Lesson: Linear regression diagnostic methods

Module: Regression Diagnostics

Photo by Janine Robinson on Unsplash

The Linear Regression Model

<u>Definition/Assumptions</u> of the linear regression model:

1. *Linearity*
2. *Independence*
3. *Homoskedasticity (constant variance)*
4. *Normality*

Motivation

So far, we've mainly taken for granted that the MLR assumptions have been met. But what if they aren't?

Diagnostic methods:

1. Graphical techniques

2. Numerical techniques

Motivation

A note on the validity assumption

Photo by Jason Leung on Unsplash

Lesson: Violations of the linearity assumption

Module: Regression Diagnostics

Photo by Janine Robinson on Unsplash

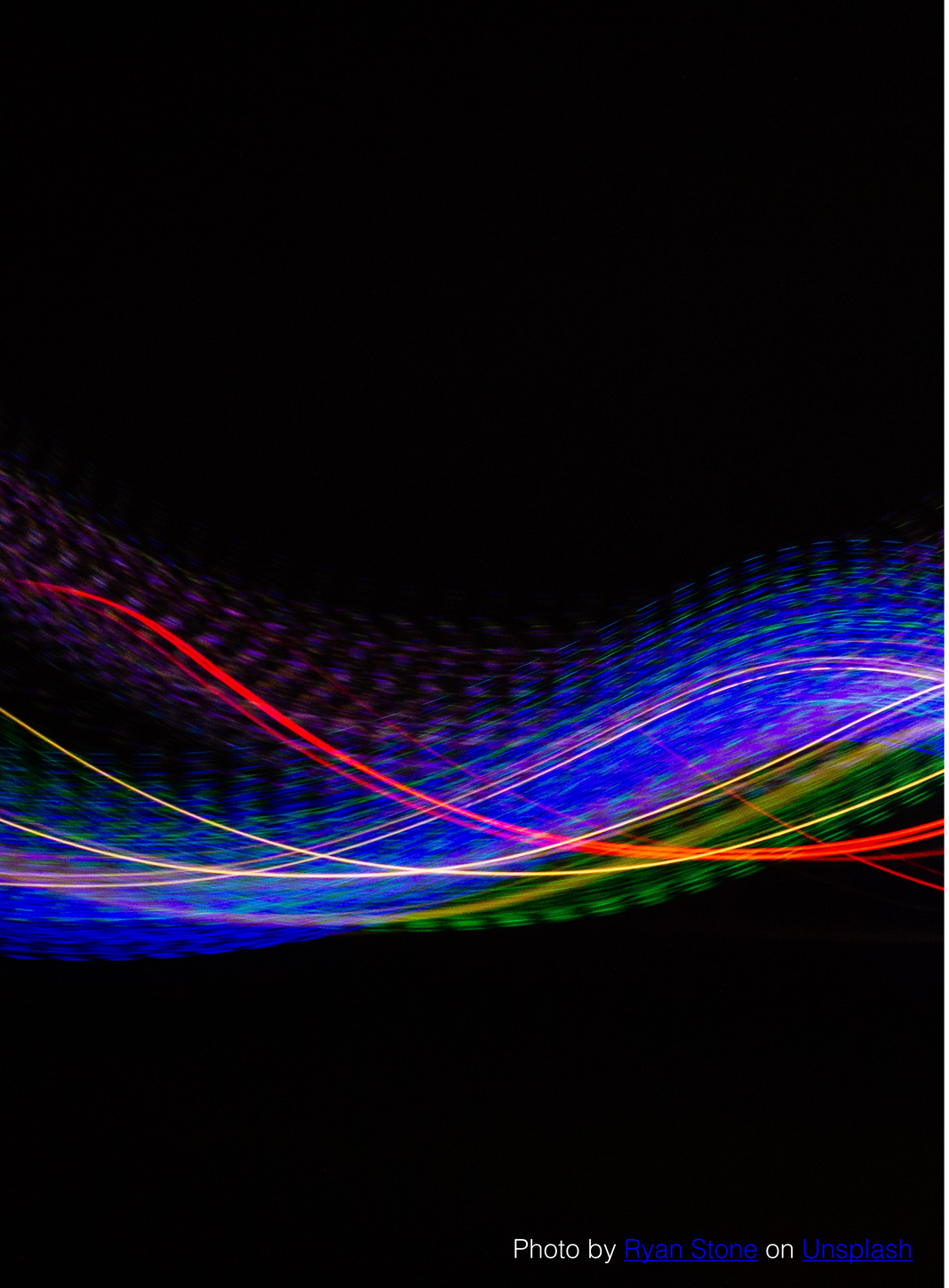


Photo by Ryan Stone on Unsplash

$\mathbf{Y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where $E(\boldsymbol{\varepsilon}) = 0$.

Or equivalently, that for $i = 1, \ldots, n$,

$Y_i = \beta_0 + \beta_1 x_{i,1} + \ldots + \beta_p x_{i,p} + \varepsilon_i$, where $E(\varepsilon_i) = 0$.

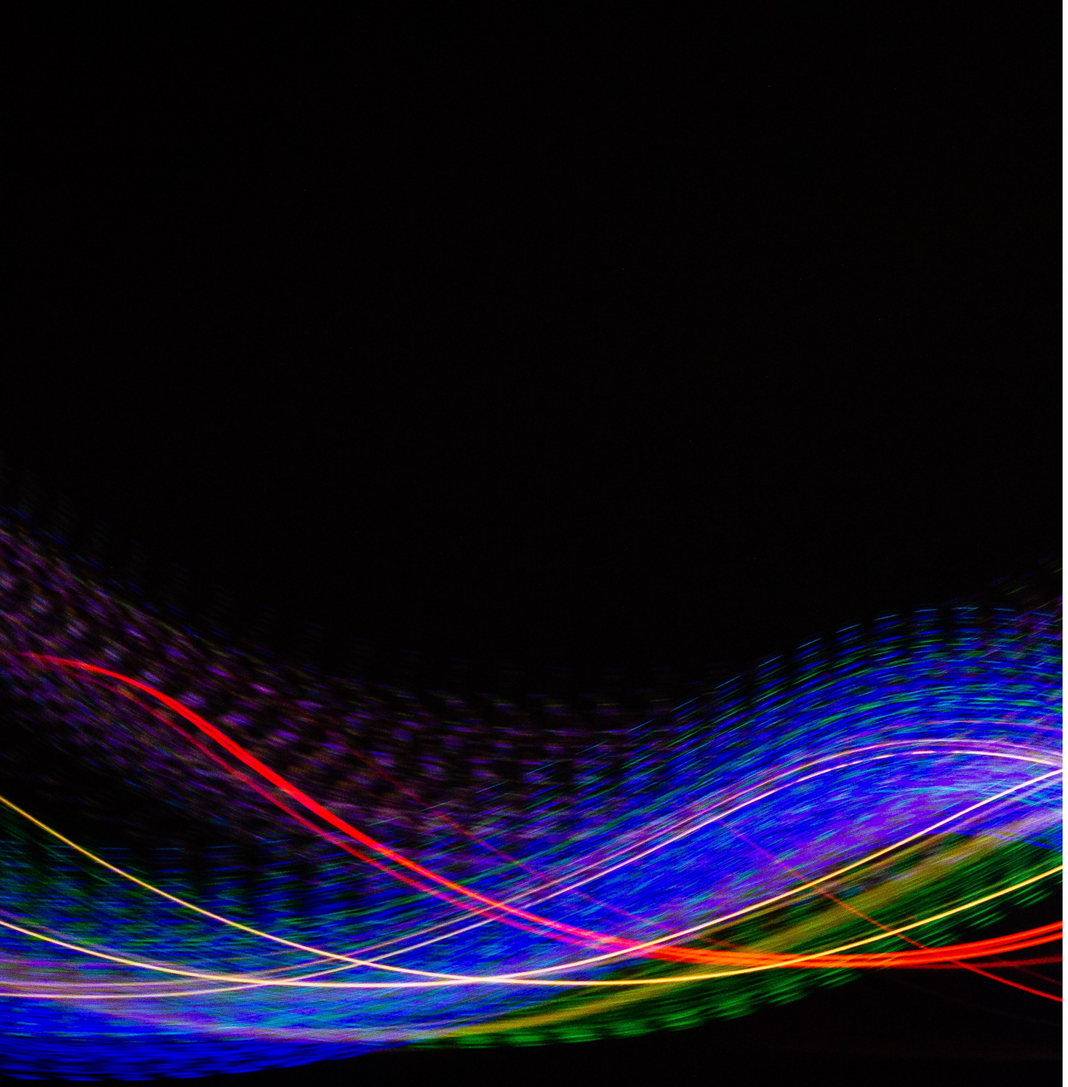


Photo by Ryan Stone on Unsplash

This specification means that, for each $i = 1, \ldots, n$ and $j = 1, \ldots, p$:

1. $E(Y_i)$ is a linear function of $x_{i,j}$, holding all other $x_{i,k}$, $k \neq j$ fixed.

2. The slope of the line that relates $E(Y_i)$ to $x_{i,j}$ does not depend on the values of any other $x_{i,k}$, $k \neq j$.

3. The effects of each $x_{i,j}$ on the

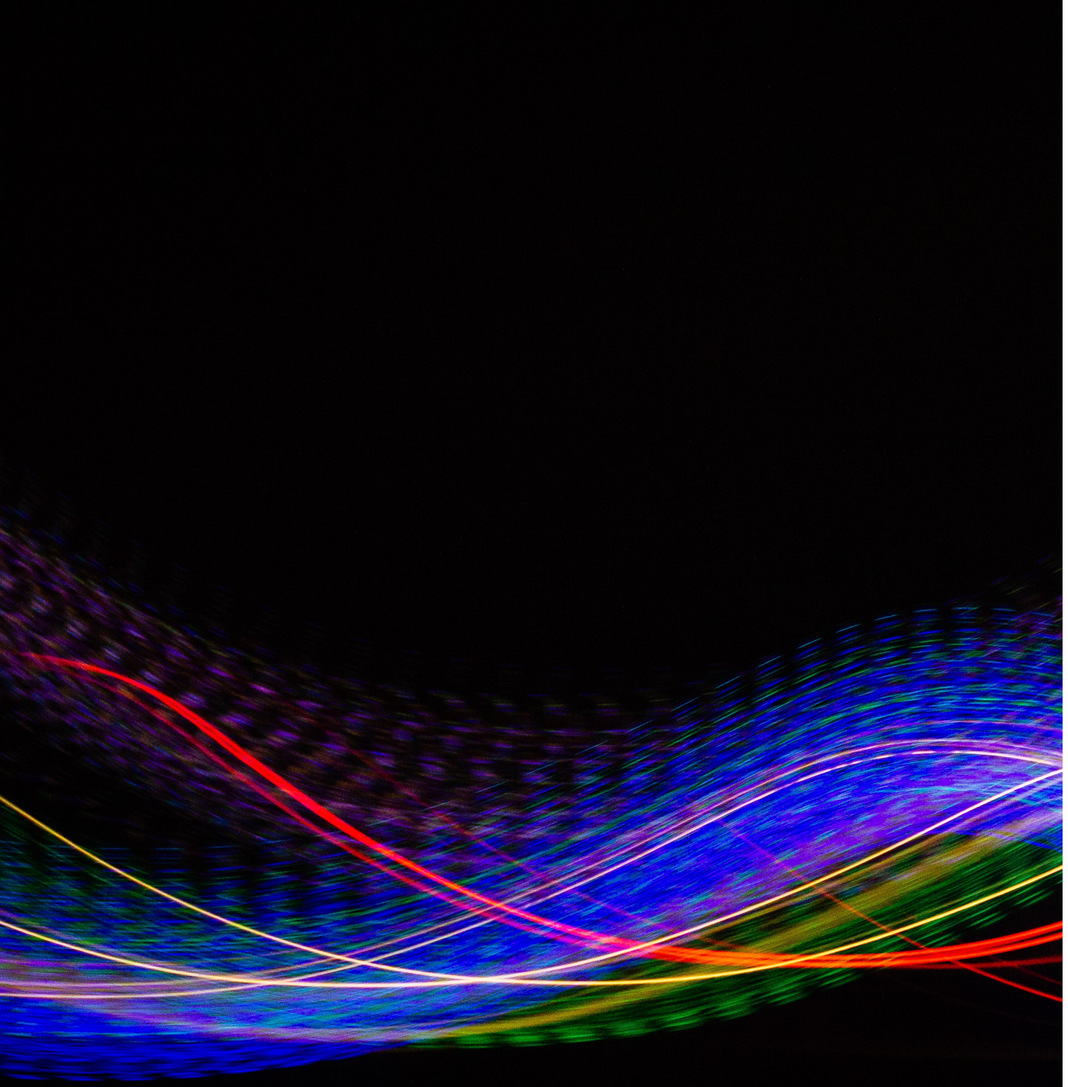


Photo by Ryan Stone on Unsplash

If this assumption is not met:

1. The LS estimator is not unbiased.
2. The model will have low explanatory power.
3. The inferences (e.g., t-tests, F-test) based on the misspecified regression model will be biased and misleading.
4. The predictions from the model are less likely to be accurate.

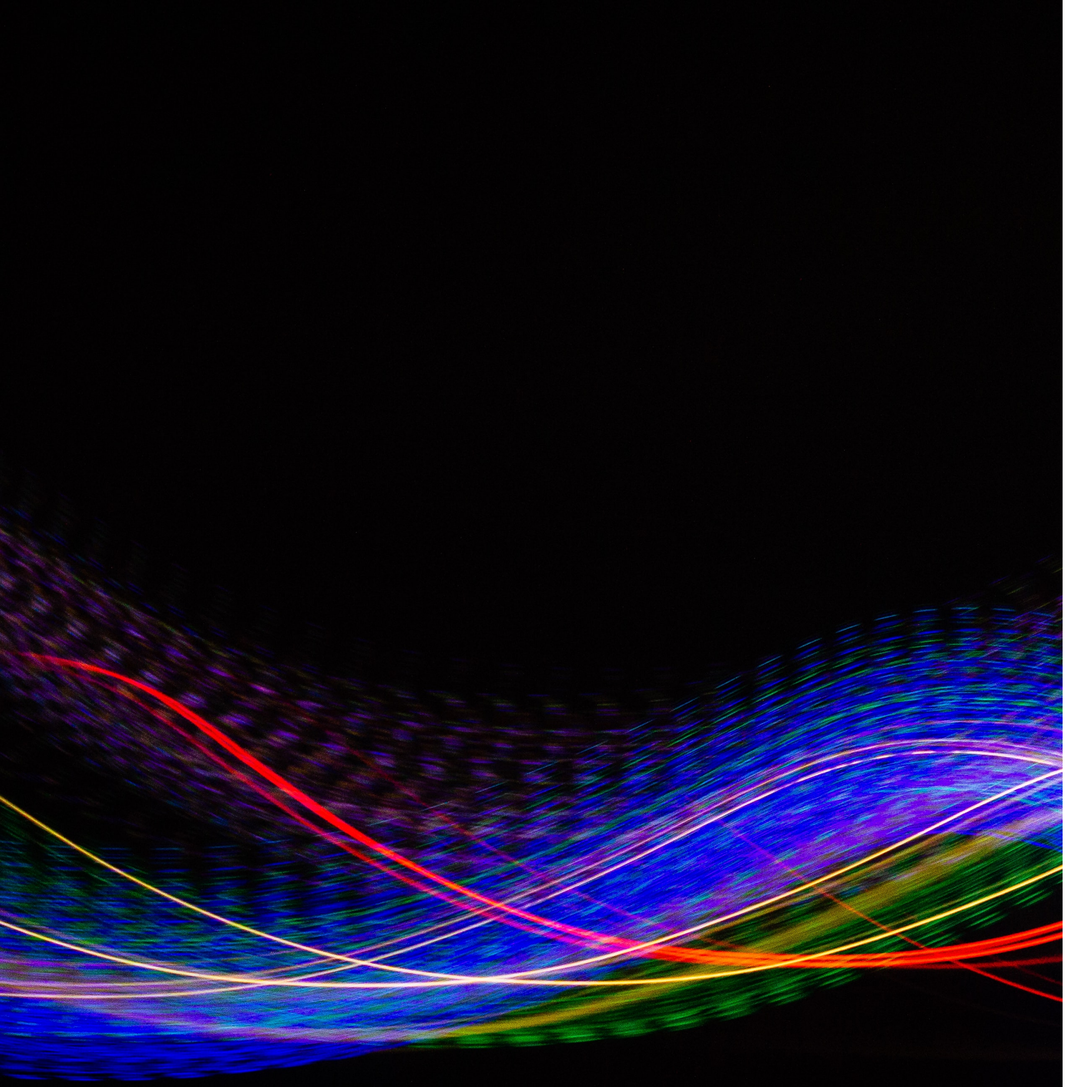
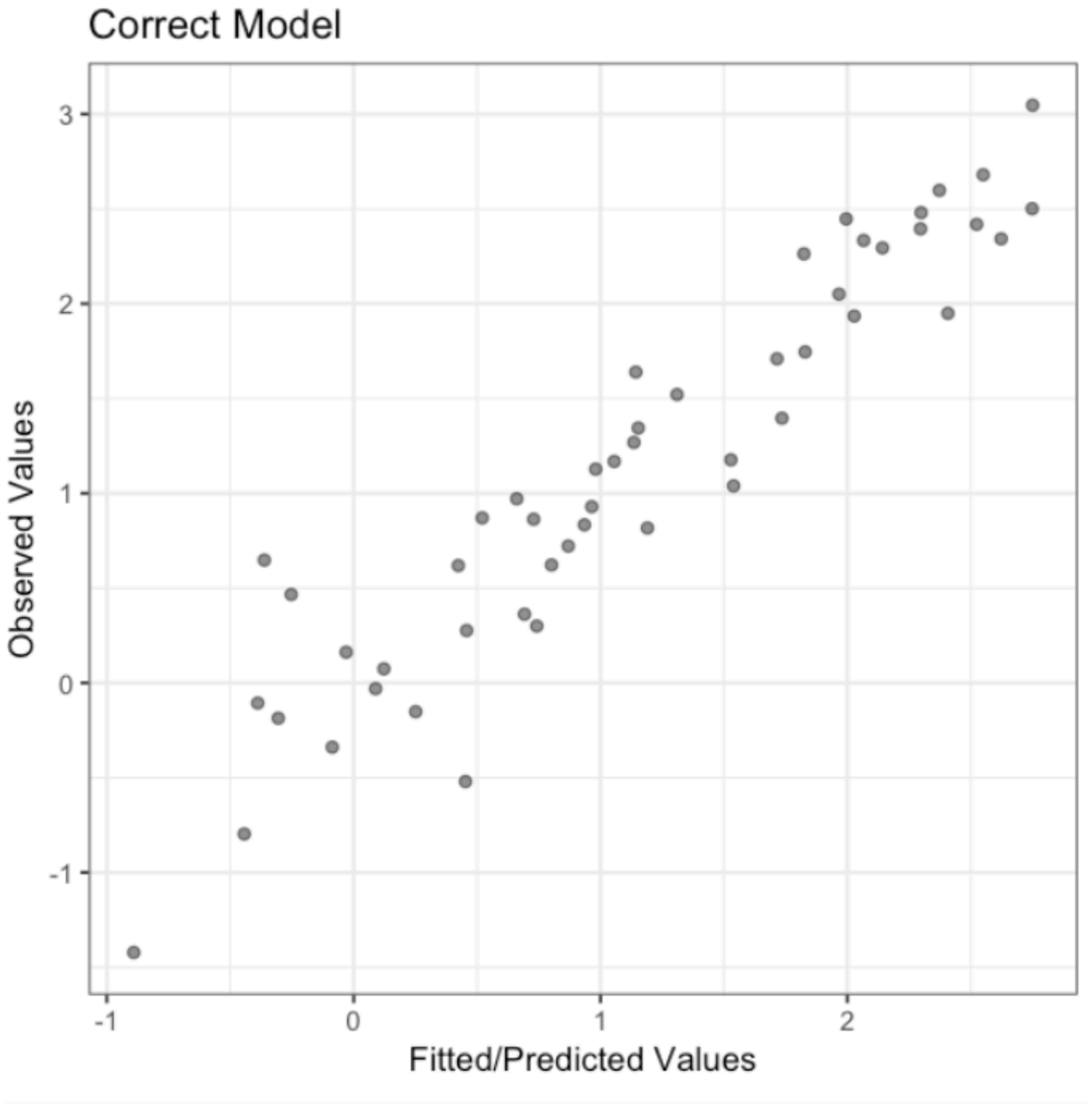
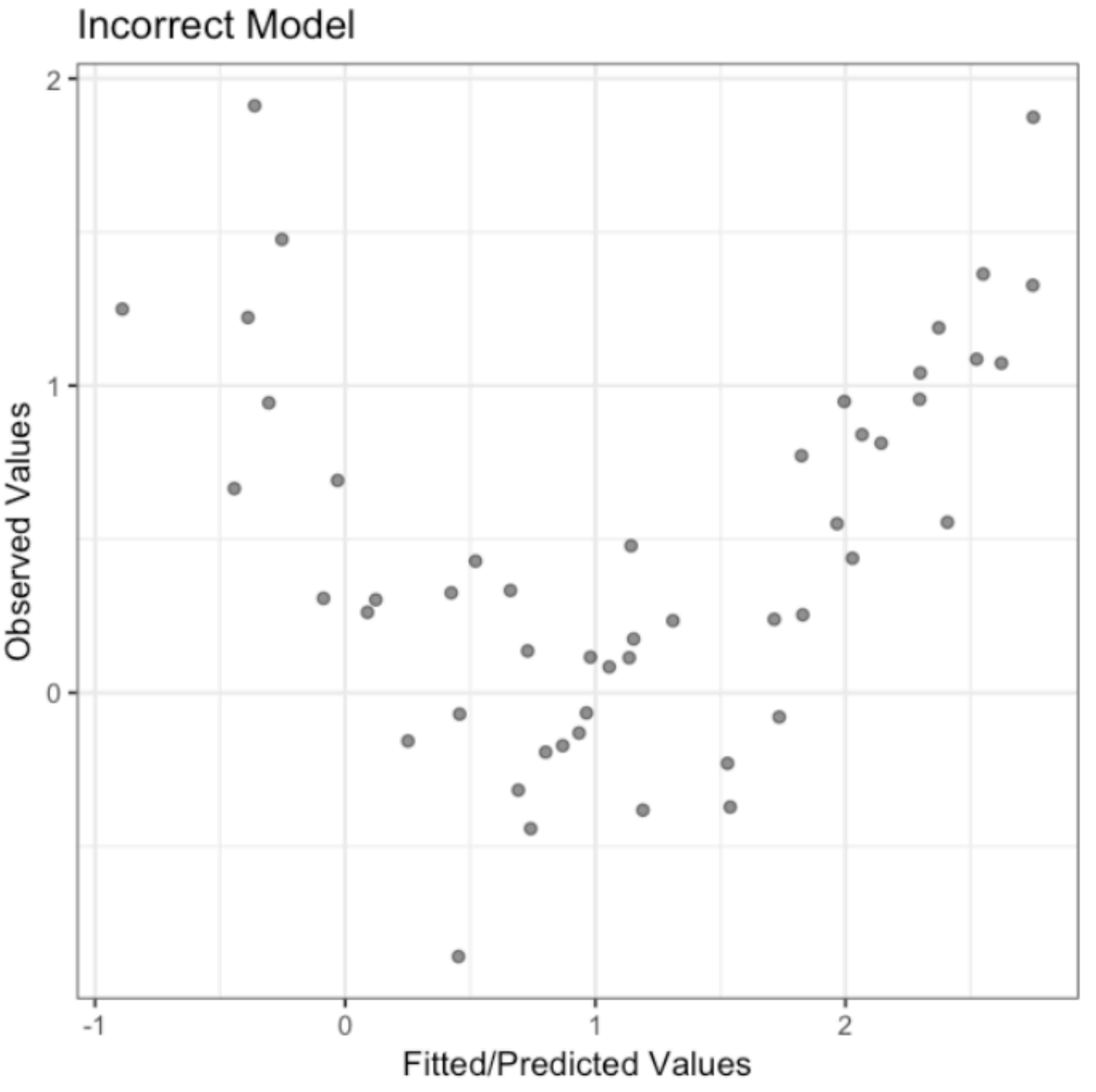


Photo by Ryan Stone on Unsplash

Fitted/Predicted vs. observed value plots


Correct Model
Observed Values
Fitted/Predicted Values
3
2
1
0
-1
-1
0
1
2


Incorrect Model
Observed Values
Fitted/Predicted Values
2
1
0
-1
0
1
2

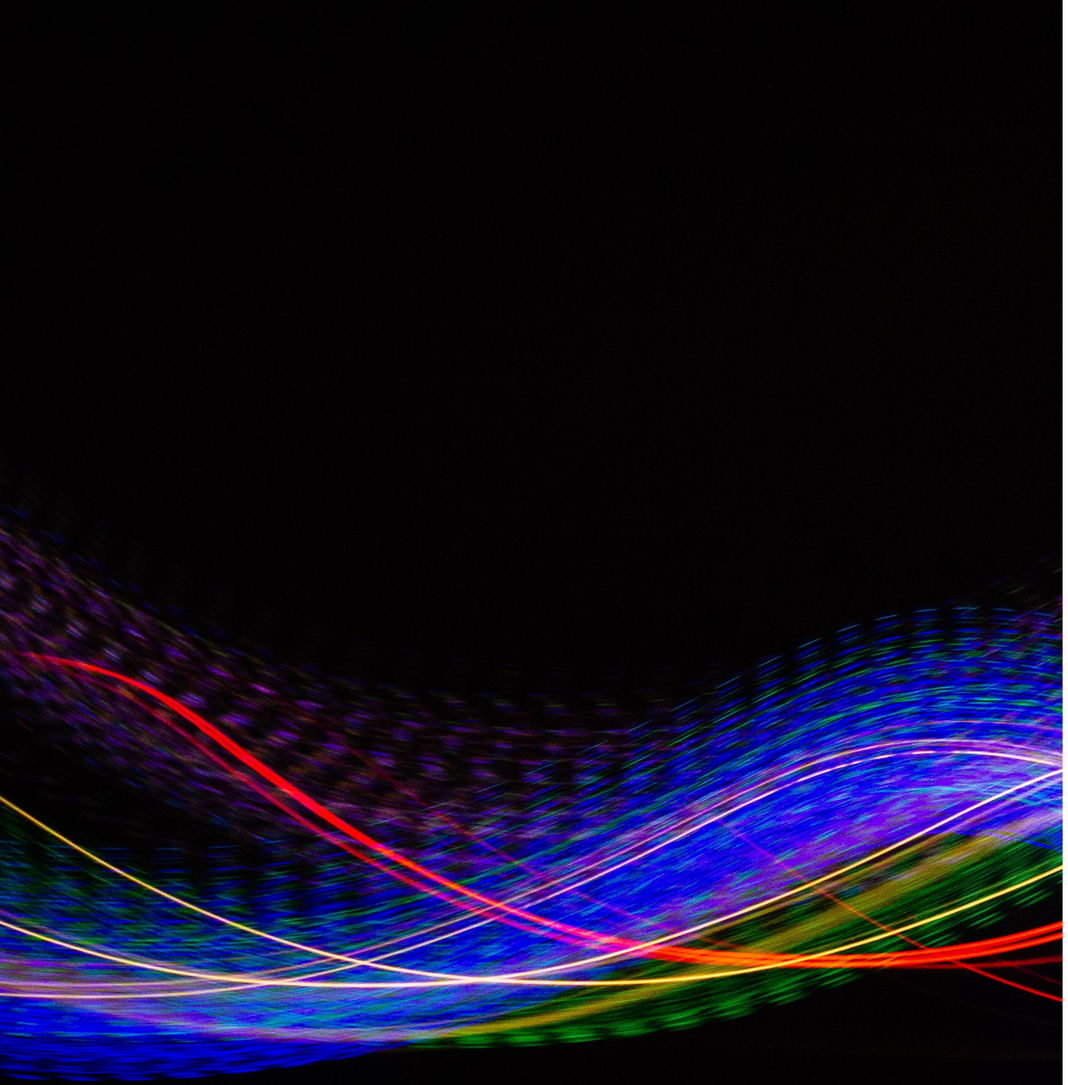
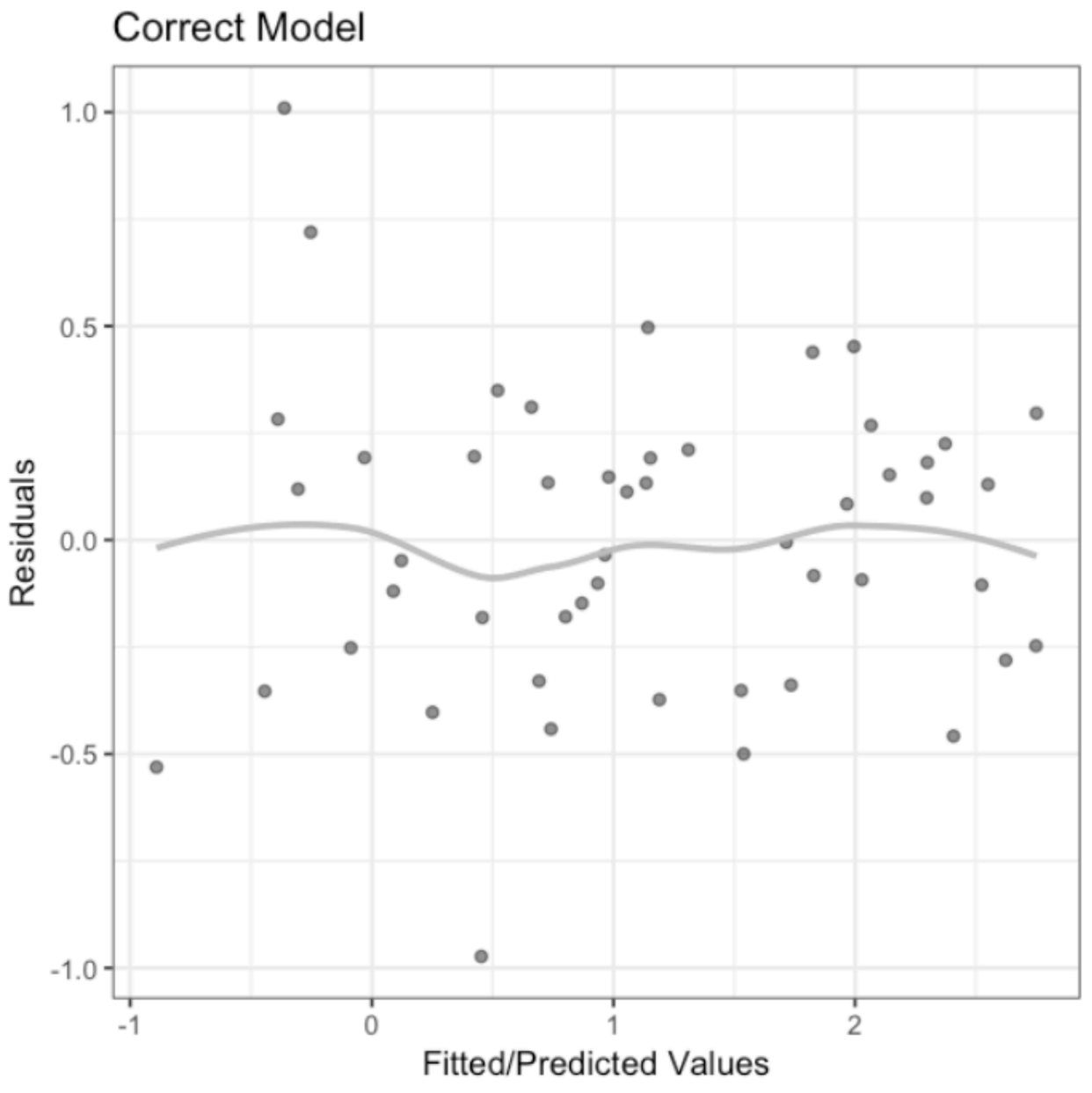
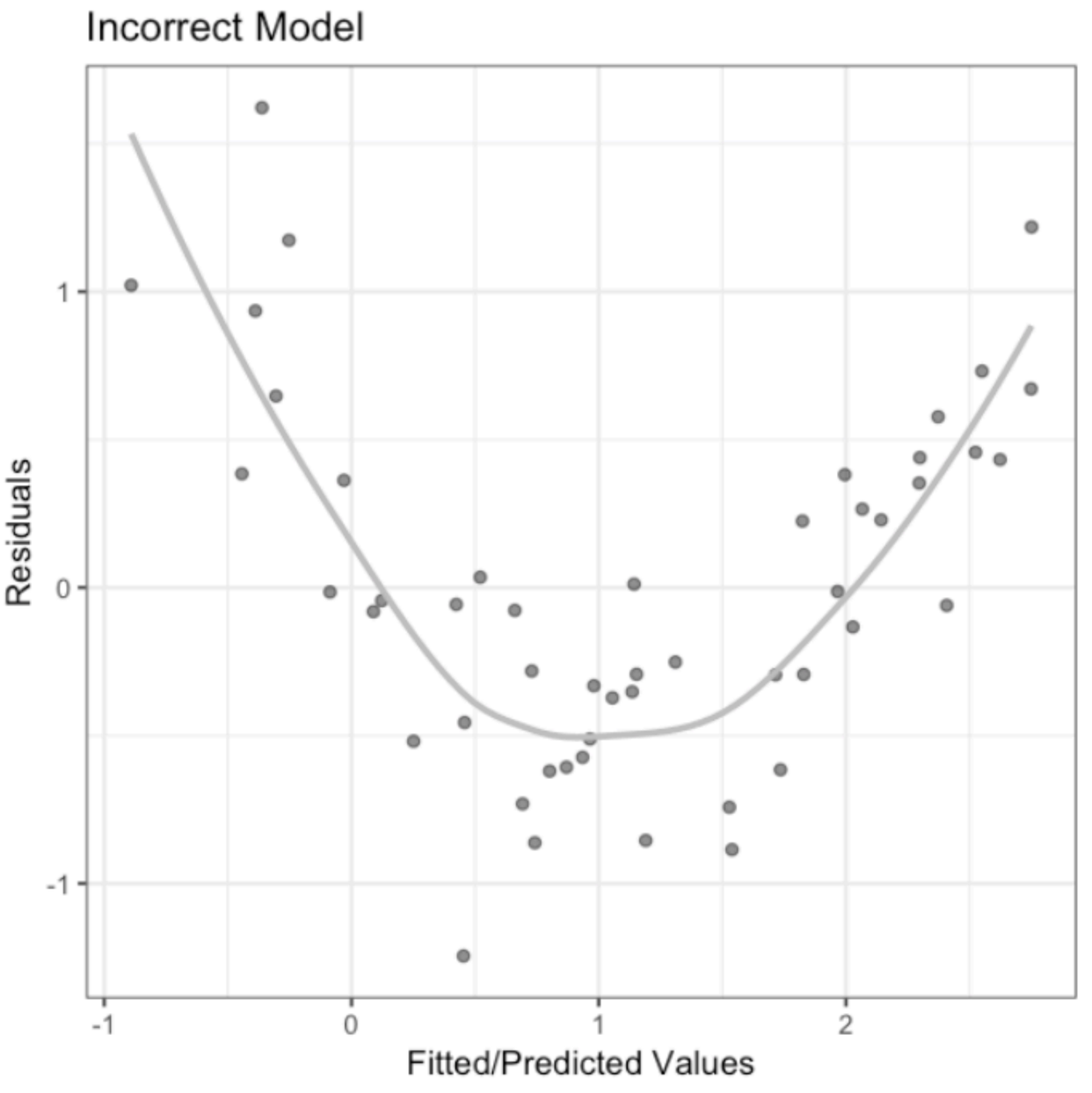


Photo by Ryan Stone on Unsplash

Residual vs. fitted/predicted value plots


Correct Model
Residuals
1.0
0.5
0.0
-0.5
-1.0
-1
0
1
2
Fitted/Predicted Values


Incorrect Model
Residuals
1
0
-1
-1
0
1
2
Fitted/Predicted Values

Lesson: Violations of the independence assumption

Module: Regression Diagnostics

Photo by Janine Robinson on Unsplash

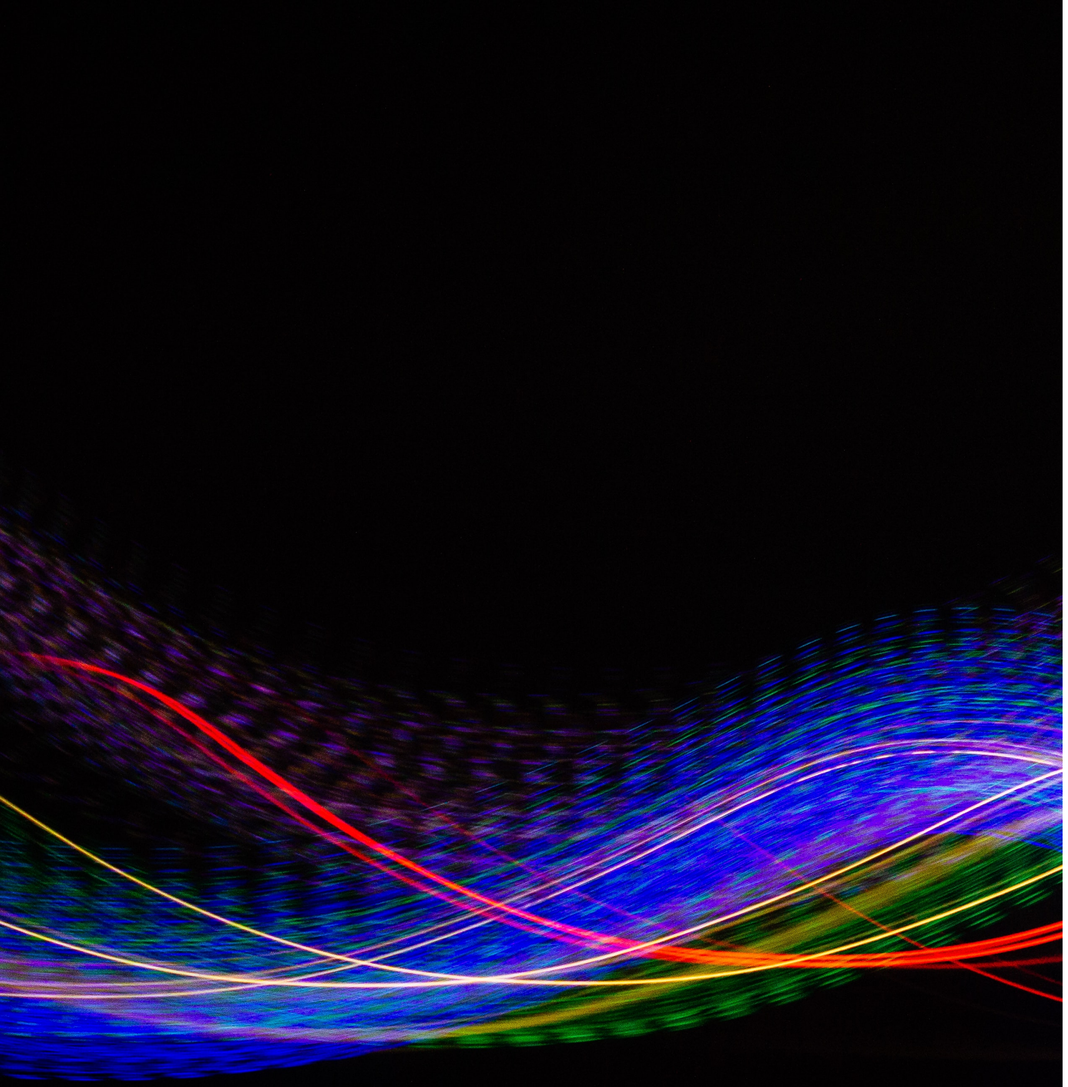


Photo by Ryan Stone on Unsplash

Uncorrelated/Independent Errors. This violation might occur, and be serious, when measuring ***spatial*** or ***temporal*** data.

1. Residuals vs time/index.

2. Successive residual plot.

3. Durbin-Watson test.

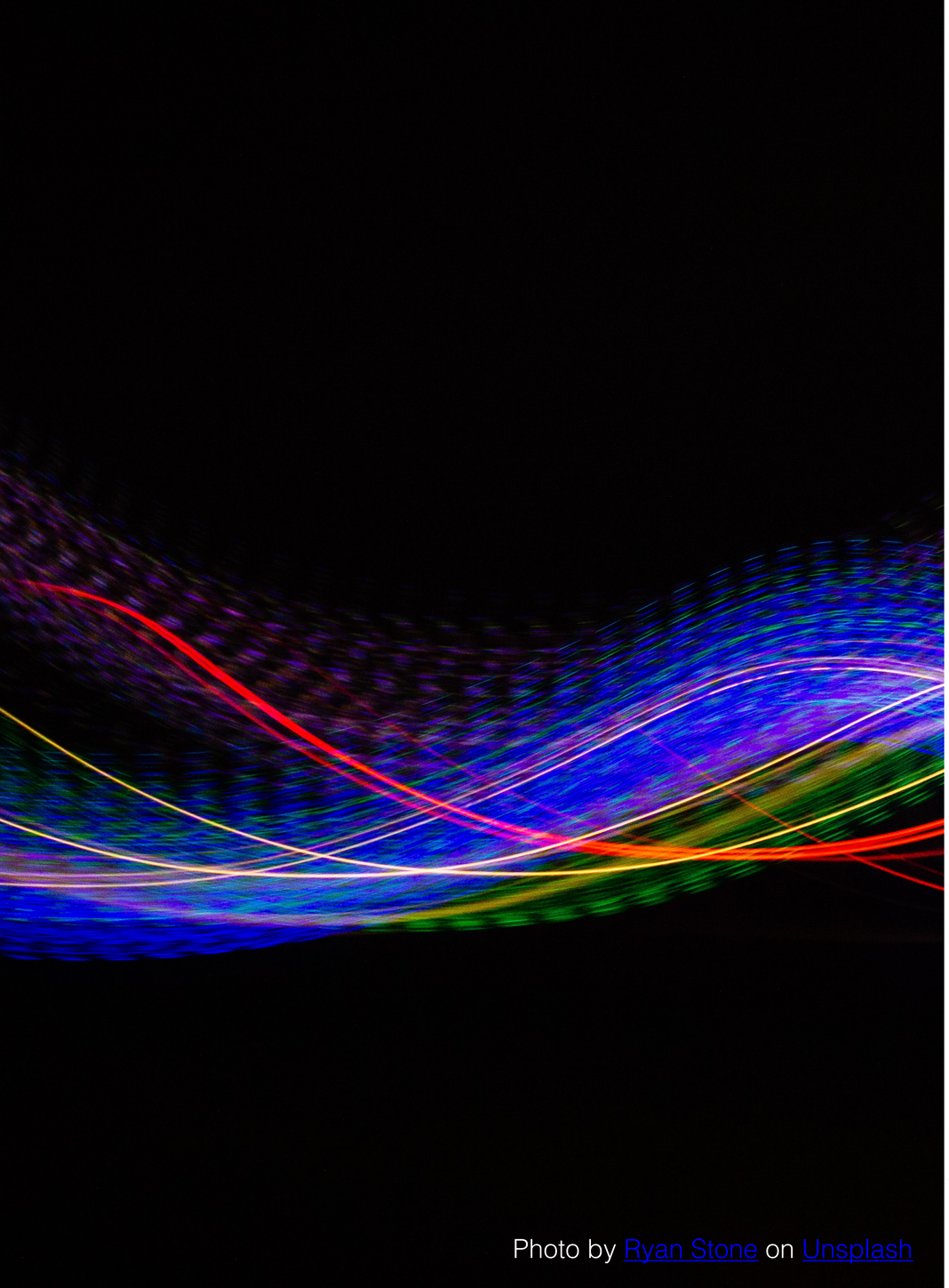
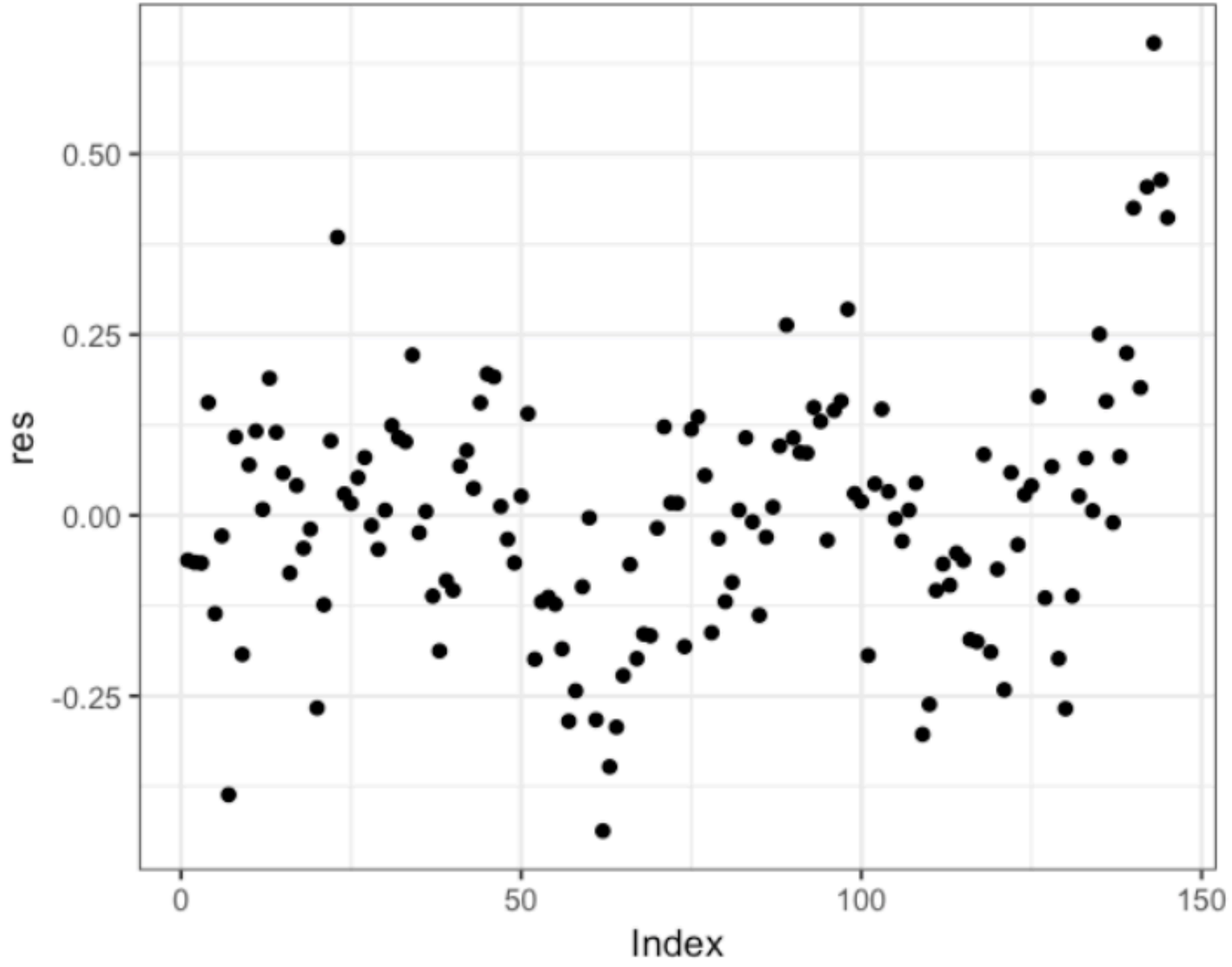


Photo by Ryan Stone on Unsplash

Residual vs index plot



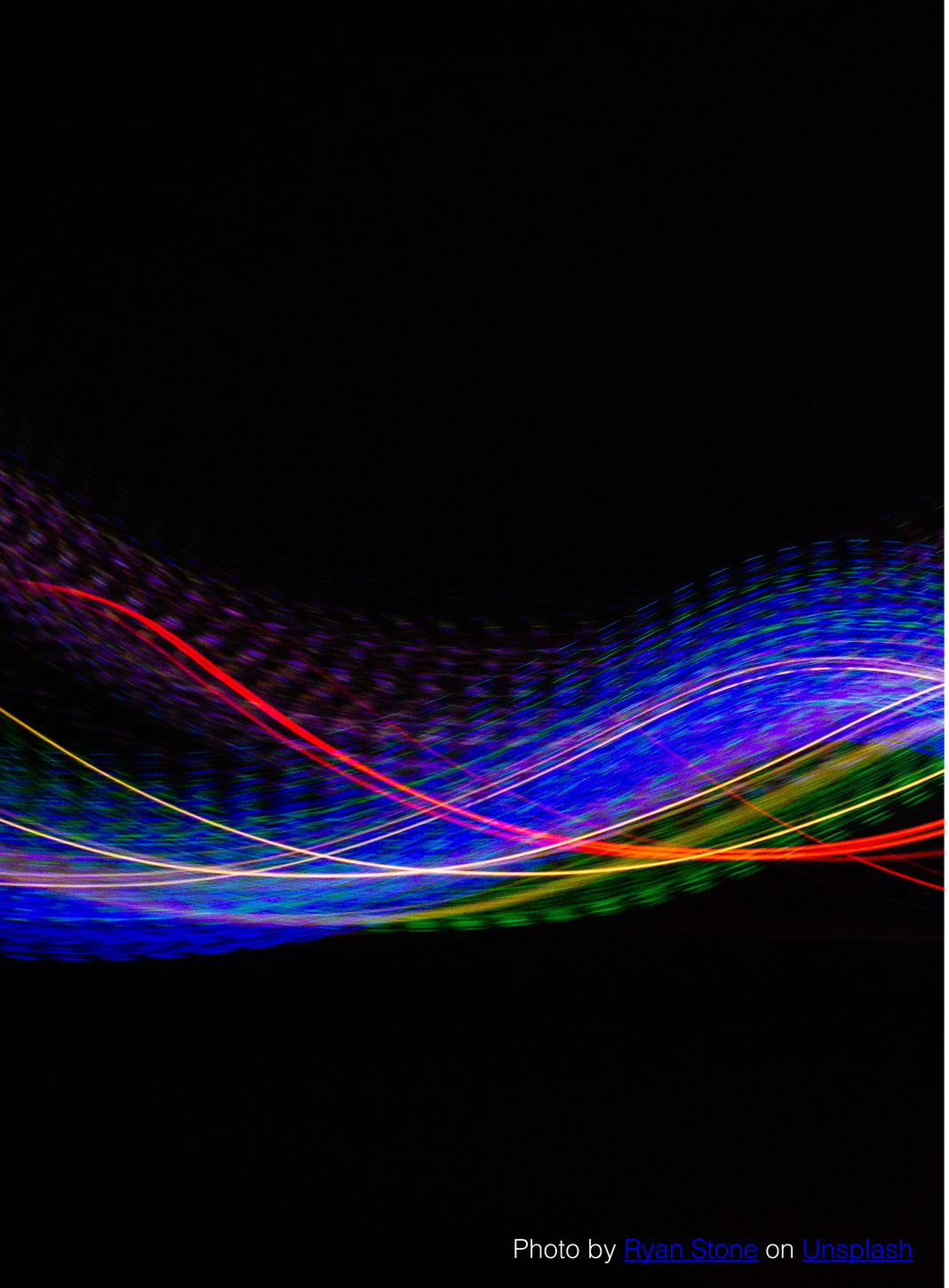
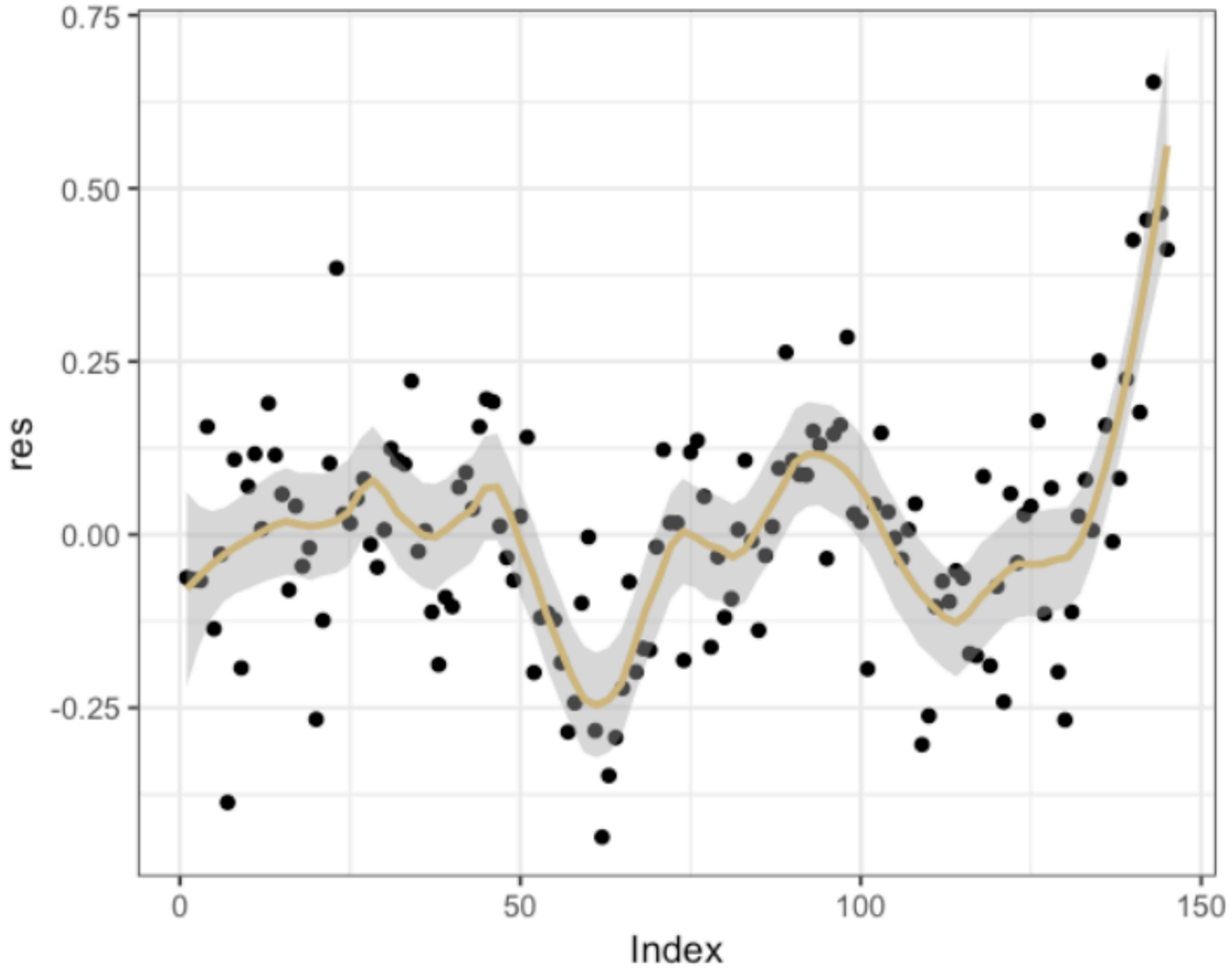


Photo by Ryan Stone on Unsplash

Residual vs index plot



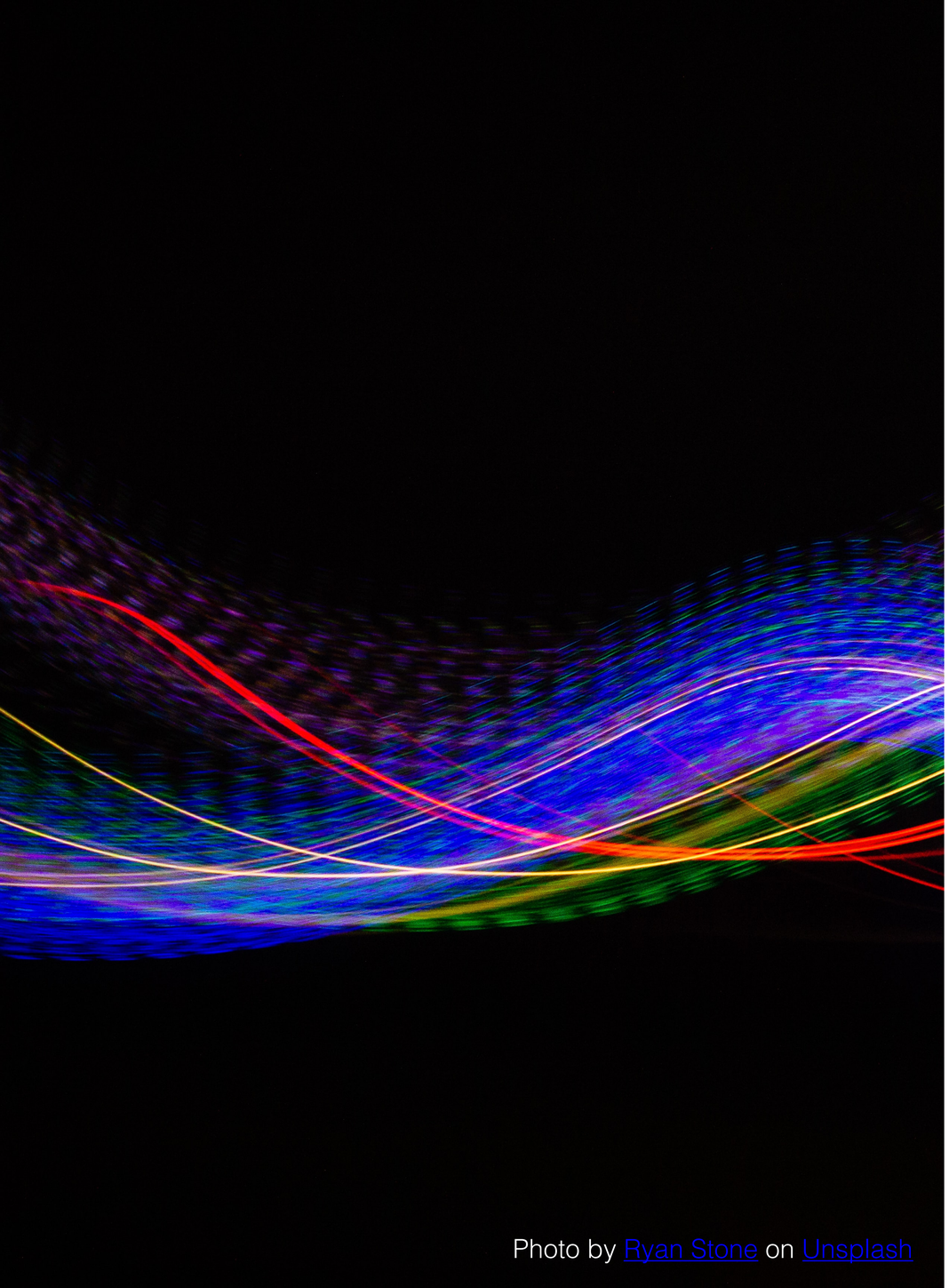
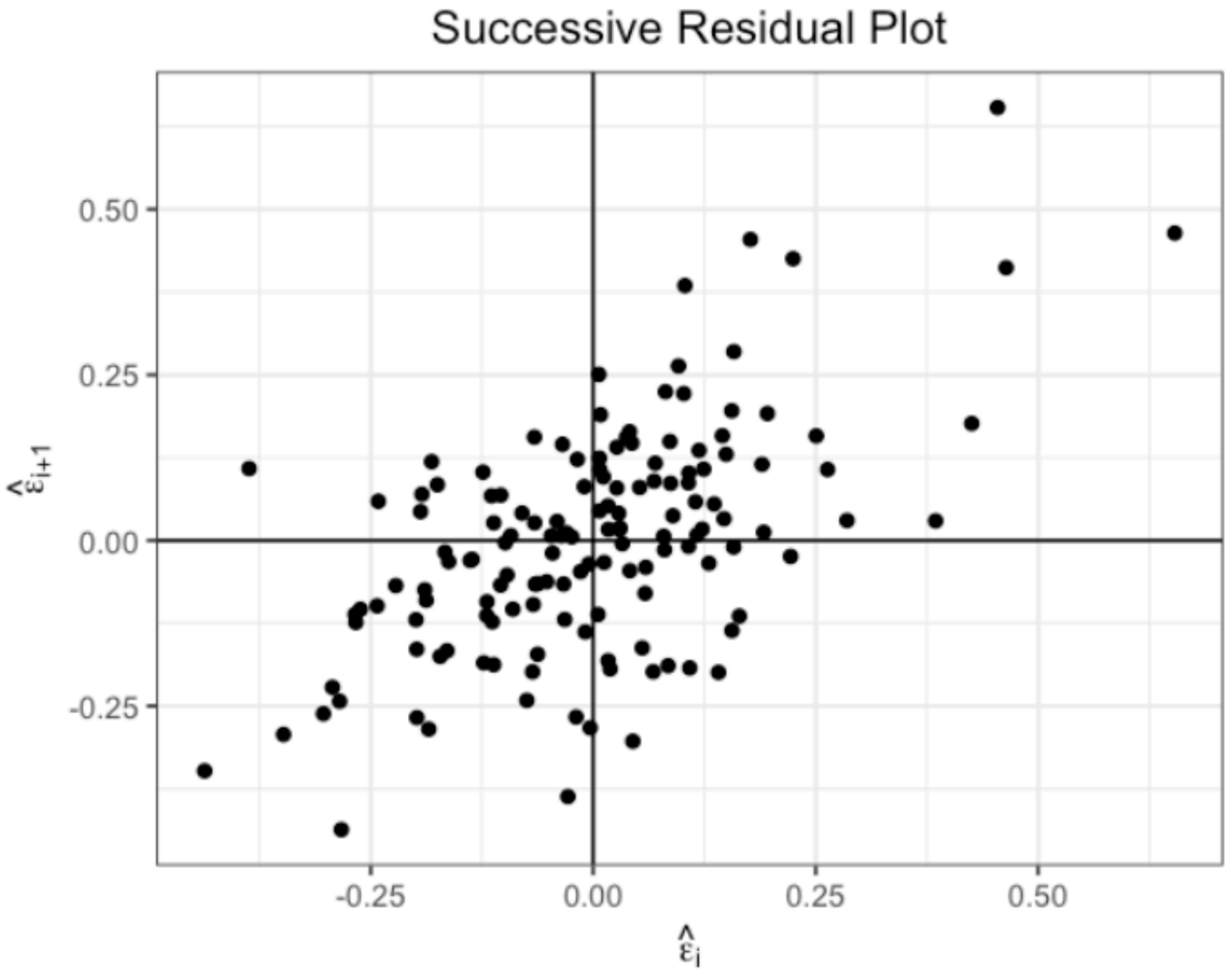


Photo by Ryan Stone on Unsplash



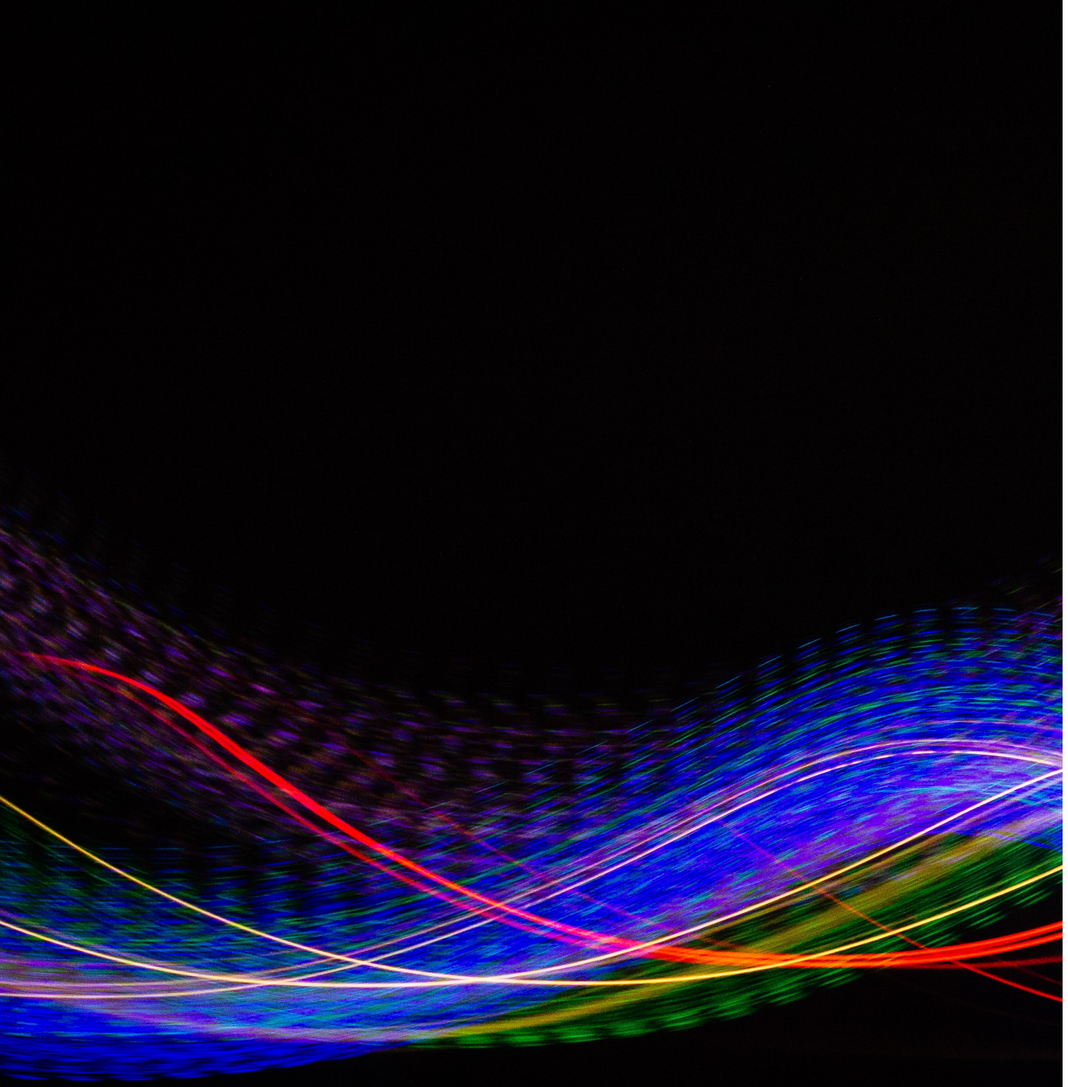


Photo by Ryan Stone on Unsplash

What solutions are available for violations of this assumption?

1. Generalized least squares

2. Time Series/Spatial Statistics

Lesson: Violations of the constant variance assumption

Module: Regression Diagnostics

Photo by Janine Robinson on Unsplash

Constant Variance Assumption. Recall that one of our assumptions is:

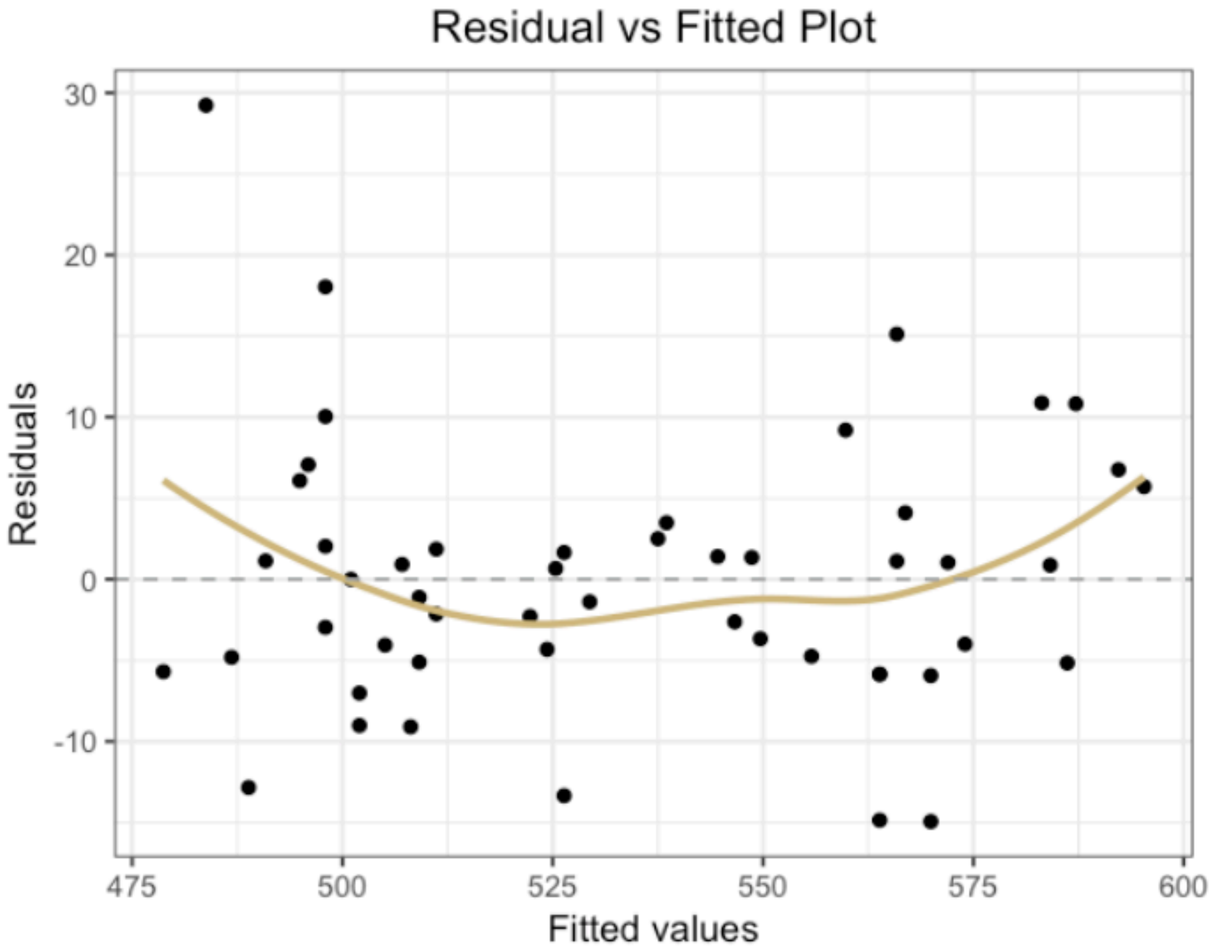
Even under this assumption, the residuals—our estimator of the error—are not constant variance:

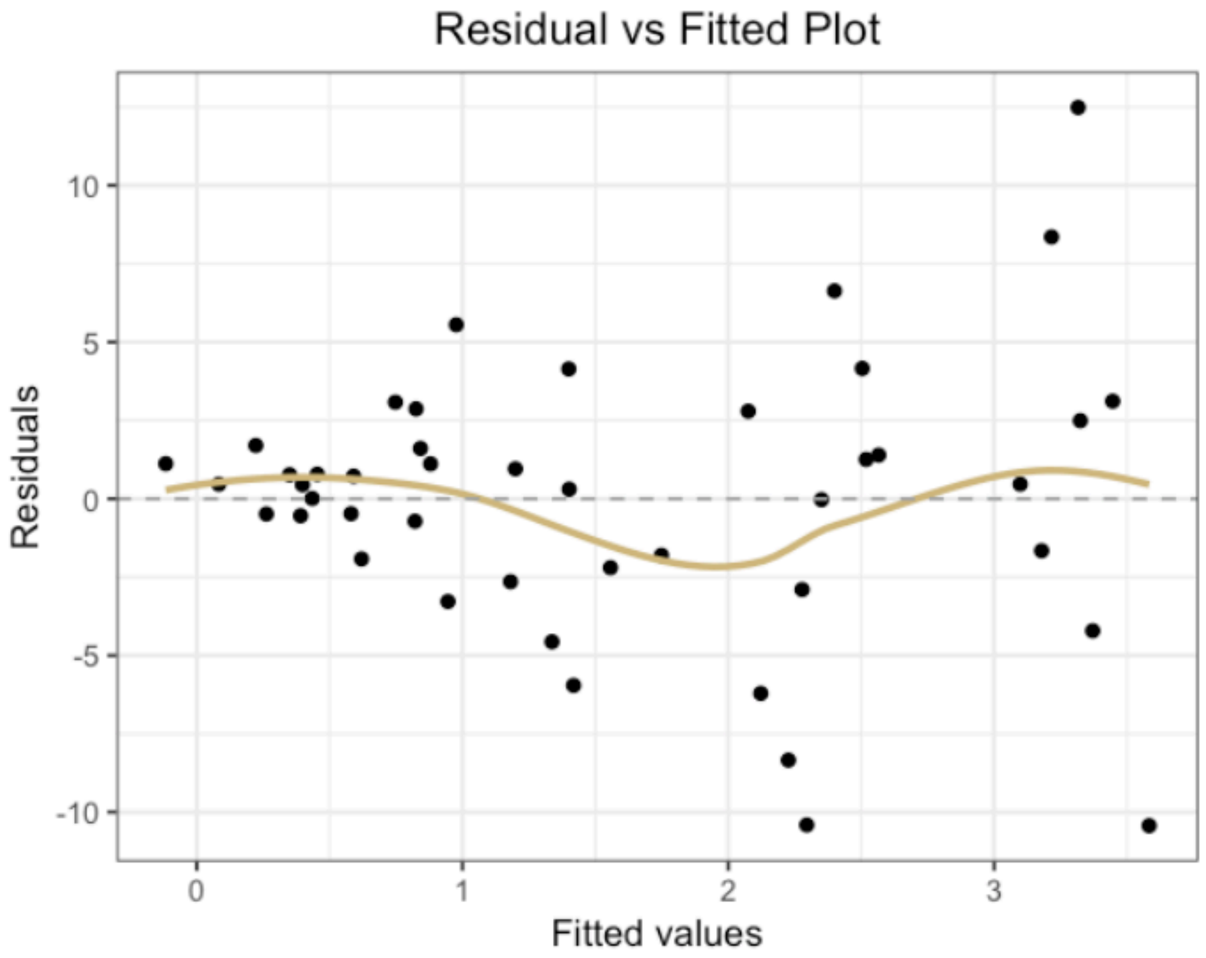
Constant Variance Diagnostics:

1. Residual vs. Fitted plot.

2. Residual vs. potential predictor plot.

Checking Error Assumptions





Solutions to non constant variance:

1. Transformation of the response.

2. Weighted/Generalized Least Squares.

Lesson: Violations of the normality assumption

Module: Regression Diagnostics

Photo by Janine Robinson on Unsplash

(Height of Hallgrímskirkja in Reykjavik, Iceland)

Normality Assumption.

To diagnose deviations from normality, we can look at:

1. QQ plots
2. Residual vs fitted plots
3. Shapiro-Wilk test for normality

(Height of Hallgrímskirkja in Reykjavik, Iceland)

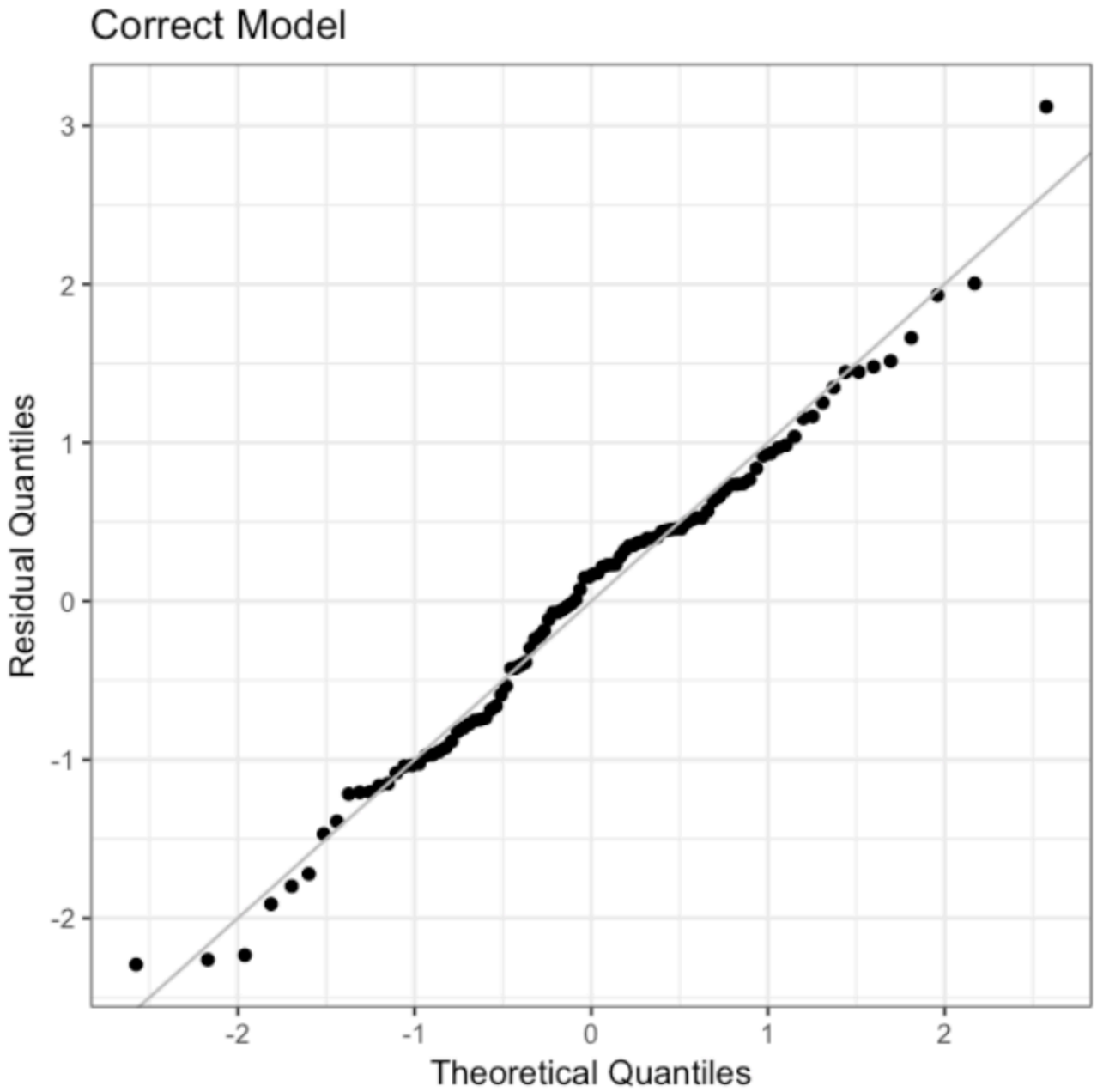


Photo by Cassidy Dickens on Unsplash

(Height of Hallgrímskirkja in Reykjavik, Iceland)

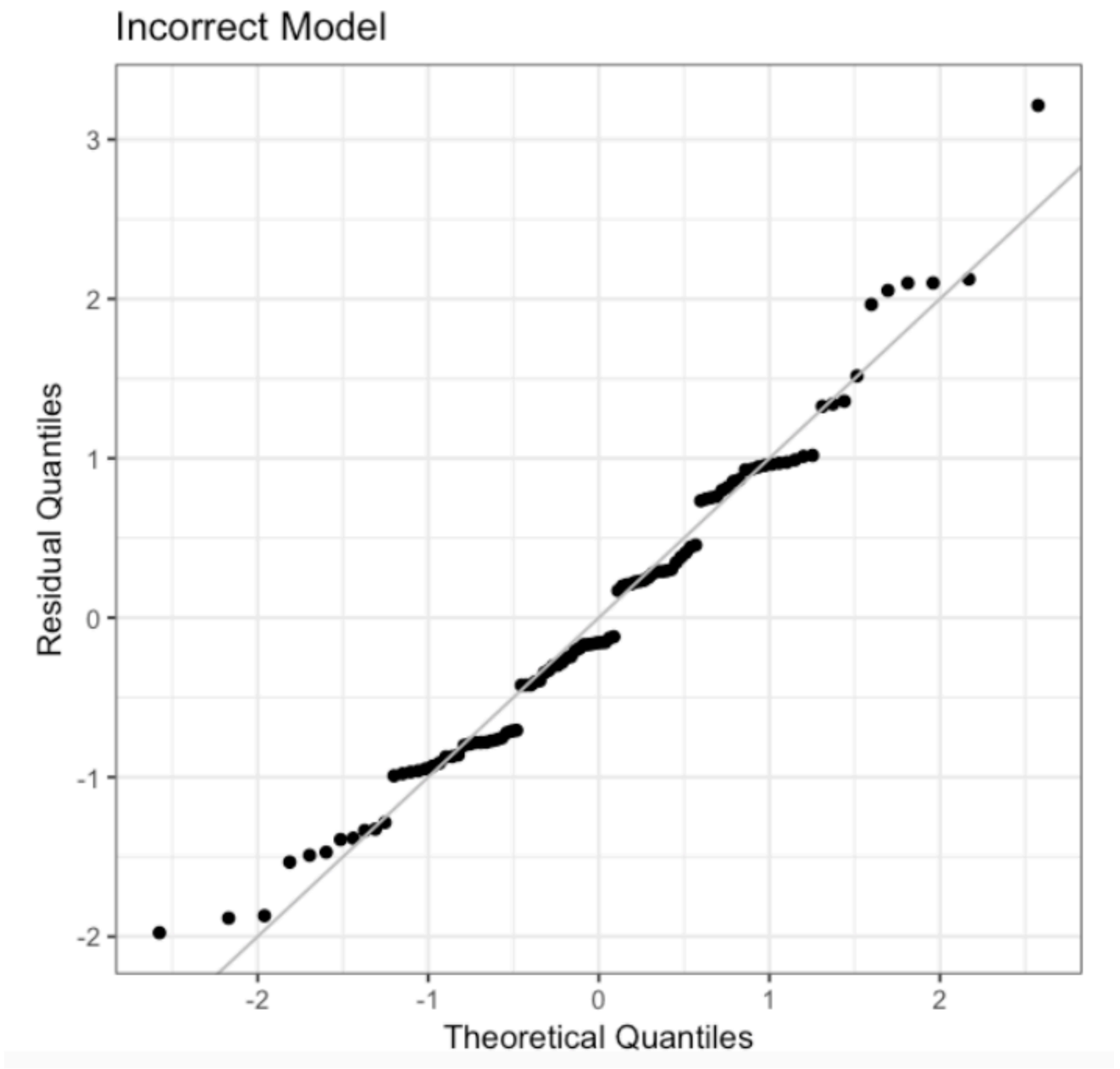


Photo by Cassidy Dickens on Unsplash

(Height of Hallgrímskirkja in Reykjavik, Iceland)

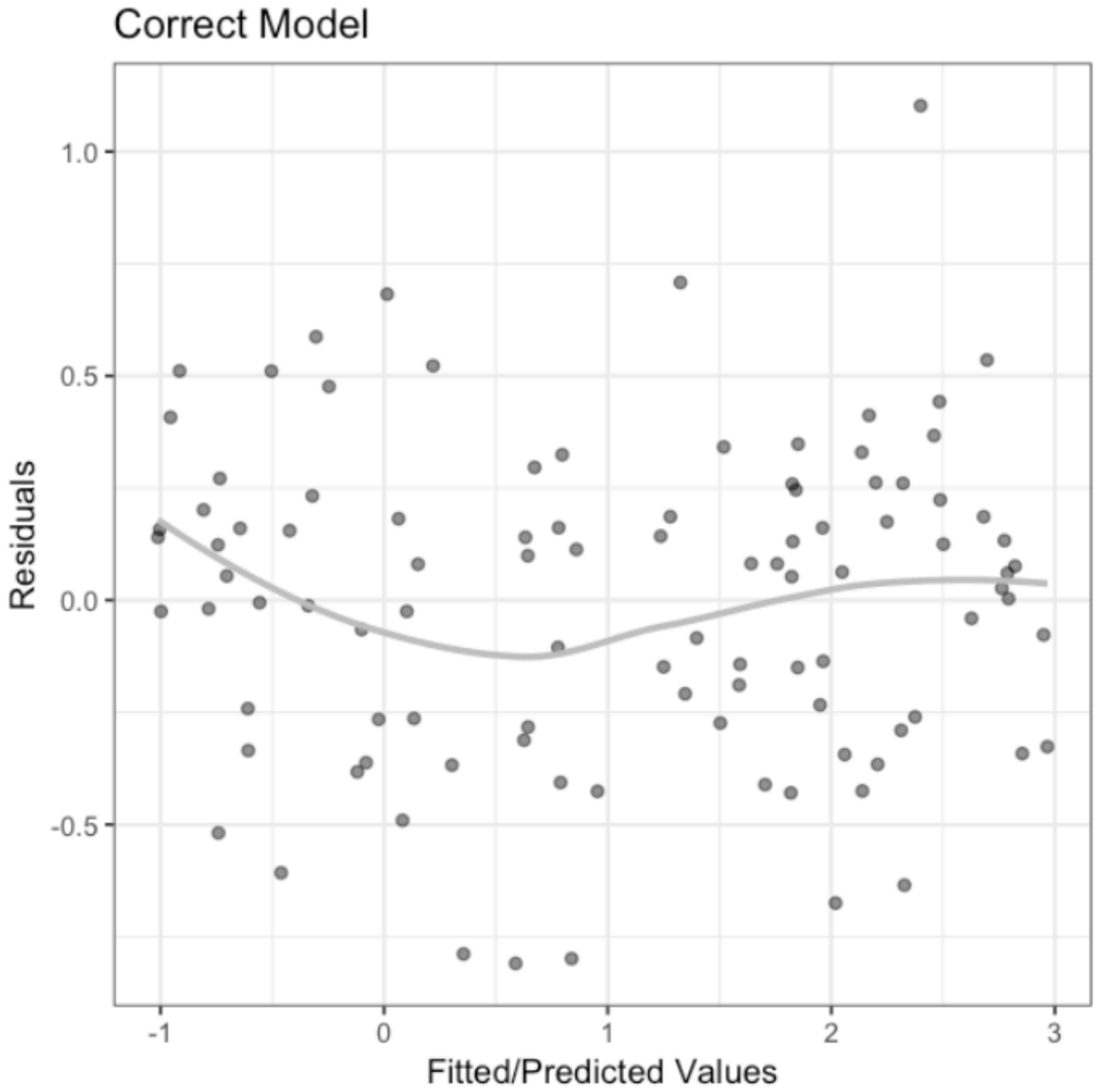


Photo by Cassidy Dickens on Unsplash

(Height of Hallgrímskirkja in Reykjavik, Iceland)

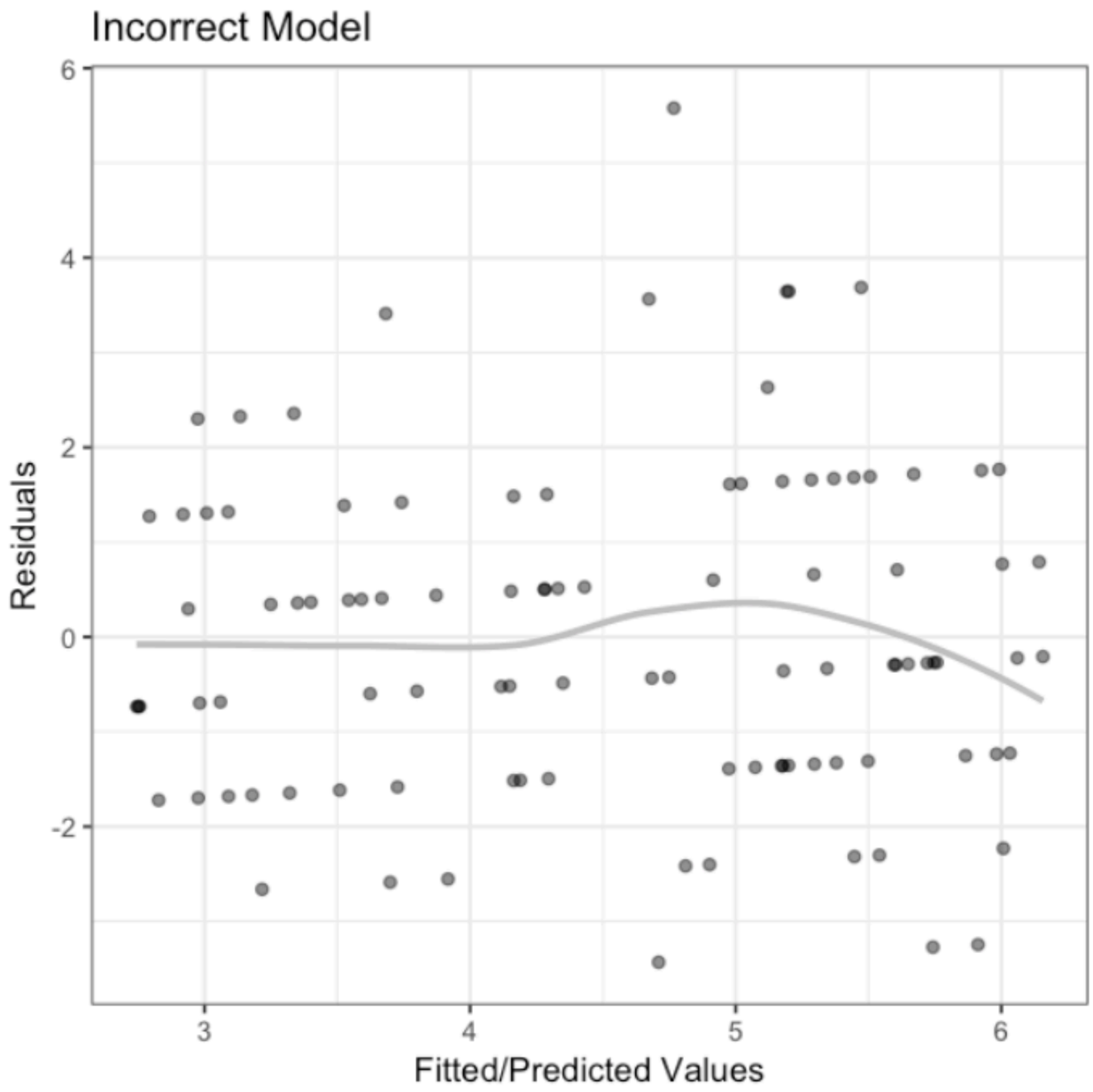


Photo by Cassidy Dickens on Unsplash

H_0: The sample (residuals) came from a normal distribution

H_1: The sample (residuals) came from a non-normal distribution

```
shapiro.test(df.diagnostics$r_correct)
```

```
        Shapiro-Wilk normality test

data:  df.diagnostics$r_correct
W = 0.98711, p-value = 0.4449
```

Photo by Cassidy Dickens on Unsplash

(Height of Hallgrímskirkja in Reykjavik, Iceland)

Solutions:

1. Transformations.

2. *Generalized linear models (GLMs)*

Photo by Cassidy Dickens on Unsplash