CSCI 3022: Intro to Data Science

Lecture 21: Inference in Regression

Rachel Cox

Department of Computer Science

Announcements & Reminders

HW 5 - Due April 17th at 11:59 pm

What will we learn today?

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

- ❑ *A Modern Introduction to Probability and Statistics, Chapter 27.3*

Simple Linear Regression (SLR)

Given data $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, 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

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

SSE – sum of squared errors

The least-squares estimates of the parameters are:

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

from our book

Simple Linear Regression (SLR)

- $\hat{\beta} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$ ← from book

$$= \frac{\sum_{i=1}^{n}(x_i y_i - x_i\bar{y} - \bar{x}y_i + \bar{x}\bar{y})}{\sum_{i=1}^{n}(x_i^2 - 2x_i\bar{x} + (\bar{x})^2)}$$

$$= \frac{n\,\overline{xy} - n\,\bar{x}\bar{y} - n\,\bar{x}\bar{y} + n\,\bar{x}\bar{y}}{n\,\overline{x^2} - 2n(\bar{x})^2 + n\,(\bar{x})^2}$$

$$= \frac{n\,\overline{xy} - n\bar{x}\bar{y}}{n\overline{x^2} - n(\bar{x})^2}$$

$$= \frac{\overline{xy} - \bar{x}\bar{y}}{\overline{x^2} - (\bar{x})^2}$$

← we derived this in the last lecture

Estimating the variance

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

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:

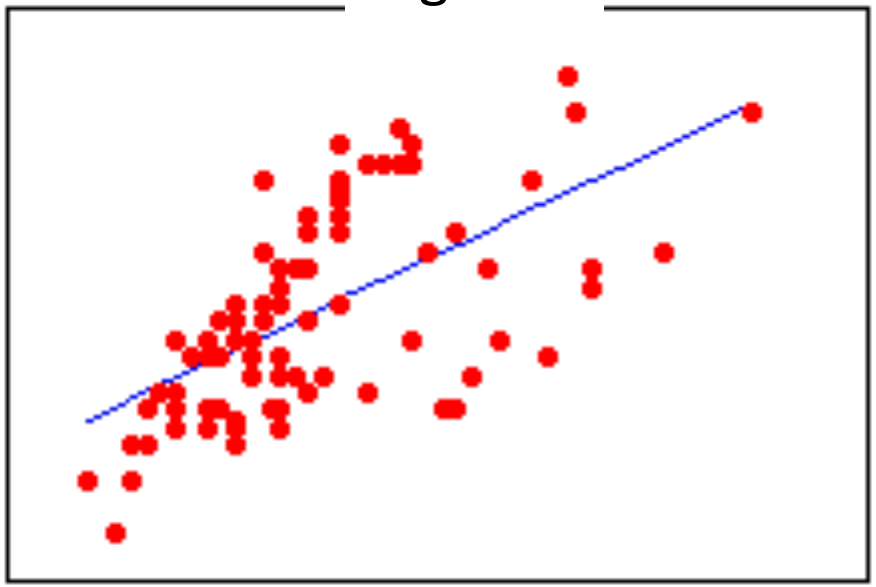
- $SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$

observed data, model. Because $\hat{y} = \hat{\alpha} + \hat{\beta}x$

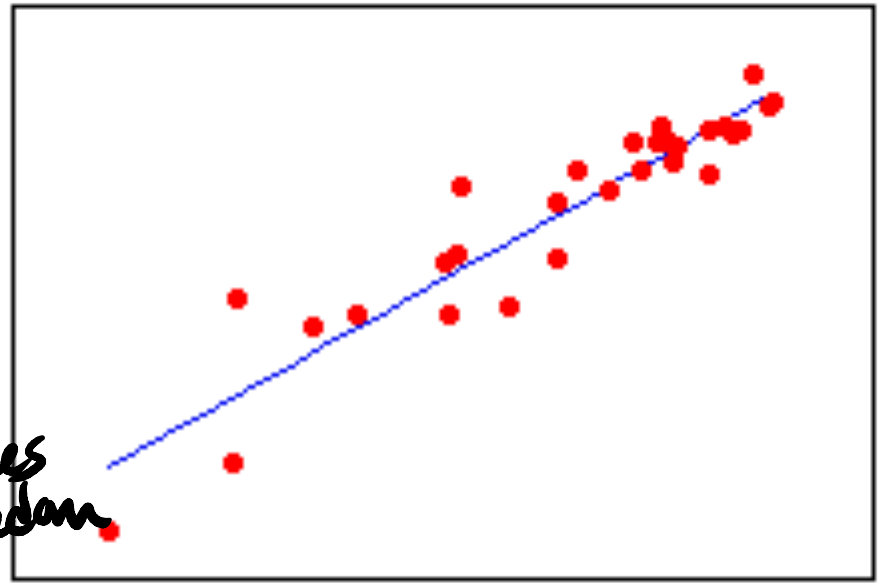
Our estimate or the variance $\hat{\sigma}^2$ is given by: $\hat{\sigma}^2 = \frac{SSE}{n-2}$

$n-2$ degrees of freedom

High σ^2



Low σ^2



The Coefficient of Determination - R^2

The coefficient of determination, R^2, quantifies how well the model explains the data.

- R^2 is a value between 0 and 1

The **sum of squared errors (SSE)** can be thought of as a measure of how much variation in Y is left unexplained by the model.

model mean

The **regression sum of squares** is given by: $\boldsymbol{SSR} = \sum_{i=1}^{n}(\hat{y}_i - \overline{y})\text{^}\mathbf{2}$

It gives a sense of how much variation in Y is explained by our model.

A quantitative measure of the total amount of variation in observed Y values is given by the **total sum of squares**: $\boldsymbol{SST} = \sum_{i=1}^{n}(y_i - \overline{y})^2$

data mean

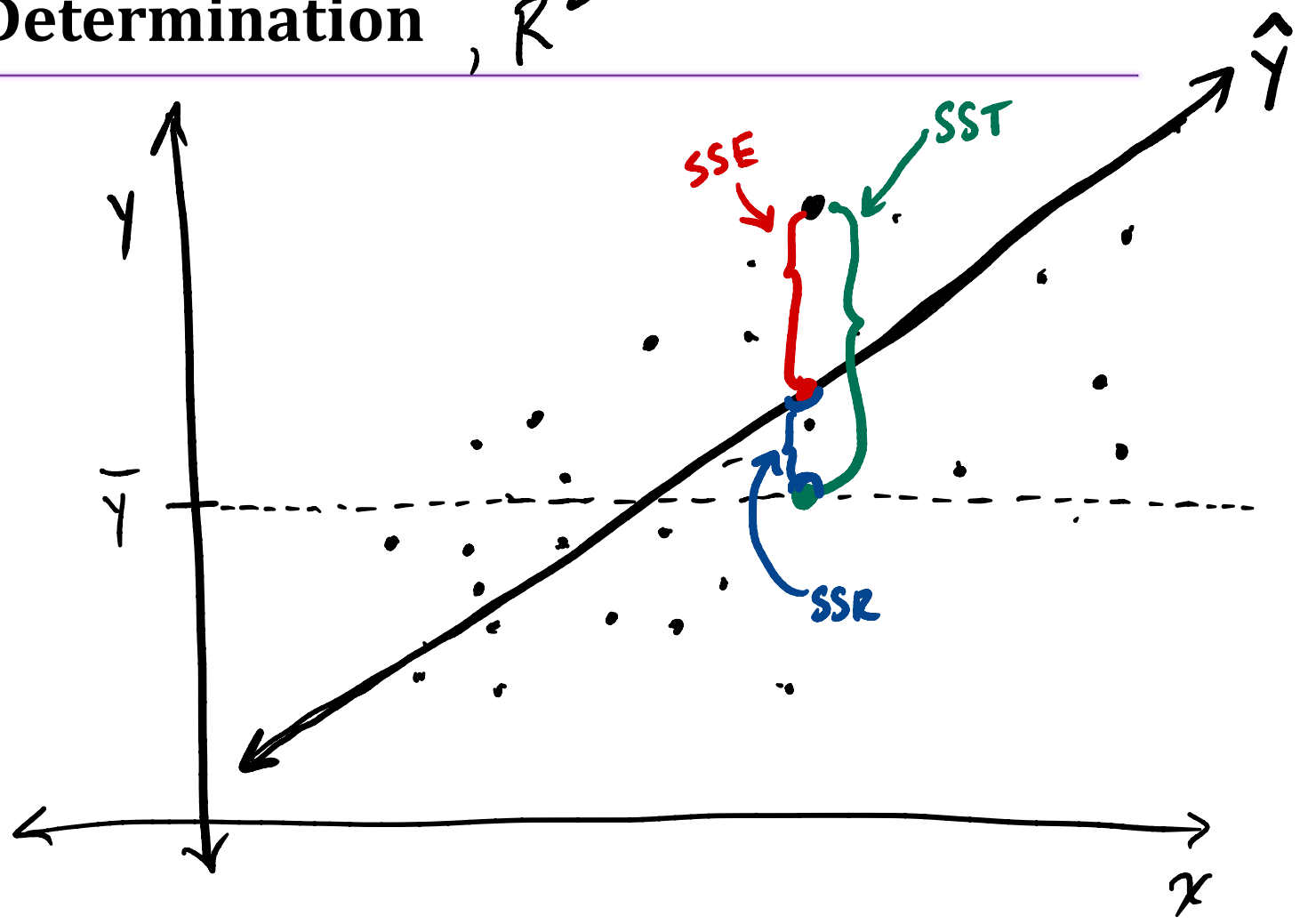
SST is what we would get for SSE if we used the mean of the data as our model.

The Coefficient of Determination, R^2

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

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

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



R^2 is a way to quantify "goodness of fit"

$$R^2 = \frac{\text{Variance explained by model}}{\text{Total Variance}} = \frac{SST - SSE}{SST} = 1 - \frac{SSE}{SST}$$

* In general, the larger R^2 is, the better the regression model fits the data.

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.

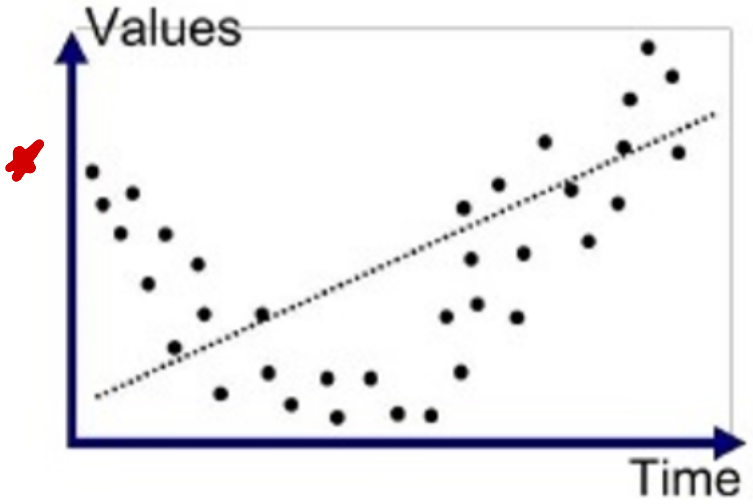
$$SSE \leq SST$$

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.

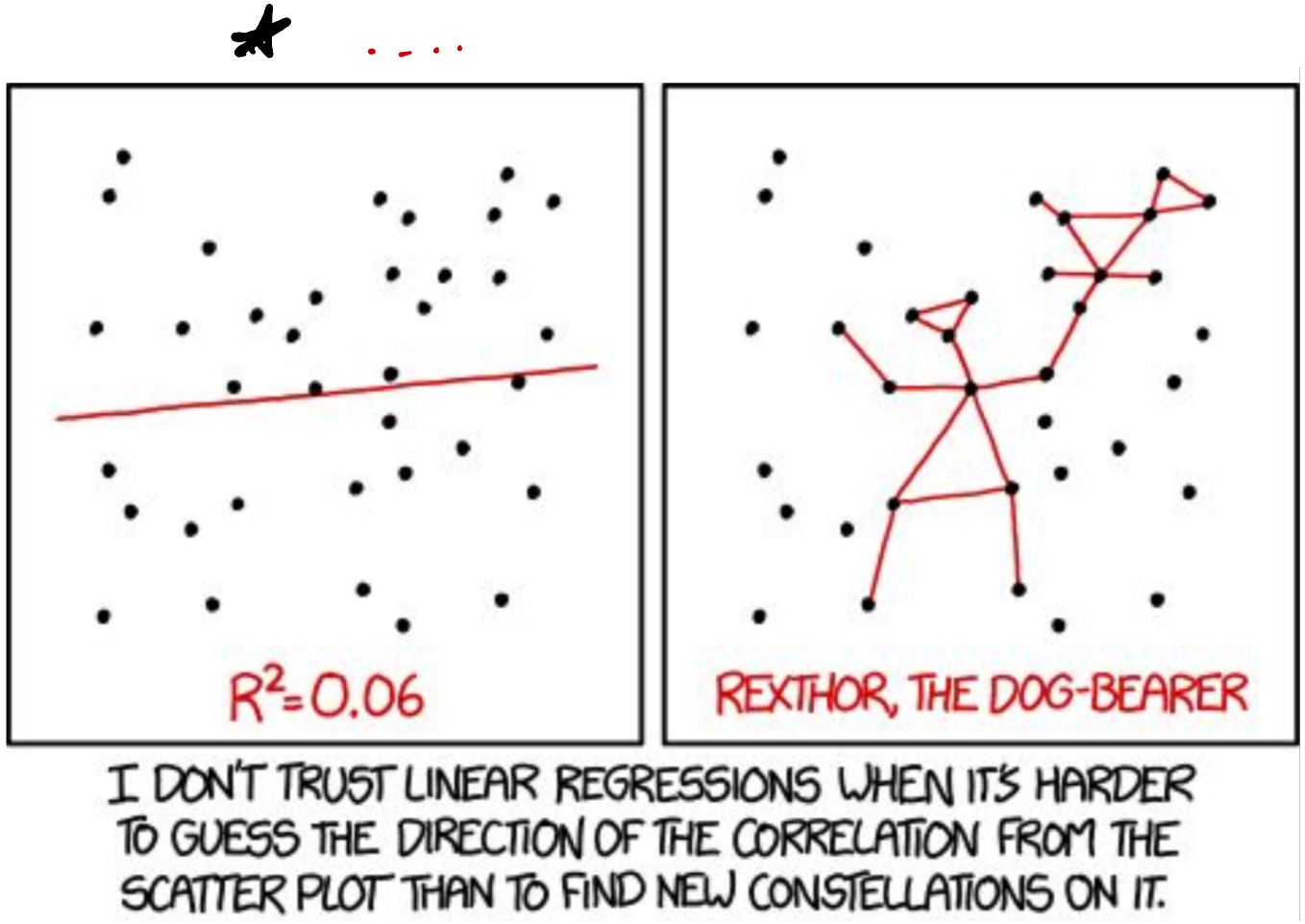
$$\boldsymbol{R^2 = 1 - \frac{SSE}{SST}}$$

The Coefficient of Determination

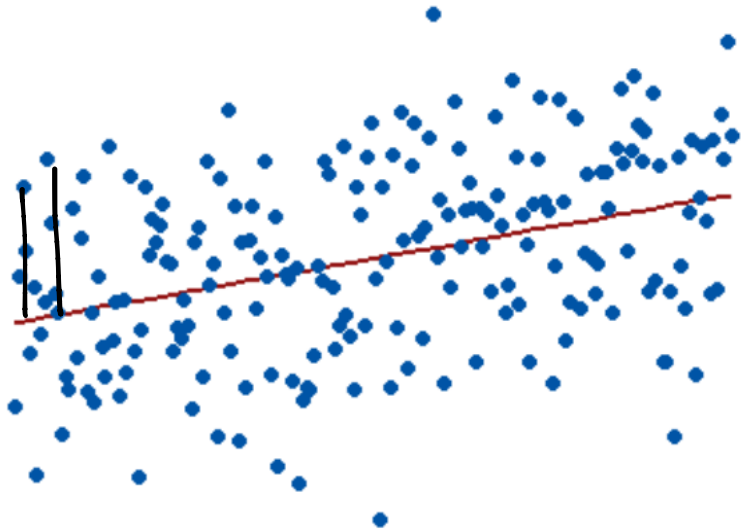
R^2 is the proportion of total variation in the data that is explained by the model. It does not tell you whether you have the correct model.



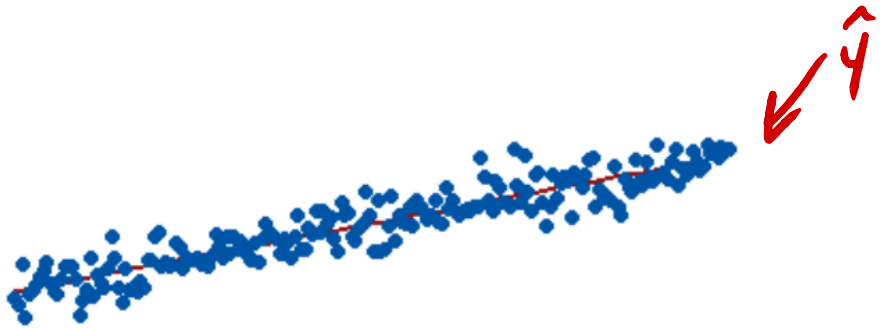
maybe we should have used a quadratic model for this data



The Coefficient of Determination



$R^2 = .15$



$R^2 = .85$

R^2 is a number between 0 and 1

- Another way to examine goodness of fit is to examine a plot of the residuals.

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, etc.

We will focus mainly on the slope parameter β - Is there really a relationship between the feature and the response?

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

$$\hat{\beta} \sim N\left(\beta, \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right) \rightarrow SE(\hat{\beta}) = \frac{\hat{\sigma}}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

Inference about SLR parameters

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

- ❖ The confidence intervals for β are given by:

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

$SE(\hat{\beta})$

t-distribution

- ❖ Hypothesis Testing:

H_0: $\beta = c$
H_1: $\beta \neq c$

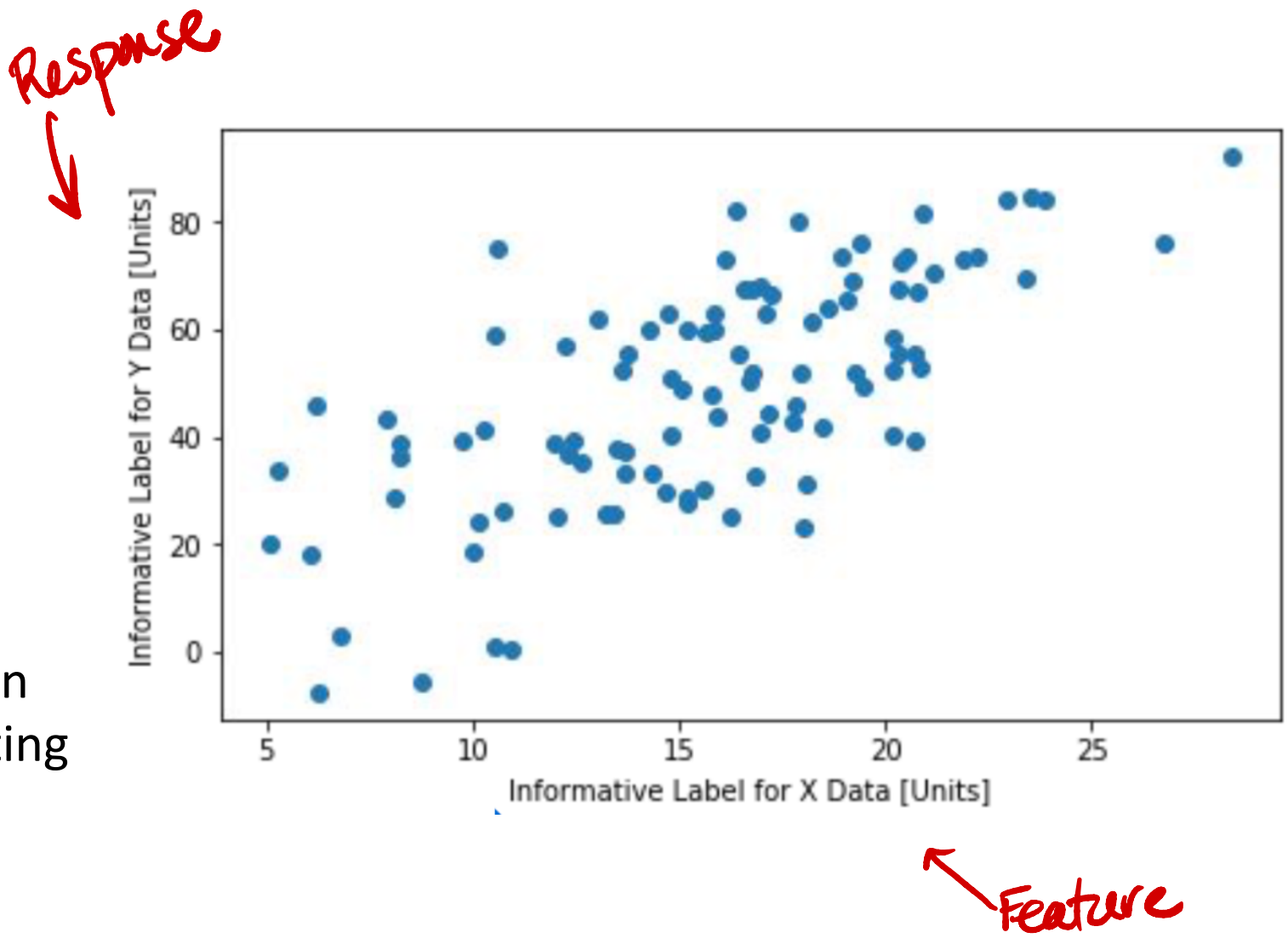
Test statistic: $t = \frac{\hat{\beta} - c}{SE(\hat{\beta})}$ → Compare to $t_{\alpha/2, n-2}$ or compute p-value

See notebook 21 for examples!

Inference about SLR parameters

Given data (x, y)

- Explore/plot/histograms
- Formulate null and alternative hypotheses
- Fit a regression line
- Compute CI or p-value/rejection region and test statistic for testing your hypotheses.
- Make conclusions



Next Time:

- Notebook Day