Simple Linear Regression Analysis: an example

Ex:

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

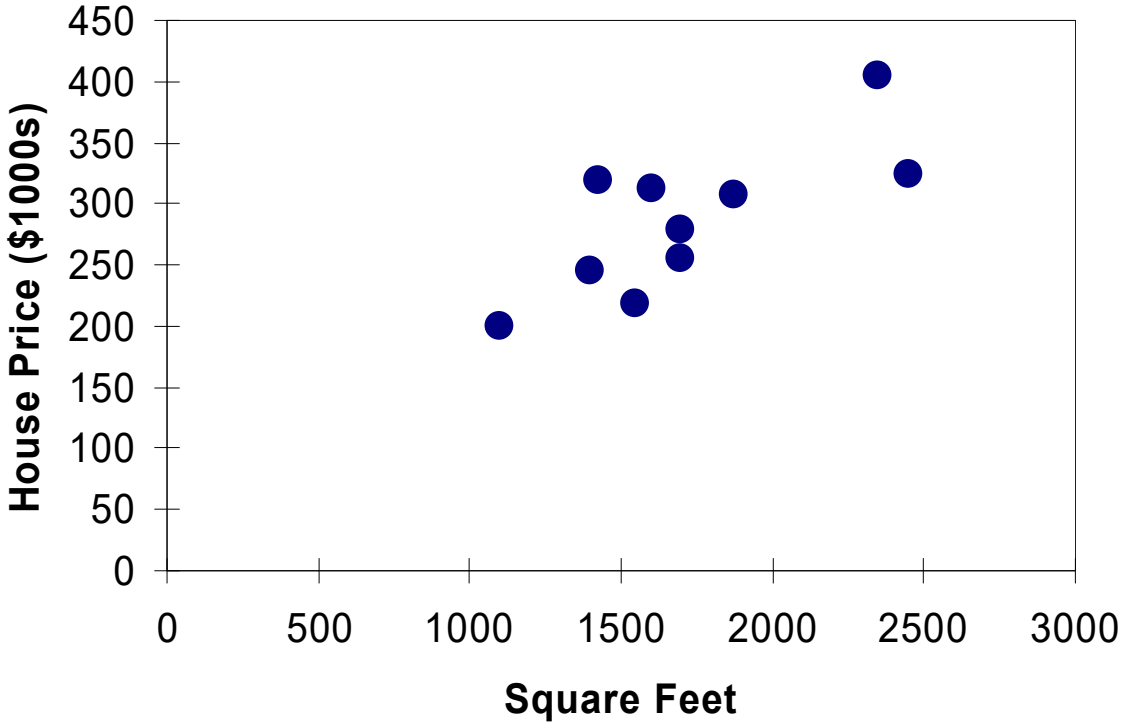
- A random sample of 10 houses is selected
  - Dependent variable (Y) = house price in $1000s
  - Independent variable (X) = square feet

Simple Linear Regression Analysis: an 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

Simple Linear Regression Analysis: an example

House price model: Scatter Plot



Simple Linear Regression Analysis: an example

	Y	X	$(Y-\bar{Y})$	$(X-\bar{X})$	$(Y-\bar{Y})^2$	$(X-\bar{X})^2$	$(X-\bar{X})(Y-\bar{Y})$
	245	1400	-41.5	-315	1722.25	99225	13072.5
	312	1600	25.5	-115	650.25	13225	-2932.5
	279	1700	-7.5	-15	56.25	225	112.5
	308	1875	21.5	160	462.25	25600	3440
	199	1100	-87.5	-615	7656.25	378225	53812.5
	219	1550	-67.5	-165	4556.25	27225	11137.5
	405	2350	118.5	635	14042.25	403225	75247.5
	324	2450	37.5	735	1406.25	540225	27562.5
	319	1425	32.5	-290	1056.25	84100	-9425
	255	1700	-31.5	-15	992.25	225	472.5
sum	2865	17150	0	0	32600.5	1571500	172500
mean	286.5	1715			3260.05	157150	17250

$$b_1 = \hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{172500}{1571500} = 0.109768$$

$$b_0 = \hat{\beta}_0 = \bar{y} - b_1\bar{x} = 286.5 - 0.109768 \cdot 1715 = 98.24833$$

Simple Linear Regression Analysis: an example

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:

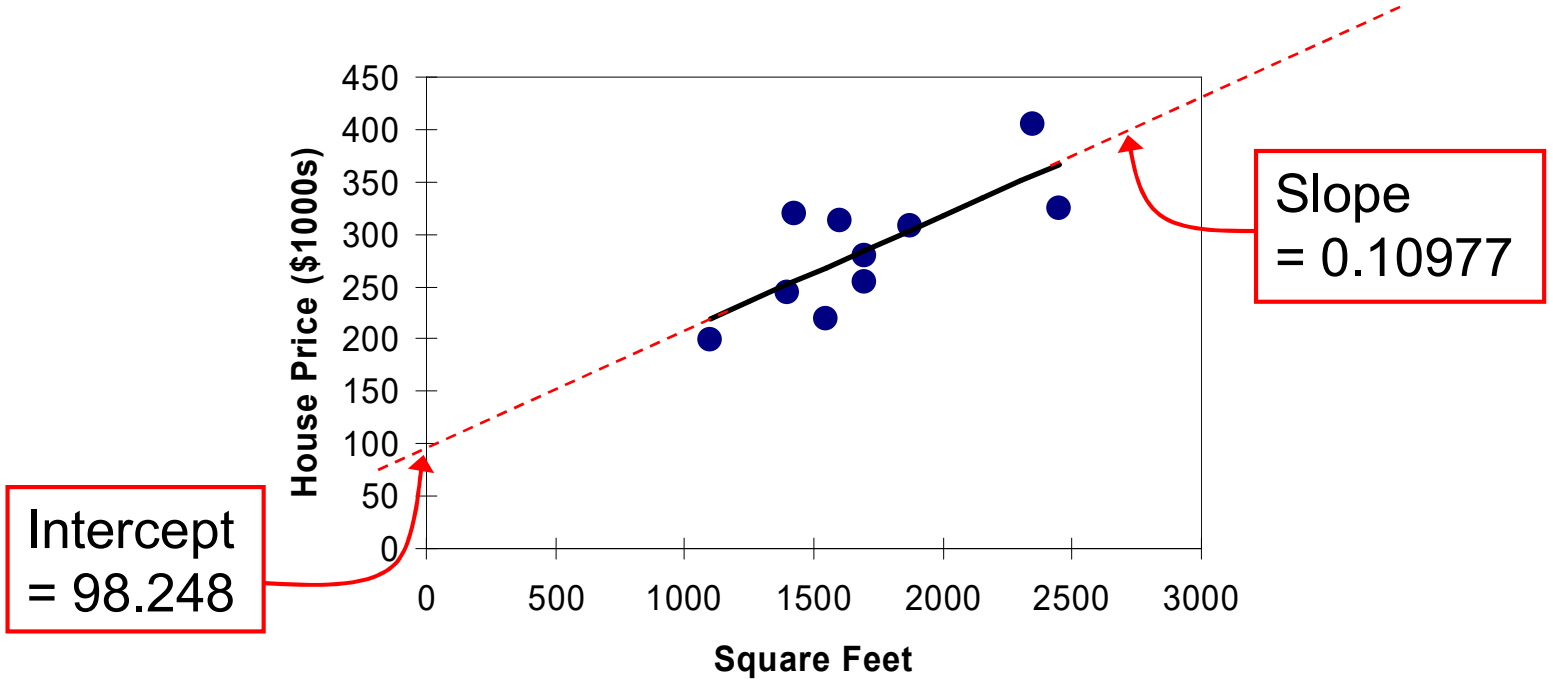
$$\widehat{\text{house price}} = 98.24833 + 0.10977\ (\text{square feet})$$

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

Simple Linear Regression Analysis: an example

House price model: Scatter Plot and Prediction Line



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

Simple Linear Regression Analysis: an example

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

Simple Linear Regression Analysis: the total variation

- Total variation is made up of two parts:

$$SST = SSR + SSE$$

Total Sum of Squares	Regression Sum of Squares	Error Sum of Squares
$SST = \sum (Y_i - \overline{Y})^2$	$SSR = \sum (\hat{Y}_i - \overline{Y})^2$	$SSE = \sum (Y_i - \hat{Y}_i)^2$

where:

$\overline{Y}$ = Mean value of the dependent variable

Y_i = Observed value of the dependent variable

$\hat{Y}_i$ = Predicted value of Y for the given X_i value

Simple Linear Regression Analysis: the coefficient of determination

- 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

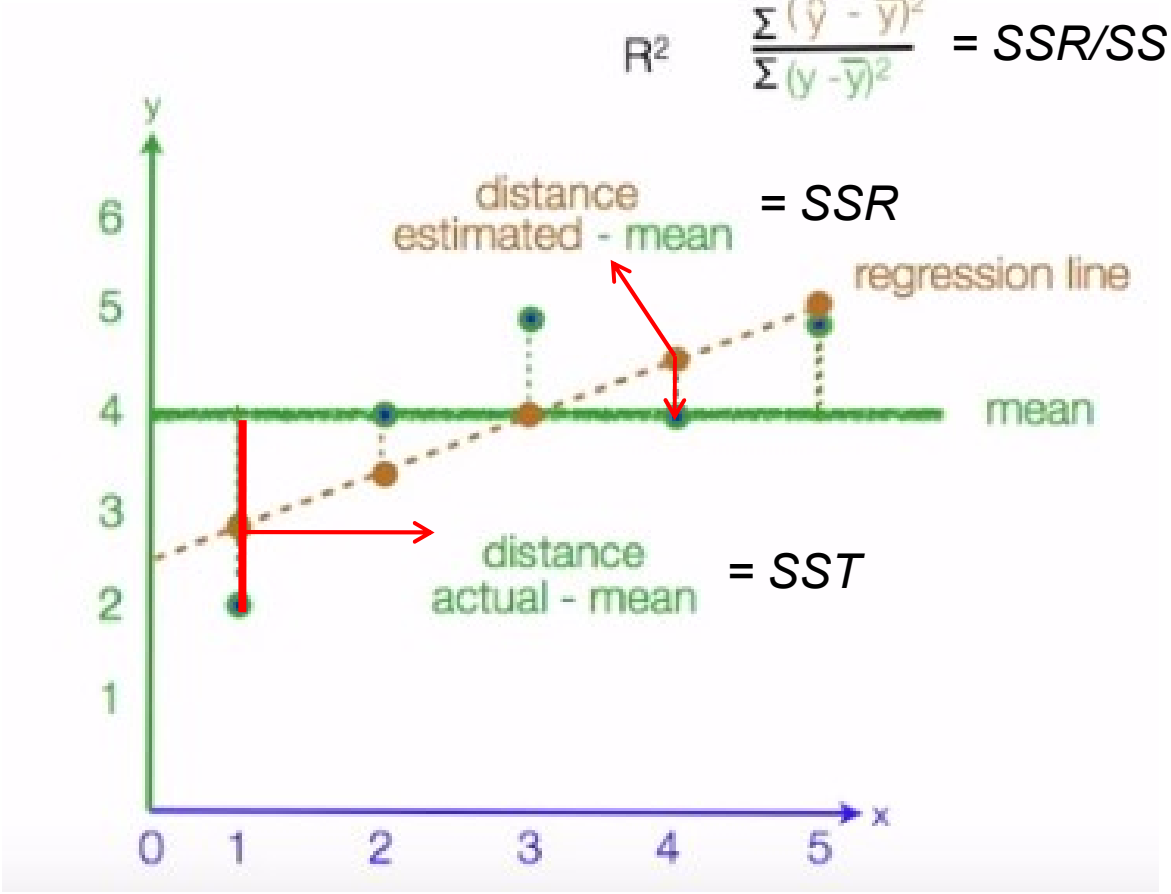
$$r^2 = \frac{SSR}{SST} = \frac{\text{regression } sum \text{ of squares}}{total \text{ sum of squares}}$$

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

Simple Linear Regression Analysis: the coefficient of determination

How well the regressed values estimated the real/actual values

$$R^2 \quad \frac{\Sigma (\hat{y} - \bar{y})^2}{\Sigma (y - \bar{y})^2} = SSR/SST$$



Simple Linear Regression Analysis: the coefficient of determination

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

Coefficient of Determination, R-squared

Definition

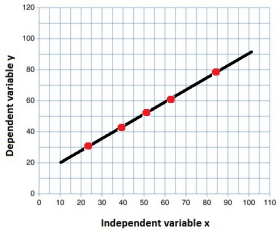
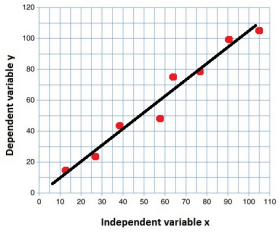
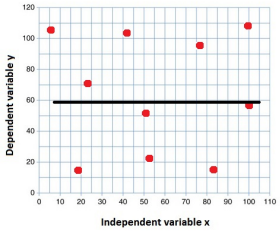
The *coefficient of determination*, or R^2, is a measure that provides information about the goodness of fit of a model. In the context of regression it is a statistical measure of how well the regression line approximates the actual data. It is therefore important when a statistical model is used either to predict future outcomes or in the testing of hypotheses. There are a number of variants (see comment below); the one presented here is widely used

$$\begin{aligned} R^2 &= 1 - \frac{\text{sum squared regression (SSR)}}{\text{total sum of squares (SST)}}, \\ &= 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}. \end{aligned}$$

The *sum squared regression* is the sum of the [residuals] squared, and the *total sum of squares* is the sum of the distance the data is away from the mean all squared. As it is a percentage it will take values between 0 and 1.

Interpretation of the R^2 value

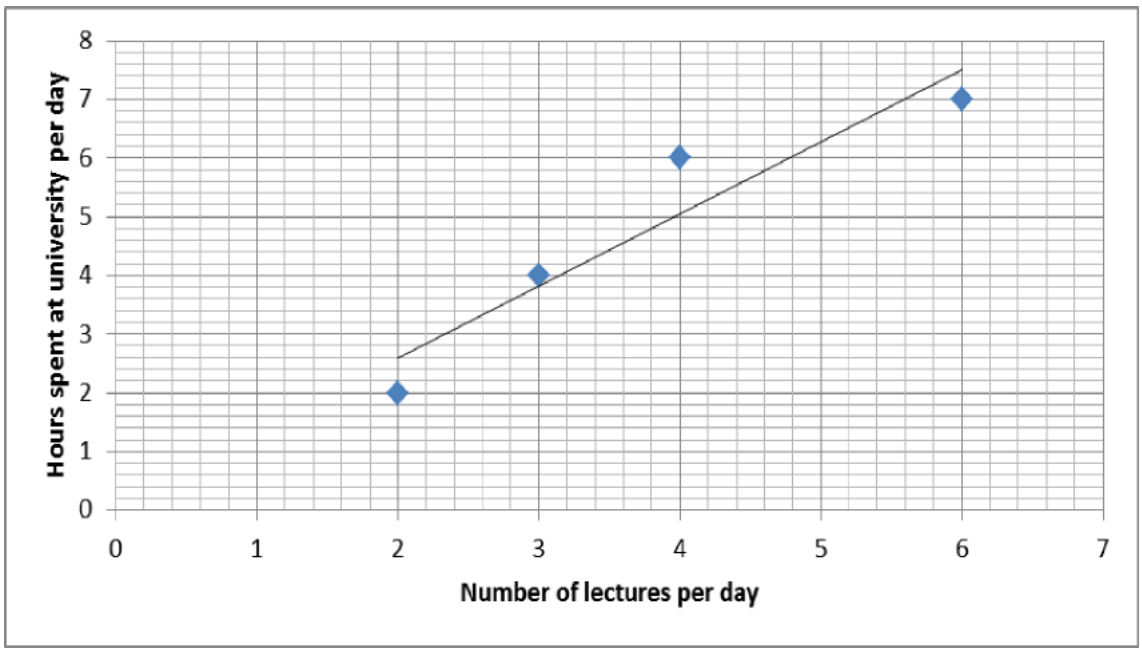
Here are a few examples of interpreting the R^2 value:

R^2 Values	Interpretation	Graph
$R^2 = 1$	All the variation in the y values is accounted for by the x values	
$R^2 = 0.83$	83% of the variation in the y values is accounted for by the x values	
$R^2 = 0$	None of the variation in the y values is accounted for by the x values	

Worked Example

Worked Example

Below is a graph showing how the number lectures per day affects the number of hours spent at university per day. The equation of the [regression line] is drawn on the graph and it has equation $\hat{y} = 0.143 + 1.229x$. Calculate R^2.



Solution

To calculate R^2 you need to find the sum of the residuals squared and the total sum of squares.

Start off by finding the residuals, which is the distance from regression line to each data point. Work out the predicted y value by plugging in the corresponding x value into the regression line equation.

- For the point $(2, 2)$

$$\begin{aligned}\hat{y} &= 0.143 + 1.229x \\ &= 0.143 + (1.229 \times 2) \\ &= 0.143 + 2.458 \\ &= 2.601\end{aligned}$$

The actual value for y is 2.

$$\begin{aligned}\text{Residual} &= \text{actual } y \text{ value} - \text{predicted } y \text{ value} \\ r_1 &= y_i - \hat{y}_i \\ &= 2 - 2.601 \\ &= -0.601\end{aligned}$$

As you can see from the graph the actual point is below the regression line, so it makes sense that the residual is negative.

- For the point $(3, 4)$

$$\begin{aligned}\hat{y} &= 0.143 + 1.229x \\ &= 0.143 + (1.229 \times 3) \\ &= 0.143 + 3.687 \\ &= 3.83\end{aligned}$$

The actual value for y is 4.

$$\begin{aligned}\text{Residual} &= \text{actual } y \text{ value} - \text{predicted } y \text{ value} \\ r_2 &= y_i - \hat{y}_i \\ &= 4 - 0.3.83 \\ &= 0.17\end{aligned}$$

As you can see from the graph the actual point is above the regression line, so it makes sense that the residual is positive.

- For the point $(4, 6)$

$$\begin{aligned}\hat{y} &= 0.143 + 1.229x \\ &= 0.143 + (1.229 \times 4) \\ &= 0.143 + 4.916 \\ &= 5.059\end{aligned}$$

The actual value for y is 6.

$$\begin{aligned}\text{Residual} &= \text{actual } y \text{ value} - \text{predicted } y \text{ value}\\ r_3 &= y_i - \hat{y}_i\\ &= 6 - 5.059\\ &= 0.941\end{aligned}$$

- For the point $(6, 7)$

$$\begin{aligned}\hat{y} &= 0.143 + 1.229x\\ &= 0.143 + (1.229 \times 6)\\ &= 0.143 + 7.374\\ &= 7.517\end{aligned}$$

The actual value for y is 7.

$$\begin{aligned}\text{Residual} &= \text{actual } y \text{ value} - \text{predicted } y \text{ value}\\ r_4 &= y_i - \hat{y}_i\\ &= 7 - 7.517\\ &= -0.517\end{aligned}$$

To find the residuals squared we need to square each of r_1 to r_4 and sum them.

$$\begin{aligned}\sum(y_i - \hat{y}_i)^2 &= \sum r_i\\ &= {r_1}^2 + {r_2}^2 + {r_3}^2 + {r_4}^2\\ &= (-0.601)^2 + (0.17)^2 + (0.941)^2 - (-0.517)^2\\ &= 1.542871\end{aligned}$$

To find $\sum(y_i - \bar{y})^2$ you first need to find the mean of the y values.

$$\begin{aligned}\bar{y} &= \frac{\sum y}{n}\\ &= \frac{2 + 4 + 6 + 7}{4}\\ &= \frac{19}{4}\\ &= 4.75\end{aligned}$$

Now we can calculate $\sum(y_i - \bar{y})^2$.

$$\begin{aligned}\sum(y_i - \bar{y})^2 &= (2 - 4.75)^2 + (4 - 4.75)^2 + (6 - 4.75)^2 + (7 - 4.75)^2\\ &= (-2.75)^2 + (-0.75)^2 + (1.25)^2 + (2.25)^2\\ &= 14.75\end{aligned}$$

Therefore;

$$\begin{aligned}R^2 &= 1 - \frac{\text{sum squared regression (SSR)}}{\text{total sum of squares (SST)}}\\ &= 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}\\ &= 1 - \frac{1.542871}{14.75}\\ &= 1 - 0.105 \text{ (3.s.f)}\\ &= 0.895 \text{ (3.s.f)}\end{aligned}$$

This means that the number of lectures per day account for 89.5% of the variation in the hours people spend at university per day.

An odd property of R^2 is that it is increasing with the number of variables. Thus, in the example above, if we added another variable measuring mean height of lecturers, R^2 would be no lower and may well, by chance, be greater - even though this is unlikely to be an improvement in the model. To account for this, an adjusted version of the coefficient of determination is sometimes used. For more information, please see [http://www.statstutor.ac.uk/resources/uploaded/correlation.pdf