



# **PARTITIONING THE VARIANCE OF Y**

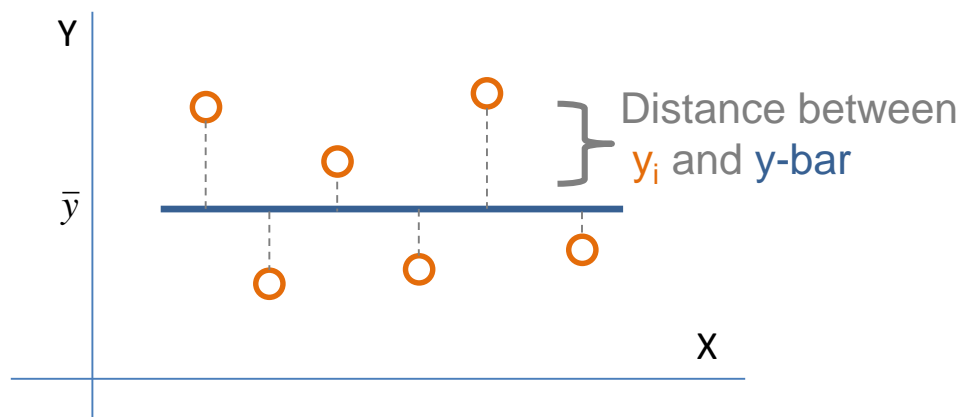
Fundamentals of  
**PROGRAM EVALUATION**

JESSE LECY

# THE VARIANCE CALCULATION

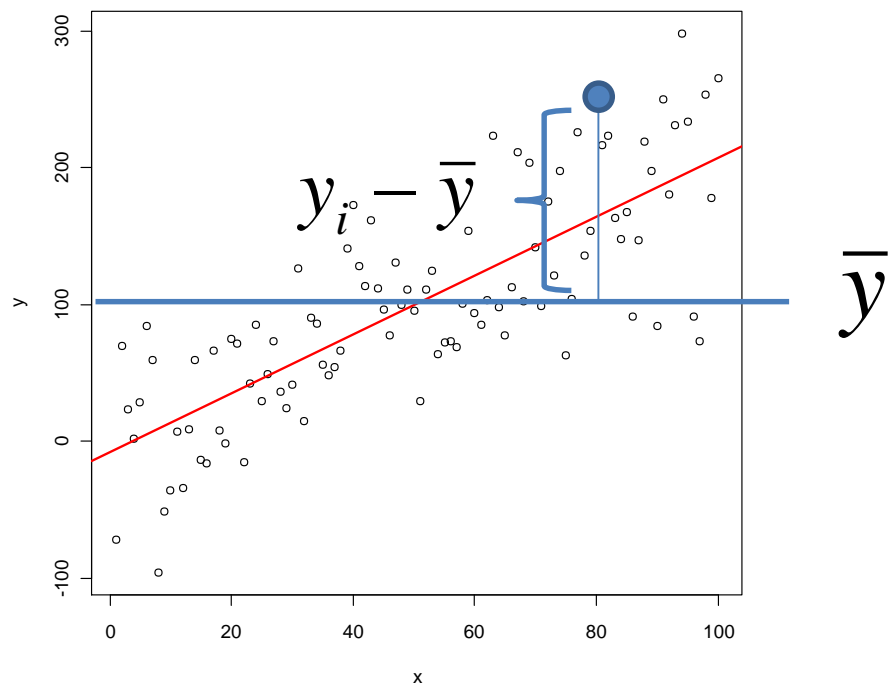
Distance between  
 $y_i$  and  $\bar{y}$

$$\text{var}(y) = \frac{\sum (y_i - \bar{y})^2}{n-1}$$



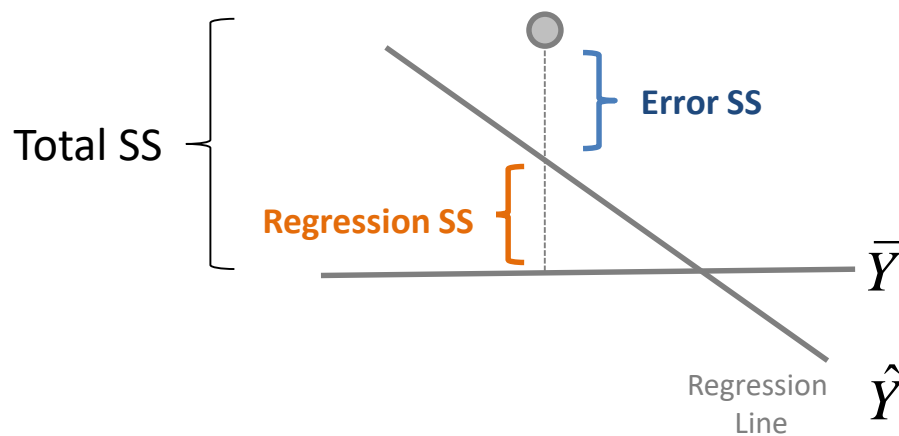
**Variance:** square the distances, add them up, divide by n-1

# PARTITIONING THE VARIANCE OF Y



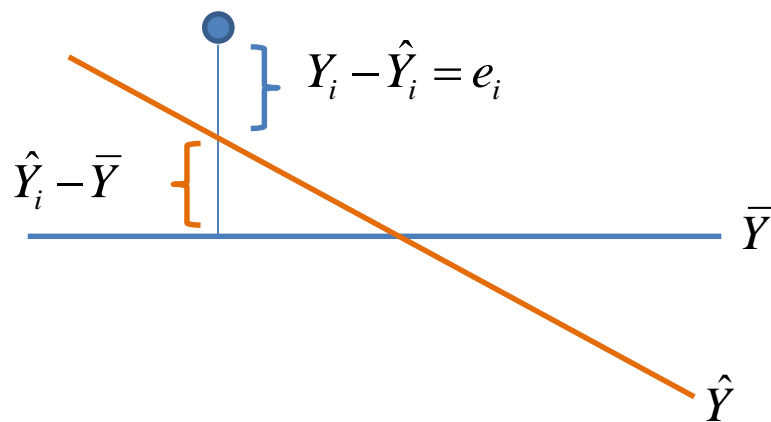
$$\text{var}(y) = \frac{\sum (y_i - \bar{y})^2}{n-1}$$

# PARTITIONING THE VARIANCE OF Y

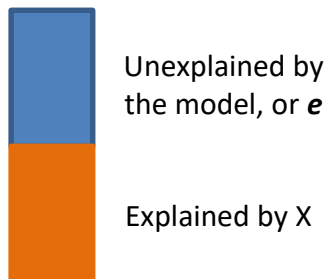


We want to split total variance into explained and unexplained portions.

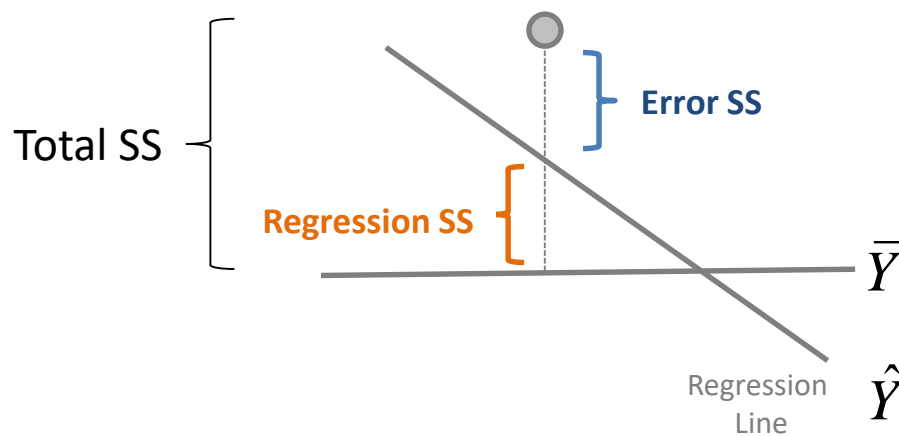
# PARTITIONING THE VARIANCE OF Y



Total Variance Y



# PARTITIONING THE VARIANCE OF Y



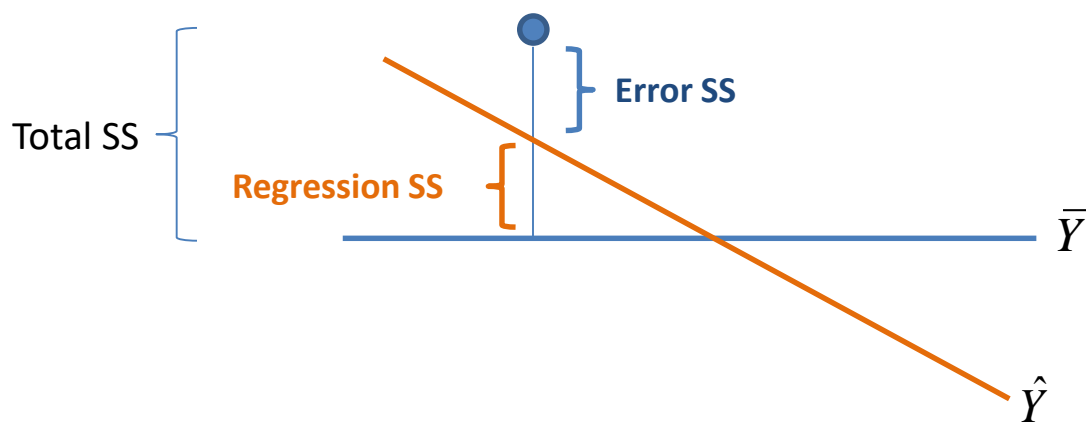
$$TotalSS = \sum (y_i - \bar{y})^2$$

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

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

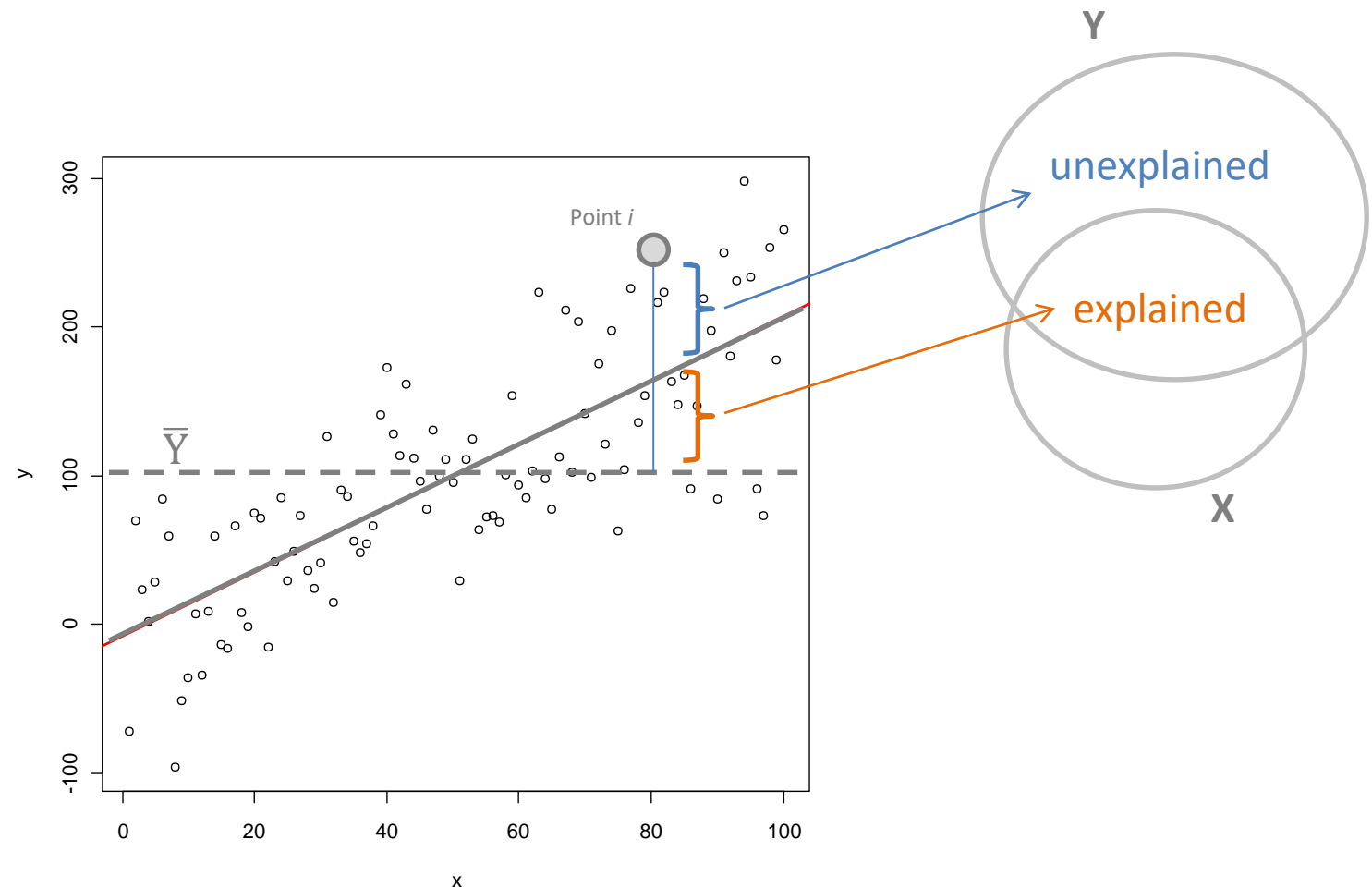
$$TSS = RSS + ESS$$

# PARTITIONING THE VARIANCE OF Y



TSS = Regression/Explained SS + Error/Residual SS

$$R^2 = \frac{\hat{Y}_i - \bar{Y}}{Y_i - \bar{Y}} = \frac{\text{Explained SS}}{\text{Total SS}}$$



## Two parts of the variance of Y

*Unexplained:*  $Y_i - \hat{Y}_i = e_i$

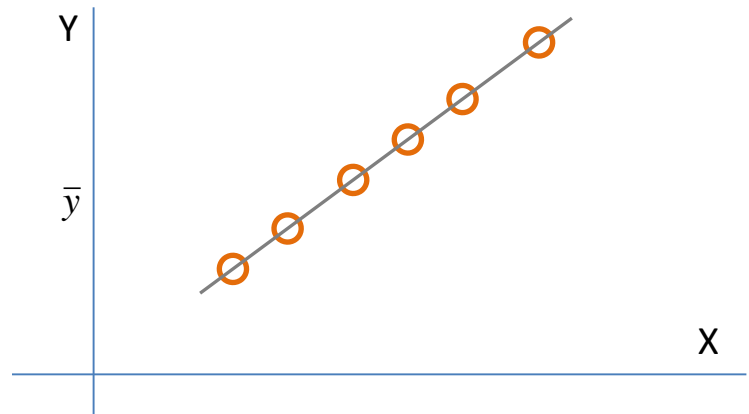
*Explained:*  $\hat{Y}_i - \bar{Y}$

The Venn diagram is a simplified representation of the regression model. In our regression, the explained portion of the variance of outcome will always be the distance from the mean to the predicted value of Y (which always falls on the regression line), and the unexplained portion is the distance between the regression line and the actual data point, also called the residual or the error  $e$ .



# PARTITIONING THE VARIANCE OF Y

Total Variance Y

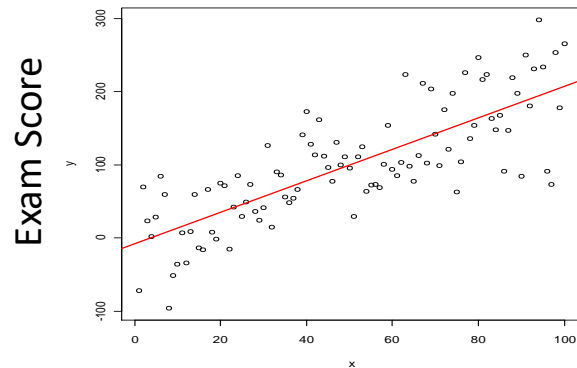


If all points lie on the regression line then we can explain everything about the variance with our model.

# PARTITIONING THE VARIANCE OF Y

Final Exam Scores

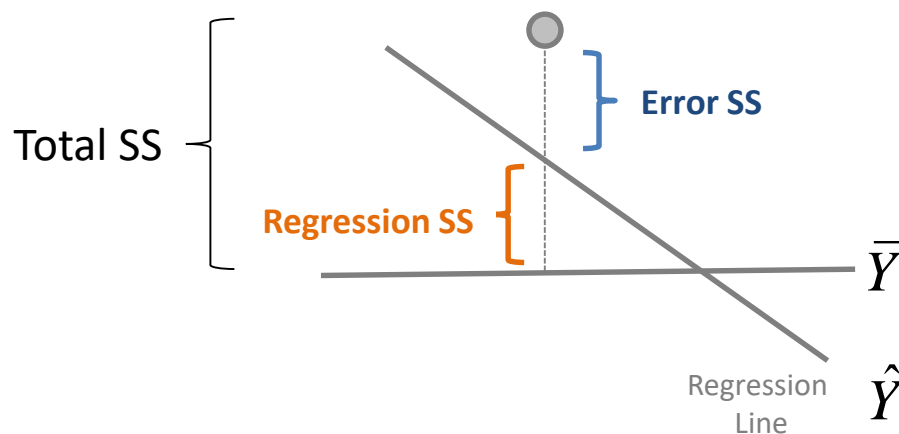
Residual  
(unexplained portion)



Hours of Preparation

The typical case is where the model explains some but not all of the variance.

# PARTITIONING THE VARIANCE OF Y



## Two parts of the variance of Y

Recall that the variance is just a sum of squared deviations from the mean divided by the sample size (minus a couple degrees of freedom). We sometimes just work with the sum of squares directly for the ease of calculation.

We can split the total variance ( $TSS/n-1$ ) into an explained and an error portion. These portions are then manipulated separately, and also used in important calculations like the R-square.

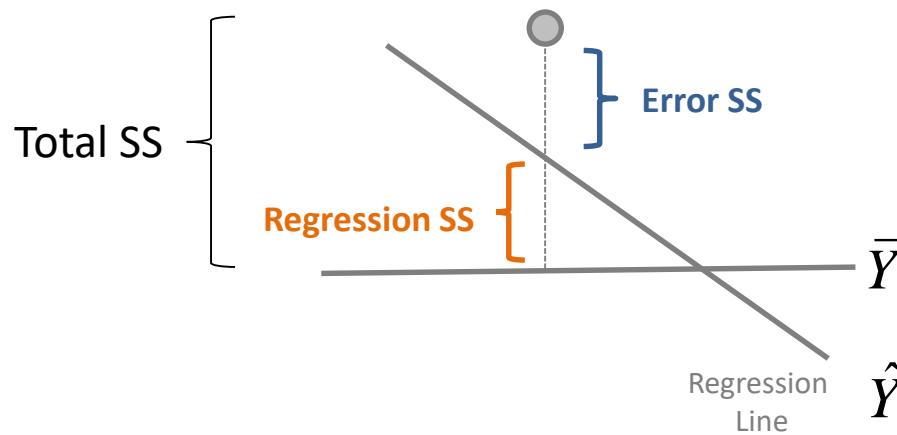
$$TotalSS = \sum (y_i - \bar{y})^2$$

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

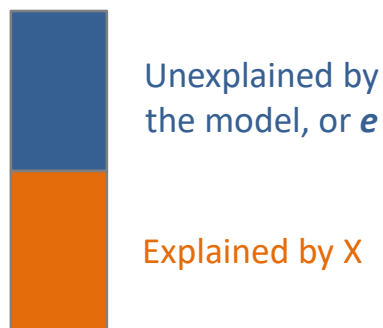
$$ErrorSS = \sum (y_i - \hat{y}_i)^2$$

$$TSS = RSS + ESS$$

# PARTITIONING THE VARIANCE OF Y



Total Variance Y



$$R^2 = \frac{\text{Explained Variance of } Y}{\text{Total Variance of } Y}$$

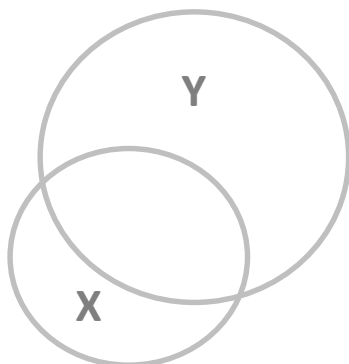
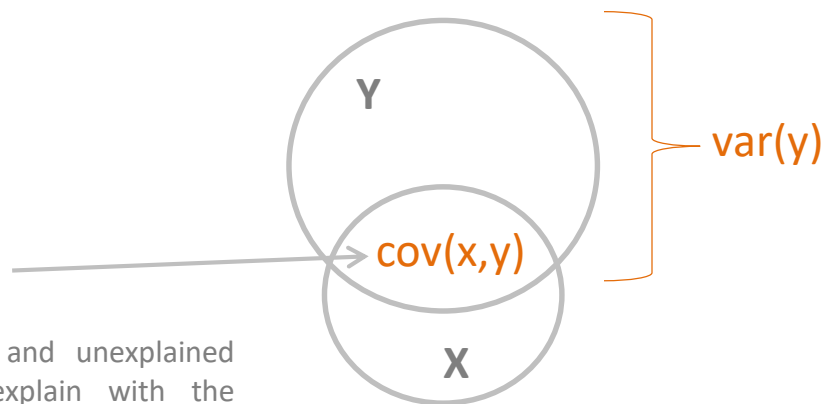
$$R^2 = \frac{RSS/n-1}{TSS/n-1}$$

$$R^2 = \frac{RSS}{TSS} = 1 - \frac{ESS}{TSS}$$

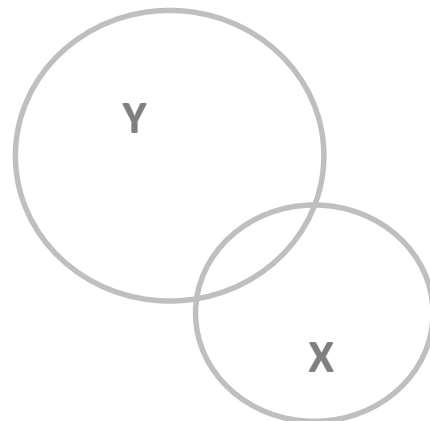
## Venn Diagram Version

### X “explains” this part of Y

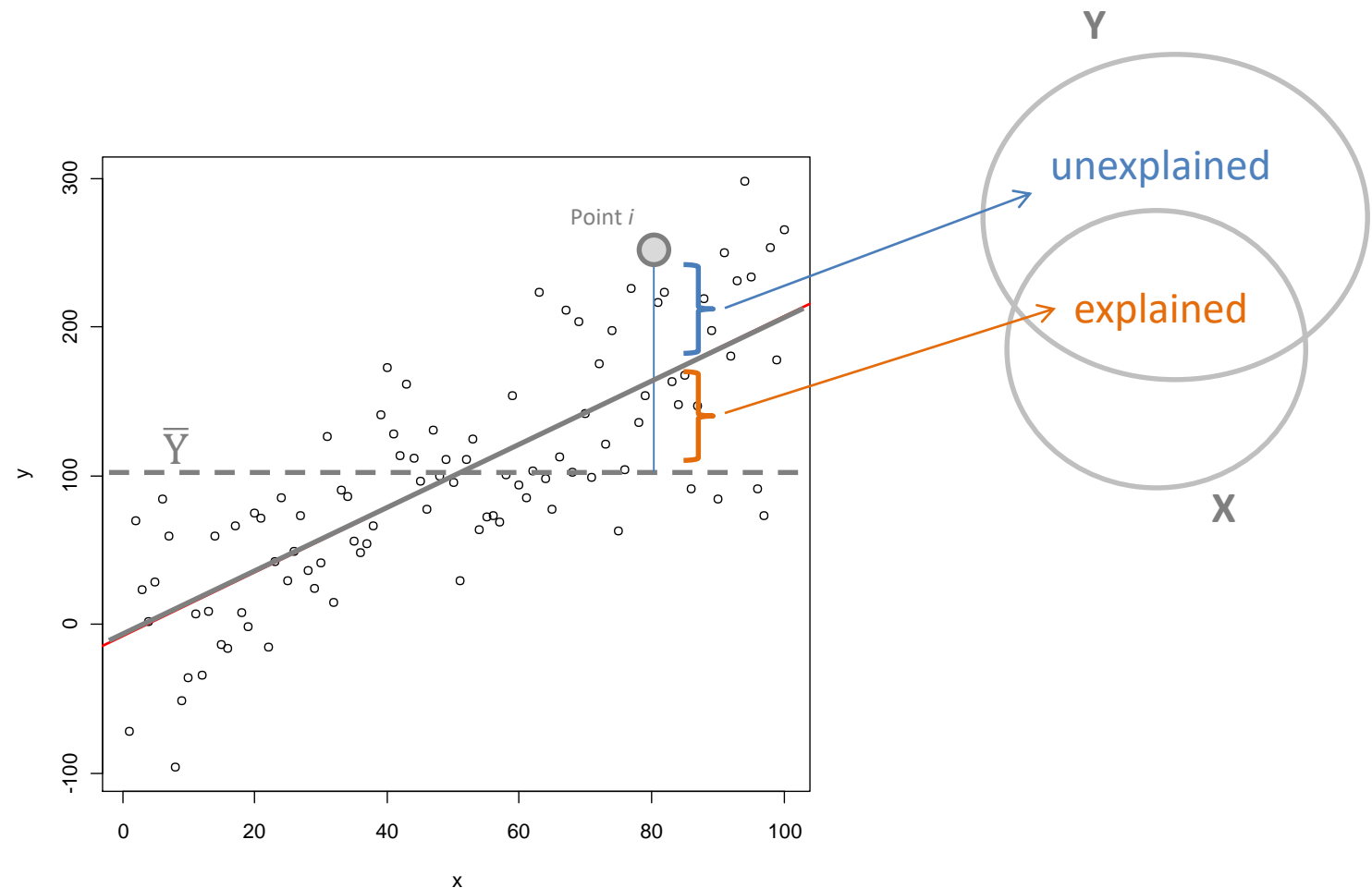
We can partition Y into the explained and unexplained portions. The portions that we can explain with the independent variable X will be the portion of Y that co-varies with X. We can often refer to the overlap region also as the correlation between X and Y. When two variables have more covariance, the correlation is stronger. Less covariance equates to weaker correlation.



more  
correlation



less  
correlation



*Unexplained:*  $Y_i - \hat{Y}_i = e_i$

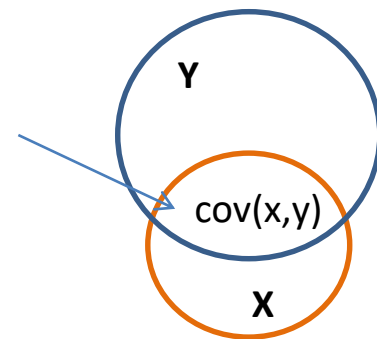
*Explained:*  $\hat{Y}_i - \bar{Y}$

## Two parts of the variance of Y

The Venn diagram is a simplified representation of the regression model. In our regression, the explained portion of the variance of outcome will always be the distance from the mean to the predicted value of Y (which always falls on the regression line), and the unexplained portion is the distance between the regression line and the actual data point, also called the residual or the error  $e$ .

$$R^2 = \frac{RSS}{TSS}$$

x “explains”  
this part of y

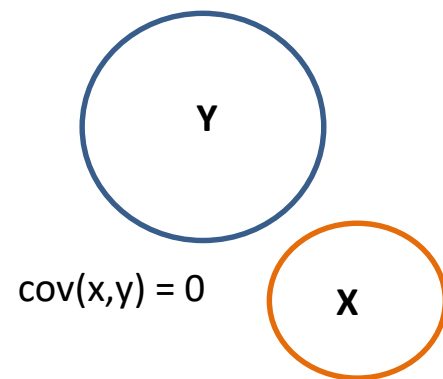
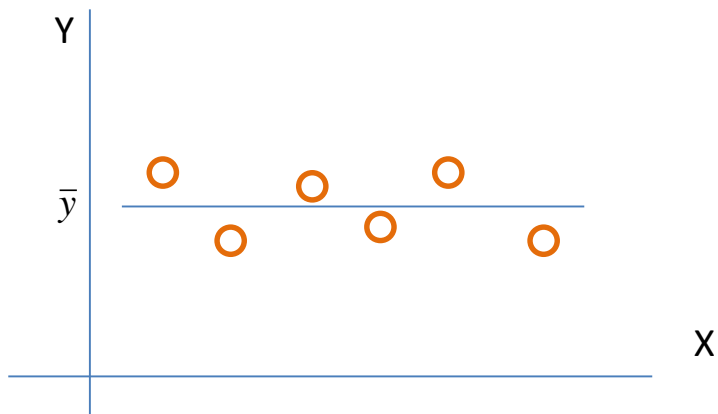
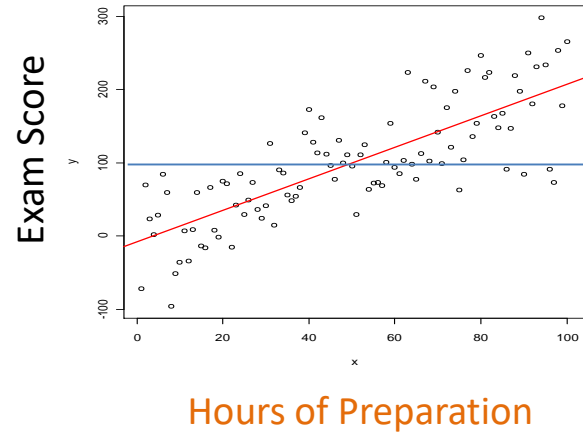
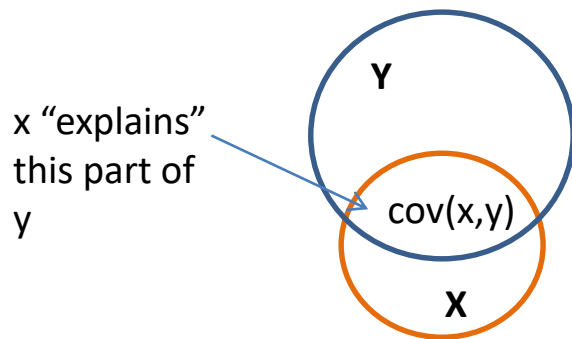


RSS (regression sum of squares)

TSS (total sum of squares)

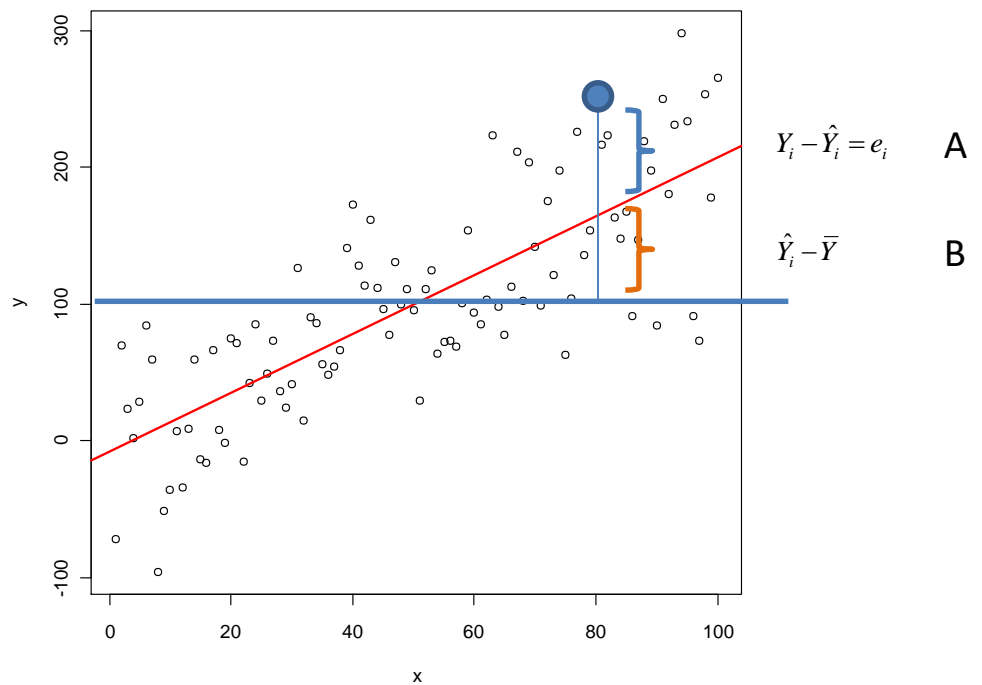
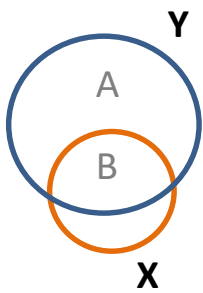
ESS (error sum of squares)

\* Note that sometimes RSS stands for  
“residual” SS and ESS sometimes for  
“explained” SS



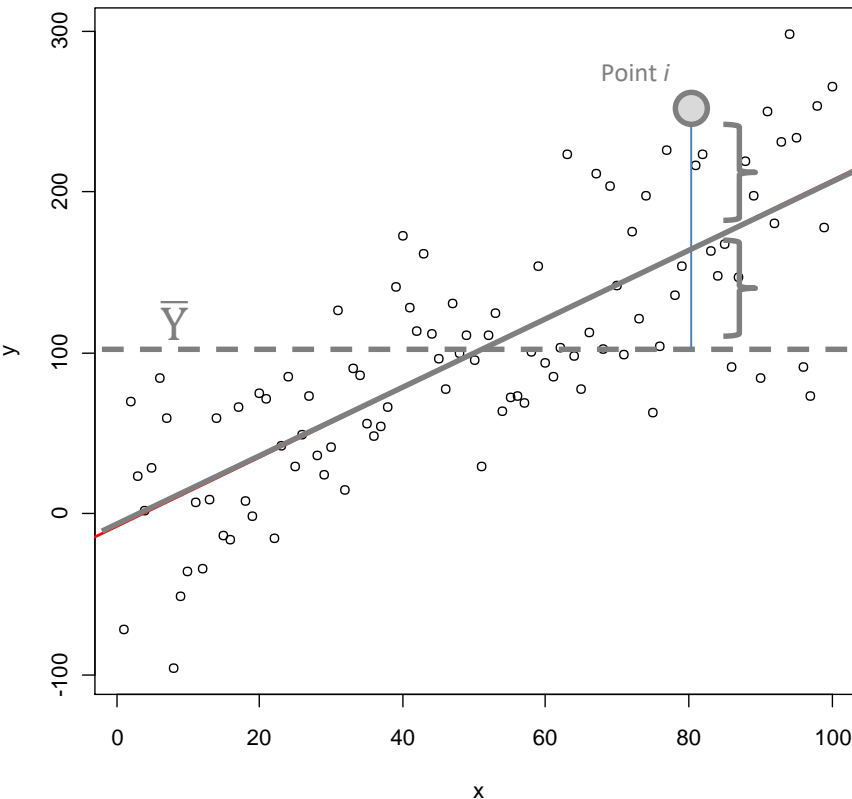


# IN VENN DIAGRAM TERMS



# **STANDARD ERROR OF THE SLOPE**

## Standard Error in Regression



$$Y_i - \hat{Y}_i = e_i$$

$$Y_i - \bar{Y}$$

$$SSE = \sum e_i^2$$

Sum of Squared Error Terms

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

Variance of the residual

The standard error of the slope is one of the most important concepts in regression because it determines the size of the confidence interval and thus the statistical significance of our study. We want standard errors to be as small as possible. We see here that the size of the standard error will be directly related to the amount of unexplained variance we have in our model, the residual  $e$ . The important thing to note is the unexplained portion shows up in the denominator of the standard error. As a result, the size of the standard error will be proportional to the amount of unexplained variance (plus a couple of other considerations to be covered later).

$$SE_{b_1} = \sqrt{\frac{\hat{\sigma}_\varepsilon^2}{\sum (x_i - \bar{x})^2}}$$

Standard error of the slope

# STANDARD ERROR IN REGRESSION

$$SE_{\bar{x}} = \frac{s_x}{\sqrt{n}}$$

The size of the standard error of the mean is driven by the variance of the variable, and the sample size.

$$\text{var}(x) = \frac{\sum (x_i - \bar{x})^2}{n-1} \Rightarrow$$

$$(n-1) \cdot \text{var}(x) = \sum (x_i - \bar{x})^2$$

We can write the formula for the standard error of the slope in a couple of ways. I prefer the top because it is explicit about sample size and  $\text{var}(x)$ .

$$SE_{b_1} = \frac{s_\varepsilon}{\sqrt{(n-1) \cdot \text{var}(x)}}$$



$$SE_{b_1} = \frac{s_\varepsilon}{\sqrt{\sum (x_i - \bar{x})^2}}$$

Similarly, the standard error of the slope is a function of the variance of the residual (the amount of unexplained variance in the outcome), the sample size (n-1), AND the variance of the explanatory variable.

$$SE_{b_1} = \sqrt{\frac{\hat{\sigma}_\varepsilon^2}{\sum (x_i - \bar{x})^2}}$$

### Standard error of the Slope

Don't get too caught up with the math. The formula for the standard error of a regression coefficient is actually quite simple when you break it down. There are three moving parts – three things that can affect the size of the standard error. The portion of unexplained variance of the dependent variable (the residual), the sample size of the regression, and the amount of variance in the variable X associated with the regression slope.

$$\text{Standard Error of the Slope} \approx \frac{\text{residual}}{\text{sample size} \cdot \text{variance X}}$$

## THE ROAD MAP

### Of the Mean:

### Of the Slope:

Sampling  
Variance:

$$\sigma_x^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}$$

(for x)

$$\sigma_\varepsilon^2 = \frac{SSE}{n-2} = \frac{\sum e_i^2}{n-2}$$

(using the residual)



Standard  
Deviation:

$$\sigma_x = \sqrt{\sigma_x^2}$$

$$\sigma_\varepsilon = \sqrt{\sigma_\varepsilon^2}$$



Standard  
Error:

$$SE_{\bar{x}} = \frac{\sigma_x}{\sqrt{n}}$$

$$SE_{b_1} = \sqrt{\frac{\sigma_\varepsilon^2}{\sum (x_i - \bar{x})^2}}$$



Confidence  
Interval

$$\mu = \bar{x} \pm t \cdot SE_{\bar{x}}$$

(of the mean)

$$\beta_1 = b_1 \pm t \cdot SE_{b_1}$$

(of the slope)

What should be clear in my mind?

1. We split the variance of Y into **explained** and **unexplained** portions with a trick, inserting the regression line  $\hat{y}$ .
2. The **standard error of the slope** is derived from the unexplained portion of Y, the **residual**.