Regression

Linear Regression

Example: The Mountain Lion population in Arizona is dependent on the Antelope Population in Arizona. We want to develop a model to help predict the Mountain Lion population in Arizona based on the Antelope Population. The following data gives the population (in hundreds) in a given year of Antelopes and the Mountain Lions.

Year	1	2	3	4	5	6	7
Antelope (x)	30	34	27	25	17	23	20
Mountain Lion (y)	66	79	70	60	48	55	60

We want to use this information to find a Linear Regression Line.

The regression line is of the form

$$\hat{y} = a + bx$$

The slope of the line is $b = \frac{n\sum xy - (\sum x)(\sum y)}{n\sum x^2 - (\sum x)^2}$ and the y coordinate of the y intercept

is $a = \overline{y} - b\overline{x}$. Remember that a bar over a random variable signifies the sample mean.

$$\overline{x} = \frac{\sum x}{n}$$
 and $\overline{y} = \frac{\sum y}{n}$.

To find the equation of the regression line, start by finding the slope, b. Making a table can be very helpful with this.

x	У	x^2	y^2	ху
30	66	900	4356	1980
34	79	1156	6241	2686
27	70	729	4900	1890
25	60	625	3600	1500
17	48	289	2304	816
23	55	529	3025	1265
20	60	400	3600	1200
$\sum x = 176$	$\sum y = 438$	$\sum x^2 = 4,628$	$\sum y^2 = 28,026$	$\sum xy = 11,337$

The table includes a column for each variable. X is the population size of the antelope and Y is the population size of the Mountain Lion. The next column is the square of the x values and then the square of the y values. Finally find the product of the x and y terms.

The last row in the table is the sum of each column. These values are plugged in to the formula to find the slope of the regression line.

$$b = \frac{n\sum xy - (\sum x)(\sum y)}{n\sum x^2 - (\sum x)^2}$$
$$b = \frac{7(11337) - (176)(438)}{7(4628) - (176)^2}$$
$$b = \frac{2271}{1420}$$
$$b = 1.599295775$$
$$b \approx 1.6$$

Next we need to calculate a, which is the y value of the y-intercept.

For this calculation we need to find the mean of the x values and the mean of the y values.

$$\overline{x} = \frac{\sum x}{n} = \frac{176}{7} = 25.14285714 \approx 25.14$$
$$\overline{y} = \frac{\sum y}{n} = \frac{438}{7} = 62.57142857 \approx 62.57$$

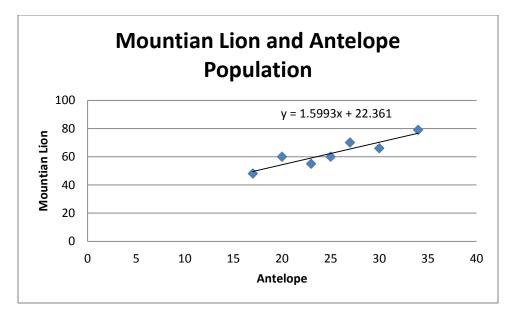
Plug in the means and the slope to get a.

$$a = \overline{y} - b\overline{x} \approx 62.57 - 1.6(25.14) \approx 22.35$$

Plug the values for b and a into the equation of the regression line.

$$\hat{y} = a + bx$$
$$\hat{y} = 22.35 + 1.6x$$

A good visual representation of the data is a scatter plot.



Plotting the regression line on the scatter plot is a good way of seeing how good the fit is. This regression line clearly fits the data.

The least square regression line can be used for prediction. If we know the Antelope population then we can predict the Mountain Lion Population.

The equation of the regression line is given by $\hat{y} = 22.35 + 1.6x$. If we know the Antelope population is 21 (hundred), then plugging this in for x in the regression equation will give an estimate of what the Mountain Lion population should be.

$$\hat{y} = 22.35 + 1.6x$$

 $\hat{y} = 22.35 + 1.6(21)$
 $\hat{y} = 22.35 + 33.6$
 $\hat{y} = 55.95$

Rounding to a whole number gives a prediction of 56 (hundred) for the size of the Mountain Lion population.

Multiple Regression

Multiple regression is an extension of linear regression for situations that involve two or more variables.

Example

The Mountain Lion population in Arizona is dependent on the Antelope Population in Arizona as well as the big horn sheep. We want to develop a model to help predict the Mountain Lion population in Arizona based on the Antelope Population and the big horn sheep. The following data gives the population (in hundreds) in a given year of the Mountain Lions, the Antelope, and the Big Horn Sheep.

Year	1	2	3	4	5	6	7
Antelope (x)	30	34	27	25	17	23	20
Mountain Lion (y)	66	79	70	60	48	55	60
Big Horn Sheep	25	32	54	43	12	25	22

The regression line for two variables is of the form

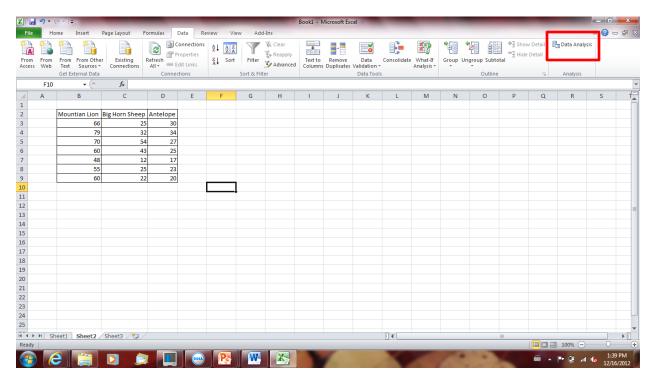
$$\hat{y} = a + b_1 x_1 + b_2 x_2$$

The calculations of a, b_1 , and b_2 are mathematically complicated; therefore they are generally done using a statistical software package. Microsoft Excel has a tool that will perform these calculations.

To use Microsoft Excel, input the data into the spreadsheet as follows.

e 5	+ (Home	Insert	Page Layout F							licrosoft Exc		-								
m Fro	om Fi	rom From Othe Text Sources •	er Existing Connections	Refresh	Connections Properties Edit Links	Ž↓ Ž Ž Ž↓ Sort	Filter	Clear	Text to	Remove	Data	Consolidate	What-If	Group Un	group Subto	Hid	w Detail e Detail	💾 Data Analysis	1	
-55 VV		et External Data	Connections	Conne	ctions		Sort & Fil	ter	Columns	Dupilcates	Data Tools		Analysis *				G.	Analysis		
F	10	- (e	f_x																	_
А		В	С	D	E	F	G	н	1	J	К	L	м	N	0	Р	Q	R	S	
	M	Iountian Lion	Big Horn Sheep																	
		66																		
		79																		
		70	-																	
	_	60 48	43																	
		40																		
		60																		
							1													
							•													
	Shee	t1 Sheet2	Sheet3 / 💱 /	/									1							
1																		Ⅲ 100% —		

Once the data is in the spreadsheet use the Data Analysis toolpak by clicking on the data tab at the top of the screen. The Data Analysis button is displayed on the right side of the screen.



Click on the Data Analysis button and a control box will appear.

File	¶ ▼ (≃ Hom		Page Layout F	ormulas	Data Re	eview Vi	ew Add	-Ins	Book1 - M	icrosoft Ex	cel			-					A <	
		From From Other Text Sources	r Existing Connections	Refresh	Connections Properties Edit Links	A↓ AZA Z↓ Sort	C104-0-0	😵 Clear 📡 Reapply Advanced	Text to Columns	Remove Duplicates	Data Validation •	Consolidate	What-If Analysis *		group Subt	Hid		💾 Data Anal	ysis	
		Get External Data		Conne	ctions		Sort & Filt	er			Data Tool	5			Outline	<u>.</u>	G.	Analysis		
	F10	• (*	f_{x}								1	1	1							_
	A	В	С	D	E	F	G	H	1	J	К	L	M	N	0	Р	Q	R	S	
		Acuption Lion	Big Horn Sheep	Antolono																
	n	66	Big Horn Sneep 25																	
		79	32																	
		70	54									(D					8	×		
		60	43									Data Anal					(Co Co			
		48	12	2 17								Analysis					ОК			
		55	25	5 23								Histogra Moving	am Average			^	Cancel			
		60	22	2 20								Random	Average Number Ger	neration			concer			
												Regress					Help			
												Sampling	g Paired Two S	ample for Mea	ans.	_				
												t-Test: 1	Two-Sample	Assuming Equ	al Variances	-				
												z-Test:	Two-Sample Two Sample	Assuming Une for Means	qual Variance	s •				
																	_	-		
	Shee	et1 Sheet2	Sheet3 / 💱 /	/								[] ◀								
• •																		100% (-)		
) V																		□ 100‰ (=	V	

Scroll down to Regression and click OK.

ie	Hom	i v v ne Insert P	age Layout F	ormulas	Data Re	view Vie	w Add		Book1 - M	icrosoft Exc	el	_			_	~ () =	- 6
	Veb	From From Other Text Sources ~ Get External Data		Refresh	Connections Properties Edit Links	$\begin{array}{c} \mathbb{A} \downarrow \\ \mathbb{Z} \downarrow \\ \mathbb{Z} \\ \mathbb{A} \downarrow \\ \mathbb{A} \downarrow \\ \end{array}$ Sort		K Clear Reapply Advanced	Text to Columns	Remove Duplicates	Data Validation Data Too	Consolidate What-If Analysis ~	Group Ungroup Subto	Hide Detail Hide	il 💾 Data Anal	ysis	
	F10		f _x	conne			3011 001110				Data 100		Oddime		Analysis		_
А		В	С	D	E	F	G	н	1	J	К	L M	N O	P (S	
												Regression		8	×		
	P	Mountian Lion	Big Horn Sheep	Antelope								Input					
		66	25	5 30								Input Y Range:	\$8\$2:\$8\$9		ОК		
		79	32												Cancel		
		70	54								_	Input <u>X</u> Range:	\$C\$2:\$D\$9				
		60	43								_	Labels	Constant is Zero		Help		
		48	12								_	Confidence Level:	95 %				
		55	25								_	Output options					
	L	60	22	2 20							- 1	Output Bange:	\$F\$10	1			
											-						
											- 1	New Worksheet Ph	/:				
											- 1	New Workbook Residuals					
											- 1	Residuals	Residual Pla	ots			
											- 1	Standardized Resid					
											- 1	Normal Probability					
											- 1	Normal Probability	Plots				
▶ ₩	She	et1 Sheet2	Sheet3 🦯 🞾 /														•
r									-	-		_			I III 100% (-		
44			0			R	W	X						~	- 🖻 🗑 a	1:45	PM

The next box will ask for the Y (dependent variable) and X (independent variables). Make sure if you include labels in the range entered in the Y and X data that you check the box for labels. Then click output range and select a cell for the results to be saved to. Finally click OK.

		[" ▲] ➡	-		-			-	-		ook1 - Microsoft E	kcel	_	-	-	-		- 0	-
File	Ho	ome Insert	Pag	ge Layout	Formul	las	Data R	eview View	/ Add	I-Ins								△ 🕜 ⊏	- ē
A	From	From From O	ther	Existing	Refres	sh area	Connections Properties	Ž↓ AZ Z↓ Sort	Filter	Reapply	Text to Remove	Data C	onsolidate What			 ●	💾 Data Analysis		
cess	Web	Text Source		Connections	All	69	Edit Links	A+ Son	T incer	Mathematical Advanced	Text to Remove Columns Duplicates	s Validation *	Analy	sis *	*				
		Get External Da	ita			Conne	ctions		Sort & Filt	er		Data Tools			Outline	5	Analysis		
	D12	• (*		f_{x}															
1	А	В		С		D	E	F		G	н	1	J	К	L	M	N	0	
			79		32	34													
			70		54	27													
			60	4	43	25													
			48		12	17													
			55		25	23													
			60		22	20													
0								SUMMARY (DUTPUT										
1																			
2									ession St										
3								Multiple R		0.925813263									
4								R Square		0.857130199									
5								Adjusted R		0.785695298									
5								Standard En		4.704743797									
7								Observation	15	7									
8																			
9								ANOVA											
D										df	SS	MS	F	Significance F					
1								Regression		2			11.99875959	0.02041178					
2								Residual		4	88.5384568	22.1346142							
3								Total		6	619.7142857								
4																			
5 6										Coefficients	Standard Error	t Stat	P-value	Lower 95%		Lower 95.0%			
								Intercept		22.20330554	8.496150485		0.059212047			-1.385789889			
7 8								Big Horn Sh	eep	0.11606348	0.1555/9061 0.376136969		0.497116816		0.548020203	-0.315893243 0.420763512			
	N CH	eet1 Sheet	Ch	100t2 /\$1		_		Antelope		1.405087159	0.376136969		0.017614516	0.420763512	2.509410806	0.420703512	2.509410806		•
eady	-/ <u>5</u>	Succession and an and an	د (١١								_						四 100% —		
				-							-			-					1 PM
5-							••••		W						-	<u></u>	🔺 🏲 📴 anto	12/10	

The summary output gives many useful pieces of information.

a = 22.2033055381857 $b_1 = 0.116063480133195$ $b_2 = 1.46508715888824$

The multiple least squares regression equation is

$$\hat{y} = 22.203 + 0.116x_1 + 1.465x_2$$

The multiple least square regression line can be used for prediction. If we know the Antelope population and the Big Horn Sheep population then we can predict the Mountain Lion Population.

The equation of the regression line is given by $\hat{y} = 22.203 + 0.116x_1 + 1.465x_2$. If we know the Antelope population is 21(hundred) and the Big Horn Sheep population is 32 (Hundred), then plugging this in for x_1 and x_2 respectively in the regression equation will give an estimate of what the Mountain Lion population should be.

$$\hat{y} = 22.203 + 0.116x_1 + 1.465x_2$$
$$\hat{y} = 22.203 + 0.116(21) + 1.465(32)$$
$$\hat{y} = 22.203 + 2.436 + 46.88$$
$$\hat{y} = 71.519$$

Rounding to a whole number gives a prediction of 52 (hundred) for the size of the Mountain Lion population.