**Are the Error Terms Normally Distributed in a Linear Regression Model?**

**Justification for the Normality Assumption**

In a linear regression model, the **normality assumption** (i.e., the error term is normally distributed) NOT required for calculating **unbiased** estimates. In this post, we’ll discuss under what situations we would need this normality assumption, why it is reasonable to make such an assumption, and how to check if the errors are normally distributed.

**What are the error terms in a linear regression model?**

The following is what a typical linear regression model would look like in the population.

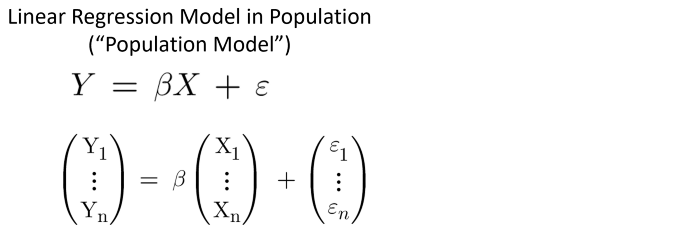


Image by author

* The response variable (Y) can be written as a linear combination of explanatory variables (X)
* **β** are the unknown **population parameters** (**fixed values**) that we would estimate from the sample data.
* **ε** is the error term that represents the difference between the true value (expressed by **βX**) and the observed response value in the population. We assume that there could be many different response values (Y) for a given value of X in the population. In other words, **conditional on X**, both Y and ε could take different values. Therefore, **Both the response variable and the error term are random variables.**

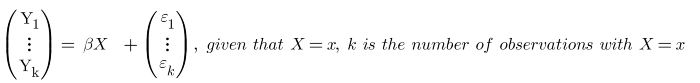


Image by author

If there is only a single value of the response variable (Y) with X= x in the sample, we’ll still assume that there are many **unobserved** different response values (Y) for a given value of X in the population.

In a given observation, **εi** is a random variable. Based on the assumptions of classical linear regression models, we assume that

* The error terms (i.e., ε1, ε2, …, εn) have a zero mean.

E(**εi**) = 0, i = 1, 2, …, n



Image by author

* The error terms have a **constant variance** (𝜎2 ) and they are NOT correlated with each other.



Image by author

**How to estimate the variance of the error term in a Linear Regression Model?**

σ2 is the variance of the error term. σ is the standard deviation of the error. This is an **unknown** parameter in the population model. We typically estimate this value using the **residuals** **in the sample data**. The residual, e is defined as

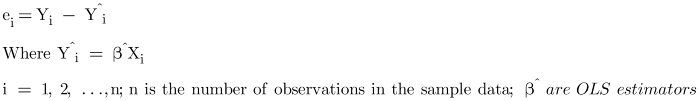


Image by author

S (aka, the residual standard deviation, the residual standard error, or the standard error of the regression) is the **unbiased estimator** of the standard deviation of the error.

S2 (aka, the residual variance, the mean square error, or MSE) is the **unbiased estimator** of the variance of the error.

S2 can be computed as the residual sum of squares divided by the number of degrees of freedom.

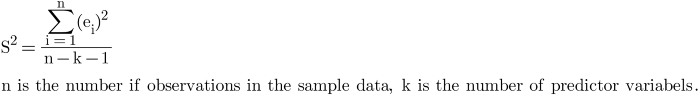


Image by author

**What is the motivation for assuming the errors are normally distributed in a Linear Regression Model?**

At this point, we know the error is a random variable with a mean of zero and a variance of σ2. and we can estimate σ2 with S2. We haven’t yet assumed any distribution (e.g., normal distribution) on the error term.

We know that OLS doesn’t require a normality assumption (e.g., the error term follows a normal distribution) to produce **unbiased estimates with the minimum variance** (aka, [**BLUE**](https://towardsdatascience.com/linear-regression-with-ols-unbiased-consistent-blue-best-efficient-estimator-359a859f757e)). Then

**Why are we motivated to assume the error term is normally distributed?**

One of the objectives of a linear regression model is to estimate the population parameter **β** using **β^** (OLS estimator), which is computed from the sample data. β^ itself is a random variable since it varies in different sample data. Therefore, knowing the **sampling distribution** of β^ allows us to calculate p-values for significance testing and generates reliable confidence intervals for OLS estimators.

With a little bit of math, we can show that if we assume that the errors are normally distributed in a linear regression model, the OLS estimators will be normally distributed as well.

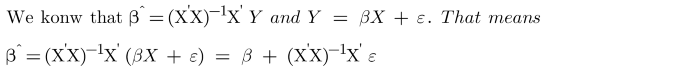


Image by author

In the above equation, **β** are fixed values. Conditional on X, the OLS estimators, **β^ are just a linear function of the error terms.** By assuming the error terms have a multivariate normal distribution, we are also implying the OLS estimators have a multivariate normal distribution.

https://miro.medium.com/max/700/1*ymQAX66xiM2TpnPKMzuFNA.png

Image by author

**Why is it reasonable to assume the errors are normally distributed in a Linear Regression Model?**

The normality assumption (i.e., the errors are normally distributed) is NOT as strong as it seems, it is often reasonable based on **Central Limit Theorem**.

The “**Central Limit Theorem for Sums**” states that the sum of many independent variables approximates a **normal distribution**, even if each independent variable follows different distributions.

You can apply this theorem to a linear regression model. We know that if we repeatedly draw different samples of size n (i.e., **repeated samples**), rerun the linear regression model, and compute the error value for the same X, we likely have different error values. If we put these error values together and draw a histogram. The distribution should look like a normal distribution.

We can think of it this way. Conditional on X=Xi, we can view εi (a variable) as **the sum of many other independent errors** from omitting important variables or pure randomness. Each of these other independent errors follows an unknown distribution.

Therefore, based on the Central Limit Theorem for Sums, the **sampling distribution** of the individual error term (**εi**) is a normal distribution. The error terms (i.e., ε1, ε2, …, εn) follow a “multivariate normal distribution” with a mean of zero and a constant variance of 𝜎2.

The normality assumption of the error terms is **NOT** related to the **sample size** in a linear regression model. It is simply due to the sources of many other independent errors influencing **a single observation.**

**How can we test whether or not the errors are normally distributed in a Linear Regression Model?**

We can implement **graphical residual analysis** to test if a linear regression model fits the data properly and if the errors are normally distributed. The following graphs are the common tools.

The **scatter plots** of the residuals vs the included explanatory variables and vs other potential explanatory variables (which are NOT included in the model) will allow us to access the **sufficiency** of the linear regression model. If a model fits the data well, the scatter plots will consist of random dots, which don’t show any **systematic structure** (e.g., trend and non-constant variation). Otherwise, any systematic structure might indicate the existing model can be improved in some way.

The **scatter plots** of the residuals vs predicted response variable also allow us to detect non-linearity, unequal error variances, and outliers in a linear regression model.

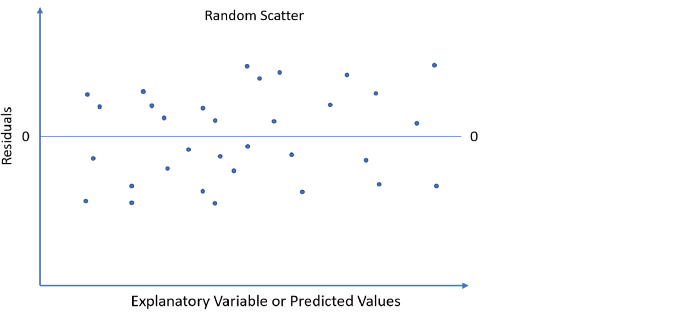


Image by author

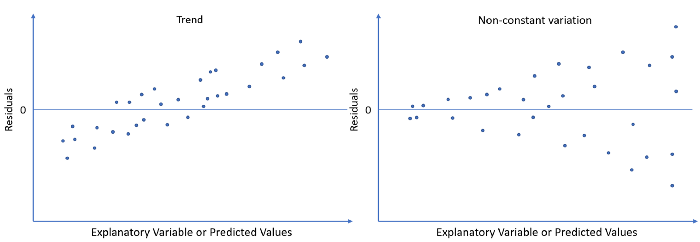


Image by author

The **histogram** and the **normal probability plot** of the residuals are often used to check if it is reasonable to assume the errors have a normal distribution and detect outliers.

This histogram is the most commonly used to show the frequency distribution of the residuals. If the error has a normal distribution, we should expect it to be more or less **bell-shaped**.

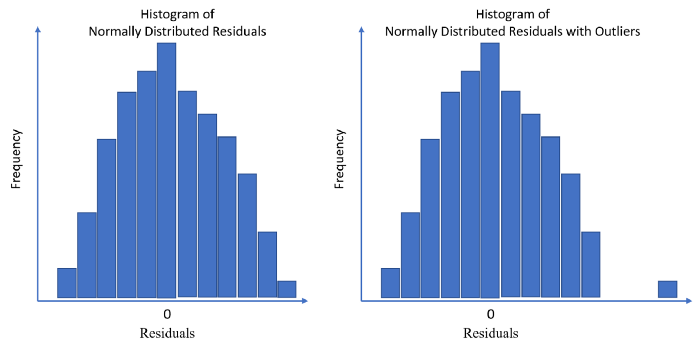


Image by author

The normal probability plot is also helpful to check the normality of a variable. It is constructed by graphing the sorted values of the residuals against the corresponding theoretical values from the standard normal distribution. **If the error is normally distributed, the plotted points should lie close to the straight line at a 45-degree angle.** Points that are far off the line might be outliers.

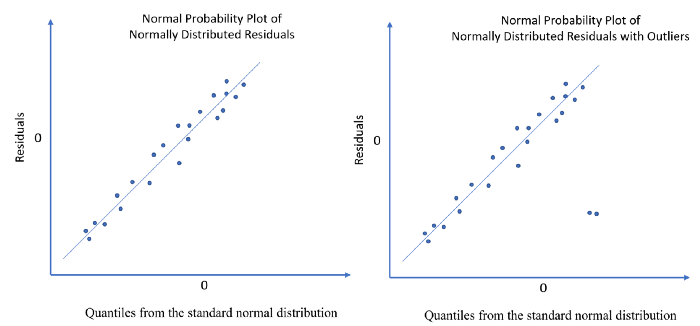


Image by author

Let’s investigate this topic with an example. In this example, we would like to study the relationship between “high school GPA” (X, i.e., 1.0, 2.0, 3.0, 4.0) and “college entrance test score” (Y). We can write their relationship in a linear regression model as

https://miro.medium.com/max/700/1*9XclnHVw0ski-EusdtTI3g.png

By applying the linear least squares method, we solve for the OLS estimators, **β^**. Then we can predict the fitted values for each X to construct the fitted line.

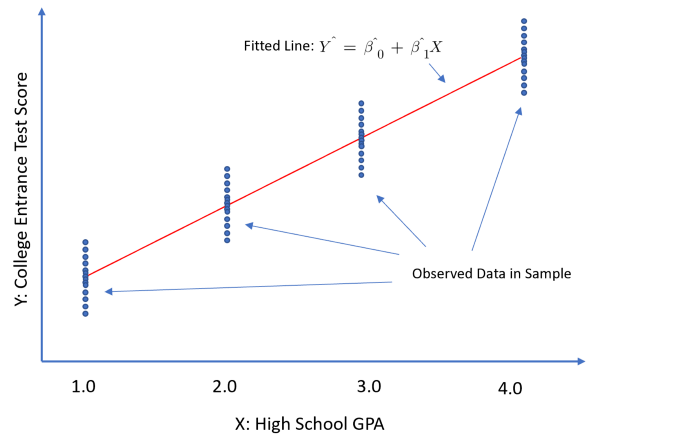


Image by author

If the relationship is properly captured by the linear regression model, we expect the scatter plot of residuals vs GPA to have random patterns, e.g., roughly the same center point and constant variation across GPAs. Conditional on X (i.e., for each X), the histogram of residuals should indicate a normal distribution.

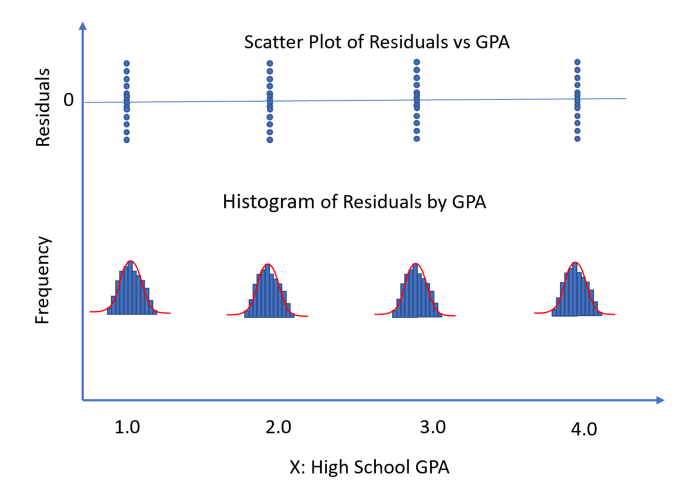


Image by author

Ideally, we would like to check the normality of residuals in each **group** (observations with the same values of X, or conditional on X). In practice, often there are not enough observations to draw a meaningful histogram per group. In that case, we can pool all residuals across groups to test normality. The aggregate histogram should also give us lots of information regarding the normality of the errors.

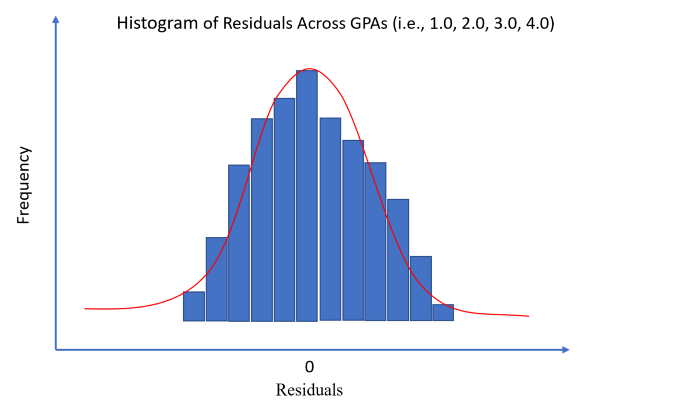


Image by author

**Conclusion**

Although normality assumption is optional to compute OLS estimates in a linear regression model, when we assume the errors are normally distributed, we can have a better idea about the precious of our estimates from the sample data.

Using the residuals to estimate the errors and implementing residuals analysis could help us to confirm our normality assumption and assure the appropriateness of the model.

If you would like to explore more posts related to **Statistics**, please check out my articles:

* [**7 Most Asked Questions on Central Limit Theorem**](https://towardsdatascience.com/7-most-asked-questions-on-central-limit-theorem-82e95eb7d964)
* [**Standard Deviation vs Standard Error: What’s the Difference?**](https://towardsdatascience.com/standard-deviation-vs-standard-error-whats-the-difference-ae969f48adef)
* [**3 Most Common Misinterpretations: Hypothesis Testing, Confidence Interval, P-Value**](https://towardsdatascience.com/the-most-common-misinterpretations-hypothesis-testing-confidence-interval-p-value-4548a10a5b72)
* [**Are the Error Terms Normally Distributed in a Linear Regression Model?**](https://towardsdatascience.com/are-the-error-terms-normally-distributed-in-a-linear-regression-model-15e6882298a4)
* [**Machine Learning: Bias-Variance Tradeoff and Regularization**](https://towardsdatascience.com/machine-learning-bias-variance-tradeoff-and-regularization-94846f945131)

**Thank you for reading !!!**

If you enjoy this article and would like to **Buy Me a Coffee,** please [click here](https://ko-fi.com/aaronzhu).

You can sign up for a [**membership**](https://aaron-zhu.medium.com/membership) to unlock full access to my articles, and have unlimited access to everything on Medium. Please [**subscribe**](https://aaron-zhu.medium.com/subscribe) if you’d like to get an email notification whenever I post a new article.