

Assumptions of linear regression

Video → <https://youtu.be/-qXMA7mOecg>

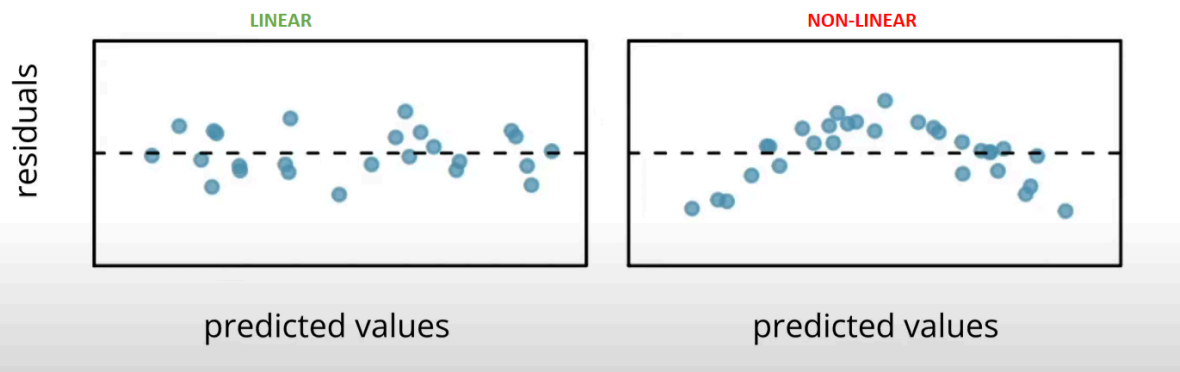
<https://www.youtube.com/watch?v=EmSNAtcHLm8> (CampuX)



1. **L**: Linear
2. **I**: Independent
3. **N**: Normal
4. **E**: Equal variance (homoscedasticity)

When you use linear regression, you assume that:

Residuals/Errors:



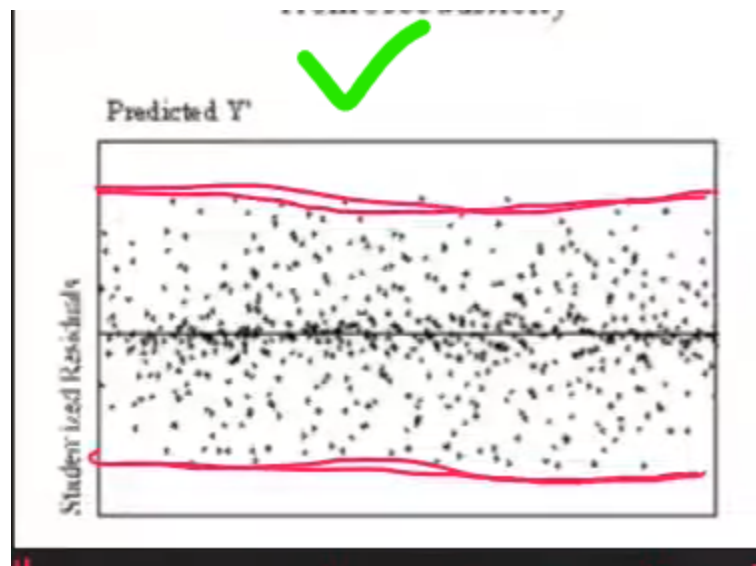
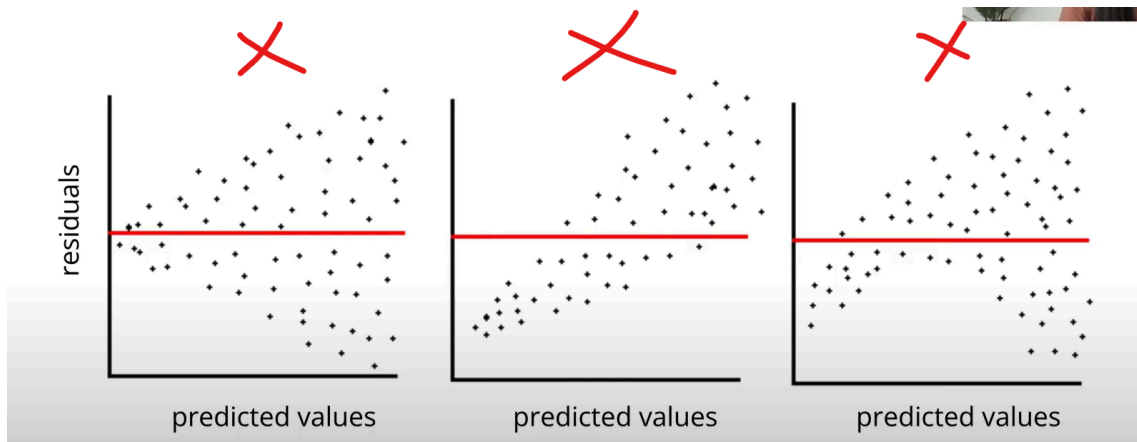
1. **Linearity:**

- **Assumption:** The relationship between the independent variables (predictors) and the dependent variable (target) is linear.
- **Why It Matters:**
 - If the true relationship is non-linear, the model's predictions will be biased.
- **How to Check:**
 - Scatter plots of each predictor vs. the response.
 - Residual plots (residuals should appear random, not curved).

2. **Independence:**

- **Assumption:** The **residuals (errors) are independent** of each other.
- **Why It Matters:**
 - Violations (like autocorrelation in time-series data) can lead to underestimated standard errors, affecting confidence intervals and hypothesis tests.
- **How to Check:**
 - Durbin-Watson test.
 - Examining residual plots.

3. **Homoscedasticity (Constant Variance):**



- **Assumption:** The residuals have constant variance at all levels of the independent variables.
- **Why It Matters:**
 - If variance changes (heteroscedasticity), the reliability of the coefficient estimates (and p-values) can be compromised.
- **How to Check:**
 - Plot residuals vs. predicted values; they should scatter randomly around zero with a constant spread.
 - Breusch-Pagan test.

4. Normality of Residuals:

- **Assumption:** The residuals are normally distributed.
- **Why It Matters:**
 - Normality is critical for constructing valid confidence intervals and hypothesis tests.
- **How to Check:**
 - Q-Q plot (quantile-quantile plot).
 - Histogram of residuals.
 - Shapiro-Wilk test.

5. No Multicollinearity:

- **Assumption:** The independent variables are not highly correlated with each other.
- **Why It Matters:**
 - High multicollinearity can make it difficult to isolate the effect of each predictor, leading to unstable coefficient estimates.
- **How to Check:**
 - **Variance Inflation Factor (VIF)**; values above 5 (or 10, by some rules) indicate a problem.
 - Correlation matrix among predictors.

```
from statsmodels.stats.outliers_influence import variance_inflation_factor
```

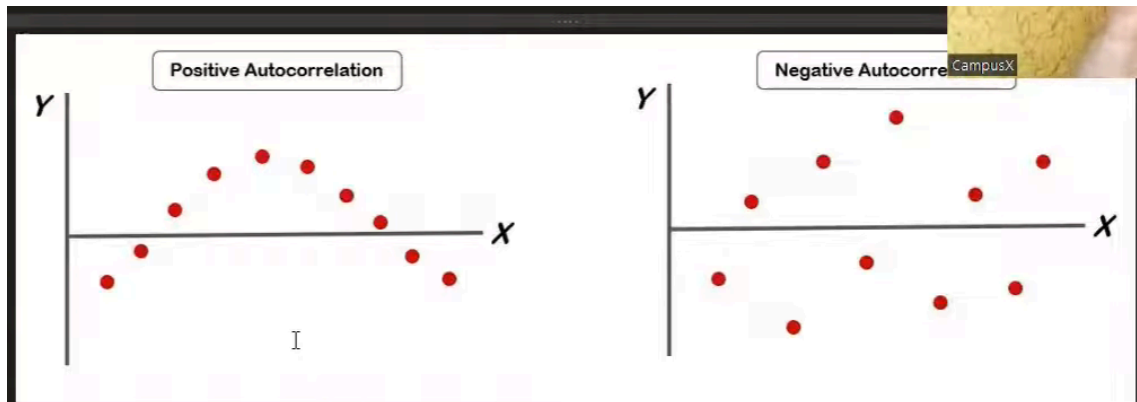
```
vif = []
```

```
for i in range(X_train.shape[1]):
    vif.append(variance_inflation_factor(X_train, i))
```

`variance_inflation_factor(X_train, i)` : This function from `statsmodels` calculates the VIF for the i-th feature in the `X_train` dataset.

6. Exogeneity (No Endogeneity):

- **Assumption:** The independent variables are uncorrelated with the error term.
- **There should not be any pattern when you plot errors.**



- **Why It Matters:**
 - If predictors are correlated with the error term (omitted variable bias, measurement error), the estimates become biased and inconsistent.
- **How to Check:**
 - It often requires domain knowledge or instrumental variable techniques to test.

Additional Interview Insights

- **Checking Assumptions in Practice:**
 - **Residual Plots:** Useful for checking linearity and homoscedasticity.
 - **Q-Q Plots:** Assess the normality of residuals.
 - **Correlation Matrix & VIF:** Identify multicollinearity.
- **Consequences of Violating Assumptions:**
 - **Linearity:** Leads to poor predictions and systematic bias.

- **Independence:** In time-series, autocorrelation can result in misleading statistical inferences.
- **Homoscedasticity:** If violated, standard errors can be incorrect, leading to unreliable hypothesis tests.
- **Normality:** Affects the validity of t-tests and F-tests.
- **Multicollinearity:** Makes it difficult to interpret coefficients; can inflate standard errors.
- **Endogeneity:** Results in biased estimates that cannot be remedied by simply increasing sample size.
- **Interviewer Expectations:**
 - Explain each assumption clearly.
 - Discuss how you would diagnose each assumption.
 - Mention remedies if an assumption is violated (e.g., transforming variables, using robust standard errors, applying ridge or lasso regression to mitigate multicollinearity, or using instrumental variables for endogeneity).

Summary Table

Assumption	What It Means	How to Check	Why It Matters
Linearity	Relationship is linear	Scatter/residual plots	Non-linear relationships cause bias
Independence	Residuals are independent	Durbin-Watson, residual plots	Violations lead to underestimated errors
Homoscedasticity	Constant variance of residuals	Residuals vs. predicted plots, BP test	Heteroscedasticity affects reliability of tests
Normality	Residuals are normally distributed	Q-Q plot, histogram, Shapiro-Wilk	Essential for valid statistical inferences

Assumption	What It Means	How to Check	Why It Matters
No Multicollinearity	Predictors not highly correlated	VIF, correlation matrix	Ensures stable and interpretable coefficients
Exogeneity	Predictors uncorrelated with the error term	Domain knowledge, IV techniques	Biases estimates if violated

DeepSeek Response:



Assumptions of Linear Regression (Interview Perspective)

Linear regression is a foundational technique in statistics and machine learning. To ensure valid and reliable results, the following assumptions must be met. Here's a structured breakdown:

1. Linearity

Assumption: The relationship between the independent variables (IVs) and dependent variable (DV) is linear.

Why It Matters: If the true relationship is nonlinear, the model will underfit.

How to Check:

- Scatter plots of DV vs. each IV.
- Residual plots (residuals vs. predicted values should show no pattern).
- **Solutions:**
 - Transform IVs (e.g., log, square, polynomial terms).
 - Use nonlinear models (e.g., polynomial regression, decision trees).

2. Independence of Errors

Assumption: Residuals (errors) are uncorrelated with each other.

Why It Matters: Correlated errors (e.g., in time-series data) inflate standard errors, leading to unreliable p-values.

How to Check:

- **Durbin-Watson test** (values near 2 indicate no autocorrelation; <1 or >3 signal issues).
 - **Solutions:**
 - Use autoregressive models (e.g., ARIMA) for time-series data.
 - Cluster standard errors for grouped data.
-

3. Homoscedasticity

Assumption: Residuals have constant variance across all levels of the DV.

Why It Matters: Heteroscedasticity (uneven variance) biases standard errors, affecting hypothesis tests.

How to Check:

- Residual vs. fitted value plots (look for "funnel" or "fan" shapes).
 - **Breusch-Pagan test** (statistical test for heteroscedasticity).
 - **Solutions:**
 - Transform the DV (e.g., log, square root).
 - Use robust standard errors (e.g., Huber-White estimator).
 - Apply weighted least squares.
-

4. Normality of Errors

Assumption: Residuals are approximately normally distributed.

Why It Matters: Critical for confidence intervals, hypothesis tests, and prediction intervals.

How to Check:

- **Q-Q plots** (points should align with the diagonal line).
 - Shapiro-Wilk or Kolmogorov-Smirnov tests.**Solutions:**
 - Transform the DV (e.g., Box-Cox transformation).
 - Increase sample size (Central Limit Theorem mitigates non-normality).
-

5. No Multicollinearity

Assumption: Independent variables are not highly correlated with each other.

Why It Matters: High correlation inflates standard errors, making coefficients unstable and hard to interpret.

How to Check:

- **Variance Inflation Factor (VIF):** $VIF > 5-10$ indicates severe multicollinearity.
 - Correlation matrix (look for $|r| > 0.8$).
 - **Solutions:**
 - Remove redundant variables.
 - Combine correlated variables (e.g., PCA).
 - Use regularization (e.g., Ridge Regression).
-

6. No Perfect Multicollinearity

Assumption: No exact linear relationship between IVs (e.g., one IV is a linear combination of others).

Why It Matters: The model cannot estimate unique coefficients (matrix inversion fails).

How to Check:

- Check for singularities in the design matrix (e.g., using rank checks).**Solutions:**
- Drop variables causing perfect collinearity.

- Check for dummy variable traps (e.g., one-hot encoding without dropping a category).
-

7. Exogeneity

Assumption: Independent variables are uncorrelated with the error term.

Why It Matters: Violation (endogeneity) causes omitted variable bias, leading to biased coefficients.

Common Causes:

- Omitted variables (e.g., excluding a key predictor).
 - Measurement error in IVs.
 - Simultaneity (DV and IV influence each other).
- Solutions:**
- Include omitted variables if possible.
 - Use instrumental variables (IV regression).
 - Apply control function approaches.
-

8. Additivity (No Interaction Effects)

Assumption: The effect of IVs on the DV is additive (unless interaction terms are explicitly included).

Why It Matters: If interactions exist but are omitted, the model is misspecified.

How to Check:

- Domain knowledge (e.g., does variable A's effect depend on variable B?).
 - Test interaction terms for significance.
- Solutions:**
- Include interaction terms (e.g., $(X_1 \times X_2)$).
-

9. Sample Size

Guideline: At least 10–20 observations per predictor variable.

Why It Matters: Small samples lead to overfitting and unstable estimates.

Solutions:

- Collect more data.
 - Use regularization (e.g., Lasso) for feature selection.
-

10. Measurement of Variables

Assumption: IVs are measured without error.

Why It Matters: Measurement error attenuate coefficients (biases them toward zero).

Solutions:

- Use measurement error models (e.g., errors-in-variables regression).
 - Improve data collection processes.
-

Interview Tips

1. **Link to Consequences:** Always explain how violating an assumption impacts results (e.g., "Heteroscedasticity biases standard errors").
2. **Practical Checks:** Mention tools like Q-Q plots, VIF, or Durbin-Watson tests.
3. **Real-World Example:**
 - *Multicollinearity:* Predicting house prices using both "square footage" and "number of rooms" (they're correlated).
 - *Endogeneity:* Modeling salary with education but omitting "ability" (unobserved variable).
4. **Mitigation Strategies:** Show you understand fixes (e.g., transformations, regularization).