Supervised Learning

Supervised vs. Unsupervised Learning

- Supervised learning (classification)
  - Supervision: The training data (observations, measurements, etc.) are accompanied by **labels** indicating the class of the observations
  - New data is classified based on the training set
- Unsupervised learning (clustering)
  - The class labels of training data is unknown
  - Given a set of measurements, observations, etc. with the aim of establishing the existence of classes or clusters in the data

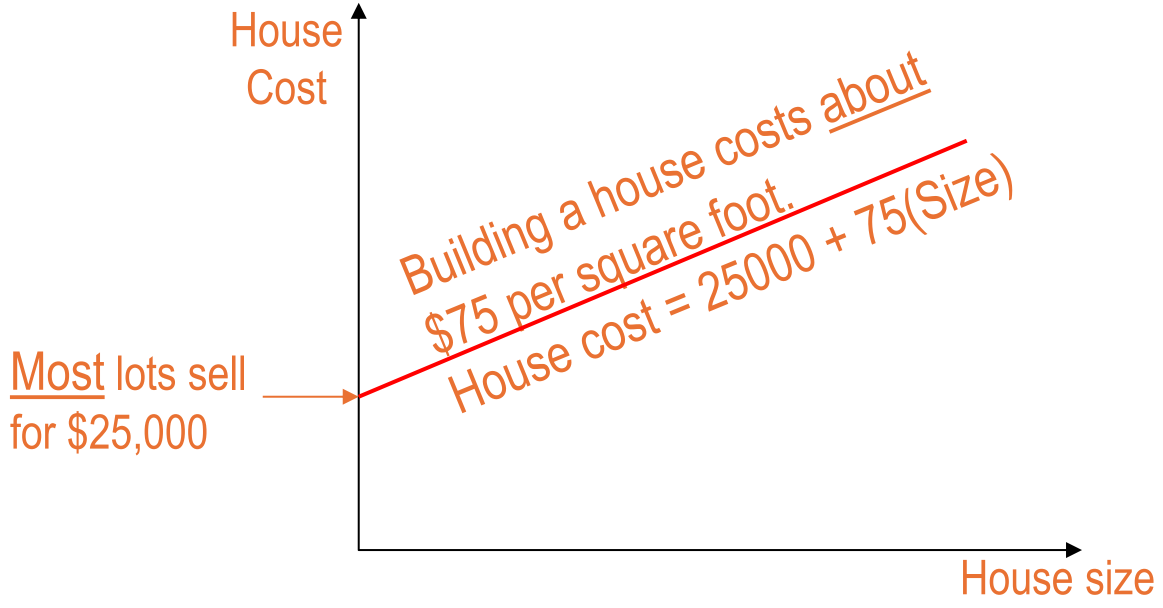
Prediction: Classification vs. Numeric Prediction

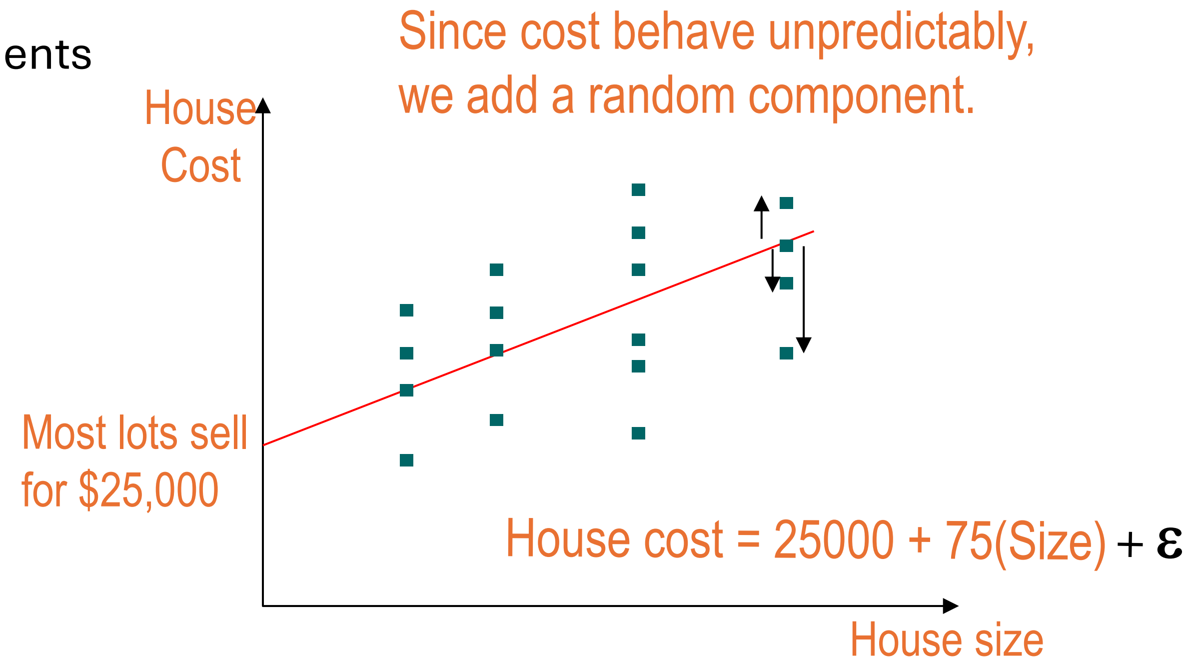
- Classification
  - predicts categorical class labels (discrete or nominal)
  - classifies data (constructs a model) based on the training set and the values (class labels) in a classifying attribute and uses it in classifying new data
  - E.g. Credit/loan approval, Medical diagnosis: if a tumor is cancerous or benign, Fraud detection: if a transaction is fraudulent or not, etc.
  - **Algorithms:** Decision trees, support vector machines (SVMs), Naive Bayes.
- Numeric Prediction
  - models continuous-valued functions, i.e., predicts unknown or missing values
  - E.g. Predicting house prices, Forecasting stock market values, Estimating temperature, etc.
  - **Algorithms:** Linear regression, polynomial regression, support vector regression (SVR).

Simple linear regression

It is statistical method that allows us to summarize and study relationships between two continuous (quantitative) variables:

- One variable, denoted *x*, is regarded as the **predictor**, **explanatory**, or **independent** variable.
- The other variable, denoted *y*, is regarded as the **response**, **outcome**, or **dependent** variable.
- We will examine the relationship between quantitative variables x and y via a mathematical equation.
- The model has a deterministic and a statistical components





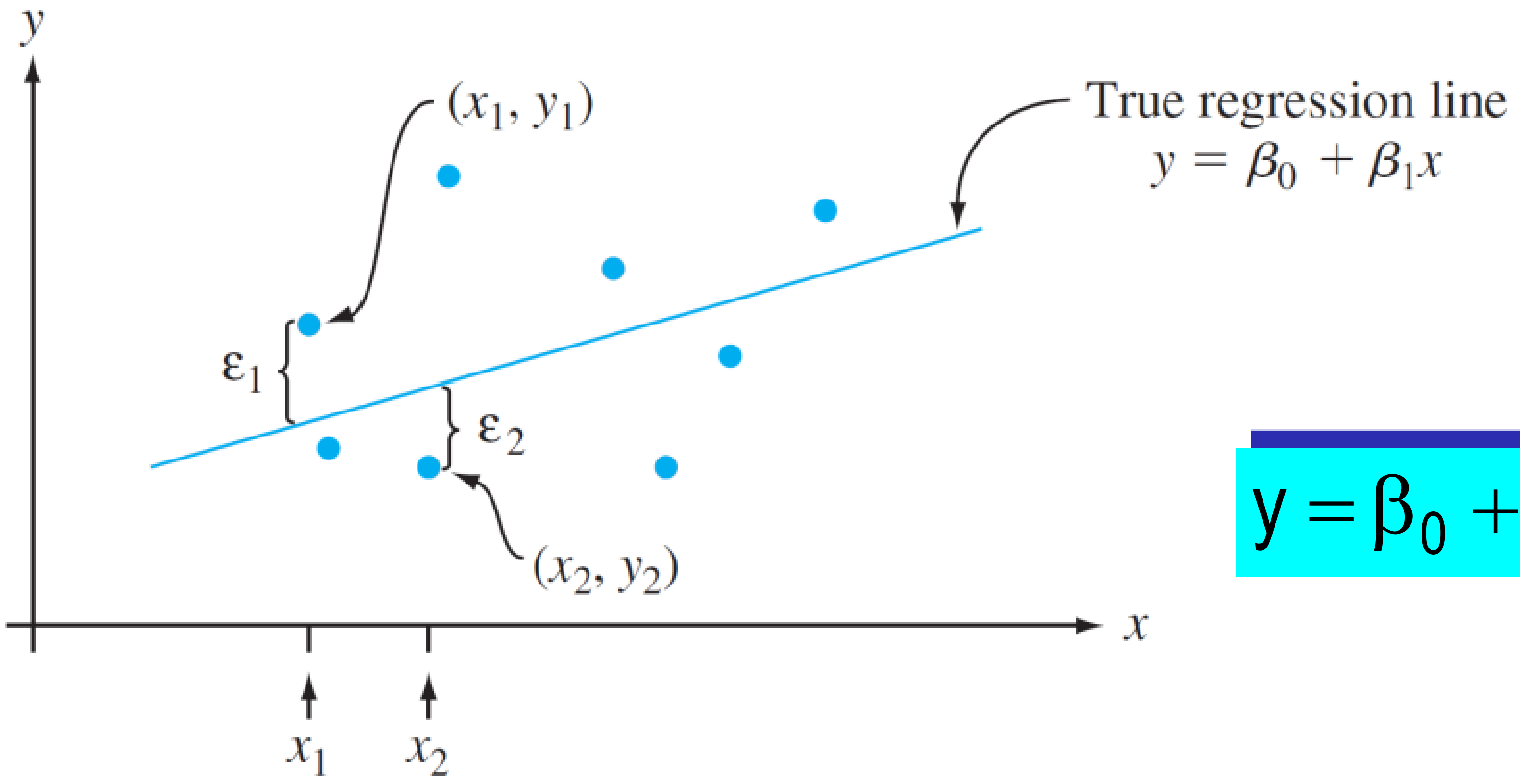
Simple linear regression

- The simplest deterministic mathematical relationship between two variables x and y is a linear relationship: $y = \beta_0 + \beta_1 x$. (True regression line)
- The objective is to develop an equivalent linear probabilistic model.
- If the two (random) variables are probabilistically related, then for a fixed value of x, there is uncertainty in the value of the second variable.
- So, we assume $y = \beta_0 + \beta_1 x + \varepsilon$, where ε is a random variable.

$$b_1 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \qquad b_0 = \bar{y} - b_1\bar{x}$$

Simple linear regression

- The points (x1, y1), …, (xn, yn) resulting from n independent observations will then be scattered about the true regression line:



$$y = \beta_0 + \beta_1 x + \varepsilon$$

Simple linear regression

Estimating Model parameters:

- The values of $\beta_{0,}$ β_1 and ε will almost never be known to an investigator.
- Instead, sample data consists of **n** observed pairs $(x_1, y_1), \ldots, (x_n, y_n)$, from which the model parameters and the true regression line itself can be estimated.
- Where $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ for $i = 1, 2, \ldots, n$ and the ***n*** *deviations* $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ are independent r.v.' s.
- Aim is to find the **Best Fit Line:** the sum of the squared vertical distances (deviations) from the observed points to that line is as small as it can be.

Simple linear regression

The sum of squared vertical deviations from the points $(x_1, y_1), \ldots, (x_n, y_n)$, to the line is then

$$f(b_0, b_1) = \sum_{i=1}^{n} [y_i - (b_0 + b_1 x_i)]^2$$

The point estimates of β_0 and β_1, denoted by b_1 and b_0, are called the least squares estimates – they are those values that minimize using partial derivatives.

$$b_1 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \qquad b_0 = \bar{y} - b_1\bar{x}$$

$$b_1 = \frac{SS_{xy}}{SS_{xx}}$$

$$b_0 = \bar{y} - b_1\bar{x}$$

$$SS_{xy} = \sum x_i y_i - \frac{\left(\sum x_i\right)\left(\sum y_i\right)}{n}$$

$$SS_{xx} = \sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n} = (n-1)s_x^2$$

The predicted values are obtained using:

$$\hat{y} = b_0 + b_1 x$$

Simple linear regression

We interpret the fitted value as the value of *y* that we would predict or expect when using the estimated regression line with *x* = x_i; thus $\hat{y}_i$ is the ***estimated true mean*** for that population when x = x_i (based on the data).

The residual $y_i - \hat{y}_i$ is a positive number if the point lies above the line and a negative number if it lies below the line.$(x_i, \hat{y}_i)$

The residual can be thought of as a measure of deviation and we can summarize the notation in the following way:

$$Y_i - \hat{Y}_i = \hat{\epsilon}_i$$

Simple linear regression

Suppose we have the following data on filtration rate (x) versus moisture content (y):

x	125.3	98.2	201.4	147.3	145.9	124.7	112.2	120.2	161.2	178.9
y	77.9	76.8	81.5	79.8	78.2	78.3	77.5	77.0	80.1	80.2
x	159.5	145.8	75.1	151.4	144.2	125.0	198.8	132.5	159.6	110.7
y	79.9	79.0	76.7	78.2	79.5	78.1	81.5	77.0	79.0	78.6

Relevant summary quantities (*summary statistics*) are

$\Sigma x_i = 2817.9$, $\Sigma y_i = 1574.8$, $\Sigma x^2_i = 415{,}949.85$,

$\Sigma x_i\, y_i = 222{,}657.88$, and $\Sigma y^2_i = 124{,}039.58$,

From $S_{xx} = 18{,}921.8295$, $S_{xy} = 776.434$.

Calculation of residuals?

Simple linear regression

x	y	x^2	xy
3	8	9	24
9	6	81	54
5	4	25	20
3	2	9	6
∑x = 20	**∑y = 20**	**$\sum x^2$ = 124**	**∑xy = 104**

$$b_1 = \frac{SS_{xy}}{SS_{xx}}$$

$$b_0 = \overline{y} - b_1\overline{x}$$

$$SS_{xy} = \sum x_i y_i - \frac{\left(\sum x_i\right)\left(\sum y_i\right)}{n}$$

$$SS_{xx} = \sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n} = (n-1)s_x^2$$

Using formula,

b1 = {4(104) – 20*20} / {4*(124) – 20^2} = 16/96 = 0.166*

b0 = 20/4 - 0.166(20/4) = 4.17*

So, linear regression equation is, y= b0 + b1x => y = 4.17 + 0.166x

Simple linear regression

Linear regression, while a powerful tool, has certain limitations that should be considered:

• **Linearity:** Assumes a linear relationship between the dependent and independent variables. If the relationship is non-linear, the model may not accurately capture the underlying pattern.

• **Independence:** Assumes that the errors are independent of each other. If there is autocorrelation in the errors, the model's estimates may be biased and inefficient.

• **Homoscedasticity:** Assumes that the variance of the errors is constant across all levels of the independent variable. If the variance is not constant (heteroscedasticity), the model's estimates may be inefficient.

• **Normality:** Assumes that the errors are normally distributed. If the errors are not normally distributed, the model's inferences may be invalid.

• **Sensitivity to Outliers:** Linear regression can be sensitive to outliers, which can have a significant impact on the model's estimates. Outliers can distort the relationship between the variables and lead to biased results.

• **Limited Flexibility:** Linear regression can only model linear relationships. If the relationship between the variables is complex or non-linear, linear regression may not be able to adequately capture the pattern.

Regression Metrics

Some common regression metrics are

- **Mean Absolute Error (MAE):** $MAE = \frac{1}{n}\sum_{i=1}^{n}|x_i - y_i|$
- **Mean Squared Error (MSE):** $MSE = \frac{1}{n}\sum_{i=1}^{n}(x_i - y_i)^2$
- **Root Mean Squared Error (RMSE):** $RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - y_i)^2}$
- **R-squared (R^2) Score:** $R^2 = 1 - (SSR / SST)$

$$\text{r2\_score} = 1 - \frac{\text{total\_error\_model}}{\text{total\_error\_baseline}} = 1 - \frac{\sum_{i=1}^{N}(\text{predicted}_i - \text{actual}_i)^2}{\sum_{i=1}^{N}(\text{average\_value} - \text{actual}_i)^2}$$

where,

- x_i represents the actual or observed value for the i-th data point.
- y_i represents the predicted value for the i-th data point.
- SSR (Sum of Squared Residuals) and SST (Total Sum of Squares).

Regression Metrics

Q. A real estate company is trying to predict the selling price of houses based on their size (in square feet). They trained a regression model and obtained the following predicted prices and actual selling prices for a sample of five houses:

House	Actual Price (in $1000)	Predicted Price (in $1000)
1	300	280
2	350	360
3	420	410
4	280	310
5	500	480

Calculate the MAE, MSE, RMSE, R2 Score.