Understand the Assumptions of OLS Regression

Understand and empirically test OLS assumptions. Dealing with breaches of OLS assmptions.

Tomáš Oleš
Department of Economic Policy
Faculty of Economics and Finance

February 1, 2025

Agenda

- **Understand the Assumptions of OLS Regression**
  - Linearity, independence, homoscedasticity, normality of residuals.
- **Testing Assumptions in Stata**
  - Nonlinearity: `rvfplot`, `lowess`
  - Heteroscedasticity: `hettest`, `imtest`
  - Distribution of errors: `swilk`, `qnorm`
- **Addressing Violations of Assumptions**
  - Transformations (quadratic, log, square root).
  - Robust or clustered standard errors.

Gauss–Markov Assumptions

The Ordinary Least-Squares (OLS) Estimator

- The least-squares method is attributed to Karl Friedrich Gauss (1821).
- The Gauss–Markov theorem states that OLS provides the Best Linear Unbiased Estimator (BLUE) under specific conditions.

Gauss–Markov Assumptions

1. **Zero Conditional Mean of Errors**
    - $E(\varepsilon|X) = 0$ ensures no omitted variable bias.
2. **Homoscedasticity (Constant Variance of Errors)**
    - Variance of errors should not depend on X.
3. **No Autocorrelation**
    - Errors should not be correlated across observations.

Additional Assumptions for OLS

4. **Correct Model Specification**
   - Omitted variables or incorrect functional form can bias results.
5. **Absence of Multicollinearity**
   - Explanatory variables should not be highly correlated.
6. **Normally Distributed Residuals**
   - Important for inference, particularly for small samples.

Next Steps:

- Investigate influential cases and potential outliers.
- Use robust methods if assumptions are violated.

Correct Model

Two Parts:

- **Model Specification:** Ensuring the correct functional form and inclusion of relevant variables.
- **Residuals Assumptions:** Ensuring homoscedasticity, normality, and independence of residuals.

Model Specification

All X-variables are Relevant, and None Irrelevant

You should not include X-variables that you have no theoretical or logical reason to include (Mehmetoglu and Jakobsen, 2022).

Testing Model Improvements in Stata:

- Run the restricted (smaller) model.
- Run the unrestricted (larger) model with additional variables.
- Use an **F-test** to check significance of added variables.

```
*Run the restricted model
reg Y X1 X2 X3
*Run the unrestricted model
reg Y X1 X2 X3 X4 X5
*Conduct an F-test to check if X_4 and X_5 jointly improve the model
test X4 X5
```

Detecting Model Misspecification with `linktest` in Stata

Purpose of `linktest`:

- Checks if the regression model is correctly specified.
- Identifies whether the wrong functional form is used.
- Detects omitted variables.

How It Works:

- Runs a regression including:
  - _hat: the predicted values of Y.
  - _hatsq: the squared predicted values.
- If _hatsq is significant, the model is misspecified.

Interpreting Results:

- If _hatsq is NOT significant: The model is correctly specified.
- If _hatsq is significant: The model has omitted variables or incorrect functional form.

```
*Estimate the factors that are associated with the trust in the legal
    system (0  10  )
use ESSGBdiagnostics.dta, clear
quietly regress trstlgl age woman political_interest religious
linktest
```

Interpreting Output:

- If _hatsq is significant ($p < 0.05$), the model is misspecified.
- Consider adding relevant variables or transforming predictors.

However, passing a diagnostic test like `linktest` (or `ovtest`) not mean that we have specified the best possible model, either statistically or substantively.

Linearity Assumption in Regression

Definition:

- A one-unit increase in X_i results in a constant change in Y, holding other variables constant.
- The effect of X on Y does not depend on the level of X.

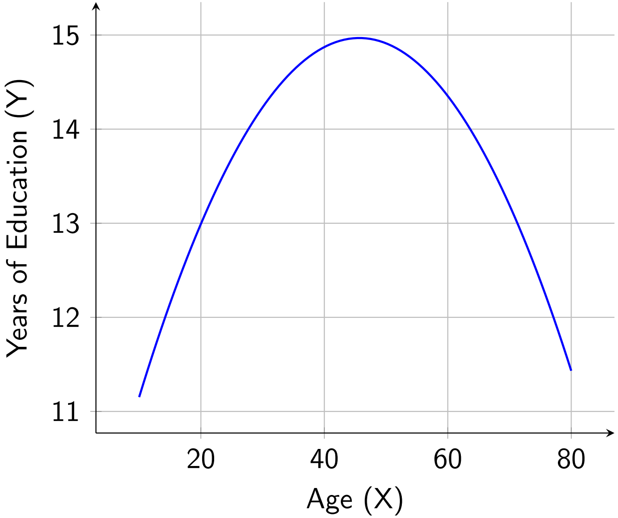
When Linearity is Violated:

- In real-world data, the impact of X on Y often changes at different levels of X.
- Example:
  - Moving from $X = 3$ to $X = 4$ increases Y.
  - Moving from $X = 45$ to $X = 46$ decreases Y.
- This indicates a nonlinear relationship.

Consequences of Misspecification:

- Incorrect slope estimates lead to misleading interpretations.
- Standard errors may be biased, affecting hypothesis testing.
- Model predictions may not accurately reflect reality.

Nonlinear Relationship: Quadratic Curve



Equation: $Y = 8.712 + 0.274X - 0.003X^2$

Equation: $Y = 8.712 + 0.274X - 0.003X^2$
Interpretation:

- Initially, Y (education) increases with X (age).
- A turning point occurs where education peaks.
- After a certain age, the negative X^2 term dominates, causing Y to decline.

Finding the Maximum/Minimum of $Y = 8.712 + 0.274X - 0.003X^2$

Equation:

$$Y = 8.712 + 0.274X - 0.003X^2$$

Step 1: Compute the First Derivative

$$\frac{dY}{dX} = 0.274 - 0.006X$$

To find the critical point, set the derivative equal to zero:

$$0.274 - 0.006X = 0$$

Step 2: Solve for X

$$X = \frac{0.274}{0.006} = 45.67$$

This means the function reaches a turning point at approximately $X = 45.67$ years.

Step 3: Second Derivative Test

$$\frac{d^2Y}{dX^2} = -0.006$$

Since the second derivative is negative, the function has a maximum at $X = 45.67$.

Interpretation of the Maximum Point

Key Findings:

- Education (Y) increases with age (X) until $X = 45.67$.
- After this age, the negative effect of X^2 dominates, causing education levels to decline.
- The maximum value of Y occurs around age 46.

Practical Implication:

- The quadratic term in regression helps capture nonlinearity in relationships.
- Ignoring nonlinear effects could lead to misleading interpretations.

Detecting and Addressing Nonlinearity

How to Detect Nonlinearity:

- Scatterplots: Visualize Y vs. X to check for patterns.
- Residual Plots: Look for systematic patterns in residuals.
- Ramsey's RESET Test (`ovtest` in Stata) detects omitted nonlinear terms.

Solutions to Nonlinearity:

- Polynomial Regression: Include X^2 or X^3 terms.
- Log Transformations: Use log-linear or log-log models - we will talk about them later :).
- Nonparametric Methods: Consider splines or kernel regressions (advanced, not covered in this course).

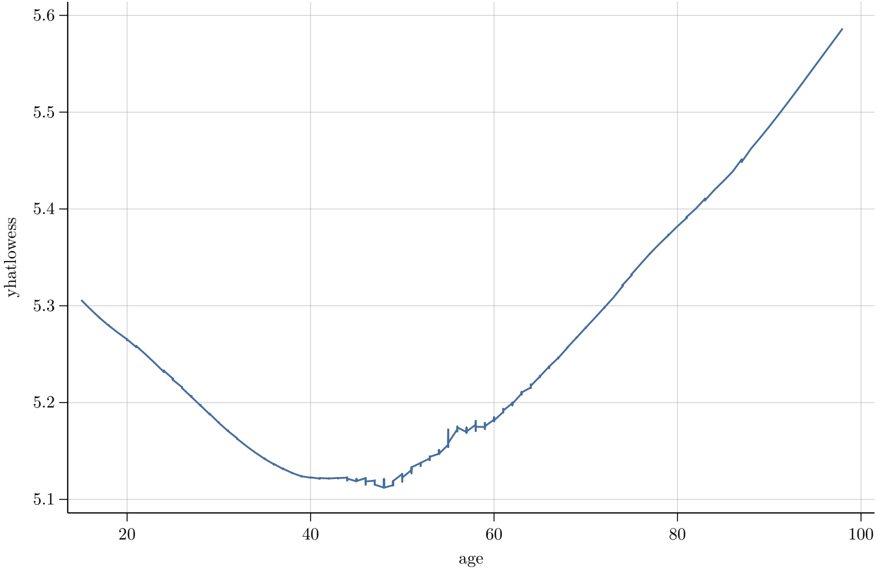
```
*Detect potential nonlinearity
scatter Y X       *Check for nonlinearity
ovtest            *Ramseys RESET test
gen X2 = X^2      *Create squared term
reg Y X X2        *Fit polynomial model
 \end{verbatim}
```

```
. *Modelling curvilinearity when there is an interaction
. use ESS5GBdiagnostics.dta, clear
. regress trstlgl age woman political_interest religious

      Source |       SS           df       MS      Number of obs   =     1,902
-------------+----------------------------------   F(4, 1897)      =     25.18
       Model |  532.142701         4  133.035675   Prob > F        =    0.0000
    Residual |  10023.8426     1,897  5.28404986   R-squared       =    0.0504
-------------+----------------------------------   Adj R-squared   =    0.0484
       Total |  10555.9853     1,901  5.55285917   Root MSE        =    2.2987

------------------------------------------------------------------------------------
           trstlgl |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------------+----------------------------------------------------------------
               age |  -.0045918   .0028796    -1.59   0.111    -.0102393    .0010557
             woman |  -.3239065    .108797    -2.98   0.003    -.5372809   -.1105321
political_interest |   .4534565   .0581223     7.80   0.000     .3394662    .5674468
         religious |   .0892736   .0217054     4.11   0.000     .0467046    .1318426
             _cons |   4.172909   .2254221    18.51   0.000     3.730808    4.615011
------------------------------------------------------------------------------------
```

```
*Curvilinearity
lowess trstlgl age, nograph gen(yhatlowess)  // predicts value of
   regression
line yhatlowess age, sort   // graph bivariate relationship
```


5.6
5.5
5.4
5.3
5.2
5.1
yhatlowess
20
40
60
80
100
age

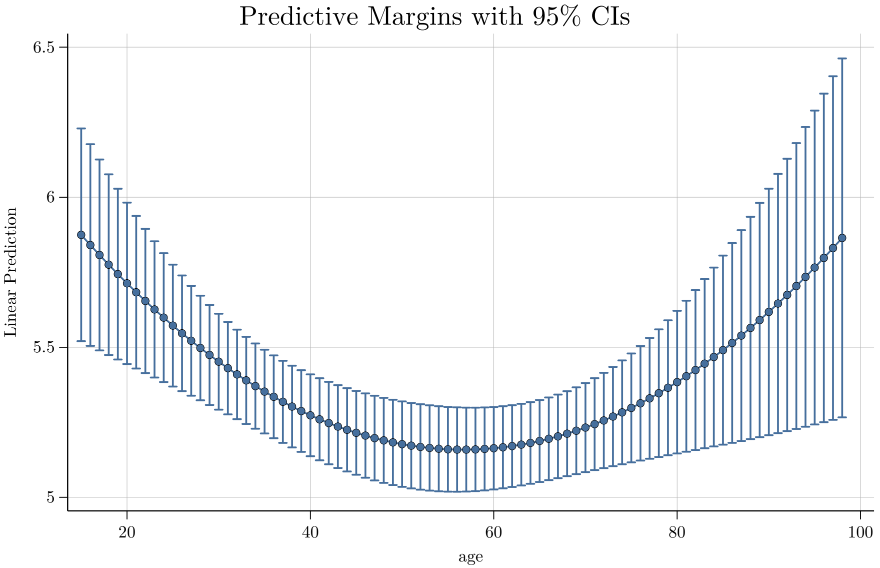
```
*Regression including squared term
regress trstlgl c.age##c.age woman political_interest religious

margins, at(age=(15(1)98))
marginsplot
```

```
. *Regression including squared term
. regress trstlgl c.age##c.age woman political_interest religious

      Source |       SS           df       MS      Number of obs   =     1,902
-------------+----------------------------------   F(5, 1896)      =     21.84
       Model |   574.94511         5  114.989022   Prob > F        =    0.0000
    Residual |  9981.04017     1,896  5.26426169   R-squared       =    0.0545
-------------+----------------------------------   Adj R-squared   =    0.0520
       Total |  10555.9853     1,901  5.55285917   Root MSE        =    2.2944

------------------------------------------------------------------------------------
           trstlgl |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------------+----------------------------------------------------------------
               age |  -.0467456   .0150601    -3.10   0.002    -.0762818   -.0172095
                   |
       c.age#c.age |   .0004126   .0001447     2.85   0.004     .0001288    .0006964
                   |
             woman |  -.3257212    .108595    -3.00   0.003    -.5386995    -.112743
political_interest |   .4723812   .0583918     8.09   0.000     .3578623       .5869
         religious |   .0864412   .0216875     3.99   0.000     .0439074    .1289751
             _cons |   5.074485   .3880667    13.08   0.000     4.313403    5.835568
------------------------------------------------------------------------------------
```


Predictive Margins with 95% CIs
Linear Prediction
6.5
6
5.5
5
20
40
60
80
100
age

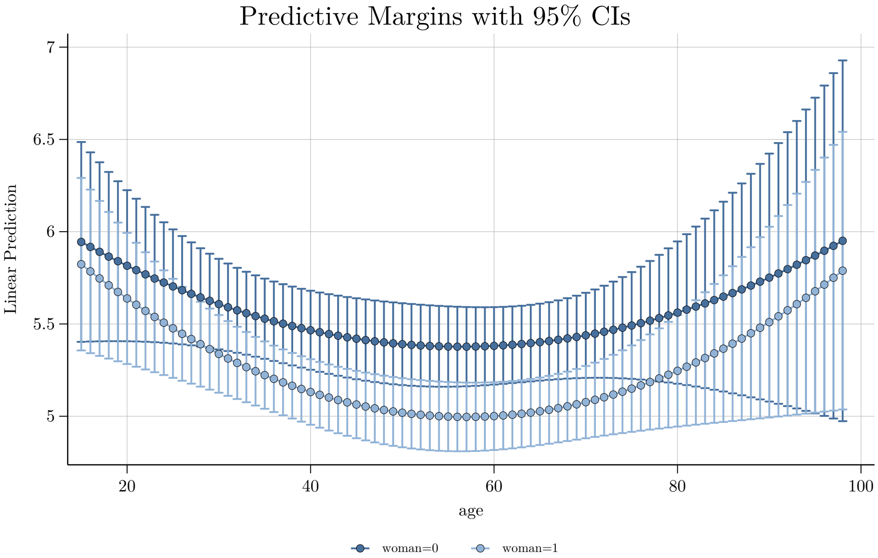
```
*Regression including squared term and interaction effect
regress trstlgl c.age##c.age##i.woman political_interest religious

margins, at (age=(15(1)98) woman=(0 1))
marginsplot
```

```
. *Regression including squared term and interaction effect
. regress trstlgl c.age##c.age##i.woman political_interest religious

      Source |       SS           df       MS      Number of obs   =     1,902
-------------+----------------------------------   F(7, 1894)      =     15.64
       Model |  576.785799         7  82.3979713   Prob > F        =    0.0000
    Residual |  9979.19948     1,894  5.26884872   R-squared       =    0.0546
-------------+----------------------------------   Adj R-squared   =    0.0511
       Total |  10555.9853     1,901  5.55285917   Root MSE        =    2.2954

------------------------------------------------------------------------------------
           trstlgl |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------------+----------------------------------------------------------------
               age |  -.0373657   .0236385    -1.58   0.114    -.0837259    .0089946
                   |
       c.age#c.age |   .0003313   .0002306     1.44   0.151    -.0001209    .0007836
                   |
           1.woman |   .0910325   .7326664     0.12   0.901    -1.345885     1.52795
                   |
       woman#c.age |
                1  |  -.0161969   .0305791    -0.53   0.596    -.0761692    .0437754
                   |
 woman#c.age#c.age |
                1  |   .0001389   .0002951     0.47   0.638    -.0004399    .0007178
                   |
political_interest |   .4723383   .0584197     8.09   0.000     .3577645    .5869122
         religious |   .0870871   .0217246     4.01   0.000     .0444804    .1296938
             _cons |   4.831897    .577209     8.37   0.000     3.699865    5.963929
------------------------------------------------------------------------------------
```


Predictive Margins with 95% CIs
Linear Prediction
7
6.5
6
5.5
5
20
40
60
80
100
age
woman=0
woman=1

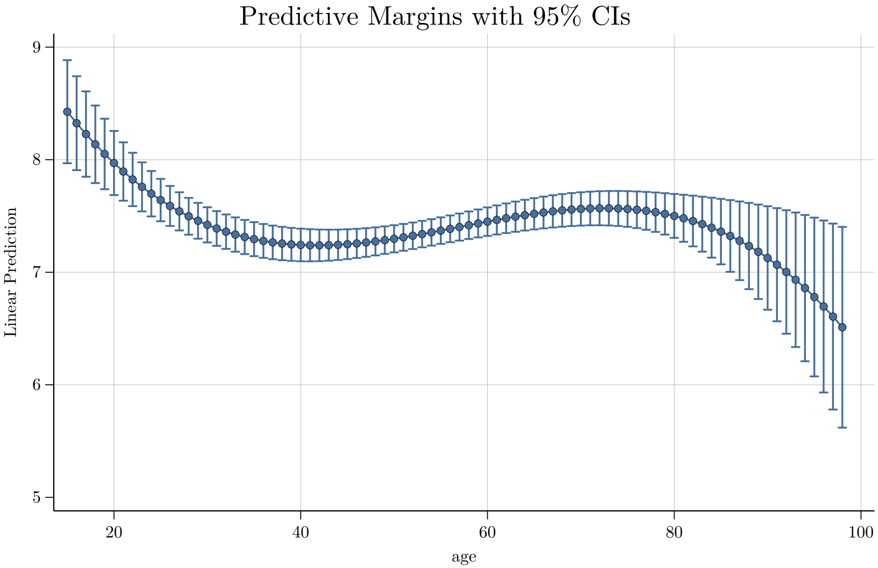
```
*Curvilinear effect with two bends
regress happy c.age##c.age##c.age woman political_interest religious
    leftright

margins, at(age=(15(1)98))
marginsplot, noci
```

```
. *Curvilinear effect with two bends
. regress happy c.age##c.age##c.age woman political_interest religious leftright

      Source |       SS           df       MS      Number of obs   =     1,649
-------------+----------------------------------   F(7, 1641)      =      9.18
       Model |  205.511812         7  29.3588303   Prob > F        =    0.0000
    Residual |   5250.7344     1,641  3.19971627   R-squared       =    0.0377
-------------+----------------------------------   Adj R-squared   =    0.0336
       Total |  5456.24621     1,648  3.31082901   Root MSE        =    1.7888

------------------------------------------------------------------------------------
             happy |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------------+----------------------------------------------------------------
               age |   -.206582   .0491139    -4.21   0.000    -.3029145   -.1102495
                   |
       c.age#c.age |   .0038972   .0009895     3.94   0.000     .0019563    .0058381
                   |
 c.age#c.age#c.age |  -.0000227   6.21e-06    -3.66   0.000    -.0000349   -.0000106
                   |
             woman |  -.0239869   .0906505    -0.26   0.791    -.2017898     .153816
political_interest |   .1368392   .0509382     2.69   0.007     .0369284    .2367499
         religious |   .0942766   .0183395     5.14   0.000     .0583054    .1302479
         leftright |   .0520857   .0257479     2.02   0.043     .0015835    .1025879
             _cons |   9.651854   .7685288    12.56   0.000     8.144453    11.15925
------------------------------------------------------------------------------------
```



Predictive Margins with 95% CIs
Linear Prediction
age
9
8
7
6
5
20
40
60
80
100

Additivity in Regression

Definition:

- The assumption of additivity means that the effect of an independent variable (X_i) on Y is constant, regardless of the values of other independent variables.
- This implies that each X-variable has a separate, independent effect on Y.

When Additivity is Violated:

- If the effect of one X-variable depends on the level of another X-variable, the assumption is breached.
- This situation is known as an interaction effect.

Solution: Introducing Interaction Terms (See Previous Lecture)

- Interaction effects can be modeled by including interaction terms in the regression:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 (X_1 \times X_2) + \varepsilon$$

- The coefficient β_3 captures the interaction effect between X_1 and X_2.

Absence of Multicollinearity

Definition:

- Multicollinearity occurs when two or more X-variables in a regression model are highly correlated.
- Perfect multicollinearity means that one X-variable can be perfectly explained by a linear combination of other X-variables.

Problems Caused by Multicollinearity:

- High correlation (above 0.8) makes it difficult to assess the individual impact of explanatory variables.
- Inflated standard errors, leading to unreliable significance tests.
- Explanatory power is shared, making it difficult to interpret coefficients.

How to Detect Multicollinearity in Stata:

- Correlation Matrix:

```
correlate X1 X2 X3
```

- Variance Inflation Factor (VIF):

```
regress  Y X1 X2 X3
vif
```

- High VIF values (> 10) indicate severe multicollinearity.

Addressing Multicollinearity

How to Fix Multicollinearity:

- **Remove one of the correlated variables** if they measure the same phenomenon.
- **Combine variables into an index or scale** using factor analysis or reliability tests (advanced not covered in this course).
- **Center variables** (for interaction terms) to reduce multicollinearity.

```
*Multicollienarity
quietly regress trstlgl age woman political_interest religious
estat vif
estat vce
```

Residuals Assumptions

Zero Conditional Mean Assumption

$$E[\varepsilon | X_1, X_2, ..., X_N] = 0 \qquad (1)$$

Why This Assumption Holds in Sample:

- Due to the least-squares method, the residuals in an OLS model always balance out in the sample.
- If the errors have a non-zero mean, this will be absorbed by the constant term.

When OLS May Not Be the Best Estimator:

- The estimated coefficients may not represent the true population relationships.
- This can happen when explanatory variables are related to the error term.

Addressing the Problem: ▸ This is subject to the Applied Economics!

Constant Variance Assumption Holds - (aka Homoscedasticity)

Homoskedasticity means that the variance of the error term remains constant across all values of the independent variables.

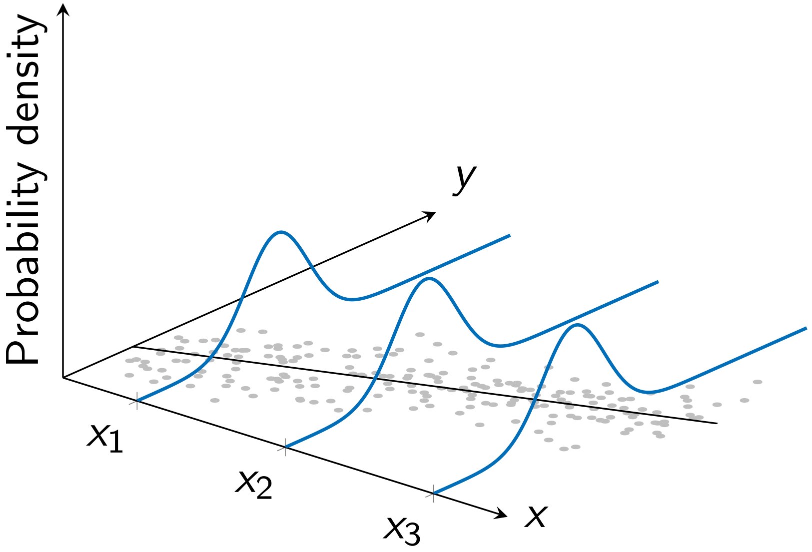
$$\text{Var}(\varepsilon_i | X) = \sigma^2$$

where σ^2 is finite and positive.

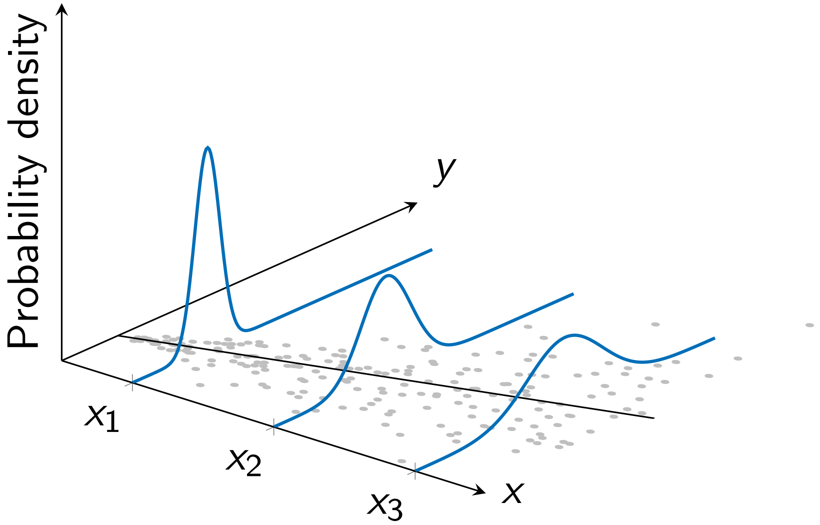
Why This Assumption Matters:

- Ensures valid statistical inference and generalization from sample to population.
- If the variance of residuals changes across levels of X, the model has heteroskedasticity, which leads to:
  - Biased standard errors.
  - Inefficient OLS estimates.
  - Incorrect hypothesis test results.

Constant Variance Assumption Holds - (aka Homoscedasticity)



Constant Variance Assumption Does Not Hold - (aka Heteroscedasticity)



▸ What's that?

Detecting and Testing for Heteroskedasticity in Stata

Step 1: Run the Regression Model

```
quietly regress trstlgl age woman political_interest religious
```

Step 2: Residual vs. Fitted Plot

```
rvfplot
```

- A funnel shape in the plot suggests heteroskedasticity.

Step 3: Breusch-Pagan / Cook-Weisberg Test

```
estat hettest
```

- A significant test result indicates heteroskedasticity.

Solutions for Heteroskedasticity:

- Use robust standard errors:

  ```
  regress trstlgl age woman political_interest religious, vce(robust)
  ```

- Consider log transformation if appropriate.
- Use weighted least squares (WLS) if needed.

Uncorrelated Errors

The errors in a regression model should be uncorrelated across observations.

$$E(\varepsilon_i \varepsilon_j | X_1, \ldots, X_n) = 0, \quad i \neq j$$

If errors are correlated, we call this autocorrelation.

When This Assumption is Violated:

- Often occurs in time series or geographically nested data.
- Example: Values from the previous year influence the current year.
- In cross-sectional data, autocorrelation is usually not a concern.

Implications of Autocorrelation:

- Standard errors are underestimated, leading to inflated significance.
- Model predictions may be biased if time dependence is ignored.

Testing for Autocorrelation in Stata

Dataset: `datasets/Durbin_Watson.dta`

Step 1: Set the Data for Time Series Analysis

```
tsset year
```

Step 2: Run the Regression Model

```
regress FDI GDPperCapita GDPGrowth incidence
```

Step 3: Perform Durbin-Watson Test

```
estat dwatson
```

Interpreting the Durbin-Watson Statistic:

- A value near 2 suggests no autocorrelation.
- A value near 0 suggests positive autocorrelation.
- A value near 4 suggests negative autocorrelation.

Addressing Autocorrelation

Solutions for Autocorrelation:

- Use robust standard errors for time-series data:

```
regress FDI GDPperCapita GDPGrowth incidence, vce(robust)
```

- Apply a first-difference transformation:

```
gen d_FDI = D.FDI
regress d_FDI GDPperCapita GDPGrowth incidence
```

- Use Generalized Least Squares (GLS) or Newey-West standard errors for time dependence.

Normally Distributed Errors

The residuals in an OLS regression should follow a normal distribution:

$$\varepsilon_i \sim N(0, \sigma^2), \quad \text{for all } i$$

This assumption ensures valid statistical inference, particularly in small samples.

Why This Assumption Matters:

- Normal errors are not required for OLS to be unbiased, but they affect:
  - The accuracy of t-tests and F-tests.
  - The efficiency of OLS estimates.
- Highly skewed distributions of the dependent variable or residuals can be problematic.

Testing for Normality in Stata

Step 1: Run the Regression Model

```
use ESSGBdiagnostics.dta, clear
quietly regress trstlgl age woman political_interest religious
```

Step 2: Generate Residuals

```
    predict res, residual
```

Step 3: Visual Inspection with Histogram

```
histogram res, normal
```

- This plots a histogram of residuals overlaid with a normal curve.

Step 4: Statistical Tests for Normality

```
summarize res, detail
sktest res
```

- The skewness/kurtosis test (sktest) checks if residuals follow a normal distribution.

Addressing Non-Normal Residuals

Solutions for Non-Normal Residuals:

- Use robust standard errors:

```
regress trstlgl age woman political_interest religious, vce(robust)
```

- Apply transformations to the dependent variable (e.g., log transformation for right-skewed distributions).
- Check for outliers and high-leverage points (we will talk about them on the next lecture) that distort normality.

When Normality is Less Important:

- In large samples, the Central Limit Theorem ensures that standard errors remain valid even when residuals are not perfectly normal.
- Focus should be on heteroskedasticity and omitted variable bias rather than strict normality.

References I

Laffers, L. (2021). *Draft poznámok k predmetu Moderná Aplikovaná regresia 1*. UMB Banská Bystrica.

Mehmetoglu, M. and Jakobsen, T. G. (2022). *Applied Statistics using Stata: a Guide for the Social Sciences*. Sage.