Linear Regression Assumptions

- Linearity
- Constant Variance
- Independence
- Normality

How to check?

Using Residual Analysis.

$$Residual = Observed - Predicted$$

$$\hat{e} = y - \hat{y}$$

We need to use the standardized residuals r_i for assessing the model assumptions.

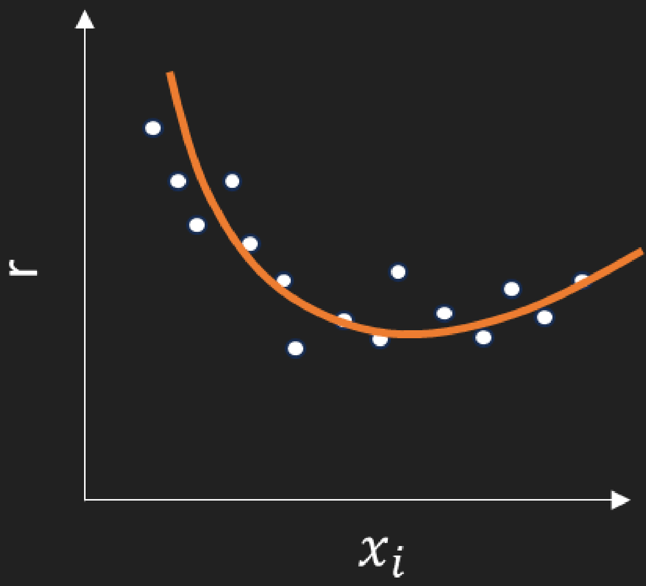
$$r_i = \frac{\hat{e}}{\hat{\sigma}\sqrt{1 - h_{i,i}}}$$

Where:

- $\hat{\sigma}$: standard deviation of the residuals.
- $h_{i,i}$: leverage value for observation.

Linearity / Mean zero Assumption

The relationship between the response and each predicting variable is linear.

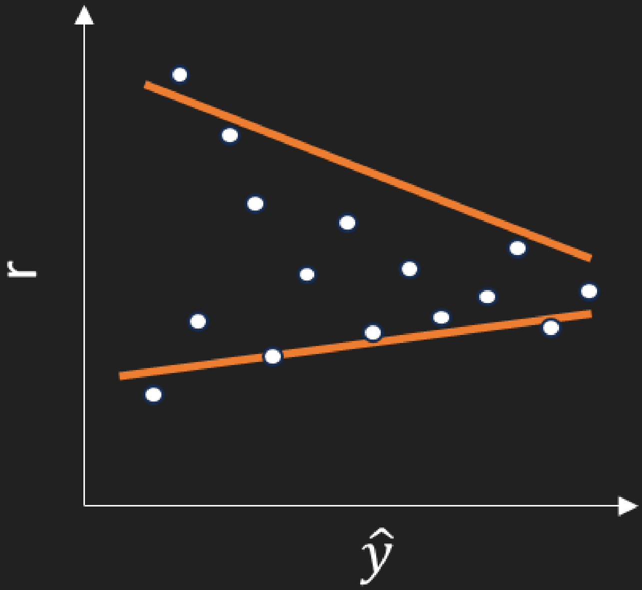


How to check: Plot residuals against each predicting variable.

The plot shows there might be a non-linear relationship between "y" and "x_1"

Constant Variance Assumption

Also called a homoscedasticity check. Linear regression assumes the variance of the residuals is constant.

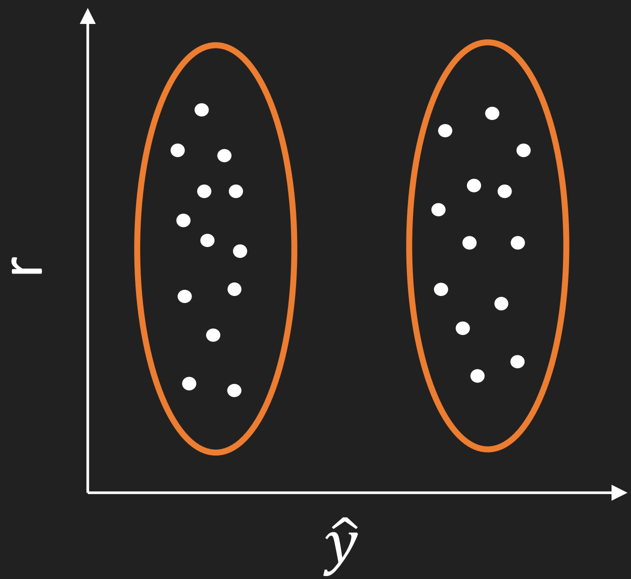


How to check: Plot residuals against fitted values.

The above plot is an example of heteroscedasticity.

Independence Assumption (1/2)

Complicated to check. When using residual analysis we are checking for uncorrelated errors, not independence.



How to check: Plot residuals against fitted values.

The above plot shows clusters of residuals which can be interpreted as correlated.

Independence Assumption (2/2)

We can also use the Durbin-Watson test to check for autocorrelation at lag 1 of the residuals.

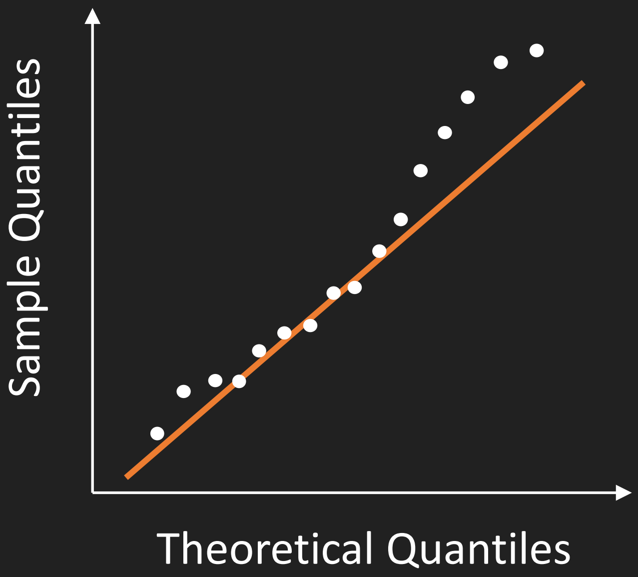
How to check: Calculate the Durbin-Watson statistic "d". If $1.5 < d < 2.5$, we can conclude there is NO first-order correlation between the residuals.

If $d < 1.5$, there is presence of positive autocorrelation.

If $d > 2.5$, there is presence of negative autocorrelation.

Normality Assumption

Important especially when it comes to t-tests / F-test (hypothesis testing) and confidence intervals.



How to check:
Create a **normal Q-Q plot** of the residuals.

The residuals should follow the straight line if they are normally distributed.

What's next??

In the next post, we'll see how to check these assumptions using Python!

```
#%% 1. Linearity Assumption
fig, axes = plt.subplots(nrows=3, ncols=4, figsize=(15, 10))
for i, ax in enumerate(axes.flatten()):
    if i < len(X.columns) - 1:  # Exclude the constant term
        sns.scatterplot(x=X.iloc[:, i + 1], y=standardized_residuals, ax=ax)
        ax.set_xlabel(X.columns[i + 1])  # Skip constant column
        ax.set_ylabel('Standardized Residuals')
        ax.axhline(0, color='r', linestyle='--')
plt.tight_layout()
plt.show()
```