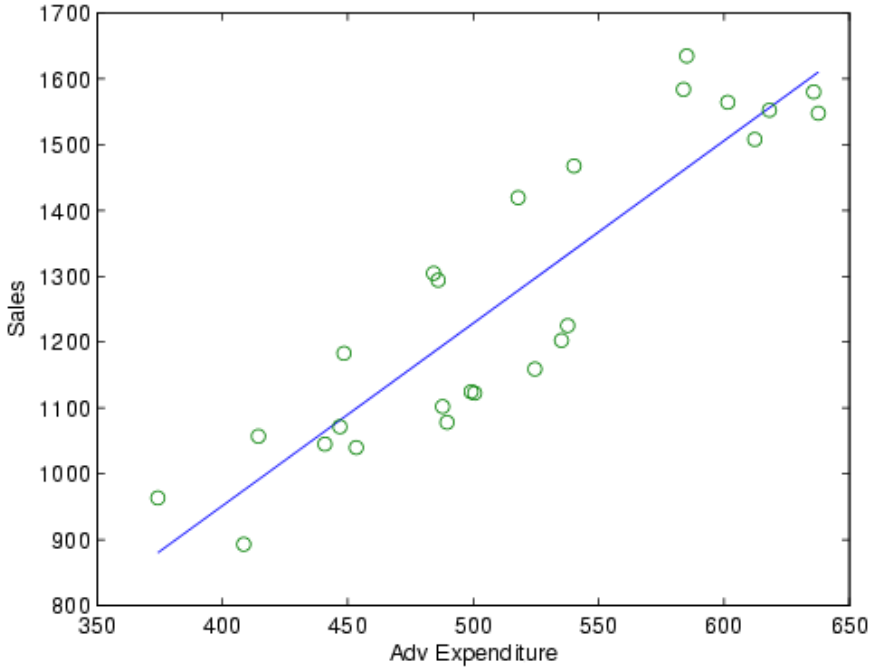
- Goal: determine the mathematical relationship between response variables and regressors

- $$Y = h(X_1, \ldots, X_n)$$

- Function can be non-linear
- In this class, we will focus on the case where Y is a linear function of $\{X_1, \ldots, X_n\}$

$$Y = h(X_1,\ldots,X_n) = \beta_0+\beta_1X_1+\ldots+\beta_nX_n$$

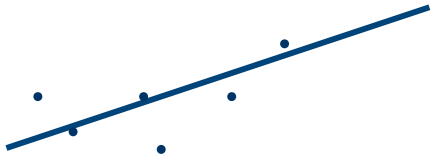




Different forms of regression

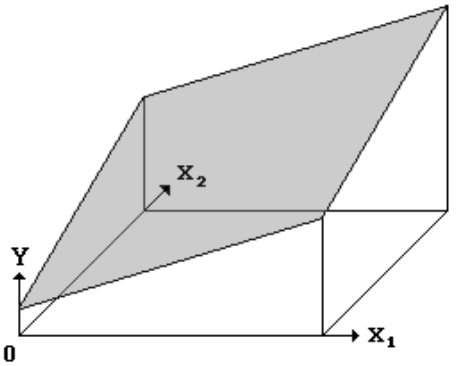
- Simple linear regression

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



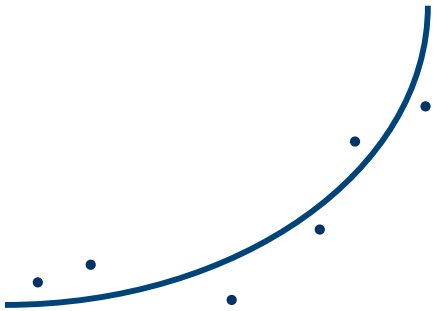
- Multiple linear regression

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$



- Polynomial regression

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \varepsilon$$



Basics of regressions

Which is the RESPONSE and which is the PREDICTOR?

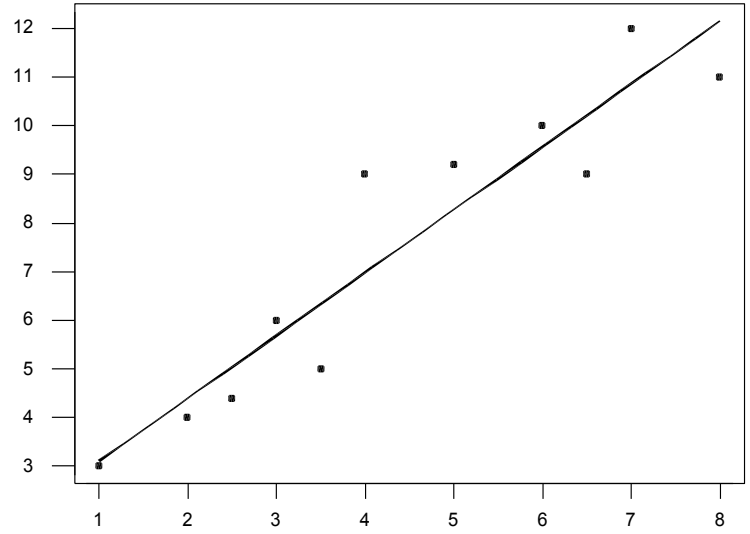
The response or dependent variable varies with different values of the regressor/predictor.

The predictor values are fixed: we observe the response for these fixed values

The focus is in explaining the response variable in association with one or more predictors

Simple linear regression

Our goal is to find the best line that describes a linear relationship:



Find (β_0,β_1) where

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

Unknown parameters:

1. β_0 *Intercept* (where the line crosses y-axis)
2. β_1 *Slope* of the line

Basic idea

a. Plot observations (X,Y)
b. Find best line that follows plotted points

Class activity

1. In the Meddicorp Company example, the **response** is:
   A. Sales **B. Advertisement Expenditure**

2. In the Meddicorp Company example, the **predictor** is:
   A. Sales **B. Advertisement Expenditure**

3. To learn about the association between sales and the advertisement expenditure we can use simple linear regression:
   A. True **B. False**

4. If the association between response and predictor is positive then the slope is
   A. Positive **B. Negative** **C. We cannot identify the slope sign**

Simple linear regression: model

With observed data $\{(X_1,Y_1),\ldots,(X_n,Y_n)\}$, we model the linear relationship

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i,\ i = 1,\ldots,n$$

$E(\varepsilon_i) = 0$

$Var(\varepsilon_i) = \sigma^2$

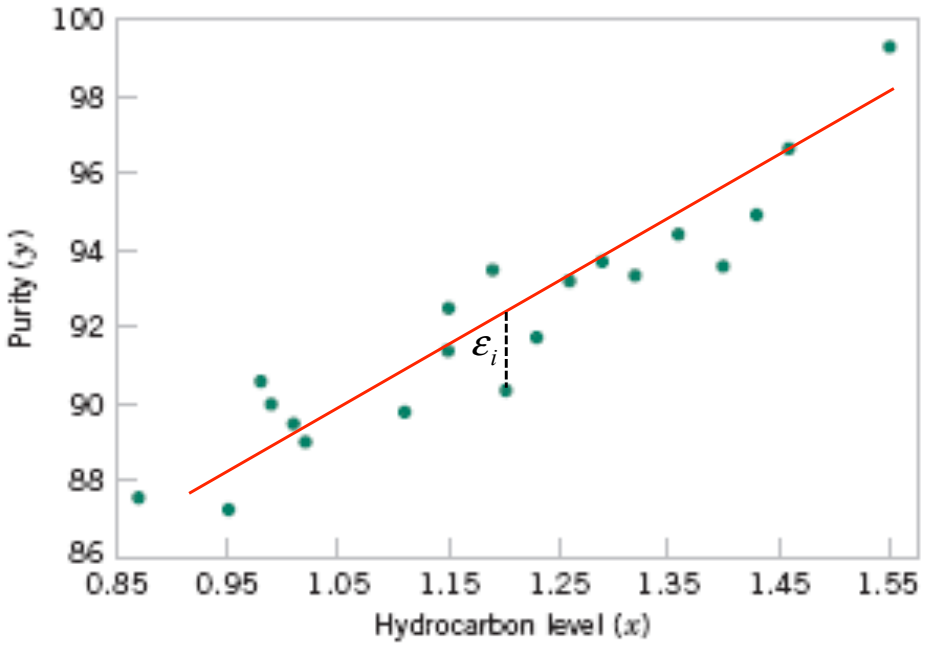
$\{\varepsilon_1,\ldots,\varepsilon_n\}$ are independent random variables

(Later we assume $\varepsilon_i \sim$ Normal)

Later, we will check these assumptions when we check "model adequacy"

Summary: simple linear regression

Based on the scatter diagram, it is probably reasonable to assume that the mean of the random variable Y is related to X by the following simple linear regression model:



Response

Regressor or Predictor

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \qquad i = 1,2,\cdots,n$$

$$\varepsilon_i \sim \mathrm{N}(0, \sigma^2)$$

Intercept

Slope

Random error

where the slope and intercept of the line are called regression coefficients.

•The case of simple linear regression considers a single regressor or predictor x and a dependent or response variable Y.

Estimate regression parameters

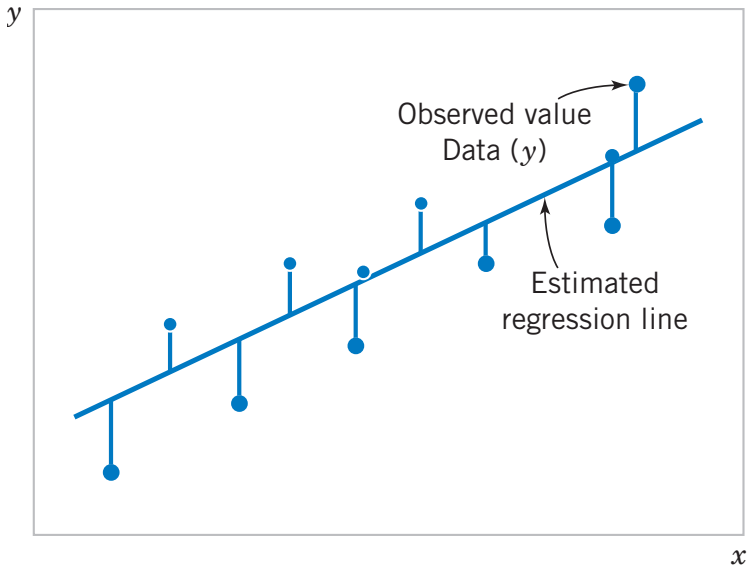
To estimate (β_0,β_1) , we find values that minimize squared error:

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

- derivation: method of least squares

Method of least squares

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \qquad i = 1, 2, \ldots, n$$



To estimate (β_0,β_1) , we find values that minimize squared error:

$$L = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

The least squares estimators of β_0 and β_1, say, $\hat{\beta}_0$ and $\hat{\beta}_1$, must satisfy

$$\left.\frac{\partial L}{\partial \beta_0}\right|_{\hat{\beta}_0,\hat{\beta}_1} = -2\sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

$$\left.\frac{\partial L}{\partial \beta_1}\right|_{\hat{\beta}_0,\hat{\beta}_1} = -2\sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) x_i = 0$$

Least square normal equations

$$n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i$$

$$\hat{\beta}_0 \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} y_i x_i$$

Least square estimates

The **least squares estimates** of the intercept and slope in the simple linear regression model are

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \tag{11-7}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} y_i x_i - \frac{\left(\sum_{i=1}^{n} y_i\right)\left(\sum_{i=1}^{n} x_i\right)}{n}}{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}} \tag{11-8}$$

where $\bar{y} = (1/n)\sum_{i=1}^{n} y_i$ and $\bar{x} = (1/n)\sum_{i=1}^{n} x_i$.

Alternative notation

$$S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n} \qquad (11\text{-}10)$$

$$S_{xy} = \sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x}) = \sum_{i=1}^{n} x_i y_i - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n} \qquad (11\text{-}11)$$

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

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$$

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

Fitted (estimated) regression model

Example: oxygen and hydrocarcon level

Table 11-1 Oxygen and Hydrocarbon Levels

Observation Number	Hydrocarbon Level x (%)	Purity y (%)
1	0.99	90.01
2	1.02	89.05
3	1.15	91.43
4	1.29	93.74
5	1.46	96.73
6	1.36	94.45
7	0.87	87.59
8	1.23	91.77
9	1.55	99.42
10	1.40	93.65
11	1.19	93.54
12	1.15	92.52
13	0.98	90.56
14	1.01	89.54
15	1.11	89.85
16	1.20	90.39
17	1.26	93.25
18	1.32	93.41
19	1.43	94.98
20	0.95	87.33

Question: fit a simple regression model to related purity (y) to hydrocarbon level (x)

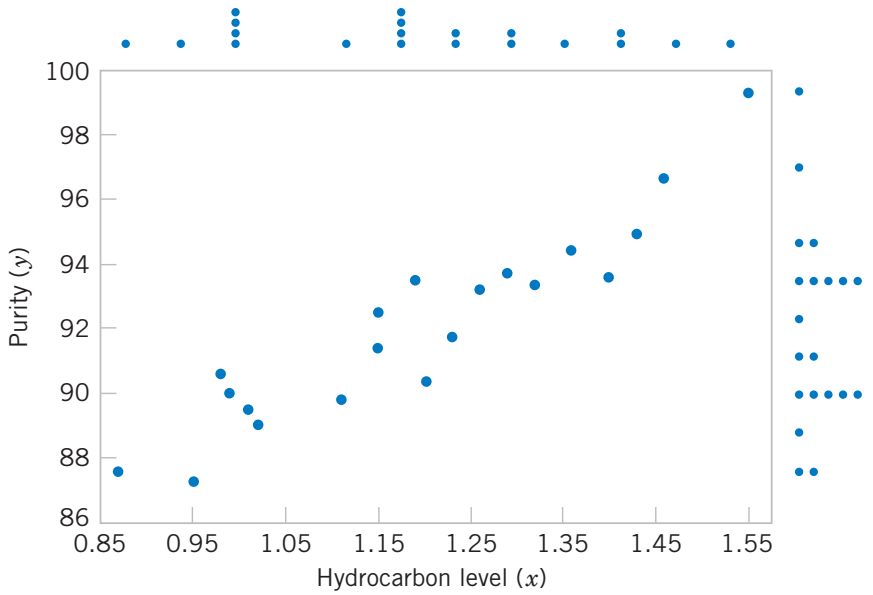


Figure 11-1 Scatter diagram of oxygen purity versus hydrocarbon level from Table 11-1.

$$n = 20 \quad \sum_{i=1}^{20} x_i = 23.92 \quad \sum_{i=1}^{20} y_i = 1{,}843.21$$

$$\bar{x} = 1.1960 \quad \bar{y} = 92.1605$$

$$\sum_{i=1}^{20} y_i^2 = 170{,}044.5321 \quad \sum_{i=1}^{20} x_i^2 = 29.2892$$

$$\sum_{i=1}^{20} x_i y_i = 2{,}214.6566$$

$$S_{xx} = \sum_{i=1}^{20} x_i^2 - \frac{\left(\sum_{i=1}^{20} x_i\right)^2}{20} = 29.2892 - \frac{(23.92)^2}{20}$$

$$= 0.68088$$

and

$$S_{xy} = \sum_{i=1}^{20} x_i y_i - \frac{\left(\sum_{i=1}^{20} x_i\right)\left(\sum_{i=1}^{20} y_i\right)}{20}$$

$$= 2{,}214.6566 - \frac{(23.92)(1{,}843.21)}{20} = 10.17744$$

Therefore, the least squares estimates of the slope and intercept are

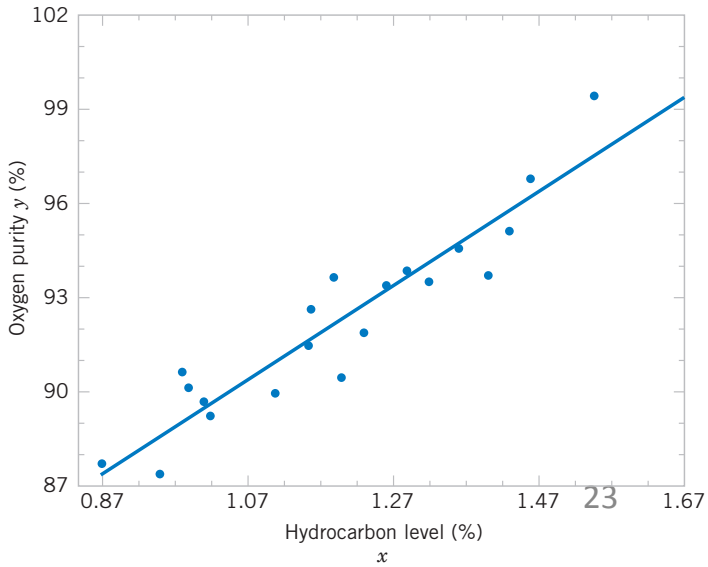
$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{10.17744}{0.68088} = 14.94748$$

and

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 92.1605 - (14.94748)1.196 = 74.28331$$

The fitted simple linear regression model (with the coefficients reported to three decimal places) is

$$\hat{y} = 74.283 + 14.947x$$



Interpretation of regression model

- Regression model

$$\hat{y} = 74.283 + 14.947x$$

$$\hat{y} = 89.23\% \text{ when the hydrocarbon level is } x = 1.00\%.$$

- This may be interpreted as an estimate of the true population **mean** purity when $x = 1.00\%$
- The estimates are subject to error
- later: we will use confidence intervals to describe the error in estimation from a regression model

Estimation of variance

- Using the fitted model, we can estimate value of the response variable for given predictor

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

- Residuals: $r_i = y_i - \hat{y}_i$
- Our model: $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$, $i = 1,\ldots,n$, $Var(\varepsilon_i) = \sigma^2$
- Unbiased estimator (MSE: Mean Square Error)

$$\hat{\sigma}^2 = MSE = \frac{\sum_{i=1}^{n} r_i^2}{n-2}$$

- oxygen and hydrocarcon level example $\hat{\sigma}^2 = 1.18$

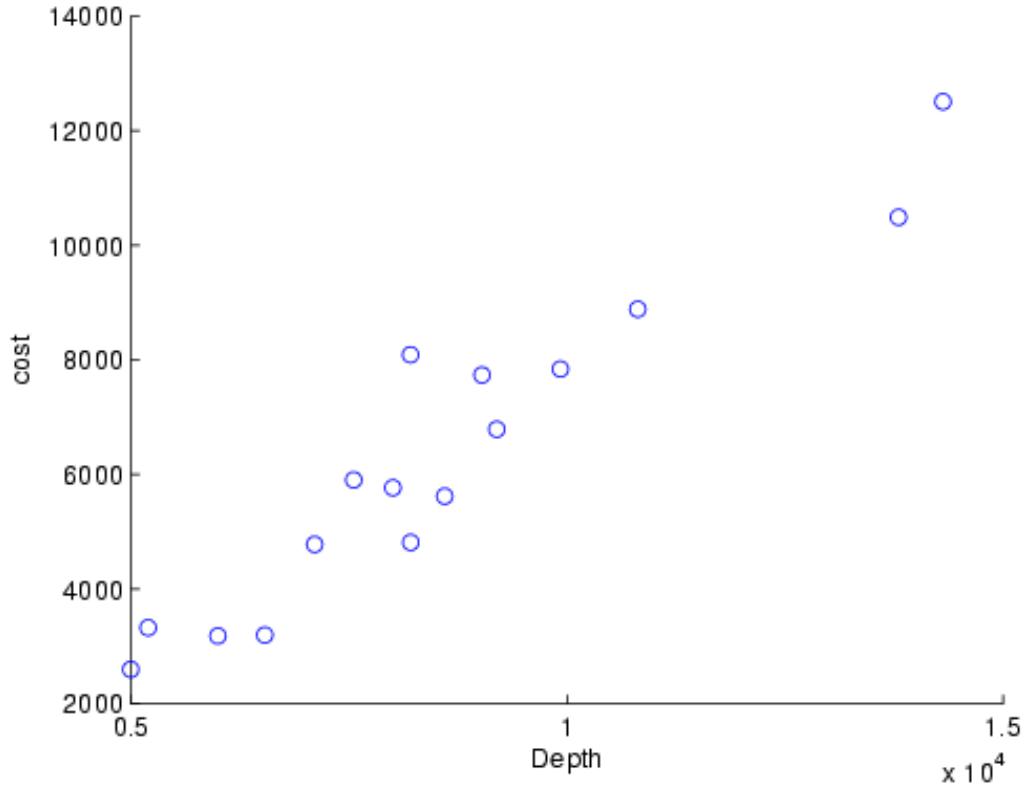
Example: Oil Well Drilling Costs

Estimating the costs of drilling oil wells is an important consideration for the oil industry.

Data: the **total costs** and the **depths** of 16 off-shore oil wells located in Philippines.

Depth	Cost
5000	2596.8
5200	3328.0
6000	3181.1
6538	3198.4
7109	4779.9
7556	5905.6
8005	5769.2
8207	8089.5
8210	4813.1
8600	5618.7
9026	7736.0
9197	6788.3
9926	7840.8
10813	8882.5
13800	10489.5
14311	12506.6

- Step 1: graphical display of the data



- R code: plot(Depth, Cost, xlab= “Depth”, ylab = “Cost”)

Class activity

1. In this example, the **response** is:

A. The drilling cost **B. The well depth**

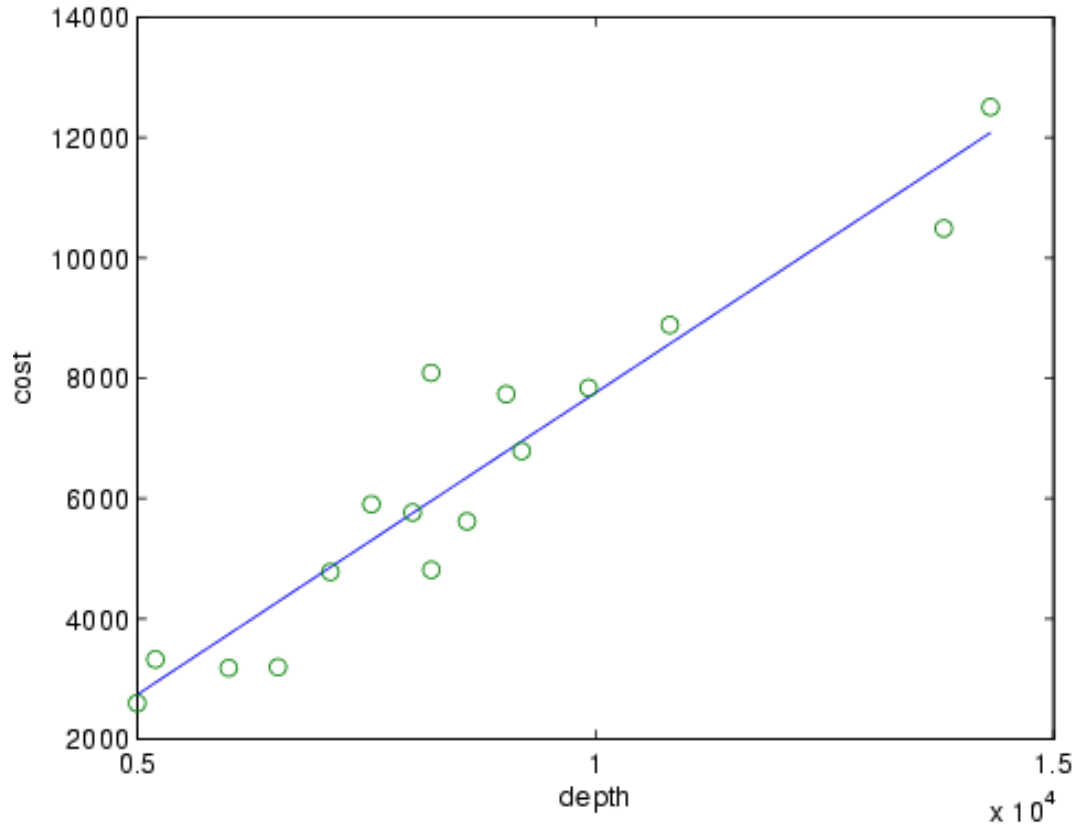
2. In this example, the **dependent** variable is:

A. The drilling cost **B. The well depth**

3. Is there a linear association between the drilling cost and the well depth?

A. Yes and positive B. **Yes and negative** **C. No**

- Step 2: find the relationship between Depth and Cost



Results and use of regression model

1. Fit a linear regression model:

Estimates (β_0,β_1) are (-2277.1, 1.0033)

2. What does the model predict as the cost increase for an additional depth of 1000 ft?

If we increase X by 1000, we increase Y by $1000\beta_1$ = \$1003

3. What cost would you predict for an oil well of 10,000 ft depth?

X = 10,000 ft is in the range of the data, and

estimate of the line at x=10,000 is $\hat{\beta}_0 + (10{,}000)\hat{\beta}_1$ = -2277.1 + 10,033 = \$7753

4. What is the estimate of the error variance? Estimate $\sigma^2 \approx 774{,}211$

5.What could you say about the cost of an oil well of depth 20,000 ft?

X=20,000 ft is much greater than all the observed values of X

We should not extrapolate the regression out that far.

Summary

- Simple linear regression

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

- Estimate coefficients from data: method of least squares

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

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$$

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

Fitted (estimated) regression model

- Estimate of variance