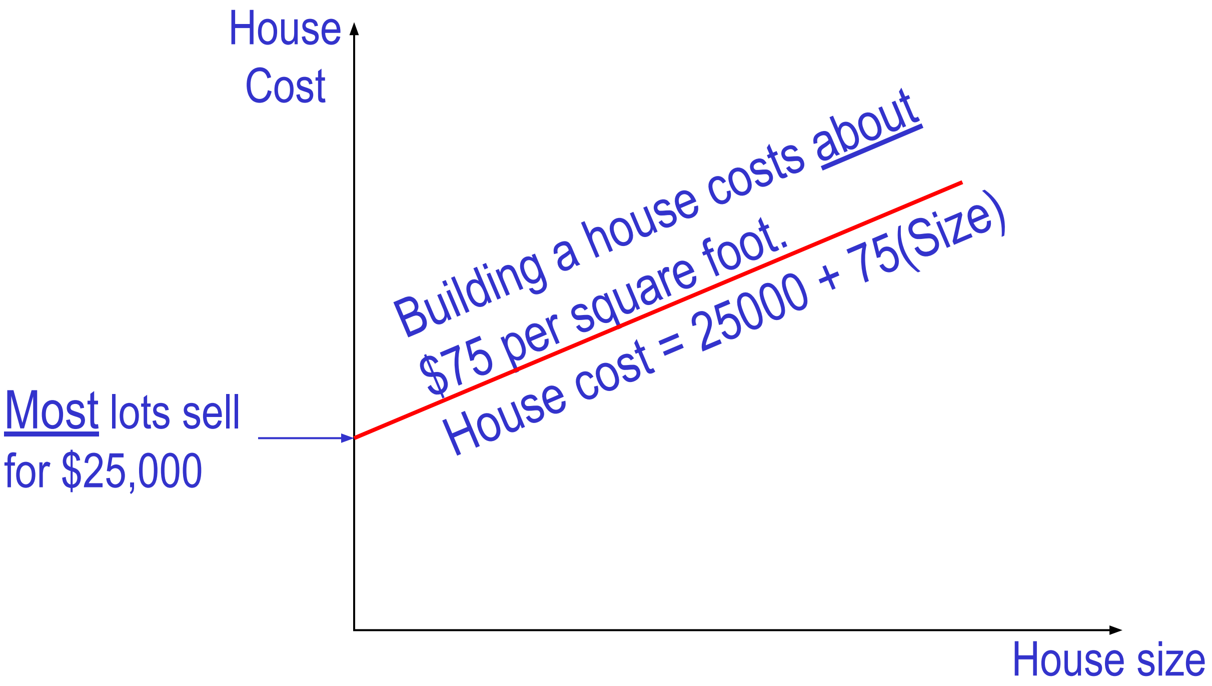
Simple Linear Regression

Introduction

- In Chapters 17 to 19, we examine the relationship between interval variables 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



House Cost

Building a house costs about $75 per square foot.

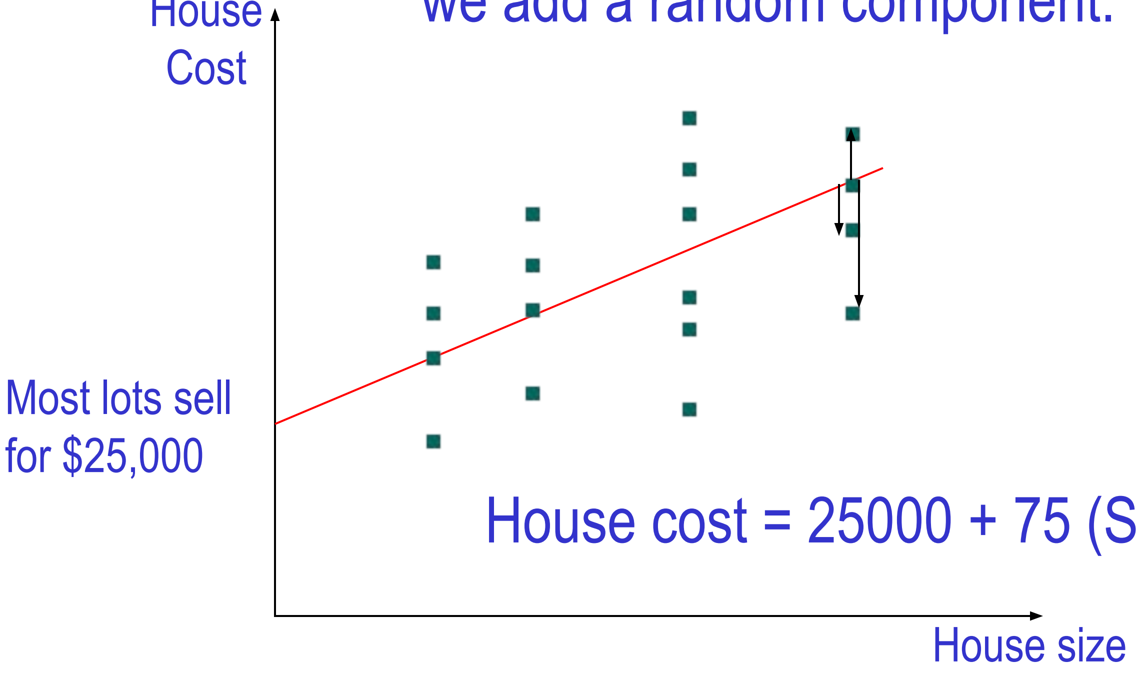
House cost = 25000 + 75(Size)

Most lots sell for $25,000

House size

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 first order linear model

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

Y = dependent variable

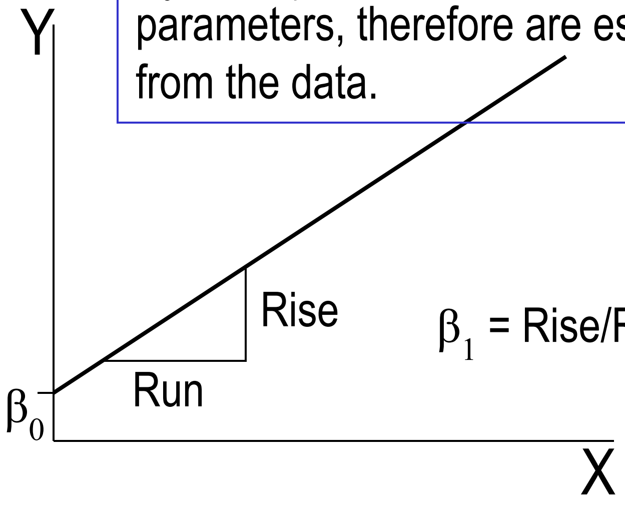
X = independent variable

β_0 = Y-intercept

β_1 = slope of the line

ε = error variable

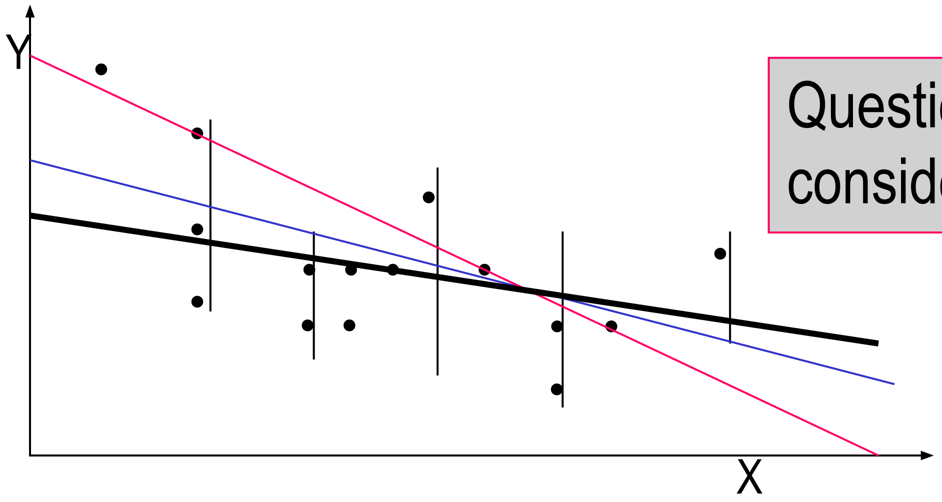
β_0 and β_1 are unknown population parameters, therefore are estimated from the data.



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

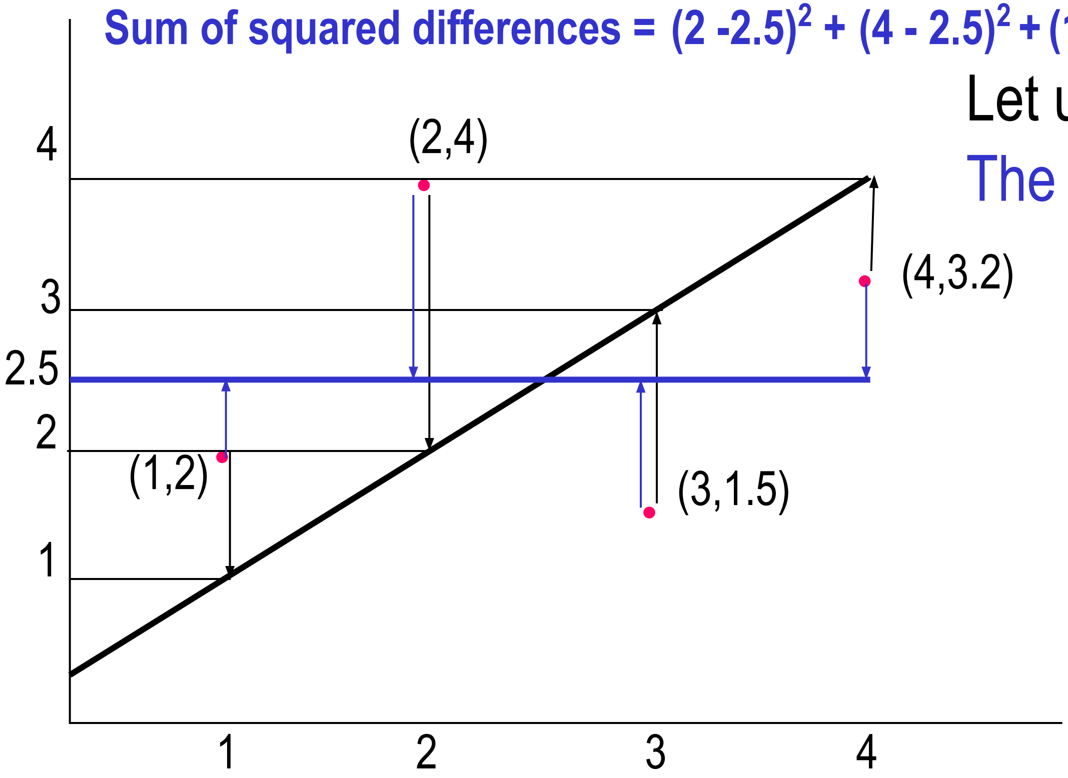
A good line is one that minimizes
the sum of squared differences between the points and the line.

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 line coefficients, that minimize the differences between the data points and the line, use the formulas:

$$b_1 = \frac{\operatorname{cov}(X,Y)}{s_X^2}\left(= \frac{s_{XY}}{s_X^2}\right)$$

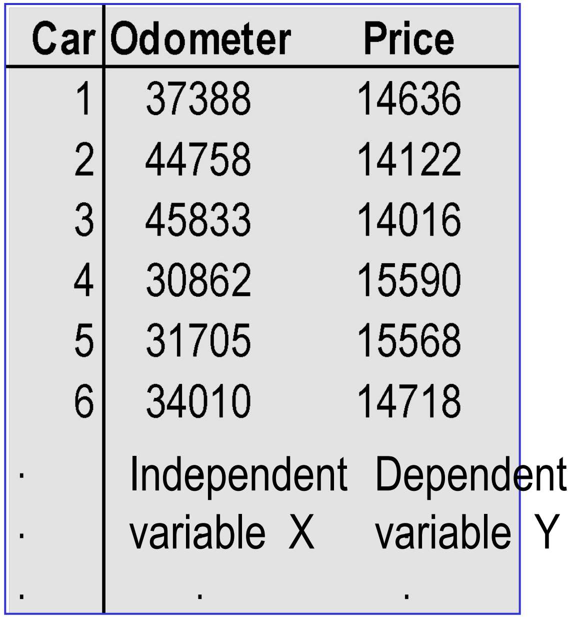
$$b_0 = \bar{Y} - b_1\bar{X}$$

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

$$\hat{Y} = b_0 + b_1X$$

The Simple Linear Regression Line

- Example 17.2 (Xm17-02)
  - 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 variable X

Dependent variable Y

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

$$\bar{X} = 36{,}009.45; \quad s_X^2 = \frac{\sum (X_i - \bar{X})^2}{n-1} = 43{,}528{,}690$$

$$\bar{Y} = 14{,}822.823; \quad \text{cov}(X,Y) = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{n-1} = -2{,}712{,}51$$

where n = 100.

$$b_1 = \frac{\text{cov}(X,Y)}{s_X^2} = \frac{-1{,}712{,}511}{43{,}528{,}690} = -.06232$$

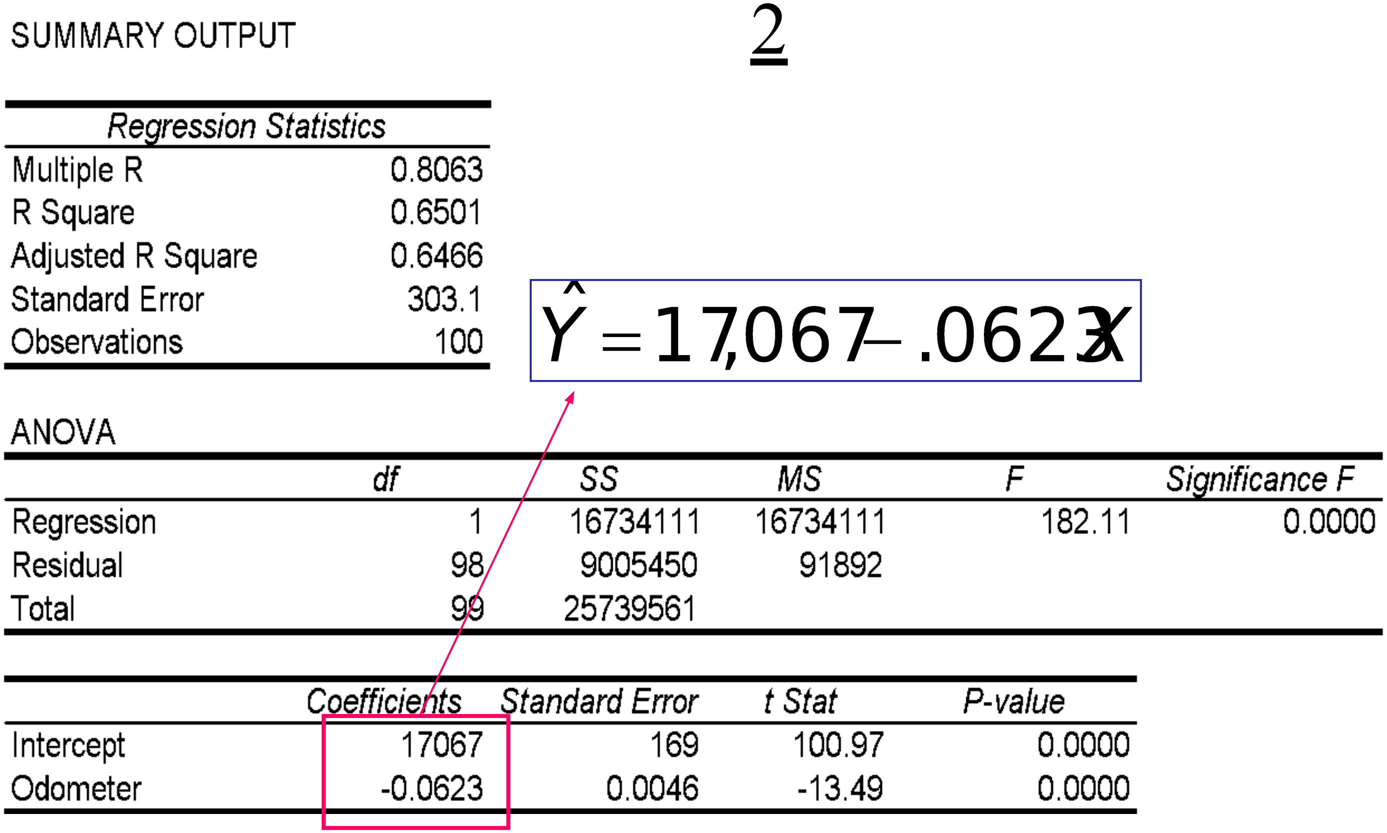
$$b_0 = \bar{Y} - b_1\bar{X} = 14{,}822.82 - (-.06232)(36{,}009.45) = 17{,}06$$

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

- Solution – continued
  - Using the computer (Xm17-02)

    Tools > Data Analysis > Regression > [Shade the Y range and the X range] > OK

Xm17-02



SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.8063
R Square	0.6501
Adjusted R Square	0.6466
Standard Error	303.1
Observations	100

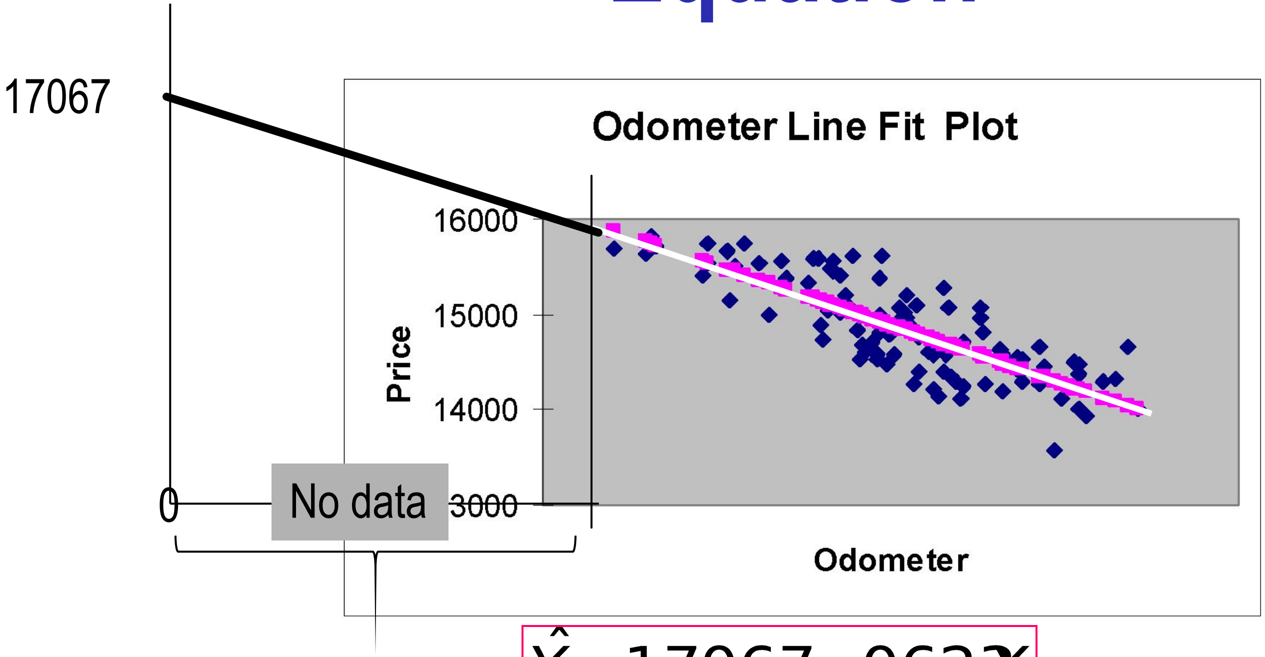
ANOVA

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

	Coefficients	*Standard Error*	*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



$$\hat{Y} = 17{,}067 - .0623X$$

The intercept is b_0 = $17067.

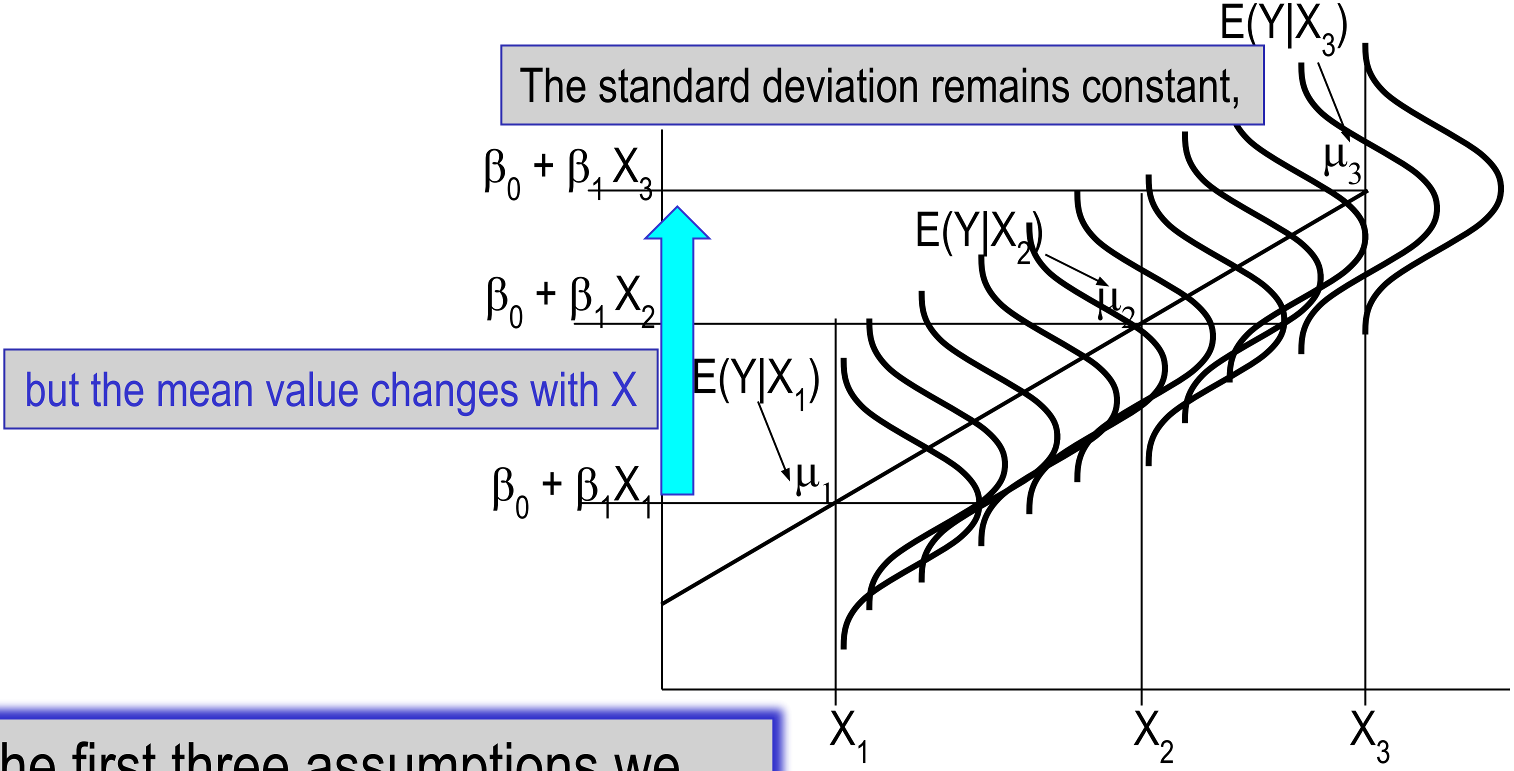
Do not interpret the intercept as the "Price of cars that have not been driven"

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

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 are 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

$$\boxed{SSE = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2.}$$

- A shortcut formula

$$\boxed{SSE = (n-1)s_Y^2 - \frac{[\mathrm{cov}(X,Y)]^2}{s_X^2}}$$

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

StandardErrorof Estimat

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

- Example 17.3
  - Calculate the standard error of estimate for Example 17.2, and describe what does it tell you about the model fit?
- Solution

$$s_Y^2 = \frac{\sum (Y_i - \hat{Y}_i)^2}{n-1} = 259{,}996$$

Calculated before

$$\text{SSE} = (n-1)s_Y^2 - \frac{[\text{cov}(X,Y)]^2}{s_X^2} = 99(259{,}996) - \frac{(-2{,}712{,}511)^2}{43{,}528{,}690} = 9{,}005{,}450$$

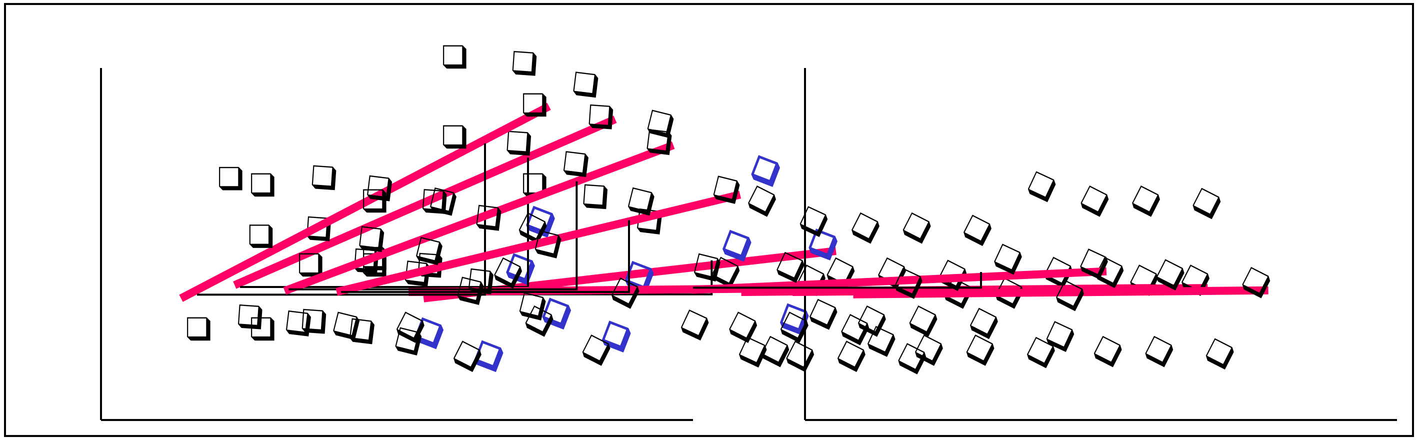
$$s_\varepsilon = \sqrt{\frac{\text{SSE}}{n-2}} = \sqrt{\frac{9{,}005{,}450}{98}} = 303.13$$

It is hard to assess the model based on s_ε even when compared with the mean value of Y.

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

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

- We can draw inference about β_1 from b_1 by testing

  H_0: $\beta_1 = 0$

  H_1: $\beta_1 \neq 0$ (or < 0,or > 0)

  - The test statistic is

$$\boxed{t = \frac{b_1 - \beta_1}{s_{b_1}}} \quad \text{where} \quad \boxed{s_{b_1} = \frac{s_\varepsilon}{\sqrt{(n-1)s_X^2}}}$$

The standard error of b_1.

  - If the error variable is normally distributed, the statistic has Student t distribution with d.f. = n-2.

- Example 17.4
  - Test to determine whether there is enough evidence to infer that there is a linear relationship between the car auction price and the odometer reading for all three-year-old Tauruses, in Example 17.2.
    Use α = 5%.

- Solving by hand
  - To compute “t” we need the values of b_1 and s_{b1}.

$$b_1 = -.0623$$

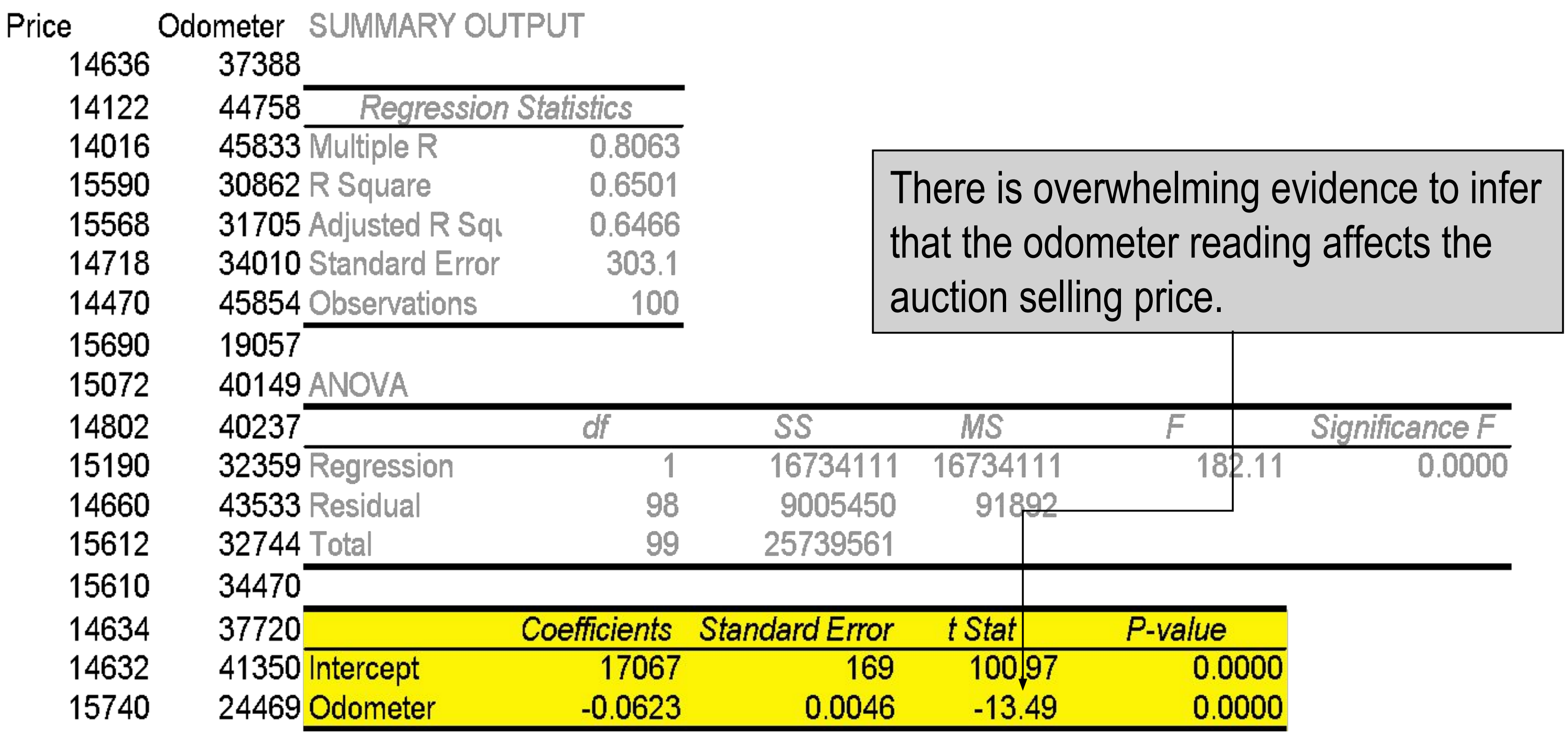
$$s_{b_1} = \frac{s_\varepsilon}{\sqrt{(n-1)s_X^2}} = \frac{3031}{\sqrt{(99)(43,528,690)}} = .0046$$

$$t = \frac{b_1 - \beta_1}{s_{b_1}} = \frac{-.0623 - 0}{.00462} = -13.49$$

  - The rejection region is t > $t_{.025}$ or t < -$t_{.025}$ with ν = n-2 = 98. Approximately, $t_{.025}$ = 1.984

- Using the computer

Xm17-0
2



Price	Odometer
14636	37388
14122	44758
14016	45833
15590	30862
15568	31705
14718	34010
14470	45854
15690	19057
15072	40149
14802	40237
15190	32359
14660	43533
15612	32744
15610	34470
14634	37720
14632	41350
15740	24469

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.8063
R Square	0.6501
Adjusted R Squ	0.6466
Standard Error	303.1
Observations	100

ANOVA

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

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

There is overwhelming evidence to infer that the odometer reading affects the auction selling price.

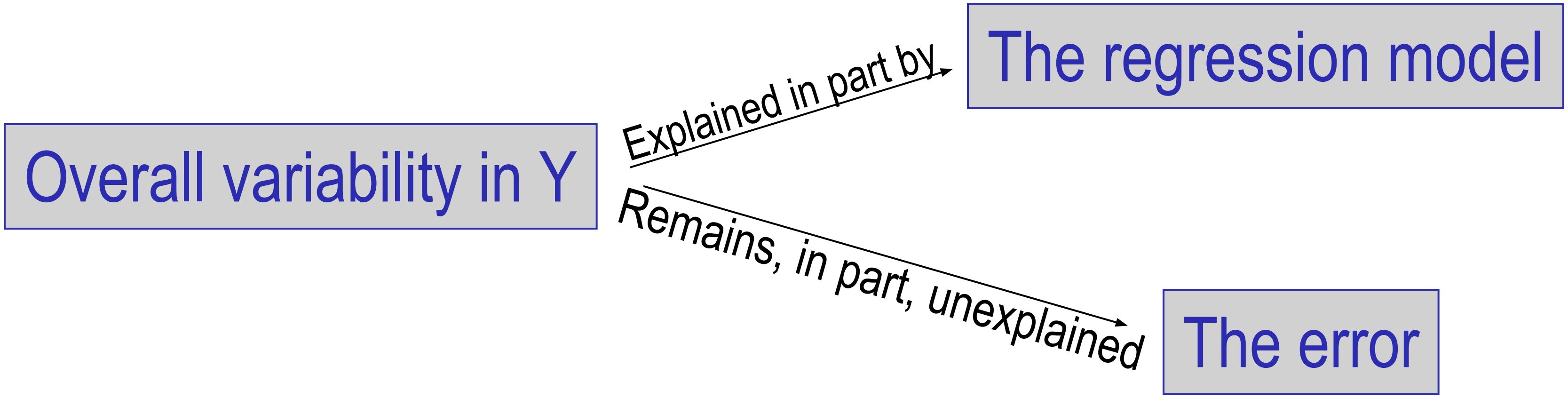
Coefficient of Determination

– To measure the strength of the linear relationship we use the coefficient of determination:

$$R^2 = \frac{[\mathrm{cov}(X,Y)]^2}{S_X^2 S_Y^2} \quad \left(\text{or,} = r_{XY}^2\right);$$

$$\text{or, } R^2 = 1 - \frac{\mathrm{SSE}}{\sum (Y_i - \bar{Y})^2} \qquad \text{(see p. 18 ab}$$

- To understand the significance of this coefficient note:



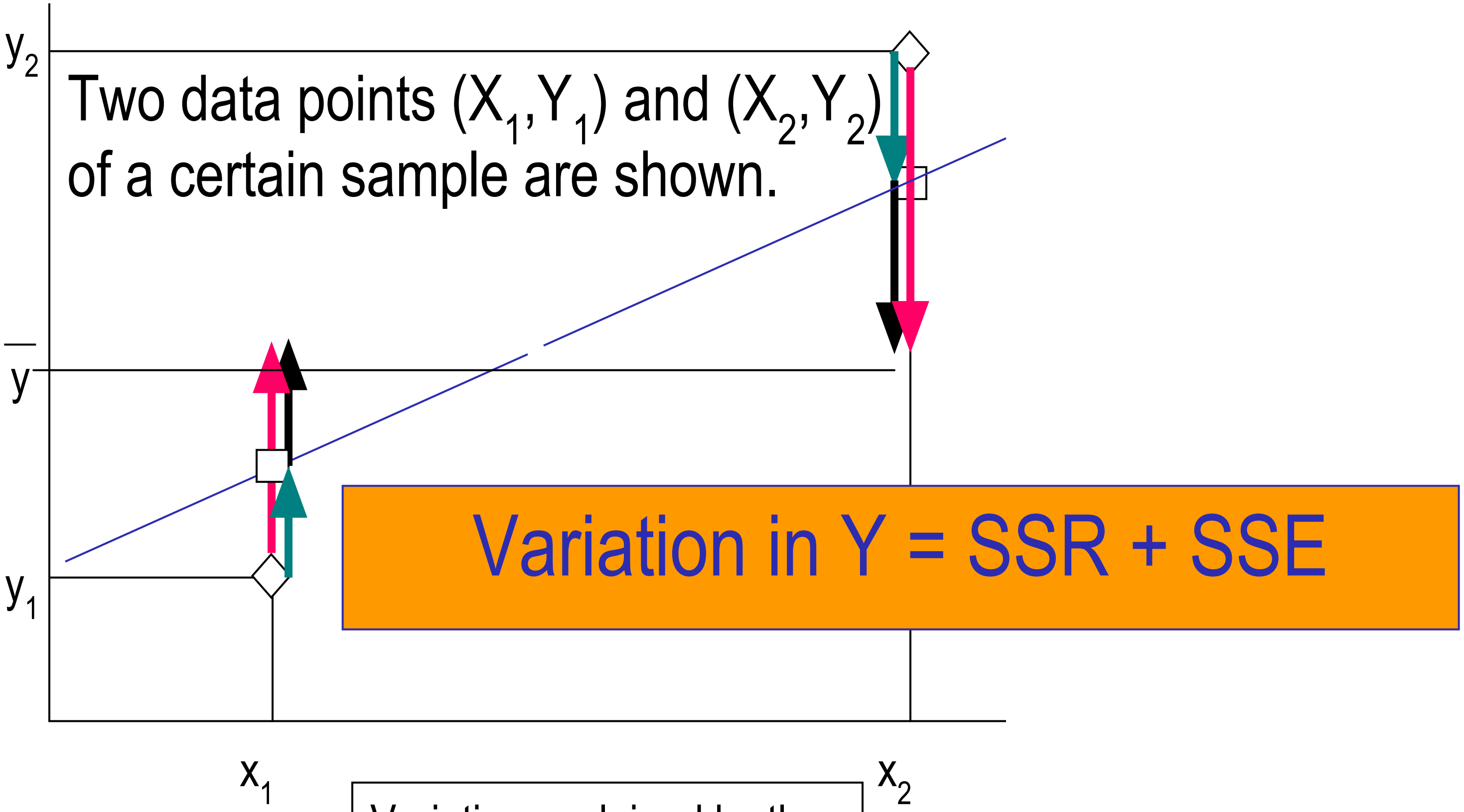
Overall variability in Y

Explained in part by

The regression model

Remains, in part, unexplained

The error



Two data points (X_1,Y_1) and (X_2,Y_2) of a certain sample are shown.

Variation in Y = SSR + SSE

Total variation in Y =

Variation explained by the regression line

+ Unexplained variation (error)

$$(Y_1-\bar{Y})^2+(Y_2-\bar{Y})^2 = \quad (\hat{Y}_1-\bar{Y})^2+(\hat{Y}_2-\bar{Y})^2 \quad +(Y_1-\hat{Y}_1)^2+(Y_2-\hat{Y}_2)^2$$

- R^2 measures the proportion of the variation in Y that is explained by the variation in X.

$$R^2 = 1 - \frac{SSE}{\sum (Y_i - \bar{Y})^2} = \frac{\sum (Y_i - \bar{Y})^2 - SSE}{\sum (Y_i - \bar{Y})^2} = \frac{SSR}{\sum (Y_i - \bar{Y})^2}$$

- R^2 takes on any value between zero and one.

  R^2 = 1: Perfect match between the line and the data points.

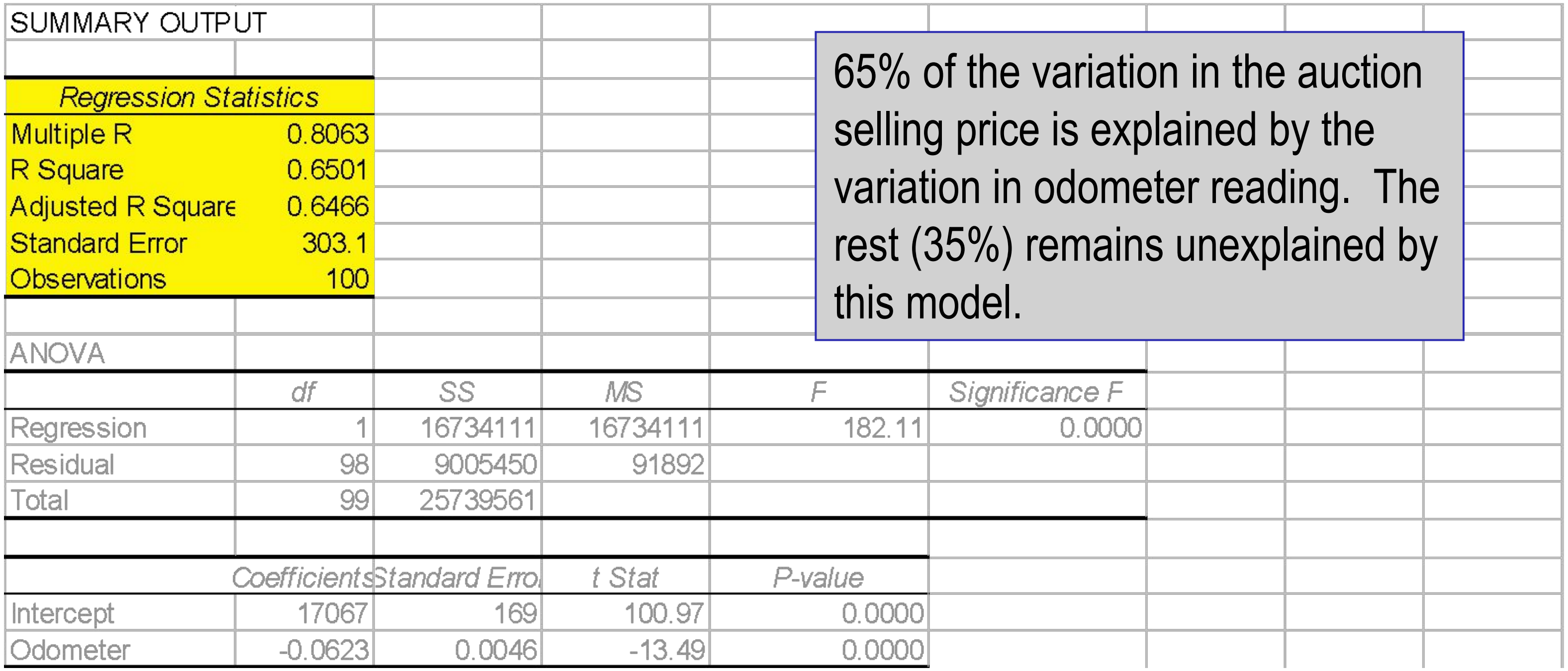
  R^2 = 0: There are no linear relationship between X and Y.

- Example 17.5
  - Find the coefficient of determination for Example 17.2; what does this statistic tell you about the model?
- Solution
  - Solving by hand;

$$R^2 = \frac{[\text{cov}(X,Y)]^2}{s_X^2 s_Y^2} = \frac{[-2{,}712{,}511]^2}{(43{,}528{,}688)(259{,}996)} = .650$$

– Using the computer

From the regression output we have



SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.8063
R Square	0.6501
Adjusted R Square	0.6466
Standard Error	303.1
Observations	100

ANOVA

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

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

65% of the variation in the auction selling price is explained by the variation in odometer reading. The rest (35%) remains unexplained by this model.