Agenda

→ Feature Importance Code

→ Assumption of Linear Regression

→ OLS using stat model

→ Code

$$\sum \left(y^{(i)} - \hat{y}^{(i)}\right)^2 = SSE$$

$$\frac{1}{m} \sum \left(y^{(i)} - \hat{y}^{(i)}\right)^2 = MSE$$

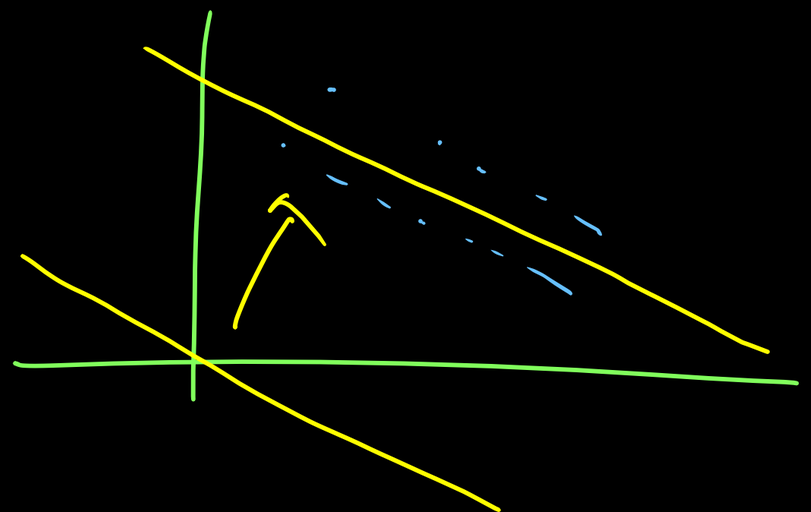
$$\hat{y} = w_1 x_1 + w_2 x_2 - \cdots - w_d x_d + w_0$$

Model weights / Coefficients

model intercept

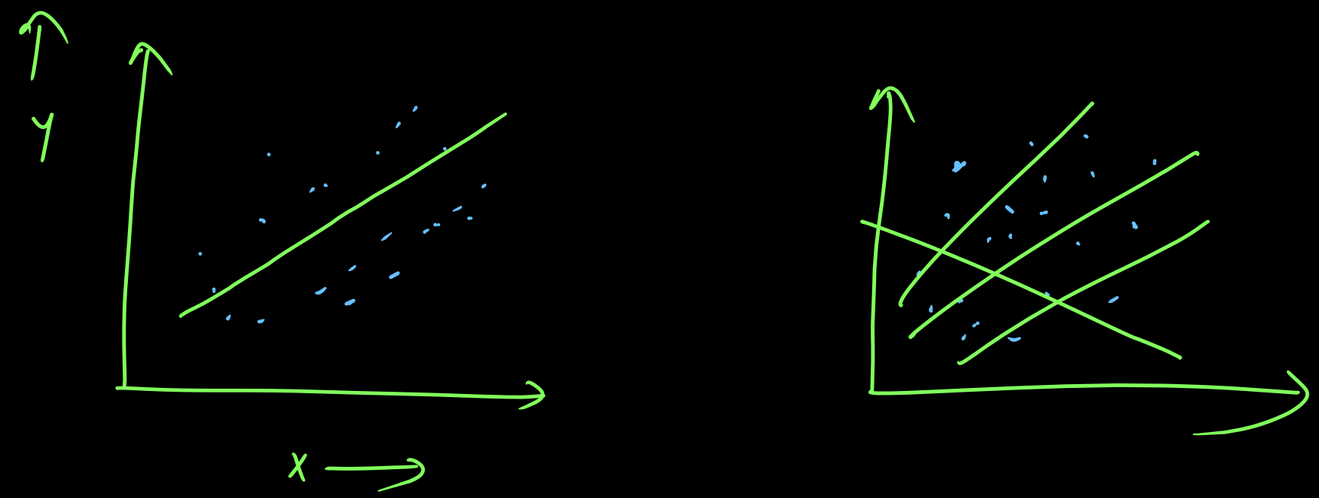
$\hat{y} = w_1 x + w_0$

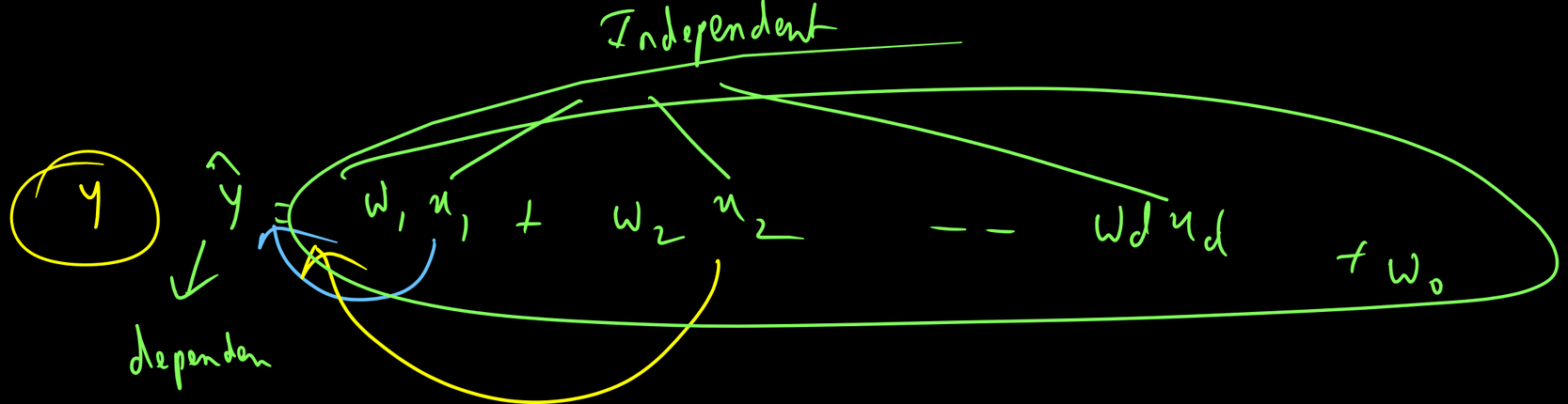
w_0



→ Assumption for linear Regression

1 Linearity





① Pearson test / Spearman Correlation

$r = 1 \longrightarrow$ Perfect +ve linear

$r = -1 \longrightarrow$ Perfect -ve linear relation

$r = 0 \longrightarrow$ No linear relation

② Scatter plot / Line plot

(2) No Multicolinearity

→ features should be independent

x_1	x_2	- - -	x_d	y

Km driven, Miles → Constant

$x_1 = \alpha x_2 + \beta$

→ Collinear

1 mile = 1.6 km

$\alpha = \frac{1}{1.6}$

→ Multicolinear

$$\begin{cases} x_1 = \alpha x_2 + \beta x_3 + \gamma x_4 \\ \end{cases}$$

$$\hat{y} = \underset{1}{x_1} + \underset{1.5}{2x_2} + \underset{1}{3x_3} + 5$$

$$= 1 + 2\times1.5 + 3 + 5$$

$$= 1 + 3 + 3 + 5$$

$$= 12$$

$x_3 > x_2 > x_1$

$x_2 = 1.5\, x_1$

Feature importance get messed up ∵ of Multicolinearity

$$\hat{y} = x_1 + 2\times1.5\,x_1 + 3x_3 + 5$$

$$= \underset{1}{4}x_1 + \underset{0}{0}x_2 + \underset{1}{3}x_3 + \underset{5}{5}$$

↳ $x_1 > x_3 > x_2$

$$= 4 + 3 + 5 = 12$$

→ How to remove Multicolinearity ?

Repeat converge

→ Regression x_1, x_2, x_3, x_4 y

↳ R^2

→ VIF for each features
(Variance Inflation factor)

↳ Check for feature with highest VIF x_1

→ Remove the feature with highest VIF

x_2, x_3, x_4, y

Logistic
↓
Cross

Linear Regression
↓
MSE

VIF

$y \approx w_1 x_1 + w_2 x_2 + w_3 x_3 + w_4 x_4$

$\hat{y} \approx y$ → $R^2 \approx 1$

$x_1 = \alpha x_2 + \beta x_3 + \gamma x_4$

y

$x_2\ x_3\ x_4$ | x_1

$R_1^2 \approx 1$

$VIF_1 = \frac{1}{1 - R^2}$

$= \frac{1}{1-1} = \infty$

$R^2 = 0$

$VIF = 1$

$R^2 = -\infty$

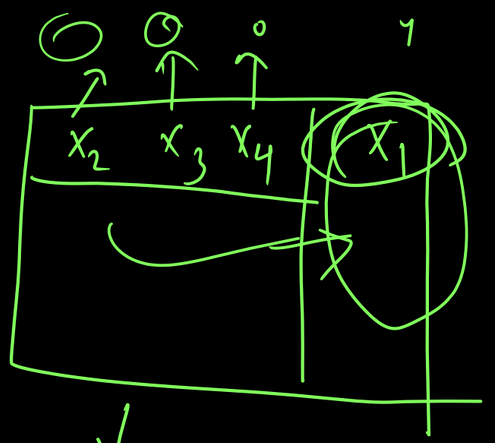
$\frac{1}{\infty}$

$VIF = 0$

VIF > 10 → very High Multicolinearity

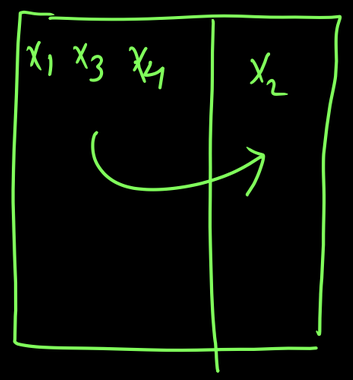
5 < VIF < 10 : → Multicol exist

VIF < 5 : low Multicolinearity



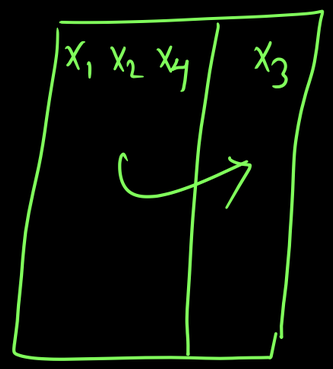
R_1^2

VIF_1



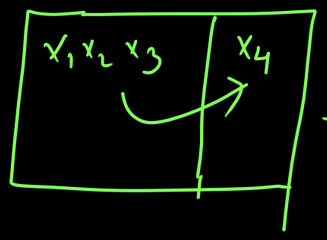
R_2^2

VIF_2



R_3^2

VIF_3



R_4^2

VIF_4

VIF

X_1	∞
X_2	10
X_3	2
X_4	0.1

→ X_1, X_2, X_3, X_4, Y → $R^2 = 0.85$

→ Calculate VIF

	VIF
x_1	α
x_2	
x_3	
x_4	

→ Remove X_1

→ feature remaining x_2 x_3 x_4

→ X_2, X_3, X_4, Y → $R^2 = 0.84$

→ Calculate VIF

	VIF
x_2	~~4.5~~
x_3	~~4~~
x_4	1

→ Remove X_2

VIF < 5

X_2	4.5
X_3	4
X_4	1

$R^2 \leq 0.75$

→ X_3, X_4, Y → $R^2 = 0.75$

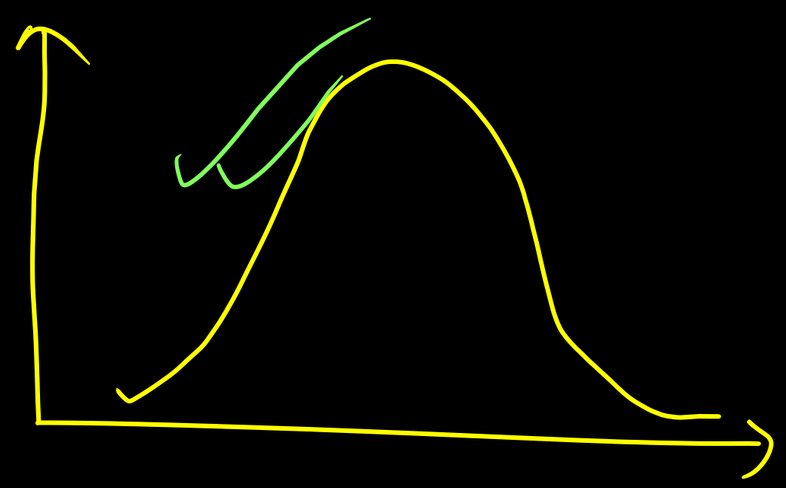
→ Stop it

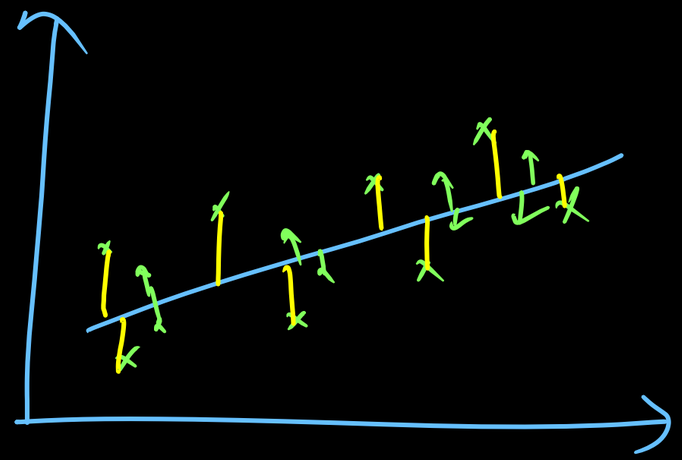
→ Errors are normally distributed

$$\hat{y}^{(i)} = w_1 x_1^{(i)} + w_2 x_2^{(i)} - - w_d x_d^{(i)}$$

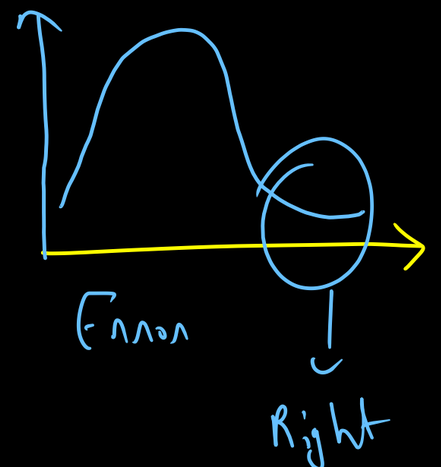
(Residual $y^{(i)}$

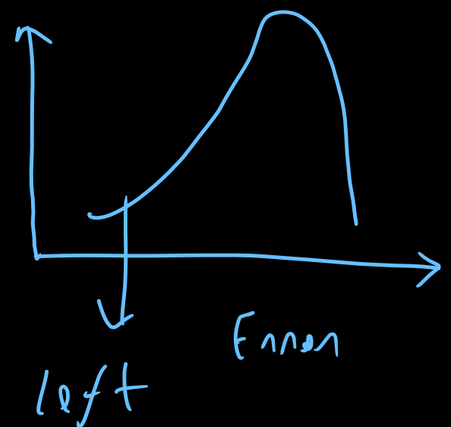
$e^{(i)} = y^{(i)} - \hat{y}^{(i)}$

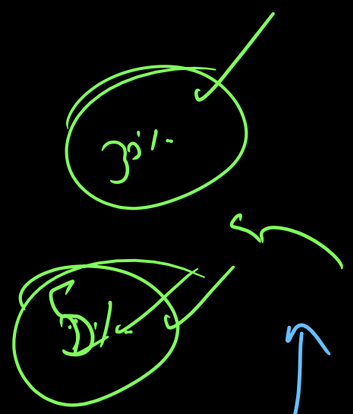


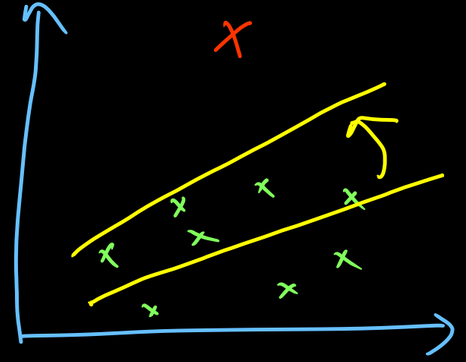


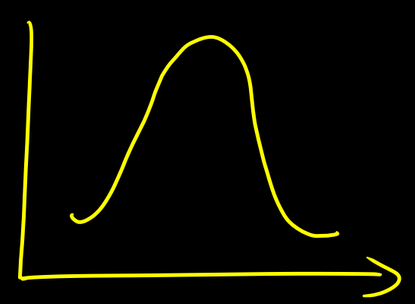
→ Shapiro, QQ plot

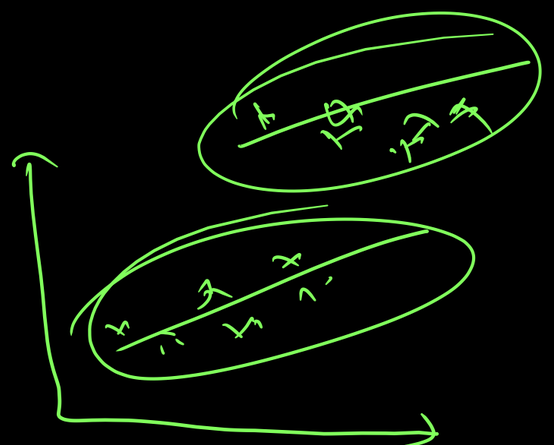

Error
Right


left
Error









Ordinary least Square = Regression

$\frac{1}{m}\sum (y^{(i)} - \hat{y}^{(i)})^2$

$m - K - 1$
$= 13874 - 16 - 1$

→ Break until 22:25

SKlearn → $\hat{y} = w_1 x_1 \cdots w_d x_d + w_0 \cdot 1$ (x_0)

$= W^T x + w_0$ (bias)

Stats model → $W^T x$

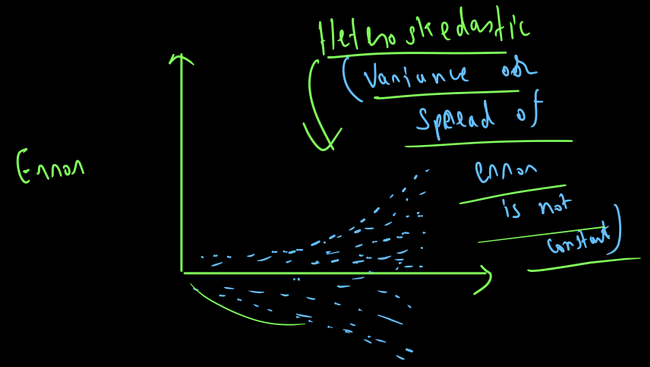
$W' = [w_1 \cdots w_d \; w_0]$

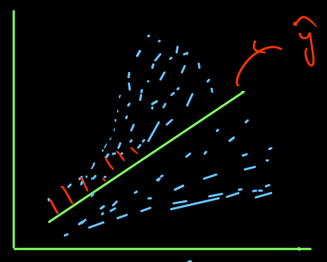
$x = [x_1 \cdots x_d \; 1]$

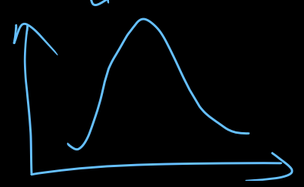
$w_1 x_1 + w_2 x_2 \cdots w_d x_d + w_0 \cdot 1$

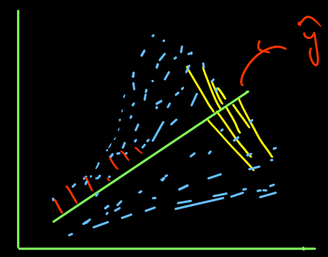
(4) → No Heteroskedascity

Heteroskedastic (Variance or Spread of Error is not constant)

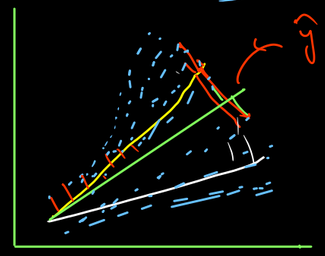


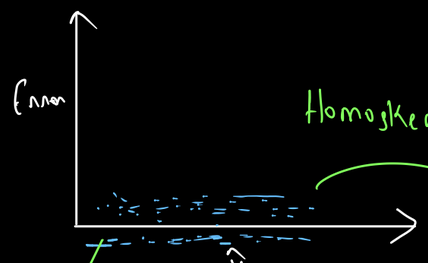






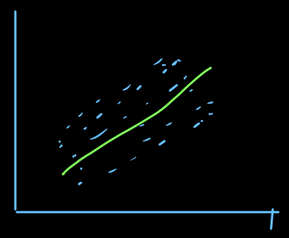
$\sum (y^{(i)} - \hat{y}^{(i)})^2$





Homoskedastic

Variance of Error is constant



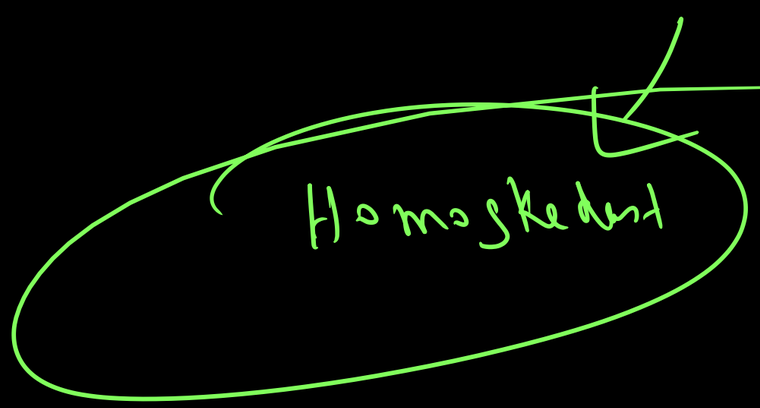
```python
# Performing the Goldfeld-Quandt test to check for Homoscedas
from statsmodels.compat import lzip
import statsmodels.stats.api as sms

name = ['F statistic', 'p-value']
test = sms.het_goldfeldquandt(y_train, X2_sm)
lzip(name, test)
```

```
[('F statistic', np.float64(1.0176613865185007)),
 ('p-value', np.float64(0.21807877686818042))]
```

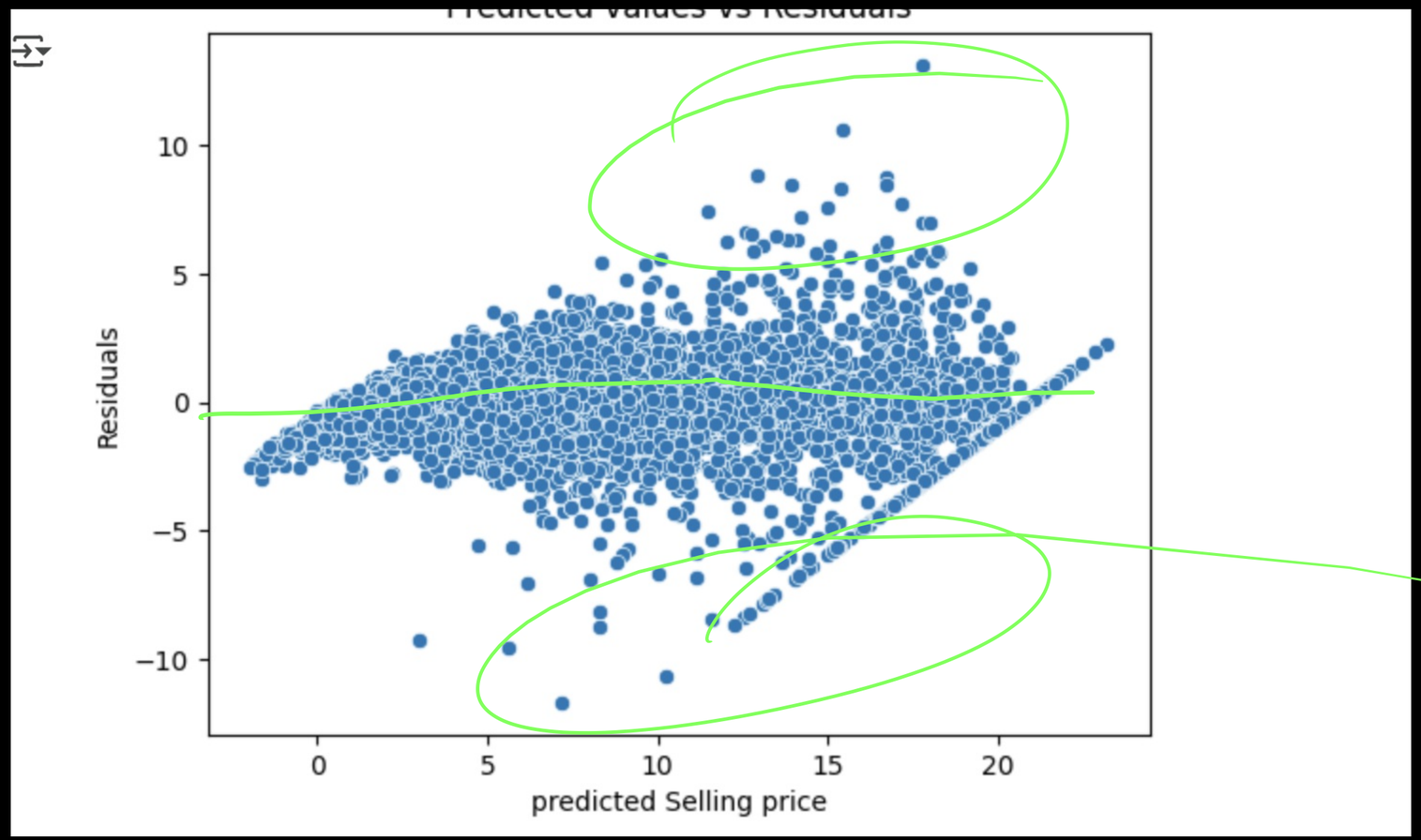
H_0 : Homoskedas

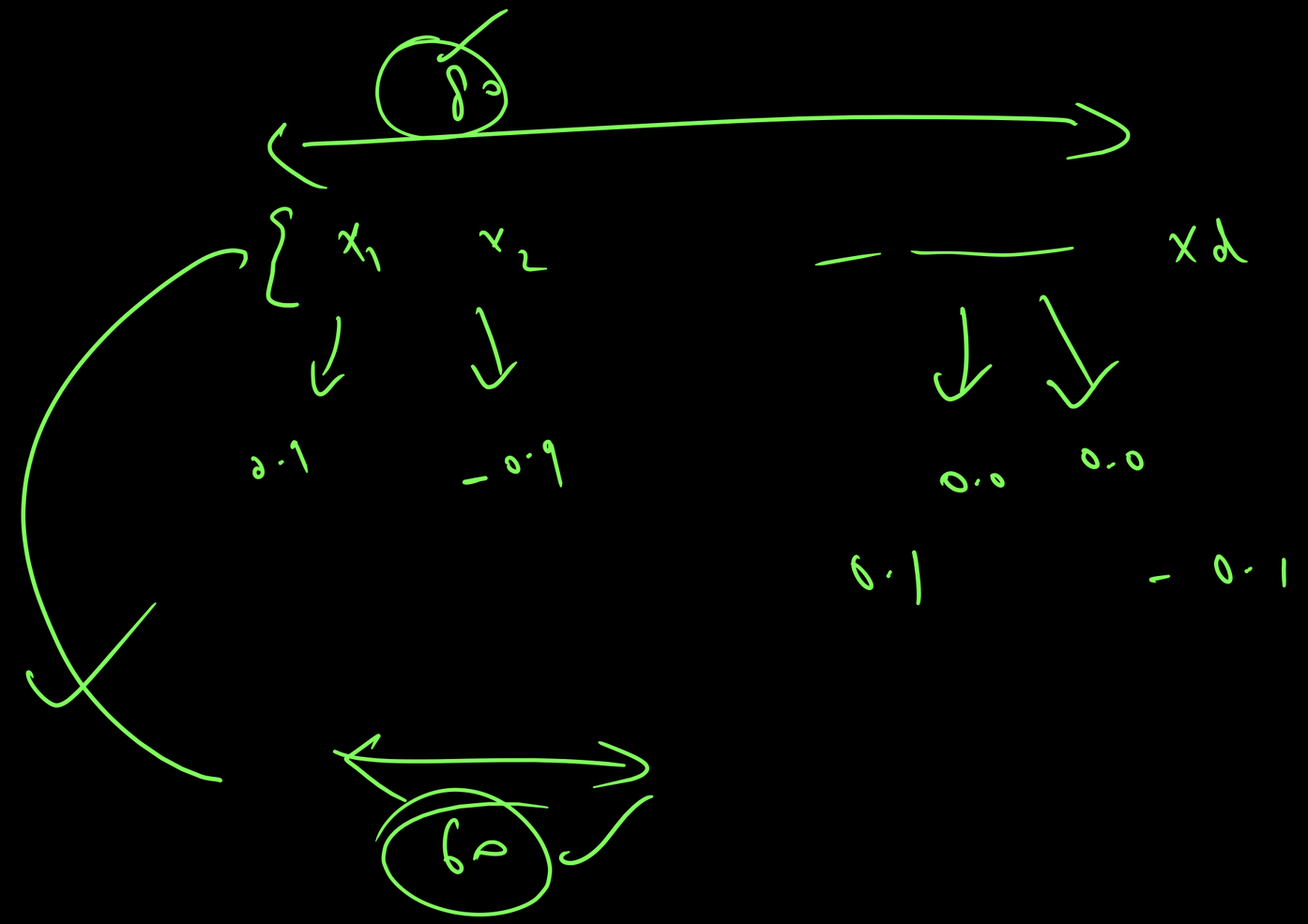
H_1 : Not Homo

0.218 → We don't Reject Homoske



$p < 0.05$ → Reject Null


Predicted Values vs Residuals
Residuals
predicted Selling price
10
5
0
−5
−10
0
5
10
15
20


x_1
x_2
x_d
0.1
-0.9
0.0
0.0
0.1
-0.1