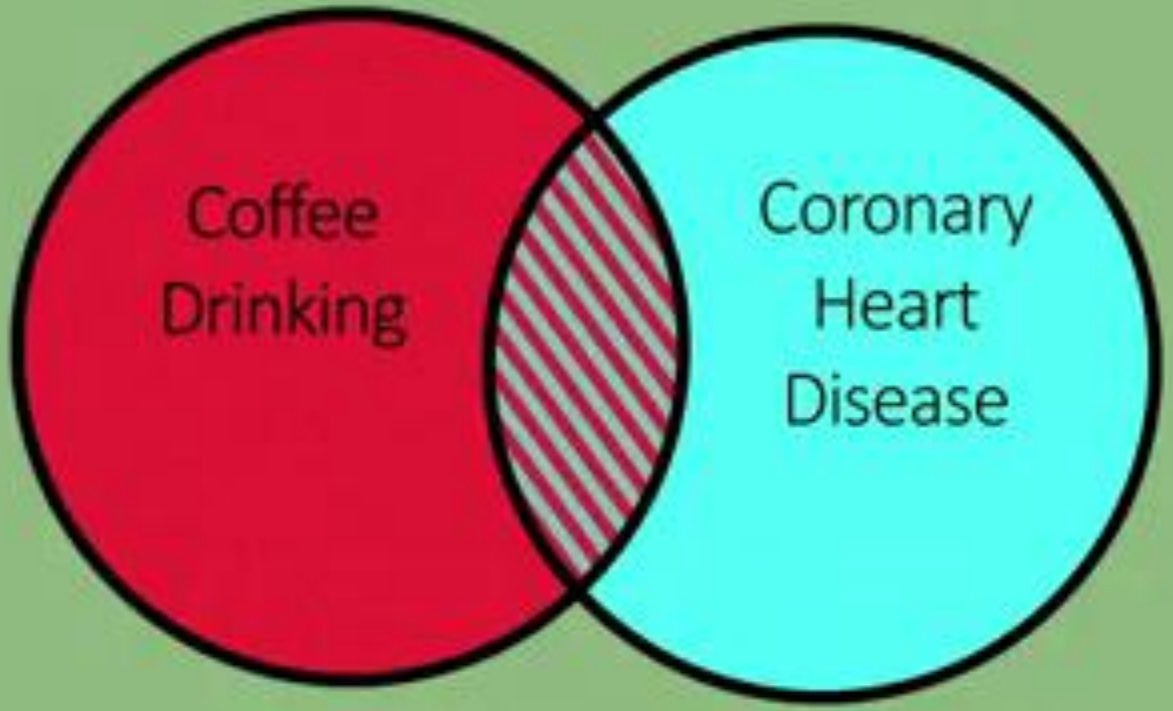
Module 2
Lesson 1 - More on Confounding

WESLEYAN
UNIVERSITY


Are there other
variables that might
explain or confound
the relationship?

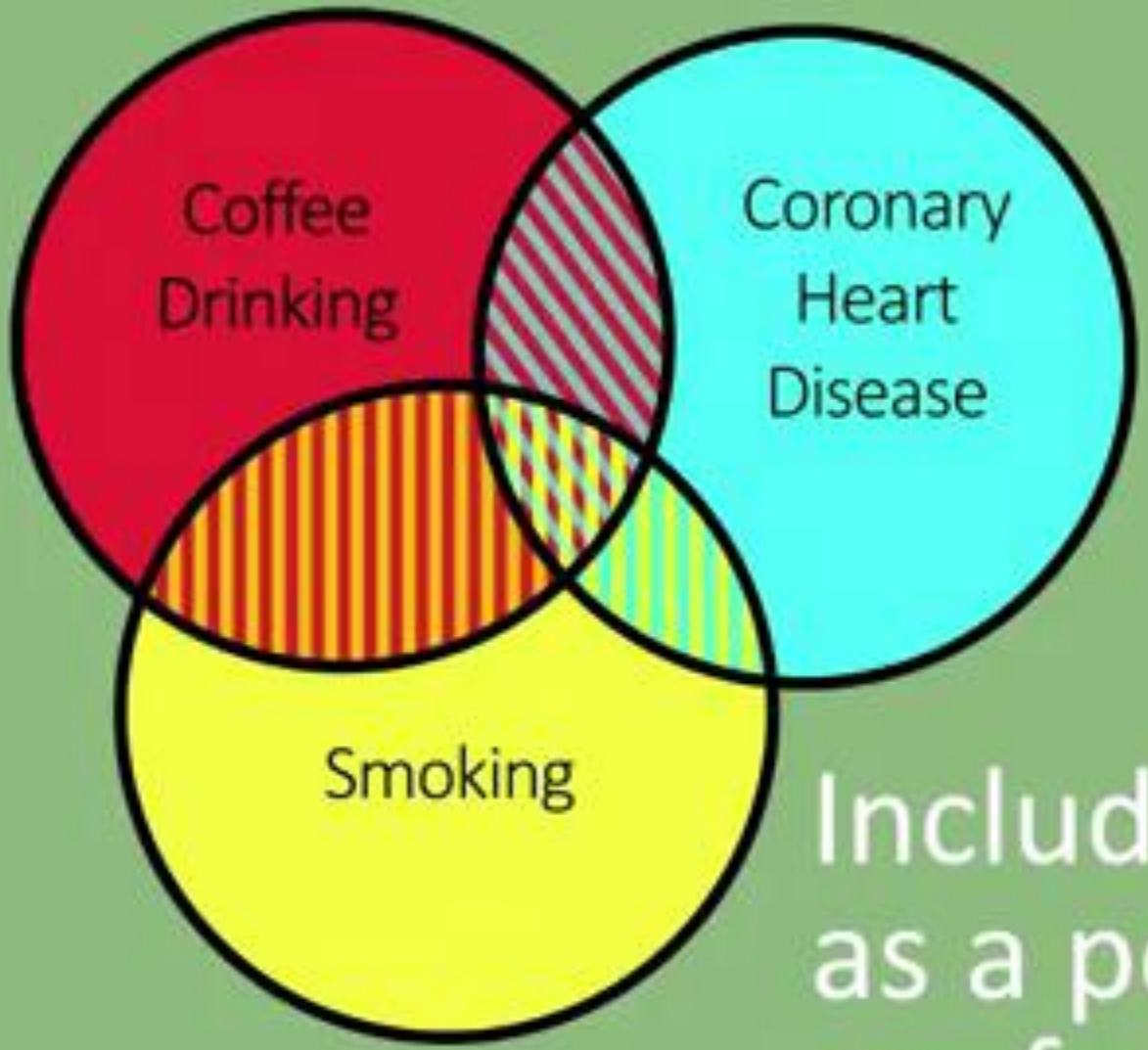
Explanatory Variable

Response Variable



Explanatory Variable

Response Variable

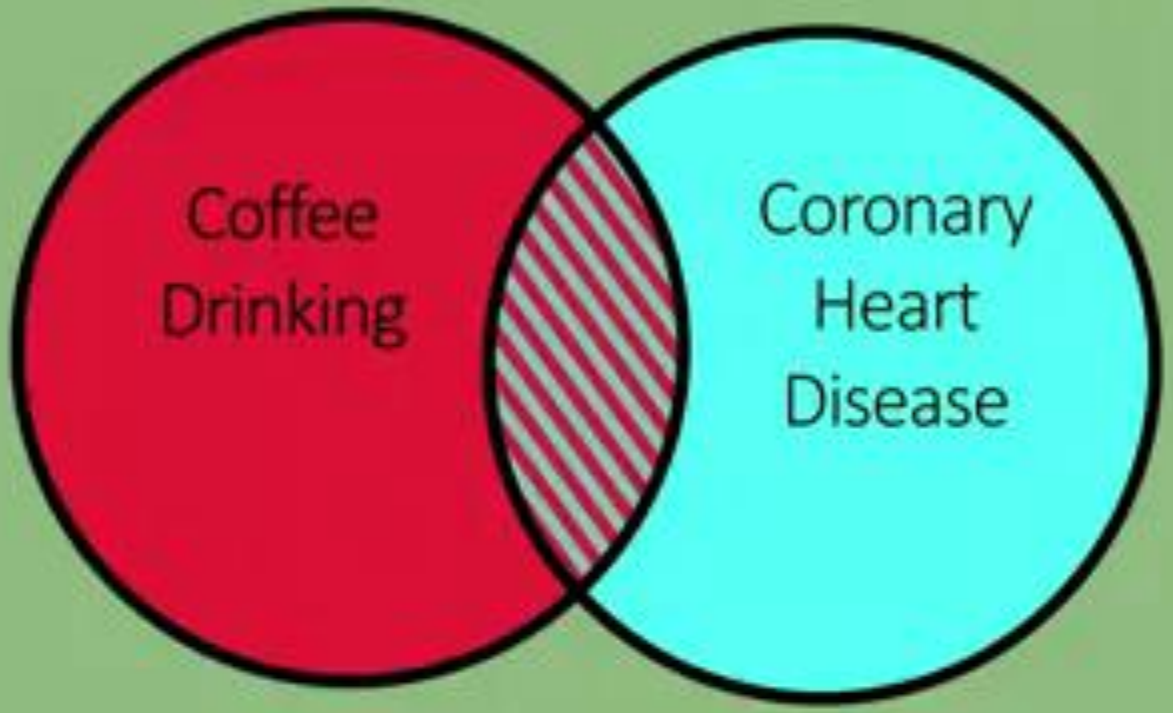


Include smoking as a possible confounder.

-Confounder
-Control Variable
-Covariate
-Third Variable
-Lurking Variable

Explanatory Variable

Response Variable

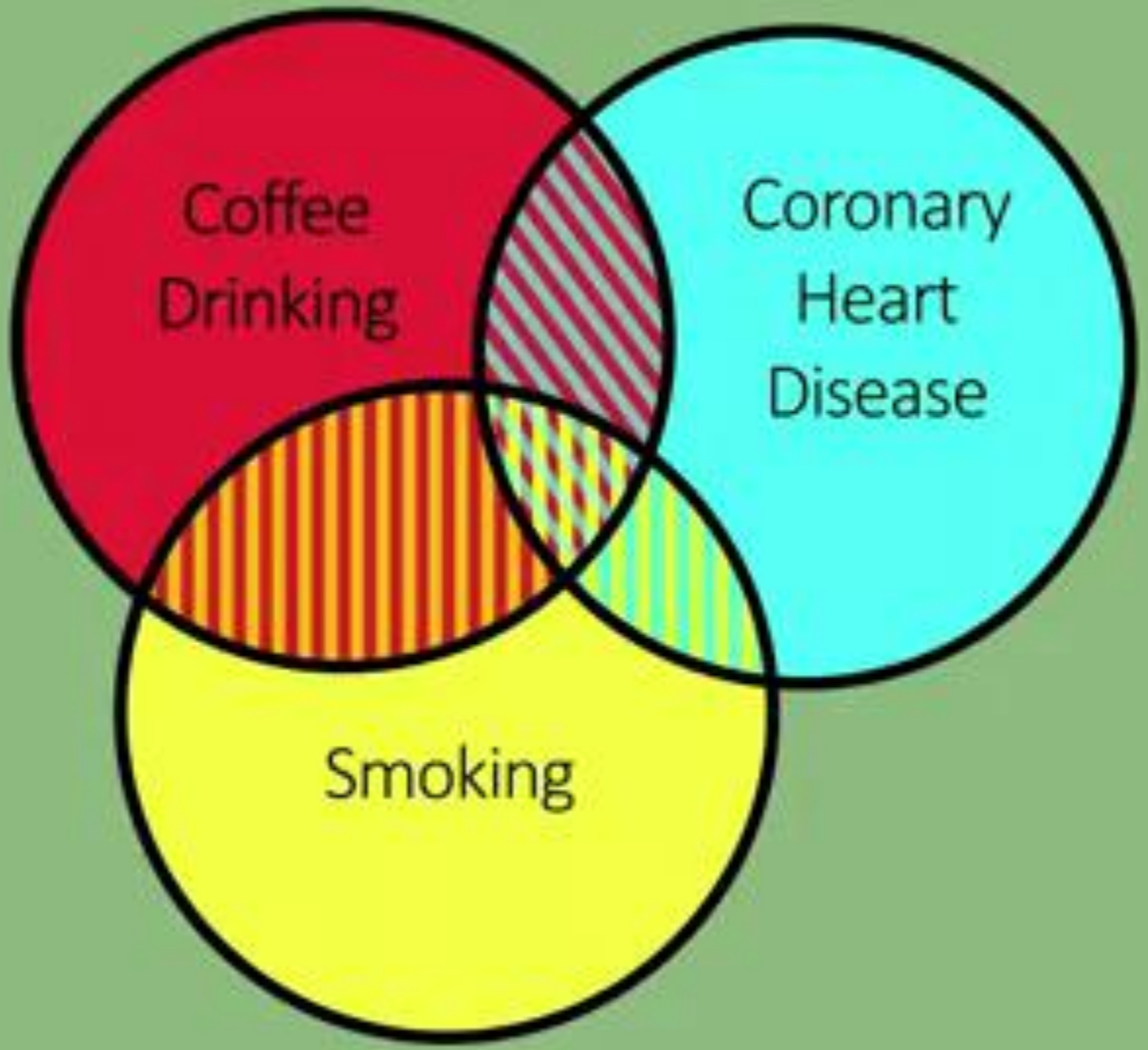


Explanatory Variable

Response Variable

Explanatory Variable

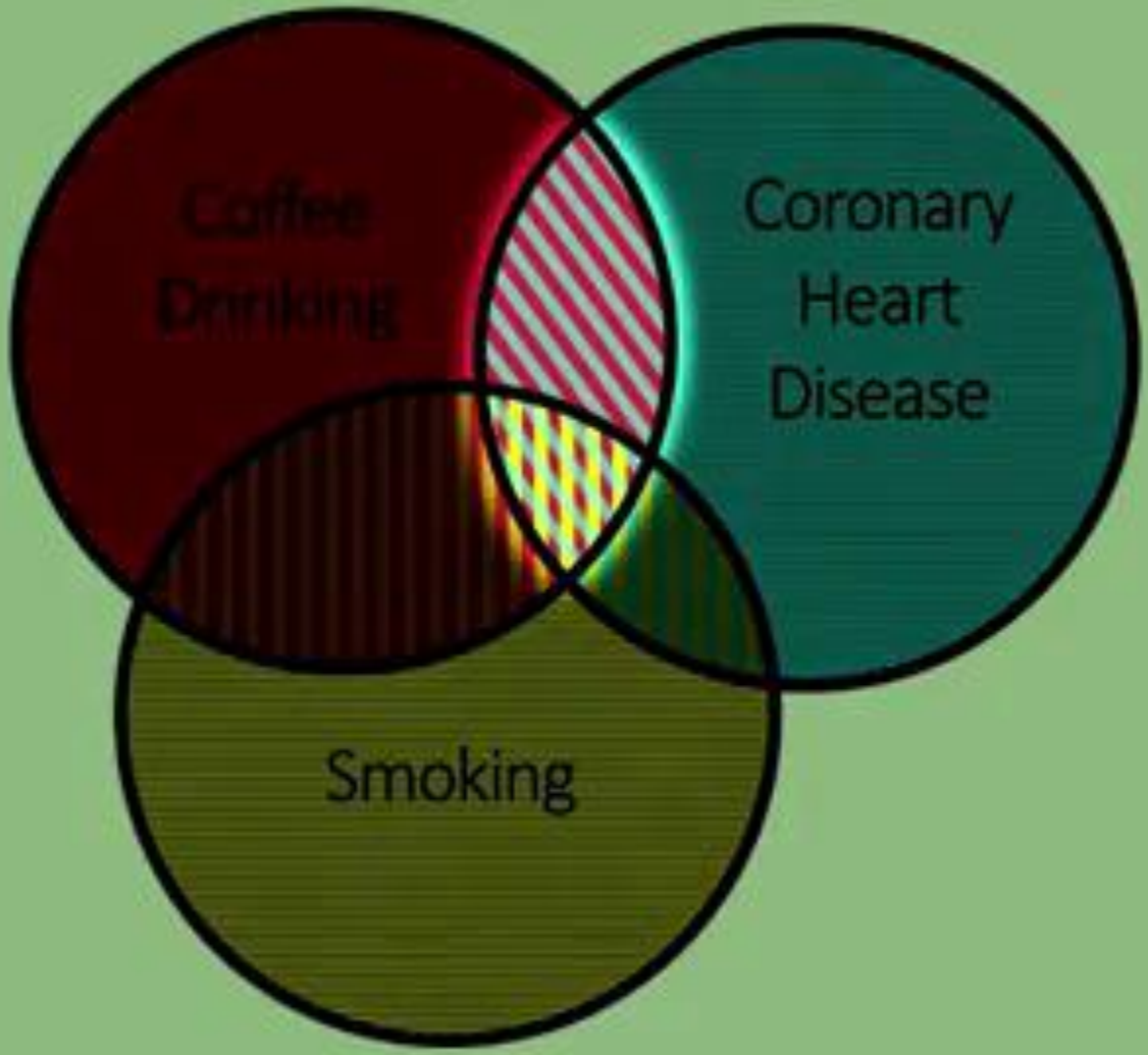
Response Variable



Confounding Variable

Explanatory Variable

Response Variable



Confounding Variable

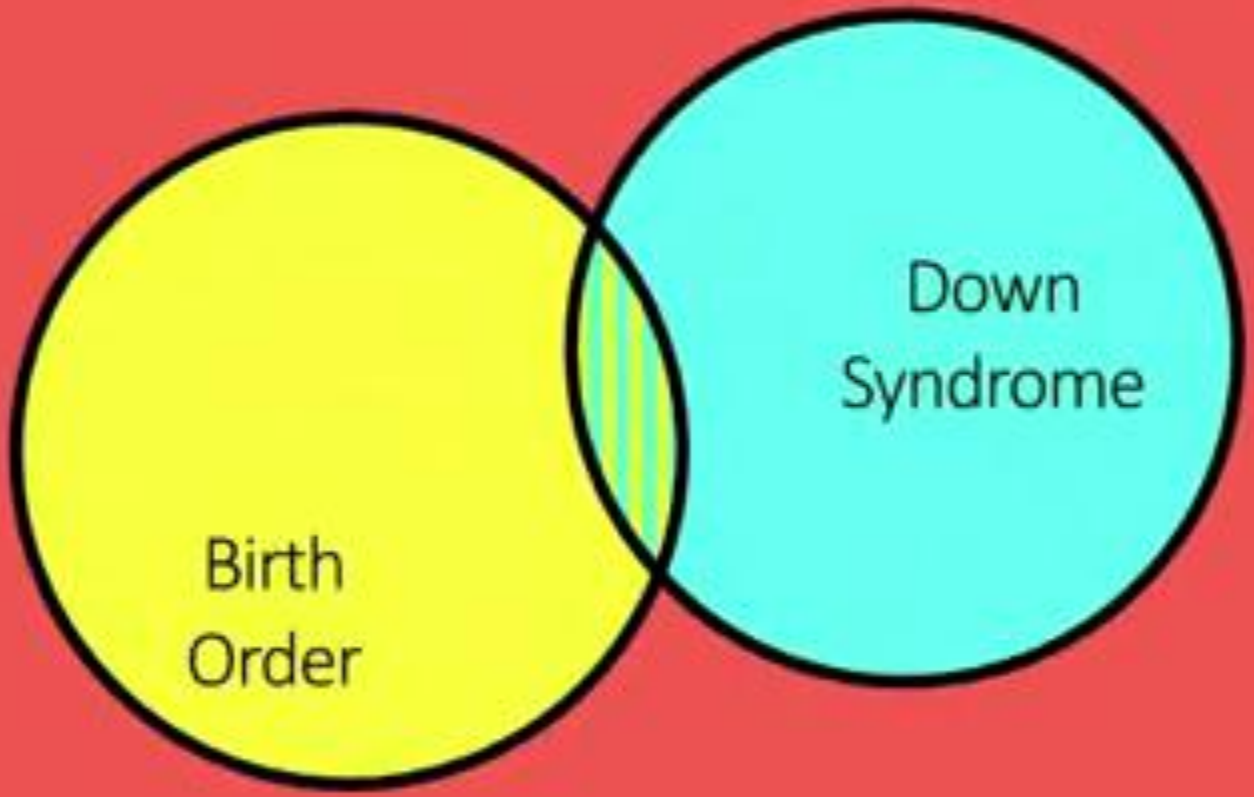
Two Types of Multivariate Model

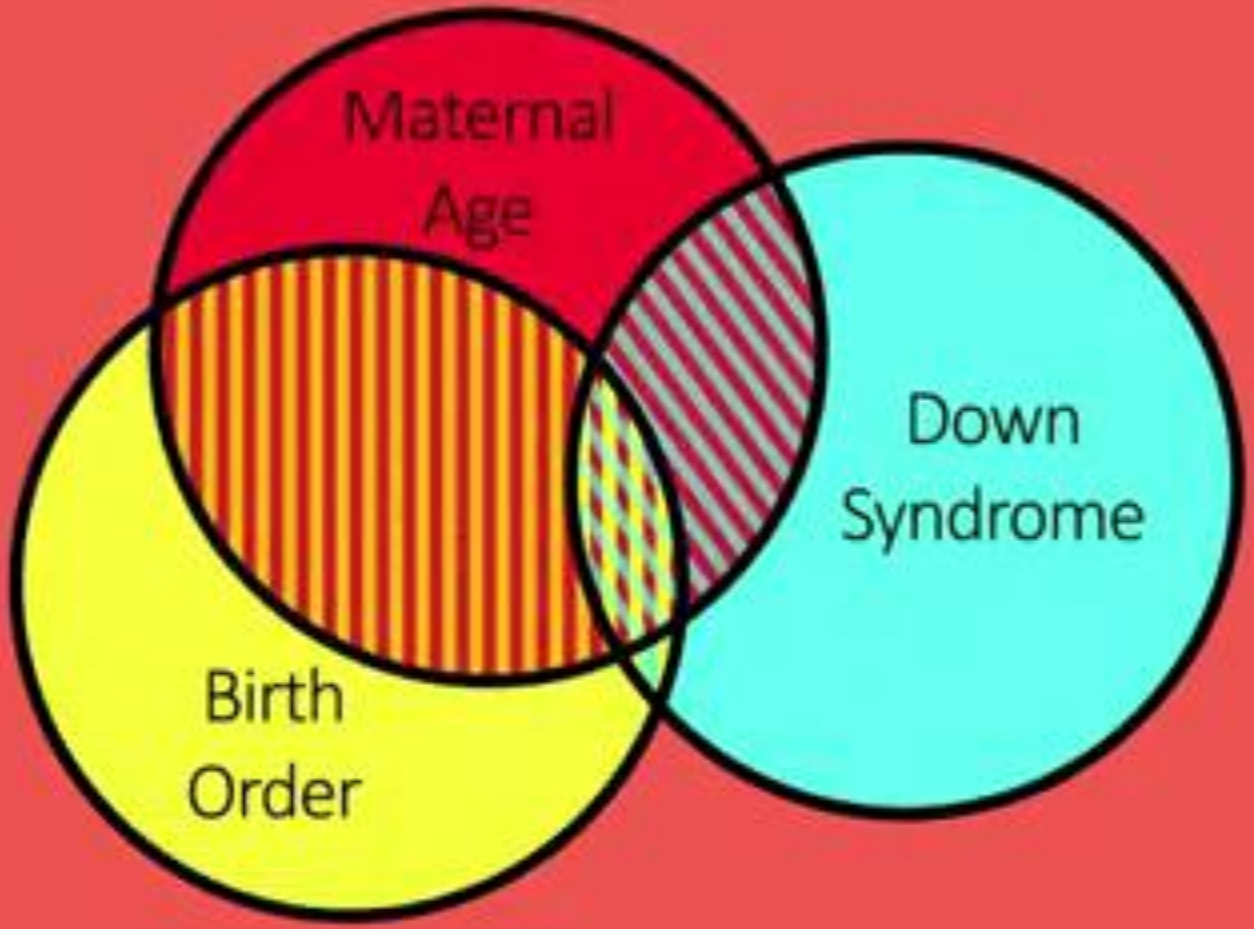
MULTIPLE REGRESSION

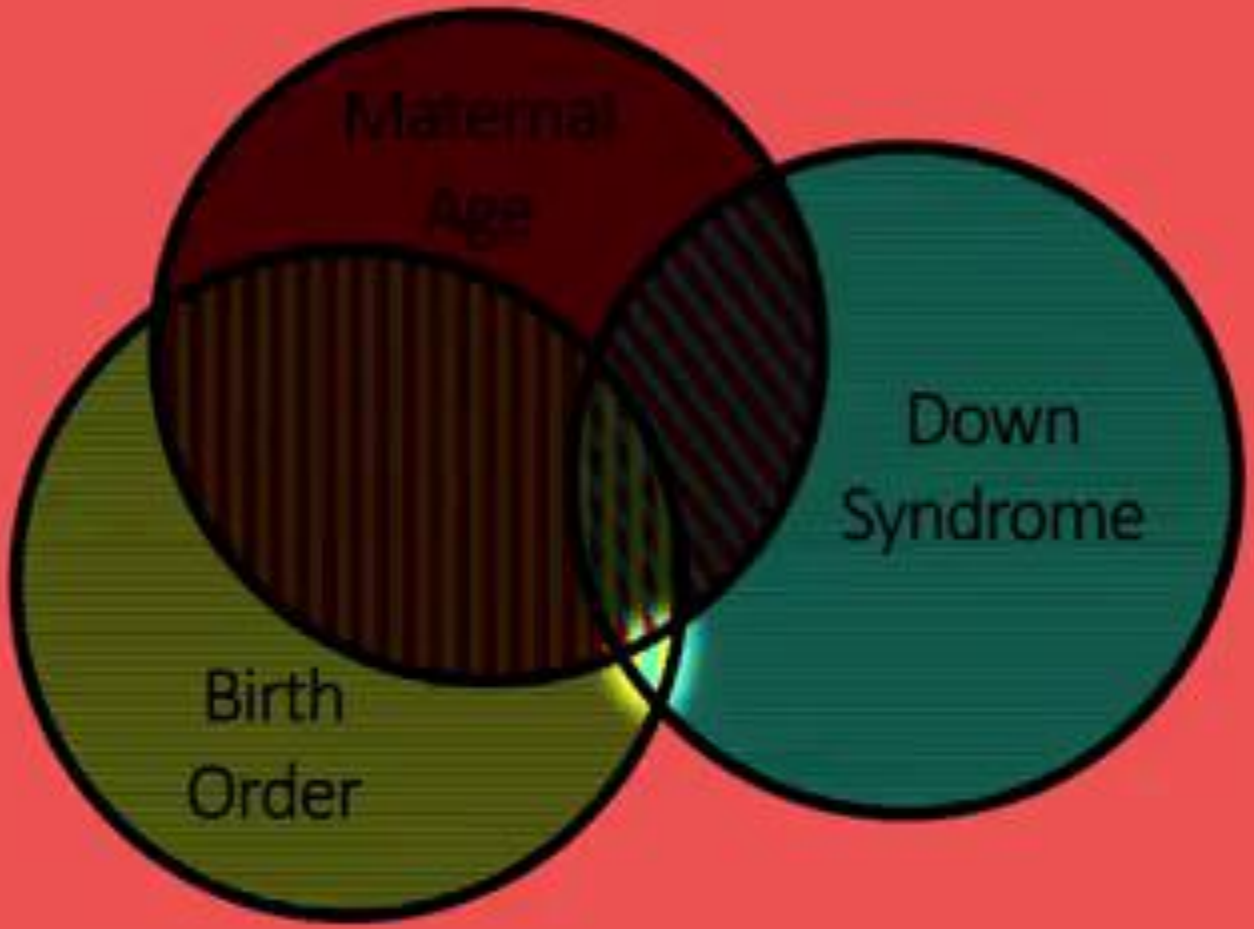
Quantitative Response Variable

Two Types of Multivariate Model

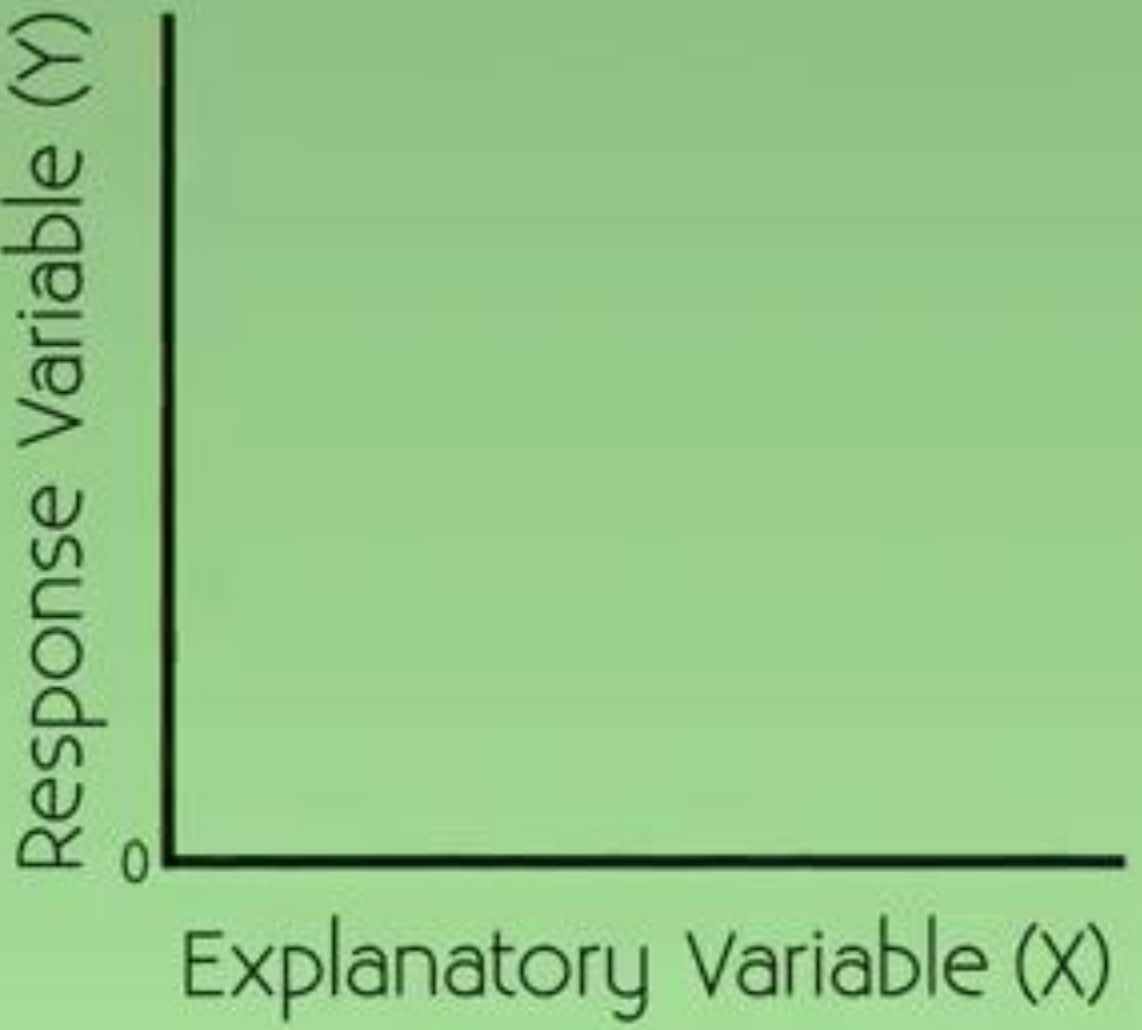
MULTIPLE REGRESSION	LOGISTIC REGRESSION
Quantitative Response Variable	Binary Response Variable


Down
Syndrome
Birth
Order


Maternal Age
Down Syndrome
Birth Order


Maternal Age
Down Syndrome
Birth Order


How does a
linear regression
analysis work?


Response Variable (Y)
0
Explanatory Variable (X)

Object inspector | Variable explorer | File explorer

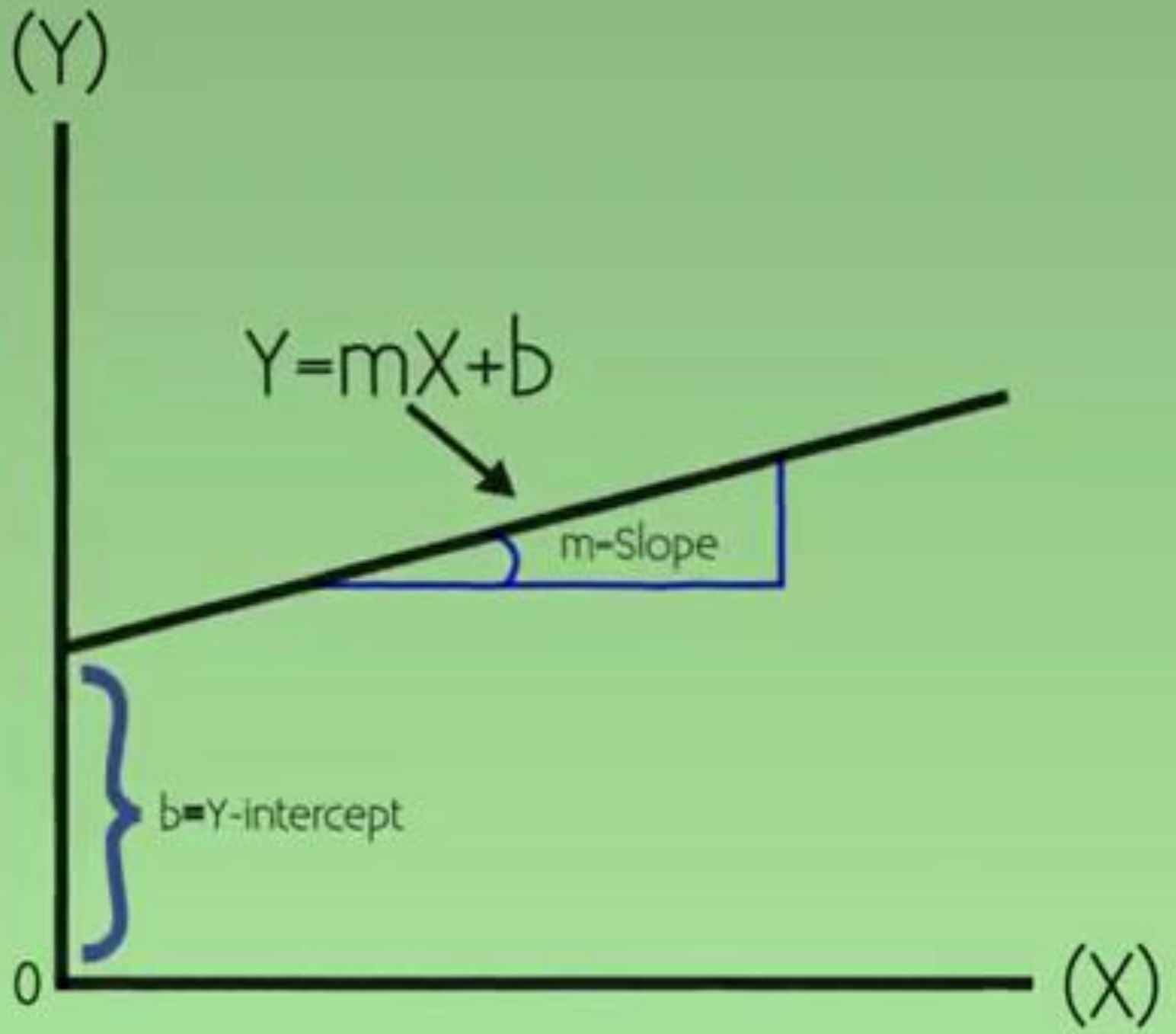
IPython console

Console 1/A

Is the rate of urbanization associated with the rate of people who use the internet?

In [3]:

Console | IPython console | History log


(Y)
Y=mX+b
m-Slope
b=Y-intercept
0
(X)

Module 2

Lesson 2 - Testing a Basic Linear Regression Model with Python

WESLEYAN
UNIVERSITY

```
modelname = smf.ols(formula='QUANT_RESPONSE ~ QUANT_EXPLANATORY',
        data=dataframe).fit()
print (modelname.summary())
```

GAPMINDER_polynomial_regression_and_diagnostic_plots.py | NESARC_data_linear_regression_modeling.py

```
# -*- coding: utf-8 -*-
"""
Created on Wed Oct 14 14:49:10 2015

@author: jml
"""

import pandas
import numpy
import statsmodels.api as sm
import statsmodels.formula.api as smf

data = pandas.read_csv('gapminder.csv')

# numeric variables that are read into python
# from the csv file as strings (objects) with empty cells should be
# converted back to numeric format using convert_objects function
data['internetuserate'] = data['internetuserate'].convert_objects(convert_numeric=True)
data['urbanrate'] = data['urbanrate'].convert_objects(convert_numeric=True)

print('OLS regression model for the association between urbanrate and internet use rate')
reg1 = smf.ols('internetuserate ~ urbanrate', data=data).fit()
print (reg1.summary())

```

Console 1/A

OLS Regression Results

Dep. Variable:	internetuserate	R-squared:	0.377
Model:	OLS	Adj. R-squared:	0.374
Method:	Least Squares	F-statistic:	113.7
Date:	Wed, 04 Nov 2015	Prob (F-statistic):	4.56e-21
Time:	15:21:51	Log-Likelihood:	-856.14
No. Observations:	190	AIC:	1716.
Df Residuals:	188	BIC:	1723.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[95.0% Conf.	Int.]
Intercept	-4.9037	4.115	-1.192	0.235	-13.021	3.213
urbanrate	0.7202	0.068	10.665	0.000	0.587	0.853

Omnibus:	10.750	Durbin-Watson:	2.097
Prob(Omnibus):	0.005	Jarque-Bera (JB):	10.990
Skew:	0.574	Prob(JB):	0.00411
Kurtosis:	3.262	Cond. No.	157.

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

In [2]:

Console 1/A

OLS Regression Results

Dep. Variable:	internetuserate	R-squared:	0.377
Model:	OLS	Adj. R-squared:	0.374
Method:	Least Squares	F-statistic:	113.7
Date:	Wed, 04 Nov 2015	Prob (F-statistic):	4.56e-21
Time:	15:21:51	Log-Likelihood:	-856.14
No. Observations:	190	AIC:	1716.
Df Residuals:	188	BIC:	1723.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[95.0% Conf.	Int.]
Intercept	-4.9037	4.115	-1.192	0.235	-13.021	3.213
urbanrate	0.7202	0.068	10.665	0.000	0.587	0.853

Omnibus:	10.750	Durbin-Watson:	2.097
Prob(Omnibus):	0.005	Jarque-Bera (JB):	10.990
Skew:	0.574	Prob(JB):	0.00411
Kurtosis:	3.262	Cond. No.	157.

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

In [2]:

Console 1/A

OLS Regression Results

Dep. Variable:	internetuserate	R-squared:	0.377
Model:	OLS	Adj. R-squared:	0.374
Method:	Least Squares	F-statistic:	113.7
Date:	Wed, 04 Nov 2015	Prob (F-statistic):	4.56e-21
Time:	15:21:51	Log-Likelihood:	-856.14
No. Observations:	190	AIC:	1716.
Df Residuals:	188	BIC:	1723.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[95.0% Conf.	Int.]
Intercept	-4.9037	4.115	-1.192	0.235	-13.021	3.213
urbanrate	0.7202	0.068	10.665	0.000	0.587	0.853

Omnibus:	10.750	Durbin-Watson:	2.097
Prob(Omnibus):	0.005	Jarque-Bera (JB):	10.990
Skew:	0.574	Prob(JB):	0.00411
Kurtosis:	3.262	Cond. No.	157.

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

In [2]:

Dependent Variable

Population y Intercept

Population Slope Coeficient

$$y = \beta_0 + \beta_1 x$$

Explanatory Variable

Variables Related
in Some Meaningful Way

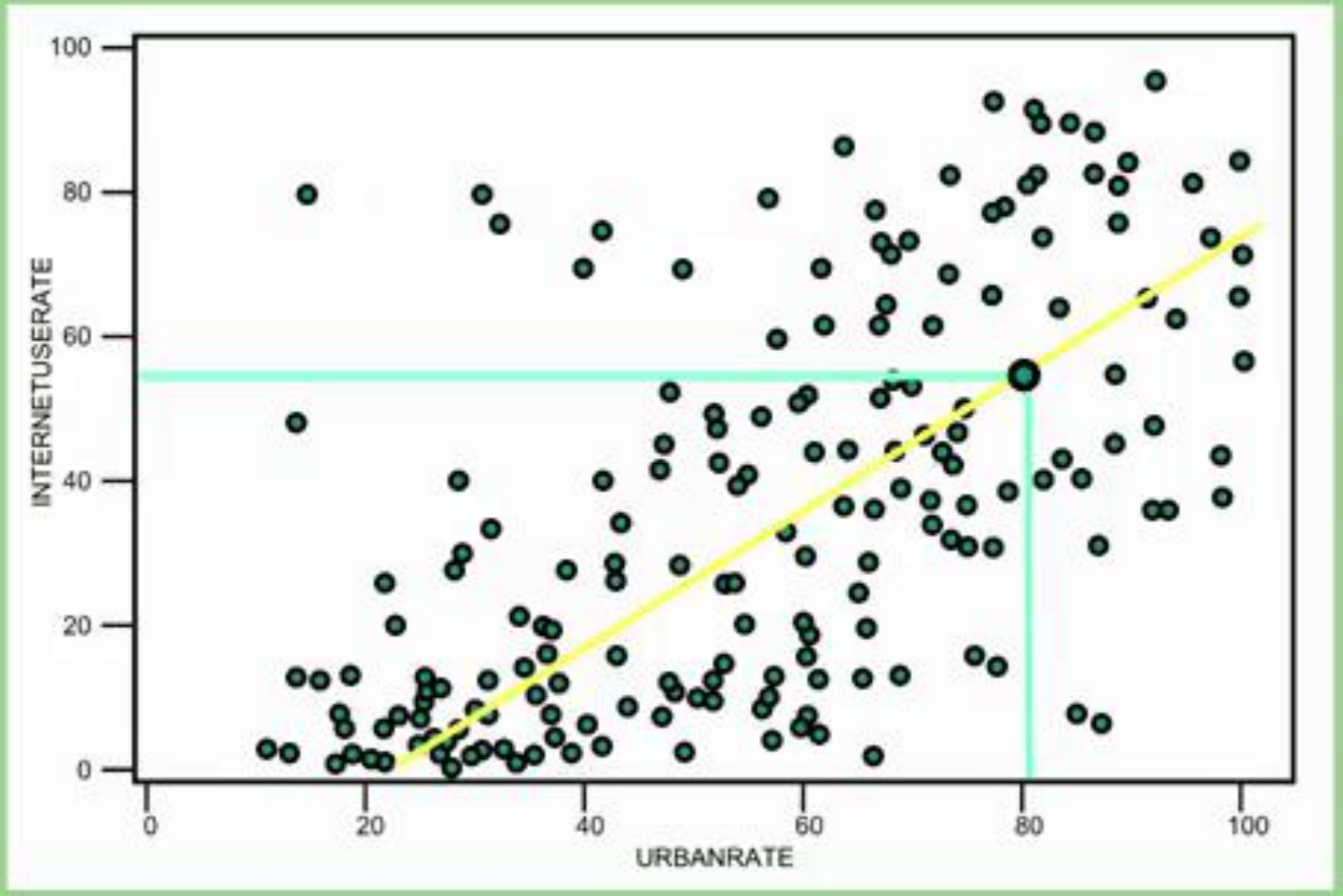
$$y = \beta_0 + \beta_1 x$$

$$\hat{y} = \beta_0 + \beta_1 x$$

β_0 =-4.90

β_1 =0.72

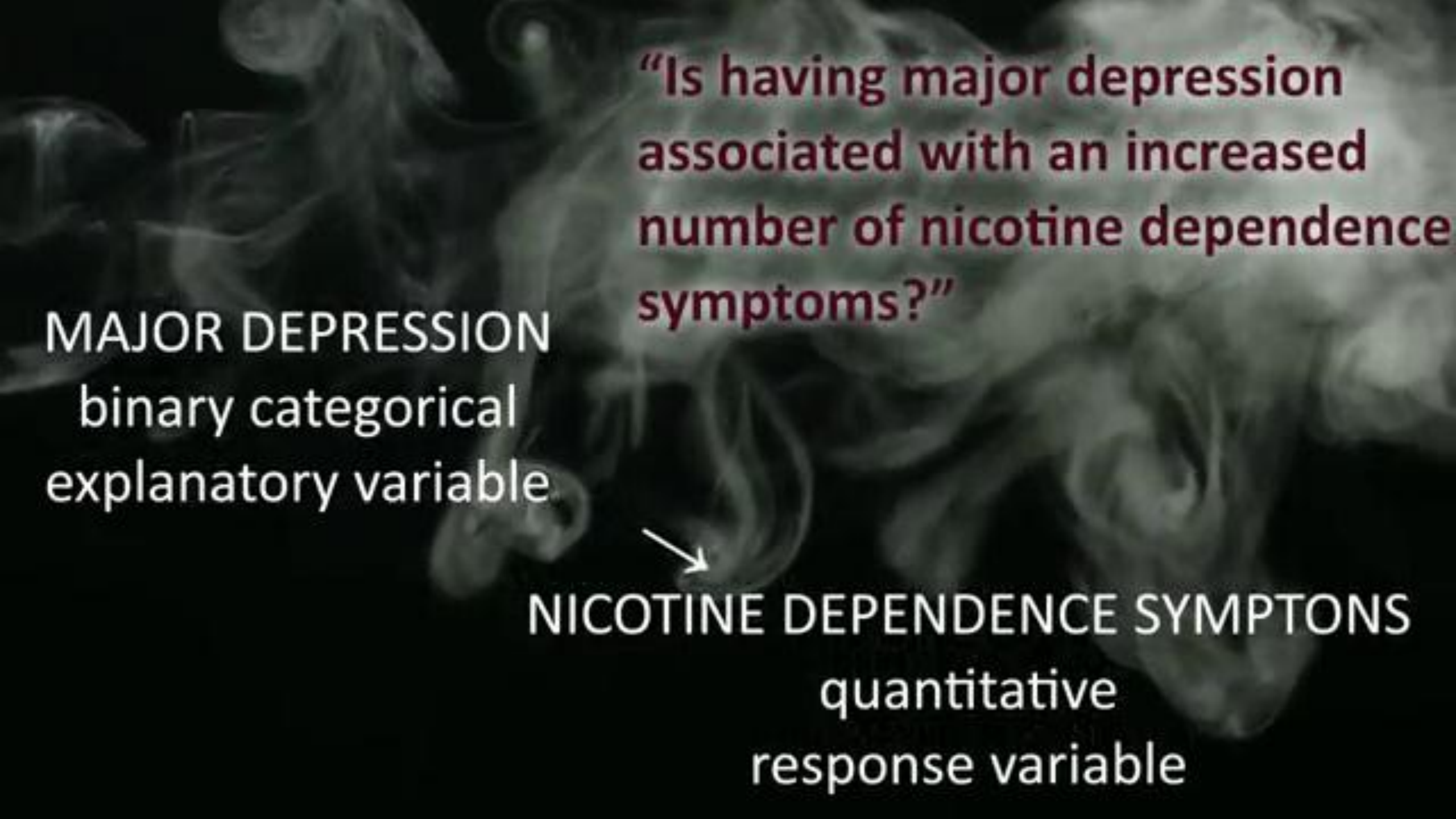
URBANRATE = 80


100
80
60
40
20
0
INTERNETUSERATE
0
20
40
60
80
100
URBANRATE

Module 2
Lesson 3 - Categorical Explanatory Variables

WESLEYAN
UNIVERSITY


Relationship Between Depression
and Number of
Nicotine Dependence Symptoms


"Is having major depression associated with an increased number of nicotine dependence symptoms?"
MAJOR DEPRESSION
binary categorical
explanatory variable
NICOTINE DEPENDENCE SYMPTONS
quantitative
response variable

NDSymptoms = 2.19 + 1.36 (majordeplife)

Without Depression:

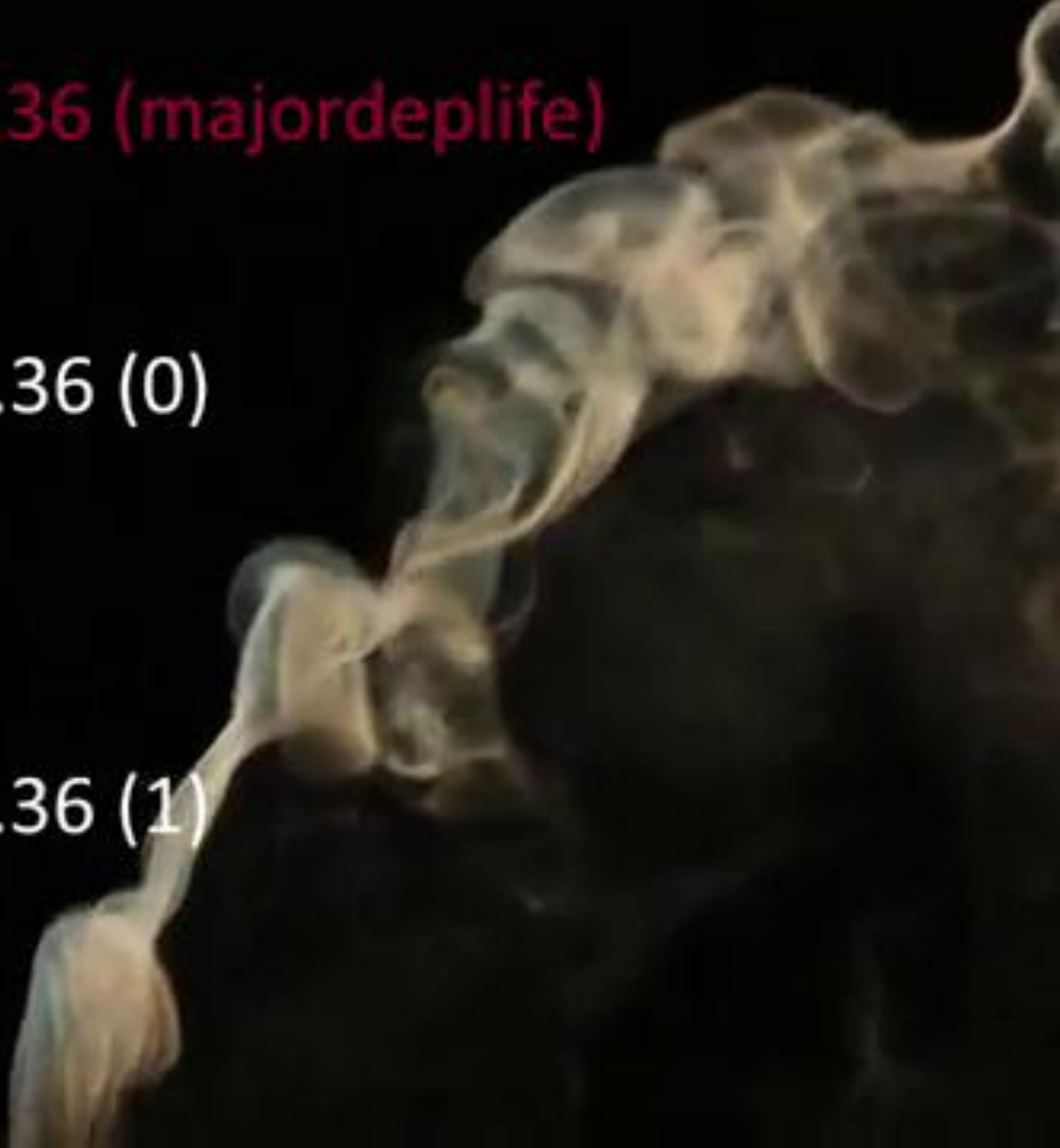
NDSymptoms = 2.19 + 1.36 (0)

NDSymptoms = 2.19

With Depression:

NDSymptoms = 2.19 + 1.36 (1)

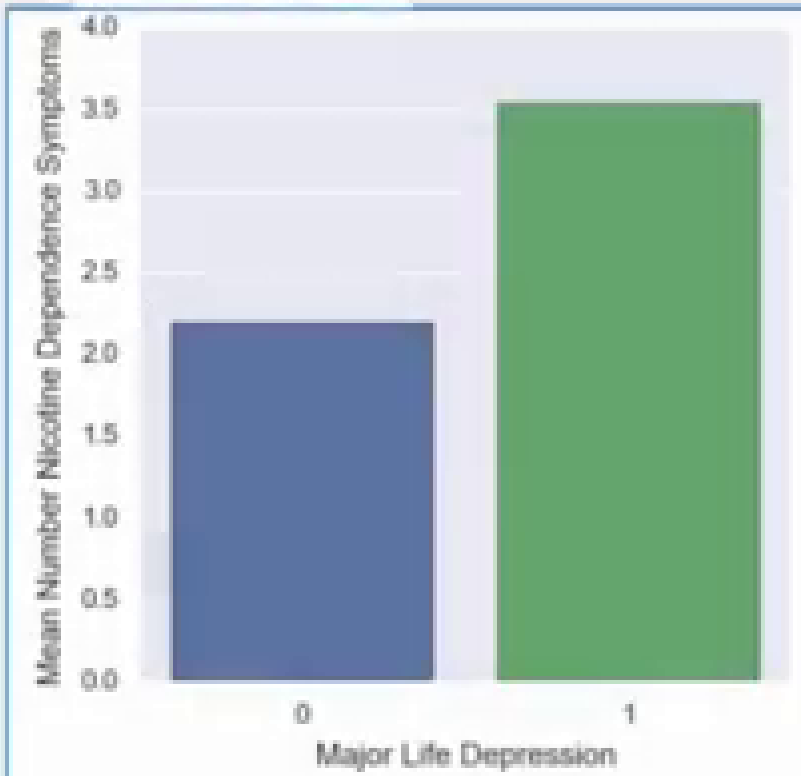
NDSymptoms = 3.55



Object inspector | Variable explorer | File explorer

IPython console

Console 1/A




```
seaborn.factorplot(x="MAJORDEPLIFE",y="NDSymptoms",
                   data=sub2, kind="bar", ci-None)
plt.xlabel('Major Life Depression')
plt.ylabel('Mean Number Nicotine Dependence')
```

In [9]:

Console | IPython console | History log

It is possible to include several explanatory or predictor variables to evaluate the independent contribution of multiple explanatory variables in predicting our response variable and to evaluate whether specific variables confound the relationship between our explanatory variable of interest and our response variable.

Module 2
Lesson 4 - Linear Regression Assumptions

WESLEYAN
UNIVERSITY

Linear Regression Assumptions

Normality

Linearity

Homoscedasticity

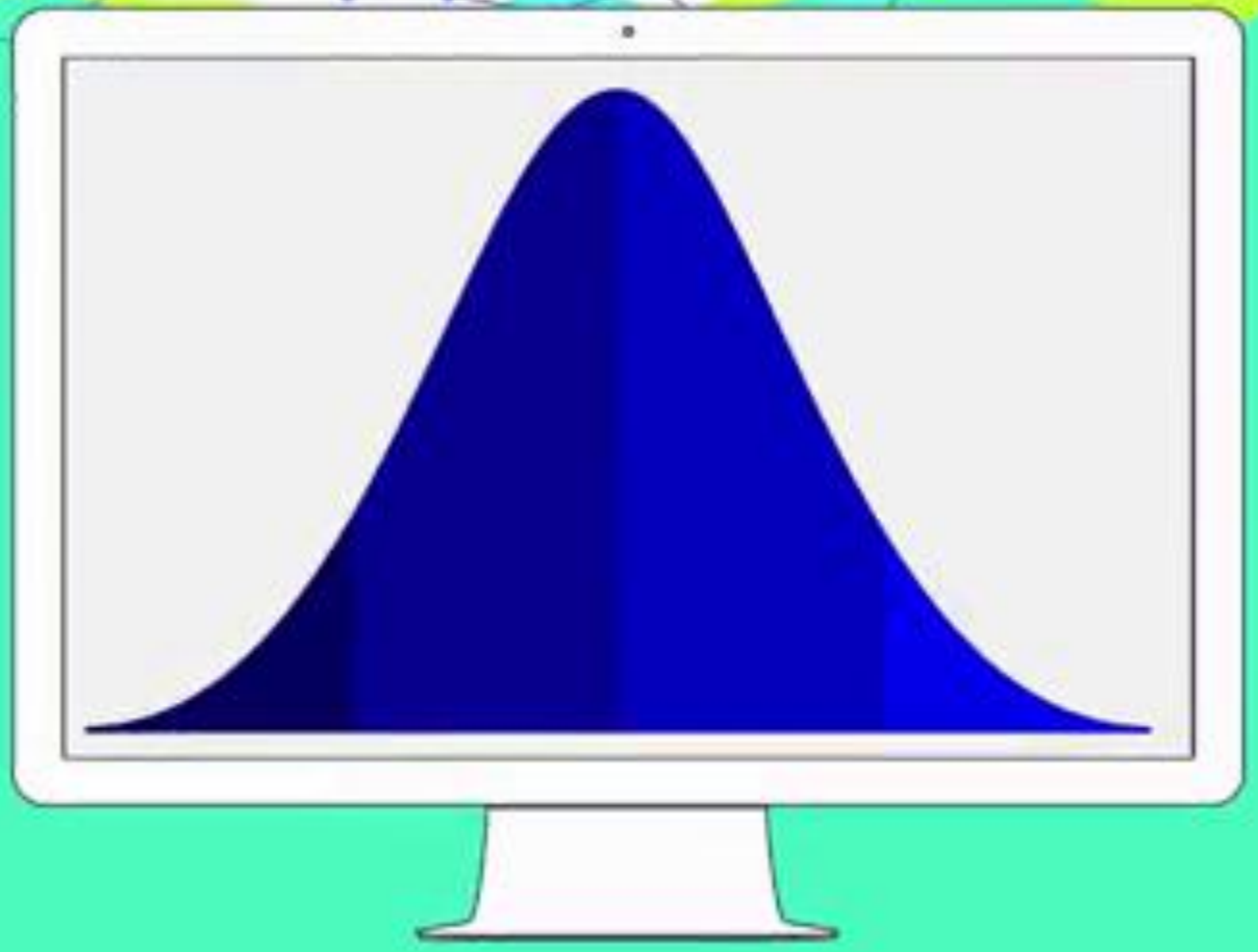
Independence

Multicollinearity

Outliers

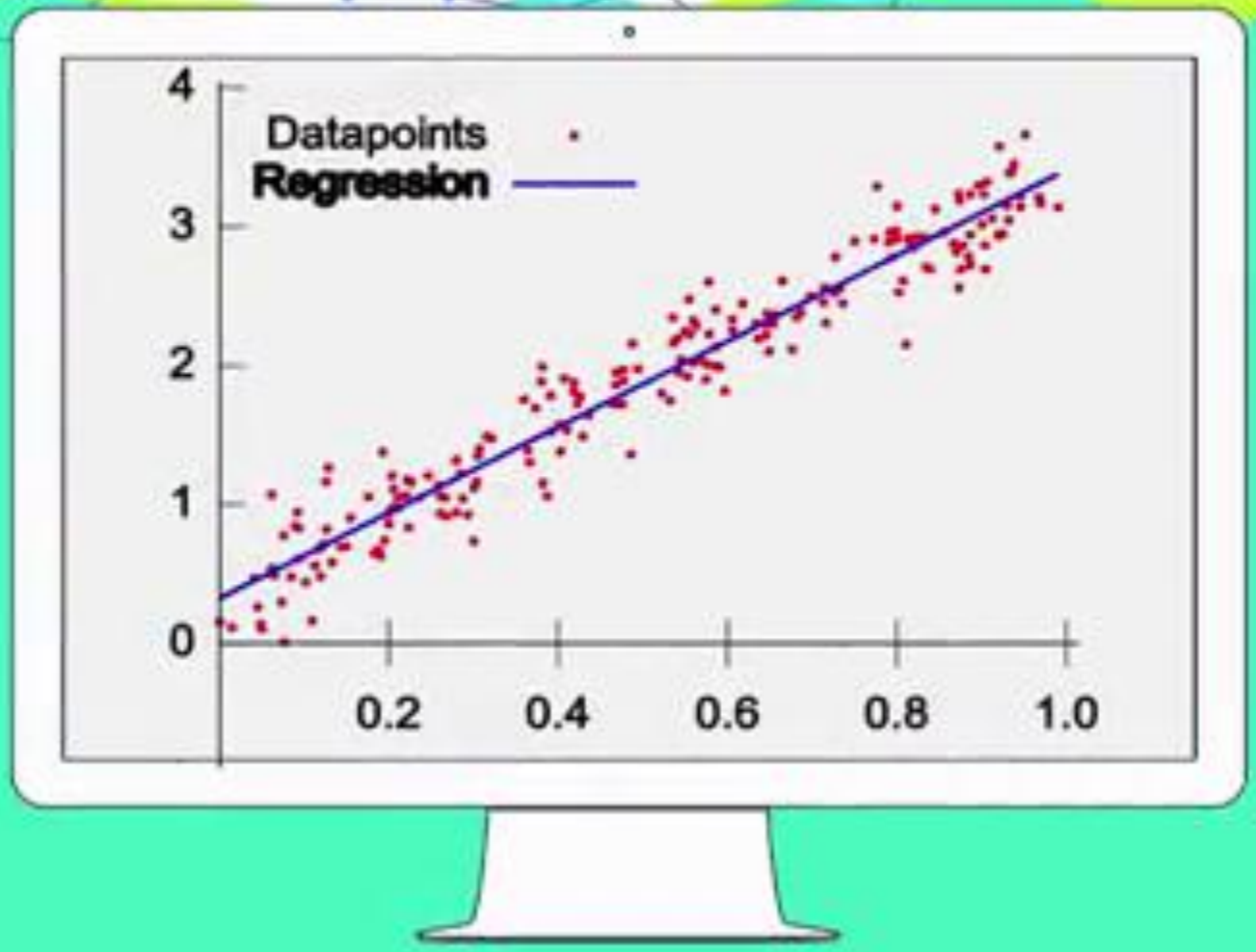
Normality

Residuals are normally distributed.



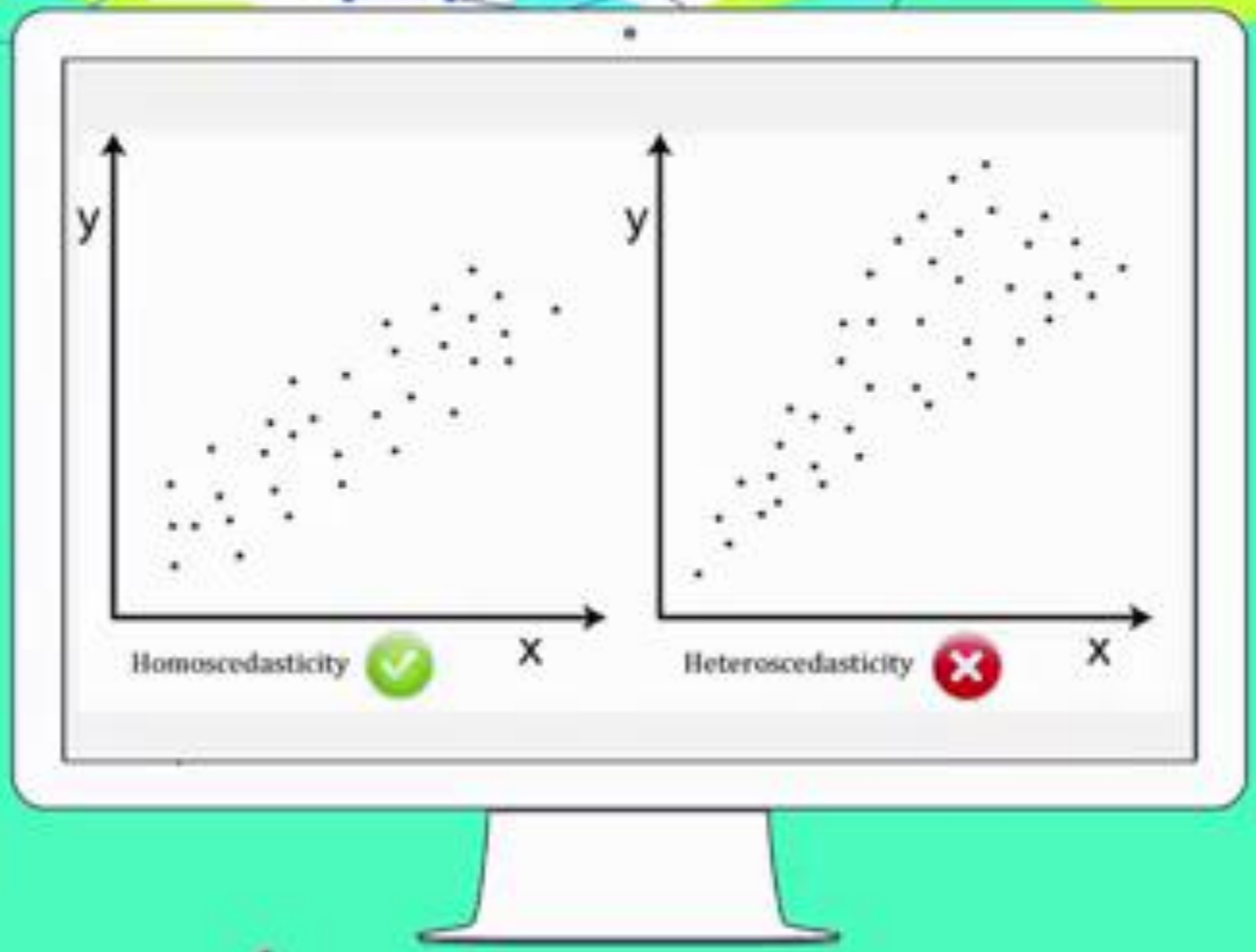
Linearty

Associations between explanatory variables and response variable are linear.



Homoscedasticity

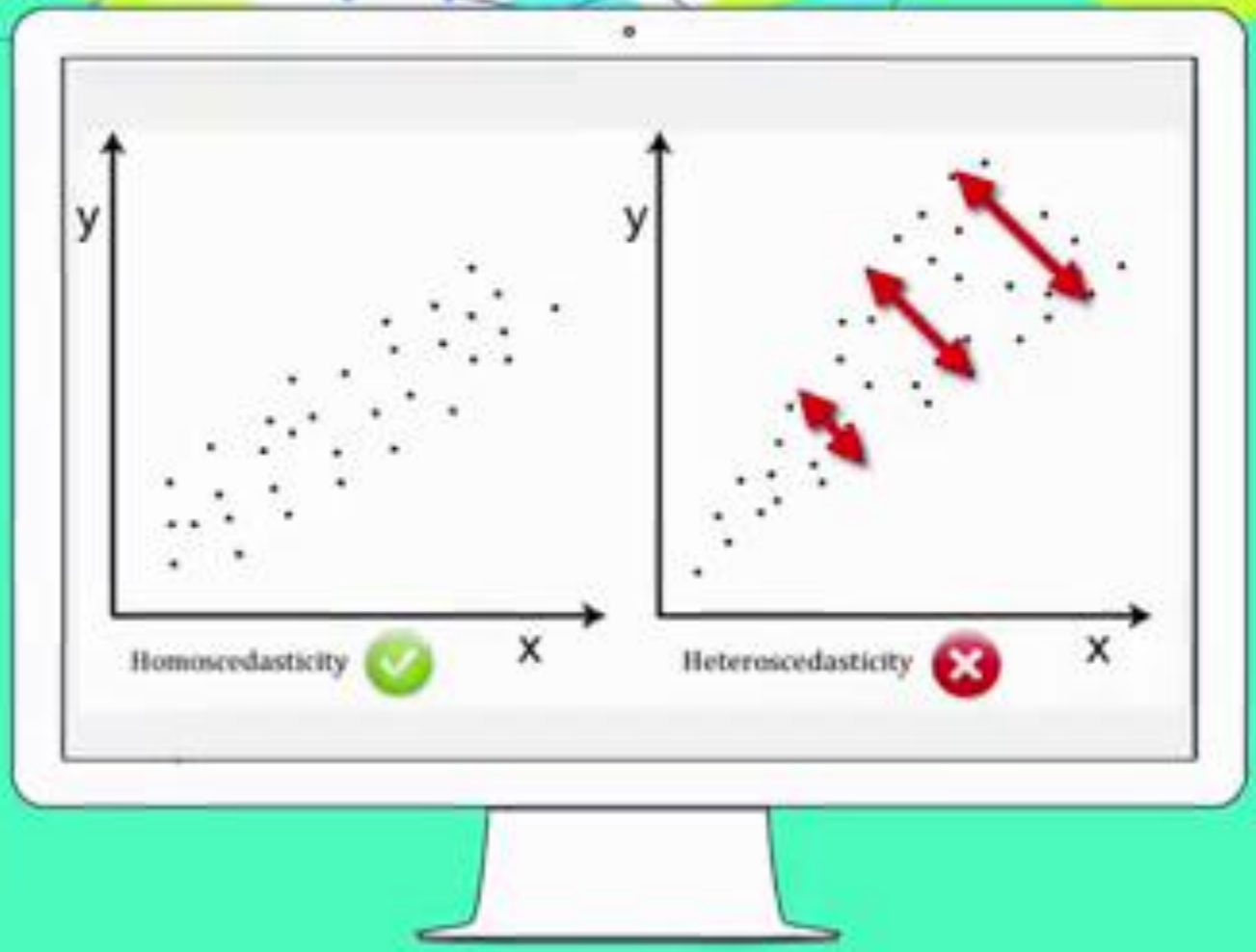
The variability in the response variable is the same at all levels of the explanatory variable



Residual = observed score (y) - predicted score ($\hat{y}$)

Homoscedasticity

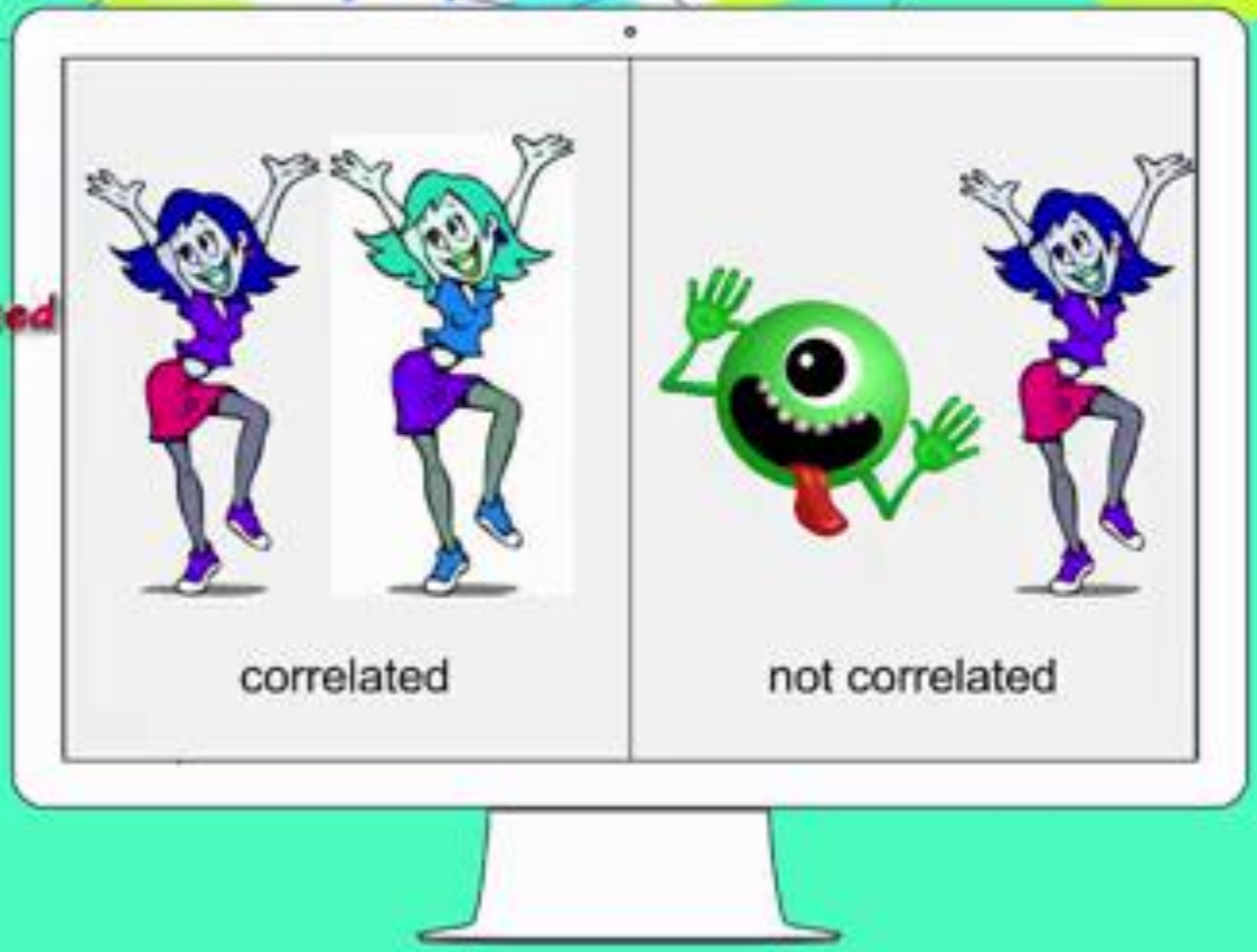
The variability in the response variable is the same at all levels of the explanatory variable



Clustered data and repeated measures data have correlated observations

Independence

Observations are not correlated with each other.



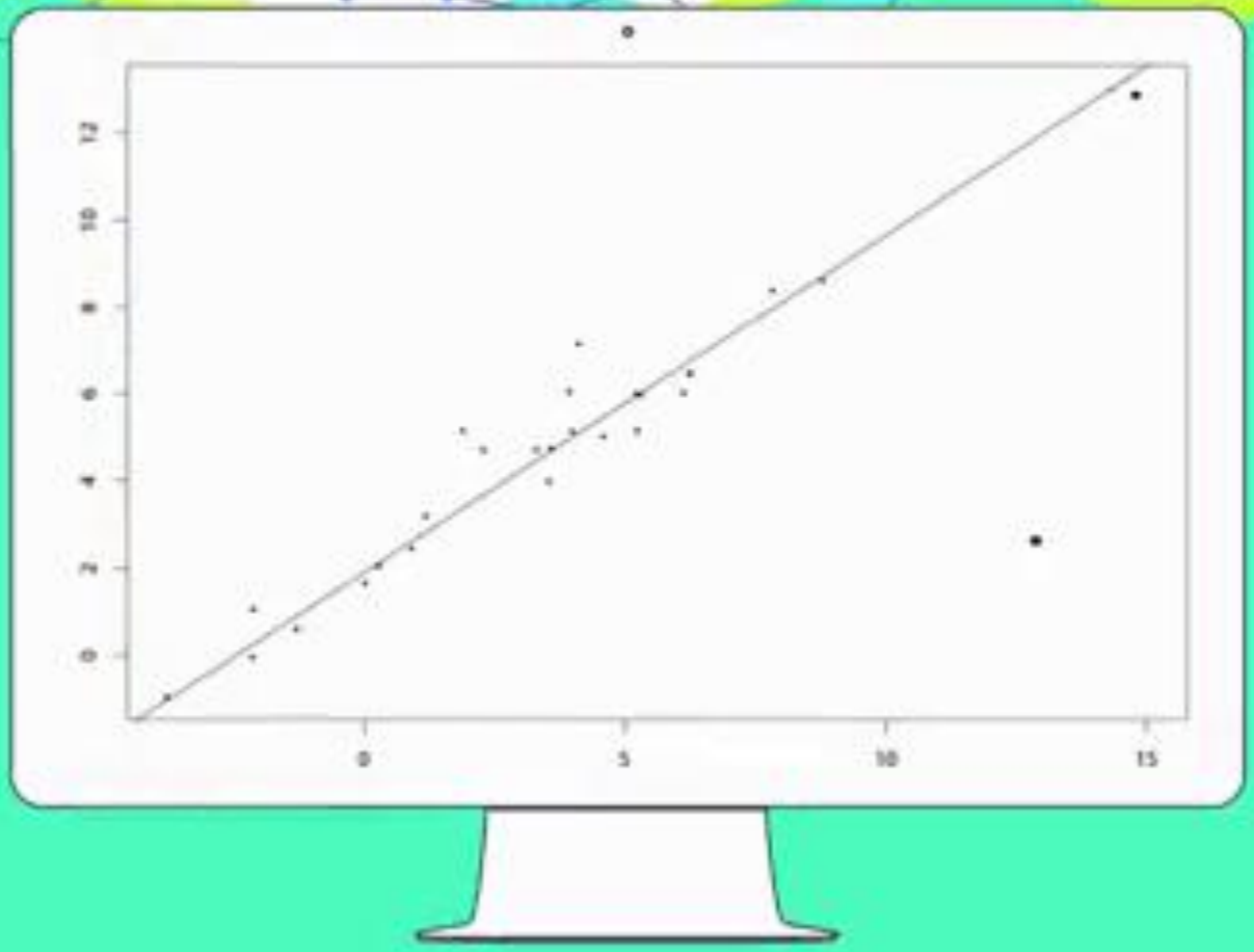

Repeated
Measures/Longitudinal Data


Hierarchical/Nested/
Clustered Data


Data structure (study design)
produces correlated observations

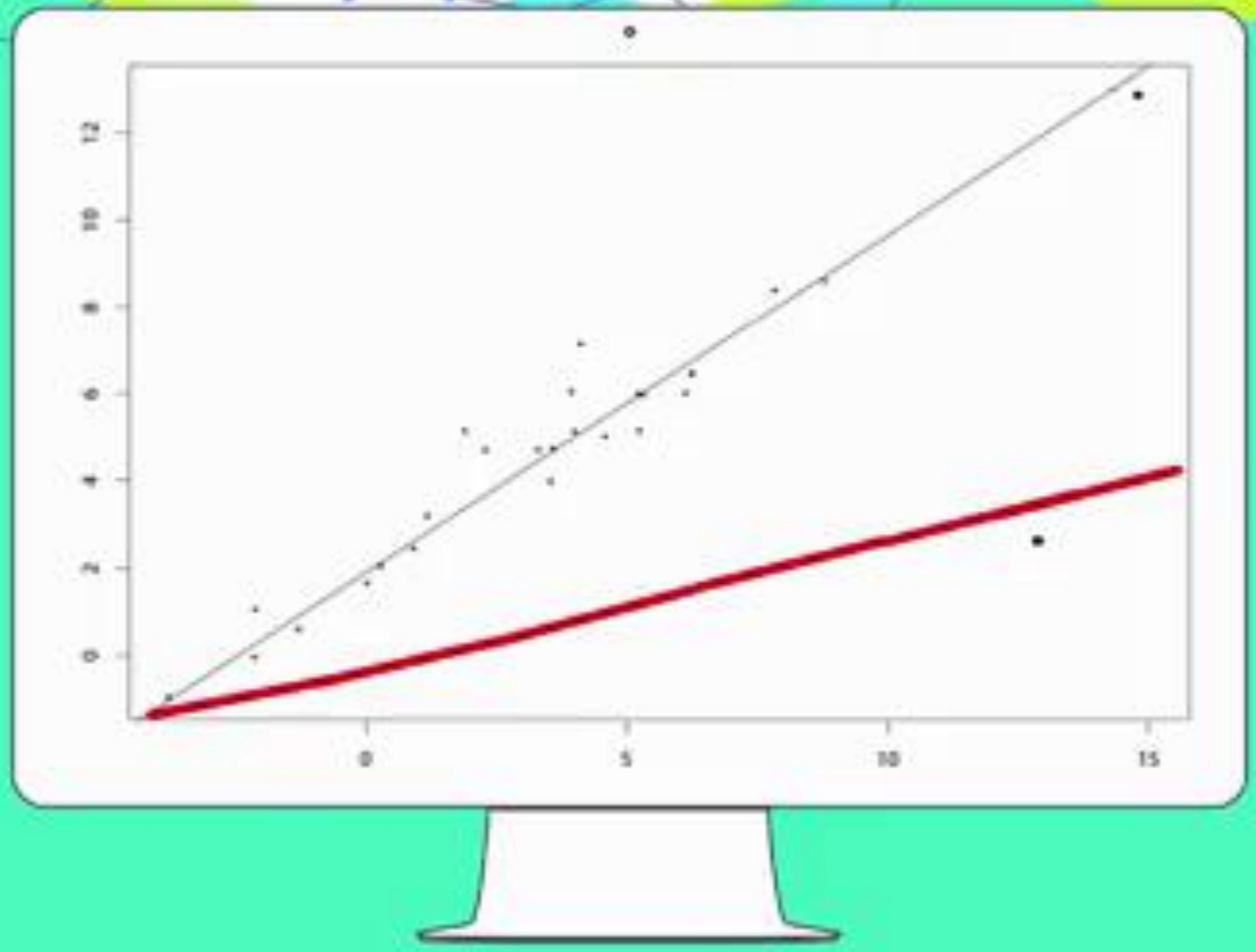
Outliers

One or more observations has extreme values on variable(s) relative to other observations.



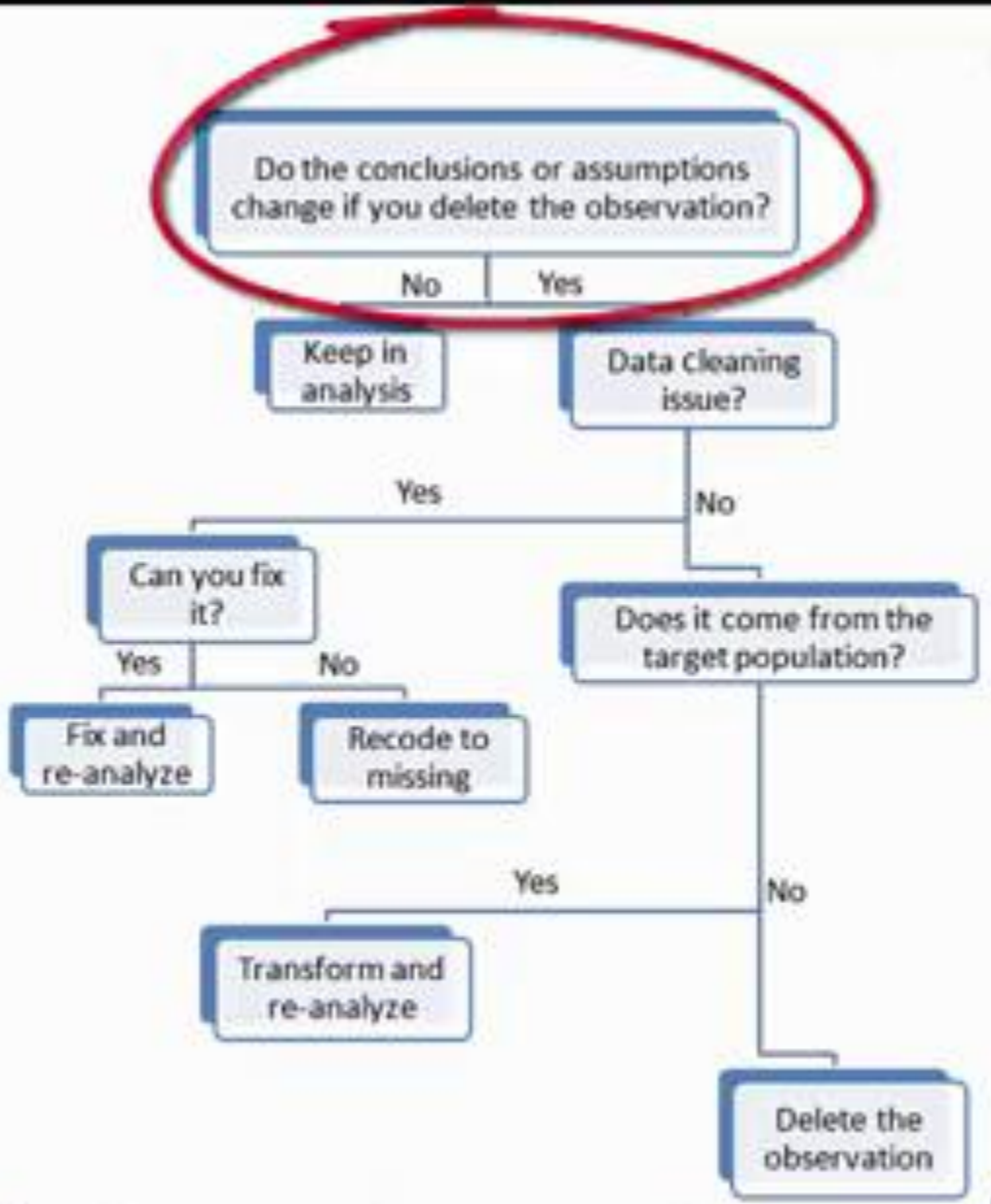
Outliers

One or more observations has extreme values on variable(s) relative to other observations.



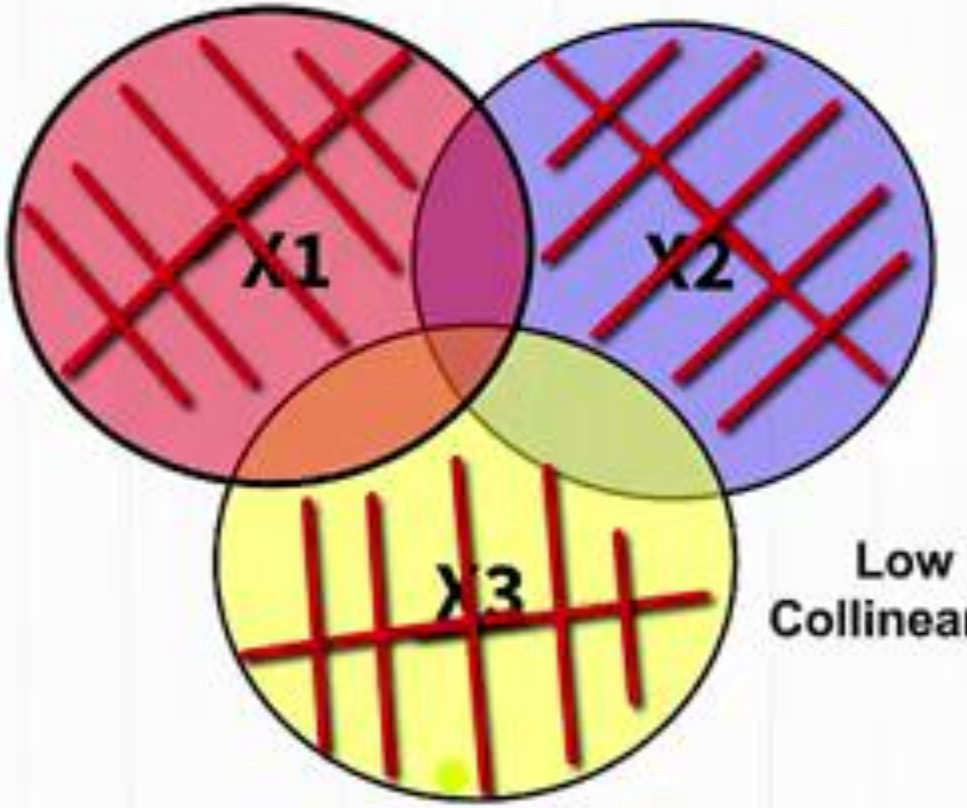
Q: WHAT DO YOU DO WITH OUTLIERS?

A: TRY USING THIS DECISION FLOWCHART

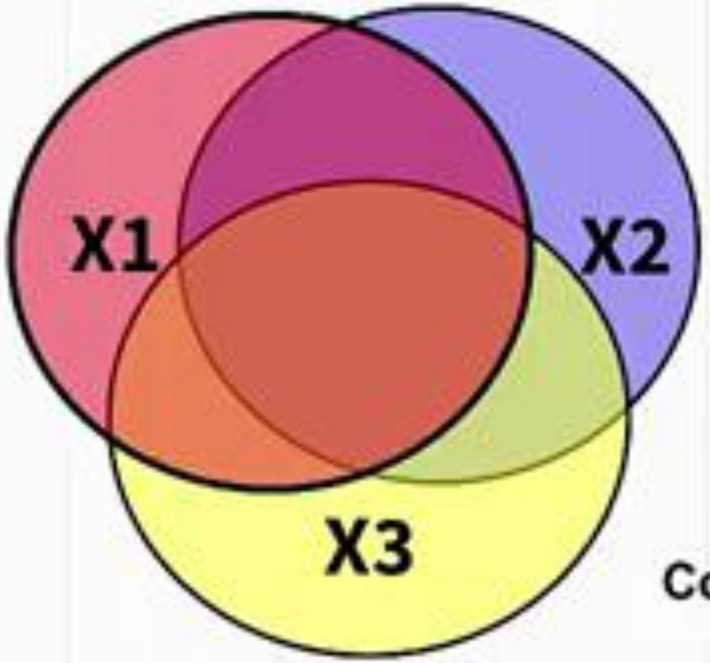


Multicollinearity

Explanatory variables are highly correlated with each other



Low Collinearity



High Collinearity

Multicollinearity

Signs:

1) highly associated explanatory variable not significant

2) negative regression coefficient that should be positive

3) taking out an explanatory variable drastically changes results

Module 2
Lesson 5 - Centering Explanatory Variables

WESLEYAN
UNIVERSITY

Age range in the data set is 18-25 years


Age = 0??

1 = Male

2 = Female

0 =  ??

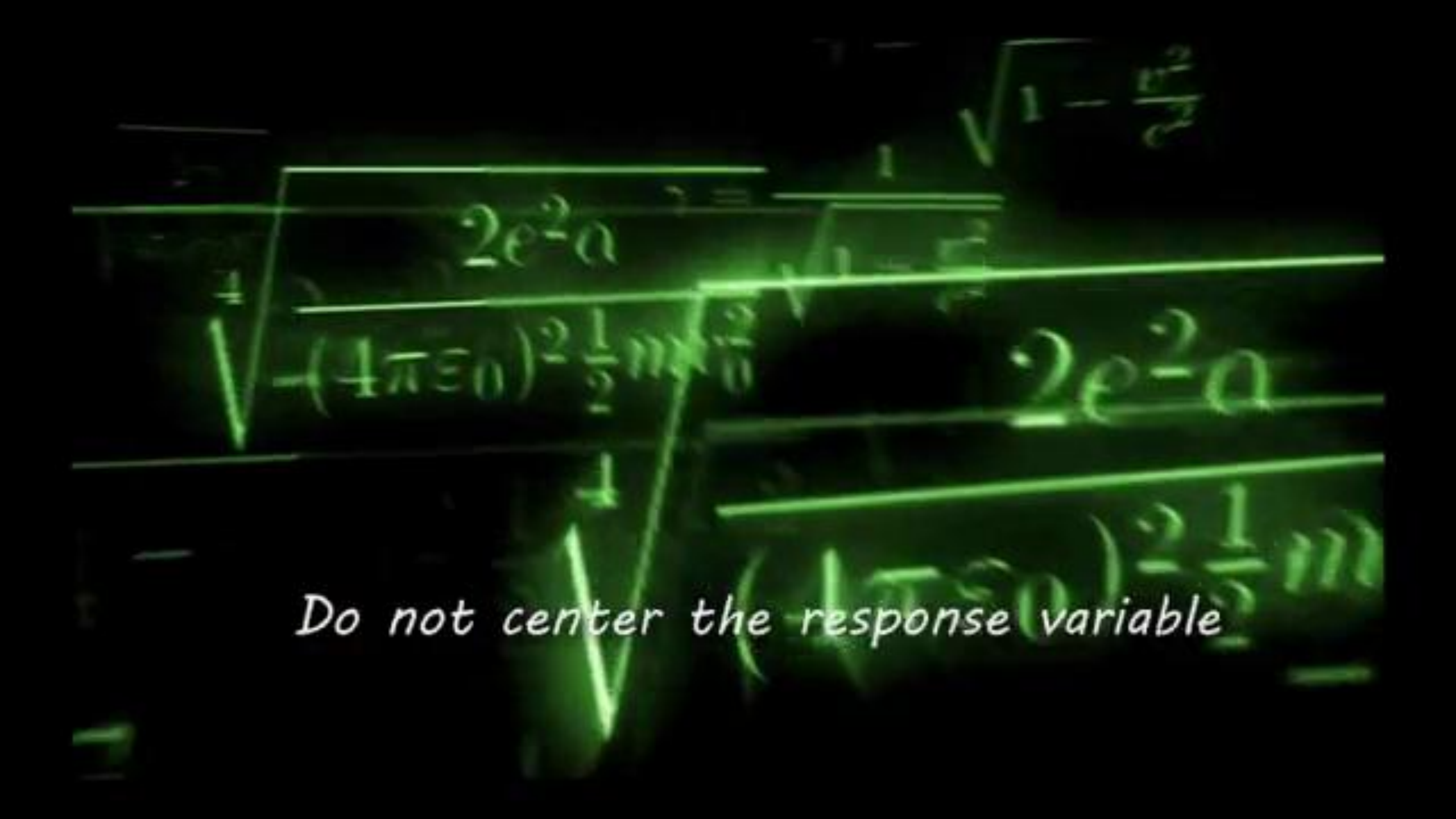
1 = Male

0 = Female

0 =  ??

Centering = subtracting the mean of a variable from the value of the variable


Mean = 0


Do not center the response variable