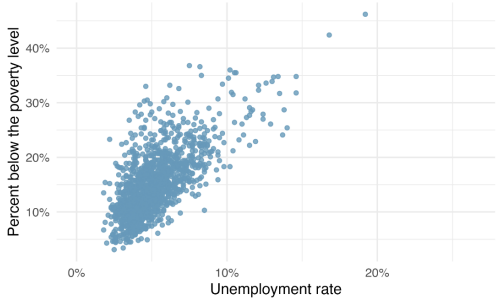
# SDS 220 - Lecture 8 Handout

IMS Chapter 7 (Part 2)

1. **[IMS 7.23] Poverty and unemployment.** The following scatterplot shows the relationship between percent of population below the poverty level (poverty) from unemployment rate among those ages 20-64 (`unemployment_rate`) in counties in the US, as provided by data from the 2019 American Community Survey. The regression output for the model for predicting poverty from `unemployment_rate` is also provided.

| term | estimate | std.error | statistic | p.value |
|---|---|---|---|---|
| (Intercept) | 4.60 | 0.349 | 13.2 | <0.0001 |
| unemployment_rate | 2.05 | 0.062 | 33.1 | <0.0001 |




   (a) Write out the linear model.

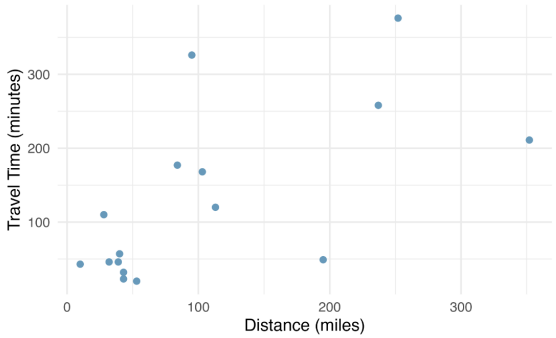
   (b) Interpret the intercept.

   (c) Interpret the slope.

   (d) For this model $R^2 = 46\%$. Interpret this value.

   (e) Calculate the correlation coefficient.

2. **[Adapted from IMS 7.21] The Coast Starlight, regression.** The Coast Starlight Amtrak train runs from Seattle to Los Angeles. The scatterplot below displays the distance between each stop (in miles) and the amount of time it takes to travel from one stop to another (in minutes). The mean travel time from one stop to the next on the Coast Starlight is 129 mins, with a standard deviation of 113 minutes. The mean distance traveled from one stop to the next is 108 miles with a standard deviation of 99 miles. The correlation between travel time and distance is 0.636.



(a) Write the equation of the regression line for predicting travel time.

(b) Interpret the slope and the intercept in this context.

(c) Calculate $R^2$ of the regression line for predicting travel time from distance traveled for the Coast Starlight, and interpret it in the context of the application.

(d) The distance between Santa Barbara and Los Angeles is 103 miles. Use the model to estimate the time it takes for the Starlight to travel between these two cities.

(e) It actually takes the Coast Starlight about 168 mins to travel from Santa Barbara to Los Angeles. Calculate the residual and explain the meaning of this residual value.

(f) Suppose Amtrak is considering adding a stop to the Coast Starlight 500 miles away from Los Angeles. Would it be appropriate to use this linear model to predict the travel time from Los Angeles to this point?

(g) If we instead measured distance in kilometers what (if anything) would change about the association and the corresponding regression line?

(h) Does a low $R^2$ mean we have a bad model?