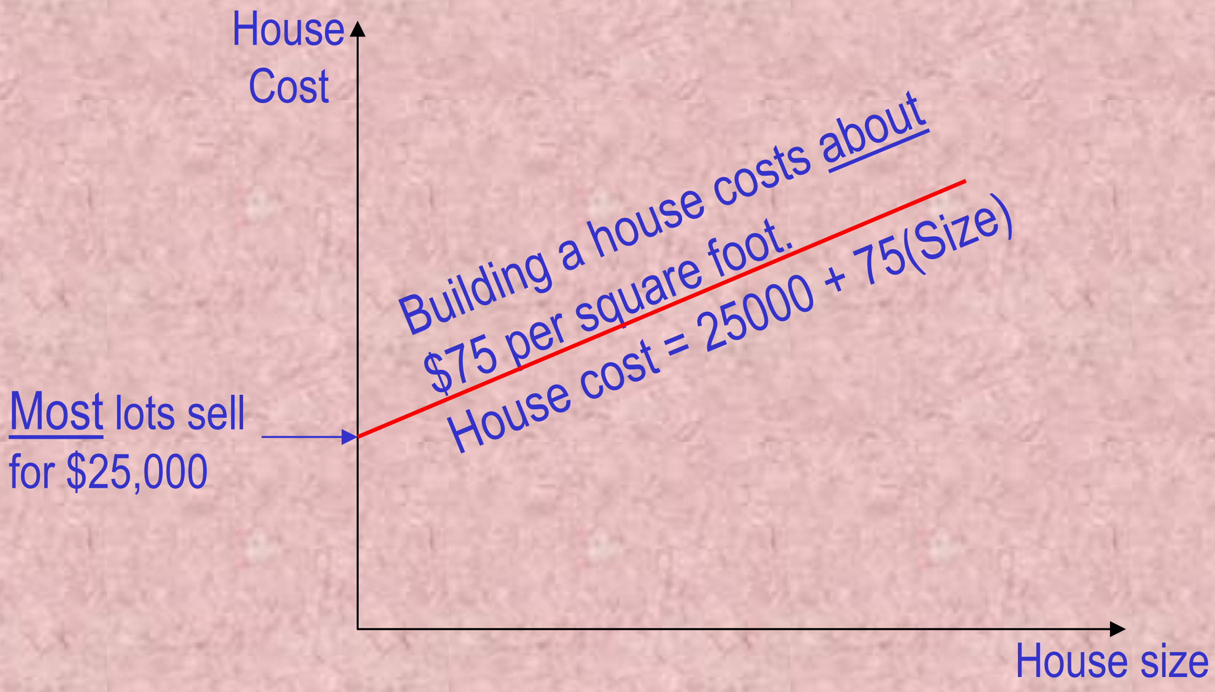
Simple Linear Regression

Introduction

- We will examine the relationship between quantitative variables x and y via a mathematical equation.
- The motivation for using the technique:
  - Forecast the value of a dependent variable (y) from the value of independent variables (x_1, x_2,…x_k.).
  - Analyze the specific relationships between the independent variables and the dependent variable.

The Model

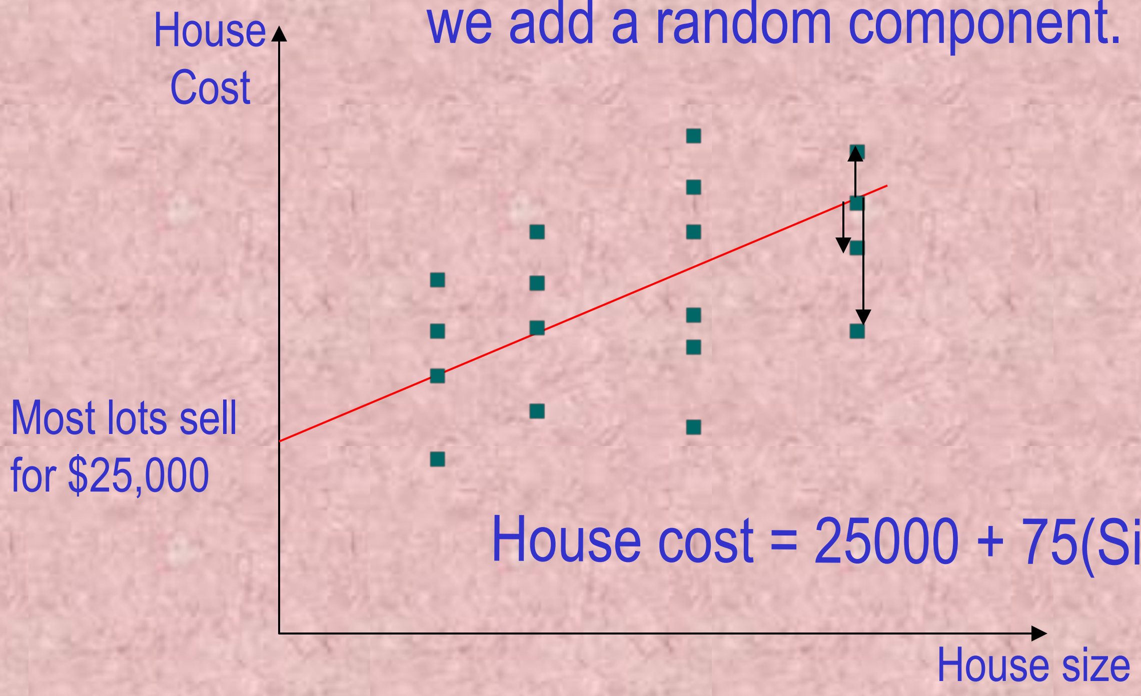
The model has a deterministic and a probabilistic components



The Model

However, house cost vary even among same size houses!

Since cost behave unpredictably, we add a random component.



House cost = 25000 + 75(Size) + ε

The Model

- The first order linear model

$$\boxed{y = \beta_0 + \beta_1 x + \varepsilon}$$

y = dependent variable

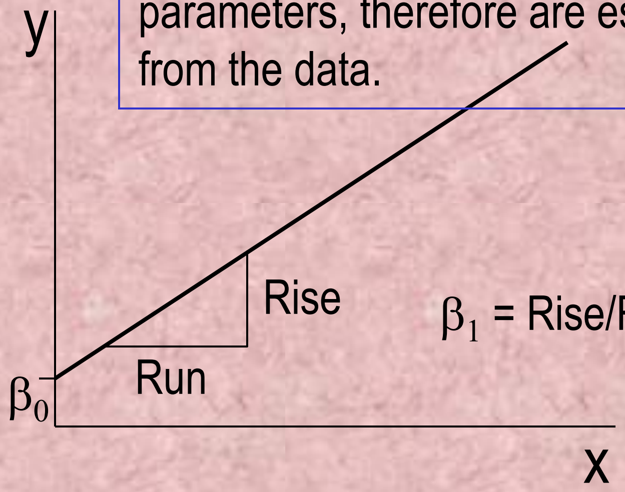
x = independent variable

β_0 = y-intercept

β_1 = slope of the line

ε = error variable

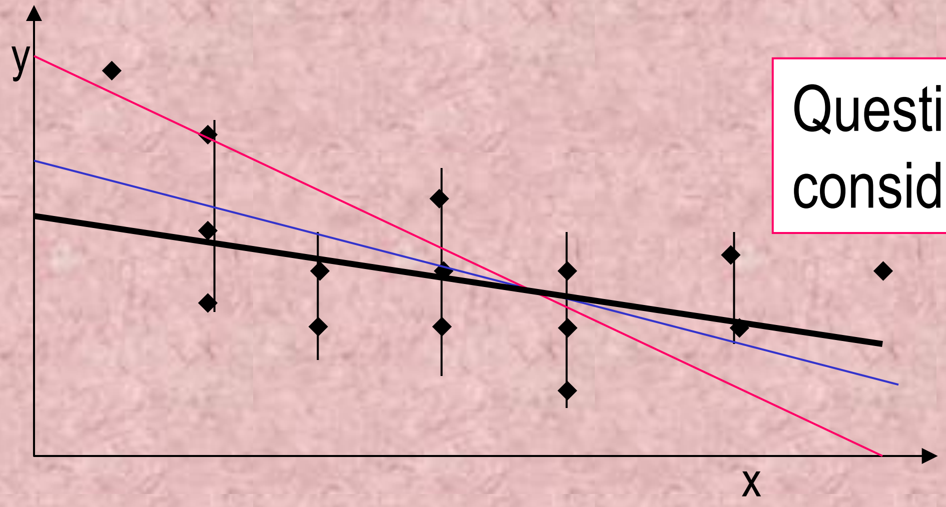
β_0 and β_1 are unknown population parameters, therefore are estimated from the data.



β_1 = Rise/Run

Estimating the Coefficients

- The estimates are determined by
  - drawing a sample from the population of interest,
  - calculating sample statistics.
  - producing a straight line that cuts into the data.



Question: What should be considered a good line?

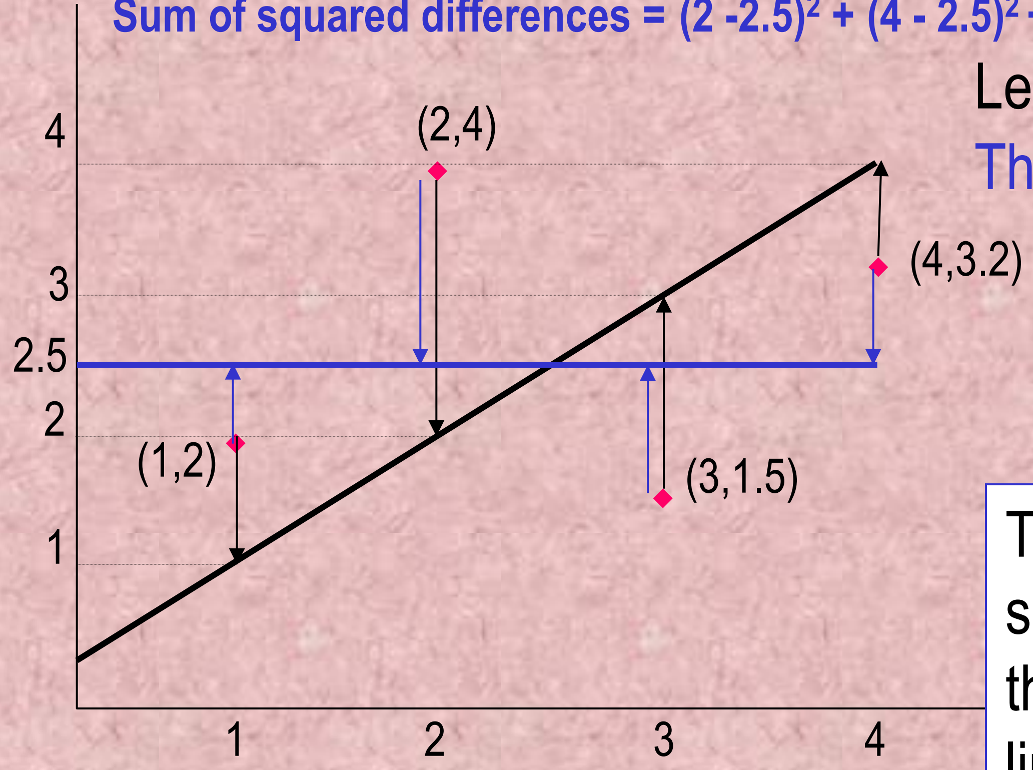
The Least Squares (Regression) Line

A good line is one that minimizes
the sum of squared differences between the points and the line.

The Least Squares (Regression) Line

Sum of squared differences = $(2 - 1)^2 + (4 - 2)^2 + (1.5 - 3)^2 + (3.2 - 4)^2 = 6.89$

Sum of squared differences = $(2 - 2.5)^2 + (4 - 2.5)^2 + (1.5 - 2.5)^2 + (3.2 - 2.5)^2 = 3.99$



Let us compare two lines

The second line is horizontal

The smaller the sum of squared differences the better the fit of the line to the data.

The Estimated Coefficients

To calculate the estimates of the slope and intercept of the least squares line , use the formulas:

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

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

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

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

Alternate formula for the slope b_1

$$b_1 = r\frac{s_y}{s_x}$$

The regression equation that estimates the equation of the first order linear model is:

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

The Simple Linear Regression Line

- Example:
  - A car dealer wants to find the relationship between the odometer reading and the selling price of used cars.
  - A random sample of 100 cars is selected, and the data recorded.
  - Find the regression line.

Car	Odometer	Price
1	37388	14636
2	44758	14122
3	45833	14016
4	30862	15590
5	31705	15568
6	34010	14718
.	Independent	Dependent
.	variable x	variable y
.	.	.

The Simple Linear Regression Line

- Solution
  - Solving by hand: Calculate a number of statistics

$\bar{x} = 36{,}009.45;$ $$SS_{xx} = \sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n} = 43{,}528{,}690$$

$\bar{y} = 14{,}822.823;$ $$SS_{xy} = \sum (x_i y_i) - \frac{\sum x_i \sum y_i}{n} = -2{,}712{,}511$$

where n = 100.

$$b_1 = \frac{SS_{xy}}{(n-1)s_x^2} = \frac{-2{,}712{,}511}{43{,}528{,}690} = -.06232$$

$$b_0 = \bar{y} - b_1\bar{x} = 14{,}822.82 - (-.06232)(36{,}009.45) = 17{,}067$$

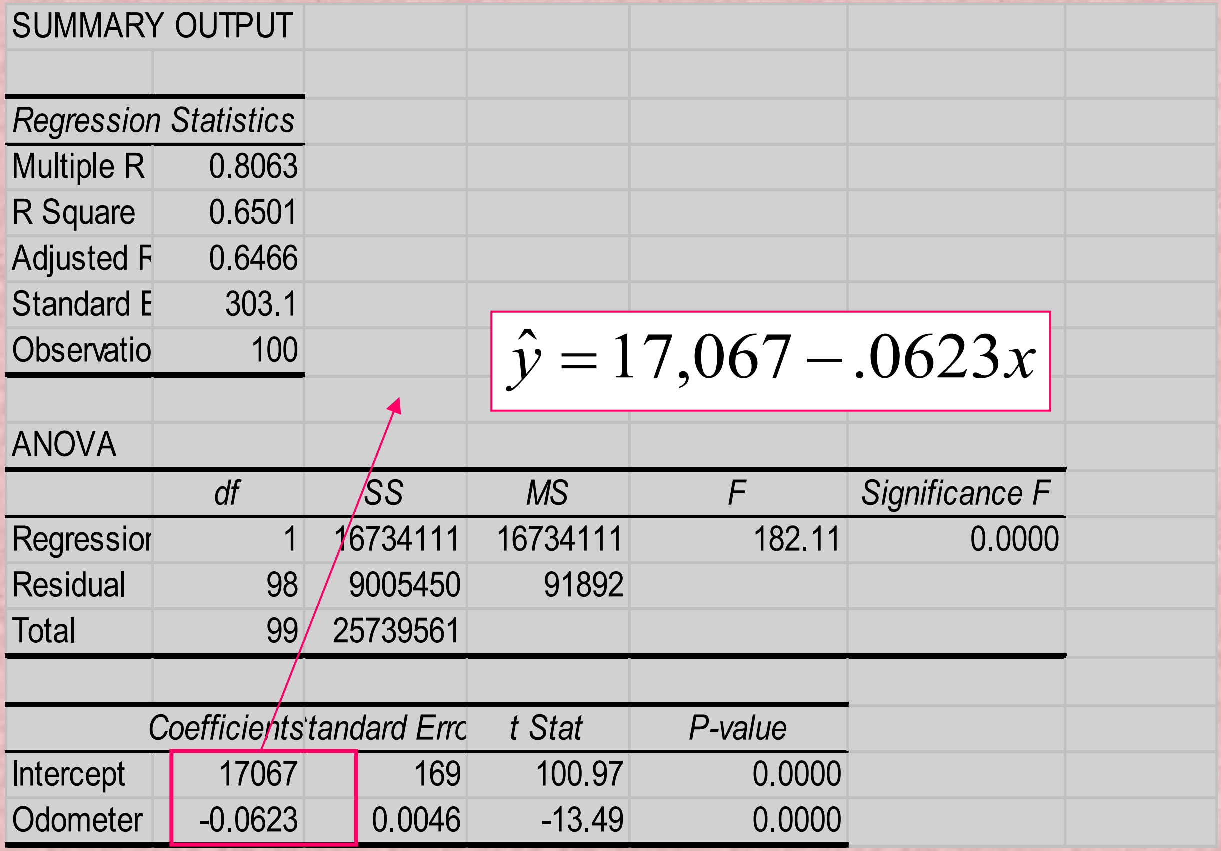
$$\boxed{\hat{y} = b_0 + b_1 x = 17{,}067 - .0623x}$$

The Simple Linear Regression Line

- Solution – continued
  - Using the computer

    1. Scatterplot
    2. Trend function
    3. Tools > Data Analysis > Regression

The Simple Linear Regression Line



SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.8063
R Square	0.6501
Adjusted F	0.6466
Standard E	303.1
Observatio	100

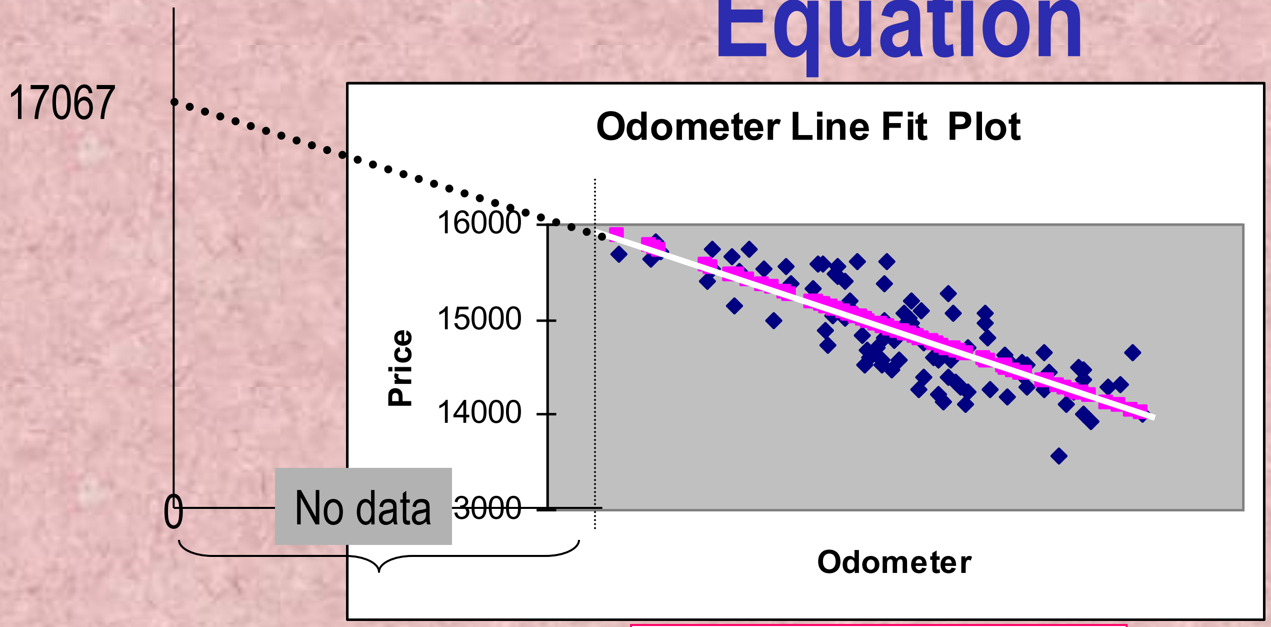
ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regressior	1	16734111	16734111	182.11	0.0000
Residual	98	9005450	91892		
Total	99	25739561			

	Coefficients	*tandard Errc*	*t Stat*	*P-value*
Intercept	17067	169	100.97	0.0000
Odometer	-0.0623	0.0046	-13.49	0.0000

$$\hat{y} = 17{,}067 - .0623x$$

Interpreting the Linear Regression - Equation



$$\hat{y} = 17{,}067 \boxminus .0623x$$

The intercept is b_0 = \$17067.

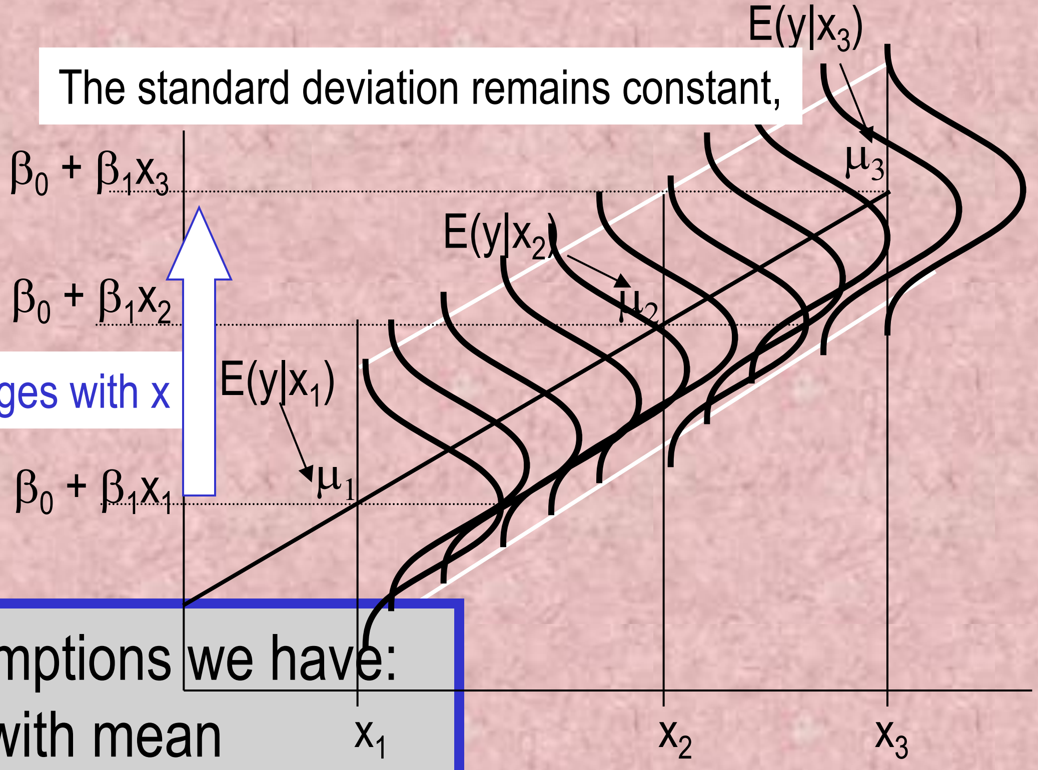
This is the slope of the line.
For each additional mile on the odometer,
the price decreases by an average of \$0.0623

Do not interpret the intercept as the "Price of cars that have not been driven"

Error Variable: Required Conditions

- The error ε is a critical part of the regression model.
- Four requirements involving the distribution of ε must be satisfied.
  - The probability distribution of ε is normal.
  - The mean of ε is zero: $E(\varepsilon) = 0$.
  - The standard deviation of ε is σ_ε for all values of x.
  - The set of errors associated with different values of y are all independent.

The Normality of ε



The standard deviation remains constant,

but the mean value changes with x

From the first three assumptions we have:
y is normally distributed with mean
$E(y) = \beta_0 + \beta_1 x$, and a constant standard deviation σ_ε

Assessing the Model

- The least squares method will produces a regression line whether or not there is a linear relationship between x and y.
- Consequently, it is important to assess how well the linear model fits the data.
- Several methods are used to assess the model. All are based on the sum of squares for errors, SSE.

Sum of Squares for Errors

- This is the sum of differences between the points and the regression line.
- It can serve as a measure of how well the line fits the data. SSE is defined by

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 .$$

- A shortcut formula

$$SSE = \sum y_i^2 - b_0 \sum y_i - b_1 \sum x_i y_i$$

Standard Error of Estimate

- The mean error is equal to zero.
- If σ_ε is small the errors tend to be close to zero (close to the mean error). Then, the model fits the data well.
- Therefore, we can, use σ_ε as a measure of the suitability of using a linear model.
- An estimator of σ_ε is given by s_ε

$$S\tan dard\ \ Error\ \ of\ \ Estimate$$

$$s_\varepsilon = \sqrt{\frac{SSE}{n-2}}$$

Standard Error of Estimate, Example

- Example:
  - Calculate the standard error of estimate for the previous example and describe what it tells you about the model fit.
- Solution

$$SSE = 9,005,450$$

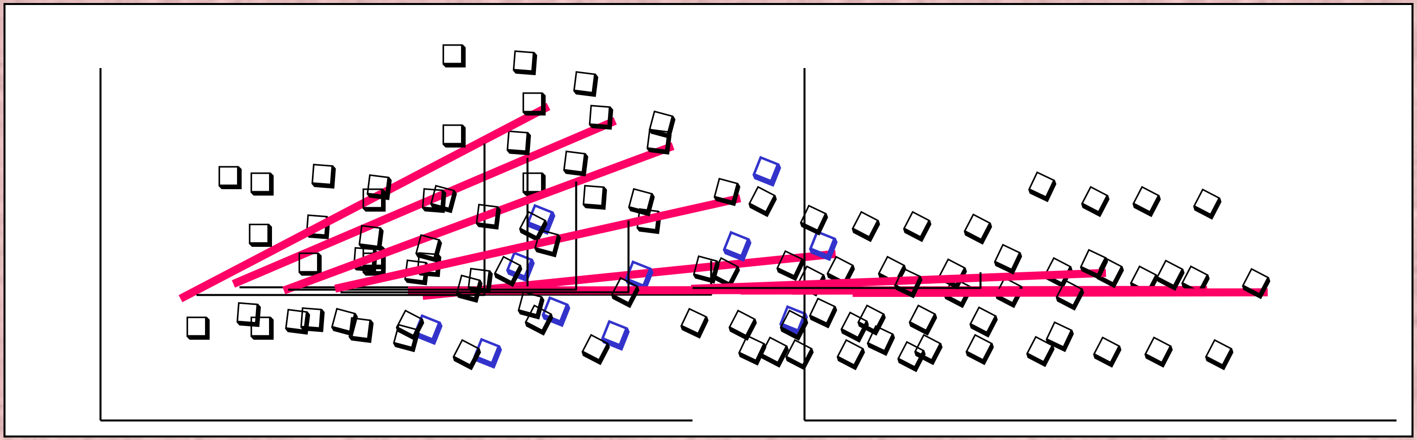
$$s_{\varepsilon} = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{9,005,450}{9}} = 303.13$$

It is hard to assess the model based on s_{ε} even when compared with the mean value of y.

$s_{\varepsilon} = 303.1$ $\bar{y} = 14,823$

Testing the slope

– When no linear relationship exists between two variables, the regression line should be horizontal.



Linear relationship.
Different inputs (x) yield different outputs (y).

The slope is not equal to zero

No linear relationship.
Different inputs (x) yield the same output (y).

The slope is equal to zero

THANK YOU