A Deep Dive into Assumptions of Linear Regression. Clearly Explained!

Unraveling the foundation of classical assumptions, part 1 of 2

Manoj Mangam · Follow

19 min read · Apr 15, 2023

Listen Share More

Let's understand through a Pizza Party Case! - Photo by Chad Montano on Unsplash

"*What are the assumptions of Linear Regression?*"

Sounds familiar ?!

Yes, it is a classic data-science interview question. On the face of it, linear regression may seem intuitive, but the devil is in the details.

Well, let's get under the hood and understand the details, in 2 parts, through a case. In part-1, let's look at the assumptions & their mathematical foundation, and in part-2 let's get a good grasp of the assumptions and their effects through a simulation in Python.

Ok! Let's dive right in.

> Ladies & Gentlemen, introducing the "**Pizza Party!**"

Meet **Alice** and **Salt-Bae**!

Alice is a data scientist who loves to build predictive models, while Salt-Bae is a pizza chef who wants to optimize his pizza production for parties in his beach resort.

> ***Case:*** *Alice and Salt-bae have been working together to build a model that predicts the* number of pizzas *that will be consumed at a party based on the* number of attendees*(consider 1 feature for simplicity). Salt-bae has hosted 100 parties (say the population) so far, but unfortunately, his boys kept track of the data of 10 parties. Now, Alice will see what she could do with the available sample (10 data points) to fit a linear regression model.*

Mathematically, the actual relationship between the dependent/target variable y(no. of pizzas) & the independent variable/feature/regressor X(no. of attendees) is shown below. It is known as the Population Regression Function(PRF). It well represents the complete population data (N=100 here).

$$y = f(X) + \epsilon \quad \{True\ or\ Population\ Regression\ Function\ (PRF)\}$$

The true relation f(X) between y(no. of pizzas consumed) & X(no. of attendees) is not known, and Alice's aim is to estimate/model the f(X). The ϵ refers to a random error or noise that cannot be systematically captured. We'll look at it in detail in the assumptions section. The estimate is denoted by f_hat(X) and is given below,

$$\hat{y} = \hat{f}(X) \quad \{Estimate\ of\ PRF\}$$

Alice: Salt-Bae can we have the party data?

Salt-Bae: "Sure! boys, furnish the data, please. Thank you. So, Alice, we have these 10 data points."

Note here that the available data or more formally the **sample** that we have is n=10.

Alice: "Ok. Let's see what we can do with the available data(sample)"

Alice: "Bae, first things first! We need to make some assumptions as we start to build our prediction model."

Salt-Bae: "Why do we need these assumptions?"

Alice: "Our predictions will be reliable under these assumptions"

> The assumptions are important to ensure that the linear regression model is reliable. Violations of these assumptions lead to some problems, which we will look into. So, it's important to check their validity before modeling, which we'll see in part 2 of 2.

> *There are many ways(estimators) to fit a linear regression function to the sample. The most commonly used one is **OLS (Ordinary Least Squares) estimator** because of its simplicity and some nice properties. It is a **Linear Least Squares (LLS)** estimator. Other estimators include MLE (Maximum Likelihood Estimation), Non-Linear Least Squares (NLLS), other LLS estimators such as Weighted Least Squares(WLS), Generalized Least Squares (GLS), etc.*

***Gauss–Markov theorem** states that if the linear regression model satisfies the classical assumptions, the Ordinary Least Squares (OLS) regression produces estimators that have the lowest sampling variance within the class of Linear Unbiased Estimators(LUE), which are known as **BLUE**(Best Linear Unbiased Estimator).*

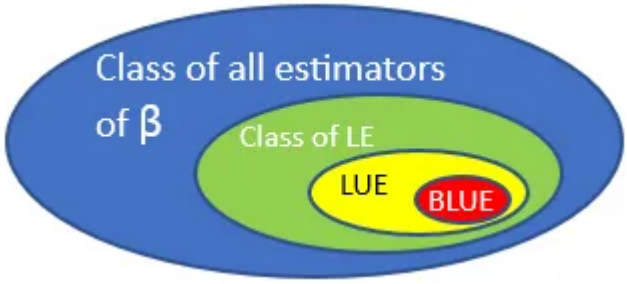

Image by author

*Btw, why are we interested in **BLUE (Best Linear Unbiased Estimator)**?*

It is because Alice & Salt-Bae will get the best possible predictions with BLUE. Let's drill it down,

At the end of the day, we want to accurately predict y for *unknown* Xs (usually known as *test set* in ML lingo). Mean Squared Error (MSE) is one of the metrics to quantify how good our predictions are. Since we do not know the true ys, the **expected test MSE** for the **predictor** is given below. So, we try to balance out the Bias & Variance of the model to get better & better predictions. This is famously known as the **Bias-Variance tradeoff.**

E(test MSE of predictor) = Variance(prediction) + [Bias(prediction)]^2 + Variance(error)

Formally,

$$E(test\ MSE\ of\ predictor) = E[(y_0 - \hat{f}(x_0))^2] = Var(\hat{f}(x_0)) + [Bias(\hat{f}(x_0))]^2 + Var(\epsilon)$$

So, the expected prediction error comprises of 3 components variance, bias, and variance of the random error, which is an irreducible error that we will come to terms with shortly, and it can be shown that the expected test MSE for the predictor is composed of **expected MSE of the estimators** of βs, and intuitively it makes sense as β_hats come closer to the true βs the prediction errors decrease. The **expected MSE of each estimator** is given by,

E(MSE of estimator) = Variance(estimator) + [Bias(estimator)]^2

Formally,

$$E(MSE\ of\ estimator) = E[(\hat{\theta} - \theta)^2] = Var(\hat{\theta}) + [Bias(\hat{\theta}, \theta)]^2$$

Now, the expected test MSE or expected prediction error can be deduced as below,

Expected Prediction Error α (Bias of estimators)² + (Variance of estimators)

We can observe that our prediction error decreases as the bias and/or variance of the estimators decreases. So, we get the best predictions if bias = 0 i.e., **Unbiased(U)** & Variance is minimum i.e., **Best(B)**. Hence our interest in **B_UE**. The missing **L** finds its way in BLUE because **Linear estimators** are intuitive & mathematically simple to handle, which we will see shortly.

> It does not necessarily mean that OLS estimates (predictions) are the absolute best, because there are other estimates such as ridge, lasso, etc., whose expected test MSEs are lower than that of OLS, under certain conditions such as multicollinearity & overfitting.

Ok! Time for the curtain raiser!

Assumptions of (OLS) Linear Regression:

There are **7 assumptions** of OLS regression, out of which **6 assumptions are necessary for OLS estimators to be BLUE**, and the 7th one is not necessary but it helps us in some ways, which we'll see. Note that the assumptions are generally made about the population.

> *First,* let's derive the assumptions *so that they seem more coherent than some discrete set of rules.* Then, let's discuss each one in detail.

Let's use the **BLUE framework** as a guiding light to derive & retain the assumptions.

> *Breaking **BLUE** (Best Linear Unbiased Estimator)*

Best — It means that an estimator(OLS) has the Least Variance among a class of estimators (LUE in our case). It means that the spread of β_hat is lower(precise) if we were to find it from different samples.

$$Var(\hat{\beta}_{OLS}) \leq Var(\tilde{\beta}), where\ \tilde{\beta}\ is\ any\ LUE$$

Linear- It means that an estimator(OLS) can be expressed as a linear function of observations (yis) as below,

$$\hat{\beta} = C\,y\,, where\ C\ is\ constant\ matrix\ of\ order\ (p+1)xn$$

Note that we will stick with the standard matrix-vector notations throughout.

Unbiased- It means that on average the β_hat estimates are equal to the true βs. Formally,

$$E(\hat{\beta}) = \beta$$

Now, let's **derive the assumptions** (through BLUE Framework) so that it will be easier for us to get into the details of each assumption later. *Let's start with the* ***L*** *in BLUE.*

L-Linearity of the estimator:

The OLS Regression yields a closed-form solution(meaning we can have a deterministic solution given in the form of an equation) when the Population Regression Function(PRF) y = β*X + ε is **linear in parameters,** which is the **assumption-1.** We will see why OLS regression doesn't yield a closed-form solution if PRF is non-linear in parameters. The solution of OLS regression is given below,

$$\hat{\beta} = (X^T X)^{-1} X^T y = C\,y$$

We can see that the OLS estimators are a linear combination of yis. It means that it is a Linear(L) estimator as per the Linear estimator definition.

Now, let's move to the **U** in BLUE.

U-Unbiasedness of OLS estimator:

First, let's prove that OLS estimators are unbiased under certain assumptions. Then, let's take a more generic path to prove the same, which will be useful later to prove the B in BLUE.

$$E[\hat{\beta}|X] = E[Cy\,|X] \qquad \{\because\ \hat{\beta} = Cy\}$$

$$E[\hat{\beta}|X] = E[C\,(X\beta + \epsilon)\,|X] \qquad \{\because y = X\beta + \epsilon\}$$

$$E[\hat{\beta}|X] = E[(\,CX\beta + C\epsilon\,)\,|X]$$

$$E[\hat{\beta}|X] = E[CX\beta\,|X] + E[C\epsilon\,|X]$$

Observe that we have considered expectation conditioned on X because in reality, the *regressors (Xs) can be stochastic*. Also, the math gets easier because we can consider X as a known or fixed quantity since conditioned on X means X is given. We have 2 terms in the last equation above. The first term is a constant because it is conditioned on X & note that the population parameter β is not random because it is characteristic of the population or data generation process, while β_hat (estimator of β) is stochastic because its value depends on the sample (some randomness involved).

$$E[\hat{\beta}|X] = CX\beta + C\,E[\epsilon\,|X]$$

Here comes our **2nd & 3rd assumptions E[ε|X] = 0**, which is called as the assumption of **Zero Conditional Mean of Errors.** It is often listed as **2 assumptions(Strict Exogeneity of Independent Variables** & **Expectation of each error is 0**) in literature. Let's get to them in the assumptions section.

$$E[\hat{\beta}|X] = (X^T X)^{-1} X^T X\beta \ \{\because C = (X^T X)^{-1} X^T\}$$

Here comes our **4th assumption**, (X^T X)^-1 is **invertible**, which is possible only when X is a **full-column rank** matrix or has **No Perfect Multicollinearity**.

We saw that we need 4 assumptions for the OLS estimator to be LUE (Linear Unbiased Estimator)

Now, let's consider a more generic approach to prove the U, whose result will be useful to prove B in BLUE.

Let's say β_tilde is any general Linear estimator of β, which has a different linear combination of y from that of the OLS estimator as shown below,

$$\tilde{\beta} = (C + D)y$$

where D is a (p+1) x n non-zero matrix. Now, take conditioned expectation on both sides,

$$E[\tilde{\beta}|X] = E[(C + D)y|X]$$

$$E[\tilde{\beta}|X] = E[(C + D)(X\beta + \epsilon)|X]$$

$$E[\tilde{\beta}|X] = E[(C + D)(X\beta + \epsilon)|X] \quad \{\because y = X\beta + \epsilon\}$$

$$E[\tilde{\beta}|X] = E[((C + D)X\beta + (C + D)\epsilon)|X] \quad \{\because E(X + Y) = E(X) + E(Y)\}$$

$$E[\tilde{\beta}|X] = E[(C + D)X\beta|X] + E[(C + D)\epsilon|X]$$

$$E[\tilde{\beta}|X] = (C + D)X\beta + (C + D)E[\epsilon|X]$$

Using the Zero Conditional Mean of Errors assumption,

$$E[\tilde{\beta}\,|X] = (C + D)X\beta$$

$$E[\tilde{\beta}\,|X] = ((X^TX)^{-1}X^T + D)X\beta$$

$$E[\tilde{\beta}\,|X] = ((X^TX)^{-1}X^T X\beta + DX\beta)$$

Using No Perfect Multicollinearity assumption,

$$E[\tilde{\beta}\,|X] = (I\beta + DX\beta) \quad \{\because (AB)^{-1} = B^{-1}A^{-1}\ \&\ AA^{-1} = I\}$$

$$E[\tilde{\beta}\,|X] = (I\beta + DX\beta) \quad \{\because (AB)^{-1} = B^{-1}A^{-1}\ \&\ AA^{-1} = I\}$$

$$E[\tilde{\beta}\,|X] = (I + DX)\beta$$

$$\Rightarrow E[\tilde{\beta}\,|X] = \beta \quad iff\ DX = 0_{(p+1)x(p+1)}$$

We will use the result DX=0 below.

Let's move to the **B** in **B**LUE. To do so, we need 2 more assumptions.

Best-ness (least variance among LUEs) of OLS estimator:

Let's see why the Gauss-Markov theorem says, OLS is the Best LUE. *Note that we cannot relax the LUE part as there can be biased estimators with variances less than that of the OLS estimator.*

$$Var(\tilde{\beta}|X) = Var((C + D)y\,|X)$$

$$Var(\tilde{\beta}|X) = (C + D)\,Var(y\,|X)(C + D)^T \qquad \{\because Var(aA) = aVar(A)a^T\}$$

We know that,

$$y = X\beta + \epsilon \Rightarrow Var(y|X) = 0 + Var(\epsilon|X) = \sigma^2 I$$

Here comes our **5th & 6th assumption(s)** i.e., **Var(ε|X) = σ^2 I**, which is known as the assumption of **Spherical Errors**. This assumption leads to 2 results which are often listed as separate assumptions in the literature. They are **Homoskedasticity** & **No autocorrelation of errors**. Let's dive into them shortly.

> *C'mon!! We have derived all the 6 necessary assumptions of (OLS) Linear Regression.*

Let's continue and prove the B in BLUE.

$$Var(\tilde{\beta}|X) = (C+D)\sigma^2 I(C+D)^T$$

$$Var(\tilde{\beta}|X) = \sigma^2((X^TX)^{-1}X^T + D)(X(X^TX)^{-1} + D^T) \ \{using\ (AB)^T = B^TA^T\}$$

$$Var(\tilde{\beta}|X) = \sigma^2((X^TX)^{-1}X^TX\,(X^TX)^{-1} + (X^TX)^{-1}(DX)^T + DX(X^TX)^{-1} + DD^T)$$

$$Var(\tilde{\beta}|X) = \sigma^2((X^TX)^{-1} + (X^TX)^{-1}(DX)^T + DX(X^TX)^{-1} + DD^T)$$

We saw that DX = 0 for LUE. Let's plug that in.

$$Var(\tilde{\beta}|X) = \sigma^2(X^TX)^{-1} + \sigma^2 DD^T$$

We can show that,

$$Var(\hat{\beta}_{OLS}|X) = \sigma^2(X^TX)^{-1}$$

Proofs involving ordinary least squares — Wikipedia

$$Var(\tilde{\beta}|X) = Var(\hat{\beta}_{OLS}|X) + \sigma^2 DD^T$$

$$Var(\tilde{\beta}|X) \geq Var(\hat{\beta}_{OLS}|X) \quad \{\because AA^T\ is\ a\ positive\ semi-definite\ matrix\}$$

Yes, we have made it to **BLUE**!

Note that an unbiased estimator with the least variance is known as an **efficient estimator**.

> Hence, OLS estimator is an efficient estimator under the assumptions.

Observe that we have not assumed anywhere so far that the random errors are normally distributed.

> *N*ow, let's look at each of these assumptions in detail.

Assumption 1: Linearity — This is an assumption about the PRF(Population Regression Function) i.e., about the **f(X)** in **y=f(X)+ε.** It means that the true underlying relationship between the target variable & the independent variables is **linear in parameters.** I know this sounds a bit weird, but it simply means that the PRF should be as given below,

$$y = parameter_0 + parameter_1 * X_1 + parameter_2 * X_2 + \cdots + \epsilon$$

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \epsilon$$

Some examples of *non-linear models* are as below,

$$y = \beta_0 + \beta_1 X_1 + \beta_1 \beta_2 X_2 + \epsilon \quad \ldots not\ linear\ in\ parameters$$

$$y = \beta_0 + (0.4 - \beta_0) e^{-\beta_1 (X_1 - 5)} + \epsilon \quad \ldots not\ linear\ in\ parameters$$

Note that the true relationship should be linear in parameters, not in Xs. It means that the models that have Xs raised to an exponent as X^2, X^3, etc., do not violate the Linearity assumption. Moreover, such models result in non-linear looking plots between y vs X as shown in the right image below. So, non-linear looking plots do NOT necessarily violate the Linearity assumption. In fact, both the models below are linear in parameters (IKR! Looks can be deceiving!)

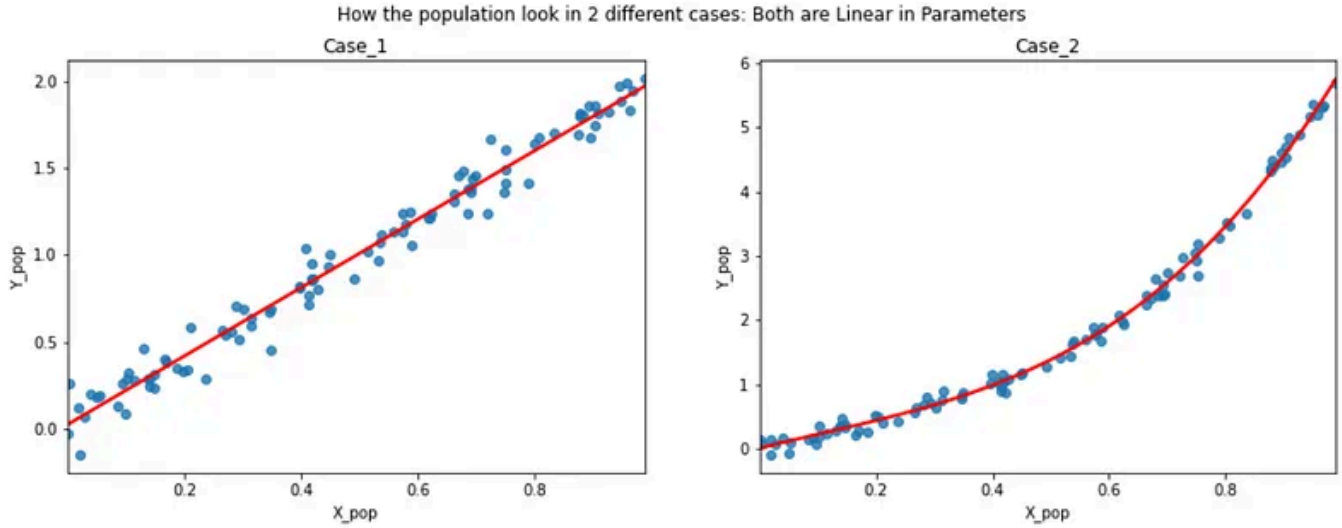

Image by author

So, how in the world are we going to know from the sample whether this assumption holds true? And when to choose non-linear regression over linear regression. Let's look at it in the Python simulation in part-2.

What happens if we violate the "Linearity" assumption?

First & foremost, we cannot get a nice-looking closed-form solution for the estimators as below because while minimizing the loss function w.r.t β_hats the first-order partial derivatives get complicated and cannot be algebraically manipulated to get a closed form, and in general we obtain the parameters (β_hats) in this case with successive approximation, which is in an iterative manner called. And ofcourse, we cannot prove the OLS estimator to be LUE (refer derivation section above).

$$\hat{\beta}_{OLS} = (X^T X)^{-1} X^T y$$

Assumptions 2 & 3: Zero Conditional Mean of Errors (Strict Exogeneity)

As shown below, it means that the expectation of the random error conditioned on X is 0. It leads to 2 results that we will consider as assumptions 2 & 3. First, let's list down the assumptions, and then we will get to the basics of the random error.

$$E[\varepsilon \mid \mathbf{X}] = \begin{bmatrix} E[\varepsilon_1 \mid \mathbf{X}] \\ E[\varepsilon_2 \mid \mathbf{X}] \\ \vdots \\ E[\varepsilon_n \mid \mathbf{X}] \end{bmatrix} = \mathbf{0}$$

Assumption 2: (Unconditional) Expectation of Error is Zero

$$E[\epsilon_i] = 0 \ \ \forall i = 1 \dots n$$

It can be shown using the law of total expectation that the unconditional expectation of the random error is 0. Intuitively, we do not want the random error to be non-zero so that we can have more certainty in our predictions.

$$E[\epsilon_i] = E\big[E[\epsilon_i|X]\big] = E[0] = 0 \ \{using\ law\ of\ total\ expectation\}$$

Assumption 3: Strict Exogeneity of Independent Variables

$$Cov[X, \epsilon_i] = 0 \ \ \forall i = 1 \dots n$$

Strict exogeneity means that the **X is not correlated with the error**, meaning we should not get any hint of the error based on the value(s) of X. Intuitively, we do not want the random error to change(correlated) with the independent variable. Let's derive it,

$$Cov[X, \epsilon_i] = Cov[X, [E[\epsilon_i|X]] \ \ \{using\ law\ of\ total\ covarinace\}$$
$$Cov[X, \epsilon_i] = Cov[X, 0] \ \ \{\because \ E[\epsilon_i|X] = 0\}$$
$$Cov[X, \epsilon_i] = 0 \ \ \forall i = 1 \dots n$$

The Assumption can also be presented as,

$$E[X^T \epsilon] = 0 \ \ \{\because Cov(A, V) = E[A^T V] - (E[A])^T E[V]\}$$

Now, let's get to the basics of the stochastic error.

Salt-Bae: "Firstly, what does this error mean?"

Alice: "The random error simply means that however good our model is, it cannot perfectly predict. So, there's always an error in our predictions that the model simply cannot account for."

Formally, the random error is the epsilon(ε) in the population regression function below,

$$y = X\beta + \epsilon$$

Visually, we can see the blue dots dancing around the population line (red) below because of the random error/noise.

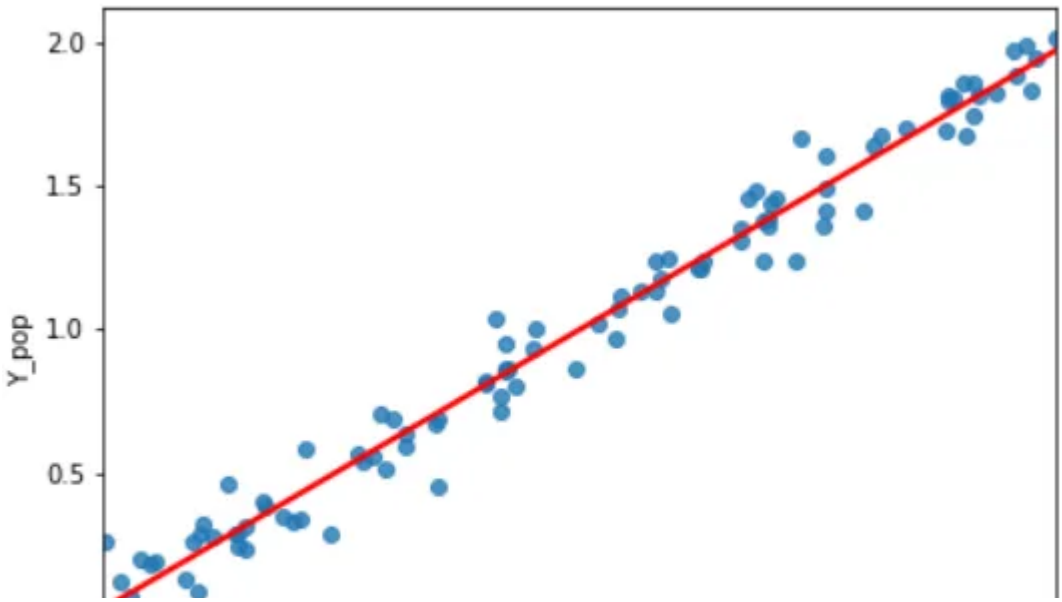

Image by author

As shown below, at best the model on average, given the assumptions, can completely account for Xβ but not for the ε because it is totally random and it is not systematic. So, it means that on average our predictions will be equal to the true values if E(ε|X) = 0. Otherwise, the model will be biased.

$$E[y|X] = X\beta + E[\epsilon|X] \quad \{using\ y = X\beta + \epsilon\}$$

$$E[y|X] = X\beta$$

$$\hat{y} = X\hat{\beta} \quad \{(OLS)linear\ regression\ model\}$$

$$E[\hat{y}|X] = X\,E[\hat{\beta}\,|X] = X\beta \ \{\because OLS\ estimator\ is\ Unbiased\}$$

Observe below that the model's expected test MSE has an irreducible component due to the random error. It means the expected model's error cannot be below Var(ε).

$$E(test\ MSE) = E\left[\left(y_0 - \hat{f}(x_0)\right)^2\right] = \underbrace{Var\left(\hat{f}(x_0)\right) + \left[Bias\left(\hat{f}(x_0)\right)\right]^2}_{\text{Reducible error}} + \underbrace{Var(\epsilon)}_{\text{Irreducible error}}$$

Salt-Bae: "But, how does the random error creep into our system?"

Alice: "It can creep in due to various reasons. Say, the boys messed up counting the no. of attendees at a party. That's a measurement error which the model cannot account for."

Salt-Bae: "Oh! Looks like I need to have some guidelines in place"

Alice: "Yes. Also, we are trying to predict the no. of pizzas consumed based on 1 variable i.e., no. of attendees at a party. Other omitted variables such as proportion of children, other food options (appetizers, drinks, deserts etc.), time, day of the week, type of party (birthday, farewell, etc), weather, type of music, etc., can result in the error."

> So, essentially ε captures the inherent noise present in the data. The noise can be due to measurement errors, omitted variables (called as Omitted Variable Bias), etc.

Salt-Bae: "Looks like I need to have a data acquisition system in place."

Alice: "Yes. That would help us limit bias in our predictions."

Salt-Bae: "What does it mean to have the average error equal to 0"?

Alice: "It means that on an avg. the boys are prone to equally over-account or under-account the no. of attendees for all kinds of pizza parties (small to big)"

> What happens if we violate the "Zero Conditional Mean of Errors" assumption?

It leads to biased estimates or predictions. It is because we cannot prove that OLS estimators are Unbiased unless we have this assumption in place (refer to the assumptions derivation section).

Assumption 4: No Multicollinearity (Full Column Rank)

It means that no feature can be derived from a linear combination of other features, simply we should not have redundant information. Mathematically, all the features should be linearly independent, or X must have full column rank.

$$X_p = \alpha_1 X_{p-1} + \alpha_m X_{p-m} + \alpha_k X_{p-k} + \ldots \quad \{\text{under Multicollinearity}\}$$

$$\alpha_0 X_p + \alpha_1 X_{p-1} + \alpha_2 X_{p-2} + \ldots = 0 \quad \{\text{under No Multicollinearity}\}$$
$$\text{iff } \alpha_0 = \alpha_1 = \alpha_2 = \cdots = 0$$

I discussed about Multicollinearity in detail in the article "Multicollinearity problems in Linear Regression. Clearly Explained."

> What happens if we violate the "No Multicollinearity" assumption?

Firstly, OLS estimators will be biased, and the violation would result in Exact Multicollinear Condition where the OLS estimators become unreliable for a variety of reasons. Multicollinearity leads to high variance of β_hats, incorrect inferencing as the p-values may paint a false picture, and interestingly the confidence intervals may not even include the true β value. Also, the matrix (X^T X)^-1 is not invertible under exact multicollinearity, which means that there's no analytical solution. In practice however, the β_hats are not solved for using the normal equation, but using gradient descent algorithm. Multicollinearity is a problem even in this case as it leads to solution convergence issues because the cost function tends to have more than one local minima or is flatter.

Assumptions 5 & 6: Spherical Errors (Homoskedasticity & No Autocorrelation of Errors)

$$Var[\epsilon|X] = \sigma^2 I_n = \begin{bmatrix} \sigma^2 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma^2 \end{bmatrix}$$

Where σ^2 is the variance of each error ε_i & I*n* is identity matrix of size n = no. of observations in the sample.

Spherical Errors leads to 2 results. Let's call them assumptions 5 & 6. Spherical Errors is a condition in which all the errors have the same conditional variance (diagonal terms), while the conditional covariance of the errors are all zero (off-diagonal terms) as shown in the conditional variance-covariance of the random error above.

Let's derive assumptions 5 & 6 from Spherical Errors assumption,

Assumption 5: Homoskedasticity

Homoskedasticty is a condition in which the conditional variance of error remains constant across the instances of X. This condition is evident from spherical errors that the diagonal terms are equal & constant.

$$Var[\epsilon|X] = E[\epsilon\epsilon^T|X] = \sigma^2 I_n \quad \{\because Var(A) = E[(A - E(A))(A - E(A))^T]\}$$

$$Var[\epsilon_i|X] = E[\epsilon_i^2|X] = \sigma^2 \quad \forall i = 1 \dots n$$

Salt-Bae: "What does it mean to us?"

Alice: "In case of violating the assumption, for bigger parties things can go quite out of hands and the boys can report the *no. of attendees* more erroneous than that of for the smaller parties. It means that the error spread for bigger parties will be higher. It is a concern that we need to address."

Salt-Bae: "Hmm. Looks like I need to find a way to automate the data-acquisition process."

What happens if we violate the "*Homoskedasticity*" assumption?

This leads to a situation called ***heteroskedasticity*** *that results in OLS estimators with higher variance. Hence, they will no longer be Best-LUE or efficient. In this case, a more efficient estimator would be Weighted Least Squares(WLS) or we can do certain transformations on the sample if we know the form of heteroskedasticity. We'll discuss more on this in part-2.*

Assumption 6: Errors Are Non-Autocorrelated

This condition is evident from the error conditional variance-covariance matrix i.e., Var(ε|X) that the off-diagonal terms are all 0, which means that the covariance between any 2 given errors is 0, implying that the errors are not correlated.

$$Cov[\epsilon_i, \epsilon_j|X] = E[\epsilon_i \epsilon_j|X] = 0 \; \forall i \neq j \quad \{using \; Cov(a,b) = E[(a - E(a))(b - E(b))]\}$$

It means that the errors should be completely random i.e., if the error for one observation is positive that should not increase the probability for the next error to be positive (positive correlation), similarly on the negative side(negative correlation).

Salt-Bae: "How it may be violated in our case?"

Alice: "Say, if the boys are over-accounting the *no. of attendees* for parties that are beyond a certain limit(party size) and under-accounting for parties below a certain limit(party size)"

> What happens if we violate the "Errors are Non-autocorrelated" assumption?

The violation would result in imprecise predictions as we get β_hats with higher variance. Simply, we can't obtain Best-LUE.

Assumption 7: Errors are Normally Distributed

This is the only assumption that is not necessary for the OLS estimators to be BLUE, but it will help us perform hypothesis testing, find out reliable confidence intervals, and hence reliable inferencing about the population. Also, it extends the list of good properties of the OLS estimators. The assumption is,

$$\epsilon|X \sim \mathcal{N}(\vec{0}, \sigma^2 I_n)$$

This means that the random error of each observation yi is normally distributed with mean 0 and variance σ^2.

But, why the errors are assumed to be normally distributed and not any other distribution?

The random errors can be thought of as a cumulative effective of lot of independent noises (omitted variables, measurement errors, etc). We know from the **Central**

Limit Theorem (CLT) that the sum of many independent random variables (irrespective of their distribution) follows normal distribution. That's why it makes sense to assume Normal distribution and not any other distribution.

Why will it help us with hypothesis testing & inferencing?

It helps because this would result in the β_hats sampling distribution to be normal, which simplifies the statistical inferencing. Infact, it makes it possible to use familiar & simpler parametric tests such as t-tests and F-tests.

$$y = X\beta + \epsilon$$

$$\Rightarrow y|X \sim \mathcal{N}(X\beta, \sigma^2 I_n)$$

Now, since β_hat is a linear combination of y i.e., β_hat = Cy, it follows that

$$\Rightarrow \hat{\beta}|X \sim \mathcal{N}(\beta, \sigma^2(X^TX)^{-1}) \quad \{\because \hat{\beta} \text{ is Unbiased \& } Var(\hat{\beta}|X) = \sigma^2(X^TX)^{-1}\}$$

Now that we have the sampling distribution of β_hat, we can construct statistical significance tests to infer the importance of a particular β_hat and thereby the corresponding feature significance. Since, we do not the actual σ^2, we estimate it using MSE of the model using the residuals(y-y^), which are normally distributed because we can show that they are a linear combination of ε. After estimating σ^2, we'd end up with a t-statistic for testing any given β_hat's significance. This is the base for the p-values that we see in the model's summary in python.

Additional benefits of this assumption:

Under the normality assumption, the OLS estimators reach the Cramer-Rao bound, meaning they have the lowest variance among all the Unbiased Estimators (both Linear & Non-Linear Estimators) i.e., BUE (Best Unbiased Estimator). Moreover, the normality assumption allows the OLS estimators to be equivalent to MLE (Maximum Likelihood Estimator), because here OLS estimation essentially maximizes the likelihood of observing y given X and the parameter β_hat. MLE has its merits (asymptotic properties) as the sample size grows bigger and bigger.

> What happens if we violate the "Errors are Normally Distributed" assumption?

The violation will lead to unreliable hypothesis tests and hence unreliable inferencing about the population. Though the OLS estimators will still be BLUE, but they will lose the additional benefit of being the Best among all the Unbiased Estimators i.e., BUE (Best Unbiased Estimator).

Phew! Finally, we have made it!

> In this part 1 of 2, we looked at the foundation of each assumption, derived them, and looked at each assumption closely. In part 2, let's look at these assumptions in action on the pizza party's dance floor through a python simulation, where we will look at how to check their validity, remedial measures under violation, etc., and help Alice & Salt-Bae plan their Pizza production for the next party.

Conclusion

There are 7 assumptions in the classical (OLS) linear regression. While the first 6 assumptions are required for the OLS estimators to be **BLUE** (Best Linear Unbiased Estimators), the last Normality assumption helps us with reliable hypothesis testing & inferencing. If any of these 3 assumptions viz., Linearity, Zero Conditional Mean of Errors, & No Perfect Multicollinearity is violated, we *cannot attain LUE in BLUE*. If the Spherical Errors assumption is violated, we *cannot attain B in BLUE*.

I hope you've enjoyed reading the article.

Until Next Time!

Photo by Max van den Oetelaar on Unsplash

References

[1] Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction (2nd ed.). Springer.

[2] James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An Introduction to Statistical Learning with Applications in R (Springer Texts in Statistics) (1st ed.). Springer.

[3] Montgomery, D. C., Peck, E. A., & Vining, G. G. (2012). Introduction to Linear Regression Analysis (5th ed.). John Wiley & Sons.

[4] Greene, W. H. (2012). Econometric Analysis (7th ed.). Pearson.

[5] Wooldridge, J. M. (2012). Introductory Econometrics: A Modern Approach (5th ed.). Cengage Learning.

[6] Sheather, S. J. (2009). A Modern Approach to Regression with R. Springer.

[7] 7 Classical Assumptions of Ordinary Least Squares (OLS) Linear Regression — Statistics By Jim