# Statistics One

## Lecture 7

## Introduction to Regression

# Two segments

- Intro to regression
- NHST: A closer look

# Lecture 7
# Segment 1

## Intro to Regression

# Regression

- Important concepts/topics
  - Regression equation and “model”
  - Ordinary least squares estimation
  - Unstandardized regression coefficients
  - Standardized regression coefficients

# Regression

- A statistical analysis used to predict scores on an outcome variable, based on scores on one or more predictor variables
  - For example, we can predict how many runs a baseball player will score (Y) if we know the player's batting average (X)

# Regression equation

- $Y = m + bX + e$ # Y is a linear function of X, b = slope

# Regression equation

- $Y = m + bX + e$ # Y is a linear function of X, $b$ = slope
- $Y = a + BX + e$ # same as above, notation changed

# Regression equation

- $Y = m + bX + e$ # Y is a linear function of X, b = slope
- $Y = a + BX + e$ # same as above, notation changed
- $Y = B_0 + B_1X_1 + e$ # same, another notation change

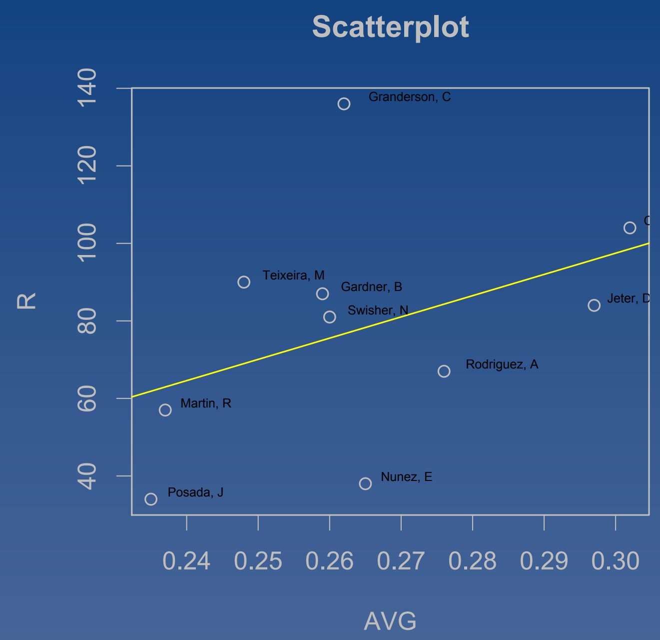
# Regression equation

- $Y = m + bX + e$ # Y is a linear function of X, b = slope
- $Y = a + BX + e$ # same as above, notation changed
- $Y = B_0 + B_1X_1 + e$ # same, another notation change
- $\hat{Y} = B_0 + B_1X_1$ # $\hat{Y}$ is the predicted score on Y

# Regression equation

- $Y = m + bX + e$ # Y is a linear function of X, b = slope
- $Y = a + BX + e$ # same as above, notation changed
- $Y = B_0 + B_1X_1 + e$ # same, another notation change
- $\hat{Y} = B_0 + B_1X_1$ # $\hat{Y}$ is the predicted score on Y
- $Y - \hat{Y} = e$ # e is the prediction error (residual)

# Scatterplot: plot(R~AVG)




$r = +.40$

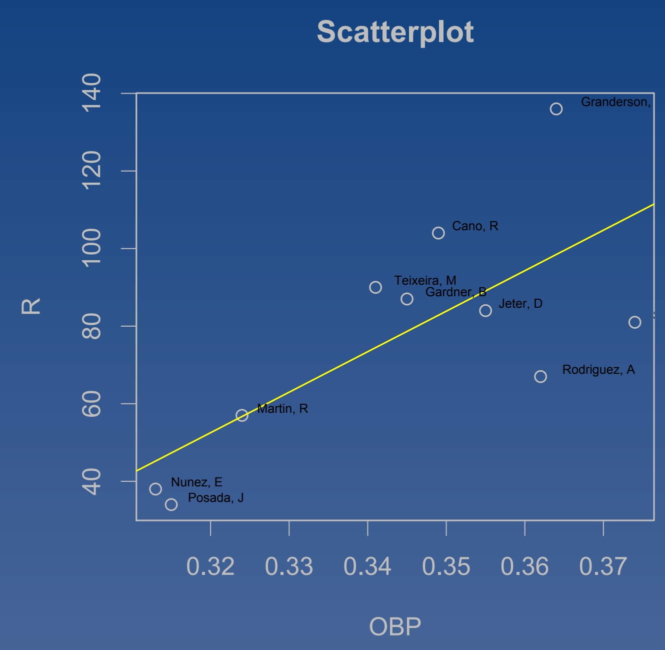
# Regression “model”

- The regression model is used to “model” or predict future behavior
  - The “model” is just the equation

# Regression: It gets better

- The goal is to produce better models so we can generate more accurate predictions
  - Add more predictor variables, and/or…
  - Develop better predictor variables
    - For example, we can better predict how many runs a baseball player will score (Y) if we know the player’s on-base-percentage (X)

# Scatterplot: plot(R~OBP)




$$r = +.72$$

# Why did it get better?

- OBP is a “model” that takes into account walks, aka “base on balls” (BB)
- The predictions improved, particularly for Granderson and Nunez, because:
    - For Granderson, BB = 85
    - For Nunez, BB = 22

# Why did it get better?

- Lesson: Examine residuals!
- Plot in a histogram
- Scatterplot residuals with X
  - Good way to test assumptions
    - Linear relationship between X and Y
    - Homoscedasticity
      - More on this next lecture

# Estimation of coefficients

- Regression equation:
  - $\hat{Y} = B_0 + B_1X_1$ # $\hat{Y}$ is the predicted score on Y
  - $Y - \hat{Y} = e$ # e is the prediction error (residual)

# Estimation of coefficients

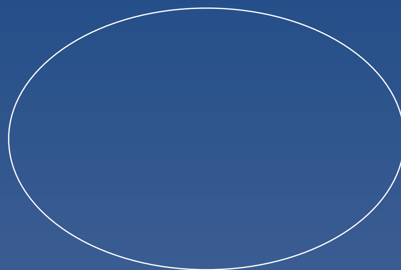
- The values of the coefficients (B) are estimated such that the model yields optimal predictions.
  - Minimize the residuals!
  - The sum of the squared (SS) residuals is minimized
  - SS.RESIDUAL = $\Sigma(\acute{Y} - Y)^2$

  - ORDINARY LEAST SQUARES estimation

# Estimation of coefficients

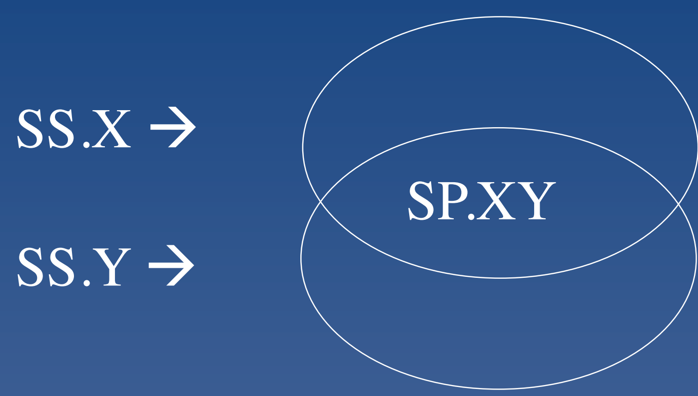
- Sum of Squared deviation scores (SS) in variable Y
  - SS.Y

SS.Y →



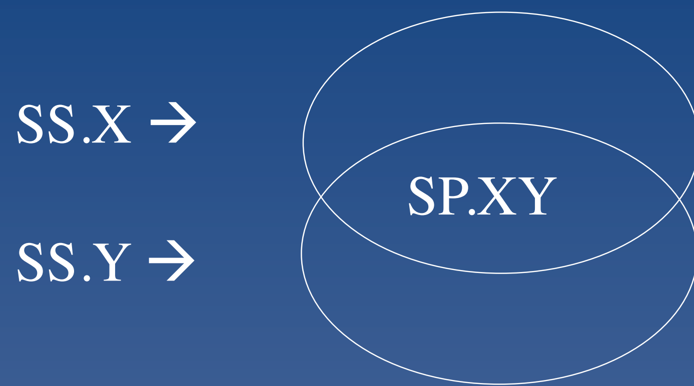
# Estimation of coefficients

- Sum of Cross Products (SP.XY)



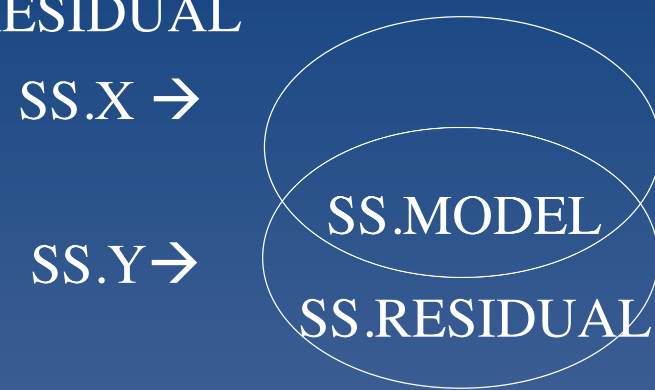
# Estimation of coefficients

- Sum of Cross Products (SP.XY)
  - Also called SS.MODEL



# Estimation of coefficients

- Sum of Squared deviation scores (SS) in variable Y
  - SS.Y = SS.MODEL + SS.RESIDUAL



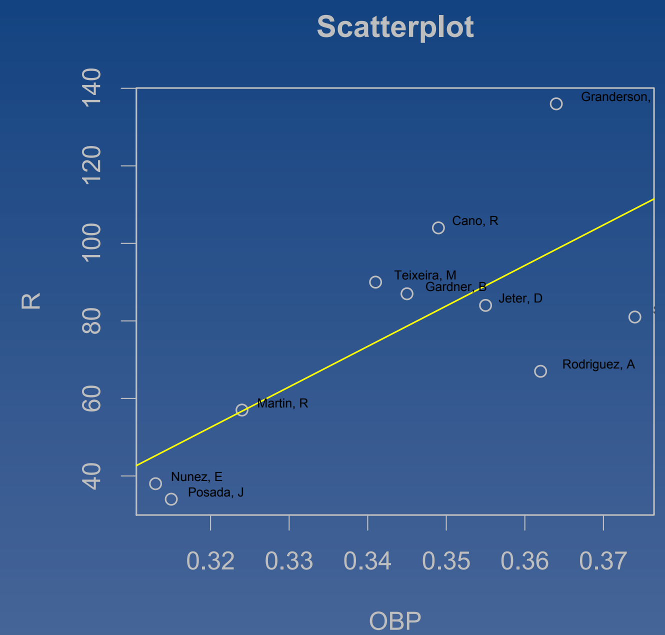
# Estimation of coefficients

- How to calculate B (unstandardized)
  - $B = r \times (SD_y / SD_x)$

# Estimation of coefficients

- Standardized regression coefficient = $\beta$ = $r$
  - If X and Y are standardized then:
    - $SD_y = SD_x = 1$
    - $B = r \times (SD_y / SD_x)$
    - $\beta = r$

# Scatterplot: plot(R~OBP)




$$r = +.72$$

# Estimation of coefficients

- In R: lm(R~OBP)
  - $\hat{Y} = -282 + (1044)X$
  - Let $X = .35$
  - $\hat{Y} = 83$

# Estimation of coefficients

- Why is B so large? (B = 1044)
- Because $SD_y$ is so much greater than $SD_x$
- $SD_y = 31$
- $SD_x = .02$
  - $B = r \times (SD_y / SD_x)$

# Regression

- Important concepts
  - Regression equation and “model”
  - Ordinary least squares estimation
  - Unstandardized regression coefficients
  - Standardized regression coefficients