Regression Equation & Analysis

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

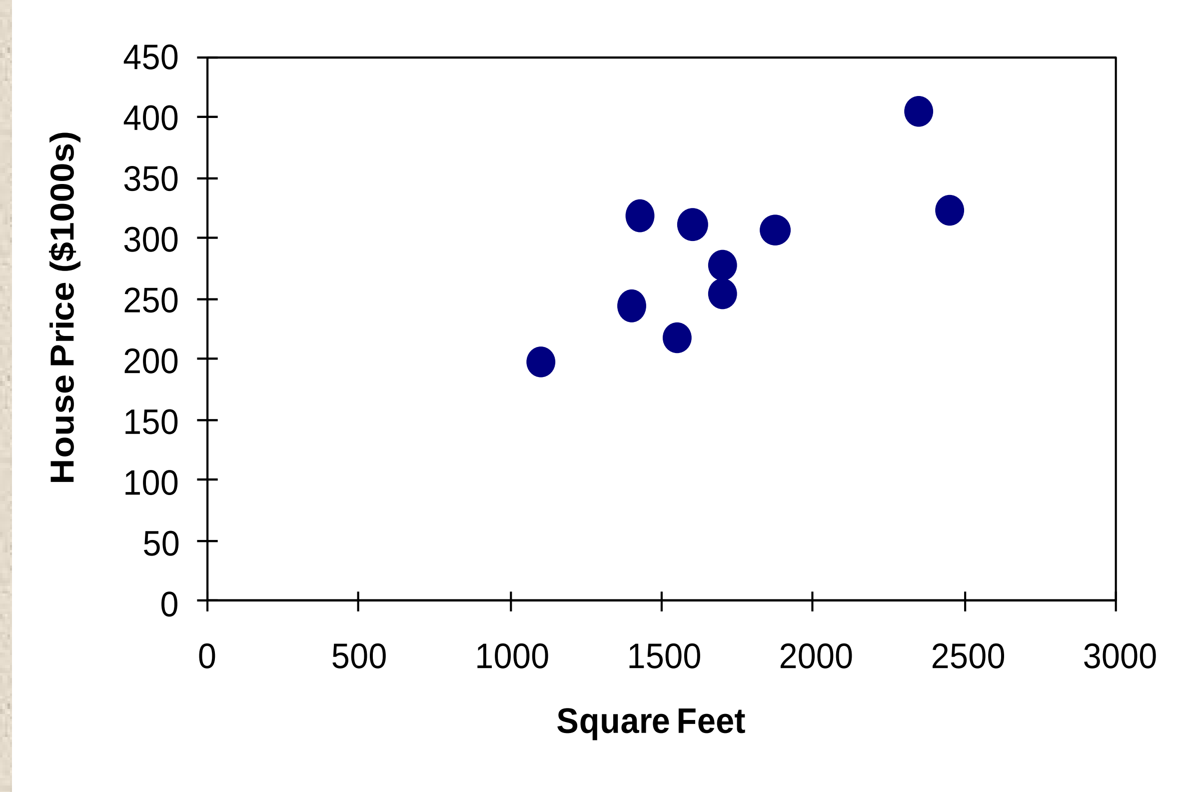
*Independent variable (X) = square feet

Simple Linear Regression Example: Data

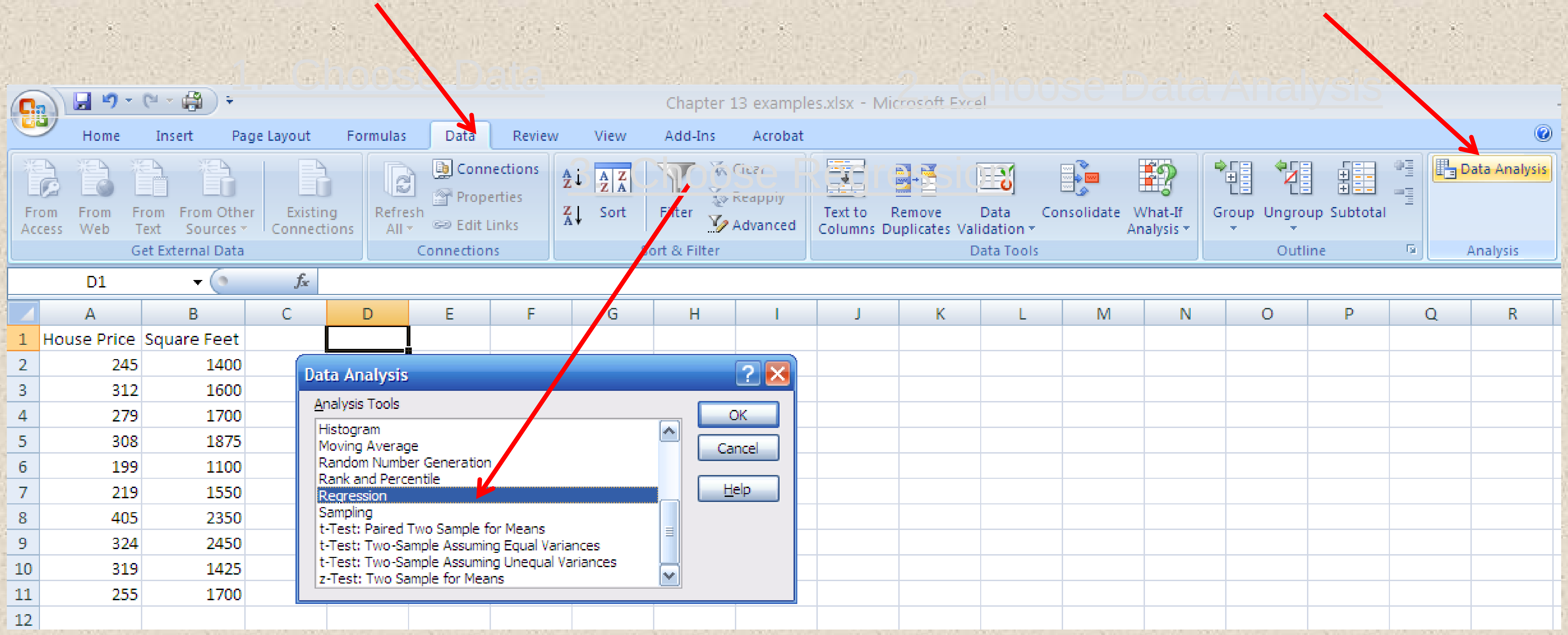
House Price in $1000s (Y)	Square Feet (X)
245	1400
312	1600
279	1700
308	1875
199	1100
219	1550
405	2350
324	2450
319	1425
255	1700

Simple Linear Regression Example: Scatter Plot

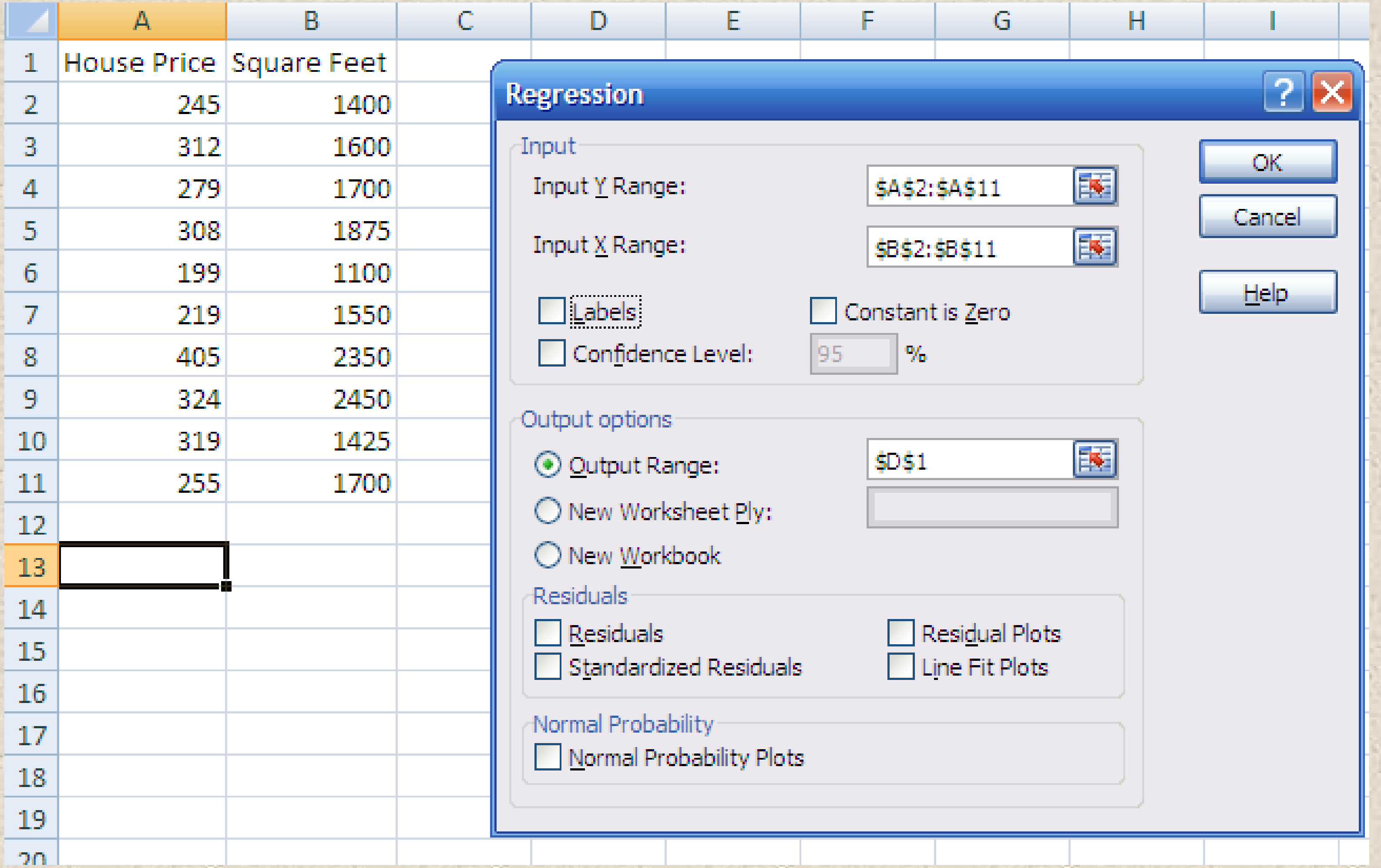
House price model: Scatter Plot



Simple Linear Regression Example: Using Excel Data Analysis Function



Simple Linear Regression Example: Using Excel Data Analysis Function



	A	B
1	House Price	Square Feet
2	245	1400
3	312	1600
4	279	1700
5	308	1875
6	199	1100
7	219	1550
8	405	2350
9	324	2450
10	319	1425
11	255	1700

Simple Linear Regression Example: 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

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

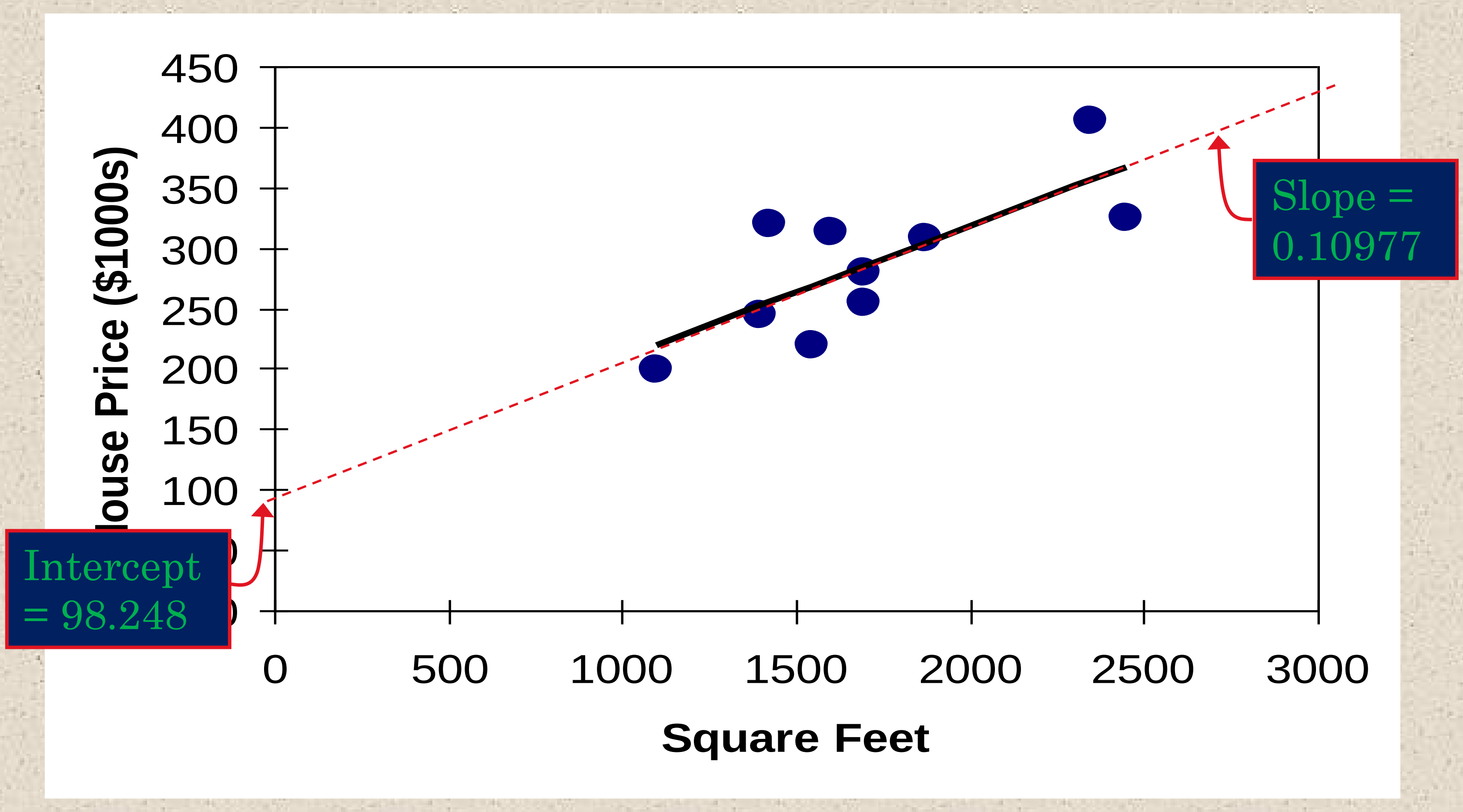
ANOVA

	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

Simple Linear Regression Example: Graphical Representation

House Price Model: Scatter Plot and Prediction Line



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

Simple Linear Regression Example: Excel Output

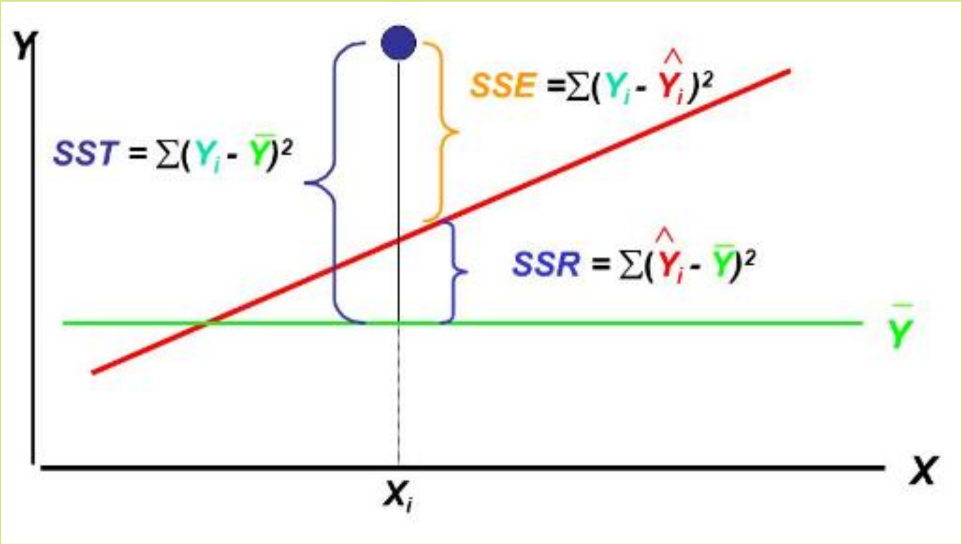
Regression Statistics	
Multiple R	0.76211
R Square	0.58082
Adjusted R Square	0.52842
Standard Error	41.33032
Observations	10

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

R-squared (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

$$R^2 = \frac{SSR}{SST}$$

$$R^2 = \frac{SST-SSE}{SST}$$

$$R^2 = 1 - \frac{SSE}{SST}$$

Best when: Zero Regression error

$R^2 = 1 - 0/SST = 1$

Simple Linear Regression Example: Interpretation of b_o

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

- b_0 is the estimated mean value of Y when the value of X is zero (if $X = 0$ is in the range of observed X values)
- Because a house cannot have a square footage of 0, b_0 has no practical application

Simple Linear Regression Example: Interpreting b_1

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

- b_1 estimates the change in the mean value of Y as a result of a one-unit increase in X
- Here, $b_1 = 0.10977$ tells us that the mean value of a house increases by .10977($1000) = $109.77, on average, for each additional one square foot of size

Simple Linear Regression Example: Making Predictions

Predict the price for a house with 2000 square feet:

$$\begin{aligned} \text{house price} &= 98.25 + 0.1098\,(\text{sq.ft.}) \\ &= 98.25 + 0.1098(2000) \\ &= 317.85 \end{aligned}$$

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

Assumptions of Regression L.I.N.E

Linearity

The relationship between X and Y is linear

Independence of Errors

Error values are statistically independent

Particularly important when data are collected over a period of time

Normality of Error

Error values are normally distributed for any given value of X

Equal Variance (also called homoscedasticity)

The probability distribution of the errors has constant variance

Inferences About the Slope: t Test Example

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:

house price = 98.25 + 0.1098 (sq. ft.)

The slope of this model is 0.1098

Is there a relationship between the square footage of the house and its sales price?

Inferences About the Slope: t Test Example

$H_0: \beta_1 = 0$

$H_1: \beta_1 \neq 0$

From Excel output:

	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

b_1

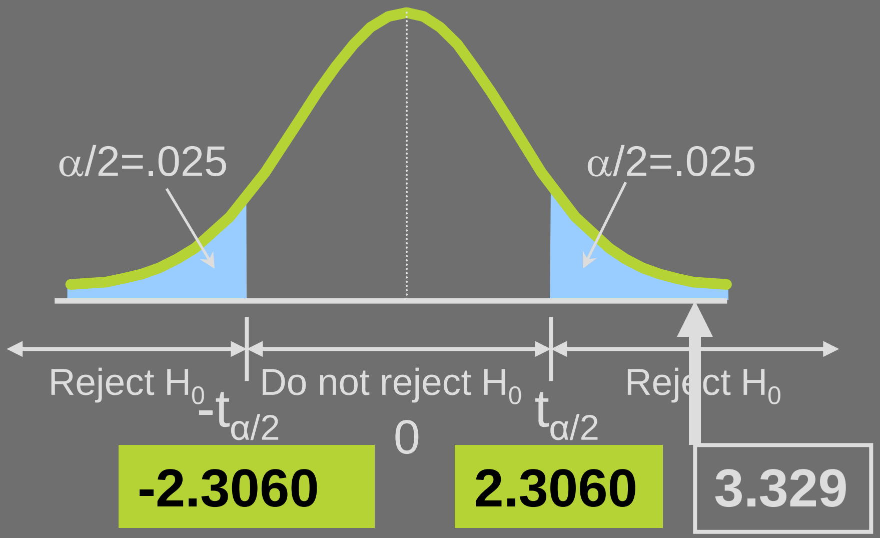
S_{b_1}

$$t_{STAT} = \frac{b_1 - \beta_1}{S_{b_1}} = \frac{0.10977 - 0}{0.03297} = 3.32938$$

Inferences About the Slope: t Test Example

Test Statistic: $t_{STAT} = 3.329$

$H_0: \beta_1 = 0$
$H_1: \beta_1 \neq 0$



Decision: Reject H_0

There is sufficient evidence that square footage affects house price

Inferences About the Slope: t Test Example

$$H_0: \beta_1 = 0$$
$$H_1: \beta_1 \neq 0$$

	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**

p-value

Decision: Reject H_0, since p-value < α

There is sufficient evidence that square footage affects house price.

Confidence Interval Estimate for the Slope

Confidence Interval Estimate of the Slope:

$$\boxed{b_1 \pm t_{\alpha/2} S_{b_1}}$$

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

	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 $33.74 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

Evaluation: Linear Regression

Evaluation metrics for linear regression

Metric	Space	Pros	Cons	When to Use
R^2	[0, 1]	Does not require comparison with other metrics to explain model fit.	Increases with the number of predictor variables, regardless of usefulness.	Simple linear regression
Adjusted R^2	[0, 1]	Adjusts the coefficient of determination for the number of predictors in the model.	The same as R^2 when there is only one predictor variable.	Multiple linear regression
MAE	≥ 0	Robust to outliers.	Does not penalise errors as extremely as other metrics.	When treating all errors equally
MSE	≥ 0	Maximises performance of linear regression.	Sensitive to outliers; magnifies large errors due to squaring.	When finding best fit models
RMSE	≥ 0	Maximises performance of linear regression.	Sensitive to outliers; magnifies large errors due to squaring.	When penalising large errors

Mean squared error	$\mathrm{MSE} = \frac{1}{n}\sum_{t=1}^{n} e_t^2$
Root mean squared error	$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{t=1}^{n} e_t^2}$
Mean absolute error	$\mathrm{MAE} = \frac{1}{n}\sum_{t=1}^{n} \lvert e_t \rvert$
Mean absolute percentage error	$\mathrm{MAPE} = \frac{100\%}{n}\sum_{t=1}^{n} \left\lvert \frac{e_t}{y_t} \right\rvert$

THANK YOU!