دورهٔ آموزشی «علم داده»

Data Science Course

جلسهٔ سیزدهم- (بخش اول):
رگرسیون چندگانه

Multiple Regression

مدرس: محمد فزونی
عضو هیات علمی دانشگاه گنبد کاووس
پائیز ۱۳۹۹

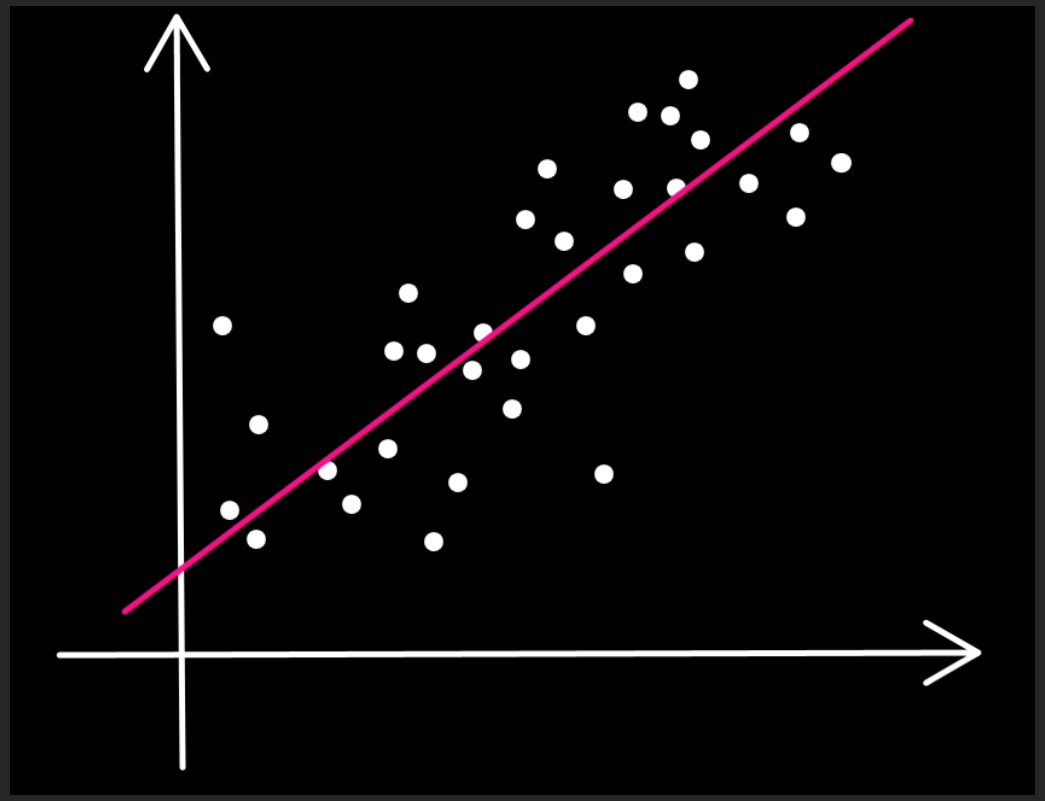
آنچه در هفته سیزدهم خواهیم آموخت:

1. مروری از هفته قبل (مقدماتی در خصوص رگرسیون خطی)
2. رگرسیون خطی چندگانه چیست؟
3. معرفی ضریب تعیین (***R- squared***) و ضریب تعیین تعدیل شده (***Adjusted R-squared***) برای مقایسه عملکرد مدل‌ها
4. تشخیص کارآمدی (اهمیت) کلی مدل با استفاده از آزمون اف (***F-test***)
5. مفروضات رگرسیون (۵ شرط باید محیا باشد)

(x_1, y_1)
(x_2, y_2)
(x_3, y_3)
$\vdots$
(x_n, y_n)

$\hat{y} = b_0 + b_1 x$

خط $\hat{y}$ در برآورد مقدار *GPA* برای سایر اشخاص در جامعه، بما کمک می‌کند. در واقع تقریبی از این نمره‌ها را فراهم می‌نماید



Step 1: Import related libraries

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import statsmodels.api as sm
import seaborn as sns
sns.set()
```

Step 2: Load your dataset

```
data = pd.read_csv('Simple Linear Regression.csv')
data.head(5)
```

```
data.describe()
```

Step 3: Pick appropriate variables

```
y = data['GPA']
x1 = data['SAT']
x1.head()
```

Step 4: Plot your dataset and see what's happening

```
plt.scatter(x1, y)
plt.xlabel('SAT', fontsize = 20)
plt.ylabel('GPA', fontsize = 20)
plt.show()
```

Step 5: Select and create your model

```
x = sm.add_constant(x1)
model = sm.OLS(y,x).fit()
model.summary()
```

Step 6: Plot your regression line

```
plt.scatter(x1,y)
yhat = 0.0017*x1 + 0.275
fig = plt.plot(x1,yhat,  lw=3, c='red')
plt.xlabel('SAT', fontsize = 20)
plt.ylabel('GPA', fontsize = 20)
plt.show()
```

Fundamentals of Multiple Linear Regression

Good models require multiple regressions, in order to address the higher complexity of problems

College GPA = 0.275+0.0017*SAT

MULTIPLE REGRESSION EQUATION

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

IT'S ABOUT THE BEST FITTING MODEL

After 3 dimensions, there is no visual way to represent the data

```
y = data['GPA']
x1 = data[['SAT', 'INC', 'MS']]
```

$$\hat{y} = b_0 + b_1 * SAT + b_2 * INC + b_3 * MS$$

خیلی مهم:

دقت کنید که همیشه افزایش ویژگی‌ها (متغیرهای مستقل) باعث بهبود مدل نمی‌شود

R-squared

دورهٔ آموزشی «علم داده»

Data Science Course

جلسهٔ سیزدهم- (بخش دوم):
ضریب تعیین و کاربردش

R-squared and its application

مدرس: محمد فزونی
عضو هیات علمی دانشگاه گنبدکاووس

Variability?

In data or statistics this is often a measurement of distance from the mean or a description of data range.

SST SUM OF SQUARES TOTAL

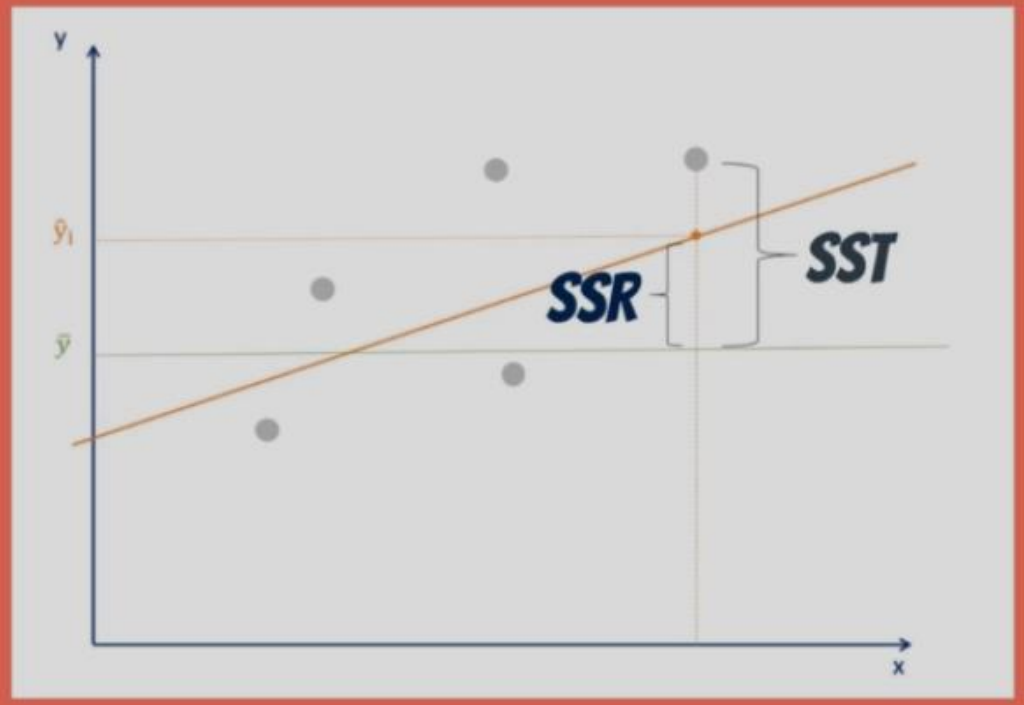
Measures the total variablity of the dataset

$$\sum_{i=1}^{n}(y_i - \bar{y})^2$$

SSR SUM OF SQUARES REGRESSION

Measures the explained variablity by your line

$$\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$



SSE

SUM OF SQUARES ERROR

Measures the unexplained variablity by the regression

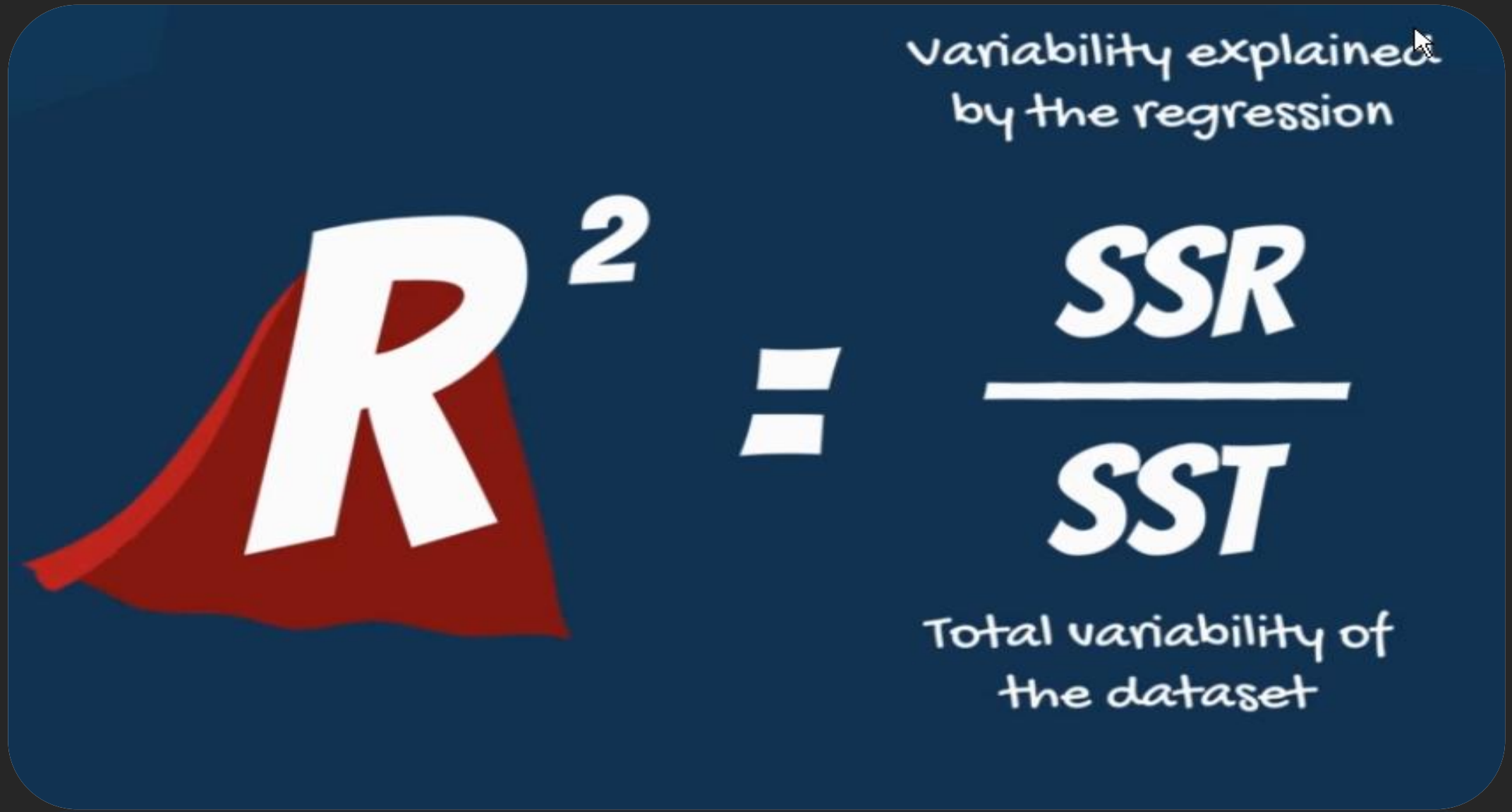
$$\sum_{i=1}^{n} e_i^2$$

SST = SSR + SSE

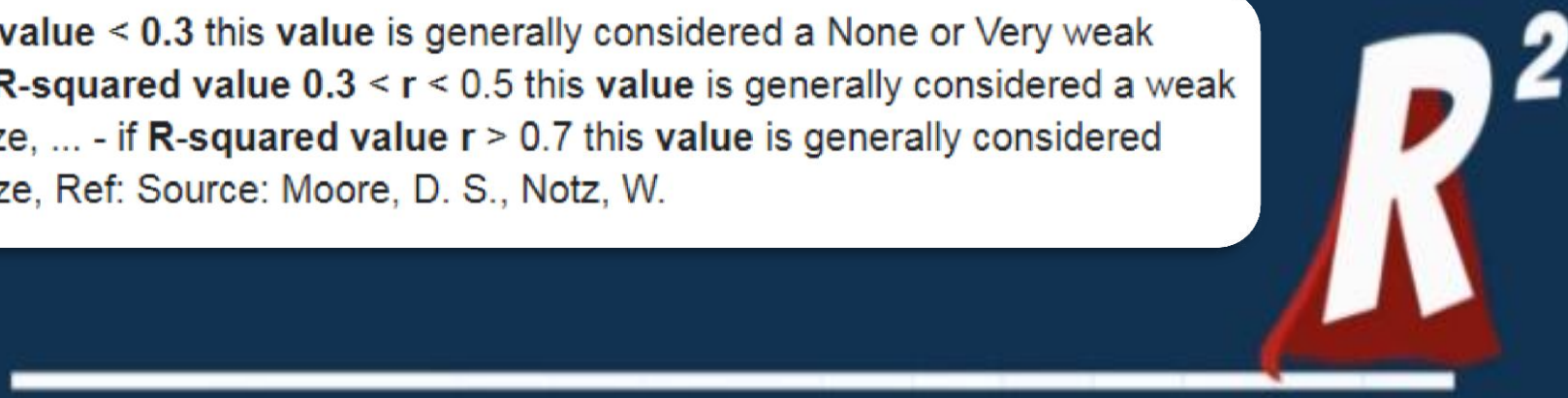
CONNECTION?

Total variability = Explained variability + Unxplained variability

$$\sum_{i=1}^{n}(y_i-\bar{y})^2=\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2+\sum_{i=1}^{n}e_i^2$$


Variability explained by the regression
$R^2 = \frac{SSR}{SST}$
Total variability of the dataset

- if **R-squared value** < **0.3** this **value** is generally considered a None or Very weak effect size, - if **R-squared value 0.3** < **r** < 0.5 this **value** is generally considered a weak or low effect size, ... - if **R-squared value r** > 0.7 this **value** is generally considered strong effect size, Ref: Source: Moore, D. S., Notz, W.



Step 5: Select and create your model

In [6]:
```
x = sm.add_constant(x1)
model = sm.OLS(y,x).fit()
model.summary()
```

Out[6]:

OLS Regression Results

Dep. Variable:	GPA	**R-squared:**	0.406
Model:	OLS	**Adj. R-squared:**	0.399
Method:	Least Squares	**F-statistic:**	56.05
Date:	Sun, 27 Dec 2020	**Prob (F-statistic):**	7.20e-11
Time:	16:48:13	**Log-Likelihood:**	12.672
No. Observations:	84	**AIC:**	-21.34
Df Residuals:	82	**BIC:**	-16.48

در ویدیوی بعدی نشان خواهیم داد که برای مقایسه دو مدل، ضریب تعیین، فاکتور تاثیرگذاری نیست و باید از ابزار دیگری استفاده نمائیم.

این کار را با اضافه نمودن یک ویژگی کاملاً نامربوط به مسئله GPA, SAT انجام می‌دهیم

دورهٔ آموزشی «علم داده»

Data Science Course

جلسهٔ سیزدهم- (بخش سوم):

پیاده‌سازی یک رگرسیون چندگانه در پایتون

Implementation of a Multiple Regression in Python

مدرس: محمد فزونی
عضو هیات علمی دانشگاه گنبدکاووس

دورهٔ آموزشی «علم داده»

Data Science Course

جلسهٔ سیزدهم- (بخش چهارم):
ضریب تعیین تعدیل شده و آزمون اف

Adjusted R-squared and F-test

مدرس: محمد فزونی
عضو هیات علمی دانشگاه گنبدکاووس

$\bar{R}^2$ or Adjusted R-squared

Difference between R-square and Adjusted R-square

Every time you add a independent variable to a model, the **R-squared** increases, even if the independent variable is insignificant. It never declines. Whereas **Adjusted R-squared** increases only when independent variable is significant and affects dependent variable.

Dep. Variable:	GPA	**R-squared:**	0.407
Model:	OLS	**Adj. R-squared:**	0.392
Method:	Least Squares	**F-statistic:**	27.76
Date:	Tue, 29 Dec 2020	**Prob (F-statistic):**	6.58e-10
Time:	12:30:35	**Log-Likelihood:**	12.720
No. Observations:	84	**AIC:**	-19.44
Df Residuals:	81	**BIC:**	-12.15
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.2960	0.417	0.710	0.480	-0.533	1.125
SAT	0.0017	0.000	7.432	0.000	0.001	0.002
Rand 1,2,3	-0.0083	0.027	-0.304	0.762	-0.062	0.046

Dep. Variable:	GPA	**R-squared:**	0.406
Model:	OLS	**Adj. R-squared:**	0.399
Method:	Least Squares	**F-statistic:**	56.05
Date:	Sun, 27 Dec 2020	**Prob (F-statistic):**	7.20e-11
Time:	16:48:13	**Log-Likelihood:**	12.672
No. Observations:	84	**AIC:**	-21.34
Df Residuals:	82	**BIC:**	-16.48

F - test
and
its application

Dep. Variable:	GPA	**R-squared:**	0.406
Model:	OLS	**Adj. R-squared:**	0.399
Method:	Least Squares	**F-statistic:**	56.05
Date:	Sun, 27 Dec 2020	**Prob (F-statistic):**	7.20e-11
Time:	16:48:13	**Log-Likelihood:**	12.672
No. Observations:	84	**AIC:**	-21.34
Df Residuals:	82	**BIC:**	-16.48

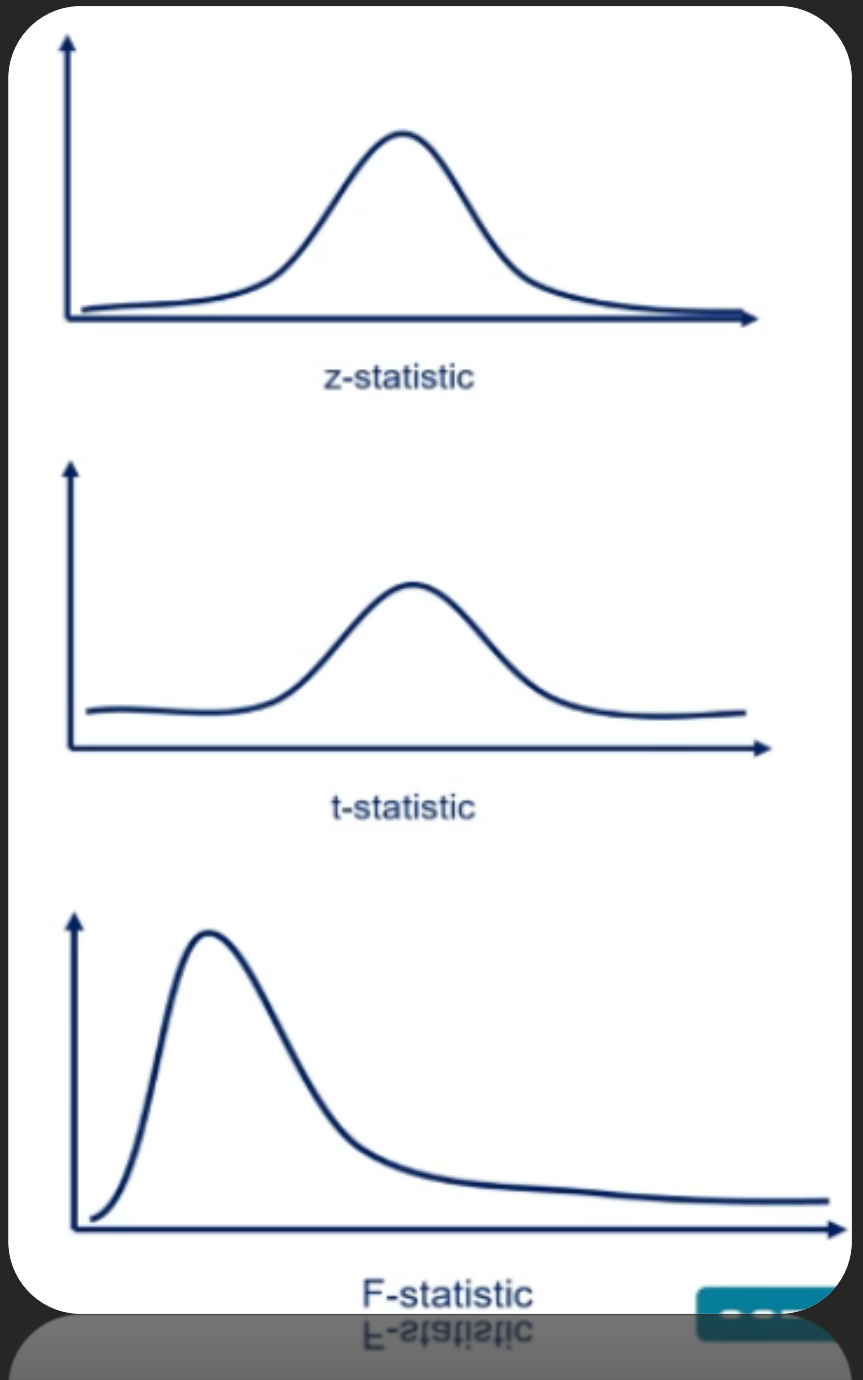
F-test:

$$H_0: \beta_1 = \beta_2 = \ldots = \beta_k = 0$$

$$H_1: \text{at least one } \beta_i \neq 0$$

If all betas are 0, then none of the Xs matter => our model has no merit

An *F*-test is any statistical test in which the test statistic has an *F*-distribution under the null hypothesis. It is most often used when comparing statistical models that have been fitted to a data set, in order to identify the model that best fits the population from which the data were sampled.



Dep. Variable:	GPA	**R-squared:**	0.407
Model:	OLS	**Adj. R-squared:**	0.392
Method:	Least Squares	**F-statistic:**	27.76
Date:	Tue, 29 Dec 2020	**Prob (F-statistic):**	6.58e-10
Time:	12:30:35	**Log-Likelihood:**	12.720
No. Observations:	84	**AIC:**	-19.44
Df Residuals:	81	**BIC:**	-12.15
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.2960	0.417	0.710	0.480	-0.533	1.125
SAT	0.0017	0.000	7.432	0.000	0.001	0.002
Rand 1,2,3	-0.0083	0.027	-0.304	0.762	-0.062	0.046

Dep. Variable:	GPA	**R-squared:**	0.406
Model:	OLS	**Adj. R-squared:**	0.399
Method:	Least Squares	**F-statistic:**	56.05
Date:	Sun, 27 Dec 2020	**Prob (F-statistic):**	7.20e-11
Time:	16:48:13	**Log-Likelihood:**	12.672
No. Observations:	84	**AIC:**	-21.34
Df Residuals:	82	**BIC:**	-16.48

Regression Assumptions

دورهٔ آموزشی «علم داده»

Data Science Course

جلسهٔ سیزدهم- (بخش پنجم):

مفروضات رگرسیون

Regression Assumptions

مدرس: محمد فزونی

عضو هیات علمی دانشگاه گنبدکاووس

1. Linearity $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon$

2. No endogeneity $\sigma_{X\varepsilon} = 0 : \forall\, x, \varepsilon$

3. Normality and homoscedasticity $\varepsilon \sim N(0, \sigma^2)$

4. No autocorrelation $\sigma_{\varepsilon_i \varepsilon_j} = 0 : \forall\, i \neq j$

5. No multicollinearity $\rho_{x_i x_j} \neq 1 : \forall\, i, \mathrm{j};\ \mathrm{i} \neq \mathrm{j}$

The biggest mistake you can make is to perform a regression that violates one of these assumptions!

LINEARITY

$$\gamma = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon$$

Fixes

1. Run a non-linear regression
2. Exponential transformation
3. Log transformation

NO ENDOGENEITY

$$\sigma_{X\varepsilon} = 0 : \forall\, x, \varepsilon$$

OMITTED VARIABLE BIAS

You forget to include a relevant variable

y is explained (somewhat correlated) by xs

y is explained (somewhat correlated) by omitted x

=>

x and ε are somewhat correlated

LONDON

Price = F (Size)

$$\hat{y} = 11342786 - 132100x_1$$

$$\sigma_{X\varepsilon} \neq 0$$

Where did we draw the sample from?

Can we get a better sample?

Why is bigger real estate cheaper?

Where did we draw the sample from? Central London

Can we get a better sample? Big enough

Why is bigger real estate cheaper? Not in the City of London

What is it about smaller size that is making it so expensive? Location

Where are the small houses?  City of London

$$\hat{y} = 11342786 - 132100x_1$$

$$\hat{y} = 520365 + 78210x_{size} + 7126579\,x_{city}$$

Size is with a positive sign once again

1 if in city, 0 if out

OMITTED VARIABLE BIAS

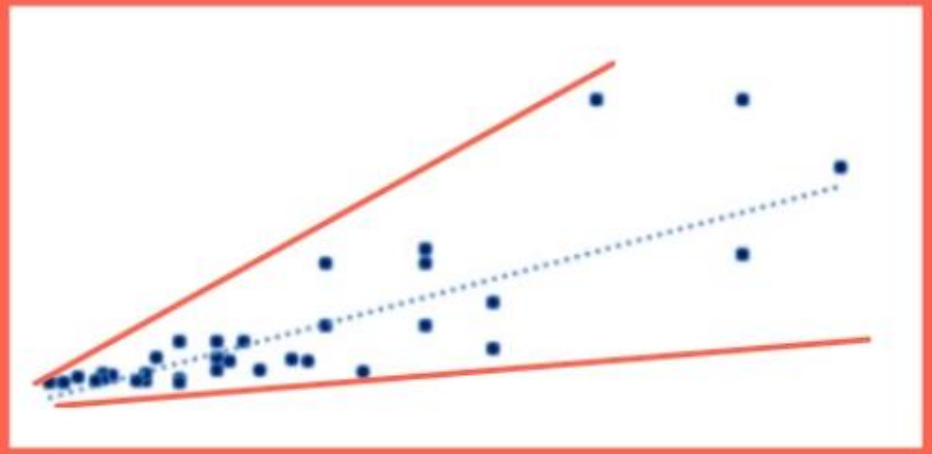
- Always different
- Always sneaky
- Only experience and advanced knowledge can help
- Don't hesitate to ask for a hand if you can't figure it out!

NORMALITY AND HOMOSCEDASTICITY

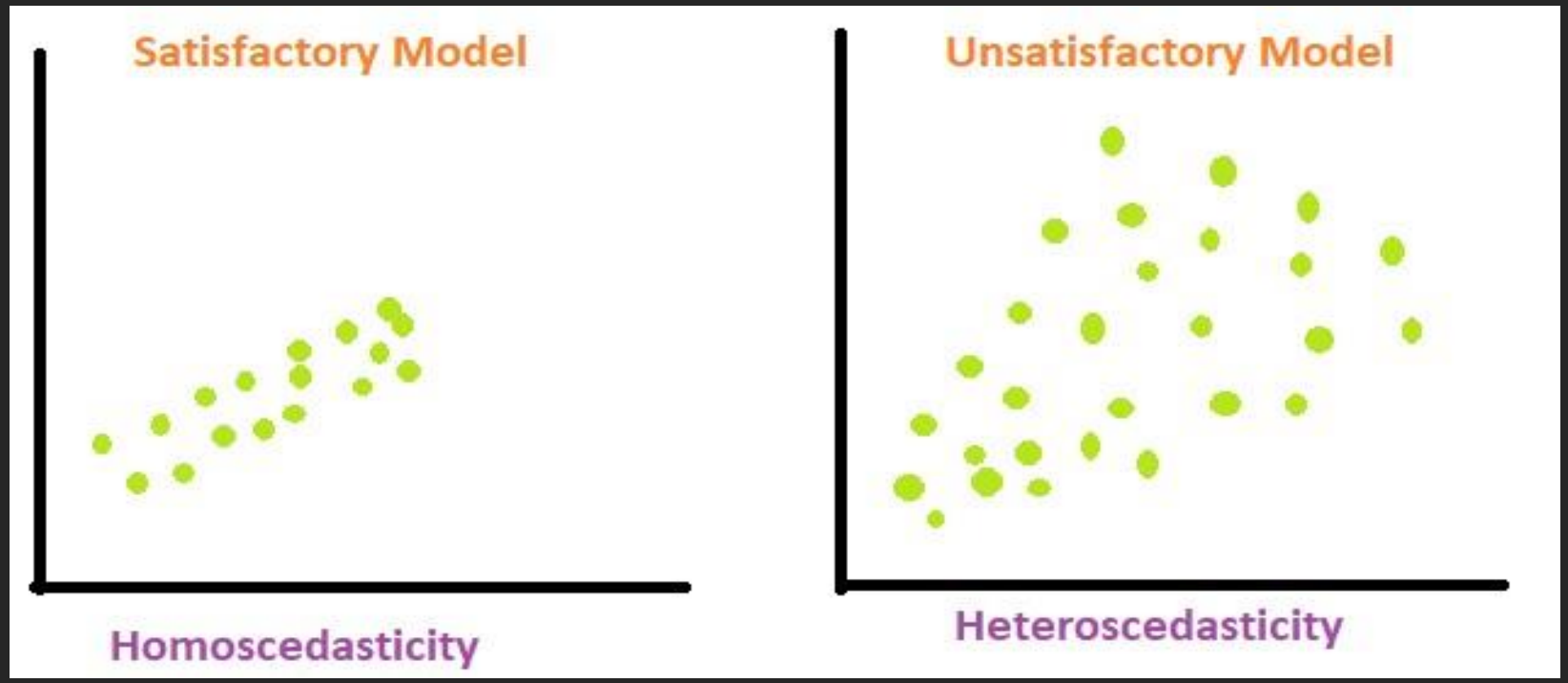
$$\varepsilon \sim N(0, \sigma^2)$$

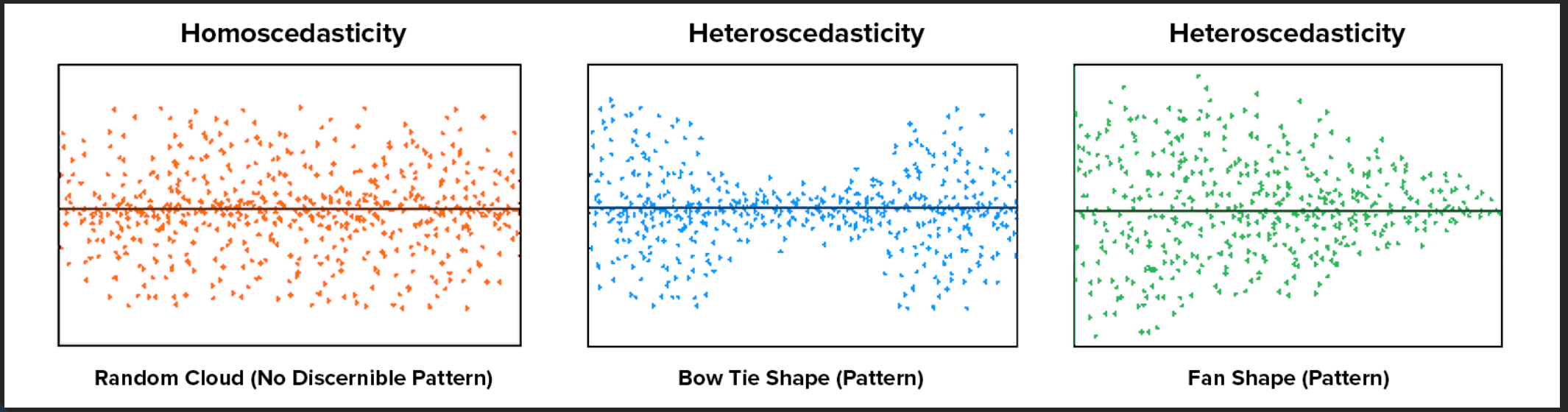
- homoscedasticity $\sigma^2_{\varepsilon_1} = \sigma^2_{\varepsilon_2} = \ldots = \sigma^2_{\varepsilon_k} = \sigma^2$

  (to have equal variance)



heteroscedastic dataset


Satisfactory Model
Homoscedasticity
Unsatisfactory Model
Heteroscedasticity

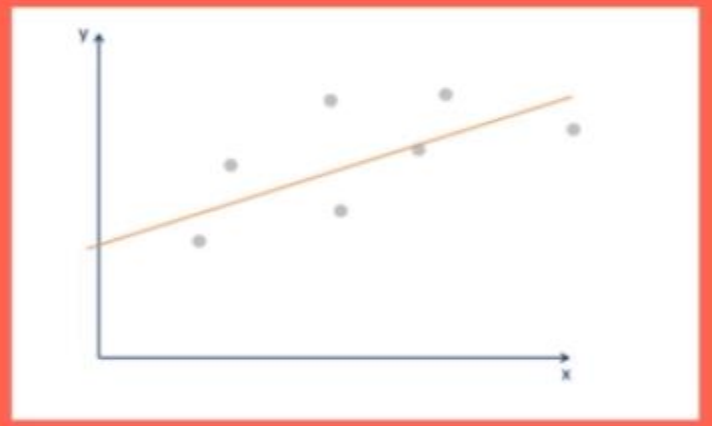

Homoscedasticity
Random Cloud (No Discernible Pattern)
Heteroscedasticity
Bow Tie Shape (Pattern)
Heteroscedasticity
Fan Shape (Pattern)

Coding time in Python to get a better understanding of Homoscedasticity

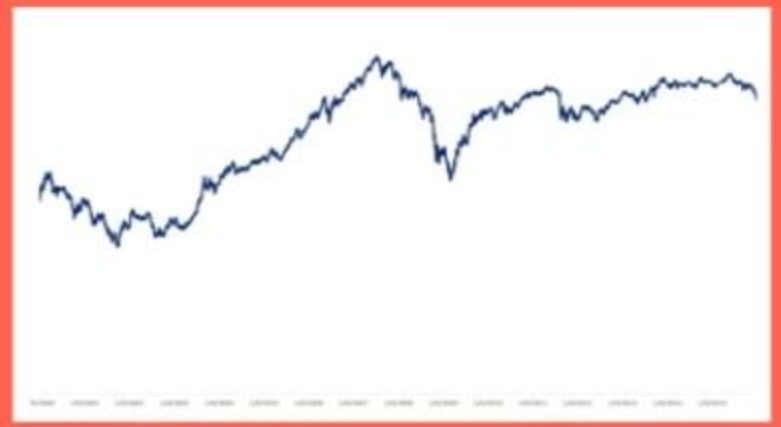
NO AUTOCORRELATION

a.k.a. no serial correlation

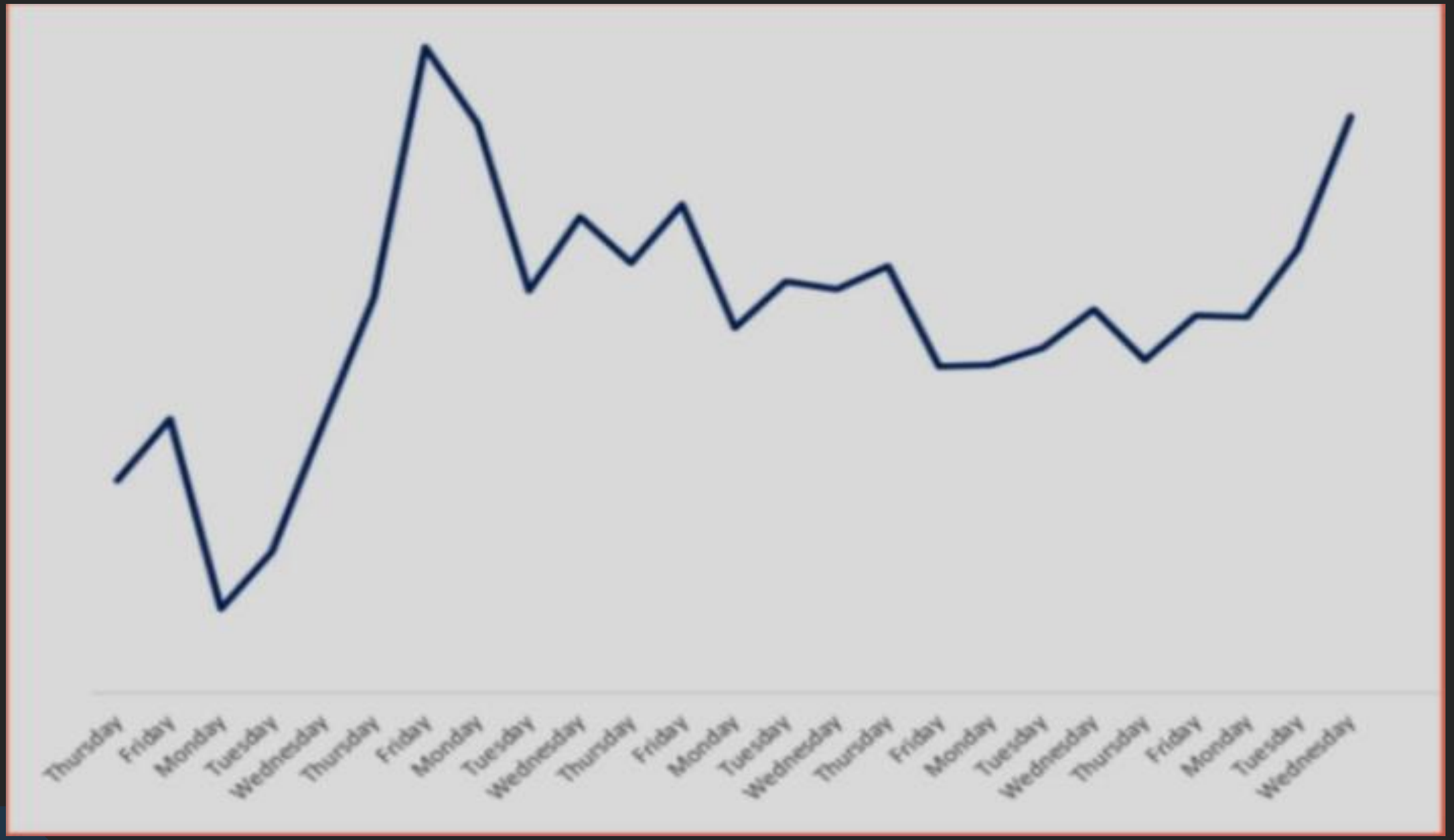
$$\sigma_{\varepsilon_i \varepsilon_j} = 0 : \forall\, i \neq j$$



cross-sectional data



time series data


Thursday
Friday
Monday
Tuesday
Wednesday
Thursday
Friday
Monday
Tuesday
Wednesday
Thursday
Friday
Monday
Tuesday
Wednesday
Thursday
Friday
Monday
Tuesday
Wednesday
Thursday
Friday
Monday
Tuesday
Wednesday

DETECTION

Dep. Variable:	GPA	R-squared:	0.406
Model:	OLS	Adj. R-squared:	0.399
Method:	Least Squares	F-statistic:	56.05
Date:	Thu, 05 Apr 2018	Prob (F-statistic):	7.20e-11
Time:	15:24:09	Log-Likelihood:	12.672
No. Observations:	84	AIC:	-21.34
Df Residuals:	82	BIC:	-16.48
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.2750	0.409	0.673	0.503	-0.538	1.088
SAT	0.0017	0.000	7.487	0.000	0.001	0.002

Omnibus:	12.839	Durbin-Watson:	0.950
Prob(Omnibus):	0.002	Jarque-Bera (JB):	16.155
Skew:	-0.722	Prob(JB):	0.000310
Kurtosis:	4.590	Cond. No.	3.29e+04

Durbin-Watson falls between 0 and 4

2 -> no autocorrelation

<1 and >3 cause an alarm

ALTERNATIVES

- Autoregressive model
- Moving average model
- Autoregressive moving average model
- Autoregressive integrated moving average model

NO MULTICOLLINEARITY

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

$$a = 2 + 5 * b \qquad b = \frac{a - 2}{5}$$

$$\rho_{ab} = 1$$ perfect multicollinearity

Rationale: If a can be represented using b, there's no point in using both

NO MULTICOLLINEARITY

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

c, d

$\rho_{cd} = 0.9$ imperfect multicollinearity

Rationale: If c can be ALMOST represented using d, there's no point in using both

NO MULTICOLLINEARITY

Prevention:

Find the correlation between each two pairs of independent variables

ρ_{x_i, x_j} for $\forall\ i, j;\ i \neq j$

Fixes

1. Drop one of the two variables
2. Transform them into one (e.g. average price)
3. Keep them both 

1. خطی بودن، یعنی بین Y و هر کدام از متغیرهای مستقل یک رابطه خطی موجود باشد
2. یک ویژگی در رگرسیون موجود نیست که باعث مدل بد می‌شود (مثال خانه در لندن)
3. تمایز واریانس (Hetroscedastisity)
4. وجود مدل در خطاها (مثال استاک مارکت)
5. دو ویژگی رابطه خطی نسبت بهم دارند. باید یکی را حذف کرد و یا به یک ویژگی تبدیل کرد.