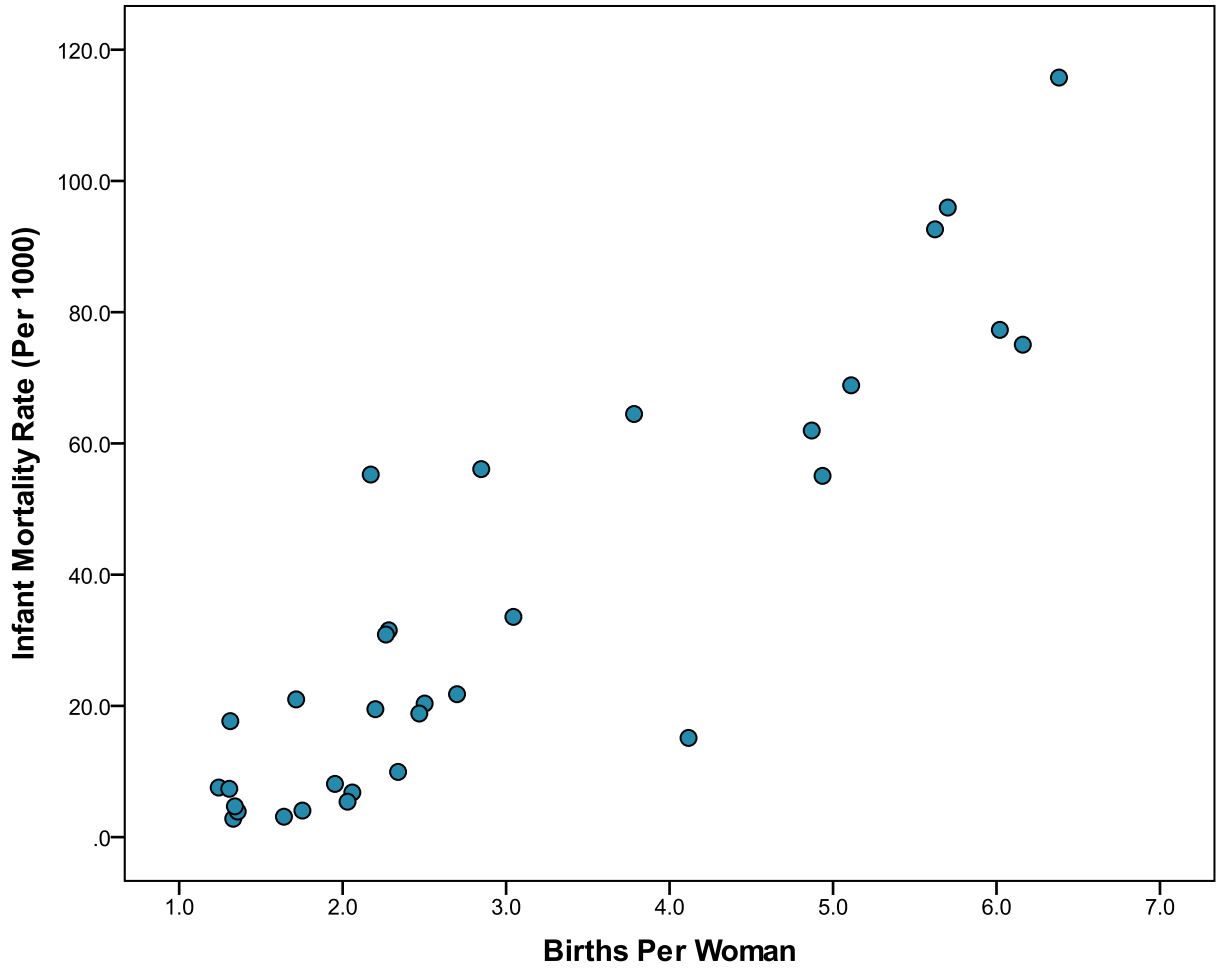
Regression Analysis

World Bank Data: 2013

Country	*Infant Mortality (per 1000)*	*Births per Woman*	*Country*	*Infant Mortality (per 1000)*	*Births per Woman*
El Salvador	20.4	2.5	Malawi	77.3	6
Haiti	64.5	3.8	Nigeria	95.9	5.7
Japan	2.8	1.3	Zambia	75.0	6.2
Bosnia	7.6	1.2	Peru	21.8	2.7
Hungary	7.4	1.3	Chile	8.1	2
Romania	17.7	1.3	India	56.1	2.8
Kuwait	9.9	2.3	Indonesia	31.5	2.3
Turkey	19.5	2.2	Myanmar	55.3	2.2
Eritrea	55.1	4.9	Kazakhstan	30.9	2.3
Sudan	62.0	4.9	Armenia	21.0	1.7
United States	6.8	2.1	Belgium	4.0	1.8
New Zealand	5.4	2	Germany	3.9	1.4
Botswana	33.6	3	Luxembourg	3.1	1.6
Congo, Dem. Rep.	115.8	6.4	Spain	4.7	1.3
Ethiopia	68.9	5.1	Tonga	15.1	4.1
Guinea	92.6	5.6	Jamaica	18.8	2.5


Infant Mortality Rate (Per 1000)
120.0
100.0
80.0
60.0
40.0
20.0
.0
1.0
2.0
3.0
4.0
5.0
6.0
7.0
Births Per Woman

There are 2 variables:

- Infant mortality rate (per 1000)
- Births per woman

Which is the independent and which is dependent?

Dependent = Infant mortality rate (per 1000)
Independent = Births per woman

or

Independent = Infant mortality rate (per 1000)
Dependent = Births per woman

It depends on how you frame your research question.

If your hypothesis is how infant mortality influences the number of births per woman, then:

Independent = Infant mortality rate (per 1000)
Dependent = Births per woman

The key here is that your research is attempting to determine whether increased infant mortality is forcing families to have more children.

Conversely, framing your research question differently results in different variable assignments.

If your hypothesis is how the number of births per woman influences infant mortality, then:

Dependent = Infant mortality rate (per 1000)
Independent = Births per woman

The key here is that your research is attempting to determine whether having more children results in higher levels of infant mortality.

The null hypothesis for the F test is that the proportion of variation in *y* explained by *x* is zero. Therefore:

$H_o : r^2 = 0$

$H_a : r^2 \neq 0$

The null hypothesis for the t test is that the slope of the regression line is not zero. Therefore:

$H_o : \beta = 0$

$H_a : \beta \neq 0$

The F statistic measures the probability that the independent variable(s) in the model are correlated with the dependent variable beyond what could be explained by pure chance (due random sampling error).

Null hypothesis for the F test:

H_o : There is no association between infant mortality and the number of births per woman.

H_a : There is an association between infant mortality and the number of births per woman.

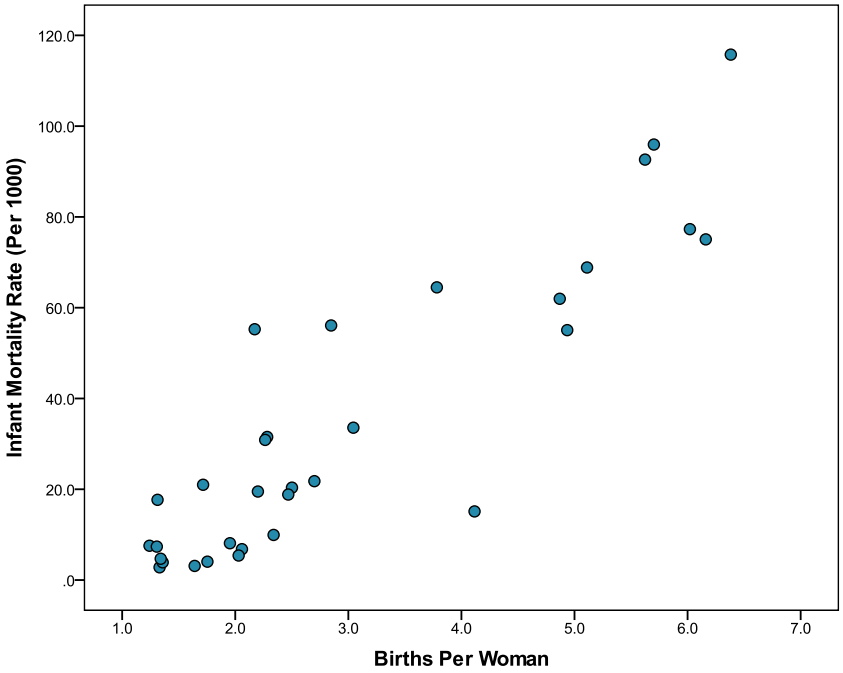
The t statistic measures the probability that the slope of the bet fit line is not zero. In other words, that there is no linear relationship between the variables. The Null hypothesis for the t test:

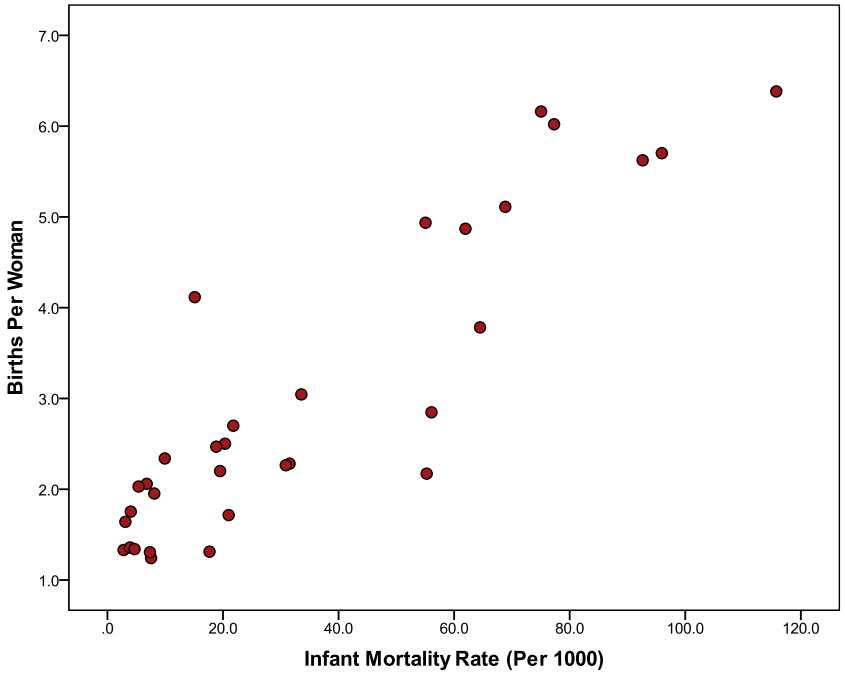
H_o : There is no positive relationship between infant mortality rates and the number of births per woman.

H_a : There is no positive relationship between infant mortality rates and the number of births per woman.

Also, be sure to state the *direction* of the relationship in your summary statement.

Note that the scatterplot are mirror images of each other, depending on how you assign the variables.





This will change the slope of the regression line, but the relationship between the variables will remain the same.

Information that remains consistent

Dependent **= Births per woman**
Independent **= Infant mortality rate**

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.899[a]	.808	.801	.7459

a. Predictors: (Constant), InfMort

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	70.071	1	70.071	125.956	.000[b]
	Residual	16.690	30	.556		
	Total	86.761	31			

a. Dependent Variable: BirthPerWoman

b. Predictors: (Constant), InfMort

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	1.396	.196		7.138	.000
	InfMort	.047	.004	.899	11.22	.000

a. Dependent Variable: BirthPerWoman(10yravg)

Dependent **= Infant mortality rate**
Independent **= Births per woman**

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.899[a]	.808	.801	14.3686

a. Predictors: (Constant), BirthPerWoman

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	26004.465	1	26004.465	125.956	.000[b]
	Residual	6193.722	30	206.457		
	Total	32198.187	31			

a. Dependent Variable: InfMort

b. Predictors: (Constant), BirthPerWoman

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-17.486	5.303		-3.29	.003
	BirthPerWoman	17.313	1.543	.899	11.22	.000

a. Dependent Variable: InfMort

Information that changes

Dependent **= Births per woman**
Independent **= Infant mortality rate**

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.899[a]	.808	.801	.7459

a. Predictors: (Constant), InfMort

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	70.071	1	70.071	125.956	.000[b]
	Residual	16.690	30	.556		
	Total	86.761	31			

a. Dependent Variable: BirthPerWoman

b. Predictors: (Constant), InfMort

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	1.396	.196		7.138	.000
	InfMort	.047	.004	.899	11.22	.000

a. Dependent Variable: BirthPerWoman

Dependent **= Infant mortality rate**
Independent **= Births per woman**

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.899[a]	.808	.801	14.3686

a. Predictors: (Constant), BirthPerWoman

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	26004.465	1	26004.465	125.956	.000[b]
	Residual	6193.722	30	206.457		
	Total	32198.187	31			

a. Dependent Variable: InfMort

b. Predictors: (Constant), BirthPerWoman

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-17.486	5.303		-3.29	.003
	BirthPerWoman	17.313	1.543	.899	11.22	.000

a. Dependent Variable: InfMort

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	InfMort[b]	.	Enter

a. Dependent Variable: BirthPerWoman

b. All requested variables entered.

List of dependent and independent variables

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.899[a]	.808	.801	.7459

a. Predictors: (Constant), InfMort

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	70.071	1	70.071	125.956	.000[b]
	Residual	16.690	30	.556		
	Total	86.761	31			

a. Dependent Variable: BirthPerWoman

b. Predictors: (Constant), InfMort

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	1.396	.196		7.138	.000
	InfMort	.047	.004	.899	11.223	.000

a. Dependent Variable: BirthPerWoman

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	InfMort[b]	.	Enter

a. Dependent Variable: BirthPerWoman

b. All requested variables entered.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.899[a]	.808	.801	.7459

a. Predictors: (Constant), InfMort

List of r, r^2, adjusted r^2, and standard error

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	70.071	1	70.071	125.956	.000[b]
	Residual	16.690	30	.556		
	Total	86.761	31			

a. Dependent Variable: BirthPerWoman

b. Predictors: (Constant), InfMort

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	1.396	.196		7.138	.000
	InfMort	.047	.004	.899	11.223	.000

a. Dependent Variable: BirthPerWoman

Adjusted R^2

$$R^2_{Adj} = \frac{R^2 - (1 - R^2)\,p}{n - p - 1}$$

where *p* is the number of independent variables and *n* is the sample size.

$$R^2_{Adj} = \frac{0.676 - (1 - 0.676)(1)}{13 - 1 - 1} = 0.03$$

$$Adjusted\ \ R^2 = 0.676 - 0.03 = 0.646$$

The adjusted r^2 penalizes the r^2 for small sample sizes and large numbers of independent variables.

Standard Error of the Estimate

The standard error of the estimate is analogous to the standard deviation. It is the average distance that all of the observations fall from the regression line.

$$SE = \sqrt{\frac{\sum (y - \hat{y})^2}{n}}$$

where y is the observed value, y-hat is the predicted value and n is the sample size.

The standard error of the estimate is in the original units. A lower SE means that the data group more tightly around the regression line.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.899[a]	.808	.801	.7459

a. Predictors: (Constant), InfMort

This standard error of the estimate means that the average residual value (prediction error) is ±0.7459.

- In other words, on average our predictions of the number of births per woman will be off by about 3/4 of a birth.
- Given that the range of births per woman is 5.1, our predictions will be off by about 15% (0.75/5.1).
- This may or may not be acceptable.

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	InfMort[b]	.	Enter

a. Dependent Variable: BirthPerWoman

b. All requested variables entered.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.899[a]	.808	.801	.7459

a. Predictors: (Constant), InfMort

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	70.071	1	70.071	125.956	.000[b]
	Residual	16.690	30	.556		
	Total	86.761	31			

a. Dependent Variable: BirthPerWoman

b. Predictors: (Constant), InfMort

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	1.396	.196		7.138	.000
	InfMort	.047	.004	.899	11.223	.000

a. Dependent Variable: BirthPerWoman

The f test for the significance of the model.

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	InfMort[b]	.	Enter

a. Dependent Variable: BirthPerWoman

b. All requested variables entered.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.899[a]	.808	.801	.7459

a. Predictors: (Constant), InfMort

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	70.071	1	70.071	125.956	.000[b]
	Residual	16.690	30	.556		
	Total	86.761	31			

a. Dependent Variable: BirthPerWoman

b. Predictors: (Constant), InfMort

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	1.396	.196		7.138	.000
	InfMort	.047	.004	.899	11.223	.000

a. Dependent Variable: BirthPerWoman

Constant = intercept (a)
Named variable = slope (b)

Intercept and slope parameters and t-tests of those parameters.

Regression Analysis Example: Nitrate Productivity

Model Summary[b]

Model	*R*	*R Square*	*Adjusted R Square*	*Std. Error of the Estimate*	*Durbin-Watson*
1	.762[a]	.581	.562	39712.788	1.922

a. Predictors: (Constant), Workers

b. Dependent Variable: Volume of Machinery (ft3)

ANOVA[a]

Model		*Sum of Squares*	*df*	*Mean Square*	*F*	*Sig.*
1	***Regression***	48102432118.133	1	48102432118.133	30.500	.000[b]
	Residual	34696321978.825	22	1577105544.492		
	Total	82798754096.958	23			

a. Dependent Variable: Volume of Machinery (ft3)

b. Predictors: (Constant), Workers

Coefficients[a]

		Unstandardized Coefficients		*Standardized Coefficients*		
Model		*B*	*Std. Error*	*Beta*	*t*	*Sig.*
1	***(Constant)***	14294.606	14758.875		.969	.343
	Workers	190.306	34.459	.762	5.523	.000

a. Dependent Variable: Volume of Machinery (ft3)

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	*Std. Error*	*Beta*	*t*	*Sig.*
1	*(Constant)*	14294.606	14758.875		.969	.343
	Workers	190.306	34.459	.762	5.523	.000

a. Dependent Variable: Volume of Machinery (ft3)

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

$$\hat{y} = 14294.6 + (190.3)x$$

$$\hat{y} = 14294.6 + (190.3)x$$

The regression equation above would read:

For every unit (1 worker) increase in the workers, the volume of machinery increases by 190.3 ft^3. What this suggests is that for each additional 190 ft^3 of added production capacity the company would have to hire an additional worker.

Model Summary[b]

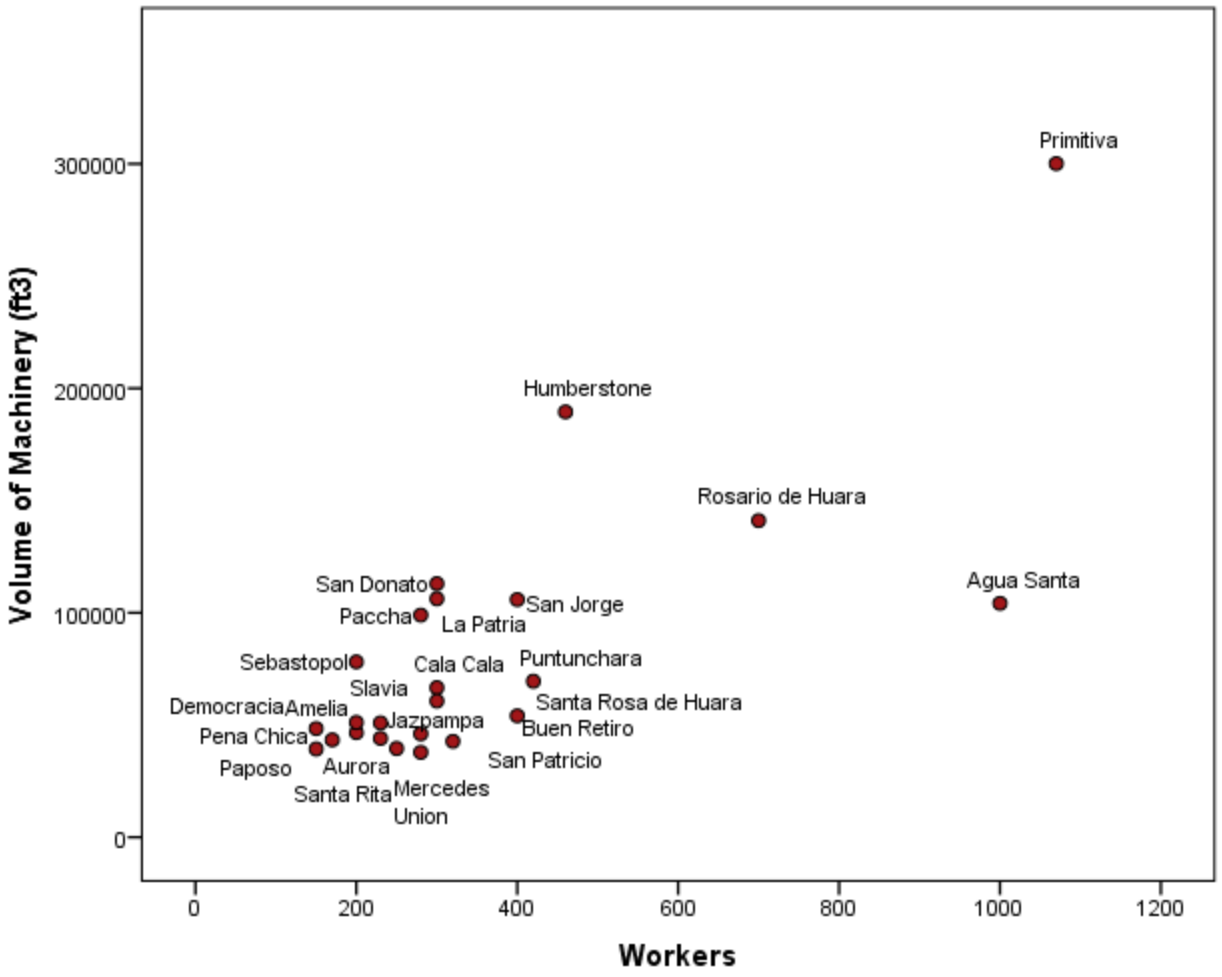
Model	*R*	*R Square*	*Adjusted R Square*	*Std. Error of the Estimate*	*Durbin-Watson*
1	.762[a]	.581	.562	39712.788	1.922

a. Predictors: (Constant), Workers

b. Dependent Variable: Volume of Machinery (ft3)

However, notice that the r2 is rather low, meaning that the relationship between nitrate production capacity and the number of workers is somewhat weak.

Plots are helpful here.



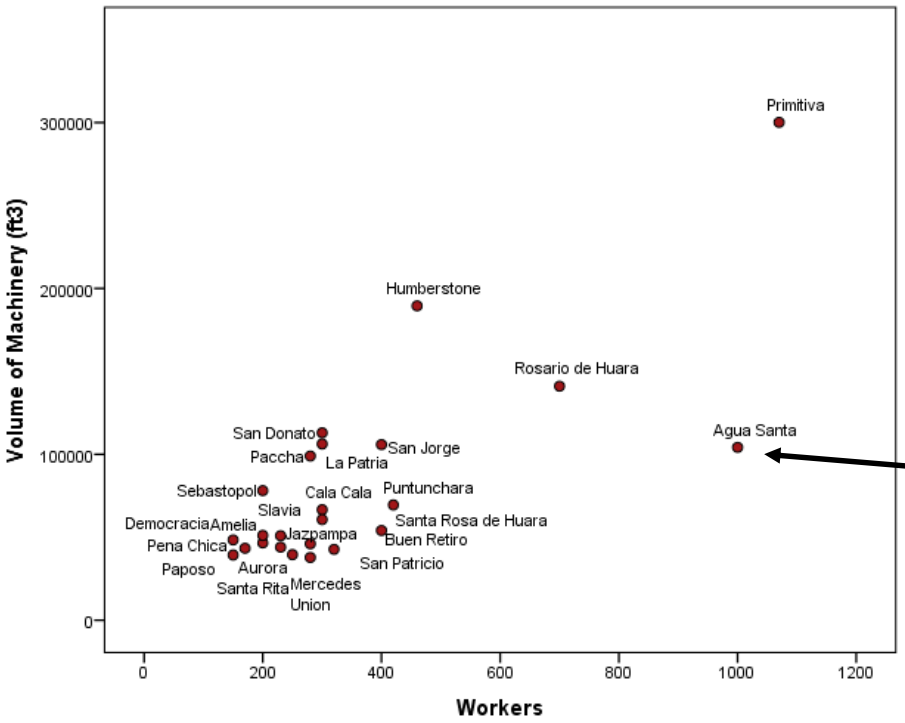
Notice how one production facility (Agua Santa) is much different than the others.

The questions becomes:

1. Is there a data entry error?

2. If not, why is this particular observation different than the trend?

	Machinery Volume	Workers
Agua Santa	104,210 ft^3	1000



Agua Santa has a lot more workers than machinery. Why?

Predicting y from x

$$\hat{y} = 14294.6 + (190.3)x$$

Slavia:
Workers= 230
Actual Machine Volume = 50949

$$\hat{y} = 14294.6 + (190.3)230$$

$$\hat{y} = 58063.6$$

Primitiva:
Workers = 1070
Actual Machine Volume= 300120

$$\hat{y} = 14294.6 + (190.3)1070$$

$$\hat{y} = 217915.6$$

Residual Values

$$\varepsilon = y - \hat{y}$$

Slavia:
Workers= 230
Actual Machine Vol = 50949
Predicted Vol = 58063.6

$$\varepsilon = 50949 - 58063.6$$

$$\varepsilon = -7114.6$$

Primitiva:
Workers = 1070
Actual Machine Vol = 300120
Predicted Vol = 217915.6

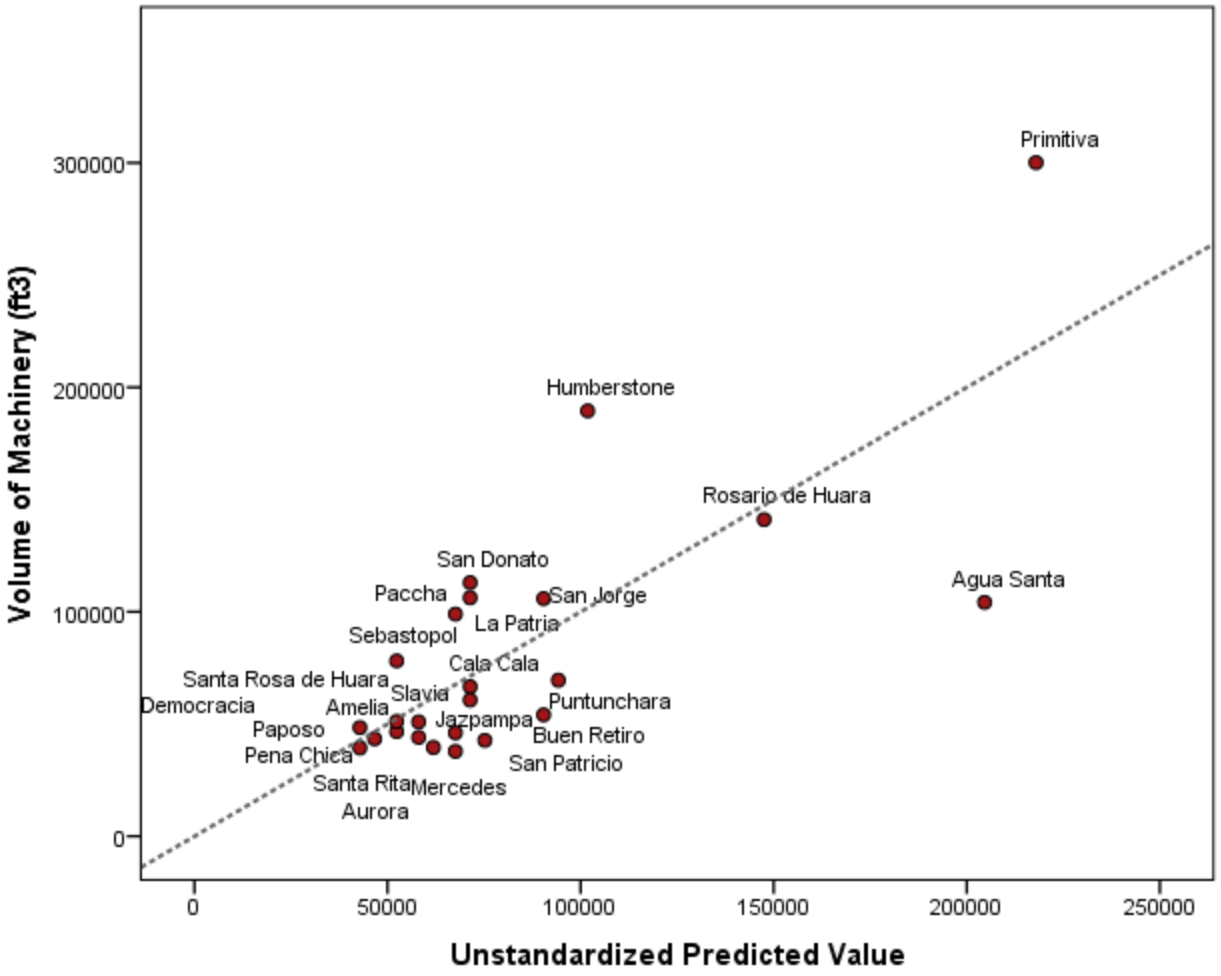
$$\varepsilon = 300120 - 217915.6$$

$$\varepsilon = 82204.4$$

In SPSS, negative residuals equate to “over-prediction.”

Graph Representation of Residuals

(The error is the distance from the observation to the line)



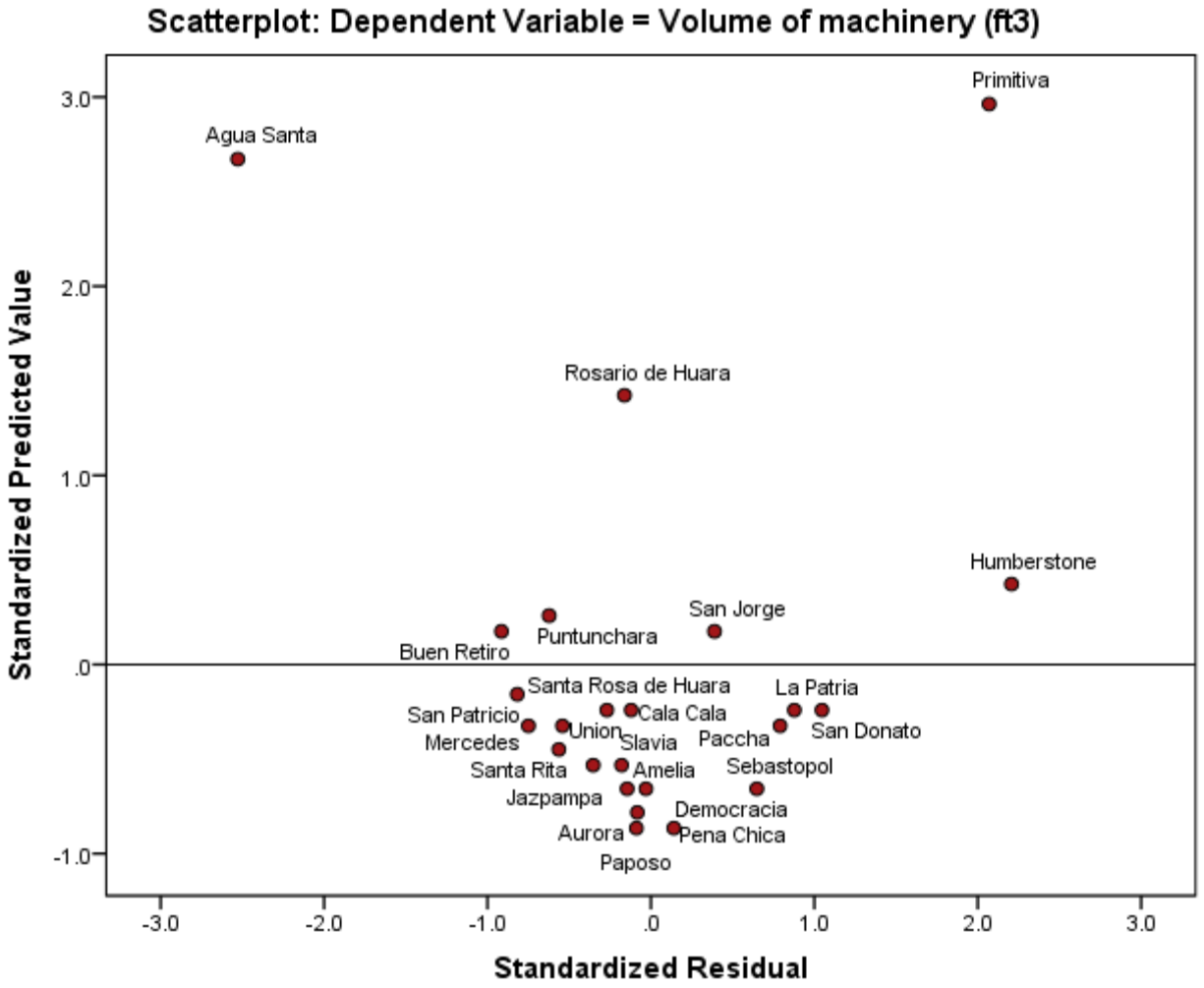
Observed residual values should be approximately equally distributed about the mean residual values.

About ½ the residuals should be positive and ½ negative.

Residuals should be normally distributed.

Residual outliers (those values far from the mean) may be of interest, since the model predicted them poorly.

Never remove an observation solely due to it having a high residual value.


Scatterplot: Dependent Variable = Volume of machinery (ft3)
Standardized Predicted Value
Standardized Residual
3.0
2.0
1.0
.0
-1.0
-3.0
-2.0
-1.0
.0
1.0
2.0
3.0
Primitiva
Agua Santa
Rosario de Huara
Humberstone
San Jorge
Puntunchara
Buen Retiro
Santa Rosa de Huara
La Patria
San Patricio
Cala Cala
Union
Slavia
Paccha
San Donato
Mercedes
Santa Rita
Amelia
Sebastopol
Jazpampa
Democracia
Aurora
Pena Chica
Paposo

Our theory is that the ability to produce nitrate was primarily a function of the number of workers and the volume of the machinery.

Our analyses showed that while there is a relationship between these two variables ($r^2 = 0.56$), other factors are operating.

Perhaps the there are differences in the purity of the nitrate ore among these production facilities that is having an effect on their ability to produce nitrate.

Our analyses has made us reexamine, and hopefully improve, our original hypothesis.

Optional Regression Output:

Durbin-Watson statistic – tests for serial correlation between residuals. Higher Durbin-Watson statistics means there is less serial correlation. The statistic has a range of 0 – 4. A value near 2 is considered good.

Model Summary[b]

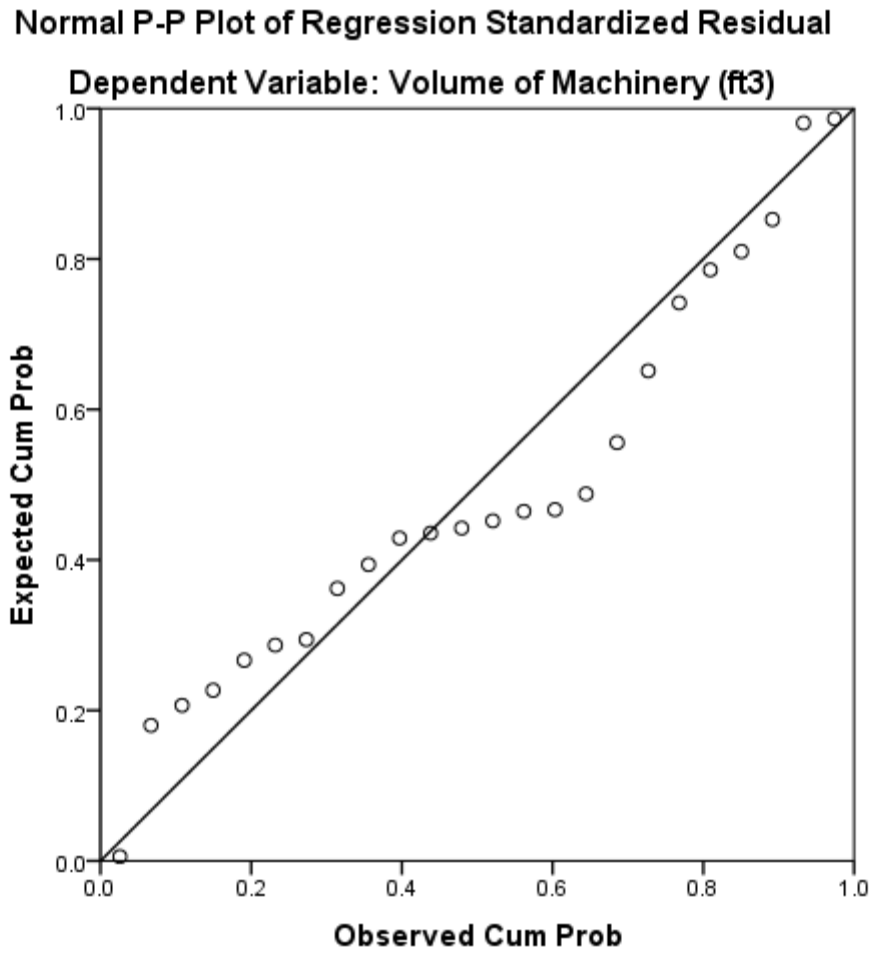
Model	*R*	*R Square*	*Adjusted R Square*	*Std. Error of the Estimate*	*Durbin-Watson*
1	.762[a]	.581	.562	39712.788	1.922

a. Predictors: (Constant), Workers

b. Dependent Variable: Volume of Machinery (ft3)

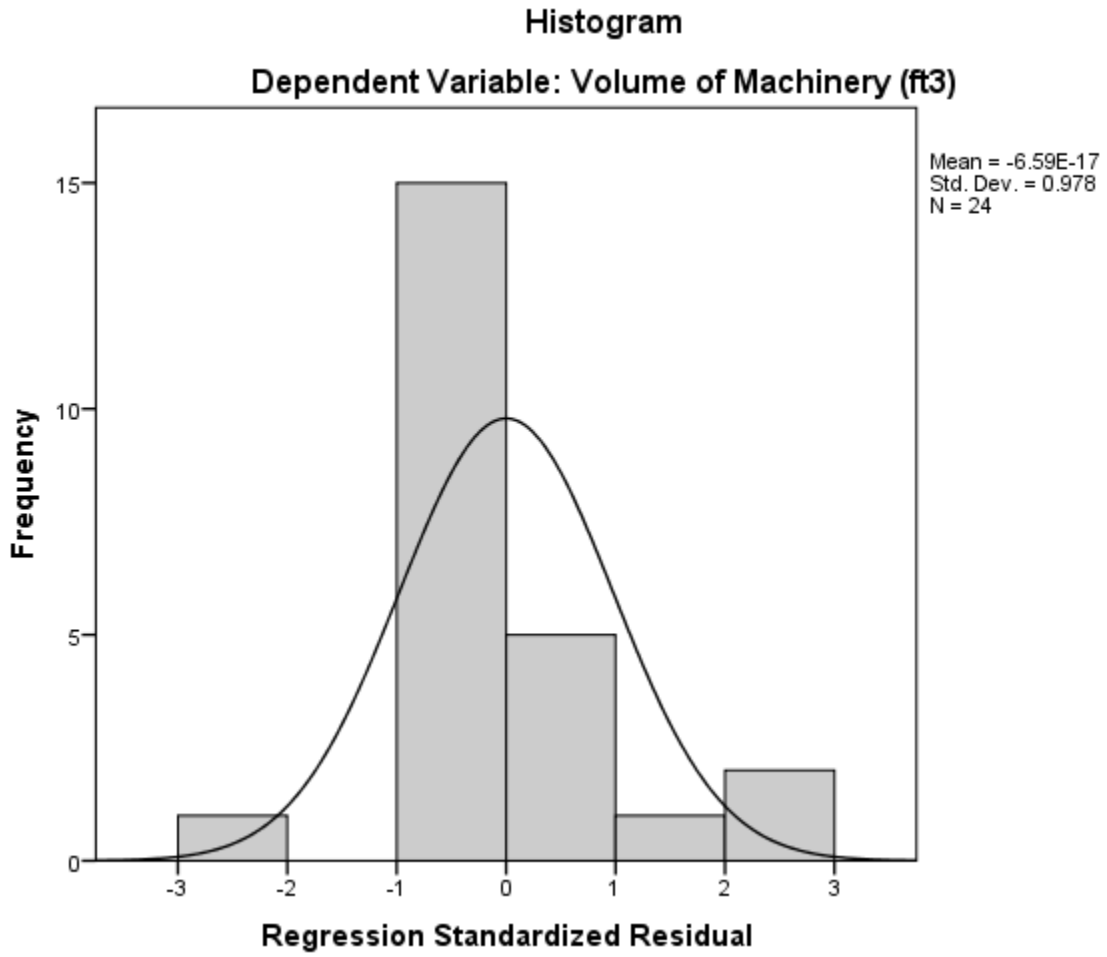
Optional Regression Output:

P-P Residual Plot – used to check for normally distributed residuals.



Optional Regression Output:

Residual Histogram – used to check for normally distributed residuals.



Optional SPSS Regression Output:

Predicted v Residual Scatterpot – used to check for outliers and that there are no patterns in the residuals. *Choose y=zpred and x=zresid*.