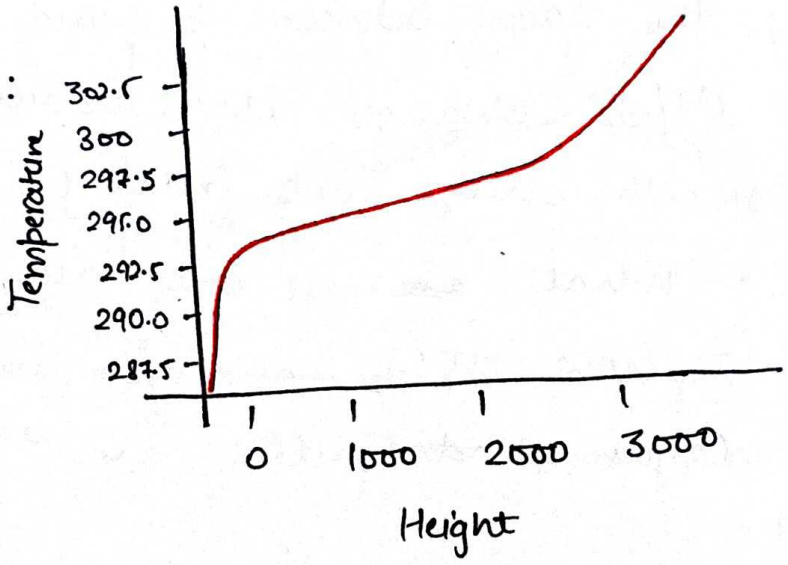
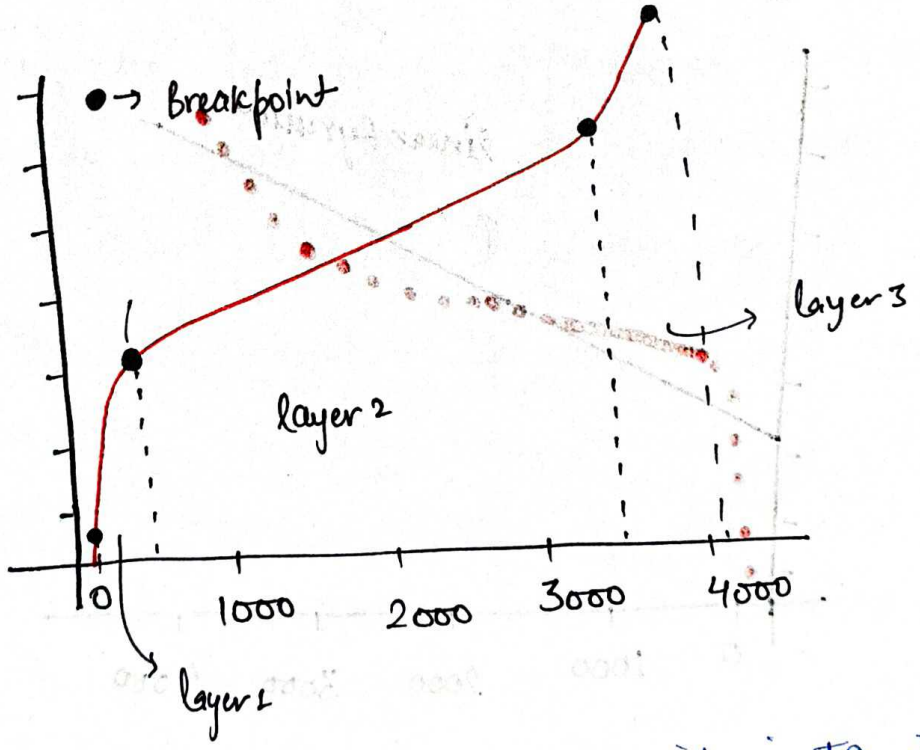
Piecewise Regression

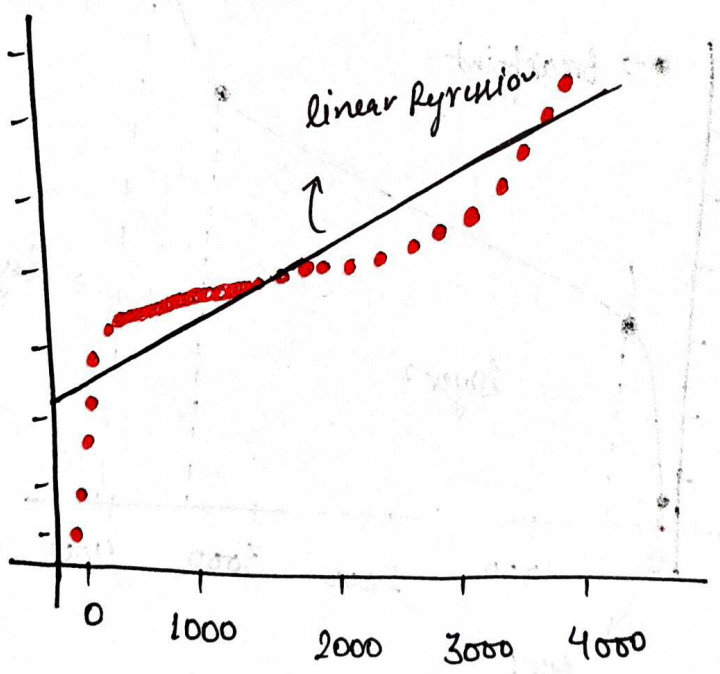
Problem:





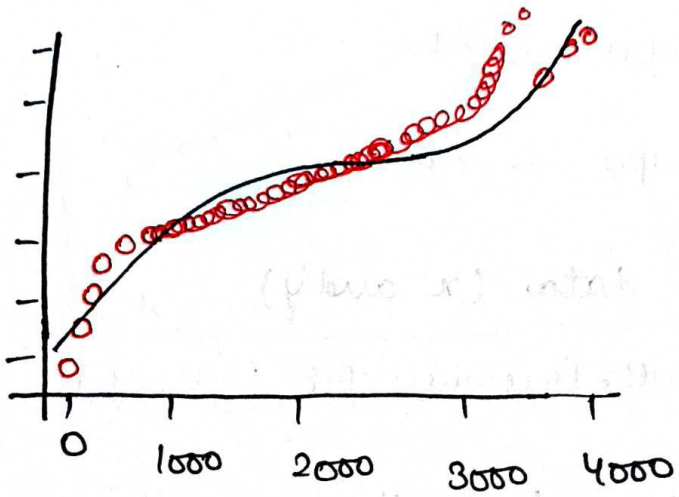
In fact, we can visually divide it into 3 different layer (or piece). Above we can see the presentation of these layers and its approximate height (x-axis). The break points are the points when the data

completely changes its behaviour. For instance, break point 1 (let's denote it b1) is our start point; the layer between b1 and b2 has a similar dT/dz changes its behaviour to a more smooth change not increasily exponentially like before. What we are trying to understand is how dT/dz changes inside of each these layers and where are the break points.



if we plot our data and try to fit a regression on it we fail miserably. However, it doesn't mean that a linear regression is not the best choice for our problem,

Polynomail Regression



The problem of polynomail regression is that you lose the Interpretability (human can understand the cause of a decision.) of the model when adding the polynomail terms (quadratic, cubic, etc). So if you dont care so much about interpretability, you can stop reading here. However if you need interpretability to deeply understand the problem, piecewise linear regression is your buddy.

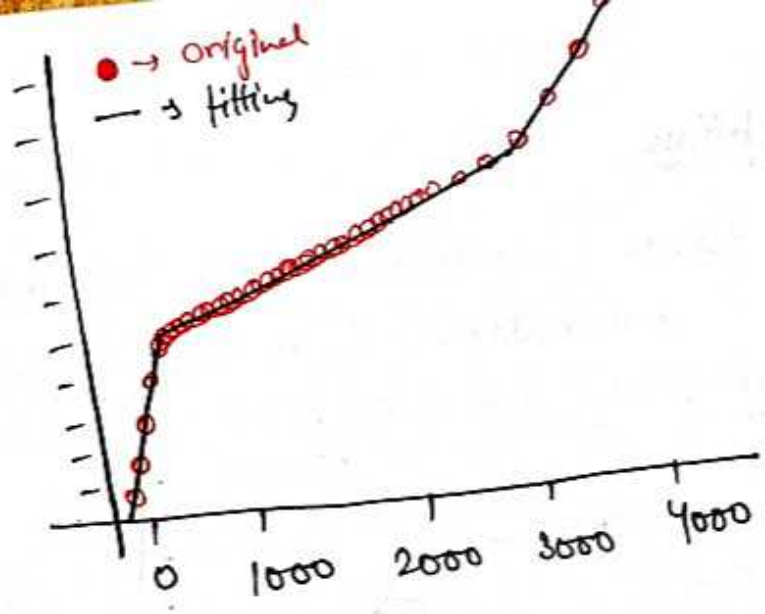
Piecewise Linear Regression

$$y(x) = \begin{cases} n_1 + \beta_1 (x - b_1), & b_1 < x \leq b_2 \\ n_2 + \beta_2 (x - b_2), & b_2 < x \leq b_3 \\ \dots \\ n_m + \beta_m (x - b_{m-1}), & b_{m-1} < x \leq b_m \end{cases}$$

Code:

```
import pwlf
import pandas as pd
import numpy as np

# fit your data (x and y)
myPWLF = pwlf.PiecewiseLinearFit(x, y)   # x -> df[col_name], y -> df[col_name]

# fit the data for n line segments
z = myPWLF.fit(3)

# Calculate slopes
slopes = myPWLF.calc_slopes()

# predict for the determined points
xHat = df[column_name]
yHat = myPWLF.predict(xHat)

# Calculate statistic
p = myPWLF.p_value(method='non-linear', step_size=1e-4)

se = myPWLF.se
```

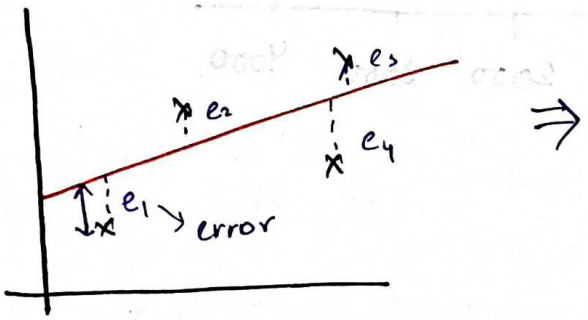
Standard error

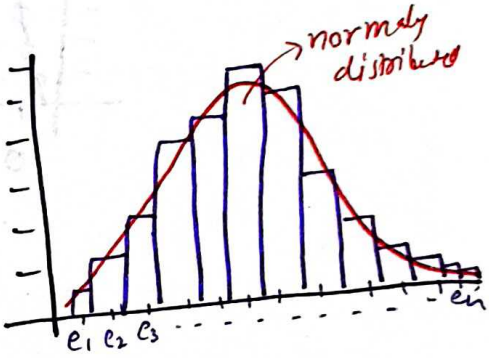

● → Original
— → Fitting
0
1000
2000
3000
4000

2. Normality Of Residual

The Assumption

The error term (residual) are assumed to follow a normal distribution with a mean of zero and a constant variance.





What happens when this assumption is violated ?

1. **Inaccurate hypothesis tests** : The t-tests and f-tests used to assess the significance of the regression coefficients and the overall model rely on the normality assumption. If the residuals are not normally distributed, these tests may produce inaccurate results, leading to incorrect inferences about the significance of the independents variables.

2. **Invalid Confidence Intervals** : The confidence interval for the regression coefficients

and the overall are based on the assumption of normally distributed residuals. If the normality assumption is violated, the confidence intervals may not be accurate, affecting the interpretation of the effect sizes and the precision of the estimates.

3. **Model performance**: The violation of the normality assumption may indicate that the chosen model is model is not the best fit for the data, potentially leading to reduced predictive accuracy.

How to check this assumption?

1. **Histogram of Residuals**: Plot a histogram of a residual to visually access their visually access their distribution. If the histogram resembles a bell shaped curve. It suggested that the residual are normally distributed.

2. **Q-Q plot**

3. **Statistical Test**: Statistical tests like Omnibus test, Jarque – Bera test or even shapiro wilk test can test this assumption.

What to do when the assumption fail?

1. **Model selection technique**: Employ model selection techniques like Cross-validation, AIC, or BIC to choose the best model among different candidate models that can handle non-normal residuals.

2. **Robust regression**: Use robust regression techniques that are less sensitive to the distribution of the residuals, such as M-estimation, least Median of square (LMS), or least Terminal Square (LTS). (Transformation may also help)

3. **Non-parametric or semi-parametric methods**:

4. **Use bootstrapping**

Remember that the normality of Residual assumption is not always critical for Linear Regression, especially, when the sample size is large, due to Central limit theorem (n more than 30)

↳ Automatically Normal if $n > 30$

↳ no. of observations.

Omnibus Test

The omnibus test is a statistical test used to check if the residuals from a linear regression model follow normal distribution. The test is based on the skewness and kurtosis of the residuals. Here is step by step guide on how to conduct the omnibus test.

1. **Decide the Null and Alternate Hypothesis**: The Null Hypothesis states that the residuals are normally distributed and the Alternate Hypothesis says that the residual are not normally distributed.

2. **Fit the linear Regression model**: Fit the linear regression model to your data to obtain the predicted values.

3. **Calculate the Residuals**: Compute the residuals (error terms) by substracting the predicted values from the observed values of the dependent variable.

4. **Calculate the skewness**: Calculate the skewness of the residuals. Skewness measures the asymmetry of the distribution. For a normal distribution skewness is expected to the close to zero.

5. **Calculate the kurtosis**: Calculate the kurtosis of the residuals, kurtosis the 'tailedness' of the distribution. For a normal distribution, kurtosis is expected to be close to zero (in excess kurtosis terms).

6. **Calculate the ~~Omnib~~ Omnibus test statistic**: Compute the Omnibus test statistic (K^2) using the skewness and kurtosis values. The formula for the Omnibus test statistic is:

chi-square $df=2$

$$K^2 = n\left[\frac{(\text{Skewness})^2}{6} + \frac{(\text{Kurtosis})^2}{24}\right]$$

n = number of observation

7. **Determine the p-value**: The Omnibus test statistic follows a chi-square distribution with 2 degree of freedom. Use this distribution to calculate the p-value corresponding to the test statistic.

8. **Compare the p-value to the significance level**: Compare the p-value obtained in step 6 to your chosen significance level (e.g., 0.05). if the p-value is greater than the significance

level, you can accept the null hypothesis that the residuals are normally distributed. If the p-value is similar than the significance level, you reject the null hypothesis suggesting that the residual may not follow a normal distribution.

Shapiro - Wilk Test

The Shapiro-wilk test is a hypothesis test that is applied to a sample with a null hypothesis that the sample has been generated from a normal distribution. If the p-value is low, we can reject such a null hypothesis and say that the sample has not been generated from a normal distribution.

But it has one flaw: It doesn't work well with large datasets. The maximum allowed size for a data set depends on the implementation (5000)

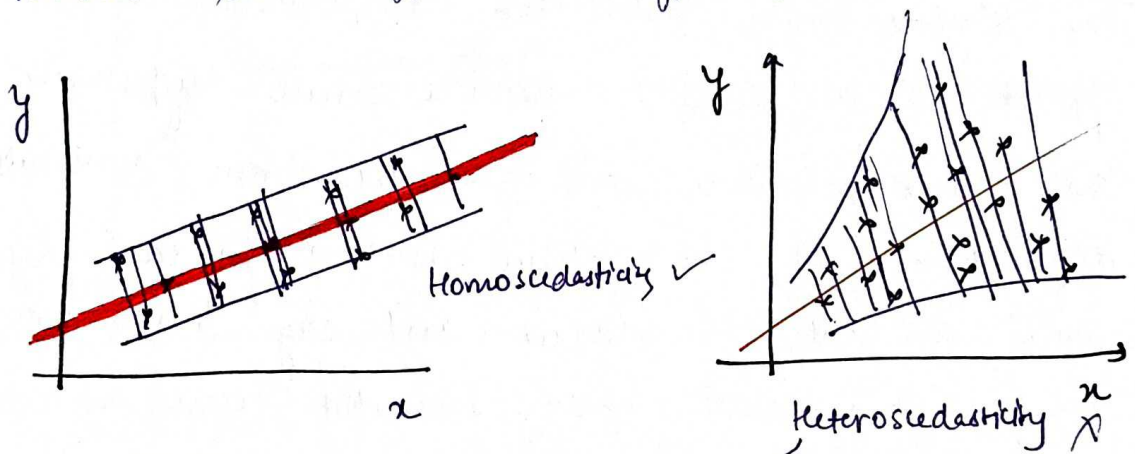
(i) P-Value

(ii) Statistic is nearly zero for normality.

3. Homoscedasticity

The Assumption

The spread the error term (residuals) should be constant across all level of the independent variable. If the spread of the residuals changes systematically. It leads to Heteroscedasticity, which can affect the efficiency of the estimate.



→ 1. Problem: Coefficient of Standard Error (SE) is no reliable. If Standard Error is not reliable then t-statistic is reliable and P-value also reliable.

2. Problem: Using not reliable standard error to find confidence Interval.
(not reliable)

What happens when this assumption is violated?

1. **Inefficient estimates**: While the parameter estimate (coefficient) are still unbiased, they are no longer the best linear unbiased estimator (BLUE) under heteroscedasticity. The inefficiency of the estimates implies that the standard errors are than they should be, which may reduce the statistical power of hypothesis.

2. **Inaccurate hypothesis tests**: The t-tests and f-tests used to assess the significance of the regression coefficient and the overall model rely on the assumption of heteroscedasticity, then tests may produce misleading results, leading to incorrect inferences about the significance of the independent variables.

3. **Invalid Confidence Intervals**: The confidence intervals for the regression coefficients are based on the assumption of homoscedastic residuals if the homoscedasticity assumption is violated, the confidence intervals may not be accurate, affecting the interpretation of the effect size and precision of the estimate.

How to check this assumption ?

1. Residual Plot:

2. Breusch-Pagan test: This is a formal statistical test for heteroscedasticity. The null hypothesis is that the error variance

What to do when the Assumption fails?

1. Transformation

2. Weighted least Square (WLS): Use a weighted least squares approch, which assigns different weights to the observations based on the magnitude of their residuals. This methods can be help account for heteroscedasticity by giving more importance to observation with smaller residuals and less importance to those with larger residuals.

3- Robust standard errors: Calculate robust (heteroscedasticity - consistent) standard errors for the regression coefficients. These standard errors are more reliable under heteroscedasticity and can be used to perform more accurate hypothesis tests and construct valid confidence intervals.

Standard Error

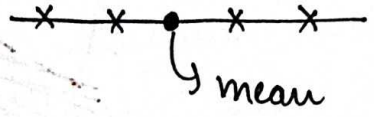
Database 3000 = μ

↳ Population mean

$$\mu = \frac{\sum x_i}{n}$$

Std Deviation

$$\sigma^2 = \frac{\sum (x_i - \bar{x})}{n}$$



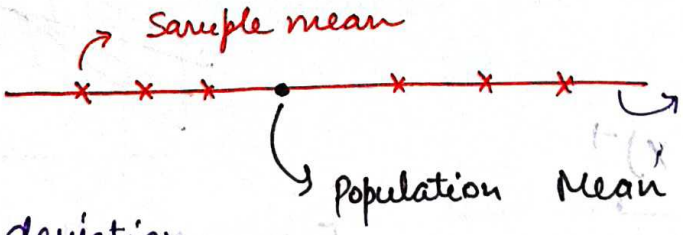
avg dist betn the mean

* If standard deviation ↓ (small) means that all points are close (****)
* If standard deviation ↑ (large) means not all points are not close (* * - *)

100 = $\bar{x}_1$ (sample data 1)

100 = $\bar{x}_2$ (Sample data 2)

100 = $\bar{x}_3$ (Sample data 3 . . .)



Find Standard deviation of this line is called Standard error.

* Standard deviation of the Sample means is standard error

$$SE = \frac{\sigma}{\sqrt{n}}$$

σ → pop std deviation

n → no. of observation

Central Limit theorem

↳ when extracting sample from Population at many times and find of mean of each sample individually and That distribution is called sample distribution because we are assum

Sample distribution more than 30 and also called Normal Distribution ($\mu, \sigma/\sqrt{n}$)

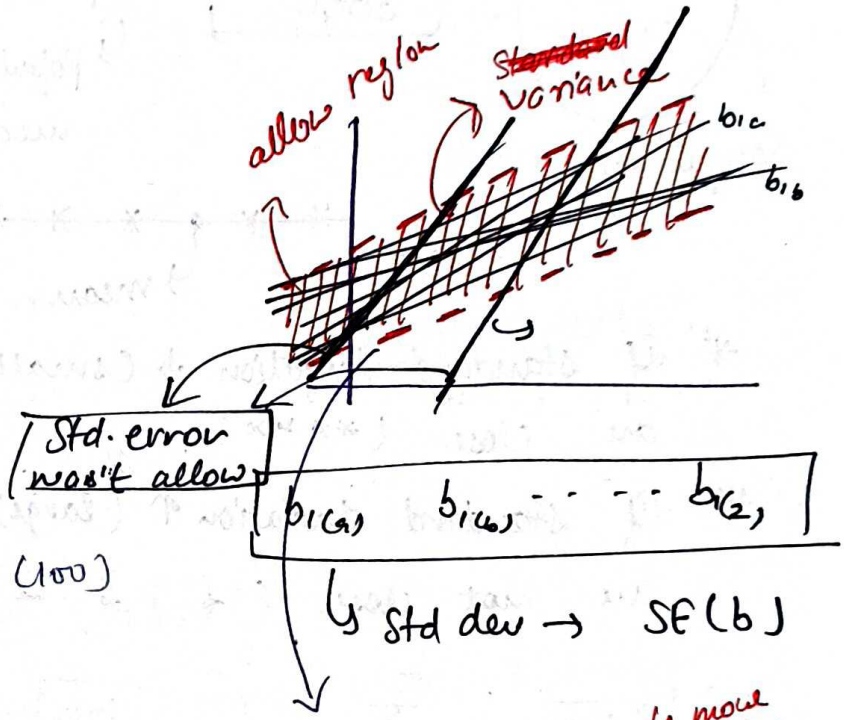
↳ pop mean ↳ stderror

Example

* Sample dataset (100)

exp	salary
–	–
–	–
–	–
–	–

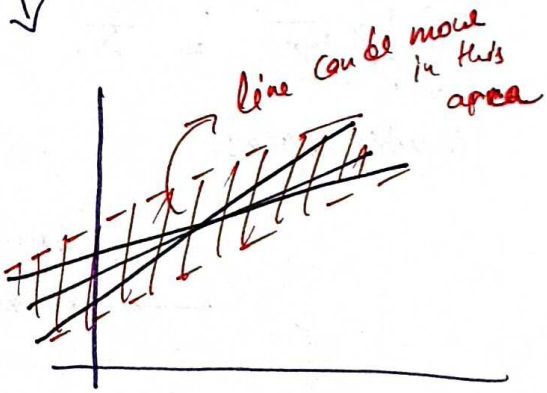
SE(b_2)



* Another sample dataset (100)

exp	salary
–	–
–	–
–	–

SE(b_1)



$$Var(\beta) = \sigma^2 \underbrace{(X^T X)^{-1}}_{3\times 3}$$

$$\sigma^2 \begin{bmatrix} - & - & - \\ - & - & - \\ - & - & - \end{bmatrix} \xrightarrow{\text{Another}} \begin{bmatrix} a & - & - \\ - & b & - \\ - & - & c \end{bmatrix}$$

$$SE = \sqrt{a \to Var(\beta_0)}$$

$$SE = \sqrt{b \to Var(\beta_1)}$$

$$SE = \sqrt{c \to Var(\beta_2)}$$

$$SE(\beta) = \sqrt{diagonal\left(\sigma^2 (X^T X)^{-1}\right)}$$