Standard Error of Estimate

The standard deviation of an estimate is called the standard error. The standard error of the coefficient measures how precisely the model estimates the coefficient's unknown value. The standard error of the coefficient is always positive.

Use the standard error of the coefficient to measure the precision of the estimate of the coefficient. The smaller the standard error, the more precise the estimate. Dividing the coefficient by its standard error calculates a t-value. If the p-value associated with this t-statistic is less than your alpha level, you conclude that the coefficient is significantly different from zero.

For example, a materials engineer at a furniture manufacturing site wants to assess the strength of the particle board that they use. The engineer collects stiffness data from particle board pieces with various densities at different temperatures and produces the following linear regression output. The standard errors of the coefficients are in the third column.

Coefficients

Term Coef SE Coef T-Value P-Value VIF

Constant 20.1 12.2 1.65 0.111

Stiffness 0.2385 0.0197 12.13 0.000 1.00

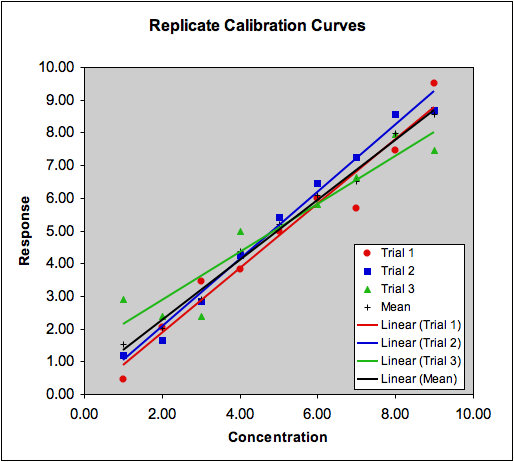
Temp -0.184 0.178 -1.03 0.311 1.00

The standard error of the *Stiffness* coefficient is smaller than that of *Temp*. Therefore, your model was able to estimate the coefficient for *Stiffness* with greater precision. In fact, the standard error of the *Temp* coefficient is about the same as the value of the coefficient itself, so the t-value of -1.03 is too small to declare statistical significance. The resulting p-value is much greater than common levels of α, so that you cannot conclude this coefficient differs from zero. You remove the *Temp* variable from your regression model and continue the analysis.

***Errors in the Regression Equation:***

There is always some error associated with the measurement of any signal. Earlier, we saw how this affected replicate measurements, and could be treated statistically in terms of the [mean and standard deviation](https://sites.chem.utoronto.ca/chemistry/coursenotes/analsci/stats/MeasMeanVar.html).

The same phenomenon applies to each measurement taken in the course of constructing a calibration curve, causing a variation in the slope and intercept of the calculated regression line. This can be reduced - though never completely eliminated - by making replicate measurements for each standard.



*Multiple calibrations with single values compared to the mean of all three trials. Note how all the regression lines pass close to the centroid of the data.*

Even with this precaution, we still need some way of estimating the likely error (or uncertainty) in the slope and intercept, and the corresponding uncertainty associated with any concentrations determined using the regression line as a calibration function.

Exc ***The Uncertainty of the Regression:***

We saw earlier that the spread of the actual calibration points either side of the line of regression of *y* on *x* (which we are using as our calibration function) can be expressed in terms of the [regression residuals](https://sites.chem.utoronto.ca/chemistry/coursenotes/analsci/stats/RegrEqn.html#section3-4-3), (*yi* − https://sites.chem.utoronto.ca/chemistry/coursenotes/analsci/stats/images/yhat.gif): The greater these resdiuals, the greater the uncertainty in where the true regression line actually lies. The uncertainty in the regression is therefore calculated in terms of these residuals. Technically, this is the standard error of the regression, *sy/x*:

https://sites.chem.utoronto.ca/chemistry/coursenotes/analsci/stats/images/sxy.gif

Note that there are (*n* − 2) degrees of freedom in calculating *sy/x*. This is because we are making **two** assumptions in this equation: a) that the sample population is representative of the entire population, and b) that the https://sites.chem.utoronto.ca/chemistry/coursenotes/analsci/stats/images/yhat.gif values are representative of the true *y*-values. For each assumption, we remove one degree of freedom, and our estimated standard deviation becomes larger.

Another way of understanding the [degrees of freedom](https://sites.chem.utoronto.ca/chemistry/coursenotes/analsci/stats/DegFree.html) is to note that we are estimating two parameters from the regression – the slope and the intercept. Therefore, *ν = n − 2* and we need at least three points to perform the regression analysis.

***The Uncertainty of the Slope:***

The slope of the regression line is obviously important, as it determines the *sensitivity* of the calibration function; that is, the rate at which the signal changes with concentration. The higher (steeper) the slope, the easier it is to distinguish between concentrations which are close to one another. (Technically, the greater the resolution in concentration terms.) The uncertainty in the slope is expressed as the standard error (or deviation) of the slope, *sb*, and is calculated in terms of the standard error of the regression as:

sb
     equals sy/x over the root of the squared x residuals

The corresponding confidence interval for the slope is calculated using the *t*-statistic for (*n* − 2) degress of freedom as:

*b* ± *tn−2sb*

**Remember**: here *n* is the *number of calibration points* used in the regression calculation.

***The Uncertainty of the Intercept:***

The intercept of the regression line has implications for both the smallest detectable signal (measured response) *and* the corresponding lowest detectable concentration. The uncertainty in the intercept is also calculated in terms of the standard error of the regression as the standard error (or deviation) of the intercept, *sa*:

https://sites.chem.utoronto.ca/chemistry/coursenotes/analsci/stats/images/sinter.gif

The corresponding confidence interval for the intercept is calculated in the same way as that for the slope, namely:

*a* ± *tn−2sa*

In this tutorial, we will cover the difference between r-squared and adjusted r-squared. It includes detailed theoretical and practical explanation of these two statistical metrics in R.

**R-squared (R²)**   
  
It measures the proportion of the variation in your dependent variable explained by all of your independent variables in the model. It assumes that every independent variable in the model helps to explain variation in the dependent variable. In reality, some independent variables (predictors) don't help to explain dependent (target) variable. In other words, some variables do not contribute in predicting target variable.  
  
Mathematically, R-squared is calculated by dividing sum of squares of residuals (**SSres**) by total sum of squares (**SStot**) and then subtract it from 1. In this case, SStot measures total variation. **SSreg**measures explained variation and SSres measures unexplained variation.  
  
As **SSres + SSreg = SStot, R² = Explained variation / Total Variation**

|  |
| --- |
| <https://3.bp.blogspot.com/-MFqQLwbGwd4/WMO7tz39beI/AAAAAAAAF9I/Bjpfohgv_5E4lu7HzHJBQtXsBM--byqPwCLcB/s1600/rsquared.png> |
| R-squared Equation |

R-Squared is also called **coefficient of determination**. It lies between **0%**and **100%.**A r-squared value of 100% means the model explains all the variation of the target variable. And a value of 0% measures zero predictive power of the model. **Higher R-squared value, better the model.**  
  
**Adjusted R-Squared**  
  
It measures the proportion of variation explained by only those independent variables that really help in explaining the dependent variable. It penalizes you for adding independent variable that do not help in predicting the dependent variable.  
  
Adjusted R-Squared can be calculated mathematically in terms of sum of squares. The only difference between R-square and Adjusted R-square equation is degree of freedom.

|  |
| --- |
| <https://4.bp.blogspot.com/-pyczHisQzi4/WMPAvcEyiOI/AAAAAAAAF9Y/ad-JKQBeszkpVrI-Cf3YtSK8GzhdLvdOwCLcB/s1600/adjusted%2Br-squared.png> |
| Adjusted R-Squared Equation |

In the above equation, dft is the degrees of freedom n– 1 of the estimate of the population variance of the dependent variable, and dfe is the degrees of freedom n – p – 1 of the estimate of the underlying population error variance.  
  
Adjusted R-squared value can be calculated based on value of r-squared, number of independent variables (predictors), total sample size.

|  |
| --- |
| <https://4.bp.blogspot.com/-qEGt3DaQIF0/V2meLITZj3I/AAAAAAAAEp4/WKCs0FrI1JsovDMwaw1r1iUboULfRI7MwCLcB/s1600/stb1.png> |
| Adjusted R-Squared Equation 2 |

**Difference between R-square and Adjusted R-square**

1. Every time you add a independent variable to a model, the **R-squared** **increases**, even if the independent variable is insignificant. It never declines. Whereas **Adjusted R-squared** increases only when independent variable is significant and affects dependent variable.

In the table below, adjusted r-squared is maximum when we included two variables. It declines when third variable is added. Whereas r-squared increases when we included third variable. It means third variable is insignificant to the model.

|  |
| --- |
| <https://1.bp.blogspot.com/-ixZb6IYiS6c/U_4DRNokxnI/AAAAAAAADPc/hDelIMR8yEs/s1600/image_rsq.png> |
| R-Squared vs. Adjusted R-Squared |

1. Adjusted r-squared can be negative when r-squared is close to zero.
2. Adjusted r-squared value always be less than or equal to r-squared value.

**Which is better?**

*Adjusted R-square should be used to compare models with different numbers of independent variables. Adjusted R-square should be used while selecting important predictors (independent variables) for the regression model.*

**R : Calculate R-Squared and Adjusted R-Squared**  
  
Suppose you have **actual**and **predicted dependent variable**values. In the script below, we have created a sample of these values. In this example, **y**refers to the observed dependent variable and **yhat**refers to the predicted dependent variable.

*y = c(21, 21, 22.8, 21.4, 18.7, 18.1, 14.3, 24.4, 22.8, 19.2)  
yhat = c(21.5, 21.14, 26.1, 20.2, 17.5, 19.7, 14.9, 22.5, 25.1, 18)  
R.squared = 1 - sum((y-yhat)^2)/sum((y-mean(y))^2)  
print(R.squared)*

**Final Result :**R-Squared = 0.6410828  
Let's assume you have three independent variables in this case.

*n = 10  
p = 3  
adj.r.squared = 1 - (1 - R.squared) \* ((n - 1)/(n-p-1))  
print(adj.r.squared)*

In this case, **adjusted r-squared value is 0.4616242** assuming we have 3 predictors and 10 observations.

**Python : Calculate Adjusted R-Squared and R-Squared**

*import numpy as np  
y = np.array([21, 21, 22.8, 21.4, 18.7, 18.1, 14.3, 24.4, 22.8, 19.2])  
yhat = np.array([21.5, 21.14, 26.1, 20.2, 17.5, 19.7, 14.9, 22.5, 25.1, 18])  
R2 = 1 - np.sum((yhat - y)\*\*2) / np.sum((y - np.mean(y))\*\*2)  
R2  
n=y.shape[0]  
p=3  
adj\_rsquared = 1 - (1 - R2) \* ((n - 1)/(n-p-1))  
adj\_rsquared*

**SAS : Adjusted R-Squared Calculation**

*data temp;*

*input y yhat;*

*cards;*

*21 21.5*

*21 21.14*

*22.8 26.1*

*21.4 20.2*

*18.7 17.5*

*18.1 19.7*

*14.3 14.9*

*24.4 22.5*

*22.8 25.1*

*19.2 18*

*;*

*run;*

*data out2;*

*set temp ;*

*d=y-yhat;*

*absd=abs(d);*

*d2 = d\*\*2;*

*run;*

*/\* Residual Sum of Square \*/*

*proc means data = out2 ;*

*var d2;*

*output out=rss sum=;*

*run;*

*data \_null\_;*

*set rss;*

*call symputx ('rss', d2);*

*run;*

*%put &RSS.;*

*/\* Total Sum of Square \*/*

*proc means data = temp ;*

*var y;*

*output out=avg\_y mean=avg\_y;*

*run;*

*data \_null\_;*

*set avg\_y;*

*call symputx ('avgy', avg\_y);*

*run;*

*%put &avgy.;*

*data out22;*

*set temp ;*

*diff = y - &avgy.;*

*diff2= diff\*\*2;*

*run;*

*proc means data = out22 ;*

*var diff2;*

*output out=TSS sum=;*

*run;*

*data \_null\_;*

*set TSS;*

*call symputx ('TSS', diff2);*

*run;*

*/\* Calculate the R2 \*/*

*%LET rsq = %SYSEVALF(1-&RSS./&TSS);*

*%put &RSQ;*

*/\* Calculate the Adj R2 \*/*

*%LET N = 10;*

*%LET P = 3;*

*%let AdjRsqrd= %SYSEVALF(1 -((1-&rsq)\*(&N-1)/(&N-&P-1)));*

*%PUT &AdjRsqrd;*