R: The correlation between the observed values of the response variable and the predicted values of the response variable made by the model. R2: The proportion of the variance in the response variable that can be explained by the predictor variables in the regression model.

Is R-squared the same as R?

Coefficient of correlation is “R” value which is given in the summary table in the Regression output. **R square is also called coefficient of determination**. Multiply R times R to get the R square value. In other words Coefficient of Determination is the square of Coefficeint of Correlation.

## What Is R-Squared?

R-squared (R2) is a statistical measure that represents the proportion of the variance for a dependent variable that's explained by an independent variable or variables in a [regression](https://www.investopedia.com/terms/r/regression.asp) model. Whereas correlation explains the strength of the relationship between an independent and dependent variable, R-squared explains to what extent the variance of one variable explains the variance of the second variable. So, if the R2 of a model is 0.50, then approximately half of the observed variation can be explained by the model's inputs.

### **KEY TAKEAWAYS**

* R-Squared is a statistical measure of fit that indicates how much variation of a dependent variable is explained by the independent variable(s) in a regression model.
* In investing, R-squared is generally interpreted as the percentage of a fund or security's movements that can be explained by movements in a benchmark index.
* An R-squared of 100% means that all movements of a security (or other dependent variables) are completely explained by movements in the index (or the independent variable(s) you are interested in).

## Formula for R-Squared

R2=1−(Total Variation/Unexplained Variation)​​)

The actual calculation of R-squared requires several steps. This includes taking the data points (observations) of dependent and independent variables and finding the [line of best fit](https://www.investopedia.com/terms/l/line-of-best-fit.asp), often from a regression model. From there you would calculate predicted values, subtract actual values and square the results. This yields a list of errors squared, which is then summed and equals the unexplained variance.

To calculate the total variance, you would subtract the average actual value from each of the actual values, square the results and sum them. From there, divide the first sum of errors (unexplained variance) by the second sum (total variance), subtract the result from one, and you have the R-squared.

## What R-Squared Can Tell You

In investing, R-squared is generally interpreted as the percentage of a fund or security's movements that can be explained by movements in a benchmark index. For example, an R-squared for a [fixed-income security](https://www.investopedia.com/terms/f/fixed-incomesecurity.asp) versus a bond index identifies the security's proportion of price movement that is predictable based on a price movement of the index.

The same can be applied to a stock versus the S&P 500 index, or any other relevant index. It may also be known as the [coefficient of determination](https://www.investopedia.com/terms/c/coefficient-of-determination.asp).

R-squared values range from 0 to 1 and are commonly stated as percentages from 0% to 100%. An R-squared of 100% means that all movements of a security (or another dependent variable) are completely explained by movements in the index (or the independent variable(s) you are interested in).

In investing, a high R-squared, between 85% and 100%, indicates the stock or fund's performance moves relatively in line with the index. A fund with a low R-squared, at 70% or less, indicates the security does not generally follow the movements of the index. A higher R-squared value will indicate a more useful [beta](https://www.investopedia.com/terms/b/beta.asp) figure. For example, if a stock or fund has an R-squared value of close to 100%, but has a beta below 1, it is most likely offering higher [risk-adjusted returns](https://www.investopedia.com/terms/r/riskadjustedreturn.asp).

## R-Squared vs. Adjusted R-Squared

R-Squared only works as intended in a simple linear regression model with one explanatory variable. With a multiple regression made up of several independent variables, the R-Squared must be adjusted.

The adjusted R-squared compares the descriptive power of regression models that include diverse numbers of predictors. Every predictor added to a model increases R-squared and never decreases it. Thus, a model with more terms may seem to have a better fit just for the fact that it has more terms, while the adjusted R-squared compensates for the addition of variables and only increases if the new term enhances the model above what would be obtained by probability and decreases when a predictor enhances the model less than what is predicted by chance.

In an [overfitting](https://www.investopedia.com/terms/o/overfitting.asp) condition, an incorrectly high value of R-squared is obtained, even when the model actually has a decreased ability to predict. [This is not the case](https://www.investopedia.com/ask/answers/012615/whats-difference-between-rsquared-and-adjusted-rsquared.asp) with the adjusted R-squared.

## R-Squared vs. Beta

[Beta](https://www.investopedia.com/terms/b/beta.asp) and R-squared are two related, but different, measures of correlation but the beta is a measure of relative riskiness. A mutual fund with a high R-squared correlates highly with a [benchmark](https://www.investopedia.com/terms/b/benchmark.asp). If the beta is also high, it may produce higher returns than the benchmark, particularly in [bull markets](https://www.investopedia.com/terms/b/bullmarket.asp). R-squared measures how closely each change in the price of an asset is correlated to a benchmark.

Beta measures how large those price changes are relative to a benchmark. Used together, R-squared and beta give investors a thorough picture of the performance of asset managers. A beta of exactly 1.0 means that the risk (volatility) of the asset is identical to that of its benchmark. Essentially, R-squared is a statistical analysis technique for the practical use and trustworthiness of betas of securities.

## Limitations of R-Squared

R-squared will give you an estimate of the relationship between movements of a dependent variable based on an independent variable's movements. It doesn't tell you whether your chosen model is good or bad, nor will it tell you whether the data and predictions are biased. A high or low R-square isn't necessarily good or bad, as it doesn't convey the reliability of the model, nor whether you've chosen the right regression. You can get a low R-squared for a good model, or a high R-square for a poorly fitted model, and vice versa.

## What Is a Good R-Squared Value?

What qualifies as a “good” R-Squared value will depend on the context. In some fields, such as the social sciences, even a relatively low R-Squared such as 0.5 could be considered relatively strong. In other fields, the standards for a good R-Squared reading can be much higher, such as 0.9 or above. In finance, an R-Squared above 0.7 would generally be seen as showing a high level of correlation, whereas a measure below 0.4 would show a low correlation. This is not a hard rule, however, and will depend on the specific analysis.

## What Does an R-Squared Value of 0.9 Mean?

Essentially, an R-Squared value of 0.9 would indicate that 90% of the variance of the dependent variable being studied is explained by the variance of the independent variable. For instance, if a mutual fund has an R-Squared value of 0.9 relative to its benchmark, that would indicate that 90% of the variance of the fund is explained by the variance of its benchmark index.

## Is a Higher R-Squared Better?

Here again, it depends on the context. Suppose you are searching for an [index fund](https://www.investopedia.com/terms/i/indexfund.asp) that will track a specific index as closely as possible. In that scenario, you would want the fund’s R-Squared to be as high as possible since its goal is to match—rather than exceed—the index. If on the other hand, you are looking for actively managed funds, a high R-Squared might be seen as a bad sign, indicating that the funds’ managers are not adding sufficient value relative to their benchmarks.

[What is Regression? Definition, Calculation, and Example](https://www.investopedia.com/terms/r/regression.asp)

Regression is a statistical measurement that attempts to determine the strength of the relationship between one dependent variable and a series of other variables.

[more](https://www.investopedia.com/terms/r/regression.asp)

[Multiple Linear Regression (MLR) Definition, Formula, and Example](https://www.investopedia.com/terms/m/mlr.asp)

Multiple linear regression (MLR) is a statistical technique that uses several explanatory variables to predict the outcome of a response variable.

[more](https://www.investopedia.com/terms/m/mlr.asp)

[Beta: Definition, Calculation, and Explanation for Investors](https://www.investopedia.com/terms/b/beta.asp)

Beta is a measure of the volatility, or systematic risk, of a security or portfolio in comparison to the market as a whole. It is used in the capital asset pricing model.

[more](https://www.investopedia.com/terms/b/beta.asp)

[What Is Negative Correlation?](https://www.investopedia.com/terms/n/negative-correlation.asp)

Negative correlation is a relationship between two variables in which one variable increases as the other decreases, and vice versa.

[more](https://www.investopedia.com/terms/n/negative-correlation.asp)

[Heteroscedasticity Definition: Simple Meaning and Types Explained](https://www.investopedia.com/terms/h/heteroskedasticity.asp)

In statistics, heteroskedasticity happens when the standard deviations of a variable, monitored over a specific amount of time, are nonconstant.

[more](https://www.investopedia.com/terms/h/heteroskedasticity.asp)

[Index Hugger](https://www.investopedia.com/terms/i/indexhugger.asp)

An index hugger is a managed mutual fund that tends to perform much like a benchmark index.

[more](https://www.investopedia.com/terms/i/indexhugger.asp)

## What Is a Regression?

Regression is a statistical method used in finance, investing, and other disciplines that attempts to determine the strength and character of the relationship between one dependent variable (usually denoted by Y) and a series of other variables (known as independent variables).

Also called simple regression or ordinary least squares (OLS), linear regression is the most common form of this technique. Linear regression establishes the [linear relationship](https://www.investopedia.com/terms/l/linearrelationship.asp) between two variables based on a [line of best fit](https://www.investopedia.com/terms/l/line-of-best-fit.asp). Linear regression is thus graphically depicted using a straight line with the slope defining how the change in one variable impacts a change in the other. The y-intercept of a linear regression relationship represents the value of one variable when the value of the other is zero. [Non-linear regression](https://www.investopedia.com/terms/n/nonlinear-regression.asp) models also exist, but are far more complex.

Regression analysis is a powerful tool for uncovering the associations between variables observed in data, but cannot easily indicate causation. It is used in several contexts in business, finance, and economics. For instance, it is used to help investment managers value assets and understand the relationships between factors such as [commodity prices](https://www.investopedia.com/terms/c/commodity.asp) and the stocks of businesses dealing in those commodities.

Regression as a statistical technique should not be confused with the concept of regression to the mean ([mean reversion](https://www.investopedia.com/terms/m/meanreversion.asp)).

### **KEY TAKEAWAYS**

* A regression is a statistical technique that relates a dependent variable to one or more independent (explanatory) variables.
* A regression model is able to show whether changes observed in the dependent variable are associated with changes in one or more of the explanatory variables.
* It does this by essentially fitting a best-fit line and seeing how the data is dispersed around this line.
* Regression helps economists and financial analysts in things ranging from asset valuation to making predictions.
* In order for regression results to be properly interpreted, several assumptions about the data and the model itself must hold.

## Understanding Regression

Regression captures the correlation between variables observed in a data set, and quantifies whether those correlations are [statistically significant](https://www.investopedia.com/terms/s/statistical-significance.asp) or not.

The two basic types of regression are simple linear regression and [multiple linear regression](https://www.investopedia.com/terms/m/mlr.asp), although there are non-linear regression methods for more complicated data and analysis. Simple linear regression uses one independent variable to explain or predict the outcome of the dependent variable Y, while multiple linear regression uses two or more independent variables to predict the outcome (while holding all others constant).

Regression can help finance and investment professionals as well as professionals in other businesses. Regression can also help predict sales for a company based on weather, previous sales, GDP growth, or other types of conditions. The [capital asset pricing model](https://www.investopedia.com/terms/c/capm.asp) (CAPM) is an often-used regression model in finance for pricing assets and discovering costs of capital.

### **Regression and Econometrics**

[Econometrics](https://www.investopedia.com/terms/e/econometrics.asp) is a set of statistical techniques used to analyze data in finance and economics. An example of the application of econometrics is to study the income effect using observable data. An economist may, for example, hypothesize that as a person increases their income their spending will also increase.

If the data show that such an association is present, a regression analysis can then be conducted to understand the strength of the relationship between income and consumption and whether or not that relationship is statistically significant—that is, it appears to be unlikely that it is due to chance alone.

Note that you can have several explanatory variables in your analysis—for example, changes to GDP and inflation in addition to unemployment in explaining stock market prices. When more than one explanatory variable is used, it is referred to as [multiple linear regression](https://www.investopedia.com/terms/m/mlr.asp). This is the most commonly used tool in econometrics.

Econometrics is sometimes criticized for relying too heavily on the interpretation of regression output without linking it to economic theory or looking for causal mechanisms. It is crucial that the findings revealed in the data are able to be adequately explained by a theory, even if that means developing your own theory of the underlying processes.

## Calculating Regression

Linear regression models often use a least-squares approach to determine the line of best fit. The least-squares technique is determined by minimizing the [sum of squares](https://www.investopedia.com/terms/s/sum-of-squares.asp) created by a mathematical function. A square is, in turn, determined by squaring the distance between a data point and the regression line or mean value of the data set.

Once this process has been completed (usually done today with software), a regression model is constructed. The general form of each type of regression model is:

**Simple linear regression:**

***Y = a + bX + u***

***​Y=a+bX+u​***

**Multiple linear regression:**

Y = a + b\_1X\_1 + b\_2X\_2 + b\_3X\_3 + ... + b\_tX\_t + u

*Y*=*a*+*b*1​*X*1​+*b*2​*X*2​+*b*3​*X*3​+...+*bt*​*Xt*​+*u*

**where:***Y*=The dependent variable you are trying to predictor explain*X*=The explanatory (independent) variable(s) you are using to predict or associate with Y*a*=The y-intercept*b*=(beta coefficient) is the slope of the explanatoryvariable()*u*=The regression residual or error term​

**Example of How Regression Analysis Is Used in Finance**

Regression is often used to determine how many specific factors such as the price of a commodity, interest rates, particular industries, or sectors influence the price movement of an asset. The aforementioned CAPM is based on regression, and it is utilized to project the expected returns for stocks and to generate costs of capital. A stock's returns are regressed against the returns of a broader index, such as the S&P 500, to generate a [beta](https://www.investopedia.com/terms/b/beta.asp) for the particular stock.

Beta is the stock's risk in relation to the market or index and is reflected as the slope in the CAPM model. The return for the stock in question would be the dependent variable Y, while the independent variable X would be the market risk premium.

Additional variables such as the market capitalization of a stock, valuation ratios, and recent returns can be added to the CAPM model to get better estimates for returns. These additional factors are known as the Fama-French factors, named after the professors who developed the multiple linear regression model to better explain asset returns.1

Why Is It Called Regression?

Although there is some debate about the origins of the name, the statistical technique described above most likely was termed "regression" by Sir Francis Galton in the 19th century to describe the statistical feature of biological data (such as heights of people in a population) to regress to some mean level. In other words, while there are shorter and taller people, only outliers are very tall or short, and most people cluster somewhere around (or "regress" to) the average.2

What Is the Purpose of Regression?

In statistical analysis, regression is used to identify the associations between variables occurring in some data. It can show both the magnitude of such an association and also determine its statistical significance (i.e., whether or not the association is likely due to chance). Regression is a powerful tool for statistical inference and has also been used to try to predict future outcomes based on past observations.

How Do You Interpret a Regression Model?

A regression model output may be in the form of Y = 1.0 + (3.2)*X1*- 2.0(*X2*) + 0.21.

Here we have a multiple linear regression that relates some variable Y with two explanatory variables X1 and X2. We would interpret the model as the value of Y changes by 3.2x for every one-unit change in X1 (if X1 goes up by 2, Y goes up by 6.4, etc.) *holding all else constant* (all else equal). That means controlling for X2, X1 has this observed relationship. Likewise, holding X1 constant, every one unit increase in X2 is associated with a 2x *decrease*in Y. We can also note the y-intercept of 1.0, meaning that Y = 1 when X1 and X2 are both zero. The [error term](https://www.investopedia.com/terms/e/errorterm.asp) (residual) is 0.21.

What Are the Assumptions That Must Hold for Regression Models?

In order to properly interpret the output of a regression model, the following main assumptions about the underlying data process of what you analyzing must hold:

* The relationship between variables is linear
* [Homoskedasticity](https://www.investopedia.com/search?q=heteroskedasticity), or that the variance of the variables and error term must remain constant
* All explanatory variables are independent of one another
* All variables are [normally-distributed](https://www.investopedia.com/terms/n/normaldistribution.asp)

## What Is Multiple Linear Regression (MLR)?

Multiple linear regression (MLR), also known simply as multiple regression, is a statistical technique that uses several explanatory variables to predict the outcome of a response variable. The goal of multiple linear regression is to model the [linear relationship](https://www.investopedia.com/terms/l/linearrelationship.asp) between the explanatory (independent) variables and response (dependent) variables. In essence, multiple regression is the extension of ordinary least-squares (OLS) [regression](https://www.investopedia.com/terms/r/regression.asp) because it involves more than one explanatory variable.

### **KEY TAKEAWAYS**

* Multiple linear regression (MLR), also known simply as multiple regression, is a statistical technique that uses several explanatory variables to predict the outcome of a response variable.
* Multiple regression is an extension of linear (OLS) regression that uses just one explanatory variable.
* MLR is used extensively in econometrics and financial inference.

## What Multiple Linear Regression Can Tell You

Simple linear regression is a function that allows an analyst or statistician to make predictions about one variable based on the information that is known about another variable. Linear regression can only be used when one has two continuous variables—an independent variable and a dependent variable. The independent variable is the parameter that is used to calculate the dependent variable or outcome. A multiple regression model extends to several explanatory variables.

The multiple regression model is based on the following assumptions:

* There is a [linear relationship](https://www.investopedia.com/terms/l/linearrelationship.asp) between the dependent variables and the independent variables
* The independent variables are not too highly [correlated](https://www.investopedia.com/terms/c/correlation.asp) with each other
* yi observations are selected independently and randomly from the population
* Residuals should be [normally distributed](https://www.investopedia.com/terms/n/normaldistribution.asp) with a mean of 0 and [variance](https://www.investopedia.com/terms/v/variance.asp) *σ*

The [coefficient of determination](https://www.investopedia.com/terms/c/coefficient-of-determination.asp) (R-squared) is a statistical metric that is used to measure how much of the variation in outcome can be explained by the variation in the independent variables. R2 always increases as more predictors are added to the MLR model, even though the predictors may not be related to the outcome variable.

R2 by itself can't thus be used to identify which predictors should be included in a model and which should be excluded. R2 can only be between 0 and 1, where 0 indicates that the outcome cannot be predicted by any of the independent variables and 1 indicates that the outcome can be predicted without error from the independent variables.

When interpreting the results of multiple regression, beta coefficients are valid while holding all other variables constant ("all else equal"). The output from a multiple regression can be displayed horizontally as an equation, or vertically in table form.

## The Difference Between Linear and Multiple Regression

[Ordinary linear squares](https://www.investopedia.com/terms/l/least-squares-method.asp) (OLS) regression compares the response of a dependent variable given a change in some explanatory variables. However, a dependent variable is rarely explained by only one variable. In this case, an analyst uses multiple regression, which attempts to explain a dependent variable using more than one independent variable. Multiple regressions can be linear and nonlinear.

Multiple regressions are based on the assumption that there is a linear relationship between both the dependent and independent variables. It also assumes no major correlation between the independent variables.

## What Makes a Multiple Regression Multiple?

A multiple regression considers the effect of more than one explanatory variable on some outcome of interest. It evaluates the relative effect of these explanatory, or independent, variables on the dependent variable when holding all the other variables in the model constant.

## Why Would One Use a Multiple Regression Over a Simple OLS Regression?

A dependent variable is rarely explained by only one variable. In such cases, an analyst uses multiple regression, which attempts to explain a dependent variable using more than one independent variable. The model, however, assumes that there are no major correlations between the independent variables.

## What Does It Mean for a Multiple Regression to Be Linear?

In multiple linear regression, the model calculates the [line of best fit](https://www.investopedia.com/terms/l/line-of-best-fit.asp) that minimizes the variances of each of the variables included as it relates to the dependent variable. Because it fits a line, it is a linear model. There are also non-linear regression models involving multiple variables, such as logistic regression, quadratic regression, and probit models.

* Both linear and nonlinear regression predict Y responses from an X variable (or variables).
* Nonlinear regression is a curved function of an X variable (or variables) that is used to predict a Y variable
* Nonlinear regression can show a prediction of population growth over time.

# **What Do Correlation Coefficients Positive, Negative, and Zero Mean?**

By

[STEVEN NICKOLAS](https://www.investopedia.com/contributors/53895/)

Updated May 31, 2021

Reviewed by

[JULIUS MANSA](https://www.investopedia.com/julius-mansa-4799781)

Fact checked by

[KATRINA MUNICHIELLO](https://www.investopedia.com/katrina-munichiello-5078531)

Correlation coefficients are indicators of the strength of the [linear relationship](https://www.investopedia.com/terms/l/linearrelationship.asp) between two different variables, x and y. A linear correlation coefficient that is greater than zero indicates a positive relationship. A value that is less than zero signifies a negative relationship. Finally, a value of zero indicates no relationship between the two variables x and y.

* Correlation coefficients are used to measure the strength of the linear relationship between two variables.
* A correlation coefficient greater than zero indicates a positive relationship while a value less than zero signifies a negative relationship.
* A value of zero indicates no relationship between the two variables being compared.
* A negative correlation, or inverse correlation, is a key concept in the creation of diversified portfolios that can better withstand portfolio volatility.
* Calculating the correlation coefficient is time-consuming, so data are often plugged into a calculator, computer, or statistics program to find the coefficient.

## Understanding Correlation

The correlation coefficient (ρ) is a measure that determines the degree to which the movement of two different variables is associated. The most common correlation coefficient, generated by the Pearson product-moment correlation, is used to measure the [linear relationship](https://www.investopedia.com/terms/l/linearrelationship.asp) between two variables. However, in a non-linear relationship, this correlation coefficient may not always be a suitable measure of dependence.

The possible range of values for the correlation coefficient is -1.0 to 1.0. In other words, the values cannot exceed 1.0 or be less than -1.0. A correlation of -1.0 indicates a perfect [negative correlation](https://www.investopedia.com/terms/n/negative-correlation.asp), and a correlation of 1.0 indicates a perfect [positive correlation](https://www.investopedia.com/terms/p/positive-correlation.asp). If the correlation coefficient is greater than zero, it is a positive relationship. Conversely, if the value is less than zero, it is a negative relationship. A value of zero indicates that there is no relationship between the two variables.

### **Correlation and the Financial Markets**

In the [financial markets](https://www.investopedia.com/terms/f/financial-market.asp), the correlation coefficient is used to measure the [correlation between two securities](https://www.investopedia.com/articles/financial-theory/11/calculating-covariance.asp). For example, when two stocks move in the same direction, the correlation coefficient is positive. Conversely, when two [stocks](https://www.investopedia.com/terms/s/stock.asp) move in opposite directions, the correlation coefficient is negative.

If the correlation coefficient of two variables is zero, there is no linear relationship between the variables. However, this is only for a linear relationship. It is possible that the variables have a strong curvilinear relationship. When the value of ρ is close to zero, generally between -0.1 and +0.1, the variables are said to have no linear relationship (or a very weak linear relationship).

## Calculating ρ

The [covariance](https://www.investopedia.com/terms/c/covariance.asp) of the two variables in question must be calculated before the correlation can be determined. Next, each variable's [standard deviation](https://www.investopedia.com/terms/s/standarddeviation.asp) is required. The correlation coefficient is determined by dividing the covariance by the product of the two variables' standard deviations.

Standard deviation is a measure of the [dispersion](https://www.investopedia.com/terms/d/dispersion.asp) of data from its average. Covariance is a measure of how two variables change together. However, its magnitude is unbounded, so it is difficult to interpret. The normalized version of the statistic is calculated by dividing covariance by the product of the two standard deviations. This is the correlation coefficient.

# **What Is the Pearson Coefficient? Definition, Benefits, and History**

By

[WILL KENTON](https://www.investopedia.com/contributors/53661/)

Updated May 06, 2022

Reviewed by

[THOMAS BROCK](https://www.investopedia.com/thomas-j-brock-4799774)

Fact checked by

[TIMOTHY LI](https://www.investopedia.com/timothy-li-5198835)

## What Is the Pearson Coefficient?

The Pearson coefficient is a type of [correlation coefficient](https://www.investopedia.com/terms/c/correlationcoefficient.asp) that represents the relationship between two variables that are measured on the same interval or ratio scale. The Pearson coefficient is a measure of the strength of the association between two continuous variables.

## Understanding the Pearson Coefficient

To find the Pearson coefficient, also referred to as the Pearson correlation coefficient or the Pearson product-moment correlation coefficient, the two variables are placed on a scatter plot. The variables are denoted as X and Y. There must be some linearity for the coefficient to be calculated; a scatter plot not depicting any resemblance to a linear relationship will be useless. The closer the resemblance to a straight line of the scatter plot, the higher the strength of association. Numerically, the Pearson coefficient is represented the same way as a correlation coefficient that is used in linear regression, ranging from -1 to +1. A value of +1 is the result of a perfect positive relationship between two or more variables. Positive correlations indicate that both variables move in the same direction. Conversely, a value of -1 represents a perfect negative relationship. Negative correlations indicate that as one variable increases, the other decreases; they are inversely related. A zero indicates no correlation.

### **KEY TAKEAWAYS**

* The Pearson coefficient is a mathematical correlation coefficient representing the relationship between two variables, denoted as X and Y.
* Pearson coefficients range from +1 to -1, with +1 representing a positive correlation, -1 representing a negative correlation, and 0 representing no relationship.
* The Pearson coefficient shows correlation, not causation.
* English mathematician and statistician Karl Pearson is credited for developing many statistical techniques, including the Pearson coefficient, the chi-squared test, p-value, and linear regression.

[Correlation: What It Means in Finance and the Formula for Calculating It](https://www.investopedia.com/terms/c/correlation.asp)

Correlation is a statistical measure of how two securities move in relation to each other.

Interpreting the result from a Hausman test is fairly straightforward: **if the p-value is small (less than 0.05), reject the null hypothesis**. The problem comes with the fact that many versions of the test — with different hypothesis and possible conclusions — exist.

A large and significant Hausman statistic means **a large and significant difference**, and so you reject the null that the two methods are OK in favour of the alternative hypothesis that one is OK (fixed effects) and one isn't (random effects).

How do you test for Hausman test?

One performs a Hausman test by comparing the Hausman statistic to a critical value obtained from its sampling distribution, and rejecting the null hypothesis of correct specification if the Hausman statistic exceeds its critical value.

How do you read a Hausman test in eviews?

Re: hausman test  
  
Fixed effects allows for endogeneity. The Hausman test checks this assumption. Thus it follows that the **you can test the null hypothesis that there is no endogeneity in the random effects model and if it is found is thus rejected**.

How do you interpret random and fixed effects?

A fixed-effects model supports prediction about only the levels/categories of features used for training. A random-effects model, by contrast, allows predicting something about the population from which the sample is drawn.

Which variables are included in Hausman test?

For running the Hausman test, you should use the complete model including the **control variables**. It is VERY important that you include control variables, and you have to pay particular attention to the fact that the a Hausman test yields different results DEPENDING on the control variables that you account for.

How do you detect endogeneity problems?

We can test for the endogeneity of suspect independent variables **using a Hausman test**. H0: the regressor is exogenous; H1: it is endogenous. Thus, a higher value of the test statistic indicates a more serious endogeneity problem.

What are the three sources of endogeneity?

In summary, each of the three sources of endogeneity bias (i.e., **measurement error, omitted variables, and simultaneity**) leads to questionable causal inferences.

Which is an example of endogeneity?

For example, **if they think a customer will buy even without a coupon, they did not send it or if they think a person might buy, they sent them more coupons**.

How do you fix endogeneity?

The best way to deal with endogeneity concerns is through **instrumental variables (IV) techniques**. The most common IV estimator is Two Stage Least Squares (TSLS). IV estimation is intuitively appealing, and relatively simple to implement on a technical level.22 Sept 2009

Endogeneity is a first-moment problem, while heteroskedasticity is a second-moment problem. where σ2 is a constant number. would imply Var(ui|xi)=σ2.4 Sept 2015

What is the difference between endogeneity and Multicollinearity?

For my under-standing, **multicollinearity is a correlation of an independent variable with another independent variable.** **Endogeneity is the correlation of an independent variable with the error term**.7 Sept 2015

How do you test for fixed effects?

Tests of fixed effects are typically done with either **Wald or likelihood ratio (LRT) tests**. With the assumptions of asymptotic distributions and independent predictors, Wald and LRT tests are equivalent.

**The fixed-effects model is the appropriate model when the number of studies is small**. Random-effects models are appropriate when the number of studies is large enough, that is, enough studies to support generalization inferences beyond the included studies.

What is random effect model in statistics?

What are fixed and random effects in panel data?

Random? Panel data models examine cross-sectional (group) and/or time-series (time) effects. These effects may be fixed and/or random. Fixed effects assume that individual group/time have different intercept in the regression equation, while random effects hypothesize individual group/time have different disturbance.

What is a good p-value for interaction?

The p-value for the test for a significant interaction between factors is 0.562. This p-value is **greater than 5% (α)**, therefore we fail to reject the null hypothesis.1 May 2021

Should I use fixed or random effects?

**As the size of the dataset increases, and the variance of the fixed-effects estimates falls, there is less support for the random-effects model**. In larger datasets, correlation of greater than 0.2 to 0.3 between the independent variable and unit effects leads to a preference for the fixed-effects model.

How do you interpret p trends?

**The p-value represents a probability of the error when expecting, that the trend differs from zero** (i.e. probability, that there is no time change and the value is based on random fluctuations only).

What are fixed effects variables?

Fixed effects are **variables that are constant across individuals**; these variables, like age, sex, or ethnicity, don't change or change at a constant rate over time. They have fixed effects; in other words, any change they cause to an individual is the same.

What is panel data fixed effects?

Fixed effects is **a statistical regression model in which the intercept of the regression model is allowed to vary freely across individuals or groups**. It is often applied to panel data in order to control for any individual-specific attributes that do not vary across time.

Why is random effects more efficient?

Additionally, random effects is estimated using GLS while fixed effects is estimated using OLS and as such, **random Page 3 effects estimates will generally have smaller variances**. As a result, the random effects model is more efficient.

What does p-value of 0.05 tell us?

**P > 0.05 is the probability that the null hypothesis is true**. 1 minus the P value is the probability that the alternative hypothesis is true. A statistically significant test result (P ≤ 0.05) means that the test hypothesis is false or should be rejected. A P value greater than 0.05 means that no effect was observed.10 Aug 2016

What does panel data tell us?

Panel data can **detect and measure statistical effects that pure time series or cross-sectional data can't**. Panel data can minimize estimation biases that may arise from aggregating groups into a single time series.29 Nov 2019

What is the Hausman test for endogeneity?

What is the Hausman Test? The Hausman Test (also called the Hausman specification test) **detects endogenous regressors (predictor variables) in a regression model**. Endogenous variables have values that are determined by other variables in the system.

Problem: One of the major problems faced during the panel data analysis was **data management**. If the data is not arranged properly then it is very difficult to get the regression results. Even if the results are obtained, they will not be robust.

Heteroskedasticity refers to **a situation where the variance of the residuals is unequal over a range of measured values**. If heteroskedasticity exists, the population used in the regression contains unequal variance, the analysis results may be invalid.3 Nov 20

What are the two ways we can check for heteroskedasticity?

Residual Plots  
  
One informal way of detecting heteroskedasticity is by creating a residual plot where you plot the least squares residuals against the explanatory variable or ˆy if it's a multiple regression. If there is an evident pattern in the plot, then heteroskedasticity is present.6 Jun 2016

Is autocorrelation the same as Multicollinearity?

Autocorrelation is used for signals or time series. Autocorrelation is the correlation of the signal with a delayed copy of itself. Multicollinearity, which should be checked during MLR, is a phenomenon in which at least two independent variables are linearly correlated (one can be predicted from the other).11 Nov 2020

How do you interpret homoscedasticity?

You can tell if a regression is homoskedastic by **looking at the ratio between the largest variance and the smallest variance**. If the ratio is 1.5 or smaller, then the regression is homoskedastic.

What is the difference between homoscedasticity and heteroscedasticity?

Simply put, homoscedasticity means “having the same scatter.” For it to exist in a set of data, the points must be about the same distance from the line, as shown in the picture above. The opposite is heteroscedasticity (“different scatter”), where points are at widely varying distances from the regression line.

What causes heteroskedasticity?

Heteroscedasticity is mainly due to **the presence of outlier in the data**. Outlier in Heteroscedasticity means that the observations that are either small or large with respect to the other observations are present in the sample. Heteroscedasticity is also caused due to omission of variables from the model.

**An autocorrelation of +1 represents a perfect positive correlation, while an autocorrelation of negative 1 represents a perfect negative correlation**. Technical analysts can use autocorrelation to measure how much influence past prices for a security have on its future price.

Empirical analyses in social science frequently confront quantitative data that are clustered or grouped. To account for group-level variation and improve model fit, researchers will commonly specify either a fixed- or random-effects model. But current advice on which approach should be preferred, and under what conditions, remains vague and sometimes contradictory. This study performs a series of Monte Carlo simulations to evaluate the total error due to bias and variance in the inferences of each model, for typical sizes and types of datasets encountered in applied research. The results offer a typology of dataset characteristics to help researchers choose a preferred model.

P value of test statistic

The p value is **a number, calculated from a statistical test, that describes how likely you are to have found a particular set of observations if the null hypothesis were true**. P values are used in hypothesis testing to help decide whether to reject the null hypothesis.16 Jul 2020

## What exactly is a p value?

The **p value**, or probability value, tells you how likely it is that your data could have occurred under the null hypothesis. It does this by calculating the likelihood of your [**test statistic**](https://www.scribbr.com/statistics/test-statistic/), which is the number calculated by a [statistical test](https://www.scribbr.com/statistics/statistical-tests/) using your data.

The p value tells you how often you would expect to see a test statistic as extreme or more extreme than the one calculated by your statistical test if the null hypothesis of that test was true. The p value gets smaller as the test statistic calculated from your data gets further away from the [range](https://www.scribbr.com/statistics/range/) of test statistics predicted by the null hypothesis.

The p value is a proportion: if your p value is 0.05, that means that 5% of the time you would see a test statistic at least as extreme as the one you found if the null hypothesis was true.

Example: Test statistic and p valueIf the mice live equally long on either diet, then the test statistic from your t test will closely match the test statistic from the null hypothesis (that there is no difference between groups), and the resulting p value will be close to 1. It likely won’t reach exactly 1, because in real life the groups will probably not be perfectly equal.

If, however, there is an average difference in longevity between the two groups, then your test statistic will move further away from the values predicted by the null hypothesis, and the p value will get smaller. The p value will never reach zero, because there’s always a possibility, even if extremely unlikely, that the patterns in your data occurred by chance.

**How do you calculate the *p*value?**

*P*values are usually automatically calculated by your statistical program (R, SPSS, etc.).

You can also find tables for estimating the *p*value of your test statistic online. These tables show, based on the test statistic and [**degrees of freedom**](https://www.scribbr.com/statistics/degrees-of-freedom/) (number of observations minus number of independent variables) of your test, how [frequently](https://www.scribbr.com/frequently-asked-questions/main-types-of-descriptive-statistics/)you would expect to see that test statistic under the null hypothesis.

The calculation of the *p*value depends on the statistical test you are using to test your [hypothesis](https://www.scribbr.com/methodology/hypothesis/):

* Different statistical tests have different assumptions and generate different test statistics. You should choose the [statistical test](https://www.scribbr.com/statistics/statistical-tests/) that best fits your data and matches the effect or relationship you want to test.
* The number of [independent variables](https://www.scribbr.com/methodology/independent-and-dependent-variables/) you include in your test changes how large or small the test statistic needs to be to generate the same *p*value.

Example: Choosing a statistical testIf you are comparing only two different diets, then a two-sample *t*test is a good way to compare the groups. To compare three different diets, use an [ANOVA](http://www.sthda.com/english/wiki/one-way-anova-test-in-r)instead – doing multiple pairwise comparisons will result in artificially low *p*values and lets you overestimate the significance of the difference between groups.

No matter what test you use, the *p*value always describes the same thing: how often you can expect to see a test statistic as extreme or more extreme than the one calculated from your test.

***P*values and statistical significance**

*P*values are most often used by researchers to say whether a certain pattern they have measured is statistically significant.

[**Statistical significance**](https://www.scribbr.com/statistics/statistical-significance/) is another way of saying that the *p*value of a statistical test is small enough to reject the null hypothesis of the test.

How small is small enough? The most common threshold is *p <* 0.05; that is, when you would expect to find a test statistic as extreme as the one calculated by your test only 5% of the time. But the threshold depends on your field of study – some fields prefer thresholds of 0.01, or even 0.001.

The threshold value for determining statistical significance is also known as the alpha value.

Example: Statistical significanceYour comparison of the two mouse diets results in a *p*value of less than 0.01, below your alpha value of 0.05; therefore you determine that there is a statistically significant difference between the two diets.

**Reporting *p*values**

*P*values of statistical tests are usually reported in the[results section](https://www.scribbr.com/dissertation/results/) of a [research paper](https://www.scribbr.com/category/research-paper/), along with the key information needed for readers to put the *p*values in context – for example, [correlation coefficient](https://www.scribbr.com/statistics/correlation-coefficient/) in a [linear regression](https://www.scribbr.com/statistics/simple-linear-regression/), or the average difference between treatment groups in a *t*-test.

Example: Reporting the resultsIn our comparison of mouse diet A and mouse diet B, we found that the lifespan on diet A (*M* = 2.1 years; *SD* = 0.12) was significantly shorter than the lifespan on diet B (*M*= 2.6 years; *SD*= 0.1), with an average difference of 6 months (*t*(80) = -12.75; *p <* 0.01).

**Caution when using *p*values**

*P*values are often interpreted as your risk of rejecting the [null hypothesis](https://www.scribbr.com/statistics/null-and-alternative-hypotheses/) of your test when the null hypothesis is actually true.

In reality, the risk of rejecting the null hypothesis is often higher than the *p*value, especially when looking at a single study or when using small sample sizes. This is because the smaller your frame of reference, the greater the chance that you stumble across a statistically significant pattern completely by accident.

*P*values are also often interpreted as supporting or refuting the alternative hypothesis. This is not the case.**The *p*value can only tell you whether or not the null hypothesis is supported.** It cannot tell you whether your alternative hypothesis is true, or why.