• RASIO

Analyzing Causal Relationships and Structural Effects of Cigarette Consumption on Income: Statistical and Machine Learning Approach

Team Dodoco

006RS24

Analysis Flow

01 Hypothesis & Objective

02 Data Exploration

03 Methodology - Statistical Approach

04 Methodology - Machine Learning Approach

05 Conclusions

Hypothesis

Factor Affecting **Income**

1. **Cigarette consumption** negatively effects income.
2. **Education** positively effects income.
3. **Age** positively effects income.
4. The effect of **age** on income decreases as age increases.

Factor Affecting **Cigarette Consumption**

1. **Income** positively effects cigarette consumption.
2. **Education** negatively effects cigarette consumption.
3. **Age** positively effects cigarette consumption.
4. **Cigarette price** negatively effects cigarette consumption.
5. **Smoking restriction** regulation negatively effects cigarette consumption.
6. The effect of **age** on cigarette consumption decreases as age increases.

Objective

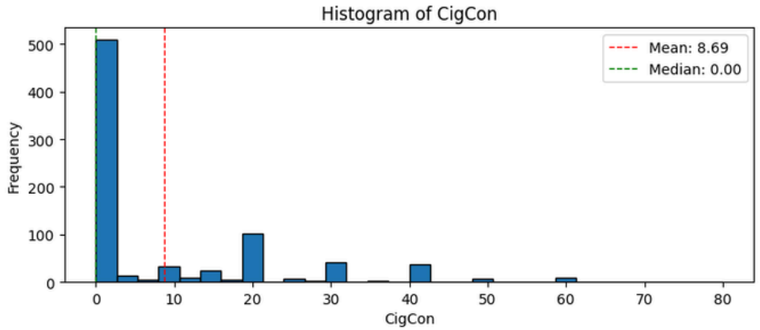
- To **estimate the effects of various factors** on income and cigarette consumption.
- To **evaluate** the related hypotheses.
- To **provide interpretations** of these effects to better understand the interrelationships.

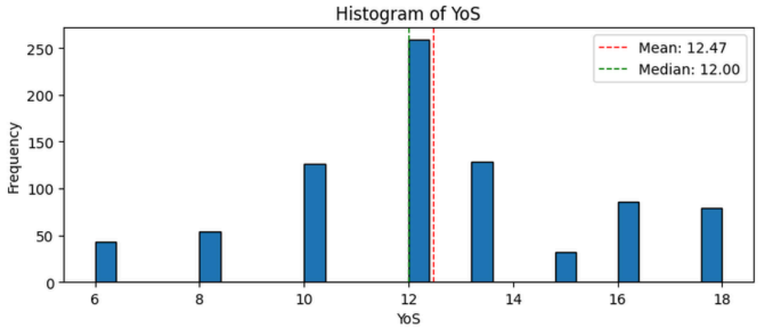
Data Descriptions

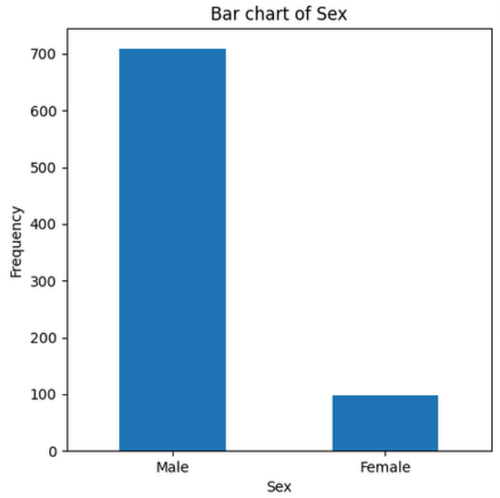
Variable		Units
CigCon	Cigarette Consumption	Package per month
YoS	Education Level	Year
Age	Age	Year
Linc	Income	Ln(Million Rupiah)
Sex	Sex	Male/Female
LCigPri	Cigarette Price	Ln(Thousand Rupiah)
CigRes	Presence of smoking restriction	Yes/No

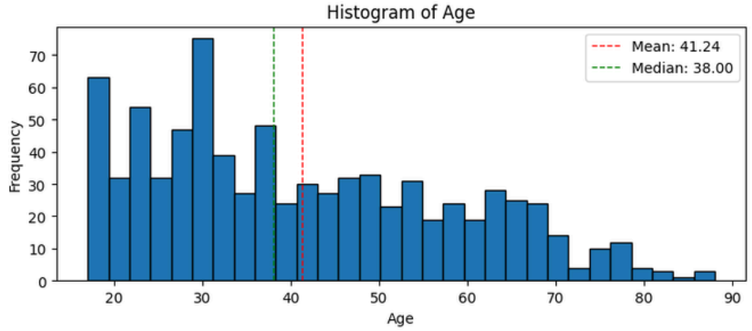
Exploratory Data Analysis

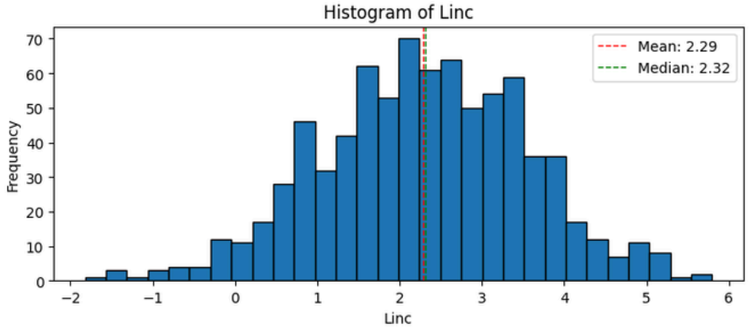
✓ No missing value and no duplicate value.

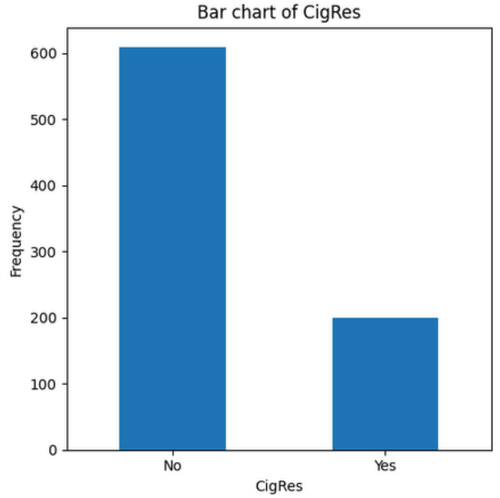


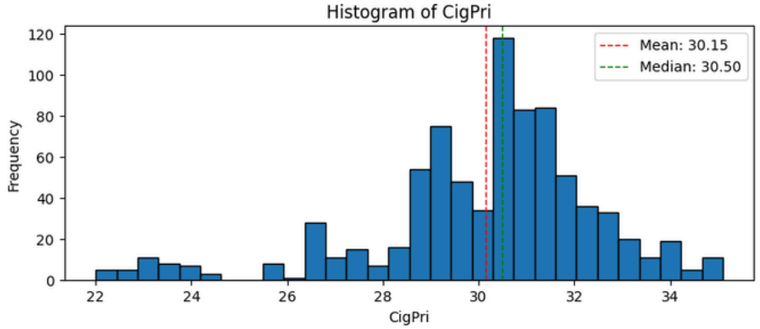


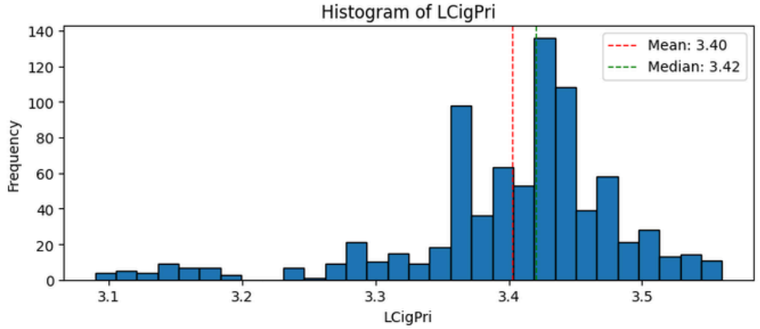




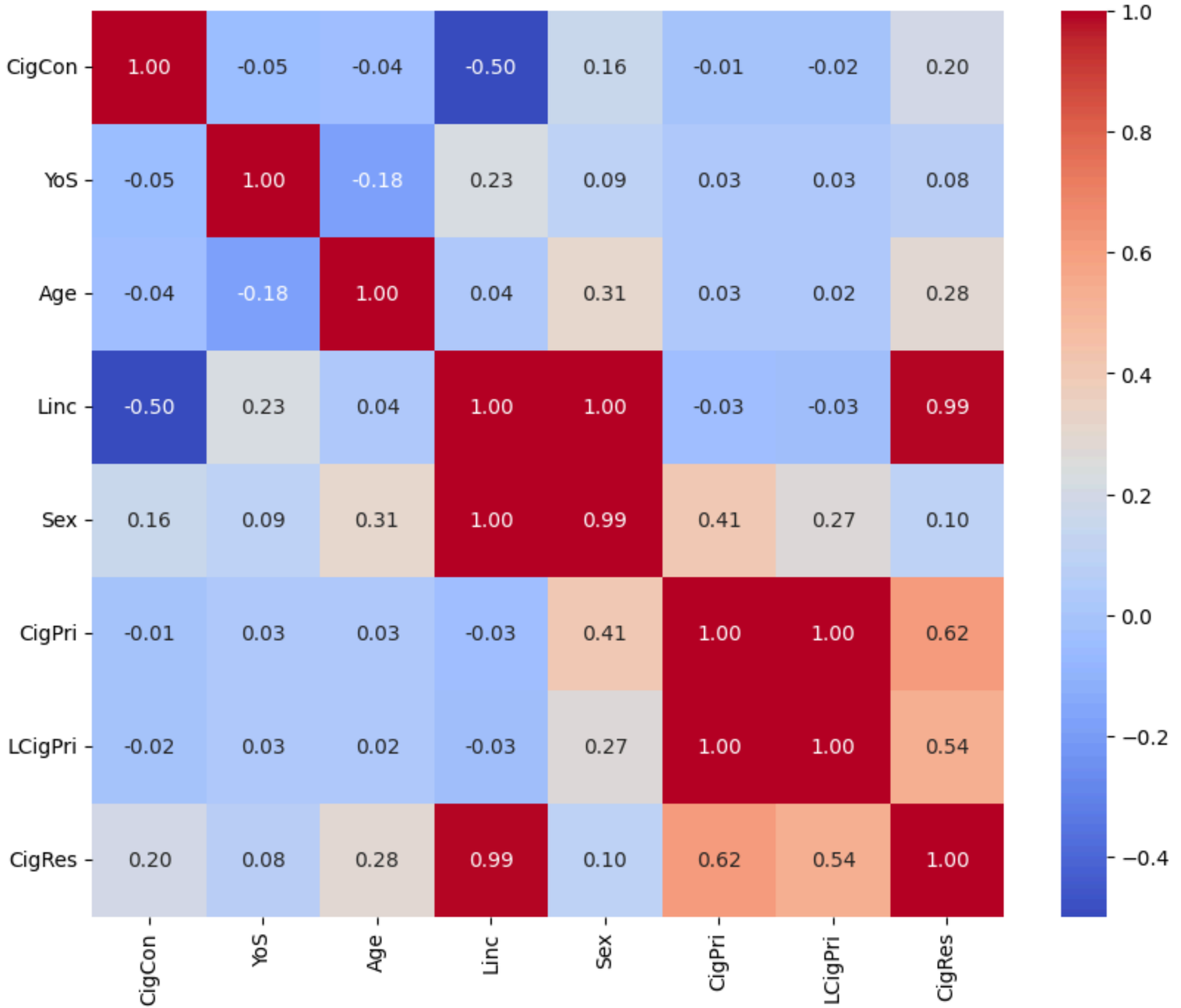




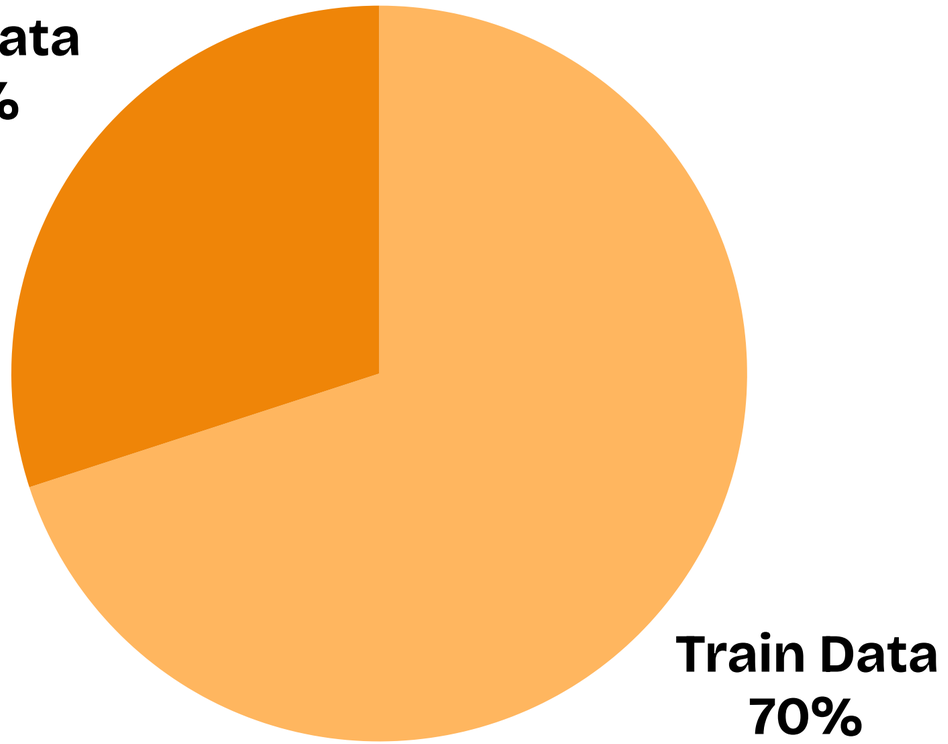




How the variables correlate each other?



Splitting Data



Structural Equation Modeling

Statistical Approach

Analysis using the library

`lavaan`

in RStudio

Assumptions Check

- **Multivariate Normality (Mardia Test)**

0.00	0.00
Skewness p-value	Kurtosis p-value

Assumption Violated!

- **No Systematic Missing Data**
  Assumption satisfied.
- **Large Sample**
  Satisfied, since the sample size is 807.
- **No Multicollinearity**
  Satisfied, VIF of all variables around 1.

Normality
of RMSE

1/5

Normally Distributed

Based on Shapiro-Wilk Test

Homogeneity of Variance

Bartlett Test

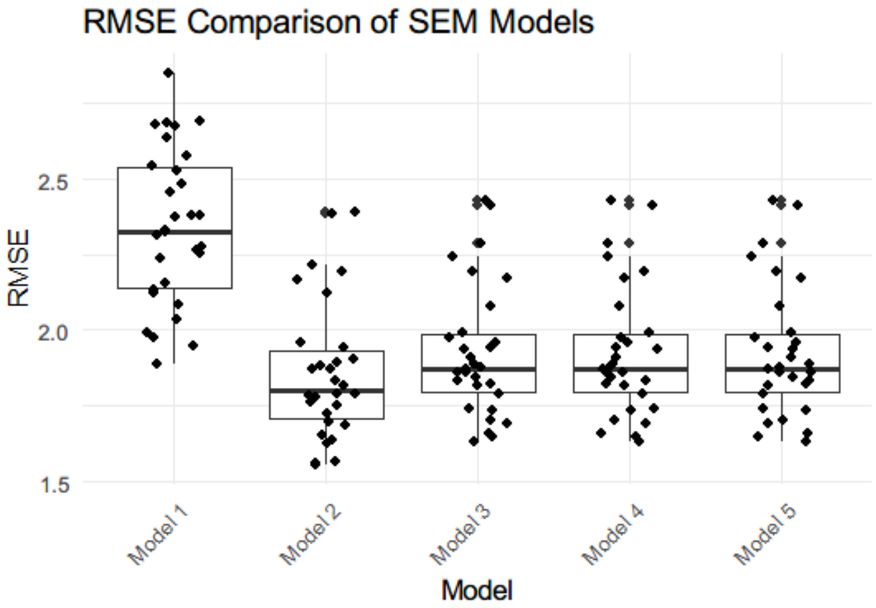
0.86

p-Value

Best Model Determination

Evaluation Metrics using **RMSE** on **k-Fold Cross Validation**

30 Folds

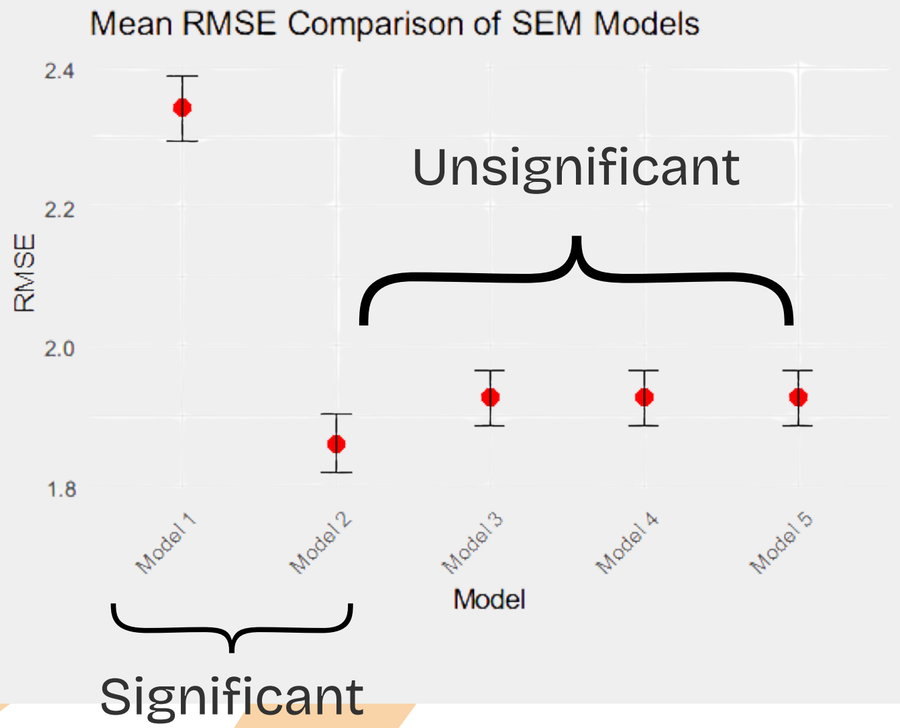


Kruskal-Wallis Test on RMSE

47.6
X2-Statistic

0.00
p-Value

There’s any significant difference between the models



Best Model: Model 5 with accounting for Parsimony

Modelling

The fifth model

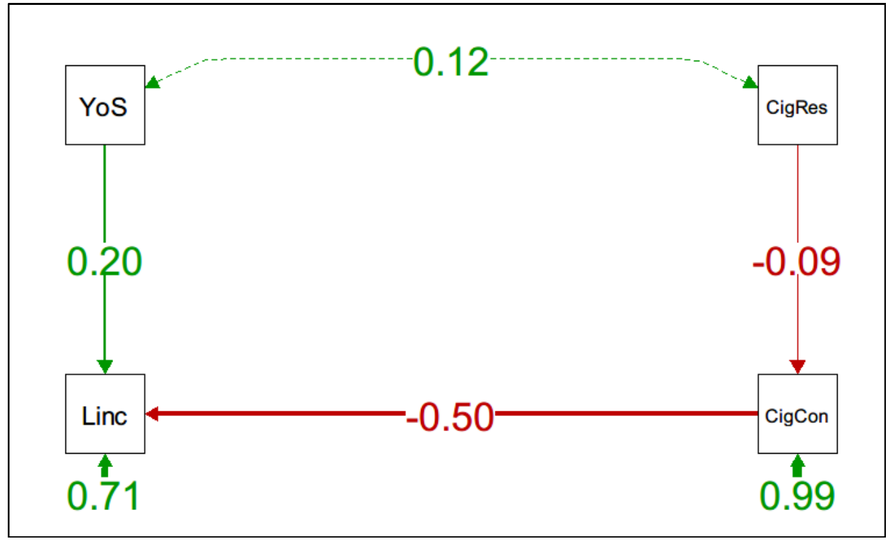
- Income ~ Cigarette Consumption + Year of School
- Cigarette ~ Cigarette Restriction

RMSE

2.001781 (Linc)
6.211374 (CigCon)

Variable	p-value
Linc~CigCon	0.000
Linc~YoS	0.000
CigCon~CigRes	0.024

Structural Equation Model (SEM) Path Diagram



Goodness-of-Fit Measurement

- Chi-Square Test of GoF

2.46
X2-Statistic

0.29
p-Value

Good Fit!

- Comparative Fit Index (CFI)

0.998 ~ 1
CFi

Good Fit!

- RMSE of Approximation (RMSEA)

0.017
Standard RMSEA

Good Fit!

Assumption is violated!

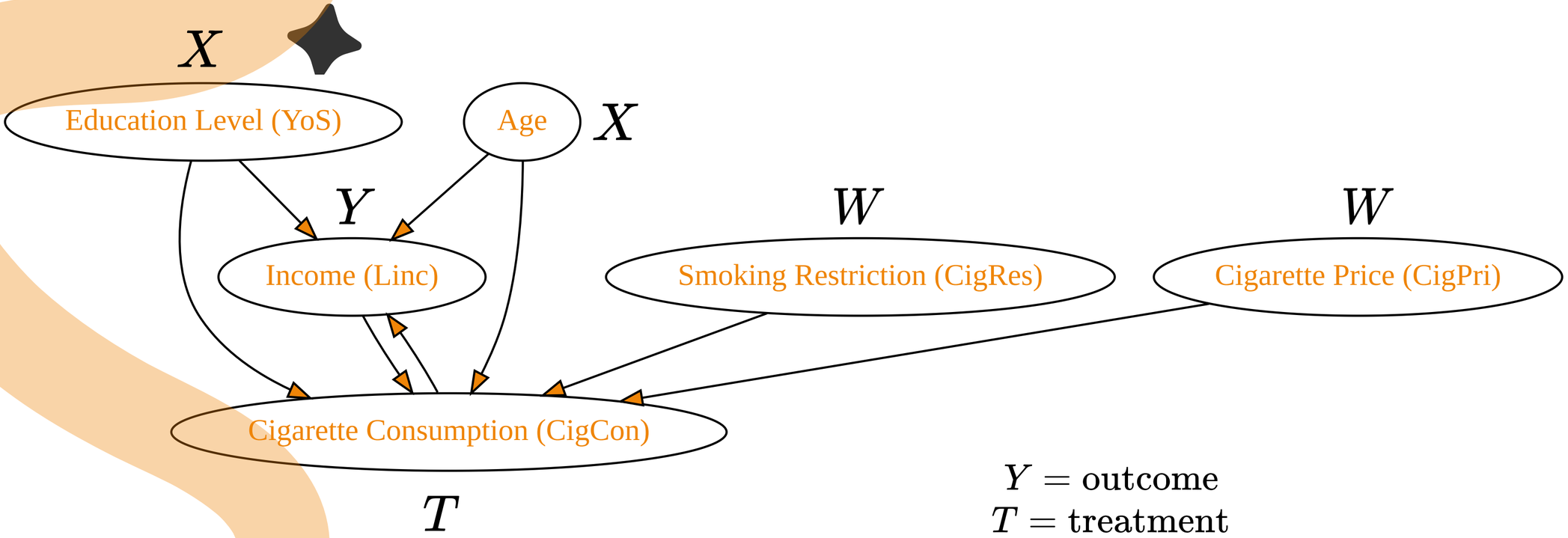
Even though the data fit well.

Another approach: Machine Learning method!

Causal Inference

Machine Learning Approach

Relationship of Variables



$Y = \text{outcome}$
$T = \text{treatment}$
$X = \text{covariate}$
$W = \text{confounder}$

Algorithm

Double Machine Learning

- Linear DML
- Sparse Linear DML
- Causal Forest DML

Machine Learning Algorithm Determination using PyCaret

Top 3 Outcome Regressor

Model	MSE	R2	MAPE
AdaBoost	1.1378	0.2879	1.0364
Gradient Boosting	1.1677	0.2673	0.9542
Linear Regression	1.1819	0,2580	1.0955

Top 3 Treatment Regressor

Model	MSE	R2	MAPE
Ridge Regression	185.3285	-0.0235	0.7844
Least Angle Regression	185.4286	-0.0240	0.7849
Linear Regression	185.4286	-0.0240	0.7849

Best Model Determination

Evaluation Metrics using **RMSE** on **k-Fold Cross Validation**

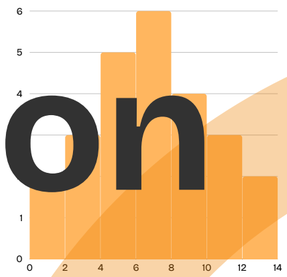
30 Folds

Homogeneity of Variance

Levene Test

0.91

p-Value

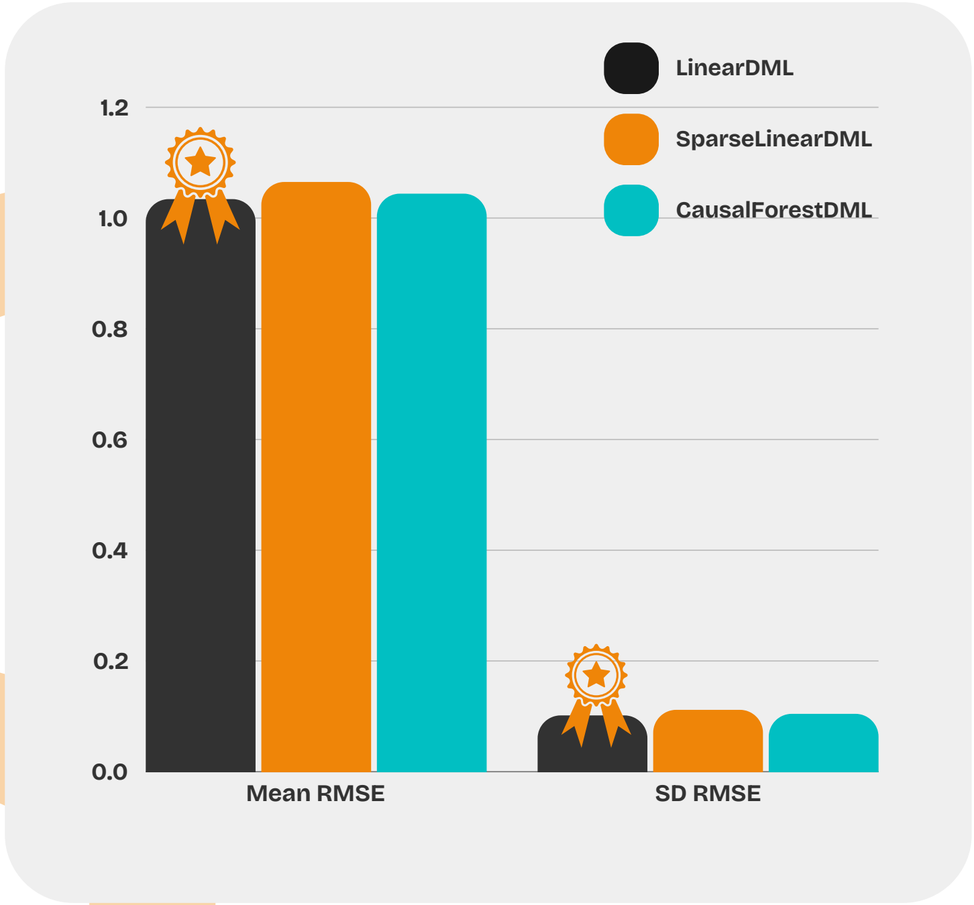


Normality of RMSE

ALL

Normally Distributed

Based on Shapiro-Wilk Test

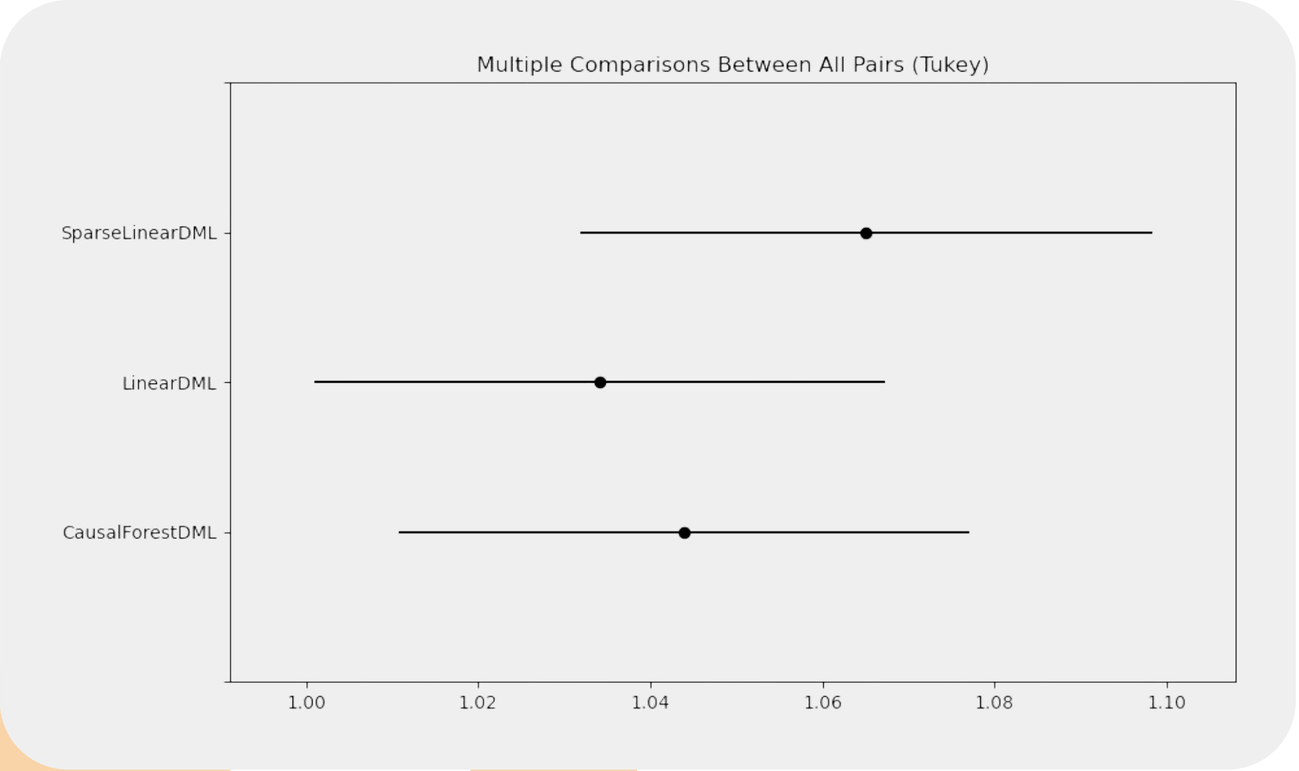


ANOVA Test on RMSE

0.65
F-Statistic

0.53
p-Value

No significant difference between the models



Best Model: Linear Double Machine Learning
with mean RMSE of 1.034

Linear DML Summary

Coefficient Results

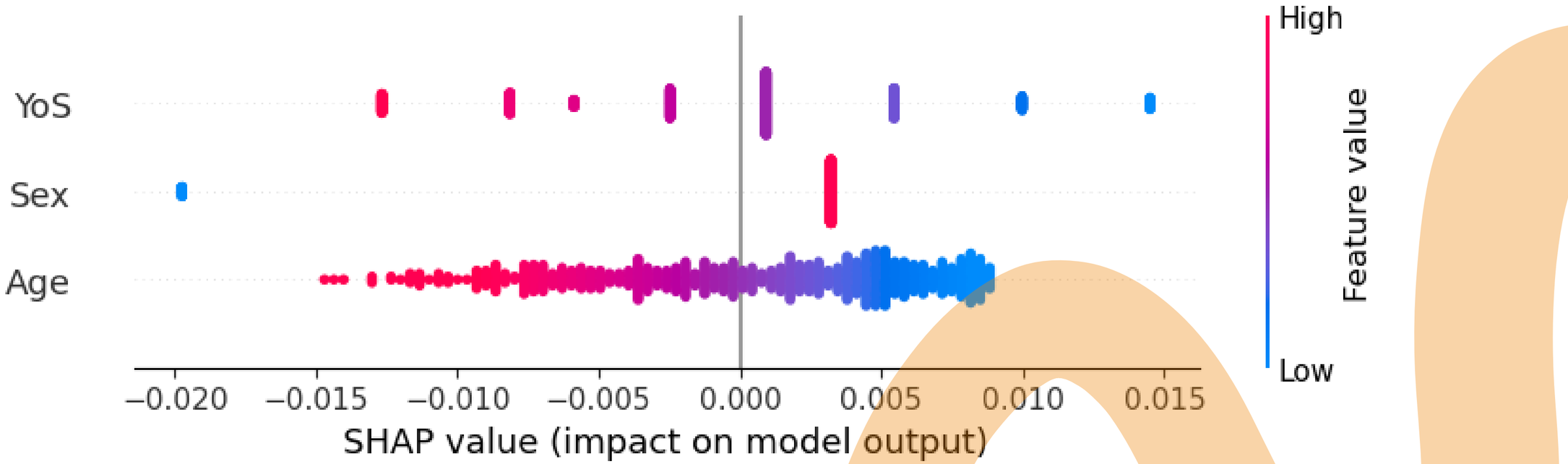
	point_estimate	stderr	zstat	pvalue	ci_lower	ci_upper
YoS	-0.002	0.001	-2.383	0.017	-0.004	-0.0
Age	-0.0	0.0	-1.488	0.137	-0.001	0.0
Sex	0.023	0.009	2.695	0.007	0.006	0.04

CATE Intercept Results

	point_estimate	stderr	zstat	pvalue	ci_lower	ci_upper
cate_intercept	-0.022	0.019	-1.181	0.238	-0.058	0.014

- Only Year of School and Sex significantly impact Income.
- YoS is negatively correlated to Income and Sex is positively correlated to Income.
- Treatment (Cigarette Consumption) effect is not significant.

SHAP Values


YoS
Sex
Age
−0.020
−0.015
−0.010
−0.005
0.000
0.005
0.010
0.015
SHAP value (impact on model output)
High
Feature value
Low

Conclusions

Based on RMSE value, Linear DML performs better than Structural Equation Modeling.

Relationship between variables:

Based on Structural Equation Modeling:

- Year of School positively correlate Income
- No Cigarette Restriction tends to make Cigarette Consumption higher
- Cigarettte Consumption negatively correlate Income

This make more sense!

Based on Linear Double Machine Learning:

- Age negatively correlate Income, but unsignificant
- Male tends to have higher Income
- Year of School negatively correlate Income
- Cigarette Consumption negatively correlate Income, but unsignificant

Thank You

canva link:
https://www.canva.com/design/DAGTH5-FU2w/evK1qVNIGiGssb-qeF7Jtw/edit?utm_content=DAGTH5-FU2w&utm_campaign=designshare&utm_medium=link2&utm_source=sharebutton