Extensions of the Linear Model

AU STAT-615

Emil Hvitfeldt

2021-04-07

Extensions of the Linear Model

Why?

Standard linear regression model provides interpretable results and works quite well on many real-world problems

However, using such a model makes strong assumptions:

The relationship between predictors and response are

- Additive
- Linear

Removing the additive assumption

Let's consider the linear regression model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 + \varepsilon$$

This model does only have additive effects

Removing the additive assumption

One way of relaxing the additive assumption and allow for an interaction term is by

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \varepsilon$$

This interaction effect enables the following

$$Y = \beta_0 + (\beta_1 + \beta_3 X_2) X_1 + \beta_2 X_2 + \varepsilon$$
$$Y = \beta_0 + \tilde{\beta}_1 X_1 + \beta_2 X_2 + \varepsilon$$

Removing the additive assumption

Now the effect of X_1 is no longer constant

Adjusting X_2 will change the impact of X_1 on Y.

It is sometimes the case that an interaction term has a very small p-value but the main associated effects do not

Hierachical Principle

If we include an interaction term in a model, we should also include the main effects, even if the p-values associated with their coefficients are not significant

If the p-value associated with X_1 and X_2 are not very small we should not worry and we should include them if the p-value associated with X_1X_2 is very small

Non-linear relationships

The reality may be that the relationship between the response and predictors is non-linear

One of the ways we have looked at is to do polynomial regression

Polynomial Regression

Turning

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

into

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \varepsilon$$

Which is extensible to

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \ldots + \beta_p X^p + \varepsilon$$

(although we rarely use $p > 3$)

Polynomial Regression

Matrix notation

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} \qquad \mathbf{Y} = \begin{bmatrix} 1 & X_1 & X_1^2 & \dots & X_1^p \\ 1 & X_2 & X_2^2 & \dots & X_2^p \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & X_n & X_n^2 & \dots & X_n^p \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

Polynomial Regression

Assuming $n > p$, since $\mathbf{X}$ is a Vandermonde matrix, the invertibility condition is guaranteed to hold if all the X_i values are distinct and we get a unique least-squares solution

Potential Problems

When we fit a linear regression model to a particular data set, many problems may occur

Non-normality

The Relationship between Y and X is not linear

Indicator

- QQ-plot
- Shapiro–Wilk test
- Skewness, Kurtosis
- Histogram; Boxplot

Remedy

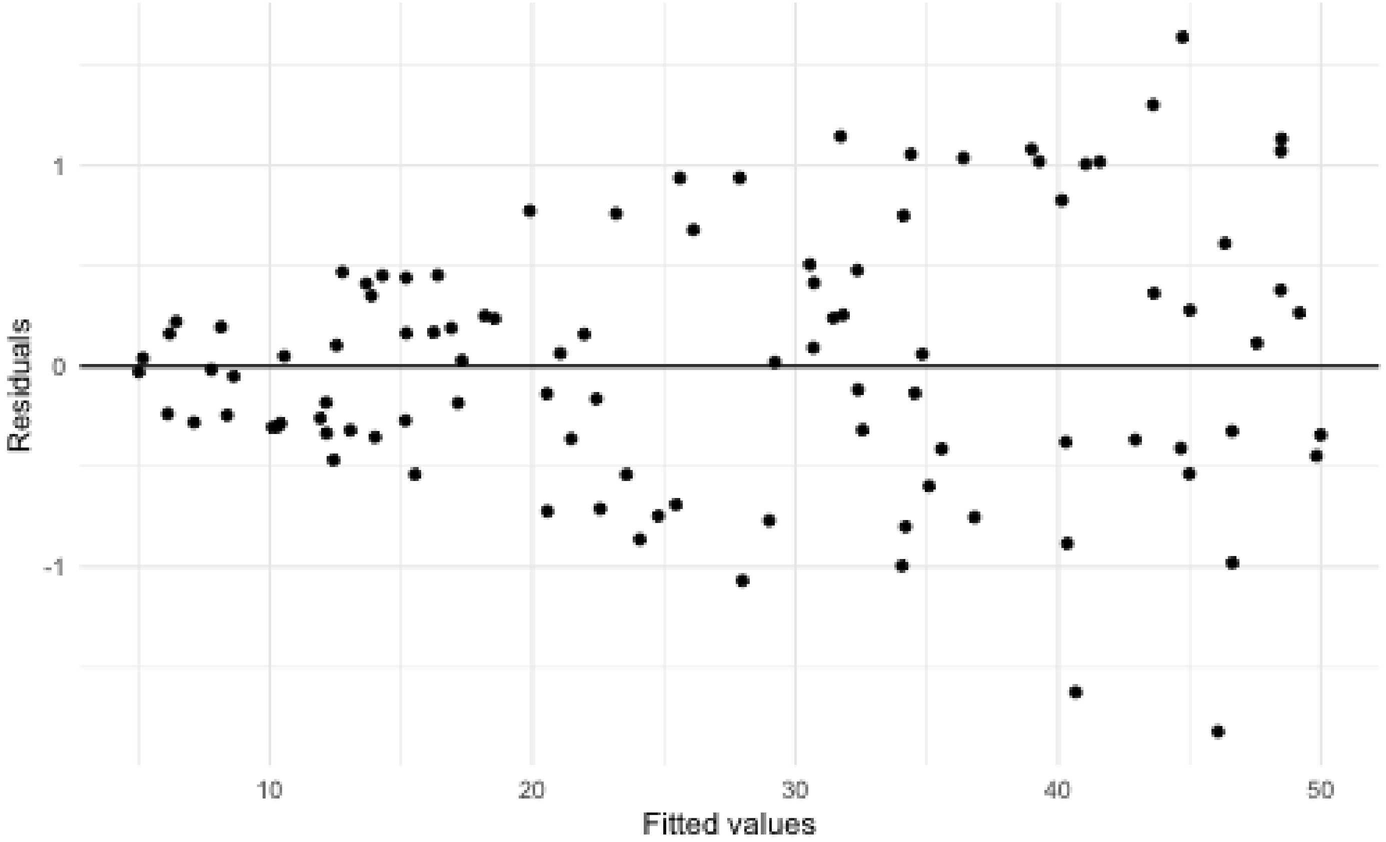
Transformations

or

non-parametric metrics

Heteroscedasticity

Non-equal variance



Heteroscedasticity

Indicator

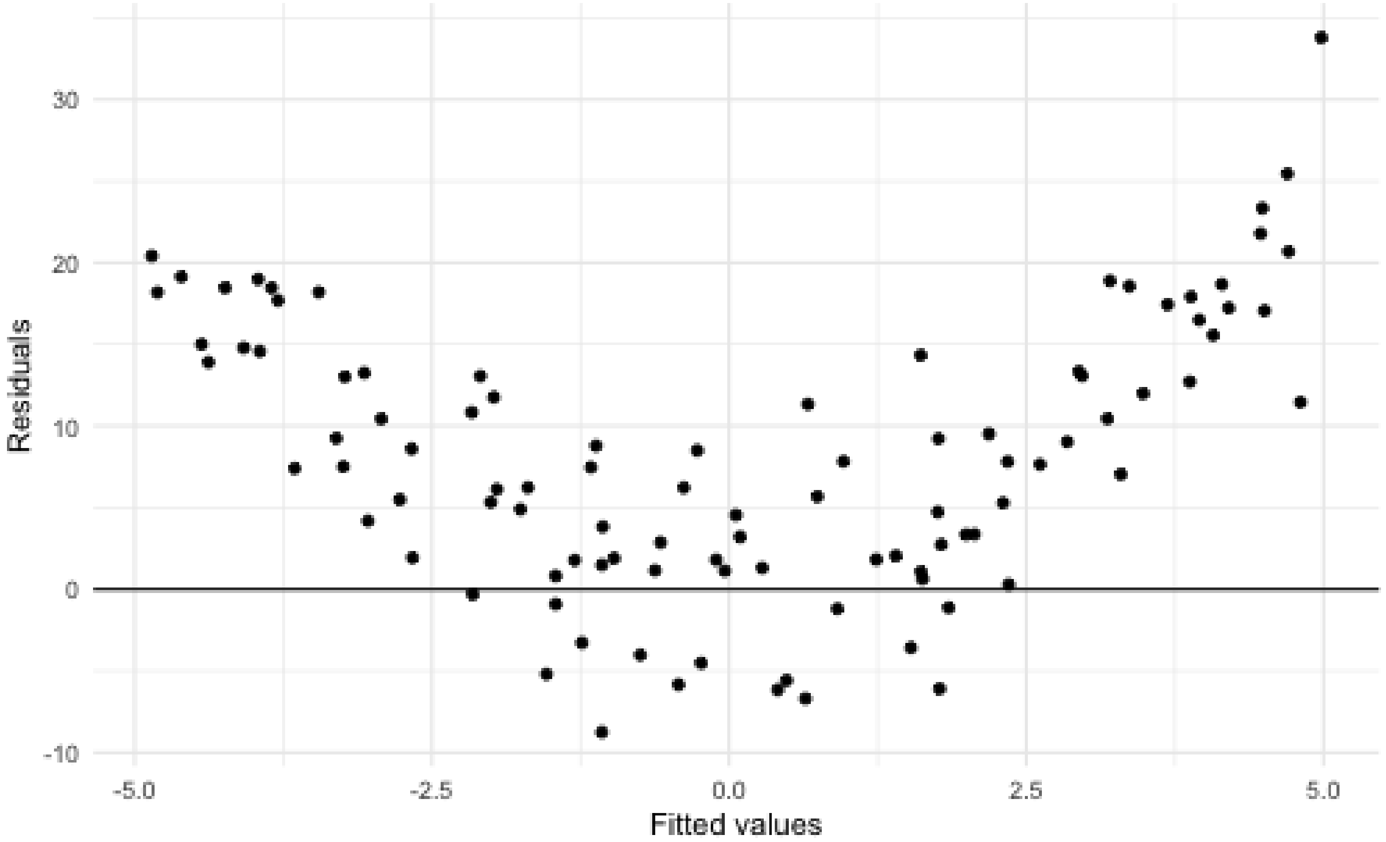
- Residual plots
- Levene
- Breusch-Pagan test

Remedy

Apply transformations to Y, such as $\log Y$ or $\sqrt{Y}$

or do weighted least squares

Linearity



Linearity

Indicator

- Residual plots
- Lack-of-fit test

Remedy

- Add predictors
- Use non-linear transformation of the predictors such as $\log Y$, $\sqrt{Y}$, or X^2

Indepence

Correlation of error terms

The error terms $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ should be correlated

if they are correlated then we may have an unwarranted sense of confidence in our model (narrower confidence bands)

Indepence

Indicator

- Residual plots, look for trends

Remedy

- Fit time series models
- Improve experimental design

Outliers / high leverage point

An **outlier** is a point for which Y_i is far from the value predicted by the model

A **high leverage point** point is a point that has extreme predictor values

high leverage observations tend to have a sizeable impact on the estimated regression line

Outliers / high leverage point

Indicator

- Residual plots
- Studentized residuals plots

Remedy

- Find the reason why they are the way they are
- Delete or reweight (you need a good reason to do this)

Collinearity

Two or more predictor variables are closely related to one another

Collinearity

Indicator

Look at the correlation matrix

Access multicollinearity by computing the variance inflation factor (VIF)

VIF is the ratio of the variance of $\hat{\beta}_j$ when fitting full model divided by variance of $\hat{\beta}_j$ if fit on its own

The smallest value is 1 which is great

if VIF > 10 we have problems

Collinearity

Remedy

- Variable selection
- Ridge regression