Course Name :Basic Statistics using GUI-R (RKWard)

Module : Example of Anova

Week 5 Lecture : 1

Harsh Pradhan, Assistant Professor,
Institute of Management Studies, BHU
https://bhu.ac.in/Site/FacultyProfile/1_5?FA000562

Model (between-group) sum of squares (SS_M)

1. Calculate the difference between the mean of each group and the grand mean.

   The grand mean is the mean of all scores

2. Square each of these differences

3. Multiply each result by the number of participants within that group – this is a correction (or “weighting”): a smaller sample will have less “weight” in the equation, a larger sample will have more “weight”.

4. Add the values for each group together.

$$SS_M = \sum n_i (M_i - M_{grand})^2$$

One-Way ANOVA

- After filling in the sum of squares, we have …

Source	SS	df	MS	F	p
Between	461.64				
Within	167.42				
Total	629.08				

Degrees of freedom

- The between group df is one less than the number of groups, k - 1
  - We have four groups, so $df_M = 3$
- The within group df is the sum of the individual df's of each group, which equals N - k
  - The sample sizes are 4, 3, 3, and 3
  - $df_R = 3 + 2 + 2 + 2 = 13 – 4 = 9$
- The total df is one less than the sample size, N - 1
  - df(Total) = 13 – 1 = 12   = (n+n+n) -1

One-Way ANOVA

- Filling in the degrees of freedom gives this …

Source	SS	df	MS	F	p
Between	461.64	3			
Within	167.42	9			
Total	629.08	12			

Calculating the Mean Squares

Divide the SS by the corresponding df

- MS_M = 461.64 / 3 = 153.88
- MS_R = 167.42 / 9 = 18.60

One-Way ANOVA

- Completing the MS gives …

Source	SS	df	MS	F	p
Between	461.64	3	153.88		
Within	167.42	9	18.60		
Total	629.08	12			

The F ratio

- F test statistic
  - – An F test statistic is the ratio of MS_M and MS_R
  - – $F = MS_M / MS_R$
- For our data, F = 153.88 / 18.60 = 8.27

- A larger F ratio means a larger difference between the group means relative to the variation within the group.

An example: Fairness in different types of societies

Fairness score: proportion of money shared in a game

	Hunter-gatherer	**Farming**	**Natural resources**	**Industrial**
P1	28	32	47	40
P2	36	33	43	47
P3	38	40	52	45
P4	31			
Mean	**33.25**	**35.0**	**47.33**	**44.0**
N	*4*	*3*	*3*	*3*

Grand Mean = 39.385 (The sum of all scores divided by the total N

Hypothesis testing

- The F test statistic has an F distribution with df_M numerator df and df_R denominator df
- F(3, 9) = 8.27, p < .001

http://www.distributome.org/V3/calc/index.html

Course Name :Basic Statistics using GUI-R (RKWard)

Module : Type of anova

Week 5 Lecture : 2

Harsh Pradhan, Assistant Professor,
Institute of Management Studies, BHU
https://bhu.ac.in/Site/FacultyProfile/1_5?FA000562

Aspect	Repeated Measures ANOVA	Between-Subjects ANOVA
Experimental Design	Subjects are tested under multiple conditions.	Different groups of subjects are exposed to different conditions. Each subject participates in only one condition.
Dependency	Assumes dependency between measures on the same subject.	Assumes independence between subjects in different groups.
Statistical Power	May have lower power due to within-subject variability.	Generally has higher power when the number of subjects is large.
Control of Individual Differences	Each subject serves as their own control.	Individual differences between subjects can introduce noise into the data.
Efficiency	Often requires fewer subjects for equivalent power.	Can be less efficient in terms of sample size requirements.
Example Research Question	Does a new teaching method result in improved test scores?	Do different teaching methods result in different test scores?

Different Types of ANOVA

- Repeated Measures is build upon paired Sample t-test
- Between Subjects Measures is build upon independent sample

- **SSW (Sum of Squares Within)**: This represents the variability of scores within each subject across different conditions.

$$SSW = SSD_{A-B} + SSD_{A-C} + SSD_{B-C}$$

$$SSW = 400 + 800 + 400$$
$$SSW = 1600$$

- **dfW (Degrees of Freedom Within)**: This is calculated as the total number of observations minus the total number of subjects.

$$dfW = 3 \times (N - 1)$$
$$dfW = 3 \times (30 - 1)$$
$$dfW = 3 \times 29$$
$$dfW = 87$$

- **MSW (Mean Squares Within)**: This is the sum of squares within divided by the degrees of freedom within.

$$MSW = \frac{SSW}{dfW}$$
$$MSW = \frac{1600}{87}$$
$$MSW \approx 18.39$$

- **Mean A**:
  - Mean A = (75 + 70 + ... + 80) / 30
  - Mean A ≈ (2250 / 30)
  - Mean A ≈ 75
- **Mean B**:
  - Mean B = (80 + 75 + ... + 85) / 30
  - Mean B ≈ (2280 / 30)
  - Mean B ≈ 76
- **Mean C**:
  - Mean C = (85 + 80 + ... + 90) / 30
  - Mean C ≈ (2370 / 30)
  - Mean C ≈ 79
- **Grand Mean**:
  - Grand Mean = (75 + 76 + 79) / 3
  - Grand Mean ≈ 230 / 3
  - Grand Mean ≈ 76.67

- **Grand Mean**:
  - Grand Mean = (75 + 76 + 79) / 3
  - Grand Mean ≈ 230 / 3
  - Grand Mean ≈ 76.67
- **SSB (Sum of Squares Between)**:
  - $SSB = (10 \times (75 - 76.67)^2) + (10 \times (76 - 76.67)^2) + (10 \times (79 - 76.67)^2)$
  - $SSB = (10 \times (-1.67)^2) + (10 \times (-0.67)^2) + (10 \times (2.33)^2)$
  - $SSB = (10 \times 2.7889) + (10 \times 0.4489) + (10 \times 5.4289)$
  - $SSB = 27.889 + 4.489 + 54.289$
  - $SSB \approx 86.667$
- **dfB (Degrees of Freedom Between)**:
  - $dfB = 3 - 1$
  - $dfB = 2$
- **MSB (Mean Squares Between)**:
  - $MSB = \frac{SSB}{dfB}$
  - $MSB = \frac{86.667}{2}$
  - $MSB \approx 43.33$

- **Mean A**:
  - Mean A = (75 + 70 + ... + 80) / 30
  - Mean A ≈ (2250 / 30)
  - Mean A ≈ 75
- **Mean B**:
  - Mean B = (80 + 75 + ... + 85) / 30
  - Mean B ≈ (2280 / 30)
  - Mean B ≈ 76
- **Mean C**:
  - Mean C = (85 + 80 + ... + 90) / 30
  - Mean C ≈ (2370 / 30)
  - Mean C ≈ 79
- **Grand Mean**:
  - Grand Mean = (75 + 76 + 79) / 3
  - Grand Mean ≈ 230 / 3
  - Grand Mean ≈ 76.67

- **Grand Mean**:
  - Grand Mean = (75 + 76 + 79) / 3
  - Grand Mean ≈ 230 / 3
  - Grand Mean ≈ 76.67
- **SSB (Sum of Squares Between)**:
  - $SSB = (10 \times (75 - 76.67)^2) + (10 \times (76 - 76.67)^2) + (10 \times (79 - 76.67)^2)$
  - $SSB = (10 \times (-1.67)^2) + (10 \times (-0.67)^2) + (10 \times (2.33)^2)$
  - $SSB = (10 \times 2.7889) + (10 \times 0.4489) + (10 \times 5.4289)$
  - $SSB = 27.889 + 4.489 + 54.289$
  - $SSB \approx 86.667$
- **dfB (Degrees of Freedom Between)**:
  - $dfB = 3 - 1$
  - $dfB = 2$
- **MSB (Mean Squares Between)**:
  - $MSB = \frac{SSB}{dfB}$
  - $MSB = \frac{86.667}{2}$
  - $MSB \approx 43.33$

विद्ययाऽमृतमश्नुते
काशी हिन्दू विश्वविद्यालय
BANARAS HINDU UNIVERSITY

Source	SS	df	MS	F
Between-Subjects	86.667	2	43.3335	2.3552
Within-Subjects	1600	87	18.3908	
Total	1686.667	89		

```
>> qf(.95,2,87)
 >>3.101
```

Ftable link

- `summary(aov(Y~X))`

Aspect	MANOVA	N-way ANOVA
Use	Multiple dependent variables, one or more independent variables	Multiple independent variables affecting a single dependent variable
Objective	Determine differences between groups in a multivariate response, controlling for other variables	Examine interaction effects between multiple independent variables and main effects of each
Assumption	Multivariate normality, equal covariance matrices	Normal distribution, homogeneous variances
Example	Investigating teaching methods' effect on exam scores, class participation, and homework completion	Analyzing temperature and humidity effects on plant growth

Data nanova.csv

```
# Fit the ANOVA model
model <- aov(Yield ~ Treatment * Dose, data = data)

# View the ANOVA summary
summary(model)
```

Interactions

```
HH:: interaction2wt(Y~Z*X)
```

```
HH::interaction2wt(data$Yield~data$Treatment*data$Dose)
```

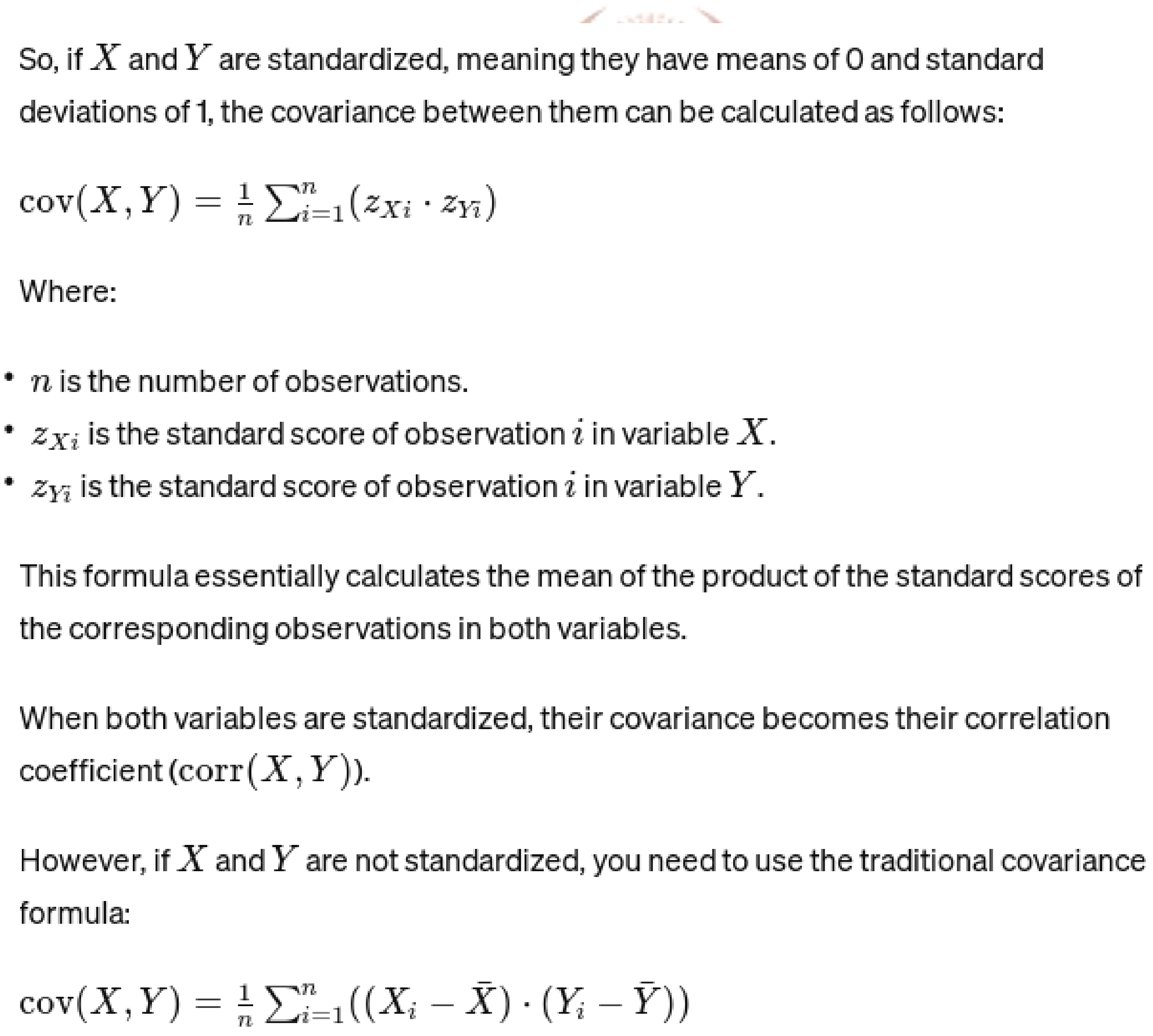
ANCOVA

Course Name :Basic Statistics using GUI-R (RKWard)

Module : Introduction to Correlation

Week 5 Lecture : 3

Harsh Pradhan, Assistant Professor,
Institute of Management Studies, BHU
https://bhu.ac.in/Site/FacultyProfile/1_5?FA000562



So, if X and Y are standardized, meaning they have means of 0 and standard deviations of 1, the covariance between them can be calculated as follows:

$$\operatorname{cov}(X, Y) = \frac{1}{n} \sum_{i=1}^{n} (z_{Xi} \cdot z_{Yi})$$

Where:

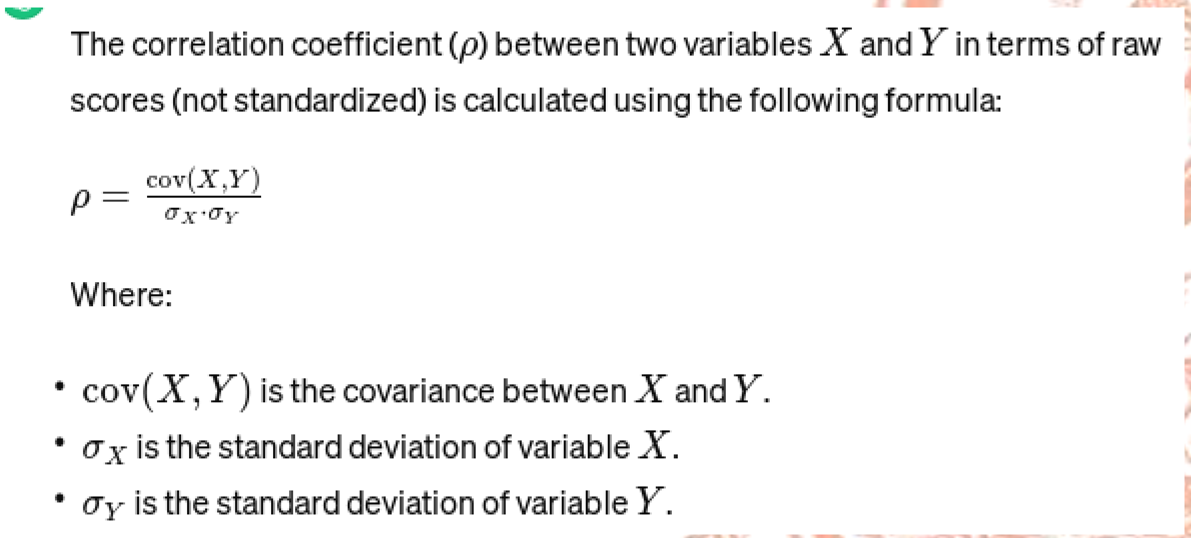
- n is the number of observations.
- z_{Xi} is the standard score of observation i in variable X.
- z_{Yi} is the standard score of observation i in variable Y.

This formula essentially calculates the mean of the product of the standard scores of the corresponding observations in both variables.

When both variables are standardized, their covariance becomes their correlation coefficient ($\operatorname{corr}(X, Y)$).

However, if X and Y are not standardized, you need to use the traditional covariance formula:

$$\operatorname{cov}(X, Y) = \frac{1}{n} \sum_{i=1}^{n} ((X_i - \bar{X}) \cdot (Y_i - \bar{Y}))$$

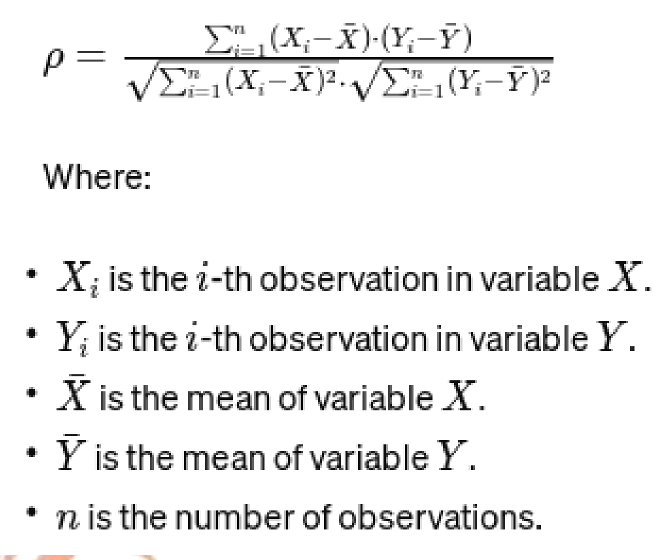


The correlation coefficient (ρ) between two variables X and Y in terms of raw scores (not standardized) is calculated using the following formula:

$$\rho = \frac{\mathrm{cov}(X,Y)}{\sigma_X \cdot \sigma_Y}$$

Where:

- $\mathrm{cov}(X,Y)$ is the covariance between X and Y.
- σ_X is the standard deviation of variable X.
- σ_Y is the standard deviation of variable Y.



$$\rho = \frac{\sum_{i=1}^{n}(X_i - \bar{X}) \cdot (Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2} \cdot \sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

Where:

- X_i is the i-th observation in variable X.
- Y_i is the i-th observation in variable Y.
- $\bar{X}$ is the mean of variable X.
- $\bar{Y}$ is the mean of variable Y.
- n is the number of observations.

Correlation

Data type	Type of Correl
Nominal	Phi
Dichotomous	Bi-serial
Ordinal/Rank	Spearmen/kendall
Ratio/Interval	Pearson

Partial Correl

Statistics⍰ Summaries⍰ Correlation matrix

```
stats::cor.test(my.csv.data$JP_01,my.csv.data$JP_02,
alternative="two.sided",conf.level=0.95)
```

```
ggm::pcor(my.csv.data$JP_01,my.csv.data$JP_02,my.c
sv.data$JP_03)
```

Teaching→ Regression→ Correlation

Course Name :Basic Statistics using GUI-R (RKWard)
Module : Correlation Continued and Introduction to Regression

Week 5 Lecture : 4

Harsh Pradhan, Assistant Professor,
Institute of Management Studies, BHU
https://bhu.ac.in/Site/FacultyProfile/1_5?FA000562

Correlation

Data type	Type of Correl
Nominal	Phi
Dichotomous	Bi-serial
Ordinal/Rank	Spearmen/kendall
Ratio/Interval	Pearson

Partial Correl

Statistics□ Summaries□ Correlation matrix

```
stats::cor.test(my.csv.data$JP_01,my.csv.data$JP_02,
alternative="two.sided",conf.level=0.95)
```

```
ggm::pcor(my.csv.data$JP_01,my.csv.data$JP_02,my.c
sv.data$JP_03)
```

Teaching→ Regression→ Correlation

	Pearson	Spearman	Kendall
Description	Measures linear relationship between two continuous variables	Measures monotonic relationship between two variables	Measures monotonic relationship between two variables
Assumptions	Normal distribution, linearity	No linearity, suitable for ordinal or ranked data	No linearity, suitable for ordinal or ranked data
Suitable for	Continuous variables	Ordinal or ranked data	Ordinal or ranked data
Sensitivity to Outliers	Sensitive	Less sensitive	Robust for small size
R Command	`cor(x, y, method = "pearson")`	`cor(x, y, method = "spearman")`	`cor(x, y, method = "kendall")`

Use of Correlation

Grouping Similar Variables
Reliability

REGRESSION

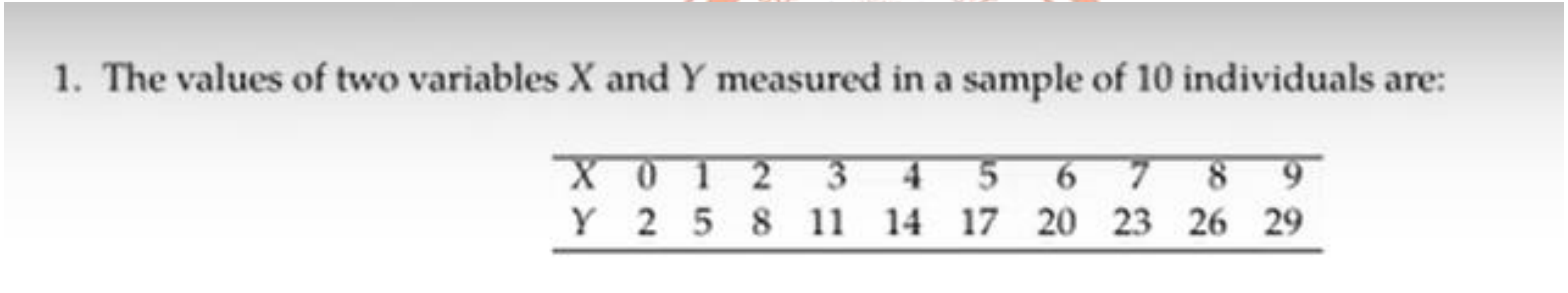
Linear regression

- Linear regression analysis is used to predict the value of a variable based on the value of another variable. The variable you want to predict is called the dependent variable. The variable you are using to predict the other variable's value is called the independent variable.

Non linear regression

- Nonlinear regression is a statistical technique that helps describe nonlinear relationships in experimental data. Nonlinear regression models are generally assumed to be parametric, where the model is described as a nonlinear equation. Typically machine learning methods are used for non-parametric nonlinear regression.

UNDERSTANDING USING AN EXAMPLE



1. The values of two variables X and Y measured in a sample of 10 individuals are:

X	0	1	2	3	4	5	6	7	8	9
Y	2	5	8	11	14	17	20	23	26	29

Do the following operations

1.Create a dataset with variables X and Y and enter the data

Course Name :Basic Statistics using GUI-R (RKWard)

Module : Correlation Continued and Introduction to Regression

Week 5 Lecture : 5

Harsh Pradhan, Assistant Professor,
Institute of Management Studies, BHU
https://bhu.ac.in/Site/FacultyProfile/1_5?FA000562

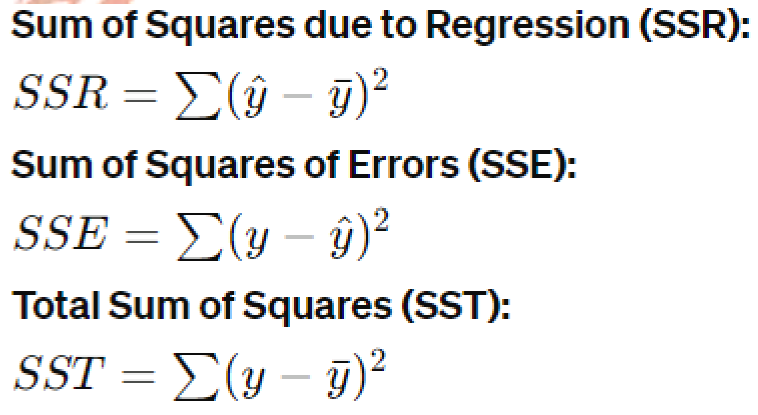
$b = SSxy / SSxx$
$a = \bar{y} - b * \bar{x} = 2.96$

$SST = \sum(y - \bar{y})2$

$SST = SSyy$

$R^2 = 1 - (SSE / SST)$

DF_regression = 1 (since there is one independent variable)
DF_total = n - 1



Sum of Squares due to Regression (SSR):

$$SSR = \sum(\hat{y} - \bar{y})^2$$

Sum of Squares of Errors (SSE):

$$SSE = \sum(y - \hat{y})^2$$

Total Sum of Squares (SST):

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