Statistics

Chapter 10: Simple Linear Regression

Where We’ve Been

- Presented methods for estimating and testing population parameters for a single sample
- Extended those methods to allow for a comparison of population parameters for multiple samples

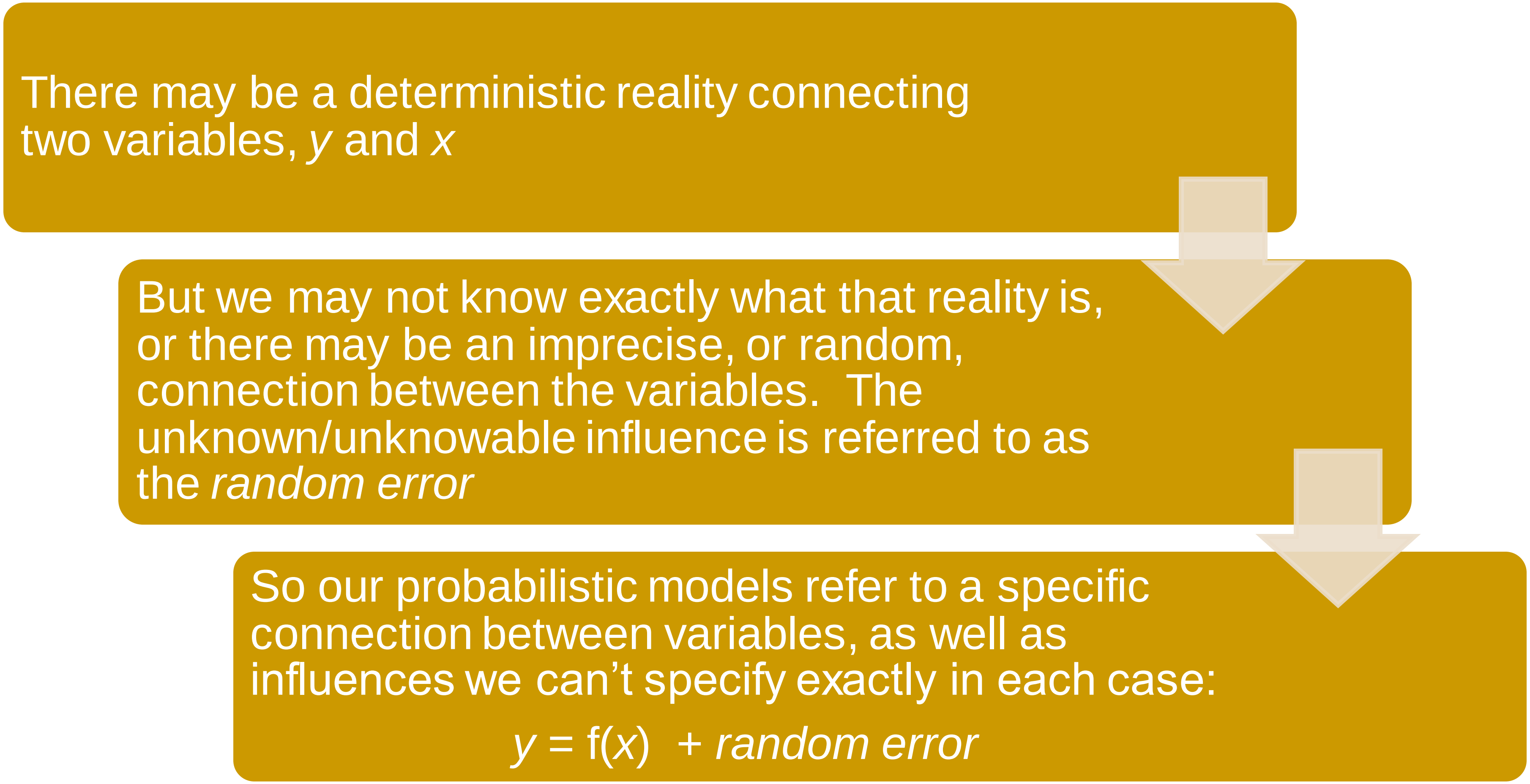
Where We're Going

- Introduce the straight-line *linear regression* model as a means of relating one quantitative variable to another quantitative variable
- Introduce the *correlation coefficient* as a means of relating one quantitative variable to another quantitative variable
- Assess how well the simple linear regression model fits the sample data
- Use the simple linear regression model to predict the value of one variable given the value of another variable

10.1: Probabilistic Models

- Regression is the bread and butter of statisticians.
- Considering bivariate qualitative data.
- Simple linear regression looks for the *best* **linear** function of one variable (x, **explanatory** variable) that fits the other (y, **response** variable).

10.1: Probabilistic Models



There may be a deterministic reality connecting two variables, *y* and *x*

But we may not know exactly what that reality is, or there may be an imprecise, or random, connection between the variables. The unknown/unknowable influence is referred to as the *random error*

So our probabilistic models refer to a specific connection between variables, as well as influences we can't specify exactly in each case:

$$y = \mathrm{f}(x) + \textit{random error}$$

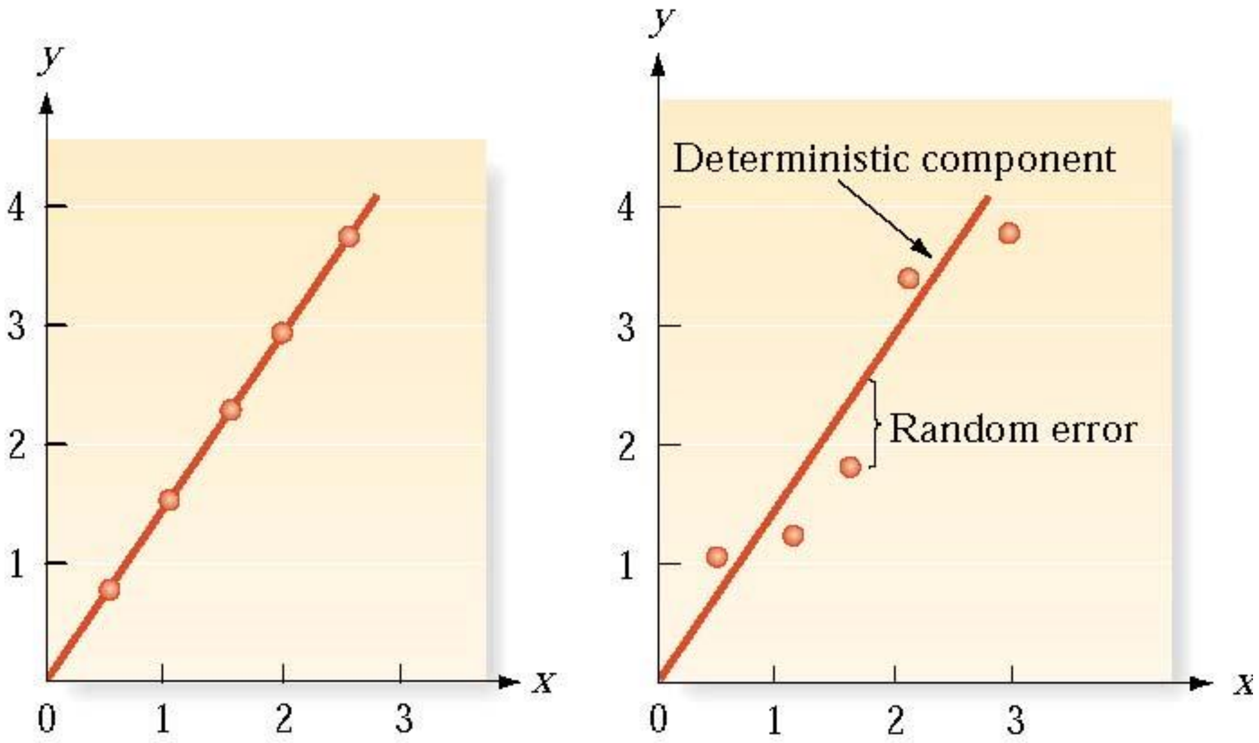
10.1: Probabilistic Models

General Form of Probabilistic Models

y = Deterministic component + Random error

where *y* is the variable of interest, and the mean value of the random error is assumed to be 0: $E(y)$ = Deterministic component.

10.1: Probabilistic Models



a. Deterministic relationship:
$y = 1.5x$

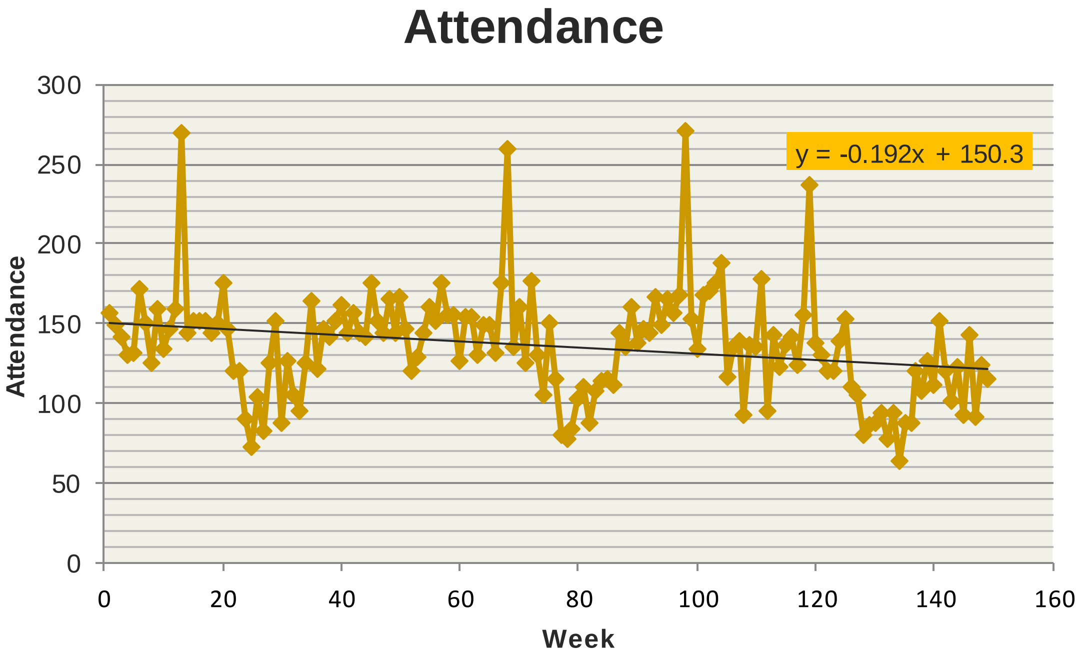
b. Probabilistic relationship:
$y = 1.5x + \text{Random error}$

10.1: Probabilistic Models

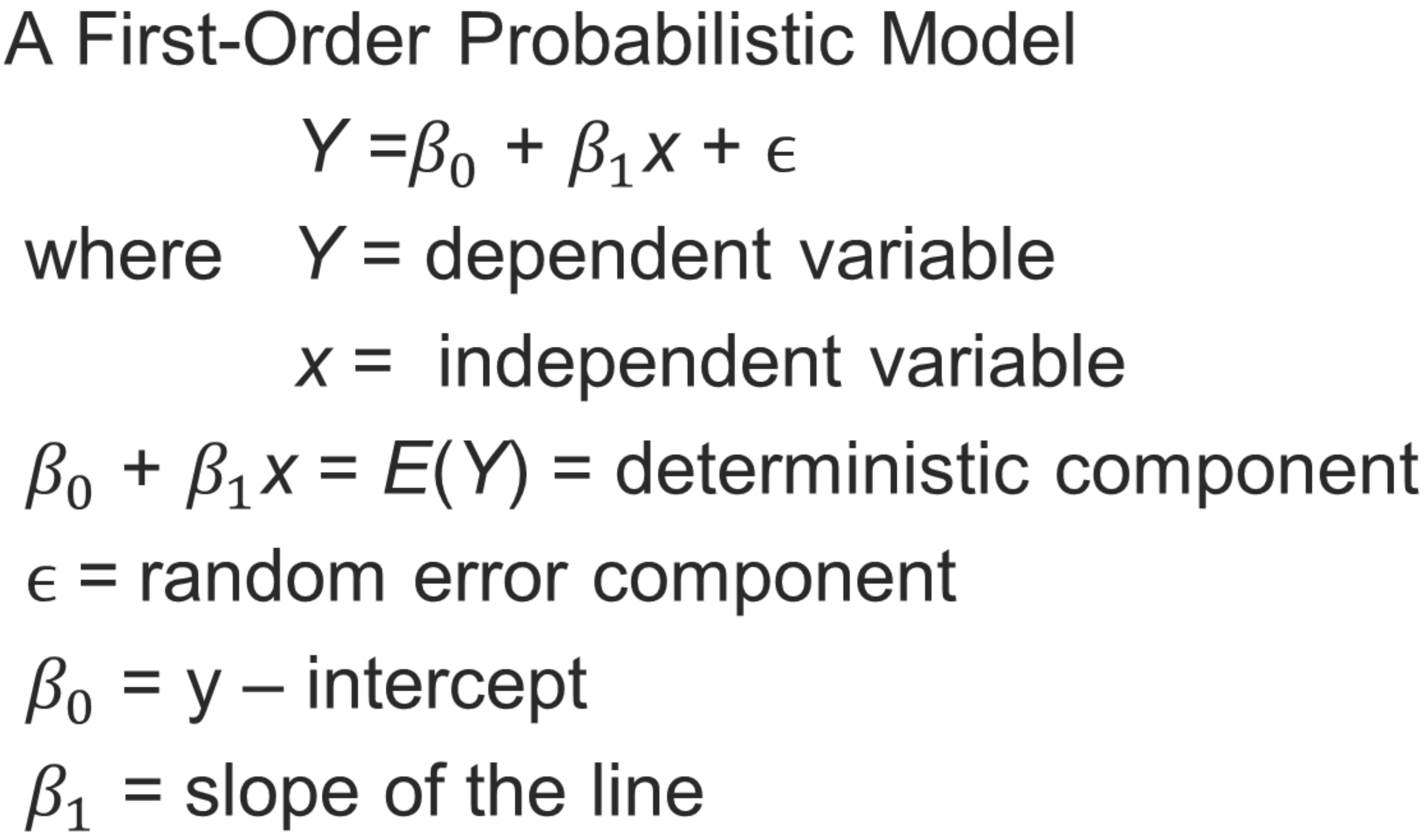
- The goal of regression analysis is to find the straight line that comes closest to all of the points in the scatter plot simultaneously.
- Closest is in terms of vertical distance between the point and the line, that is, the discrepancy in fitting the y-value.

10.1: Probabilistic Models

- The goal of regression analysis is to find the straight line that comes closest to all of the points in the scatter plot simultaneously.



10.1: Probabilistic Models



- A First-Order Probabilistic Model

$$Y = \beta_0 + \beta_1 x + \epsilon$$

where Y = dependent variable

x = independent variable

$\beta_0 + \beta_1 x = E(Y)$ = deterministic component

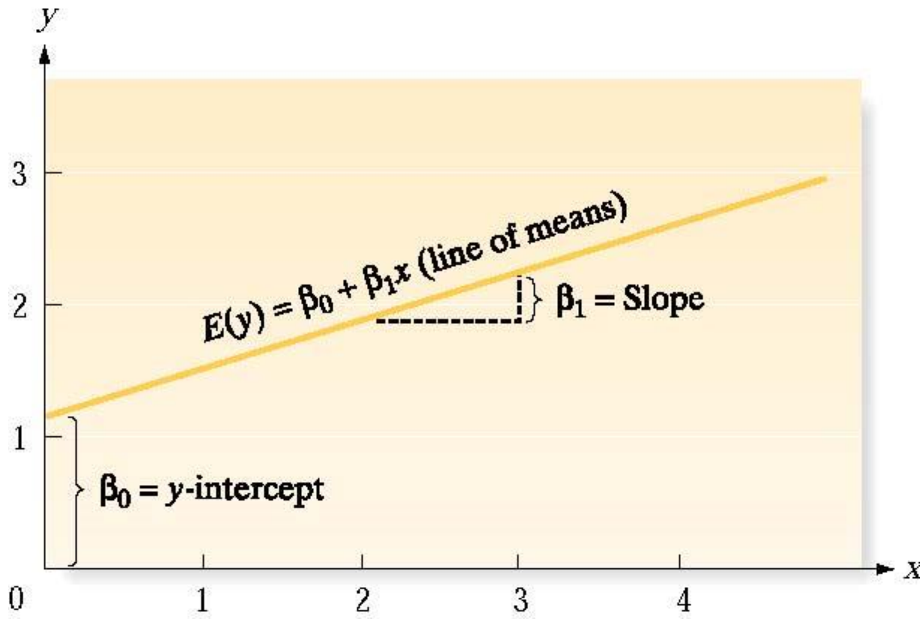
ϵ = random error component

β_0 = y – intercept

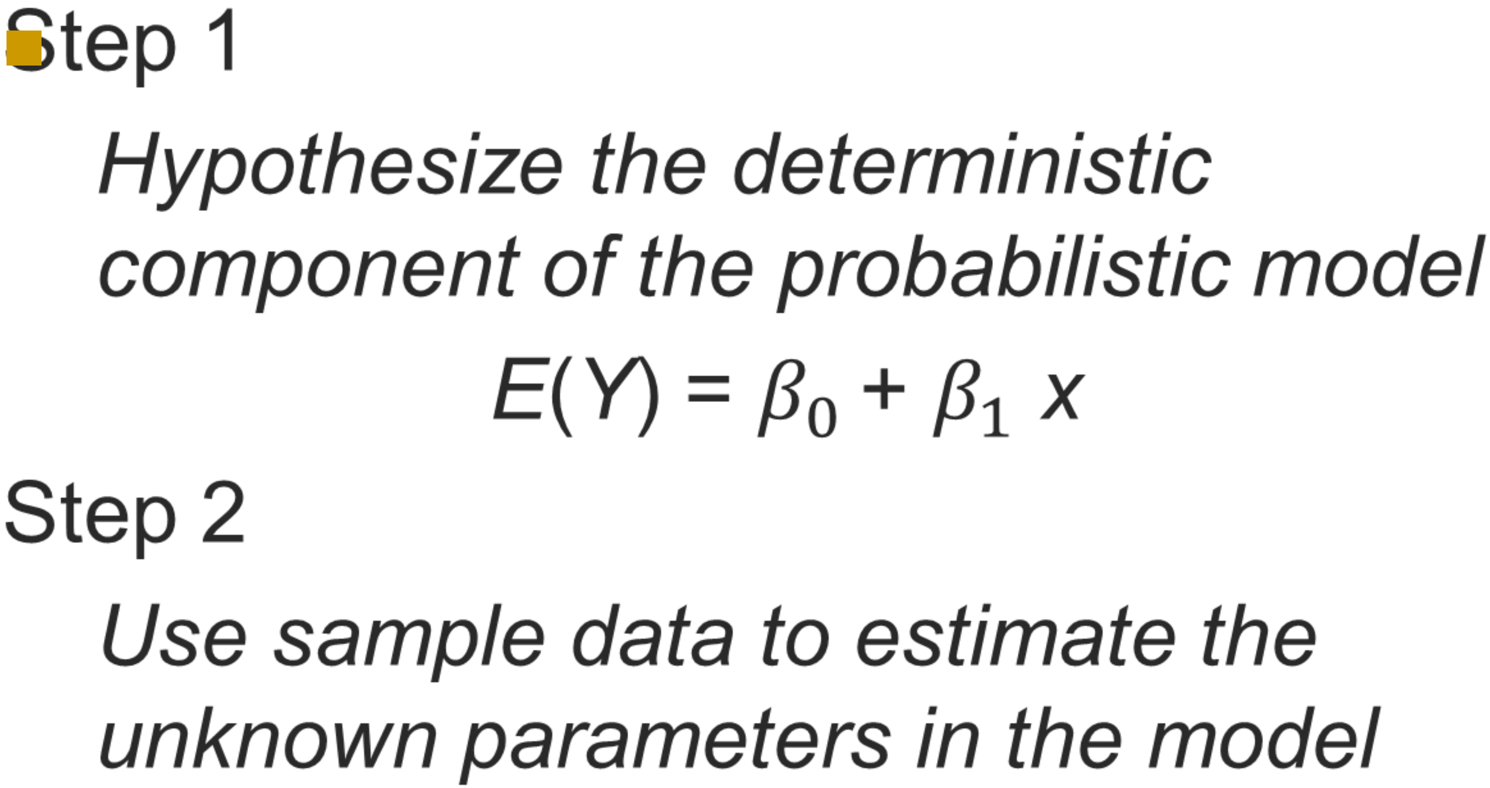
β_1 = slope of the line

10.1: Probabilistic Models

- β_0, the y – intercept, and β_1 , the slope of the line, are population parameters, and invariably unknown. Regression analysis is designed to estimate these parameters.



10.2: Fitting the Model: The Least Squares Approach



- Step 1

  Hypothesize the deterministic component of the probabilistic model

  $$E(Y) = \beta_0 + \beta_1 x$$

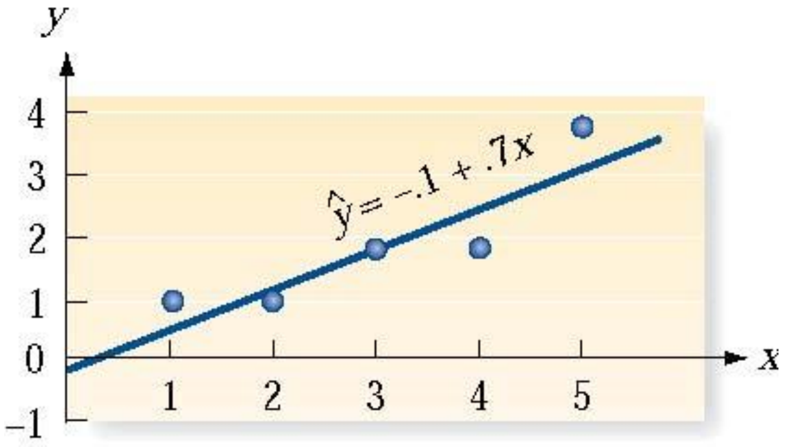
  Step 2

  Use sample data to estimate the unknown parameters in the model

10.2: Fitting the Model: The Least Squares Approach

Table 11.1 **Advertising-Sales Data**

Month	Advertising Expenditure, *x* ($100s)	Sales Revenue, *y* ($1,000s)
1	1	1
2	2	1
3	3	2
4	4	2
5	5	4



10.2: Fitting the Model: The Least Squares Approach

- Model: $y = \beta_0 + \beta_1 x$
- Estimates: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$
- Deviation: $(y_i - \hat{y}_i) = [y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)]$
- SSE: $\sum [y_i - (\hat{\beta}_0 + \hat{\beta}_1 x)]^2$

10.2: Fitting the Model: The Least Squares Approach

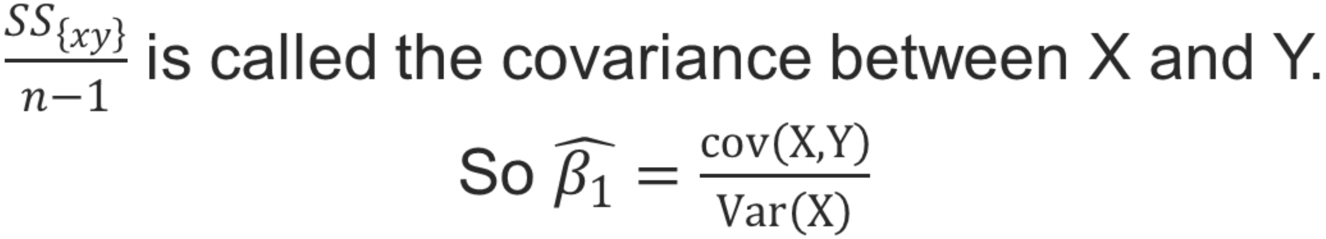
- The **least squares line** $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ is the line that has the following two properties:
  1. The sum of the errors (SE) equals 0.
  2. The sum of squared errors (SSE) is smaller than that for any other straight-line model.

10.2: Fitting the Model: The Least Squares Approach

Formulas for the Least Squares Estimates

$$\text{Slope: } \hat{\beta}_1 = \frac{\text{SS}_{xy}}{\text{SS}_{xx}} = \frac{\sum (x_i - \overline{x})(y_i - \overline{y})}{\sum (x_i - \overline{x})^2} = \frac{\sum x_i y_i - \frac{\left(\sum x_i\right)\left(\sum y_i\right)}{n}}{\sum x_i^{\ 2} - \frac{\left(\sum x_i\right)^2}{n}}$$

$$y\text{– intercept: } \beta_0 = \overline{y} - \hat{\beta}_1 \overline{x}$$



$\frac{SS_{\{xy\}}}{n-1}$ is called the covariance between X and Y.

So $\widehat{\beta_1} = \frac{\text{cov(X,Y)}}{\text{Var(X)}}$

10.2: Fitting the Model: The Least Squares Approach

```
> y<-c(1,1,2,2,4)
> x<-c(1,2,3,4,5)
> lm(y~x)

Call:
lm(formula = y ~ x)

Coefficients:
(Intercept)            x
       -0.1          0.7

> b0<-mean(y)-mean(x)*cov(x,y)/var(x)
> b0
[1] -0.1
> b1<-cov(x,y)/var(x)
> b1
[1] 0.7
```

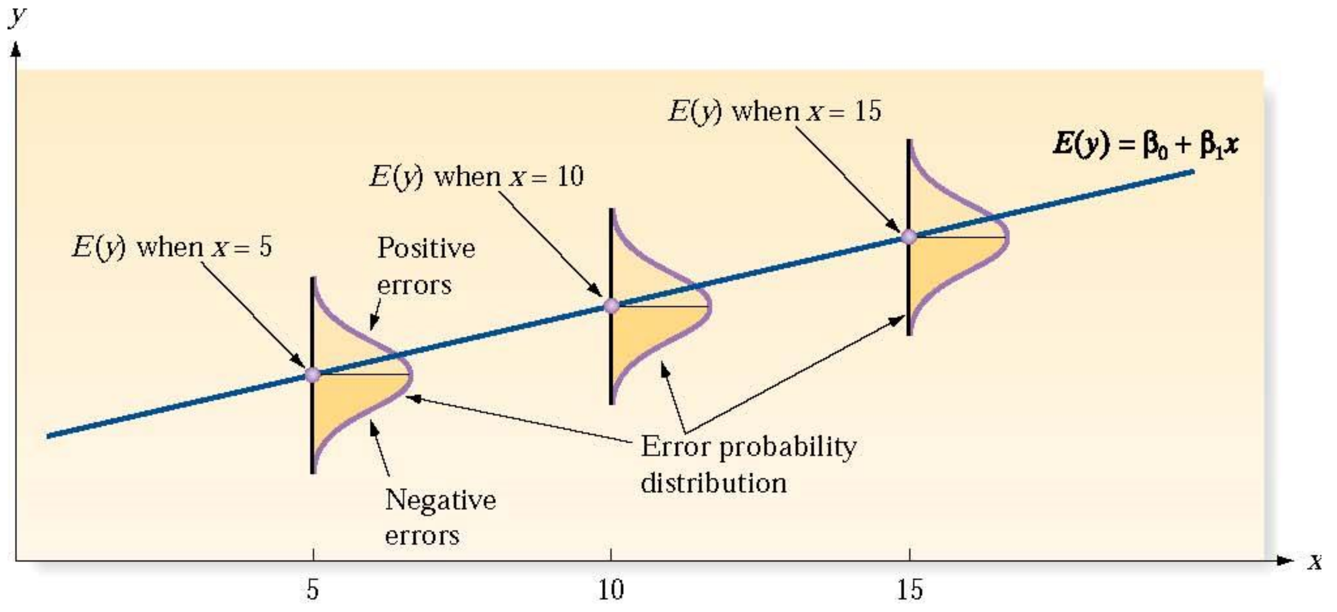
10.3: Model Assumptions

Assumptions

1. The mean of the probability distribution of ε is 0.
2. The variance, σ^2, of the probability distribution of ε is constant.
3. The probability distribution of ε is normal.
4. The values of ε associated with any two values of y are independent.

10.3: Model Assumptions

- The variance, σ^2, is used in every test statistic and confidence interval used to evaluate the model.
- Invariably, σ^2 is unknown and must be estimated.



10.3: Model Assumptions

Estimation of σ^2 for a (First-Order) Straight-Line Model

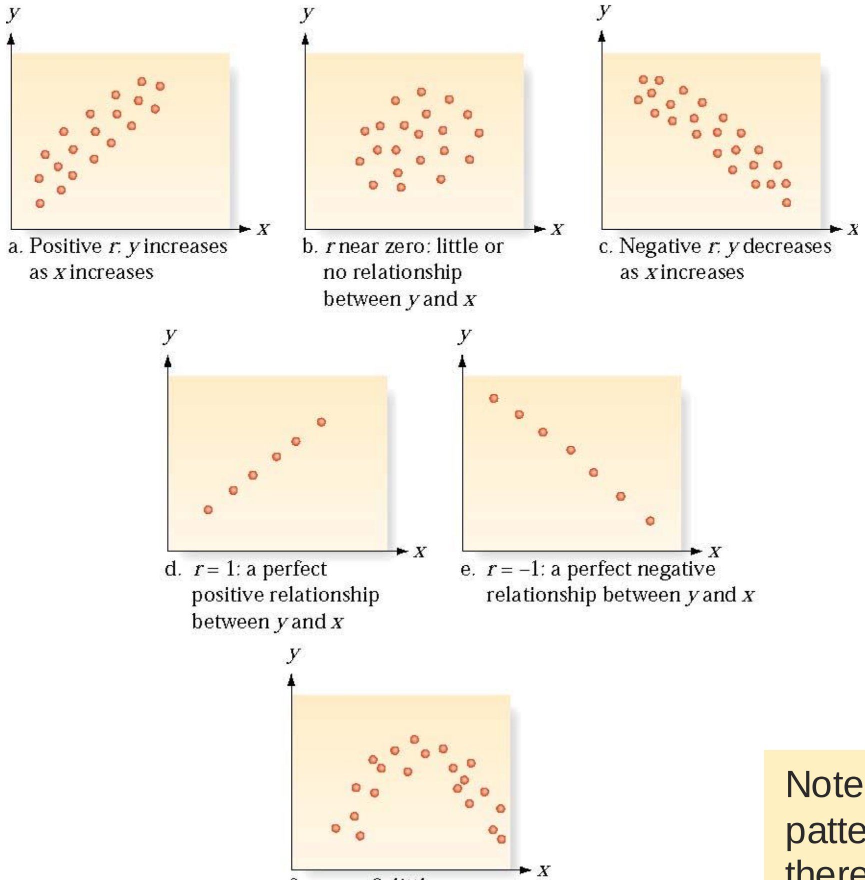
$$s^2 = \frac{SSE}{n-2} = \frac{\sum\left(y_i - \hat{y}_i\right)^2}{n-2} = \frac{SS_{yy} - \hat{\beta}_1 SS_{xy}}{n-2}$$

$$SS_{yy} = \sum\left(y_i - \overline{y}\right)^2 \text{ and } SS_{xy} = \sum\left(x_i - \overline{x}\right)\left(y_i - \overline{y}\right)$$

The estimated standard error of ε is the square root of the variance:

$$s = \sqrt{s^2} = \sqrt{\frac{SSE}{n-2}}$$

10.4: Assessing the Utility of the Model: Making Inferences about the Slope β_1



a. Positive *r*: *y* increases as *x* increases

b. *r* near zero: little or no relationship between *y* and *x*

c. Negative *r*: *y* decreases as *x* increases

d. *r* = 1: a perfect positive relationship between *y* and *x*

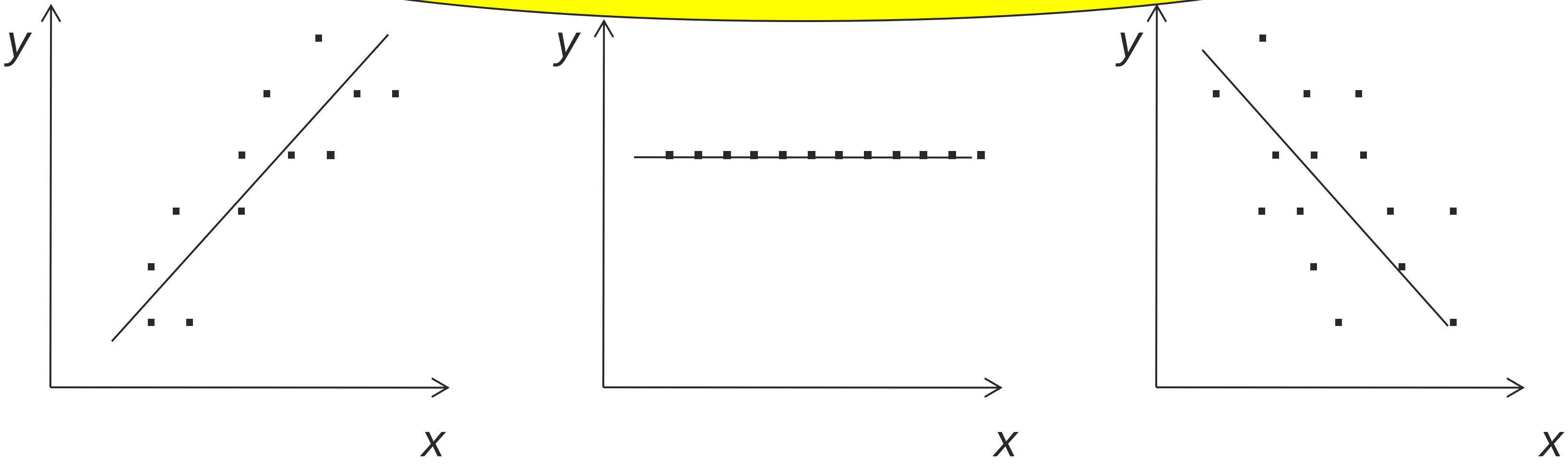
e. *r* = –1: a perfect negative relationship between *y* and *x*

f. *r* near 0: little or no linear relationship between *y* and *x*

Note: There may be many different patterns in the scatter plot when there is no linear relationship.

10.4: Assessing the Utility of the Model: Making Inferences about the Slope β_1

A critical step in the evaluation of the model is to test whether $\beta_1 = 0$



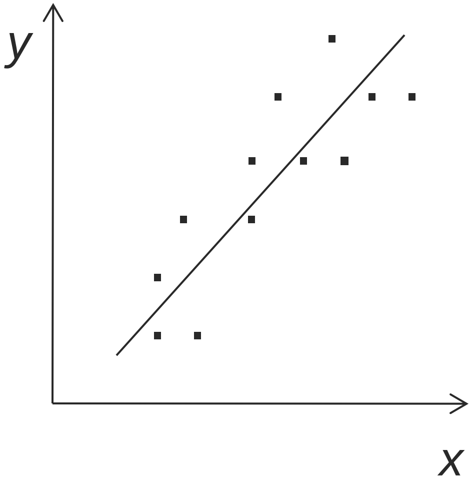
Positive Relationship
$\beta_1 > 0$

No Relationship
$\beta_1 = 0$

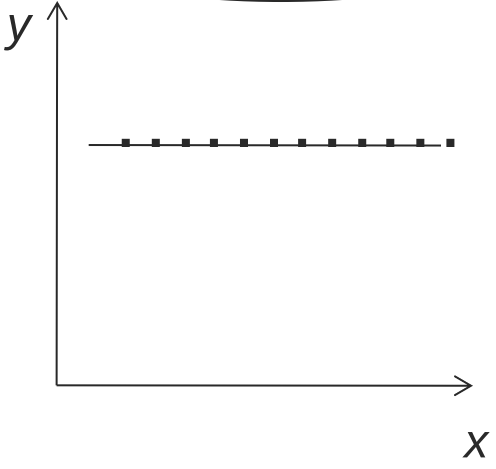
Negative Relationship
$\beta_1 < 0$

10.4: Assessing the Utility of the Model: Making Inferences about the Slope β_1

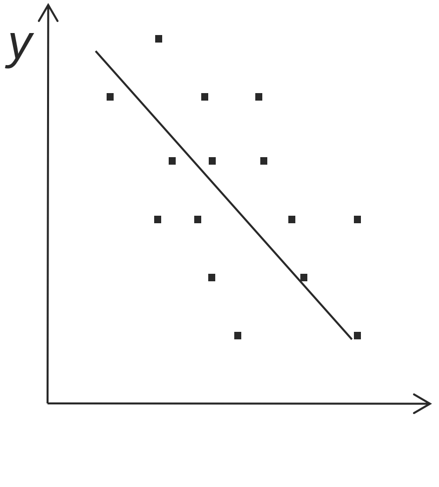
$H_0 : \beta_1 = 0$
$H_a : \beta_1 \neq 0$



Positive Relationship
$\beta_1 > 0$



No Relationship
$\beta_1 = 0$



Negative Relationship
$\beta_1 < 0$

10.4: Assessing the Utility of the Model: Making Inferences about the Slope β_1

- The four assumptions described above produce a normal sampling distribution for the slope estimate:

$$\hat{\beta}_1 \sim N(\beta_1, \sigma_{\hat{\beta}_1})$$

$$\text{where } \sigma_{\hat{\beta}_1} = \frac{\sigma}{\sqrt{SS_{xx}}}$$

$$\text{and } \hat{\sigma}_{\hat{\beta}_1} = s_{\hat{\beta}_1} = \frac{s}{\sqrt{SS_{xx}}} \quad ,$$

called the **estimated standard error of the least squares slope estimate**.

10.4: Assessing the Utility of the Model: Making Inferences about the Slope β_1

A Test of Model Utility: Simple Linear Regression

One-Tailed Test

$H_0 : \beta_1 = 0$

$H_a : \beta_1 < 0 \ (> 0)$

Two-Tailed Test

$H_0 : \beta_1 = 0$

$H_a : \beta_1 \neq 0$

$$\text{Test Statistic} : t = \frac{\hat{\beta}_1}{s_{\hat{\beta}_1}} = \frac{\hat{\beta}_1}{s / \sqrt{SS_{xx}}}$$

Rejection Region:

$t < -t_\alpha \quad (> t_\alpha)$

$|t| > t_{\alpha/2}$

Degrees of freedom = $n - 2$

10.4: Assessing the Utility of the Model: Making Inferences about the Slope β_1

```
> summary(lm(y~x))

Call:
lm(formula = y ~ x)

Residuals:
         1          2          3          4          5
 4.000e-01 -3.000e-01 -3.886e-16 -7.000e-01  6.000e-01
```

10.4: Assessing the Utility of the Model: Making Inferences about the Slope β_1

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.1000     0.6351  -0.157   0.8849
x             0.7000     0.1915   3.656   0.0354 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6055 on 3 degrees of freedom
Multiple R-squared:  0.8167,       Adjusted R-squared:
0.7556
F-statistic: 13.36 on 1 and 3 DF,  p-value: 0.03535
```

10.4: Assessing the Utility of the Model: Making Inferences about the Slope β_1

```
> n<-length(y)
> s<-sqrt(sum((y-b0-b1*x)^2)/(n-2))
> s
[1] 0.6055301
> t<-b1*sd(x)*sqrt(n-1)/s
> t
[1] 3.655631
```

- Since the *t*-value leads to rejection of the null hypothesis at 5%, we can conclude that there is a significant linear relationship between the variables.

10.4: Assessing the Utility of the Model: Making Inferences about the Slope β_1

- Interpreting *p*-Values for *β*
  - Software packages report *two-tailed p*-values.
  - To conduct *one-tailed* tests of hypotheses, the reported *p*-values must be adjusted:

$$\text{Upper-tailed test } (H_a : \beta_1 > 0): \quad p-\text{value} = \begin{cases} p/2 \text{ if } t > 0 \\ 1 - p/2 \text{ if } t < 0 \end{cases}$$

$$\text{Lower-tailed test } (H_a : \beta_1 < 0): \quad p-\text{value} = \begin{cases} p/2 \text{ if } t < 0 \\ 1 - p/2 \text{ if } t > 0 \end{cases}$$

10.4: Assessing the Utility of the Model: Making Inferences about the Slope β_1

- A Confidence Interval on β_1

$$\hat{\beta}_1 \pm t_{\alpha/2} s_{\hat{\beta}_1}$$

where the estimated standard error is

$$s_{\hat{\beta}_1} = \frac{s}{\sqrt{SS_{xx}}}$$

and $t_{\alpha/2}$ is based on (n – 2) degrees of freedom

10.4: Assessing the Utility of the Model: Making Inferences about the Slope β_1

- In the example, the estimated β_1 was 0.7, and the estimated standard error was 0.1915. With 3 degrees of freedom at 5% two-tailed, t = 3.182.
- The confidence interval is, therefore,

$$\hat{\beta}_1 \pm t_{\alpha/2} s_{\widehat{\beta_1}} = 0.7 \pm 3.182 * 0.1915 = (0.0906, 1.3094)$$

which does not include 0, so there is a linear relationship between the two variables.

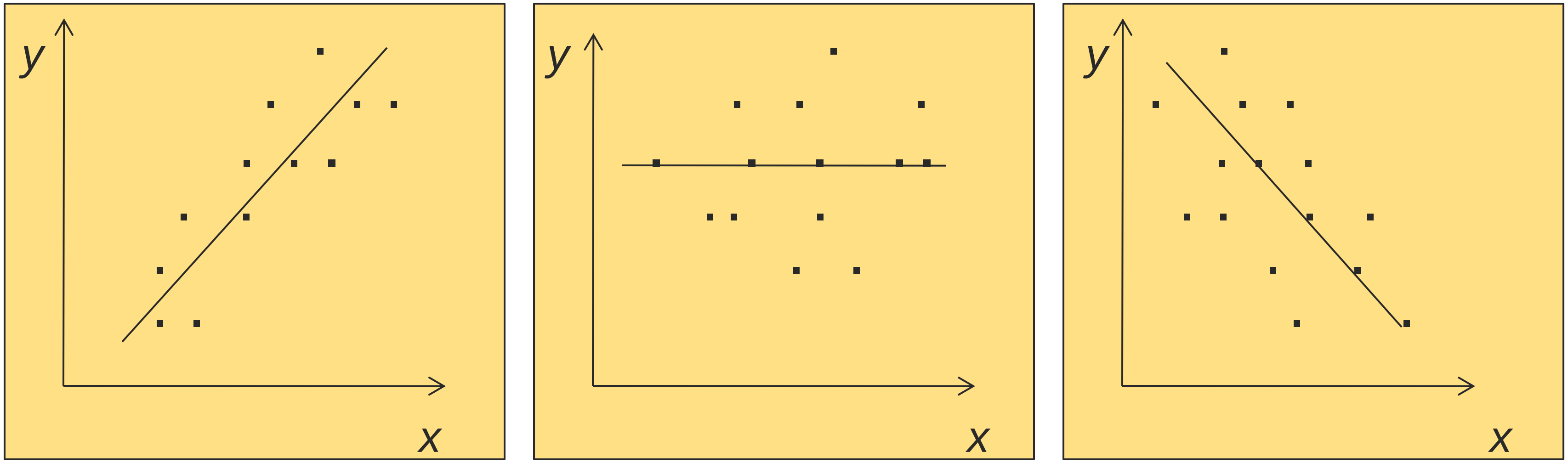
10.5: The Coefficients of Correlation and Determination

- The **coefficient of correlation**, ***r***, is a measure of the strength of the *linear* relationship between two variables. It is computed as follows:

$$r = \frac{SS_{xy}}{\sqrt{SS_{xx} SS_{yy}}}.$$

10.5: The Coefficients of Correlation and Determination

Positive linear relationship | No linear relationship | Negative linear relationship



$r \rightarrow +1$ | $r \approx 0$ | $r \rightarrow -1$

Values of *r* <u>equal</u> to +1 or -1 require each point in the scatter plot to lie on a single straight line.

10.5: The Coefficients of Correlation and Determination

- An *r* value that close to zero suggests there may not be a linear relationship between the variables, which is consistent with our earlier look at the null hypothesis and the confidence interval on β_1.

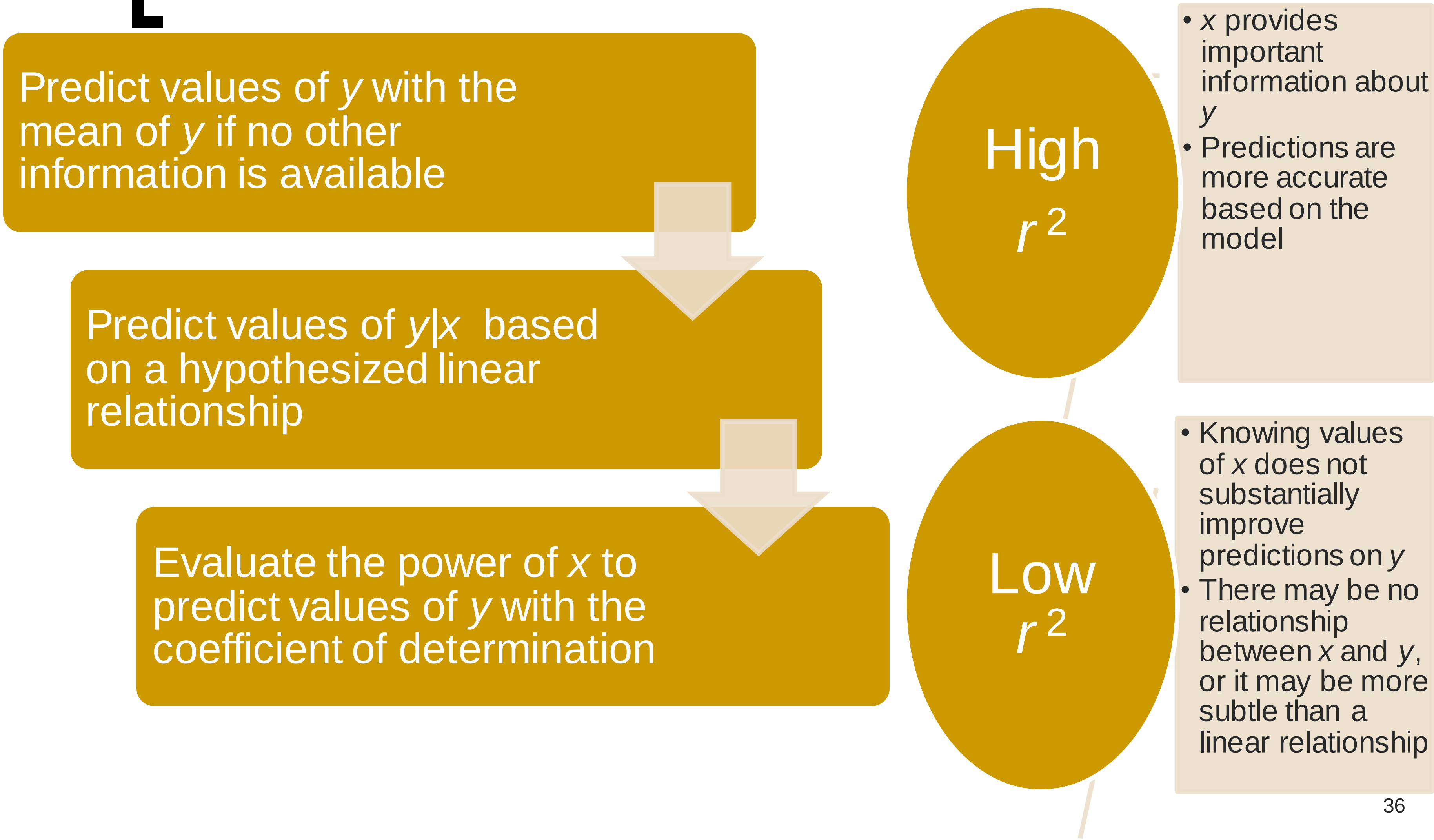
10.5: The Coefficients of Correlation and Determination

- The **coefficient of determination**, ***r*²**, represents the proportion of the total sample variability around the mean of *y* that is explained by the linear relationship between *x* and *y*.

$$r^2 = \frac{SS_{yy} - SSE}{SS_{yy}} = 1 - \frac{SSE}{SS_{yy}}$$

$$0 < r^2 < 1$$

10.5: The Coefficients of Correlation and Determination



Predict values of *y* with the mean of *y* if no other information is available

Predict values of *y*|*x* based on a hypothesized linear relationship

Evaluate the power of *x* to predict values of *y* with the coefficient of determination

High r^2

- *x* provides important information about *y*
- Predictions are more accurate based on the model

Low r^2

- Knowing values of *x* does not substantially improve predictions on *y*
- There may be no relationship between *x* and *y*, or it may be more subtle than a linear relationship