Ordinary Least Squares:

Linear Regression

$$y = \alpha + \beta x + u$$

Ch.2 Ordinary Least Squares

0. Terminology & Data
1. Describing Data
2. Ordinary Least Squares
   - The Derivation
   - Algebraic Properties
3. Goodness of Fit

0. Some Terminology

- In the simple linear regression model, where $y = \alpha + \beta x + u$, we typically refer to y as the
  - Dependent Variable,
  - Explained Variable,
  - Response Variable, or
  - Regressand

Some Terminology *cont.*

- In the simple linear regression of *y* on *x*, we typically refer to *x* as the
  - Independent Variable,
  - Explanatory Variable,
  - Control Variable, or
  - Regressor

Types of Data

Cross-sectional data

- Each observation is a new individual, firm, etc. with information at a point in time

Time Series data

- Time series data has a separate observation for each time period – e.g. stock prices

Panel data

- Panel data follow the same random individual observations over time

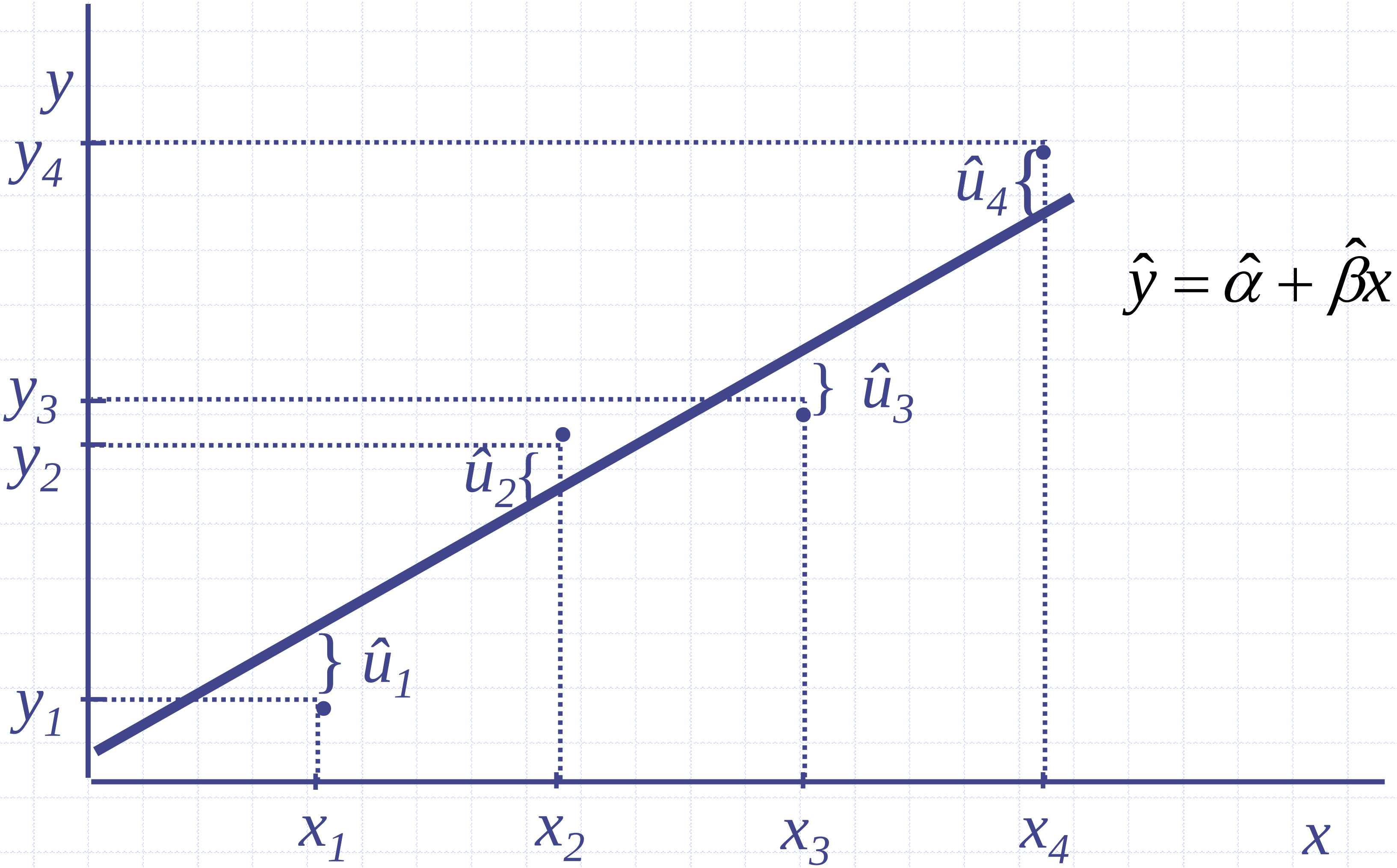
1. Describing Data

- Population & Sample
- Numerical Measures
  - Measures of location
    - Sample mean (2.2)
  - Measures of dispersion
    - Sample Variance (2.3) & Sample s.d. (2.4)
  - Correlation analysis
    - Sample Covariance (2.5)

2. Ordinary Least Squares (OLS)

- Intuitively, OLS is fitting a line through the sample points such that the sum of squared residuals is as small as possible, hence the term least squares
- The *residual*, $\hat{u}$, is the difference between the fitted line (sample regression function) and the sample point

Sample regression line, sample data points and the associated residual terms (fig.2.2)



Deriving OLS Estimators

- Given the intuitive idea of fitting a line, we can set up a formal minimization problem
- That is, we want to choose our parameters such that we minimize the following,

$$\sum_{i=1}^{n}(\hat{u}_i)^2 = \sum_{i=1}^{n}\left(y_i - \hat{\alpha} - \hat{\beta}x_i\right)^2 \qquad (2.12')$$

Deriving OLSE, *continued*

- If one uses calculus to solve the minimization problem for the two parameters, you obtain the following first order conditions.

$$\sum_{i=1}^{n}\left(y_i - \hat{\alpha} - \hat{\beta}x_i\right) = 0 \qquad (2.15'\text{a})$$

$$\sum_{i=1}^{n} x_i\left(y_i - \hat{\alpha} - \hat{\beta}x_i\right) = 0 \qquad (2.15'\text{b})$$

More Derivation of OLS

- Given the definition of a sample mean, and properties of summation, we can rewrite the first condition as follows

$$\bar{y} = \hat{\alpha} + \hat{\beta}\bar{x}, \quad \Rightarrow \quad \hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} \qquad (2.24)$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \qquad with \sum_{i=1}^{n}(x_i - \bar{x})^2 > 0 \qquad (2.23)$$

Regression Analysis

- The regression equation:

$$\hat{y} = \hat{\alpha} + \hat{\beta}x \qquad (2.17)$$

where y is the average predicted value for any x.

  - α is the y intercept. It is the estimated y value when $x = 0$
  - β is the slope of the line, or the average change in y for each change of one unit in x

Algebraic Properties of OLS

- The sum of the OLS residuals is zero. Thus, the sample average of the OLS residuals is zero as well.
- The sample covariance between the regressors and the OLS residuals is zero.
- The OLS regression line always goes through the mean of the sample.

Algebraic Properties *(precise)*

$$\sum_{i=1}^{n} \hat{u}_i = 0 \text{ and thus, } \frac{1}{n}\sum_{i=1}^{n} \hat{u}_i = 0$$

$$\sum_{i=1}^{n} x_i \hat{u}_i = 0$$

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

More terminology

We can think of each observation as being made up of an explained part, and an unexplained part, $y_i = \hat{y}_i + \hat{u}_i$ We then define the following:

$\sum (y_i - \bar{y})^2$ is the total sum of squares (SST)

$\sum (\hat{y}_i - \bar{y})^2$ is the explained sum of squares (SSE)

$\sum \hat{u}_i^2$ is the residual sum of squares (SSR)

Then, **SST = SSE + SSR**

3. Goodness-of-Fit

- How do we think about how well our sample regression line fits our sample data?
- The fraction of the total sum of squares (SST) that is explained by the model is called the R-squared of regression
- $R^2 = SSE/SST = 1 – SSR/SST$