Regression

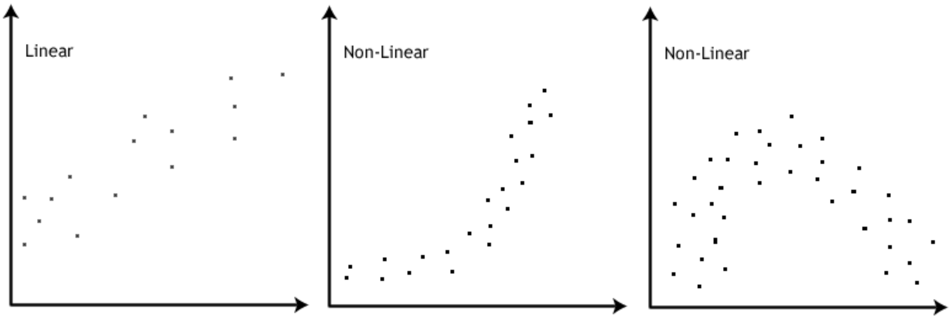
1. Simple linear regression

In simple linear regression a single independent variable is used to predict the value of a dependent variable. A linear regression model has the form

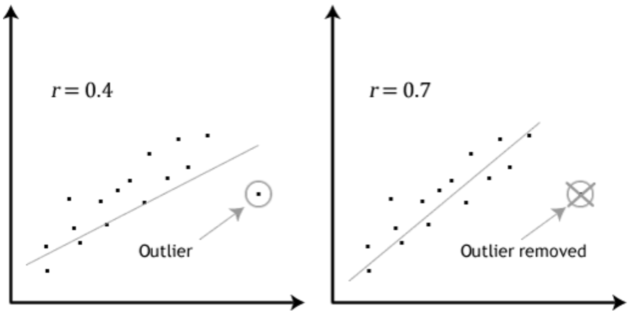
$$Y = ax + e$$

In multiple linear regression two or more independent variables are used to predict the value of a dependent variable. The difference between the two is the number of independent variables. Your data should be continuous, i.e. not nominal or ordinal. Below are the assumptions of a simple Linear regression

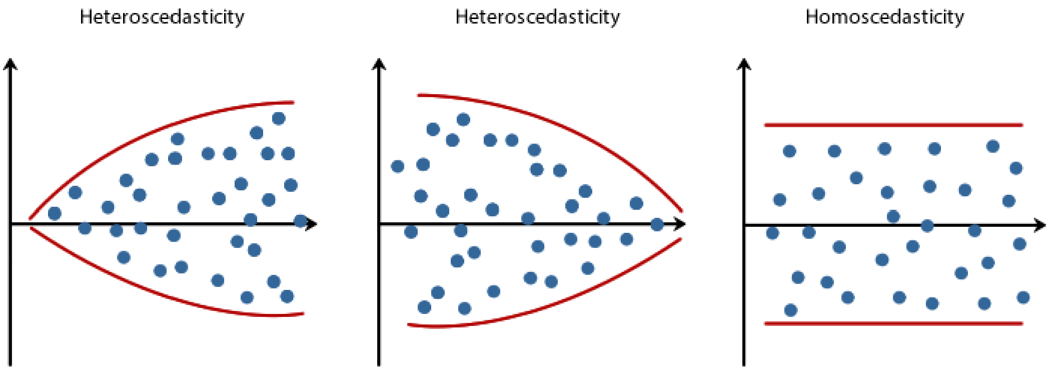
- There should be a linear relationship between your two variables.



- There should by no significant outliers.



- The data needs to show homoscedasticity of variance.



Copyright 2014. Laerd Statistics.

- The residuals of the regression need to be approximately normally distributed.
- There should be independence of observations.
- Both your variables should be interval or ratio in nature.

Example 1.1. Say we want to use a linear regression to predict drug absorption time of penicillin based on the patients gastric pH. We took a sample of 6 patients and their data is in the table below.

gastric pH	absorption time in minutes
1.5	23
1.8	25
2.0	27
2.5	30
3.0	32
3.2	33

Absorption time is our dependent variable (Y) and gastric pH is our independent variable (X). The first ting we need to do is find the mean for both our variables.

$$\overline{x} = \frac{1.5 + 1.8 + 2 + 2.5 + 3 + 3.2}{6} = 2.33$$

$$\overline{y} = \frac{23 + 25 + 27 + 30 + 32 + 33}{6} = 28.33$$

First calculate the variance of X

$$\sigma^2 = \frac{(1.5 - 2.33)^2 + (1.8 - 2.33)^2 + (2 - 2.33)^2 + (2.5 - 2.33)^2 + (3 - 2.33)^2 + (3.2 - 2.33)^2}{5}$$

$$\sigma^2 = \frac{0.6889 + 0.2809 + 0.1089 + 0.0289 + 0.4489 + 0.7569}{5} = 0.46268$$

Now we calculate the covariance of X and Y using the following formula

$$Cov(X, Y) = \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{n - 1}$$

$$Cov(X, Y) = \frac{4.4239 + 1.7649 + 0.4389 + 0.2839 + 2.4589 + 4.0629}{5}$$

$$Cov(X, Y) = 2.687$$

To get the coefficient of x in our model we divide the covariance by the variance of X.

$$a = \frac{Cov(X, Y)}{\sigma^2} = \frac{2.687}{0.46268} = 5.807$$

To estimate the error we subtract $a \times \overline{x}$ from $\overline{y}$.

$$e = \overline{y} - a\overline{x} = 28.33 - (5.807 \times 2.33) = 14.8$$

Our regression model is

$$Y' = ax + e$$

$$Y' = 5.807x + 14.8$$

Below is how to test the assumptions in Minitab:

- Make a scatter plot to check if there is a linear relationship between the variables.
  1. Go to **Graph** → **Scatterplot**.
  2. Then click on **With Regression**.
  3. Enter your dependent variable as your Y variable and the independent variable as the X variable.
  4. Click OK.
- When you run your regression Minitab will detect outliers in the **Fits and Diagnostics for Unusual Observations** and denote them with an **R** and an **X** if it's just an unusual observation.
- To check for homoscedasticity when you run your regression and go in to **Graphs** and tick all the options. Look for the **Residual vs Fitted Value** graph.
- Look at the **Normal Probability Plot** to check if the residuals are normally distributed.

Below is the output from Minitab for the above example.

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	1	78.0062	78.0062	235.12	0.0001
Error	4	1.3271	0.338		
Total	5	79.3333			

Model Summary

S	R-sq	R-sq(adj)
0.575997	98.33%	97.91%

Coefficients

Term	Coeff	SE Coeff	T-Value	P-Value
Constant	14.7839	0.9134	16.17	$<$0.0001
pH	5.8069	0.3787	15.33	0.0001

Regression Equation

$time = 14.7839 + 5.8069pH$

Here we'll go through how to interpret the Model Summary table;

- R: This represents the correlation between the independent and dependent variables.
- R^2: Means the proportion of the variation of our dependent variable which can be explained by variation in the independent variable.
  - if $R^2 = 1$ there is a perfect match between the line and the data points.
  - if $R^2 = 0$ There is no linear relationship between the independent and dependent variables.
- Adjusted R^2: Adjusts the R^2 number to account for variation which is the result of a small sample size.
- Std. estimate of the error: This is the standard deviation of the error and is the square root of the Mean Square Residual.

The Analysis of Variance table can be interpreted as follow;

- This indicates the statistical significance of the regression model that was run.
- If the regression model is statistically significant it indicates that it predicts the dependent variable and it's a good fit for the data.

The Coefficients be broken down as follows;

- **Coef:** This column gives us the regression equation.
- **SE Coef:** The error associated with each the constant term and the independent variable.
- **Sig:** This tells you if a particular independent variable has a significant relationship with the dependent variable.

2. Multiple Regression

Example 2.1. Say we want to use a linear regression to predict drug absorption time of penicillin based on the patients gastric pH and their weight in kg. We took a sample of 6 patients and their data is in the table below.

Weight	Gastric pH	Absorption time in minutes
67	1.5	23
70	1.8	25
68	2.0	27
72	2.5	30
75	3.0	32
77	3.2	33

Below are the assumptions of a multiple regression;

- These are largely the same as a simple linear regression.
- There must be at least be two independent variables.
- There must also be no multicolinearity.
  - This can be checked by making sure the numbers in the VIF column in the coefficients table do not exceed 5.

The model takes (broadly) the form as the Simple linear model, that is

$$Y' = aX_1 + bX_2 + c,$$

where, in our example, Y represents absorption time, X_1 is gastric pH and X_2 is weight.

Example 2.2. Below is the output from Minitab for the example above;

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	2	78.6613	39.3306	175.57	0.0008
Error	3	0.6721	0.2240		
Total	5	79.3333			

Model Summary

S	R-sq	R-sq(adj)
0.473308	99.15%	98.59%

Coefficients

Term	Coeff	SE Coeff	T-Value	P-Value	VIF
Constant	34.85	11.76	2.96	0.0594	
pH	7.720	1.161	6.65	0.0069	13.93
wieght	-0.3431	0.2006	-1.71	0.1858	13.93

Regression Equation

$$time = 34.85 + 7.720\text{pH} - 0.3431weight$$

We interpret the tables the same way as we did in the simple linear regression.

- From the ANOVA table we can see that the regression model is a good fit for the data.
- The coefficients table tells us there is a significant relationship between pH and absorption time. However the relationship between weight and absorption time is not significant.
- The VIF score also tells us there is a lot of multicolinearity. Therefore it may be better to remove the "weight" variable from the model since it is not significantly relate to absorption time and there is a lot of multicolinearity.

3. Logistic Regression

Definition 3.1. A (binomial/binary) logistic regression predicts a dichotomous dependent variable based on one or more independent variable.

Logistic regression makes the following assumptions:

(1) The dependent variable should be measured on a dichotomous scale, i.e. outcomes can only fall into one of two categories.
(2) You can have one or more independent variable which can be scale or categorical in nature.
(3) There should be independence of observation and you categories should be mutually exclusive.
(4) There should be a linear relationship between the independent variables and the logit transform of the dependent.
(5) There should be no multicolinearity.
(6) There should be no significant outliers or risiduals.

Example 3.2. We've taken a sample of twenty one people and recorded whether they have heart disease or not. We've also recorded the height, age and weight of the participants. We're going to use logistic regression to see if we can predict if a person has heart disease based on the above independent variables.

Age	Heart disease status	Weight in kg
34	No	56.18
35	Yes	86.13
42	Yes	87.3
37	No	76.34
45	Yes	86
47	No	80.43
50	No	77
33	No	75.81
31	No	72.56
40	No	78.9
48	Yes	87
46	Yes	86
49	Yes	91.23
30	Yes	88.76
53	Yes	81.65
50	No	73.45
43	No	68.23
35	No	67.32
47	Yes	78.48
48	Yes	85.45
47	Yes	60

You can test the assumptions in Minitab as follows:

- You can check the multicolinearity by looking at the VIF number on the Coefficients table.
- if any of these numbers are greater than 5 then there is a problem with multicolinearity.
- If there are any outliers or residuals, they will be highlighted at the end of the output in the **Fits and Diagnostics for Unusual Observations** table.

To do a logistic regression in Minitab perform thee steps:

- Go to **Stat** → **Regression** → **Binary Logistic Regression** → **Fit Binary Logistic Model**.
- Enter **Heart Disease** under **Response**.
- Enter **Age** and **Weight** under **Continuous Predictors**.
- Click ok.

Then we get the output:

Regression Equation

$P(Yes) = exp(Y')/(1 + exp(Y'))$

$Y' = -18.788 + 0.08885Age + 0.19107Weight$

Response Information

Variable	Value	Count	
Heart Disease	Yes	11	(Event)
	No	9	
	Total	21	

Deviance Table

Source	DF	Adj Dev	Adj Mean	Chi-square	P-Value
Regression	2	9.9576	4.97881	9.96	0.0069
Age	1	1.2007	1.20075	1.20	0.2732
Weight	1	7.6296	7.62958	7.63	0.0057
Error	18	19.1069	1.06150		
Total	20	29.0645			

Model Summary

Deviance R-sq	Deviance R-sq(adj)	AIC
34.26%	27.38%	25.11

Coefficients

Term	Coeff	SE Coeff	VIF
Constant	-18.79	8.38	
Age	0.0888	0.0827	1.01
Wieght	0.1911	0.0905	1.01

Odds Ratio for Continuous Predictors

	Odds Ratio	95% CI
Age	1.0929	(0.9293, 1.2853)
wieght	1.2105	(1.0137, 1.4456)

Goodness-of-Fit Tests

Test	DF	Chi-Square	P-Value
Deviance	18	19.11	0.3853
Pearson	18	32.12	0.0213
Hosmer-Lemeshow	8	25.55	0.0013

Fits and Diagnostics for Unusual Observations

Obs	Observed Probability	Fit	Resid	Std Resid	
21	1	0.0411819	2.52577	2.73	R
R	Large Residual				

Below is a break down of how to interpret this output.

- The **Goodness of Fit** table has three rows one for deviance, Pearson and Hosmer Lemeshow.
  - The Deviance row determines whether the predicted probabilities in the model deviate significantly from the observed probabilities. If the p-value is lower than your chosen level of significance, the predicted probabilities deviate from the observed in a way that the binomial distribution does not predict. This is the same for the next two rows.
  - The Person row assesses the same problem as the deviation row.
  - The Hosmer-Lemeshow goodness-of-fit test compares the observed and expected frequencies of events and non-events to assess how well the model fits the data. If the p-value is not lower than you significance level, the model is seen as a good fit for the data.
- Deviance Table
  - **Regression:** If the p-value is less than your significance level you can conclude at least one term in you model is great than 0.
  - **Age/Weight:** If the p-value is less than your significance level you can conclude there is a significant association between Age/Weight and Heart disease.
- The Model Summary table tells you how much of the variance in the data can be explained by the model.
- The Regression Equation table gives the equation for the model and how to calculate the probability of the event happening.