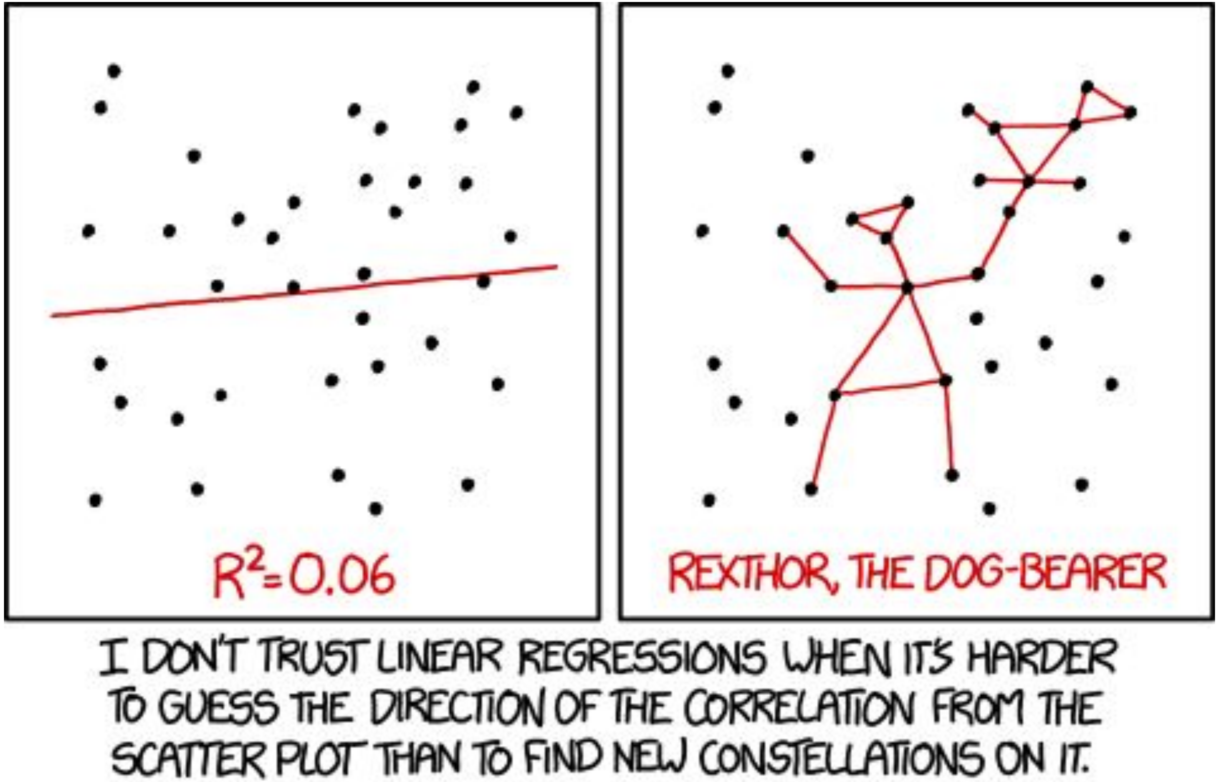
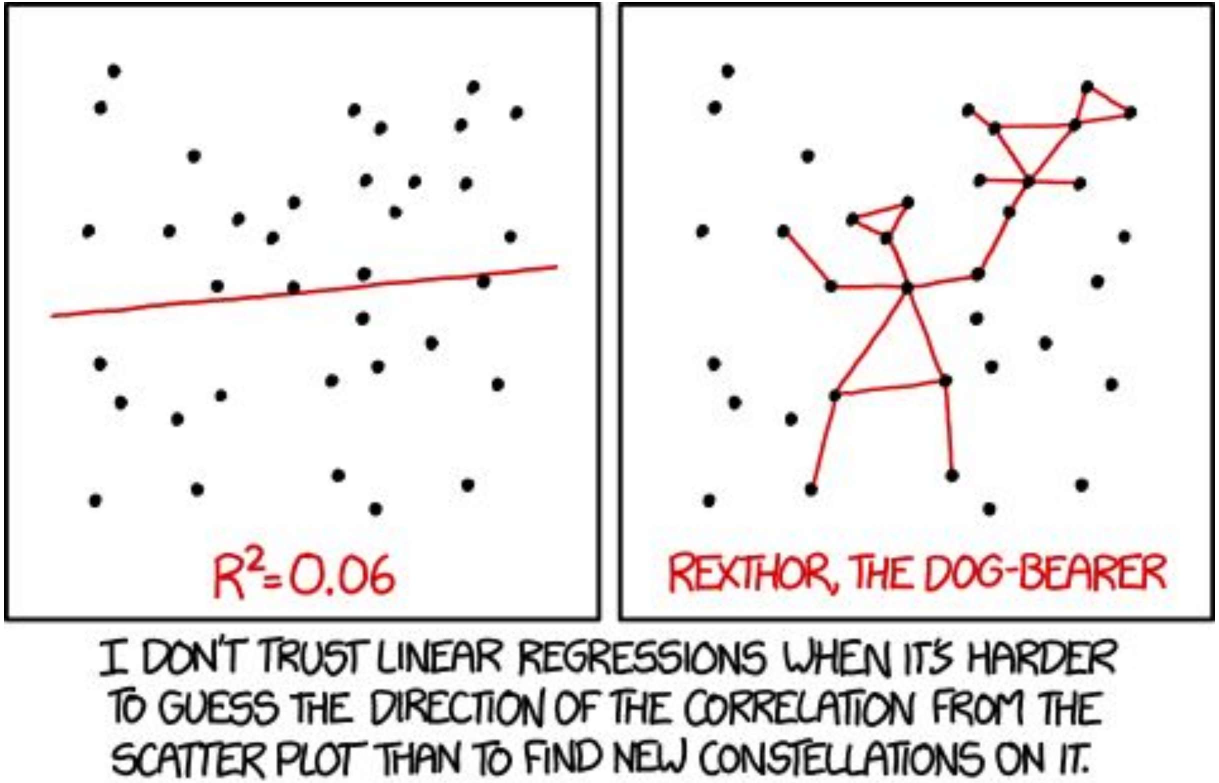
Lecture 21: Inference in Regression



Announcements and reminders

- HW 5 due next Friday



Simple linear regression (SLR)

Given data $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$, fit a simple linear regression of the form

$$y_i = \alpha + \beta x_i + \epsilon_i \quad \text{where} \quad \epsilon_i \sim N(0, \sigma^2)$$

Estimates of the intercept and slope parameters are given by minimizing

$$SSE = \sum_{i=1}^{n} (y_i - (\alpha + \beta x_i))^2$$

The least-squares estimates of the parameters are:

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} \quad \text{and} \quad \hat{\beta} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

Game plan

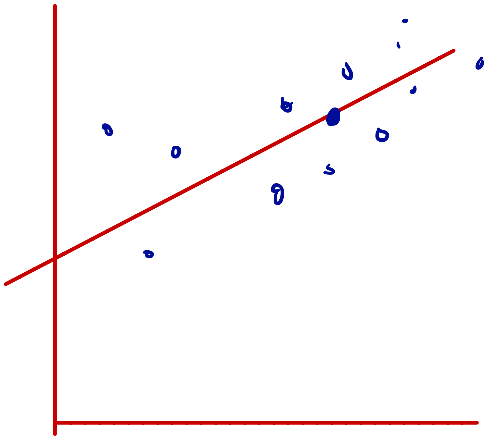
Today, we'll see how we can…

- Estimate the variance in data about the true regression line
- Quantify the goodness-of-fit in our SLR model
- Perform inference on the regression parameters

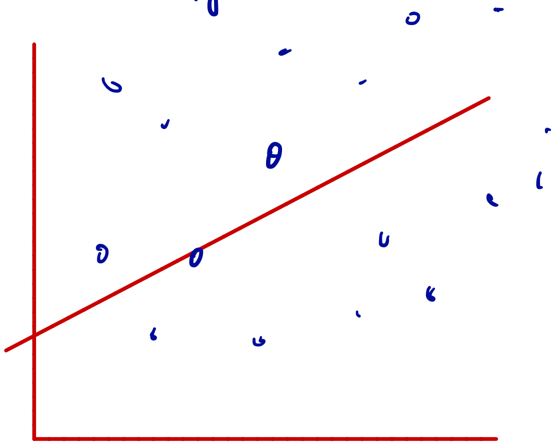
Estimating the variance

The parameter σ^2 determines the spread of the data about the true regression line

low σ^2



high σ^2



Estimating the variance

An estimate of σ^2 will be used in computing confidence intervals and doing hypothesis testing on the estimated regression parameters.

What does this mean?

We want answers to questions like: $Y = \alpha + \beta x$

- Is the slope $\beta \neq 0$?
- Is the intercept $\alpha > 0$?

Mean term = bias

Estimating the variance

An estimate of σ^2 will be used in computing confidence intervals and doing hypothesis testing on the estimated regression parameters.

$\hat{y}_i = \hat{\alpha} + \hat{\beta} x_i$

Recall that the sum of squared errors is given by:

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

Our estimate of the variance $\hat{\sigma}^2$ is given by:

σ^2 = true population var

$\hat{\sigma}^2$ = est. of σ^2

$$\boxed{\hat{\sigma}^2 = \frac{SSE}{n-2}}$$

Estimating the variance

An estimate of σ^2 will be used in computing confidence intervals and doing hypothesis testing on the estimated regression parameters.

Recall that the sum of squared errors is given by:

$$SSE = \sum_{i=1}^{n} (y_i - (\alpha + \beta x_i))^2$$

Our estimate of the variance $\hat{\sigma}^2$ is given by:

$$\hat{\sigma}^2 = \frac{SSE}{n-2}$$

↳ we lose two degree of freedom

Degrees of freedom (df) is reduced **by two** in denominator for $\hat{\sigma}^2$ … why?

had to override α & β to get SSE

Estimating the variance

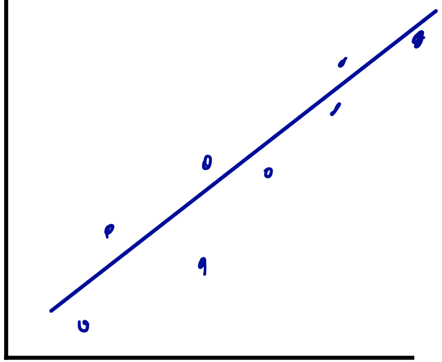
Degrees of freedom (df) is reduced **by two** in denominator for $\hat{\sigma}^2$ … why?

- Estimating each parameter requires one degree of freedom
- We had to estimate α and β first → loss of 2 df

The coefficient of determination

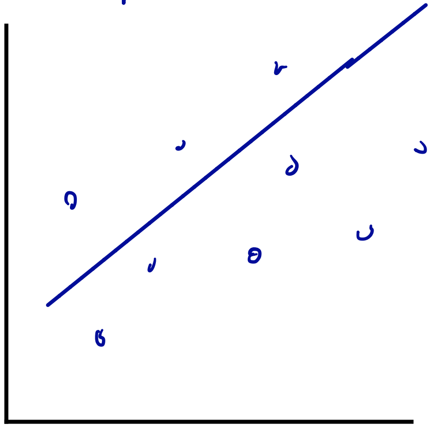
- The **coefficient of determination**, R^2, quantifies how well the model explains the data
- R^2 is a value between 0 and 1

$R^2 \approx 1$



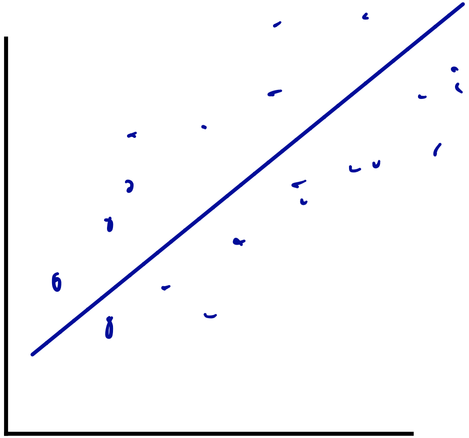
good response

$R^2 \approx 0$

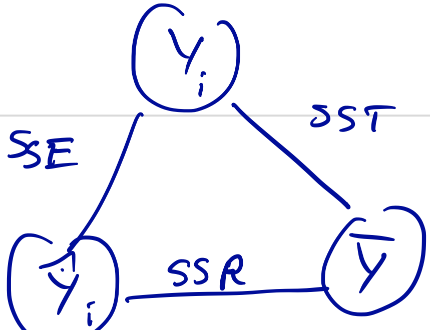


not well response

R^2 somewhere in b/w data.



The coefficient of determination



The **sum of squared errors** $SSE = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$

can be thought of as a measure of how much variation in Y is left unexplained by the model.
-- that is, how much cannot be attributed to a **linear** relationship

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

The **regression sum of squares** is given by
↳ give sense of how variation in y is explained by our mode

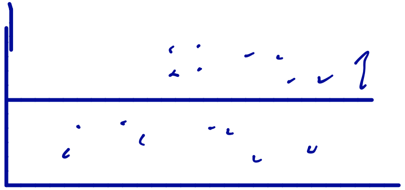
A quantitative measure of the total amount of variation in observed Y values is given by the **total sum of squares**:

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

Intuition: SST is what we would get for SSE if we just used the **mean** of the data as our model

The coefficient of determination

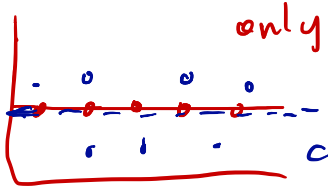
The sum of squared deviations about the least-squares line is smaller than the sum of squared deviations about any other line.



→ SSE ≤ SST

→ **Concept check:** when are they equal?

$\hat{Y}_i = \bar{Y}$



only situation where $\hat{Y}_i = \bar{Y}$? NO! come up to same horizontal slop

The ratio SSE/SST is the proportion of total variation in the data (SST) that cannot be explained by the SLR model (SSE). So we define the **coefficient of determination** R^2 to be the proportion that *can* be explained by the model:

✓ FYOG!

Form fact: $SST = SSR + SSE$

$$\hookrightarrow \frac{SST}{SST} = \frac{SSR}{SST} + \frac{SSE}{SST}$$

$$\Rightarrow 1 = R^2 + \frac{SSE}{SST}$$

$$\boxed{R^2 = 1 - \frac{SSE}{SST}}$$

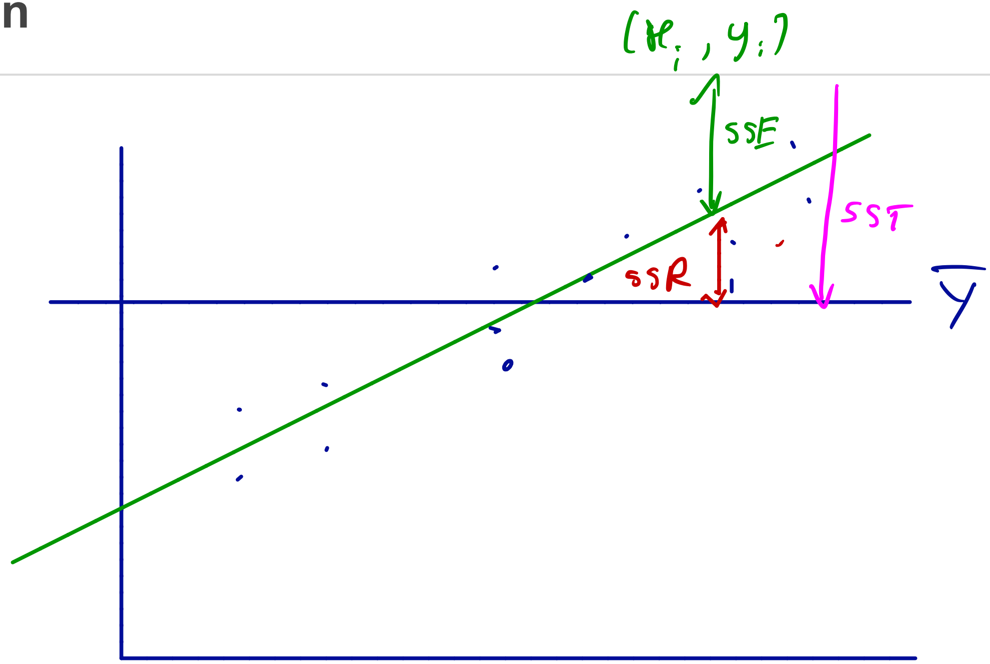
The coefficient of determination

$$SSE = \sum (y_i - \hat{y}_i)^2$$

$$SSR = \sum (\hat{y}_i - \bar{y})^2$$

$$SST = \sum (y_i - \bar{y})^2$$

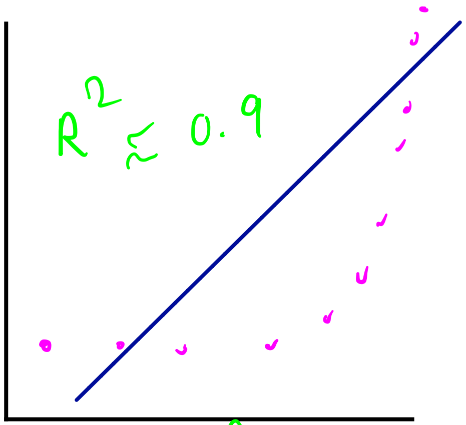
$$SST = SSR + SSE$$

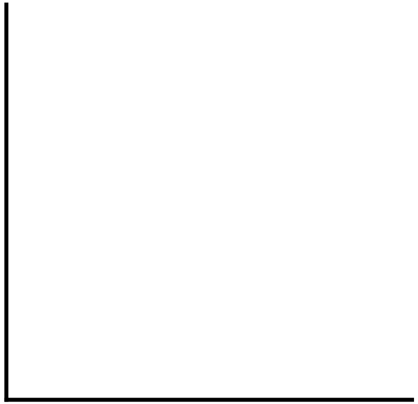


The coefficient of determination

Warning! R^2 is the proportion of total variation in the data that is explained by the model

It does **not** tell you that you necessarily have the correct model.





probably better to fit a quadratic

Inference about SLR parameters

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

$Y = \alpha + \beta x + \varepsilon$

- The parameters in the simple linear regression model have distributions.

- From these distributions, we can construct CIs for the parameters, conduct hypothesis tests, and all that groovy stuff.

$$\boxed{\sigma^2 = \frac{SSE}{n-2}}$$

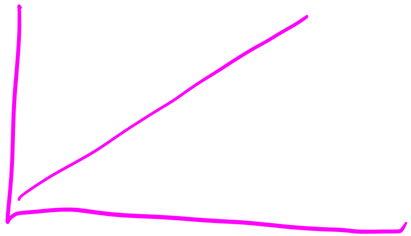
- We will focus mainly on the slope parameter β
    - β allows us to ask/answer questions like:
      Is there really a relationship between the feature and the response?

    - The distribution for the estimate of the slope is given by:

Inference about SLR parameters

- The parameters in the simple linear regression model have distributions.

- From these distributions, we can construct CIs for the parameters, conduct hypothesis tests, and all that groovy stuff.



- We will focus mainly on the slope parameter β
  - β allows us to ask/answer questions like:
    Is there really a relationship between the feature and the response?
  - The distribution for the estimate of the slope is given by:

$\sqrt{\frac{SSE}{n-2}}$

normal distribute

var

$$\hat{\beta} \sim N\left(\beta, \frac{\sigma^2}{\sum_i (x_i - \bar{x})^2}\right) \quad \rightarrow \quad SE(\hat{\beta}) = \frac{\hat{\sigma}}{\sqrt{\sum_i (x_i - \bar{x})^2}}$$

slop

Inference about SLR parameters

Then confidence intervals for β are given by:

t- distribution to get CI

$100(1-\alpha)\%$ CI for β :

SE ERR $(\hat{\beta})$

$$\hat{\beta} \pm t_{\alpha/2,\, df=n-2} \cdot \frac{\hat{\sigma}}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}}$$

where $\hat{\sigma}^2 = \frac{SSE}{n-2} = \left(\frac{\sum_{i=1}^{n}(Y_i-\hat{Y}_i)^2}{n-2}\right)$

get estimate from regression line

And hypothesis testing:

$H_0 : \beta = 0$

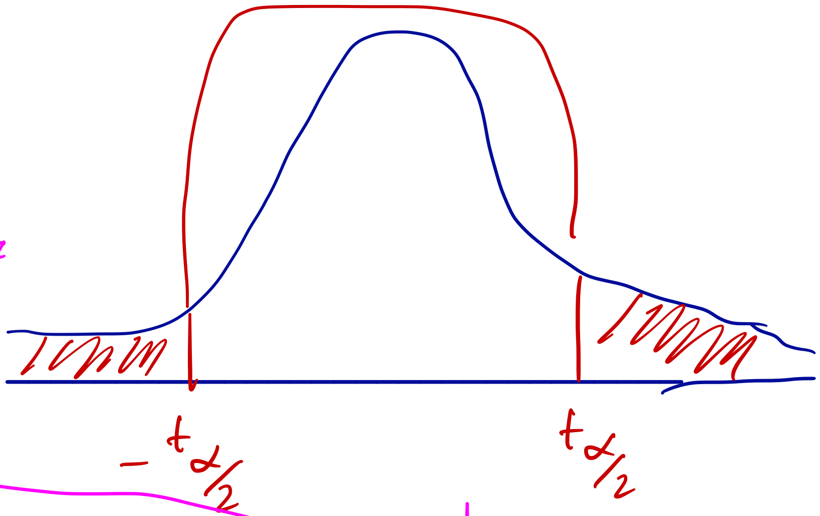
$H_1 : \beta \neq 0$ or $\beta > 0$ or $\beta < 0$

Inference about SLR parameters

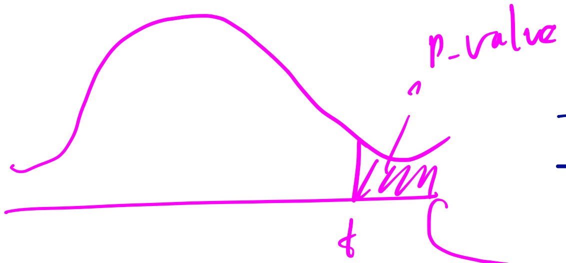
critical value

Then confidence intervals for β are given by:

100(1-α)% CI for β is $\hat{\beta} \pm t_{\alpha/2,n-2} \cdot SE(\hat{\beta})$



And hypothesis testing:



H_0: β = c
H_1: β ≠ c (or maybe something like β = c against β > c)

Test statistic: $t = \dfrac{\hat{\beta} - c}{SE(\hat{\beta})}$ → Compare to $t_{\alpha/2,\, n-2}$ or compute p-value

larger ts to favor, so p-value is on right

Concept check: What *t* critical value would we compare for the test of β = 0 against β > 0?

Inference about SLR parameters

Workflow: **Given data (x, y)...**

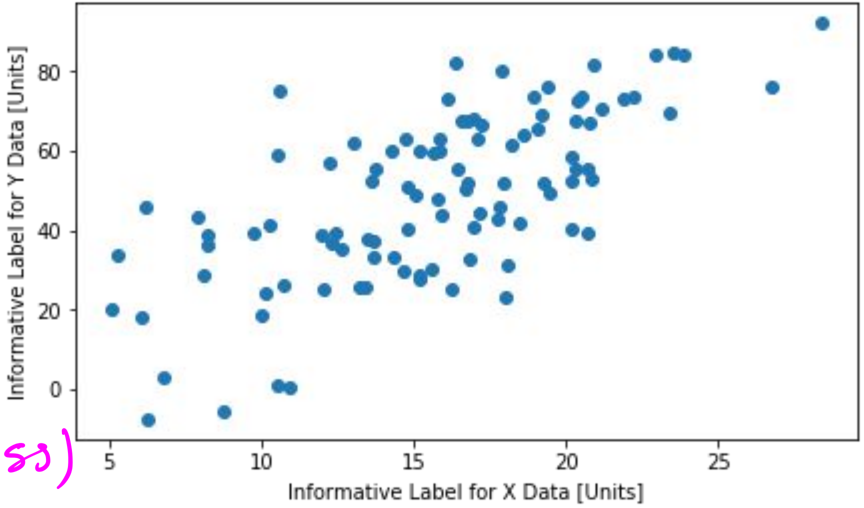
0.) Explore / plot / histograms,

1) Formular hypothesis

$H_0: \beta = 0$

$H: \beta > 0$

2) Fit regression line (stats. linregress)



3) compute CI or p-value / rejection regions & test stats for testing your hypothesis.

4) Make conclusions.

What just happened?

- **Inference for linear regression happened!**
    - How to estimate the variance, $\hat{\sigma}^2$
    - How to make inference for the slope parameter, β