Photo credit: Dave DiCello

17-803 Empirical Methods

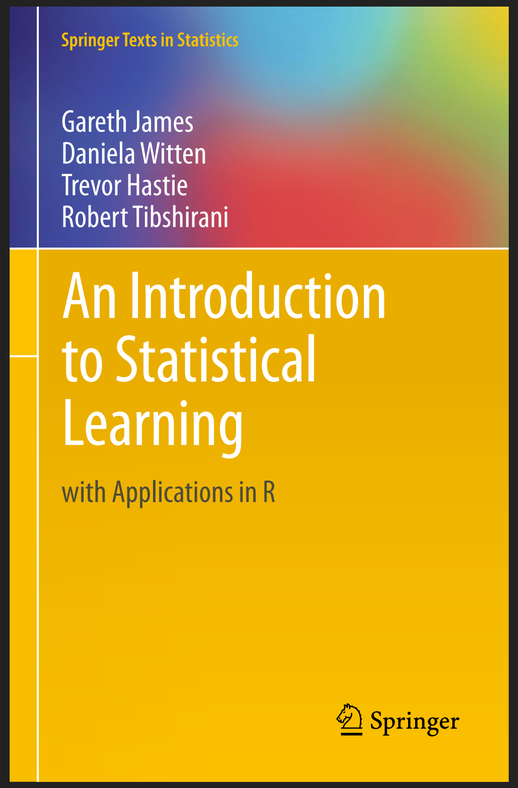
Bogdan Vasilescu, S3D

Regression Modeling (Part 2)

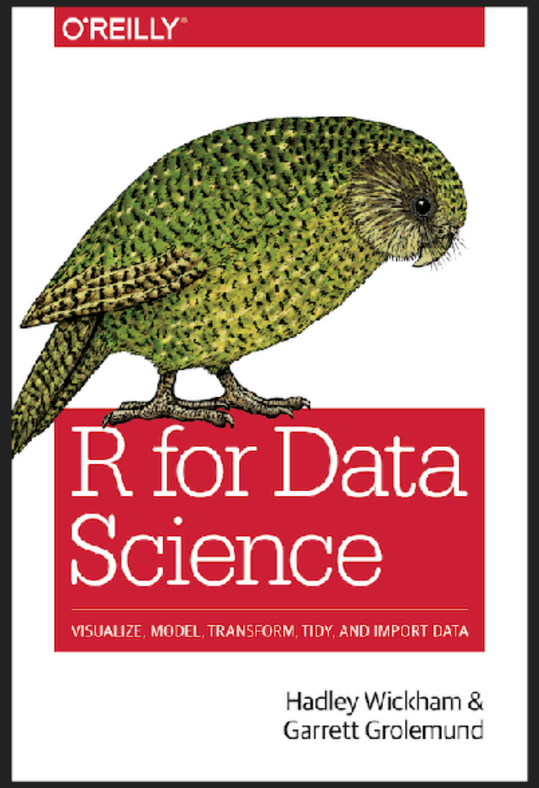
Thursday, March 21, 2024

Outline for Today

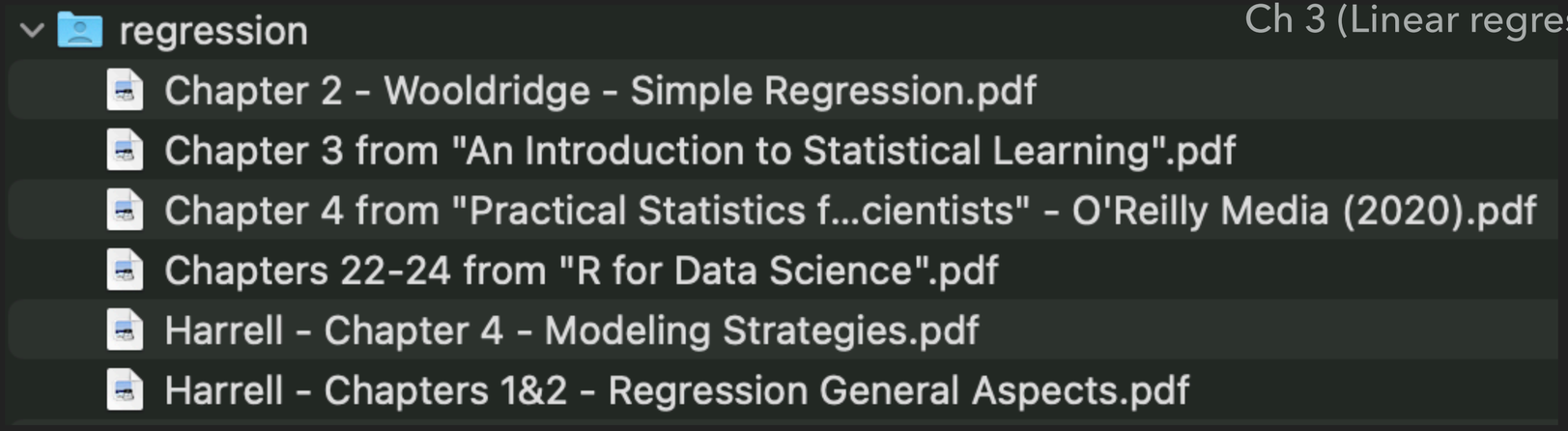
- More linear regression (Rmd + only limited slides)



Ch 3 (Linear regression)



Ch 22-24 (Modeling)



More To Read

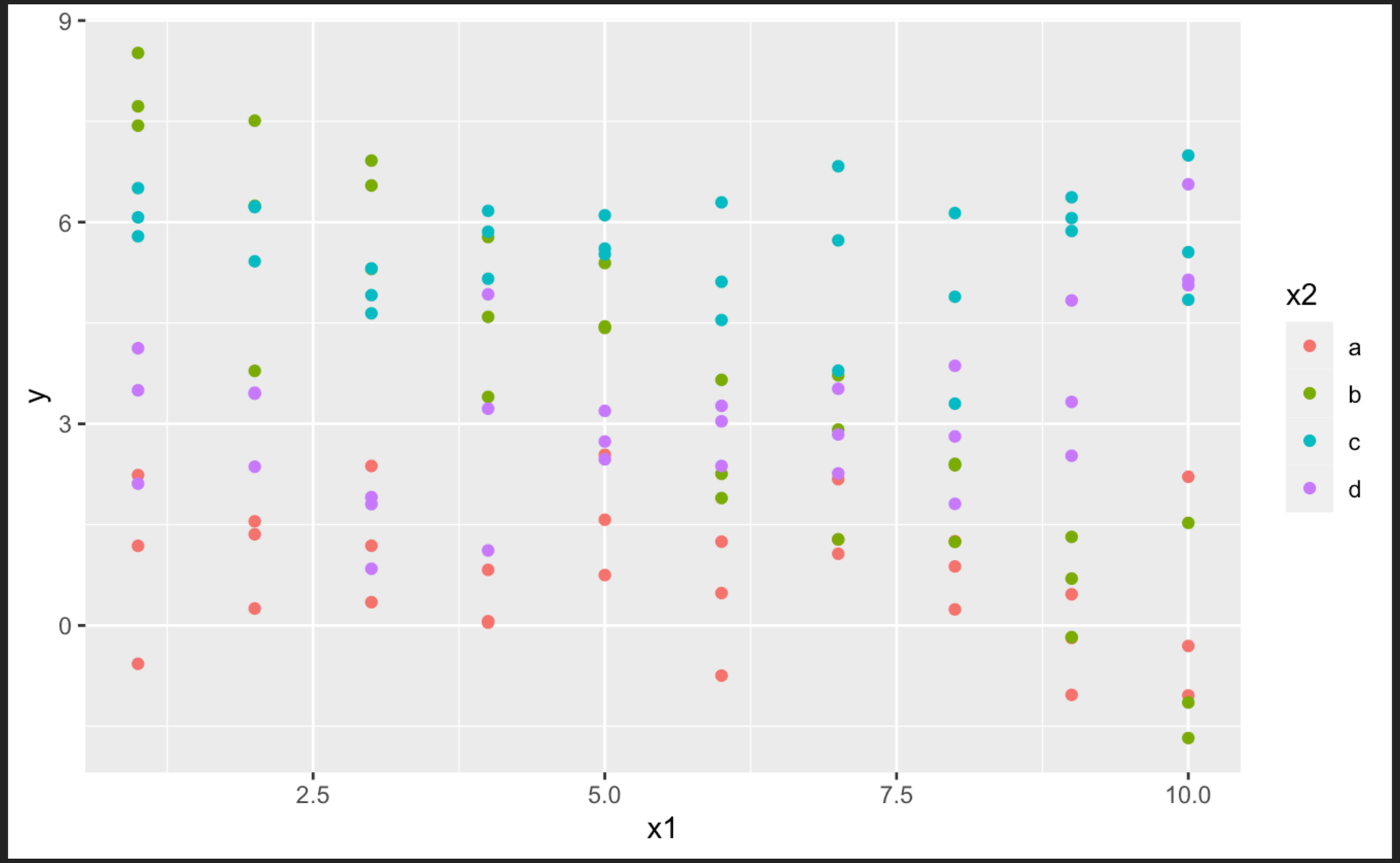
- Dealing with outliers: https://andrewpbray.github.io/reg/week6B-outliers.html
- Assumptions: https://blog.msbstats.info/posts/2018-08-30-linear-regression-assumptions/
- Diagnostics:
  - Anscombe: https://andrewpbray.github.io/reg/week6A-diagnostics.html
  - https://www.andrew.cmu.edu/user/achoulde/94842/homework/regression_diagnostics.html
  - https://data.library.virginia.edu/diagnostic-plots/
- Q-Q plots: http://seankross.com/2016/02/29/A-Q-Q-Plot-Dissection-Kit.html
- Interactive visualization: https://gallery.shinyapps.io/slr_diag/
- How to code categorical variables in a regression: https://stats.idre.ucla.edu/r/library/r-library-contrast-coding-systems-for-categorical-variables/
- Understanding model outputs: https://www.andrew.cmu.edu/user/achoulde/94842/
- Alpha vs p-value: https://rationalwiki.org/wiki/Statistical_significance#Alpha_value_versus_p-value

See Also

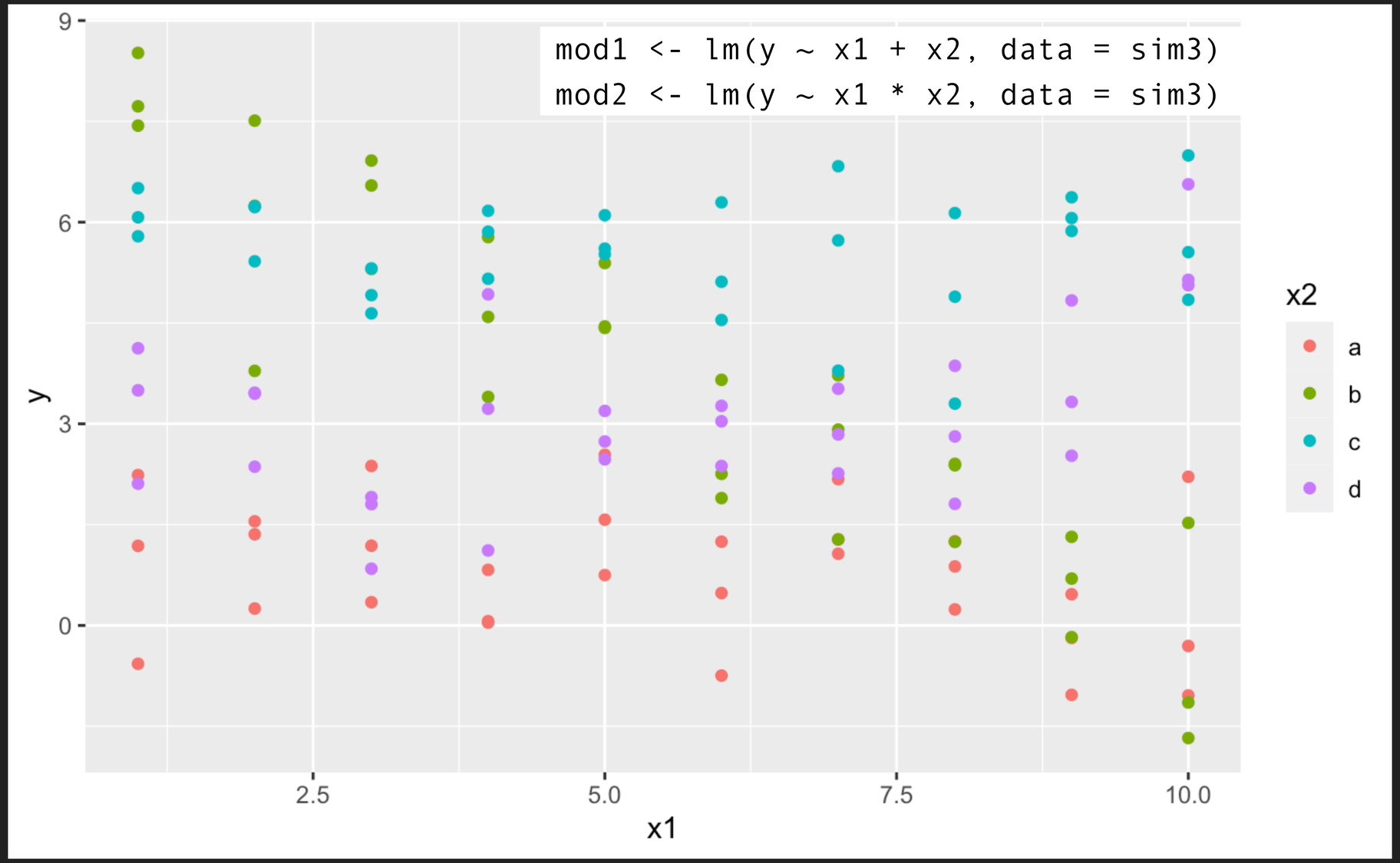
- CMU 94-842: Programming in R for Analytics:
  - https://www.andrew.cmu.edu/user/achoulde/94842/

Another interaction example

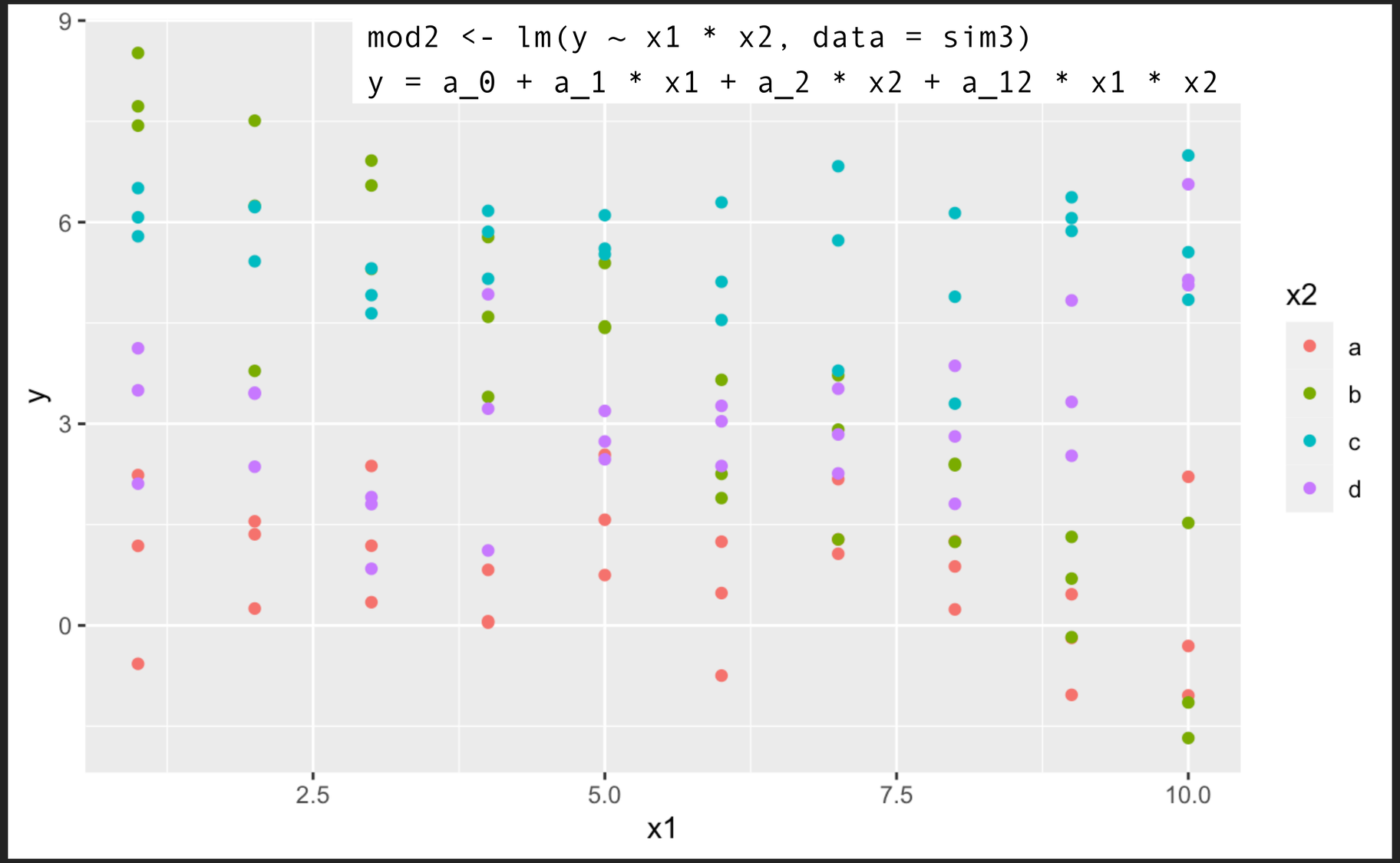
What happens when you combine a continuous and a categorical variable?



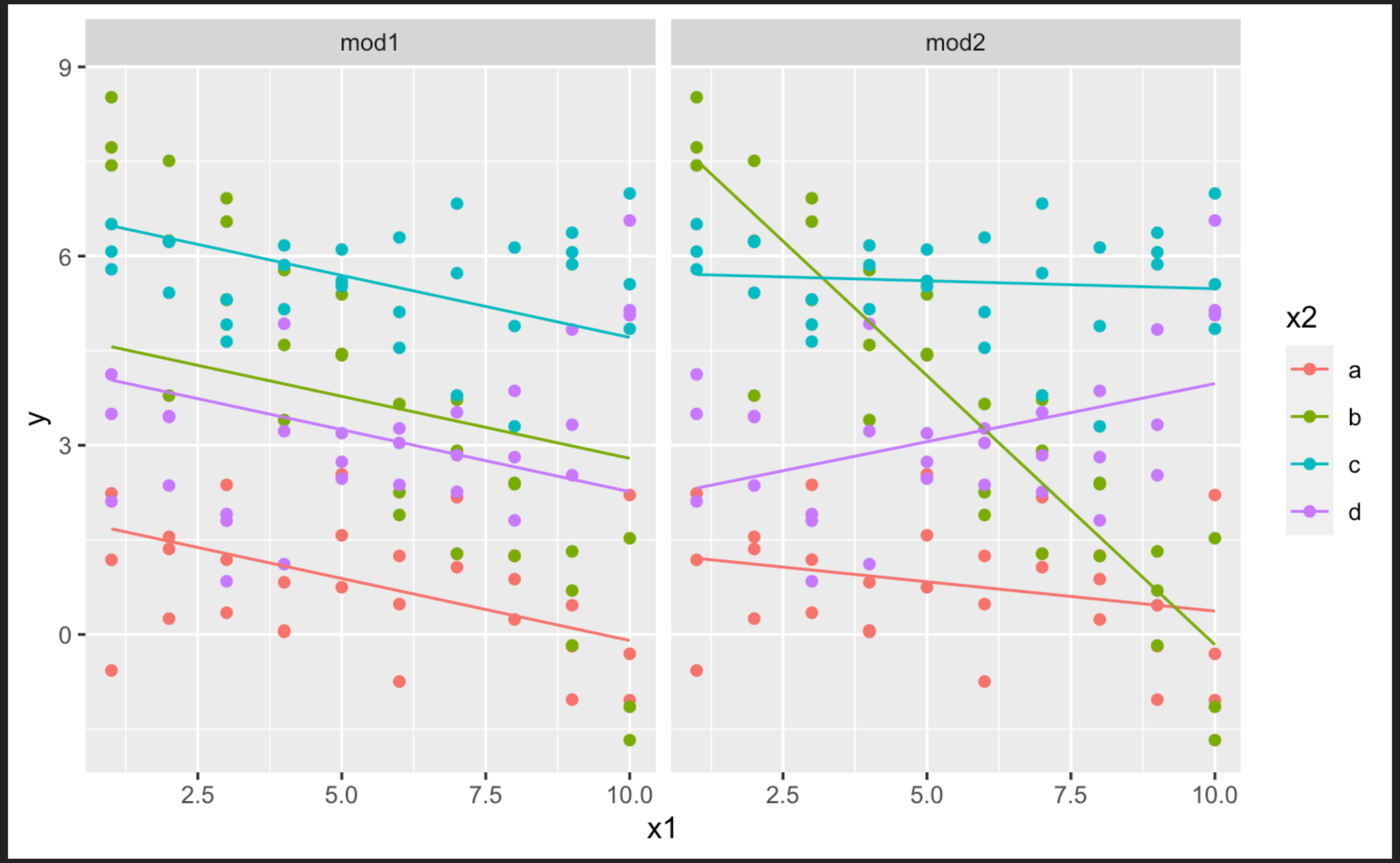
There are two possible models you could fit to this data



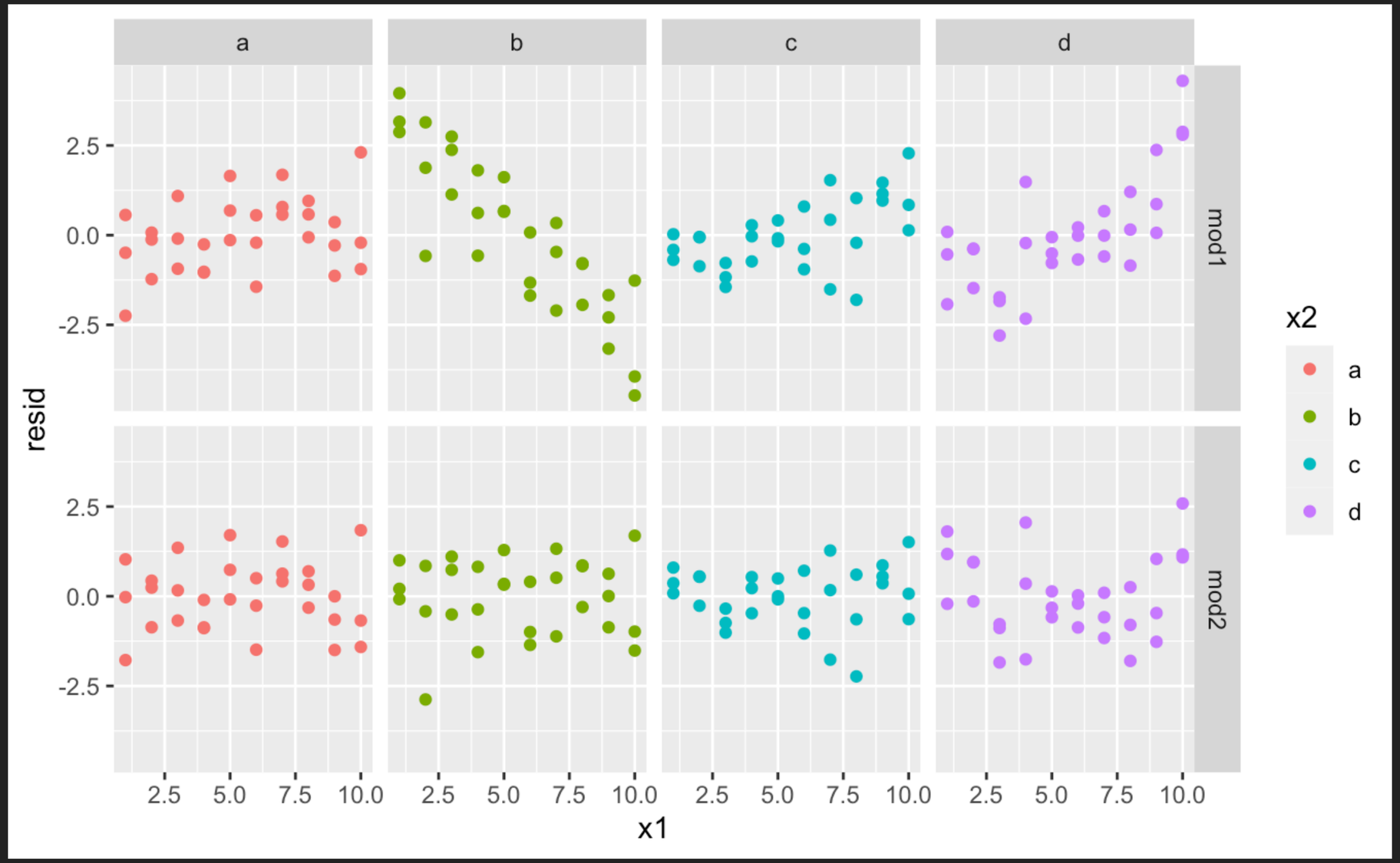
*: Both the interaction and the individual components are included in the model



The model using * has a different slope and intercept for each line

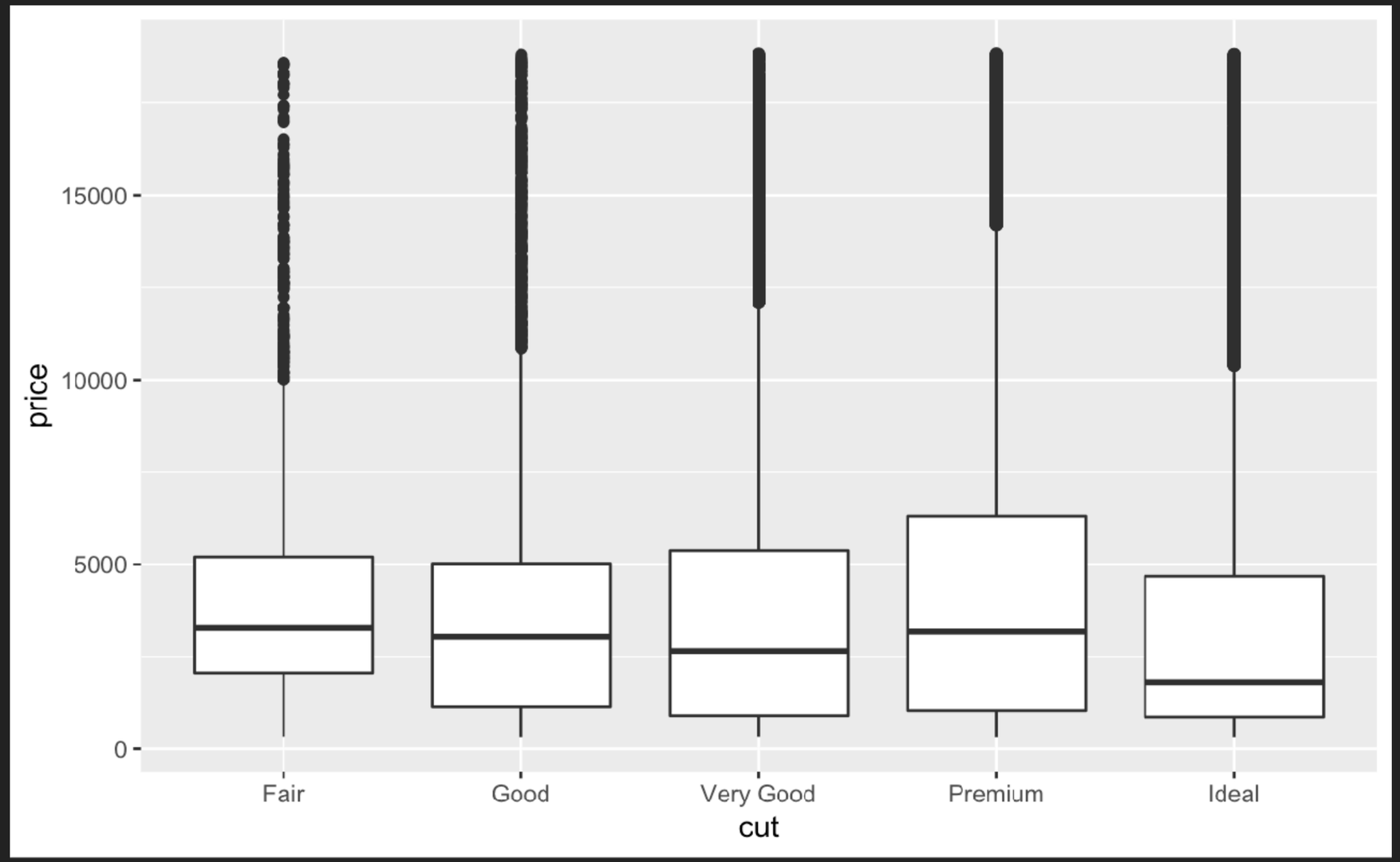


Which model is better for this data?



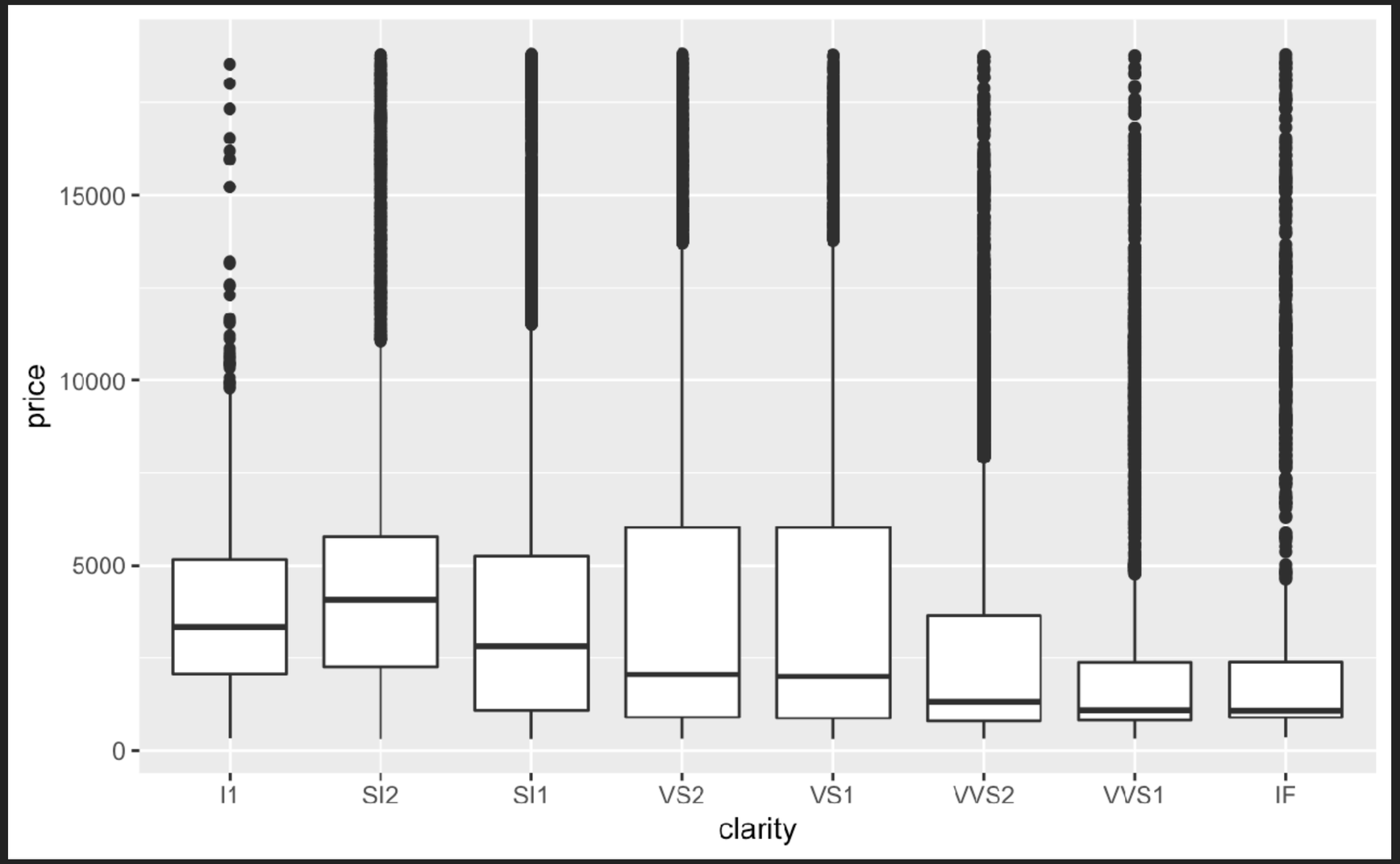
Another example

Why are low quality diamonds more expensive?



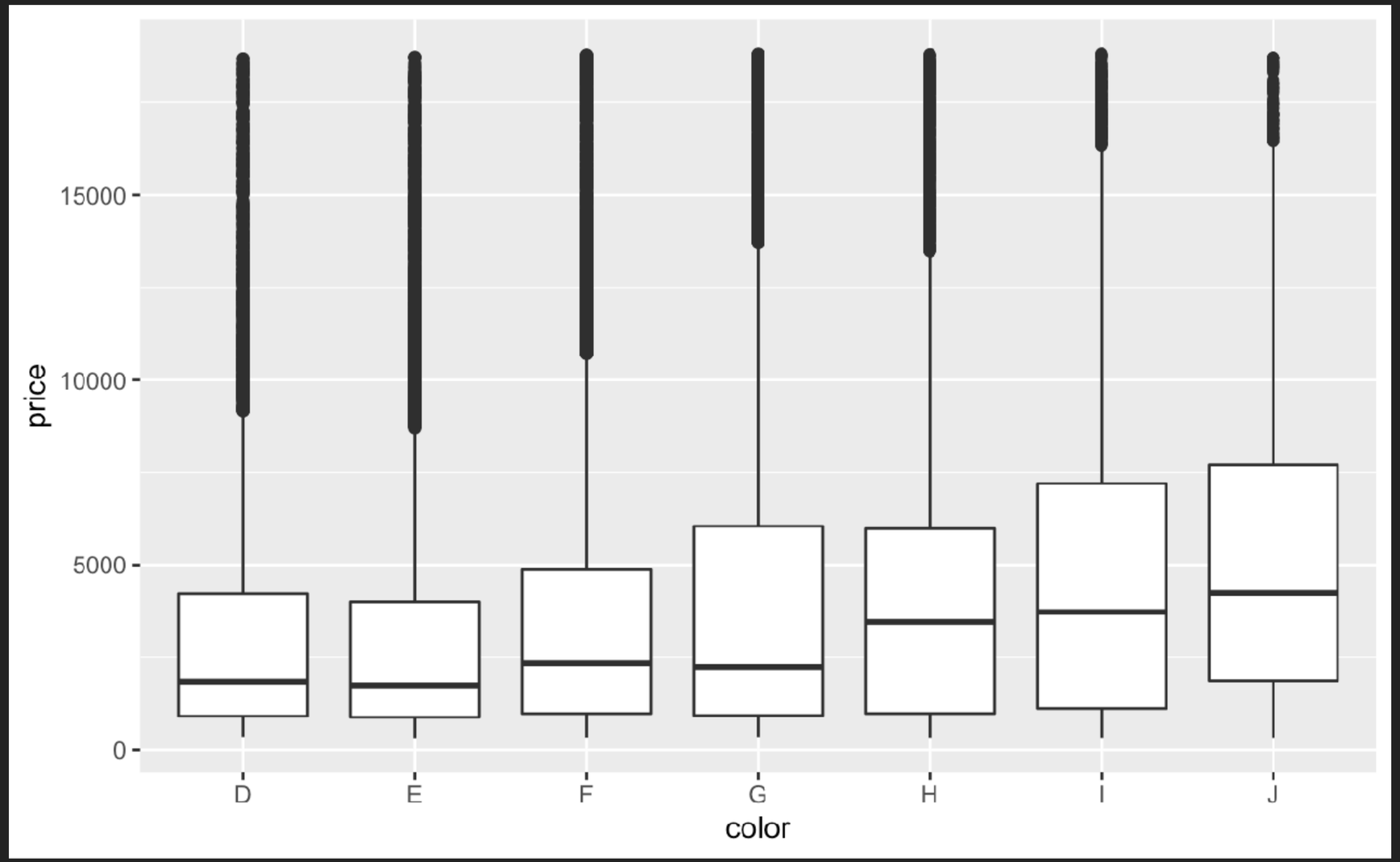
- Fair: worst cut

Why are low quality diamonds more expensive?



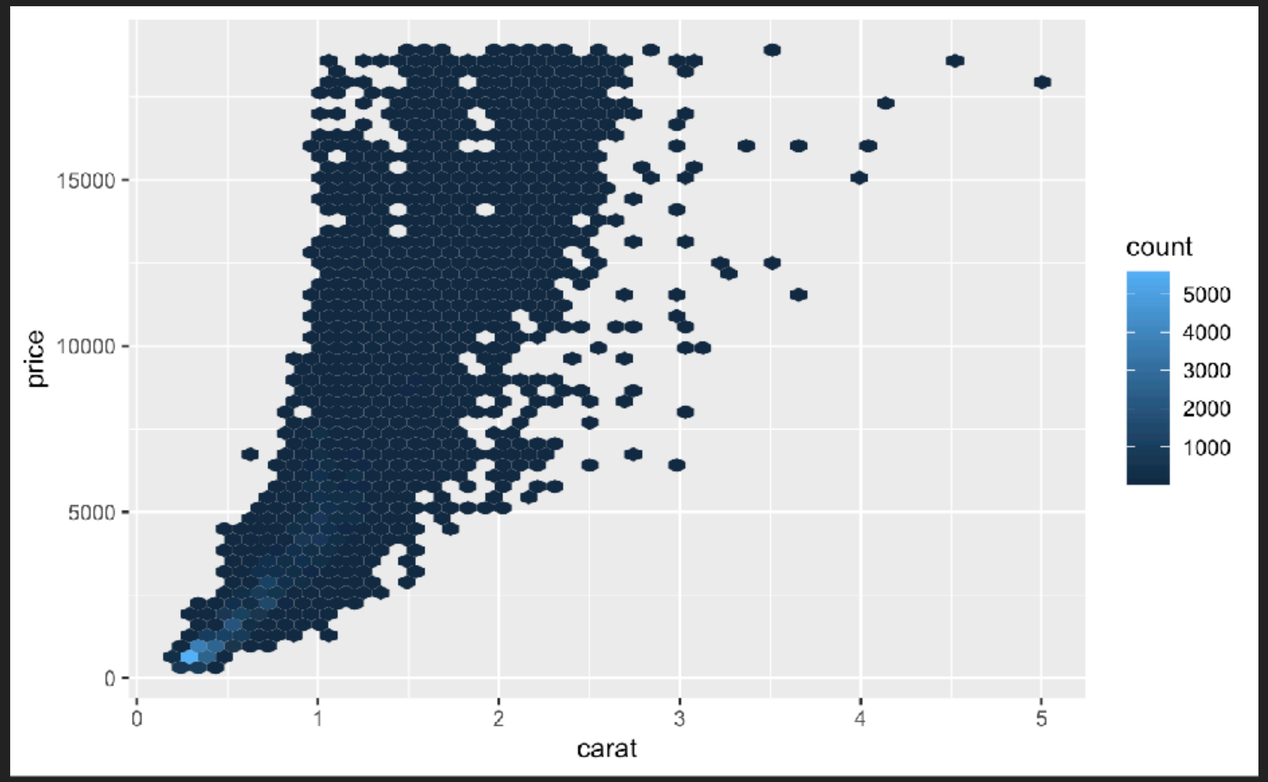
- I1: inclusions visible to the naked eye (worst clarity)

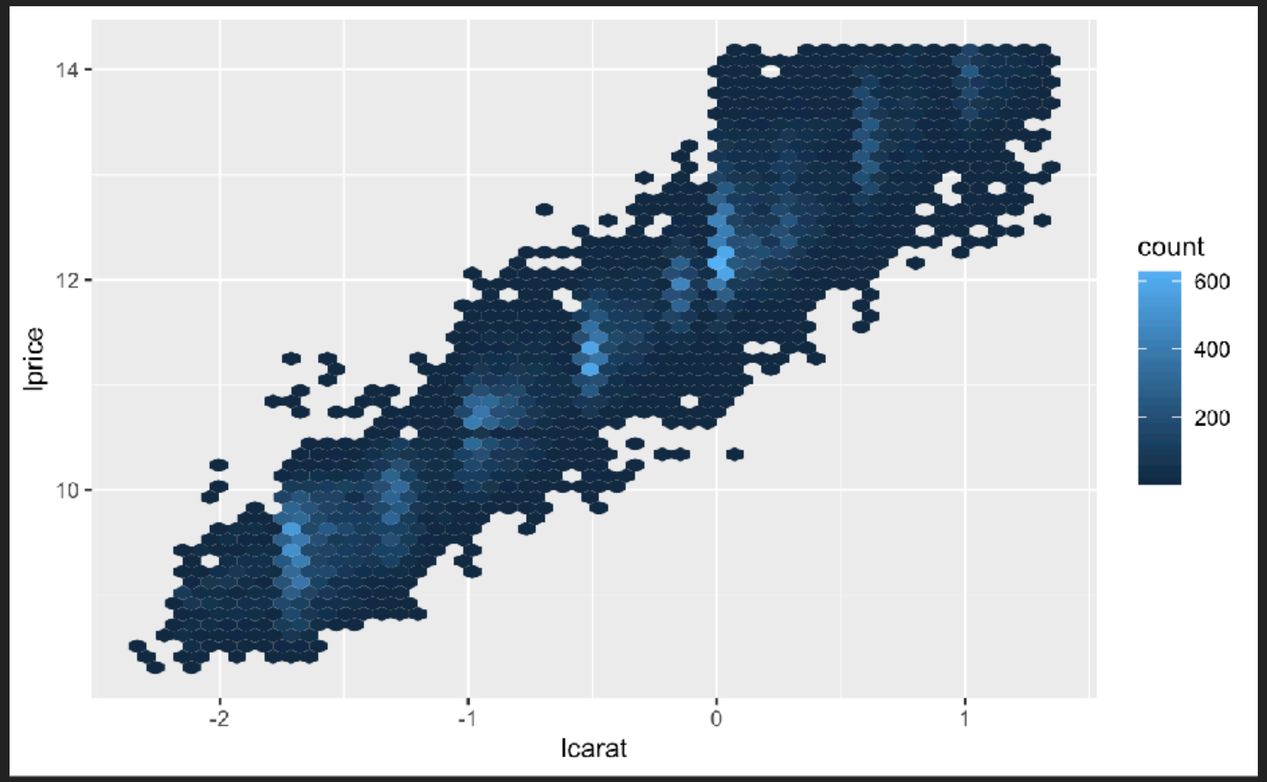
Why are low quality diamonds more expensive?



- J: slightly yellow (worst color)

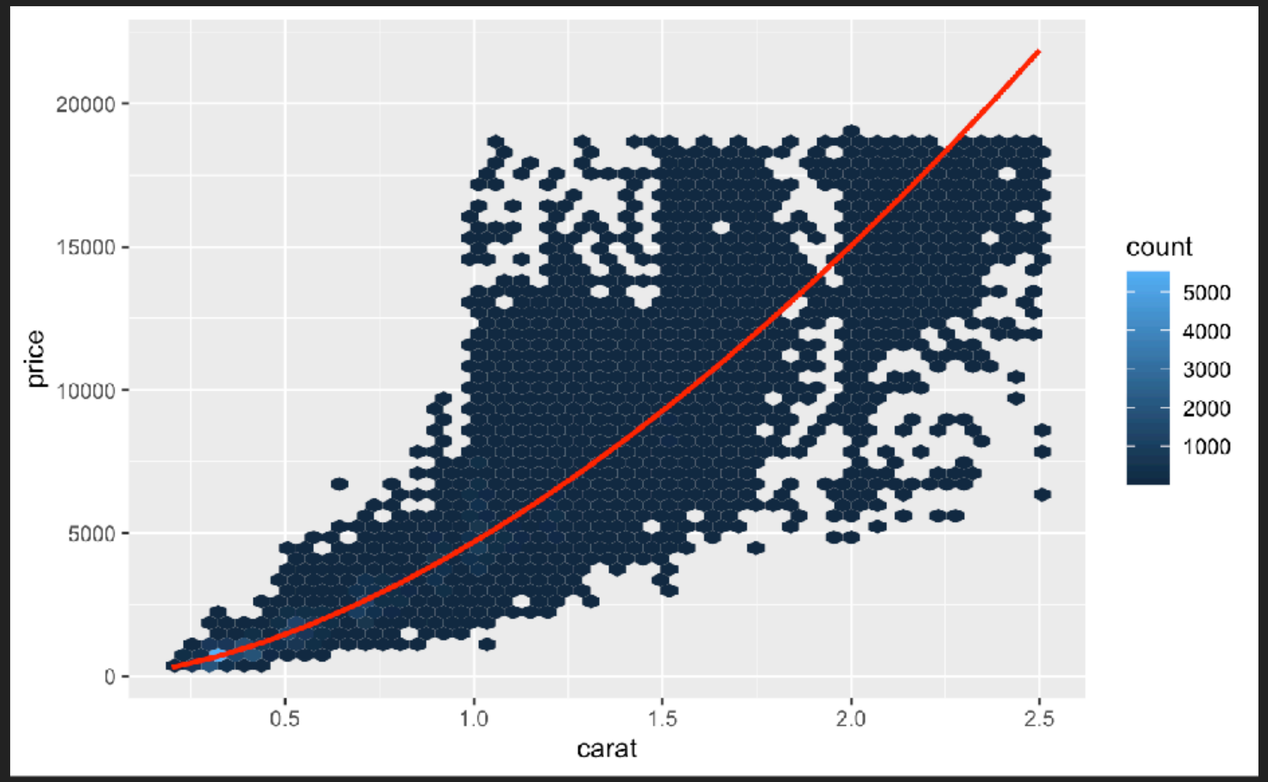
Because lower quality diamonds tend to be larger

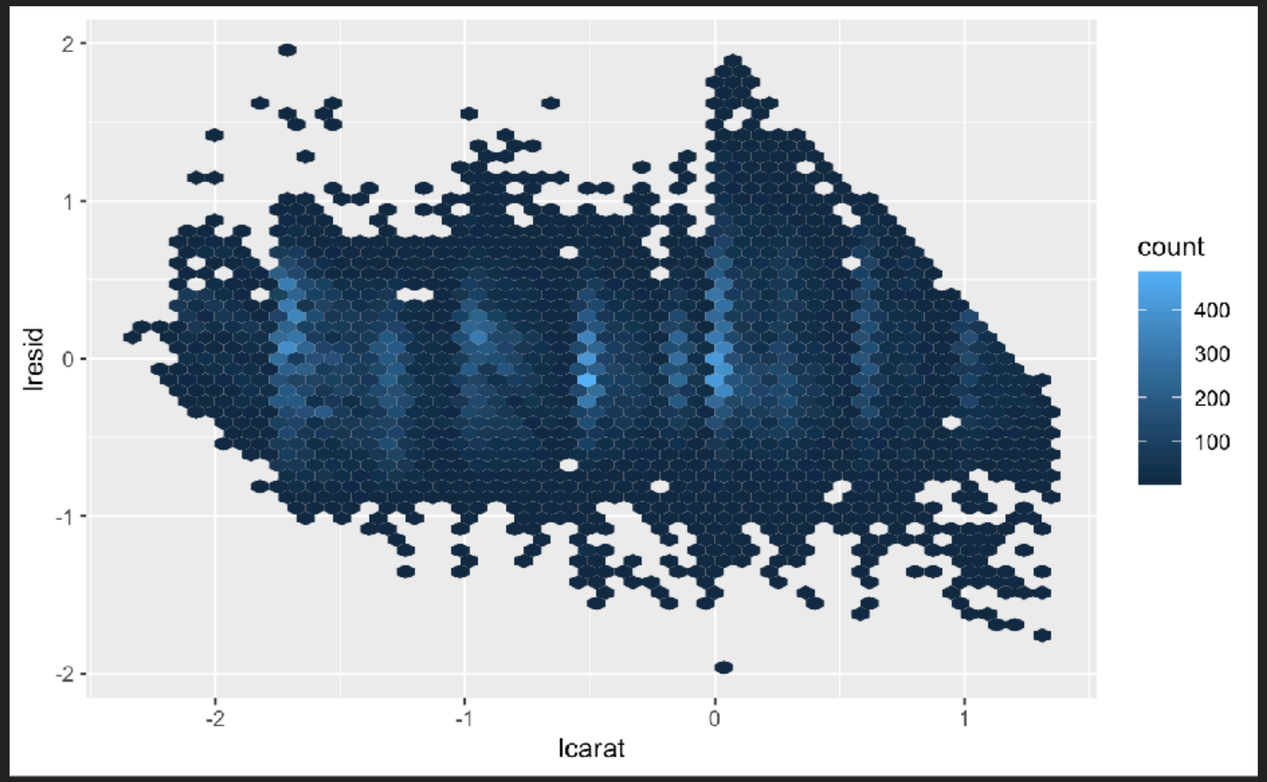




- The weight of the diamond is the single most important factor for determining the price of the diamond.
  - Left: raw
  - Right: log-transformed

Let's remove that strong linear pattern

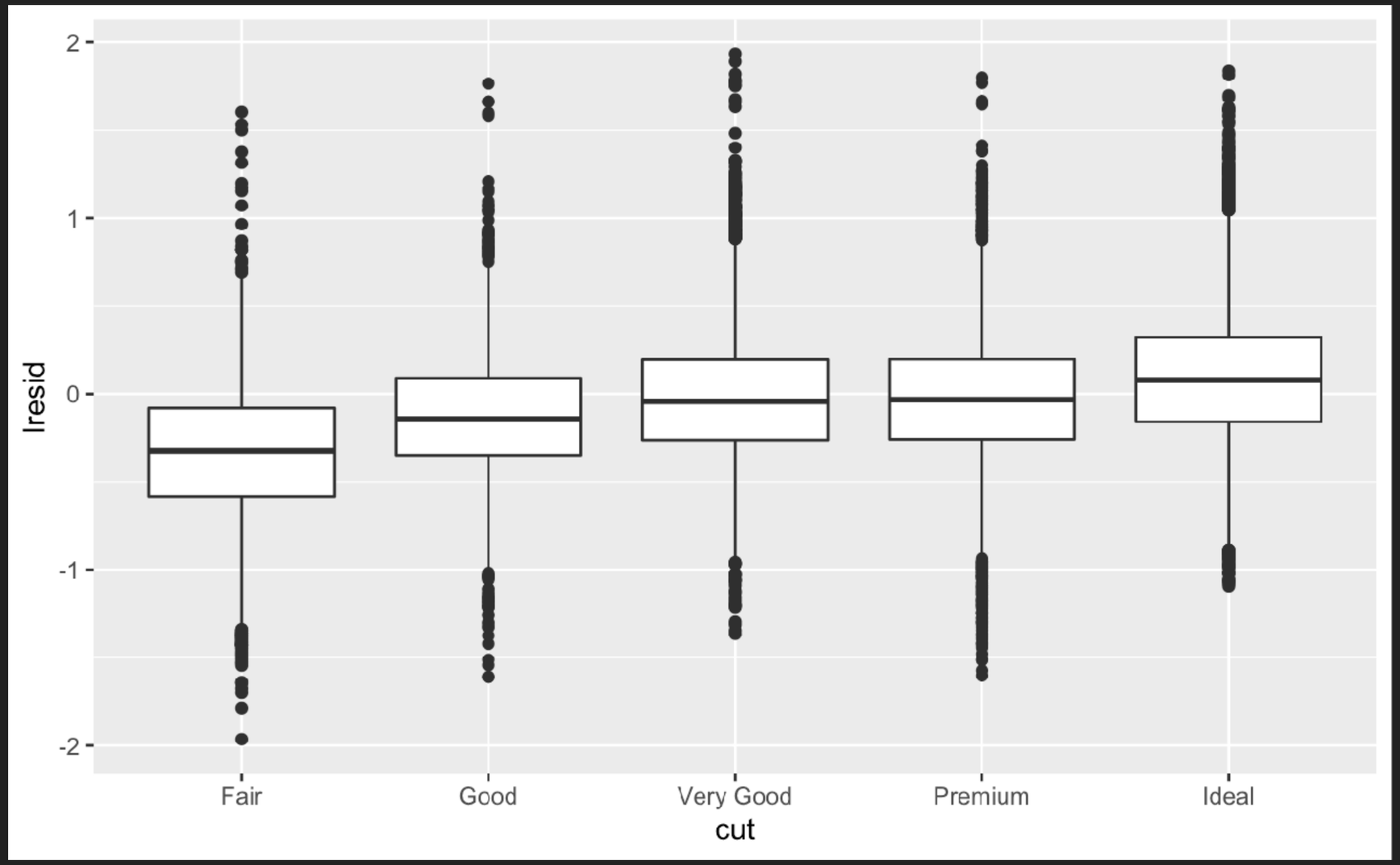





```
mod_diamond <- lm(lprice ~ lcarat, data = diamonds2)
```

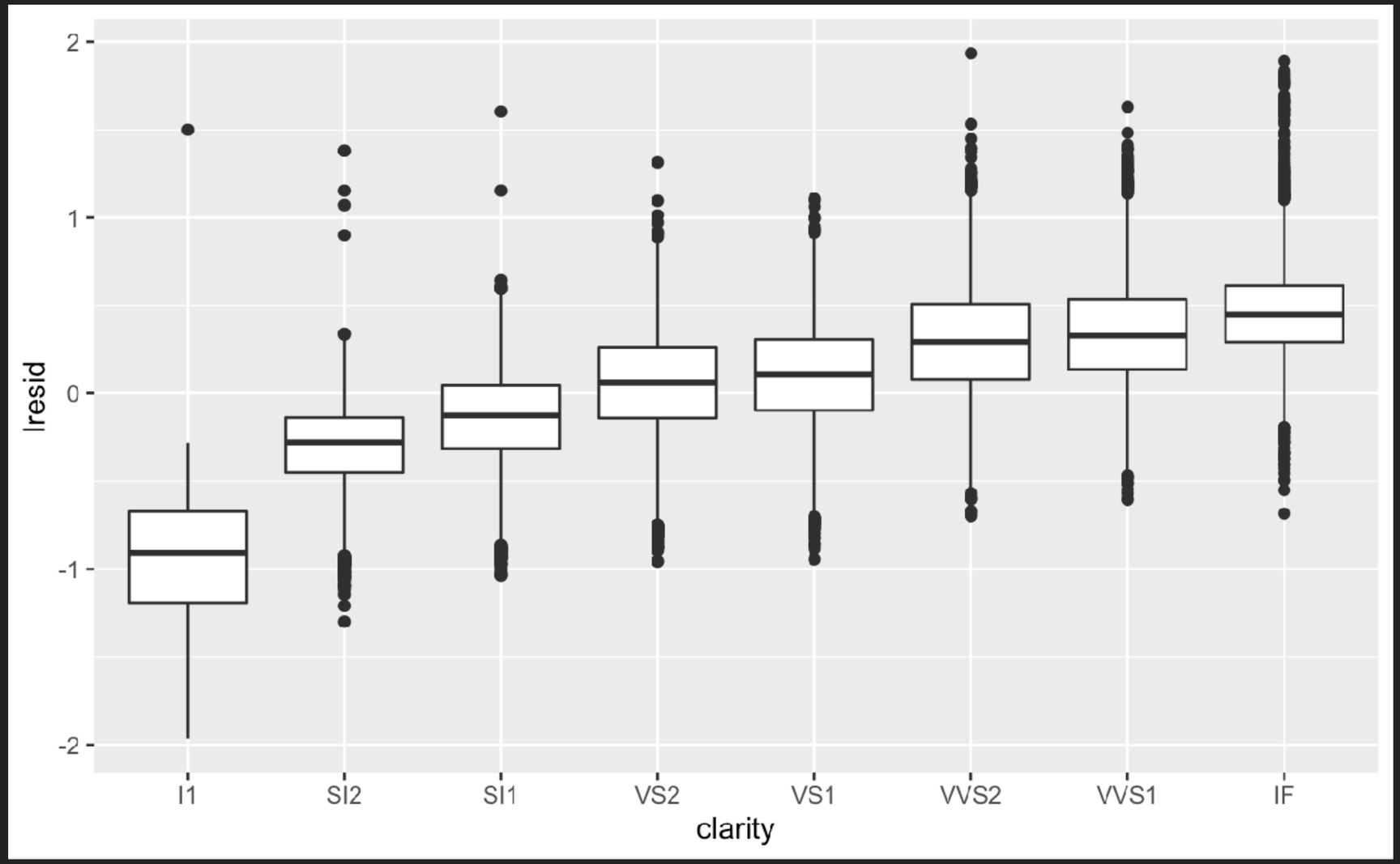
- Residuals confirm that we've successfully removed the strong linear pattern.

Now we see the relationship we expect



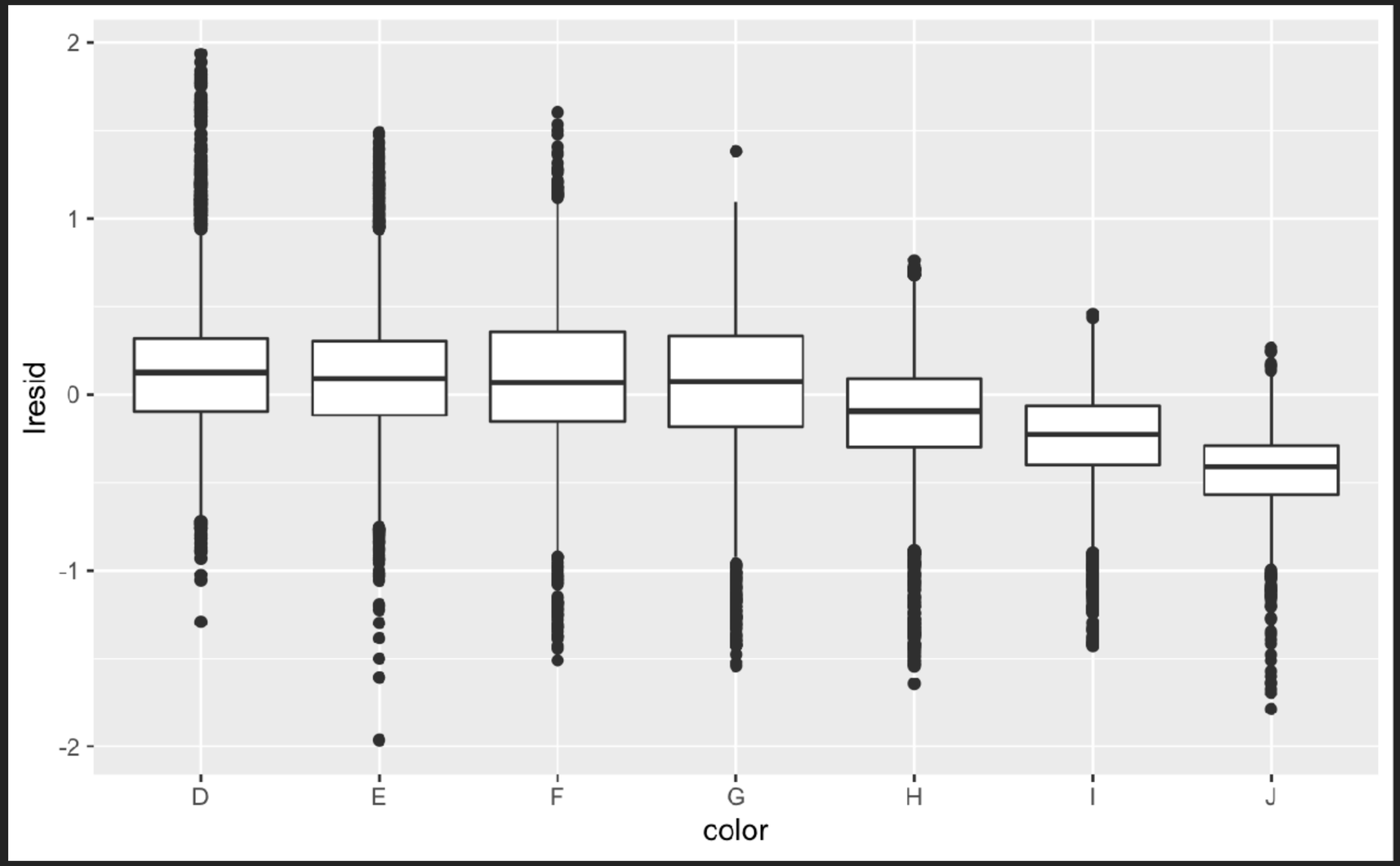
- Re-did our motivating plots using those residuals instead of price.
- Fair: worst cut

Now we see the relationship we expect



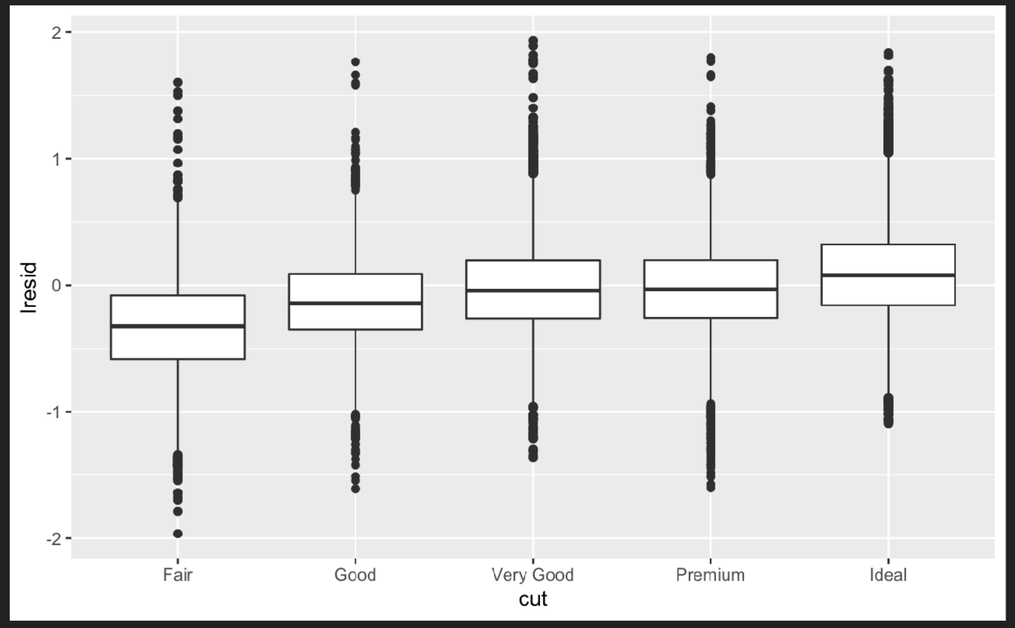
- I1: inclusions visible to the naked eye (worst clarity)

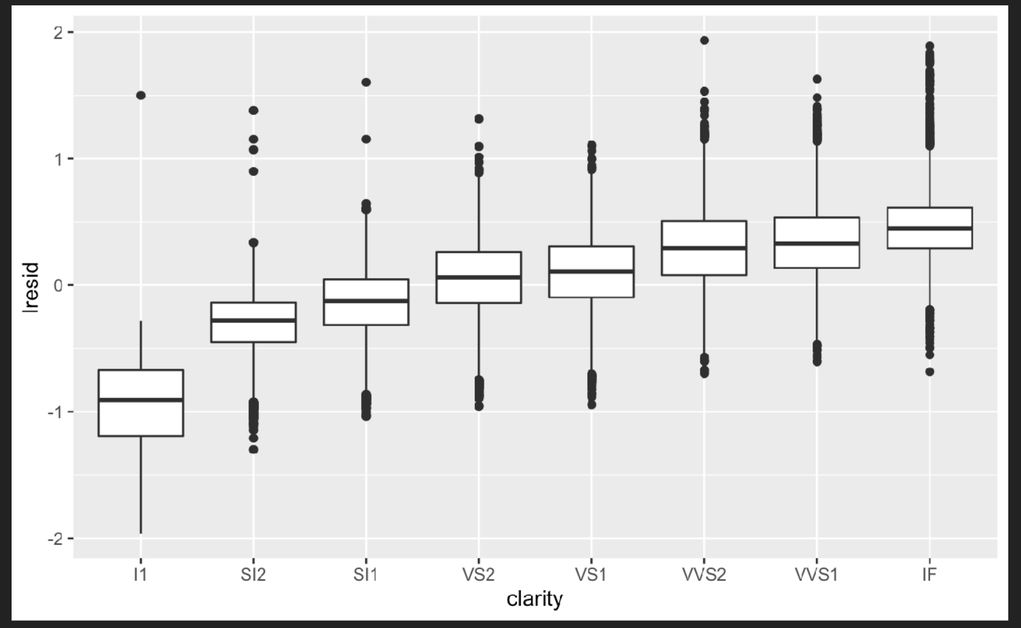
Now we see the relationship we expect

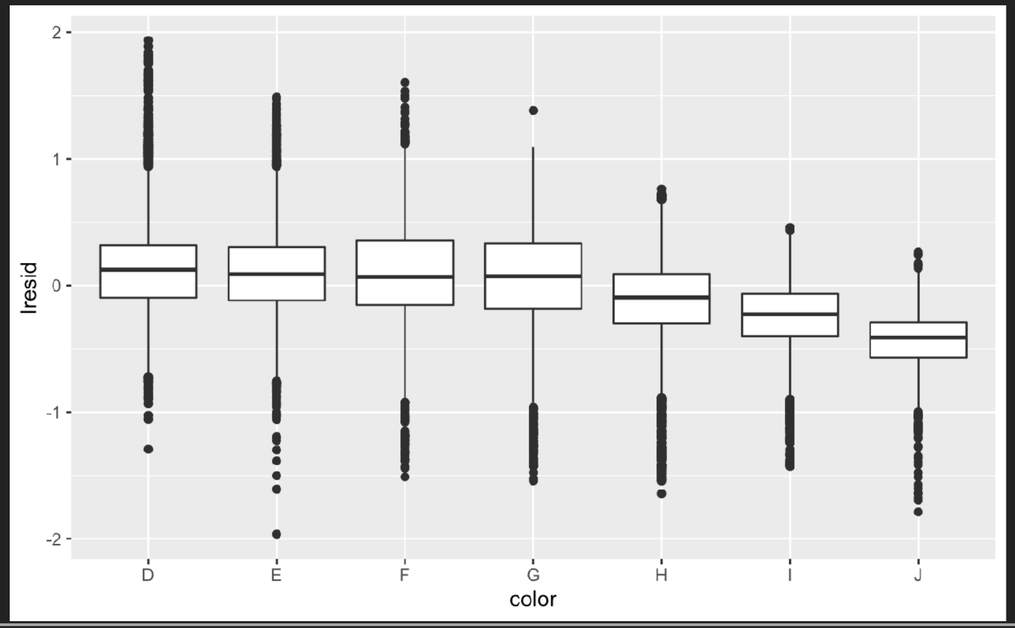


- J: slightly yellow (worst color)

Now we see the relationship we expect



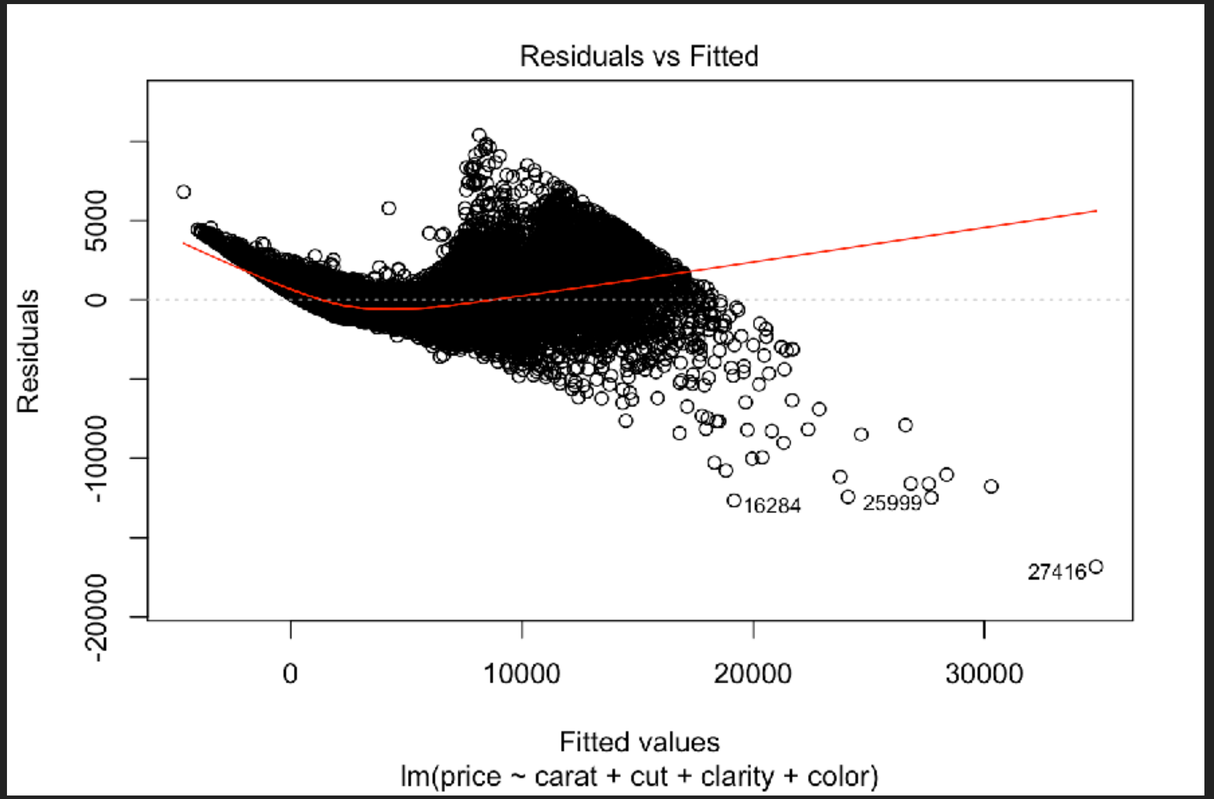




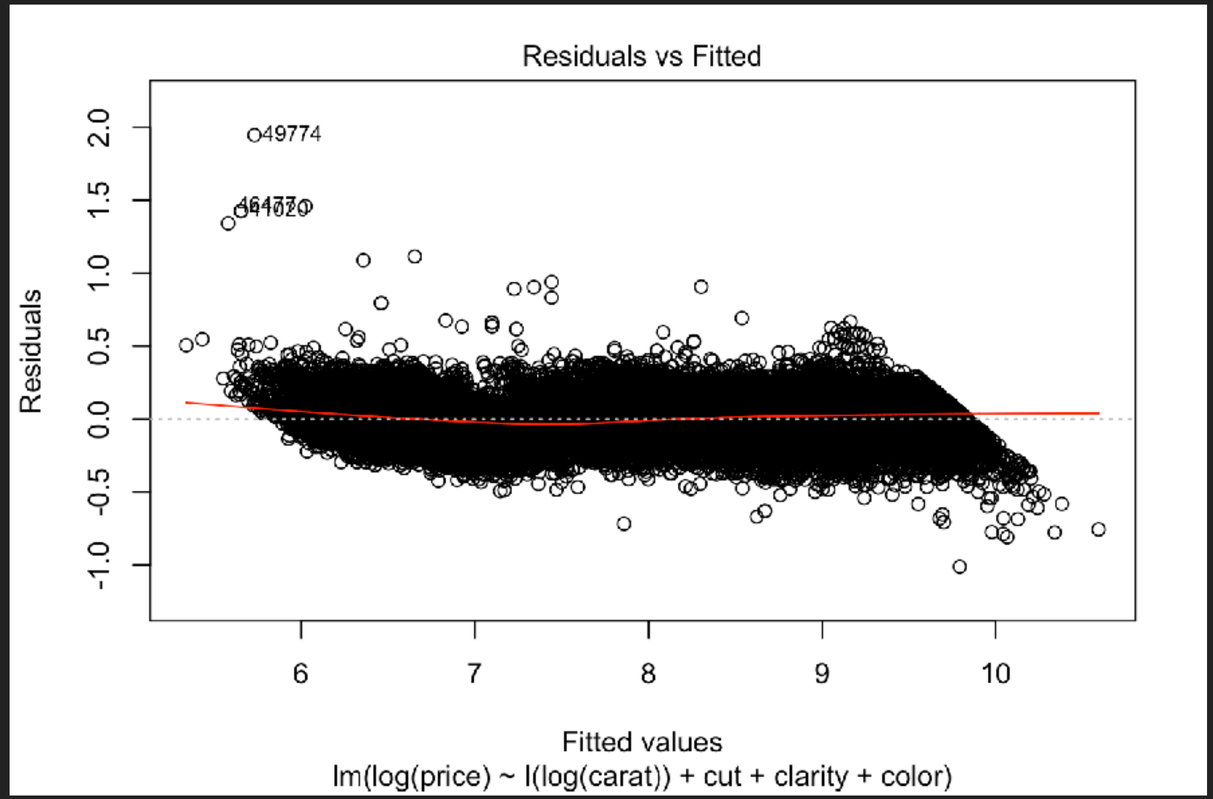
- Re-did our motivating plots using those residuals instead of price.

Regression Diagnostics on the Diamonds Example

```
diamonds.lm <- lm(price ~ carat
                        + cut
                        + clarity
                        + color,
                        data = diamonds)
```

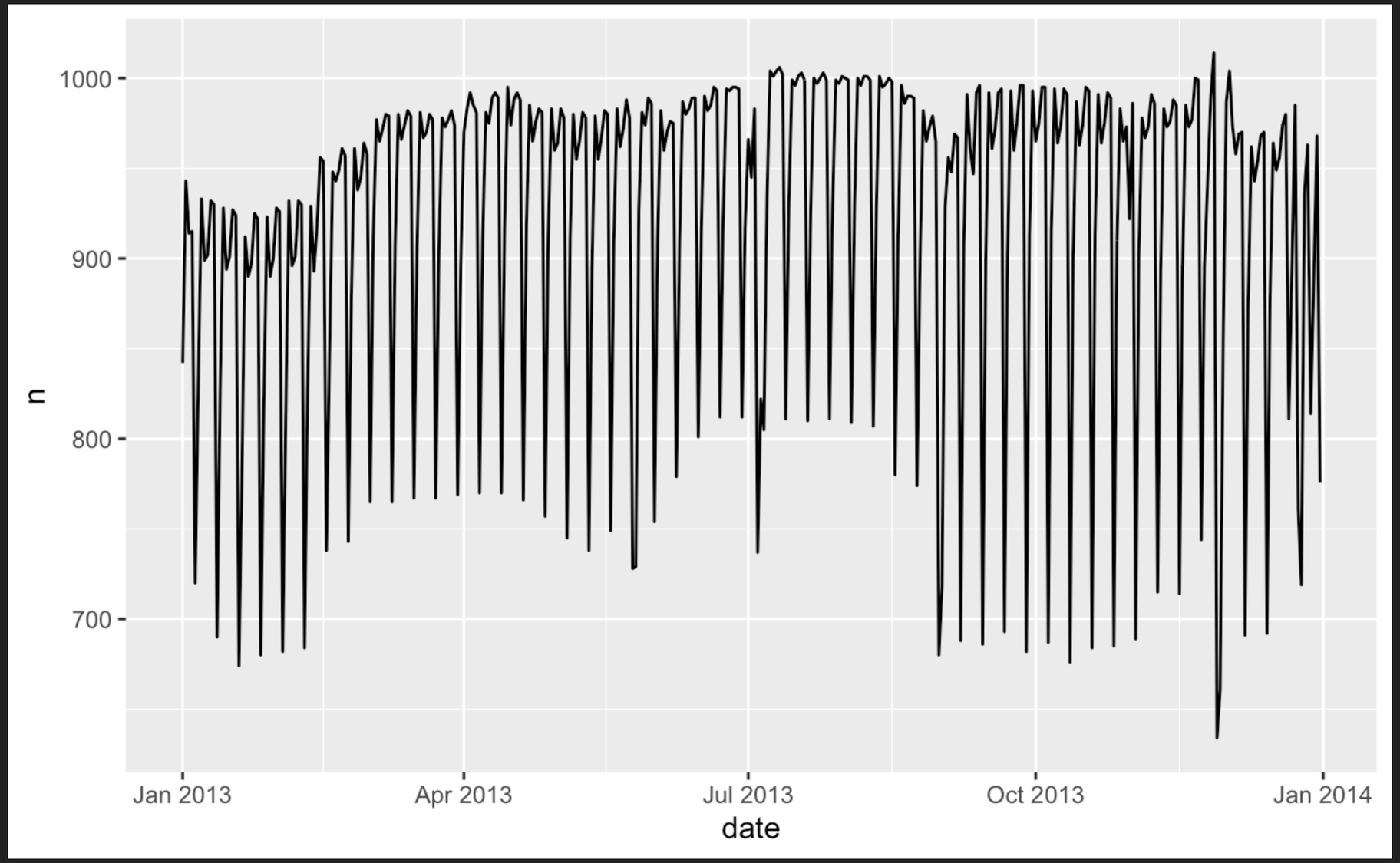



```
diamonds.lm2 <- lm(log(price) ~ I(log(carat))
                              + cut
                              + clarity
                              + color,
                              data = diamonds)
```

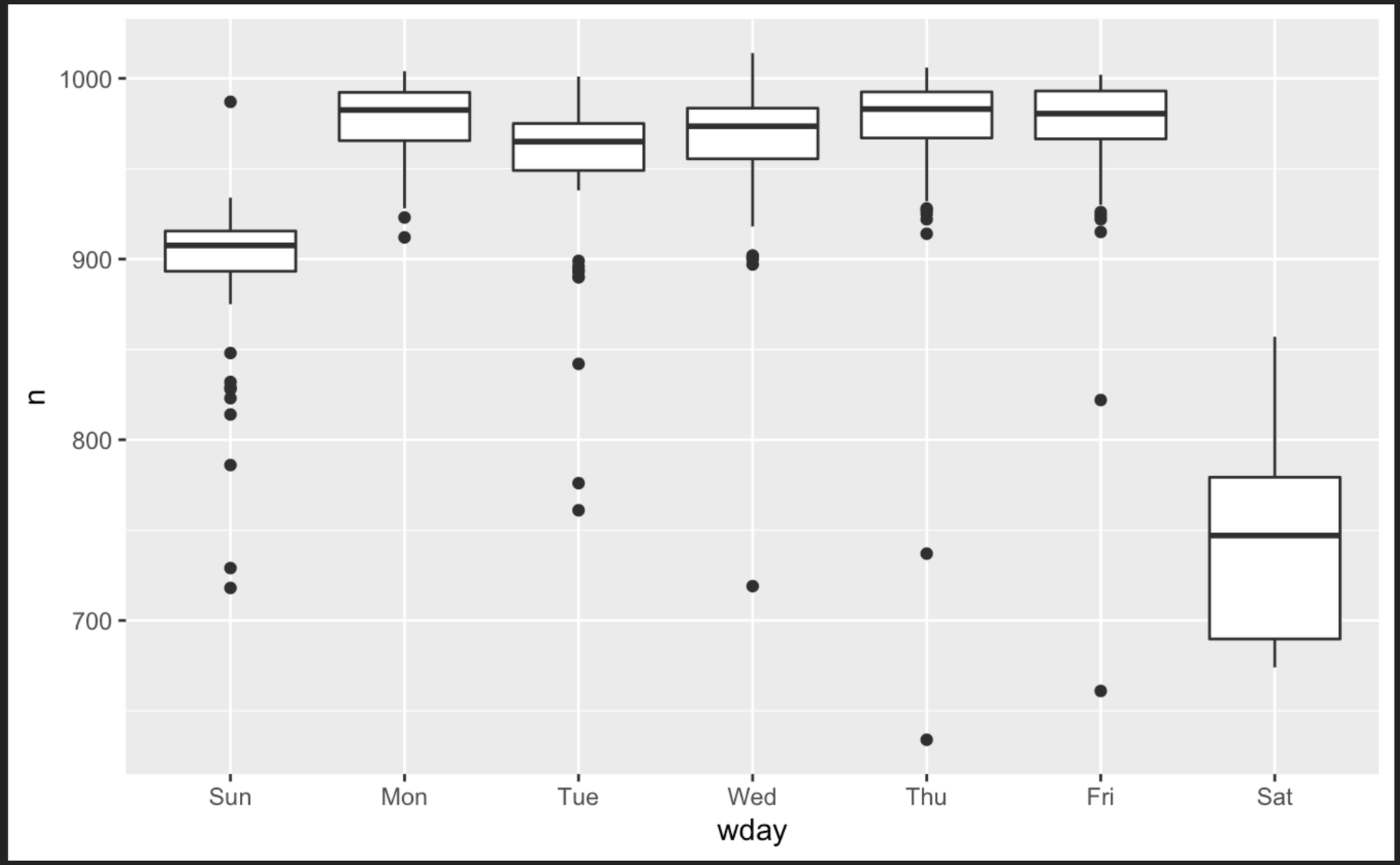


Another example

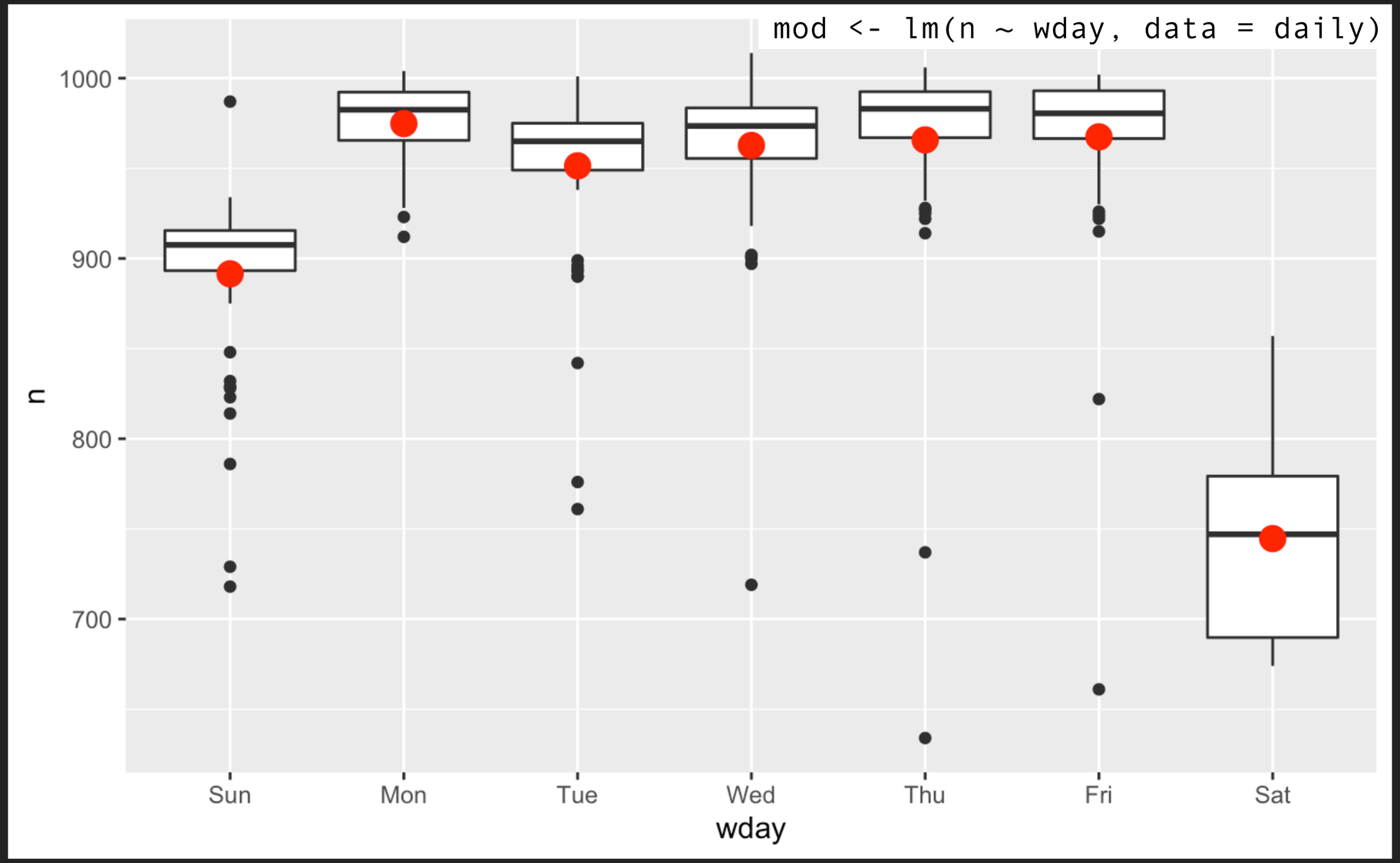
The number of flights that leave NYC per day



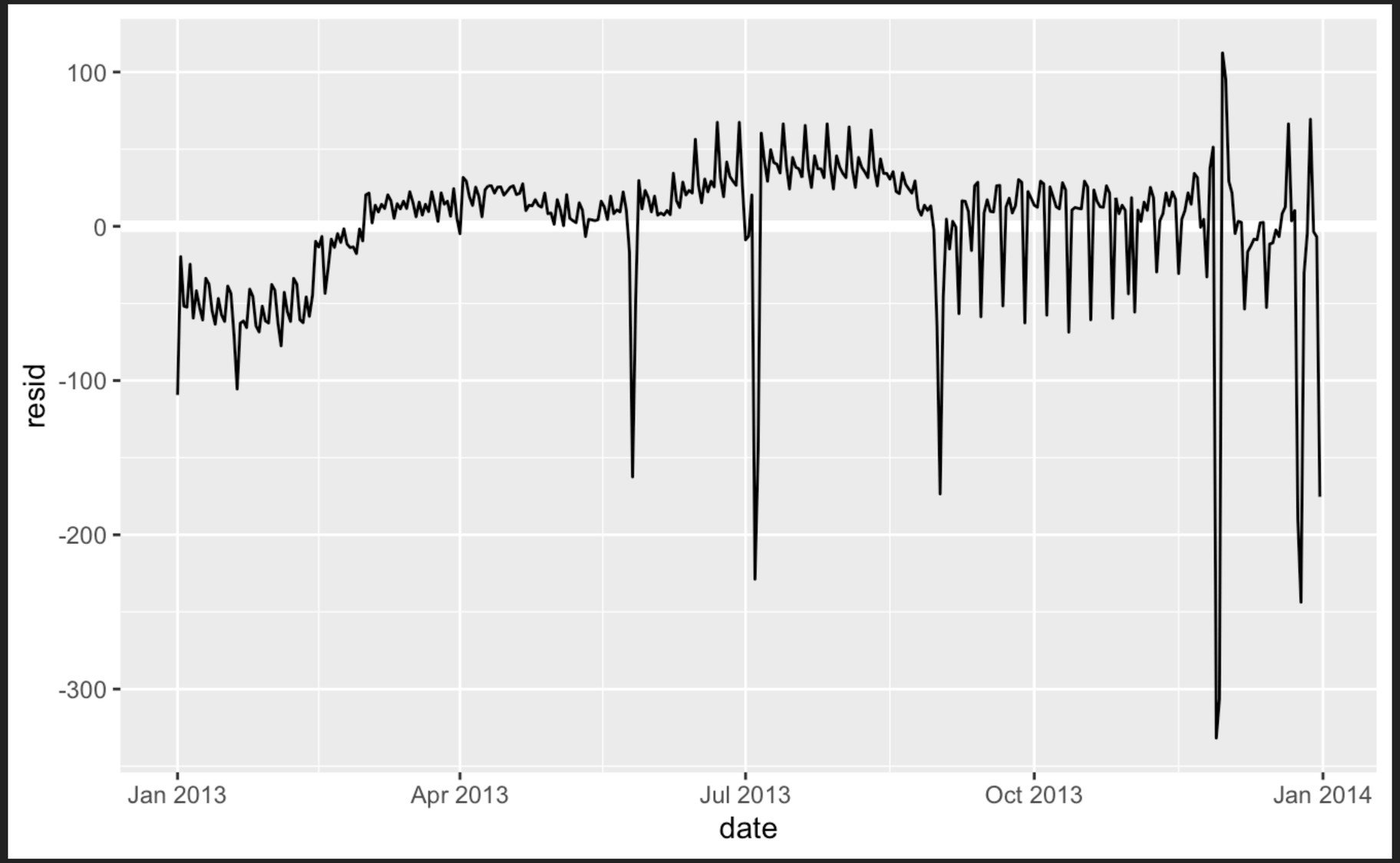
A very strong day-of-week effect dominates the subtler patterns



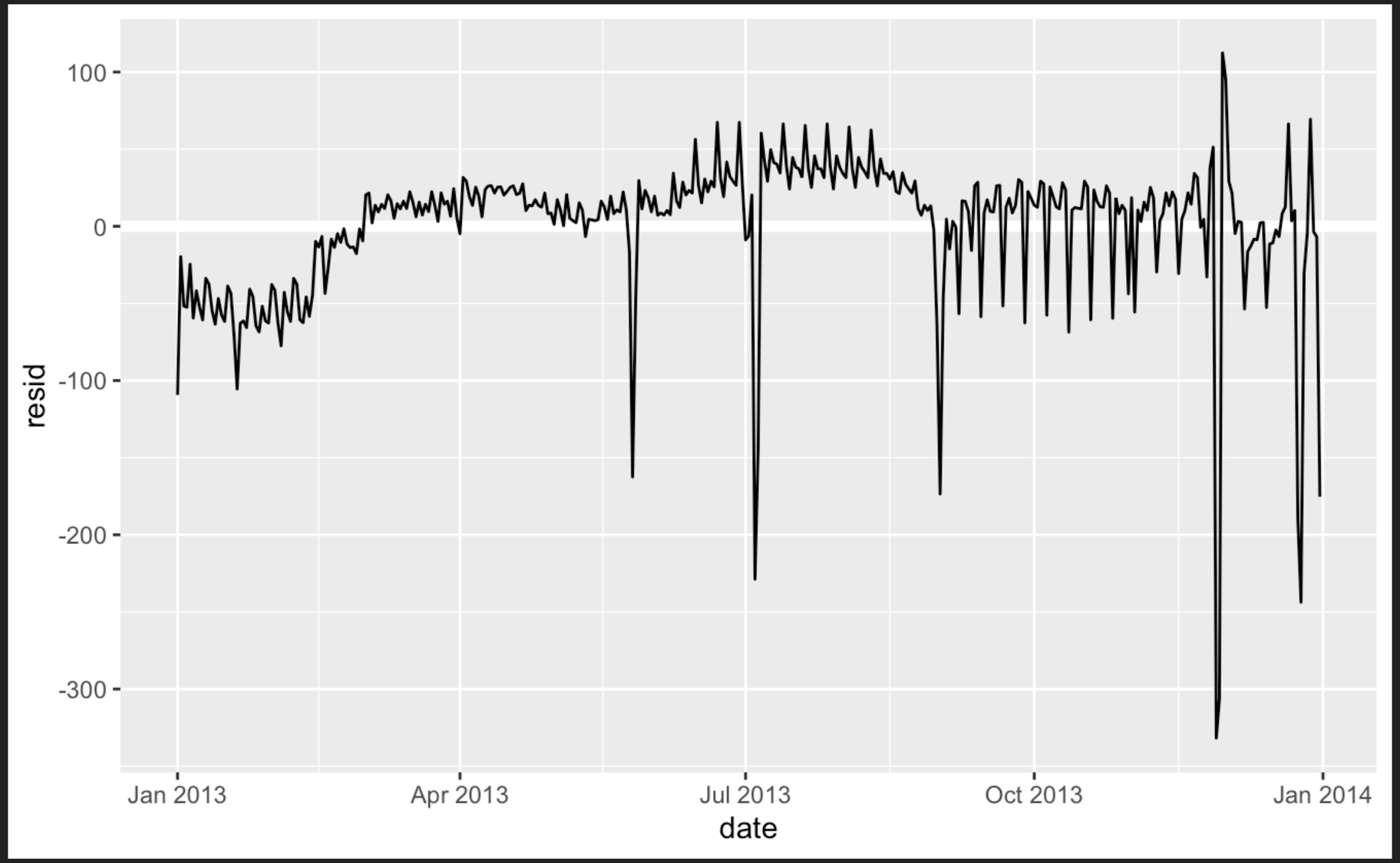
Modeling the week day effect



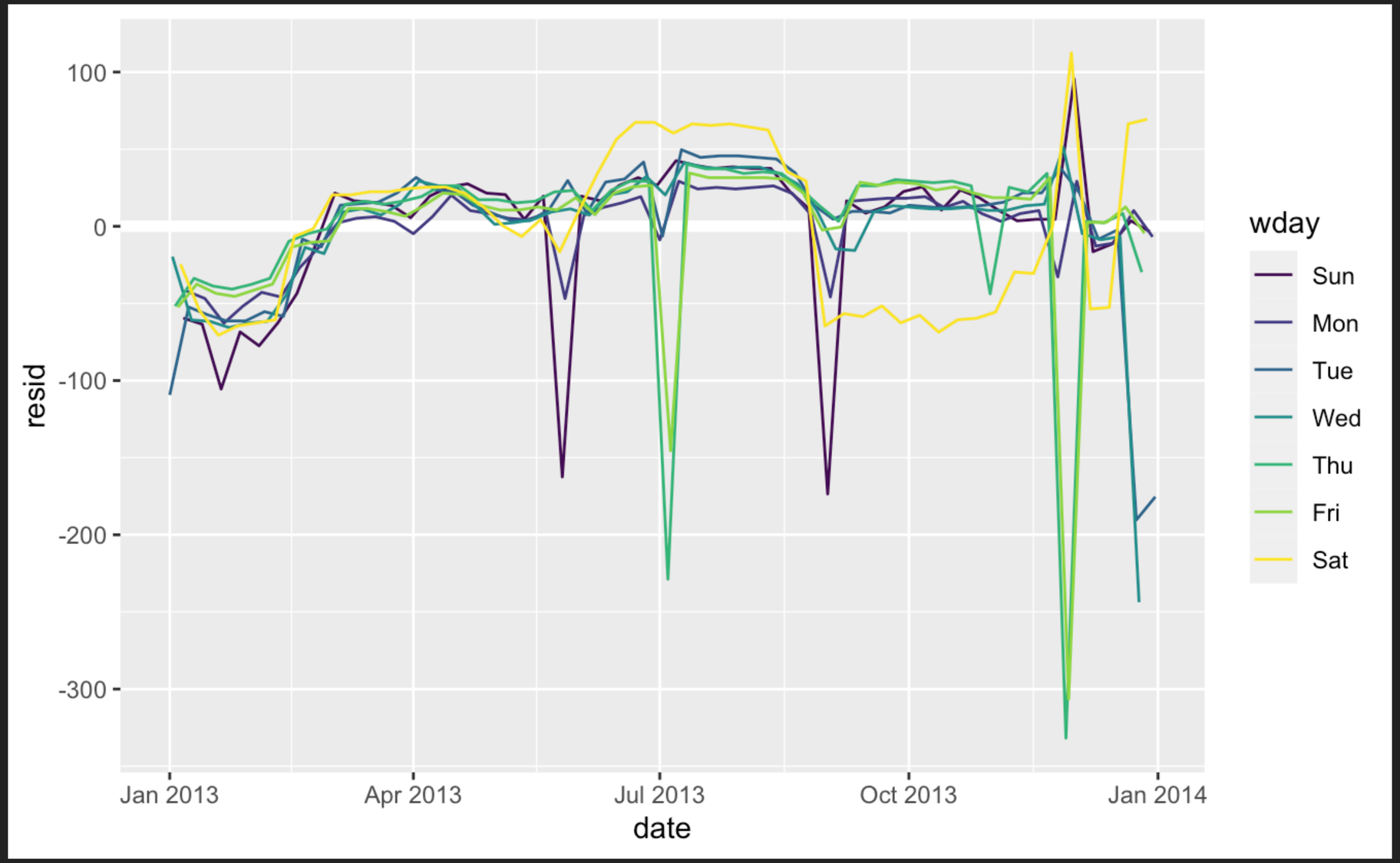
Visualizing the residuals



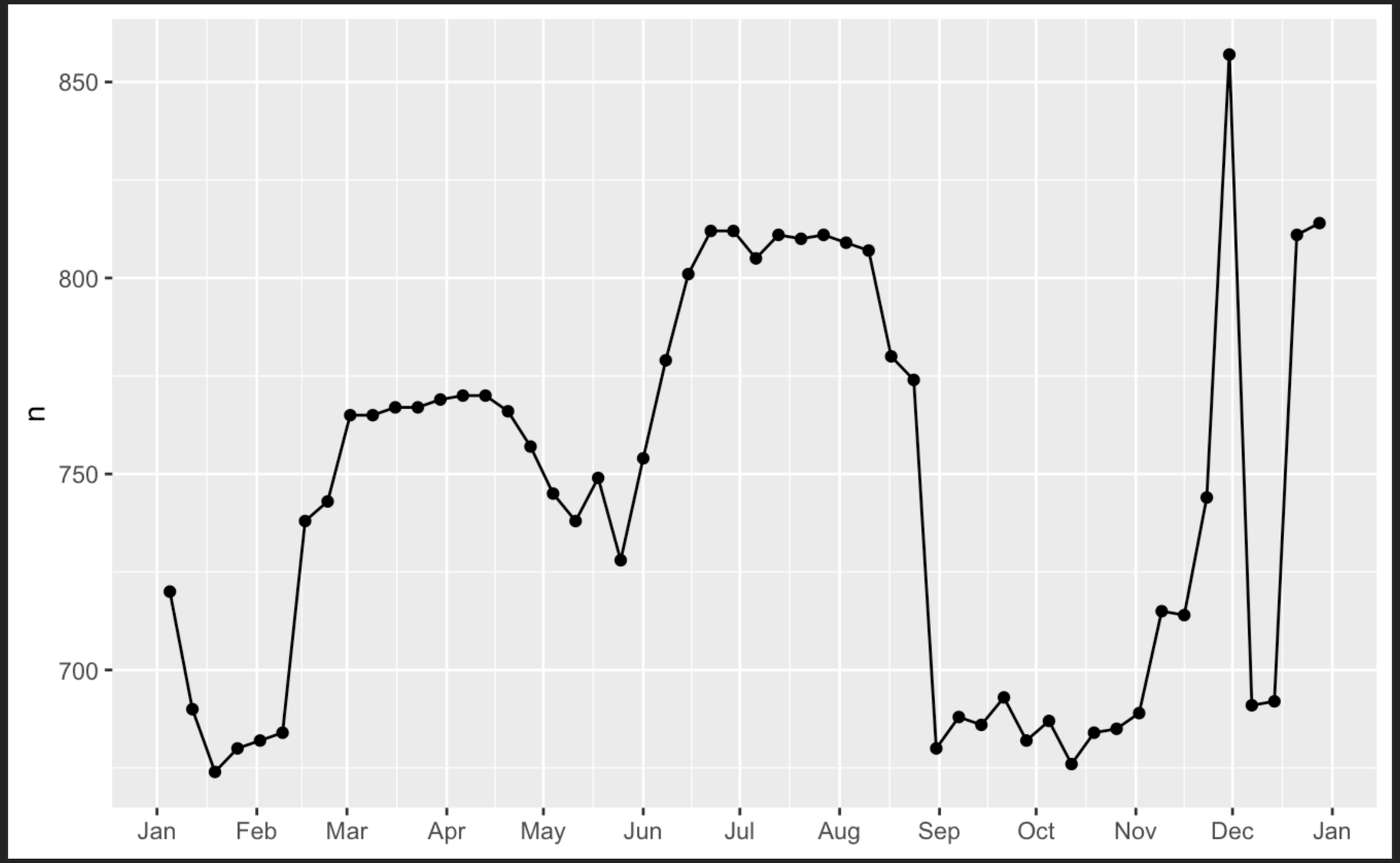
Our model seems to fail starting in June



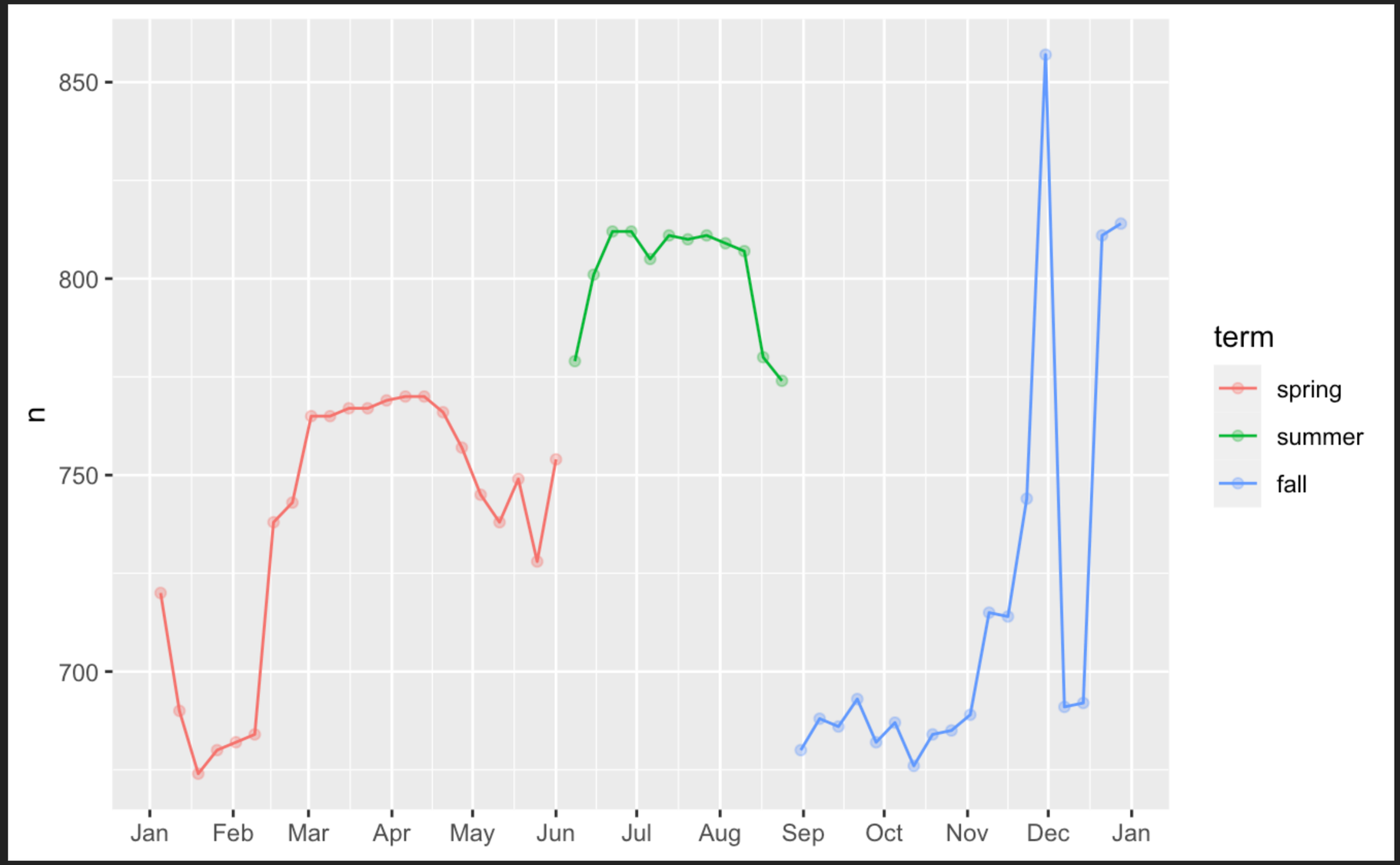
. . . especially when looking at weekend days



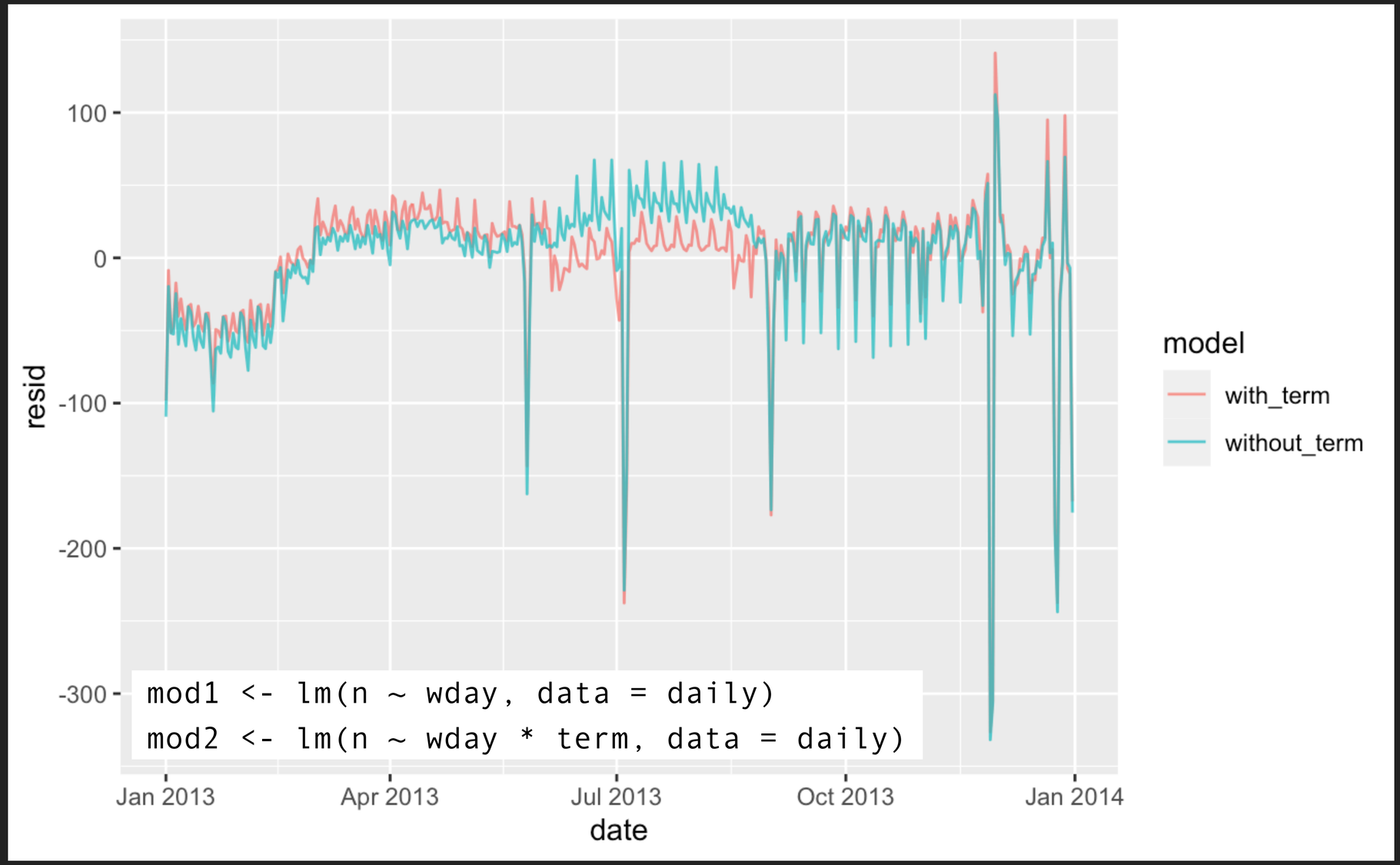
The model fails to accurately predict the number of flights on Saturday



Let's create a "term" variable that roughly captures the three school terms



Fitting a separate day of week effect for each term improves our model



Credits

- Graphics: Dave DiCello photography (cover)
- Bruce, P., Bruce, A., & Gedeck, P. (2020). Practical Statistics for Data Scientists: 50+ Essential Concepts Using R and Python. O'Reilly Media.
- Goodman, S. (2008). A dirty dozen: Twelve p-value misconceptions. In Seminars in Hematology (Vol. 45, No. 3, pp. 135-140). WB Saunders.
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning (Vol. 112, p. 18). New York: springer.
- Grolemund, G., & Wickham, H. (2018). R for data science.