# STAT 571A — Advanced Statistical Regression Analysis

## Chapter 1 NOTES

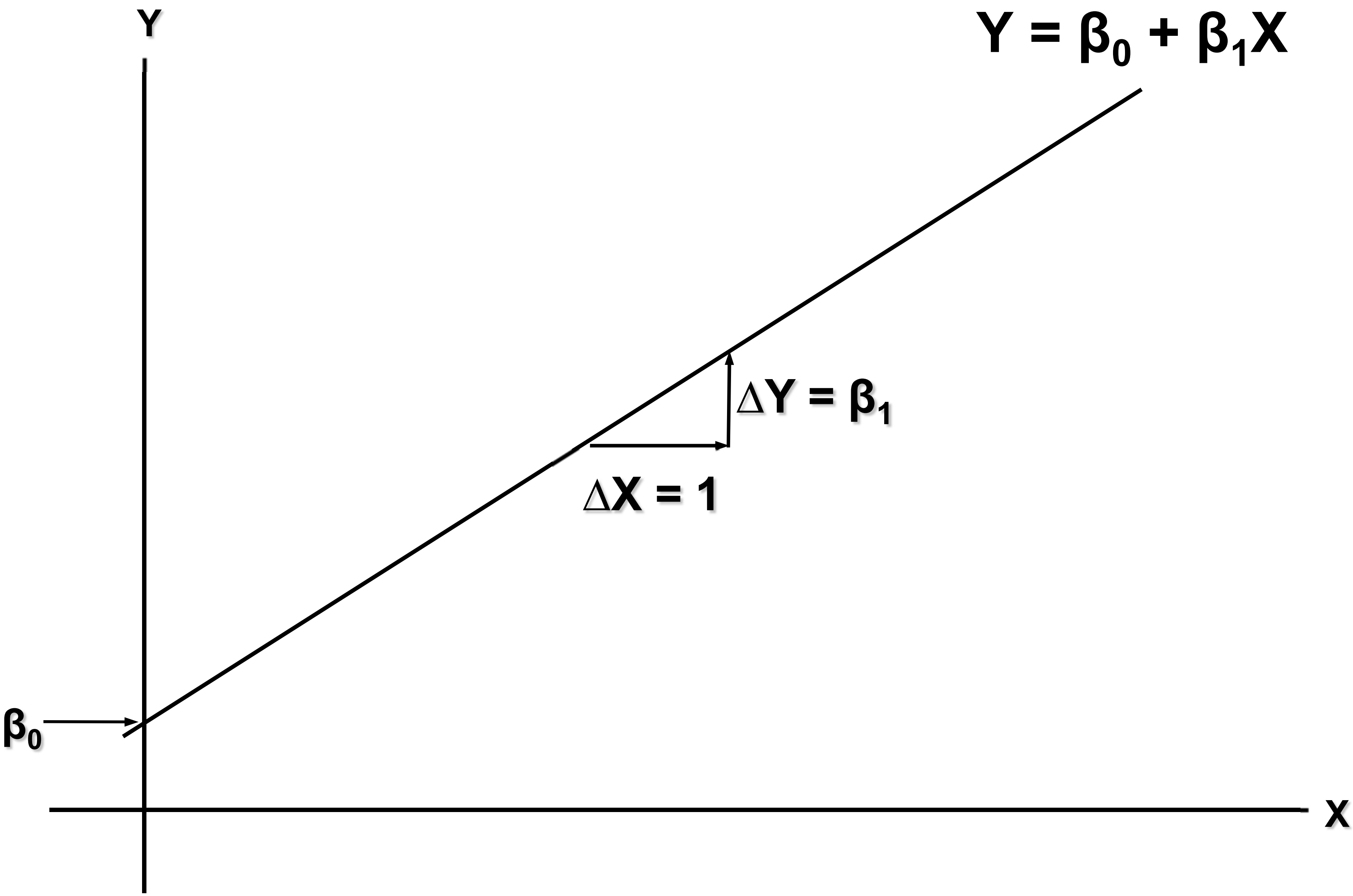
## Linear Regression with One Predictor Variable

# Linear Regression

- **Linear regression** is concerned with estimating relationships between a response variable, Y and an explanatory/predictor variable, X
- The simple linear (straight-line) relationship is

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

where $\beta_1$ is the **slope** ("rise-over-run") and $\beta_0$ is the **Y-intercept**.


Y
$Y = \beta_0 + \beta_1 X$
$\Delta Y = \beta_1$
$\Delta X = 1$
$\beta_0$
X

# Data model

**In practice, we observe data pairs $(X_i,Y_i)$, $i = 1,\ldots, n$, usually with observational/experimental error. Model this as**

$$Y_i = (\beta_0 + \beta_1 X_i) + \varepsilon_i \qquad (1.1)$$

**i.e., observed Y =**
**(simple linear model) + random error term**

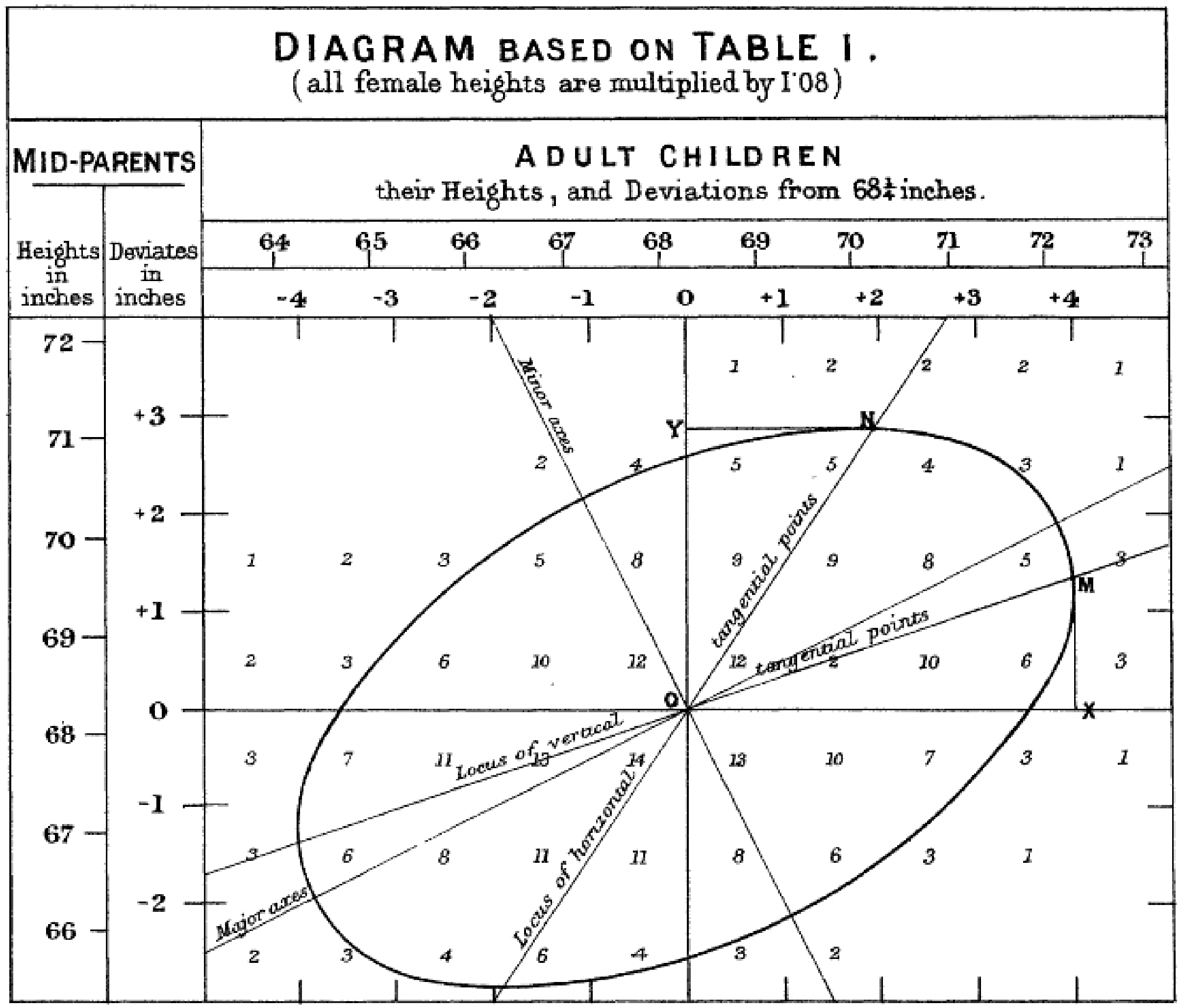
**Can think of this as Y = (signal) + noise.**

## Historical Note

- **As the text indicates, the originator of the term "regression" was Sir Francis Galton**
- **He plotted X = 'Mid-Parent' height (as the <u>vertical</u> axis), and Y = Adult child height (as the <u>horizontal</u> axis), and found that as adults, the offspring "regressed" to more central heights.**
- **Galton published the data in 1886 (in the *J. Anthropol. Inst. Gr. Brit. & Ireland*).**

# Galton's 1886 "Plate X"

## Galton's original plot:



# Model Assumptions

- For our **simple linear model**, we assume
  - $X_i$ is a known constant
  - $\beta_0$ and $\beta_1$ are unknown parameters
  - $E[\varepsilon_i] = 0$ for all i
  - $\sigma^2[\varepsilon_i] = \sigma^2$ (constant) for all i
  - $\sigma[\varepsilon_i,\varepsilon_j] = 0$ (zero!) for all $i \neq j$
- Notice: $\varepsilon_i$ is a random variable, thus so is $Y_i$.

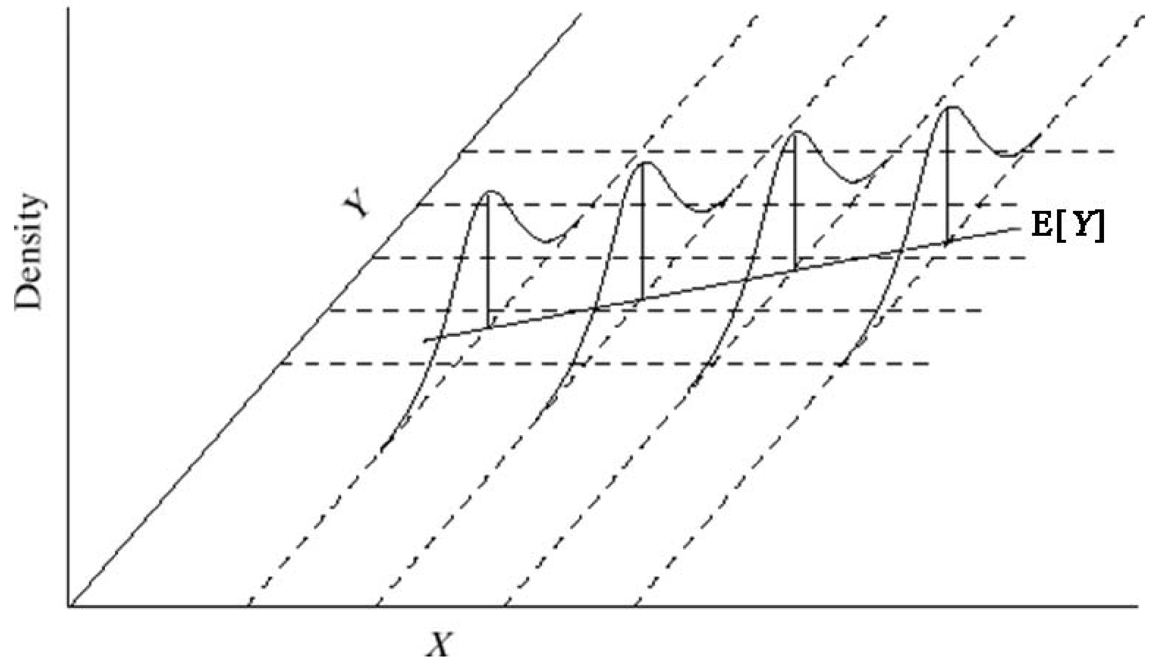
# Model Impact on $Y_i$

**We find**

- $E[Y_i] = E[\beta_0 + \beta_1 X_i + \varepsilon_i] \;=\; E[\beta_0 + \beta_1 X_i] + E[\varepsilon_i]$
  $= \;\beta_0 + \beta_1 X_i + E[\varepsilon_i] \;=\; \beta_0 + \beta_1 X_i + 0$
  $= \;\beta_0 + \beta_1 X_i$
  (since $\beta_0 + \beta_1 X_i$ is nonrandom).
  We say $E[Y_i] = \beta_0 + \beta_1 X_i$ is the **mean response**.
- $\sigma^2[Y_i] = \sigma^2[\beta_0 + \beta_1 X_i + \varepsilon_i] \;=\; \sigma^2[\varepsilon_i] \;=\; \sigma^2$
  (again, since $\beta_0 + \beta_1 X_i$ is nonrandom).
- $\sigma[Y_i, Y_j] = \cdots = 0$ for all $i \neq j$

# Probability Model

- **Graphically, there is some probability function for $Y_i$ resting at each $X_i$:**




- **Notice that each prob. function has the same variance!**

# Alternative Formulations

- Alternative (but, essentially equivalent) formulations for the simple linear model include:
  - $Y_i = \beta_0^* + \beta_1(X_i - \bar{X}) + \varepsilon_i$

    (so $\beta_0^* = \beta_0 + \beta_1\bar{X}$) for $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$
  - $Y_i = \beta_0 X_{0i} + \beta_1 X_{1i} + \varepsilon_i$

    where $X_{0i} = 1$ and $X_{1i} = X_i$ for all i.
- These can be useful in select cases.

# Data Generation Mechanisms

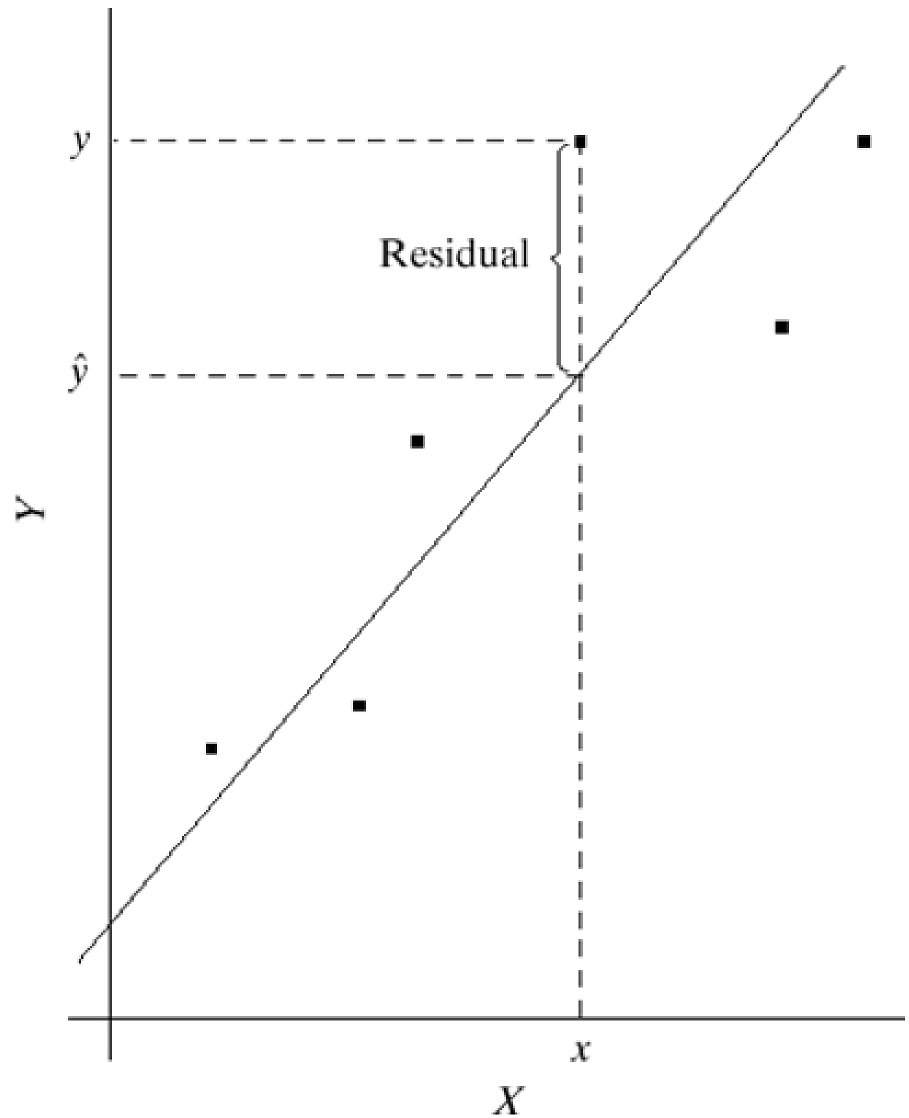
- **Note that we can observe the data pairs in two fundamentally different ways:**
  - **Observational study: data are recorded without strict experimental controls**
    **$\Rightarrow$ harder to relate cause & effect**
  - **Experimental study: data from controlled experiments**
    **$\Rightarrow$ inference is truer but conduct is more expensive**
- **We consider both forms in our data examples.**

# Least Squares (LS)

- Given data pairs $(X_i, Y_i)$, $i = 1, \ldots, n$, we estimate $\beta_0$ and $\beta_1$ using the method of **least squares** (LS) (from Appendix A.5).
- Denote these as $b_0$ and $b_1$, resp. (Equations will follow.)
- Then, the **fitted value** is $\hat{Y}_i = b_0 + b_1 X_i$. Find these by minimizing $Q = \sum(Y_i - \hat{Y}_i)^2$.
- The corresp. **residual** is $e_i = Y_i - \hat{Y}_i$. We want $\hat{Y}_i$ to be as close to $Y_i$ as possible.

# LS Line and Residuals

- Graphically, the idea is something like this →
- Points are data pairs; line is $b_0 + b_1X$



# ‘Normal’ Equations

- To minimize Q with resp. to $b_0$ and $b_1$, via calculus (see pp.17-18), we find the LS estimators solve the system of equations

$$\sum Y_i = nb_0 + b_1 \sum X_i$$

$$\sum X_i Y_i = b_0 \sum X_i + b_1 \sum X_i^2$$

- These are called the **normal equations** for the LS estimators.

# Gauss-Markov Theorem

- A result known as the **Gauss-Markov Theorem** motivates use of the LS estimators: under model (1.1), the LS solutions for $b_0$ and $b_1$ are (a) unbiased and (b) have min. variance among all unbiased linear estimators.
- (a) says that $E[b_j] = \beta_j$ for $j=0,1$
- (b) says $\sigma^2[b_0]$ and $\sigma^2[b_1]$ are minimized.

# LS Solutions: Slope

**The LS solution has, in fact, a closed form. First, the slope parameter is**

$$b_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

**or also**

$$b_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})Y_i}{\sum_{m=1}^{n}(X_m - \bar{X})^2} = \sum_{i=1}^{n} k_i Y_i$$

$$\text{for} \quad k_i = \frac{(X_i - \bar{X})}{\sum_{m=1}^{n}(X_m - \bar{X})^2}$$

# LS Solutions: Intercept

**Then, the intercept is expressed conveniently as**

$$b_0 = \overline{Y} - b_1\overline{X}$$

**Indeed, $b_0$ can also be written in the form**
**$b_0 = \sum_{i=1}^{n} k_i Y_i$ (...for a different set of $k_i$s)**

# Example CH01TA01 (p. 19)

X = lot size of refrig. parts production
Y = hours worked (labor)
at the Toluca Manufacturing Co.

The data are in Table 1.1. We could just 'do the math':

$$\sum_{i=1}^{n}(X_i - \bar{X})Y_i = 70690 \quad \& \quad \sum_{i=1}^{n}(X_i - \bar{X})^2 = 19800$$

$$\text{so } b_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})Y_i}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = \frac{70690}{19800} = 3.5702$$

# Example CH01TA01 (cont'd)

**Also, $\bar{Y}$ = 312.28 and $\bar{X}$ = 70, so**

**$b_0 = \bar{Y} - b_1\bar{X} = 312.28 - (3.5702)(70)$**
**$= 62.37$.**

**The prediction equation for $\hat{Y}_i = b_0 + b_1X_i$ is then $\hat{Y}_i = 62.37 + 3.5702X_i$.**

**But, it's so much easier in R $\longrightarrow$**

# Example CH01TA01 (cont'd)

**Toluca Co. example: LS fit for simple linear model via R:**

```
> X = c(80, 30, ... , 70)
> Y = c(399, 121, ... , 323)
> CH01TA01.lm = lm( Y ~ X )
> summary( CH01TA01.lm )
```

# summary() output for Toluca example

```
Call:
lm(formula = Y ~ X)

Residuals:
    Min      1Q  Median      3Q     Max
-83.876 -34.088  -5.982  38.826 103.528

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   62.366     26.177   2.382   0.0259 *
X              3.570      0.347  10.290 4.45e-10 ***
```

**LS estimates of the regr. parameters highlighted in red here.**

# Example CH01TA01 (cont'd)

**ALWAYS PLOT THE DATA!**

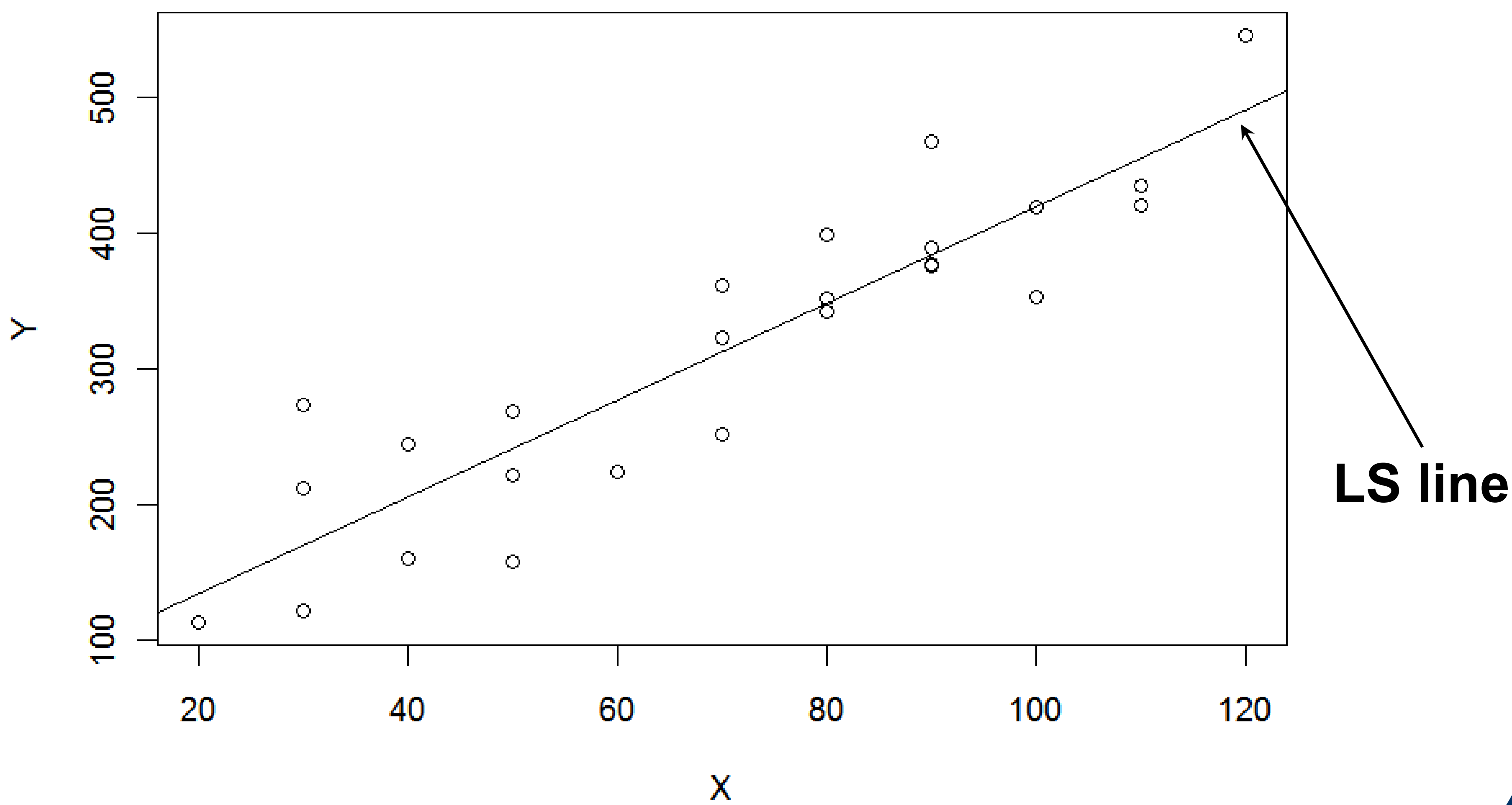
**Toluca example. Scatterplot plot in R:**

```
> plot( Y ~ X )

> abline( CH01TA01.lm )
```

`abline()` **command overlays line from LS fit**

# Toluca Data Scatterplot



# Example CH01TA01 (p.22)

- What about predicting the fitted value at a given $X_i$?
- For instance, at $X_1 = 80$ we see

  $\hat{Y}_1 = b_0 + b_1(80) = 62.37 + (3.5702)(80)$
  $= 347.98$

  The corresp. residual is
  $e_1 = Y_1 - \hat{Y}_1 = 399 - 347.98 = 51.02.$
  (See Table 1.2.)

# Consequences of the LS Fit

**The LS estimators produce the following, interesting, mathematical consequences:**

- $\sum e_i = 0$ (always!)
- $\sum e_i^2$ is a minimum (since it's LS)
- $\sum Y_i = \sum \hat{Y}_i$
- $\sum e_i X_i = 0$ (weighted sum of $e_i$'s is zero)
- $\sum e_i \hat{Y}_i = 0$
- $\hat{Y}(\overline{X}) = b_0 + b_1 \overline{X} = \overline{Y}$

# Estimating $\sigma^2$

- How to estimate a variance?
- Think of the single-sample case. For $Y_1,\ldots,Y_n$, we use the sample variance:

$$S^2 = \frac{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}{n-1} \quad \longleftrightarrow \quad \frac{\text{a “sum of squares”}}{\text{degrees of freedom}}$$

(dissection)

- Now do the same thing with the simple linear model, but replace $\overline{Y}$ with $\hat{Y}_i$ $\longrightarrow$

# SSE

The resulting sum of squares is called the **Sum of Squared Errors**, or SSE:

$$SSE = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} e_i^2$$

(Notice that it's just the sum of squared residuals, so it's also called the **Residual Sum of Squares**.)

This has d.f. = n – 2. (Why? Think of it as "# observations" – "# fitted components")

# MSE

Now, divide SSE by its d.f.:

$$\text{MSE} = \frac{\text{SSE}}{\text{d.f.}} = \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n-2}$$

We call this the **Mean Squared Error**, or MSE.

Under (1.1), can show that $E[\text{MSE}] = \sigma^2$ (unbiased!).

To estimate $\sigma$, use the **root mean squared error**: $\sqrt{\text{MSE}}$

# Normal Error Model

- Let's add one last model component: impose a formal distribution on the error terms in $\varepsilon_i$.
- Recall: $Y_i = (\beta_0 + \beta_1 X_i) + \varepsilon_i$.
- Now, let $\varepsilon_i \sim$ i.i.d. $N(\ E[\varepsilon_i], \sigma^2[\varepsilon_i]\ )$. Since we already specified $E[\varepsilon_i] = 0$ and $\sigma^2[\varepsilon_i] = \sigma^2$, this yields

$$\varepsilon_i \sim \text{i.i.d. } N(0,\sigma^2).$$

# Maximum Likelihood

- The combination of the simple linear model in (1.1) with normal errors is called a Simple Linear Regression (SLR) model.
- By imposing a formal probability distribution into the model, a form of estimation known as Maximum Likelihood is available.
- We use ML occasionally. For now, know that ML estimators and LS estimators for the SLR model actually coincide (i.e., they're identical).