LECTURE 4

CEIC6789 NOTES

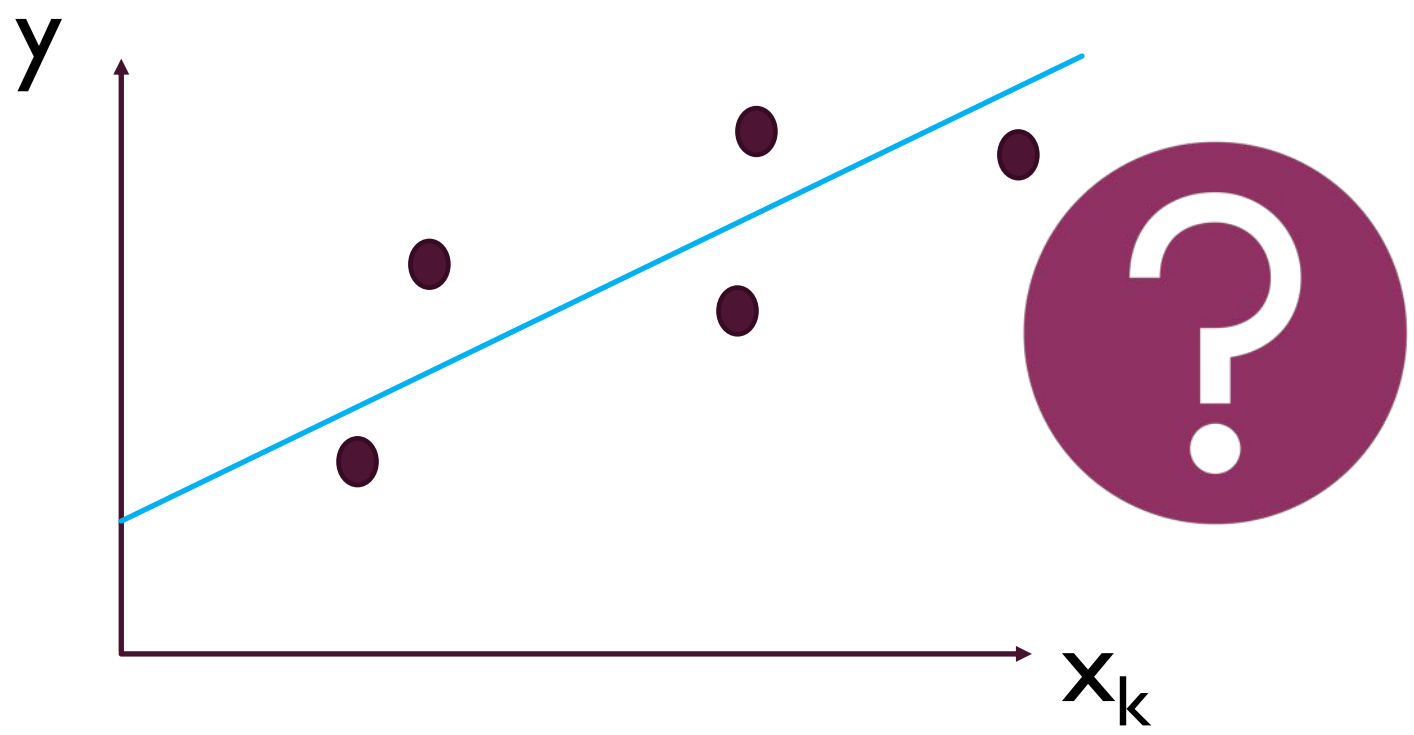
PREVIOUS WEEK

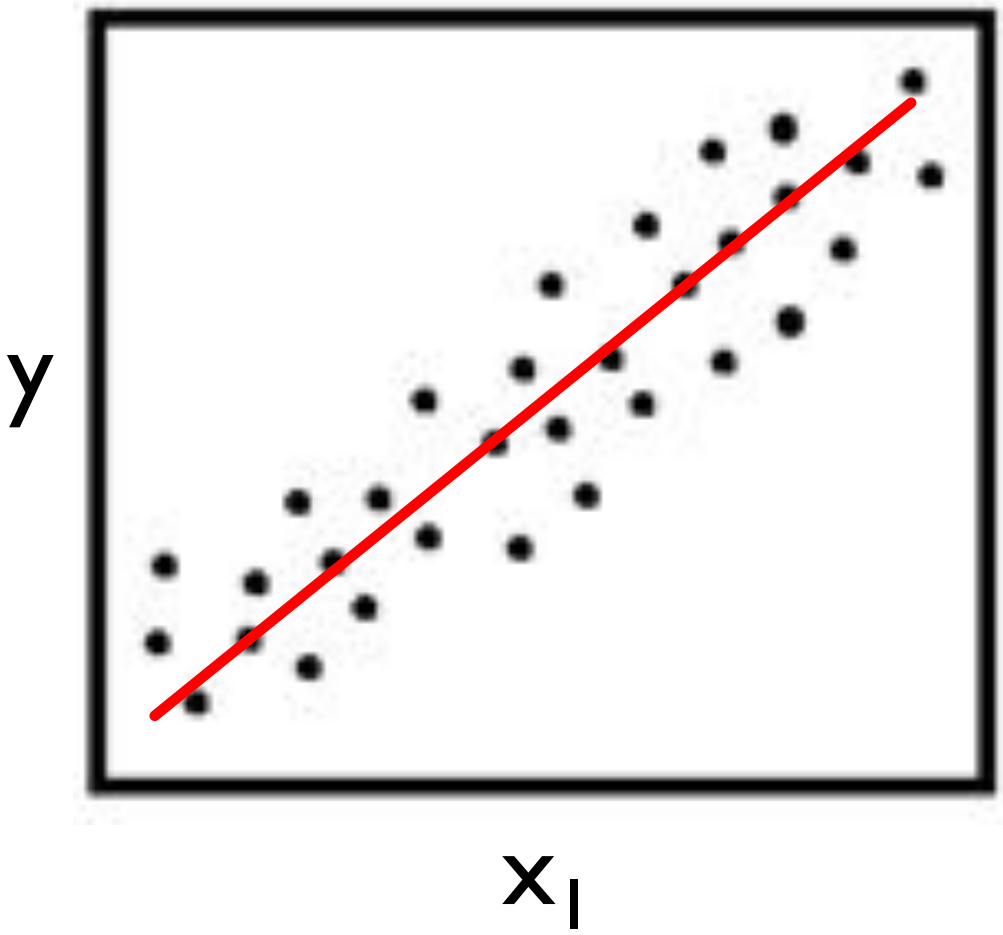
- Difference between correlation and causation
- Simple linear regression model
- Multiple linear regression model
- Decomposition of variability - SST, SSR and SSE
- Ordinary least squares
- R-squared and adjusted R-squared
- Feature selection using F-test

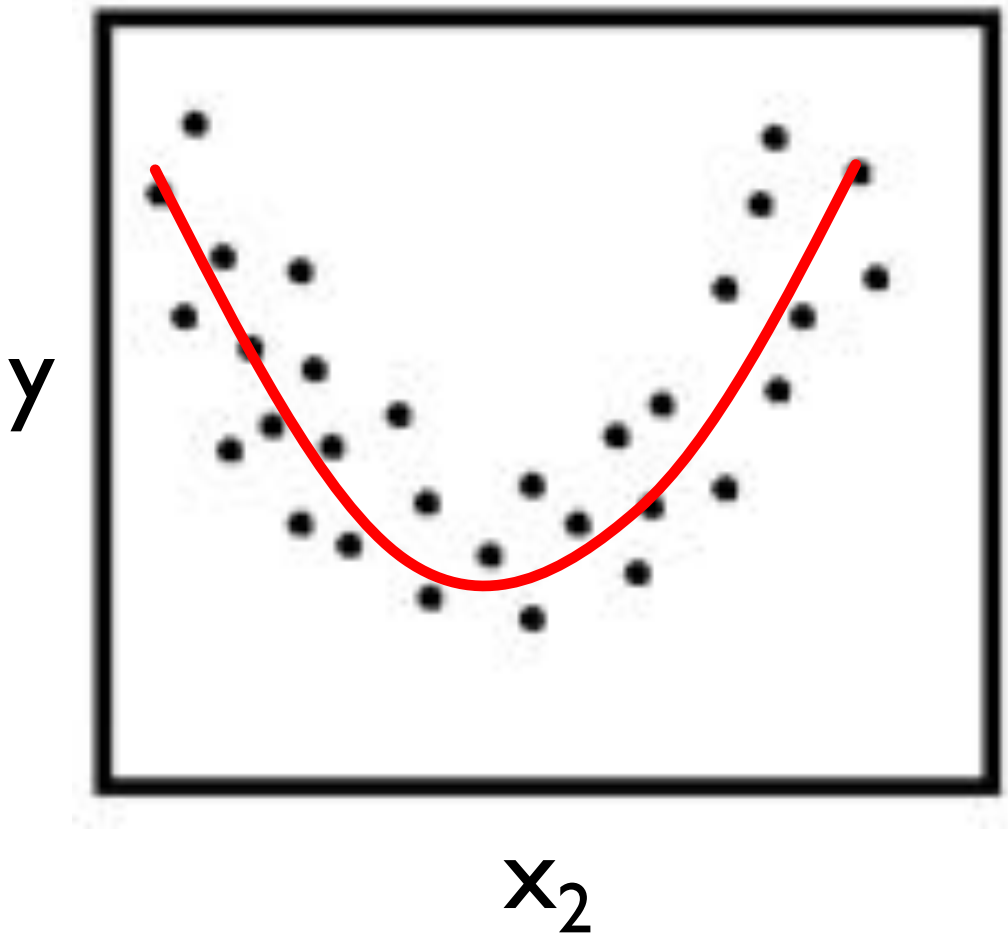
ASSUMPTIONS

LINEARITY

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_k x_k$$

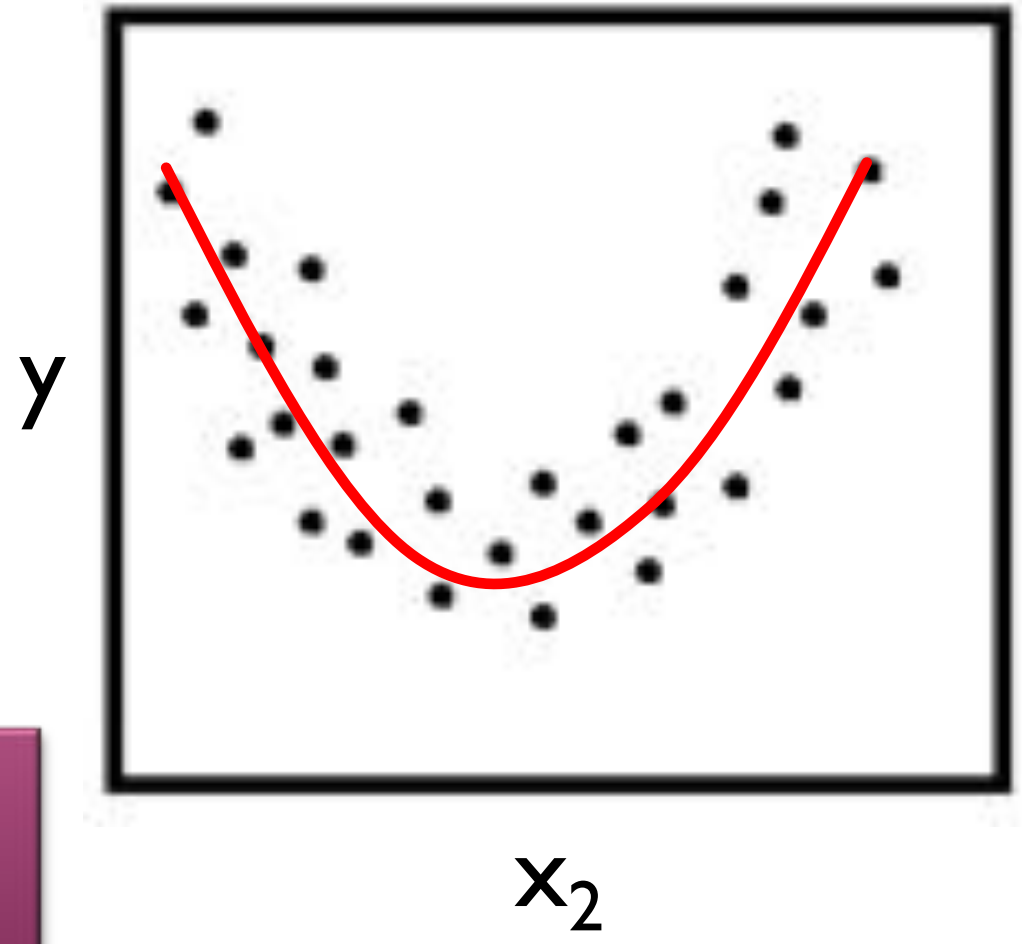



y
x_1


y
x_2

$$x_2 \rightarrow (x_2)^2$$

$$\hat{y} = b_0 + b_2(x_2)^2$$



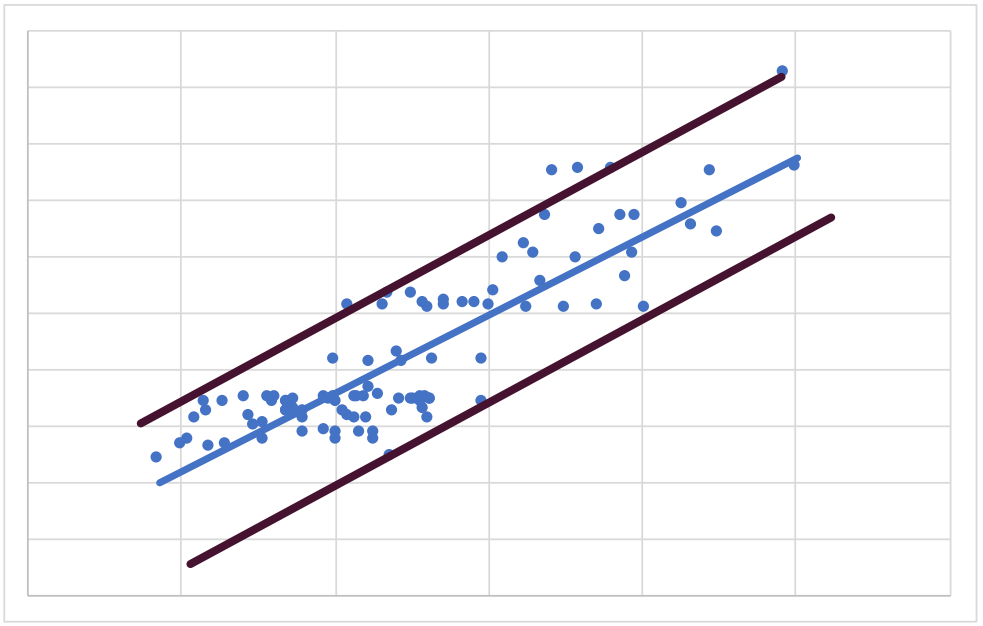
If the relationship is nonlinear, you should not use the data directly. You should transform it appropriately and then proceed with linear regression models.

HOMOSCEDASTICITY

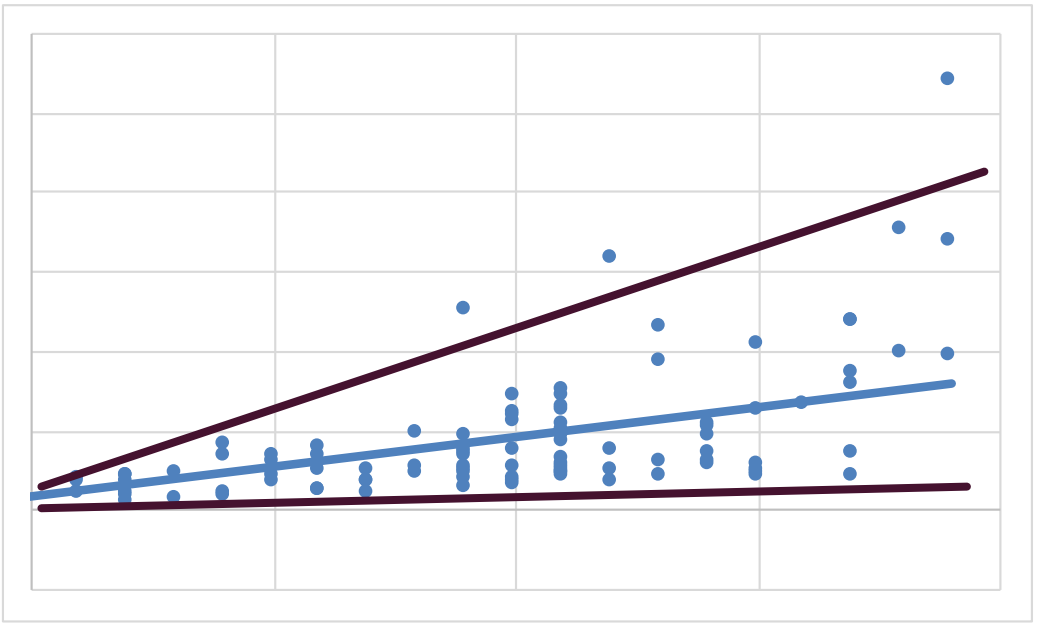
To have equal variance

SSE

$$(\sigma_{\varepsilon_1})^2 = (\sigma_{\varepsilon_2})^2 = \cdots = (\sigma_{\varepsilon_k})^2$$



Homoscedastic

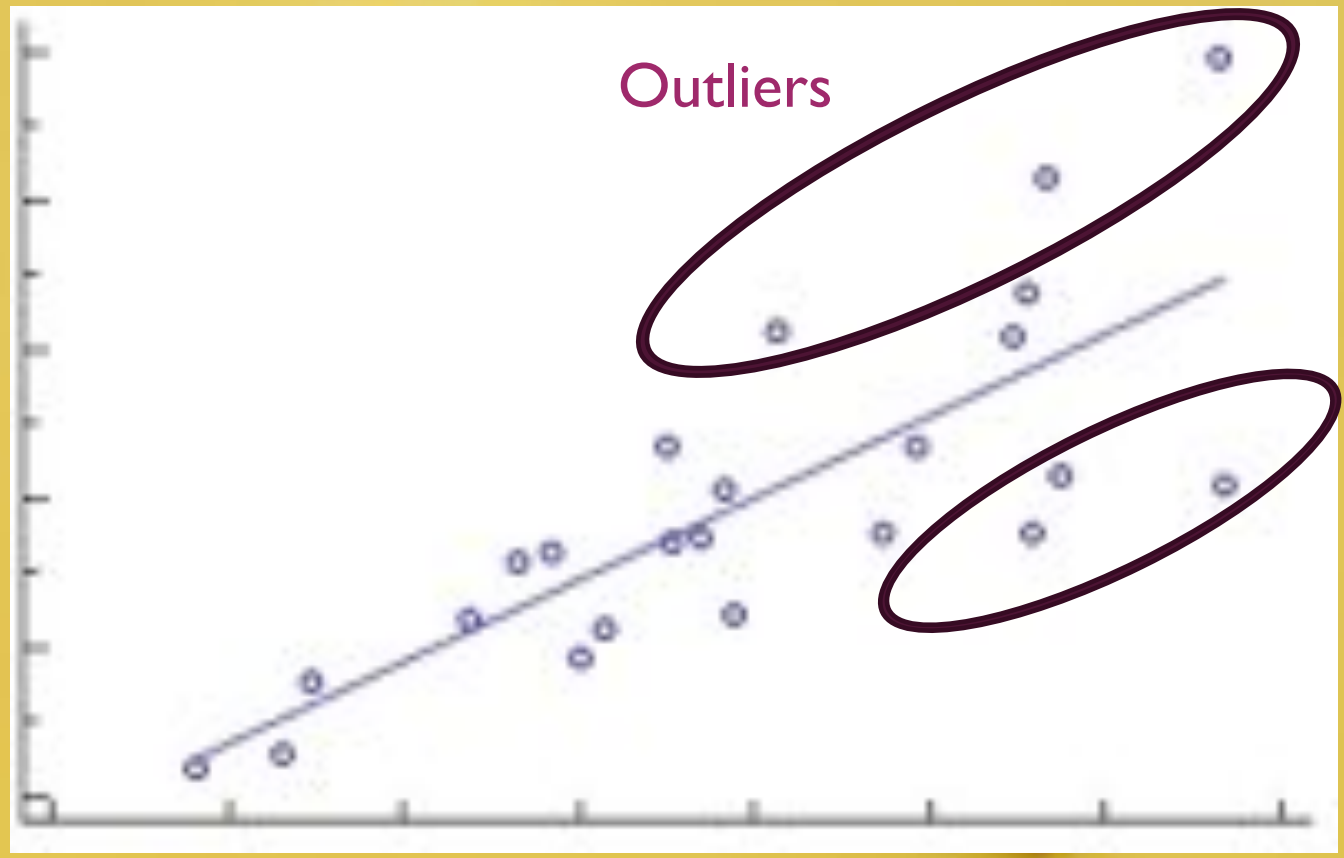


Heteroscedastic

HOW TO OVERCOME HETEROSCEDASTICITY?

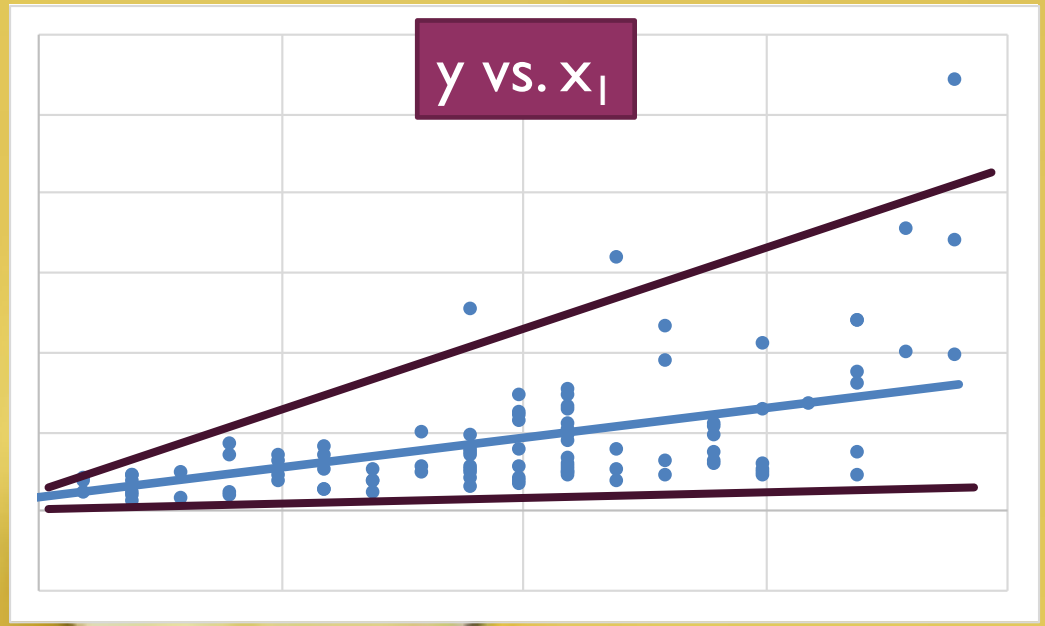
HOW TO OVERCOME HETEROSCEDASTICITY?

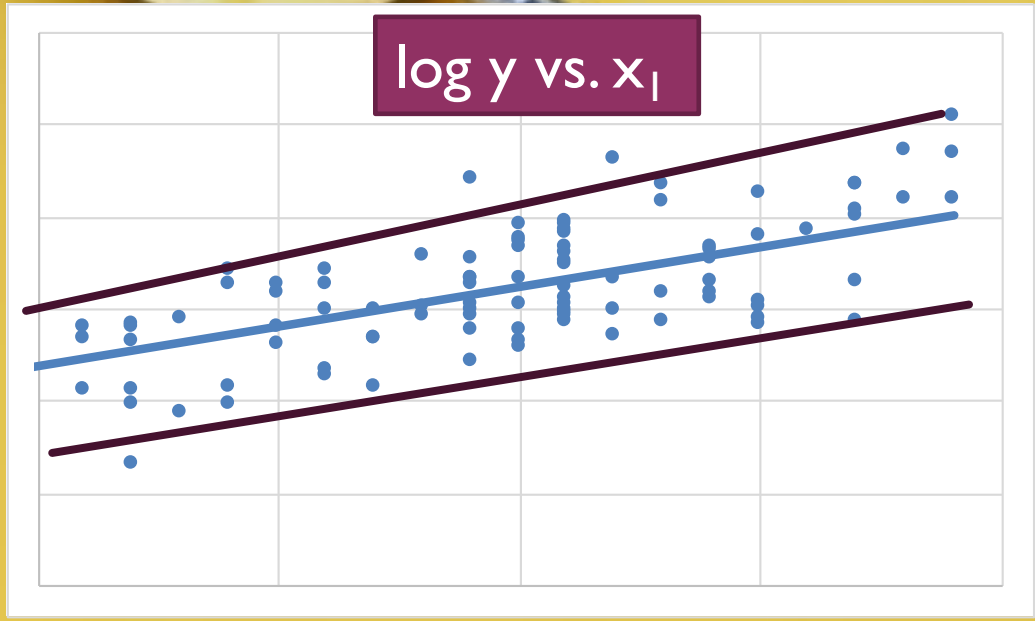
- Look for outliers and try to remove them



HOW TO OVERCOME HETEROSCEDASTICITY?

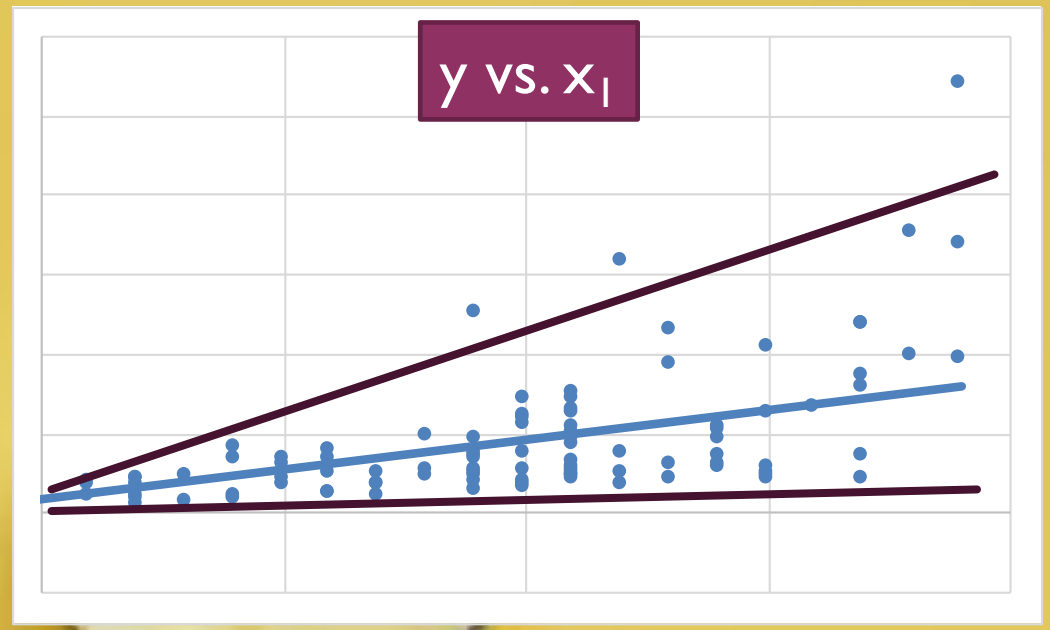
- Look for outliers and try to remove them
- Log transformation

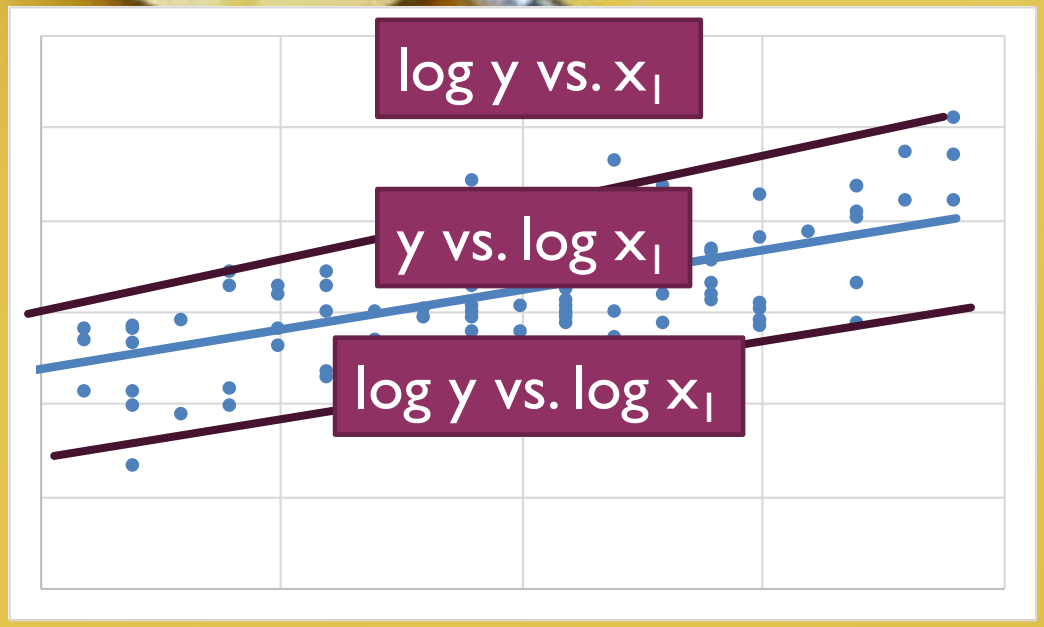




HOW TO OVERCOME HETEROSCEDASTICITY?

- Look for outliers and try to remove them
- Log transformation



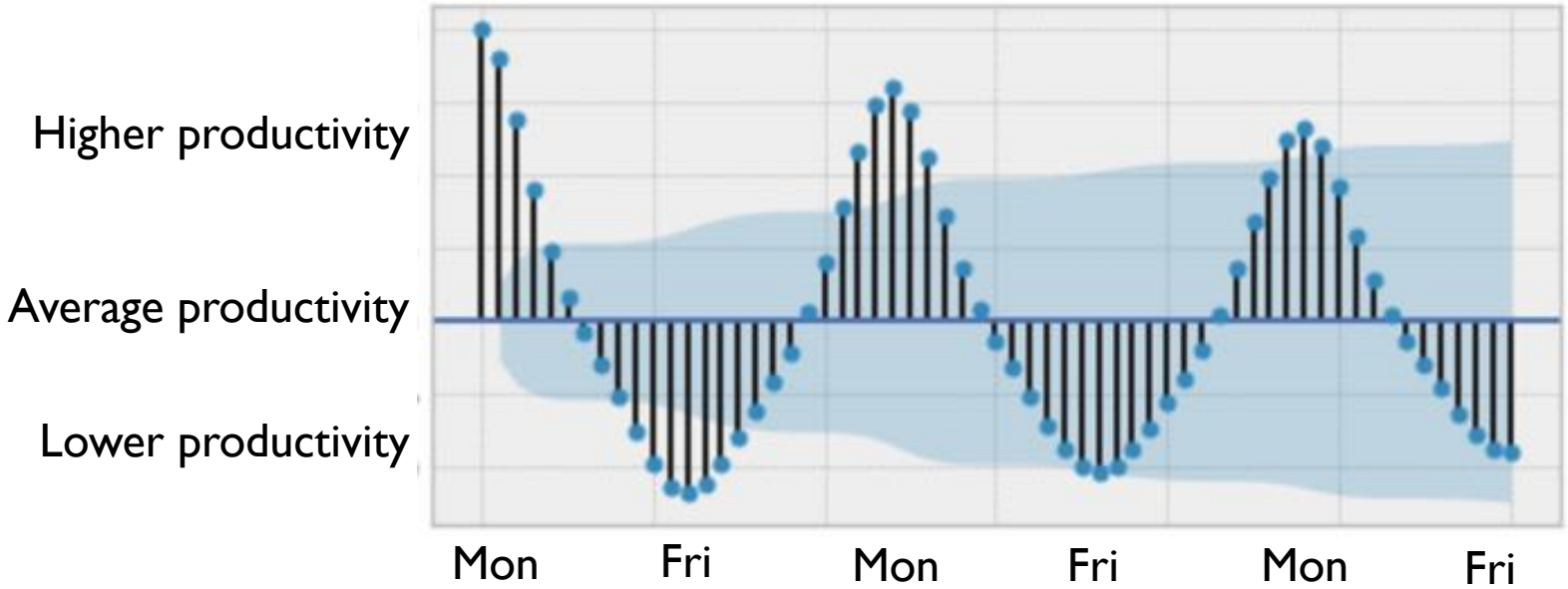


NO AUTOCORRELATION

Error terms (individual terms part of SSE) are assumed to be uncorrelated

$$\sigma_{\varepsilon_i \varepsilon_j} = 0 : \forall\, i \neq j$$

TIME SERIES DATA



FIXES

- Avoid linear regression
- Apply other regressions
  - Auto regressive model
  - Moving average model
  - Auto regressive moving average model

NO MULTICOLLINEARITY

Two or more variables have a high degree of correlation

$$\rho_{x_i x_j} \approx 1 : \forall\, i, j;\ i \neq j$$

INFANT HEALTH

Health	Age	Weight

NO MULTICOLLINEARITY

Two or more variables have a high degree of correlation

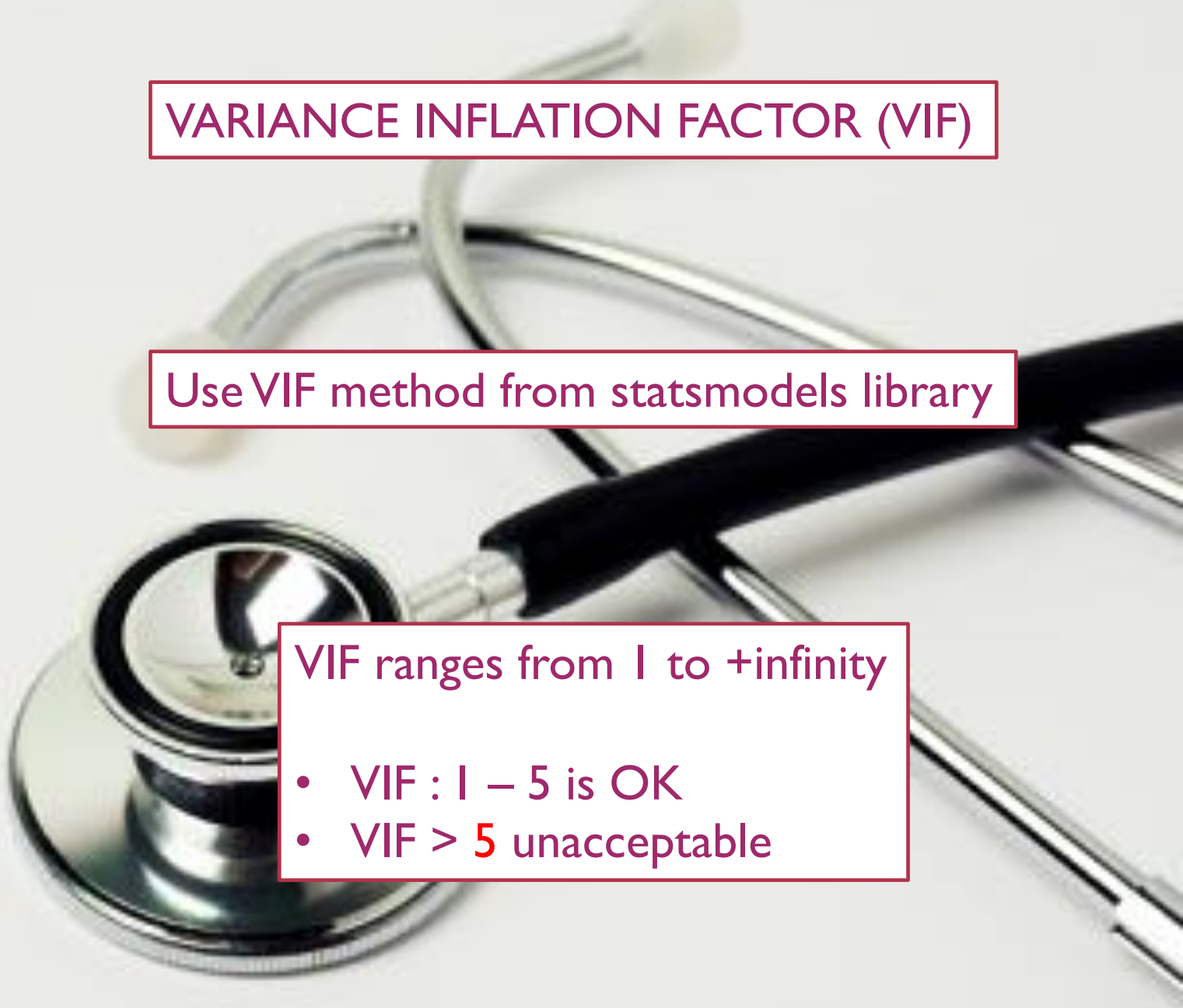
$$\rho_{x_i x_j} \approx 1 : \forall\, i, j;\ i \neq j$$

INFANT HEALTH

Health	Age	Weight

Messed-up coefficients; wrong p values in F regression etc.

HOW TO CHECK FOR MULTICOLLINEARITY?



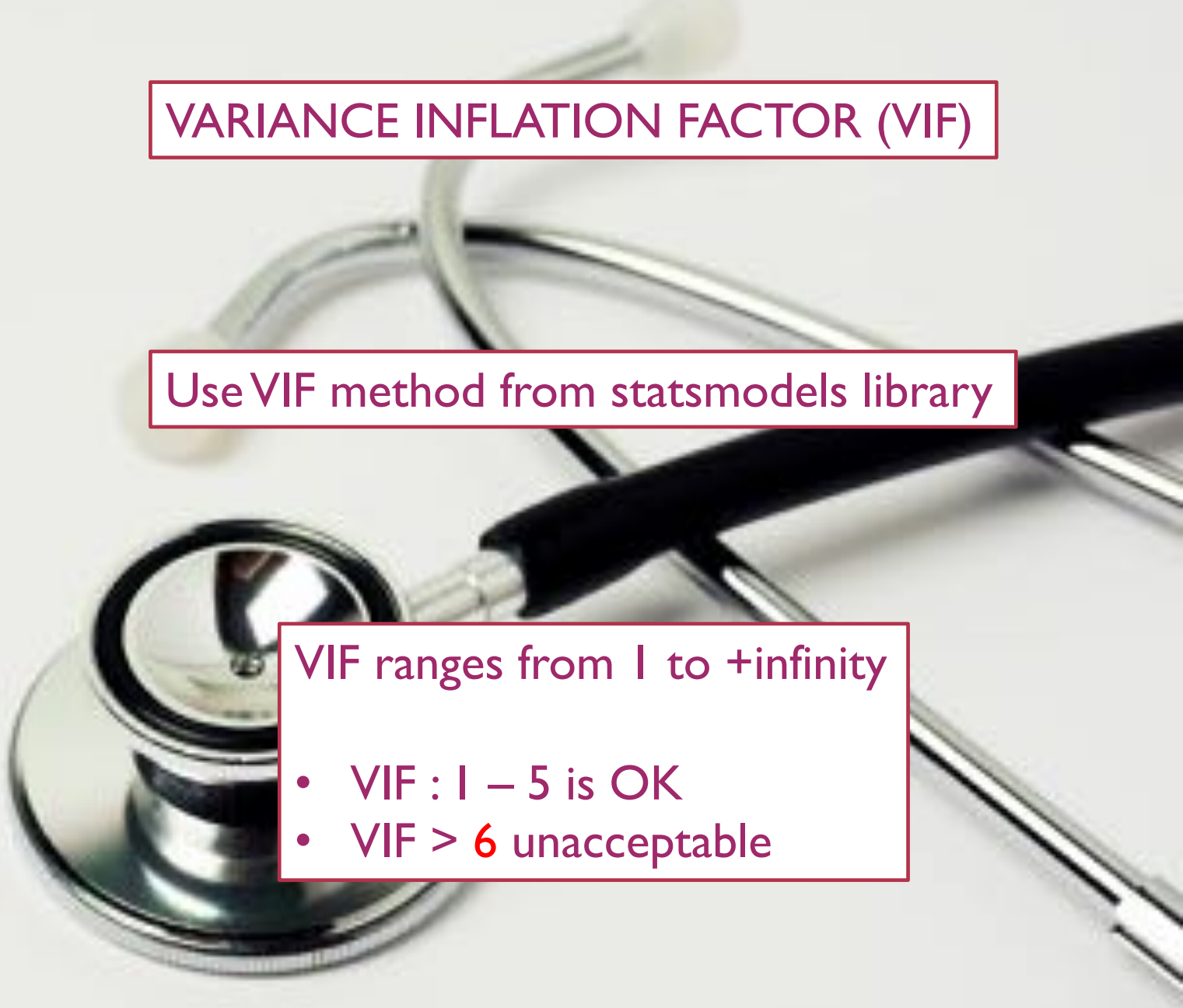
VARIANCE INFLATION FACTOR (VIF)

Use VIF method from statsmodels library

VIF ranges from 1 to +infinity

- VIF : 1 – 5 is OK
- VIF > 5 unacceptable

HOW TO CHECK FOR MULTICOLLINEARITY?



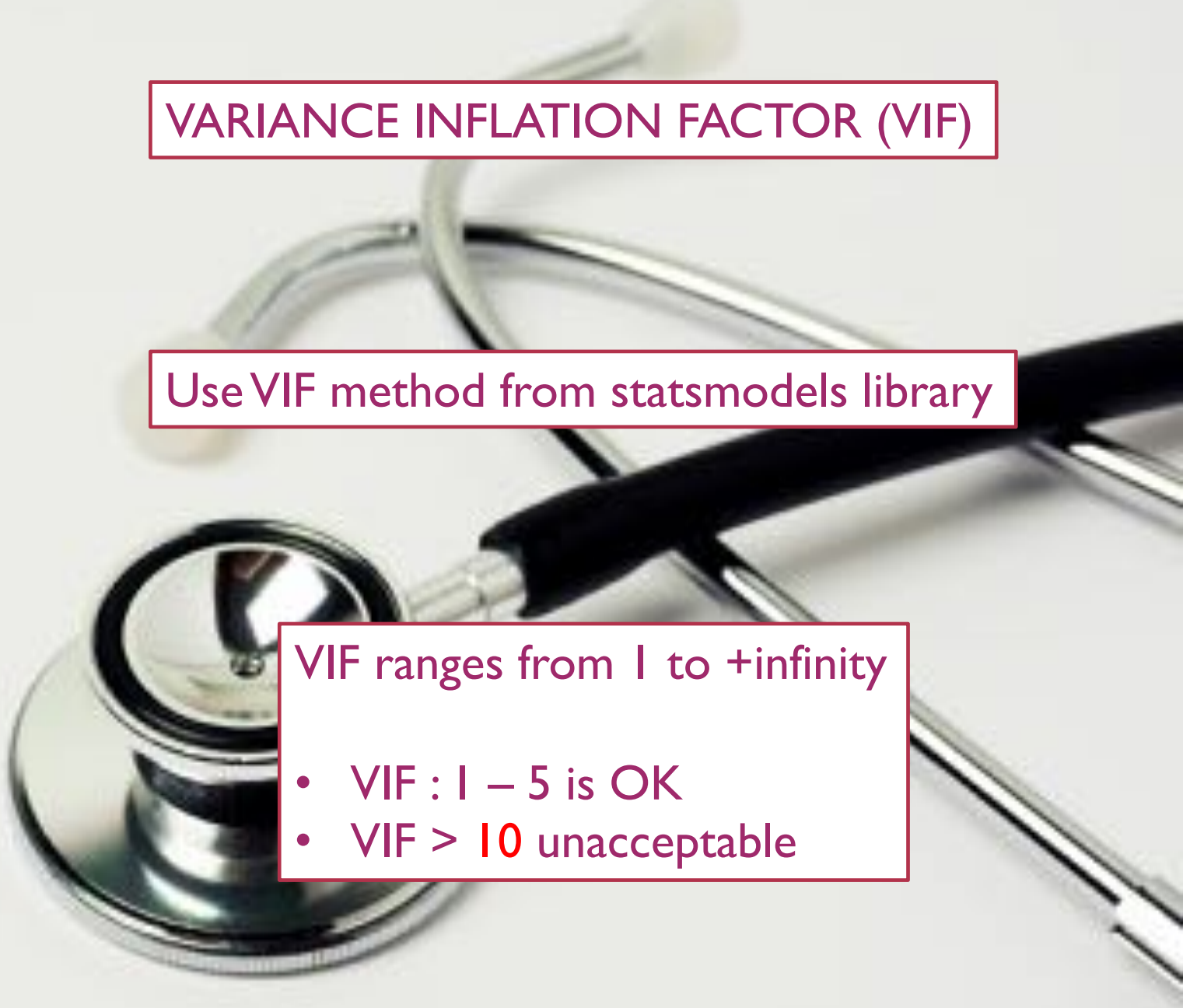
VARIANCE INFLATION FACTOR (VIF)

Use VIF method from statsmodels library

VIF ranges from 1 to +infinity

- VIF : 1 – 5 is OK
- VIF > 6 unacceptable

HOW TO CHECK FOR MULTICOLLINEARITY?



VARIANCE INFLATION FACTOR (VIF)

Use VIF method from statsmodels library

VIF ranges from 1 to +infinity

- VIF : 1 – 5 is OK
- VIF > 10 unacceptable

FIXES

Drop one of the two variables

Transform into one variable

Keep the variables in model, but handle with caution

Here is a nice article about multicollinearity: https://statisticsbyjim.com/regression/multicollinearity-in-regression-analysis/

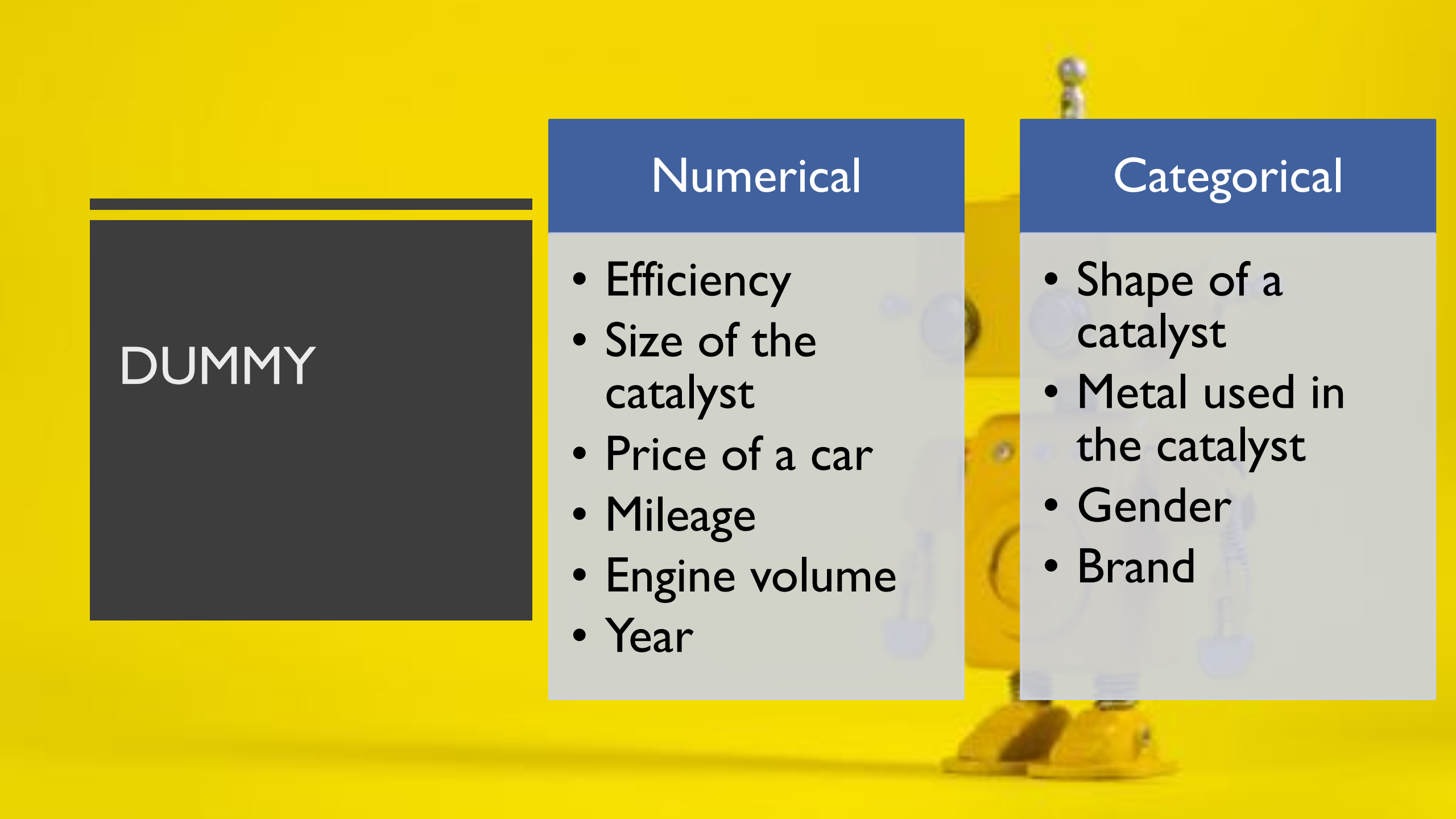
DUMMY VARIABLES

DUMMY

A variable used to include categorical data into a regression model

DUMMY

Numerical	Categorical
• Efficiency • Size of the catalyst • Price of a car • Mileage • Engine volume • Year	• Shape of a catalyst • Metal used in the catalyst • Gender • Brand



Gender	Gender_dummy
Male	0
Female	1

Metal
Platinum
Gold
Silver
Copper

Gender	Gender_dummy
Male	0
Female	1

How many dummies to create?

Metal
Platinum
Gold
Silver
Copper

Gender	Gender_dummy
Male	0
Female	1

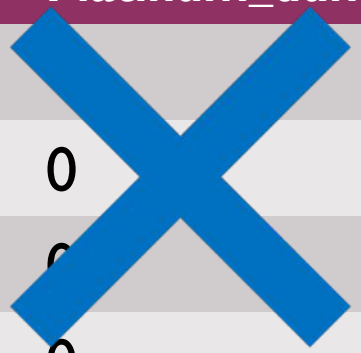
Metal
Platinum
Gold
Silver
Copper

How many dummies to create?

If we have N categories, we have to create N-1 dummies

Gender	Gender_dummy
Male	0
Female	1

Metal	Platinum_dummy	Gold_dummy	Silver_dummy	Copper_dummy
Platinum		0	0	0
Gold	0	1	0	0
Silver	0	0	1	0
Copper	0	0	0	1

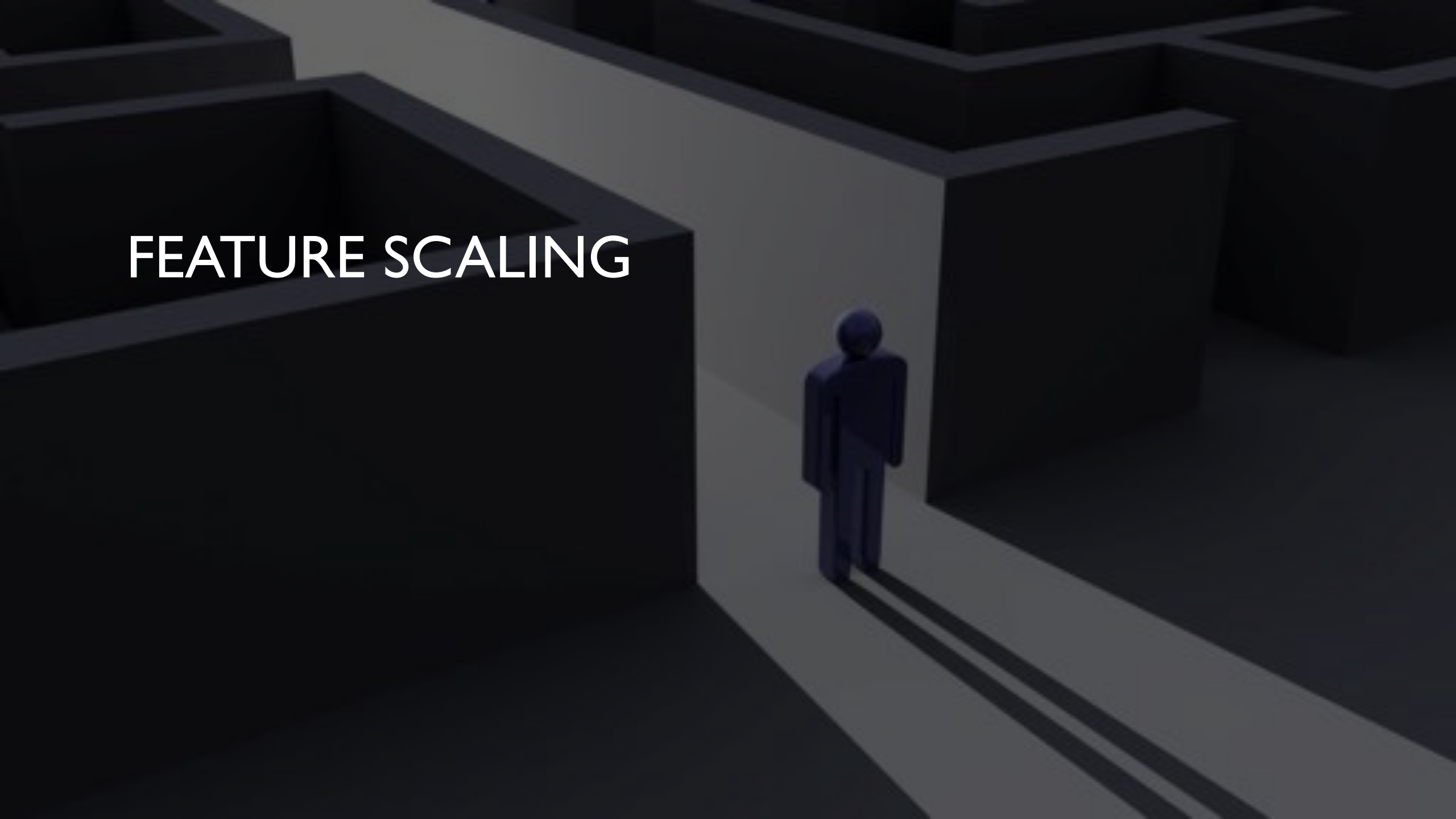


Gender	Gender_dummy
Male	0
Female	1

Metal	Platinum_dummy	Gold_dummy	Silver_dummy	Copper_dummy
Platinum	0	0	0	0
Gold	1	0	0	
Silver	0	0	1	0
Copper	0	0	0	1

- Redundant
- Multicollinearity

FEATURE SCALING



FEATURE SCALING

Transforming the data into a standard scale

$$x' = \frac{x - \mu}{\sigma}$$

Mean

Standard deviation

x':

Mean = 0

Standard deviation = 1

ACTIVITY

Given the following (simple) dataset, obtain the transformed dataset by subtracting the mean and dividing by standard deviation values for each of the columns.

Note that we will use sklearn to carry out feature scaling in practical cases, where sklearn uses the population standard deviation. Therefore, you can use the formula for the population standard deviation in your activity.

NO. OF DAYS	DISTANCE COVERED (KM)
2	8200
3	13000
7	31200
10	45800

TRAIN TEST SPLIT

ACTIVITY

Please go ahead and build a regression model using x_train as inputs and y_train as targets. You can then use a scatter plot to graph the predicted y values vs. the observed y values (contained in y_train). Visually inspect if your model did a good job.