Regression Analysis

Regression analysis is a stastistical method used to estimate and analyze the relationships between a dependent variable (called as outcomes) and one or more independent variables (also called predictors)

Inference Vs Prediction [Why regression analysis is required?]

Inference: Inference focuses on understanding the underlying relationships and structure within the data.
The goal is to uncover patterns, dependencies, and casual mechanisms that explain why outcomes occur.
Ex. Linear Regression for estimating the effect of spend on Sales, logistic regression for understanding risk factors on disease

Prediction: Prediction aims to forecast futrue outcomes or estimate unknown values for new, unseen data points.
The focus is on accuracy and reliability of these forecasts, not necessarily on understanding why prediction works.
Example: Using a ML model to predict next month's sales.

Details about the data

Summary of relationships.

Relationships of each feature on individual level

Testing the assumptions of Linear regression.

OLS Regression Results

Dep. Variable:	Sales	R-squared:	0.897
Model:	OLS	Adj. R-squared:	0.896
Method:	Least Squares	F-statistic:	570.3
Date:	Tue, 20 May 2025	Prob (F-statistic):	1.58e-96
Time:	22:42:38	Log-Likelihood:	-386.18
No. Observations:	200	AIC:	780.4
Df Residuals:	196	BIC:	793.6
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	2.9389	0.312	9.422	0.000	2.324	3.554
TV	0.0458	0.001	32.809	0.000	0.043	0.049
Radio	0.1885	0.009	21.893	0.000	0.172	0.206
Newspaper	-0.0010	0.006	-0.177	0.860	-0.013	0.011

Omnibus:	60.414	Durbin-Watson:	2.084
Prob(Omnibus):	0.000	Jarque-Bera (JB):	151.241
Skew:	-1.327	Prob(JB):	1.44e-33
Kurtosis:	6.332	Cond. No.	454.

R-squared:	0.897
Adj. R-squared:	0.896
F-statistic:	570.3
Prob (F-statistic):	1.58e-96
Log-Likelihood:	-386.18
AIC:	780.4
BIC:	793.6

→ To prove by doing hypothesis Testing

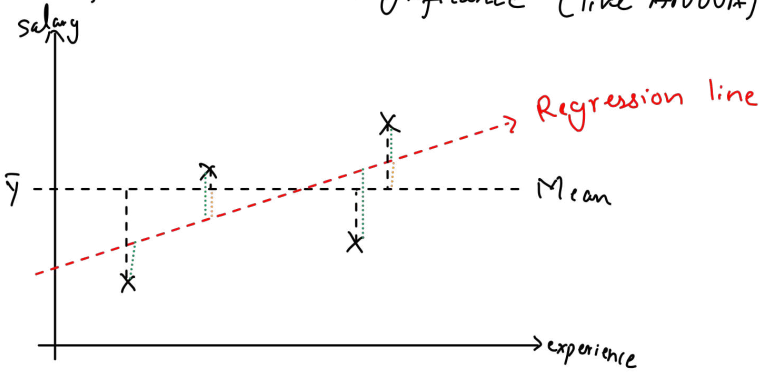
F-Test for overall significance (like ANOVA)

few terms to know before F-Test

TSS : Total Sum of squares

RSS : Residual sum of squares

ESS : Explained sum of squares.



(color based on graph dashed lines)

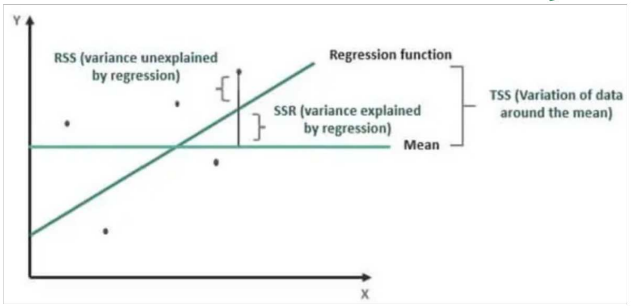
TSS : $\sum (y_i - \bar{y})^2$ (Overall variance in data)

RSS : $\sum (y_i - \hat{y}_i)^2$ (Variance with respect to regression line)

TSS → ESS + RSS → ε

ESS → explained

RSS → unexplained → Reducible, Irreducible



Degree of Freedom

In linear regression, **the total degrees of freedom** (df_total) represent the total number of data points minus 1. It represents the overall variability in the dataset that can be attributed to both the model and the residuals.
For a linear regression with n data points (observations), the total degrees of freedom can be calculated as:
df_total = n - 1
where: n is the number of data points (observations) in the dataset

The total degrees of freedom in linear regression is divided into two components:

1. **Degrees of freedom for the model** (df_model): This is equal to the number of independent variables in the model (k).
2. **Degrees of freedom for the residuals** (df_residuals): The degrees of freedom for the residuals indicate the number of independent pieces of information that are available for estimating the variability in the residuals (errors) after fitting the regression model.
This is equal to the number of data points (n) minus the number of estimated parameters, including the intercept (k+1).

The sum of the degrees of freedom for the model and the degrees of freedom for the residuals is equal to the total degrees of freedom:
df_total = df_model + df_residuals

k (# of input columns) , n-k-1 (n: # of rows)

$df\_total = k + n - k - 1 = n - 1$

$\therefore \boxed{df\_total = n - 1}$

F Statistic and Prob (F-statistic)

The **F-test for overall significance** is a statistical test used to determine whether a linear regression model is statistically significant, meaning it provides a better fit to the data than just using the mean of the dependent variable.

Here are the steps involved in conducting an F-test for overall significance:

1. State the null and alternative hypotheses:
   - Null hypothesis (H0): All regression coefficients (except the intercept) are equal to zero ($\beta 1 = \beta 2 = \ldots = \beta k = 0$), meaning that none of the independent variables contribute significantly to the explanation of the dependent variable's variation.
   - Alternative hypothesis (H1): At least one regression coefficient is not equal to zero, indicating that at least one independent variable contributes significantly to the explanation of the dependent variable's variation.

2. Fit the linear regression model to the data, estimating the regression coefficients (intercept and slopes).

3. Calculate the Sum of Squares (SS) values:
   - Total Sum of Squares (TSS): The sum of squared differences between each observed value of the dependent variable and its mean.
   - Regression Sum of Squares (ESS): The sum of squared differences between the predicted values of the dependent variable and its mean.
   - Residual Sum of Squares (RSS): The sum of squared differences between the observed values and the predicted values of the dependent variable.

4. Compute the Mean Squares (MS) values:
   - Mean Square Regression (MSR): ESS divided by the degrees of freedom for the model (df_model), which is the number of independent variables (k). This could also be called as Average Explained Variance per independent feature.
   - Mean Square Error (MSE): RSS divided by the degrees of freedom for the residuals (df_residuals), which is the number of data points (n) minus the number of estimated parameters, including the intercept (k+1). This could also be called as average unexplained variance per degree of freedom.

5. Calculate the F-statistic: F-statistic = MSR / MSE

6. Determine the p-value:
   - Compute the p-value associated with the calculated F-statistic using the F-distribution or a statistical software package.

7. Compare the calculated F-statistic to the p-value to the chosen significance level (α):
   - If the p-value $< \alpha$, reject the null hypothesis. This indicates that at least one independent variable contributes significantly to the prediction of the dependent variable, and the overall regression model is statistically significant.
   - If the p-value $\geq \alpha$, fail to reject the null hypothesis. This suggests that none of the independent variables in the model contribute significantly to the prediction of the dependent variable, and the overall regression model is not statistically significant.

Following these steps, you can perform an F-test for overall significance in a linear regression analysis and determine whether the regression model is statistically significant.

Intuition:

The f-statistic test reveals the relationship between dependent and independent columns.

$$\text{F-stat} : \frac{\frac{ESS}{K}}{\frac{RSS}{n-k-1}}$$

$\frac{ESS}{K}$ → Avg explained variance

$\frac{RSS}{n-k-1}$ → Avg Unexplained variance

On dividing average ESS with average RSS (unexplained), we will get the ratio between both, which will help us to determine the hypotheses result.

If the ratio is very large, then when we find the p-value for that ratio, it will be smaller than α (0.05). Therefore we can reject the null hypothesis. And vice versa for small ratio.

Strength of Relationship (R^2)

R-squared (R^2), also known as the coefficient of determination, is a measure used in regression analysis to assess the goodness-of-fit of a model. It quantifies the proportion of the variance in the dependent variable (response variable) that can be explained by the independent variables (predictor variables) in the regression model. R-squared is a value between 0 and 1, with higher values indicating a better fit of the model to the observed data.

In the context of a simple linear regression, R^2 is calculated as the square of the correlation coefficient (r) between the observed and predicted values. In multiple regression, R^2 is obtained from the ratio of the explained sum of squares (ESS) to the total sum of squares (TSS):

$R^2 = ESS / TSS$

where:

- ESS (Explained Sum of Squares) is the sum of squared differences between the predicted values and the mean of the observed values. It represents the variation in the response variable that can be explained by the predictor variables in the model.
- TSS (Total Sum of Squares) is the sum of squared differences between the observed values and the mean of the observed values. It represents the total variation in the response variable.

An R-squared value of 0 indicates that the model does not explain any of the variance in the response variable, while an R-squared value of 1 indicates that the model explains all of the variance. However, R-squared can be misleading in some cases, especially when the number of predictor variables is large or when the predictor variables are not relevant to the response variable.

$$R^2 = \frac{ESS}{TSS}$$

$$= \frac{TSS - RSS}{TSS}$$

$$R^2 = 1 - \frac{RSS}{TSS}$$

This simply gives you a numerical value that you can use to compare with other models. If R^2 is 0.81 → it means the model is able to explain 81% of it's part.

Disadvantage

This says that, when you add more input cols into a model, if the input col helps the model, then R^2 score increases, but if a input col does not help then the R^2 score does not decrease. It remains the same. This misleads the model.

To solve this issue, we use Adjusted R^2

Simply, if a irrelevant (misleading) column is added, adjusted R^2 penalises the score and decreases it.

Adjusted R-squared is a modified version of R-squared (R^2) that adjusts for the number of predictor variables in a multiple regression model. It provides a more accurate measure of the goodness-of-fit of a model by considering the model's complexity.

In a multiple regression model, R-squared (R^2) measures the proportion of variance in the response variable that is explained by the predictor variables. However, R-squared always increases or stays the same with the addition of new predictor variables, regardless of whether those variables contribute valuable information to the model. This can lead to overfitting, where a model becomes too complex and starts capturing noise in the data instead of the underlying relationships.

Adjusted R-squared accounts for the number of predictor variables in the model and the sample size, penalizing the model for adding unnecessary complexity. Adjusted R-squared can decrease when an irrelevant predictor variable is added to the model, making it a better metric for comparing models with different numbers of predictor variables.

The formula for adjusted R-squared is:

$$\text{Adjusted } R^2 = 1 - \left[\frac{(1 - R^2) * (n-1)}{(n-K-1)}\right]$$

n → reduce

k → # input col

where:

- R^2 is the R-squared of the model
- n is the number of observations in the dataset
- k is the number of predictor variables in the model

By using adjusted R-squared, you can more accurately assess the goodness-of-fit of a model and choose the optimal set of predictor variables for your analysis.

The choice between using R-squared and adjusted R-squared depends on the context and the goals of your analysis. Here are some guidelines to help you decide which one to use:

1. **Model comparison:** If you're comparing models with different numbers of predictor variables, it's better to use adjusted R-squared. This is because adjusted R-squared takes into account the complexity of the model, penalizing models that include irrelevant predictor variables. R-squared, on the other hand, can be misleading in this context, as it tends to increase with the addition of more predictor variables, even if they don't contribute valuable information to the model.

2. **Model interpretation:** If you're interested in understanding the proportion of variance in the response variable that can be explained by the predictor variables in the model, R-squared can be a useful metric. However, keep in mind that R-squared does not provide information about the significance or relevance of individual predictor variables. It's also important to remember that a high R-squared value does not necessarily imply causation or a good predictive model.

3. **Model selection and overfitting:** When building a model and selecting predictor variables, it's important to guard against overfitting. In this context, adjusted R-squared can be a helpful metric, as it accounts for the number of predictor variables and penalizes the model for unnecessary complexity. By using adjusted R-squared, you can avoid including irrelevant predictor variables that might lead to overfitting.

In summary, adjusted R-squared is generally more suitable when comparing models with different numbers of predictor variables or when you're concerned about overfitting. R-squared can be useful for understanding the overall explanatory power of the model, but it should be interpreted with caution, especially in cases with many predictor variables or potential multicollinearity.

T - Statistic (To study relationship on individual level)

Performing a t-test for a simple linear regression, including the intercept term and using the p-value approach, involves the following steps:

1. State the null and alternative hypotheses for the slope and intercept coefficients:

   For the slope coefficient (β1):

   - Null hypothesis (H0): β1 = 0 (no relationship between the predictor variable (X) and the response variable (y))
   - Alternative hypothesis (H1): β1 ≠ 0 (a relationship exists between the predictor variable and the response variable)

   For the intercept coefficient (β0):

   - Null hypothesis (H0): β0 = 0 (the regression line passes through the origin)
   - Alternative hypothesis (H1): β0 ≠ 0 (the regression line does not pass through the origin)

2. Estimate the slope and intercept coefficients (b0 and b1): Using the sample data, calculate the slope (b1) and intercept (b0) coefficients for the regression model.

3. Calculate the standard errors for the slope and intercept coefficients (SE(b0) and SE(b1)): Compute the standard errors of the slope and intercept coefficients using the following formulas:

$$t\text{-statistic} = \frac{b_1 - 0}{SE(b_1)} \quad (\beta_0)$$

$$SE(b_1) = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{(n-2)\sum (x_i - \bar{x}_i)^2}} \qquad SE(b_0) = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n-2}\left[\frac{1}{n} + \frac{\bar{x}^2}{\sum (x_i - \bar{x})^2}\right]}$$

4. Compute the t-statistics for the slope and intercept coefficients: Calculate the t-statistics for the slope and intercept coefficients using the following formulas:

$$t\text{-value}_{b_0} = \frac{b_0 - 0}{SE(b_0)}$$

$$t\text{-value}_{b_1} = \frac{b_1 - 0}{SE(b_1)}$$

5. Calculate the p-values for the slope and intercept coefficients: Using the t-statistics and the degrees of freedom, look up the corresponding p-values from the t-distribution table or use a statistical calculator.

6. Compare the p-values to the chosen significance level (α): A common choice for α is 0.05, which corresponds to a 95% confidence level. Compare the calculated p-values to α:

   - If the p-value is less than or equal to α, reject the null hypothesis.
   - If the p-value is greater than α, fail to reject the null hypothesis.

Intuition in simple term

1. First, we consider the null and alternative hypotheses. The Null Hypothesis (H0) says that the value of β is 0 (i.e., there's no effect). The Alternative Hypothesis (H1) says that the value is not 0 (i.e., there is some effect).

2. Then we find the coefficients of all the columns based on the linear regression model.

3. Now, we find the standard error (std err) by using the given formula for standard error.

$$SE(b_1) = \sqrt{\frac{\Sigma(y_i - \hat{y}_i)^2}{(n-2)\Sigma(x_i - \bar{x}_i)^2}} \qquad SE(b_0) = \sqrt{\frac{\Sigma(y_i - \hat{y}_i)^2}{n-2}\left[\frac{1}{n} + \frac{\bar{x}^2}{\Sigma(x_i - \bar{x})^2}\right]}$$

4. Finally, we calculate the t-statistic—we compute the t_value for each input column by using the formula for t-statistic.

$$t\_value\ b_0 = \frac{b_0 - 0}{SE(b_0)} \qquad t\_value\ b_1 = \frac{b_1 - 0}{SE(b_1)}$$

5. Once we have the t_value (for each individual column), we calculate the p_value from that particular t_value and the degree of freedom. We then compare the p_value with alpha, which is generally taken as **0.05**.

6. If the **p_value is greater than alpha**, this means that we **cannot reject** the Null Hypothesis. This suggests that the null hypothesis (H0) is likely correct, meaning the value of β is 0. Therefore, we say there is **no statistically significant relationship** between the X (input column) and Y (output column).

7. If the **p_value is smaller than alpha (0.05)**, then we **reject the Null Hypothesis**, and we can say that there **is a statistically significant relationship** between the X and Y columns.

follow the flow below

	coef	std err	t	P>|t|	[0.025	0.975]
const	2.9389	0.312	9.422	0.000	2.324	3.554
TV	0.0458	0.001	32.809	0.000	0.043	0.049
Radio	0.1885	0.009	21.893	0.000	0.172	0.206
Newspaper	-0.0010	0.006	-0.177	0.860	-0.013	0.011

→ Confidence interval.

→ Next page

Confidence Intervals

1. **Estimate the slope and intercept coefficients (b0 and b1)**: Using the sample data, calculate the slope (b1) and intercept (b0) coefficients for the regression model.

2. **Calculate the standard errors for the slope and intercept coefficients (SE(b0) and SE(b1))**:

$$SE(b_1) = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{(n-2)\sum (x_i - \bar{x})^2}} \leftarrow$$

$$SE(b_0) = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{(n-2)} \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum (x_i - \bar{x})^2} \right]}$$

3. **Determine the degrees of freedom**: In a simple linear regression, the degrees of freedom (df) is equal to the number of observations (n) minus the number of estimated parameters (2: the intercept and the slope coefficient).
   df = n - 2

4. **Find the critical t-value**: Look up the critical t-value from the t-distribution table or use a statistical calculator based on the chosen confidence level (e.g., 95%) and the degrees of freedom calculated in step 3.

5. **Calculate the confidence intervals for the slope and intercept coefficients**: Compute the confidence intervals for the slope (b1) and intercept (b0) coefficients using the following formulas:

$$CI_{b_0} = b_0 \pm t\_value * SE(b_0)$$

$$CI_{b_1} = b_1 \pm t\_value * SE(b_1)$$

These confidence intervals represent the range within which the true population regression coefficients are likely to fall with a specified level of confidence (e.g., 95%)

Assumptions of Linear Regression

1. linearity
2. Normality of Residuals
3. Homoscedasticity
4. No Autocorrelation
5. No or little multicollinearity

Linearity

The Assumption

There is a linear relationship between the independent variables and the dependent variable. The model assumes that changes in the independent variables lead to proportional changes in the dependent variable.

What happens when this assumption is violated?

1. Bias in parameter estimates: When the true relationship is not linear, the estimated regression coefficients can be biased, leading to incorrect inferences about the relationship between the independent and dependent variables.
2. Reduced predictive accuracy: A mis specified linear model may not accurately capture the underlying relationship, which can result in poor predictive performance. The model might underfit the data, missing important patterns and trends.
3. Invalid hypothesis tests and confidence intervals: The violation of the linearity assumption can affect the validity of hypothesis tests and confidence intervals, leading to incorrect inferences about the significance of the independent variables and the effect sizes.

How to check this assumption

1. Scatter plots: Create scatter plots of the dependent variable against each independent variable. If the relationship appears to be linear, the linearity assumption is likely satisfied. Nonlinear patterns or other trends may indicate that the assumption is violated.
2. Residual plots: Plot the residuals (the differences between the observed and predicted values) against the predicted values or against each independent variable. If the linearity assumption holds, the residuals should be randomly scattered around zero, with no discernible pattern. Any trends, curvature, or heteroscedasticity in the residual plots suggest that the linearity assumption may be violated.
3. Polynomial terms: Add polynomial terms to your model and compare the model fit with the original linear model. If the new model with additional terms significantly improves the fit, it may suggest that the linearity assumption is violated.

What to do when the assumption fails?

1. Transformations: Apply transformations to the dependent and/or independent variables to make their relationship more linear. Common transformations include logarithmic, square root, and inverse transformations.
2. Polynomial regression: Add polynomial terms of the independent variables to the model to capture non-linear relationships.
3. Piecewise regression: Divide the range of the independent variable into segments and fit separate linear models to each segment.
4. Non-parametric or semi-parametric methods: Consider using non-parametric or semi-parametric methods that do not rely on the linearity assumption, such as generalized additive models (GAMs), splines, or kernel regression.

Normality of Residual

The Assumption

The error terms (residuals) 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 independent variables.
2. Invalid confidence intervals: The confidence intervals for the regression coefficients 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 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 the residuals to visually assess their distribution. If the histogram resembles a bell-shaped curve, it suggests that the residuals are normally distributed.
2. Q-Q plot: A Q-Q (quantile-quantile) plot compares the quantiles of the residuals to the quantiles of a standard normal distribution. If the points in the Q-Q plot fall approximately along a straight line, it indicates that the residuals are normally distributed. Deviations from the straight line suggest deviations from normality.
3. Statistical tests: Statistical tests like Omnibus test, Jarque-Bera test or even Shapiro wilk test can test this assumption.

Next page

What to do when the assumption fails?

1. Model selection techniques: 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 Squares (LMS), or Least Trimmed Squares (LTS).(Transformation may also help)
3. Non-parametric or semi-parametric methods: Consider using non-parametric or semi-parametric methods that do not rely on the normality assumption, such as generalized additive models (GAMs), splines, or kernel regression.
4. Use bootstrapping: Bootstrap-based inference methods do not rely on the normality of residuals and can provide more accurate confidence intervals and hypothesis tests.

Remember that the normality of residuals assumption is not always critical for linear regression, especially when the sample size is large, due to the Central Limit Theorem.

Omnibus test

The Omnibus test is a statistical test used to check if the residuals from a linear regression model follow a normal distribution. The test is based on the skewness and kurtosis of the residuals. Here's a 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 residuals 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 subtracting 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 be close to zero.
5. Calculate the kurtosis: Calculate the kurtosis of the residuals. Kurtosis measures the "tailedness" of the distribution. For a normal distribution, kurtosis is expected to be close to zero (in excess kurtosis terms).
6. Calculate the Omnibus test statistic: Compute the Omnibus test statistic (K^2) using the skewness and kurtosis values. The formula for the Omnibus test statistic is:

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

$n \rightarrow$ number of observations

6. Determine the p-value: The Omnibus test statistic follows a chi-square distribution with 2 degrees of freedom. Use this distribution to calculate the p-value corresponding to the test statistic.
7. 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 smaller than the significance level, you reject the null hypothesis, suggesting that the residuals may not follow a normal distribution.

Homoscedasticity

The Assumption

The spread of the error terms (residuals) should be constant across all levels of the independent variables. If the spread of the residuals changes systematically, it leads to heteroscedasticity, which can affect the efficiency of the estimates.

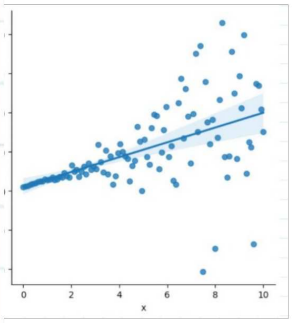
What happens when this assumption is violated?

1. **Inefficient estimates**: While the parameter estimates (coefficients) are still unbiased, they are no longer the best linear unbiased estimators (BLUE) under heteroscedasticity. The inefficiency of the estimates implies that the standard errors are larger than they should be, which may reduce the statistical power of hypothesis tests.
2. **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 assumption of homoscedasticity. If the residuals exhibit heteroscedasticity, these 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 sizes and the precision of the estimates.

t-stats
Confia

How to check this assumption

1. **Residual plot**: Create a scatter plot of the residuals (the differences between the observed and predicted values) against the predicted values or against each independent variable. If the plot shows a random scattering of points around zero with no discernible pattern, it suggests homoscedasticity. If there is a systematic pattern, such as a funnel shape or a curve, it indicates heteroscedasticity.
2. **Breusch-Pagan test**: This is a formal statistical test for heteroscedasticity. The null hypothesis is that the error variances are constant (homoscedastic). If the resulting p-value is less than a chosen significance level (e.g., 0.05), the null hypothesis is rejected, indicating heteroscedasticity.



What to do when the assumption fails?

1. **Transformations**: Apply transformations to the dependent and/or independent variables to stabilize the variance of the residuals. Common transformations include logarithmic, square root, and inverse transformations.
2. **Weighted Least Squares (WLS)**: Use a weighted least squares approach, which assigns different weights to the observations based on the magnitude of their residuals. This method can help account for heteroscedasticity by giving more importance to observations with smaller residuals and less importance to those with larger residuals.
3. **Robust standard errors**: Calculate robust (or 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.