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 $b = \dfrac{n\sum xy - \left(\sum x\right)\left(\sum y\right)}{n\sum x^2 - \left(\sum x\right)^2}$ and the y coordinate of the y intercept is $a = \bar{y} - b\bar{x}$. Remember that a bar over a random variable signifies the sample mean. $\bar{x} = \dfrac{\sum x}{n}$ and $\bar{y} = \dfrac{\sum y}{n}$.

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

x	y	x^2	y^2	xy
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.

$$b=\frac{n\sum xy-\left(\sum x\right)\left(\sum y\right)}{n\sum x^2-\left(\sum x\right)^2}$$

$$b=\frac{7(11337)-(176)(438)}{7(4628)-(176)^2}$$

$$b=\frac{2271}{1420}$$

$$b=1.599295775$$

$$b\approx 1.6$$

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.

$$\bar{x}=\frac{\sum x}{n}=\frac{176}{7}=25.14285714\approx 25.14$$

$$\bar{y}=\frac{\sum y}{n}=\frac{438}{7}=62.57142857\approx\sim 62.57$$

Plug in the means and the slope to get a.

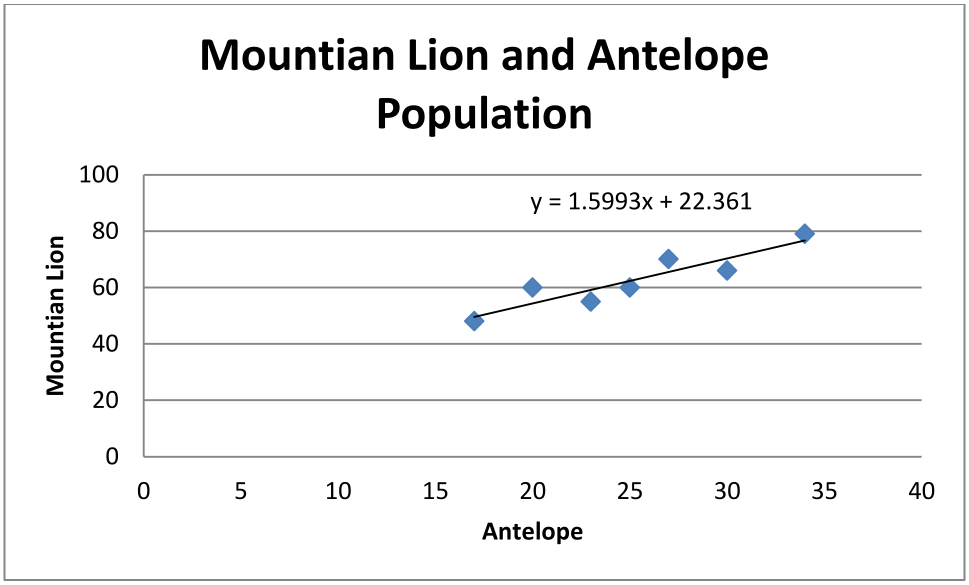
$$a=\bar{y}-b\bar{x}\approx 62.57-1.6(25.14)\approx 22.35$$

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.

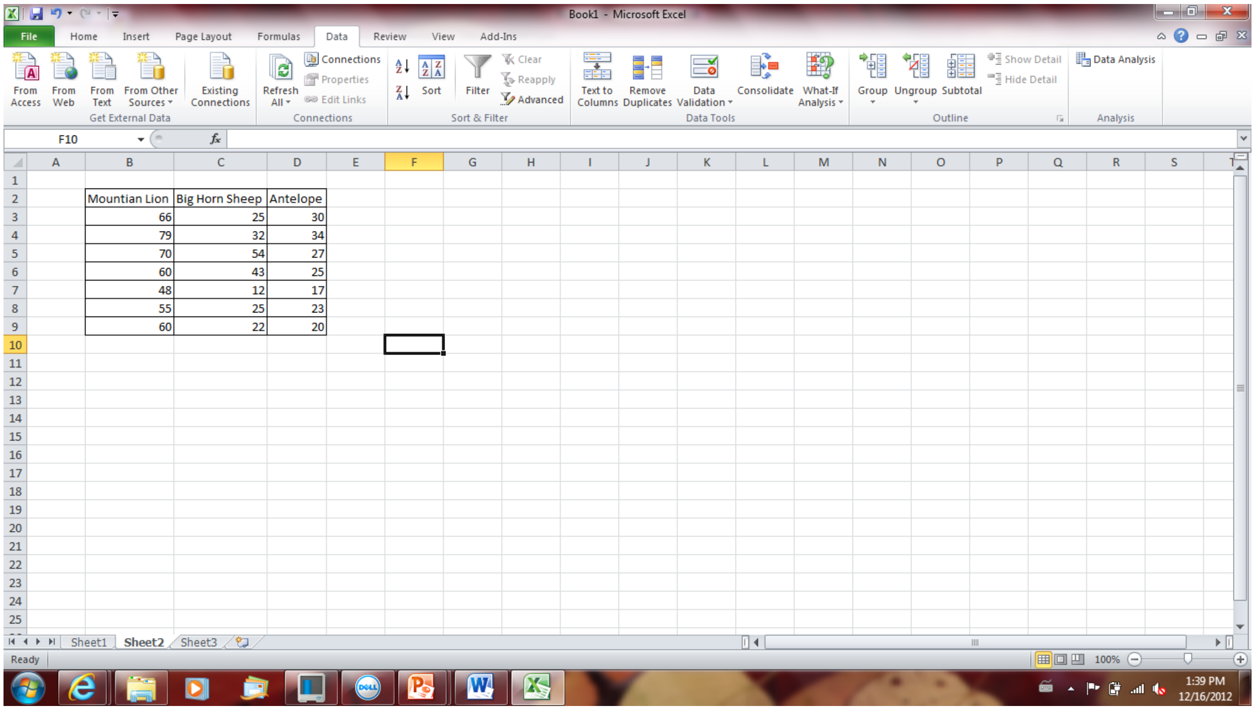
Year	**1**	**2**	**3**	**4**	**5**	**6**	**7**
Antelope (x)	30	34	27	25	17	23	20
Mountain Lion (y)	66	79	70	60	48	55	60
Big Horn Sheep	25	32	54	43	12	25	22

The regression line for two variables is of the form

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

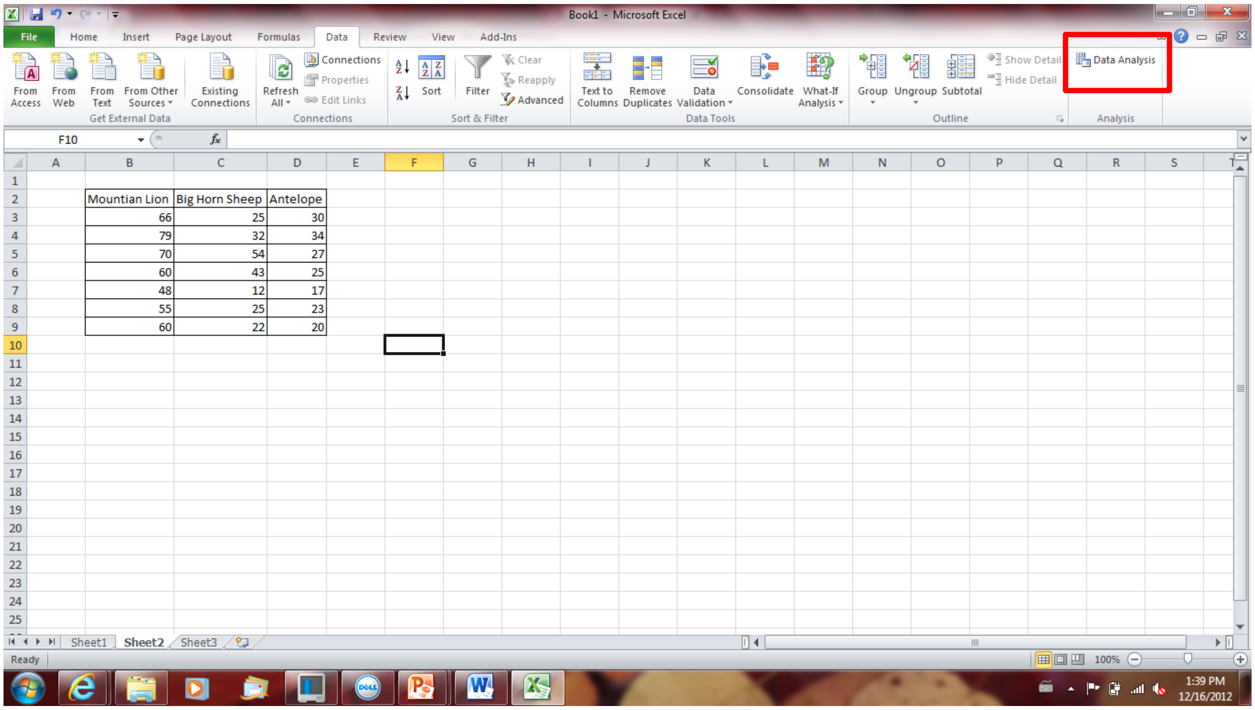
The calculations of $a, b_1,$ and b_2 are mathematically complicated; therefore they are generally done using a statistical software package. Microsoft Excel has a tool that will perform these calculations.

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



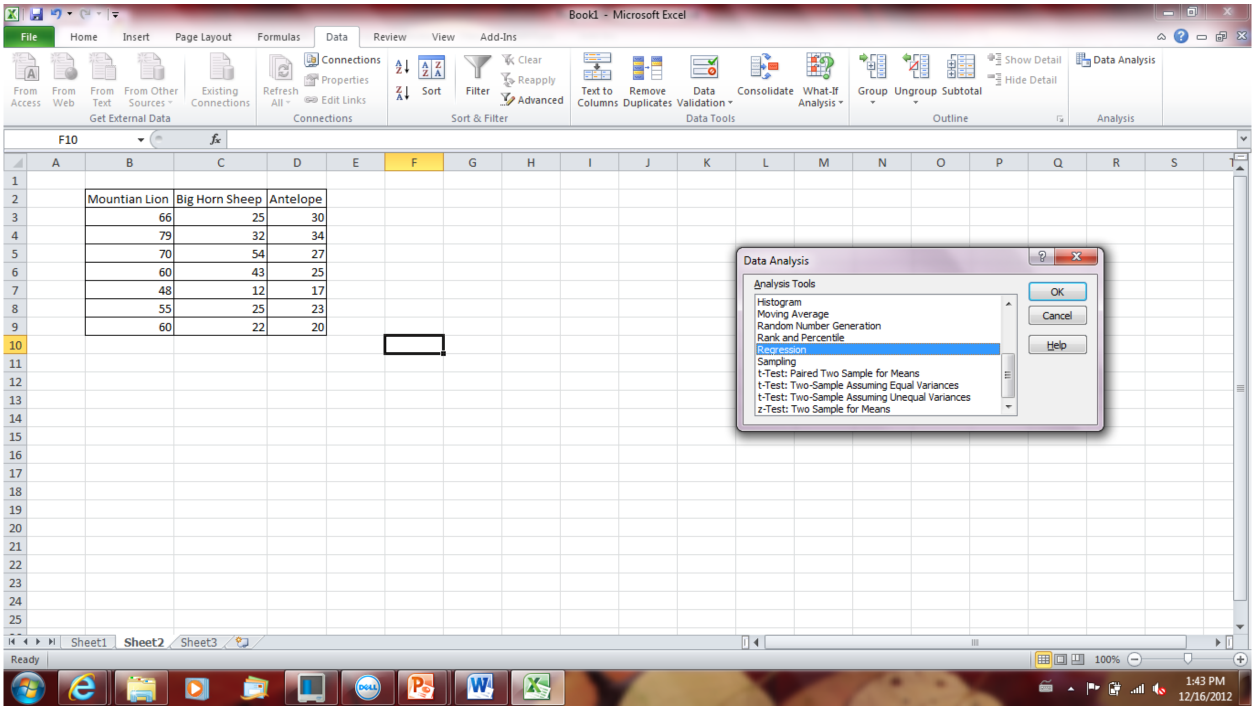
Mountian Lion	Big Horn Sheep	Antelope
66	25	30
79	32	34
70	54	27
60	43	25
48	12	17
55	25	23
60	22	20

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.



Mountian Lion	Big Horn Sheep	Antelope
66	25	30
79	32	34
70	54	27
60	43	25
48	12	17
55	25	23
60	22	20

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



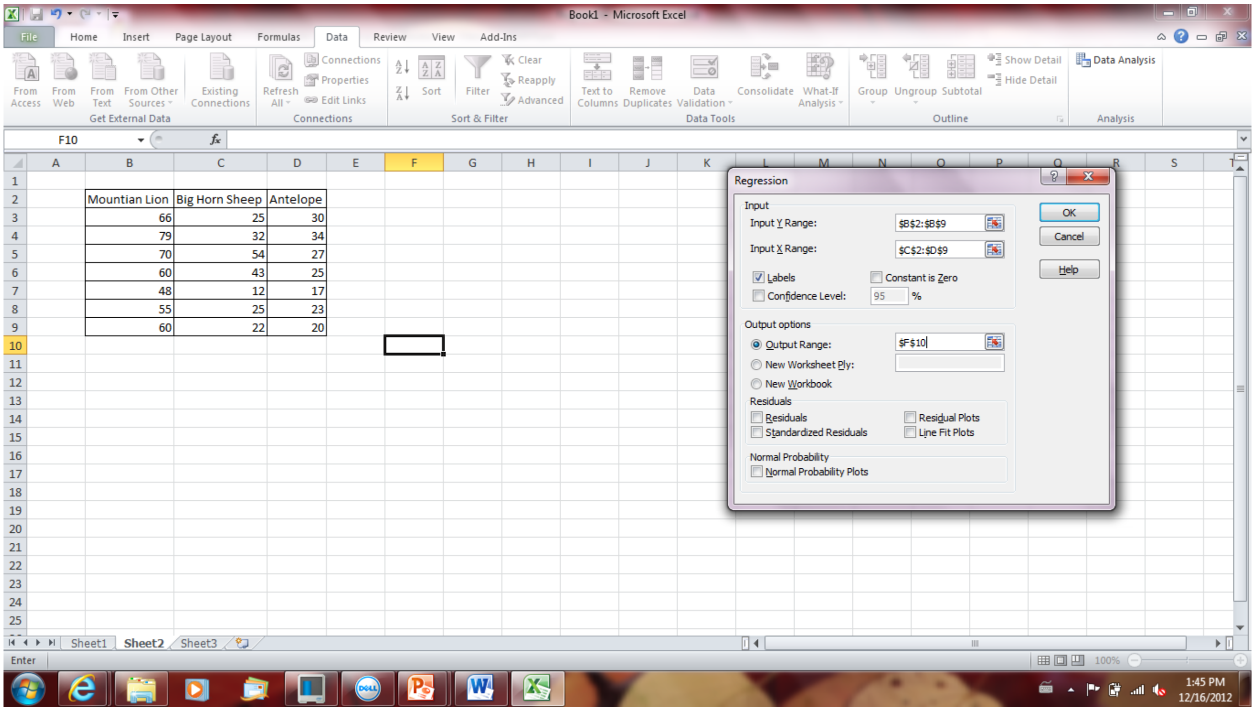
Mountian Lion	Big Horn Sheep	Antelope
66	25	30
79	32	34
70	54	27
60	43	25
48	12	17
55	25	23
60	22	20

Data Analysis

Analysis Tools

- Histogram
- Moving Average
- Random Number Generation
- Rank and Percentile
- Regression
- Sampling
- t-Test: Paired Two Sample for Means
- t-Test: Two-Sample Assuming Equal Variances
- t-Test: Two-Sample Assuming Unequal Variances
- z-Test: Two Sample for Means

Scroll down to Regression and click OK.



Mountian Lion	Big Horn Sheep	Antelope
66	25	30
79	32	34
70	54	27
60	43	25
48	12	17
55	25	23
60	22	20

Regression

Input

Input Y Range: $B$2:$B$9

Input X Range: $C$2:$D$9

☑ Labels ☐ Constant is Zero

☐ Confidence Level: 95 %

Output options

● Output Range: $F$10

○ New Worksheet Ply:

○ New Workbook

Residuals

☐ Residuals ☐ Residual Plots

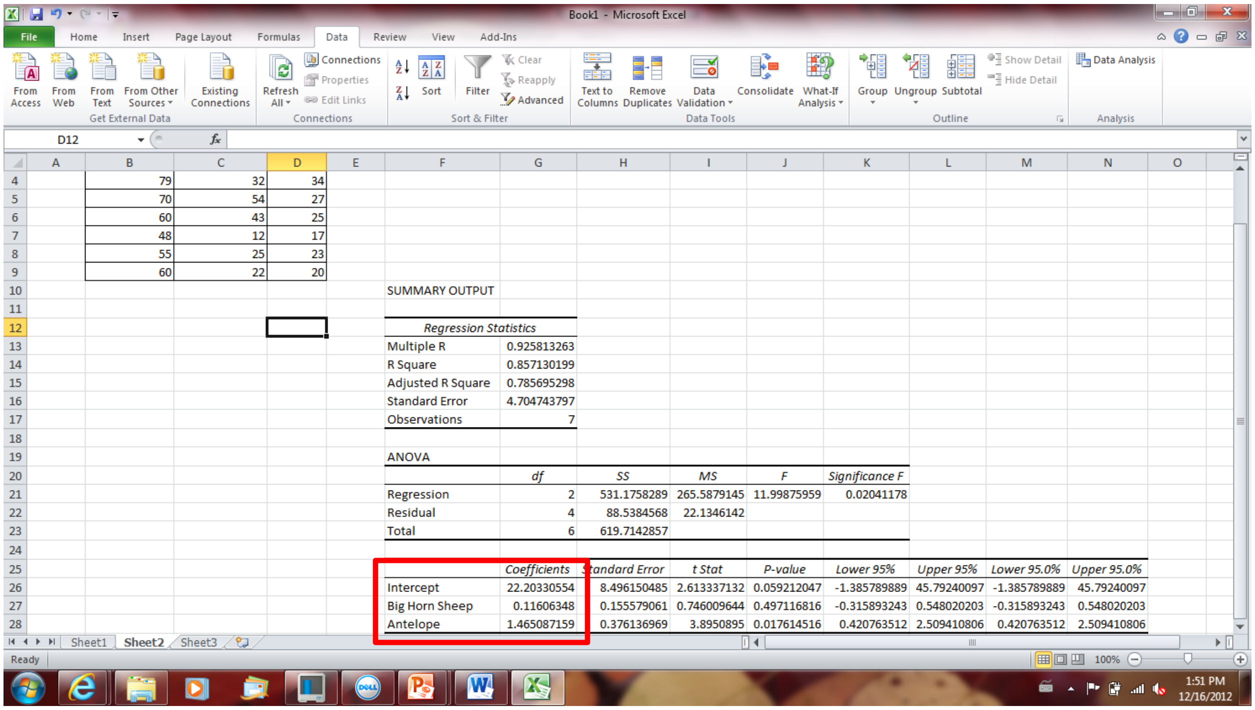
☐ Standardized Residuals ☐ Line Fit Plots

Normal Probability

☐ Normal Probability Plots

OK Cancel Help

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.



79	32	34
70	54	27
60	43	25
48	12	17
55	25	23
60	22	20

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.925813263
R Square	0.857130199
Adjusted R Square	0.785695298
Standard Error	4.704743797
Observations	7

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	531.1758289	265.5879145	11.99875959	0.02041178
Residual	4	88.5384568	22.1346142		
Total	6	619.7142857			

	Coefficients	*tandard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	22.20330554	8.496150485	2.613337132	0.059212047	-1.385789889	45.79240097	-1.385789889	45.79240097
Big Horn Sheep	0.11606348	0.155579061	0.746009644	0.497116816	-0.315893243	0.548020203	-0.315893243	0.548020203
Antelope	1.465087159	0.376136969	3.8950895	0.017614516	0.420763512	2.509410806	0.420763512	2.509410806

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 $x_1 \text{ and } x_2$ 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.