**Lineup Activity:**

**Learning to Read Residual Plots**

The goal of this activity is to learn how to distinguish problematic patterns from random noise in residual plots from simple linear regression.

**BACKGROUND**A rail trail is a segment of abandoned railroad track that has been converted to a trail for recreation and exercise (e.g., walking, running, or cycling). Advocates of rail trails suggest that they have an economic benefit to the community, such as increased property values (Hartenian and Horton, 2015). Conventional wisdom suggests that proximity to greenspace or a park impacts the value of a home, so is this the case with rail trails? Hartenian and Horton (2015) explored the relationship between the sales price and distance from a rail-trail system for 104 homes in Northampton, Massachusetts in 2007. These homes were tracked on Zillow and their estimated sales prices in 2011 and 2014 were also recorded.

In this activity, you will use a simple linear regression model to predict property value from distance to the rail-trail system (in miles) for the homes in this data set. You will also consider whether the model adequately represents this association.

**DATA SET**The RailsTrails.csv data set consists of 30 variables collected on 104 homes. In this activity, you focus on two variables in this data set:

* Price2014 (Zillow’s estimate of the property value in 2014, in thousands of dollars)
* Distance (the distance, in feet, to the nearest rail-trail entry point).

Download the RailTrails.csv data set and import it into RStudio.

**GROUP QUESTIONS**

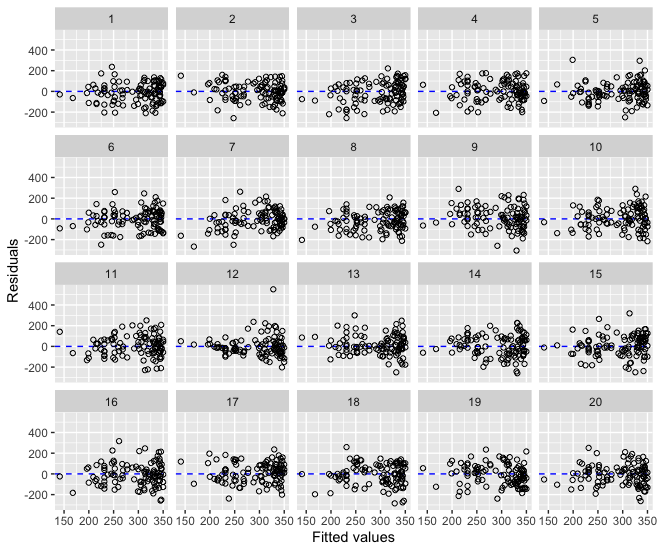
1. Which variable is the response variable? How do you know?
2. Which variable is the explanatory variable? How do you know?
3. Create a scatterplot displaying the relationship between the sales price and distance from the rail-trail system. Describe the relationship you observe in the plot. Be sure to mention form, direction, strength, and any unusual features.
4. Fit a simple linear regression model that predicts the sales price using the distance to the rail-trail system. Report the fitted regression equation below. Be sure to denote the names of the variables somewhere in your answer (e.g., use the names in the equation or define Y and X after the equation).
5. Provide an interpretation of the intercept in the context of the problem.
6. Provide an interpretation of the slope in the context of the problem.
7. The first house in our data set is 2.4 miles from the rail-trail system. Use the fitted regression equation to predict the price of this home.
8. The actual value of the home from question #7 is 210.729 thousand dollars. Calculate the residual for this home. How would you interpret this value?

In order to use the least squares regression model to describe the relationship between two variables and to use it for prediction, the following conditions must be met:

* The relationship between the variables is linear.
* The distribution of the residuals is symmetric and centered at 0.
* The spread of the residuals is constant across all values of the explanatory variable.
* The residuals are independent; thus, one point falling above/below the line does not impact any other point.

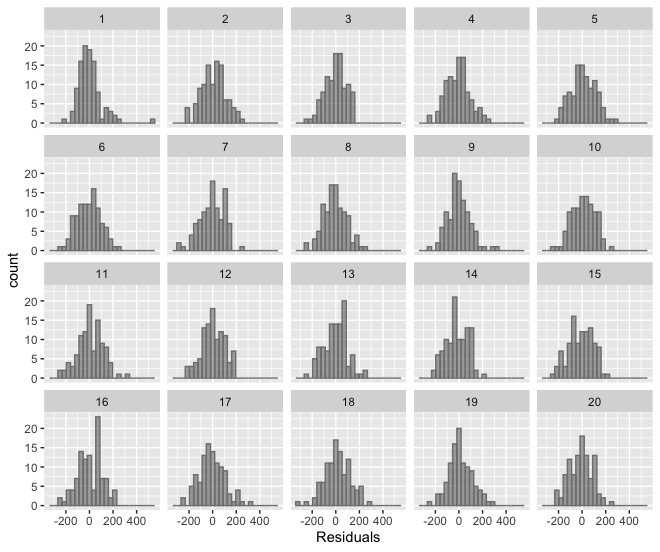
1. A residual plot is created by plotting the residuals on the y-axis and the fitted values on the x-axis. What conditions can you check using a residual plot?
2. What conditions can you check using a histogram of the residuals?

**INDIVIDUAL QUESTIONS** *Please do not discuss your answers with your group until you start question #17.*

A lineup of residual plots is created by placing the observed residual plot from your regression model in a field of 19 “decoy” residual plots that are generated from a simple linear regression model that meets all of the necessary conditions. A lineup of residual plots is shown below. 

1. Which plot do you think is the most different from the others?
2. What feature(s) of the plot led you to this choice?
3. Choose two other plots (i.e., plots that you think are decoys) in the lineup and describe any patterns that you see.

A lineup of residual histograms is shown below. Again, there are 19 decoy plots that show histograms from simple linear regression models that meet all of the necessary conditions.



1. Which histogram do you think is the most different from the others?
2. What feature(s) of the distribution led you to this choice?
3. Choose two other histograms (i.e., two decoys) in the lineup and describe the distributions.

**GROUP QUESTIONS**

Now that you have evaluated the two lineups, answer the following questions with your group.

1. Which residual plot do you think is the most different from the others? What feature(s) of the plot led you to this choice?
2. What patterns do you see in the decoy residual plots? To answer this question, be sure to describe the relationship between the residuals and fitted values in the decoy plots.
3. Which residual histogram do you think is the most different from the others? What feature(s) of the distribution led you to this choice?
4. Describe common features of the distributions of the decoy residual histograms that you see.

***STOP HERE!*** *We will have a large group discussion sharing the results and then the plots that display the observed residuals will be revealed.*

The observed residual plot in plot # \_\_\_\_\_\_.

The observed residual histogram in plot # \_\_\_\_\_\_.

1. Did your group choose the observed residual plot and histogram?
2. Based on your answer to the previous question, does it seem reasonable to use your regression model to predict the value of a home? Explain your reasoning.