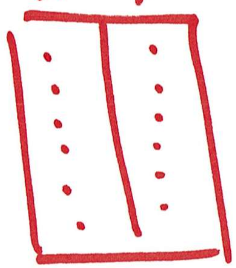
Learn from Data

P S

X Y



$\xrightarrow{\text{learn}}$ Understand the world

→ Prediction

→ Interpretation

Math model

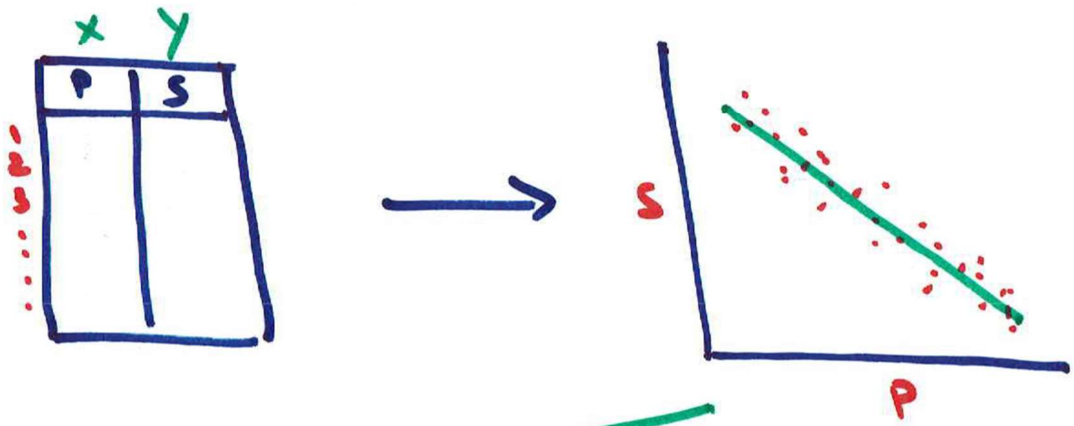
→ $S = 1000 - 20p$

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

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

$S \Rightarrow$ 

$S \Rightarrow$ 



Find the $\boxed{\text{best}}$ 'St' line

Find the $\boxed{\text{best}}$ 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

obs

$$\underbrace{y_i = \alpha + \beta x_i}_{\text{Data Generating Model}} + \underbrace{\epsilon_i}_{\text{Noise}}$$

P	S

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

Stat. Inf.

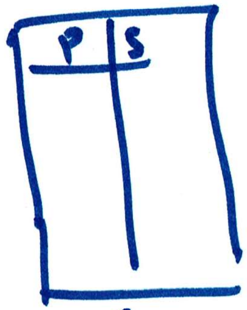
① $\hat{a}, \hat{b}$

↳ What can we say about α, β?

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

ML

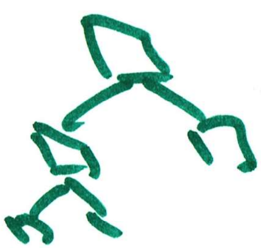
No Assumption on the Data



Find a, b ⇒ $\hat{y} = a + b x$

$y = a + bx + cx^2$

$y = a e^{bx}$

y ⇒ 

y = 

y = 

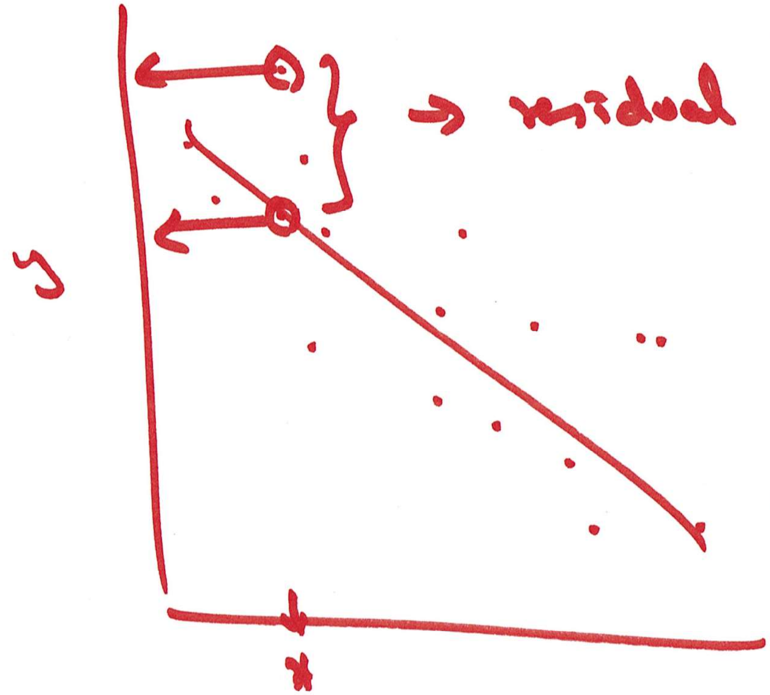
Understanding the world
(From Data)

Stats (~ 300 yrs)

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

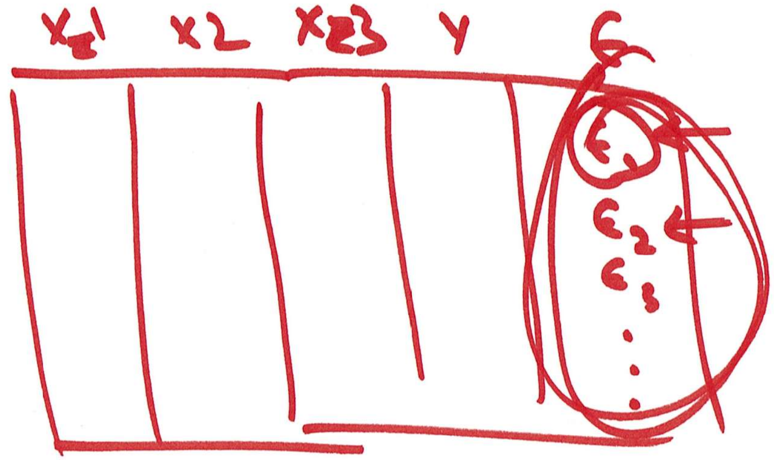
ML (~ 30 yrs)

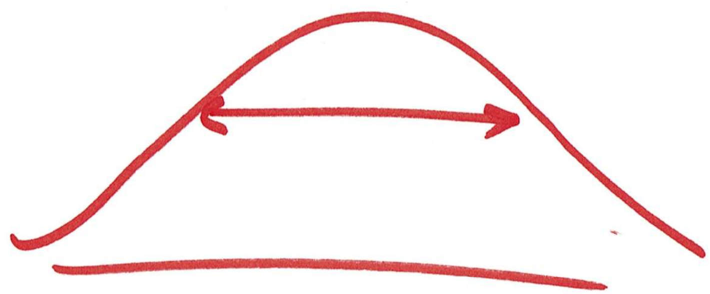
- No Assumption on DGM
- Train Vs Test
- No Stat. Inf.
- Field of CS
  - learn
  - weight
  - feature

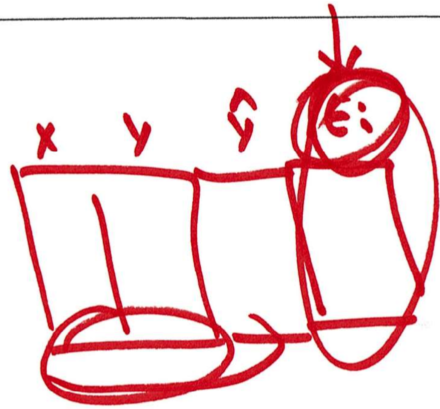


→ residual

$\epsilon_i = y_i - \hat{y}_i$



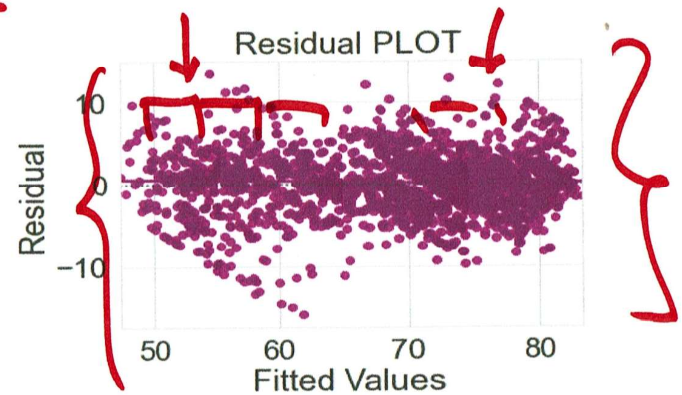




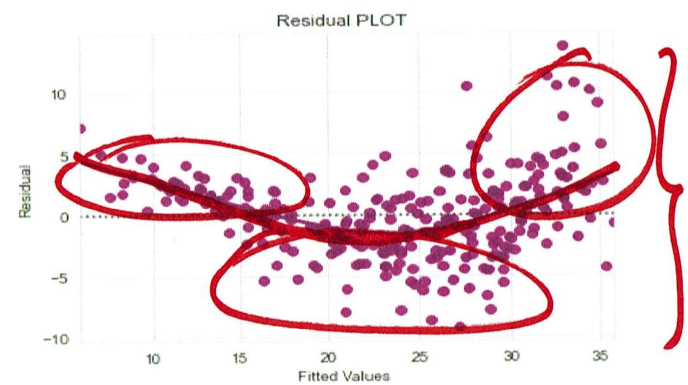
Linear Relationship

$y = a + bx + cx^2$

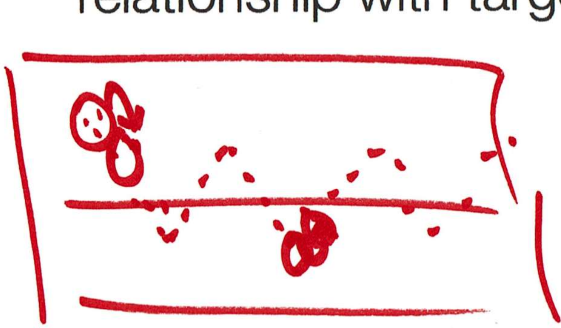
- After finding the best linear fit, a plot of the residuals will provide a good insight.
- If they don't follow any pattern, we say that the model is linear otherwise model is showing signs of non-linearity
- To deal with non-linearity, we can try transforming variables as per their relationship with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

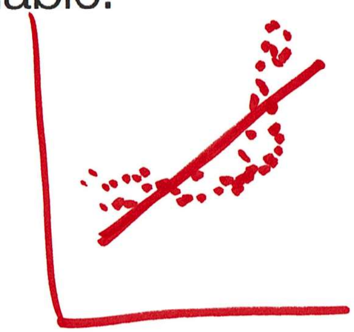


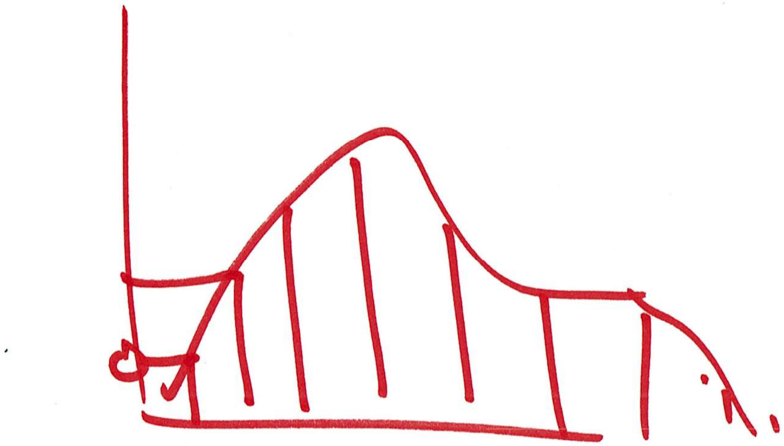
No pattern

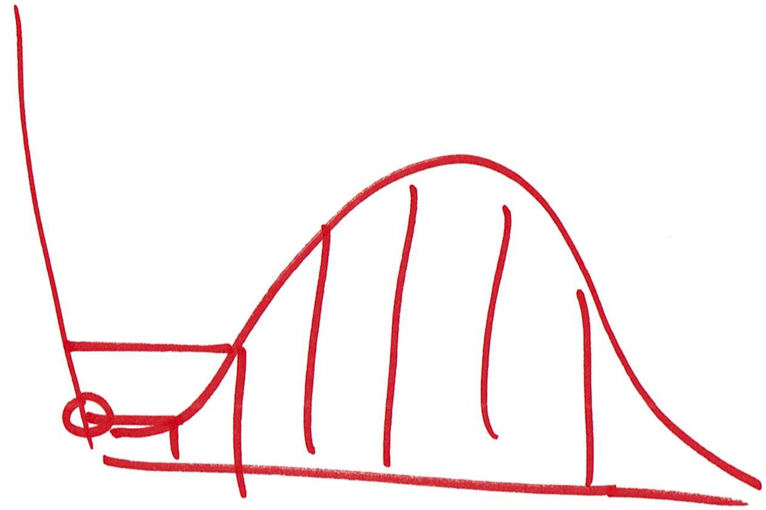


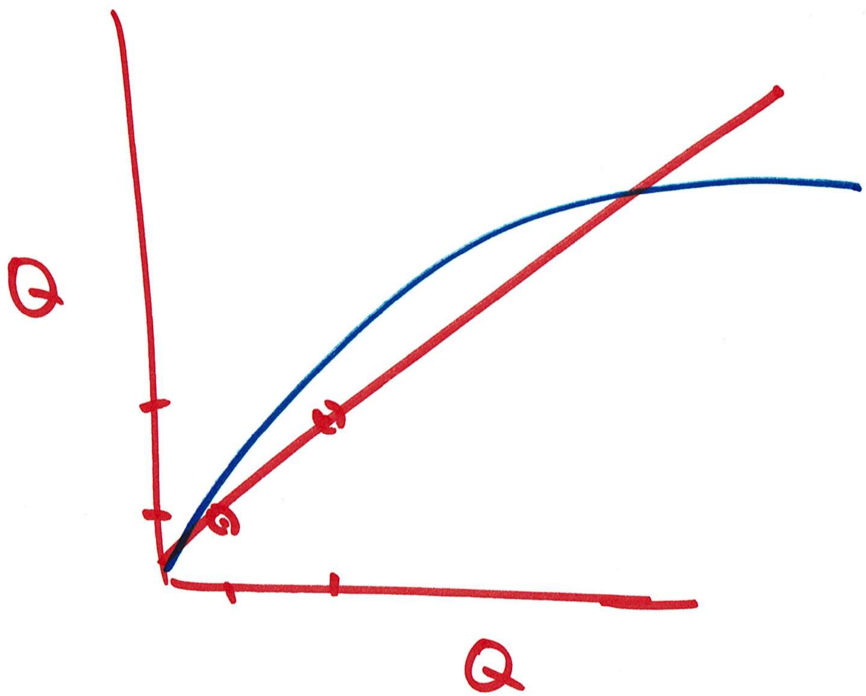
Some non-linearity





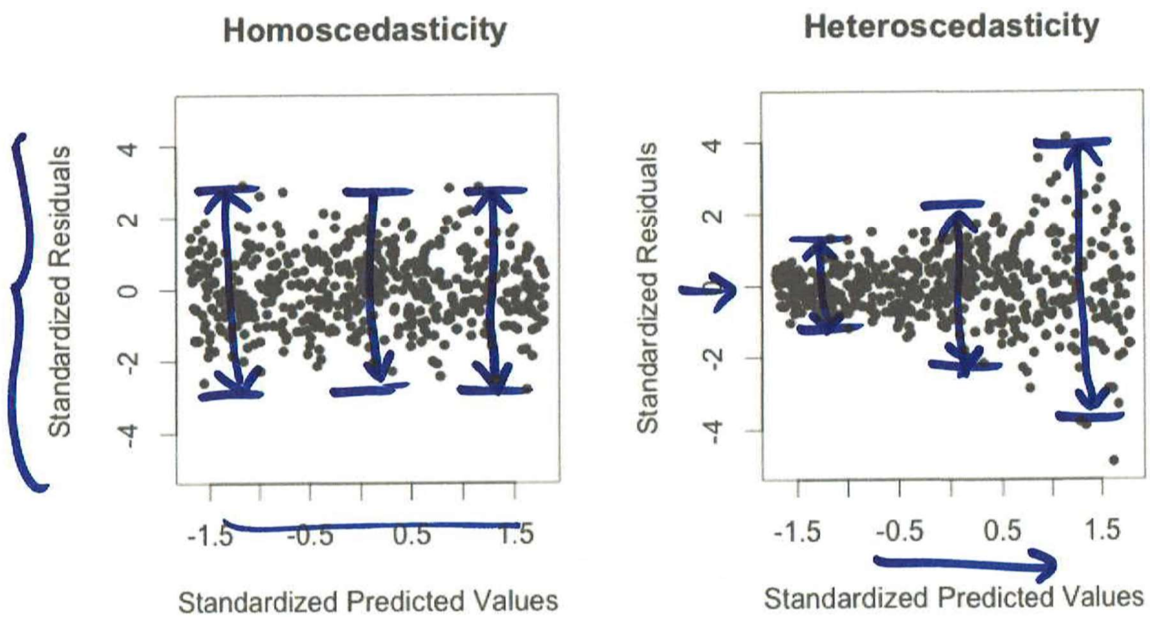


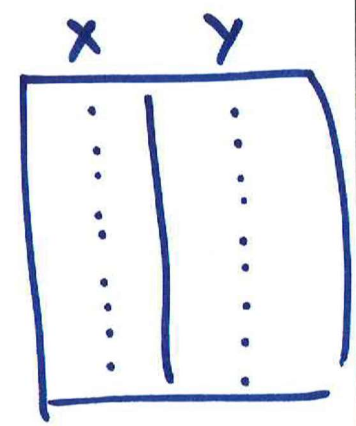



Q
Q

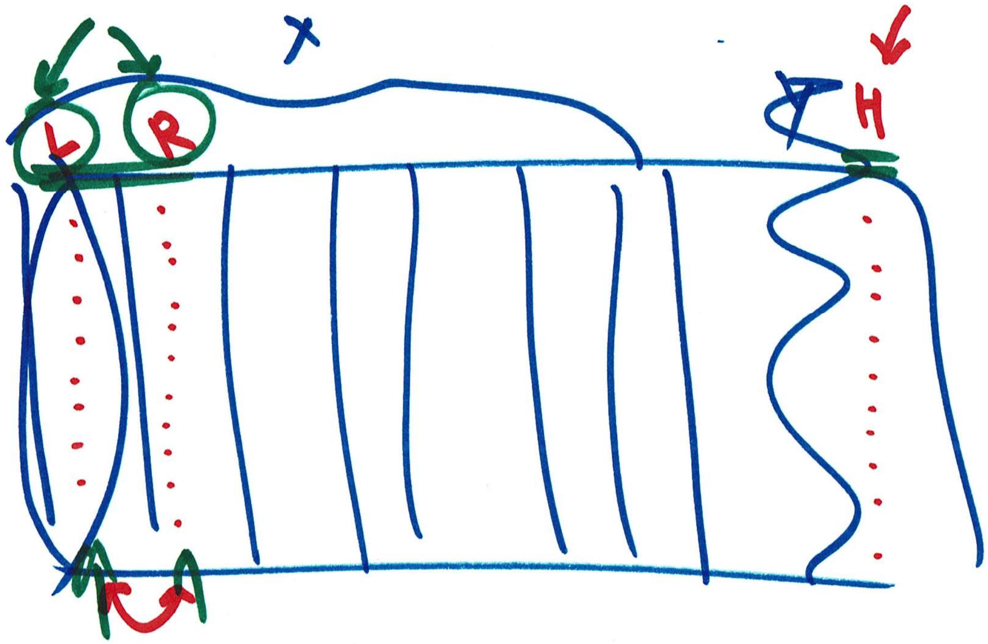
Homoscedasticity

- 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





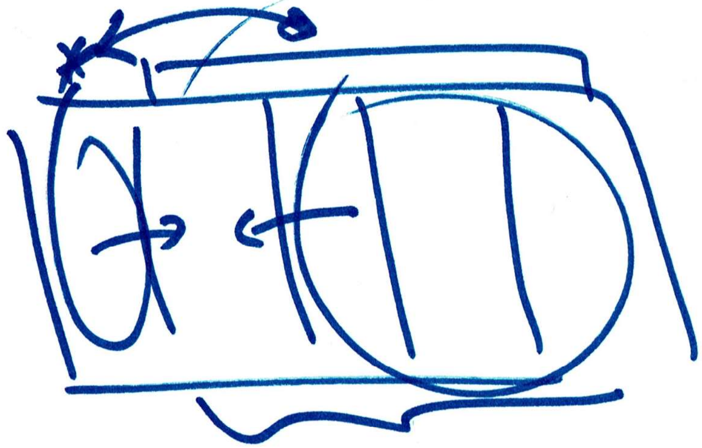
- Statistical test: The Goldfeld–Quandt test



① $H = 31.52 + 3.197\,L$

② $H = 31.55 + 3.195\,R$

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



$\rightarrow R_1^2 \rightarrow \quad VIF_i = \dfrac{1}{1-R_i^2} \quad \uparrow \infty \quad \downarrow 1$

R_2^2

R_3^2

$\vdots$

Statistical Inference

$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \dots \beta_{10} x_{10}$?

$\boxed{a, b_1, b_2 \dots b_{10}}$

$\boxed{a, b}$

- 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

$\alpha \in (\quad , \quad)$ 95%

$\beta_3 \in (\quad , \quad)$ 95%

$\boxed{\text{Is } \beta_2 = 0\ ?}$ → Yes / No

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) + ...

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

?

$$mpg = \alpha + \beta_1(cyl) + \underline{\underline{\beta_2}}(disp) + \ldots$$

$\cancel{\beta_2 = -0.39}$

$\underline{\underline{\beta_2}} = (\underline{\underline{-1.228}}, \underline{\underline{0.439}})$ $\underline{\underline{95\%}}$

$\beta_3 = (0.009, 0.049)$

Is $\beta_2 = 0$?

H_0: $\cancel{\beta_2 = 0}$? Reject

$b_2 = \underline{\underline{0.0289}} \Rightarrow \boxed{0.004}$ P-value

$\Downarrow$

$\beta_2 \neq 0$

H_0: $\boxed{\beta_1 = 0?}$ ✓ Accept

$b_1 = -0.3948$ $P = 0.352$