Photo credit: Dave DiCello

17-803 Empirical Methods

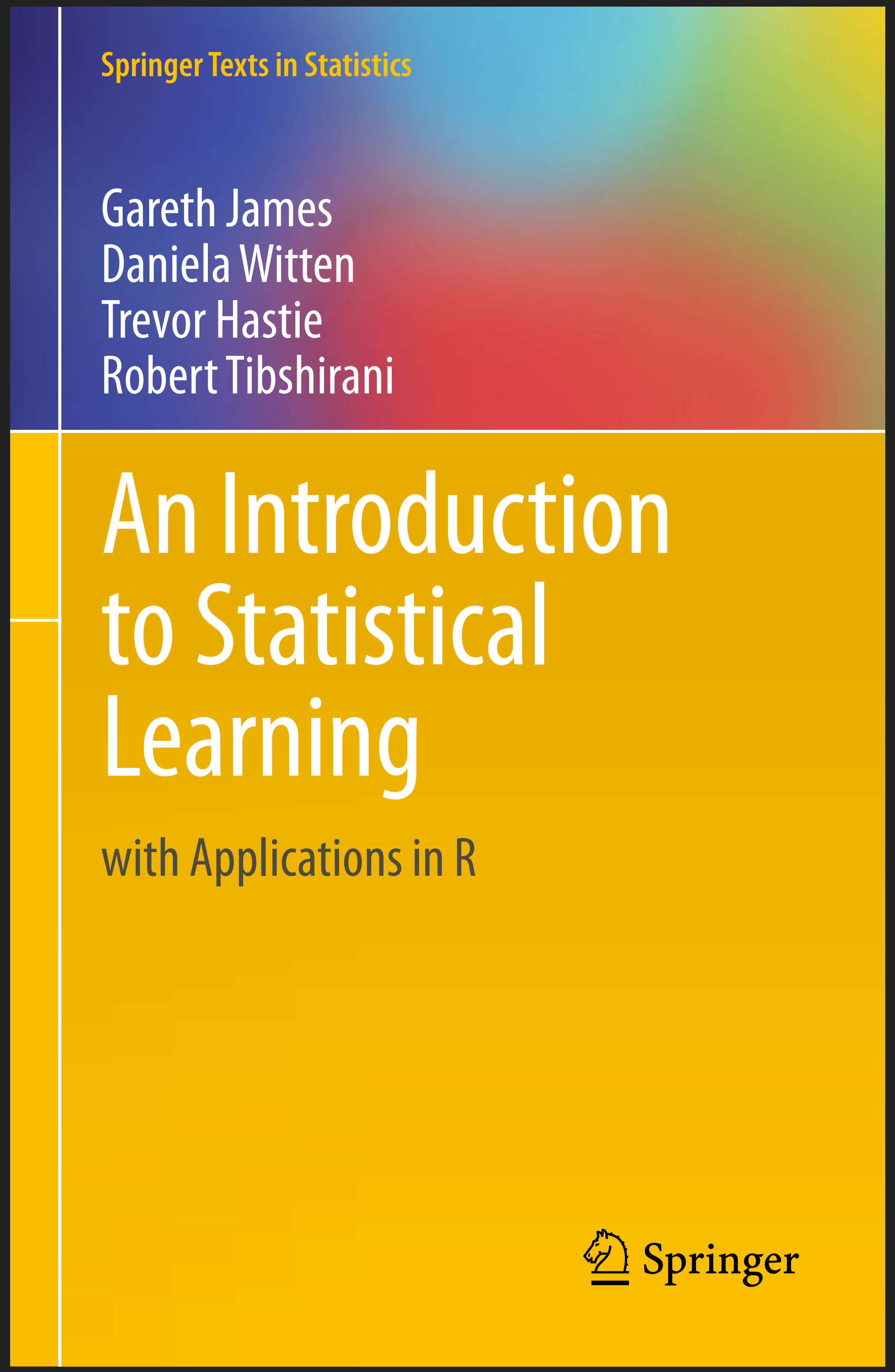
Bogdan Vasilescu, Institute for Software Research

Regression Modeling (Part 2)

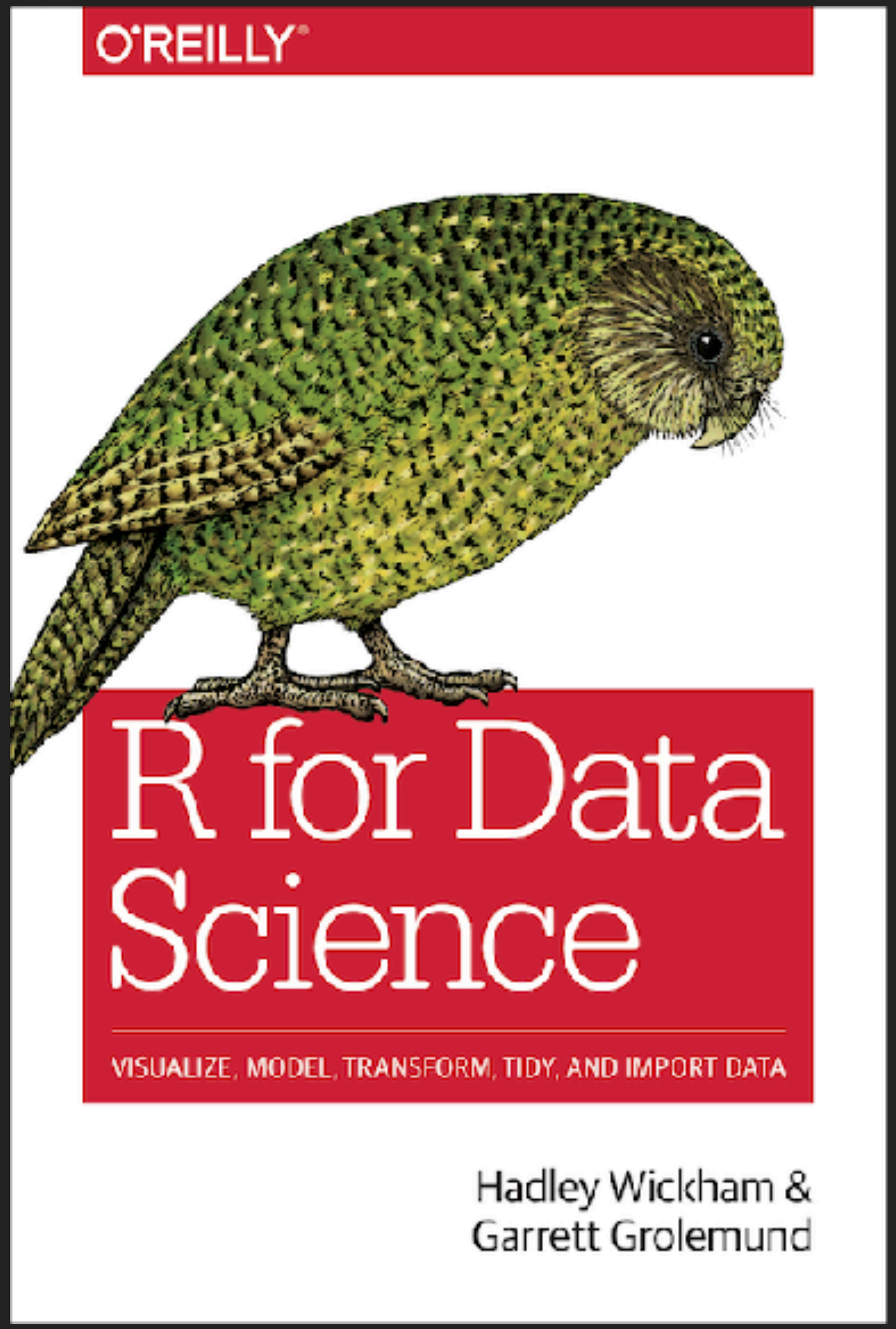
Thursday, March 25, 2021

Outline for Today

- More linear regression



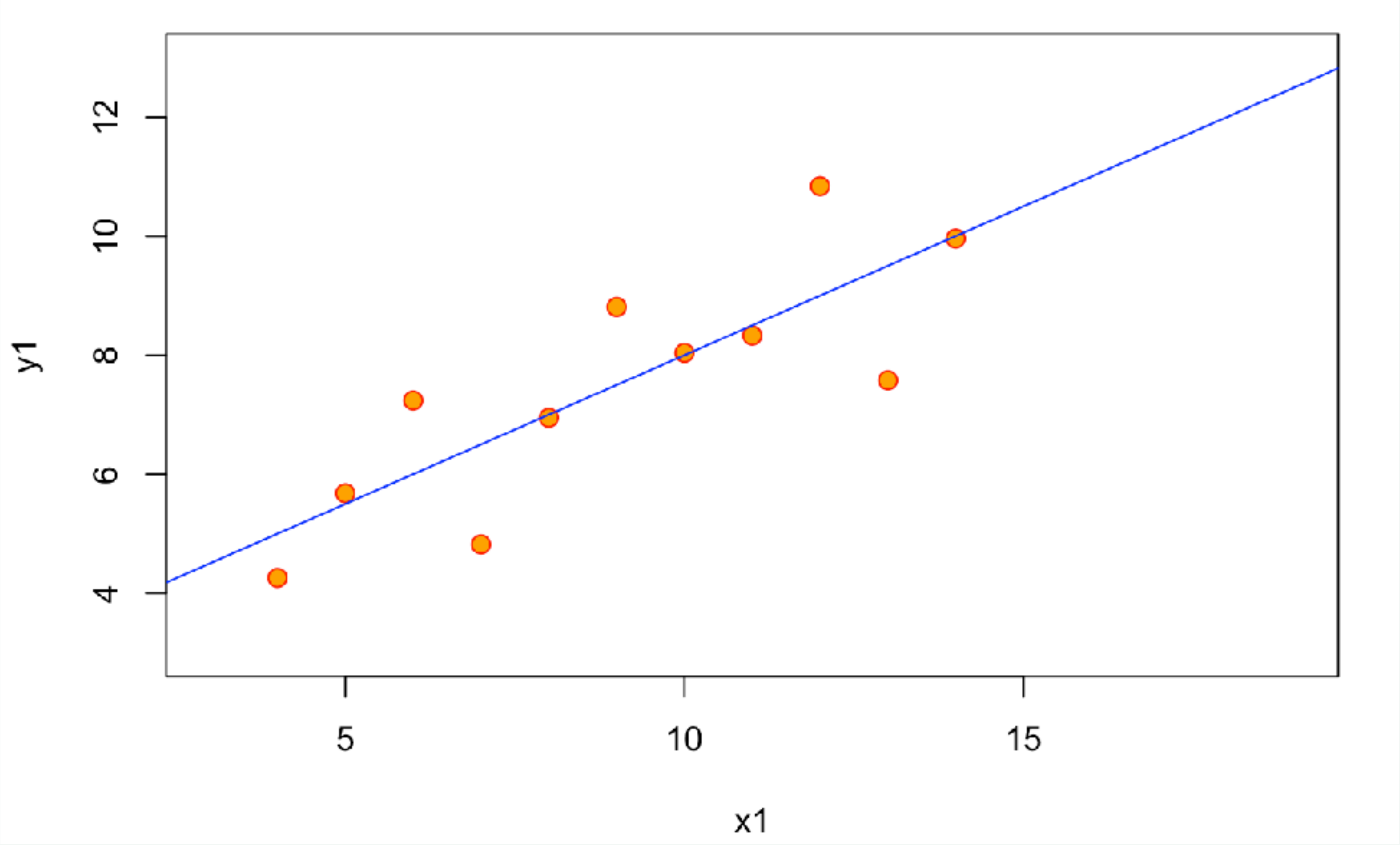
Ch 3 (Linear regression)



Ch 22-24 (Modeling)

- Remember:
  - https://bvasiles.github.io/empirical-methods/

Simple Linear Regression Example




```
##
## Call:
## lm(formula = y1 ~ x1, data = anscombe)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.92127 -0.45577 -0.04136  0.70941  1.83882
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.0001     1.1247   2.667  0.02573 *
## x1            0.5001     0.1179   4.241  0.00217 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.237 on 9 degrees of freedom
## Multiple R-squared:  0.6665, Adjusted R-squared:  0.6295
## F-statistic: 17.99 on 1 and 9 DF,  p-value: 0.00217
```

Let's make it more realistic

How To Extend our Analysis To Accommodate all Predictors?

- One option is to run three separate simple linear regressions.

	Coefficient	Std. error	t-statistic	p-value
`Intercept`	7.0325	0.4578	15.36	< 0.0001
`TV`	0.0475	0.0027	17.67	< 0.0001

	Coefficient	Std. error	t-statistic	p-value
`Intercept`	9.312	0.563	16.54	< 0.0001
`radio`	0.203	0.020	9.92	< 0.0001

	Coefficient	Std. error	t-statistic	p-value
`Intercept`	12.351	0.621	19.88	< 0.0001
`newspaper`	0.055	0.017	3.30	0.00115

How To Extend our Analysis To Accommodate all Predictors?

- A better option is to give each predictor a separate slope coefficient in a single model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon,$$

$$\texttt{sales} = \beta_0 + \beta_1 \times \texttt{TV} + \beta_2 \times \texttt{radio} + \beta_3 \times \texttt{newspaper} + \epsilon.$$

- We interpret βj as the average effect on Y of a one unit increase in Xj, *holding all other predictors fixed*.

Aside: Ingredients for Establishing a Causal Relationship

The cause preceded the effect

The cause was related to the effect

We can find no plausible alternative explanation for the effect other than the cause

Back to our Advertising Example

	Coefficient	Std. error	t-statistic	p-value
Intercept	2.939	0.3119	9.42	< 0.0001
TV	0.046	0.0014	32.81	< 0.0001
radio	0.189	0.0086	21.89	< 0.0001
newspaper	−0.001	0.0059	−0.18	0.8599

	Coefficient	Std. error	t-statistic	p-value
Intercept	7.0325	0.4578	15.36	< 0.0001
TV	0.0475	0.0027	17.67	< 0.0001

	Coefficient	Std. error	t-statistic	p-value
Intercept	9.312	0.563	16.54	< 0.0001
radio	0.203	0.020	9.92	< 0.0001

	Coefficient	Std. error	t-statistic	p-value
Intercept	12.351	0.621	19.88	< 0.0001
newspaper	0.055	0.017	3.30	0.00115

Interaction Effects

- Consider the standard linear regression model with two variables

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon.$$

  - According to this model, if we increase X1 by one unit, then Y will increase by an average of β1 units

Interaction Effects

- Extending this model with an interaction term gives:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \epsilon.$$

Interaction Effects

- Extending this model with an interaction term gives:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \epsilon.$$

$$\begin{aligned} &= \beta_0 + (\beta_1 + \beta_3 X_2) X_1 + \beta_2 X_2 + \epsilon \\ &= \beta_0 + \tilde{\beta}_1 X_1 + \beta_2 X_2 + \epsilon \end{aligned}$$

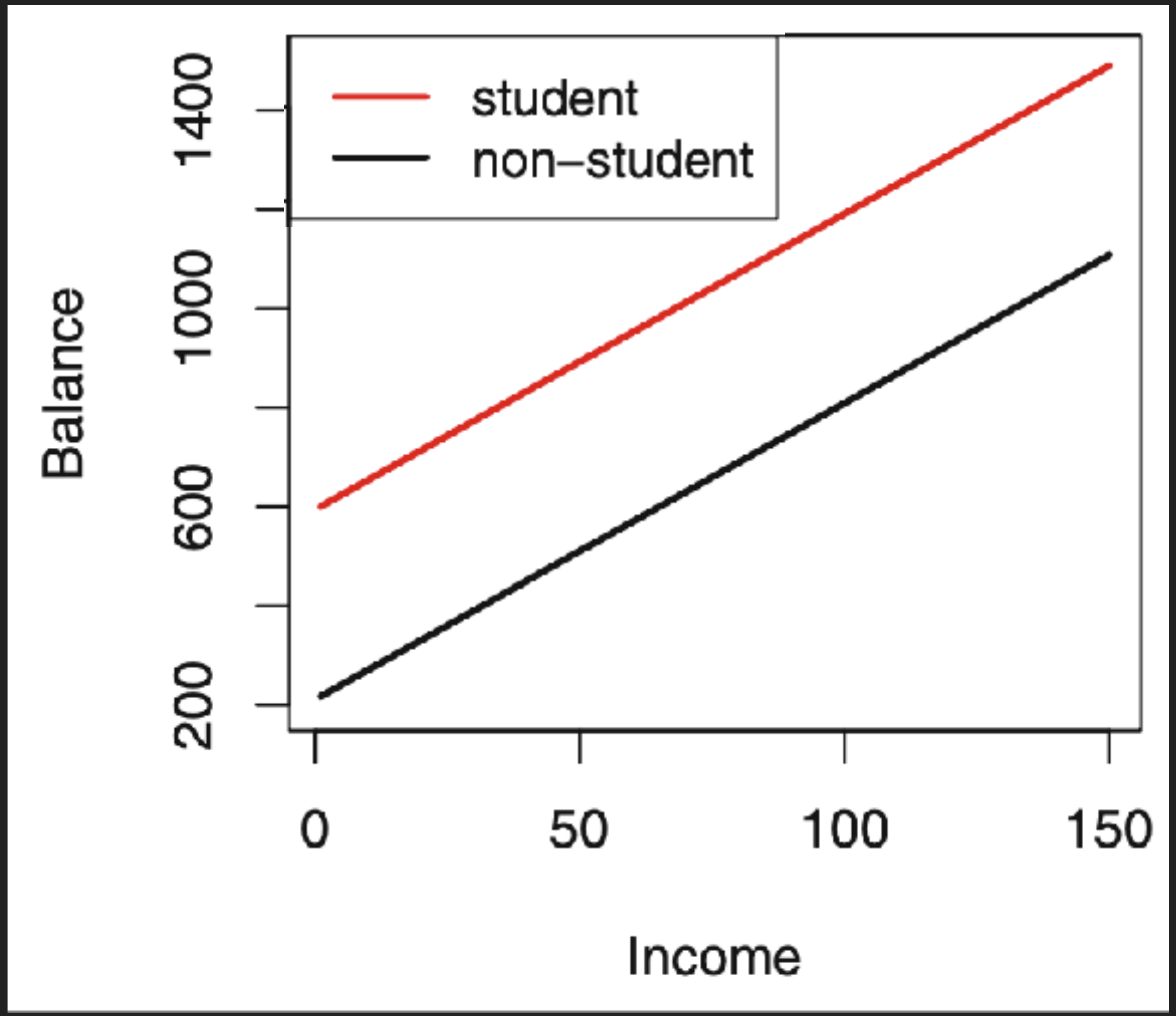
  - According to this model, adjusting X2 will change the impact of X1 on Y

Example: Model Credit Card Balance Using Income (Numerical) and Student (Categorical)

$$
\begin{aligned}
\texttt{balance}_i &\approx \beta_0 + \beta_1 \times \texttt{income}_i + \begin{cases} \beta_2 & \text{if } i\text{th person is a student} \\ 0 & \text{if } i\text{th person is not a student} \end{cases} \\
&= \beta_1 \times \texttt{income}_i + \begin{cases} \beta_0 + \beta_2 & \text{if } i\text{th person is a student} \\ \beta_0 & \text{if } i\text{th person is not a student.} \end{cases}
\end{aligned}
$$

Example: Model Credit Card Balance Using Income (Numerical) and Student (Categorical)

$$
\begin{aligned}
\texttt{balance}_i &\approx \beta_0 + \beta_1 \times \texttt{income}_i + \begin{cases} \beta_2 & \text{if } i\text{th person is a student} \\ 0 & \text{if } i\text{th person is not a student} \end{cases} \\
&= \beta_1 \times \texttt{income}_i + \begin{cases} \beta_0 + \beta_2 & \text{if } i\text{th person is a student} \\ \beta_0 & \text{if } i\text{th person is not a student.} \end{cases}
\end{aligned}
$$



- Without an interaction term: fitting two parallel lines to the data, one for students and one for non-students.
- The lines for students and non-students have different intercepts, β0 + β2 versus β0, but the same slope, β1.

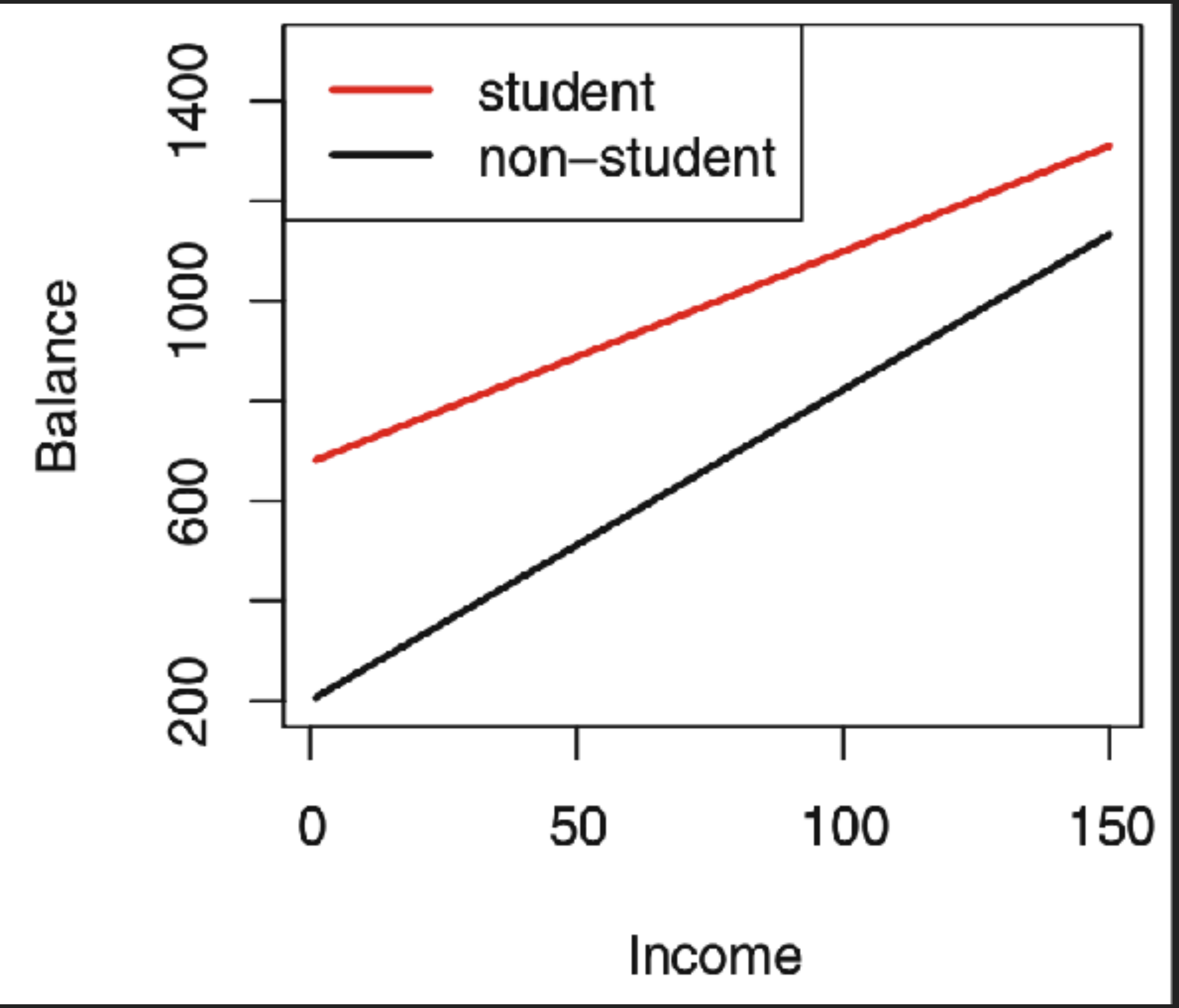
Example: Model Credit Card Balance Using Income (Numerical) and Student (Categorical)

$$\begin{aligned}
\texttt{balance}_i &\approx \beta_0 + \beta_1 \times \texttt{income}_i + \begin{cases} \beta_2 + \beta_3 \times \texttt{income}_i & \text{if student} \\ 0 & \text{if not student} \end{cases} \\
&= \begin{cases} (\beta_0 + \beta_2) + (\beta_1 + \beta_3) \times \texttt{income}_i & \text{if student} \\ \beta_0 + \beta_1 \times \texttt{income}_i & \text{if not student} \end{cases}
\end{aligned}$$

Example: Model Credit Card Balance Using Income (Numerical) and Student (Categorical)

$$\texttt{balance}_i \approx \beta_0 + \beta_1 \times \texttt{income}_i + \begin{cases} \beta_2 + \beta_3 \times \texttt{income}_i & \text{if student} \\ 0 & \text{if not student} \end{cases}$$

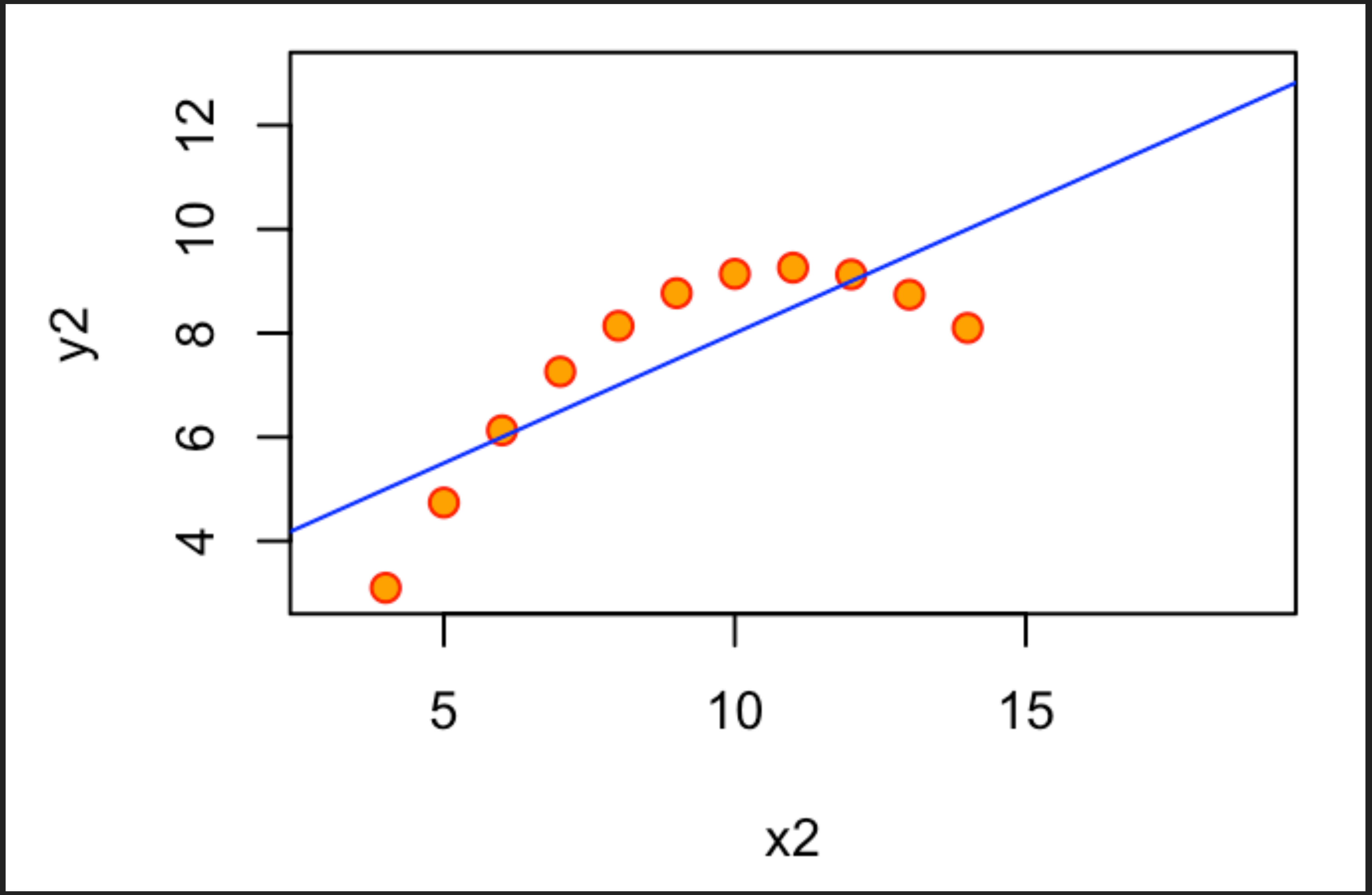
$$= \begin{cases} (\beta_0 + \beta_2) + (\beta_1 + \beta_3) \times \texttt{income}_i & \text{if student} \\ \beta_0 + \beta_1 \times \texttt{income}_i & \text{if not student} \end{cases}$$



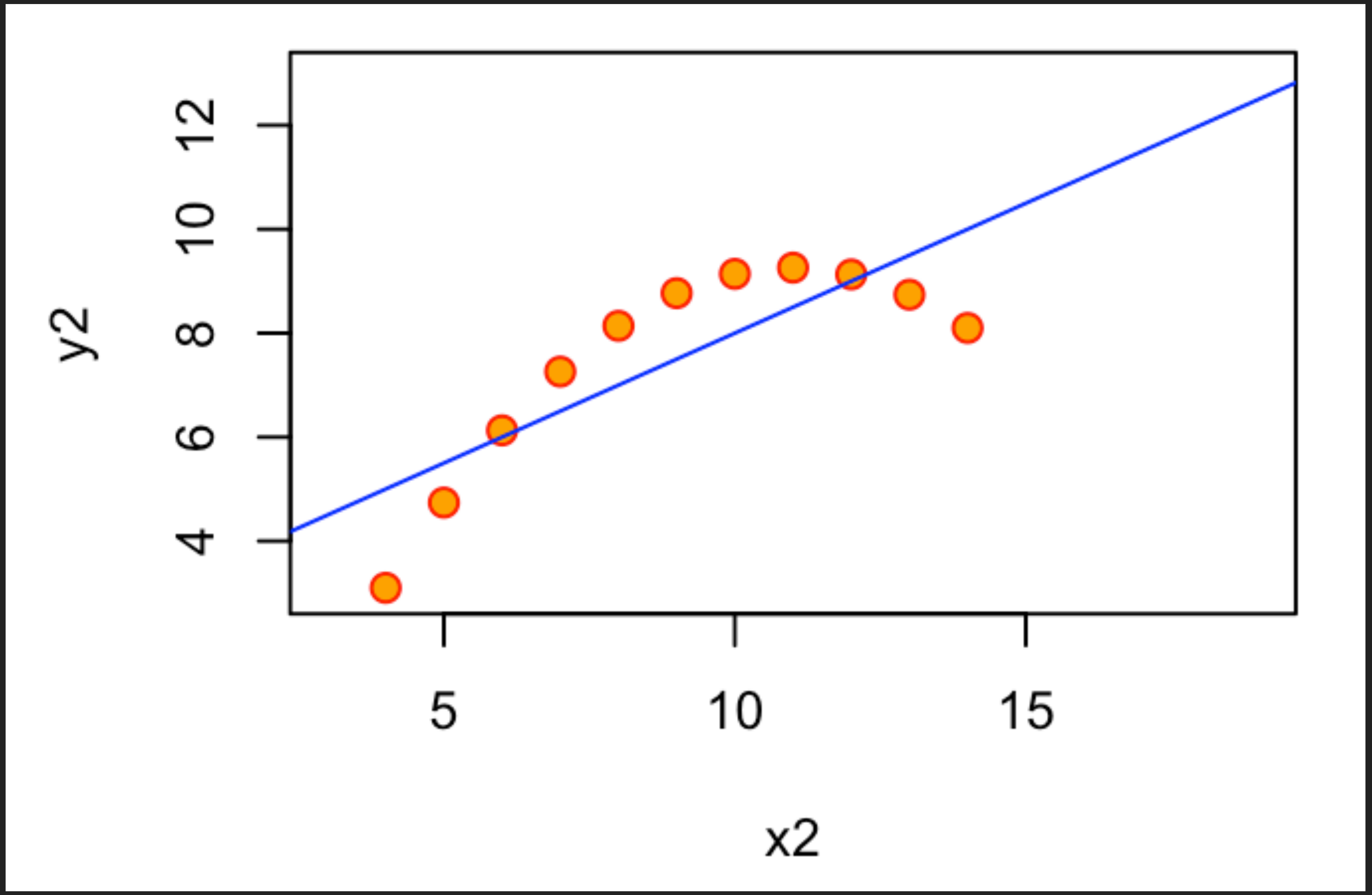
- With an interaction term: the regression lines for the students and the non-students have different intercepts, β0+β2 versus β0, and different slopes, β1+β3 versus β1.
- Allows for the possibility that changes in income may affect the credit card balances of students and non-students differently.

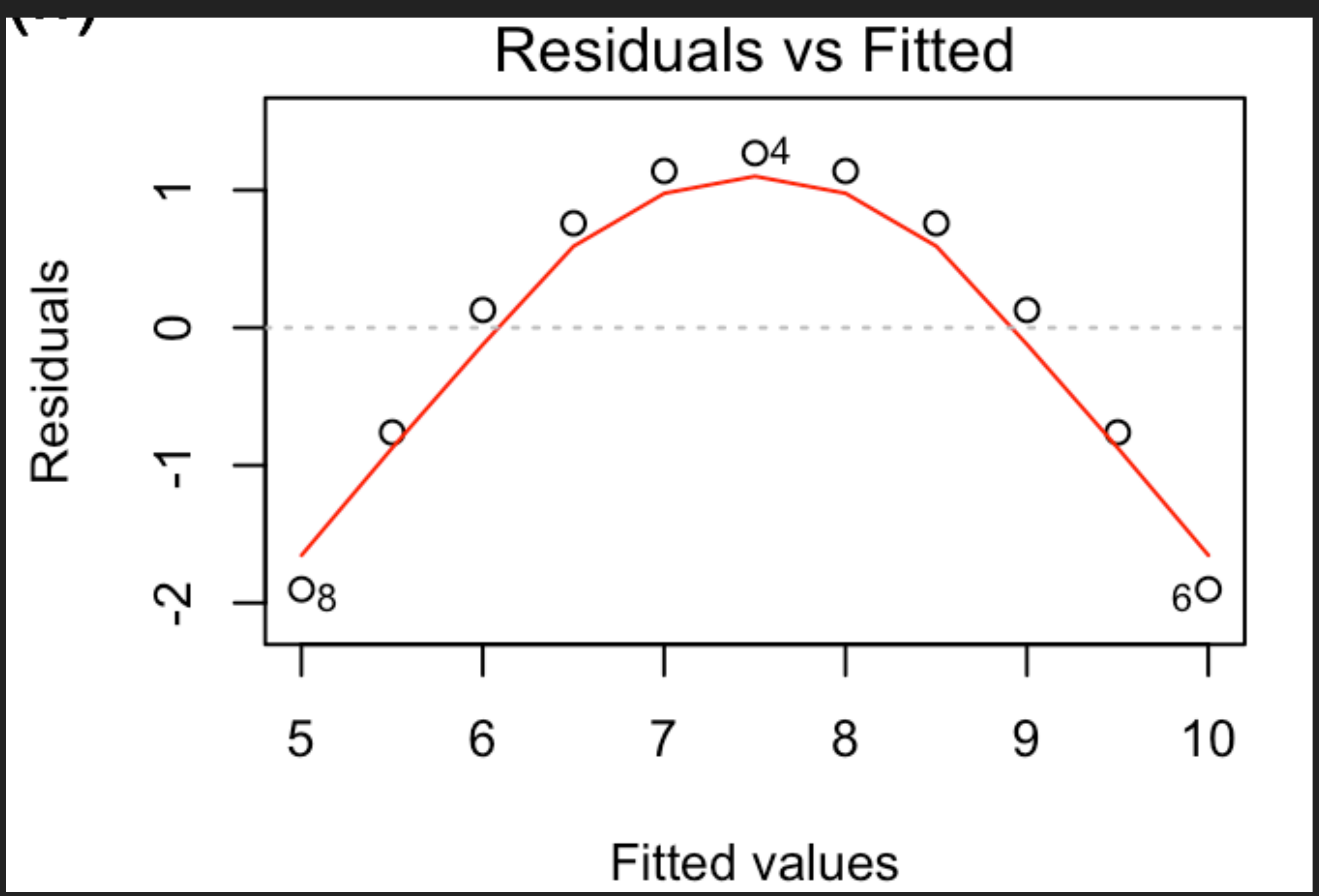
It´s complicated.

Potential Problem: Non-Linearity of the Data



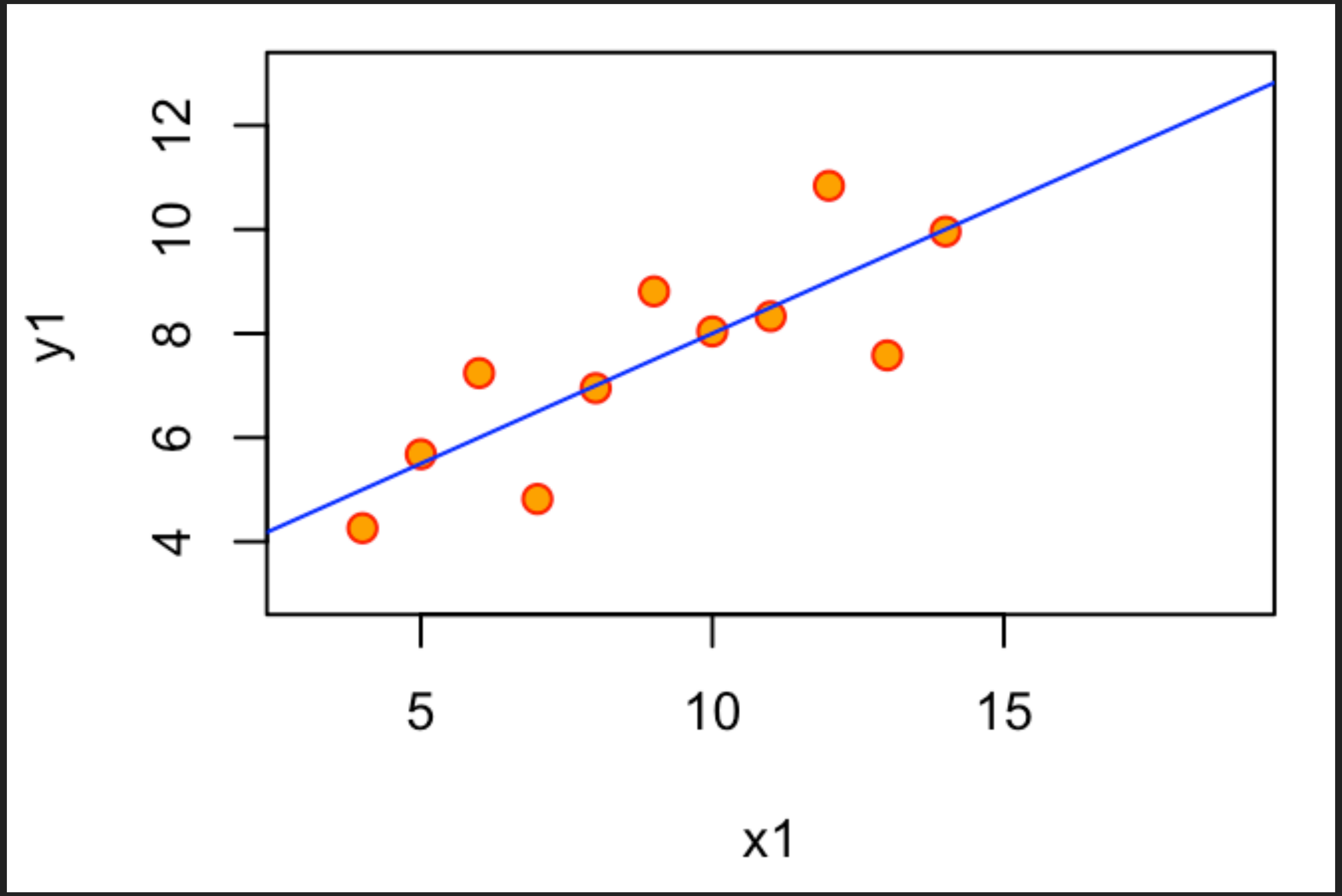
Potential Problem: Non-Linearity of the Data

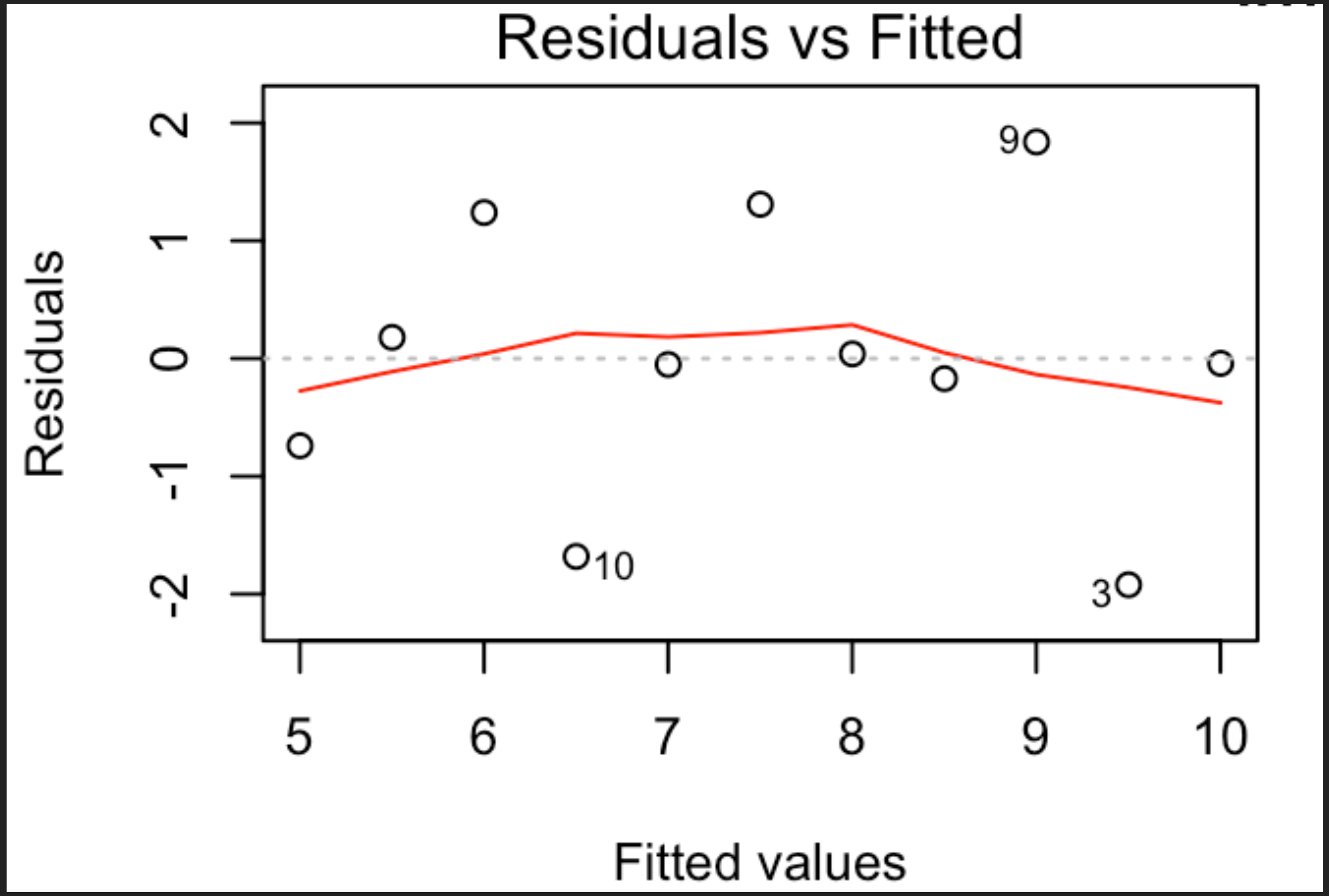




- Ideally, the residual plot will show no discernible pattern.
  - Otherwise, indicates nonlinear relationship in the data.

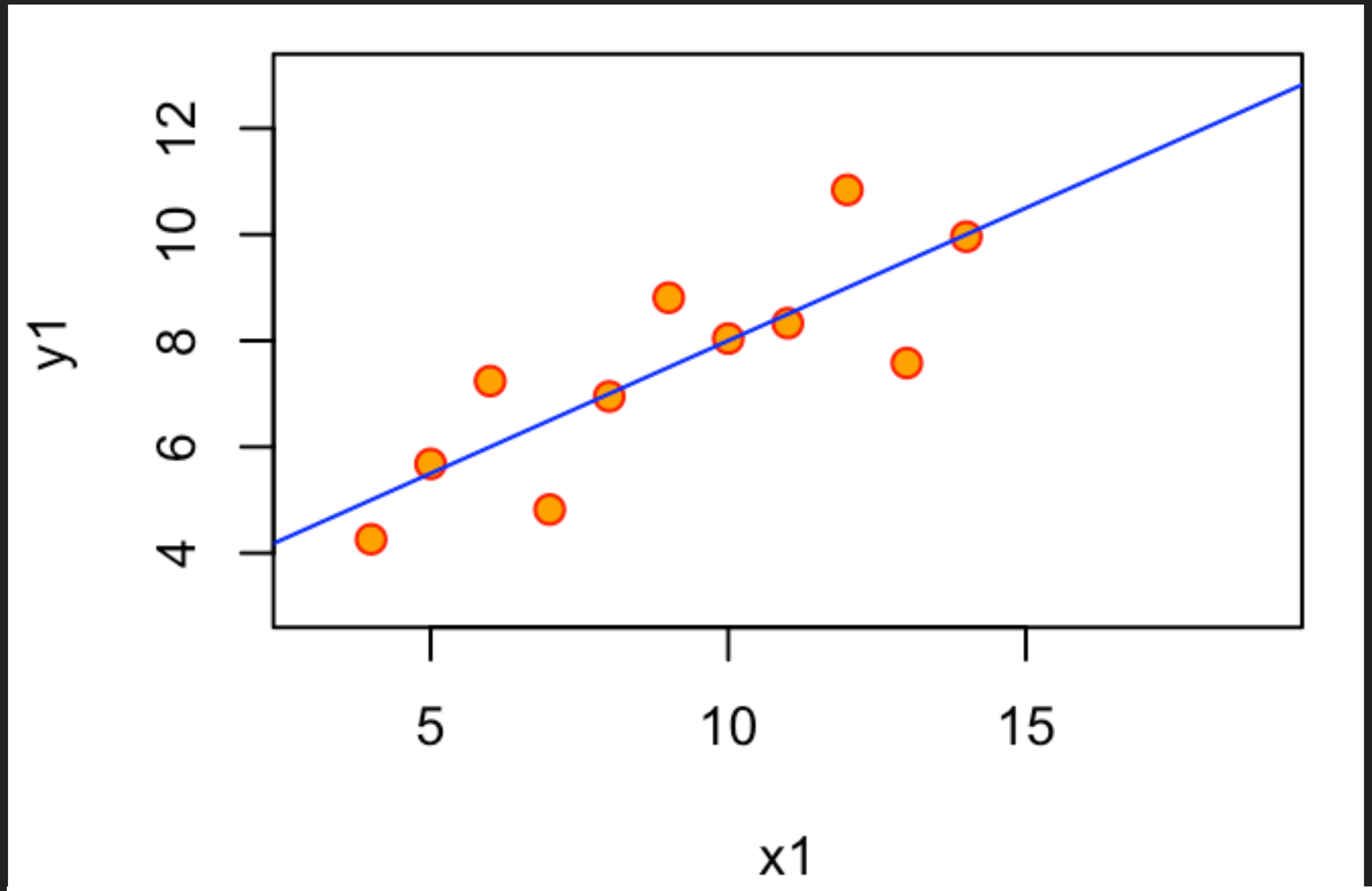
Potential Problem: Non-Linearity of the Data

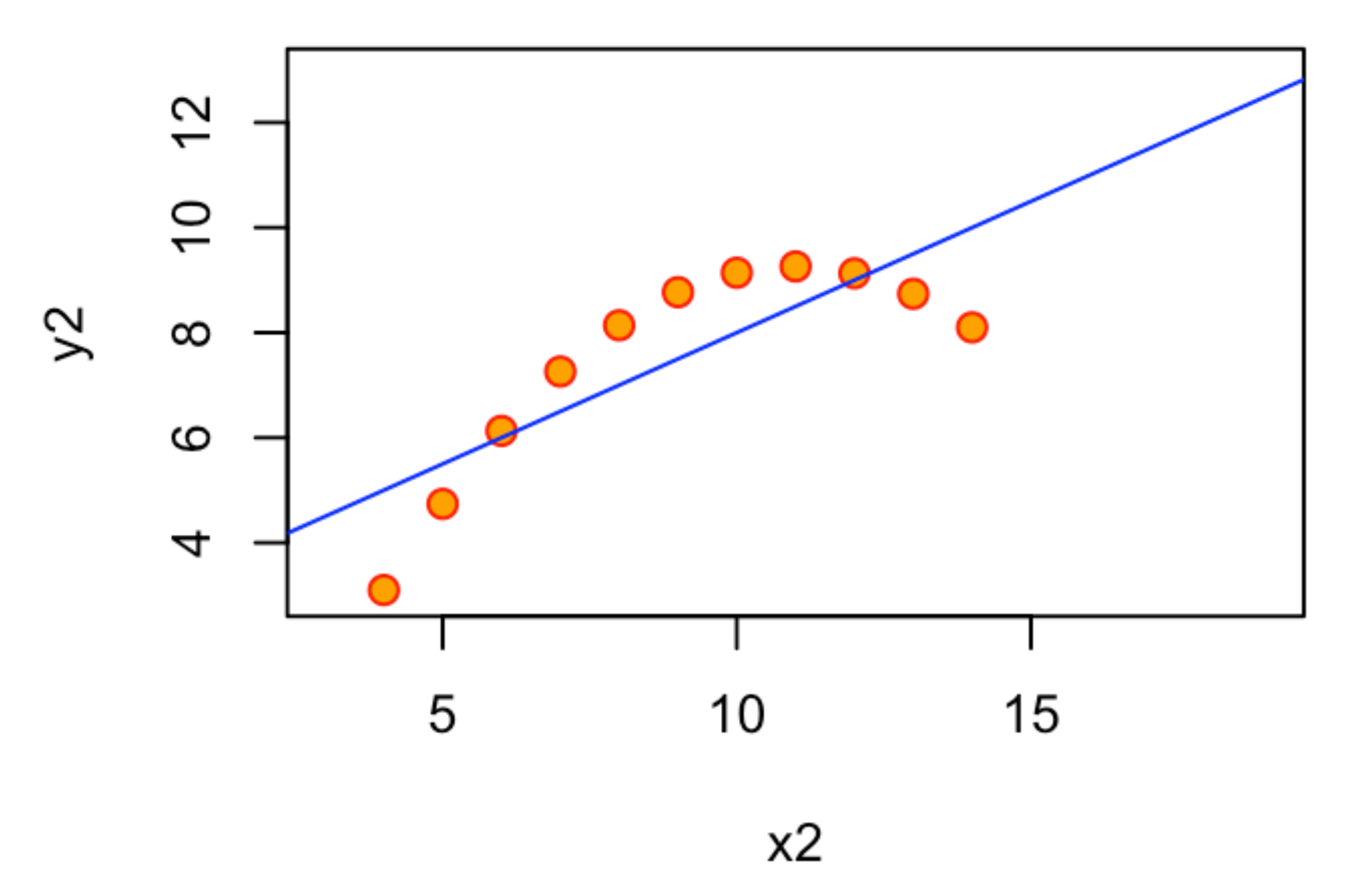




- Contrast the example on the previous slide to the one we had earlier.

Ouch!





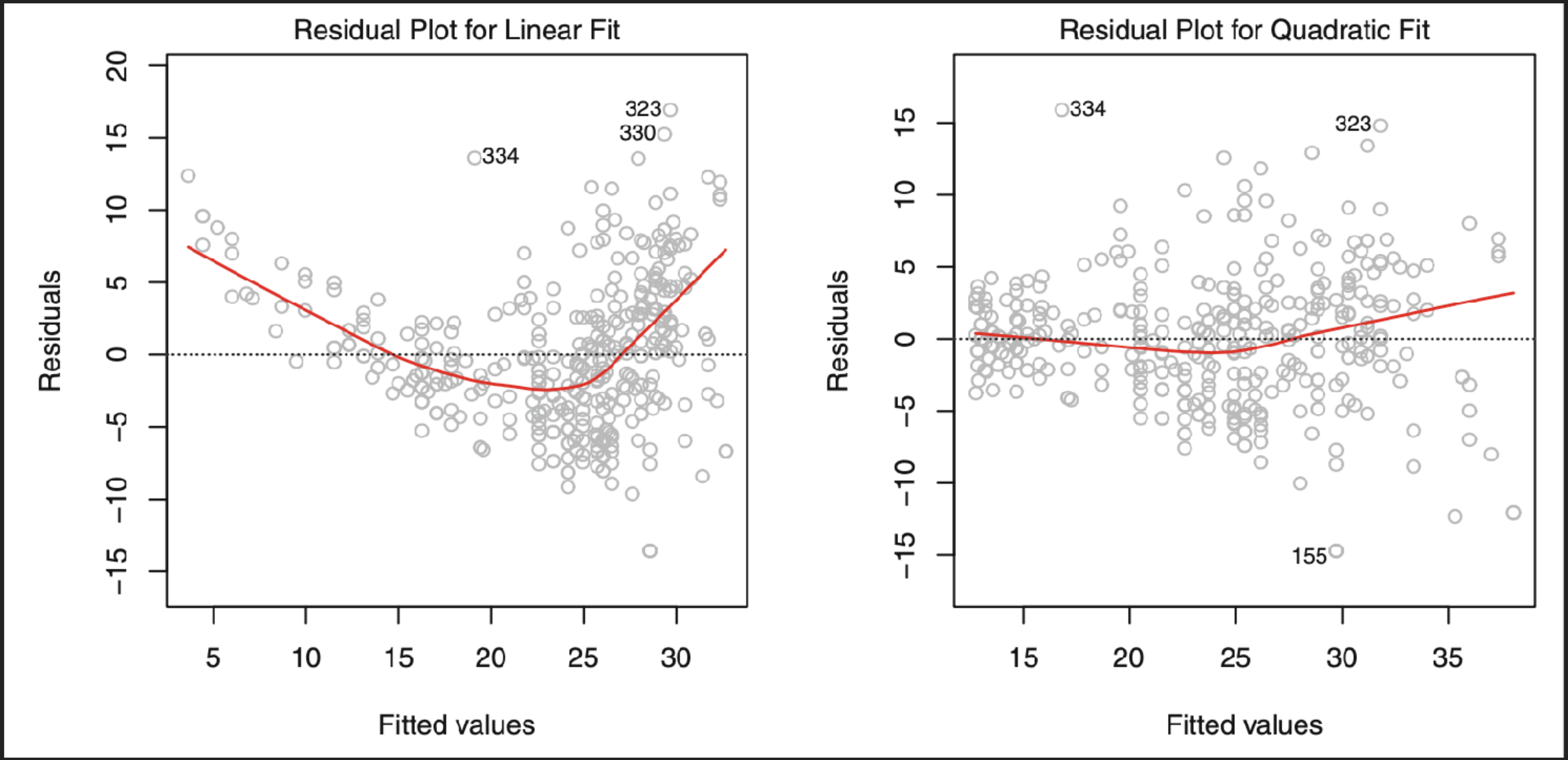

```
##
## Call:
## lm(formula = y1 ~ x1, data = anscombe)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.92127 -0.45577 -0.04136  0.70941  1.83882
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.0001     1.1247   2.667  0.02573 *
## x1            0.5001     0.1179   4.241  0.00217 **
##
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.237 on 9 degrees of freedom
## Multiple R-squared:  0.6665, Adjusted R-squared:  0.6295
## F-statistic: 17.99 on 1 and 9 DF,  p-value: 0.00217
```

```
##
## Call:
## lm(formula = y2 ~ x2, data = anscombe)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.9009 -0.7609  0.1291  0.9491  1.2691
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    3.001      1.125   2.667  0.02576 *
## x2             0.500      0.118   4.239  0.00218 **
##
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.237 on 9 degrees of freedom
## Multiple R-squared:  0.6662, Adjusted R-squared:  0.6292
## F-statistic: 17.97 on 1 and 9 DF,  p-value: 0.002179
```

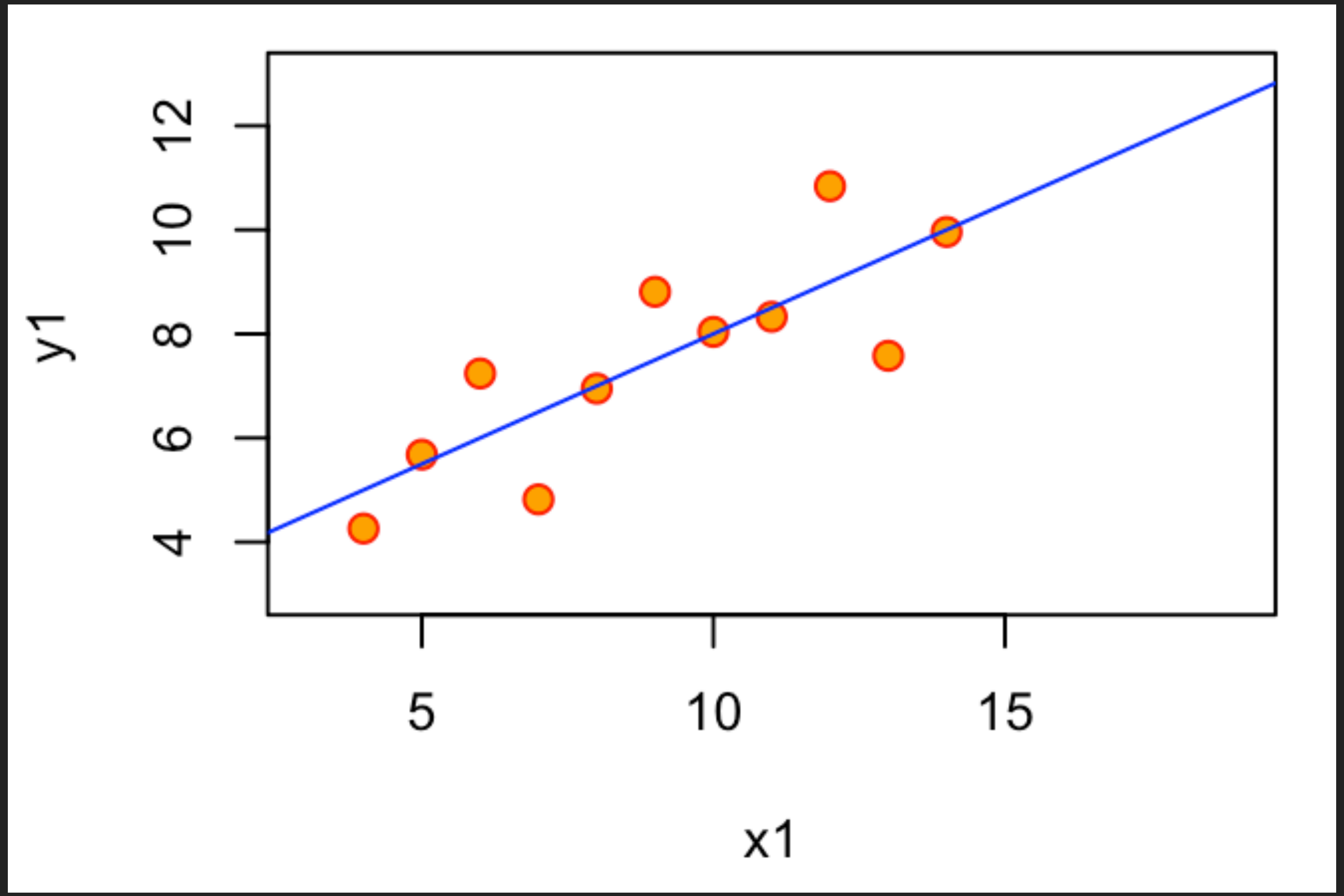
Another Example: Dealing With Non-Linearity

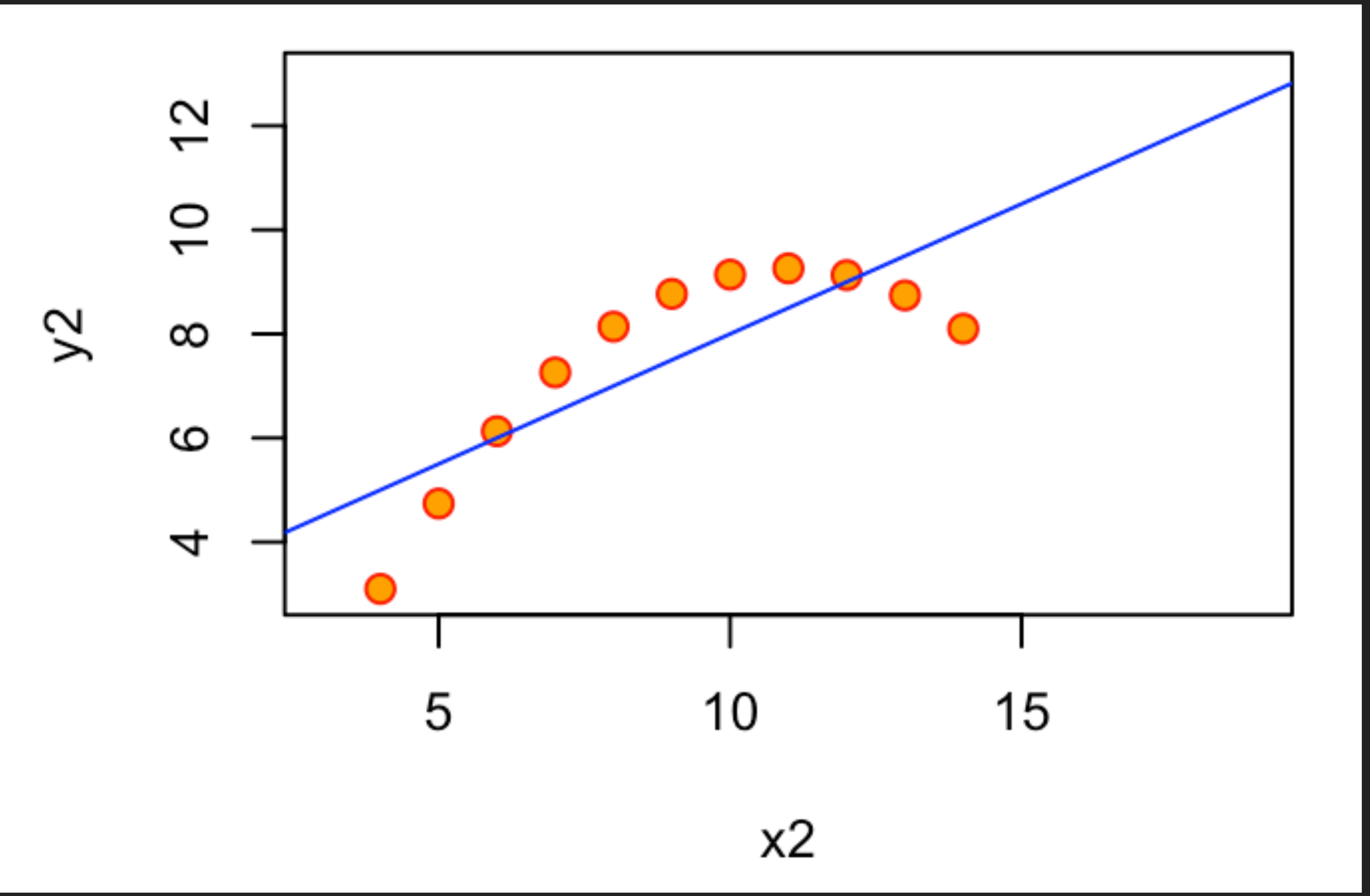
linear regression of mpg on horsepower

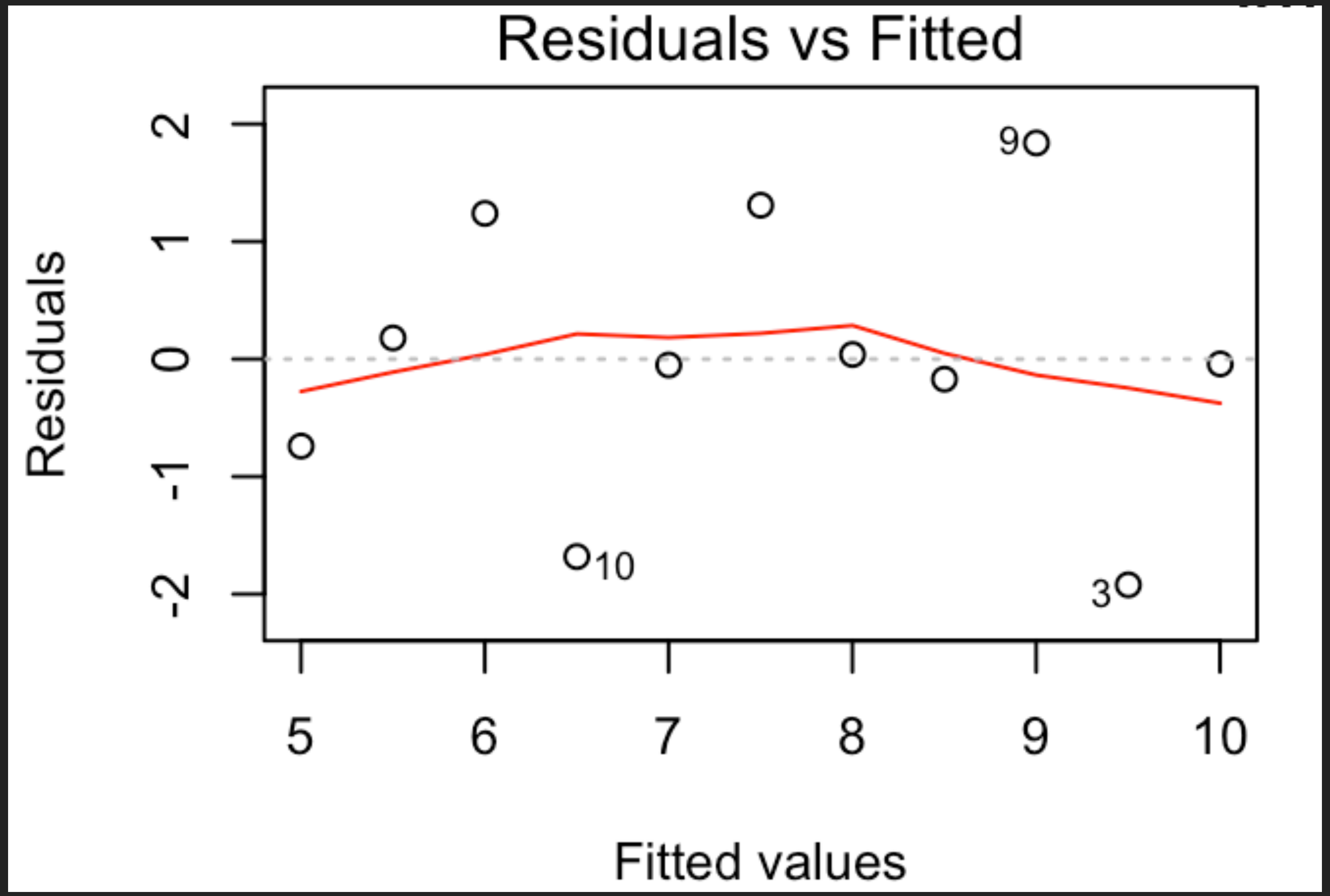
linear regression of mpg on horsepower and horsepower^2

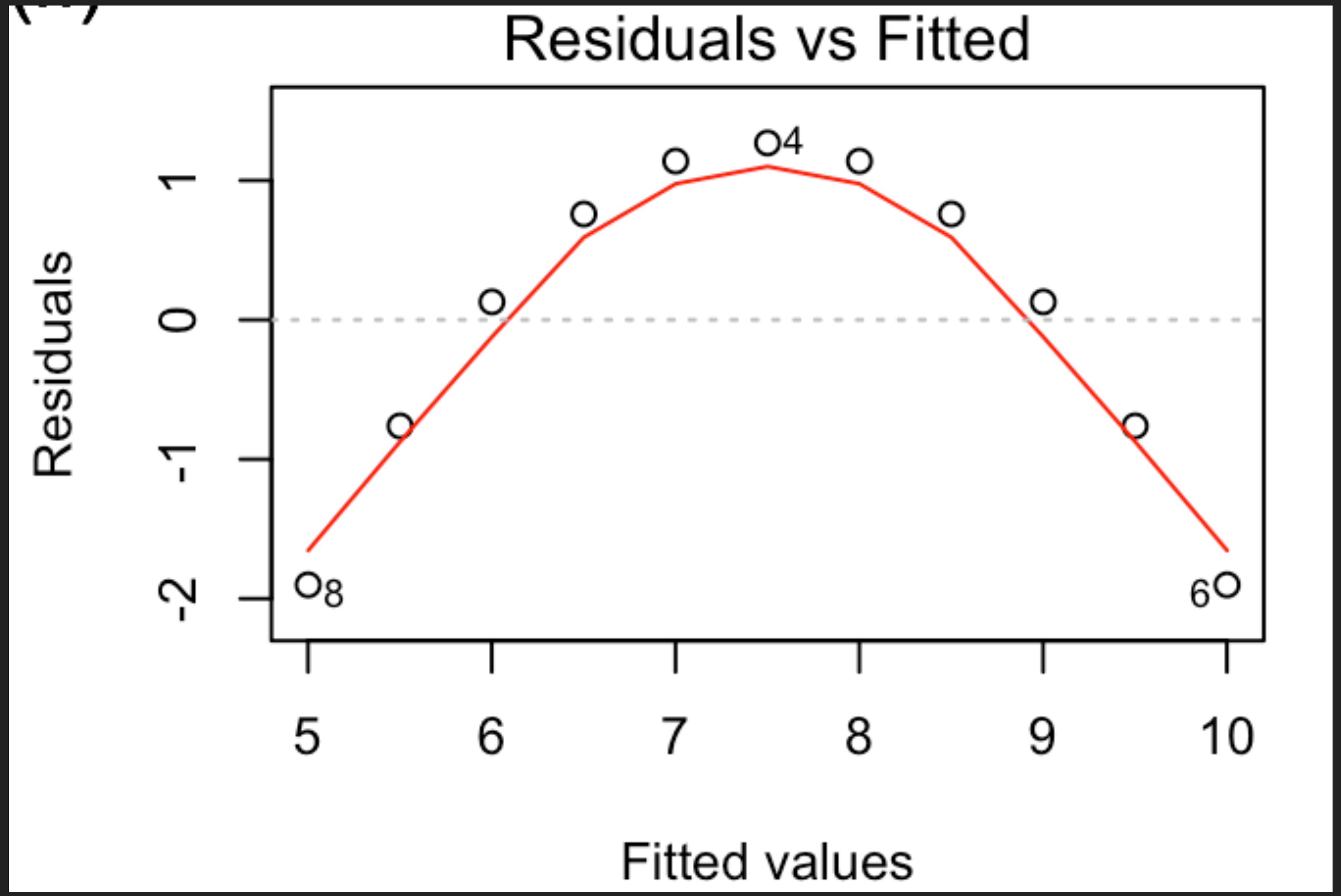


Remember This Example?



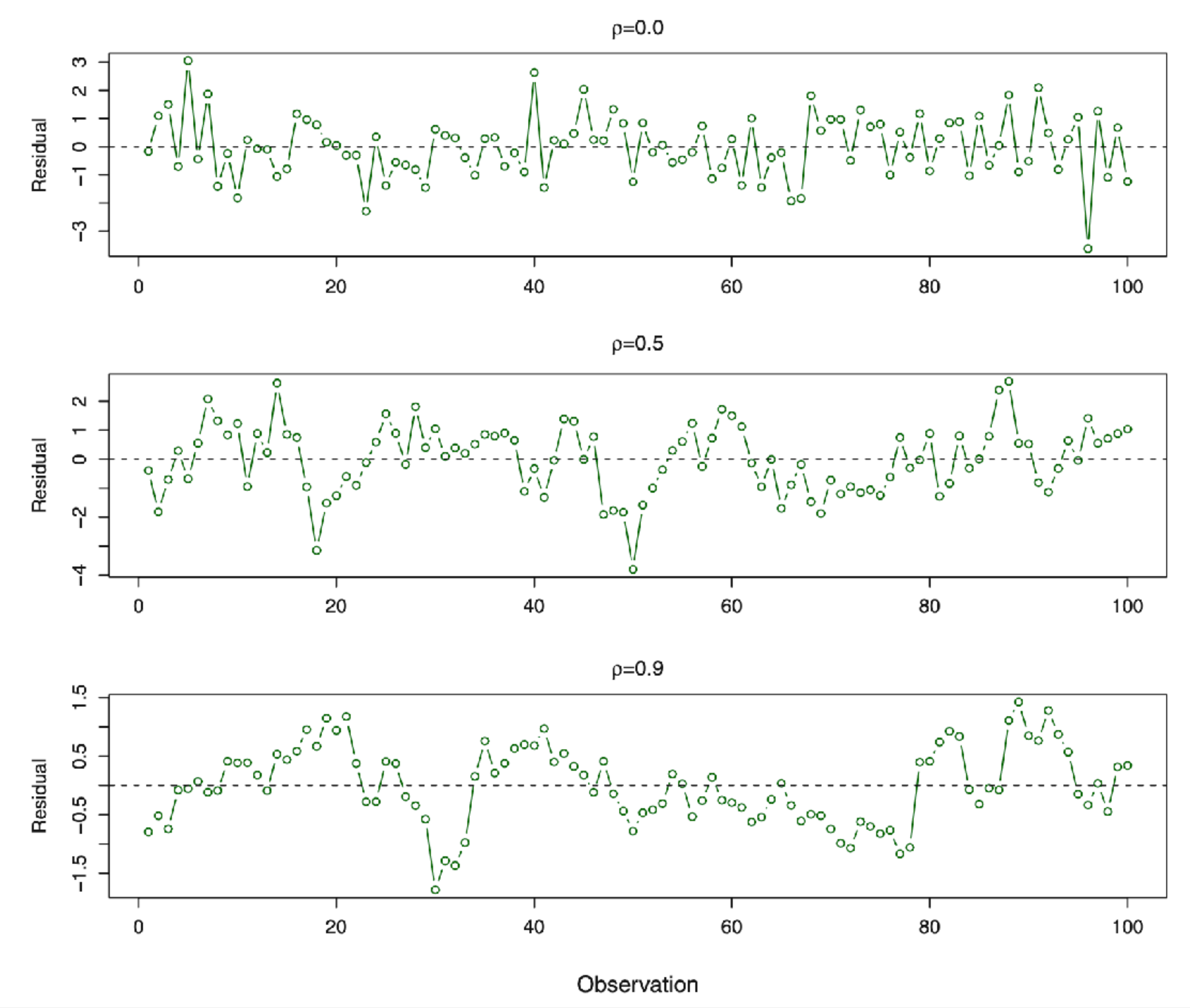






Potential Problem: Correlation of Error Terms

- Some causes:
  - Time series: observations at adjacent time points will have positively correlated errors
  - Also non time-series causes
- Effect:
  - The estimated standard errors will tend to underestimate the true standard errors.

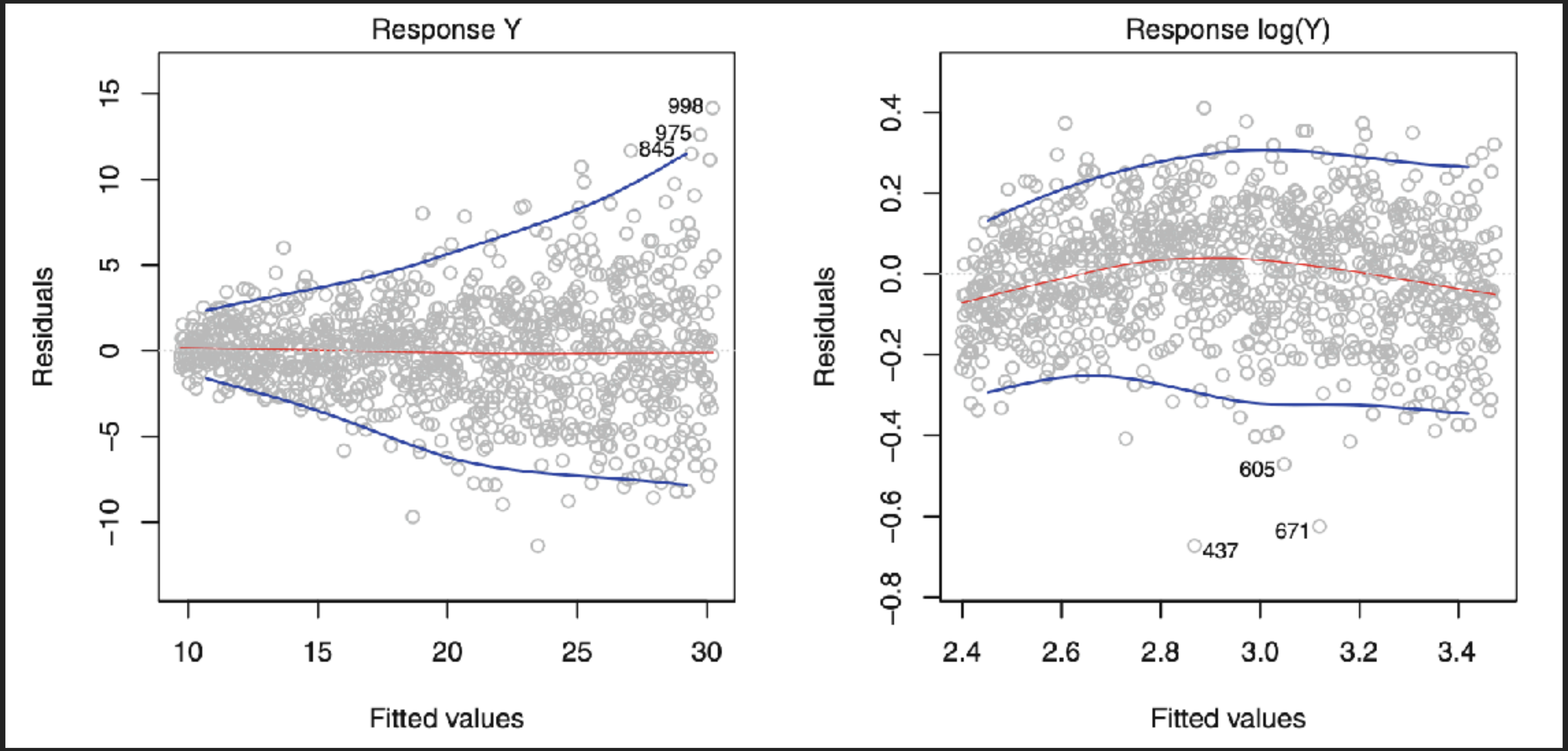


Plots of residuals from simulated time series data sets generated with differing levels of correlation between error terms for adjacent time points.

Potential Problem: Non-Constant Variance of Error Terms ("Heteroscedasticity")

- Symptom: the variances of the error terms may increase with the value of the response.

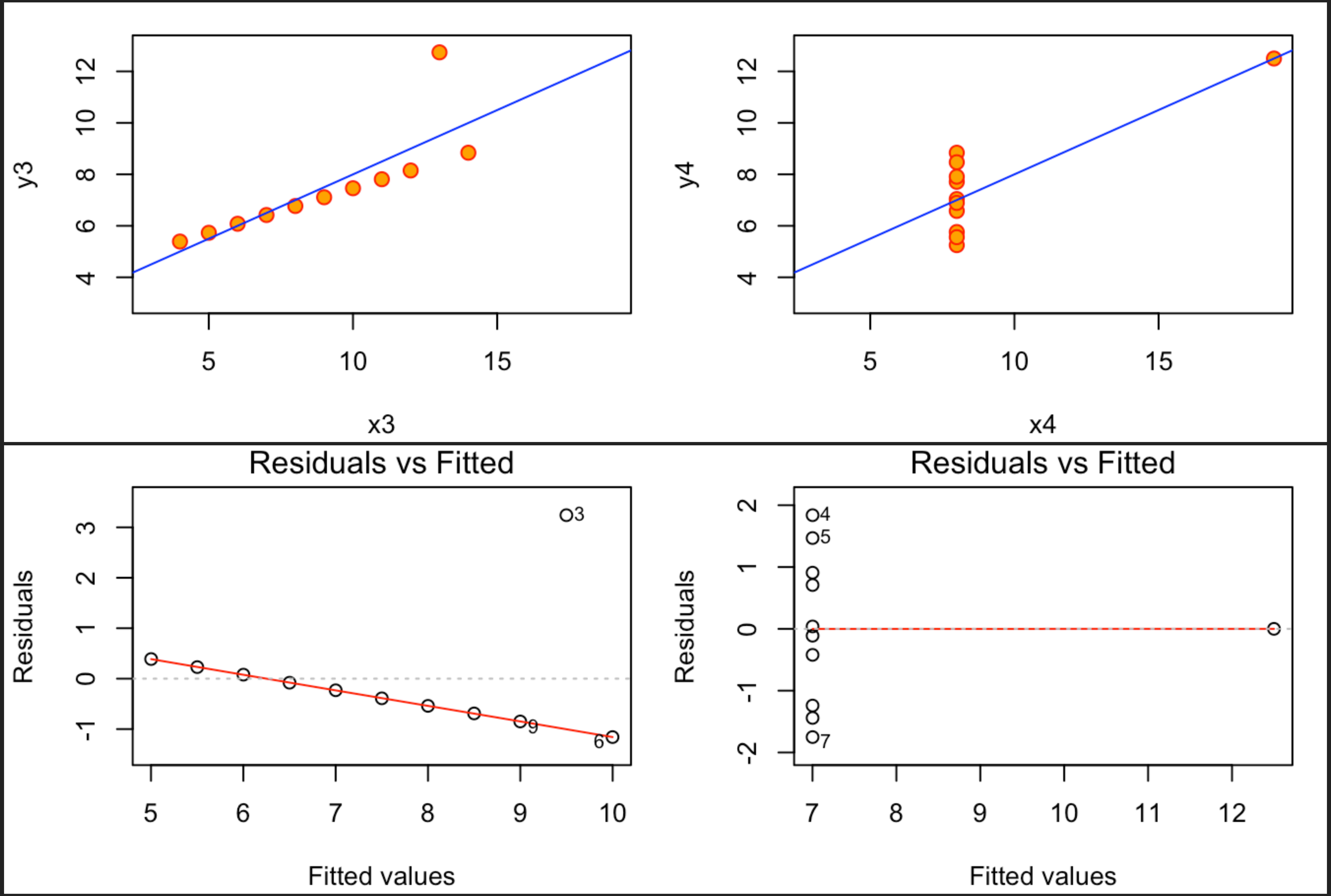
The funnel shape indicates heteroscedasticity.



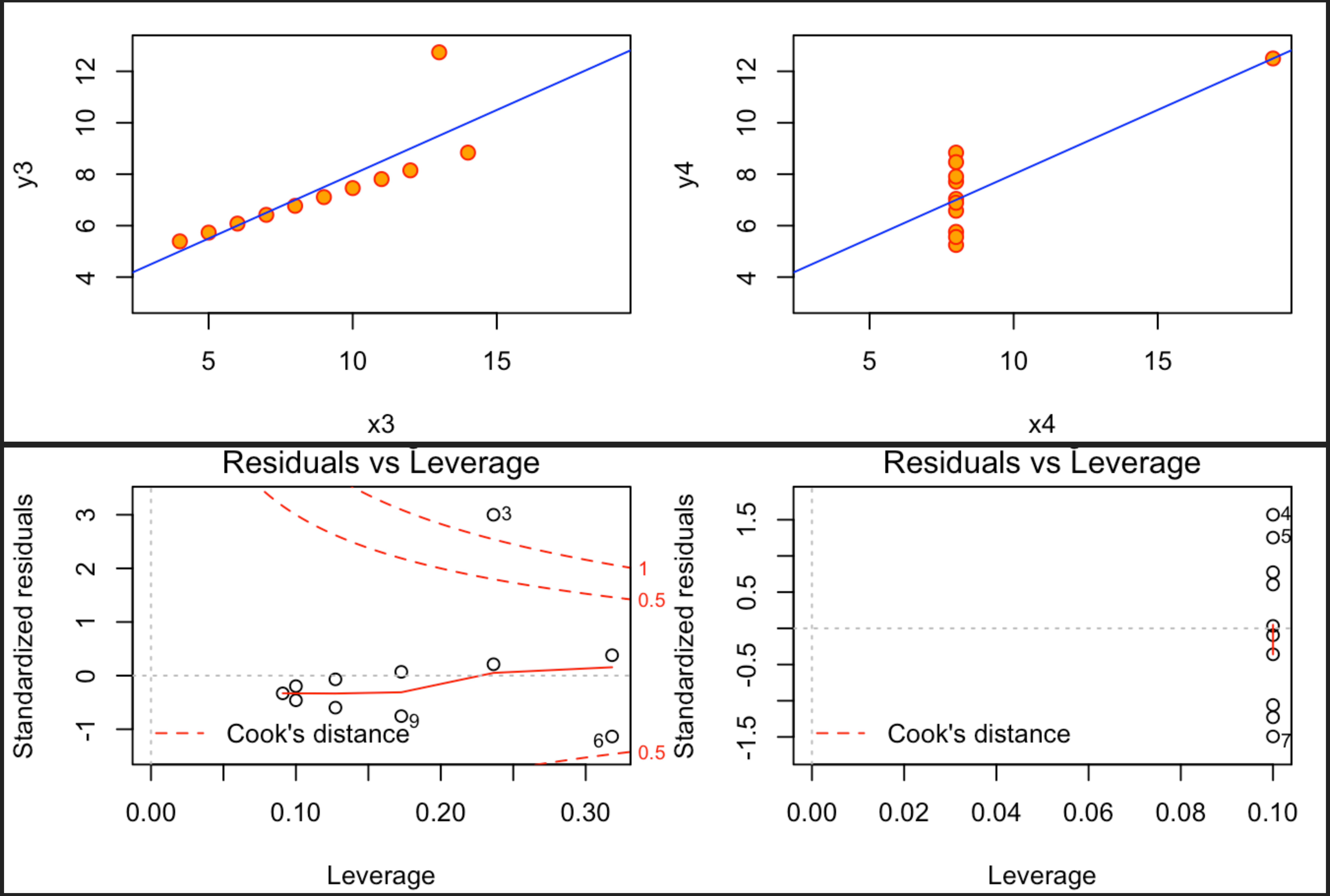
The response has been log transformed, and there is now no evidence of heteroscedasticity.

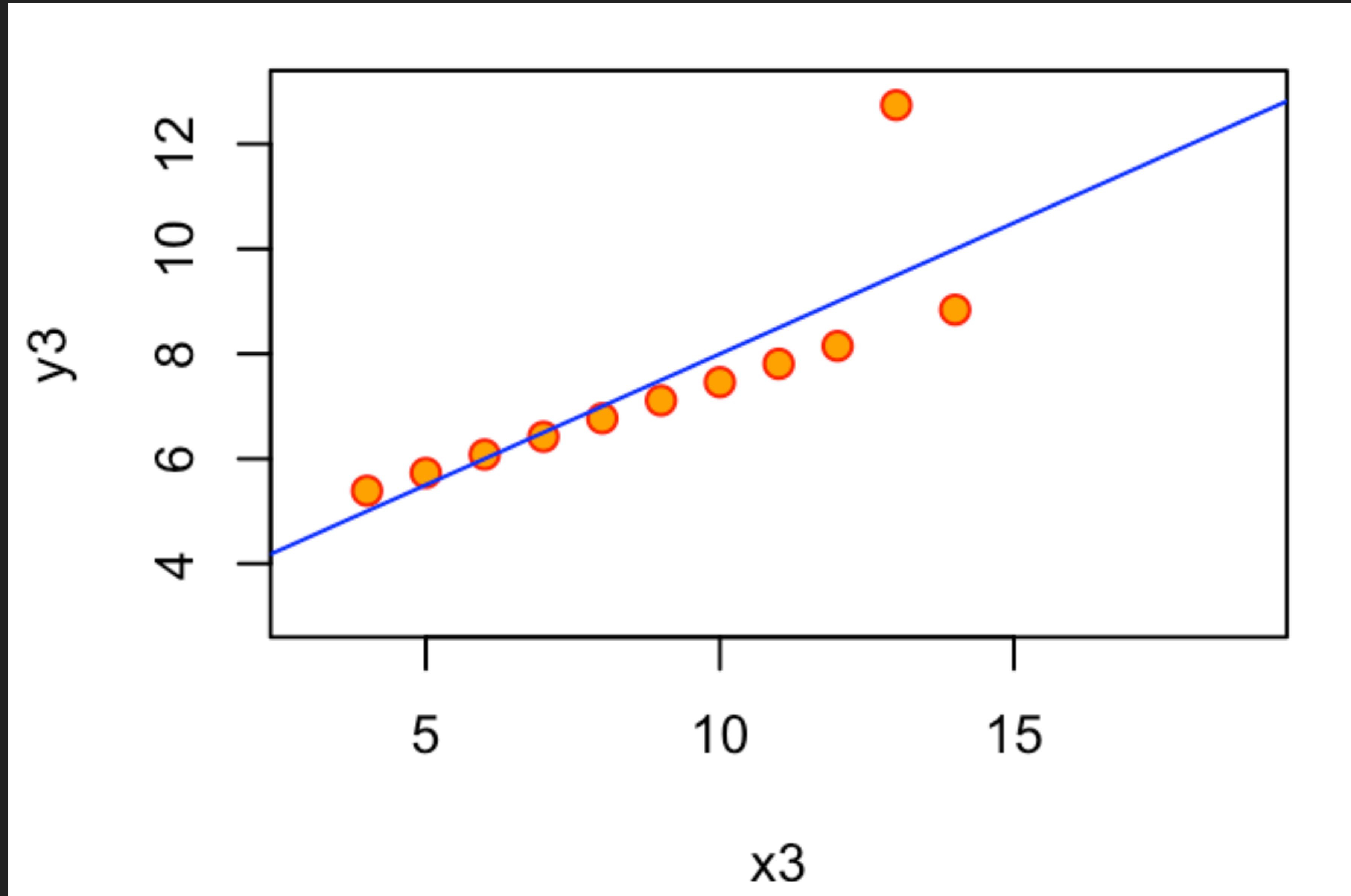
Heteroscedasticity tends to produce p-values that are smaller than they should be.

Potential Problems: Outliers and High Leverage Points

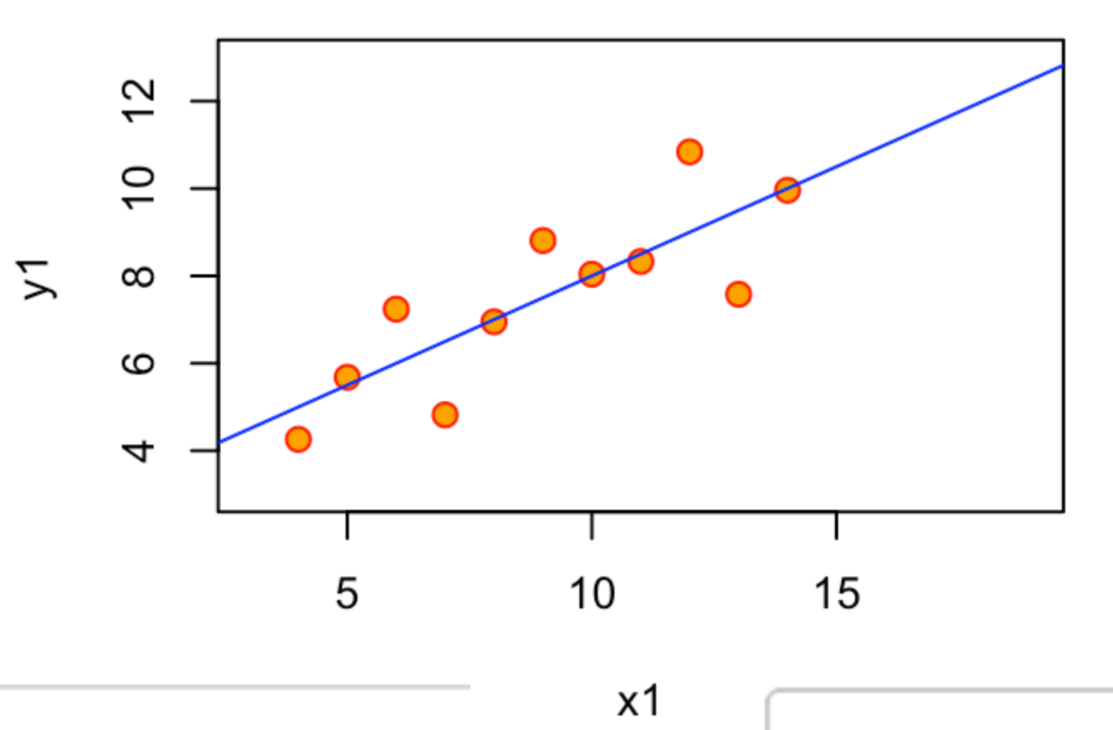


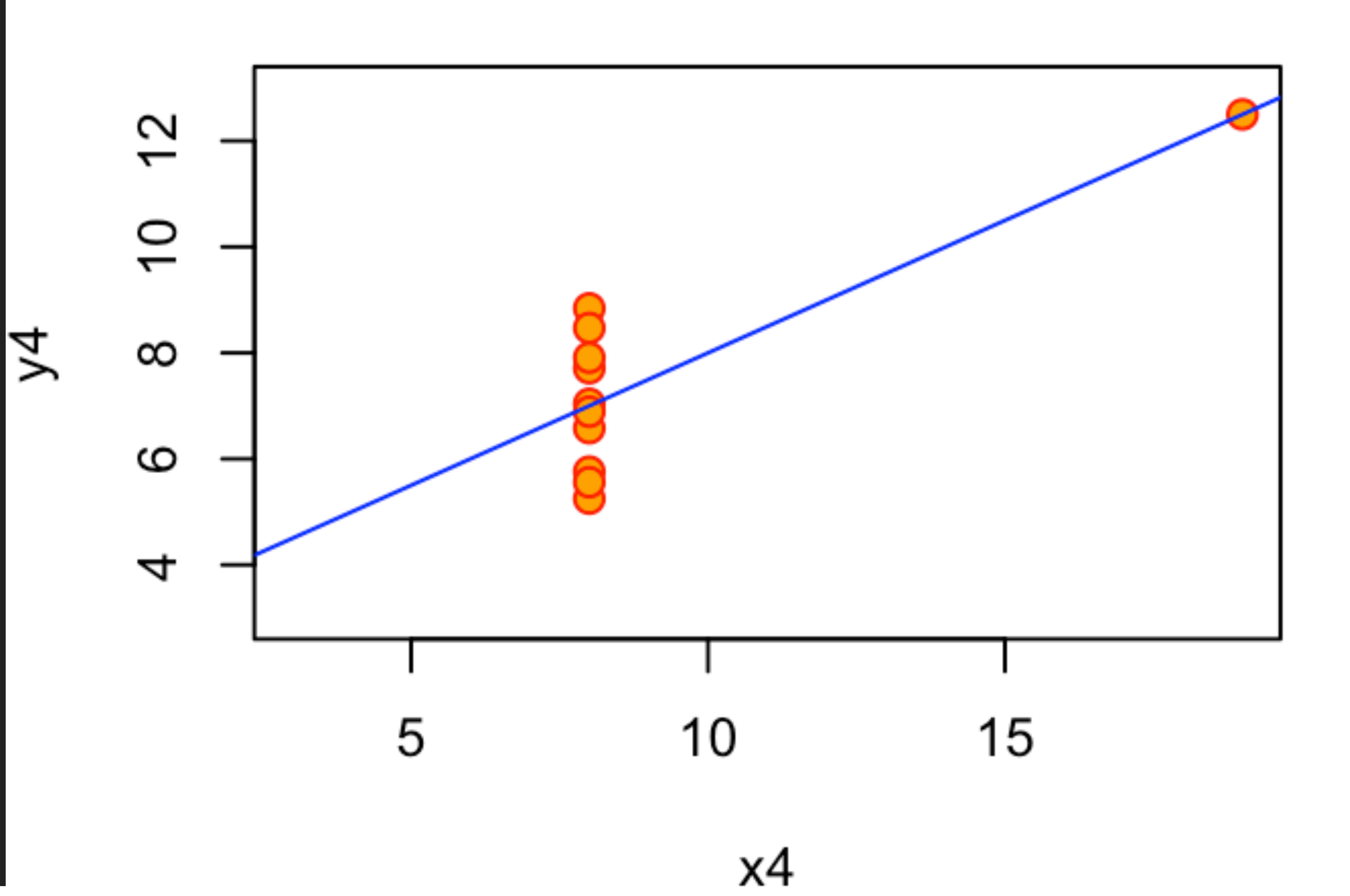
Potential Problems: Outliers and High Leverage Points





Remember the Earlier Example?






```
##
## Call:
## lm(formula = y3 ~ x3, data = anscombe)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.1586 -0.6146 -0.2303  0.1540  3.2411
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.0025     1.1245   2.670  0.02562 *
## x3            0.4997     0.1179   4.239  0.00218 **
##
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.236 on 9 degrees of freedom
## Multiple R-squared:  0.6663, Adjusted R-squared:  0.6292
## F-statistic: 17.97 on 1 and 9 DF,  p-value: 0.002176
```

```
##
## Call:
## lm(formula = y4 ~ x4, data = anscombe)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -1.751 -0.831  0.000  0.809  1.839
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.0017     1.1239   2.671  0.02559 *
## x4            0.4999     0.1178   4.243  0.00216 **
##
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.236 on 9 degrees of freedom
## Multiple R-squared:  0.6667, Adjusted R-squared:  0.6297
## F-statistic:    18 on 1 and 9 DF,  p-value: 0.002165
```

Potential Problem: Collinearity

- Here's an extreme example of perfectly collinear data.
- By construction, x1 and x2 are exactly the same variable, and the outcome y is perfectly modeled as $y = x_1 + x_2$

```
my.data <- data.frame(y =  c(12, 13, 10, 5, 7, 12, 15),
                      x1 = c(6, 6.5, 5, 2.5, 3.5, 6, 7.5),
                      x2 = c(6, 6.5, 5, 2.5, 3.5, 6, 7.5))
my.data
```

Potential Problem: Collinearity

- Here's an extreme example of perfectly collinear data.
- By construction, x1 and x2 are exactly the same variable, and the outcome y is perfectly modeled as $y = x_1 + x_2$
- But there's a problem… because the following are also true

$$y = 2x_1$$

$$y = 3x_1 - x_2$$

$$y = -400x_1 + 402x_2$$

```
my.data <- data.frame(y =  c(12, 13, 10, 5, 7, 12, 15),
                      x1 = c(6, 6.5, 5, 2.5, 3.5, 6, 7.5),
                      x2 = c(6, 6.5, 5, 2.5, 3.5, 6, 7.5))
my.data
```

Potential Problem: Collinearity

- Here's an extreme example of perfectly collinear data.
- By construction, x1 and x2 are exactly the same variable, and the outcome y is perfectly modeled as $y = x_1 + x_2$
- But there's a problem… because the following are also true

$$y = 2x_1$$

$$y = 3x_1 - x_2$$

$$y = -400x_1 + 402x_2$$

```r
my.data <- data.frame(y =  c(12, 13, 10, 5, 7, 12, 15),
                      x1 = c(6, 6.5, 5, 2.5, 3.5, 6, 7.5),
                      x2 = c(6, 6.5, 5, 2.5, 3.5, 6, 7.5))
my.data
```

Effects:

- The model is unable to accurately distinguish between many nearly equally plausible linear combinations of colinear variables.
- This can lead to large standard errors on coefficients, and even coefficient signs that don't make sense.

Potential Problem: Collinearity

- Here's an extreme example of perfectly collinear data.
- By construction, x1 and x2 are exactly the same variable, and the outcome y is perfectly modeled as $y = x_1 + x_2$
- But there's a problem… because the following are also true

$$y = 2x_1$$

$$y = 3x_1 - x_2$$

$$y = -400x_1 + 402x_2$$

```
my.data <- data.frame(y =  c(12, 13, 10, 5, 7, 12, 15),
                      x1 = c(6, 6.5, 5, 2.5, 3.5, 6, 7.5),
                      x2 = c(6, 6.5, 5, 2.5, 3.5, 6, 7.5))
my.data
```

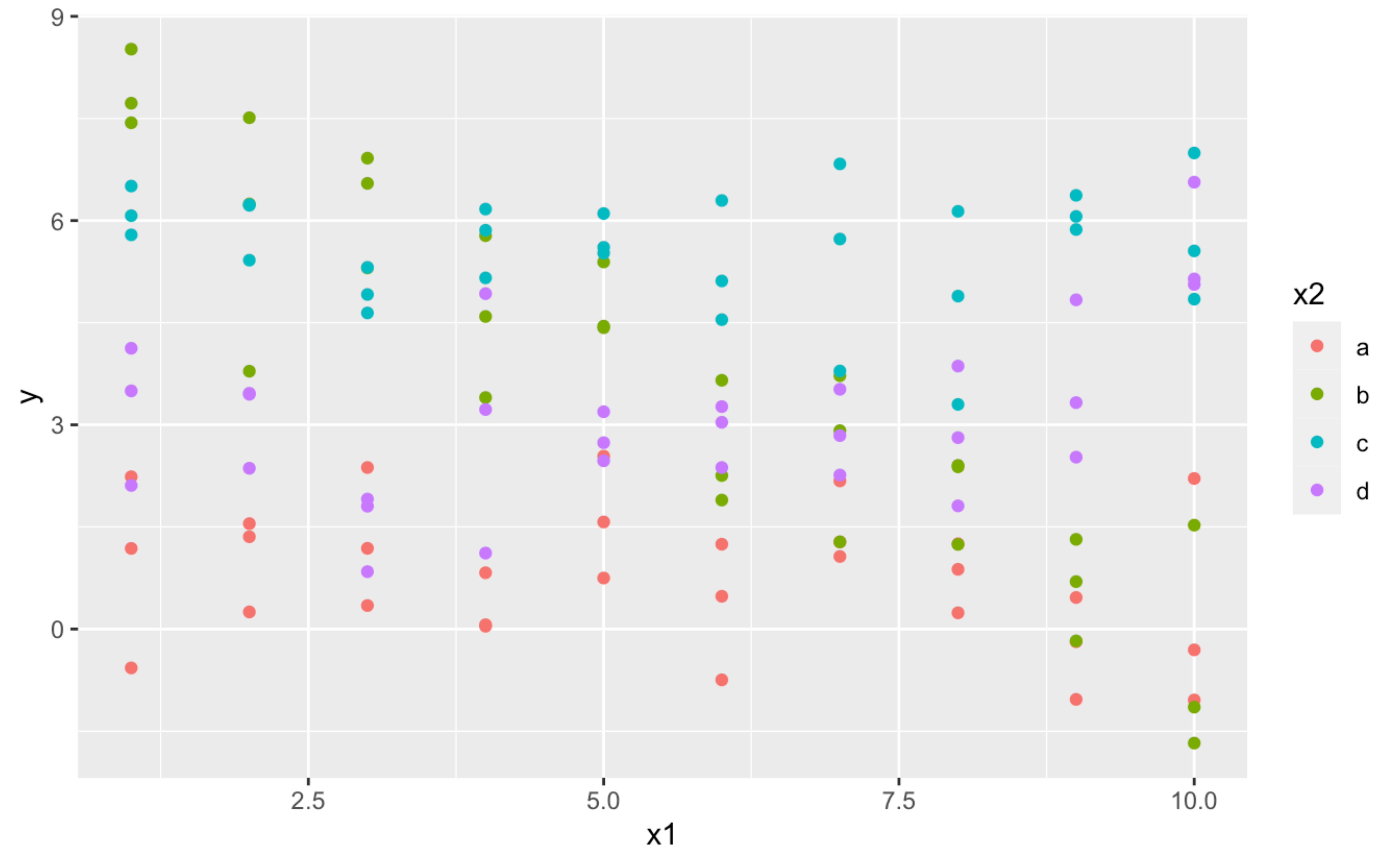
```
# Evaluate Collinearity
library(car)
vif(fit) # variance inflation factors
sqrt(vif(fit)) > 2 # problem?
```

Activity: How To Address These Questions?

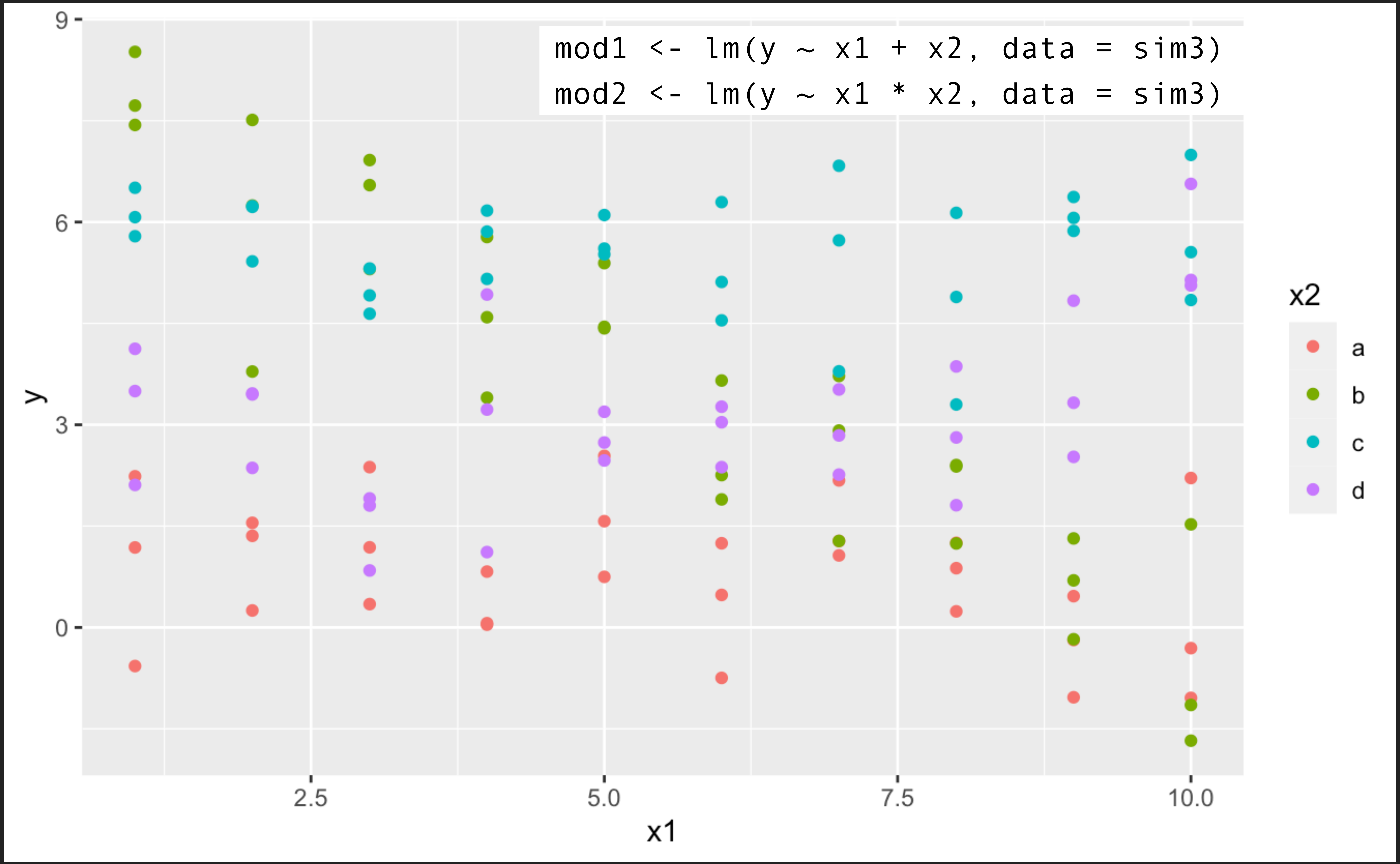
- Is there a relationship between advertising budget and sales?
- How strong is the relationship between advertising budget and sales?
- Which media contribute to sales?
- How accurately can we estimate the effect of each medium on sales?
- How accurately can we predict future sales?
- Is the relationship linear?
- Is there synergy among the advertising media?

Another interaction example

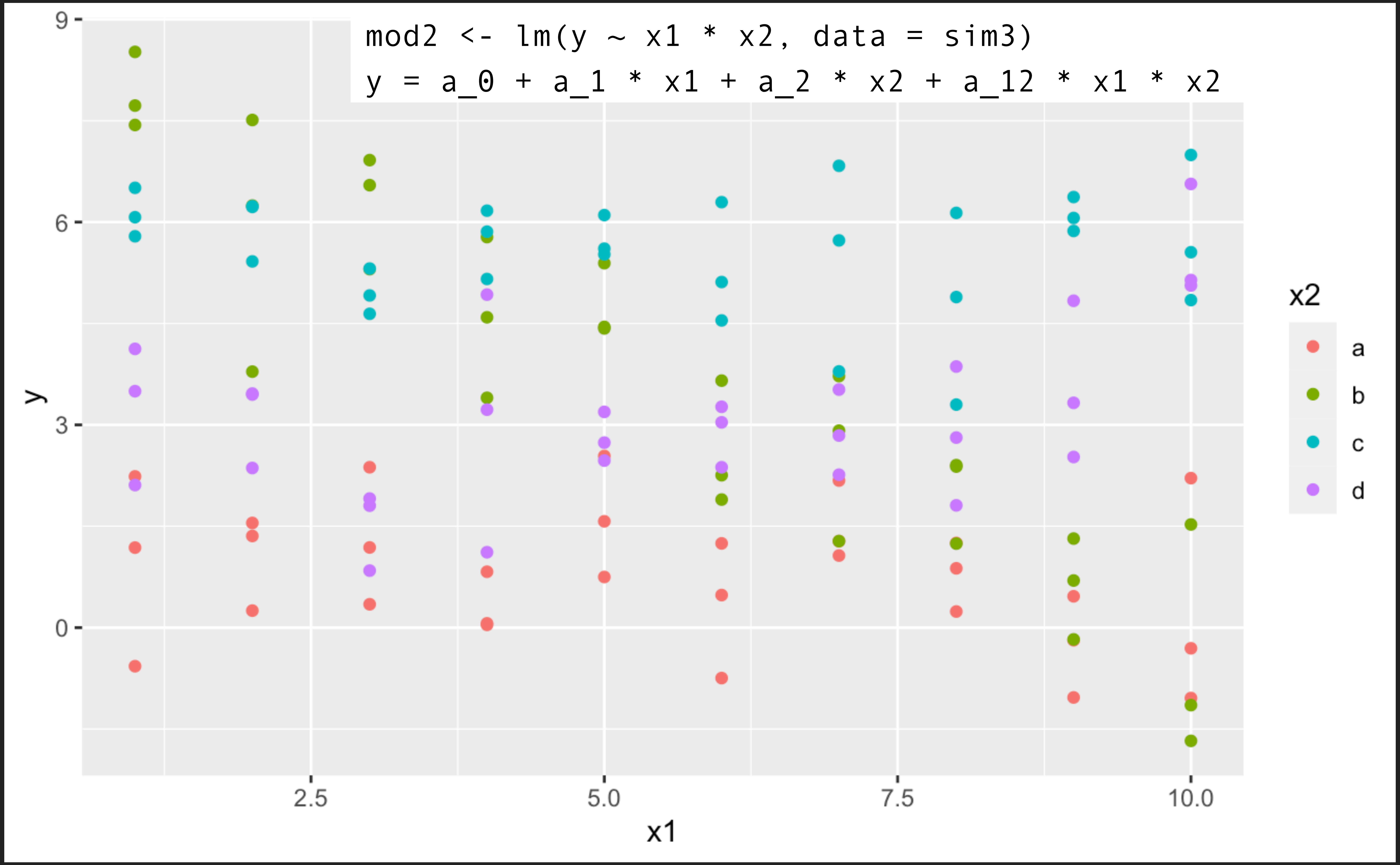
What happens when you combine a continuous and a categorical variable?



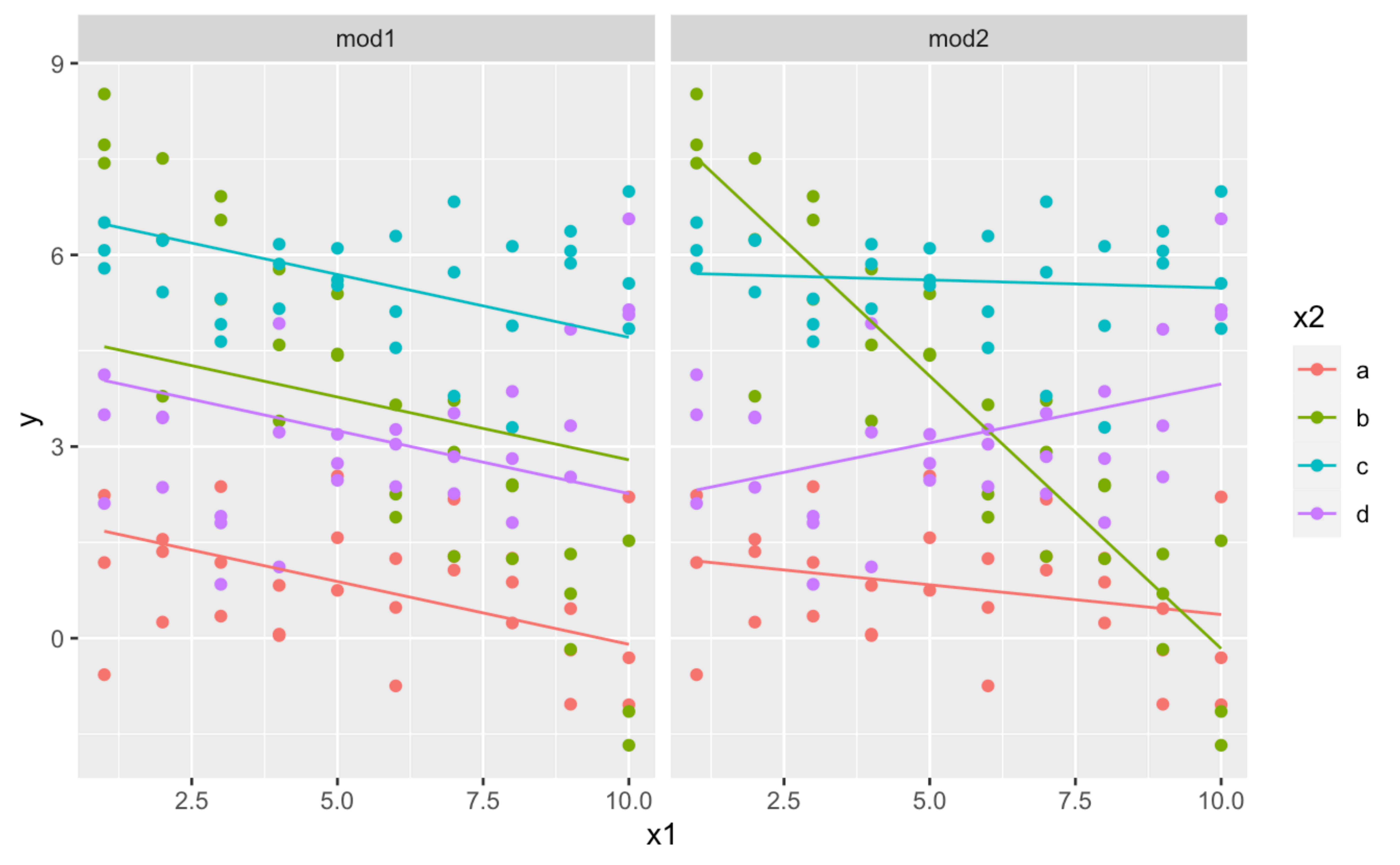
There are two possible models you could fit to this data



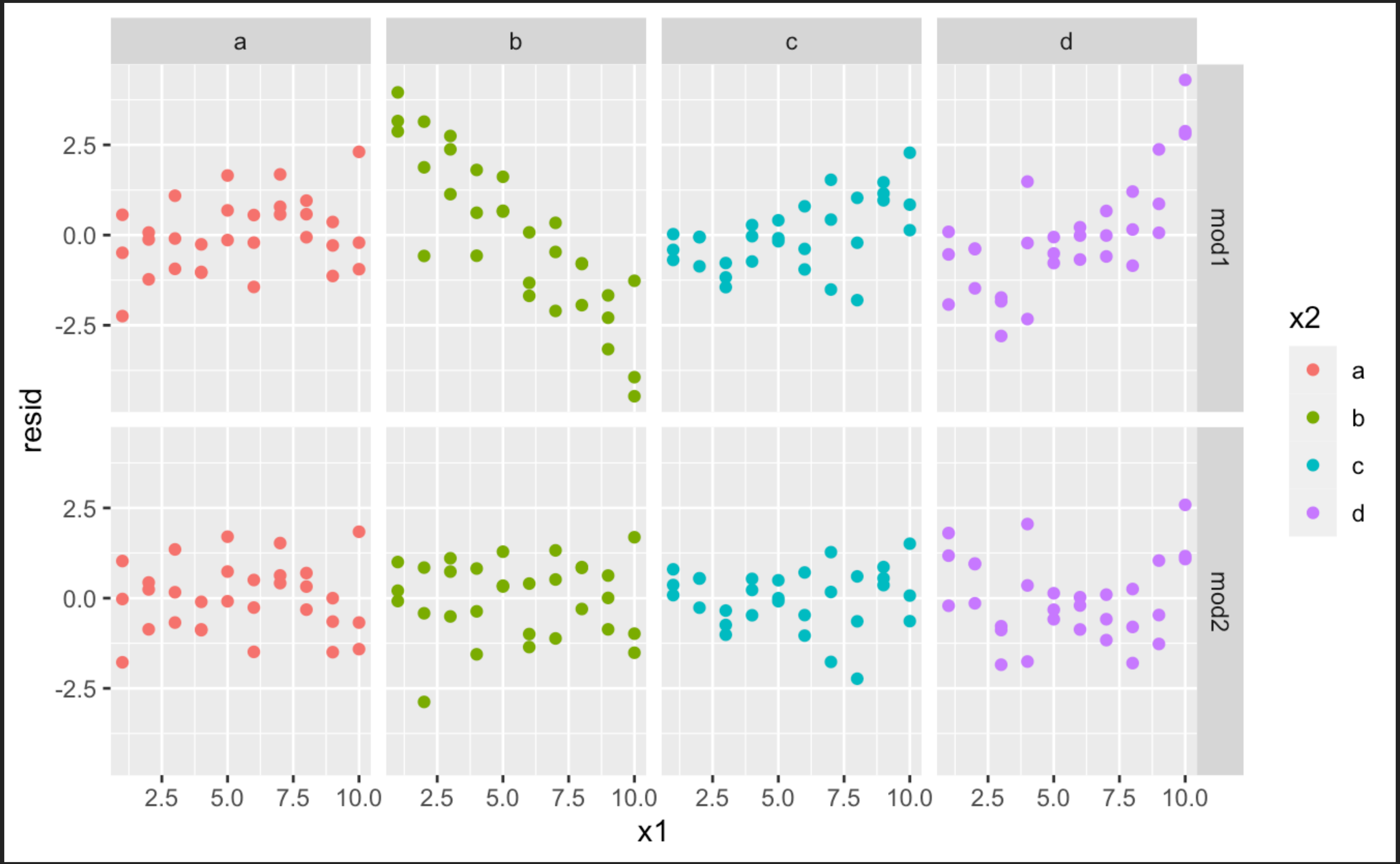
*: Both the interaction and the individual components are included in the model



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

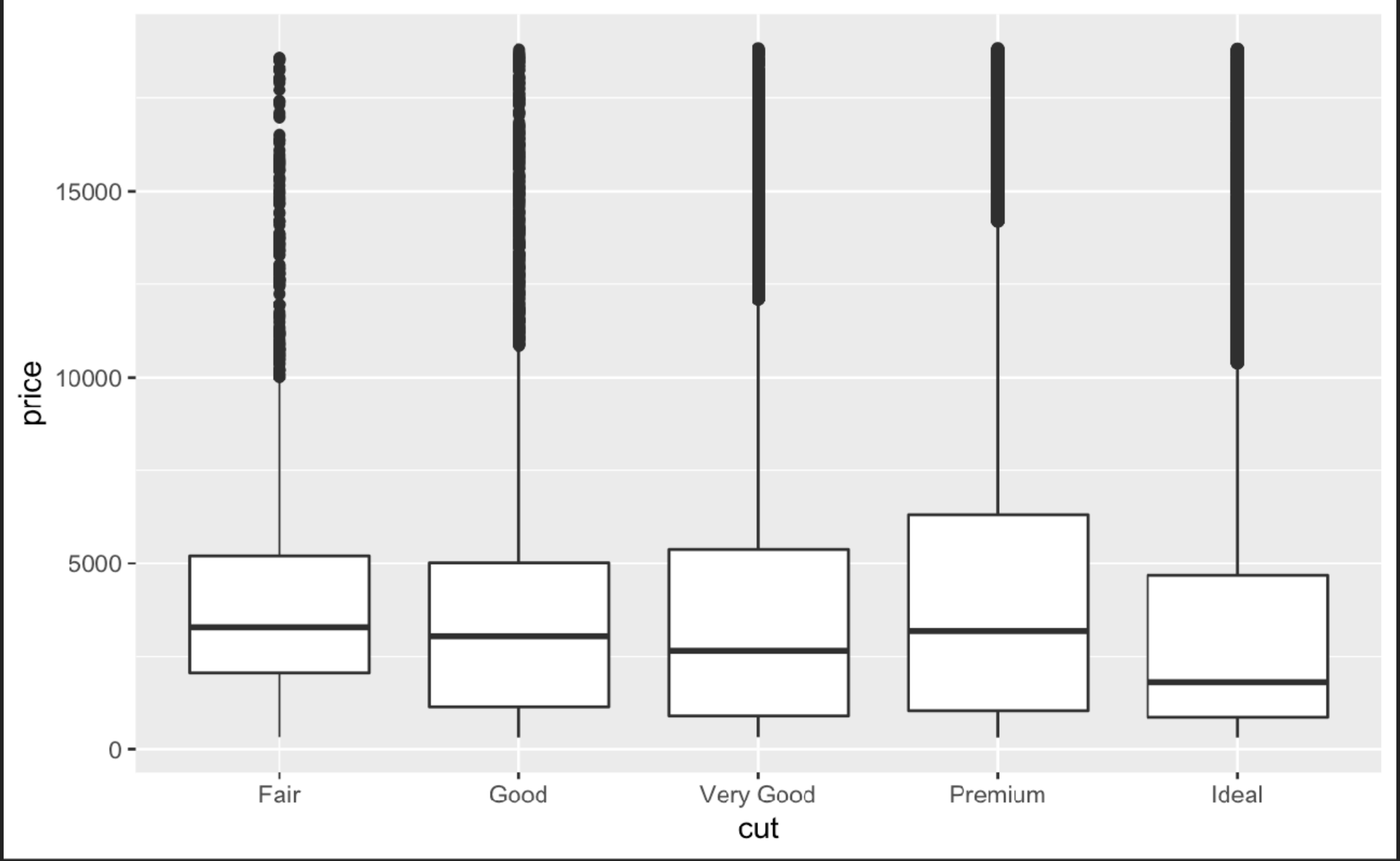


Which model is better for this data?



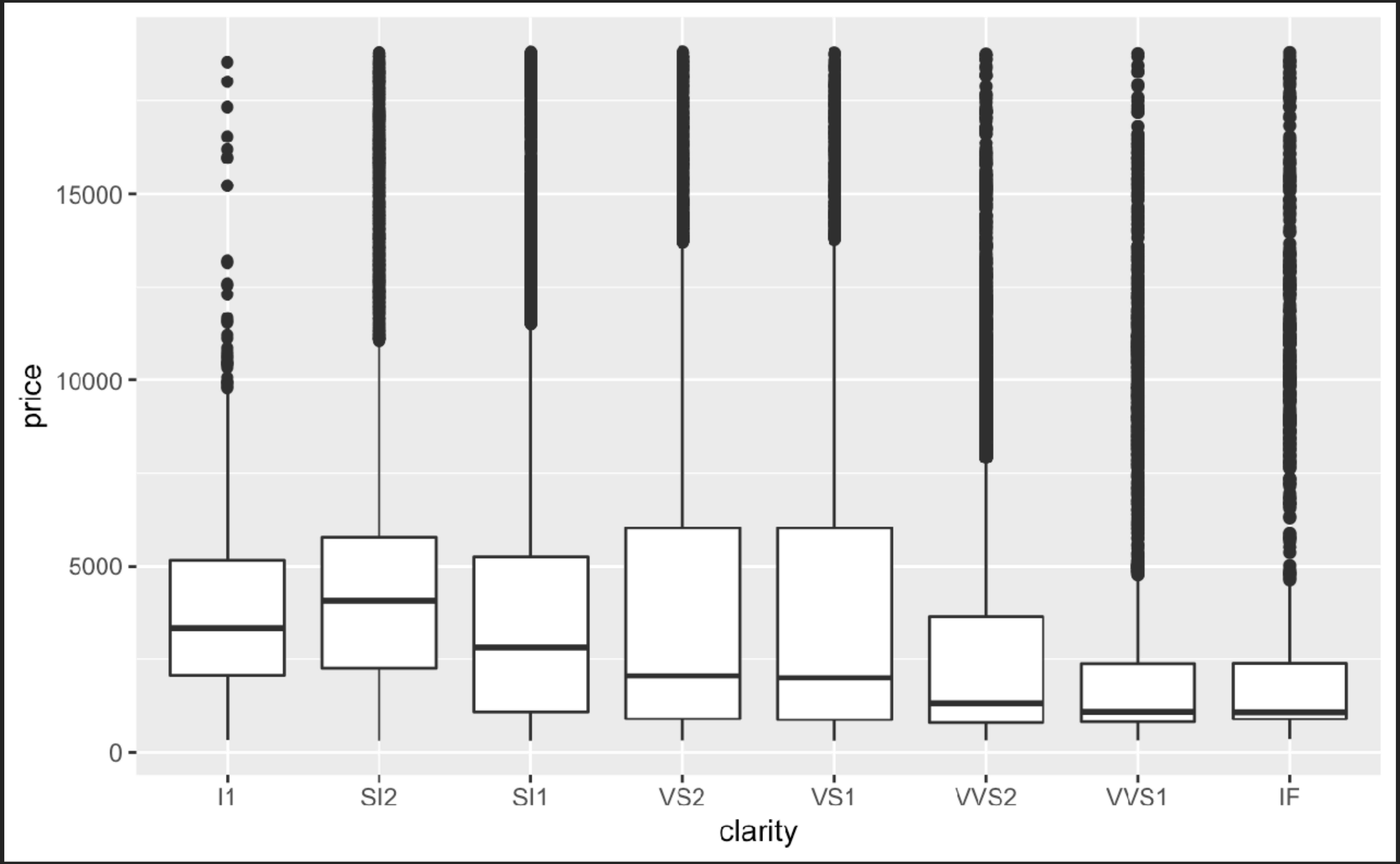
Another example

Why are low quality diamonds more expensive?



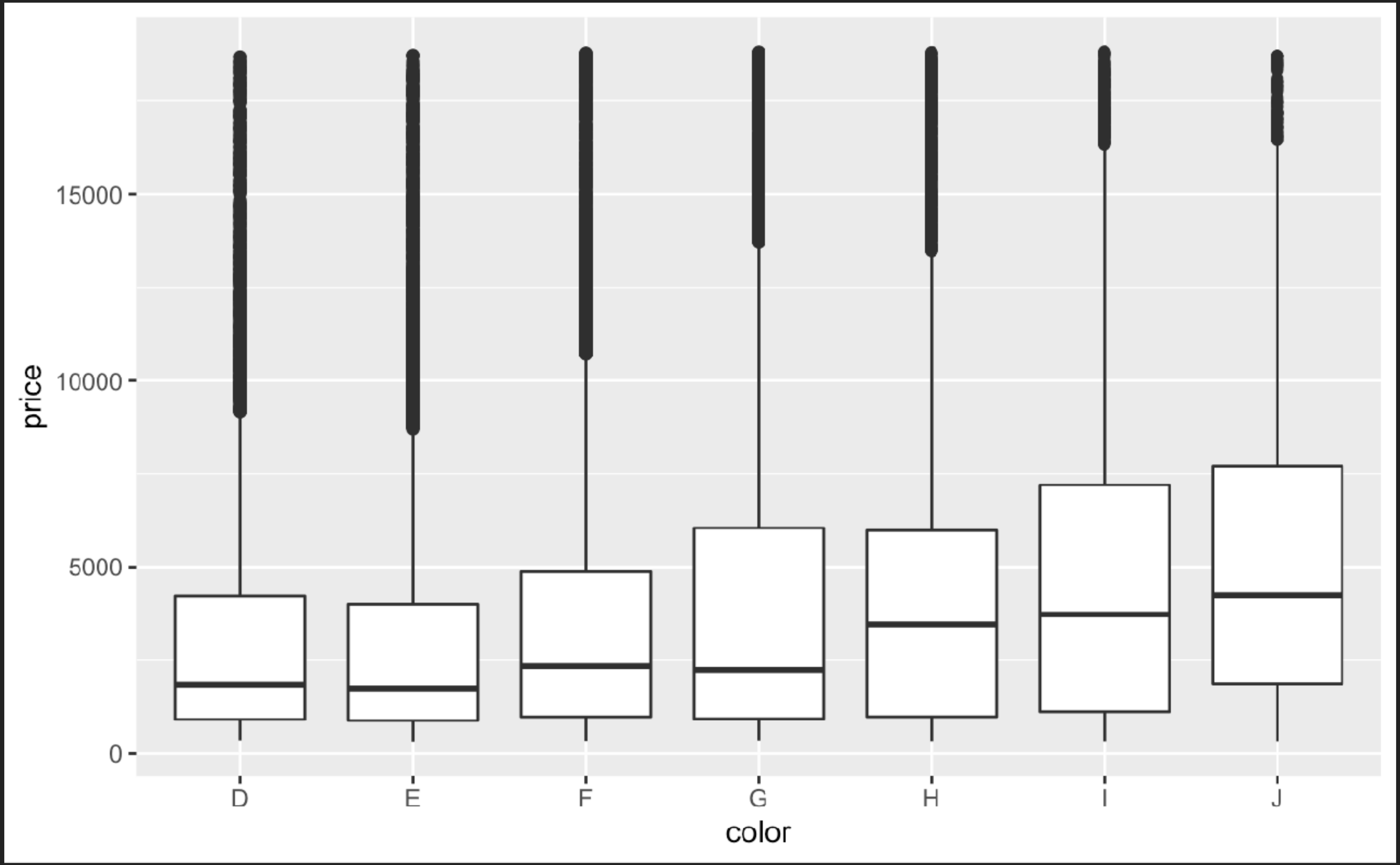
- Fair: worst cut

Why are low quality diamonds more expensive?



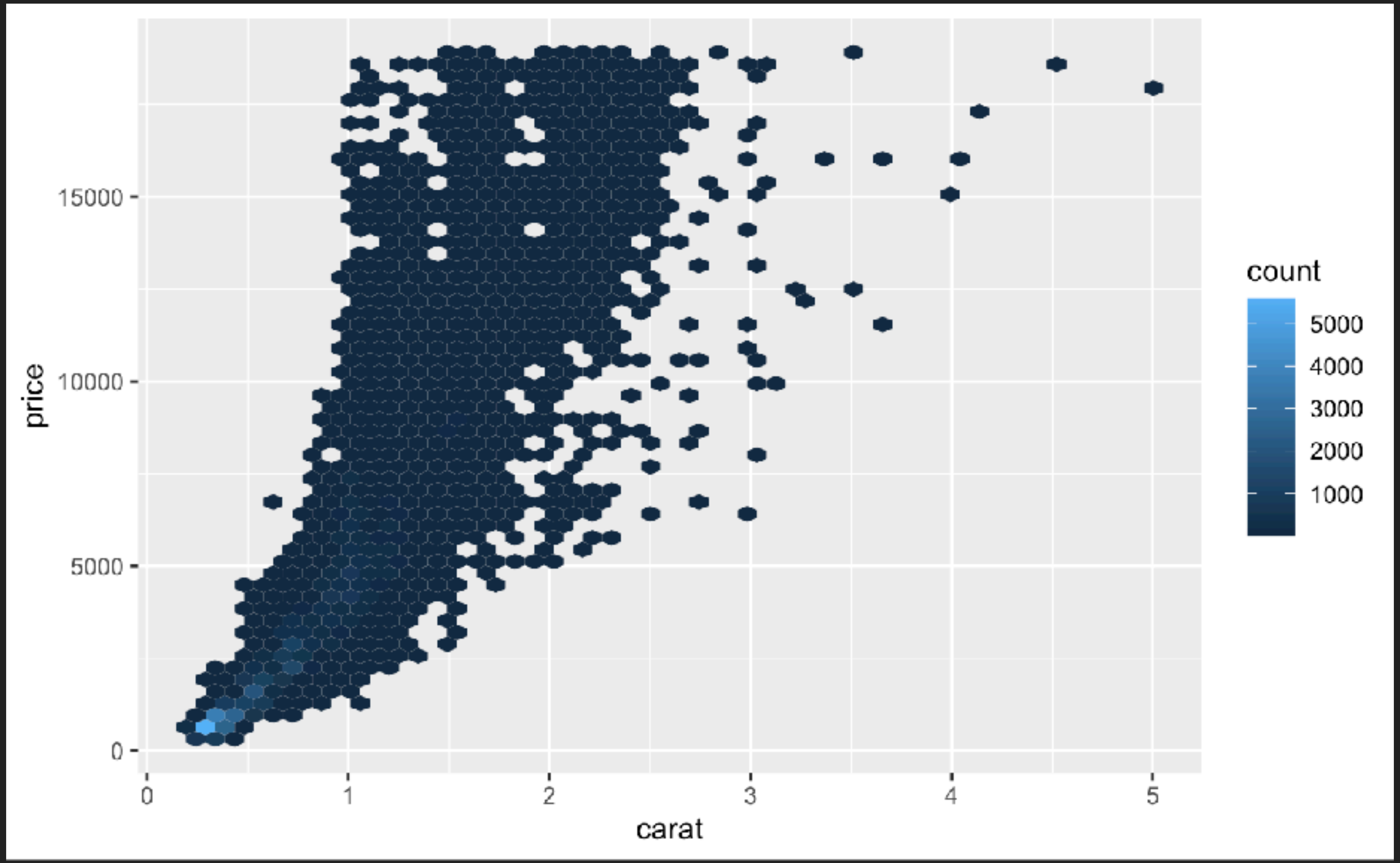
- I1: inclusions visible to the naked eye (worst clarity)

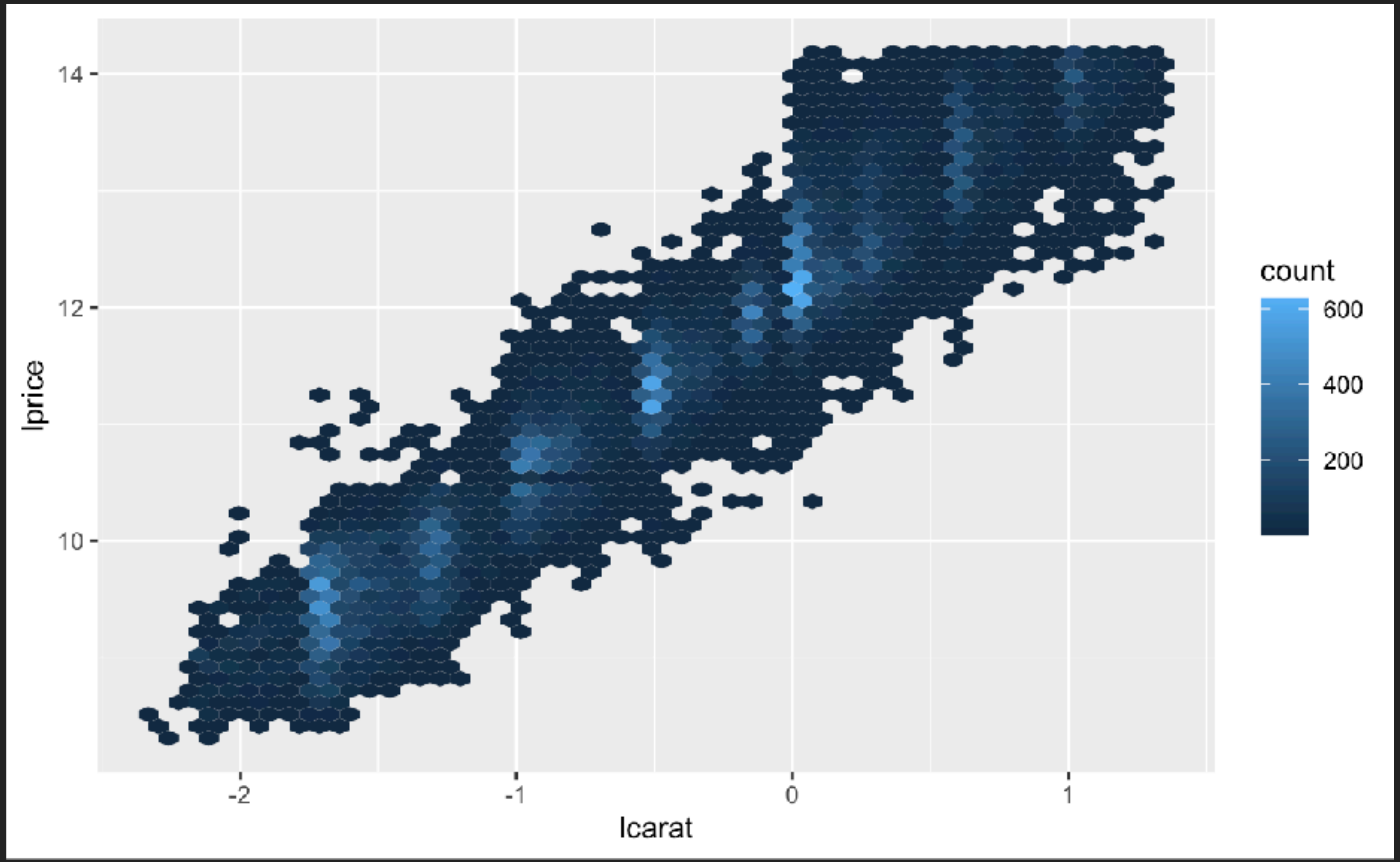
Why are low quality diamonds more expensive?



- J: slightly yellow (worst color)

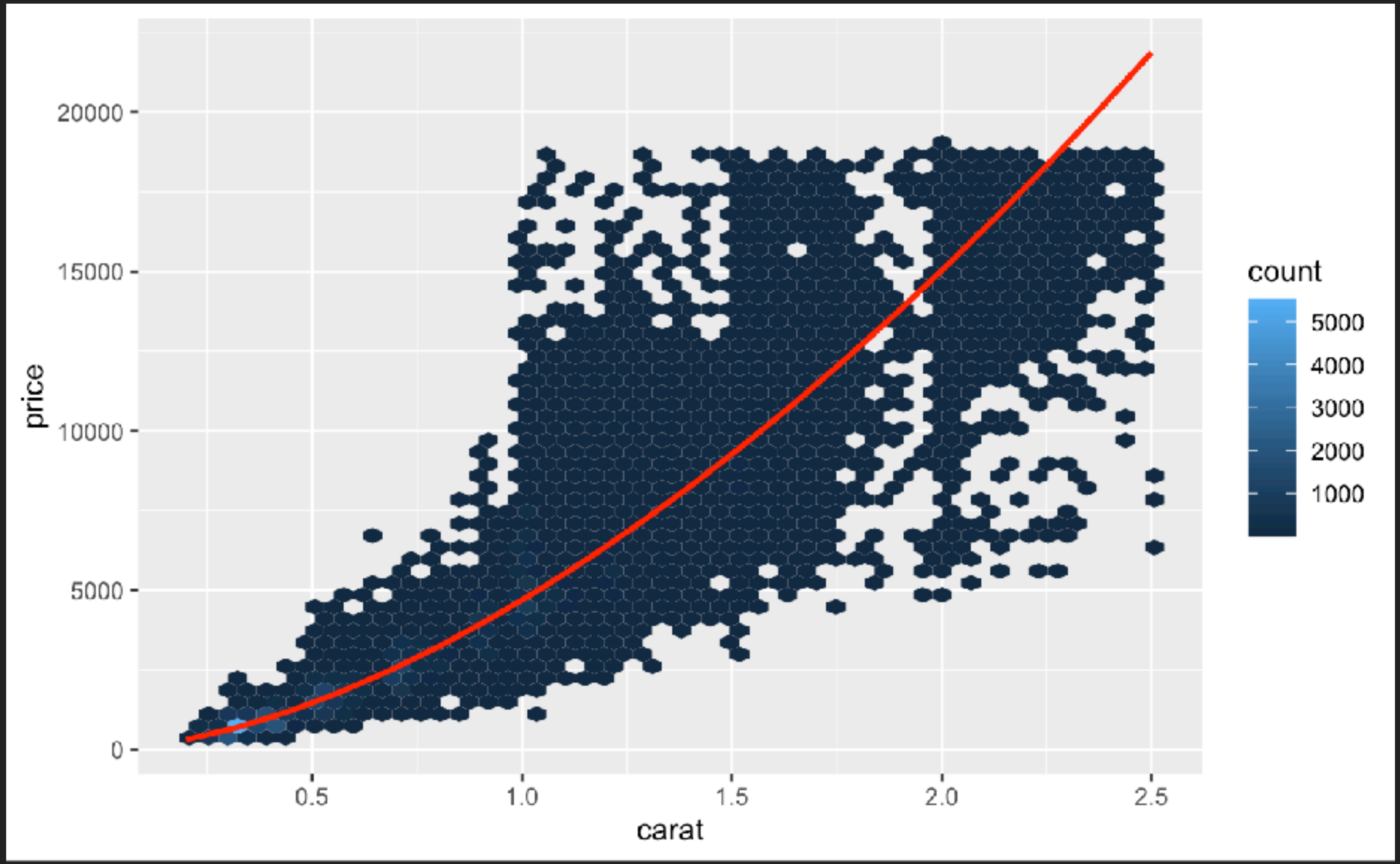
Because lower quality diamonds tend to be larger

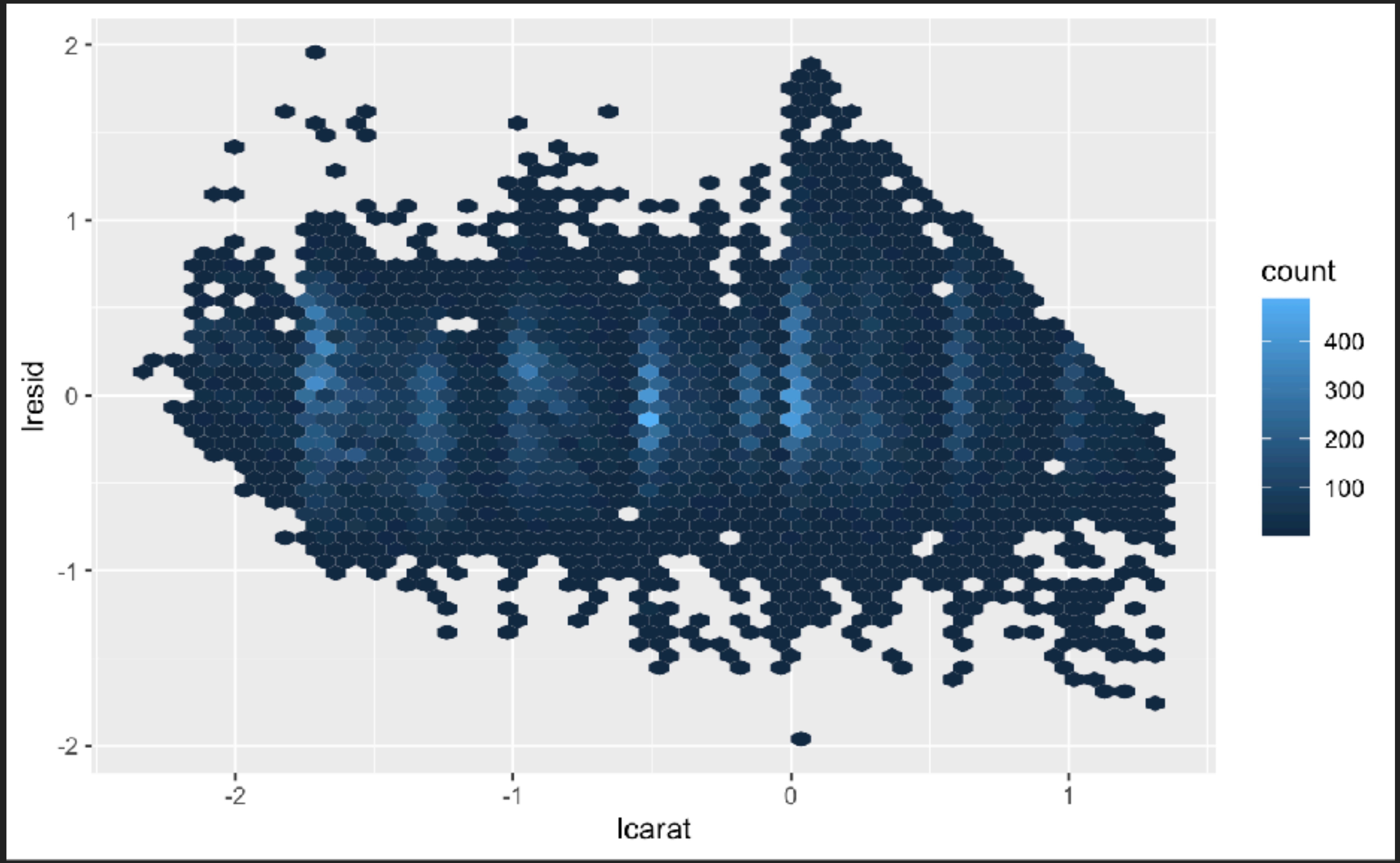




- The weight of the diamond is the single most important factor for determining the price of the diamond.
  - Left: raw
  - Right: log-transformed

Let's remove that strong linear pattern

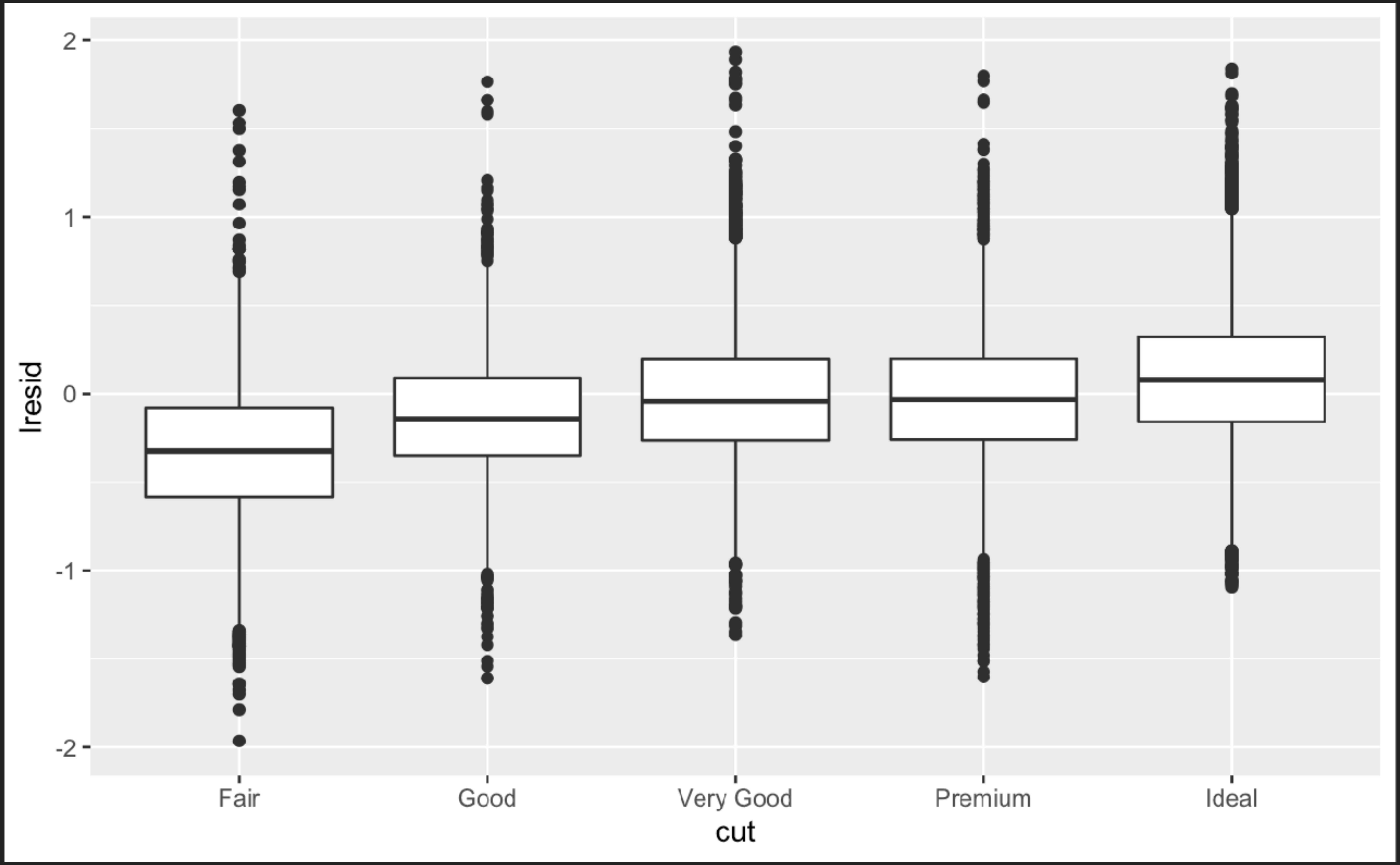





```
mod_diamond <- lm(lprice ~ lcarat, data = diamonds2)
```

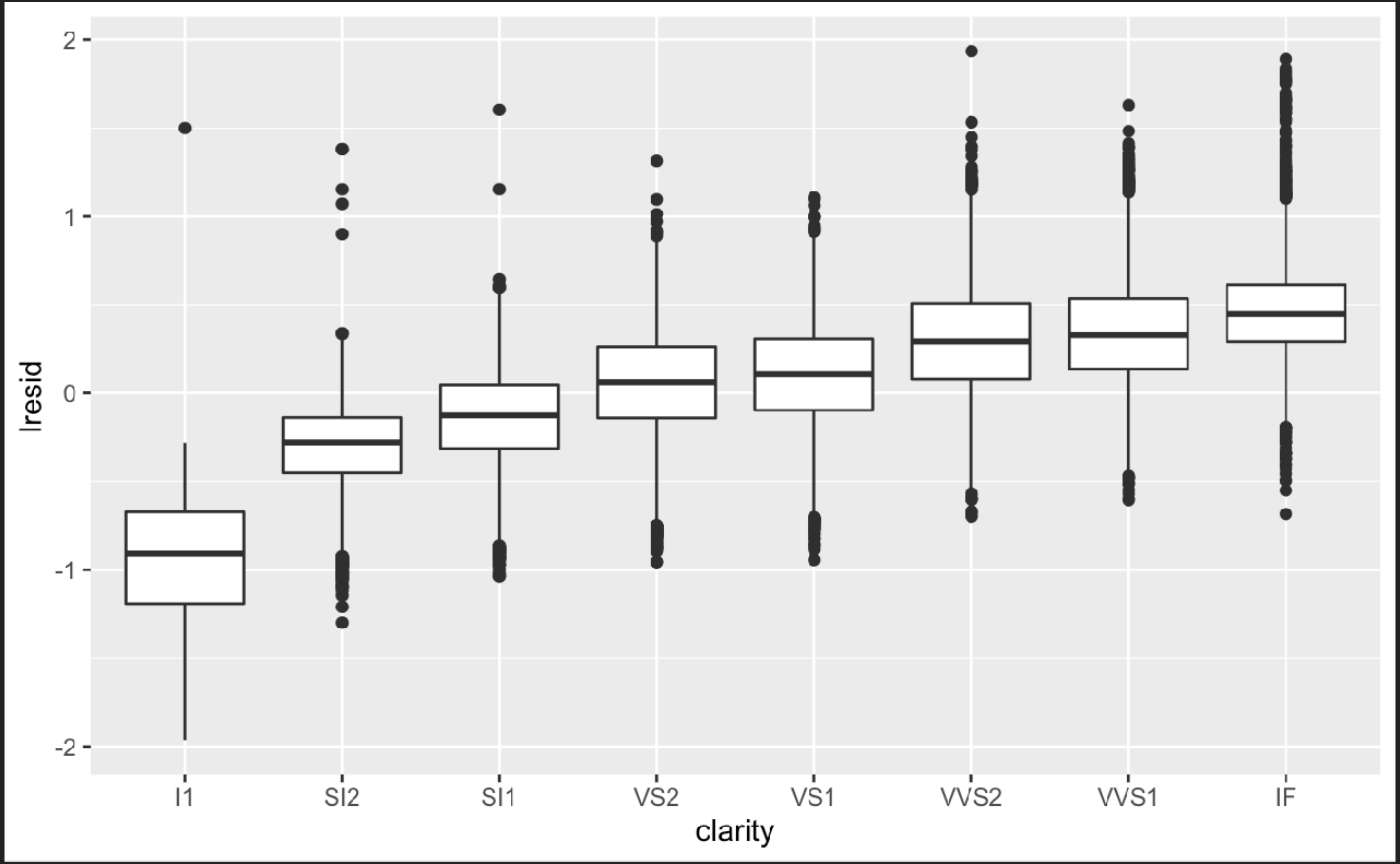
- Residuals confirm that we've successfully removed the strong linear pattern.

Now we see the relationship we expect



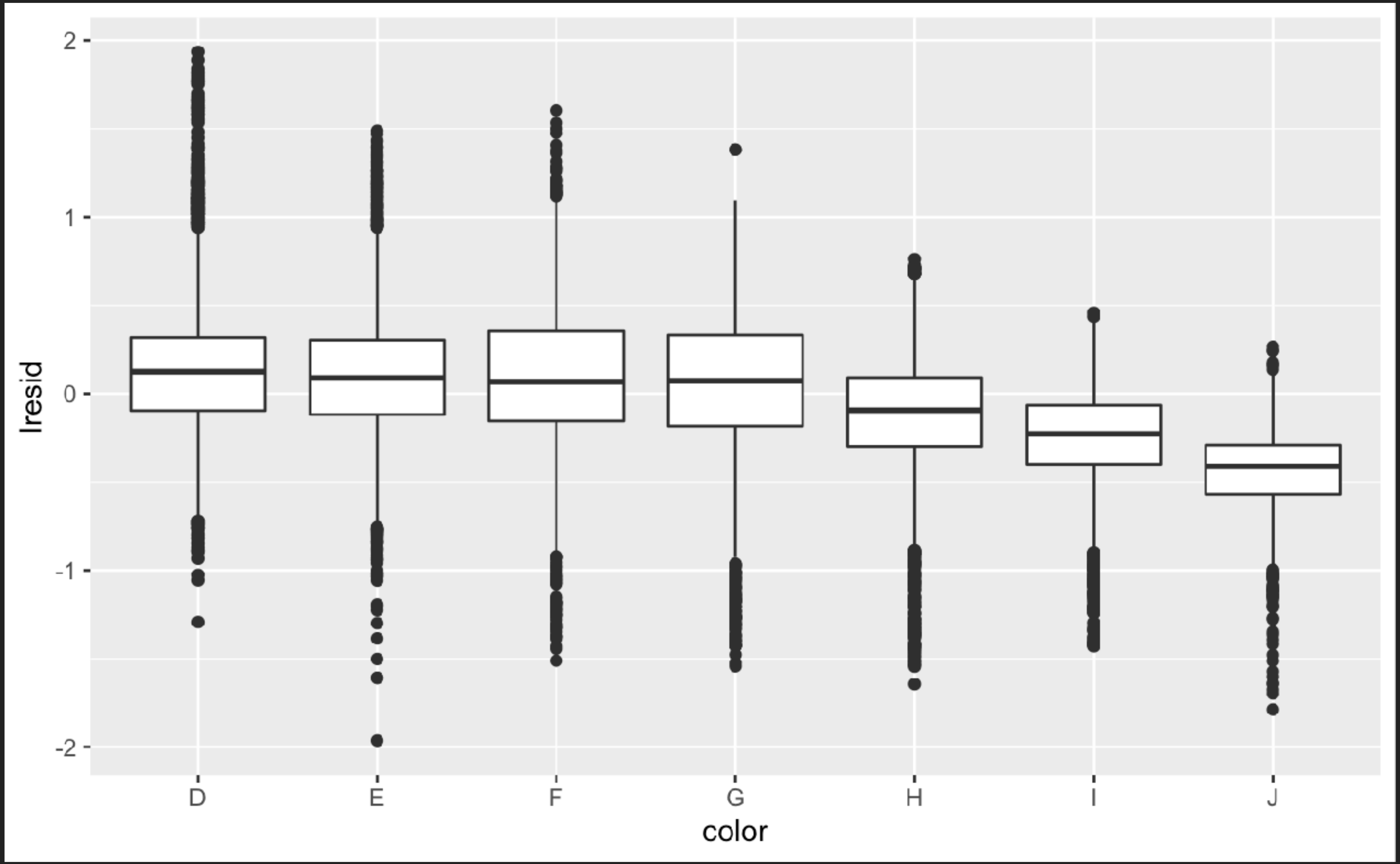
- Re-did our motivating plots using those residuals instead of price.
- Fair: worst cut

Now we see the relationship we expect



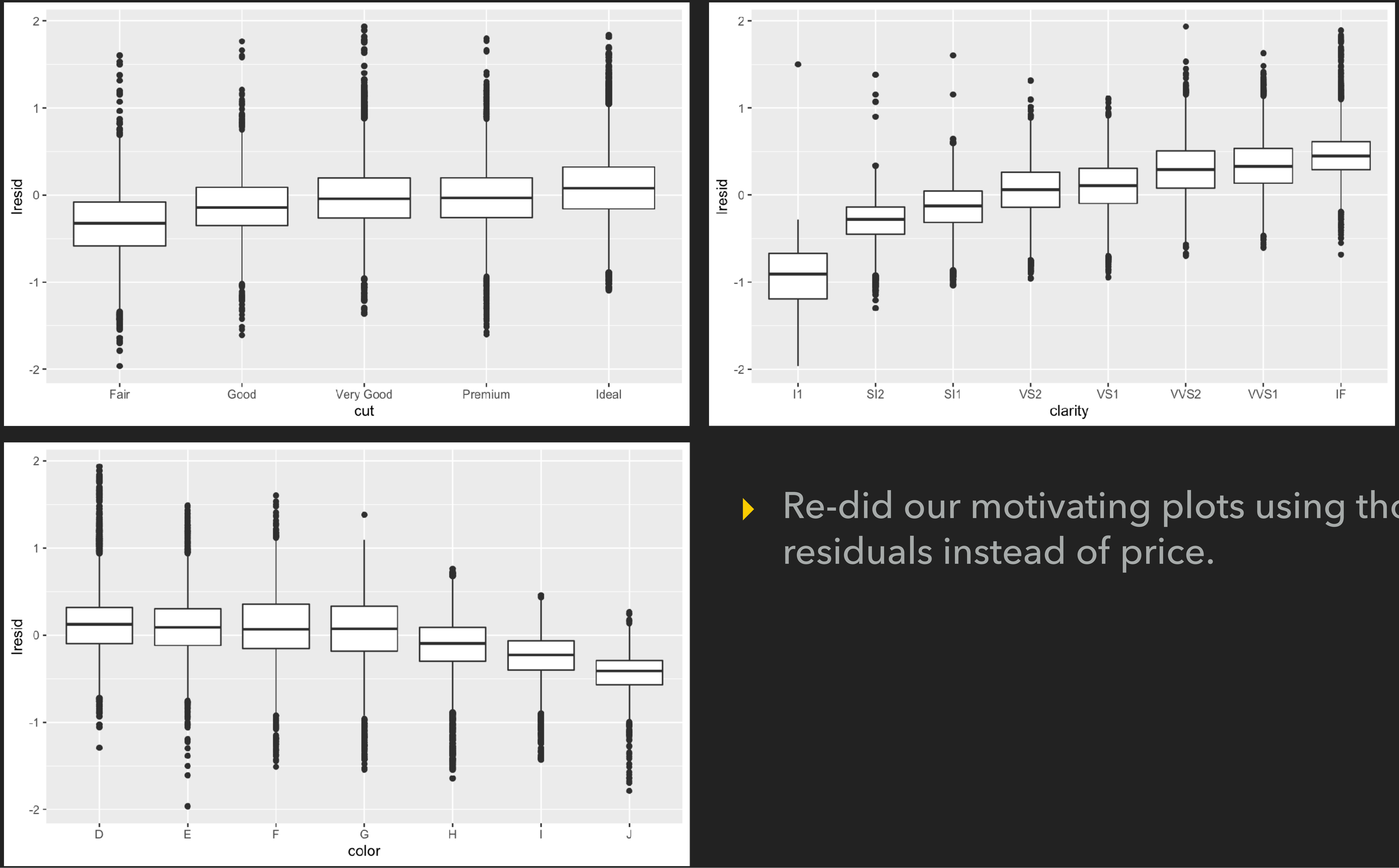
- I1: inclusions visible to the naked eye (worst clarity)

Now we see the relationship we expect



- J: slightly yellow (worst color)

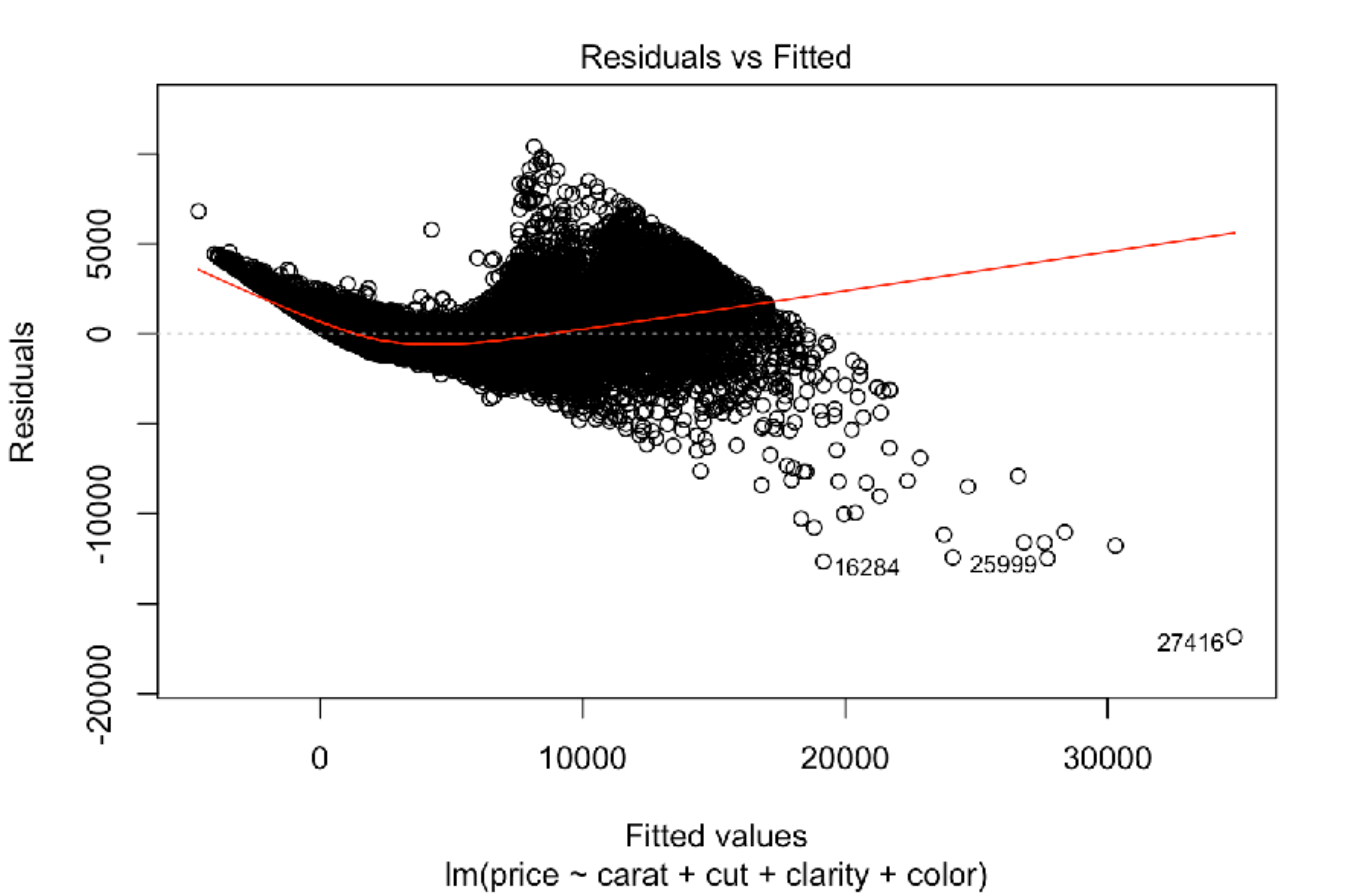
Now we see the relationship we expect



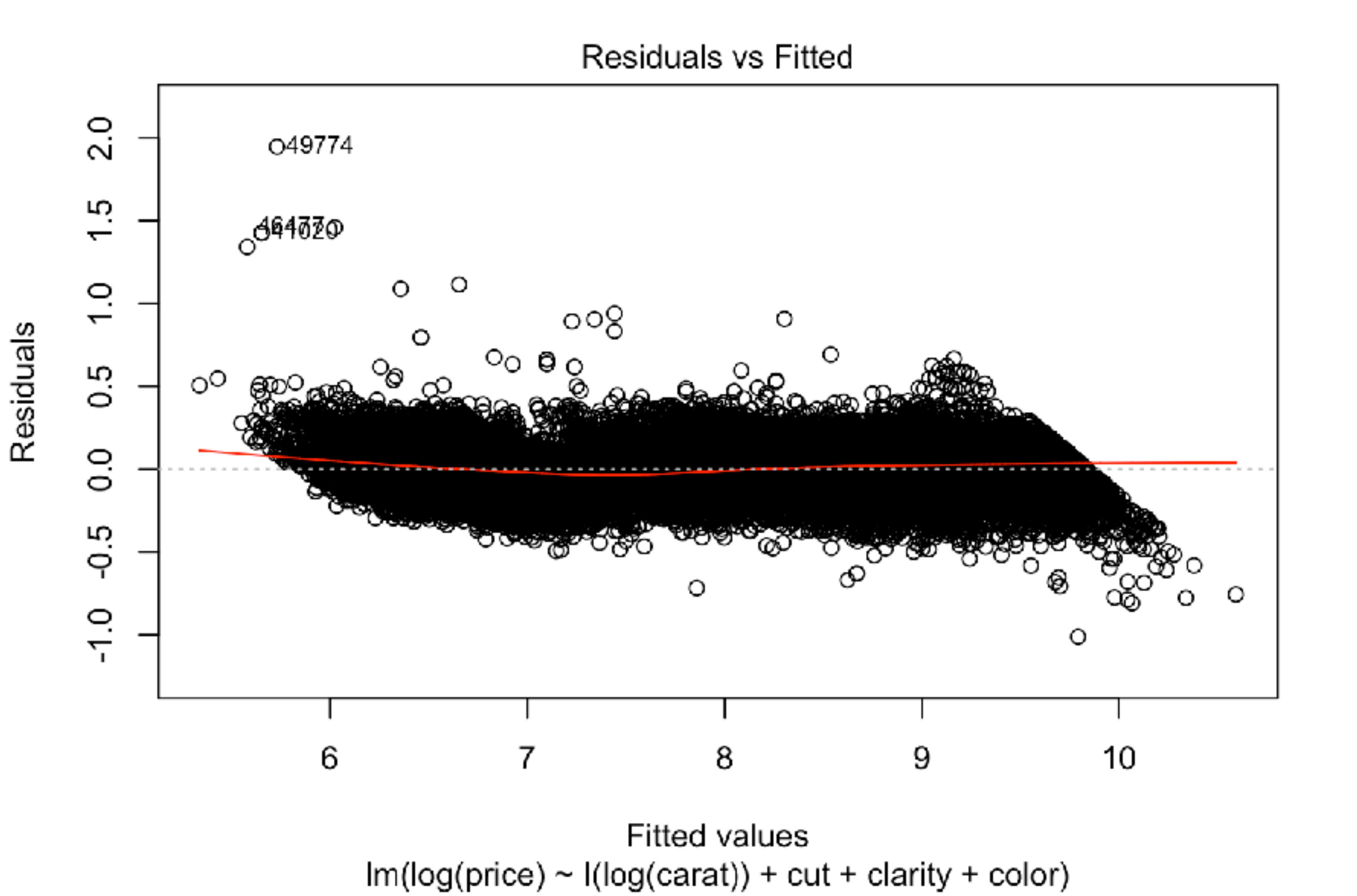
- Re-did our motivating plots using those residuals instead of price.

Regression Diagnostics on the Diamonds Example

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

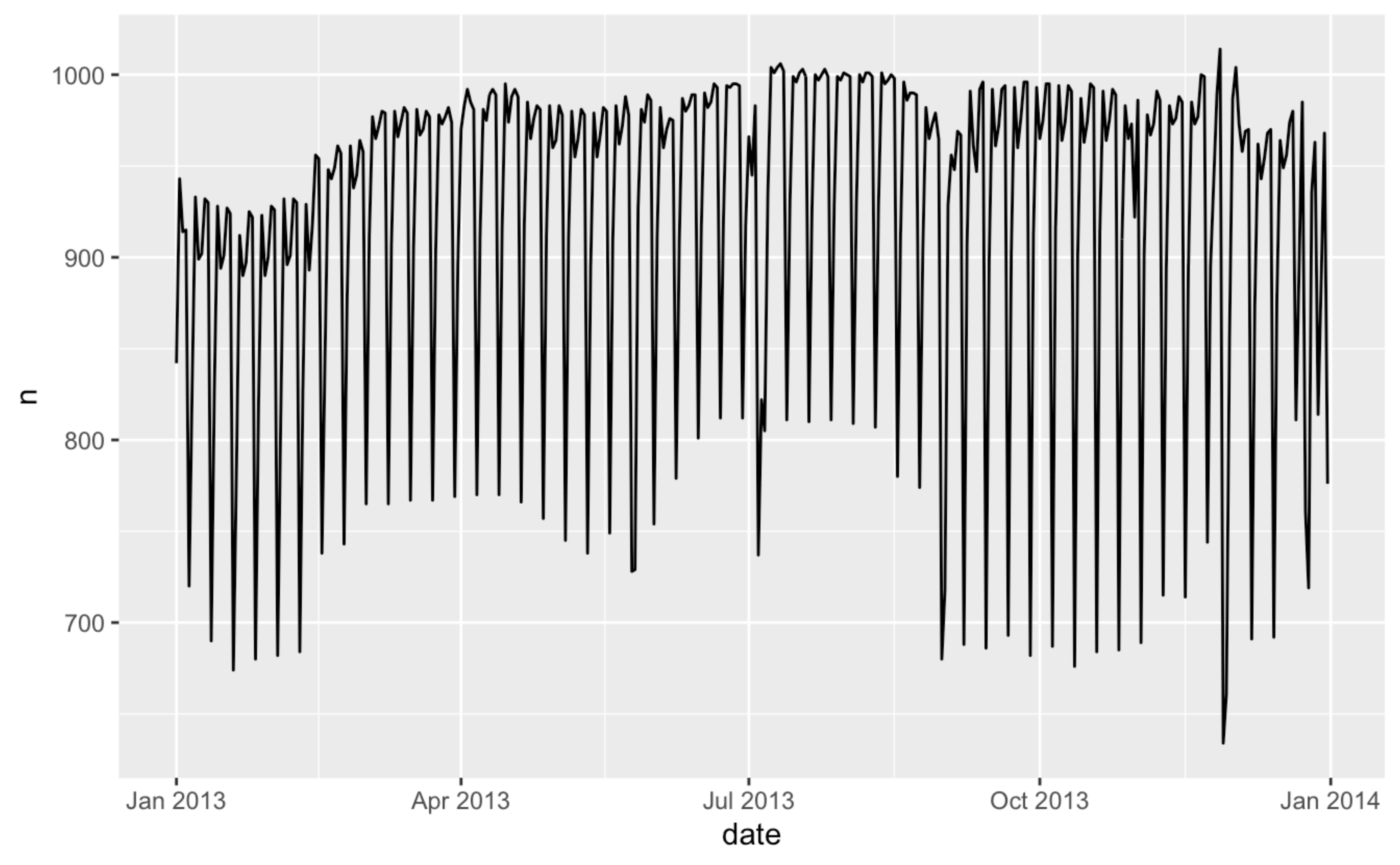



```
diamonds.lm2 <- lm(log(price) ~ I(log(carat))
                                + cut
                                + clarity
                                + color,
                                data = diamonds)
```

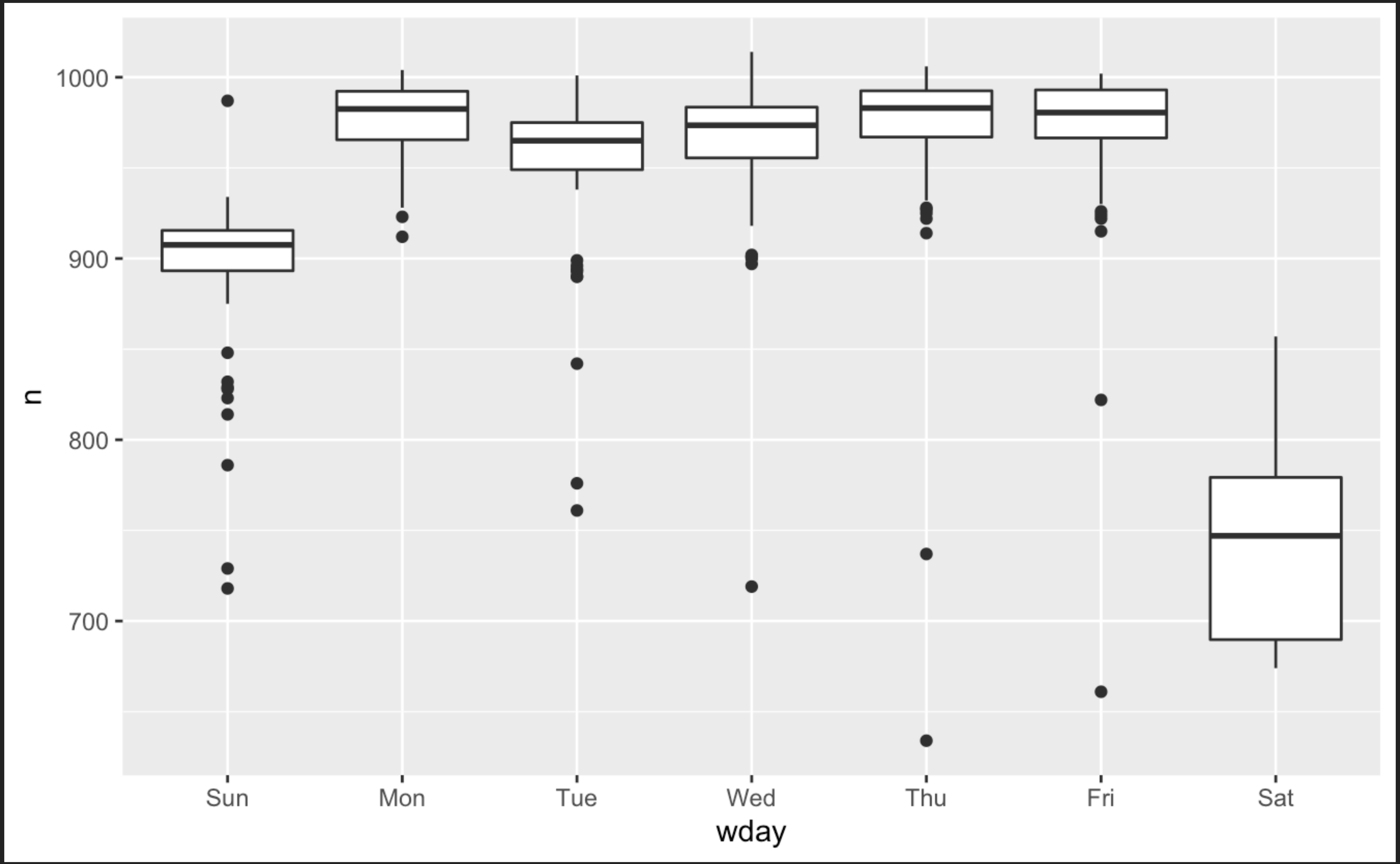


Another example

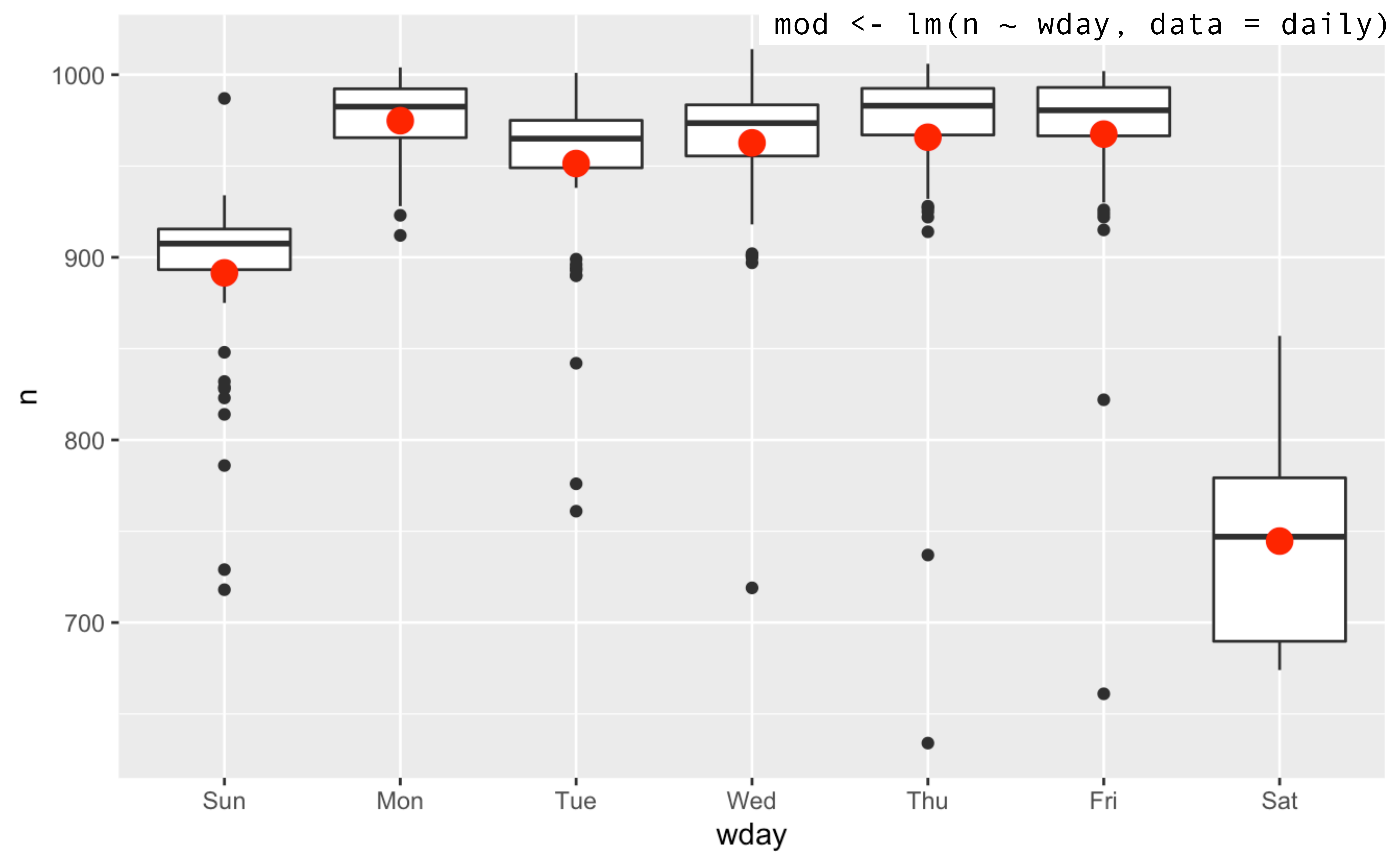
The number of flights that leave NYC per day



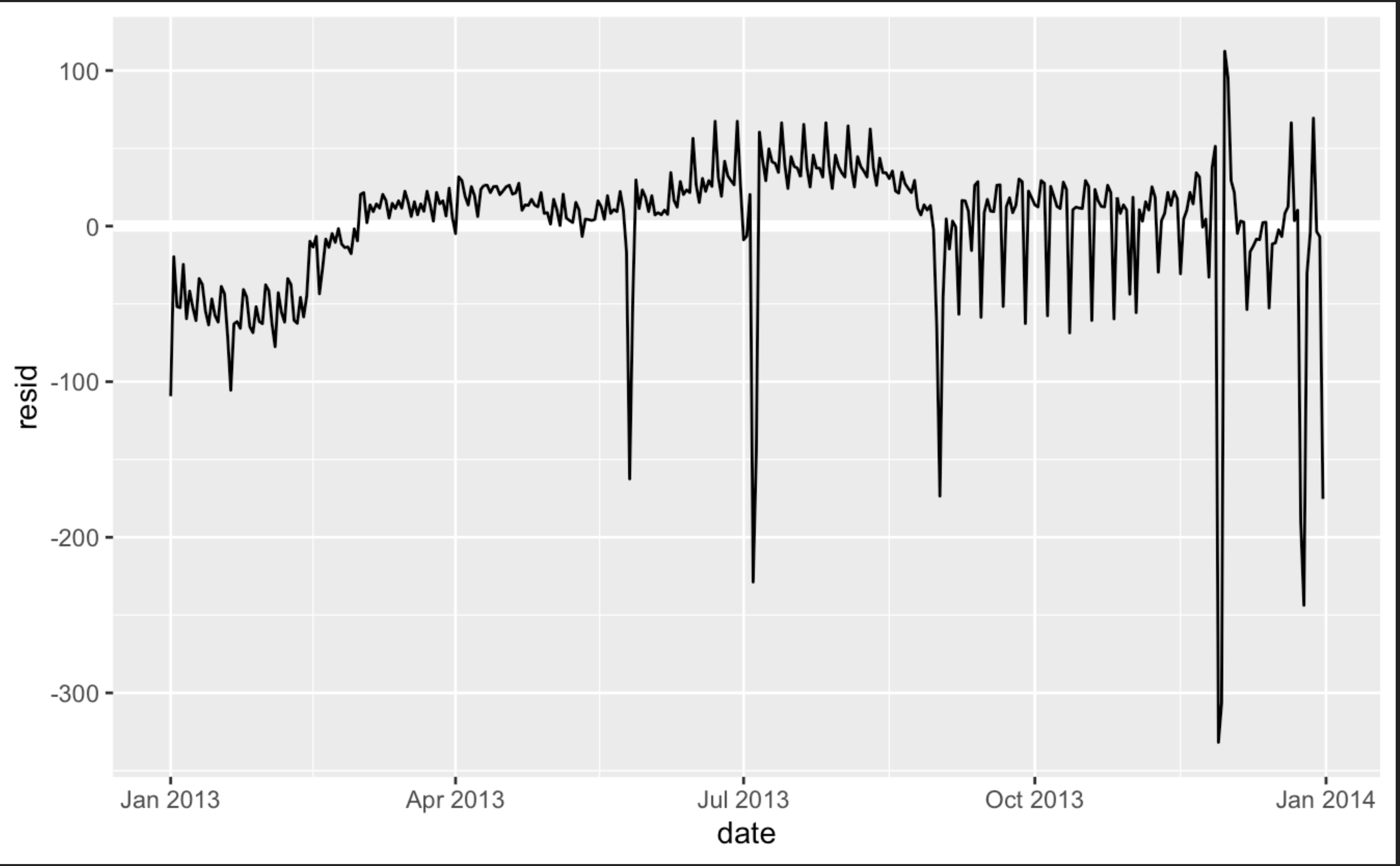
A very strong day-of-week effect dominates the subtler patterns



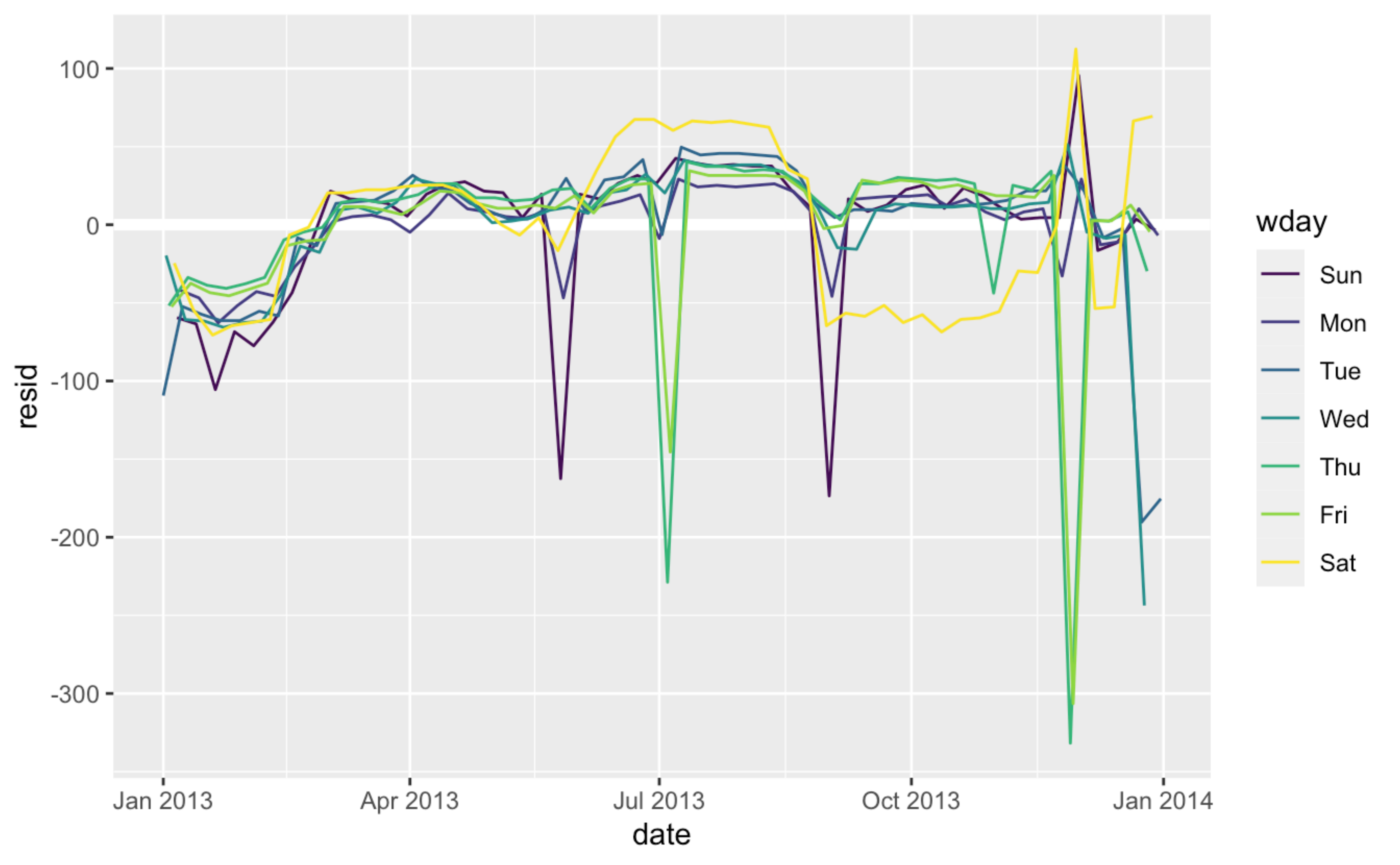
Modeling the week day effect



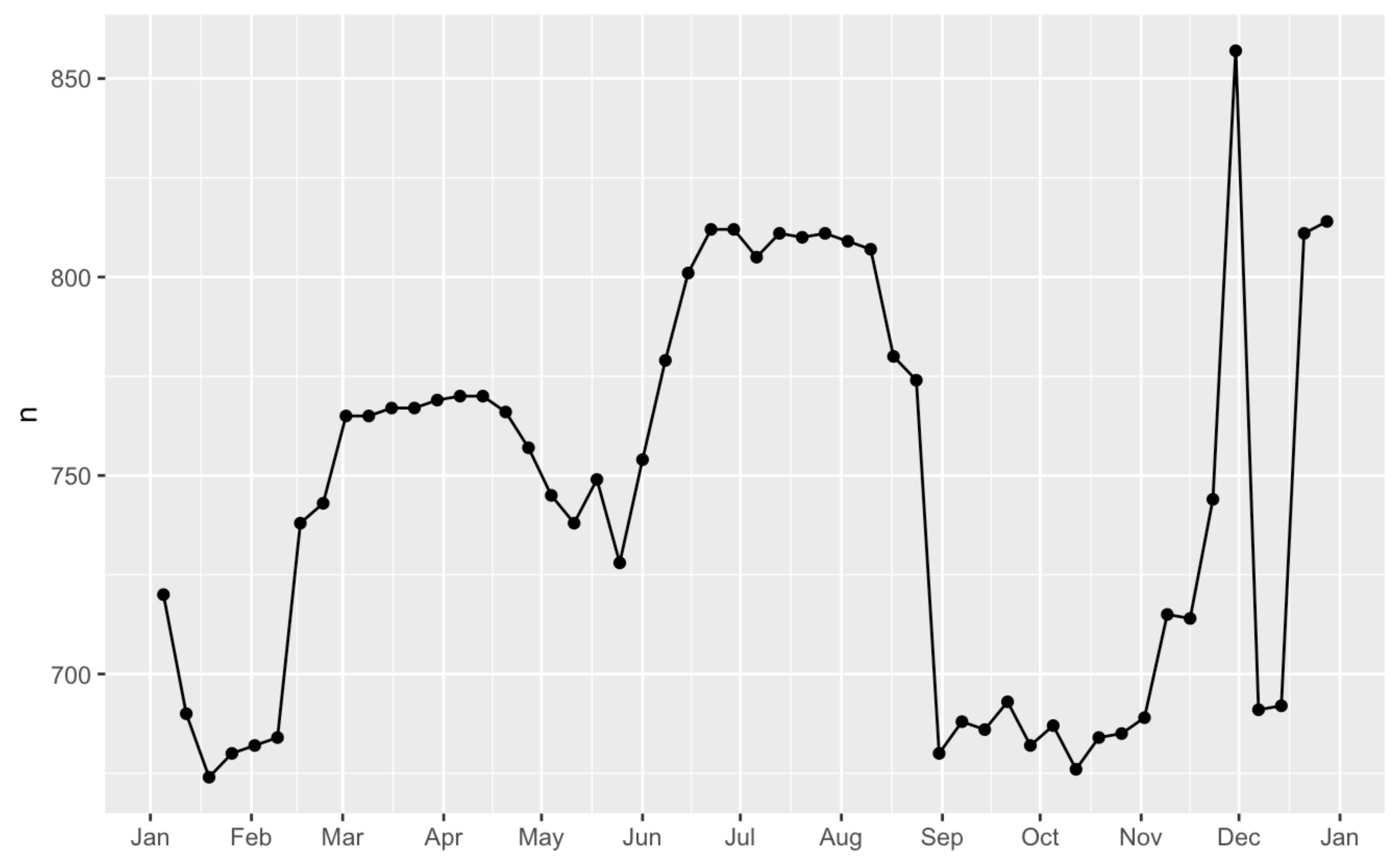
Visualizing the residuals



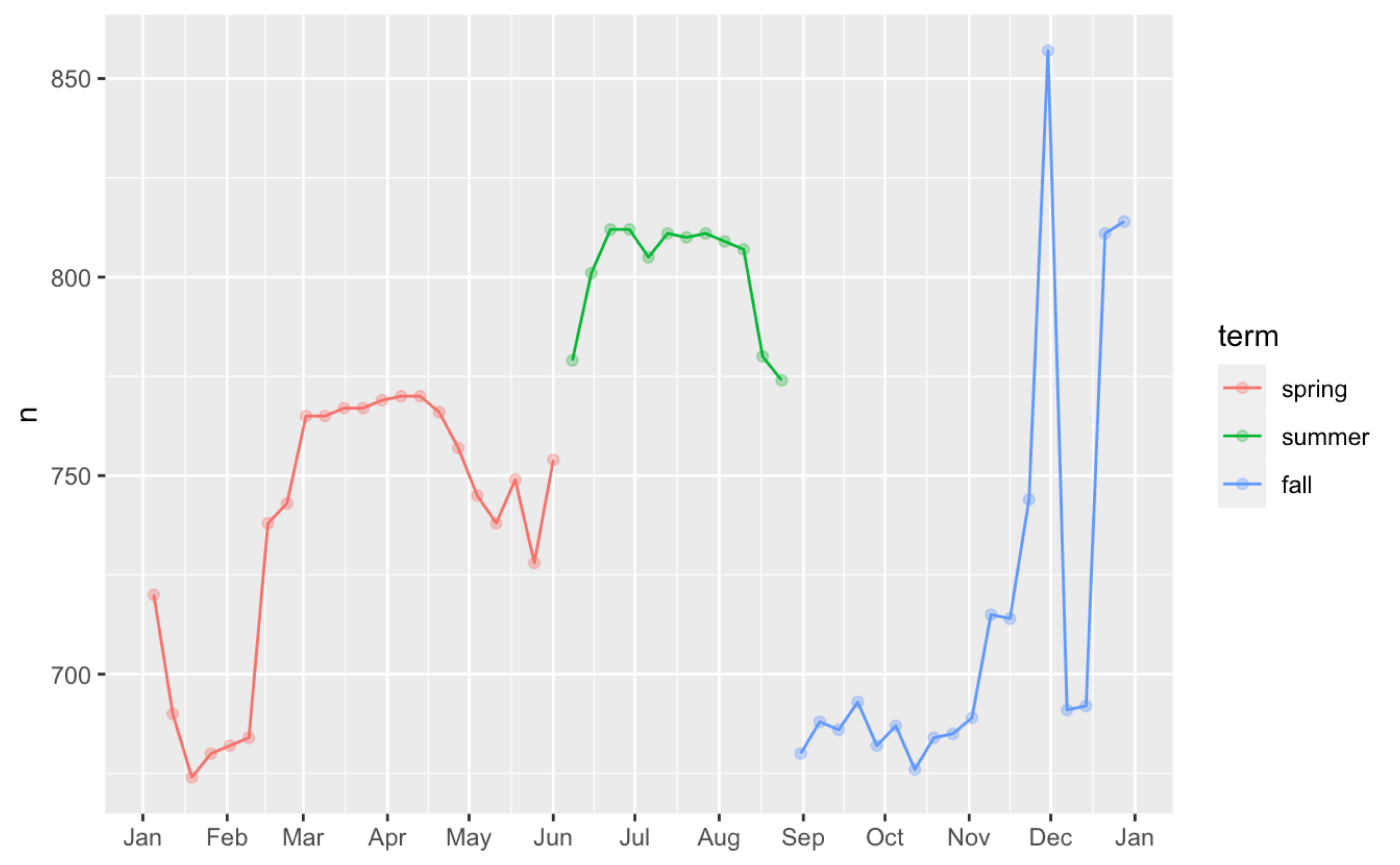
Our model seems to fail starting in June



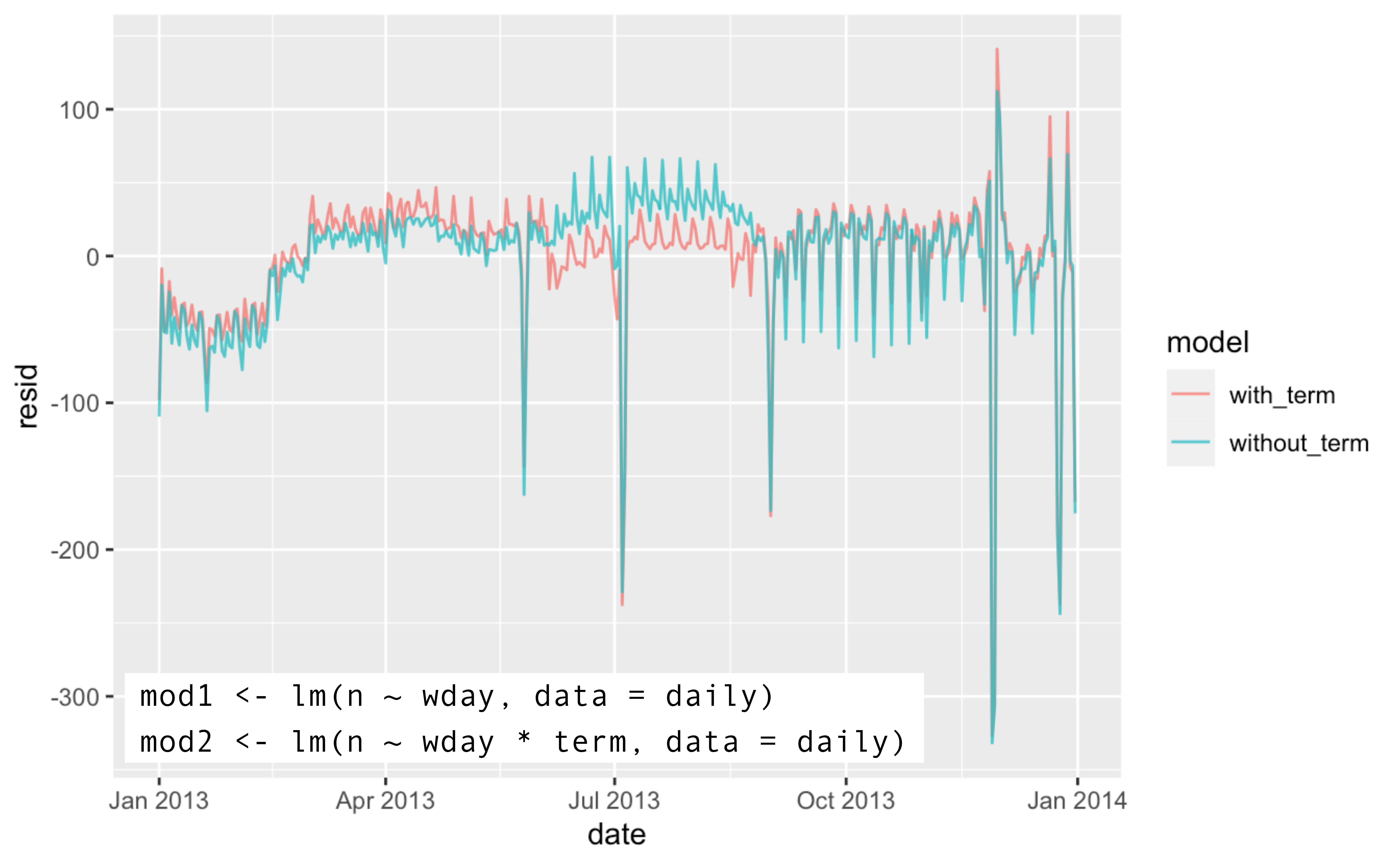
The model fails to accurately predict the number of flights on Saturday



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.