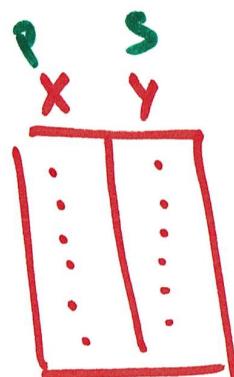


## Learn From Data



learn →

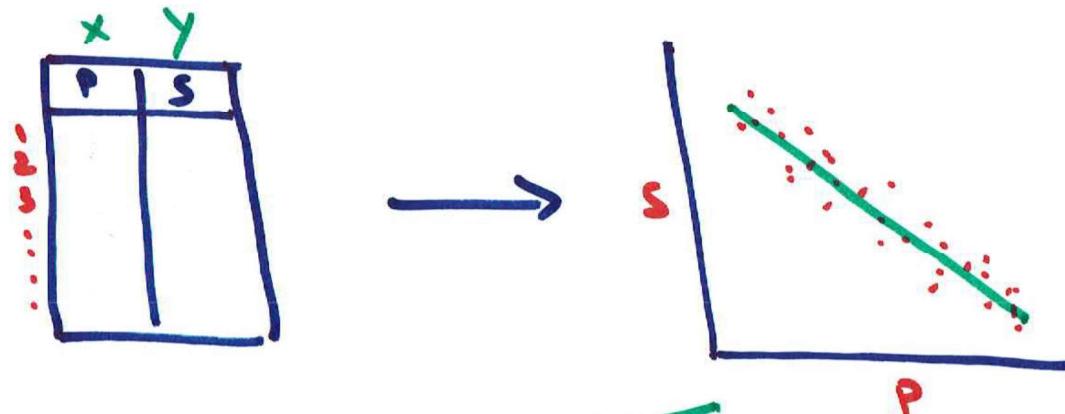
→ Prediction  
→ Interpretation  
Understand  
the world

Math model

$$\rightarrow S = 1000 - 20P$$
$$S = 100 + 20P - 30P^2$$
$$S = 10 e^{5P + 6P^2}$$

$S \Rightarrow$

$S \Rightarrow$



→ find the **best** 'S' line

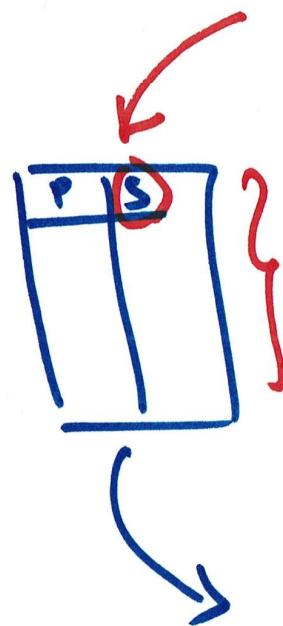
→ find the **best**  $a, b$  in

$$\hat{y}_i = \underline{\underline{a}} + \underline{\underline{b}} x_i$$

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

→ find  $\underline{\underline{a}}, \underline{\underline{b}}$  s.t.  $\min \sum_i (y_i - (a + b x_i))^2$

Stat



Assumption

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

Data Noise  
Generating Model

find  $a, b \Rightarrow \hat{y} = \hat{\alpha} + \hat{\beta} x$

Stat / fit

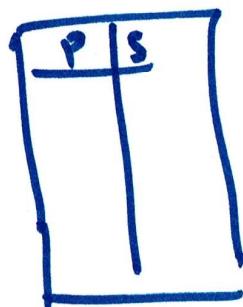
$\hat{\alpha}, \hat{\beta}$

↳ What can we say  
about  $\alpha, \beta$ ?

$\left\{ \begin{array}{l} C1 \rightarrow \alpha \in (1000 - 10, 1000 + 10) \\ H2 \rightarrow \text{Is } \beta \text{ really zero or not?} \end{array} \right.$

ML

No Assumption on the DNN



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

$$y = a + bx + cx^2$$

$$y = a e^{bx}$$

$$y \Rightarrow$$

$$y = \sqrt{a + bx + cx^2}$$

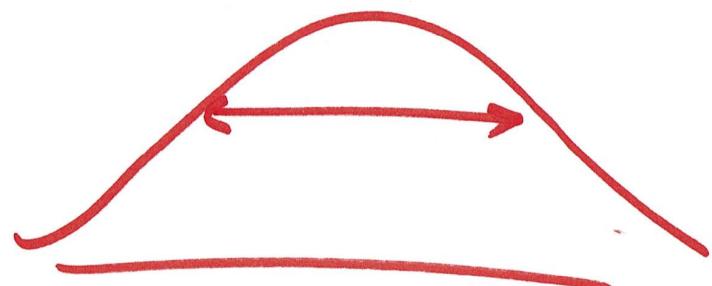
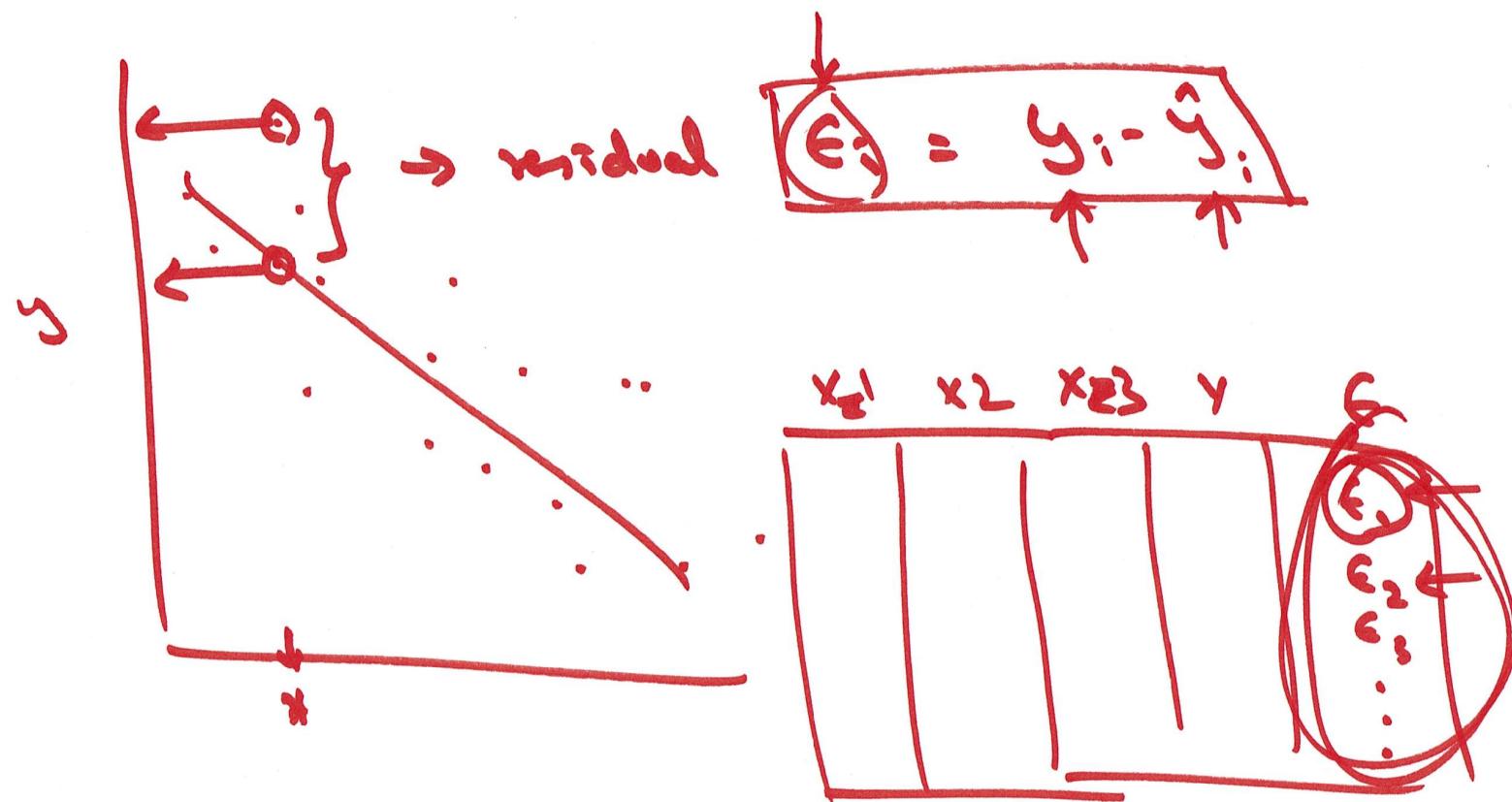
# Understanding the world (from Data)

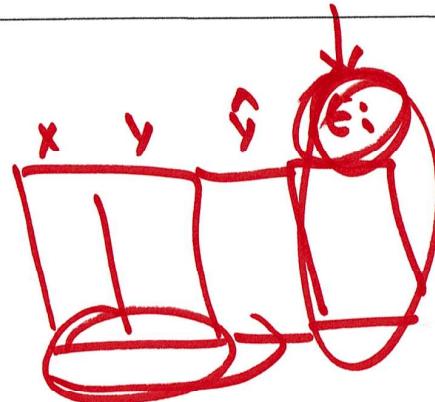
Stats ( $\sim 300$  yrs)

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

ML ( $\sim 30$  yrs)

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

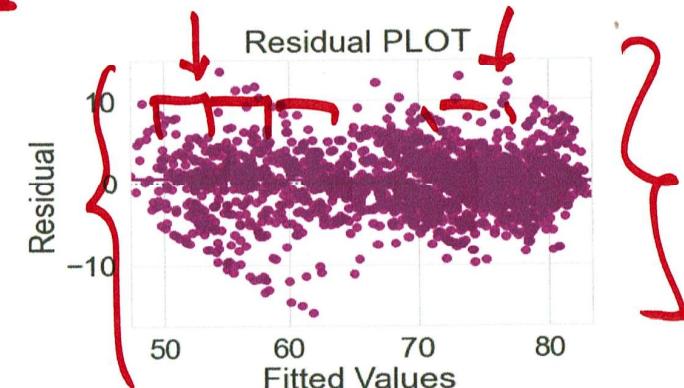
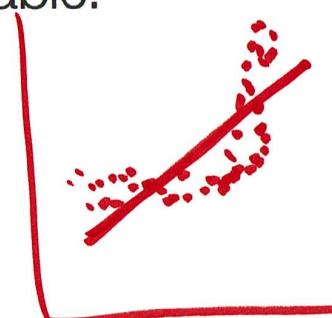
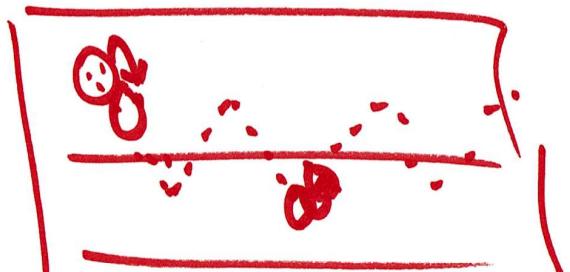




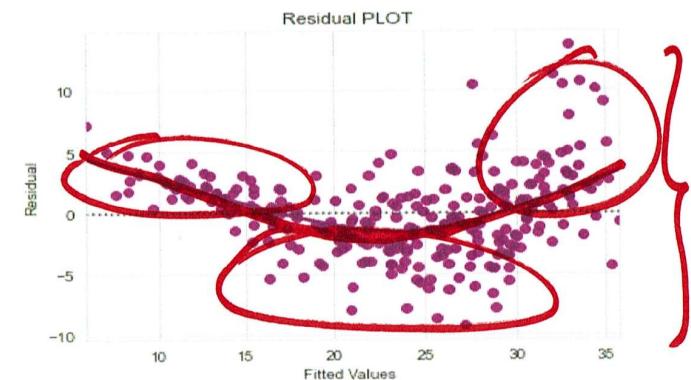
## Linear Relationship

$$y = \alpha + \beta x + \epsilon$$

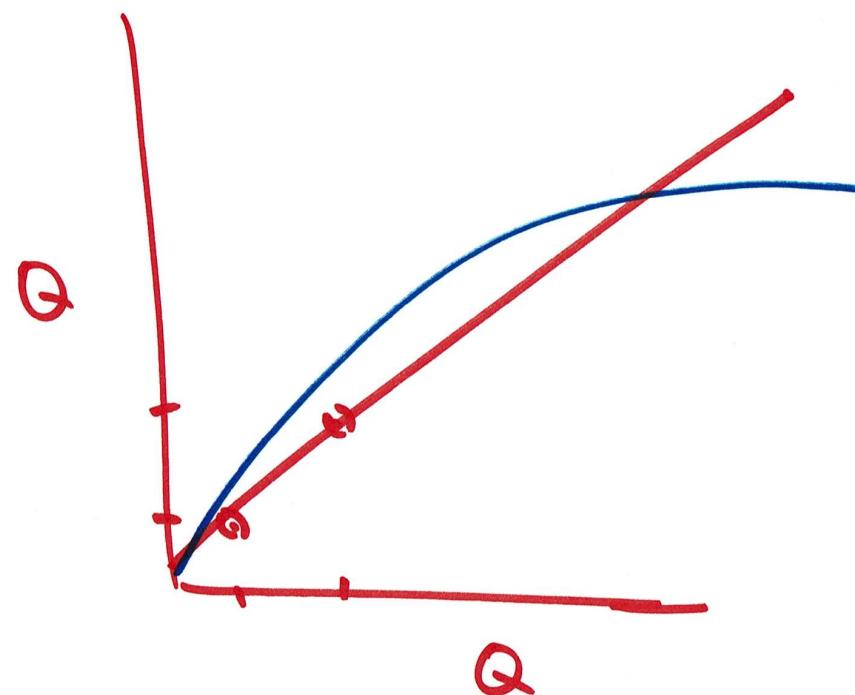
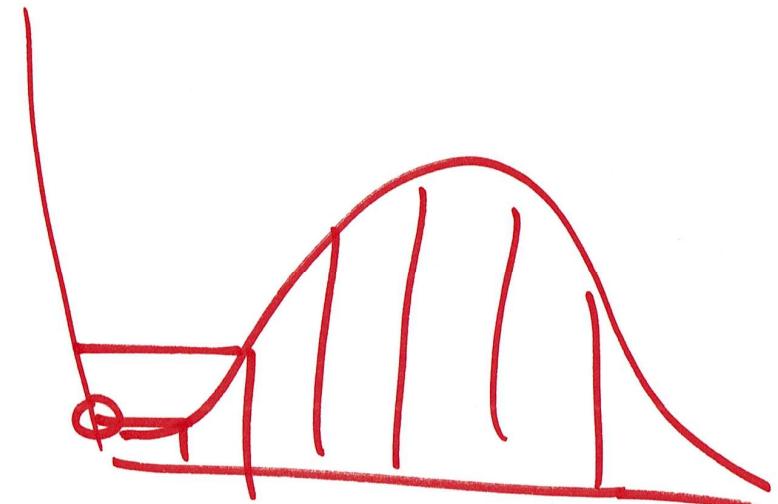
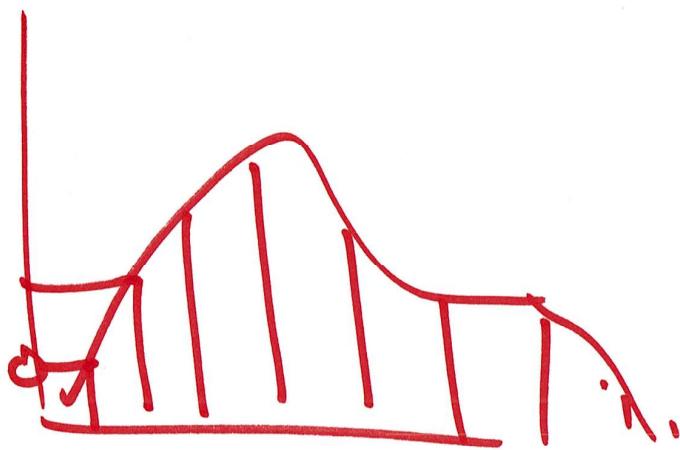
- After finding the best linear fit, a plot of the residuals will provide a good insight.
- If they don't follow any pattern, we say that the model is linear otherwise model is showing signs of non-linearity
- To deal with non-linearity, we can try transforming variables as per their relationship with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



No pattern



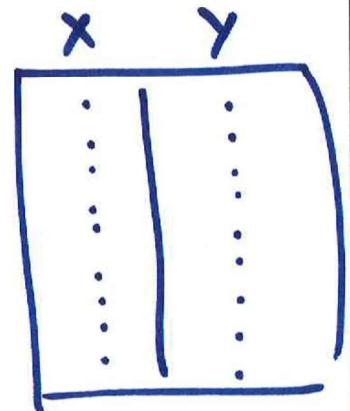
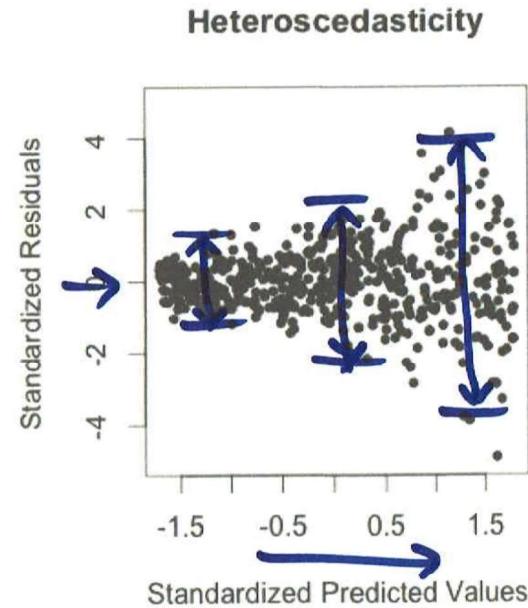
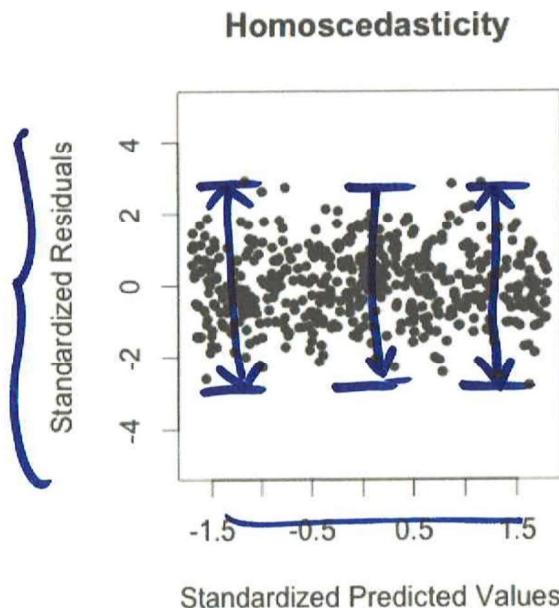
Some non-linearity



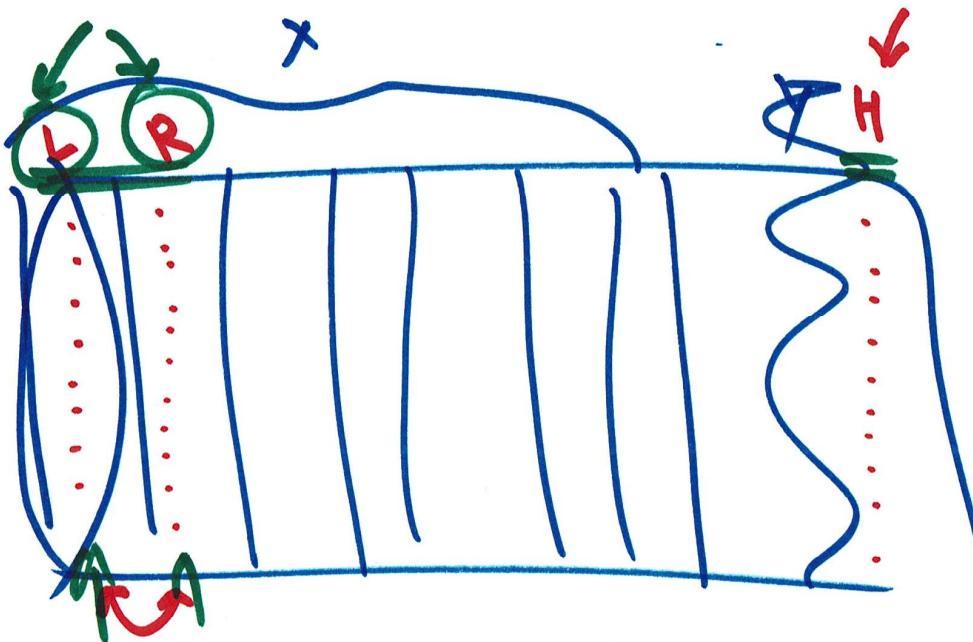
This file is meant for personal use by jemanuel.perez@gmail.com only.  
Sharing or publishing the contents in part or full is liable for legal action.

# 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 heteroscedasticity is usually the best way to reason out ways to fix it



- Statistical test: The Goldfeld–Quandt test



1

$$H = 31.52 + \underline{3.197} L \quad \{ \leftarrow$$

$$H = 31.55 + 3.195 R \quad \{$$

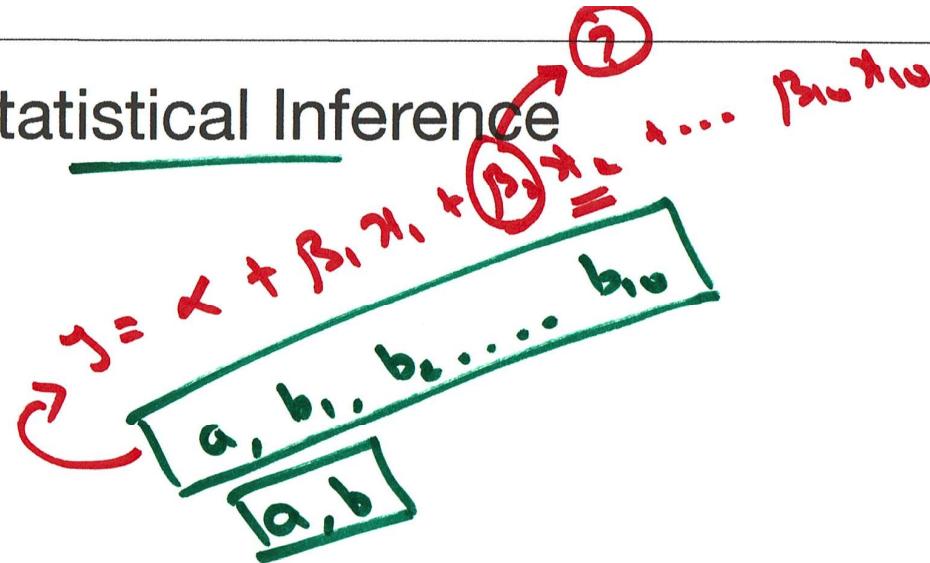
$$\left\{ \begin{array}{l} H = \frac{31.76}{1} + \frac{6.82}{R} + \frac{(-3.65)}{L} \end{array} \right.$$

$$VIF_i = \frac{1}{1 - R_i^2}$$

$R_1^2$   
 $R_2^2$   
 $R_3^2$   
 $\vdots$   
 $\vdots$   
 $\vdots$

$\uparrow^\infty$   
 $\downarrow,$

## 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

$$\begin{aligned} \alpha &\in ( , ) \quad 95\% \\ \beta_3 &\in ( , ) \quad 99\% \end{aligned}$$

$\left\{ \begin{array}{l} T \text{ vs } \beta_2 = 0 ? \\ \text{vs } \epsilon \end{array} \right.$

# 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			

$$b_1 = \underline{\underline{-0.39}}$$

?

$$m_{ij} = \alpha + \beta_1(c_{ij}) + \underline{\underline{\beta_2(d_{ij})}} + \dots$$

~~$$\beta_2 = \underline{\underline{-0.51}}$$~~

$$\underline{\underline{\beta_2}} = (-\underline{\underline{1.228}}, \underline{\underline{0.439}})$$

95%

$$\underline{\underline{\beta_3}} = (\underline{\underline{0.009}}, \underline{\underline{0.049}})$$

In  $\beta_2 = 0$ ?

I<sub>2</sub>  ~~$\beta_2 = 0$~~ ? Reject

$$b_2 = \frac{0.0289}{\text{S.E.}} \Rightarrow \boxed{0.004} \quad P\text{-value}$$

H<sub>0</sub>  
 $\beta_2 \neq 0$

---

I<sub>2</sub>  $\boxed{\beta_2 = 0?} \checkmark$  Accept

$$b_1 = -0.3948 \quad P = 0.352$$