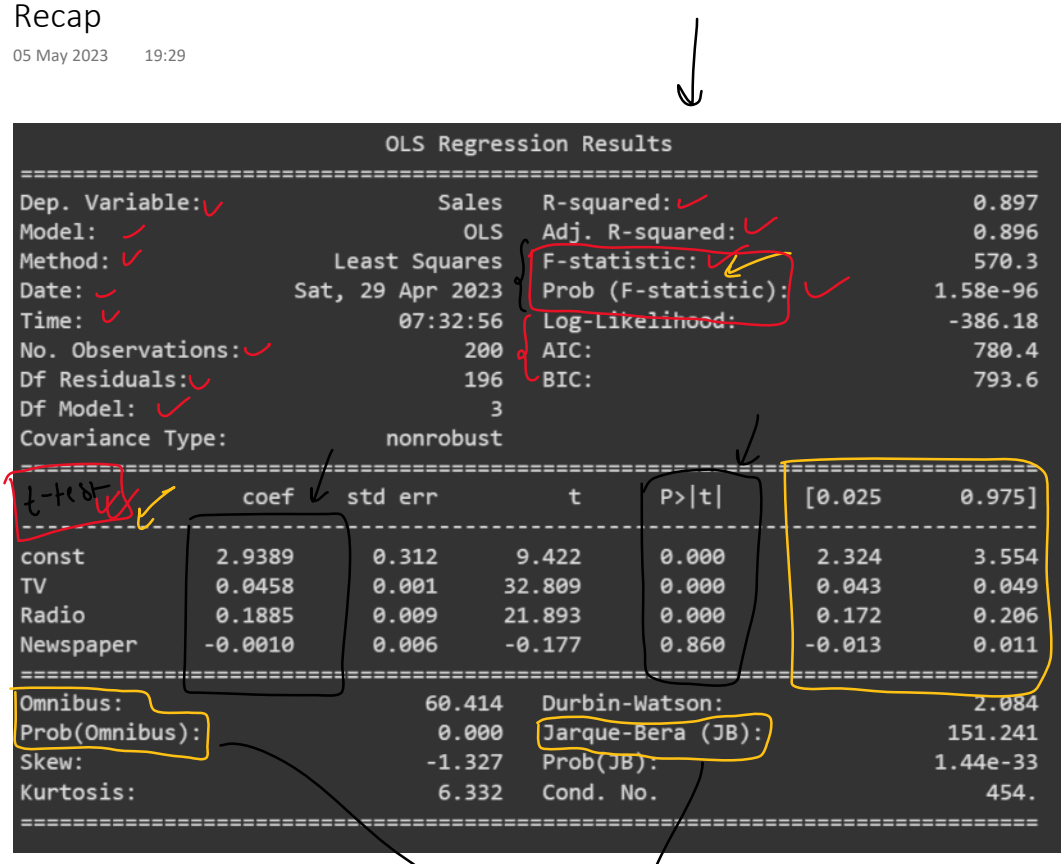
# Recap

05 May 2023 19:29



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: | Sat, 29 Apr 2023 | Prob (F-statistic): | 1.58e-96 |
| Time: | 07:32:56 | 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. |

t-test

Assumption

test of normality of residuals

Linear → Coefficient → OLS, GD → Regression Analysis ②

half → Assumption of LR → Multicollinearity

# Assumptions of Linear Regression

03 May 2023 14:17

Linear regression relies on several assumptions to ensure the validity and reliability of the estimates and inferences. The key assumptions of linear regression are:

1. Linearity
2. Normality of Residuals
3. Homoscedasticity
4. No Autocorrelation
5. No or little Multicollinearity

Tomorrow

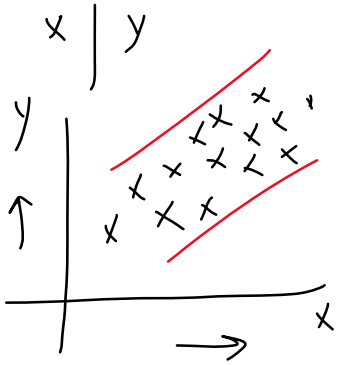
1) What
2) What if the assumption fails
3) Detect
4) Remedy

# 1. Linearity

03 May 2023 14:21

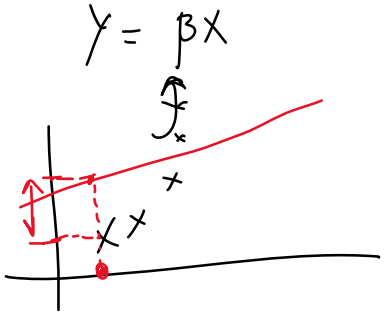
## The Assumption

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.



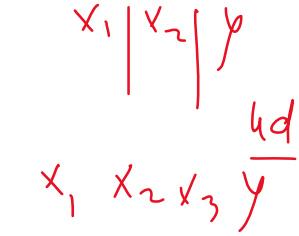
## What happens when this assumption is violated?

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.

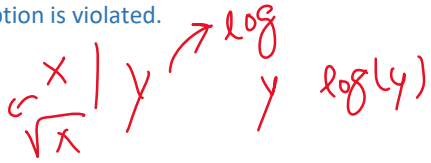


## How to check this assumption

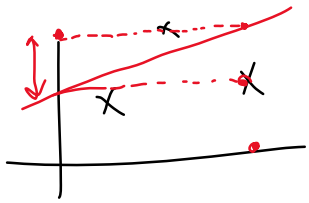
1. Scatter plots: Create scatter plots of the dependent variable against each independent variable. If the relationship appears to be linear, the linearity assumption is likely satisfied. Nonlinear patterns or other trends may indicate that the assumption is violated.
2. Residual plots: Plot the residuals (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.

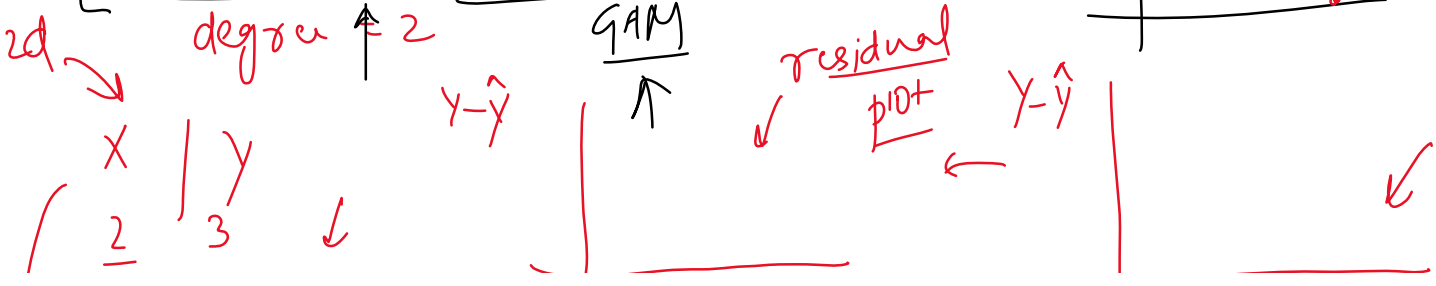


## What to do when the assumption fails?



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.





2 1 3

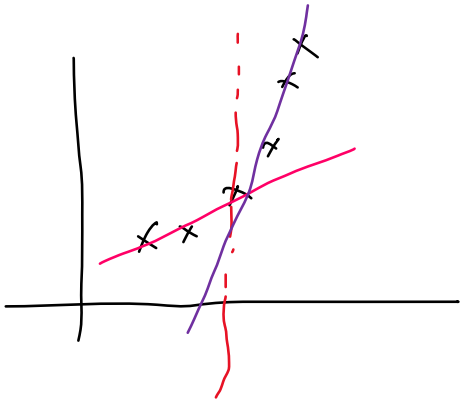
$y = \beta_0 + \beta_1 x$

| $x^0$ | $x^1$ | $x^2$ | $y$ |
|---|---|---|---|
| 1 | 2 | 4 | 3 |

4d

$\hat{y}$

$y = \beta_0 + \beta_1 x^0 + \beta_2 x^1 + \beta_3 x^2$



## 2. Normality of Residual

03 May 2023 16:36

### The Assumption

The error terms (residuals) are assumed to follow a normal distribution with a mean of zero and a constant variance.

$e_i \sim N(0, \sigma^2)$

### 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.

### 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

04 May 2023 10:03

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:

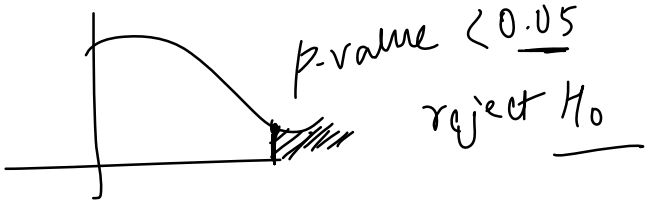
$$K^2 = n\left[\frac{(\text{skewness})^2}{6} + \frac{(\text{Kurtosis})^2}{24}\right]$$

chi square distrib

df = 2

$n \rightarrow$ 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.




0.63 > 0.05

cant reject $H_0$

< 0.05 → reject $H_0$

# 3. Homoscedasticity

03 May 2023 16:49

What is the problem

## 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.

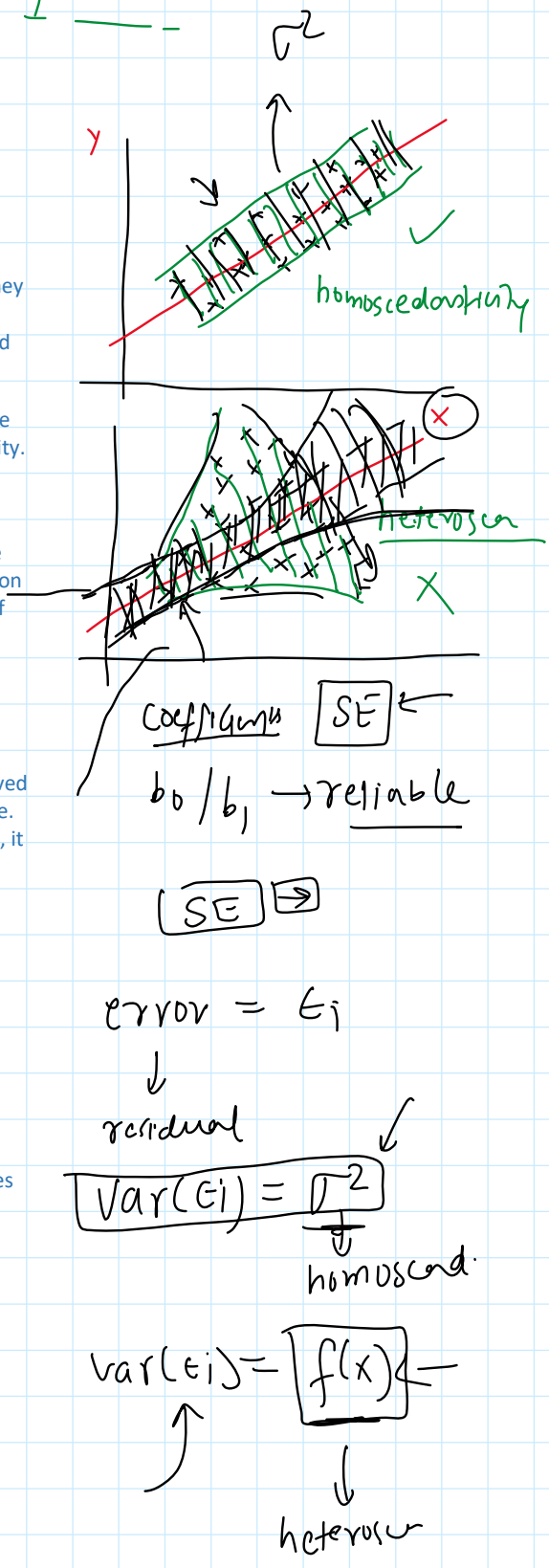
t-stats
Confia

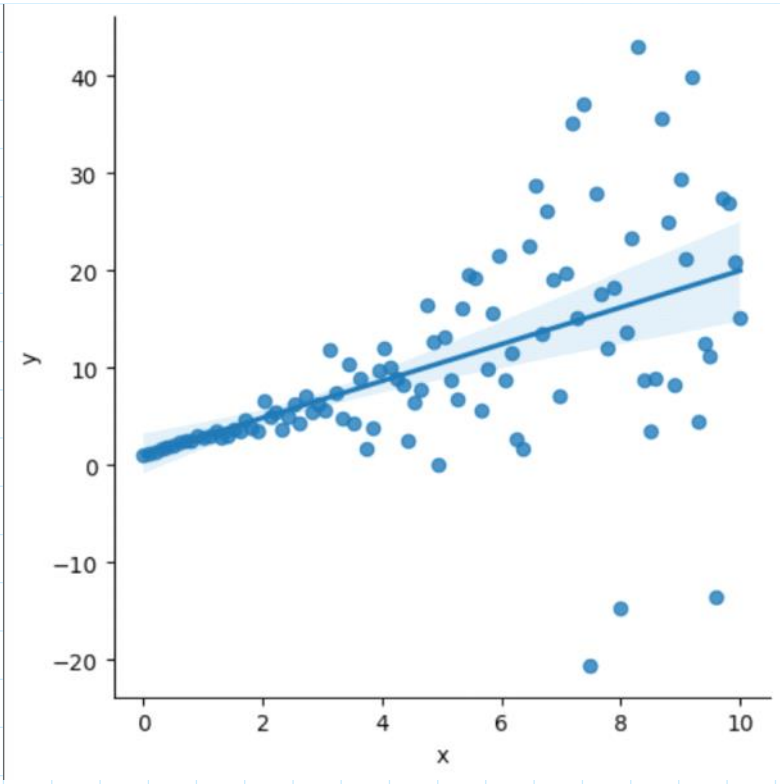
## How to check this assumption

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?

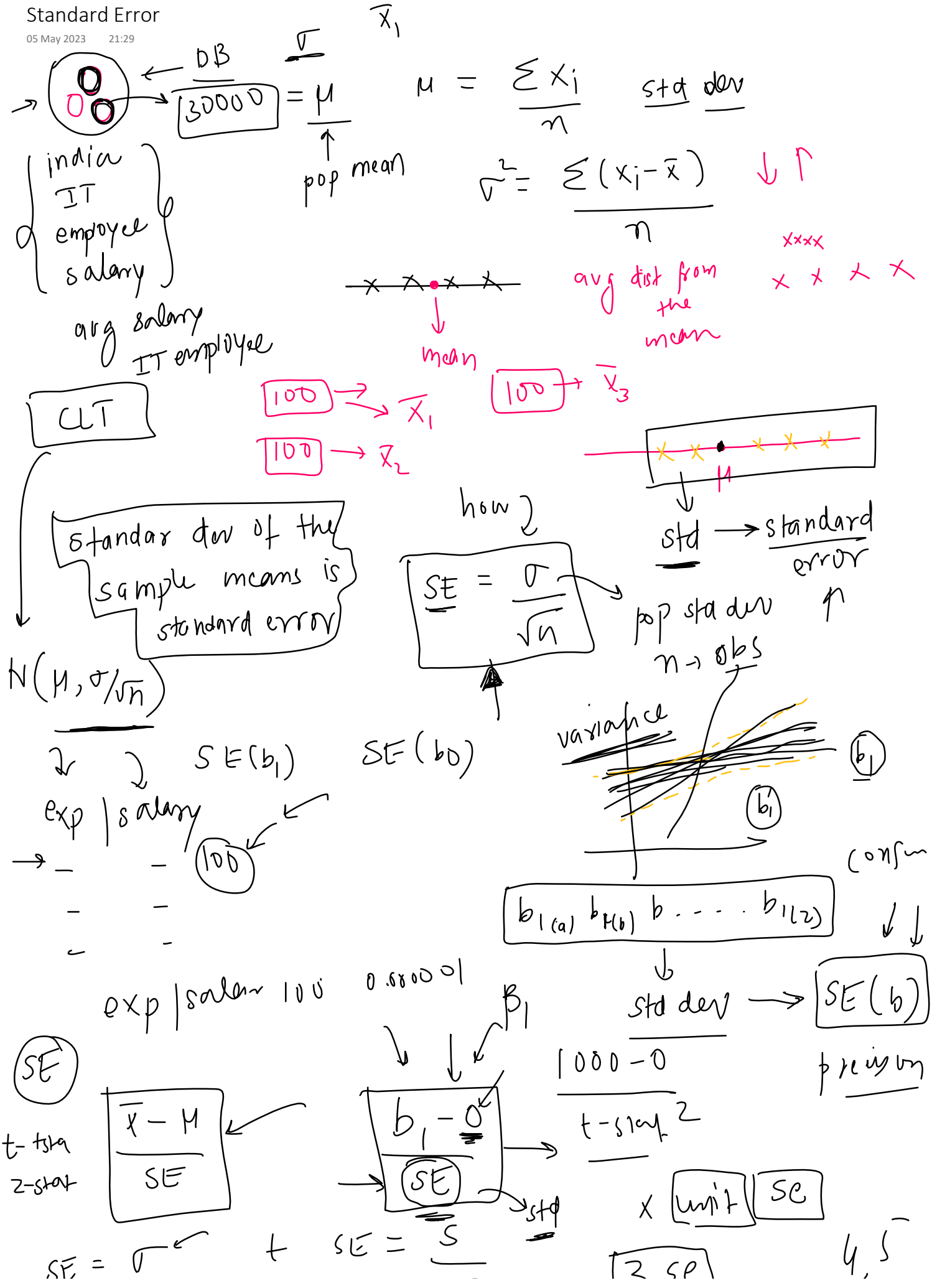
1. Transformations: Apply transformations to the dependent and/or independent variables to stabilize the variance of the residuals. Common transformations include logarithmic, square root, and inverse transformations.
2. Weighted Least Squares (WLS): Use a weighted least squares approach, which assigns 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.




40
30
20
10
0
−10
−20
y
0
2
4
6
8
10
x

Standard Error

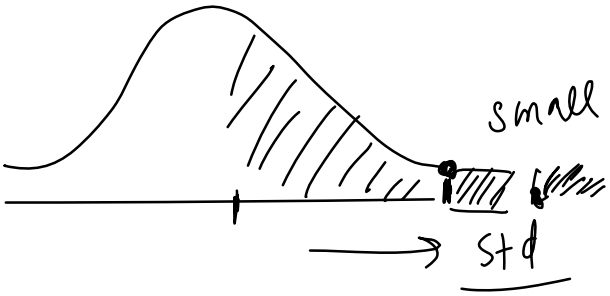
05 May 2023 21:29



$SE = \frac{\sigma}{\sqrt{n}}$ ← + $SE = \frac{s}{\sqrt{n}}$

3 se

4, 5



# Breusch-Pagan Test

05 May 2023 08:15

The Breusch-Pagan test, also known as the Cook-Weisberg test, is a statistical test used to detect heteroscedasticity in a linear regression model. The test is based on the assumption that the variance of the errors is a function of one or more independent variables. Here are the steps to perform the Breusch-Pagan test:

1. Estimate the linear regression model: Fit a linear regression model to the data using the ordinary least squares (OLS) method. Obtain the residuals (errors) from this model.
2. Calculate the squared residuals: Square each residual obtained in step 1.
3. Regress squared residuals on the independent variables: Perform another linear regression, this time with the squared residuals as the dependent variable and the same set of independent variables used in the original model. Obtain the R-squared value from this regression.
4. Calculate the test statistic: The Breusch-Pagan test statistic, known as the Lagrange Multiplier (LM) statistic, is calculated as follows:
   $LM = n * R^2$
   where n is the number of observations and $R^2$ is the R-squared value obtained in step 3.
5. Determine the p-value: The LM statistic follows a chi-squared distribution with k degrees of freedom, where k is the number of independent variables (excluding the constant term). Calculate the p-value for the LM statistic using the chi-squared distribution.
6. Make a decision based on the p-value: Compare the calculated p-value to a chosen significance level (usually $\alpha = 0.05$). If the p-value is less than or equal to $\alpha$, reject the null hypothesis and conclude that there is evidence of heteroscedasticity in the data. If the p-value is greater than $\alpha$, do not reject the null hypothesis and assume that the data exhibits homoscedasticity (constant variance of the residuals).

Note that the Breusch-Pagan test assumes a linear relationship between the independent variables and the variance of the errors. If the relationship is not linear, the test may not be appropriate, and other tests for heteroscedasticity should be considered.

# 4. No Autocorrelation

03 May 2023 16:49

## The Assumption

There should be no apparent correlation or pattern in the residuals, as this would suggest that the error terms are not independent.

## What happens when this assumption is violated?

1. Inefficient estimates: The parameter estimates (coefficients) remain unbiased, but they are no longer the best linear unbiased estimators (BLUE). The inefficiency of the estimates implies that the standard errors may be 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 no autocorrelation. If the error terms exhibit autocorrelation, 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 no autocorrelation. If this assumption is violated, the confidence intervals may not be accurate, affecting the interpretation of the effect sizes and the precision of the estimates.

## How to check this assumption

1. Durbin-Watson test: This is a formal statistical test for autocorrelation, specifically first-order autocorrelation. The Durbin-Watson test statistic ranges from 0 to 4, with a value of 2 indicating no autocorrelation. Values less than 2 suggest positive autocorrelation, while values greater than 2 indicate negative autocorrelation. It is important to note that the Durbin-Watson test is only applicable for first-order autocorrelation and may not detect higher-order autocorrelation.

## What to do when the assumption fails?

1. Lagged variables: Include lagged values of the dependent variable or the

independent variables as predictors in the model to account for the autocorrelation.

2. Differencing: Apply differencing to the dependent and/or independent variables, which can help remove the autocorrelation by focusing on the changes between consecutive observations rather than the absolute values.

3. Generalized least squares (GLS): Use a generalized least squares approach that accounts for the autocorrelation structure in the error terms, leading to more efficient and reliable estimates.

4. Time series models: Consider using specialized time series models, such as autoregressive (AR), moving average (MA), autoregressive integrated moving average (ARIMA), or seasonal decomposition of time series (STL), which are designed to handle autocorrelation.

5. Robust standard errors: Calculate robust standard errors that are more reliable under autocorrelation, such as Newey-West standard errors or HAC (heteroscedasticity and autocorrelation consistent) standard errors.

# 5. No Multicollinearity

03 May 2023 16:49

# Extra

04 May 2023 14:19