METHOD CHOOSER

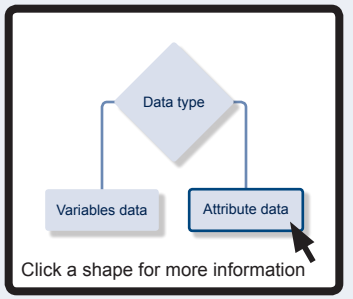
Minitab 15 Statistical Software

Regression and ANOVA

Click here to start

Minitab
QUALITY. ANALYSIS. RESULTS.

Regression and ANOVA

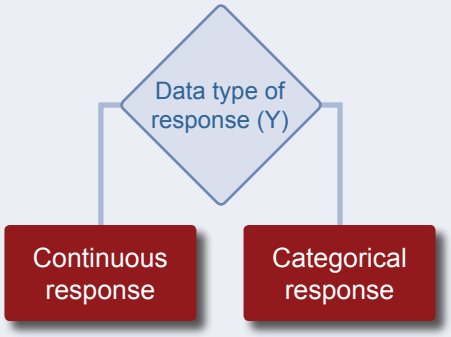


Data type of response (Y)

- Continuous response
  - Number of X (input) variables
    - One X variable
      - Data type of X variable
        - Continuous → Fitted Line Plot
        - Categorical → One-Way ANOVA
    - More than one X variable
      - Data type of primary X variables
        - Continuous → Regression
        - Categorical → General Linear Model
- Categorical response
  - Number of response categories
    - Two categories → Binary Logistic Regression
    - More than two categories
      - Order of categories
        - Natural order → Ordinal Logistic Regression
        - No natural order → Nominal Logistic Regression

Click on any page to return here.

Do you have a continuous response or a categorical response?

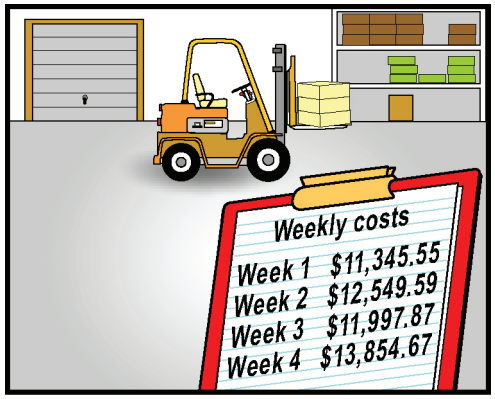


Continuous response

Measures a characteristic of a part or process. Allows you to estimate the mean response with one or more X variables.

Example

A financial analyst tracks the total weekly costs at a retail distribution center. The analyst wants to detemine how the weekly costs are related to the number of cases shipped, labor hours, and energy use.

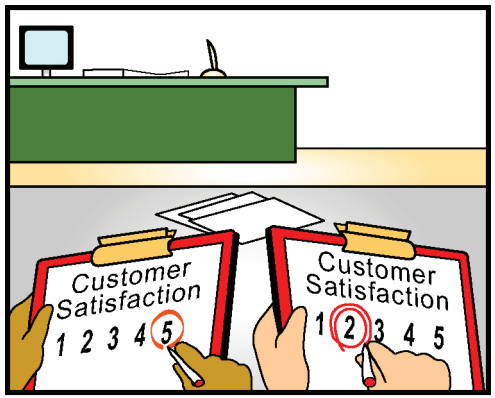


Categorical response

Classifies the response into categories, such as poor, good, excellent. Allows you to estimate the probability of each level of the response with one or more X variables.

Example

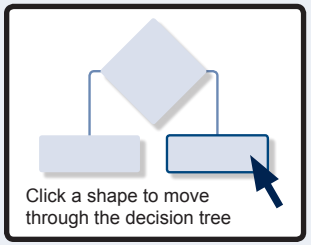
A hotel manager asks guests to rate their satisfaction on a scale of 1 to 5. The manager wants to determine the probability that a customer is unsatisfied (gives a rating of 1 or 2) for each type of room.

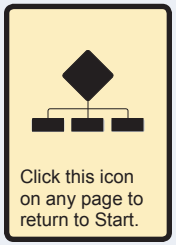


The response is the variable that you want to describe, explain, or predict with an X variable. The response is also called the Y or output variable.

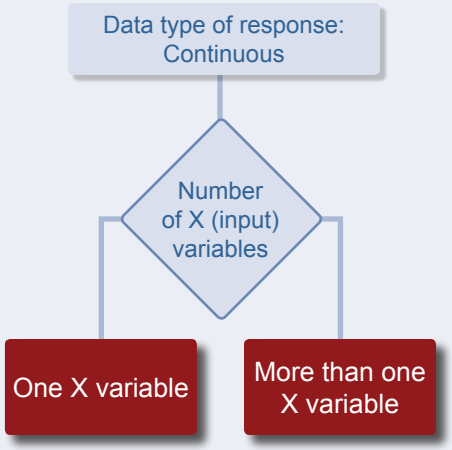
Continuous response variables are measurements, such as length, weight, and temperature, and they often include fractional (or decimal) values. Categorical response variables are characteristics, such as a grade of a raw material, a type of method, or rating, such as high, medium, or low.

If the response variable has many categories that can be naturally ordered and represented by ordinal numbers, you can treat the variable as either continuous or categorical. For example, you ask customers to rate the quality of shoe brands on a scale from 1 to 10. When you analyze the data with a continuous model, the results indicate that the average rating for brand A is 4.4. When you analyze the data with a categorical (attribute) model, the results indicate that the probability that Brand A receives a score of 3 or lower is 40%.





How many X variables do you want to include in the model?

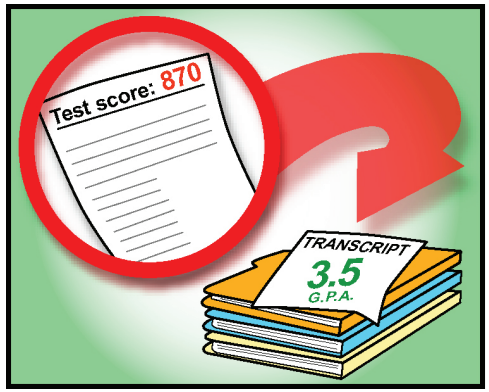


One X variable

Use only one X to describe the response because other potential X variables are not important, or you want to use the simplest model possible.

Example

Investigators want to know whether standardized test scores can predict the future grades of college students. They realize that other variables may affect student grades, but they want to focus solely on the relationship between test scores and the students' grade point averages.

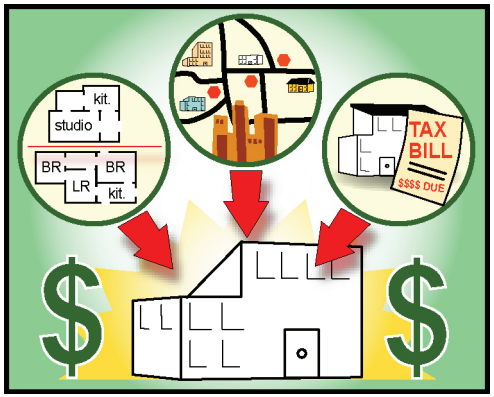


More than one X variable

Use two or more X variables in the model because you need more than one X to adequately describe the response, or you want to study the effect of one X, while accounting for the variability of other X variables.

Example

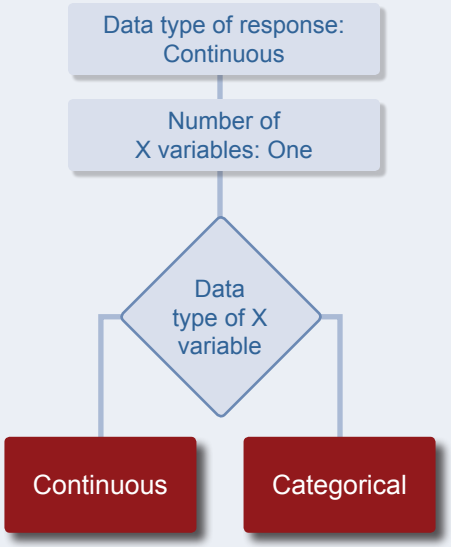
Real estate appraisers determine that the following factors significantly affect the sale price of an urban condominium: size, distance from the city center, and average property tax of nearby homes. The appraisers include all three X variables in the model to predict the sale price of a condominium.



X is an input value that is used to describe, explain, or predict the response. When the value of X changes, the response may also change. X is also called the predictor, input variable, or explanatory variable. When X is categorical, it is often called a factor.

If you are unsure of whether an X variable is important enough to include in the model, you can initially include it. The analysis that you perform helps you to determine which X variables are statistically significant and which X variables explain most of the variability in the response. However, you must still decide whether each X has practical importance based on your knowledge of the process.

Is the X variable continuous or categorical?

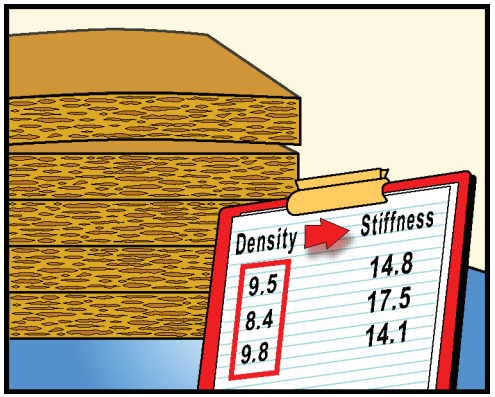


Continuous X

Measures a characteristic of a part or process, such as length, weight, or temperature. The data often includes fractional (or decimal) values.

Example

A manufacturer of particle board wants to know whether the density of the particles affects the stiffness of the board. Investigators measure the density to evaluate how stiffness changes when density increases or decreases.

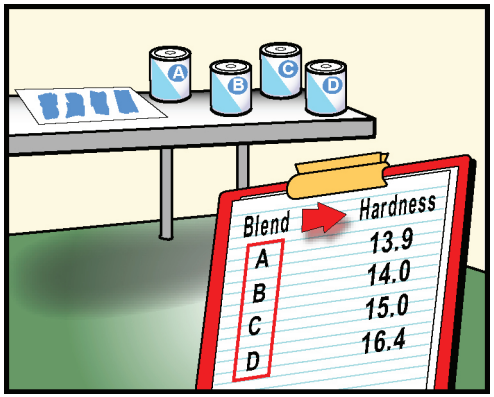


Categorical X

Classifies X into categories based on a characteristic or condition, such as a grade of a raw material, a type of method, or a rating, such as pass/fail.

Example

Investigators at a chemical company want to evaluate whether the blend of a paint (A, B, C, or D) affects its hardness after it dries. Operators apply each paint blend to a piece of metal and measure the hardness of the paint after it dries.



A continuous X is often called a predictor. A categorical X is often called a factor and its categories are called levels.

If X is categorical and contains many categories that can be naturally ordered and represented by ordinal numbers (such as a scale from 1 to 10), you can treat the variable as continuous or categorical. If you treat the variable as continuous, you can predict the response for all continuous values of X. If you treat the variable as categorical, you can estimate the mean response for each level of X.

If X contains counts, such as the number of packages in a shipment or the number of calls handled by a call center, you can generally treat the counts as a continuous even though they are whole numbers.

Data type of response: Continuous
Number of X variables: One
Data type of X variable: Continuous
Fitted Line Plot

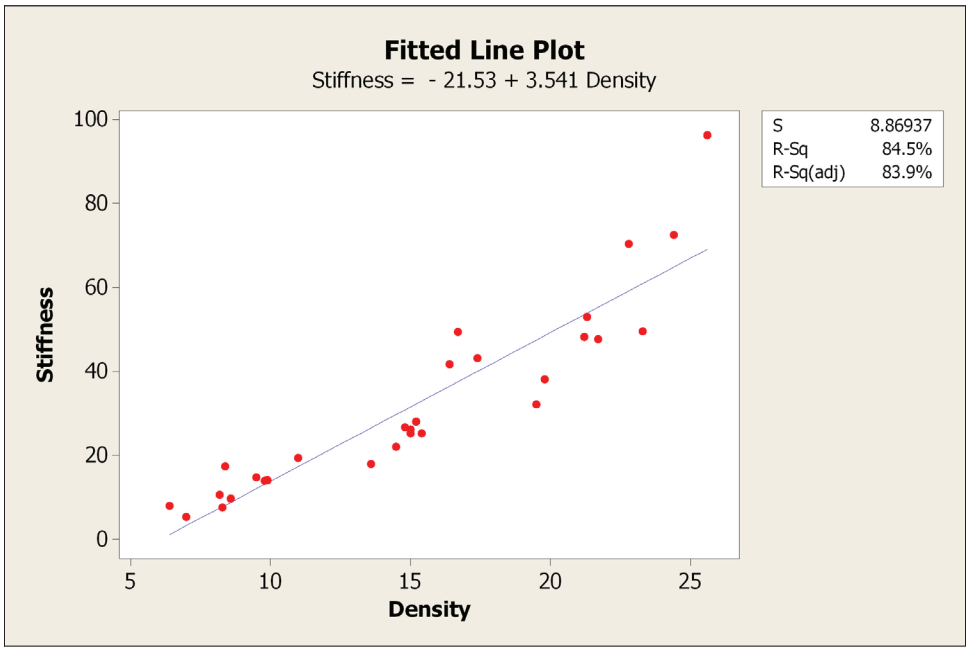
Fitted Line Plot

A fitted line plot examines the relationship between a predictor (continuous X) and a response (continuous Y).

Example

Investigators for a manufacturer of particle board measure the density and stiffness of the boards. They use a fitted line plot to examine the relationship between stiffness and density.

To display a fitted line plot in Minitab, choose **Stat** > **Regression** > **Fitted Line Plot**.



Generally, you first fit a straight line to model the data. A straight line provides the simplest model of the relationship between the response and the predictor. However, if a straight line doesn't fit the data well, you can:

- Fit a curved line with quadratic or cubic terms
- Apply a log transformation to the response or predictor variable

Data type of response: Continuous
Number of X variables: One
Data type of X variable: Categorical
One-Way ANOVA

One-Way ANOVA

One-way ANOVA examines the relationship between a factor (categorical X) and a response (continuous Y).

Example

Investigators at a chemical company measure the hardness of each paint blend after it dries. They use one-way ANOVA to examine the relationship between the blend of the paint and the hardness of the paint.

To perform a one-way ANOVA in Minitab:

- If the response is in one column and the factor levels are in a second column in the worksheet, choose **Stat** > **ANOVA** > **One-Way**.
- If you have a separate response column for each factor level, choose **Stat** > **ANOVA** > **One-Way (Unstacked)**.

One-way ANOVA: Hardness versus Paint

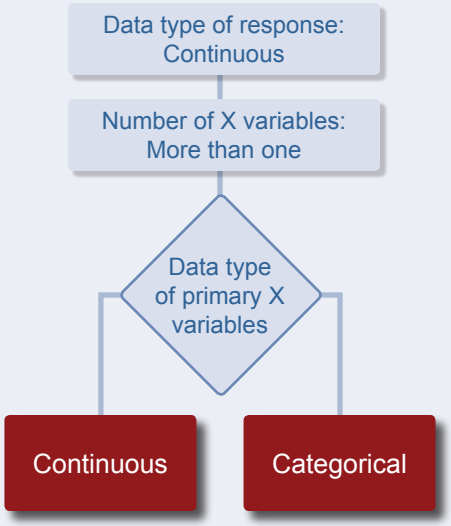
```
Source  DF     SS    MS     F      P
Paint    3  281.7  93.9  6.02  0.004
Error   20  312.1  15.6
Total   23  593.8

S = 3.950   R-Sq = 47.44%   R-Sq(adj) = 39.56%

                           Individual 95% CIs For Mean Based on Pooled StDev
Level    N    Mean  StDev    +---------+---------+---------+---------
Blend A  6  14.733  3.363                  (-----*------)
Blend B  6   8.567  5.500    (------*------)
Blend C  6  12.983  3.730              (------*------)
Blend D  6  18.067  2.636                       (------*------)
                             +---------+---------+---------+---------
                           5.0      10.0      15.0      20.0

Pooled StDev = 3.950
```

For accurate results with one-way ANOVA, your response data should be normal (or nonnormal with relatively few extreme values) at each level of the factor. To check whether the assumptions for the analysis are satisfied, click **Graphs** and under **Residual Plots**, check **Four in one**.

If your response data are nonnormal and contain many outliers at each level of the factor, consider using the Kruskal-Wallis test.

Use Power and Sample Size for one-way ANOVA to determine how much data you need to detect important differences between groups.

Are the primary X variables of interest continuous or categorical?

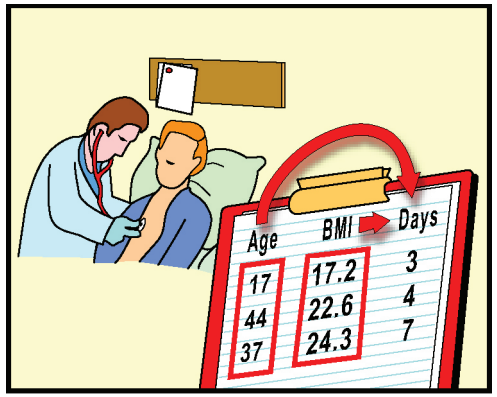


Continuous X variables

Measures a characteristic of a part or process, such as length, weight, or temperature. The data often includes fractional (or decimal) values. Use to predict the response over a continuous range of X values with an equation.

Example

Hospital administrators want to examine how the age and the body mass index (BMI) of a patient are associated with the length of stay in the hospital.

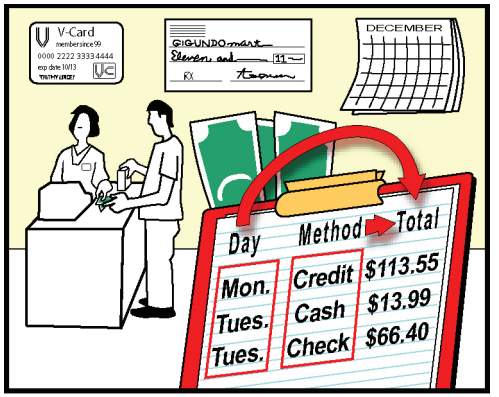


Categorical X variables

Classifies the X into categories based on a characteristic or condition, such as a grade of a raw material, a type of method, or a rating, such as pass/fail. Use to compare the response for each level of X.

Example

Analysts at a large retail store want to examine how the payment method and the day of the week are associated with the cost of a transaction.



A continuous X is often called a predictor. A categorical X is often called a factor and its categories are called levels.

In some processes, both continuous and categorical variables can serve as X variables to explain the response. If so, decide whether the primary X variables are continuous or categorical.

- If your primary X variables are continuous, you can fit a predictive model to assess how changes in X relate to changes in the response. For example, a researcher can predict how changes in blood pressure and cholesterol affect the risk of heart attack.
- If your primary X variables are categorical, you can assess the effect of each factor level and the interactions between factors on the response. For example, a researcher examines how gender (male or female) and race (African-American, Caucasian, Hispanic) affect the risk of heart attack, and whether both gender and race interact to affect the risk.

Data type of response: Continuous
Number of X variables: More than one
Data type of primary X variables: Continuous
Regression

Regression

A regression analysis examines the relationship between a response (continuous Y) and one or more predictors (continuous X variables).

Example

Hospital administrators want to predict the length of stay for a patient in the hospital based on age and body mass index (BMI). They perform a regression analysis to examine the relationship between multiple predictors and a single response.

To perform a regression analysis in Minitab, choose **Stat** > **Regression** > **Regression**.

Regression Analysis: Length of stay versus BMI, Age

```
The regression equation is
Length of stay = - 0.666 + 0.182 BMI + 0.0630 Age

Predictor      Coef   SE Coef      T      P
Constant    -0.6660    0.7084  -0.94  0.350
BMI          0.18248   0.02418   7.55  0.000
Age         0.063043  0.005510  11.44  0.000

S = 1.15395   R-Sq = 62.8%   R-Sq(adj) = 62.1%

Analysis of Variance

Source           DF      SS      MS      F      P
Regression        2  218.22  109.11  81.94  0.000
Residual Error   97  129.17    1.33
Total            99  347.39
```

A regression model can include linear and polynomial terms. For example, you can include a polynomial term, such as BMI^2 to model a curvilinear relationship between body mass index and length of hospital stay. You can also include interactions, such as a term to account for an interaction between body mass index and age.

To include categorical predictors in regression analysis, create indicator variables that use numeric values to represent each level of X. In Minitab, choose **Calc** > **Make Indicator Variables**.
For accurate results with regression, the data must satisfy certain assumptions. To check whether the assumptions for the analysis are satisfied, click **Graphs** and, under **Residual Plots**, check **Four in one**.

Regression and ANOVA
General Linear Model

Data type of response: Continuous

Number of X variables: More than one

Data type of primary X variables: Categorical

General Linear Model

General Linear Model

A general linear model examines the relationship between a response (continuous Y) and one or more factors (categorical X variables). You can also include continuous X variables in the model as covariates.

Example

Analysts at a large retail store track the payment method, the day of the week, and the price of each transaction. They use a general linear model to determine whether the payment method and the day of the week are associated with the cost of the transaction.

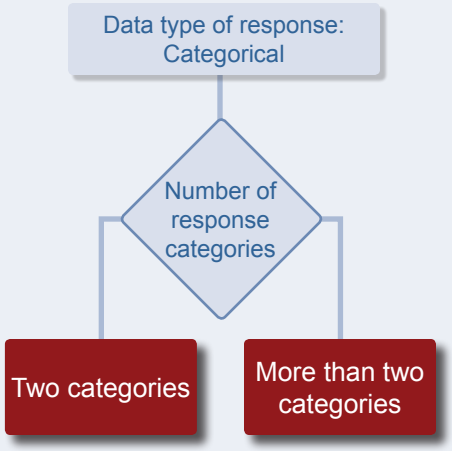
To evaluate a general linear model in Minitab, choose **Stat** > **ANOVA** > **General Linear Model**.

General Linear Model: Transaction price versus Payment Method, Weekday

```
Factor          Type   Levels  Values
Payment Method  fixed       4  Cash, Check, Credit, Debit
Weekday         fixed       7  Friday, Monday, Saturday, Sunday, Thursday,
                               Tuesday, Wednesday

Analysis of Variance for Transaction price, using Adjusted SS for Tests

Source           DF   Seq SS   Adj SS   Adj MS      F      P
Payment Method    3  59745.9  57783.8  19261.3  69.90  0.000
Weekday           6   3331.3   3331.3    555.2   2.01  0.068
Error           130  35821.8  35821.8    275.6
Total           139  98898.9

S = 16.5998   R-Sq = 63.78%   R-Sq(adj) = 61.27%

Unusual Observations for Transaction price

     Transaction
Obs        price      Fit  SE Fit  Residual  St Resid
 25       8.1000  39.7587  6.0686  -31.6587     -2.05 R
140      31.5800  63.9319  4.0800  -32.3519     -2.01 R

R denotes an observation with a large standardized residual.
```

For accurate results with general linear model, the data must satisfy certain assumptions. To check whether the assumptions for the analysis are satisfied, click **Graphs** and, under **Residual Plots**, check **Four in one**.

How many categories does your response have?

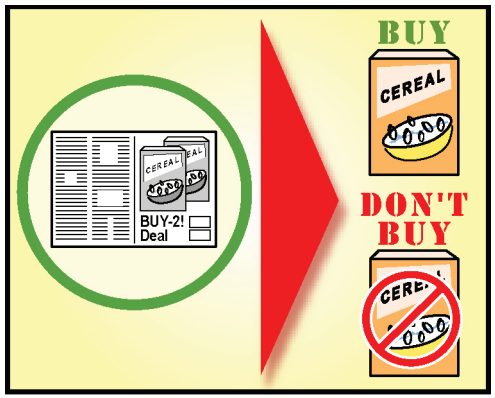


Two response categories

Response is classified into exactly two categories, such as pass/fail or yes/no.

Example

A cereal company wants to know whether customers who saw an advertisement for its new cereal are more likely to buy the product. Marketing analysts randomly sample customers and ask them whether they saw the advertisement and whether they bought the cereal.

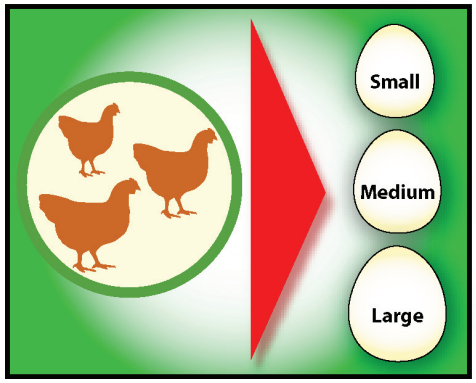


More than two response categories

Response is classified into more than two categories, such as poor, good, and excellent, or north, south, east, and west.

Example

Agricultural researchers want to know whether the weight of a hen is related to the size of its eggs. They randomly sample hens, record the weight of each hen, and classify the size of its eggs as small, medium, or large.



The response is the variable that you want to describe, explain, or predict with an X variable. The response is also called the Y or output variable.

Data type of response: Categorical
Number of response categories: Two
Binary Logistic Regression

Binary Logistic Regression

Binary logistic regression examines the relationship between one or more X variables and a categorical response with two categories.

Example

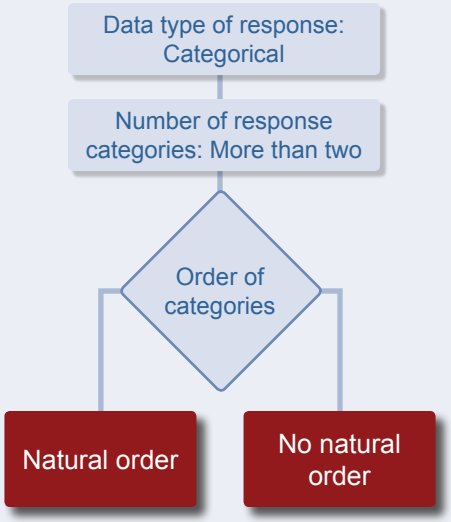
Marketing analysts at the cereal company ask customers whether they saw an advertisement for its new cereal and whether they bought the cereal. They use binary logistic regression to determine whether a customer who has seen the advertistment is more likely to buy the cereal.

To perform a binary logistic regression in Minitab, choose **Stat** > **Regression** > **Binary Logistic Regression**.

Binary Logistic Regression: Bought versus ViewAd

```
Link Function: Logit

Response Information

Variable  Value  Count
Bought    1         22  (Event)
          0         49
          Total     71

Logistic Regression Table

                                                    Odds     95% CI
Predictor      Coef   SE Coef      Z      P  Ratio  Lower  Upper
Constant   -1.45529  0.419750  -3.47  0.001
ViewAd
 Yes        1.21890  0.543589   2.24  0.025   3.38   1.17   9.82

Log-Likelihood = -41.278
Test that all slopes are zero: G = 5.341, DF = 1, P-Value = 0.021

* NOTE * No goodness of fit test performed.
* NOTE * The model uses all degrees of freedom.

Measures of Association:
(Between the Response Variable and Predicted Probabilities)

Pairs       Number  Percent  Summary Measures
Concordant     450     41.7  Somers' D              0.29
Discordant     133     12.3  Goodman-Kruskal Gamma  0.54
Ties           495     45.9  Kendall's Tau-a        0.13
Total         1078    100.0
```

Do the response categories follow a natural order?

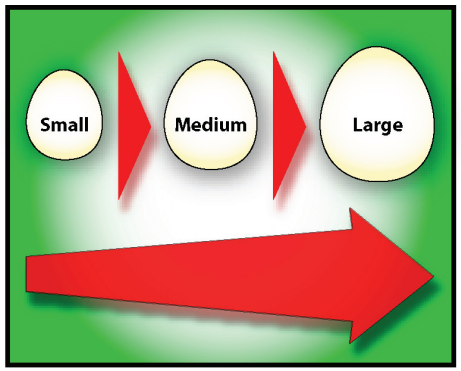


Natural order

Categories for the response can be arranged from least to greatest.

Example

Agricultural researchers want to investigate whether the weight of a hen is related to the size of its eggs. They record the weight of each hen and whether its eggs are small, medium, or large.

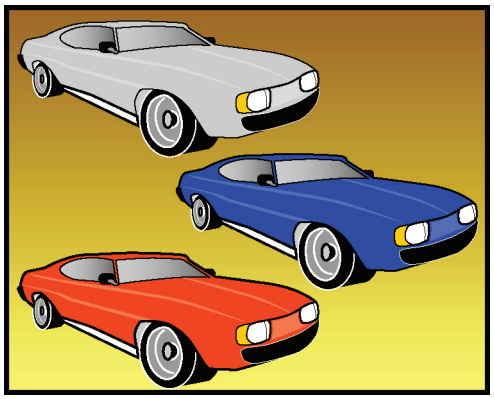


No natural order

Categories for the response cannot be arranged from least to greatest.

Example

Marketing analysts at an automotive company want to know whether the color of the vehicle that consumers purchase is related to their gender or age. Because the colors of the vehicles cannot be arranged from least to greatest, the response categories do not follow a natural order.



Regression and ANOVA
Ordinal Logistic Regression

Data type of response: Categorical

Number of response categories: More than two

Order of categories: Natural order

Ordinal Logistic Regression

Ordinal Logistic Regression

Ordinal logistic regression examines the relationship between one or more X variables and a categorical response with three or more categories that follow a natural order.

Example

Agricultural researchers record the weight of each hen and size of its eggs (small, medium, or large). The researchers use ordinal logistic regression to determine whether the weight of the hen is related to the size of its eggs.

To perform ordinal logistic regression in Minitab, choose **Stat** > **Regression** > **Ordinal Logistic Regression**.

Ordinal Logistic Regression: Egg size versus Weight

```
Link Function: Logit

Response Information

Variable  Value  Count
Egg size  1         25
          2          9
          3         26
          Total     60

Logistic Regression Table

                                                    Odds      95% CI
Predictor       Coef   SE Coef      Z      P  Ratio  Lower  Upper
Const(1)     4.62840   1.32875   3.48  0.000
Const(2)     5.53347   1.38379   4.00  0.000
Weight     -0.780097  0.212678  -3.67  0.000   0.46   0.30   0.70

Log-Likelihood = -47.834
Test that all slopes are zero: G = 25.739, DF = 1, P-Value = 0.000

Goodness-of-Fit Tests

Method    Chi-Square  DF      P
Pearson      85.8181  61  0.020
Deviance     63.4791  61  0.389

Measures of Association:
(Between the Response Variable and Predicted Probabilities)

Pairs       Number  Percent  Summary Measures
Concordant     943     85.0  Somers' D              0.72
Discordant     148     13.3  Goodman-Kruskal Gamma  0.73
Ties            18      1.6  Kendall's Tau-a        0.45
Total         1109    100.0
```

Regression and ANOVA
Nominal Logistic Regression

Data type of response: Categorical

Number of response categories: More than two

Order of categories: No natural order

Nominal Logistic Regression

Nominal Logistic Regression

Nominal logistic regression examines the relationship between one or more X variables and a categorical response with three or more categories that do not follow a natural order.

Example

Marketing analysts at an automotive company record the age and gender of each car buyer and the color of the car they purchase. They use nominal logistic regression to determine whether the color of the car is related to the gender and the age of the buyer.

To perform nominal logistic regression in Minitab, choose **Stat** > **Regression** > **Nominal Logistic Regression**.

Nominal Logistic Regression: Color Preference versus Gender, Age

Response Information

Variable	Value	Count	
Color Preference	White	108	(Reference Event)
	Silver	107	
	Red	71	
	Light blue	35	
	Green	51	
	Dark blue	43	
	Black	40	
	Total	455	

Logistic Regression Table

Predictor	Coef	SE Coef	Z	P	Odds Ratio	95% CI Lower
Logit 1: (Silver/White)						
Constant	-0.883311	0.500823	-1.76	0.078		
Gender						
Male	0.527942	0.309343	1.71	0.088	1.70	0.92
Age	0.0161152	0.0100709	1.60	0.110	1.02	1.00
Logit 2: (Red/White)						
Constant	0.763009	0.525294	1.45	0.146		
Gender						
Male	0.379432	0.327794	1.16	0.247	1.46	0.77
Age	-0.0361957	0.0124037	-2.92	0.004	0.96	0.94
Logit 3: (Light blue/White)						
Constant	-0.640298	0.824108	-0.78	0.437		
Gender						
Male	1.93182	0.504388	3.83	0.000	6.90	2.57
Age	-0.0472456	0.0199261	-2.37	0.018	0.95	0.92
Logit 4: (Green/White)						
Constant	-2.06546	0.663825	-3.11	0.002		
Gender						
Male	0.654506	0.392535	1.67	0.095	1.92	0.89
Age	0.0249973	0.0128226	1.95	0.051	1.03	1.00

Contacts

Minitab World Headquarters

Minitab Inc.

Quality Plaza
1829 Pine Hall Road
State College, PA 16801-3008
USA

Minitab is a global company with subsidiaries and representatives around the world. To find a Minitab partner in your country, visit www.minitab.com/contacts.

Training

Phone: +1 814.238.3280 x3236
Fax: +1 814.238.4383
Email: training@minitab.com
www.minitab.com/training

Training by Minitab™ maximizes your ability to improve quality. It helps you make more effective business decisions by teaching you how to analyze your data with Minitab® Statistical Software and manage your projects using Quality Companion by Minitab™.

Technical Support

Phone: +1 814.231.2682
www.minitab.com/support;
customer.minitab.com to log a question

Our specialists are highly skilled in Minitab software, statistics, quality improvement, and computer systems. Minitab subsidiaries and Independent Local Representatives around the world offer technical support by phone in their local language.

Mentoring

Phone: +1 814.238.3280 x3236
Email: mentoring@minitab.com
www.minitab.com/training/mentoring/

Mentoring by Minitab™ makes it easier to implement cost-saving quality improvement initiatives by providing the statistical support you need, just when you need it. We even begin with a free consultation.