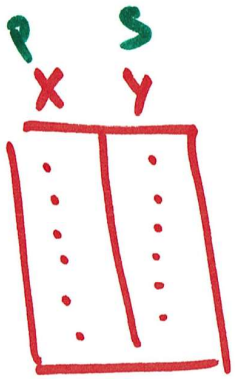
Learn From Data



learn

→ Prediction
→ Interpretation

Understand the world

Math model

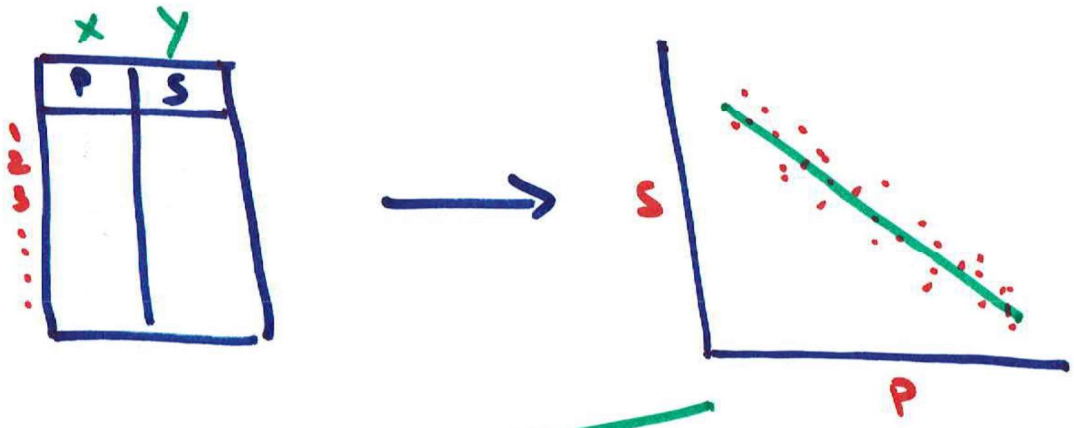
→ $S = 1000 - 20p$

$S = 100 + 20p - 30p^2$

$S = 10\, e^{5p + 6p^2}$

S =>

S =>



Find the **best** 'fit' line

Find the **best** a, b in $\hat{y}_i = \overset{?}{a} + \overset{?}{b} x_i$

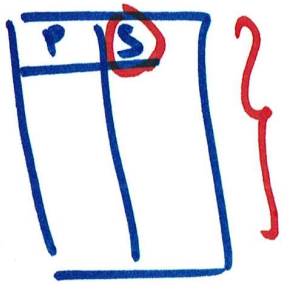
Find a, b such that $\min \sum_i (y_i - \hat{y}_i)^2$

Find a, b s.t. $\min \sum_i (y_i - (a + b x_i))^2$

Stat

Assumption : TRUTH $y_i = \alpha + \beta * x_i + \epsilon_i$

Data Generating Model



Find $a, b \Rightarrow \hat{y} = a + b * x$

Stat. Inf.

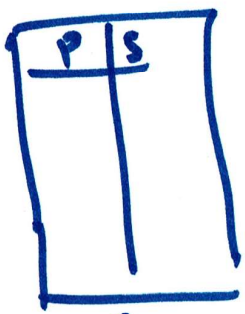
Est. $\hat{a}, \hat{b}$

↳ What can we say about α, β?

- CI → $\alpha \in (1000 \pm 10, 1000 \pm 10)$ 95%
- HT → Is β really Zero or Not?

ML

No Assumption on the DGM



Find $a, b \Rightarrow \hat{y} = \hat{a} + \hat{b}x$

$y = a + bx + cx^2$

$y = a e^{bx}$

$y \Rightarrow$

$y =$

$y =$

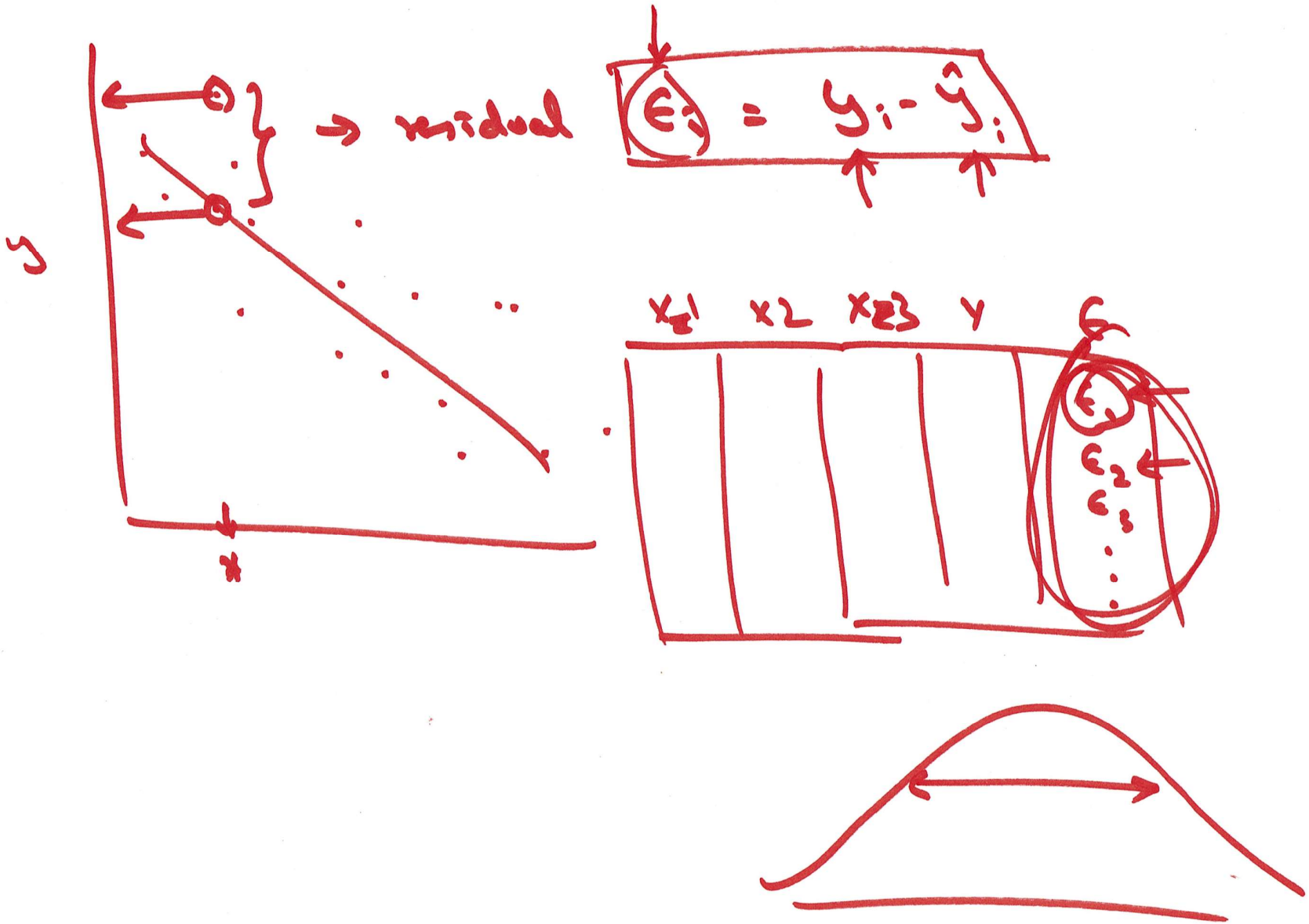
Understanding the world (from Data)

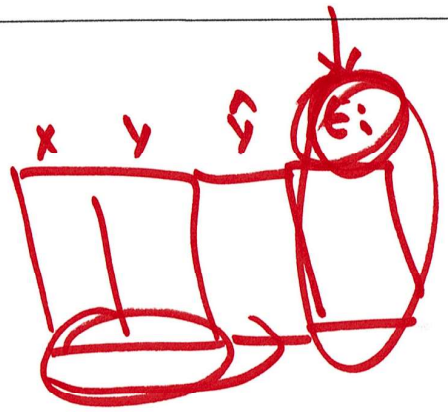
Stats (~300 yrs)

- Assumes a DGM
- In-Sample
- Stat. Inf.
- Field of Math
  - Fit a model
  - parameter
  - Covariate

ML (~30 yrs)

- No Assumption on DGM
- Train Vs Test
- No Stat. Inf.
- Field of CS
  - learn
  - weight
  - feature

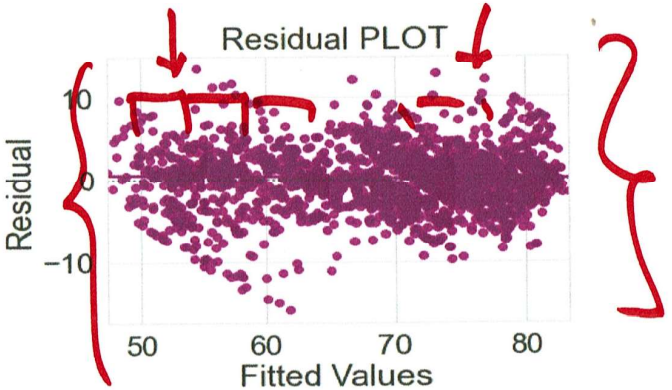




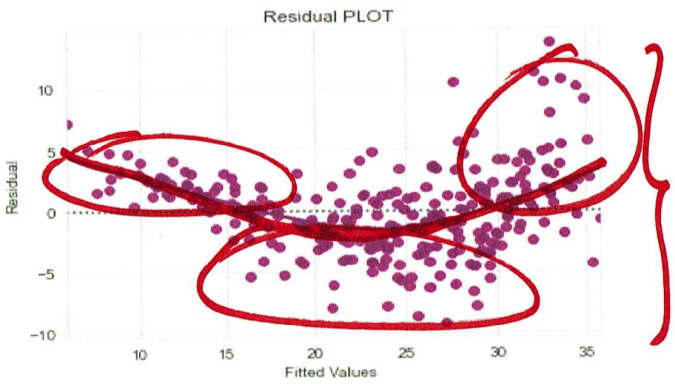
Linear Relationship

$y = a + bx + bcx^2$

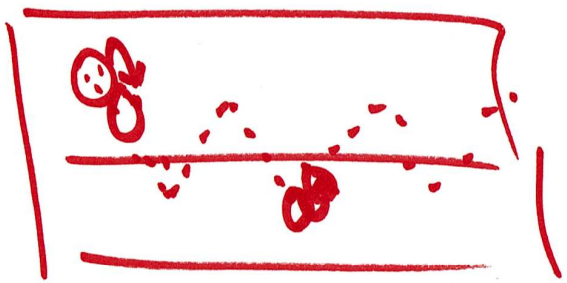
- After finding the best linear fit, a plot of the residuals will provide a good insight.
- If they don't follow any pattern, we say that the model is linear otherwise model is showing signs of non-linearity
- To deal with non-linearity, we can try transforming variables as per their relationship with target variable.

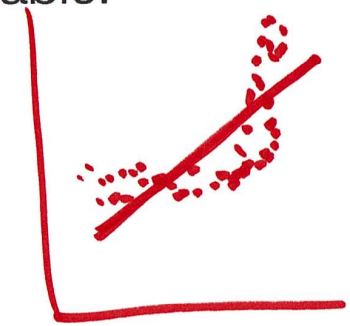


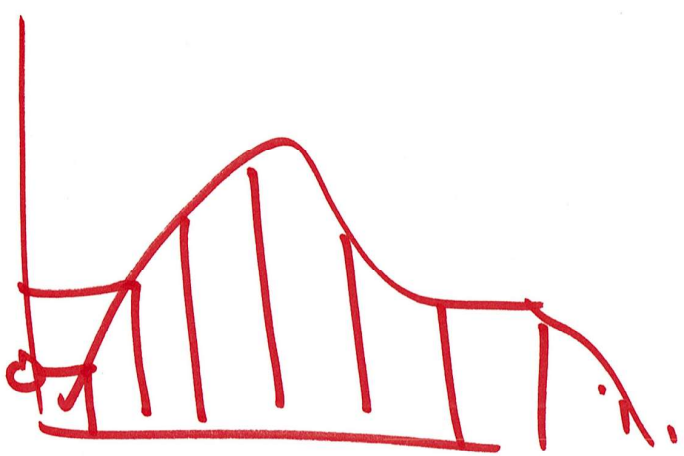
No pattern

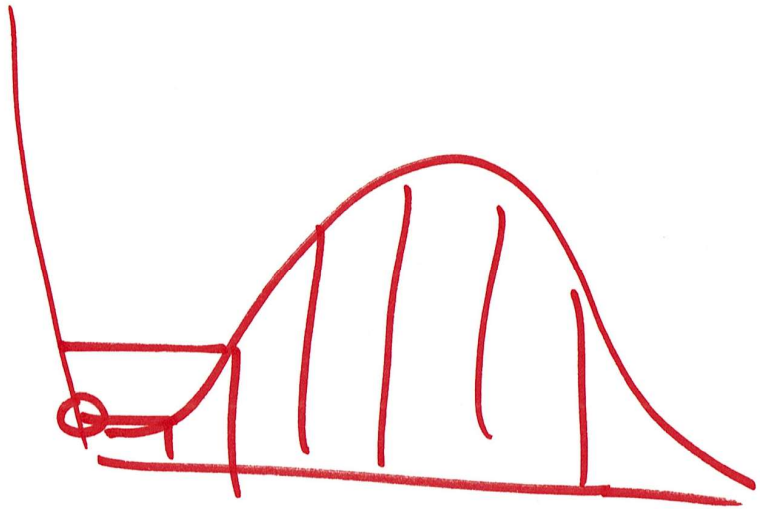


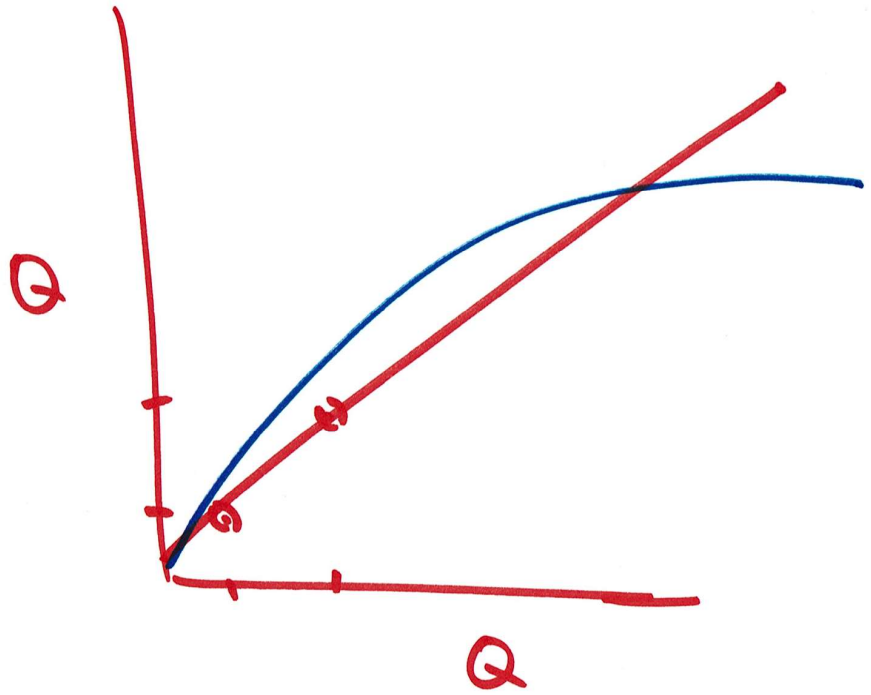
Some non-linearity





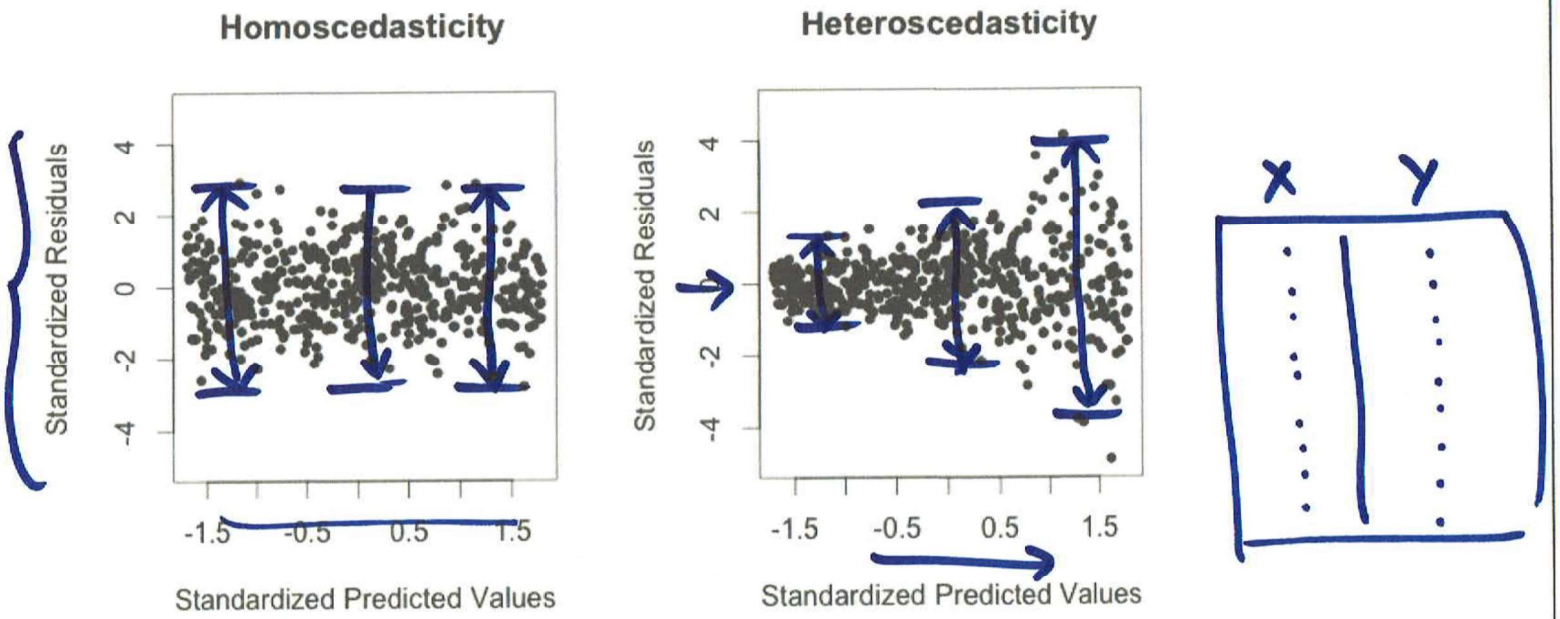




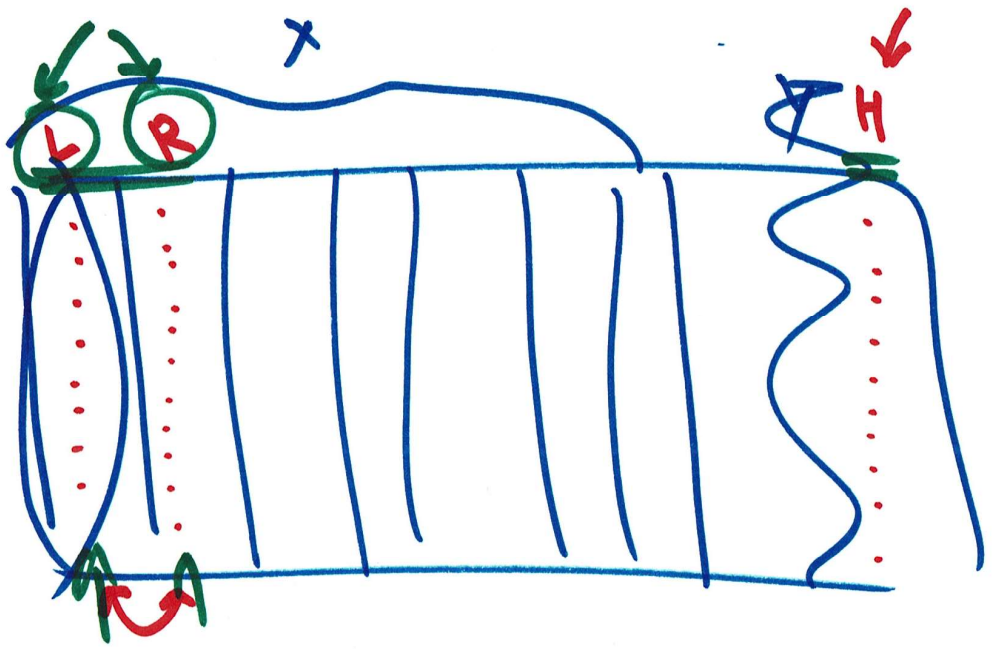


Homoscedasticity

- If the variance is not equal for the residuals across the regression line, then the data is said to be heteroscedastic.
- In this case the residuals can form a funnel shape or any other non symmetrical shape.
- Identifying the cause of hetroscedasticity is usually the best way to reason out ways to fix it



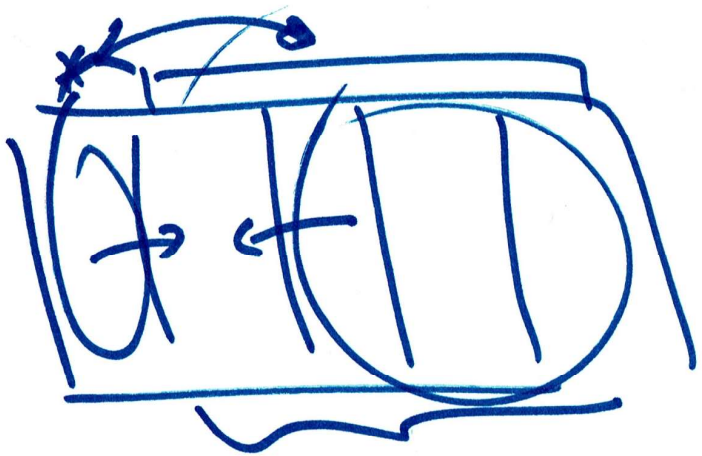
- Statistical test: The Goldfeld–Quandt test



① $H = 31.52 + 3.197\,L$ ←

② $H = 31.55 + 3.195\,R$ ←

③ $H = \underline{31.76} + \underbrace{6.82}_{\uparrow}\,R + \underbrace{(-3.65)}_{\uparrow}\,L$

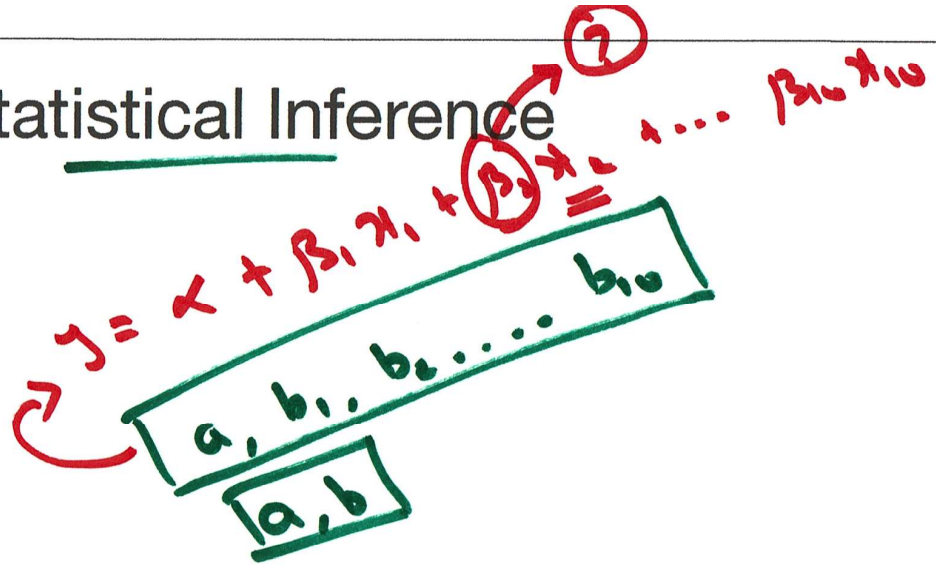


$R_1^2 \rightarrow VIF_i = \frac{1}{1-R_i^2}$ ↑ ∞ ↓ 1

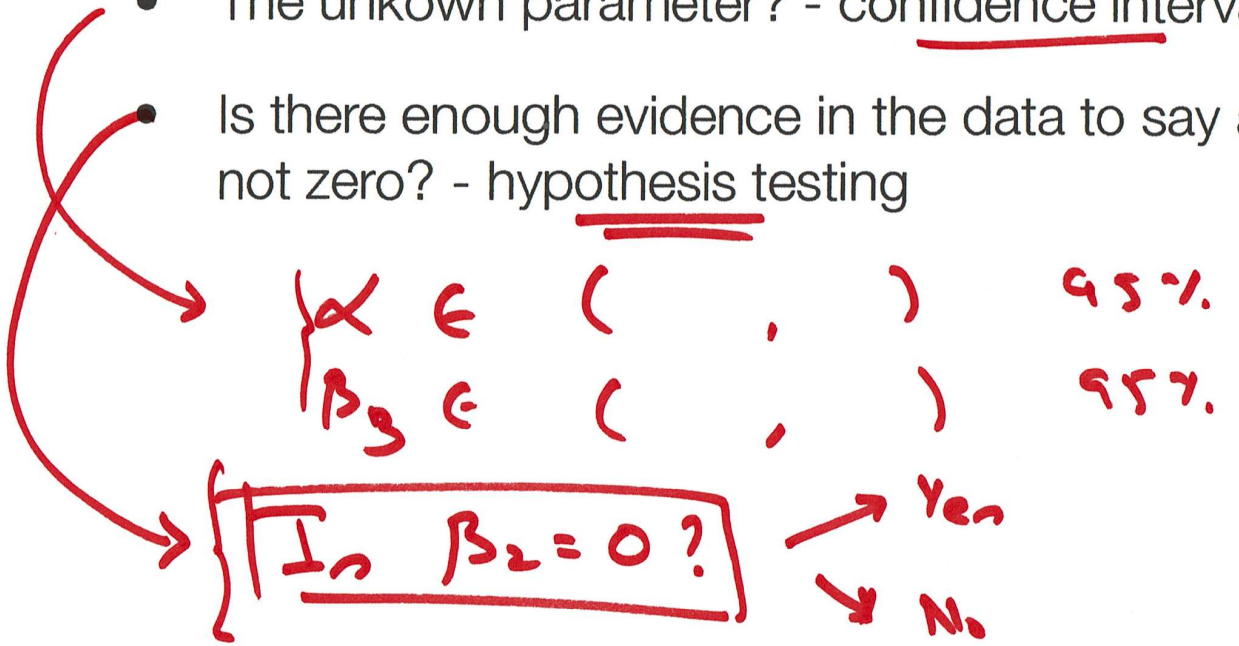
R_2^2

R_3^2

Statistical Inference



- Given the best estimates from the data, what can we say about the unkown true model?
  - The unkown parameter? - confidence interval
  - Is there enough evidence in the data to say a coefficient is not zero? - hypothesis testing



Reviewing Linear Regression

OLS Regression Results

Dep. Variable:	mpg	R-squared:	0.814
Model:	OLS	Adj. R-squared:	0.809
Method:	Least Squares	F-statistic:	147.3
Date:	Wed, 09 Dec 2020	Prob (F-statistic):	1.20e-93
Time:	12:48:42	Log-Likelihood:	-734.21
No. Observations:	278	AIC:	1486.
Df Residuals:	269	BIC:	1519.
Df Model:	8		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-18.2835	5.549	-3.295	0.001	-29.209	-7.358
cylinders	-0.3948	0.423	-0.933	0.352	-1.228	0.439
displacement	0.0289	0.010	2.870	0.004	0.009	0.049
horsepower	-0.0218	0.016	-1.330	0.185	-0.054	0.010
weight	-0.0074	0.001	-8.726	0.000	-0.009	-0.006
acceleration	0.0619	0.118	0.524	0.601	-0.171	0.295
model year	0.8369	0.064	13.149	0.000	0.712	0.962
origin_america	-3.0013	0.704	-4.262	0.000	-4.388	-1.615
origin_asia	-0.6060	0.705	-0.860	0.391	-1.994	0.782

Omnibus:	13.244	Durbin-Watson:	2.244
Prob(Omnibus):	0.001	Jarque-Bera (JB):	16.958
Skew:	0.386	Prob(JB):	0.000208
Kurtosis:	3.932	Cond. No.	8.26e+04

$a \quad b_1 \quad b_2 \quad b_3$

$mpg = -18.28 - 0.39(cyl) + 0.289(disp) + \ldots$

$b_1 = -0.39$

$mpg = \alpha + \beta_1(cyl) + \beta_2(disp) + \ldots$?

~~$\beta_2 = (-0.059$~~

$\beta_2 = (-1.228, 0.439)$ 95%.

$\beta_3 = (0.009, 0.049)$

In $\beta_2 = 0$?