Degree of Freedom

In linear regression, the total degrees of freedom (df-total) represent the total number of data point minus 1. It represent the overall variability in the dataset that can be attribute 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:

$df\_total = n - 1$ ← no. of rows

where n is the number of data points (observation) 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). ← no. of input cols

2. **Degree of freedom for the residuals (df-residuals):**
   The degrees of the freedom for the residuals indicates the ~~discount~~ number of independent pieces of information that are avaible 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 intercepts (k+1).

The sum of the degree of freedom for the model and the degree of freedom for the residuals is equal to the total degrees of freedom.

$$df\text{-}total = df\text{-}model + df\text{-}residual$$

$df\text{-}residual = n-(k+1)$

X_1 X_2 | Y

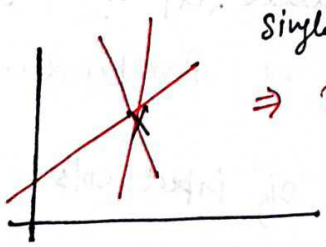
$k=2$ ↳ no. of input cols.

$n-k-1+k = (n-1)$ ↳ df-total

* Residuals

X | Y

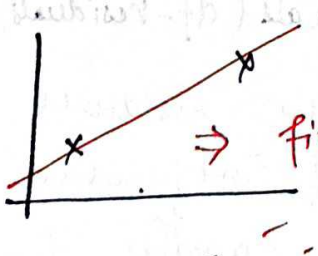
#1



single point ⇒ multiple

min dot points for reg line.

#2

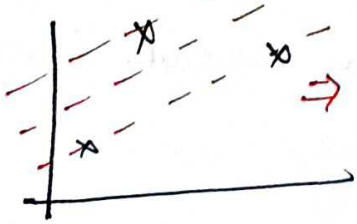


multiple point ⇒ fixed line

↳ no. of point

$n=3$ df-residual = 1

#3



⇒ some freedom

if $n=4$ df-residual = 2

if $n=5$ df-residual = 3

df_residual = n - k - 1 (k = 1 input col)

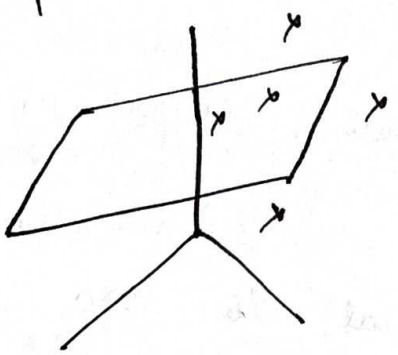
df-residual = 3 - 1 - 1 = 1

df_residual = 4 - 1 - 1 = 2

df_residual = 5 - 1 - 1 = 3

* when point is 3 then degree of freedom 1 because at point 3 have first time to move line.

example:- x_1 x_2 | Y



k = 2

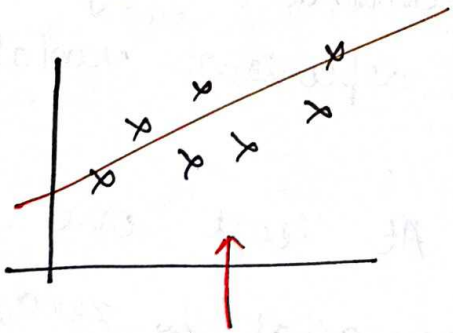
n = 4

n - k - 1

4 - 2 - 1 = 1

* if know parameter than we can easily find last term.

know n-1 !



B_0 B_1 ↳ Given

if n-1th term knows then with the help of B_0, B_1 and n-1 we can easily find last term. That's why we consider n-1 term.

F-statistic & Prob (F-statistics)

The F-test for overall significance is a statistical test used to determine whether a linear regression model is statically 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 hypothesis:

- Null hypothesis (H0): All regression coefficient (except the intercept) are equal to zero ($\beta_1 = \beta_2 \dots = \beta_k = 0$) mean a that none of the independent variables contributes significantly to the explanation of the dependent variables' 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 Square (SS) value:

- Total Sum of square (TSS): The sum of squared of different between each observed value of the dependent variable and its mean.

- Regression Sum of Square (ESS): The sum of squared difference btn the predicted values of the dependent variable and its mean.

- Residual Sum of Squares (RSS): The sum of squared difference btn the observed values and the predicted value of the dependent variable.

4. Compute the Mean Squares (MS) values:

- Mean square Regression (MSR): ESS divide by the degree 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 degree of freedom for the residuals (df-residual) which is the number of data point (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 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 $\geq \alpha$, fail to reject the null hypothesis. This suggest 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.

experience	salary
–	–
–	–

prove that linear relationship between experience and salary using Hypothesis (f-stastic)

$$\underbrace{y}_{\text{sal}} = \beta_0 + \beta_1 \underbrace{x}_{\text{exp}}$$

$H_0 \rightarrow \beta_1 \Rightarrow 0$ ⇒ * if $\beta_1 = 0$ then y doesn't depend on X.
↳ Null hypothesis

$H_0 \rightarrow \beta_1 \neq 0$ ⇒ Some relationship positive or negative
↳ Alternate Hypothesis

↓

F-stastistic

↓

F-distribution

χ_1^2 → chi-square distribution

$\frac{\chi_1^2}{df_1}$ → chi-square degree of freedom

χ_2^2 → chi-square distribution

$\frac{\chi_2^2}{df_2}$ → chi-square degree of freedom

↓ Answer

F-distribution follow kaise hua

Any Number

* Chi-square distribution is square of Normal distribution.

F-Statistic = MSR / MSE

MSR = ESS / k ← No. of independent col.

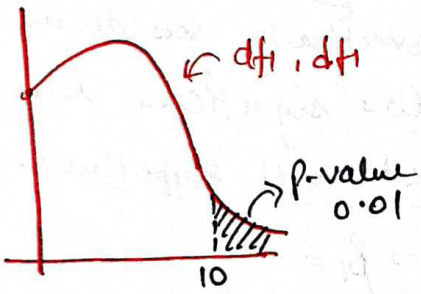
MSE = RSS / (n-k-1) ← no. of rows

MSR = TSS − ESS / k ← degree of freedom

MSE = Σ(Yi − Ŷ) / (n-k-1)

Chisquare ⇐ MSE = Σ(Yi − Ŷ)² / (n-k-1) ← degree of freedom

let.



* Let assume F-stats is 10. and find P-value.

* Let P-value is 0.01 but Normal significance level is 0.05.

0.01 < 0.05

Since significance value is less than 0.05. Reject null hypothesis.

β_1 is not zero.

F-Stat = (ESS / k → avg explained variance per df) / (RSS / (n-k-1) → avg unexplained variance per df)

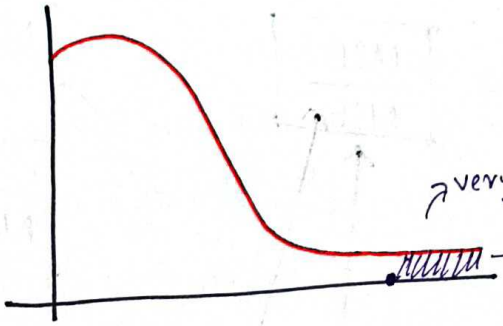
⇒ * If it is very large

$\frac{\chi^2}{df_1}$

$\frac{\chi^2}{df_2}$ → (N^2) → (Normal distribution)² → Chi-square

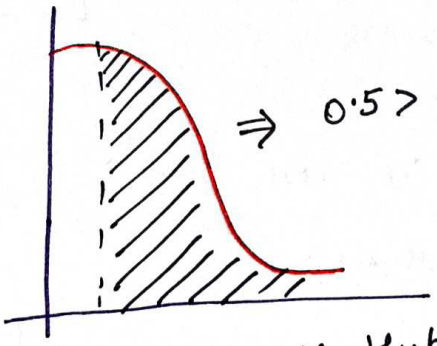
→ $\Sigma(Y_i - \hat{Y})^2$

* If avg explained variance per df is very large then f-stat increase.



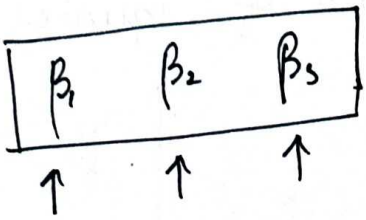
* reject null Hypothesis.
  ↳ $\beta_1 = 0$

* If f-stats is very small. Means unexplained variance is very large in comparison to explained variable.



* You can not reject null Hypothesis and proove input and output data has no relationship. And non-linear data.

Multiple Input



[atleast one of them is not 0.] ⇒ Linear relationship exist.

R-Squared

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 (prediction 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):

$$R^2 = ESS / TSS$$

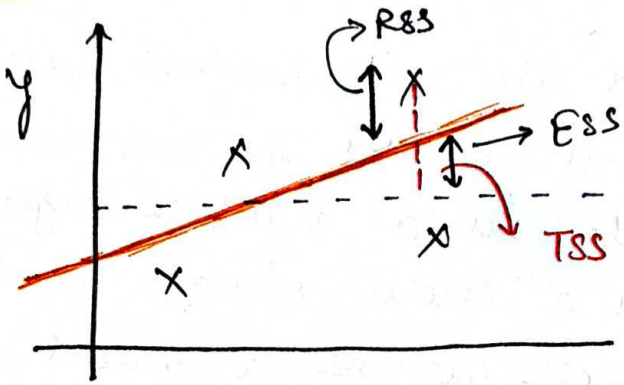
where:

- ESS (Explained sum of squares) is the sum of squared difference between the predicted values and the mean of the observation values. It represents the variation in the response variable that can be explained by the predictor variable in the model.
- TSS (Total sum of squares) is the sum of squared differences betn the observed values and the mean of the observed values. It represent the variation in the respons variable.

Disadvantage

An R-squared value of 0 indicates that the model doesn't not explain any of the variance in the response variable, while an R value of 1 indicate that the model explain all of the variance. However, R squared can be misleading in some case, especially when the number of predictor variable is large or when the

Predictor variables are not relevant to the response variable -



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

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

[Proportion] (0-1)

R^2 means how much percent you explain in Total Variance. eg :-

Module 1 = 0.81
Module 2 = 0.61

} so Module 1 is best because 0.81 (81%) explain in Total variance