**Instructor Guide: Learning to Read Residual Plots via Lineups**

**Quick Info**

Level: *Intro or intermediate undergraduate statistics*

Brief Description: *Students learn to interpret residual plots by trying to identify the observed plot in a field of decoys (i.e., a lineup).*

Topics Covered: *simple linear regression, residual analysis*

Learning Goals: *Interpret residual plots. Discern problematic patterns vs. random scatter. Review interpretations of simple linear regression models and conditions necessary for their use.*

Software Required: *Optional. 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 the overhead, via handouts, or by using a* [*Shiny app*](http://shinyapps.its.carleton.edu/multi-sample-quant-app/)*).*

Prerequisites: *simple linear regression (interpretation, definition of residual, conditions)*

Time: *50-60 minutes in class*

Instructor Resources: *Student handout and instructor guide, sample data set, Shiny app (*[*http://shinyapps.its.carleton.edu/multi-sample-quant-app/*](http://shinyapps.its.carleton.edu/multi-sample-quant-app/)*), tutorial on creating lineups in R (*[*https://aloy.github.io/classroom-vizinf/*](https://aloy.github.io/classroom-vizinf/)*)*

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

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. Lineups provide a helpful framework to build intuition using “Sesame Street logic” (i.e., “which one of these is not like the others”).

**What type of course is this lineup activity designed for?**

This activity is designed for an introductory undergraduate statistics course but could help intermediate students review residual analysis.

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

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.

**How should you conduct the activity? How much time should you expect to allocate?**

* **Introduce regression conditions** (5 minutes)**.** At the beginning of class, take a few minutes to introduce (or review) the conditions for regression. *The regression conditions could be introduced via reading or a pre-class video. This is also a good time to assign student roles in the groups.*
* **Students discuss questions #1-9 in groups** (15 minutes). 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 prime groups for productive discussion.*
* **Introduce lineups** (3 minutes)**.** 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. Don’t worry about the technical details here. *You could also take a few minutes to review the answers to questions #1-9 at this point.*
* **Questions #10 & 11: Students evaluate the residual plot and lineup individually** (4 minutes).Be sure to remind students to work individually on this evaluation. Discussion is more interesting and meaningful when each student is an unbiased evaluator of both the residual plot and the lineup.
* **Question #12: Groups discuss lineups and come to consensus** (5-10 minutes).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).
* **Regroup and reveal the true residual plot.** Have each group briefly report what panel they chose and why (1 minute per group). Display the lineup on a slide and announce and/or highlight the data panel. Ask one of the groups who selected the data panel why they made that choice and what it implies about the model (3 minutes). *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.*
* **Questions #13 & 14: Students evaluate the residual plot and lineup individually** (4 minutes). Same idea as before.
* **Question #15: Groups discuss lineups and come to consensus** (5-7 minutes). Same idea as before.
* **Regroup and reveal the true residual plot** (1 minute/group + 3 minutes).Same idea as before.
* **Debrief** (10 minutes)**.** Once both lineups have been specifically discussed, review the key ideas that were introduced and discuss the importance of model checking.

**Here are some helpful hints for instructors using lineups for the first time.**

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.

**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.

**Learning to Read Residual Plots via Lineups**

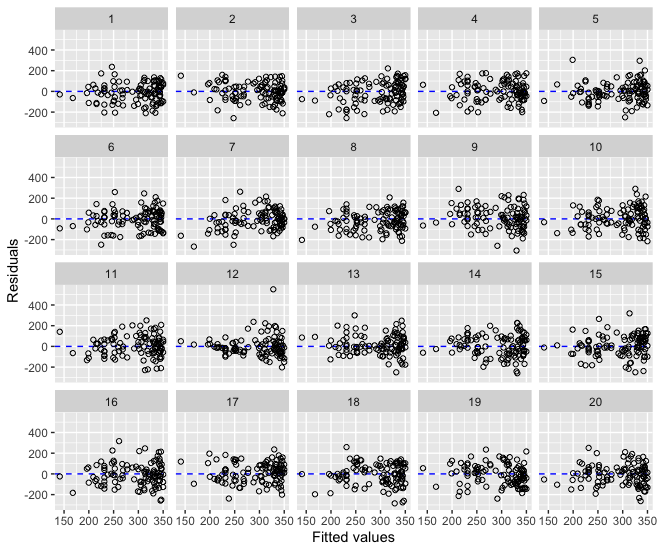
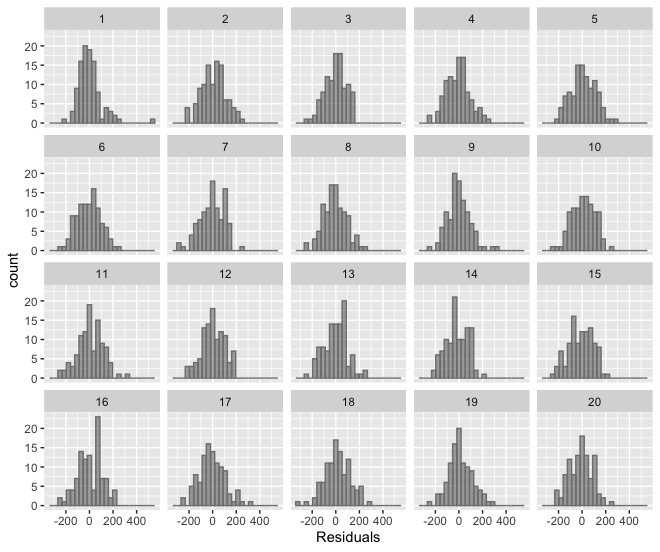
**Background**A rail trail is a segment of abandoned railroad track that has been converted to a trail for recreation and exercise (e.g., walking, running, or cycling). Hartenian and Horton (2015) explored the relationship between the sales price and distance from rail-trail system for 104 homes in Northampton, Massachusetts. In this activity you will use a simple linear regression model to describe the association between property value and 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.

The RailsTrails data set can be loaded into R via the Stat2Data package. To do this, run the following command:

data("RailsTrails", package = "Stat2Data")

The RailsTrails data set should consist of 30 variables collected on 104 cases. In this activity, you will focus your attention on two of these variables: Price2014 (Zillow’s estimate of the property value in 2014, in thousands of dollars) and Distance (the distance, in feet, to the nearest rail-trail entry point).

*Answer question #1-9 as a group.*

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. Use R to 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.
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?
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 we check using a residual plot?
10. Here is a residual plot for the model you fit in question #4. Does this plot provide any evidence that the regression model is not appropriate? Justify your answer.
11. A lineup of residual plots is created by placing the observed residual plot from question #10 in a field of 19 “decoy” residual plots that are generated from a simple linear regression model that meets all of the necessary conditions. Which panel contains the residual plot from question #10? 
12. Now that you have chosen the observed residual plot, answer the following questions with your group.
    1. Which panel contains the residual plot from question #10?
    2. Choose three decoy residual plots and describe any patterns that you see.
    3. Is the observed residual plot systematically different from the decoy residual plots?
    4. What does your answer to part iii indicate about the appropriateness of your regression model?
13. Here is a histogram of the residuals for the model you fit in question #4. Does this plot provide any evidence that the regression model is not appropriate?
14. A lineup of 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. Which panel contains the residual plot from question #10? 
15. Now that you have chosen the observed histogram, answer the following questions with your group.
    1. Which panel contains the histogram from question #10?
    2. Choose three decoy residual plots and describe any patterns that you see.
    3. Is the observed residual plot systematically different from the decoy residual plots?
    4. What does your answer to part iii indicate about the appropriateness of your regression model?



This work is licensed under a [Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License](http://creativecommons.org/licenses/by-nc-sa/4.0/).