EPSY 5261 : Introductory Statistical Methods

Day 26

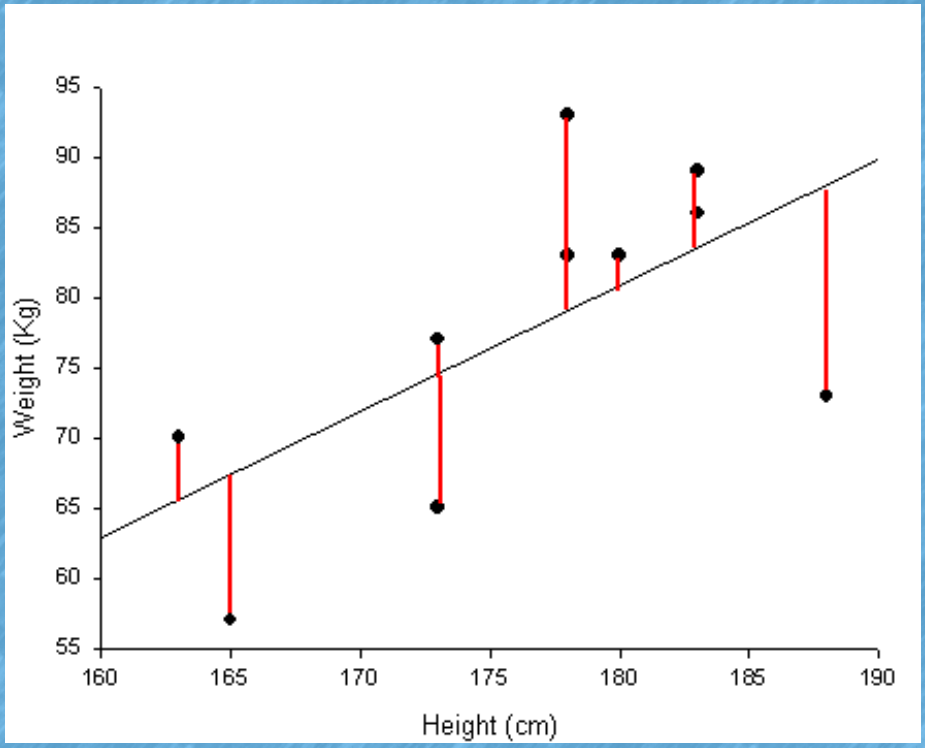
Introduction to Simple Linear Regression

Learning Goals

- At the end of this lesson, you should be able to…
  - Explain when to use linear regression
  - Write a linear regression equation
  - Interpret the coefficients in a linear regression equation
  - Calculate and interpret a residual
  - Interpret an R^2 value

If we have a linear relationship we might consider doing a linear regression

Linear Regression



Line found by minimizing the vertical distances between the points and the line (least squares)

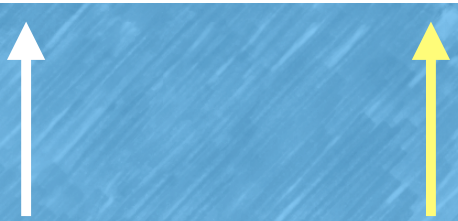
Linear Regression Equation

From high school math class:
y = mx + b

$$\hat{y} = \beta_0 + \beta_1 X$$

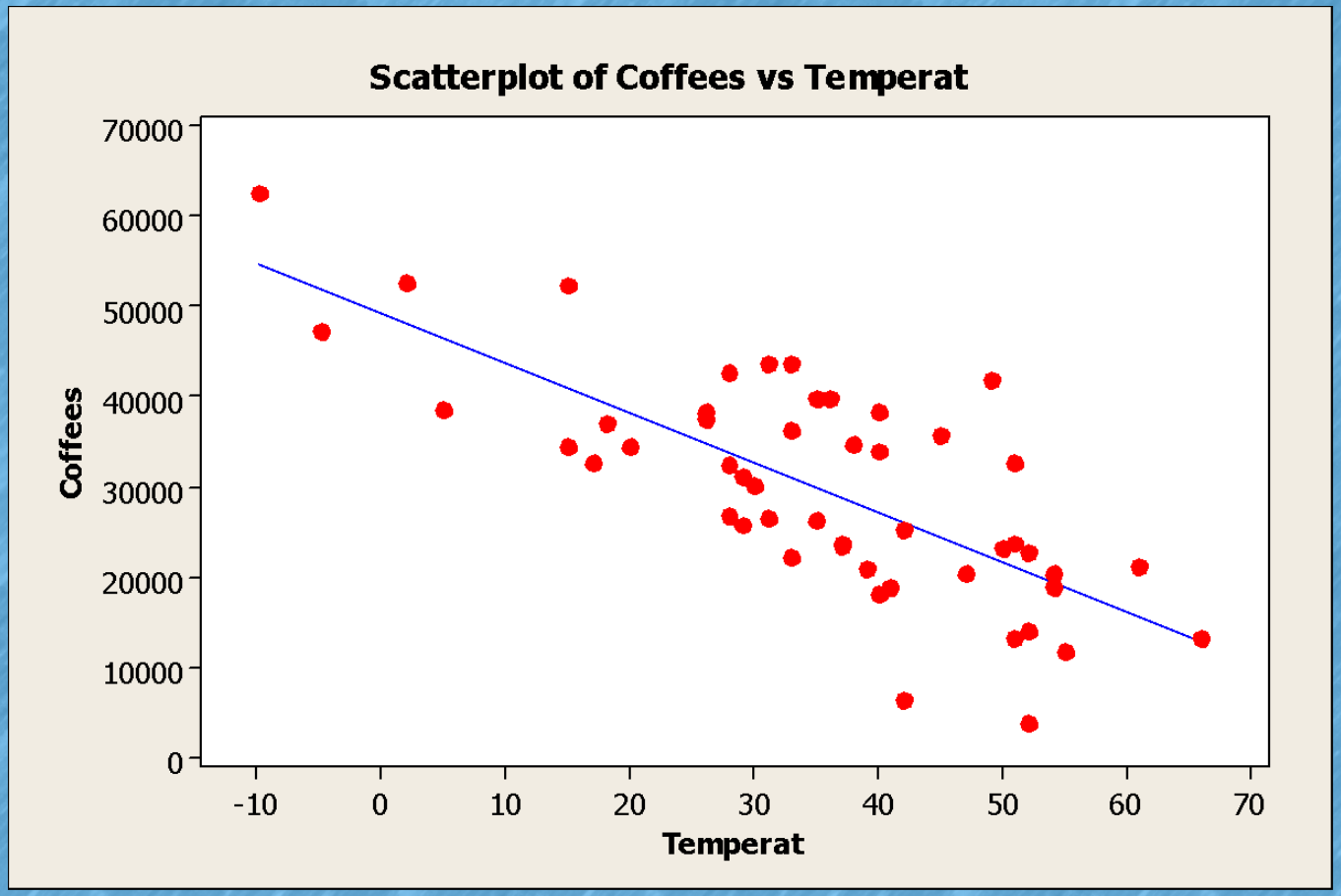
Linear Regression Equation

$$\hat{y} = \beta_0 + \beta_1 X$$



Intercept Slope

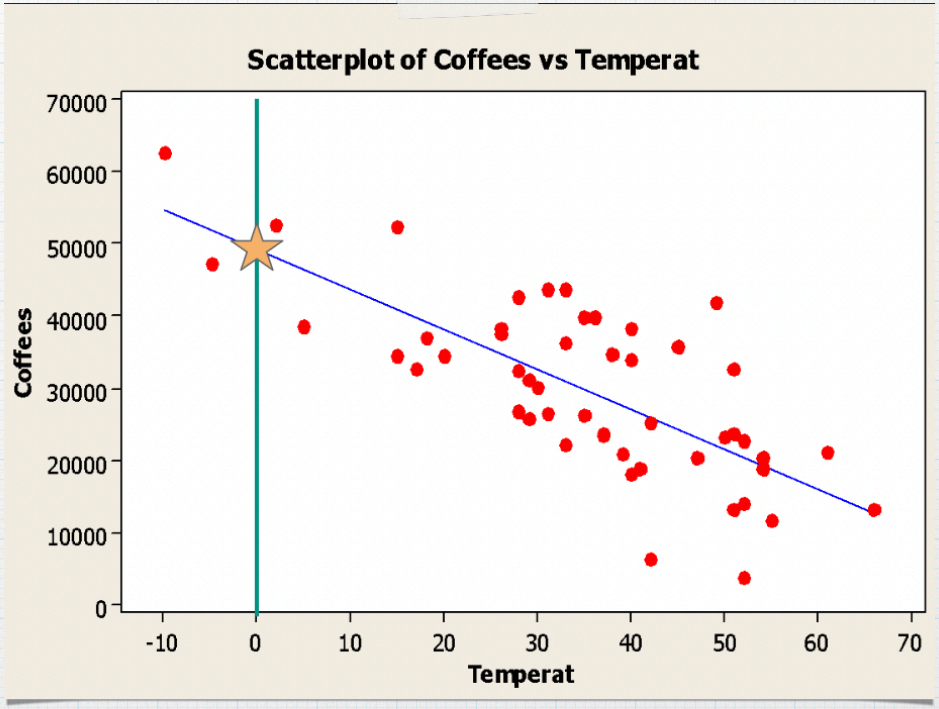
$$\hat{Coffees} = 49{,}337 - 553.7(Temperature)$$



Insert R OUTPUT SLIDE

Interpreting the Intercept

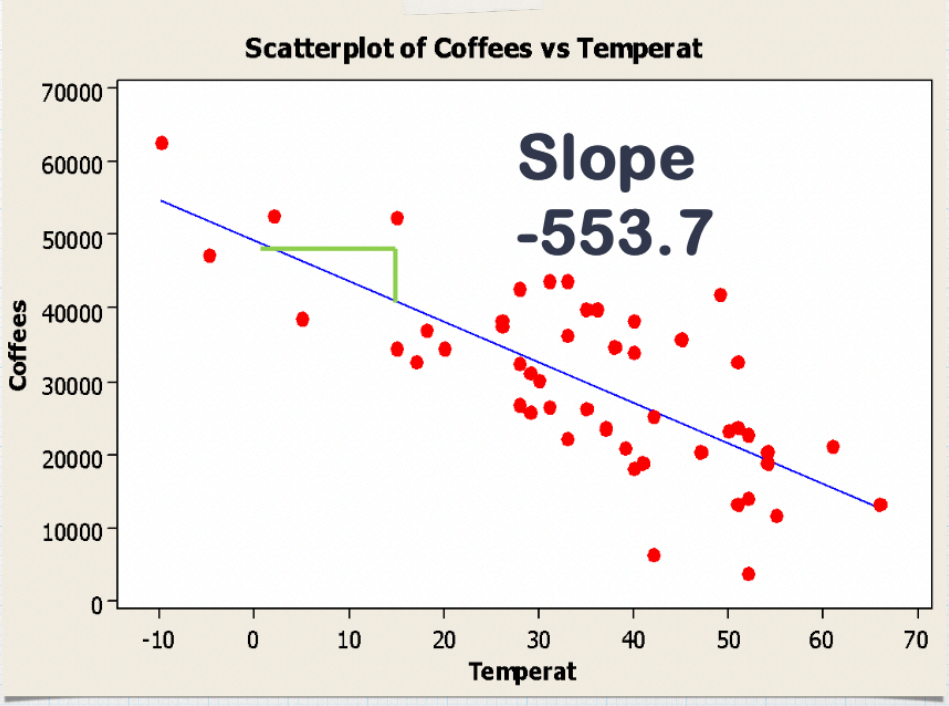
- Intercept = 49,337
- Where line crosses Y axis at (X = 0)
- For a 0°F day, we predict an average of 49,337 coffees sold



Caution: Can't interpret intercept if 0 is not within range of data

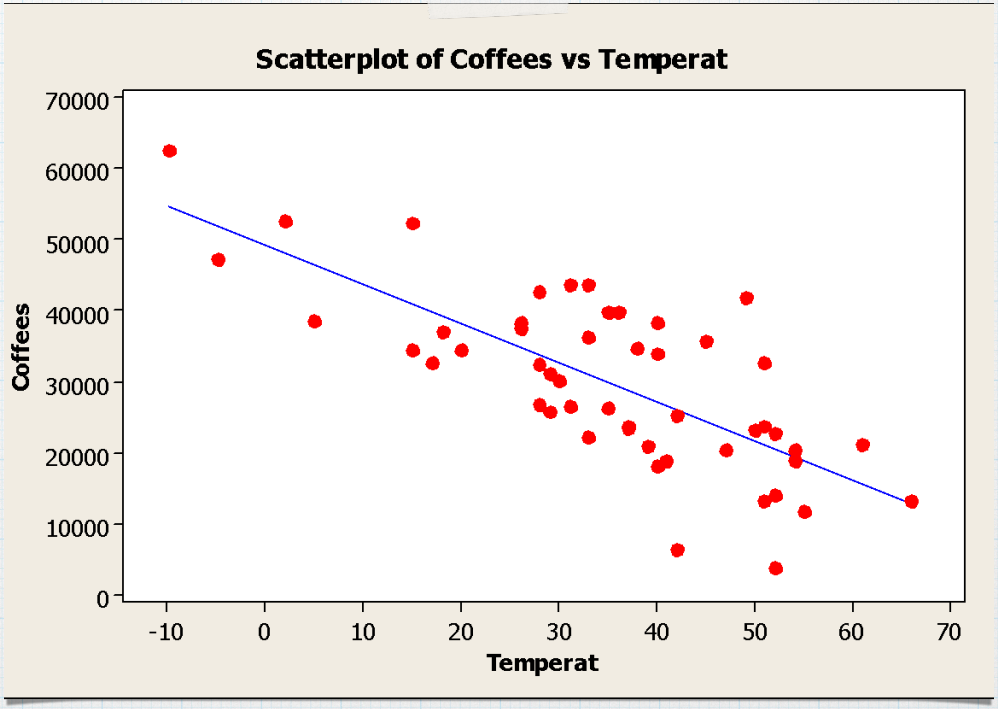
Interpreting the Slope

- "Rise over run": Predicted change in Y as X increases by 1 unit.
- For each additional degree warmer, we expect about 554 fewer coffees to be sold, on average.



Making Predictions Using Our Line

- First: Beware of extrapolation!
- Extrapolation is predicting y values for x values that are *outside* the range of the data. (NOT OK)
- Can we predict coffee sales for:
  - 40 degree days?
  - 80 degree days?



Making Predictions

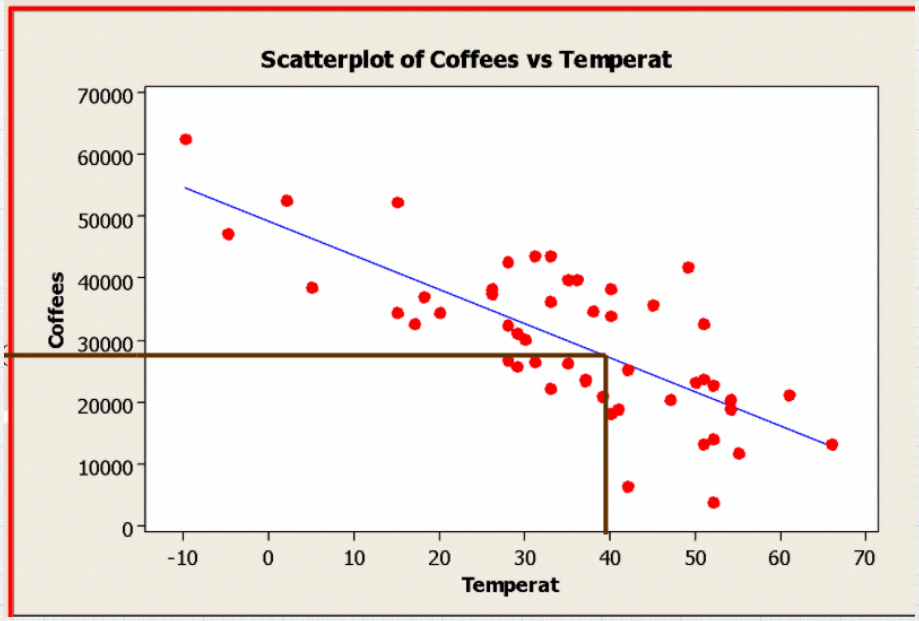
Predict about how many coffee sales to expect on a 40 degree day:

$$\hat{y} = 49337 - 553.70 * (40)$$

=

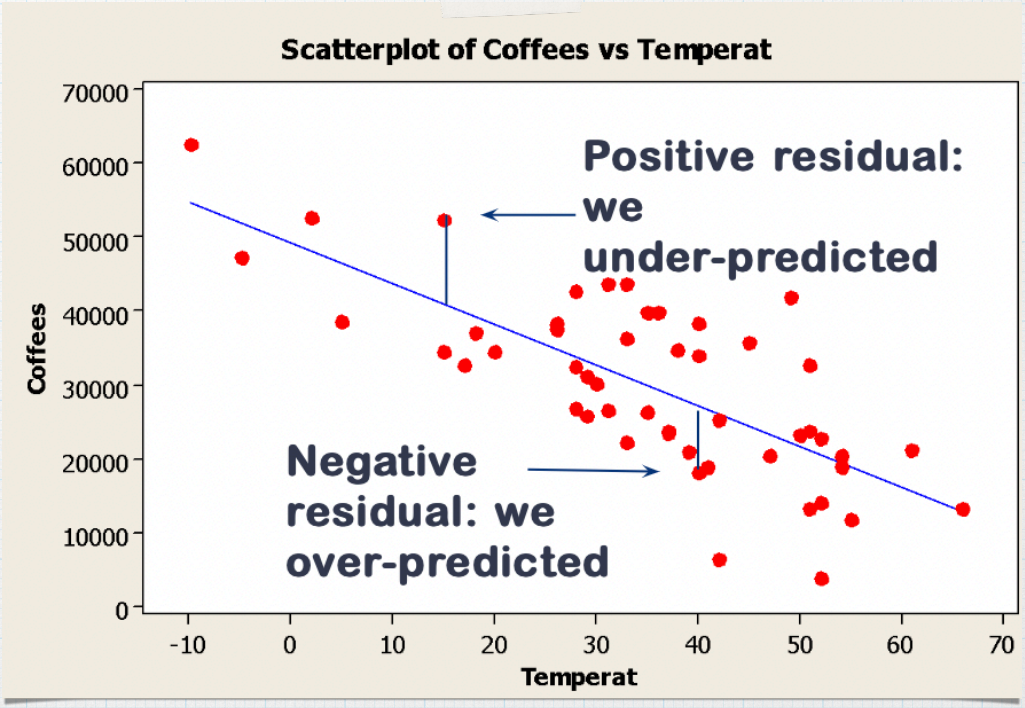
27,189

coffees expected to be sold on 40°day.



Residuals

- $Residual = y - \hat{y}$ (in that order)
- Difference between the observed and predicted values



Residual

Example:

- On a 40 degree day we predicted 27,189 ($\hat{y}$) coffees to be sold. Suppose we actually observed 27,100 (y) coffees were sold.
- What is the residual?
- Residual = 27,100 - 27,189 = -89
- The linear model over-predicted the number of coffees sold by 89.

Coefficient of Determination (R^2)

- R^2= correlation squared
- R^2= 0.549 for coffee data
- Idea: coffee sales vary

How much of this variation can be explained by the linear relationship with temperature?

R^2= 54.9%

- R^2 indicates how much of the total variability in our response variable (Y) can be explained by this linear regression on X).

Summary

- We can create a least squares regression line to learn more about the relationship between two quantitative variables
- We can interpret the slope and intercept of that line to describe that relationship
- We can use the R^2 value to tell us how much variability we are explaining in the Y variable be creating a linear regression with X