Econ 3790: Statistics Business and Economics

Instructor: Yogesh Uppal
Email: yuppal@ysu.edu

Chapter 14

- Covariance and Simple Correlation Coefficient
- Simple Linear Regression

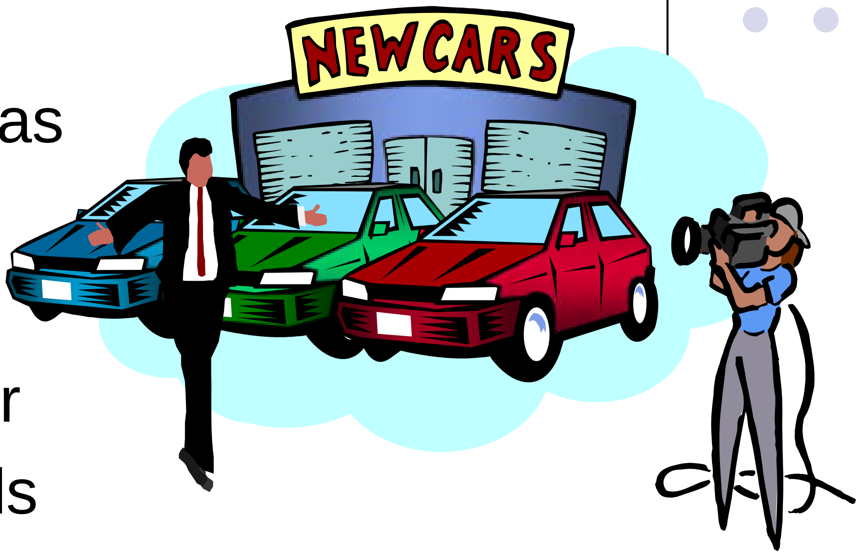
Covariance

- Covariance between x and y is a measure of relationship between x and y.

$$\operatorname{cov}(x, y) = \frac{SS_{xy}}{n-1} = \frac{\sum (y - \bar{y})(x - \bar{x})}{n-1}$$

Covariance

- Example: Reed Auto Sales

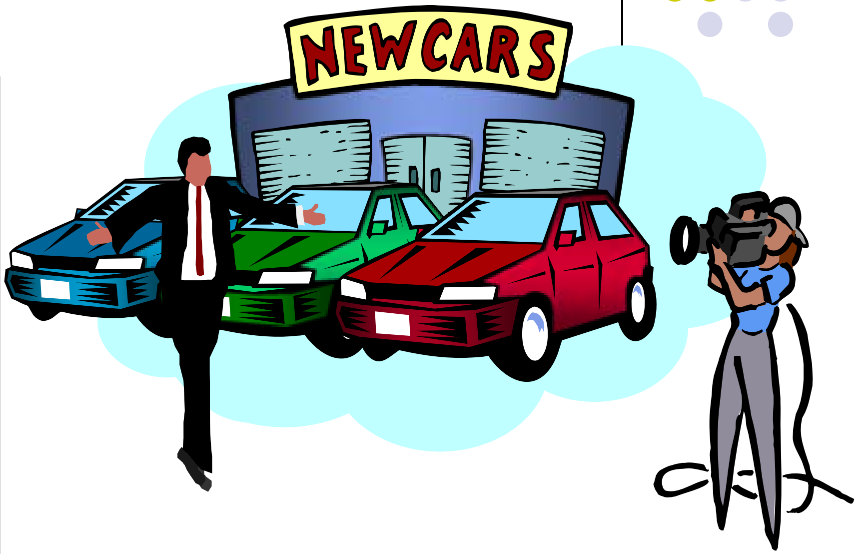


Reed Auto periodically has a special week-long sale. As part of the advertising campaign Reed runs one or more television commercials during the weekend preceding the sale. Data from a sample of 5 previous sales are shown on the next slide.

Covariance

- Example: Reed Auto Sales

Number of TV Ads	Number of Cars Sold
1	14
3	24
2	18
1	17
3	27



Covariance

x	y	$x-\bar{x}$	$y-\bar{y}$	$(x-\bar{x})(y-\bar{y})$
1	14	-1	-6	6
3	24	1	4	4
2	18	0	-2	0
1	17	-1	-3	3
3	27	1	7	7
Total=10	Total = 100			SS_{xy}=20

$$\text{cov}(x,y)=\frac{SS_{xy}}{n-1}=\frac{20}{4}=5$$

Simple Correlation Coefficient

- **Simple Population Correlation Coefficient**

$$\rho = \frac{\text{cov}(x, y)}{\sigma_x \sigma_y}$$

$$-1 \le \rho \le +1$$

- **If $\rho < 0$, a negative relationship between x and y.**
- **If $\rho > 0$, a positive relationship between x and y.**

Simple Correlation Coefficient

- Since population standard deviations of x and y are not known, we use their sample estimates to compute an estimate of ρ.

$$r = \frac{\mathrm{cov}(x, y)}{s_x s_y}$$

$$-1 \le r \le +1$$

Simple Correlation Coefficient

- Example: Reed Auto Sales

x	y	$x-\bar{x}$	$y-\bar{y}$	SS_x	SS_y
1	14	-1	-6	1	36
3	24	1	4	1	16
2	18	0	-2	0	4
1	17	-1	-3	1	9
3	27	1	7	1	49
Total=10	Total=98			Total=4	Total= 114

Simple Correlation Coefficient

$$s_x = \sqrt{\frac{(x-\bar{x})^2}{n-1}} = \sqrt{\frac{4}{4}} = 1$$

$$s_y = \sqrt{\frac{(y-\bar{y})^2}{n-1}} = \sqrt{\frac{114}{4}} = 5.34$$

$$r_{xy} = \frac{\operatorname{cov}(x,y)}{s_x s_y} = \frac{5}{1*5.34} = 0.936$$

Chapter 14 Simple Linear Regression

- Simple Linear Regression Model
- Residual Analysis
- Coefficient of Determination
- Testing for Significance
- Using the Estimated Regression Equation for Estimation and Prediction

Simple Linear Regression Model

- The equation that describes how y is related to x and an error term is called the regression model.
- The simple linear regression model is:

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

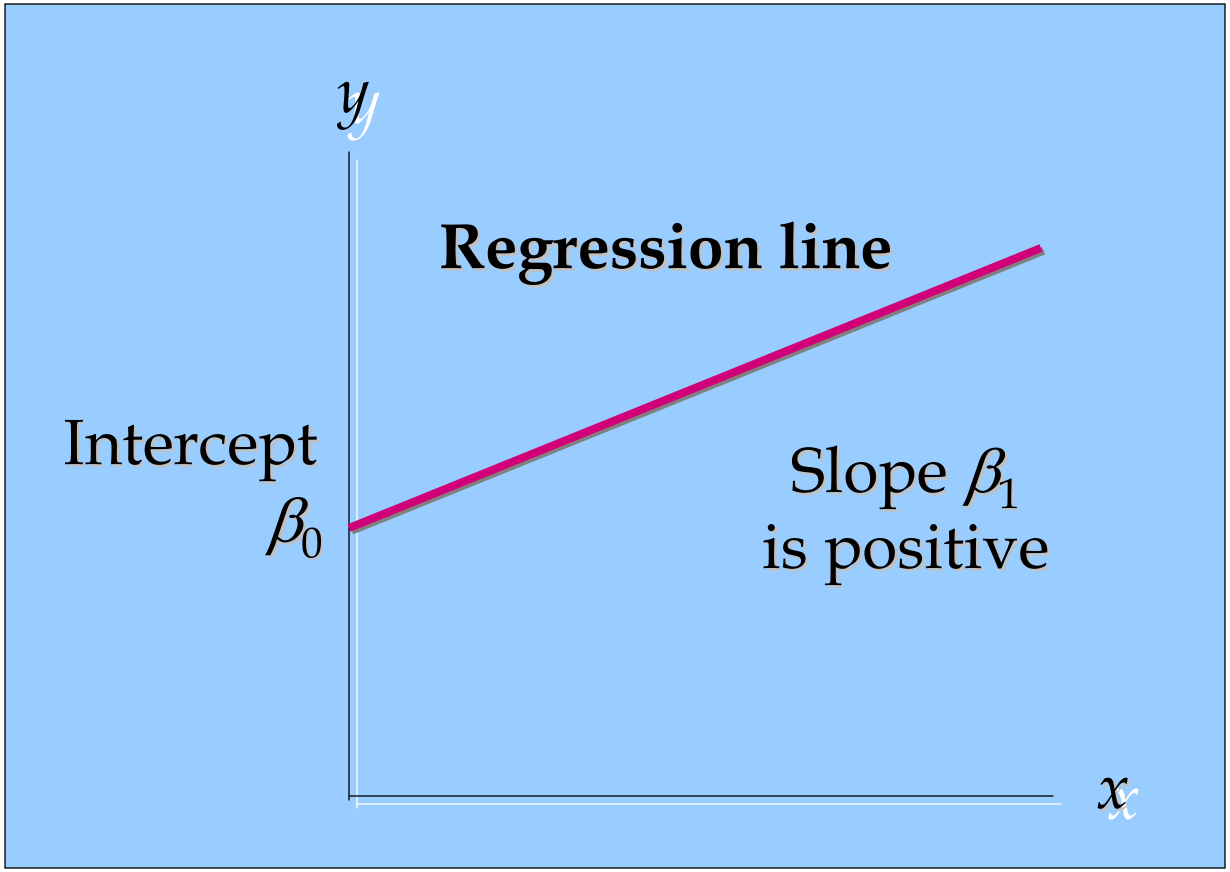
where:

β_0 and β_1 are called parameters of the model,

ε is a random variable called the error term.

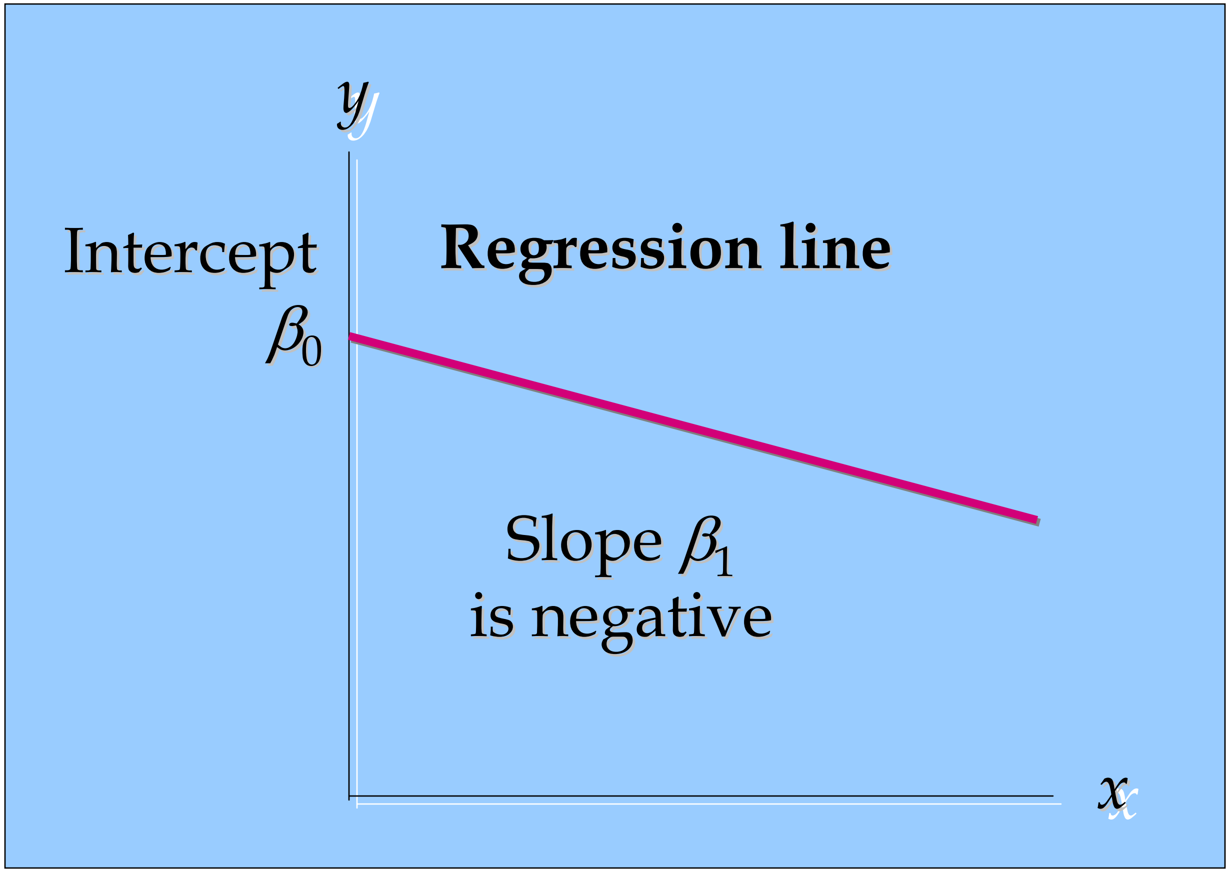
Simple Linear Regression Equation

- Positive Linear Relationship



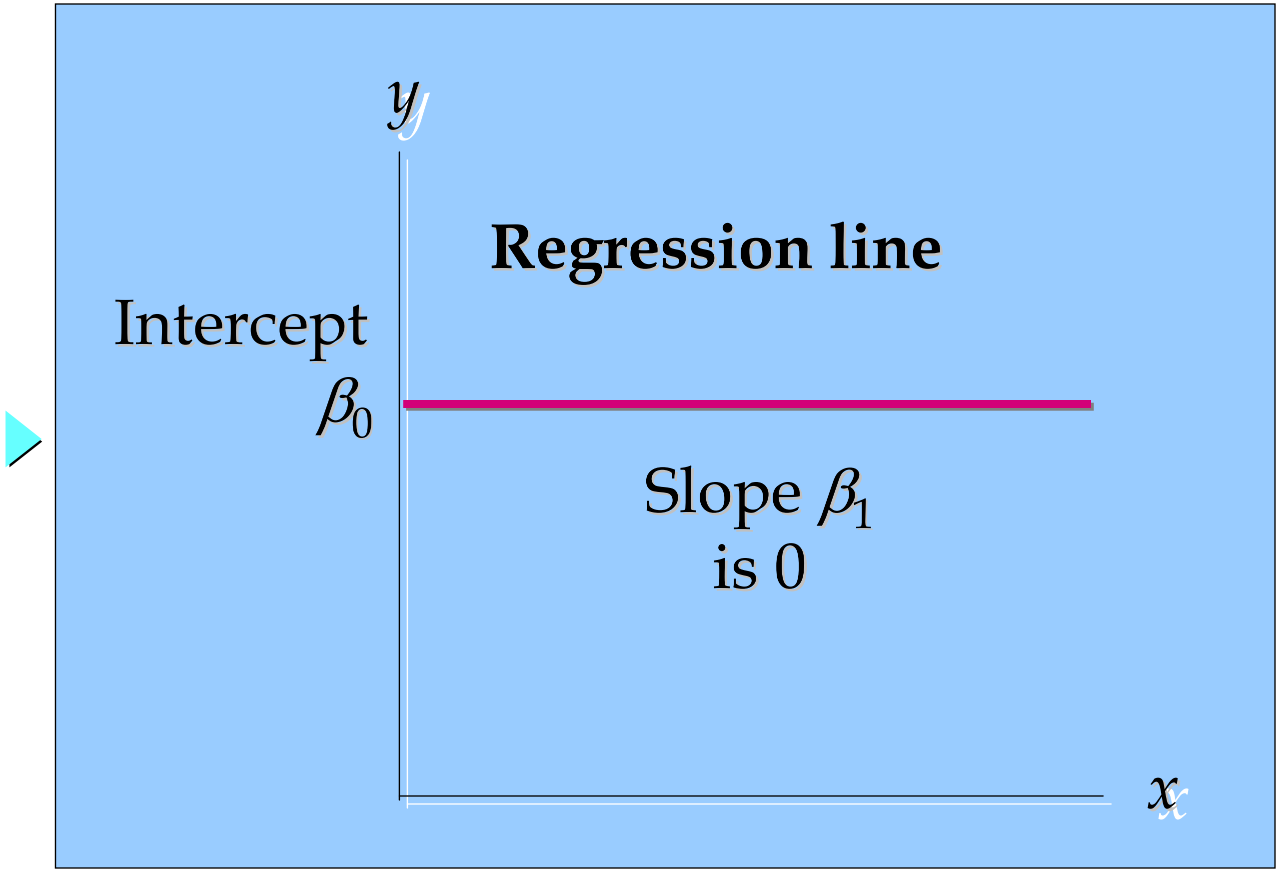
Simple Linear Regression Equation

- Negative Linear Relationship



Simple Linear Regression Equation

- No Relationship



Interpretation of β_0 and β_1

- β_0 (intercept parameter): is the value of y when x = 0.
- β_1 (slope parameter): is the change in y given x changes by 1 unit.

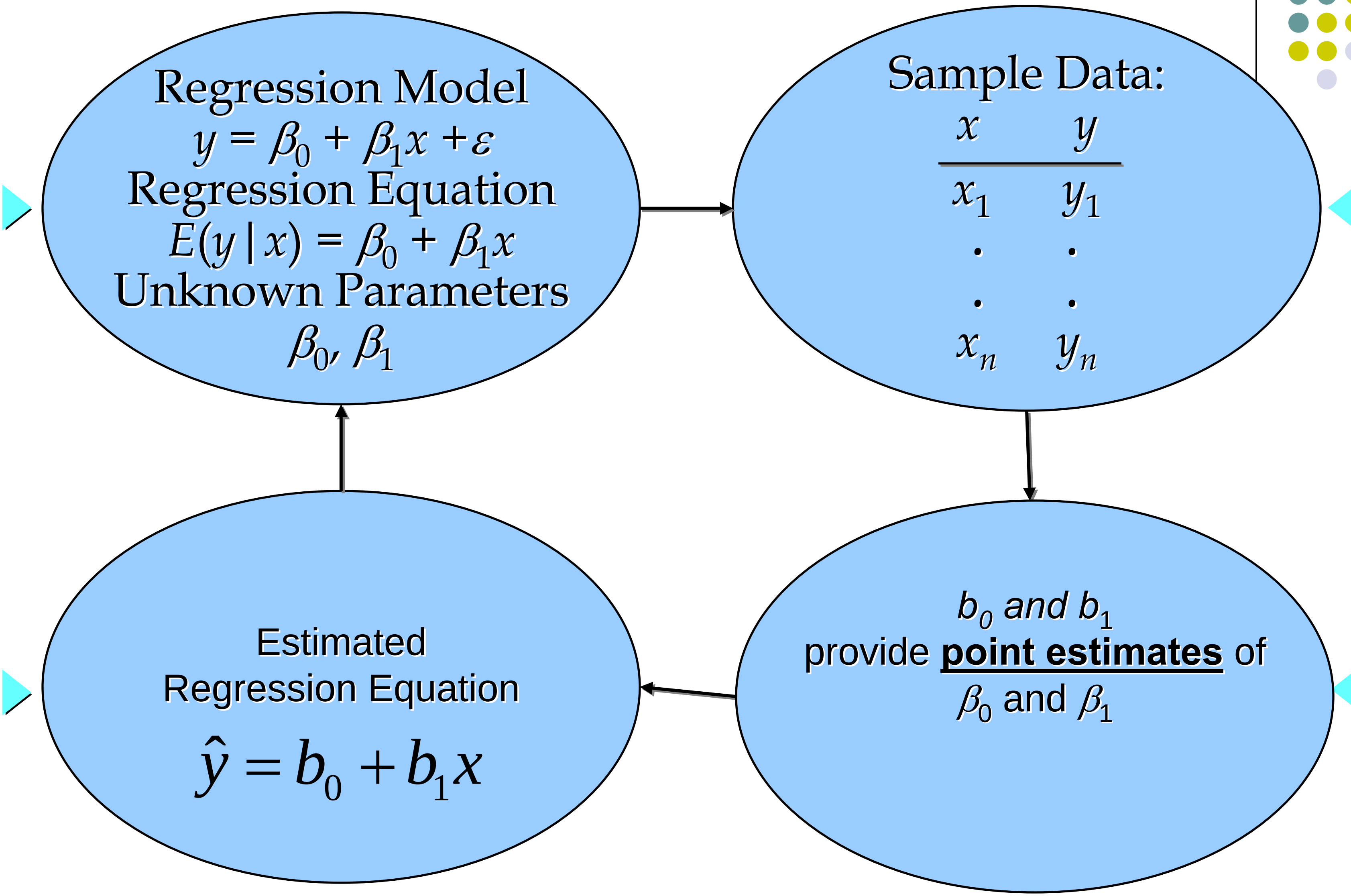
Estimated Simple Linear Regression Equation

- The estimated simple linear regression equation

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

- The graph is called the estimated regression line.
- b_0 is the y intercept of the line.
- b_1 is the slope of the line.
- $\hat{y}$ is the estimated value of y for a given value of x.

Estimation Process



Least Squares Method

- Slope for the Estimated Regression Equation

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

$$Where \quad SS_{xy} = \sum (y - \bar{y})(x - \bar{x})$$

$$and \quad SS_x = \sum (x - \bar{x})^2$$

Least Squares Method

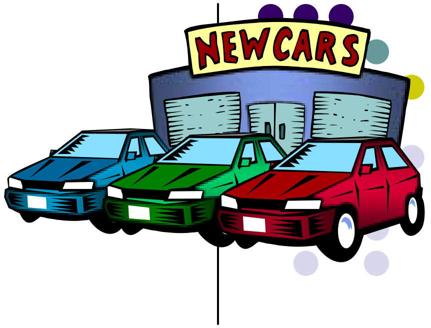
- *y*-Intercept for the Estimated Regression Equation

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

$\bar{x}$ = mean value for independent variable

$\bar{y}$ = mean value for dependent variable

Estimated Regression Equation



- Example: Reed Auto Sales

- Slope for the Estimated Regression Equation

$$b_1 = \frac{SS_{xy}}{SS_x} = \frac{20}{4} = 5$$

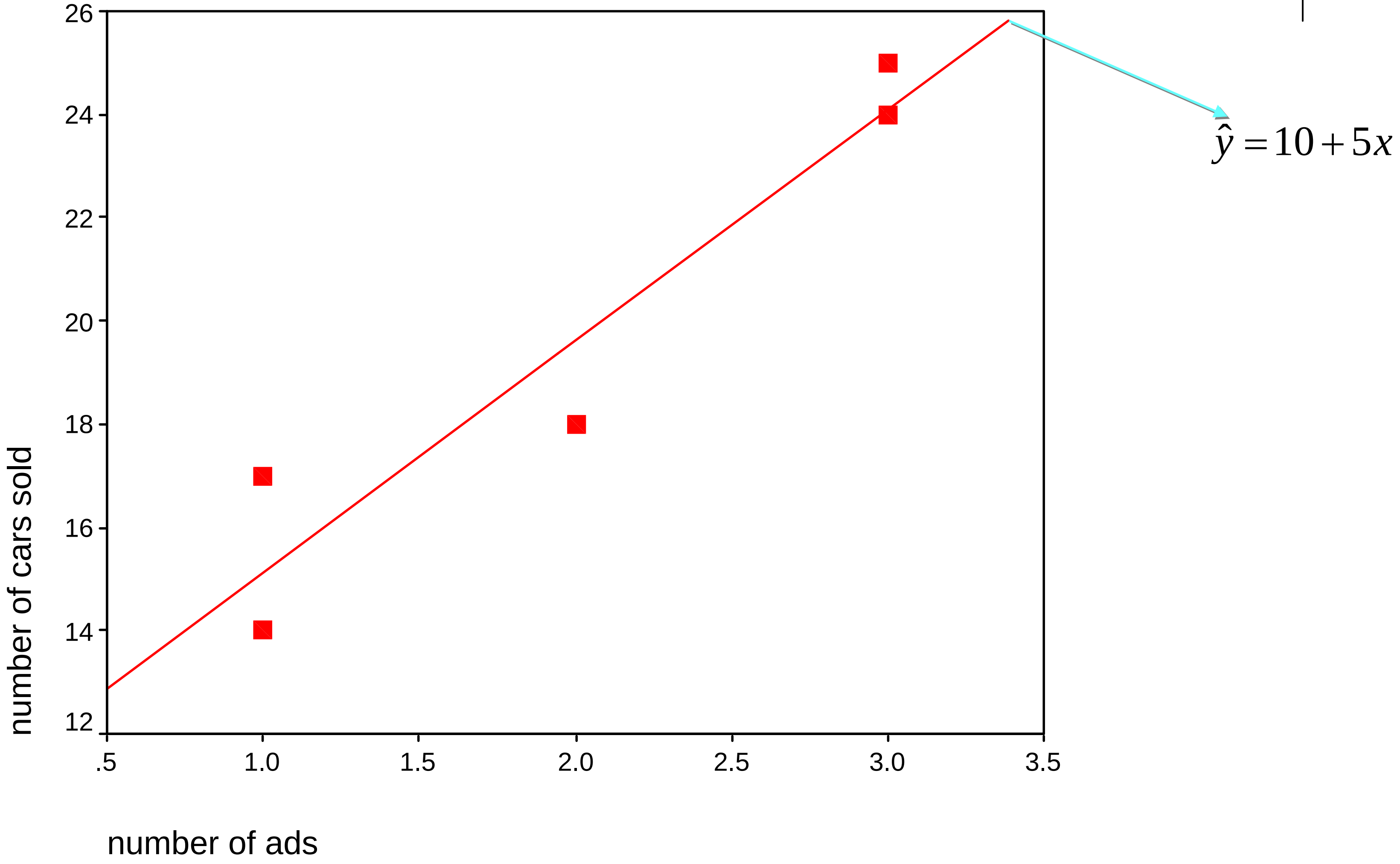
- *y*-Intercept for the Estimated Regression Equation

$$b_0 = \bar{y} - b_1\bar{x} = 20 - 5*(2) = 10$$

- Estimated Regression Equation

$$\hat{y} = 10 + 5*x$$

Scatter Diagram and Regression Line



Estimate of Residuals

x	y	$\hat{y}$	$e = y - \hat{y}$
1	14	15	-1.0
3	24	25	-1.0
2	18	20	-2.0
1	17	15	2.0
3	27	25	2.0

Decomposition of total sum of squares

- Relationship Among SST, SSR, SSE

SST = SSR + SSE

$$\sum (y_i - \bar{y})^2 = \sum (\hat{y}_i - \bar{y})^2 + \sum (y_i - \hat{y}_i)^2$$

where:

SST = total sum of squares

SSR = sum of squares due to regression

SSE = sum of squares due to error

Decomposition of total sum of squares

$e = y - \hat{y}$	$SSE = \sum (y - \hat{y})^2$	$\hat{y}$	$\hat{y} - \bar{y}$	$SSR = \sum (\hat{y} - \bar{y})^2$
-1	1	15	-5	25
-1	1	25	5	25
-2	4	20	0	0
2	4	15	-5	25
2	4	25	5	25
	SSE=14			SSR=100

- Check if SST= SSR + SSE

Coefficient of Determination

- The coefficient of determination is:

$$r^2 = SSR/SST$$

$$r^2 = SSR/SST = 100/114 = 0.8772$$

- The regression relationship is very strong; about 88% of the variability in the number of cars sold can be explained by the number of TV ads.
- The coefficient of determination (r^2) is also the square of the correlation coefficient (r).

Sample Correlation Coefficient

$$r = (\text{sign of } b_1)\sqrt{\text{Coefficient of Determination}}$$

$$r = (\text{sign of } b_1)\sqrt{r^2}$$

$$r = +\sqrt{0.8772} = 0.936$$

Sampling Distribution of b_1

- Expected value of b_1:

$$E(b_1) = b_1$$

- Variance of b_1:

$$\sigma_1^2 = Var(b_1) = \sigma^2/SS_x$$

- Standard error of b_1:

$$\sigma_1 = \sqrt{Var(b_1)}$$

Estimate of σ²

- The mean square error (MSE) provides the estimate of σ^2.

$$s^2 = \text{MSE} = \text{SSE}/(n-2)$$

where:

$$\text{SSE} = \sum (y_i - \hat{y}_i)^2$$

Interval Estimate of β_1:

- $(1-\alpha)100\%$ confidence interval for b_1 is:

$$b_1 \pm t_{\alpha/2} \times \sigma_1$$

- Where $t_{\alpha/2}$ is the value from t distribution with (n-2) degrees of freedom such that probability in the upper tail is $\alpha/2$.

Example: Reed Auto Sales

- $s^2 = \text{MSE} = \text{SSE}/(n - 2) = 14/3 = 4.67$

$$\text{Var}(b_1) = s^2/SS_x = 4.67/4 = 1.17$$

$$\sigma_1 = \sqrt{1.17} = 1.08$$

- 95% confidence interval for β_1:

$$5 \pm 3.182 \times 1.08 = 5 \pm 3.44$$

- We can say we 95% confidence that β_1 will lie between 1.56 and 8.44.

Testing for Significance: *t* Test

- Hypotheses

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

- Test Statistic

$$t = \frac{b_1 - 0}{SE(b_1)}$$

- Where b_1 is the slope estimate and $SE(b_1)$ is the standard error of b_1.

Testing for Significance: t Test

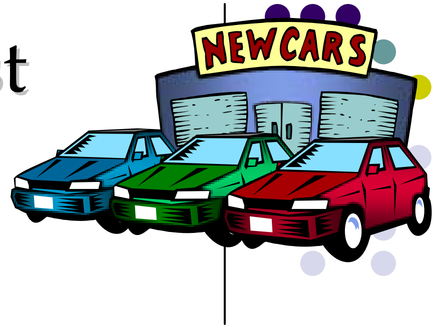
- Rejection Rule

  - Reject H_0 if p-value $\leq \alpha$
    or $t \leq -t_{\alpha/2}$ or $t \geq t_{\alpha/2}$

  - where:

    $t_{\alpha/2}$ is based on a t distribution
    with n - 2 degrees of freedom

Testing for Significance: t Test



1. Determine the hypotheses.

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

2. Specify the level of significance.

$$\alpha = .05$$

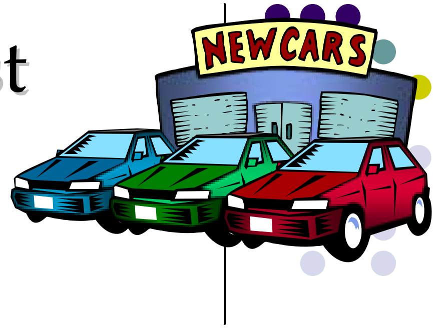
3. Select the test statistic.

$$t = \frac{b_1}{SE(b_1)}$$

4. State the rejection rule.

Reject H_0 if p-value $\leq .05$
or t ≤ 3.182 or t ≥ 3.182

Testing for Significance: t Test



- 5. Compute the value of the test statistic.

$$t = \frac{b_1}{SE(b_1)} = \frac{5}{1.08} = 4.63$$

- 6. Determine whether to reject H_0.

$t = 4.63 > t_{\alpha/2} = 3.182$. We can reject H_0.

Some Cautions about the Interpretation of Significance Tests

- Rejecting H_0: $\beta_1 = 0$ and concluding that the relationship between x and y is significant does not enable us to conclude that a cause-and-effect relationship is present between x and y.

- Just because we are able to reject H_0: $\beta_1 = 0$ and demonstrate statistical significance does not enable us to conclude that there is a linear relationship between x and y.