**How to interpret r summary output**

Et bilde som inneholder tekst

Automatisk generert beskrivelse

X-variables

* In this case, height and width.
* The different lines are comparing each models as if it is better or worse to add the independent variable:
  + The **height** line compares the multiple regression to the simple regression. It compares:
    - Weight = y-intercept + slope1 x **height** + slope 2 x **width**
    - Weight y-intercept + ~~slope1 x~~ **~~height~~** + slope 2 x **width**
    - by looking at the variable’s p-value, we see that using **height** and **width** isn’t significantly better than using **width** alone to predict **weight**.
  + The **width** line also compares the multiple regression to the simple regression. It compares:
    - Weight = y-intercept + slope1 x **height** + slope 2 x **width**
    - Weight = y-intercept + slope1 x **height** + ~~slope 2 x~~ **~~width~~**
    - By looking at the variable’s p-value, we see that using **height** and **width** is significantly better than using **height** alone to predict **weight**.
  + In summary, using **height** and **width** to predict **width** is good, but if we wanted to save time by not measuring height, we could just use the **width** to predict **weight**.

Estimate

* Used for predicting your response variable (the dependent variable).
* A one cm increase in height will result in a 4,825 increase in grams of the weight.
* A one cm increase in width will result in a 178,523 increase in grams of the weight.
* These numbers should be positive but sometimes, when doing a multiple linear regression there’s multicollinearity (the x-variables are not independent) then the estimates might be drastically different from what you expected and the interpretation might not make much sense.

Std. Error

* Average amount that the estimates vary from the actual value.
* Ideally, you would want a **lower** number relative to the coefficients.
* Can use this number to calculate the confidence interval of your estimate.
* Calculating standard error:

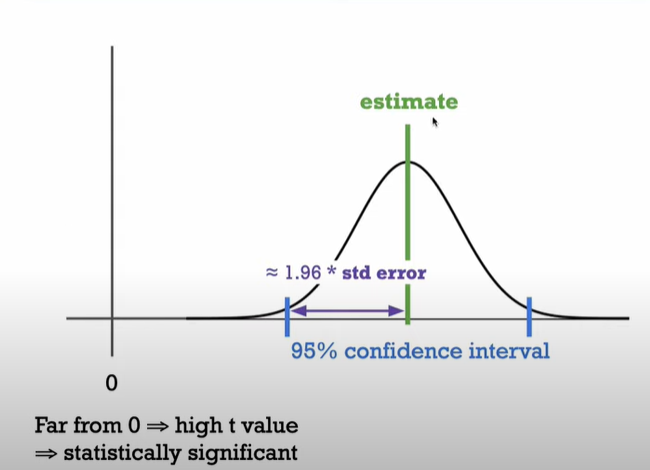
t-value

* Measure of how many standard deviations there are between the estimate and zero.
* The estimate divided by the standard error.
* In general, if the t-value has a really **high** magnitude the coefficient is going to be statistically significant.

Pr(>|t|)

* Gives us a p-value for the t-test.
* If the number is **smaller than the alpha**, it means that it’s unlikely that the relationship between the y-variable and one of the coefficients was due to chance.
* The stars represent the significance.
* The standard error of th estimates and the t-value are both provided to show how the p-values were calculated.
* <https://www.omnicalculator.com/statistics/p-value>

What does a statistically significant estimate mean?



* Estimate, but not the actual number, and that’s why we have a distribution around this estimate because we don’t know what the actual number is.
* But we are fairly sure, 95% sure, that it lies within this confidence interval.
* We calculate the confidence interval by taking the estimate and subtract about two times the standard error to get the lower bound and add about two times the standard error to get the upper bound.
* Null hypothesis: this variable has no effect on the dependent variable, i.e., this estimate should be about zero. But if our confidence interval is really far from zero, that’s gonna give us a low p-value for a t-test, implying that this effect is actually nonzero and is statistically significant.

Et bilde som inneholder tekst

Automatisk generert beskrivelse

Residual standard error

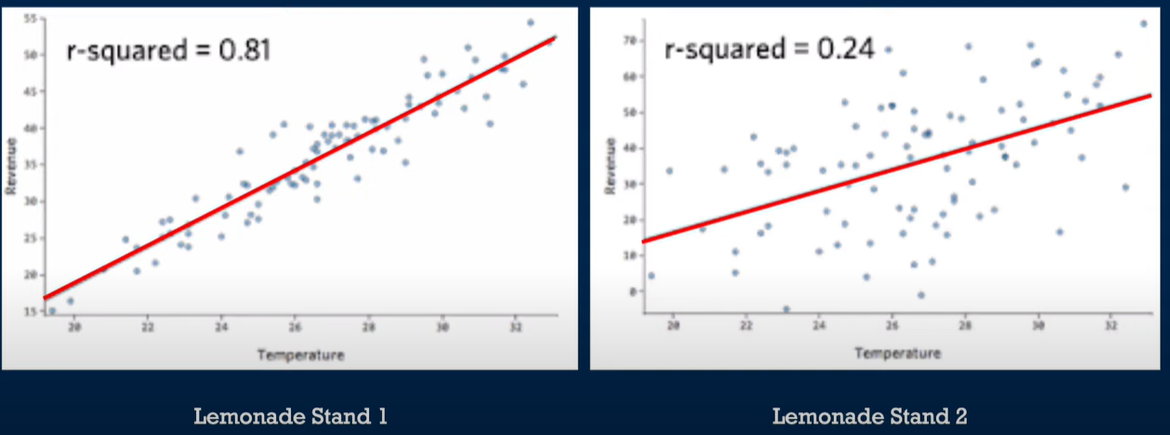
* The standard deviation for the residuals.
* Provides us a measure of how much the actual value will differ from the value we’re trying to predict. I.e., the standard error of the residuals.
* Used to measure how well a regression model fits a dataset.
* **Smaller** residual standard error means predictions are better. The model fits the data better.

Degrees of freedom

* The degrees of freedom are the number of datapoints that went into the estimation of the parameter.
* This is typically the rows of the data minus the number of variables that we’re estimating and that includes the intercept.
* 159 rows and we’re estimating three different values, so 159-3 gives us 156 degrees of freedom.
* **Higher** degrees of freedom often mean more power to reject a false null hypothesis and find a significant result. Higher degrees of freedom mean larger sample sizes.

Multiple R-squared

* Gives a measurement of what % of the variance in the response variable can be explained by the regression.
* Will show you the amount of variation in the response variable explained by the predictor variables.
* If you were to add more predictor variables the R squared would essentially always increase because predictors will always explain some portion of the variance, but this doesn’t necessarily mean the model is getting better. It could just be that you’re overfitting to your data.
* **Larger** r^2 means the model is better.

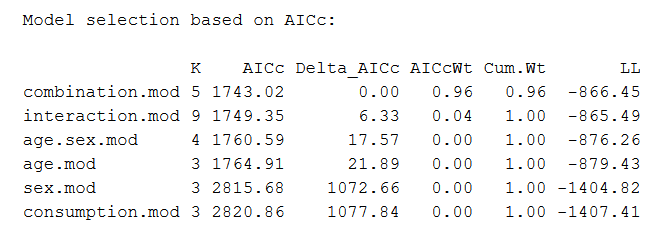


Adjusted R-squared

* Controls for each additional predictor added (to prevent overfitting), so it may not increase as you add more variables.
* It adds penalties each time you add another predictor.
* The distinction between R-squared and adjusted R-squared doesn’t matter too much in a simple linear regression where you only have one intercept and one slope, but it makes more sense in multiple linear regression where you might have multiple predictors.
* Essentially, you want your model to be parsimonious or somewhat simple but also fit the data really well and if there’s a large enough difference between your multiple R-squared and your adjusted R-squared you may have overfit your model.
* **Larger** adjusted r^2 means the model is better.
* We are more interested in this in a multiple regression with several independent variables because that’s what it’s adjusting for.

F-statistic

* Indicates if the model as a whole is statistically significant.
* A general indicator of whether there’s actually a relationship between the predictor variables and your response variable.
* A number further from one is actually better but again, you can pay attention to the p-value on the side to determine if the model as a whole is statistically significant. It needs to be **less than the alpha** to be significant.
* An F-statistic is a test for the ratio of two variances. . If the value of the statistic is 1 then the two variances are equal. If the value is much larger than 1 then the variance of the first population is greater than the second.
* To find out if means are significantly different.



<https://www.scribbr.com/statistics/akaike-information-criterion/#:~:text=In%20statistics%2C%20AIC%20is%20used,the%20model%20reproduces%20the%20data>).

AIC (Akaike information criterion)

* The **smaller** the AIC, the better fit is the model.

Et bilde som inneholder tekst

Automatisk generert beskrivelse

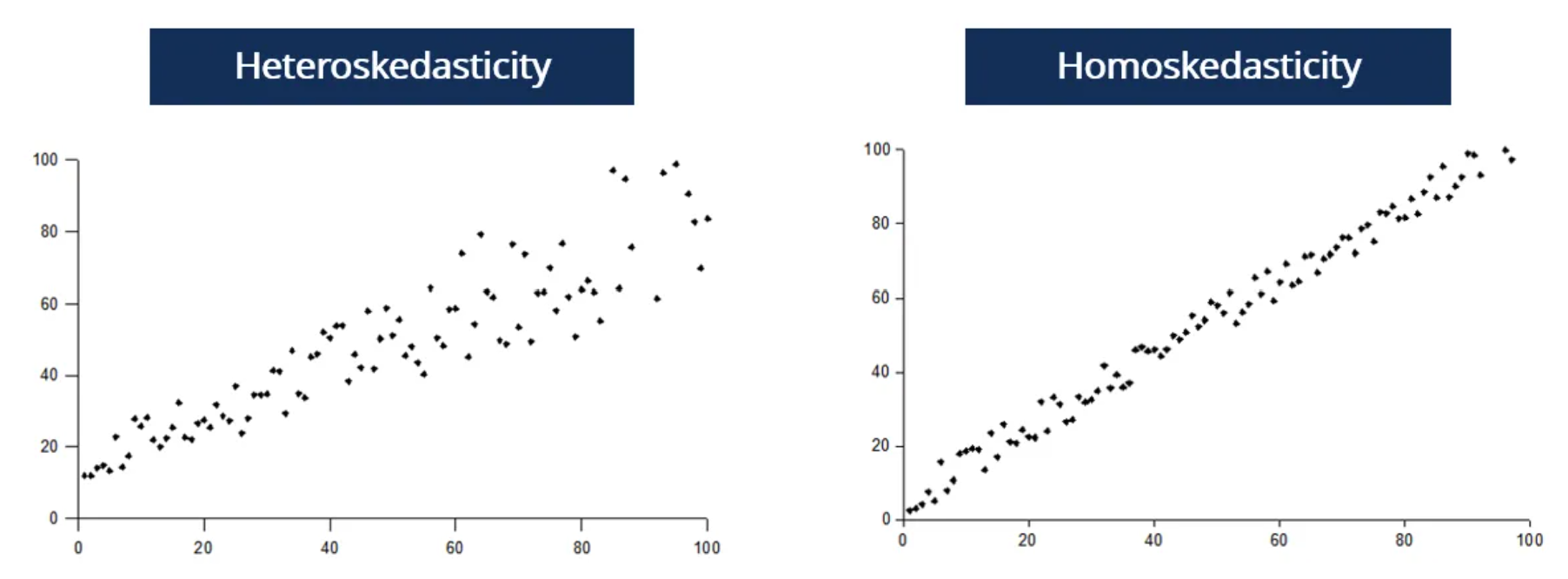
<https://www.statology.org/variance-inflation-factor-r/#:~:text=Multicollinearity%20in%20regression%20analysis%20occurs,information%20in%20the%20regression%20model>.

VIF (variance of inflation)

* We use it to test for multicollinearity, which occurs when two or more predictor variables are highly correlated with each other, such that they do not provide unique or independent information in the regression model.
* The value for VIF starts at 1 and has no upper limit. A general rule of thumb for interpreting VIFs is as follows:
  + A value of 1 indicates there is no correlation between a given predictor variable and any other predictor variables in the model.
  + A value between 1 and 5 indicates moderate correlation between a given predictor variable and other predictor variables in the model, but this is often not severe enough to require attention.
  + A value greater than 5 indicates potentially severe correlation between a given predictor variable and other predictor variables in the model. In this case, the coefficient estimates and p-values in the regression output are likely unreliable.
* By setting up a correlation matrix you can easily detect which variables are highly correlated.

Perfect multicollinearity

* Perfect multicollinearity makes it impossible to estimate a value for every coefficient in a regression model.
* The dummy variable trap: Another scenario where perfect multicollinearity can occur is known as the [dummy variable trap](https://www.statology.org/dummy-variable-trap/).
* <https://www.statology.org/perfect-multicollinearity/>



<https://corporatefinanceinstitute.com/resources/knowledge/other/heteroskedasticity/>

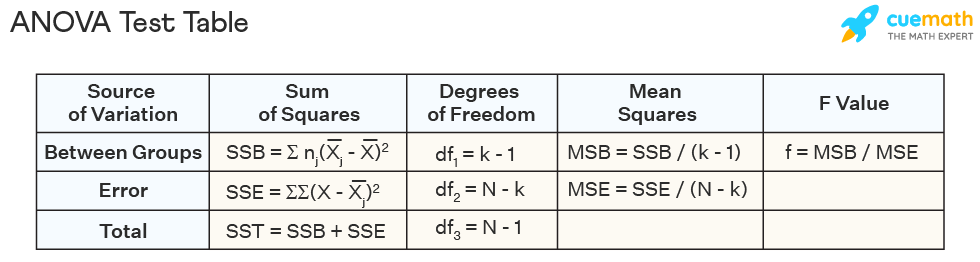
Heteroskedasticity

* Happens when the standard deviations of a predicted variable, monitored over different values of an independent variable or as related to prior time periods, are non-constant.
* Refers to situations 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.
* Models involving a wide range of values are supposedly more prone to heteroskedasticity.

**How to interpret ANOVA R output**

* If the task asks about SSE, SSM or TSS, use ANOVA
* **SSE** (sum of squared estimate of errors) = **RSS** (residual sum of squares) = **SSR** (sum of squared residuals)
  + SSE = the residuals’ Sum of Squares (SS)
* **SSM** (model sum of squares) = **ESS** (explained sum of squares)
  + SSM = the independent variable’s Sum of Squares (SS)
* **TSS** (total sum of squares) = **SST** (sum of squares total)
  + SSM + SSE = TSS

<https://community.gooddata.com/metrics-and-maql-kb-articles-43/hypothesis-testing-one-way-analysis-of-variance-anova-with-f-test-243>



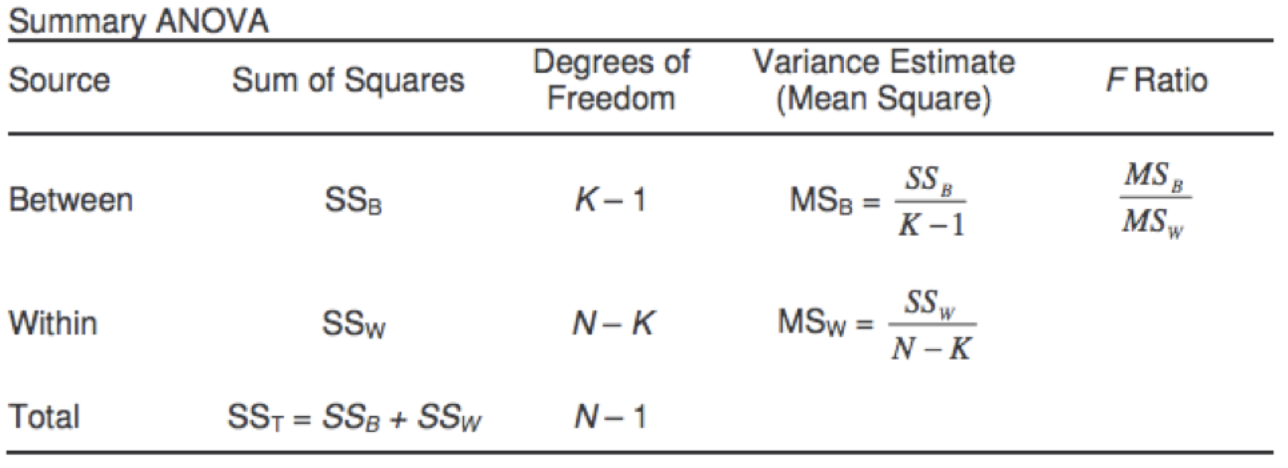
Et bilde som inneholder tekst, skilt, skjermbilde, lukk

Automatisk generert beskrivelse

<https://www.youtube.com/watch?v=reOZVODpLDk>

One-way ANOVA

* K is number of groups, N is number of records
* Involves one factor or independent variable.
* A one-way ANOVA is primarily designed to enable the equality testing between three or more means.
* In a one-way ANOVA, the one factor or independent variable analyzed has three or more categorical groups.
* <https://www.technologynetworks.com/informatics/articles/one-way-vs-two-way-anova-definition-differences-assumptions-and-hypotheses-306553#:~:text=A%20one%2Dway%20ANOVA%20only,multiple%20groups%20of%20two%20factors>.
* If the p-value turns out to be significant, we only know that at least one of the groups are different from each other, but not identified which one of them that differ. We cannot see this in the ANOVA analysis, but we can use TukeyHSD() in R or other post hoc tests to find out.
* Assumptions:
  + Normality – Each sample was drawn from a normally distributed population.
  + Equal Variances – The variances of the populations that the samples come from are equal. You can use [Bartlett’s Test](https://www.statology.org/bartletts-test/" \t "_blank) to verify this assumption.
  + Independence – The observations in each group are independent of each other and the observations within groups were obtained by a random sample.
* <https://www.statology.org/one-way-anova/>



Two-way ANOVA

* A two-way ANOVA tests the effect of two independent variables on a dependent variable.
* A two-way ANOVA test analyzes the effect of the independent variables on the expected outcome along with their relationship to the outcome itself.
* Demonstration: <https://www.youtube.com/watch?v=WdWOwKcOD6M>
* If there
* Same assumptions for two-way ANOVA as for one-way ANOVA.

F-values

* Large F-value indicates more difference between the groups than within the groups.
* If the F-value is bigger than the critical value, it is significant.
* We need a **high** F-value to reject the null-hypothesis.
* If the F-score is much greater than one, the variance between is probably the source of most of the variance in the total sample, and the samples probably come from populations with different means.

Eta squared

* A descriptive measure of the strength of association between independent and dependent variables in the sample.
* In context of a one-way ANOVA, the most typically reported measure of effect size is known as eta-squared.
* It represents the percentage of variance in the dependent variable accounted for by the independent variable.
* Similar to coefficient of determination: r^2
* The value for Eta squared ranges from 0 to 1, where values closer to 1 indicate a **higher** proportion of variance that can be explained by a given variable in the model.
* The following rules of thumb are used to interpret values for Eta squared:
  + .01: Small effect size
  + .06: Medium effect size
  + .14 or higher: Large effect size
* <https://www.statology.org/eta-squared/#:~:text=The%20value%20for%20Eta%20squared,given%20variable%20in%20the%20model>.
* Typically preferred to partial eta squared.

Partial eta squared

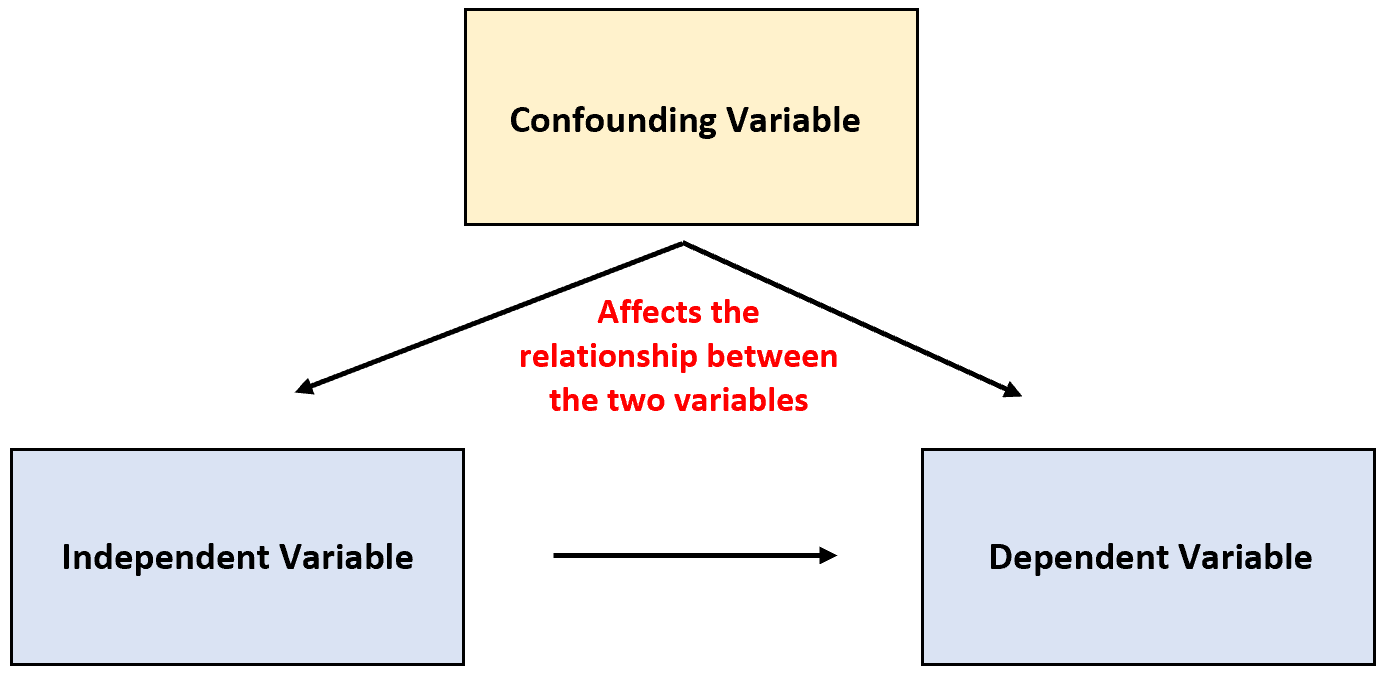
* Partial eta squared is the [ratio](https://www.statisticshowto.com/ratios-and-rates/#ratio)of variance associated with an effect, plus that effect and its associated [error variance](https://www.statisticshowto.com/residual-variance/). The formula is similar to eta2
* When you only have one independent variable, partial eta2 is the same as eta2.
* Partial etas are usually used when a person appears in more than one cell (i.e. the cells are not independent).
* A way to measure the [effect size](https://www.statology.org/effect-size/) of different variables in ANOVA models.
* The value for Partial eta squared ranges from 0 to 1, where values closer to 1 indicate a **higher** proportion of variance that can be explained by a given variable in the model after accounting for variance explained by other variables in the model.
* The following rules of thumb are used to interpret values for Partial eta squared:
  + .01: Small effect size
  + .06: Medium effect size
  + .14 or higher: Large effect size

Hovedeffekt

* Hovedeffekter vil si effekter av den ene uavhengige variabelen på den avhengige variabelen. Vi kan måle hovedeffekt isolert for hver av de uavhengige variablene. Interaksjonseffekt vil si at effekt av én uavhengig variabel avhenger av nivå på den/de andre uavhengige variabelen(e).

Chi-square ( test

* A **Chi-Square goodness of fit** test is used to determine whether or not a categorical variable follows a hypothesized distribution.
* A **Chi-Square Test of Independence** is used to determine whether or not there is a significant association between two categorical variables.
* The **larger** the Chi-square value, the greater the probability that there really is a significant difference. There is a significant difference between the groups we are studying.
* A low value for chi-square means there is a high correlation between your two sets of data.



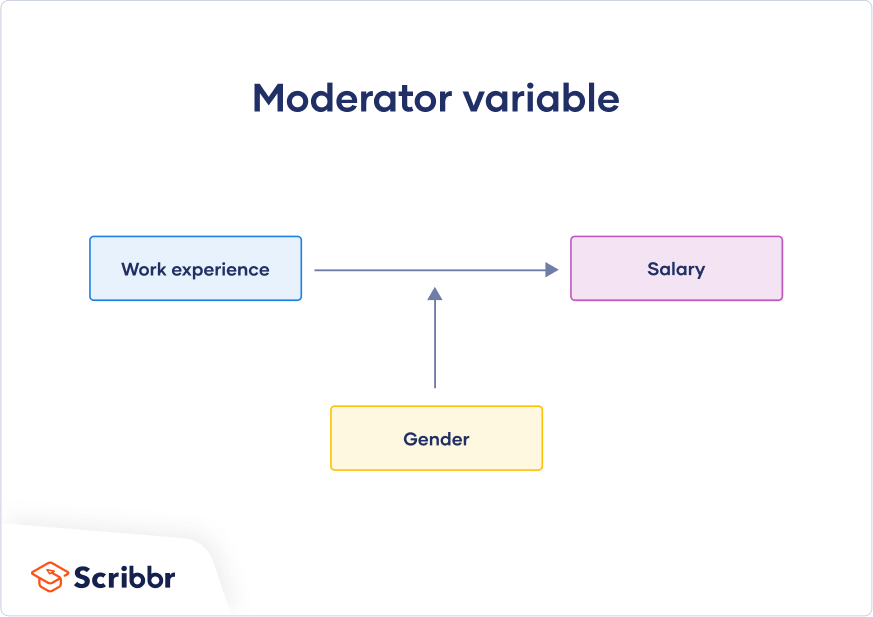
<https://www.statology.org/confounding-variable/>

Confounding variable

* A variable that is not included in an experiment, yet affects the relationship between the two variables in an experiment.
* Requirements for confounding variables:
  + It must be correlated with the independent variable.
  + It must have a causal relationship with the dependent variable.
* They are problematic because:
  + Confounding variables can make it seem that cause-and-effect relationships exist when they don’t.
  + Confounding variables can mask the true cause-and-effect relationship between variables.
* In experimental studies we remove the effect of confounding variables by keeping them constant.

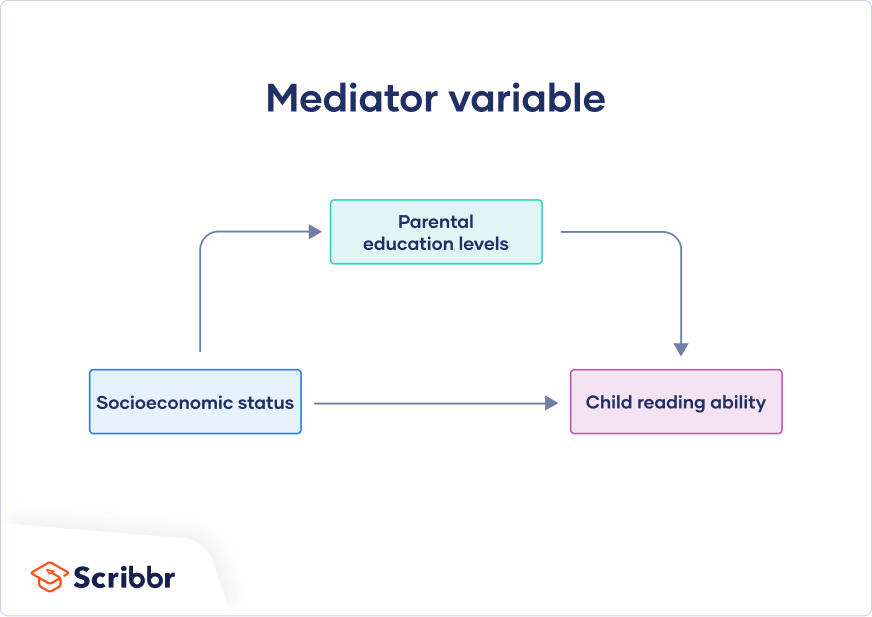
Suppressor variable

* Sometimes there is no association between two variables until you control for a third one. The third variable is a suppressor because it suppress the association between the first two variables.



Moderator variable

* A moderator influences the level, direction, or presence of a relationship between variables. It shows you for whom, when, or under what circumstances a relationship will hold.
* <https://www.scribbr.com/methodology/mediator-vs-moderator/>



Mediator variable

* A mediator is a way in which an independent variable impacts a dependent variable. It’s part of the causal pathway of an effect, and it tells you how or why an effect takes place.