Statistics Foundation

Simple Linear Regression

Digital Vidya

Correlation Analysis

- Correlation analysis is used to measure strength of the association (linear relationship) between two variables
  - Correlation is only concerned with strength of the relationship
  - No causal effect is implied with correlation

Digital Vidya

Correlation Analysis

- The population correlation coefficient is denoted ρ (the Greek letter rho)
- The sample correlation coefficient is

$$r = \frac{s_{xy}}{s_x s_y}$$

where

$$s_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

Digital Vidya

Introduction to Regression Analysis

- Regression analysis is used to:
  - Predict the value of a dependent variable based on the value of at least one independent variable
  - Explain the impact of changes in an independent variable on the dependent variable

Dependent variable: the variable we wish to explain
(also called the endogenous variable)

Independent variable: the variable used to explain the dependent variable
(also called the exogenous variable)

Digital Vidya

Linear Regression Model

- The relationship between X and Y is described by a linear function
- Changes in Y are assumed to be caused by changes in X
- Linear regression population equation model

$$\boxed{Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i}$$

- Where β_0 and β_1 are the population model coefficients and ε is a random error term.

Digital Vidya

Simple Linear Regression Model

The population regression model:

Population Y intercept

Population Slope Coefficient

Independent Variable

Random Error term

Dependent Variable

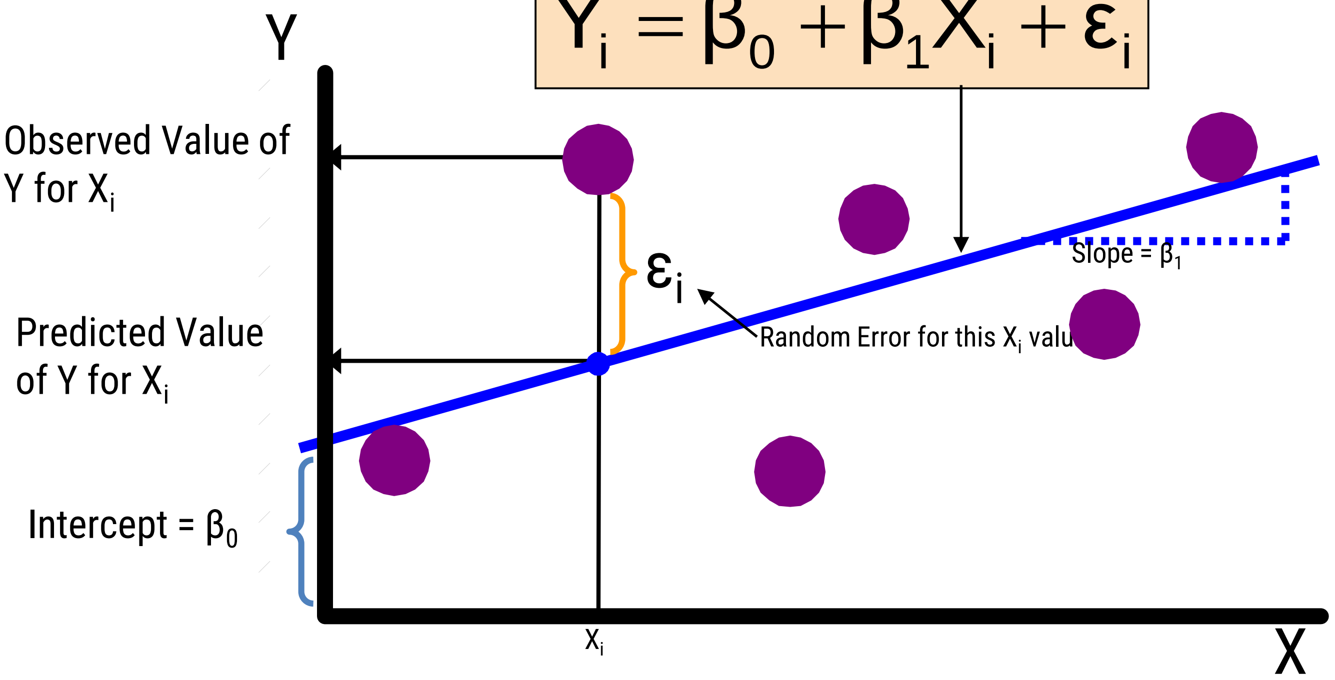
$$Y_i = \underbrace{\beta_0 + \beta_1 X_i}_{\text{Linear component}} + \underbrace{\varepsilon_i}_{\text{Random Error component}}$$

Digital Vidya

Simple Linear Regression Model

(continued)

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$



Digital Vidya

Simple Linear Regression Equation

The simple linear regression equation provides an estimate of the population regression line

Estimated (or predicted) y value for observation i

Estimate of the regression intercept

Estimate of the regression slope

Value of x for observation i

$$\hat{y}_i = b_0 + b_1 x_i$$

The individual random error terms e_i have a mean of zero

$$e_i = (y_i - \hat{y}_i) = y_i - (b_0 + b_1 x_i)$$

Digital Vidya

Least Squares Estimators

- b_0 and b_1 are obtained by finding the values of b_0 and b_1 that minimize the sum of the squared differences between y and $\hat{y}$:

$$\boxed{\begin{aligned} \min \ SSE &= \min \sum e_i^2 \\ &= \min \sum (y_i - \hat{y}_i)^2 \\ &= \min \sum [y_i - (b_0 + b_1 x_i)]^2 \end{aligned}}$$

Differential calculus is used to obtain the coefficient estimators b_0 and b_1 that minimize SSE

Least Squares Estimators

(continued)

- The slope coefficient estimator is

$$b_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = r_{xy}\frac{s_Y}{s_X}$$

- And the constant or y-intercept is

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

- The regression line always goes through the mean $\bar{x}, \bar{y}$

Digital Vidya

Finding the Least Squares Equation

- The coefficients b_0 and b_1, and other regression results in this module, will be found using a computer
  - Hand calculations are tedious
  - Statistical routines are built into Excel
  - Other statistical analysis software can be used

Linear Regression Model Assumptions

- The true relationship form is linear (Y is a linear function of X, plus random error)
- The error terms, ε_i are independent of the x values
- The error terms are random variables with mean 0 and constant variance, σ^2

(the constant variance property is called homoscedasticity)

- The random error terms, ε_i, are not correlated with one another, so that

$$\boxed{E[\varepsilon_i] = 0 \quad \text{and} \quad E[\varepsilon_i^2] = \sigma^2 \quad \text{for } (i = 1, \ldots, n)}$$

$$\boxed{E[\varepsilon_i \varepsilon_j] = 0 \quad \text{for all } i \neq j}$$

Digital Vidya

Interpretation of the Slope and the Intercept

- b_0 is the estimated average value of y when the value of x is zero (if $x = 0$ is in the range of observed x values)

- b_1 is the estimated change in the average value of y as a result of a one-unit change in x

Simple Linear Regression Example

- A real estate agent wishes to examine the relationship between the selling price of a home and its size (measured in square feet)

- A random sample of 10 houses is selected
  - Dependent variable (Y) = house price in $1000s
  - Independent variable (X) = square feet

Digital Vidya

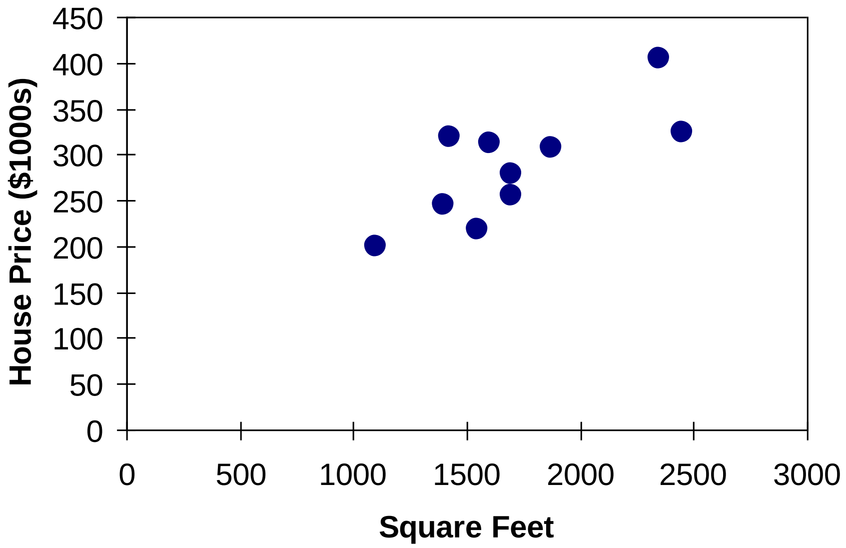
Sample Data for House Price Model

House Price in $1000s (Y)	Square Feet (X)
245	1400
312	1600
279	1700
308	1875
199	1100
219	1550
405	2350
324	2450
319	1425
255	1700

Digital Vidya

Graphical Presentation

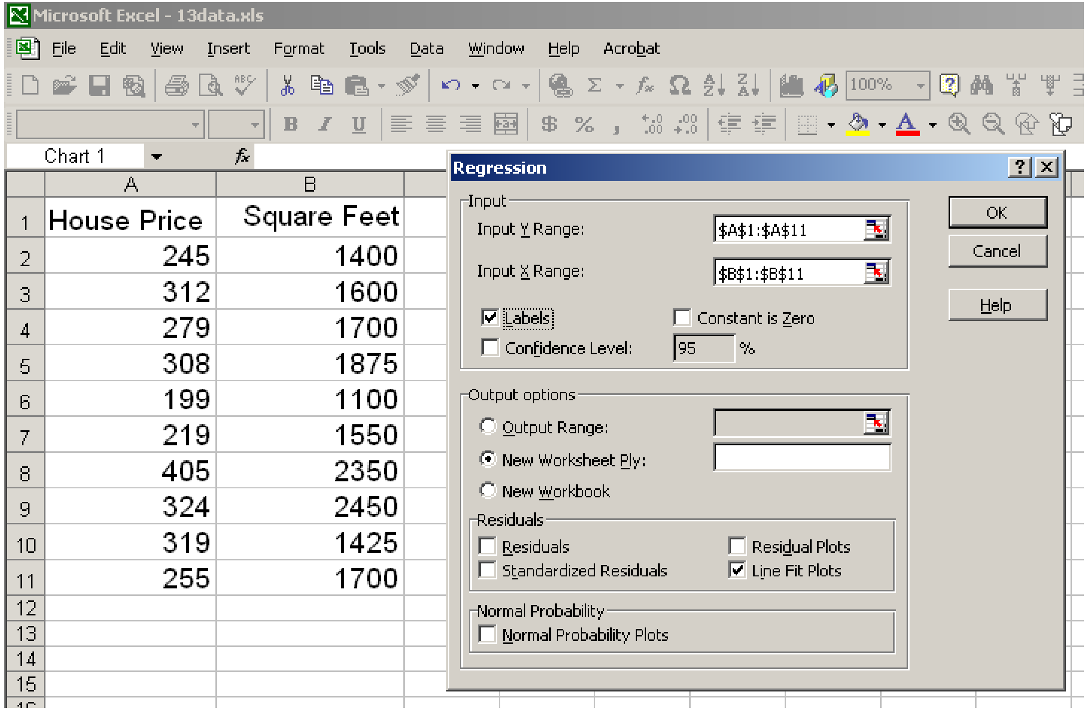
- House price model: scatter plot



Digital Vidya

Regression Using Excel

- Tools / Data Analysis / Regression



Digital Vidya

Excel Output

Regression Statistics	
Multiple R	0.76211
R Square	0.58082
Adjusted R Square	0.52842
Standard Error	41.33032
Observations	10

The regression equation is:

$$\widehat{\text{house price}} = 98.24833 + 0.10977\ (\text{square feet})$$

ANOVA

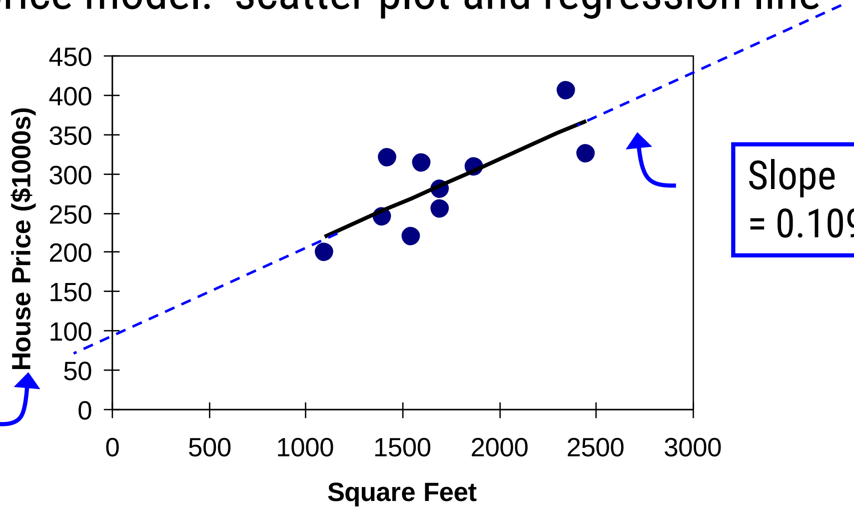
	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	18934.9348	18934.9348	11.0848	0.01039
Residual	8	13665.5652	1708.1957		
Total	9	32600.5000			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	98.24833	58.03348	1.69296	0.12892	-35.57720	232.07386
Square Feet	0.10977	0.03297	3.32938	0.01039	0.03374	0.18580

Digital Vidya

Graphical Presentation

- House price model: scatter plot and regression line



Slope
= 0.10977

Intercept
= 98.248

$$\widehat{\text{house price}} = 98.24833 + 0.10977\,(\text{square feet})$$

Digital Vidya

Interpretation of the Intercept, b_0

$$\widehat{\text{house price}} = 98.24833 + 0.10977\,(\text{square feet})$$

- b_0 is the estimated average value of Y when the value of X is zero (if X = 0 is in the range of observed X values)
  - Here, no houses had 0 square feet, so $\boxed{b_0 = 98.24833}$ just indicates that, for houses within the range of sizes observed, $98,248.33 is the portion of the house price not explained by square feet

Digital Vidya

Interpretation of the Slope Coefficient, b_1

$$\widehat{\text{house price}} = 98.24833 + 0.10977\,(\text{square feet})$$

- b_1 measures the estimated change in the average value of Y as a result of a one-unit change in X
  - Here, $\boxed{b_1 = .10977}$ tells us that the average value of a house increases by .10977($1000) = $109.77, on average, for each additional one square foot of size

Measures of Variation

- Total variation is made up of two parts:

$$SST = SSR + SSE$$

Total Sum of Squares	Regression Sum of Squares	Error Sum of Squares
$SST = \sum (y_i - \bar{y})^2$	$SSR = \sum (\hat{y}_i - \bar{y})^2$	$SSE = \sum (y_i - \hat{y}_i)^2$

where:

$\bar{y}$ = Average value of the dependent variable

y_i = Observed values of the dependent variable

$\hat{y}_i$ = Predicted value of y for the given x_i value

Digital Vidya

Measures of Variation

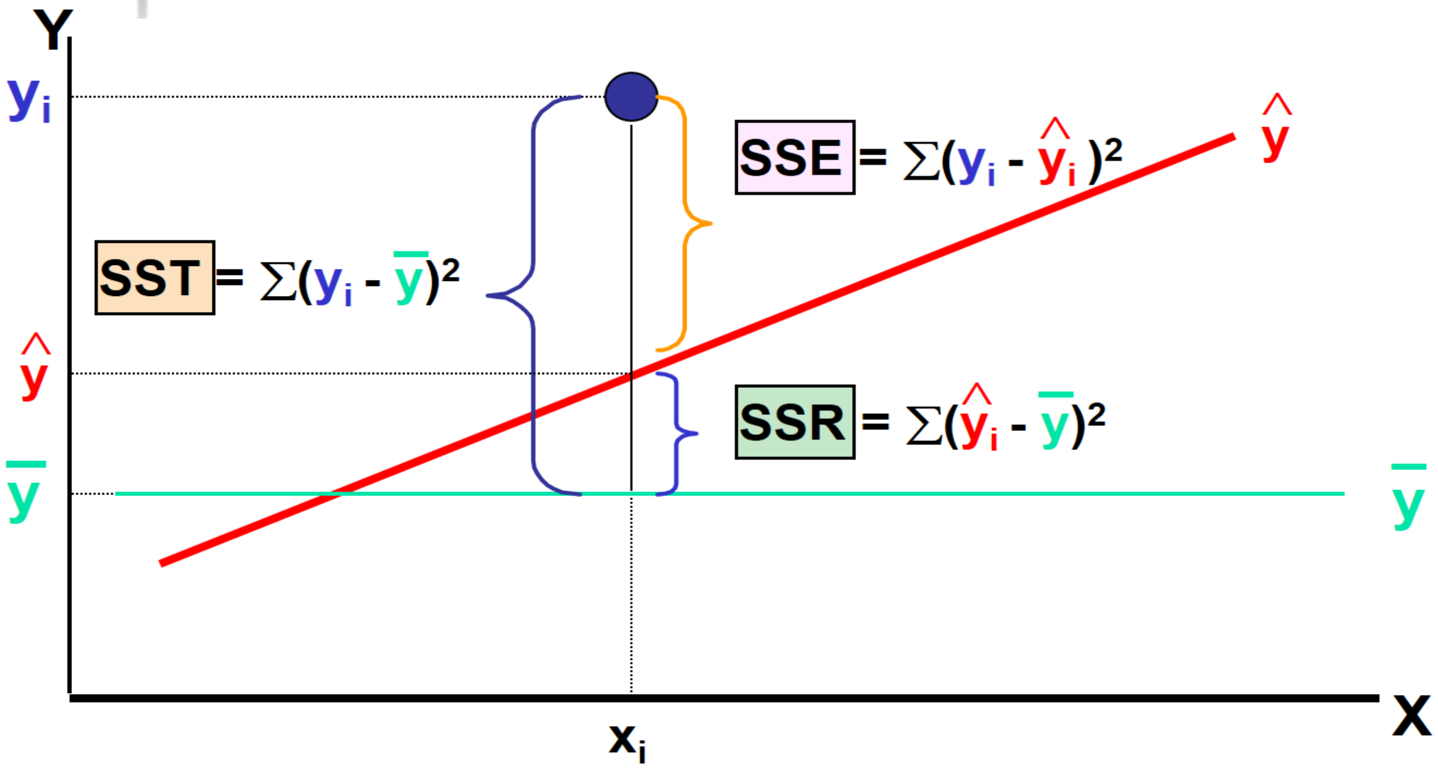
(continued)

- SST = total sum of squares
  - Measures the variation of the y_i values around their mean, $\bar{y}$
- SSR = regression sum of squares
  - Explained variation attributable to the linear relationship between x and y
- SSE = error sum of squares
  - Variation attributable to factors other than the linear relationship between x and y

Digital Vidya

Measures of Variation

(continued)



Digital Vidya

Coefficient of Determination, R^2

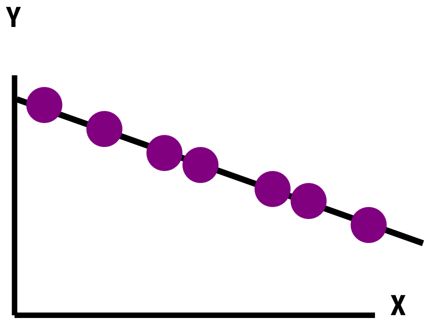
- The coefficient of determination is the portion of the total variation in the dependent variable that is explained by variation in the independent variable
- The coefficient of determination is also called R-squared and is denoted as R^2

$$R^2 = \frac{SSR}{SST} = \frac{\text{regression sum of squares}}{\text{total sum of squares}}$$

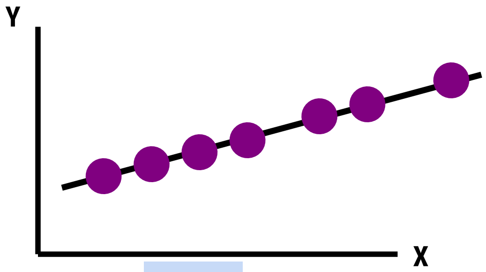
note: $0 \leq R^2 \leq 1$

Digital Vidya

Examples of Approximate r^2 Values



$r^2 = 1$



$r^2 = 1$

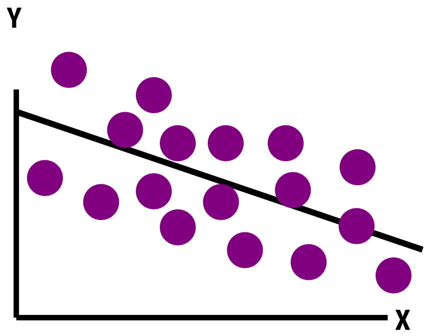
$r^2 = 1$

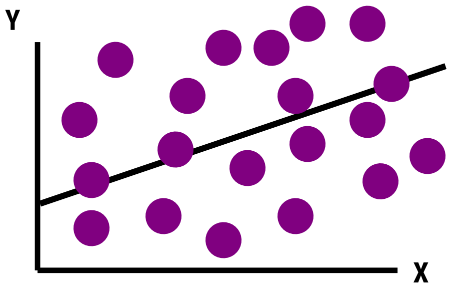
Perfect linear relationship between X and Y:

100% of the variation in Y is explained by variation in X

Digital Vidya

Examples of Approximate r2 Values





$0 < r^2 < 1$

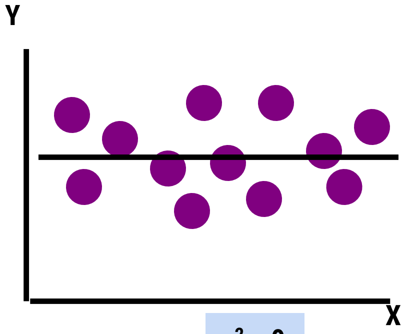
Weaker linear relationships between X and Y:

Some but not all of the variation in Y is explained by variation in X

Digital Vidya

Examples of Approximate r^2 Values

$r^2 = 0$



$r^2 = 0$

No linear relationship between X and Y:

The value of Y does not depend on X. (None of the variation in Y is explained by variation in X)

Digital Vidya

Excel Output

Regression Statistics	
Multiple R	0.76211
R Square	0.58082
Adjusted R Square	0.52842
Standard Error	41.33032
Observations	10

$$R^2 = \frac{SSR}{SST} = \frac{18934.9348}{32600.5000} = 0.58082$$

58.08% of the variation in house prices is explained by variation in square feet

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	18934.9348	18934.9348	11.0848	0.01039
Residual	8	13665.5652	1708.1957		
Total	9	32600.5000			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	98.24833	58.03348	1.69296	0.12892	-35.57720	232.07386
Square Feet	0.10977	0.03297	3.32938	0.01039	0.03374	0.18580

Correlation and R_2

- The coefficient of determination, R^2, for a simple regression is equal to the simple correlation squared

$$\boxed{R^2 = r_{xy}^2}$$

Digital Vidya

Estimation of Model Error Variance

- An estimator for the variance of the population model error is

$$\boxed{\hat{\sigma}^2 = s_e^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-2} = \frac{SSE}{n-2}}$$

- Division by n – 2 instead of n – 1 is because the simple regression model uses two estimated parameters, b_0 and b_1, instead of one

$\boxed{s_e = \sqrt{s_e^2}}$ is called the standard error of the estimate

Digital Vidya

Excel Output

Regression Statistics	
Multiple R	0.76211
R Square	0.58082
Adjusted R Square	0.52842
Standard Error	41.33032
Observations	10

$$s_e = 41.33032$$

ANOVA

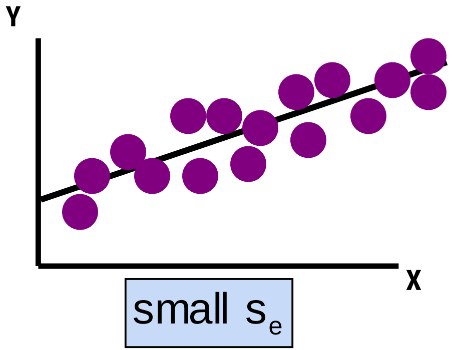
	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	18934.9348	18934.9348	11.0848	0.01039
Residual	8	13665.5652	1708.1957		
Total	9	32600.5000			

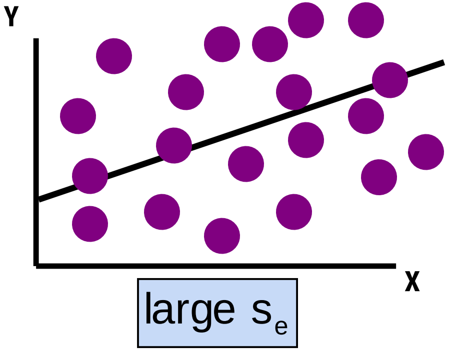
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	98.24833	58.03348	1.69296	0.12892	-35.57720	232.07386
Square Feet	0.10977	0.03297	3.32938	0.01039	0.03374	0.18580

Digital Vidya

Comparing Standard Errors

s_e is a measure of the variation of observed y values from the regression line





The magnitude of s_e should always be judged relative to the size of the y values in the sample data

i.e., s_e = \$41.33K is moderately small relative to house prices in the \$200 - \$300K range

Digital Vidya

Inferences About the Regression Model

- The variance of the regression slope coefficient (b_1) is estimated by

$$\boxed{s^2_{b_1} = \frac{s^2_e}{\sum (x_i - \bar{x})^2} = \frac{s^2_e}{(n-1)s^2_x}}$$

where:

s_{b_1} = Estimate of the standard error of the least squares slope

$s_e = \sqrt{\frac{SSE}{n-2}}$ = Standard error of the estimate

Digital Vidya

Excel Output

Regression Statistics	
Multiple R	0.76211
R Square	0.58082
Adjusted R Square	0.52842
Standard Error	41.33032
Observations	10

$$s_{b_1} = 0.03297$$

ANOVA

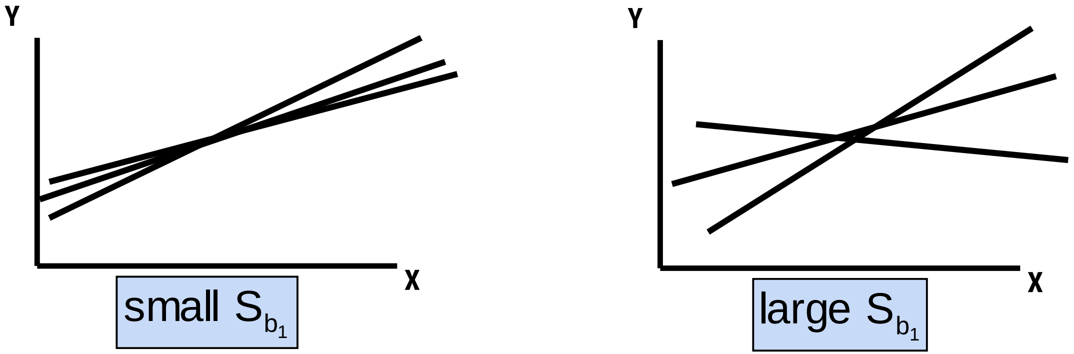
	df	SS	MS	F	Significance F
Regression	1	18934.9348	18934.9348	11.0848	0.01039
Residual	8	13665.5652	1708.1957		
Total	9	32600.5000			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	98.24833	58.03348	1.69296	0.12892	-35.57720	232.07386
Square Feet	0.10977	$\boxed{0.03297}$	3.32938	0.01039	0.03374	0.18580

Digital Vidya

Comparing Standard Errors of the Slope

S_{b_1} is a measure of the variation in the slope of regression lines from different possible samples



Inference about the Slope: t Test

- t test for a population slope
  - Is there a linear relationship between X and Y?
- Null and alternative hypotheses

H_0:	$\beta_1 = 0$	(no linear relationship)
H_1:	$\beta_1 \neq 0$	(linear relationship does exist)

- Test statistic

$$t = \frac{b_1 - \beta_1}{s_{b_1}}$$

$$d.f. = n - 2$$

where:

b_1 = regression slope coefficient

β_1 = hypothesized slope

s_{b1} = standard error of the slope

Digital Vidya

Inference about the Slope: t Test

(continued)

House Price in $1000s (y)	Square Feet (x)
245	1400
312	1600
279	1700
308	1875
199	1100
219	1550
405	2350
324	2450
319	1425
255	1700

Estimated Regression Equation:

$$\widehat{\text{house price}} = 98.25 + 0.1098\,(\text{sq.ft.})$$

The slope of this model is 0.1098

Does square footage of the house affect its sales price?

Digital Vidya

Inferences about the Slope: t Test Example

H_0: $\beta_1 = 0$

H_1: $\beta_1 <> 0$

From Excel output:

b_1

s_{b_1}

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	98.24833	58.03348	1.69296	0.12892
Square Feet	0.10977	0.03297	3.32938	0.01039

$$t = \frac{b_1 - \beta_1}{s_{b_1}} = \frac{0.10977 - 0}{0.03297} = 3.32938$$

Digital Vidya

Inferences about the Slope: t Test Example

(continued)

Test Statistic: t = 3.329

H_0: $\beta_1 = 0$

H_1: β_1 <> 0

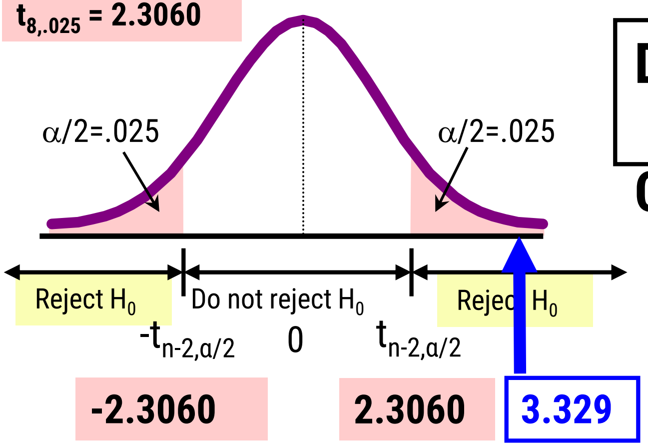
d.f. = 10-2 = 8

$t_{8,.025}$ **= 2.3060**

From Excel output:

b_1 | S_{b_1} | t

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	**98.24833**	**58.03348**	**1.69296**	**0.12892**
Square Feet	**0.10977**	**0.03297**	**3.32938**	**0.01039**



Decision:

Reject H_0

Conclusion:

There is sufficient evidence that square footage affects house price

Digital Vidya

Inferences about the Slope: t Test Example

(continued)

P-value = **0.01039**

H_0: $\beta_1 = 0$

H_1: $\beta_1 <> 0$

From Excel output:

P-value

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	**98.24833**	**58.03348**	**1.69296**	**0.12892**
Square Feet	**0.10977**	**0.03297**	**3.32938**	**0.01039**

This is a two-tail test, so the p-value is

$P(t > 3.329)+P(t < -3.329) = 0.01039$

(for 8 d.f.)

Decision: P-value < α so Reject H_0

Conclusion:

There is sufficient evidence that square footage affects house price

Digital Vidya

Confidence Interval Estimate for the Slope

Confidence Interval Estimate of the Slope:

$$b_1 - t_{n-2,\alpha/2} s_{b_1} < \beta_1 < b_1 + t_{n-2,\alpha/2} s_{b_1}$$

d.f. = n - 2

Excel Printout for House Prices:

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	98.24833	58.03348	1.69296	0.12892	-35.57720	232.07386
Square Feet	0.10977	0.03297	3.32938	0.01039	0.03374	0.18580

At 95% level of confidence, the confidence interval for the slope is (0.0337, 0.1858)

Confidence Interval Estimate for the Slope

(continued)

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	98.24833	58.03348	1.69296	0.12892	-35.57720	232.07386
Square Feet	0.10977	0.03297	3.32938	0.01039	0.03374	0.18580

Since the units of the house price variable is $1000s, we are 95% confident that the average impact on sales price is between $33.70 and $185.80 per square foot of house size

This 95% confidence interval does not include 0.

Conclusion: There is a significant relationship between house price and square feet at the .05 level of significance

F-Test for Significance

- F Test statistic: $$F = \frac{MSR}{MSE}$$

where $$MSR = \frac{SSR}{k}$$ $$MSE = \frac{SSE}{n-k-1}$$

where F follows an F distribution with k numerator and $(n - k - 1)$ denominator degrees of freedom

(k = the number of independent variables in the regression model)

Digital Vidya

Excel Output

Regression Statistics	
Multiple R	0.76211
R Square	0.58082
Adjusted R Square	0.52842
Standard Error	41.33032
Observations	10

$$F = \frac{MSR}{MSE} = \frac{18934.9348}{1708.1957} = 11.0848$$

With 1 and 8 degrees of freedom

P-value for the F-Test

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	18934.9348	18934.9348	11.0848	0.01039
Residual	8	13665.5652	1708.1957		
Total	9	32600.5000			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	98.24833	58.03348	1.69296	0.12892	-35.57720	232.07386
Square Feet	0.10977	0.03297	3.32938	0.01039	0.03374	0.18580

F-Test for Significance

(continued)

$H_0: \beta_1 = 0$

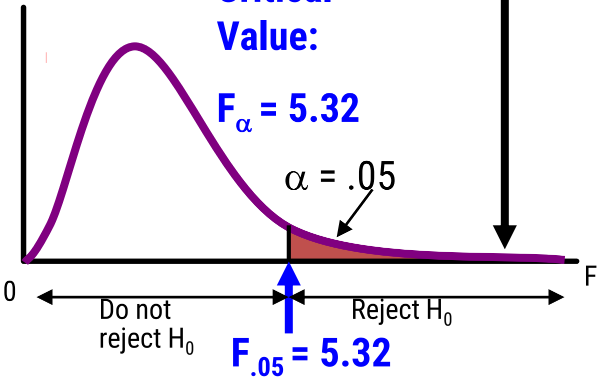
$H_1: \beta_1 \neq 0$

$\alpha = .05$

$\underbrace{df_1 = 1 \quad df_2 = 8}$

Critical Value:

$F_\alpha = 5.32$



Test Statistic:

$$F = \frac{MSR}{MSE} = 11.08$$

Decision:

Reject H_0 at $\alpha = 0.05$

Conclusion:

There is sufficient evidence that house size affects selling price

Digital Vidya

Prediction

- The regression equation can be used to predict a value for y, given a particular x

- For a specified value, x_{n+1} , the predicted value is

$$\boxed{\hat{y}_{n+1} = b_0 + b_1 x_{n+1}}$$

Digital Vidya

Predictions Using Regression Analysis

Predict the price for a house with 2000 square feet:

$$\widehat{\text{house price}} = 98.25 + 0.1098\,(\text{sq.ft.})$$

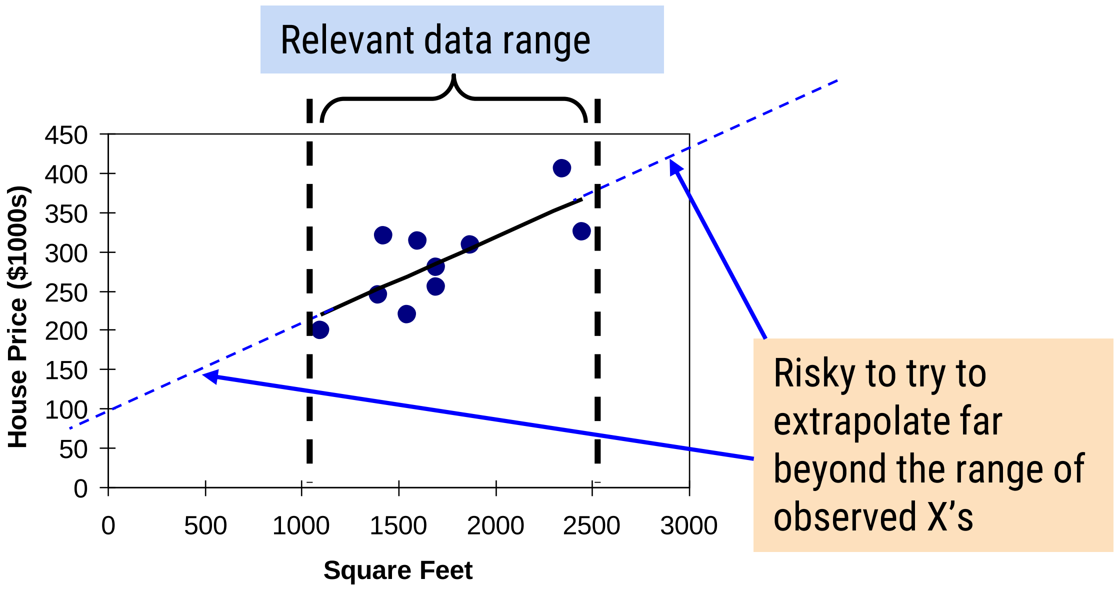
$$= 98.25 + 0.1098(2000)$$

$$= 317.85$$

The predicted price for a house with 2000 square feet is 317.85($1,000s) = $317,850

Relevant Data Range

- When using a regression model for prediction, only predict within the relevant range of data



Graphical Analysis

- The linear regression model is based on minimizing the sum of squared errors
- If outliers exist, their potentially large squared errors may have a strong influence on the fitted regression line
- Be sure to examine your data graphically for outliers and extreme points
- Decide, based on your model and logic, whether the extreme points should remain or be removed

Thank You