Why MARS

MARS (Part 1)

Lecture Video Slides

Recall the Linear Regression Model

$$y = \underbrace{b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_m x_m}_{\hat{y}} + e$$

Straight Line Equation

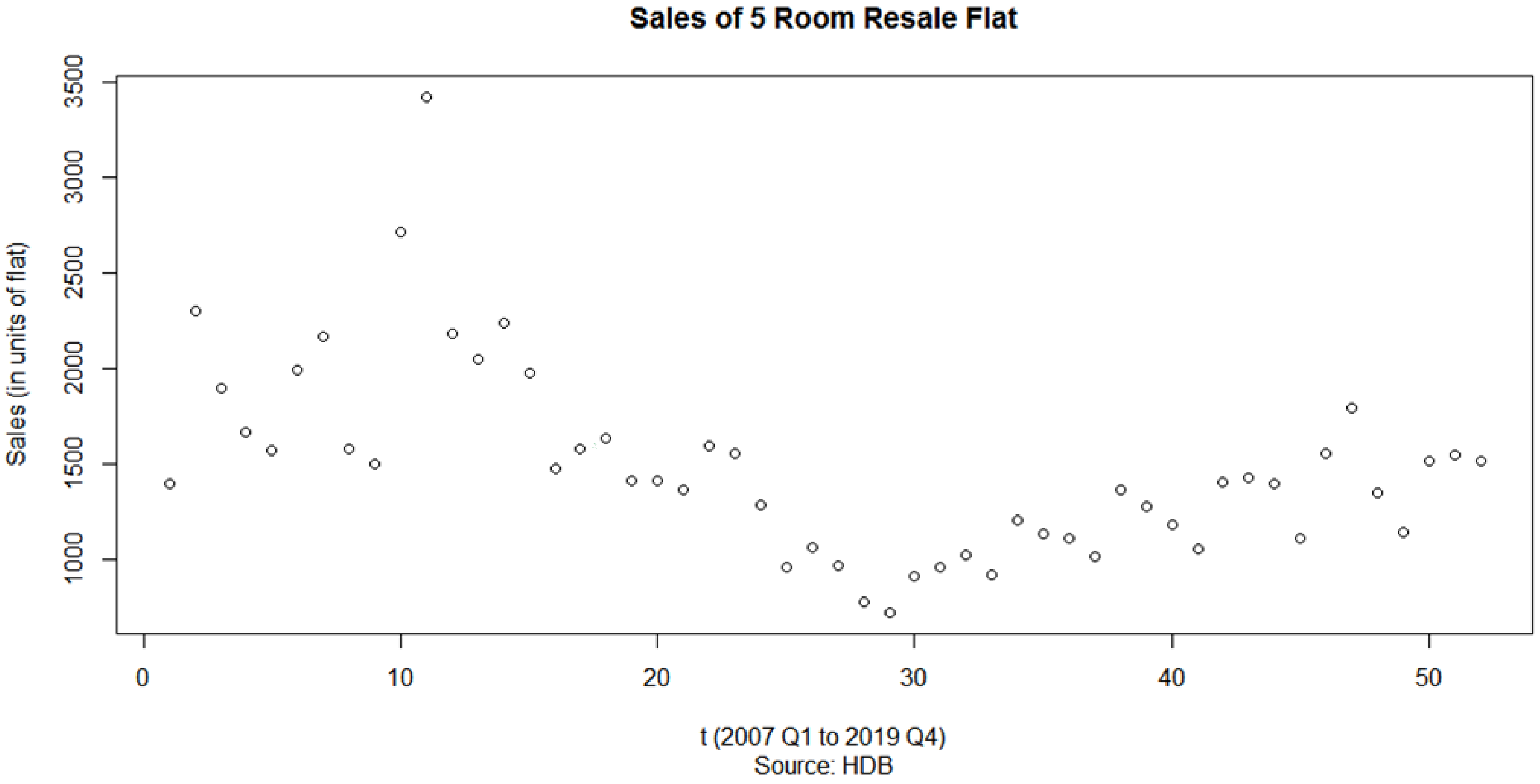
e ~ N(0, σ)

Errors (aka Residuals) follow a Normal Distribution with mean 0 and constant standard deviation.

Assumes LINEAR trend relating Y to all the Xs.
What if non-linear?
Fit Quadratic? Fit Cubic?

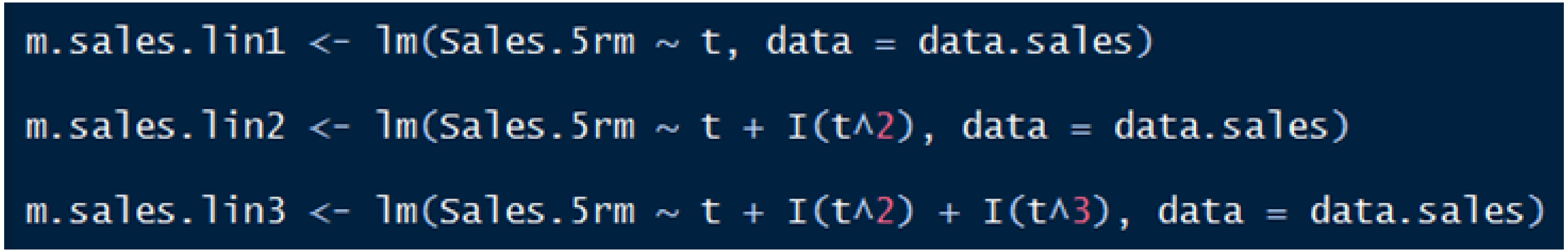
HDB 5-room Flat Resale data

Real Dataset: 5 room flat resale applications.csv



- Linear Trend? Quadratic? Cubic?

Fitting Linear, Quadratic and Cubic Trends in R



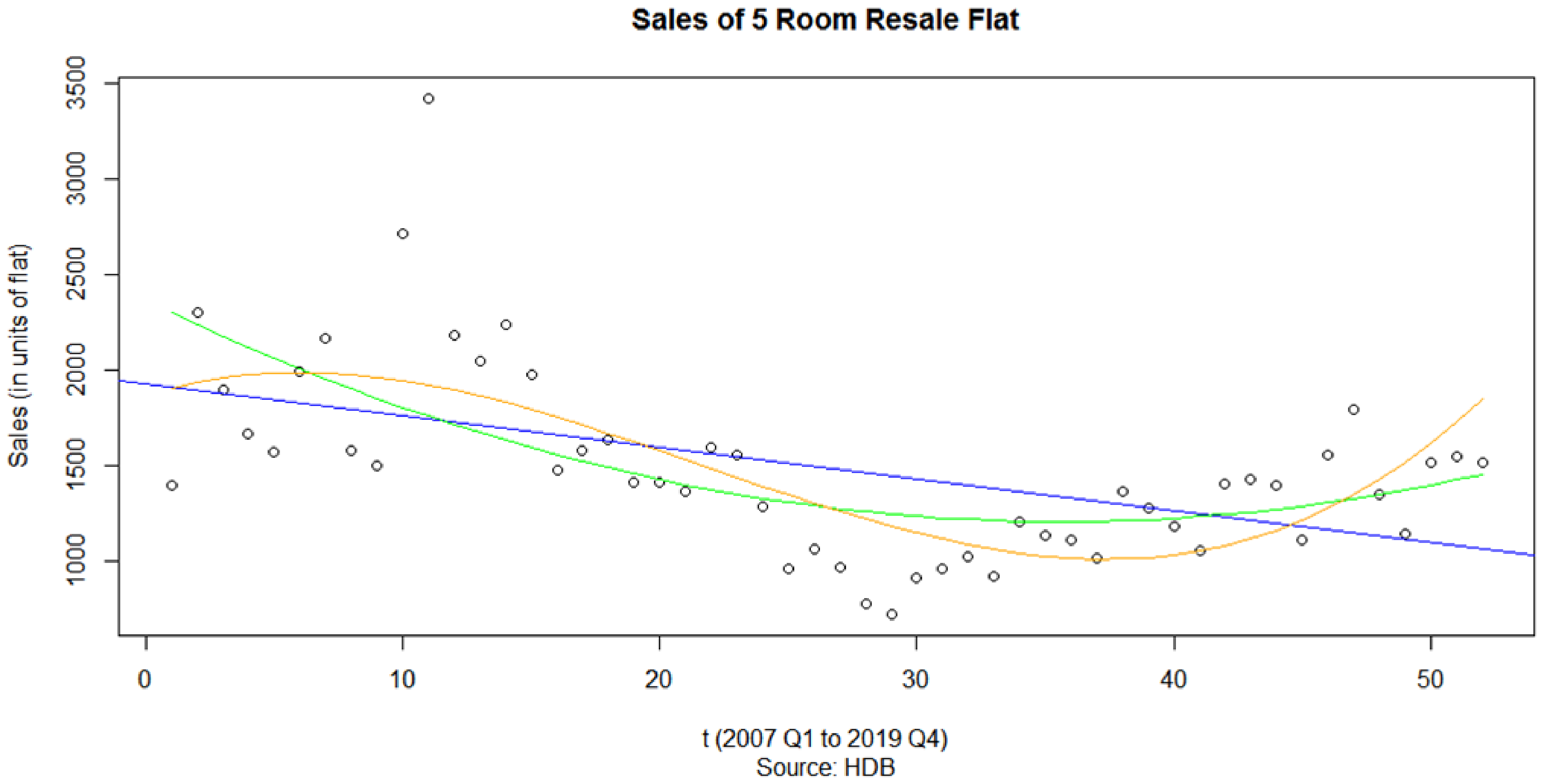
```
m.sales.lin1 <- lm(Sales.5rm ~ t, data = data.sales)

m.sales.lin2 <- lm(Sales.5rm ~ t + I(t^2), data = data.sales)

m.sales.lin3 <- lm(Sales.5rm ~ t + I(t^2) + I(t^3), data = data.sales)
```

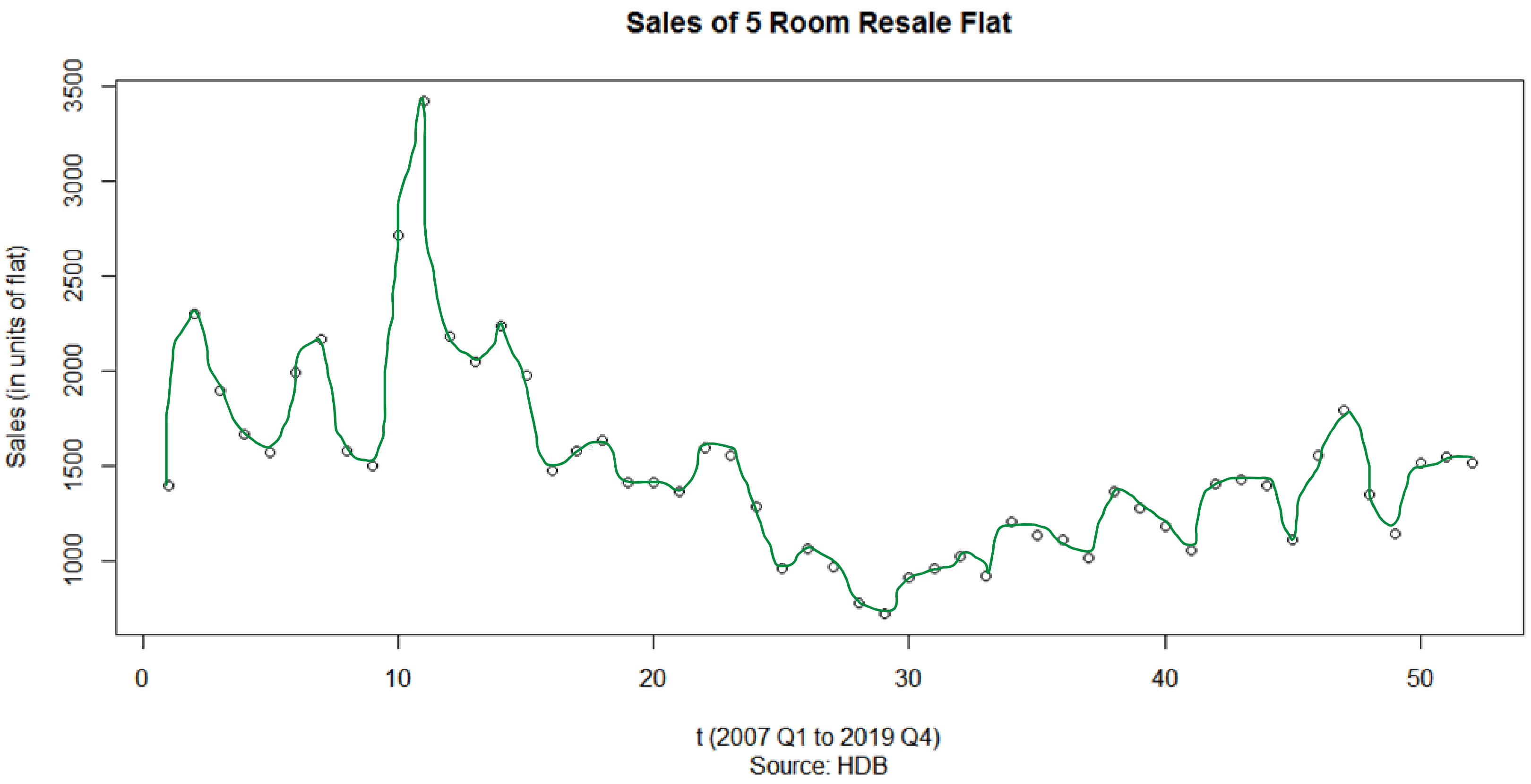
Note: Necessary to use I()

Fitted Linear, Quadratic and Cubic Trends



- Q: What is the problem with using Linear/Quadratic/Cubic reg?
- A: The “trend” applies globally throughout the entire data.

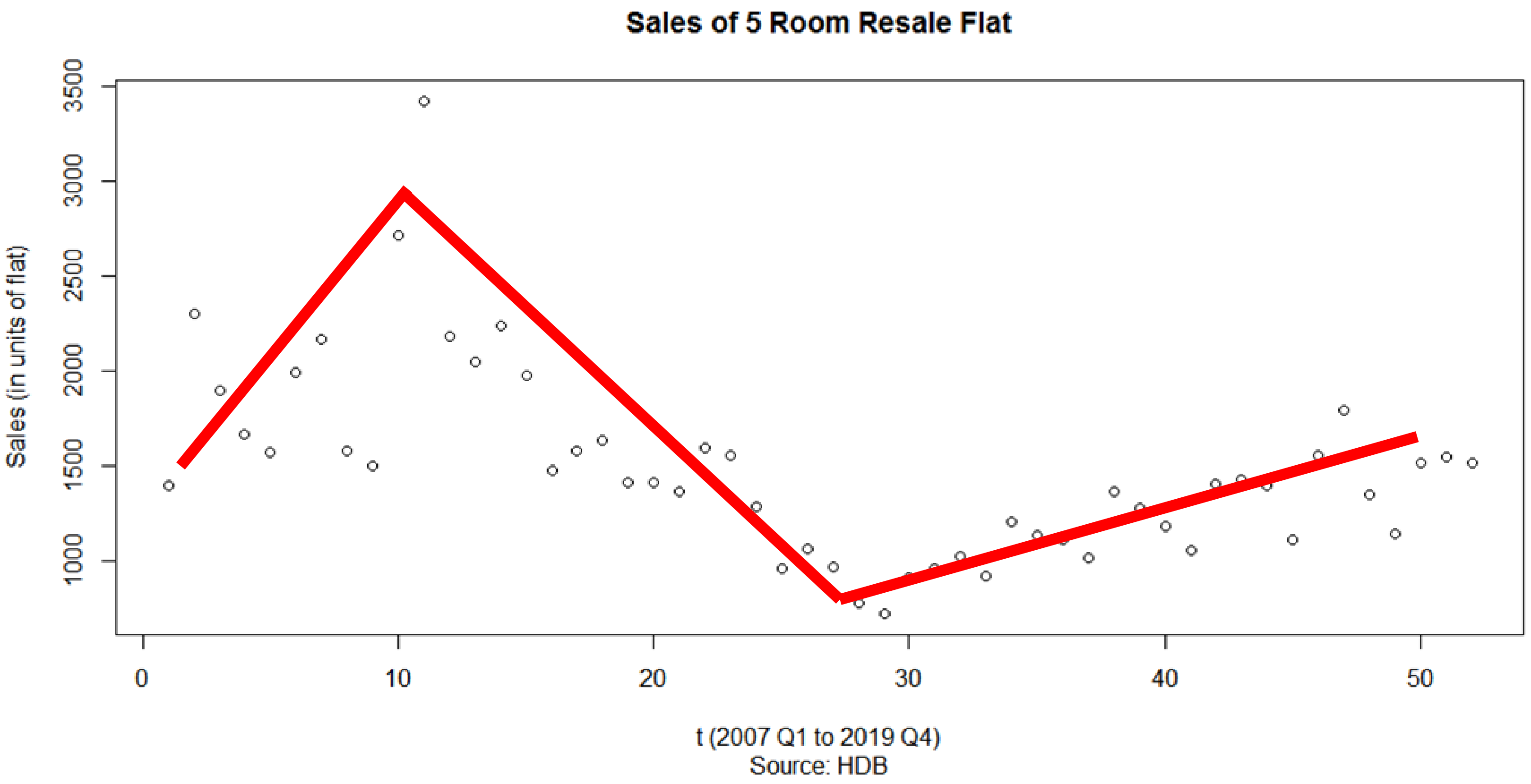
Can a data-dependent function fit the data better?



Overfitting.

- Go thru all data points (Polynomial Interpolation Theorem). Perfect Fit!

Can a data-dependent function fit broad trends in the data?

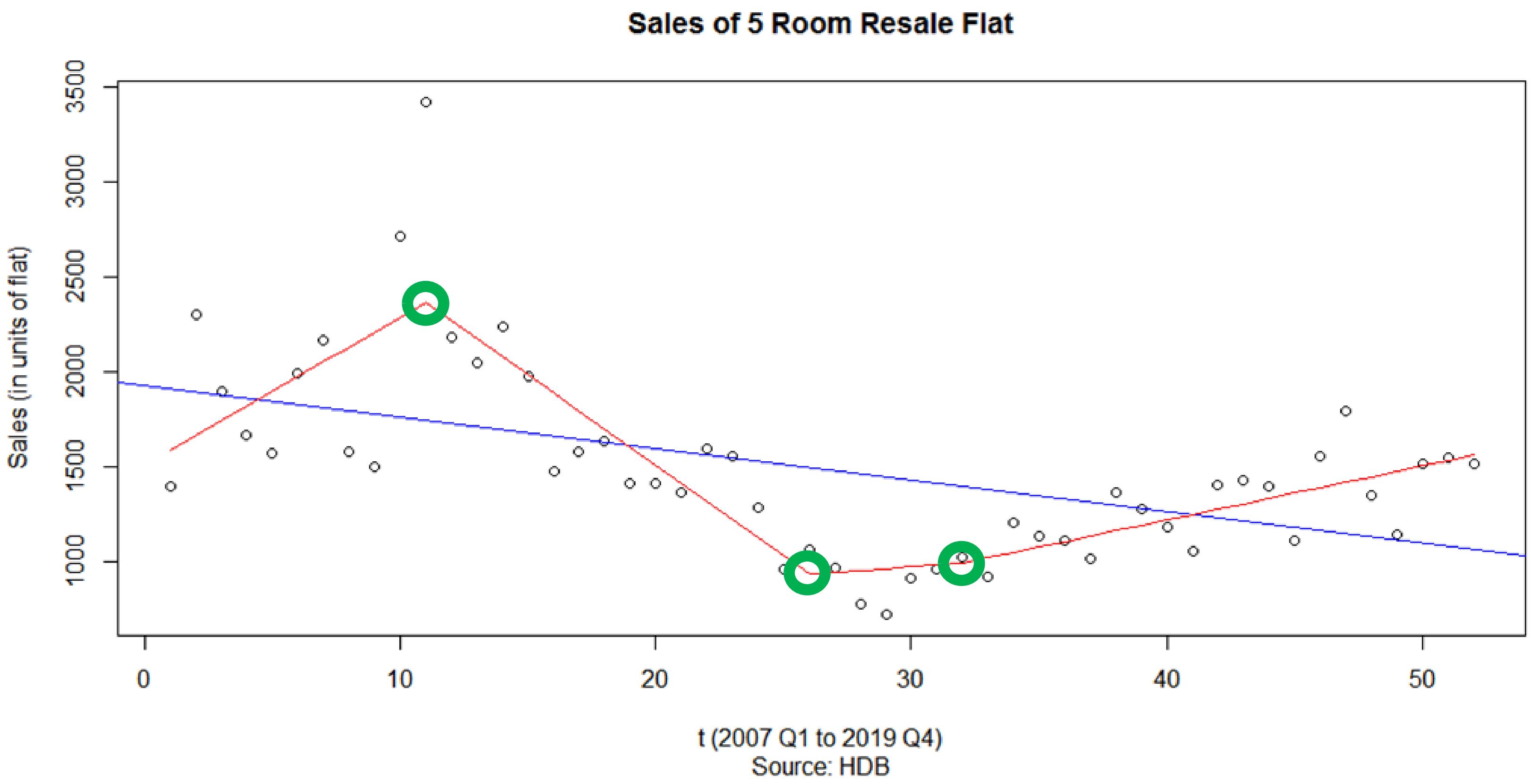


Source: HDB

How to do this automatically from data, without human intervention, especially with multiple X variables?

- Fit obvious trends and only the obvious key (not every) turning points.

Linear vs MARS in Rscript: flatsales-mars.R



MARS test and find best-fitting hinge functions

- Automatically from data
- The knots (aka cuts) t = 11, 26, 32 are found automatically in MARS. [How?]
- What is a hinge function?

Next Video

- Hinge Functions.
- Optimisation Process that determines the knots.
- Automated Variable Selection.
- Automated (Variable) Interaction Selection.