# Instructor Guide to
# Lineups Activity: Learning to Read Residual Plots

**Quick Info**

Audience: *Intro or intermediate undergraduate statistics students*

Brief Description*: Students learn to distinguish between problematic patterns vs. random noise by choosing which plot in a lineup is not like the others.*

Topics Covered: *simple linear regression, residual analysis*

Learning Goals: *Distinguish problematic patterns from random noise in residual plots.*

Prerequisites: *simple linear regression (interpreting coefficients, definition of residual, conditions/assumptions), statistical graphics (histograms, scatterplots)*

Class Resources: *Students can use R, JMP, SPSS, etc. to fit the regression model and create the residual plot(s), or the output and graphics could be included on the student handout. Students will need access to the lineup (this can be done using a slide, handout, or a Shiny app).*

Instructor Resources:

- *Student handout and instructor guide*
- *Sample data set*
- *Shiny app (BLINDED URL)*
- *Tutorial on creating lineups in R (BLINDED URL)*

Time: *50-60 minutes in class*

**Why use this lineup activity in your course?**

Students often struggle to distinguish problematic patterns vs. random noise when they first start interpreting residual plots. Some students will see patterns everywhere, while other students will hesitate to call any pattern but the most glaring a problem. Lineups provide students with a helpful framework to learn to read residual plots, only requiring students to apply "Sesame Street logic" (i.e., "which one of these is not like the others"). Lineups are created by generating a number of decoy plots and randomly situating the data plot into this grid of plots. To create a lineup of residual plots, the residual plot for your model is randomly situated amongst a field of residual plots created in a way where no model violations exist (e.g., using a parametric bootstrap). Situating the "true" residual plot amongst the decoys forces students to compare what they are seeing in their model to what they should expect to see, helping them learn to read residual plots.

**When should you use this activity in your course and what are the prerequisites?**

This activity is designed to introduce students to residual plots, so I recommend using it the first day you discuss model diagnostics. Students should know the basics behind simple linear regression (how to interpret the coefficients, fitted vs. observed values, residuals) and have been introduced to the conditions necessary for its use. It is also assumed that students have a firm understanding of EDA, particularly how to interpret histograms and scatterplots.

**What is an example lesson plan for the activity?**

Note: Italicized text indicates a teaching tip or aside.

- **Introduce regression conditions and data set** (5 minutes)**.** At the beginning of class, take a few minutes to introduce (or review) the conditions for regression and the data set that will be used for the activity. Alternatively, the regression conditions could be introduced via reading or a pre-class video*. This is a good time to assign student roles in the groups.*
- **Small group discussion** (15 minutes). Send students into their small groups to discuss questions 1-10. These questions review fundamental ideas for simple linear regression. *I include quite a few review questions to make strong connections to the previous class(es) and to prepare groups for productive discussion (i.e., breaking the ice if folks aren't as talkative initially).*
- **Introduce lineups** (3 minutes)**.** After working through questions 1-10, introduce lineup plots. The important point here is that students know one panel is the observed residual plot and the others are generated from models where there are no violations to the conditions (i.e., they are "decoys"). Don't worry about the technical details here. *You could also take a few minutes to review the answers to questions #1-10 at this point.*
- **Individual work** (6 minutes). Send students to work individually on questions 11-16 to evaluate the lineup plots and determine which plot is most different from the others. It's important to remind students to work individually because discussion is more interesting and meaningful when each student is an unbiased evaluator.
- **Small group discussion** (5-10 minutes). Bring students back into their small groups to discuss questions 17-20 and come to a consensus that can be shared during a large group discussion. Having students unpack their thoughts about the lineups in groups helps them flesh out why they chose the panel. In addition, discussing one choice per group makes it easier to hear from "everyone" in a short amount of time in a lower stress environment (it's the group's choice, not the individual's). *I recommend reminding students that they need to arrive at consensus and that the spokesperson should be ready to share the group's thoughts with the class.*
- **Large group discussion.** Regroup and have each group briefly report what plot they chose and why (1 minute per group). A less time-consuming alternative is to have each group put a post-it note on the board under their choice with a brief rationale for their choice. If you choose the post-it approach, then call on one group to provide their rationale. *You could also use some online polling platform (e.g., Google forms, Moodle, etc.) to collect this information. Even if you have each group report their choice and rationale, the post-it note approach provides you with a "progress bar" to gauge how things are going. You could extend this conversation to ask about the implications of other patterns, or you could devise another activity or homework assignment to explore those situations with lineups.*
- **Small group discussion** (5 minutes). Bring students back into their small groups to discuss questions 21-22. These questions ask students to interpret the implications of their lineup selections.
- **Debrief** (10 minutes)**.** Once the lineups have been discussed, review the key ideas that were introduced and discuss the importance of model checking.

**Do you have any hints for using lineups?**

1. Spend some time introducing the data set. This can be done while you distribute the handouts and as students get situated in their groups.
2. Have students work in small groups, perhaps 3 or 4. This provides students with the opportunity to discuss their understanding in a lower-stakes environment and to learn from each other.
3. Assign roles to each group member to help groups function efficiently and to avoid the situation where one student does all of the work. We recommend assigning roles such as
   - Facilitator: makes sure the group stays on task and that each member has room to contribute
   - Spokesperson: reports back to the class, reads from the recorder's notes
   - Recorder: completes the worksheet for the group, takes coherent notes

- Encourager/Questioner: suggests alternatives if the group gets stuck, asks for clarification, poses questions

**Follow-up Activities and Discussion Questions:**

You can follow-up this activity with homework questions, or warm-up questions for the next class, where lineups are used to explore a model without deficiencies and a model with a different deficiency. Here are a few ideas that you could try:

- Create a couple lineup plots to diagnose models with different/no model violations to explore at the start of the next class period. You could have them use the Shiny app to create the lineups and discuss what they find, or simply use them as conversation starters to reinforce the ideas from this activity.
- You could have students sketch a (small) lineup that would indicate a certain model deficiency as a class or homework exercise.
- Do embellishments help you interpret residual plots? You could have students evaluate lineups with and without LOESS smoothers and ask about the potential benefits and pitfalls of the smoothers.
- The SLR model discussed in this activity is overly simplistic, so you could have students discuss what other variables might be important to consider when exploring the value of a home. This could take many forms, such as an in-class discussion, a homework problem, a minute paper, or an exam question.

**What else is in this Instructor Guide?**

In the next section, we provide a commented version of the student activity. We suggest possible alternative formats you can use, questions that you can ask students to facilitate discussion, and possible issues you may encounter.

**References**

Hartenian, E., & Horton, N. J. (2015). Rail Trails and Property Values: Is There an Association? *Journal of Statistics Education*, *23*(2). DOI: 10.1080/10691898.2015.11889735

The format of this instructor guide was inspired by Shonda Kuiper's Stat2Labs.

# 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.

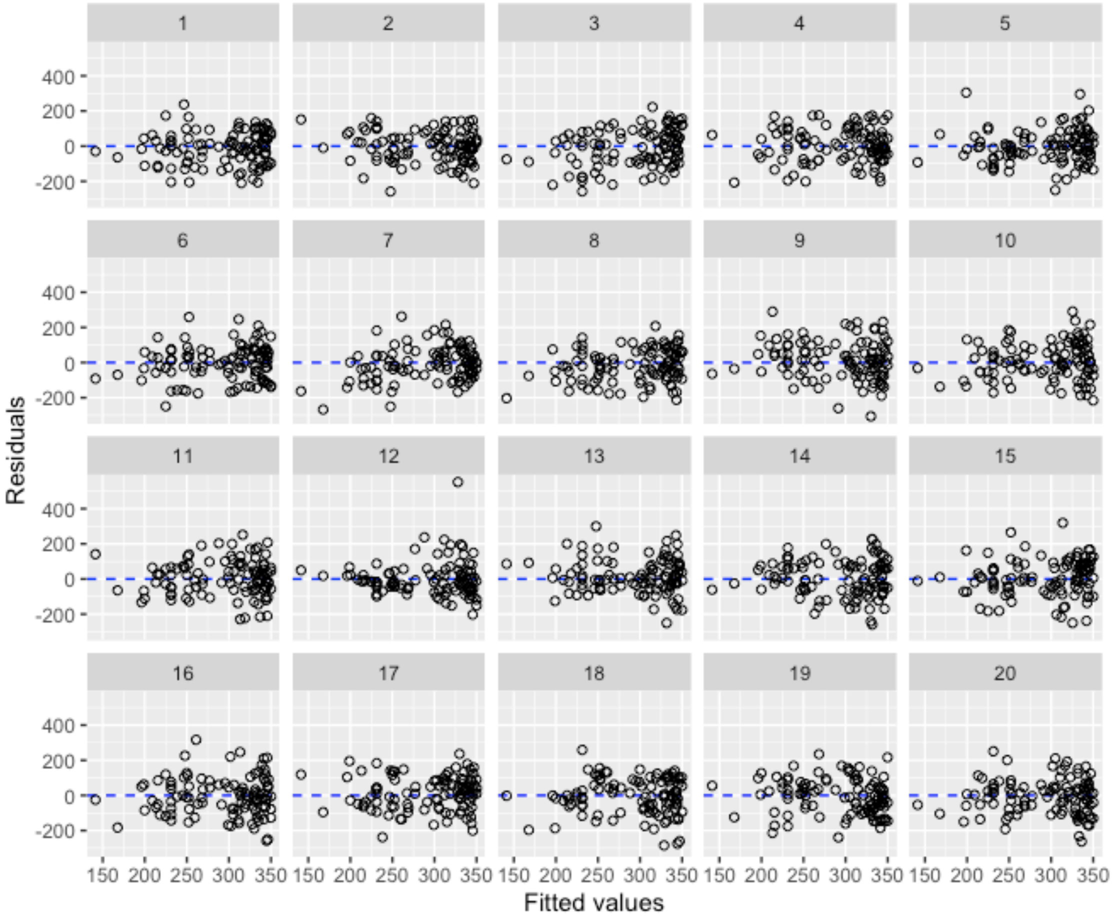
- The residuals are independent; thus, one point falling above/below the line does not impact any other point.

9. 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?
10. 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.*

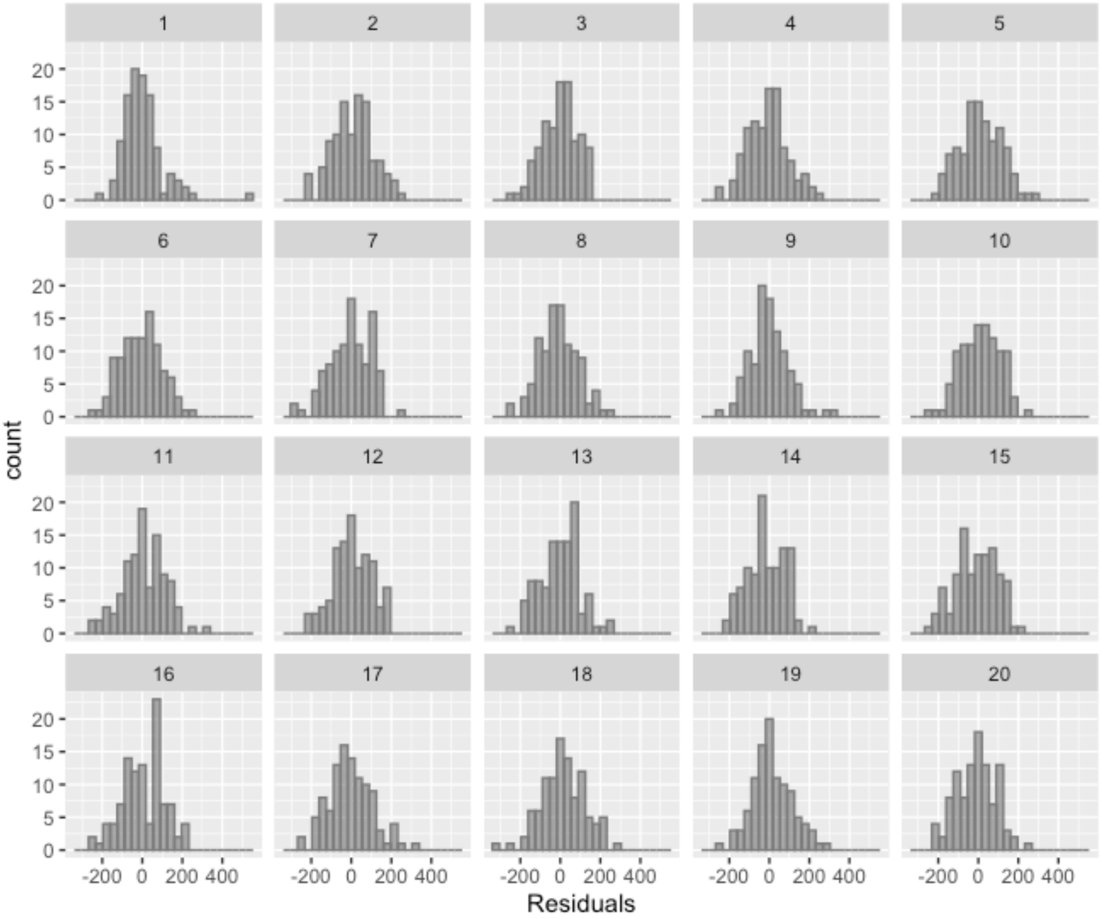
A lineup of residual plots is created by placing the observed residual plot from your regression model in a field of 19 "decoy" residual plots that are generated from a simple linear regression model that meets all of the necessary conditions. A lineup of residual plots is shown below.

11. Which plot do you think is the most different from the others?
12. What feature(s) of the plot led you to this choice?
13. 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.

14. Which histogram do you think is the most different from the others?
15. What feature(s) of the distribution led you to this choice?
16. 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.

17. Which residual plot do you think is the most different from the others? What feature(s) of the plot led you to this choice?
18. 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.
19. Which residual histogram do you think is the most different from the others? What feature(s) of the distribution led you to this choice?
20. 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 # ______.

21. Did your group choose the observed residual plot and histogram?
22. 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.

This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.