Model transformations

BIOL2022 – **B**iology **E**xperimental **D**esign and **A**nalysis (**BEDA**)

Januar Harianto

The University of Sydney

Semester 2, 2025

THE UNIVERSITY OF SYDNEY

Learning objectives

You should:

- [ ] Understand why model transformations are necessary.
- [ ] Differentiate between transforming the data and formulating a new model.
- [ ] Apply common transformations (log, square root) to the response variable.
- [ ] Interpret the results of a log-transformed model.

Why do we perform model transformations?

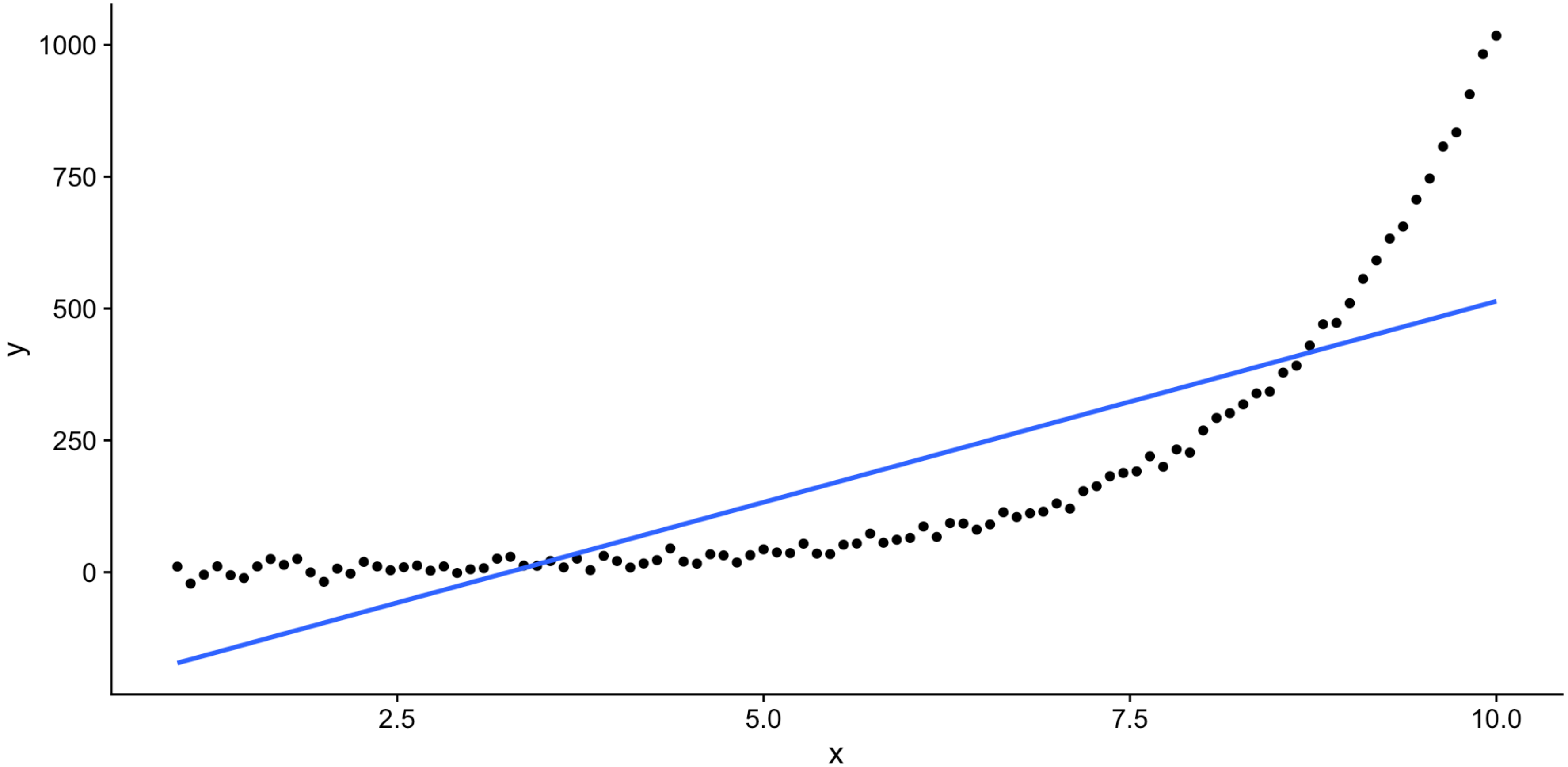
- When data does not meet LINE assumptions, we can attempt transformations to improve the model fit before considering more complex models.
- Transformations linearise (or at least, attempt to) the relationship between the response and predictor variables.
- *Not cheating!* **We are improving the model fit, not changing the data arbitrarily.**

The idea behind transformations

Given a simple model between two variables, y and x:

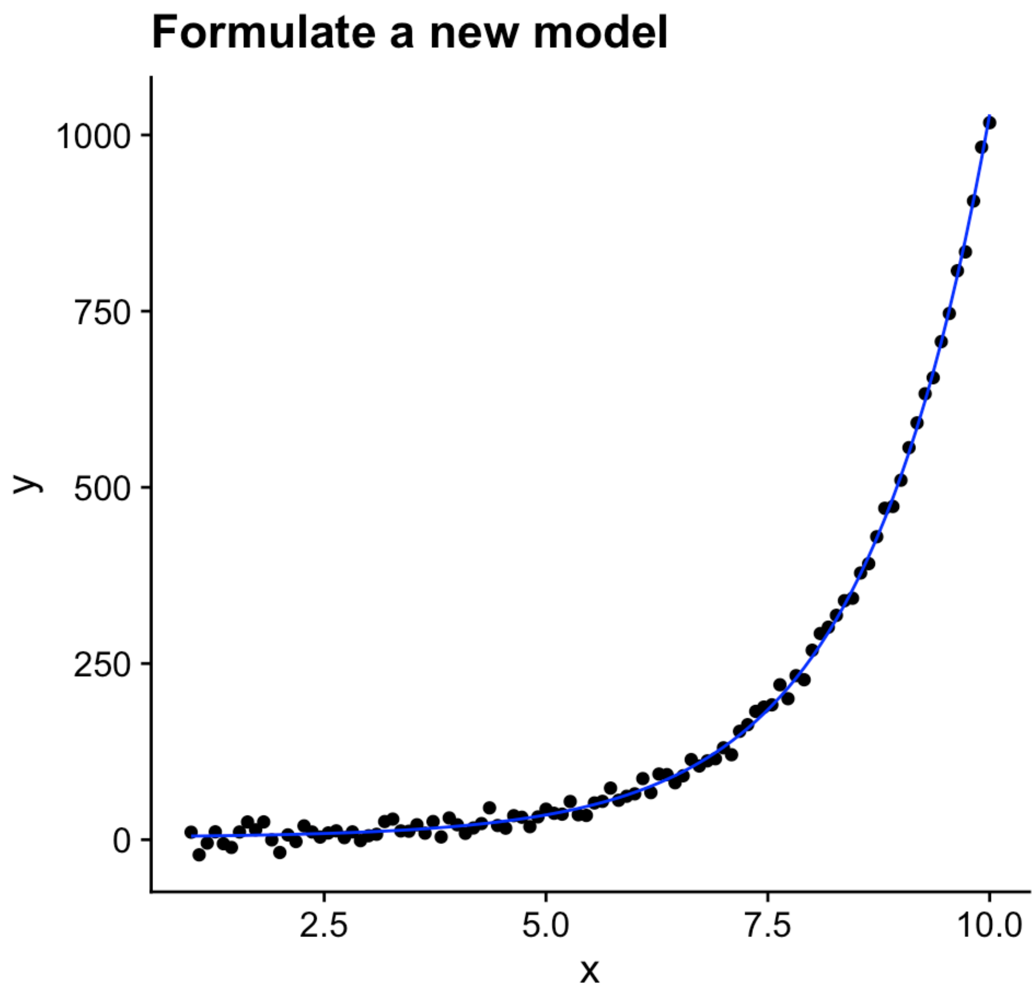
$$y \sim x$$

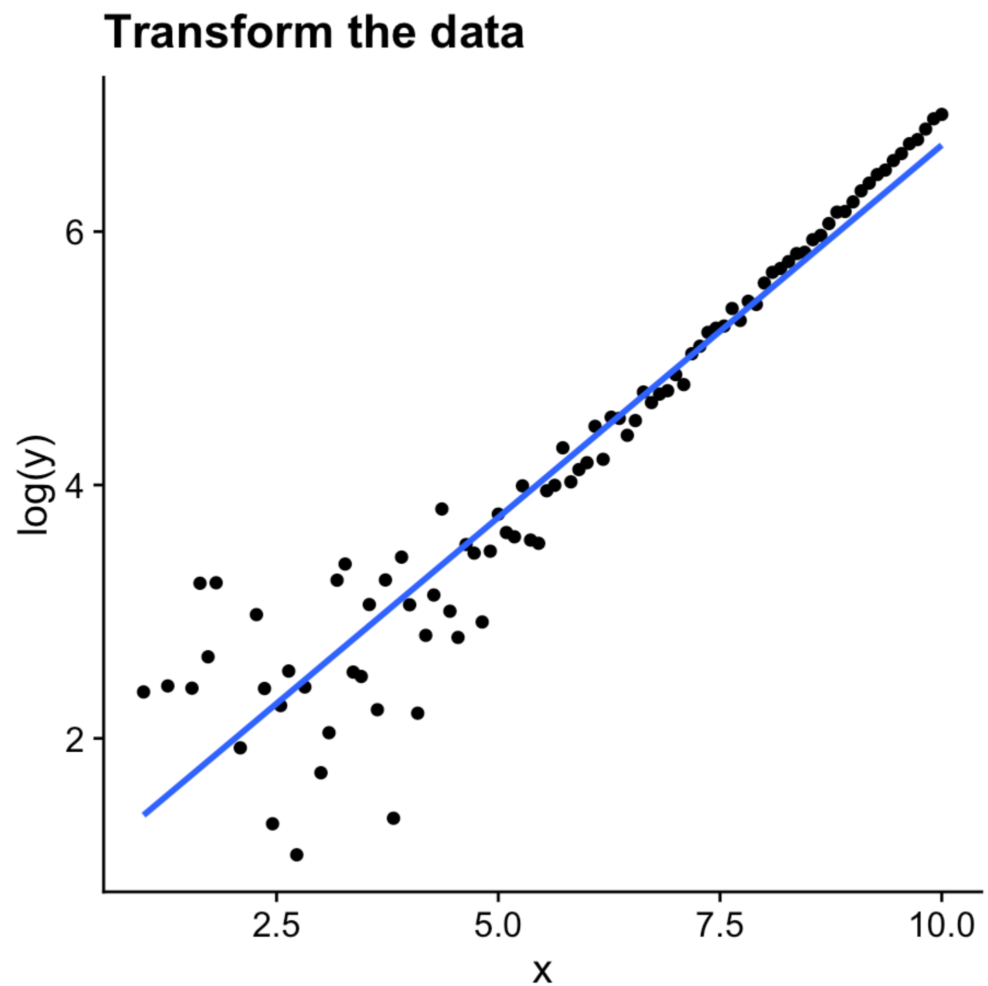
Where the relationship is not linear, we may end up with a model that does not fit the data well:



Two ways to transform

We can either formulate a new model that better fits the data, or transform the data to better fit the model. **Both methods are essentially equivalent.**





It’s not easy to formulate a new model

It turns out that you need a lot of domain knowledge to formulate a new model.

To fit the model to this particular dataset, we need to formulate:

$$y = a \times 2^x + b$$

Transforming the data is easier. Basically, we can transform the response variable and approximate the model:

$$\log(y) \sim x$$

It is not perfect and may even introduce issues, but it can be a good starting point. It is also easier to do when dealing with complex multi-factorial models.

How do we transform data?

Irregardless of the complexity of the model, apply the transformation to the response variable:

$$y \sim x_1 + x_2 + \ldots + x_n \rightarrow f(y) \sim x_1 + x_2 + \ldots + x_n$$

Depending on the relationship between the response and predictor variables, we can apply different transformations:

- **Logarithmic transformation**: $f(y) = \log(y)$: right skewed data
- **Square root transformation**: $f(y) = \sqrt{y}$: count data with many small values
- **Reciprocal transformation**: $f(y) = \frac{1}{y}$: when other transforms do not work

In most cases, a logarithmic transformation is a good starting point. It is also easier to interpret the results.

Interpretation of a log-transformed model

Given a model:

$$\log(y) = \beta_0 + \beta_1 x$$

where we are transforming the response variable y using the **natural** logarithm, $\log(y)$, then **for a one-unit increase in x, the response variable y increases by a factor of** $\beta_1 \times 100\%$.

We can also use estimated marginal means to interpret the results in R, where back-transforms can be automatically calculated to the model.

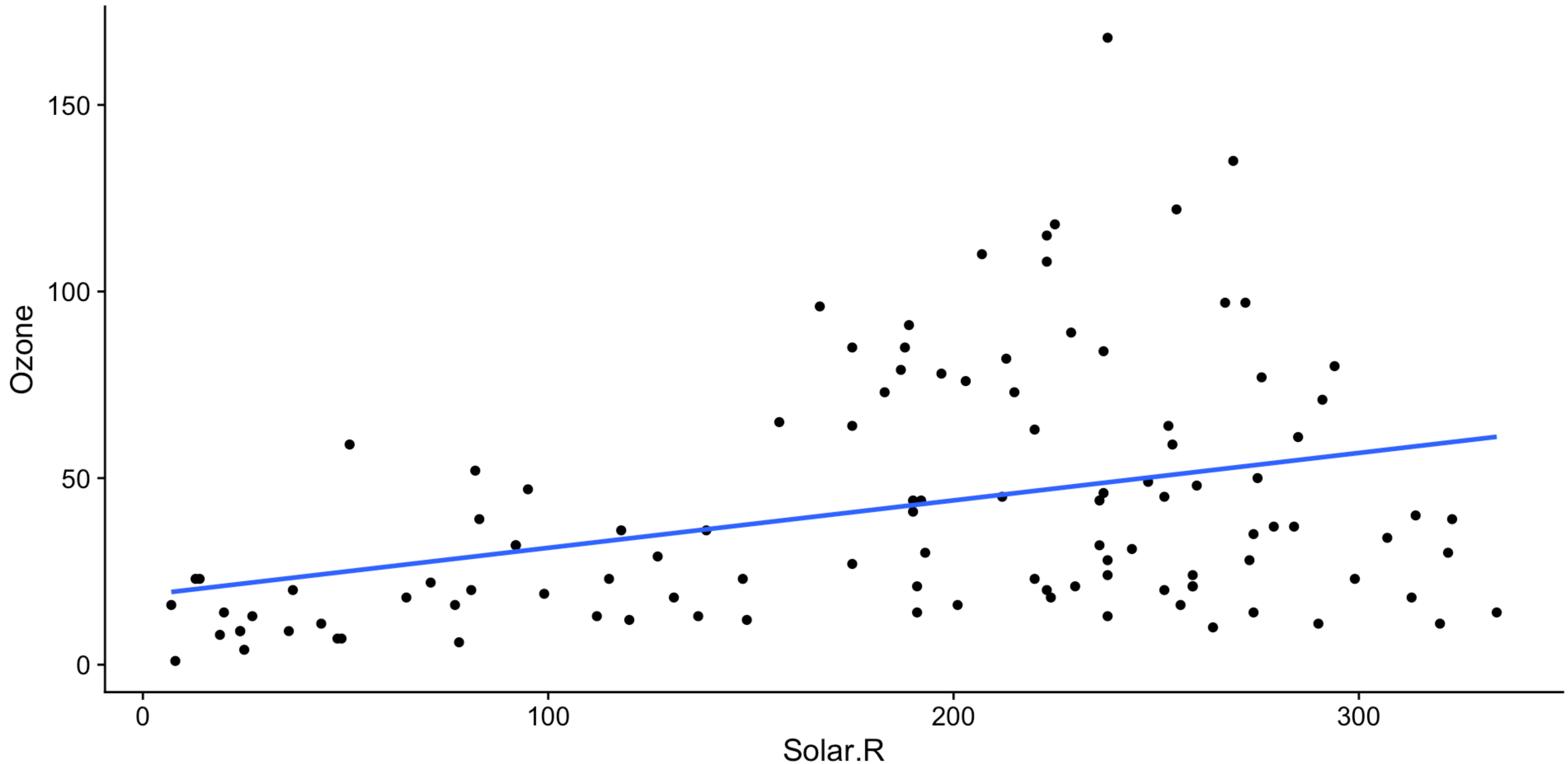
Example

New York City skyline enveloped in heavy smog, May 1973. Photo by Chester Higgins/NARA (CC BY-NC 2.0)

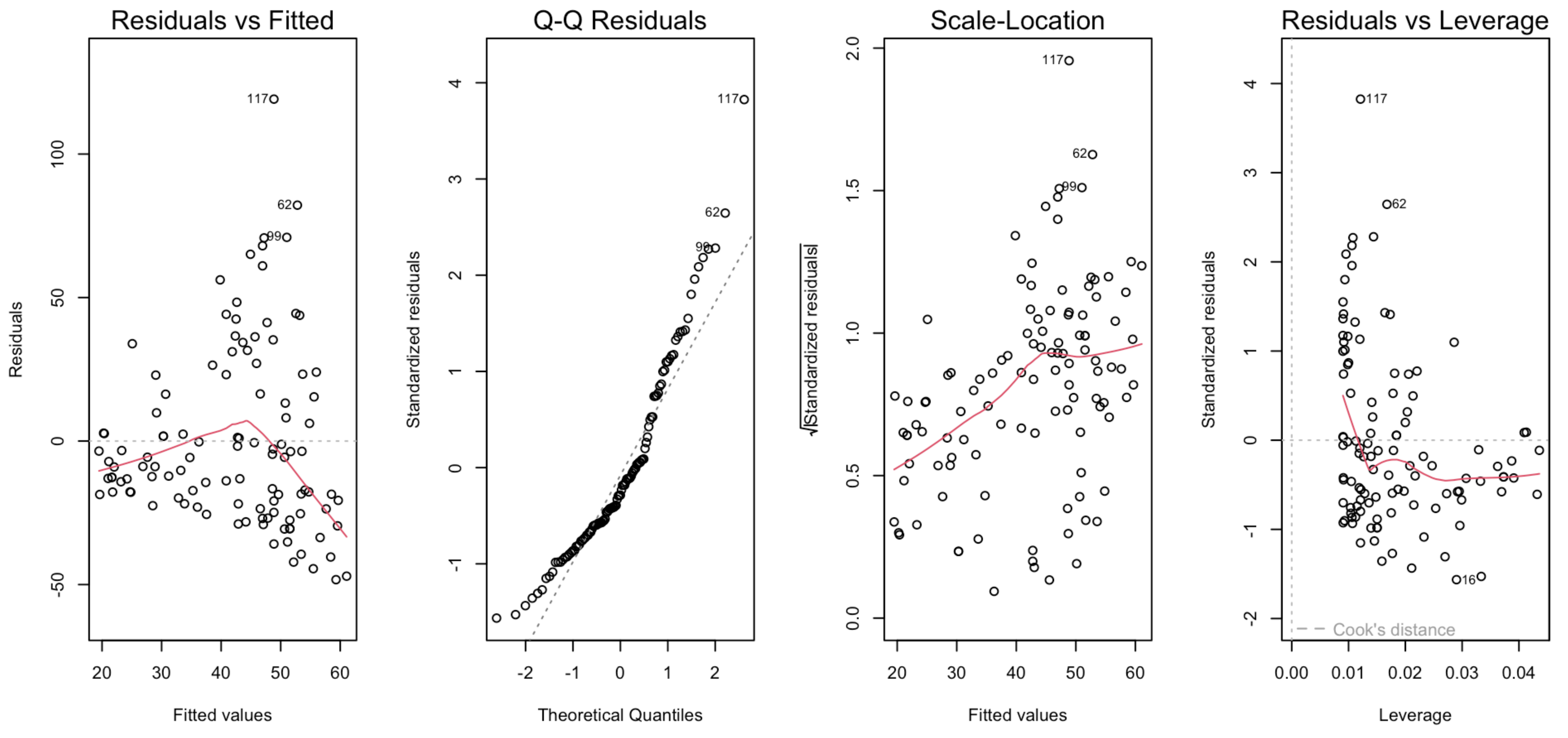
Air quality in New York City, 1973

Is air quality (ozone concentration) in New York City influenced by solar radiation? The model is:

$$\text{ozone} \sim \text{solar radiation}$$



Assumptions



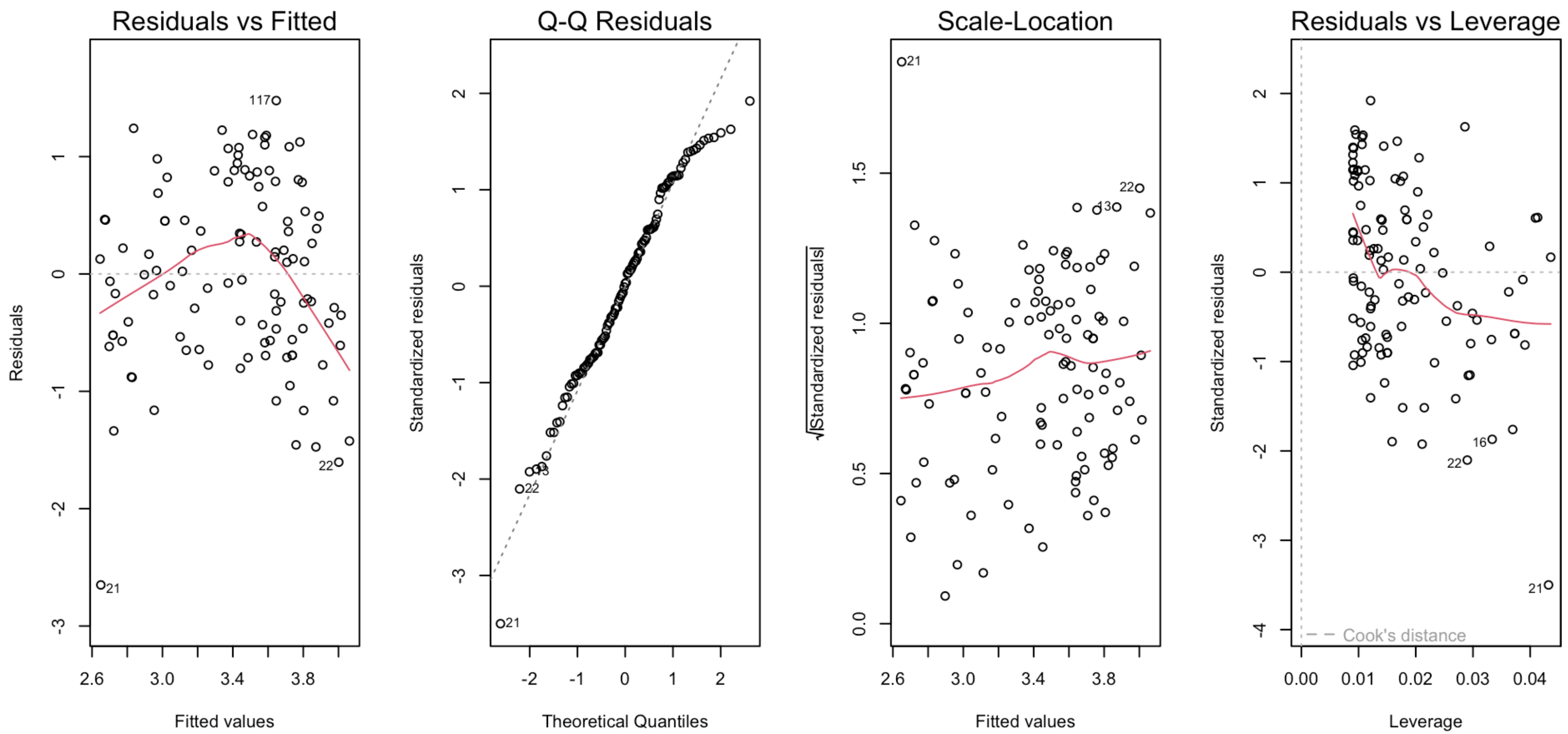
Did the model assumptions hold?

- **Linearity**: the relationship between ozone and solar radiation is not linear, evident fan-shape in the residual vs fitted plot.
- **Normality**: the residuals in the qq-plot are “u-shaped”, indicating a positively skewed distribution.
- **Equal variance**: the residuals are not homoscedastic – increasing variance with increasing fitted values seen in the scale-location plot, although it is not severe (not more than 2 standard deviations).

Transforming the data

Given the non-linear relationship between ozone and solar radiation, we can apply a logarithmic transformation to the response variable:

$$\log(\text{ozone}) \sim \text{solar radiation}$$



Is the model a better fit?

Yes! Fanning in the residual vs fitted plot is reduced, and the “u-shaped” distribution in the qq-plot is no longer evident. Scale-location plot shows a more consistent variance across the fitted values.

```
Call:
lm(formula = log(Ozone) ~ Solar.R, data = airquality)

Residuals:
     Min       1Q   Median       3Q      Max 
-2.64991 -0.56329  0.02199  0.55373  1.47755 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 2.6152491  0.1666990  15.688  < 2e-16 ***
Solar.R     0.0043326  0.0008097   5.351 4.88e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7741 on 109 degrees of freedom
  (42 observations deleted due to missingness)
Multiple R-squared:  0.208,	Adjusted R-squared:  0.2008 
F-statistic: 28.63 on 1 and 109 DF,  p-value: 4.885e-07
```

How do we interpret the results?

So, is transformation necessary?

Let's compare the summaries of both models. This is sometimes called a sensitivity analysis.

Untransformed model   Log-transformed model

```
summary(fit)
```

```
Call:
lm(formula = Ozone ~ Solar.R, data = airquality)

Residuals:
    Min      1Q  Median      3Q     Max 
-48.292 -21.361  -8.864  16.373 119.136 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 18.59873    6.74790   2.756 0.006856 ** 
Solar.R      0.12717    0.03278   3.880 0.000179 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 31.33 on 109 degrees of freedom
  (42 observations deleted due to missingness)
Multiple R-squared:  0.1213,	Adjusted R-squared:  0.1133 
F-statistic: 15.05 on 1 and 109 DF,  p-value: 0.0001793
```

Interpretation trade-offs

- The **untransformed model** is easier to interpret: for every 1 W/m^2 increase in solar radiation, ozone concentration is predicted to increase by ~0.13 ppb.
- However, this model does not fit the data well and violates several assumptions, so we cannot be confident in this prediction.

- The **log-transformed model** fits the data better, but is more difficult to interpret.
- A 1 W/m^2 increase in solar radiation is associated with a 261.52, 0.43% increase in ozone concentration.
- We can also back-transform the results to the original scale, but this gives us the *median* change in ozone concentration, not the *mean*.

Verdict: If the goal is to understand the relationship between variables and **make predictions**, the transformed model is better. If the goal is simple interpretation and the model violations are not severe, the untransformed model may be acceptable – **most biologists prioritise interpretability.**

A sensitivity analysis is quick and easy to perform, so it is worth doing as soon as you are unsure of your model's assumptions (happens often).

Questions to consider

- When should you consider transforming your data versus fitting a more complex model (e.g., a generalised linear model)?
- How do you choose the appropriate transformation for your data?
- What are the challenges in interpreting the coefficients of a log-transformed model, and how can back-transformation help?
- Can transformations fix all violations of model assumptions? When might they not be enough?

Thanks!