

Assumptions for Regression Analysis

Mgmt 230: Introductory Statistics

- Learn about the assumptions behind OLS estimation.
- Learn how to evaluate the validity of these assumptions.
- Introduce how to handle cases where the assumptions may be violated.

- Learn about the assumptions behind OLS estimation.
- Learn how to evaluate the validity of these assumptions.
- Introduce how to handle cases where the assumptions may be violated.

- Learn about the assumptions behind OLS estimation.
- Learn how to evaluate the validity of these assumptions.
- Introduce how to handle cases where the assumptions may be violated.

You're textbook says the four assumptions behind valid OLS estimation are:

- 1 Linearity.
- 2 Independence of error term. ← not completely necessary.
- 3 Normality of the error terms ϵ_i . ← not true.
- 4 Stationary variance of ϵ_i . ← not completely necessary.

You're textbook fails to mention some very important assumptions:

- The explanatory variables must be independent of the error term.
- The explanatory variables must be stationary. Often not true in financial and economics.

Your textbook says the four assumptions behind valid OLS estimation are:

- 1 Linearity.
- 2 Independence of error term. ← not completely necessary.
- 3 Normality of the error terms ϵ_i . ← not true.
- 4 Stationary variance of ϵ_i . ← not completely necessary.

Your textbook fails to mention some very important assumptions:

- The explanatory variables must be independent of the error term.
- The explanatory variables must be stationary. Often not true in financial and economics.

Your textbook says the four assumptions behind valid OLS estimation are:

- 1 Linearity.
- 2 Independence of error term. ← not completely necessary.
- 3 Normality of the error terms ϵ_i . ← not true.
- 4 Stationary variance of ϵ_i . ← not completely necessary.

Your textbook fails to mention some very important assumptions:

- The explanatory variables must be independent of the error term.
- The explanatory variables must be stationary. Often not true in financial and economics.

Your textbook says the four assumptions behind valid OLS estimation are:

- 1 Linearity.
- 2 Independence of error term. ← not completely necessary.
- 3 Normality of the error terms ϵ_i . ← not true.
- 4 Stationary variance of ϵ_i . ← not completely necessary.

Your textbook fails to mention some very important assumptions:

- The explanatory variables must be independent of the error term.
- The explanatory variables must be stationary. Often not true in financial and economics.

Your textbook says the four assumptions behind valid OLS estimation are:

- 1 Linearity.
- 2 Independence of error term. ← not completely necessary.
- 3 Normality of the error terms ϵ_i . ← not true.
- 4 Stationary variance of ϵ_i . ← not completely necessary.

Your textbook fails to mention some very important assumptions:

- The explanatory variables must be independent of the error term.
- The explanatory variables must be stationary. Often not true in financial and economics.

Your textbook says the four assumptions behind valid OLS estimation are:

- 1 Linearity.
- 2 Independence of error term. ← not completely necessary.
- 3 Normality of the error terms ϵ_i . ← not true.
- 4 Stationary variance of ϵ_i . ← not completely necessary.

Your textbook fails to mention some very important assumptions:

- The explanatory variables must be independent of the error term.
- The explanatory variables must be stationary. Often not true in financial and economics.

Your textbook says the four assumptions behind valid OLS estimation are:

- 1 Linearity.
- 2 Independence of error term. ← not completely necessary.
- 3 Normality of the error terms ϵ_i . ← not true.
- 4 Stationary variance of ϵ_i . ← not completely necessary.

Your textbook fails to mention some very important assumptions:

- The explanatory variables must be independent of the error term.
- The explanatory variables must be stationary. Often not true in financial and economics.

Your textbook says the four assumptions behind valid OLS estimation are:

- 1 Linearity.
- 2 Independence of error term. ← not completely necessary.
- 3 Normality of the error terms ϵ_i . ← not true.
- 4 Stationary variance of ϵ_i . ← not completely necessary.

Your textbook fails to mention some very important assumptions:

- The explanatory variables must be independent of the error term.
- The explanatory variables must be stationary. Often not true in financial and economics.

- Linear model must be an accurate description of the true relationship between the variables.

$$y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \dots + \beta_{k-1,i} x_{k-1} + \epsilon_i$$

- Evaluating Linearity:
 - Scatter plot with a trend line.
 - Scatter plot of the residuals.

- Linear model must be an accurate description of the true relationship between the variables.

$$y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \dots + \beta_{k-1} x_{k-1} + \epsilon_i$$

- Evaluating Linearity:
 - Scatter plot with a trend line.
 - Scatter plot of the residuals.

- Linear model must be an accurate description of the true relationship between the variables.

$$y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \dots + \beta_{k-1} x_{k-1} + \epsilon_i$$

- Evaluating Linearity:
 - Scatter plot with a trend line.
 - Scatter plot of the residuals.

- Linear model must be an accurate description of the true relationship between the variables.

$$y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \dots + \beta_{k-1,i} x_{k-1} + \epsilon_i$$

- Evaluating Linearity:
 - Scatter plot with a trend line.
 - Scatter plot of the residuals.

- Linear model must be an accurate description of the true relationship between the variables.

$$y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \dots + \beta_{k-1,i} x_{k-1} + \epsilon_i$$

- Evaluating Linearity:
 - Scatter plot with a trend line.
 - Scatter plot of the residuals.

- Quadratic relationship. y_i increases as x_i increases, but then decreases (or vice versa).
 - To account for this possibility, also put x_i^2 into the model.
 - Example: worker productivity and age.
- Exponential relationship. y_i increases as x_i increases, but at an increasing rate.
 - To account for this possibility, put e^{x_i} into the model instead of x_i .
 - Example: Total costs and total output.
- Logarithmic relationship. y_i increases as x_i increases, but at a decreasing rate.
 - To account for this possibility, put $\ln(x_i)$ into the model instead of x_i .
 - A quadratic relationship may appropriately capture this relationship as well.
 - Example: earnings and experience.

- Quadratic relationship. y_i increases as x_i increases, but then decreases (or vice versa).
 - To account for this possibility, also put x_i^2 into the model.
 - Example: worker productivity and age.
- Exponential relationship. y_i increases as x_i increases, but at an increasing rate.
 - To account for this possibility, put e^{x_i} into the model instead of x_i .
 - Example: Total costs and total output.
- Logarithmic relationship. y_i increases as x_i increases, but at a decreasing rate.
 - To account for this possibility, put $\ln(x_i)$ into the model instead of x_i .
 - A quadratic relationship may appropriately capture this relationship as well.
 - Example: earnings and experience.

- Quadratic relationship. y_i increases as x_i increases, but then decreases (or vice versa).
 - To account for this possibility, also put x_i^2 into the model.
 - Example: worker productivity and age.
- Exponential relationship. y_i increases as x_i increases, but at an increasing rate.
 - To account for this possibility, put e^{x_i} into the model instead of x_i .
 - Example: Total costs and total output.
- Logarithmic relationship. y_i increases as x_i increases, but at a decreasing rate.
 - To account for this possibility, put $\ln(x_i)$ into the model instead of x_i .
 - A quadratic relationship may appropriately capture this relationship as well.
 - Example: earnings and experience.

- Quadratic relationship. y_i increases as x_i increases, but then decreases (or vice versa).
 - To account for this possibility, also put x_i^2 into the model.
 - Example: worker productivity and age.
- Exponential relationship. y_i increases as x_i increases, but at an increasing rate.
 - To account for this possibility, put e^{x_i} into the model instead of x_i .
 - Example: Total costs and total output.
- Logarithmic relationship. y_i increases as x_i increases, but at a decreasing rate.
 - To account for this possibility, put $\ln(x_i)$ into the model instead of x_i .
 - A quadratic relationship may appropriately capture this relationship as well.
 - Example: earnings and experience.

- Quadratic relationship. y_i increases as x_i increases, but then decreases (or vice versa).
 - To account for this possibility, also put x_i^2 into the model.
 - Example: worker productivity and age.
- Exponential relationship. y_i increases as x_i increases, but at an increasing rate.
 - To account for this possibility, put e^{x_i} into the model instead of x_i .
 - Example: Total costs and total output.
- Logarithmic relationship. y_i increases as x_i increases, but at a decreasing rate.
 - To account for this possibility, put $\ln(x_i)$ into the model instead of x_i .
 - A quadratic relationship may appropriately capture this relationship as well.
 - Example: earnings and experience.

- Quadratic relationship. y_i increases as x_i increases, but then decreases (or vice versa).
 - To account for this possibility, also put x_i^2 into the model.
 - Example: worker productivity and age.
- Exponential relationship. y_i increases as x_i increases, but at an increasing rate.
 - To account for this possibility, put e^{x_i} into the model instead of x_i .
 - Example: Total costs and total output.
- Logarithmic relationship. y_i increases as x_i increases, but at a decreasing rate.
 - To account for this possibility, put $\ln(x_i)$ into the model instead of x_i .
 - A quadratic relationship may appropriately capture this relationship as well.
 - Example: earnings and experience.

- Quadratic relationship. y_i increases as x_i increases, but then decreases (or vice versa).
 - To account for this possibility, also put x_i^2 into the model.
 - Example: worker productivity and age.
- Exponential relationship. y_i increases as x_i increases, but at an increasing rate.
 - To account for this possibility, put e^{x_i} into the model instead of x_i .
 - Example: Total costs and total output.
- Logarithmic relationship. y_i increases as x_i increases, but at a decreasing rate.
 - To account for this possibility, put $\ln(x_i)$ into the model instead of x_i .
 - A quadratic relationship may appropriately capture this relationship as well.
 - Example: earnings and experience.

- Quadratic relationship. y_i increases as x_i increases, but then decreases (or vice versa).
 - To account for this possibility, also put x_i^2 into the model.
 - Example: worker productivity and age.
- Exponential relationship. y_i increases as x_i increases, but at an increasing rate.
 - To account for this possibility, put e^{x_i} into the model instead of x_i .
 - Example: Total costs and total output.
- Logarithmic relationship. y_i increases as x_i increases, but at a decreasing rate.
 - To account for this possibility, put $\ln(x_i)$ into the model instead of x_i .
 - A quadratic relationship may appropriately capture this relationship as well.
 - Example: earnings and experience.

- Quadratic relationship. y_i increases as x_i increases, but then decreases (or vice versa).
 - To account for this possibility, also put x_i^2 into the model.
 - Example: worker productivity and age.
- Exponential relationship. y_i increases as x_i increases, but at an increasing rate.
 - To account for this possibility, put e^{x_i} into the model instead of x_i .
 - Example: Total costs and total output.
- Logarithmic relationship. y_i increases as x_i increases, but at a decreasing rate.
 - To account for this possibility, put $\ln(x_i)$ into the model instead of x_i .
 - A quadratic relationship may appropriately capture this relationship as well.
 - Example: earnings and experience.

- Quadratic relationship. y_i increases as x_i increases, but then decreases (or vice versa).
 - To account for this possibility, also put x_i^2 into the model.
 - Example: worker productivity and age.
- Exponential relationship. y_i increases as x_i increases, but at an increasing rate.
 - To account for this possibility, put e^{x_i} into the model instead of x_i .
 - Example: Total costs and total output.
- Logarithmic relationship. y_i increases as x_i increases, but at a decreasing rate.
 - To account for this possibility, put $\ln(x_i)$ into the model instead of x_i .
 - A quadratic relationship may appropriately capture this relationship as well.
 - Example: earnings and experience.

- This assumption states that an error from one observation (ϵ_i) is independent of the error from another observation (ϵ_j).
- This often happens in financial and economic time series data.
- Satisfying this assumption *is not necessary* for OLS results to be consistent. But, better methods than OLS are possible.
- **Consistency:** An estimate is consistent if as the sample size gets very large, the sample estimates for the coefficients approach the true population coefficients.
- If the residuals are not independent, this most likely indicates you mis-specified the model (i.e. linearity assumption is violated).

- This assumption states that an error from one observation (ϵ_i) is independent of the error from another observation (ϵ_j).
- This often happens in financial and economic time series data.
- Satisfying this assumption *is not necessary* for OLS results to be consistent. But, better methods than OLS are possible.
- **Consistency:** An estimate is consistent if as the sample size gets very large, the sample estimates for the coefficients approach the true population coefficients.
- If the residuals are not independent, this most likely indicates you mis-specified the model (i.e. linearity assumption is violated).

- This assumption states that an error from one observation (ϵ_i) is independent of the error from another observation (ϵ_j).
- This often happens in financial and economic time series data.
- Satisfying this assumption *is not necessary* for OLS results to be consistent. But, better methods than OLS are possible.
- **Consistency:** An estimate is consistent if as the sample size gets very large, the sample estimates for the coefficients approach the true population coefficients.
- If the residuals are not independent, this most likely indicates you mis-specified the model (i.e. linearity assumption is violated).

- This assumption states that an error from one observation (ϵ_i) is independent of the error from another observation (ϵ_j).
- This often happens in financial and economic time series data.
- Satisfying this assumption *is not necessary* for OLS results to be consistent. But, better methods than OLS are possible.
- **Consistency:** An estimate is consistent if as the sample size gets very large, the sample estimates for the coefficients approach the true population coefficients.
- If the residuals are not independent, this most likely indicates you mis-specified the model (i.e. linearity assumption is violated).

- This assumption states that an error from one observation (ϵ_i) is independent of the error from another observation (ϵ_j).
- This often happens in financial and economic time series data.
- Satisfying this assumption *is not necessary* for OLS results to be consistent. But, better methods than OLS are possible.
- **Consistency:** An estimate is consistent if as the sample size gets very large, the sample estimates for the coefficients approach the true population coefficients.
- If the residuals are not independent, this most likely indicates you mis-specified the model (i.e. linearity assumption is violated).

- Why was the central limit theorem so cool?
- Correct assumption: the sample size is sufficiently large *or* the population error term is normally distributed.
- If this assumption holds:
 - The sampling distribution of the estimates for the coefficients (b 's) will be normal.
 - The residuals will be normal.
- Forget about rules of thumb like $n > 30$ for regression.
- To evaluate if this assumption holds, can do a histogram of the residuals.

- Why was the central limit theorem so cool?
- Correct assumption: the sample size is sufficiently large *or* the population error term is normally distributed.
- If this assumption holds:
 - The sampling distribution of the estimates for the coefficients (b 's) will be normal.
 - The residuals will be normal.
- Forget about rules of thumb like $n > 30$ for regression.
- To evaluate if this assumption holds, can do a histogram of the residuals.

- Why was the central limit theorem so cool?
- Correct assumption: the sample size is sufficiently large *or* the population error term is normally distributed.
- If this assumption holds:
 - The sampling distribution of the estimates for the coefficients (b 's) will be normal.
 - The residuals will be normal.
- Forget about rules of thumb like $n > 30$ for regression.
- To evaluate if this assumption holds, can do a histogram of the residuals.

- Why was the central limit theorem so cool?
- Correct assumption: the sample size is sufficiently large *or* the population error term is normally distributed.
- If this assumption holds:
 - The sampling distribution of the estimates for the coefficients (b 's) will be normal.
 - The residuals will be normal.
- Forget about rules of thumb like $n > 30$ for regression.
- To evaluate if this assumption holds, can do a histogram of the residuals.

- Why was the central limit theorem so cool?
- Correct assumption: the sample size is sufficiently large *or* the population error term is normally distributed.
- If this assumption holds:
 - The sampling distribution of the estimates for the coefficients (b 's) will be normal.
 - The residuals will be normal.
- Forget about rules of thumb like $n > 30$ for regression.
- To evaluate if this assumption holds, can do a histogram of the residuals.

- Why was the central limit theorem so cool?
- Correct assumption: the sample size is sufficiently large *or* the population error term is normally distributed.
- If this assumption holds:
 - The sampling distribution of the estimates for the coefficients (b 's) will be normal.
 - The residuals will be normal.
- Forget about rules of thumb like $n > 30$ for regression.
- To evaluate if this assumption holds, can do a histogram of the residuals.

- Why was the central limit theorem so cool?
- Correct assumption: the sample size is sufficiently large *or* the population error term is normally distributed.
- If this assumption holds:
 - The sampling distribution of the estimates for the coefficients (b 's) will be normal.
 - The residuals will be normal.
- Forget about rules of thumb like $n > 30$ for regression.
- To evaluate if this assumption holds, can do a histogram of the residuals.

- The population error term should have a constant variance.
- The variance should not increase as x_i increases.
- This often happens with data related to income or wealth.
 - Suppose you are predicting how much people spend on luxury goods.
 - Larger errors are going to be made for people with larger incomes.
- Satisfying this assumption *is not necessary* for consistency, although better methods than OLS exist for estimating models with this problem.
- To evaluate if this assumption holds, do a scatter plot of the residuals on the y-axis, and the x variable on the x-axis.

- The population error term should have a constant variance.
- The variance should not increase as x_i increases.
- This often happens with data related to income or wealth.
 - Suppose you are predicting how much people spend on luxury goods.
 - Larger errors are going to be made for people with larger incomes.
- Satisfying this assumption *is not necessary* for consistency, although better methods than OLS exist for estimating models with this problem.
- To evaluate if this assumption holds, do a scatter plot of the residuals on the y-axis, and the x variable on the x-axis.

- The population error term should have a constant variance.
- The variance should not increase as x_i increases.
- This often happens with data related to income or wealth.
 - Suppose you are predicting how much people spend on luxury goods.
 - Larger errors are going to be made for people with larger incomes.
- Satisfying this assumption *is not necessary* for consistency, although better methods than OLS exist for estimating models with this problem.
- To evaluate if this assumption holds, do a scatter plot of the residuals on the y-axis, and the x variable on the x-axis.

- The population error term should have a constant variance.
- The variance should not increase as x_i increases.
- This often happens with data related to income or wealth.
 - Suppose you are predicting how much people spend on luxury goods.
 - Larger errors are going to be made for people with larger incomes.
- Satisfying this assumption *is not necessary* for consistency, although better methods than OLS exist for estimating models with this problem.
- To evaluate if this assumption holds, do a scatter plot of the residuals on the y-axis, and the x variable on the x-axis.

- The population error term should have a constant variance.
- The variance should not increase as x_i increases.
- This often happens with data related to income or wealth.
 - Suppose you are predicting how much people spend on luxury goods.
 - Larger errors are going to be made for people with larger incomes.
- Satisfying this assumption *is not necessary* for consistency, although better methods than OLS exist for estimating models with this problem.
- To evaluate if this assumption holds, do a scatter plot of the residuals on the y-axis, and the x variable on the x-axis.

- The population error term should have a constant variance.
- The variance should not increase as x_i increases.
- This often happens with data related to income or wealth.
 - Suppose you are predicting how much people spend on luxury goods.
 - Larger errors are going to be made for people with larger incomes.
- Satisfying this assumption *is not necessary* for consistency, although better methods than OLS exist for estimating models with this problem.
- To evaluate if this assumption holds, do a scatter plot of the residuals on the y-axis, and the x variable on the x-axis.

- The population error term should have a constant variance.
- The variance should not increase as x_i increases.
- This often happens with data related to income or wealth.
 - Suppose you are predicting how much people spend on luxury goods.
 - Larger errors are going to be made for people with larger incomes.
- Satisfying this assumption *is not necessary* for consistency, although better methods than OLS exist for estimating models with this problem.
- To evaluate if this assumption holds, do a scatter plot of the residuals on the y-axis, and the x variable on the x-axis.

- For consistent and unbiased results, the X variables *must be independent* of the population error term (ϵ_j).
- That is, the errors made in the regression cannot be related to your variables.
- Omitted variable bias: when there are possible explanatory variables (that may not even be measurable) not included in the regression that are correlated with the included explanatory variables.
- The error term accounts for anything not included in the regression.

- For consistent and unbiased results, the X variables *must be independent* of the population error term (ϵ_j).
- That is, the errors made in the regression cannot be related to your variables.
- Omitted variable bias: when there are possible explanatory variables (that may not even be measurable) not included in the regression that are correlated with the included explanatory variables.
- The error term accounts for anything not included in the regression.

- For consistent and unbiased results, the X variables *must be independent* of the population error term (ϵ_j).
- That is, the errors made in the regression cannot be related to your variables.
- Omitted variable bias: when there are possible explanatory variables (that may not even be measurable) not included in the regression that are correlated with the included explanatory variables.
- The error term accounts for anything not included in the regression.

- For consistent and unbiased results, the X variables *must be independent* of the population error term (ϵ_j).
- That is, the errors made in the regression cannot be related to your variables.
- Omitted variable bias: when there are possible explanatory variables (that may not even be measurable) not included in the regression that are correlated with the included explanatory variables.
- The error term accounts for anything not included in the regression.

- The explanatory variables must be stationary.
- Economics and time series data are often not stationary, rather they grow as time goes on.
- Examples: GDP, income, price level, wages.
- It can be *very tough* to handle this problem.

- The explanatory variables must be stationary.
- Economics and time series data are often not stationary, rather they grow as time goes on.
- Examples: GDP, income, price level, wages.
- It can be *very tough* to handle this problem.

- The explanatory variables must be stationary.
- Economics and time series data are often not stationary, rather they grow as time goes on.
- Examples: GDP, income, price level, wages.
- It can be *very tough* to handle this problem.

- The explanatory variables must be stationary.
- Economics and time series data are often not stationary, rather they grow as time goes on.
- Examples: GDP, income, price level, wages.
- It can be *very tough* to handle this problem.