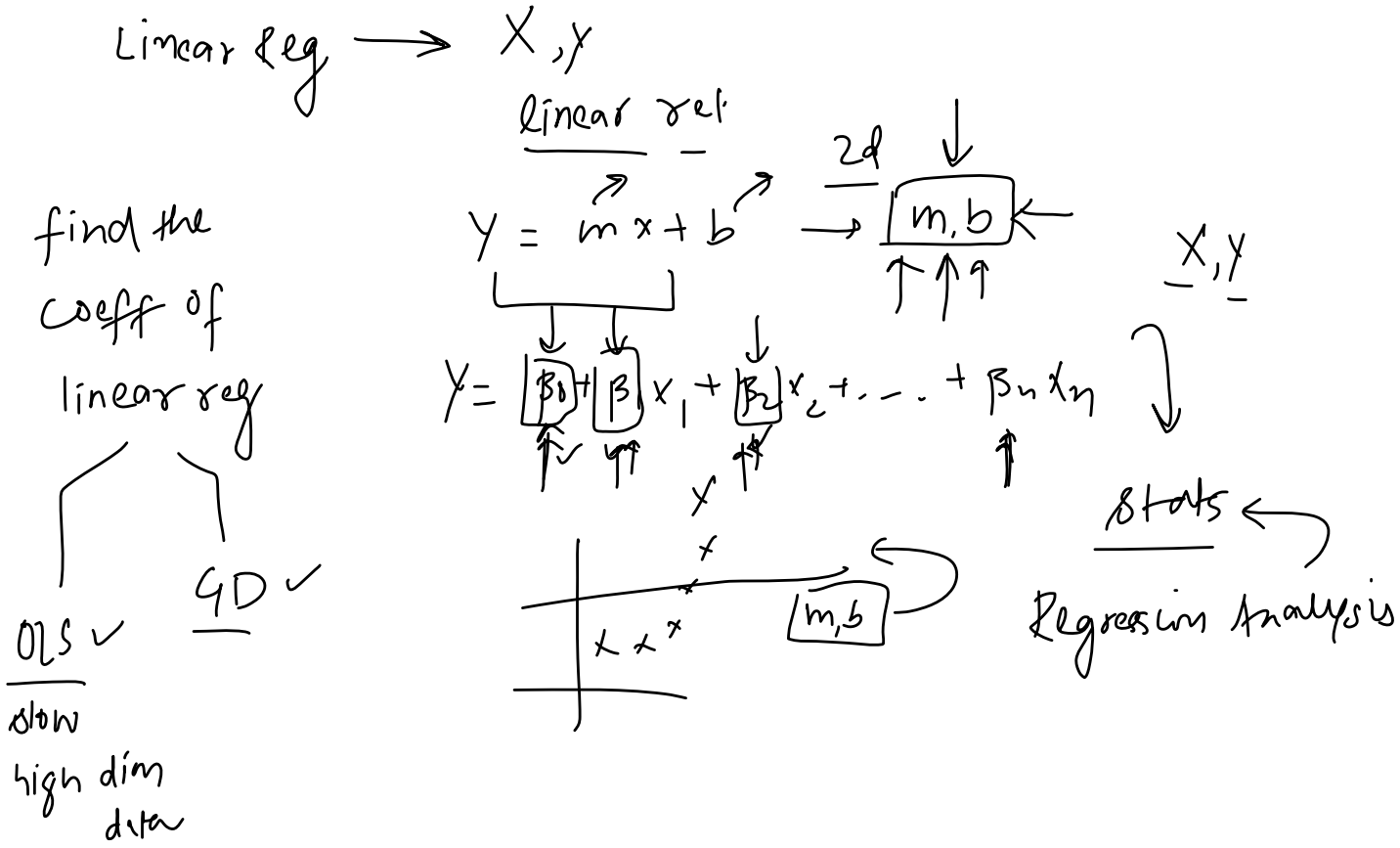
Till now

28 April 2023 06:49



Regression Analysis

29 April 2023 02:21

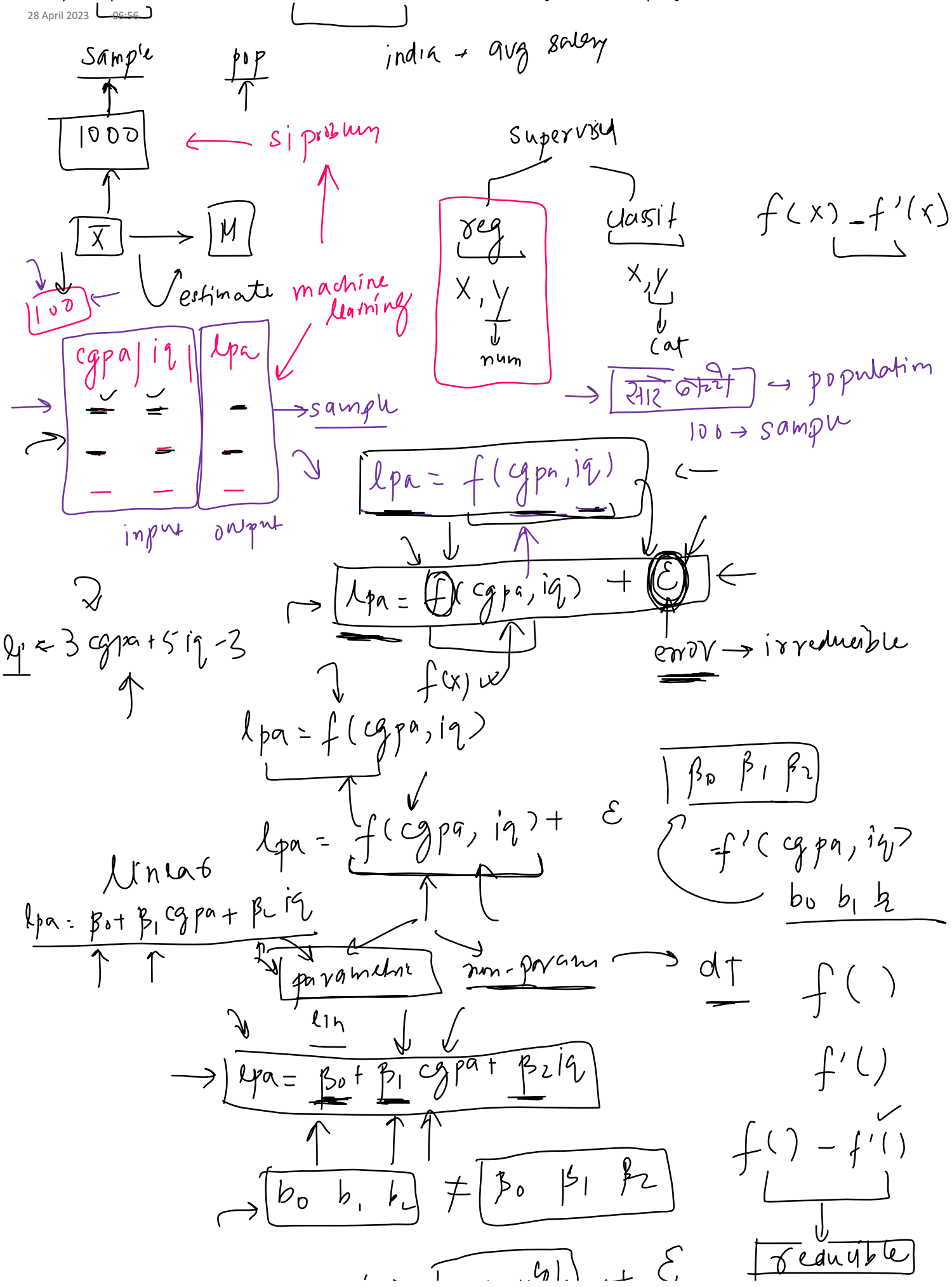
Regression analysis is a statistical method used to examine the relationship between one dependent variable and one or more independent variables. The goal of regression analysis is to understand how the dependent variable changes when one or more independent variables are altered, and to create a model that can predict the value of the dependent variable based on the values of the independent variables.

1. Define the research question: Identify the dependent variable (the variable you want to predict or explain) and the independent variable(s) (the variables that you think influence the dependent variable).
2. Collect and prepare data: Gather data for the dependent and independent variables. The data should be organized in a tabular format, with each row representing an observation and each column representing a variable. It's essential to clean and pre-process the data to handle missing values, outliers, and other potential issues that may affect the analysis.
3. Visualize the data: Before fitting a linear regression model, it's helpful to create scatter plots to visualize the relationship between the dependent variable and each independent variable. This can help you identify trends, outliers, and any potential issues with the data.
4. Check assumptions: Linear regression has some underlying assumptions, including linearity, independence of errors, homoscedasticity (constant variance of errors), and normality of errors. You can use diagnostic plots and statistical tests to check whether these assumptions hold for your data.
5. Fit the linear regression model: Use statistical software (e.g., R, Python, or Excel) to fit a linear regression model to your data. The model will estimate the regression coefficients (intercept and slope) that minimize the sum of squared residuals (i.e., the differences between the observed and predicted values of the dependent variable).
6. Interpret the model: Analyse the estimated regression coefficients, their standard errors, t-values, and p-values to determine the statistical significance of the relationship between the dependent and independent variables. The R-squared value and adjusted R-squared value can provide insights into the goodness-of-fit of the model and the proportion of variation in the dependent variable explained by the independent variables.
7. Validate the model: If you have a sufficiently large dataset, you can split it into a training and testing set. Fit the linear regression model to the training set, and then use the model to predict the dependent variable in the testing set. Calculate the mean squared error, root mean squared error, or another performance metric to assess the predictive accuracy of the model.
8. Report results: Summarize the findings of the linear regression analysis in a clear and concise manner, including the estimated coefficients, their interpretation, and any limitations or assumptions that may impact the results.

1. What's the statistics connection?
2. Why is Regression Analysis required?

Why ML problems are a Statistical Inference Problems? [with Example]

28 April 2023 06:56



reducible

$lpa = f'(cgpa, iq) + \boxed{\text{reducible}} + \varepsilon$

estimate of f using given data

$f'() \simeq f()$ ↑ true of x, y for pop

ε

$y = \boxed{2x - 5} + \boxed{\text{some randomness}}$

$f(x) = 2x - 5$

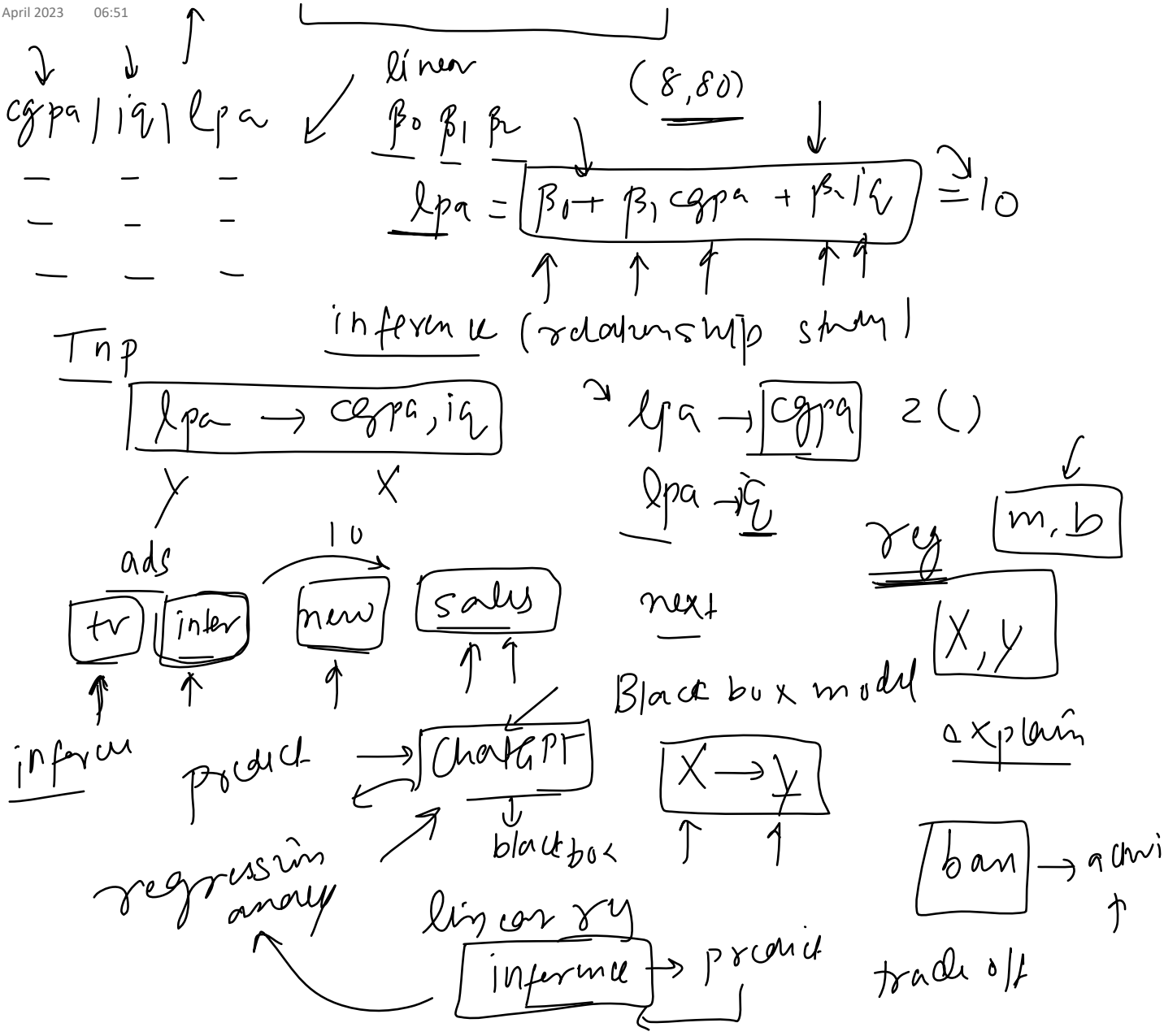
$\beta_0 = -5$ pop parameters

$\beta_1 = 2$

$\boxed{b_0 \quad b_1}$ → current set of 50 points

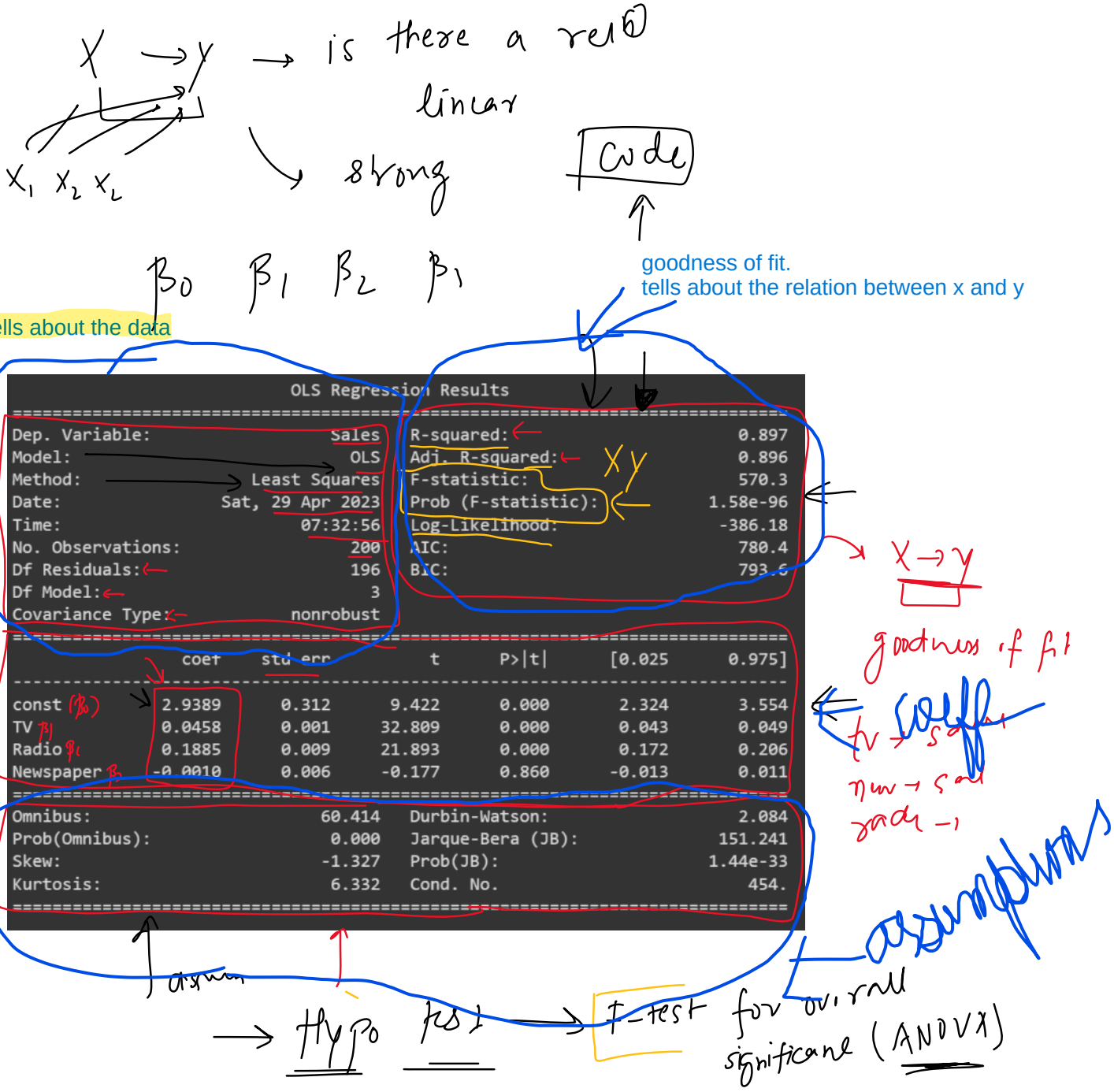
Inference Vs Prediction [Why regression analysis is required?]

28 April 2023 06:51



Statsmodel Linear Regression

28 April 2023 06:59



```
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.
==============================================================================
```

feature selection

CI

$x_1 \to y$

$x_2 \to y$

$x_3 \to y$

f-sla

SE b1

$0.0458 \pm 3.18 \times 0.01$

T-test → formula → Intuition → derivation → Assumptions

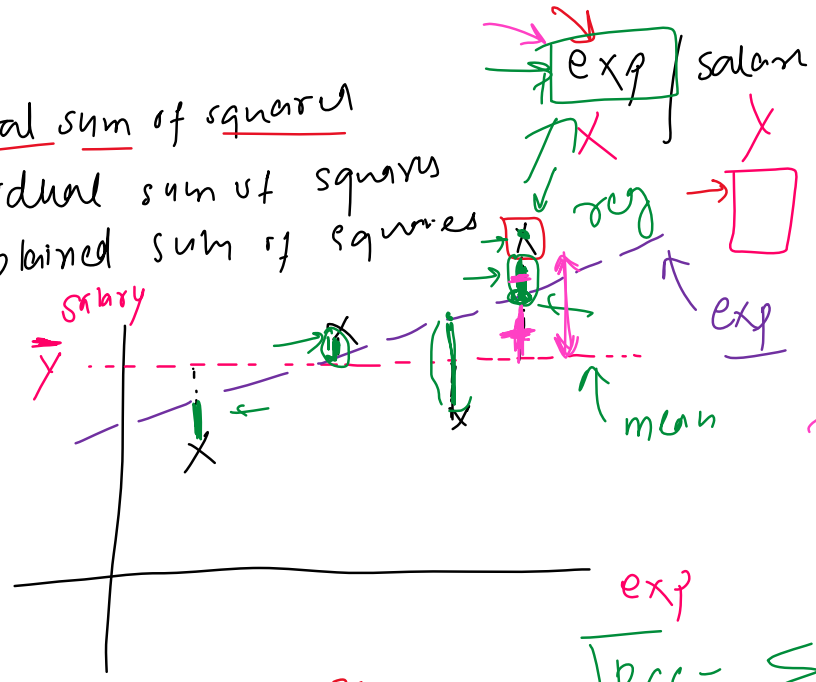
TSS, RSS and ESS

29 April 2023 04:16

TSS → Total sum of squares

RSS → Residual sum of squares

ESS → Explained sum of squares



$Y = 5X + 2$

TSS − RSS

ESS → explained variance

$$TSS = \sum (y_i - \bar{y})^2$$

$$RSS = \sum (y_i - \hat{y}_i)^2$$

overall variance in data

f(x)

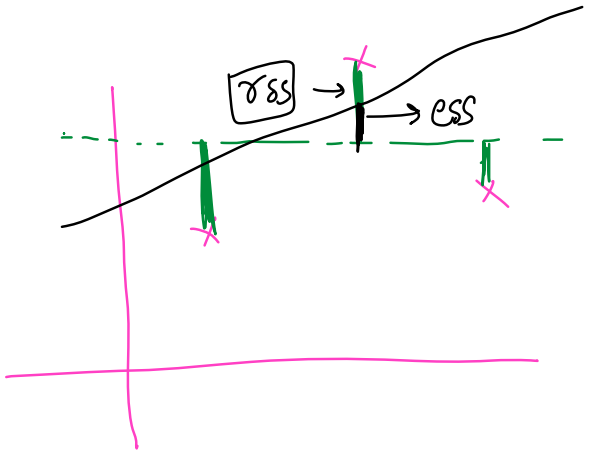
TSS → ESS + RSS + ε

Explained

unexplained

reducible

irreducible



total variance

Degree of Freedom

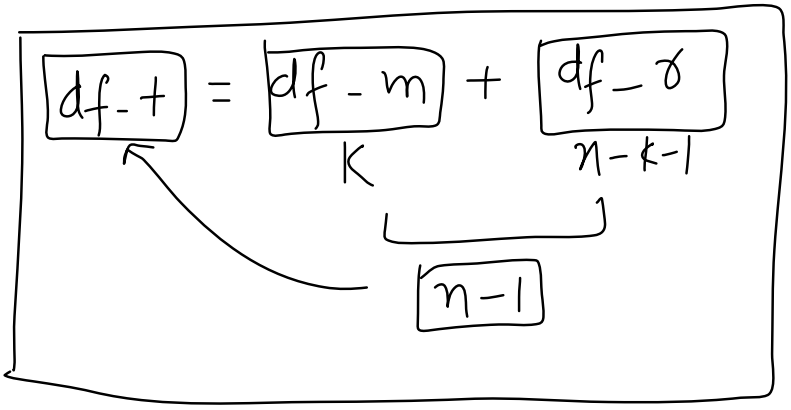
28 April 2023 07:00

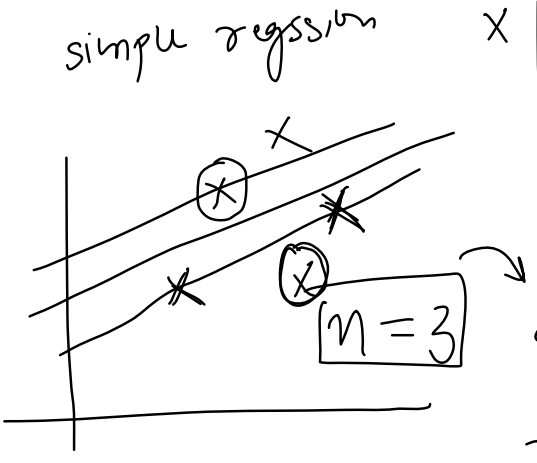
In linear regression, the total degrees of freedom (df_total) represent the total number of data points minus 1. It represents the overall variability in the dataset that can be attributed to both the model and the residuals.
For a linear regression with n data points (observations), the total degrees of freedom can be calculated as:
df_total = n - 1
where: n is the number of data points (observations) in the dataset

The total degrees of freedom in linear regression is divided into two components:

1. Degrees of freedom for the model (df_model): This is equal to the number of independent variables in the model (k).
2. Degrees of freedom for the residuals (df_residuals):
   The degrees of freedom for the residuals indicate the number of independent pieces of information that are available for estimating the variability in the residuals (errors) after fitting the regression model.
   This is equal to the number of data points (n) minus the number of estimated parameters, including the intercept (k+1).

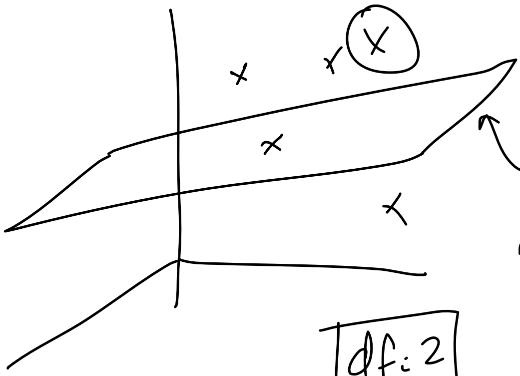
The sum of the degrees of freedom for the model and the degrees of freedom for the residuals is equal to the total degrees of freedom:
df_total = df_model + df_residuals

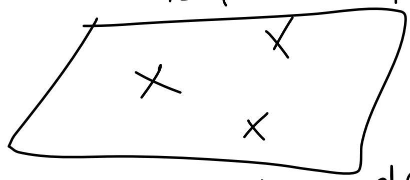




cgpa | iq | lpa

$n=4$ $\quad 4-1-1=2$





min → data points

reg plane

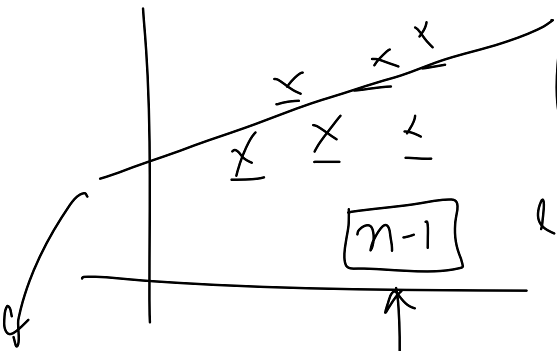
$n=4 \quad k=2$

$n-k-1$

$4-2-1=1$

df: 2

$n-1$



$t = n-1$

last nth point

$n-1$

$\beta_0 \quad \beta_1$

100 samp

$\bar{x}$

μ

F-statistic & Prob(F-statistic)

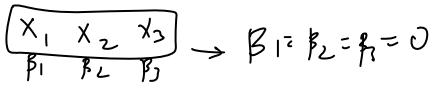
28 April 2023 07:01

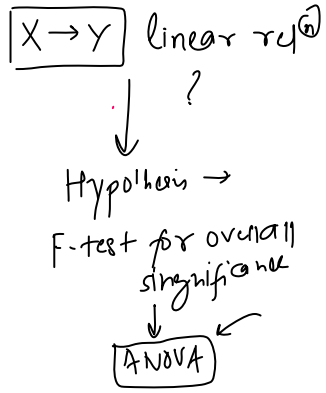
The F-test for overall significance is a statistical test used to determine whether a linear regression model is statistically significant, meaning it provides a better fit to the data than just using the mean of the dependent variable.

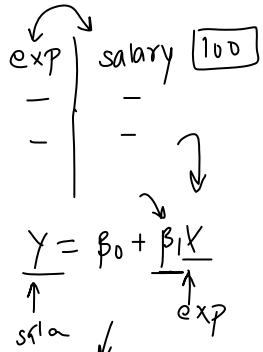
Here are the steps involved in conducting an F-test for overall significance:

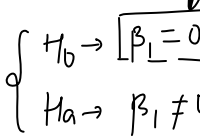
1. State the null and alternative hypotheses:
   - Null hypothesis (H0): All regression coefficients (except the intercept) are equal to zero (β1 = β2 = ... = βk = 0), meaning that none of the independent variables contribute significantly to the explanation of the dependent variable's variation.
   - Alternative hypothesis (H1): At least one regression coefficient is not equal to zero, indicating that at least one independent variable contributes significantly to the explanation of the dependent variable's variation.
2. Fit the linear regression model to the data, estimating the regression coefficients (intercept and slopes).
3. Calculate the Sum of Squares (SS) values:
   - Total Sum of Squares (TSS): The sum of squared differences between each observed value of the dependent variable and its mean.
   - Regression Sum of Squares (ESS): The sum of squared differences between the predicted values of the dependent variable and its mean.
   - Residual Sum of Squares (RSS): The sum of squared differences between the observed values and the predicted values of the dependent variable.
4. Compute the Mean Squares (MS) values:
   - Mean Square Regression (MSR): ESS divided by the degrees of freedom for the model (df_model), which is the number of independent variables (k). This could also be called as Average Explained Variance per independent feature.
   - Mean Square Error (MSE): RSS divided by the degrees of freedom for the residuals (df_residuals), which is the number of data points (n) minus the number of estimated parameters, including the intercept (k+1). This could also be called as average unexplained variance per degree of freedom.
5. Calculate the F-statistic: F-statistic = MSR / MSE
6. Determine the p-value:
   - Compute the p-value associated with the calculated F-statistic using the F-distribution or a statistical software package.
7. Compare the calculated F-statistic to the p-value to the chosen significance level (α):
   - If the p-value < α, reject the null hypothesis. This indicates that at least one independent variable contributes significantly to the prediction of the dependent variable, and the overall regression model is statistically significant.
   - If the p-value ≥ α, fail to reject the null hypothesis. This suggests that none of the independent variables in the model contribute significantly to the prediction of the dependent variable, and the overall regression model is not statistically significant.

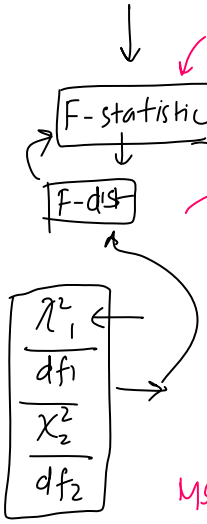
Following these steps, you can perform an F-test for overall significance in a linear regression analysis and determine whether the regression model is statistically significant.

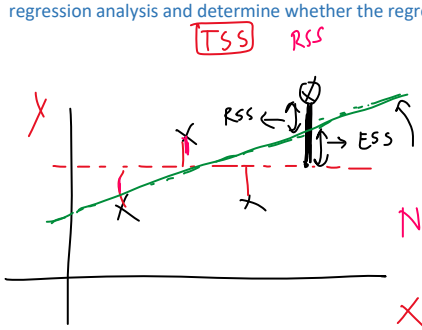


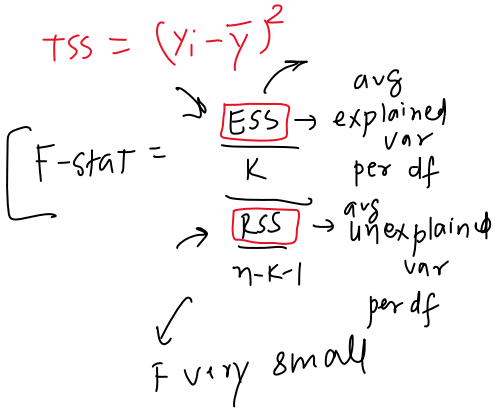


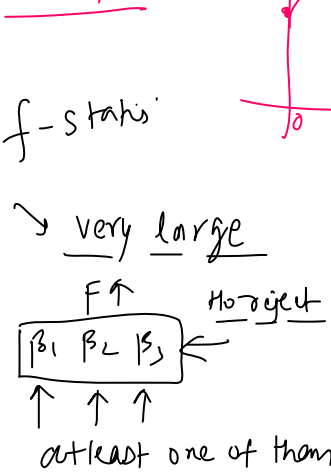


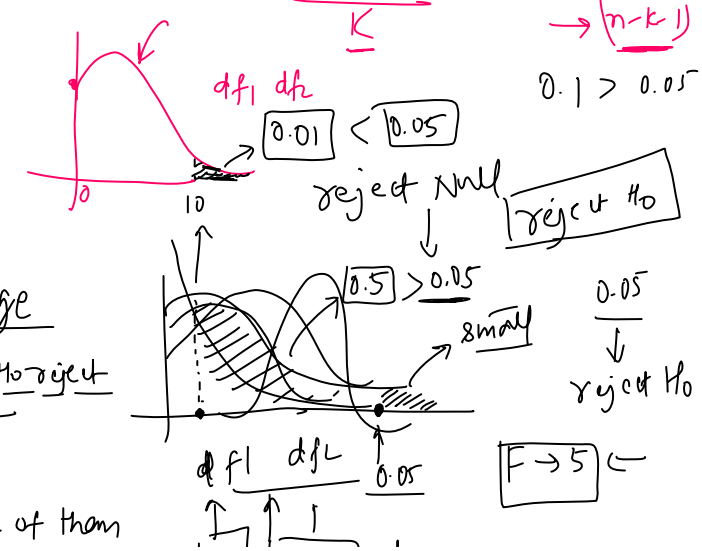


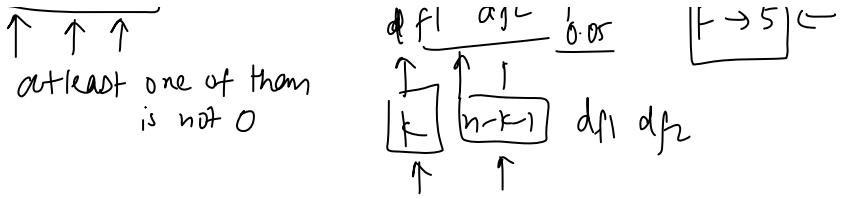








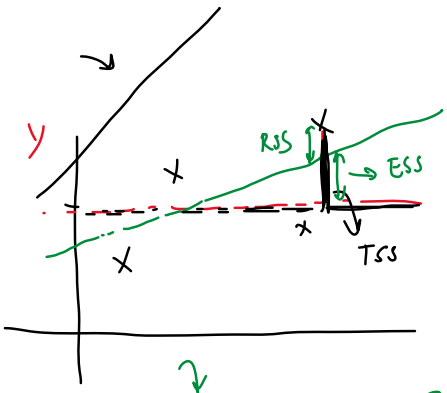



atleast one of them
is not 0
df1 and
0.05
F → 5
k
n-k-1
df1 df2

R-squared

28 April 2023 07:00

R^2 → goodness of fit



R-squared (R^2), also known as the coefficient of determination, is a measure used in regression analysis to assess the goodness-of-fit of a model. It quantifies the proportion of the variance in the dependent variable (response variable) that can be explained by the independent variables (predictor variables) in the regression model. R-squared is a value between 0 and 1, with higher values indicating a better fit of the model to the observed data.

In the context of a simple linear regression, R^2 is calculated as the square of the correlation coefficient (r) between the observed and predicted values. In multiple regression, R^2 is obtained from the ratio of the explained sum of squares (ESS) to the total sum of squares (TSS):

$R^2 = ESS / TSS$

where:

- ESS (Explained Sum of Squares) is the sum of squared differences between the predicted values and the mean of the observed values. It represents the variation in the response variable that can be explained by the predictor variables in the model.
- TSS (Total Sum of Squares) is the sum of squared differences between the observed values and the mean of the observed values. It represents the total variation in the response variable.

$R^2 = \frac{ESS}{TSS} = \frac{TSS - RSS}{TSS}$

[proportion (0-1)]

$R^2 = 1 - \frac{RSS}{TSS}$

ESS → TSS

models: 0.81 → 0.61

RSS > TSS

An R-squared value of 0 indicates that the model does not explain any of the variance in the response variable, while an R-squared value of 1 indicates that the model explains all of the variance. However, R-squared can be misleading in some cases, especially when the number of predictor variables is large or when the predictor variables are not relevant to the response variable.

Disadvantage of R^2

↑

Adjusted R2 score

$x_1 \; x_2 \; x_3 \; x_4 \; \ldots \; x_n \mid y$

input cols

$x_1 \rightarrow 0.4$

r2 score

$X_1 \mid y$

X_2

$x_2 \rightarrow y$

$x_2 \neq y$

x_3 x_4 x_5

r2_score → 0.4

r2_score → 0.6 (0.4)

(5) irr.

6 input → 0.4 (0.4)

$X_1 \rightarrow X_6$ misleading

cgpa | salary

iq

temp x

Adjusted R-squared

28 April 2023 07:01

Adjusted R-squared is a modified version of R-squared (R^2) that adjusts for the number of predictor variables in a multiple regression model. It provides a more accurate measure of the goodness-of-fit of a model by considering the model's complexity.

In a multiple regression model, R-squared (R^2) measures the proportion of variance in the response variable that is explained by the predictor variables. However, R-squared always increases or stays the same with the addition of new predictor variables, regardless of whether those variables contribute valuable information to the model. This can lead to overfitting, where a model becomes too complex and starts capturing noise in the data instead of the underlying relationships.

Adjusted R-squared accounts for the number of predictor variables in the model and the sample size, penalizing the model for adding unnecessary complexity. Adjusted R-squared can decrease when an irrelevant predictor variable is added to the model, making it a better metric for comparing models with different numbers of predictor variables.

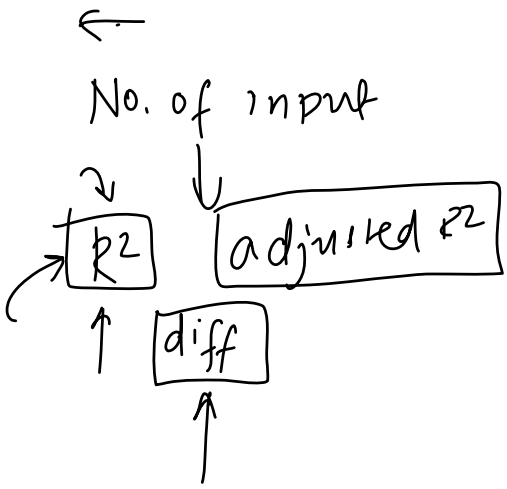
The formula for adjusted R-squared is:

$$\text{Adjusted } R^2 = 1 - \left[\frac{(1-R^2)*(n-1)}{(n-K-1)}\right]$$

where:

- R^2 is the R-squared of the model
- n is the number of observations in the dataset
- k is the number of predictor variables in the model

By using adjusted R-squared, you can more accurately assess the goodness-of-fit of a model and choose the optimal set of predictor variables for your analysis.



Which one should be used?

02 May 2023 13:43

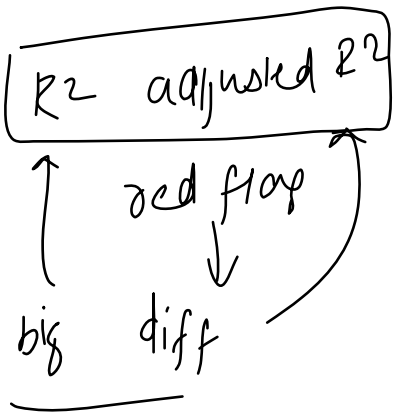
The choice between using R-squared and adjusted R-squared depends on the context and the goals of your analysis. Here are some guidelines to help you decide which one to use:

1. Model comparison: If you're comparing models with different numbers of predictor variables, it's better to use adjusted R-squared. This is because adjusted R-squared takes into account the complexity of the model, penalizing models that include irrelevant predictor variables. R-squared, on the other hand, can be misleading in this context, as it tends to increase with the addition of more predictor variables, even if they don't contribute valuable information to the model.

2. Model interpretation: If you're interested in understanding the proportion of variance in the response variable that can be explained by the predictor variables in the model, R-squared can be a useful metric. However, keep in mind that R-squared does not provide information about the significance or relevance of individual predictor variables. It's also important to remember that a high R-squared value does not necessarily imply causation or a good predictive model.

3. Model selection and overfitting: When building a model and selecting predictor variables, it's important to guard against overfitting. In this context, adjusted R-squared can be a helpful metric, as it accounts for the number of predictor variables and penalizes the model for unnecessary complexity. By using adjusted R-squared, you can avoid including irrelevant predictor variables that might lead to overfitting.

In summary, adjusted R-squared is generally more suitable when comparing models with different numbers of predictor variables or when you're concerned about overfitting. R-squared can be useful for understanding the overall explanatory power of the model, but it should be interpreted with caution, especially in cases with many predictor variables or potential multicollinearity.



T-statistic

28 April 2023 07:02

Performing a t-test for a simple linear regression, including the intercept term and using the p-value approach, involves the following steps:

1. State the null and alternative hypotheses for the slope and intercept coefficients:

   For the slope coefficient (β1):

   - Null hypothesis (H0): β1 = 0 (no relationship between the predictor variable (X) and the response variable (y))
   - Alternative hypothesis (H1): β1 ≠ 0 (a relationship exists between the predictor variable and the response variable)

   For the intercept coefficient (β0):

   - Null hypothesis (H0): β0 = 0 (the regression line passes through the origin)
   - Alternative hypothesis (H1): β0 ≠ 0 (the regression line does not pass through the origin)

2. Estimate the slope and intercept coefficients (b0 and b1): Using the sample data, calculate the slope (b1) and intercept (b0) coefficients for the regression model.

3. Calculate the standard errors for the slope and intercept coefficients (SE(b0) and SE(b1)): Compute the standard errors of the slope and intercept coefficients using the following formulas:

$$SE(b_1) = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{(n-2)\sum (x_i - \bar{x})^2}}$$

$$SE(b_0) = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{(n-2)}\left[\frac{1}{n} + \frac{\bar{x}^2}{\sum (x_i - \bar{x})^2}\right]}$$

4. Compute the t-statistics for the slope and intercept coefficients: Calculate the t-statistics for the slope and intercept coefficients using the following formulas:

$$t\text{-value}_{b_0} = \frac{b_0 - 0}{SE(b_0)}$$

$$t\text{-value}_{b_1} = \frac{b_1 - 0}{SE(b_1)}$$

5. Calculate the p-values for the slope and intercept coefficients: Using the t-statistics and the degrees of freedom, look up the corresponding p-values from the t-distribution table or use a statistical calculator.

6. Compare the p-values to the chosen significance level (α): A common choice for α is 0.05, which corresponds to a 95% confidence level. Compare the calculated p-values to α:

   - If the p-value is less than or equal to α, reject the null hypothesis.
   - If the p-value is greater than α, fail to reject the null hypothesis.

Confidence Intervals for Coefficients

29 April 2023 02:35

1. Estimate the slope and intercept coefficients (b0 and b1): Using the sample data, calculate the slope (b1) and intercept (b0) coefficients for the regression model.
2. Calculate the standard errors for the slope and intercept coefficients (SE(b0) and SE(b1)):

$$SE(b_1) = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{(n-2)\sum (x_i - \bar{x})^2}} \leftarrow$$

$$SE(b_0) = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{(n-2)}\left[\frac{1}{n} + \frac{\bar{x}^2}{\sum (x_i - \bar{x})^2}\right]}$$

3. Determine the degrees of freedom: In a simple linear regression, the degrees of freedom (df) is equal to the number of observations (n) minus the number of estimated parameters (2: the intercept and the slope coefficient).
   df = n - 2
4. Find the critical t-value: Look up the critical t-value from the t-distribution table or use a statistical calculator based on the chosen confidence level (e.g., 95%) and the degrees of freedom calculated in step 3.

t-dist

significance

↓

0.05

95% prob

5. Calculate the confidence intervals for the slope and intercept coefficients: Compute the confidence intervals for the slope (b1) and intercept (b0) coefficients using the following formulas:

$$CI_{b_0} = b_0 \pm t\_value * SE(b_0)$$

$$CI_{b_1} = b_1 \pm t\_value * SE(b_1)$$

$$b_1 \pm 3.18 \times SE(b_1)$$

lower + upper

b_1

These confidence intervals represent the range within which the true population regression coefficients are likely to fall with a specified level of confidence (e.g., 95%)

Others

28 April 2023 07:02

```
                            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.
==============================================================================
```