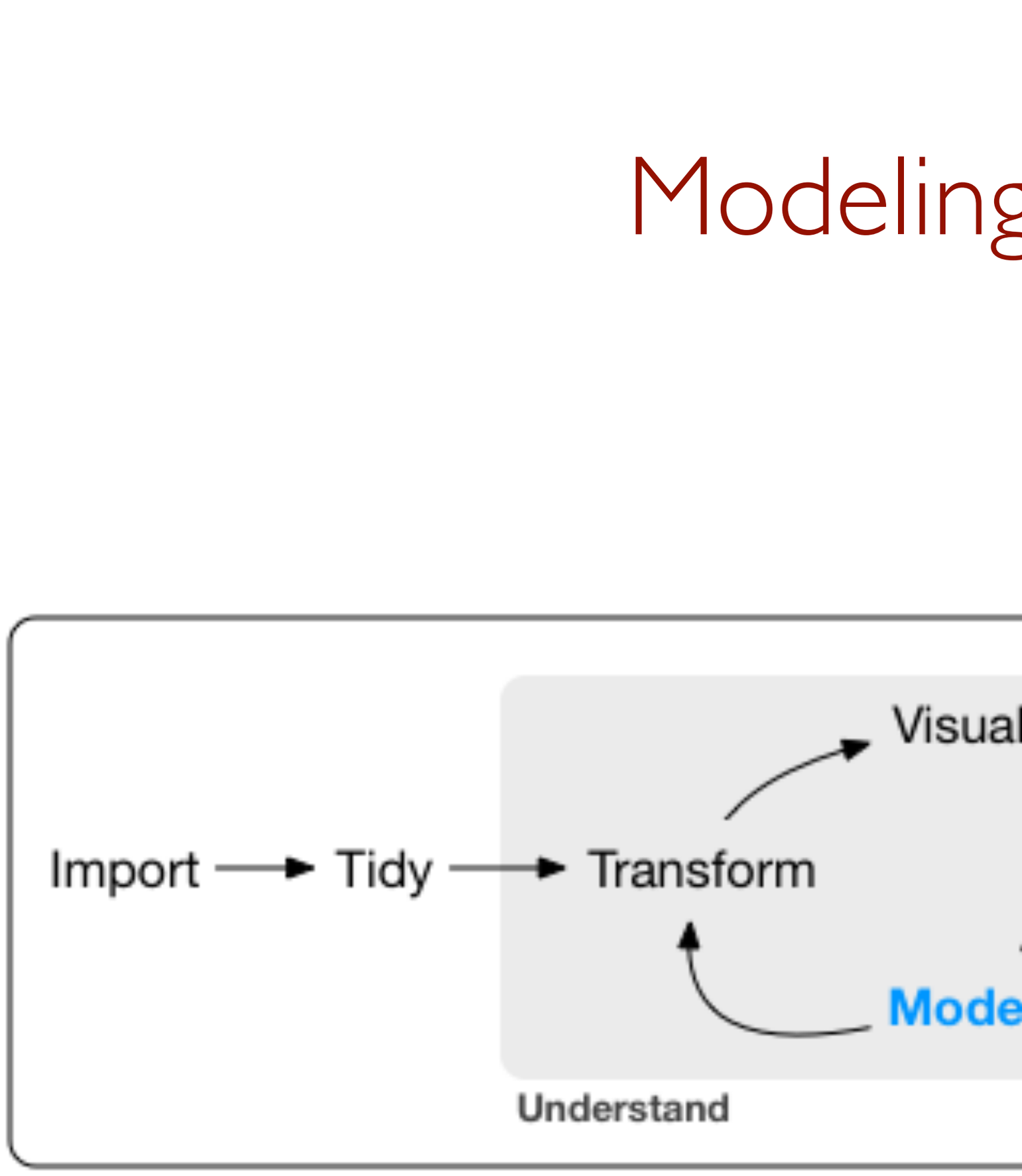
Data Model
and Linear Regr

Kylie A. Bem

Northeastern Uni
Khoury College of Comput

Northeastern U

Modeling



R for Data Science, Wickham an

Learning go

- What are the goals of r
- Why linear regression?
- Fitting linear models
- Model diagnostics

DATA MOD

Modeling go

- Predict new values
  - Regression — continuous value
  - Classification — categorical val
- Discover hidden structu
  - Cluster into groups based on s
- Interpret and understar
  - Statistical inference

Regressio

- Given predictors x_1, x_2, x_3…
- Predict continuous-valued response variable y
- Supervised: requires known response values

Classificatio

- Given predictors x_1, x_2, x_3…
- Predict categorical-valued response variable y
- Supervised: requires known category labels

Clustering

- Given features $x_1, x_2, x_3 \ldots$
- Group similar data points together into clusters
- Unsupervised: expects unlabeled data

Statistical infer

- Given sample data $x_1, x_2, x_3 \ldots$
- Infer properties of the underlying population
- Confidence intervals
- Hypothesis tests

Machine learning &

Statistics and machine learning share many n
and some models (but not all) can be u

- Supervised learning
  - Regression
  - Classification
- Unsupervised learnin
  - Clustering
  - Dimension reduction
- Statistical inference

"All models are wrong, but
— George E

- No mathematical mode
representation of reality
- But some models are m
illuminating than others
- Choice of model should
  - What are the overall goals of t
  - How useful is the model for th

LINEAR REGR

Linear regres

- Predicts a scalar continu *(dependent variable)*
- Linear function of one o *(explanatory/independer*
- Flexible model with ma
  - Linear constraint is rarely a pro
  - Useful for statistical inference: C

Linear regres

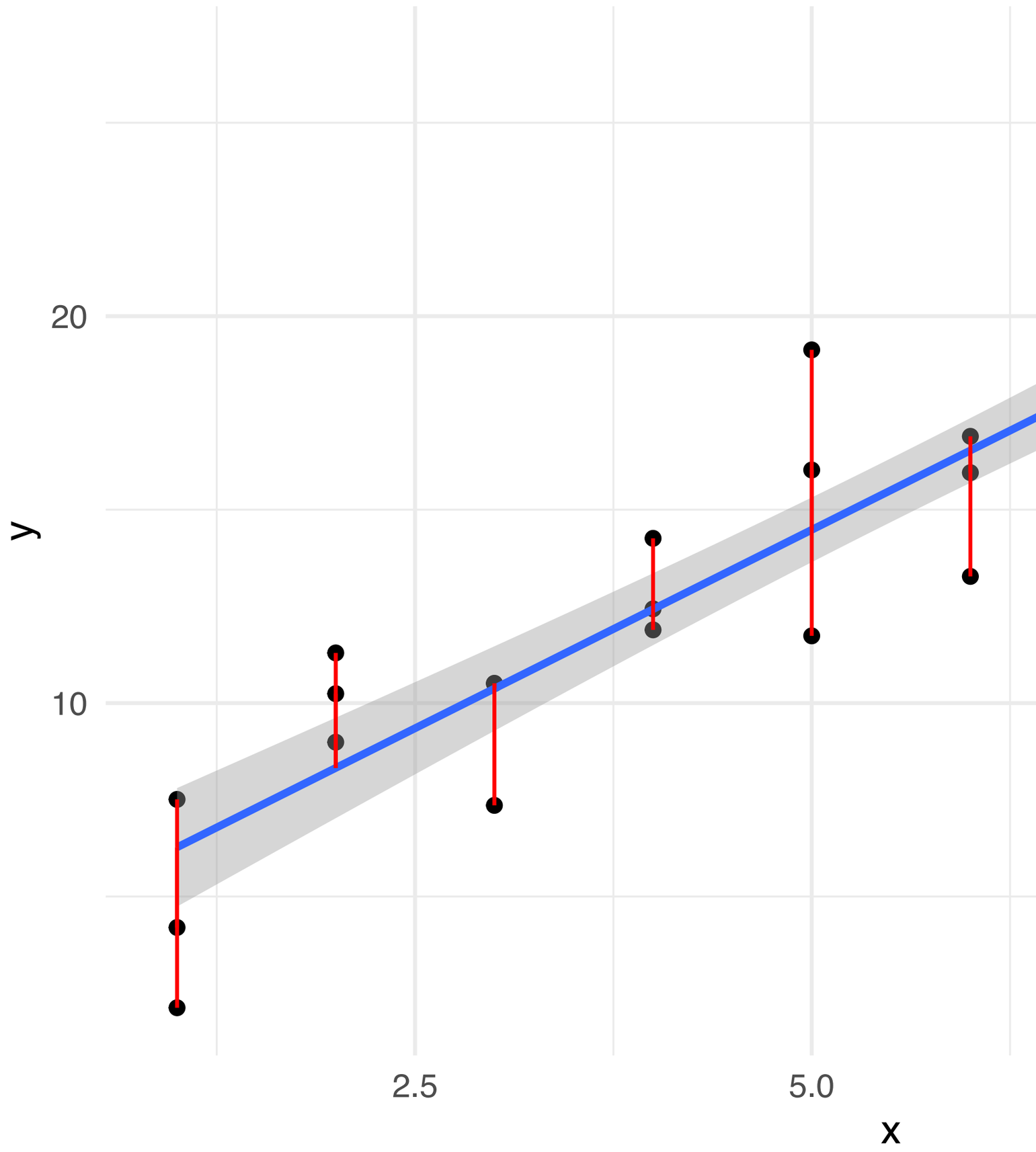
- Parametric model
  - Create a low-dimensional sum
  - Forms a constraint on possible
- Linear model form:

$$y = {}_0 + \quad x + \ldots +$$

$$error \sim N(0,$$

- Fit the model through c
  - Maximize log-likelihood (statist
  - Minimize squared errors (loss f

Fitting the regressi



Find parameters (β's) that minimiz

Assumptions of linea

- Linearity
  - The relationship must be linear
  - Transformations can help make
- Constant variance (hom
- Independent errors
  - Especially important for data co
- Lack of perfect multicol
  - Predictors can't be (too) corre

Modeling real

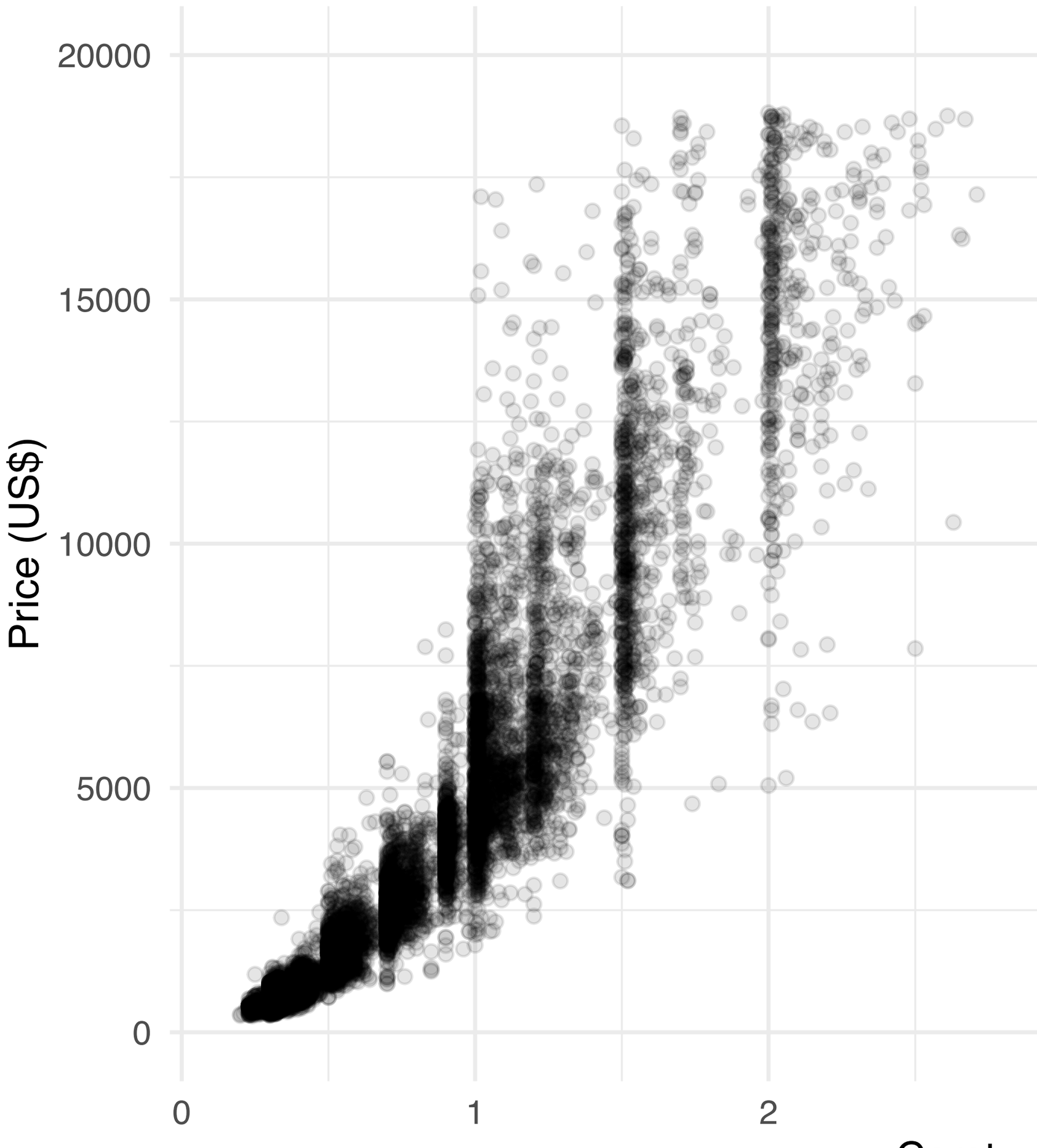
Prices and measurements of

```
diamonds
```

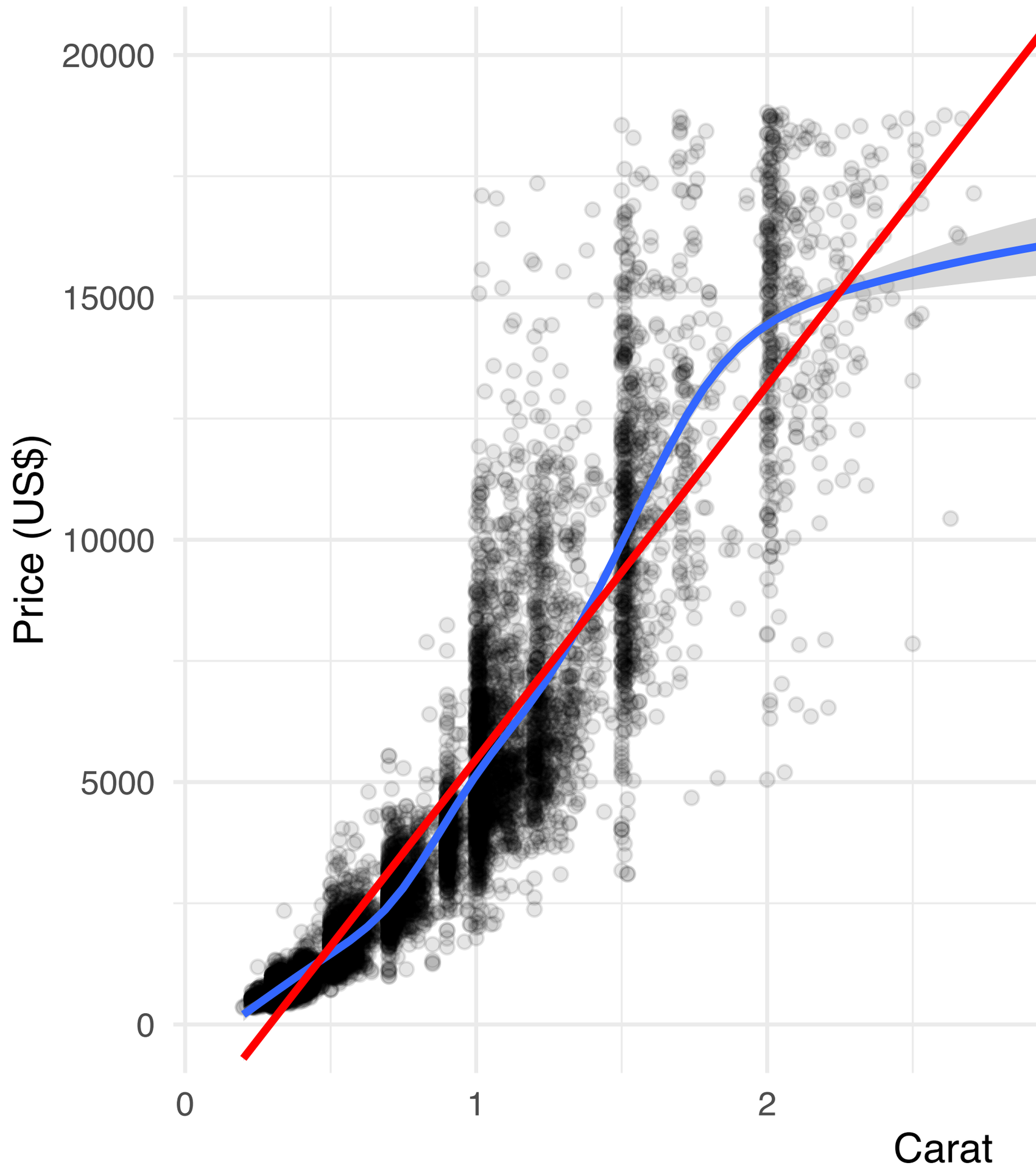
```
## #   tibble: 53,940 x 10
##     carat cut       color clarity depth t
##     <dbl> <ord>     <ord> <ord>   <dbl> <
##  1 0.23  Ideal     E     SI2      61.5
##  2 0.21  Premium   E     SI1      59.8
##  3 0.23  Good      E     VS1      56.9
##  4 0.290 Premium   I     VS2      62.4
##  5 0.31  Good      J     SI2      63.3
##  6 0.24  Very Good J     VVS2     62.8
##  7 0.24  Very Good I     VVS1     62.3
##  8 0.26  Very Good H     SI1      61.9
##  9 0.22  Fair      E     VS2      65.1
## 10 0.23  Very Good H     VS1      59.4
## # ... with 53,930 more rows
```

Available in `ggplot2` p

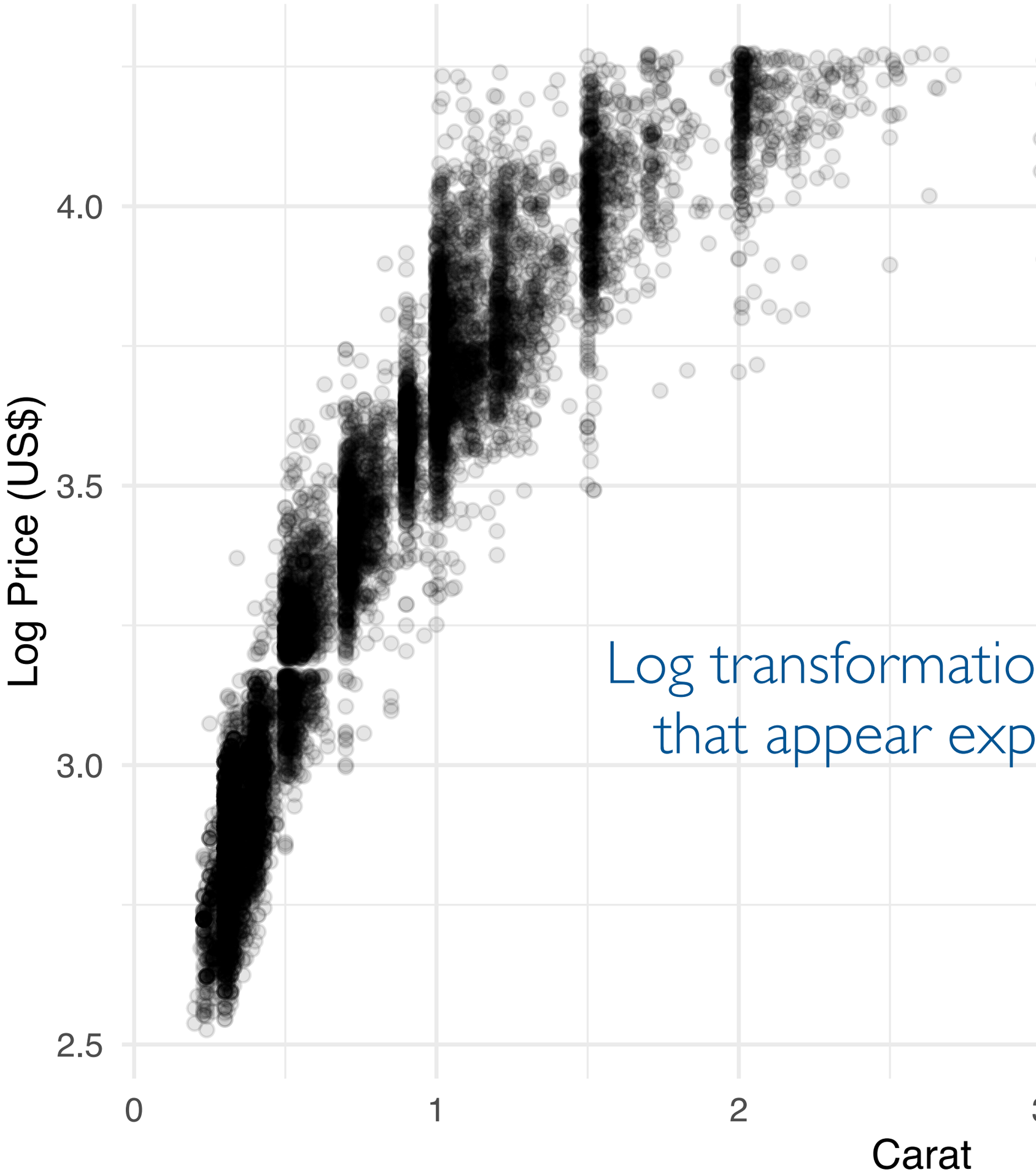
Does diamond size e



Is it linear

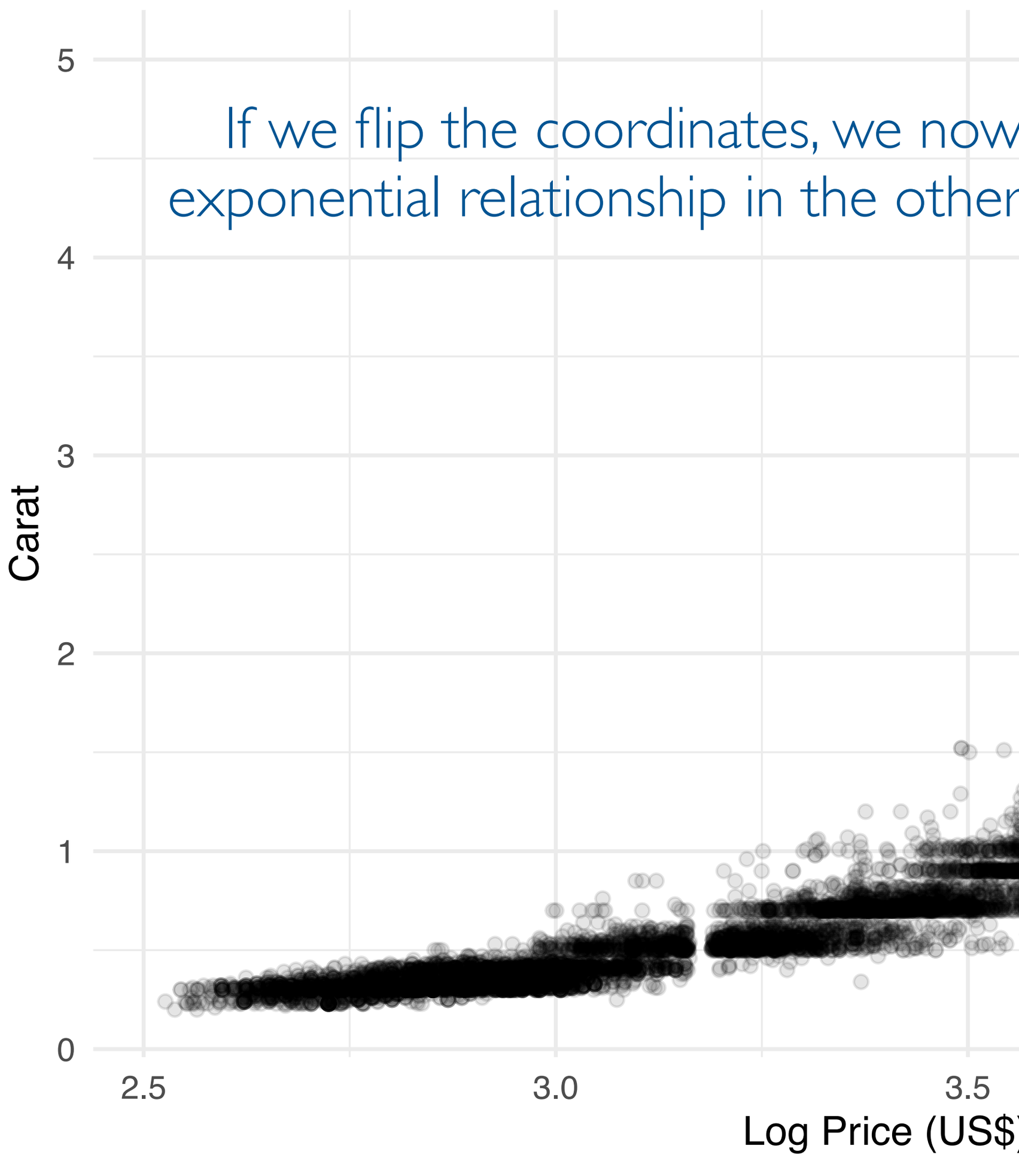


Log-transform



A different pers

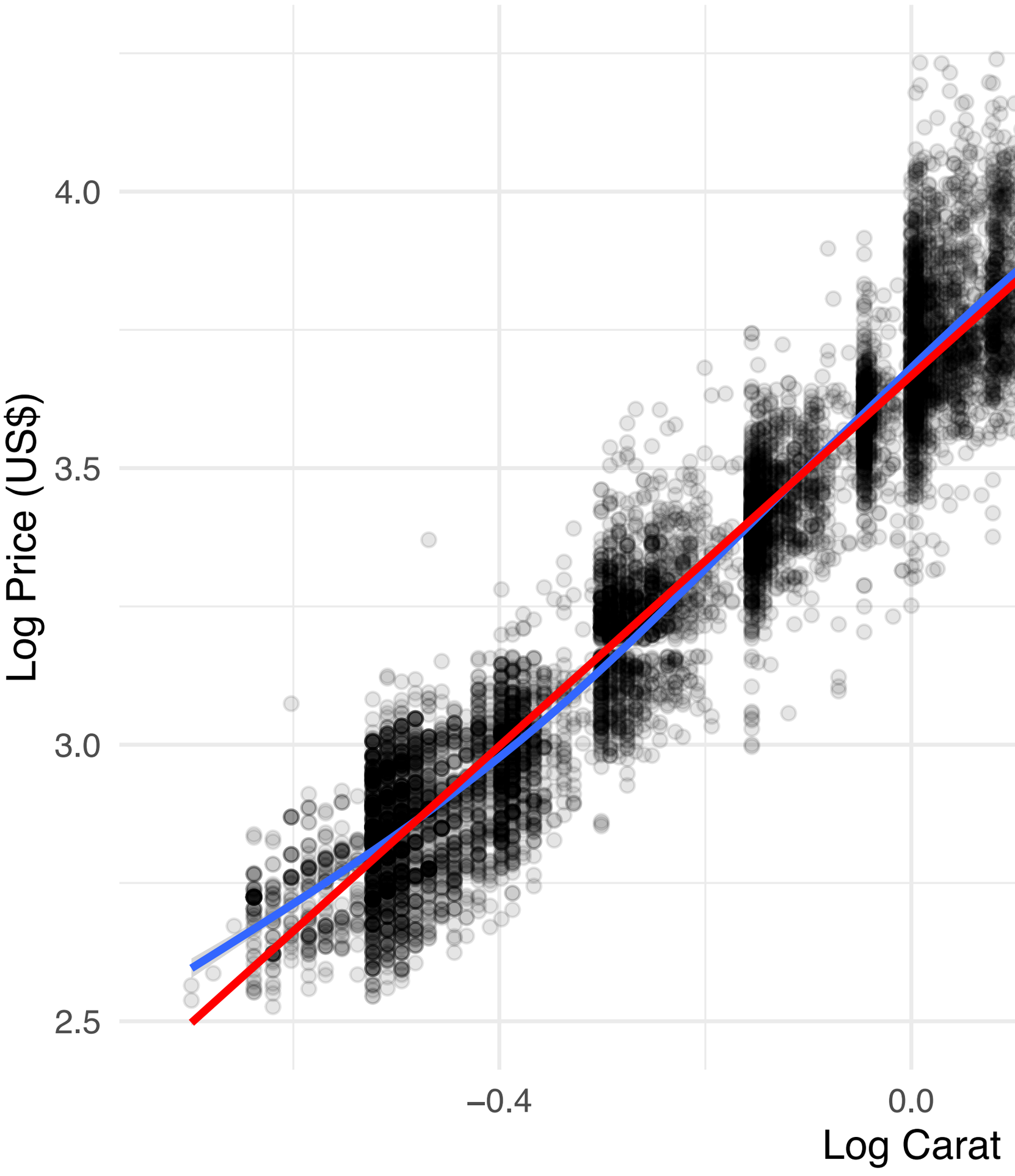
If we flip the coordinates, we now exponential relationship in the other



Log-log transfor

L
appe

Is it linear no



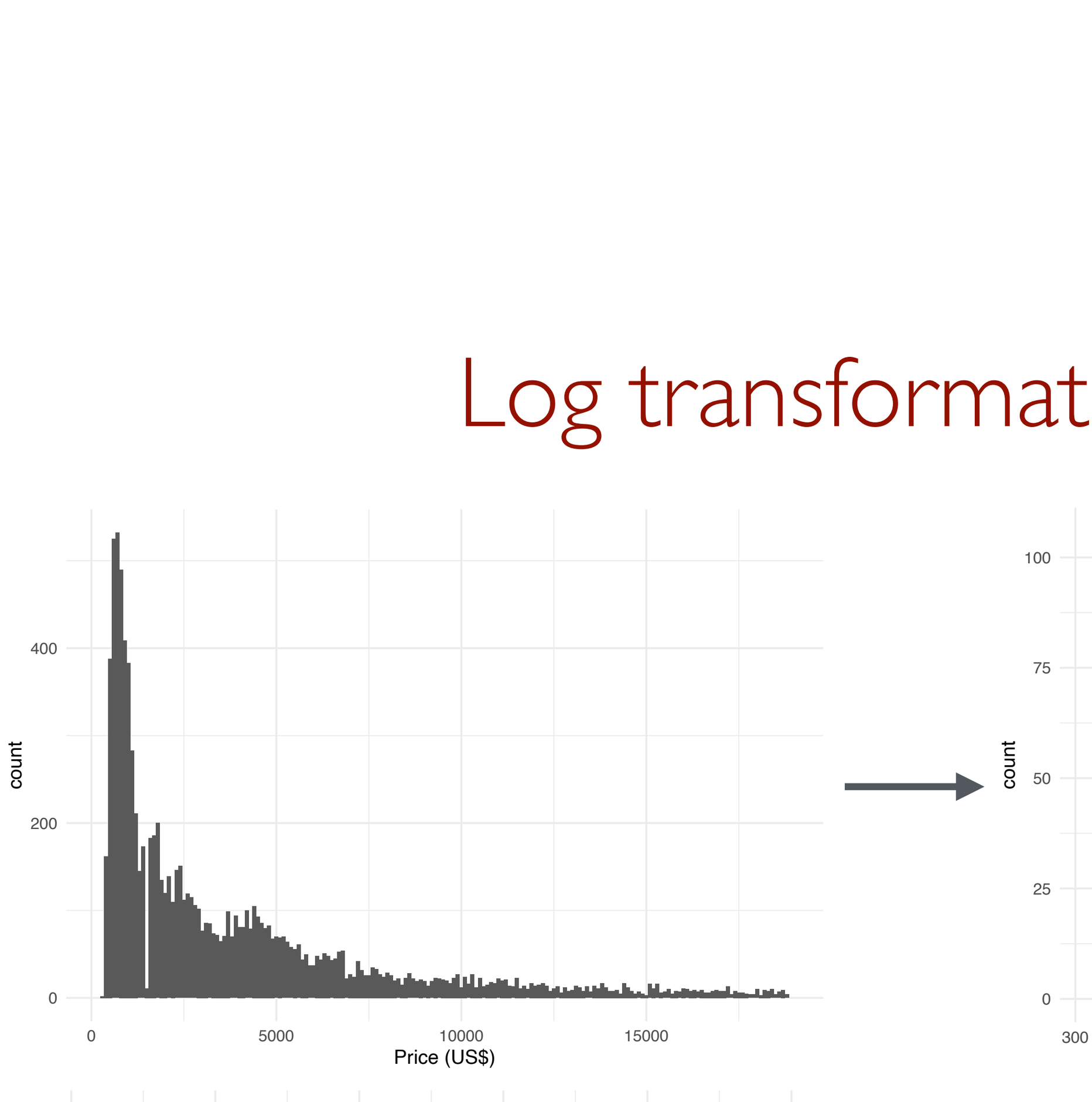
A note on log trans

- Log transformations are
  - Right-skewed data → more sy
  - Exponential relationships → m
  - Log-normal distributions → m
- Has effect of "subduing"
- Good for visualization
- Makes linear models m

Log transformat

- Square root has similar
- Log transformations hav
  - Monotonic increasing → prese
  - Constant ratios → constant dif
  - Easy to differentiate → useful f
- Transformations for left
  - Power functions
  - Exponentiation

Log transformat



Fitting the model: pr

```
fit1 <- lm(log10(price) ~ log10(carat), d
summ ry(fit1)
```

```
##
## Call:
## lm(formula = log10(price) ~ log10(car
##
## Residuals:
##       Min        1Q    Median        3Q
## -0.58268 -0.07321 -0.00186  0.07172
##
## Coefficients:
##                Estimate Std. Error t va
## (Intercept)  3.667107   0.001393    2
## log10(carat) 1.673639   0.004536
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.
##
## Residual standard error: 0.115 on 999
## Multiple R-squared:  0.9316,  djusted
## F-statistic: 1.361e+05 on 1 and 9998
```

Interpretati

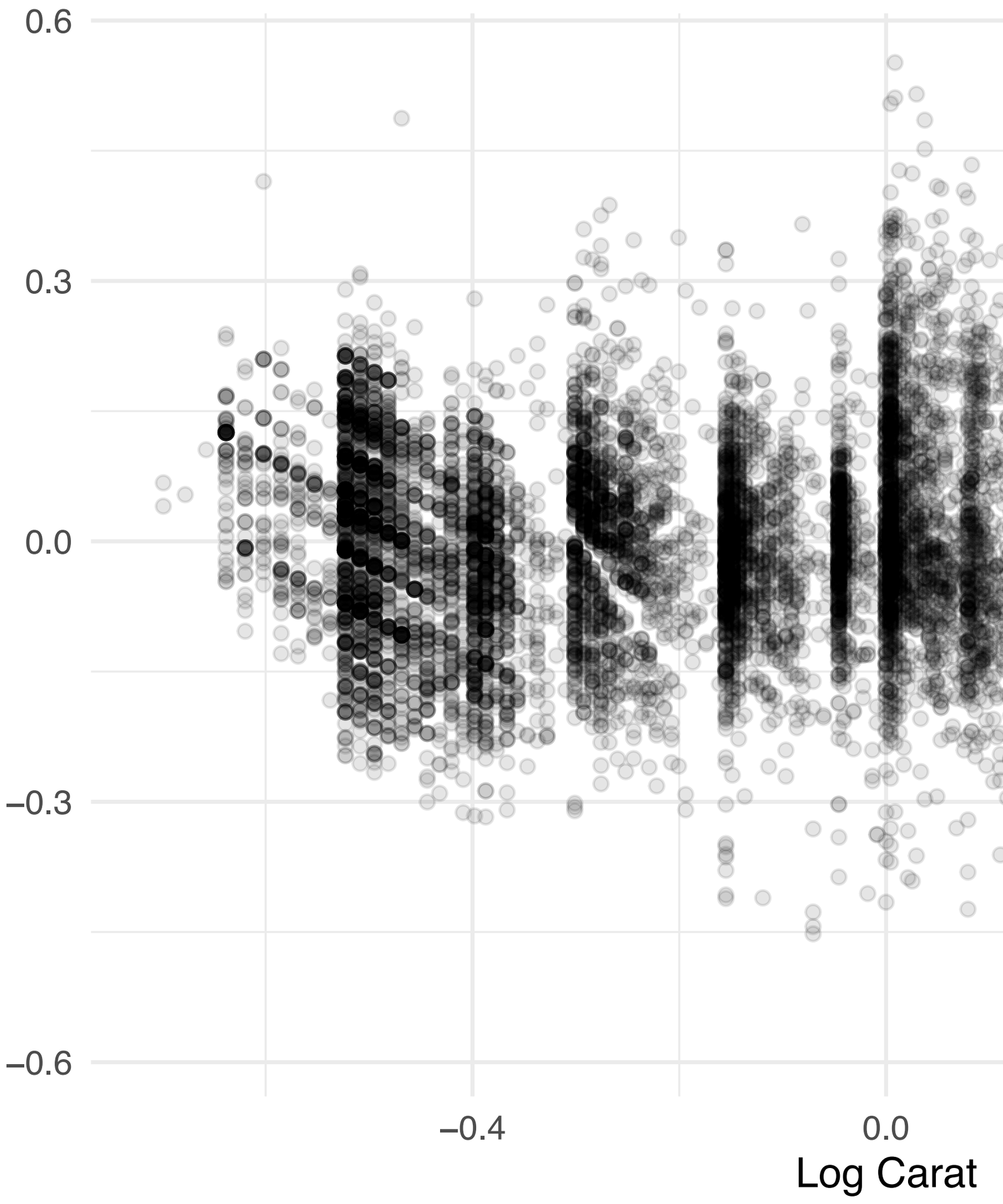
- Model coefficients desc
  between predictor and
- Assume all other predic
- For a one-unit change i
  change in **y** is the coeff

MODEL DIAGN

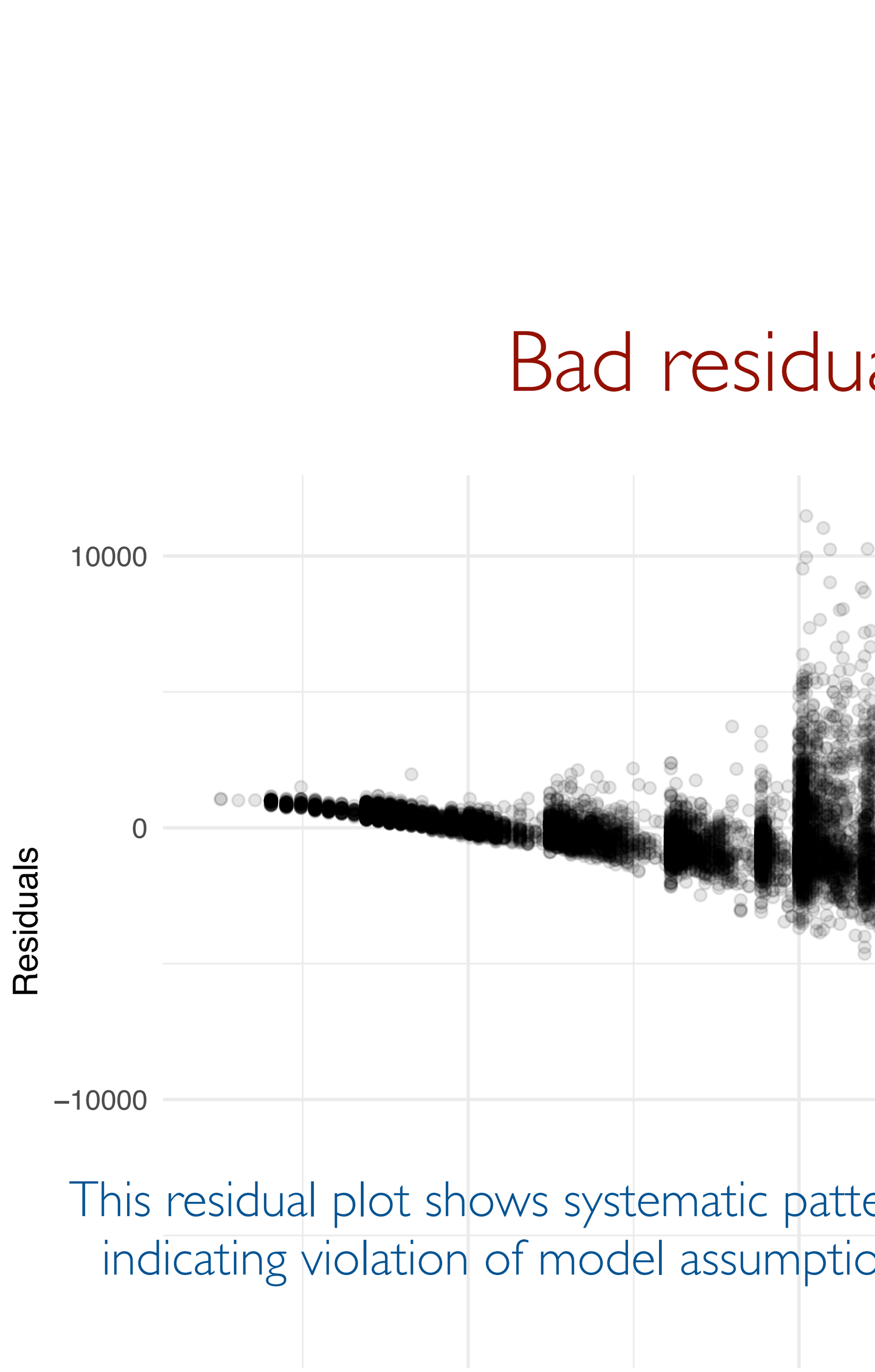
Model diagno

- Visualize the residuals (y
- Useful diagnostic step fo violations of model assu
- Can be used to identify relationships to model

Plotting the res



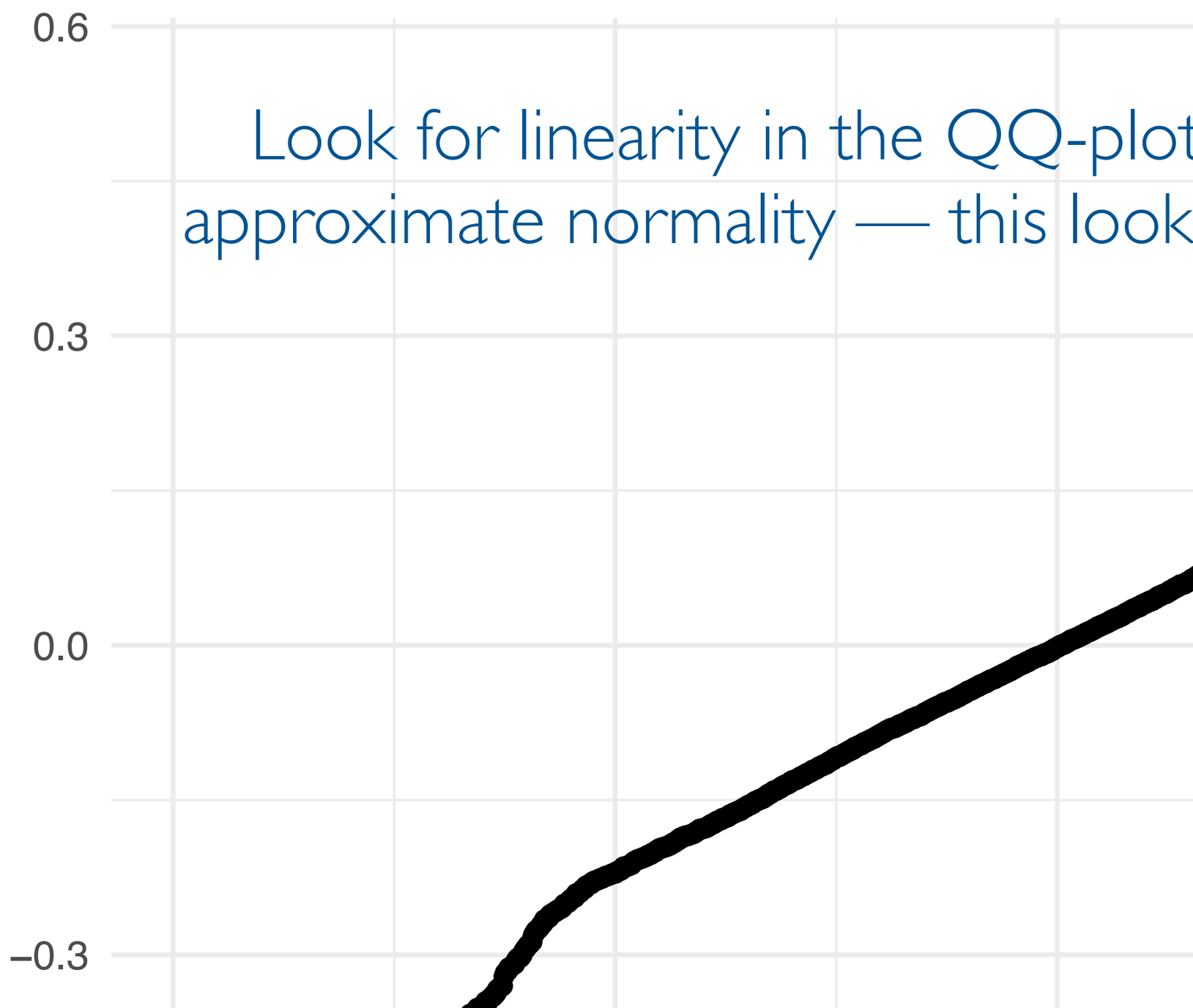
Bad residua



This residual plot shows systematic patte
indicating violation of model assumptio

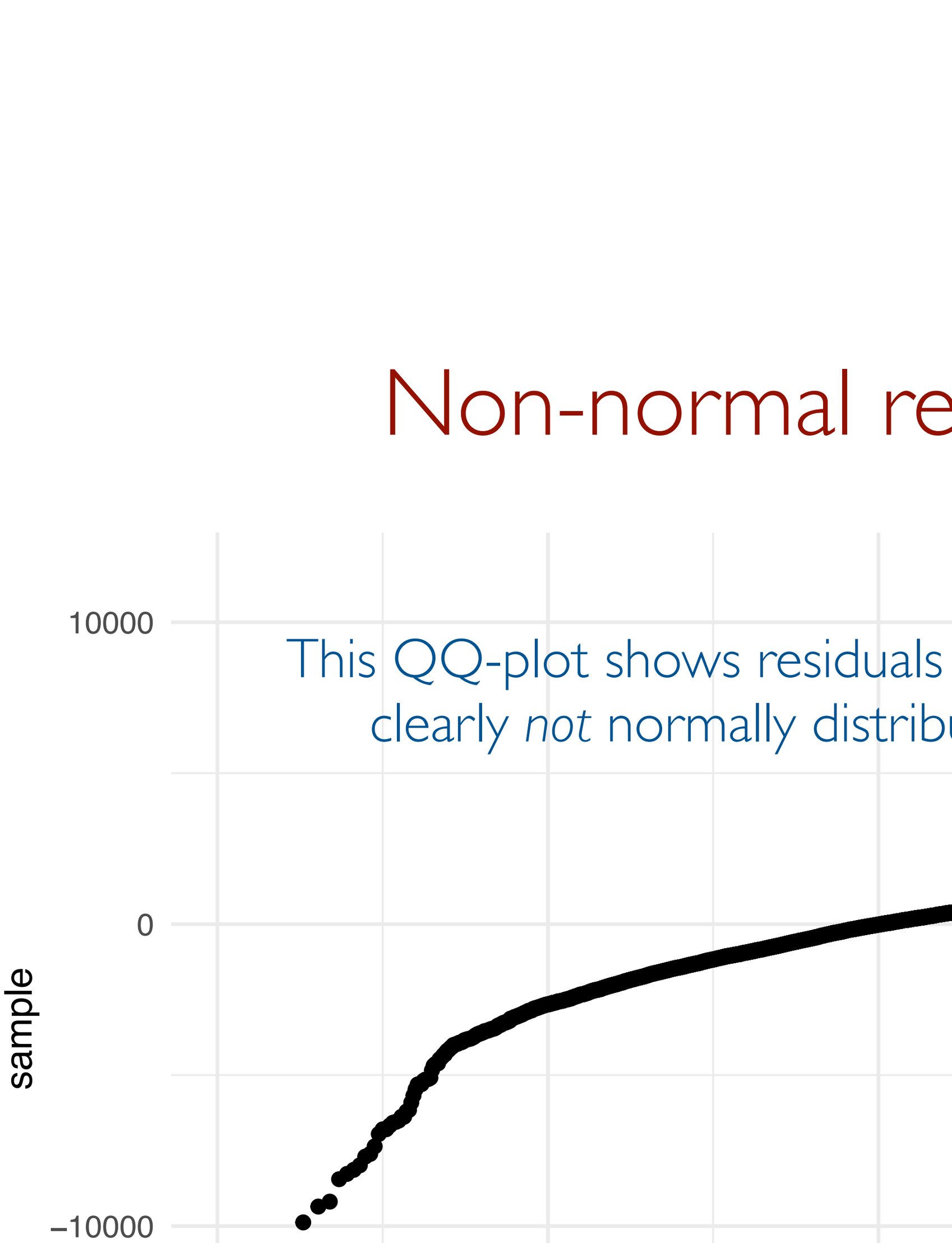
Normality of re

Look for linearity in the QQ-plot
approximate normality — this look



Non-normal re

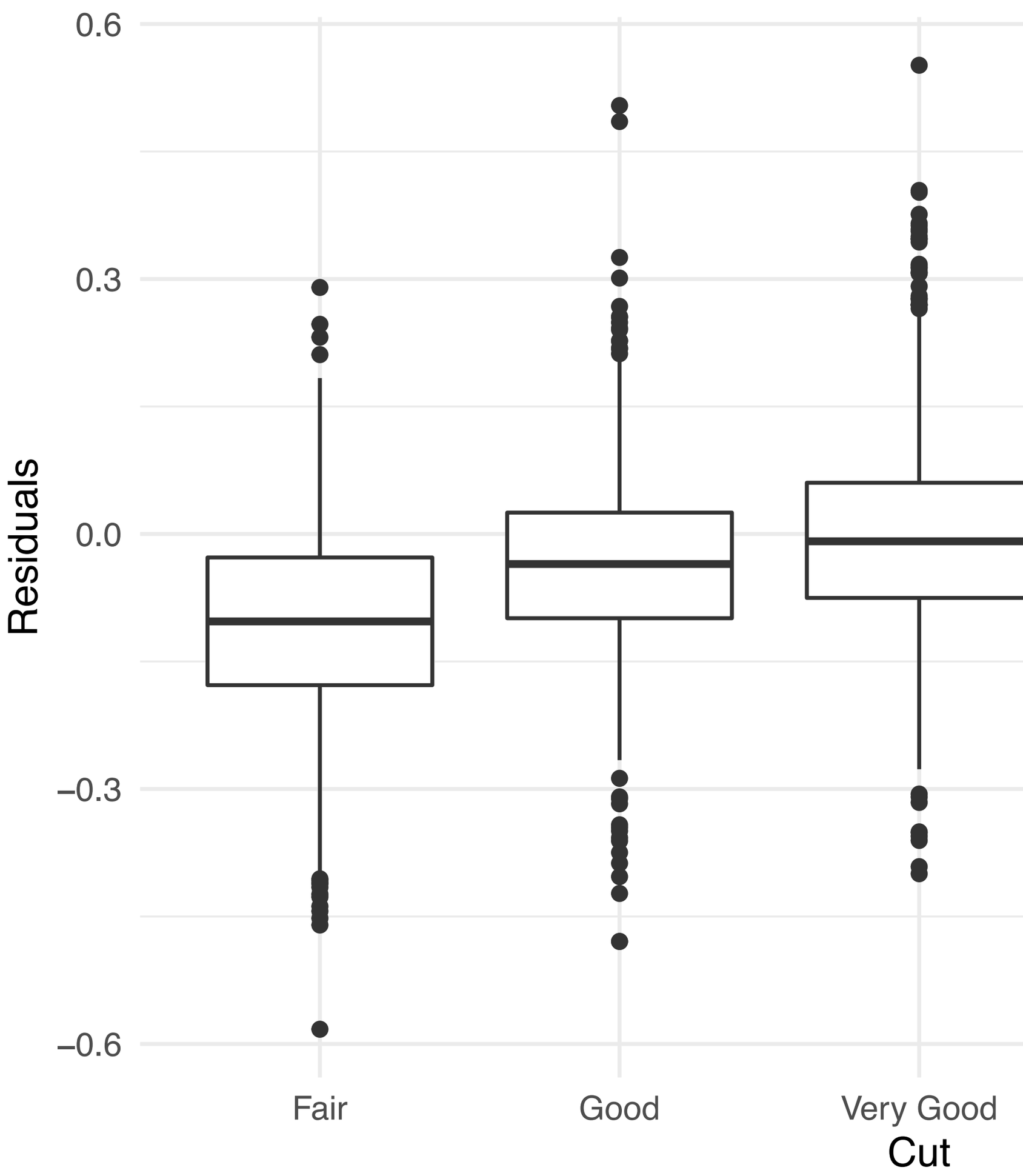
This QQ-plot shows residuals clearly *not* normally distribu



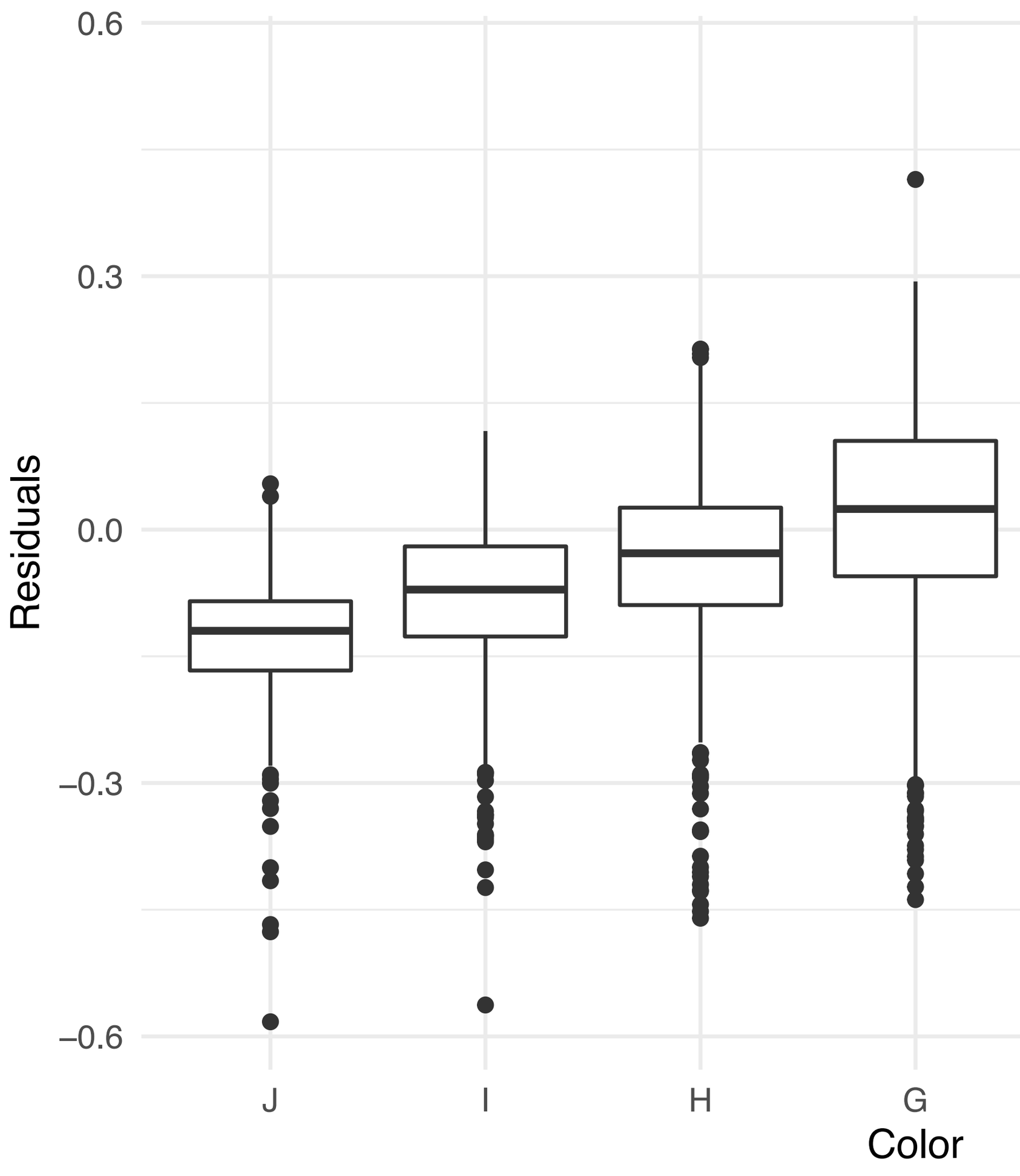
Checking other v

- The residuals can act as
variable" for the respon
- Modeling "removes" var
by the predictors in the
- Residuals contain the "le
- Look for relationship be
and unused variables

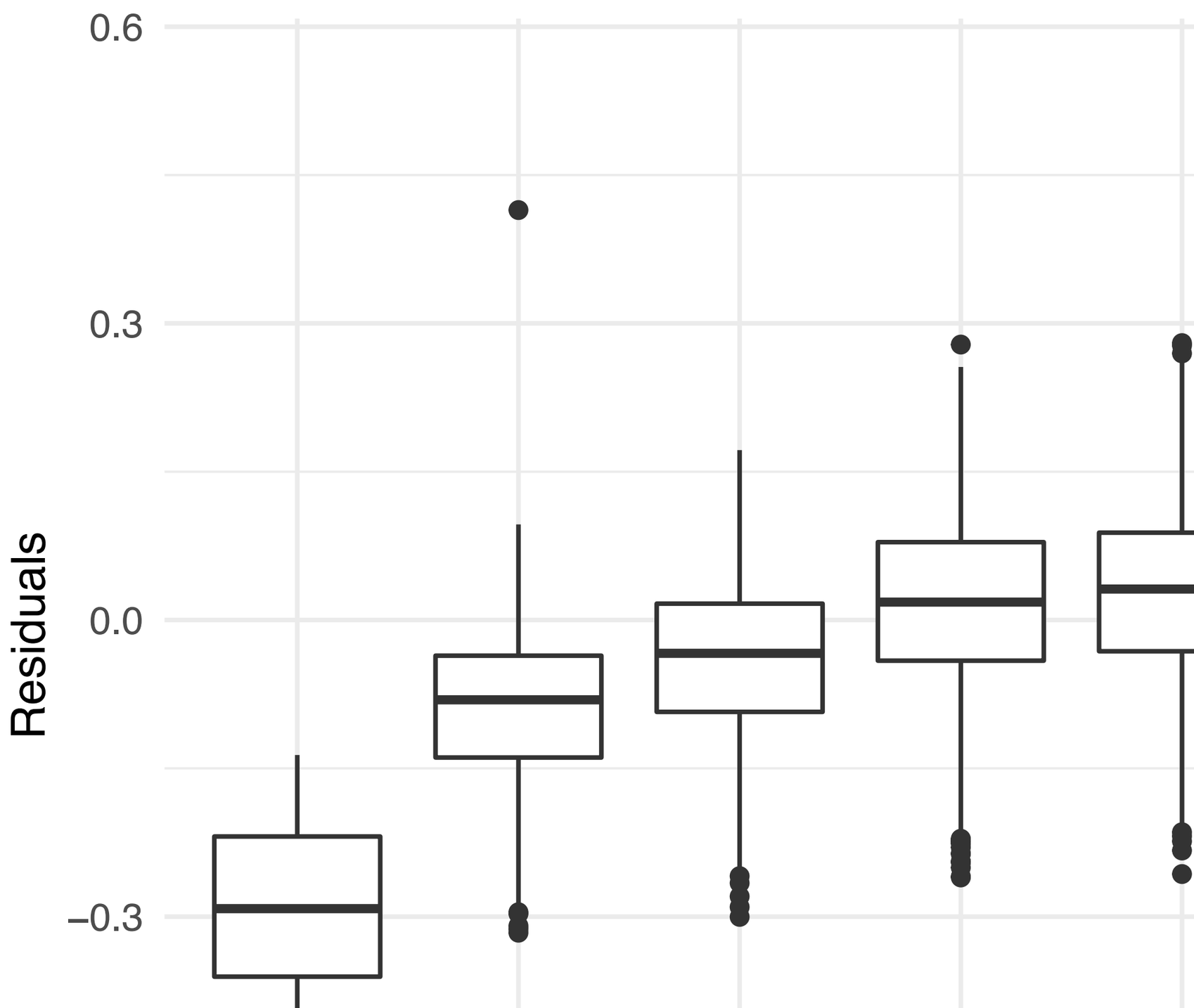
Better cuts → hig



Better color → hig



Better clarity → hi



Modeling with catego

- Categorical variables ar indicator variables
- Each indicator variable i
- For a categorical variab model it with ***m-1*** indi
- Otherwise, model is ov

A better model for di

```
fit2 <- lm(log10(price) ~ log10(carat) + cut + colo
summ ry(fit2)
```

```
##
## Call:
## lm(formula = log10(price) ~ log10(carat) + cut +
##     data = diamonds10k)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.30130 -0.03785  0.00030  0.03564  0.61635
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t
## (Intercept)    3.406711   0.005657 602.249   <2e-
## log10(carat)   1.885043   0.002622 719.022   <2e-
## cutGood        0.032996   0.003916   8.425   <2e-
## cutIdeal       0.068733   0.003574  19.233   <2e-
## cutPremium     0.061333   0.003600  17.039   <2e-
## cutVery Good   0.051455   0.003637  14.147   <2e-
## colorE        -0.025128   0.002143 -11.725   <2e-
## colorF        -0.042578   0.002166 -19.662   <2e-
                                          ...
## clarityVVS2    0.420432   0.005195  80.931    <2e
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.0
##
## Residual standard error: 0.05791 on 9981 degree
## Multiple R-squared:  0.9827,  djusted R-squared
## F-statistic: 3.148e+04 on 18 and 9981 DF,  p-va
```