

Sum of Squares: SST, SSR, SSE

The sum of squares is a statistical measure of variability. It indicates the dispersion of data points around the mean and how much the dependent variable deviates from the predicted values in regression analysis.

We decompose variability into the **sum of squares total** (SST), the **sum of squares regression** (SSR), and the **sum of squares error** (SSE). The decomposition of variability helps us understand the sources of variation in our data, assess a model's goodness of fit, and understand the relationship between variables.

We define SST, SSR, and SSE below and explain what aspects of variability each measure.

SST, SSR, SSE: Definition and Formulas

What Is SST in Statistics?

The **sum of squares total (SST)** or the **total sum of squares (TSS)** is the sum of squared differences between the observed *dependent variables* and the overall **mean**. Think of it as the dispersion of the observed variables around the **mean**—similar to the **variance** in descriptive statistics. But SST measures the total variability of a dataset, commonly used in regression analysis and ANOVA.

Mathematically, the difference between variance and SST is that we adjust for the degree of freedom by dividing by $n-1$ in the variance formula.

$$SST = n \sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (y_i - \bar{y})^2$$

Where:

y_i – observed dependent variable

\bar{y} – mean of the dependent variable

What Is SSR in Statistics?

The **sum of squares due to regression (SSR)** or **explained sum of squares (ESS)** is the sum of the differences between the *predicted value* and the **mean** of the *dependent variable*. In other words, it describes how well our line fits the data.

The SSR formula is the following:

$$SSR = n \sum_{i=1} (\hat{y}_i - \bar{y})^2 = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$$

Where:

\hat{y}_i – the predicted value of the dependent variable

\bar{y} – mean of the dependent variable

If **SSR** equals **SST**, our **regression model** perfectly captures all the observed variability, but that's rarely the case.

What Is SSE in Statistics?

The **sum of squares error (SSE)** or **residual sum of squares (RSS, where residual means remaining or unexplained)** is the difference between the *observed* and *predicted* values.

The SSE calculation uses the following formula:

$$SSE = n \sum_{i=1} \varepsilon_i^2 = \sum_{i=1}^n \varepsilon_i^2$$

Where ϵ_i is the difference between the actual value of the dependent variable and the predicted value:

$$\epsilon_i = y_i - \hat{y}_i = y_i - \beta_0 - \beta_1 x_i$$

Regression analysis aims to [minimize the SSE](#)—the smaller the error, the better the **regression's** estimation power.

The Confusion between the Abbreviations

As mentioned, the **sum of squares error (SSE)** is also known as the **residual sum of squares (RSS)**, but some individuals denote it as **SSR**, which is also the abbreviation for the sum of squares due to regression.

Although there's no universal standard for abbreviations of these terms, you can readily discern the distinctions by carefully observing and comprehending them.

The conflict regards the abbreviations, not the concepts or their application. So, remember the definitions and the possible notations (**SST**, **SSR**, **SSE** or **TSS**, **ESS**, **RSS**) and how they relate..

What Is the Relationship between SSR, SSE, and SST?

The rationale is the following:

The total variability of the dataset is equal to the variability explained by the **regression line** plus the unexplained variability, known as error.

Mathematically, **SST = SSR + SSE**.

The rationale is the following:

The total variability of the dataset is equal to the variability explained by the **regression line** plus the unexplained variability, known as error.

Given a constant total variability, a lower error means a better **regression** model. Conversely, a higher error means a less robust **regression**. And that's valid regardless of the notation you use.

Next Steps

Why do we need SST, SSR, and SSE? We can use them to calculate the [R-squared](#), conduct F-tests in regression analysis, and combine them with other goodness-of-fit measures to evaluate regression models.

Our [linear regression calculator](#) automatically generates the SSE, SST, SSR, and other relevant statistical measures. The adjacent article includes detailed explanations of all crucial concepts related to regression, such as coefficient of determination, standard error of the regression, correlation coefficient, etc.

Measuring Explanatory Power with the R-squared

If you are looking for a widely-used measure that describes how powerful a **regression** is, the **R-squared** will be your cup of tea. [A prerequisite to understanding the math behind the R-squared is the decomposition of the total variability of the observed data into explained and unexplained.](#)

A key highlight from that decomposition is that the smaller the regression error, the better the regression.

Now, it's time to introduce you to the **R-squared**. The **R-squared** is an intuitive and practical tool, when in the right hands. It is equal to variability explained by the **regression**, divided by total variability.

What Exactly is the R-squared?

It is a relative measure and takes values ranging from 0 to 1. An **R-squared** of zero means our **regression line** explains none of the variability of the data.

An **R-squared** of 1 would mean our model explains the entire variability of the data.

Unfortunately, **regressions** explaining the entire variability are rare. What we usually observe are values ranging from 0.2 to 0.9.

What's the Best Value for an R-squared?

The immediate question you may be asking: “What is a good **R-squared**? When do I know, for sure, that my **regression** is good enough?”

Unfortunately, there is no definite answer to that.

In fields such as physics and chemistry, scientists are usually looking for **regressions** with **R-squared** between 0.7 and 0.99. However, in social sciences, such as economics, finance, and psychology the situation is different. There, an **R-squared** of 0.2, or 20% of the variability explained by the model, would be fantastic.

It depends on the complexity of the topic and how many variables are believed to be in play.

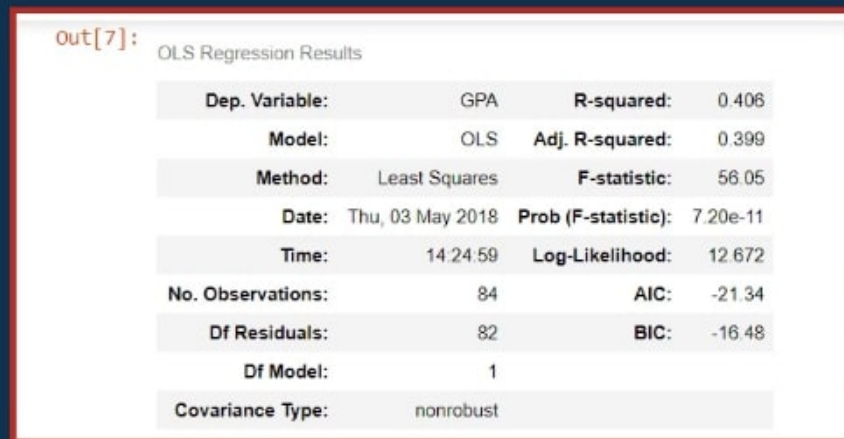
Dealing with Multiple Variables

Take your income, for example. It may depend on your household income (including your parents and spouse), your education, years of experience,

country you are living in, and languages you speak. However, this may still account for less than 50% of the variability of income.

Your salary is a very complex issue. But you probably know that.

SAT-GPA Example

A screenshot of a Jupyter Notebook output cell showing OLS Regression Results. The output is labeled 'Out[7]:'. The results are displayed in a table-like format with alternating light and dark gray rows. The table includes statistics such as R-squared, Adjusted R-squared, F-statistic, Prob (F-statistic), Log-Likelihood, AIC, BIC, and the number of observations and residuals. The covariance type is listed as 'nonrobust'.

Out[7]: OLS Regression Results			
Dep. Variable:	GPA	R-squared:	0.406
Model:	OLS	Adj. R-squared:	0.399
Method:	Least Squares	F-statistic:	56.05
Date:	Thu, 03 May 2018	Prob (F-statistic):	7.20e-11
Time:	14:24:59	Log-Likelihood:	12.672
No. Observations:	84	AIC:	-21.34
Df Residuals:	82	BIC:	-16.48
Df Model:	1		
Covariance Type:	nonrobust		

The SAT score is one of the better determinants of intellectual capacity and capability. The truth is that our regression had an **R-squared** of 0.406, as you can see in the picture below.

$$R^2 = 0.406$$

$$\text{College GPA} = 0.275 + 0.0017 * \text{SAT}$$

Out[7]:

OLS Regression Results			
Dep. Variable:	GPA	R-squared:	0.406
Model:	OLS	Adj. R-squared:	0.399
Method:	Least Squares	F-statistic:	56.05
Date:	Thu, 03 May 2018	Prob (F-statistic):	7.20e-11
Time:	14:24:59	Log-Likelihood:	12.672
No. Observations:	84	AIC:	-21.34
Df Residuals:	82	BIC:	-16.48
Df Model:	1		
Covariance Type:	nonrobust		

In other words, SAT scores explain 41% of the variability of the college grades for our sample.

Other Factors

An **R-squared** of 41% is neither good nor bad. But since it is far away from 90%, we may conclude we are missing some important information. Other determinants must be considered. Variables such as gender, income, and marital status could help us understand the full picture a little better.

$$R^2 = 0.406$$

$$\text{College GPA} = 0.275 + 0.0017 * \text{SAT}$$



Now, you probably feel ready to move on. However, you should remember one thing.

Don't jump into regressing so easily. Critical thinking is crucial. Before agreeing that a factor is significant, you should try to understand why. So, let's quickly justify that claim.

Gender

First, women are more likely to outperform men in high school.

But then in higher education, more men enter academia.

There are many biases in place here. Without telling you if female or male candidates are better, scientific research shows that a gender gap *exists* in education. Gender is an important input for any **regression** on the topic.

Income

The second factor we pointed out is income. If your household income is low, you are more likely to get a part-time job.

Thus, you'll have less time for studying and probably get lower grades.

If you've ever been to college, you will surely remember a friend who underperformed because of this reason.

Children

Third, if you get married and have a child, you'll definitely have a lower attendance.

Contrary to what most students think when in college, attendance is a significant factor for your GPA. You may think your time is better spent when skipping a lecture, but your GPA begs to differ.

When to Include More Factors?

After these clarifications, let's find the bottom line. The **R-squared** measures the goodness of fit of our model. The more factors we include in our **regression**, the higher the **R-squared**.

So, should we include gender and income in our **regression**? If this is in line with our research, and their inclusion results in a better model, we should do that.

The Adjusted R-squared

The **R-squared** seems quite useful, doesn't it? However, it is not perfect. To be more precise, we'll have to refine it. Its new version will be called the **adjusted R-squared**.

What it Adjusts for

Let's consider the following two statements:

1. The **R-squared** measures how much of the total variability is explained by our model.
2. Multiple **regressions** are always better than simple ones. This is because with each additional variable that you add, the explanatory power may only increase or stay the same.

Well, the **adjusted R-squared** considers exactly that. It measures how much of the total variability our model explains, considering the number of variables.

The **adjusted R-squared** is always smaller than the **R-squared**, as it penalizes excessive use of variables.