Learning to Read Residual Plots via Lineups

Adam Loy

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)[1] 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.

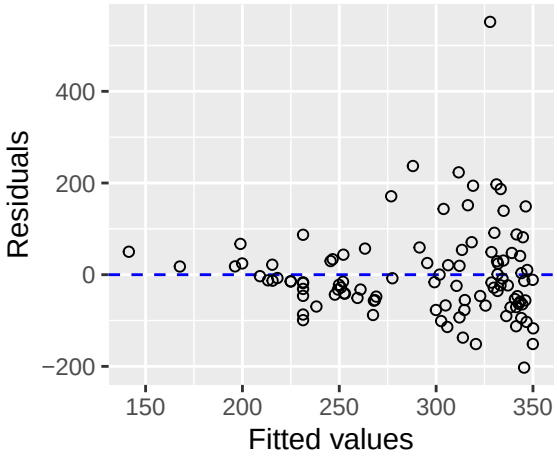
The `RailsTrails` data set can be loaded into R via the `Stat2Data` package. To do this, run the following command:

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

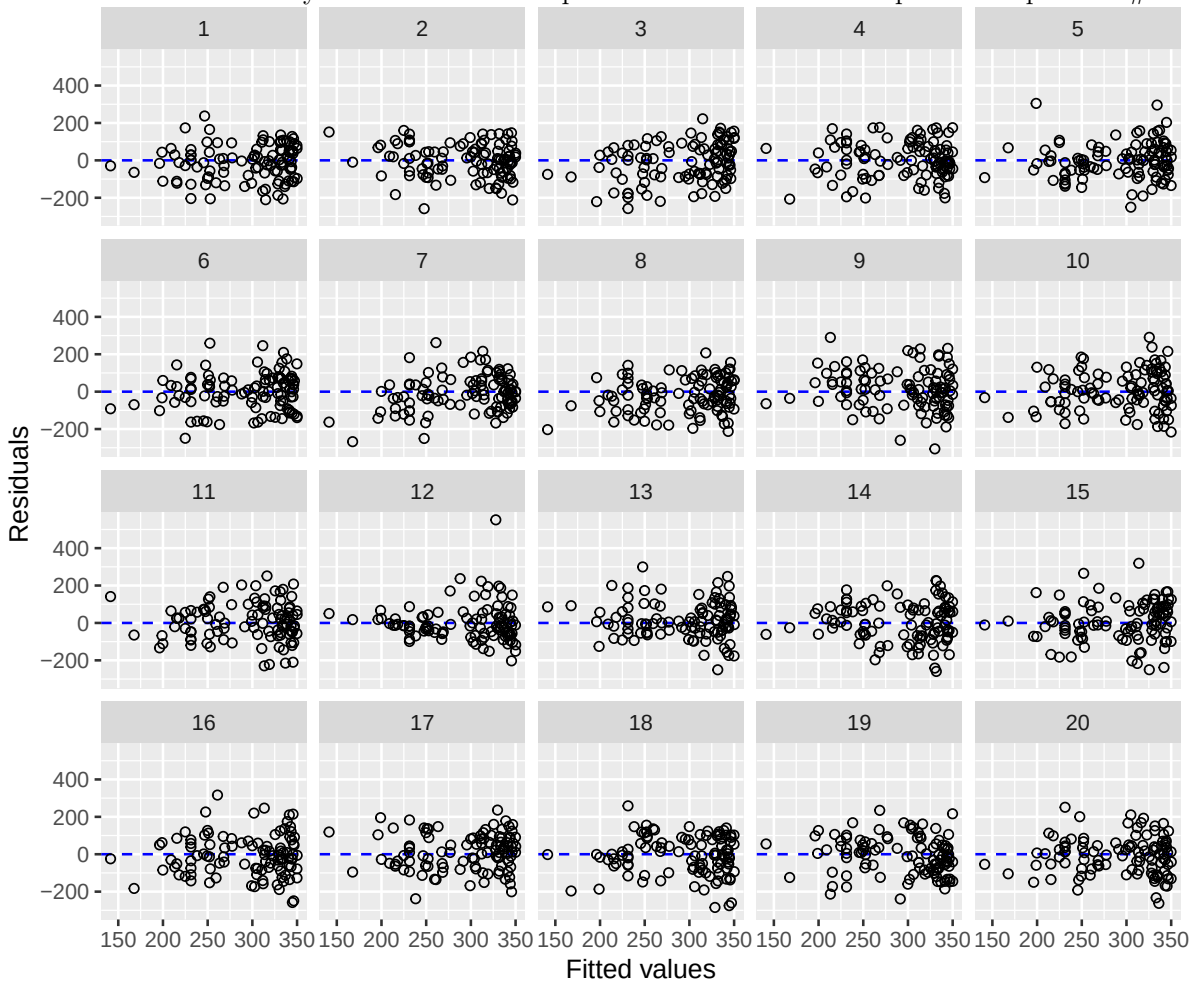
The `RailsTrails` data set should consist of 30 variables collected on 104 cases. In this activity, you will focus your attention on two of these variables: `Price2014` (Zillow's estimate of the property value in 2014, in thousands of dollars) and `Distance` (the distance, in feet, to the nearest rail-trail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

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 the 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. 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. Below 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?

[1] Data source: Hartenian, E., & Horton, N. J. (2015). Rail Trails and Property Values: Is There an Association? *Journal of Statistics Education*, 23(2).



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?

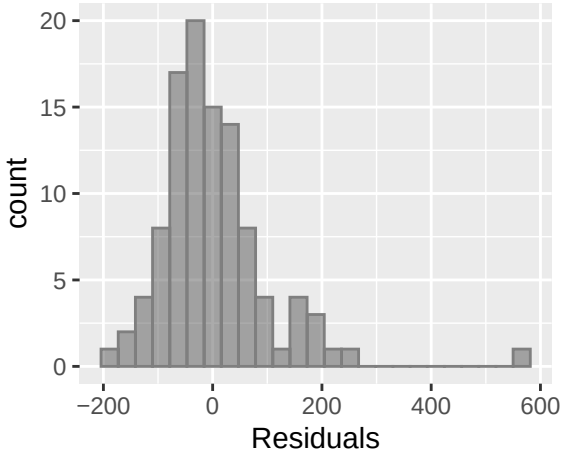


12. Now that you have chosen the observed residual plot, answer the following questions with your group.

    i. Which panel contains the residual plot from question #10?
    ii. Choose three decoy residual plots and describe any patterns that you see.

iii. Is the observed residual plot systematically different from the decoy residual plots?
iv. What does your answer to part iii indicate about the appropriateness of your regression model?

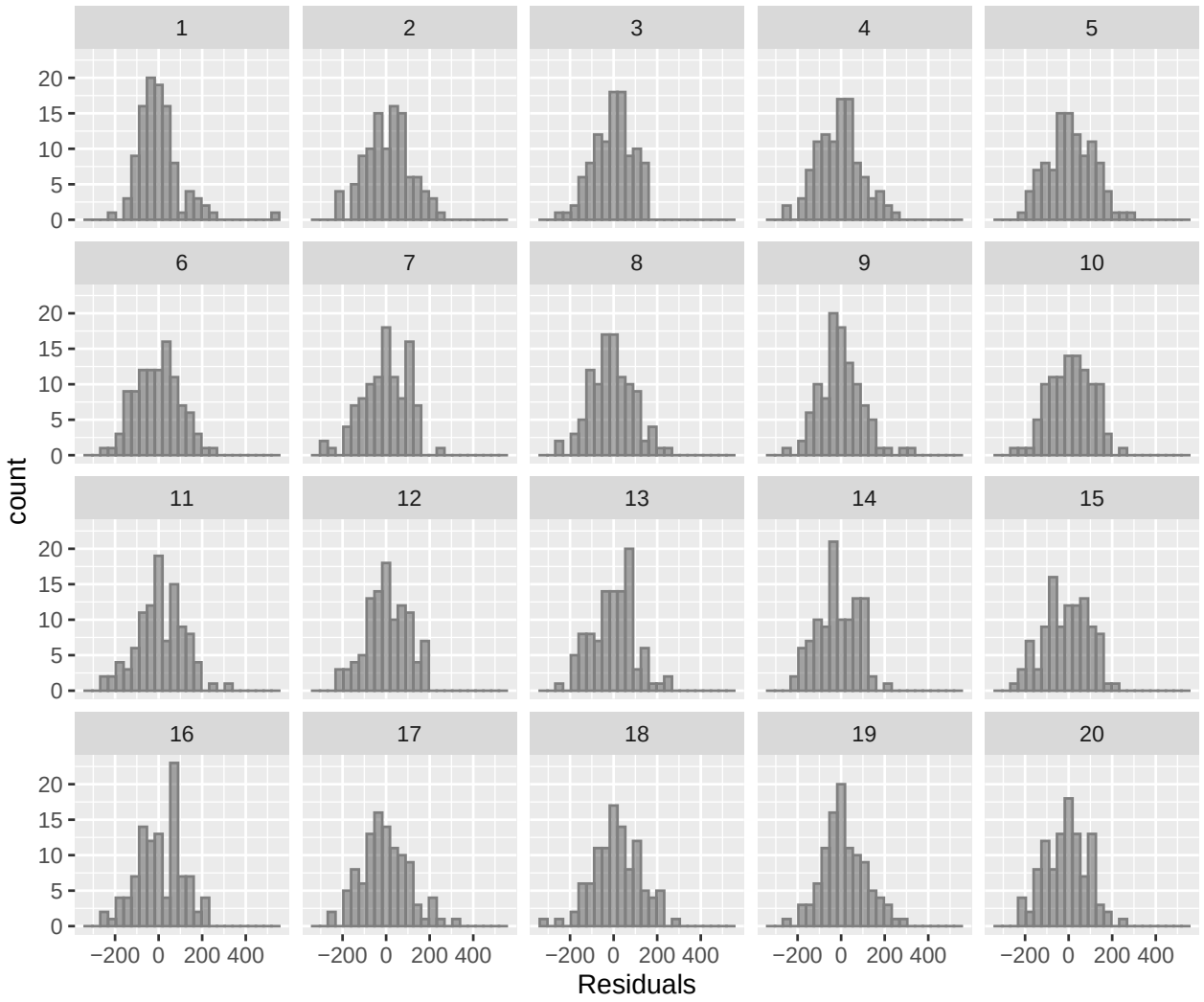
13. Below is a histogram of the residuals for the model you fit in question #4. Does this plot provide any evidence that the regression model is not appropriate?



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.

i. Which panel contains the histogram from question #10?
ii. Choose three decoy residual plots and describe any patterns that you see.
iii. Is the observed residual plot systematically different from the decoy residual plots?
iv. What does your answer to part iii indicate about the appropriateness of your regression model?


1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
count
0
5
10
15
20
−200
0
200
400
Residuals