PSY 503: Foundations of Statistical Methods in Psychological Science

Linear Models and their assumptions

Suyog Chandramouli

Zoom & 411 PSH (Princeton University)

2nd December, 2024

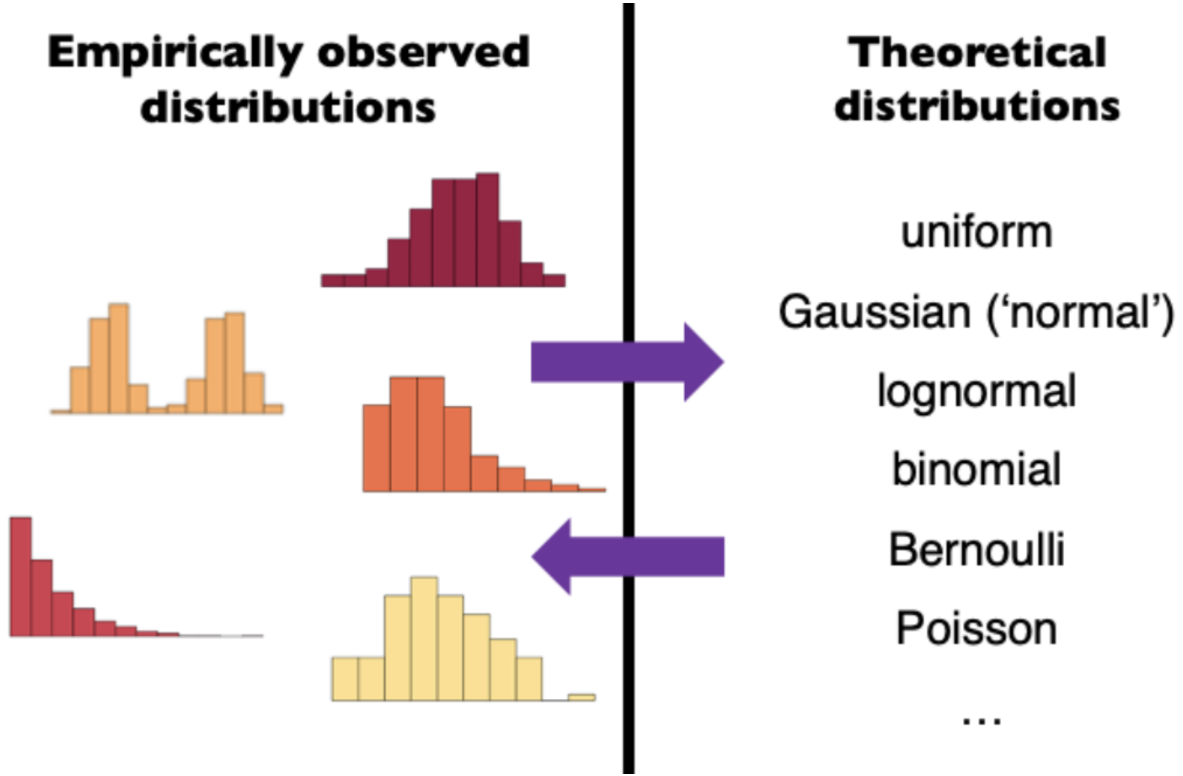
What is a model?

- Models are simplifications of things in the real world

What is a statistical modelling?

- **Statistical modeling** = “making **models** of **distributions**”

  (coming up with a plausible data generating process/ DGP)



All models are approximations. Essentially, all models are wrong, but some are useful. However, the approximate nature of the model must always be borne in mind.

— George E. P. Box —

[redacted]

Models as Golems

- Golem = animated human-like being, made from inanimate matter such as clay or mud (Clay robots)
- Powerful but mindless servants
  - Servant when used well
  - Dangerous because they follow instructions literally (no wisdom, no foresight)
- In some versions, Rabbi Judah Loew ben Bezalel built a golem to protect. But he lost control, causing innocent death

Statistical Golems

Statistical (and scientific) models are our golems

- We build them from basic parts
- They are powerful—we can use them to understand the world and make predictions
- They are animated by “truth” (data), but they themselves are neither true nor false
- The model describes the golem, not the world
    - The model doesn’t describe the world or tell us what scientific conclusion to draw—that’s on us

- We need to be careful about how we build, interpret, and apply models!

Statistical Golems

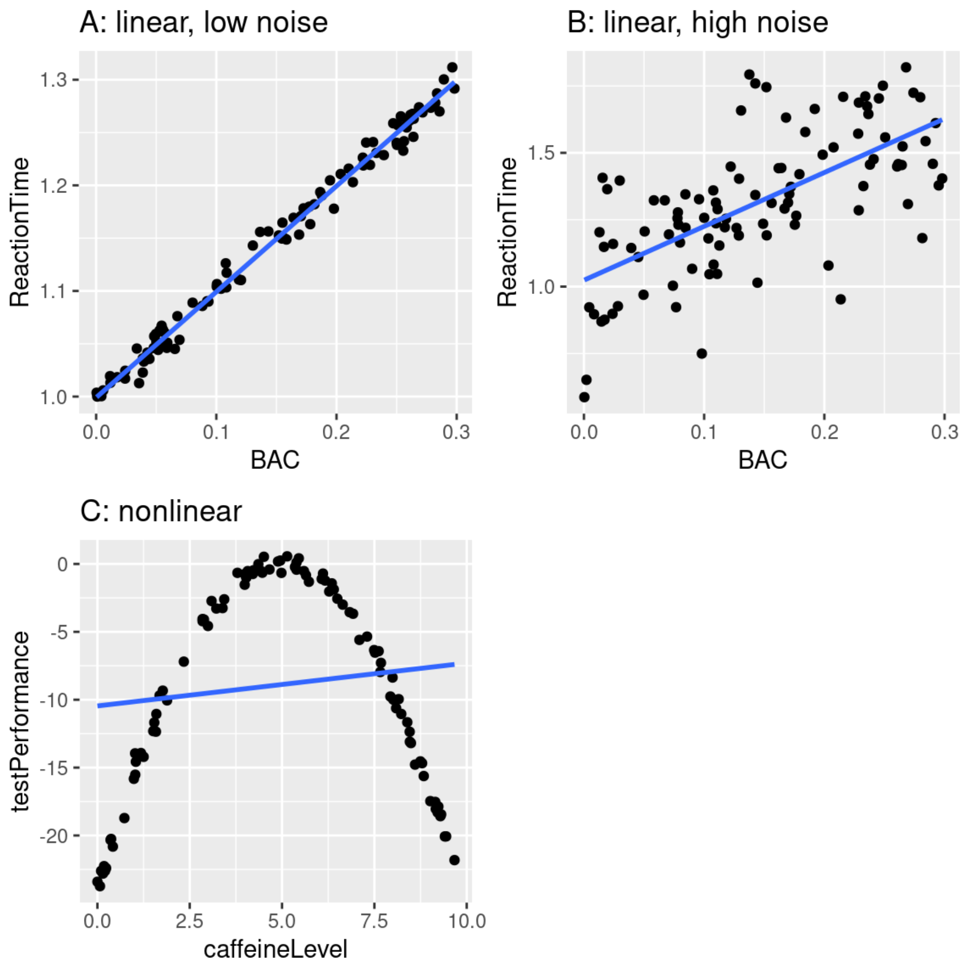
Statistical (and scientific) models are our golems

- We build them from basic parts
- They are powerful—we can use them to understand the world and make predictions
- They are animated by "truth" (data), but they themselves are neither true nor false
- The model describes the golem, not the world
    - The model doesn't describe the world or tell us what scientific conclusion to draw—that's on us

- We need to be careful about how we build, interpret, and apply models!

What Makes a Model "Good"

- We want it to describe our data well
- We want it to generalize to new datasets
- We want error to be as small as possible



Models assumptions

Model’s assumptions

- Foundation upon which its validity and usefulness rest
  - When assumptions are met, the model is more reliable and useful.
  - When assumptions are violated, the model's conclusions become questionable.

Model’s assumptions

- Foundation upon which its validity and usefulness rest
  - When assumptions are met, the model is more reliable and useful.
  - When assumptions are violated, the model's conclusions become questionable.

- **Checking model assumptions**
  - focuses specifically on verifying whether the fundamental assumptions underlying the chosen model are met by the data.

Model’s assumptions

- Foundation upon which its validity and usefulness rest
  - When assumptions are met, the model is more reliable and useful.
  - When assumptions are violated, the model's conclusions become questionable.

- **Checking model assumptions**
  - verifying whether the fundamental assumptions underlying the chosen model are met **by the data**.

4 assumptions made by regression models

4 Assumptions

- Linearity
- Constant Variance
- Normality
- Independence

All four assumptions are about the noise (ε)

- **Linearity**: ε contains no patterns - just noise
- **Independence**: Each ε is its own random draw
- **Normality**: ε follows a specific noise distribution
- **Homogeneity of Variance**: ε has consistent noisiness

All four assumptions are about the noise (ε)

- **Linearity**: ε contains no patterns - just noise
- **Independence**: Each ε is its own random draw
- **Normality**: ε follows a specific noise distribution
- **Homogeneity of Variance**: ε has consistent noisiness
  (equality of variance)
  (no heteroscedasticity)
  (constant variance)

All four assumptions are about the noise (ε)

- **Linearity**: ε contains no patterns - just noise
- **Independence**: Each ε is its own random draw
- **Normality**: ε follows a specific noise distribution
- **Homogeneity of Variance**: ε has consistent noisiness
  (equality of variance)
  (no heteroscedasticity)
  (constant variance)

It's All About the Errors

- Good model = boring residuals

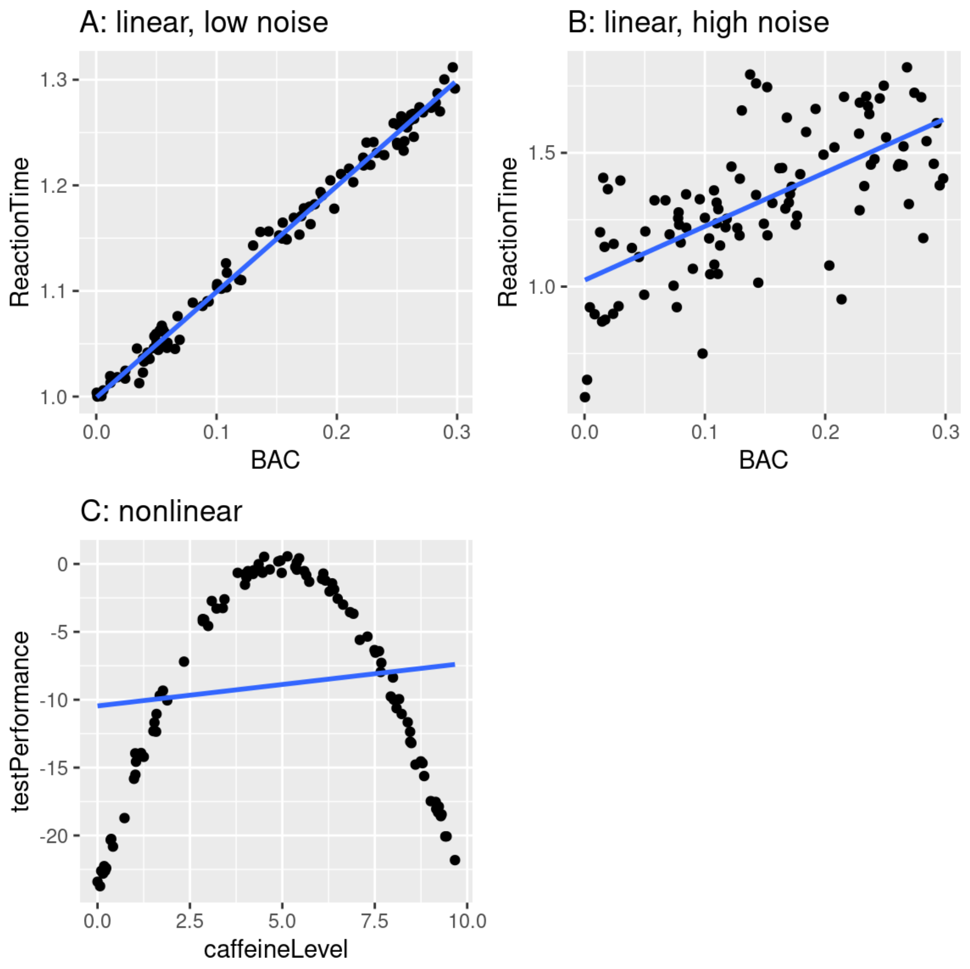
Linearity

- Terrible naming
- A better name might be
  - “No patterns (of any kind) in residuals”
  - "No systematic structure in errors”
  - “The noise is purely noise”
  - “$E(\varepsilon|X) = 0$”

- Terminology is historical artifact
  - From before linear regression was used more generally
  - May make sense with simple regression with one X

Linearity assumption

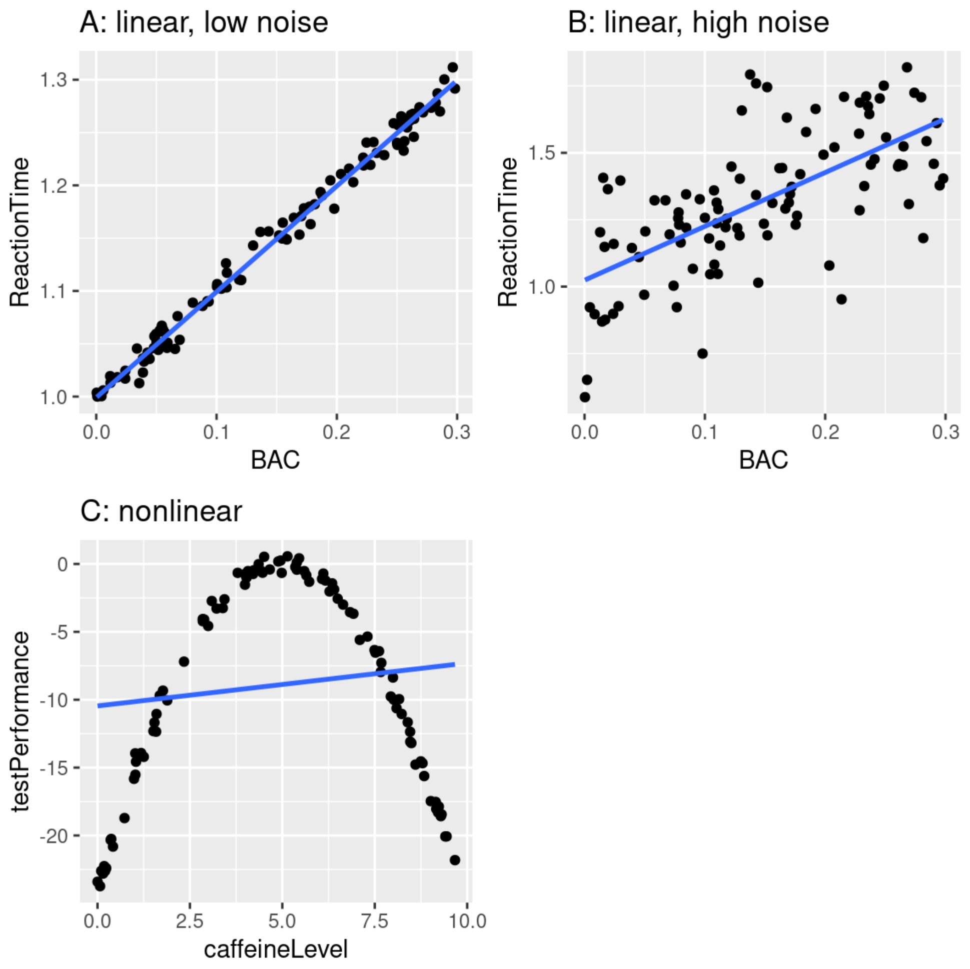
- It's about if the formulated model of the data-generating process is appropriate in its current form.



Linearity assumption

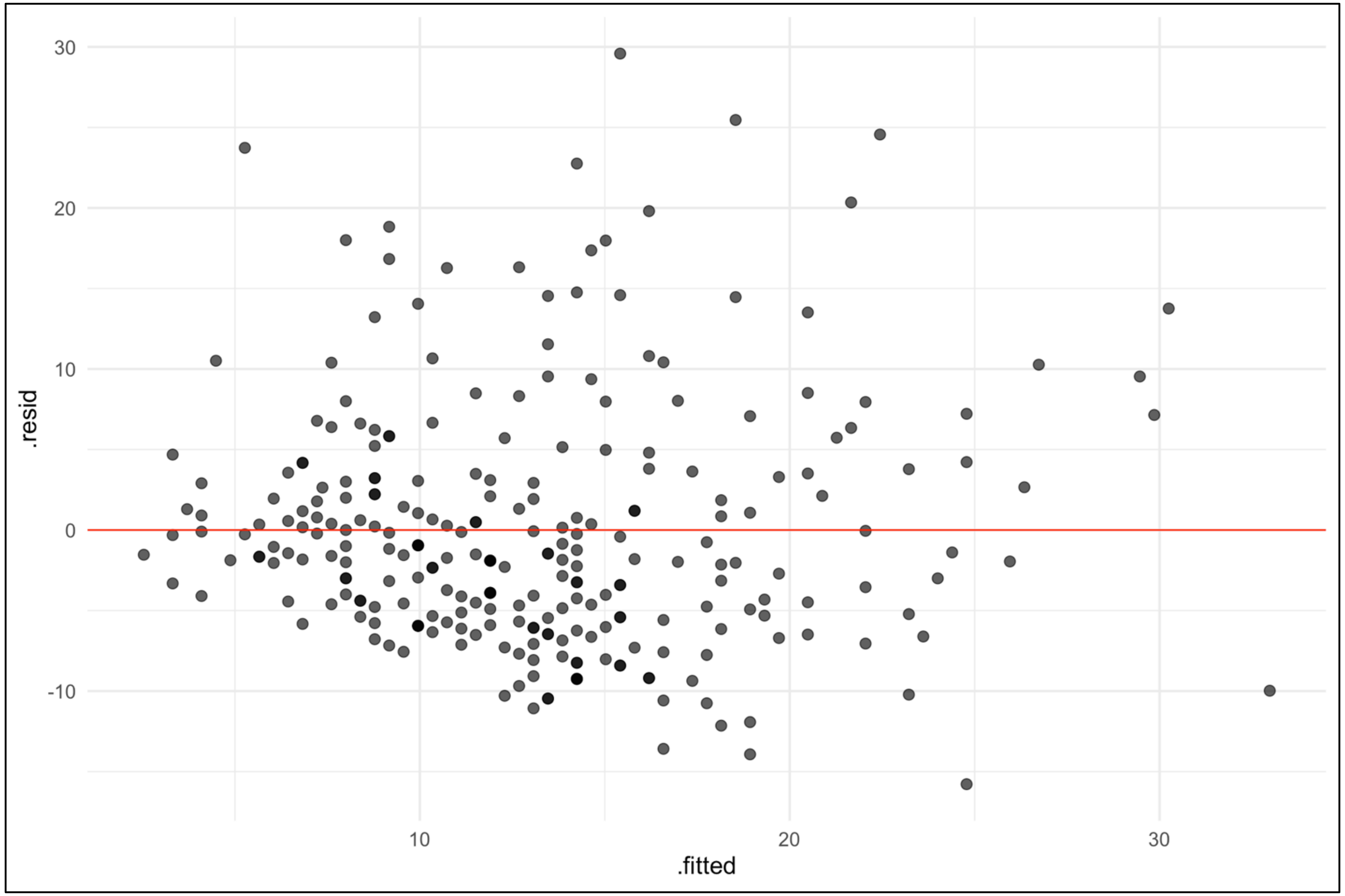
- It's about if the formulated model of the data-generating process is appropriate in its current form.

- Diagnostics
  - **1. Check via regular scatterplot**
    - Captures obvious non-linearity
      - Do you see a straight line?
      - Red flag: Curves, megaphone shapes, etc.



Linearity assumption

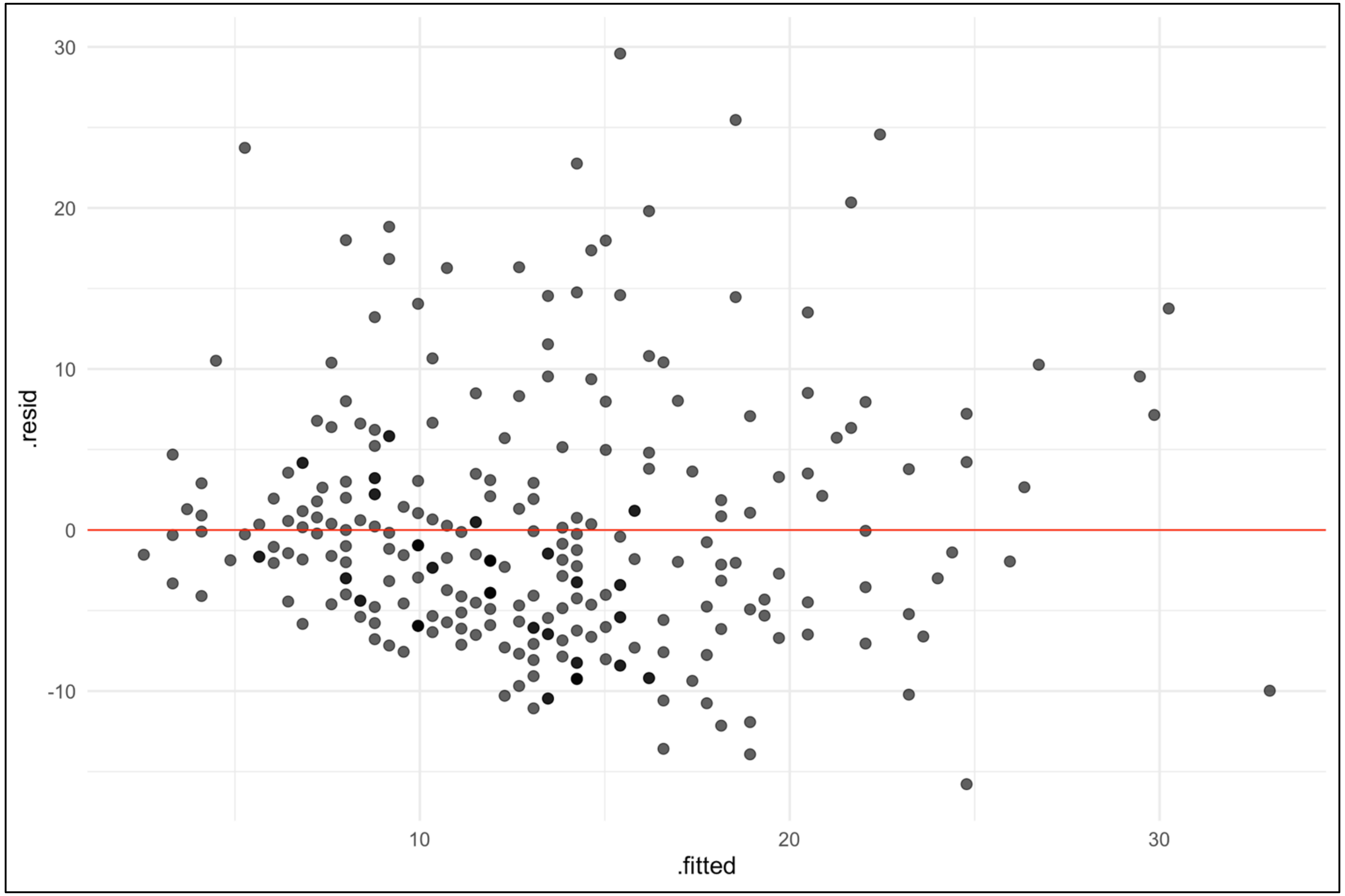
- **2. Scatterplot of fitted values vs residuals**
  - Plotted after model fit
  - Look for: random scattering around 0
  - Red flag:
    - Patterns, linear trends, etc.



Linearity assumption

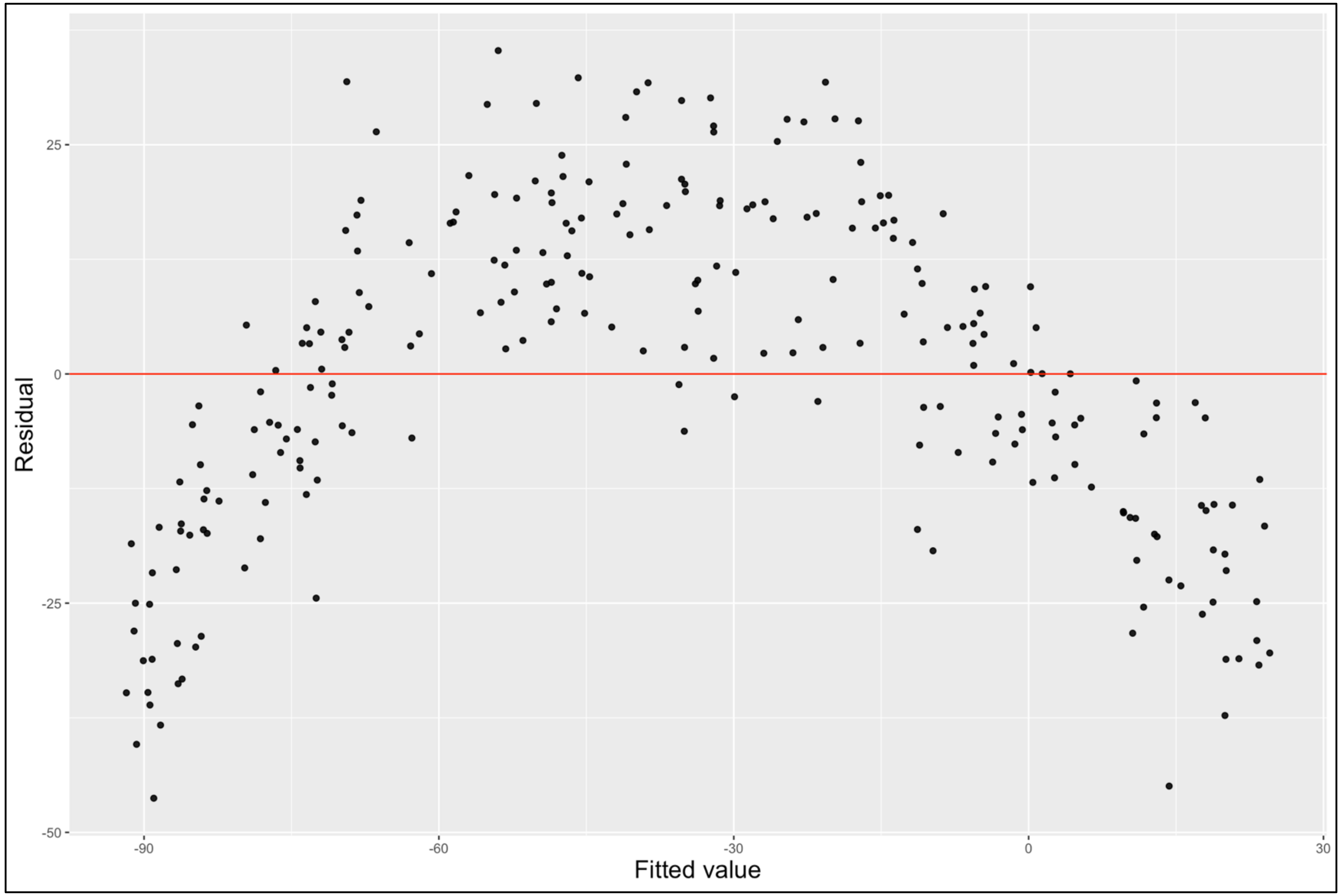
- **2. Scatterplot of fitted values vs residuals**
  - Plotted after model fit
  - Look for: random scattering around 0
  - Red flag:
    - Patterns, linear trends, etc.

✅ : random scattering around 0



✗ Violation: Non-Linear pattern

- **2. Scatterplot of fitted values vs residuals**
  - Plotted after model fit
  - Look for: random scattering around 0
  - Red flag:
    - Patterns, linear trends, etc.



2. Normality

- Model equation
  - $y = \beta_0 + \beta_1 x + \varepsilon$

Note here that we assume $\varepsilon \sim N(0, \sigma^2)$

2. Normality

- $y = \beta_0 + \beta_1 x + \varepsilon$

  Note here that we assume $\varepsilon \sim N(0, \sigma^2)$

- Problem: If errors are not Gaussian, we get
  - Invalid p-values
  - Incorrect confidence intervals
  - Unreliable hypothesis tests

- Why is this a common assumption?

2. Normality

- $y = \beta_0 + \beta_1 x + \varepsilon$

  Note here that we assume $\varepsilon \sim N(0, \sigma^2)$

- Problem: If errors are not Gaussian, we get
  - Invalid p-values
  - Incorrect confidence intervals
  - Unreliable hypothesis tests

2. Normality

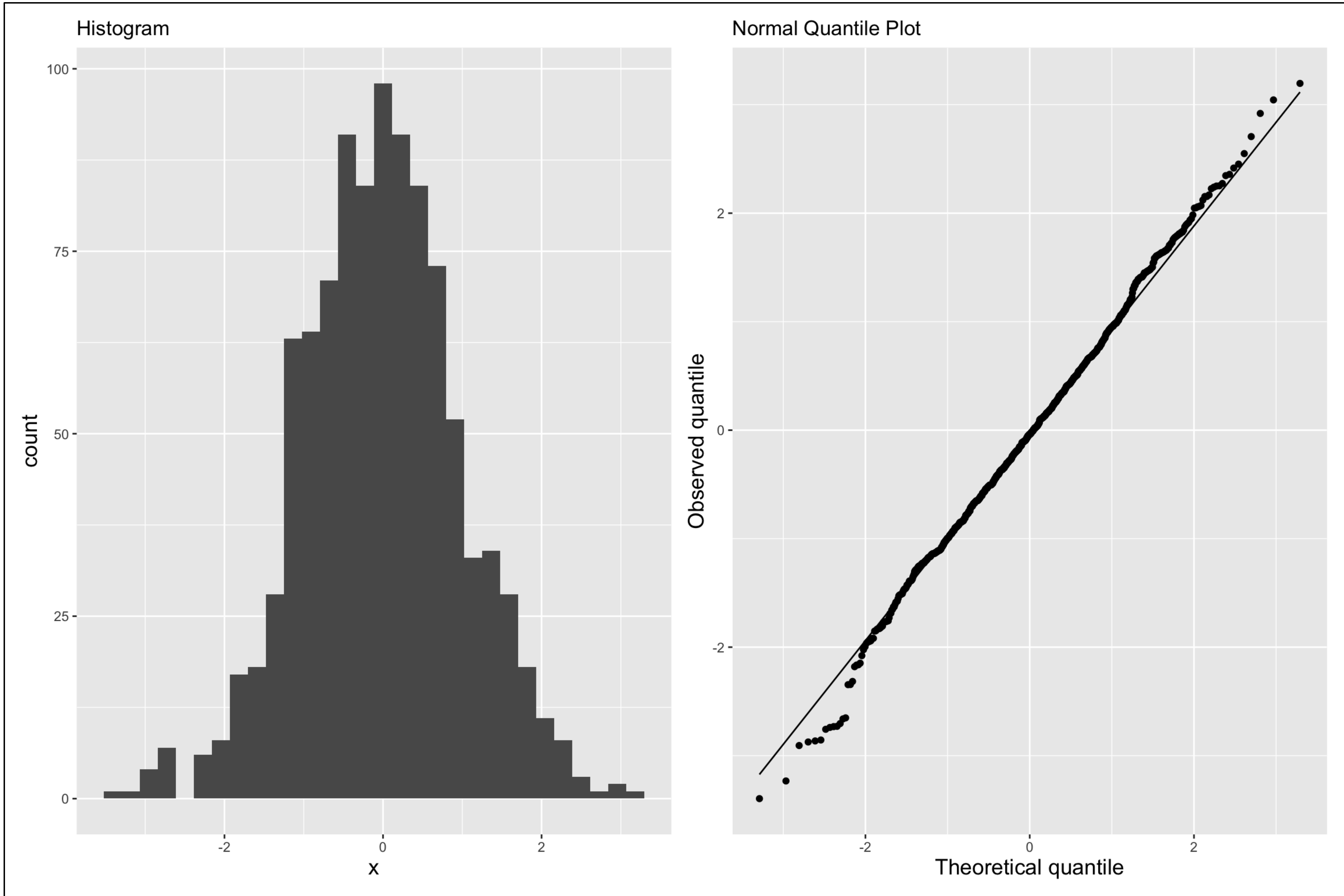
- $y = \beta_0 + \beta_1 x + \varepsilon$

  Note here that we assume $\varepsilon \sim N(0, \sigma^2)$

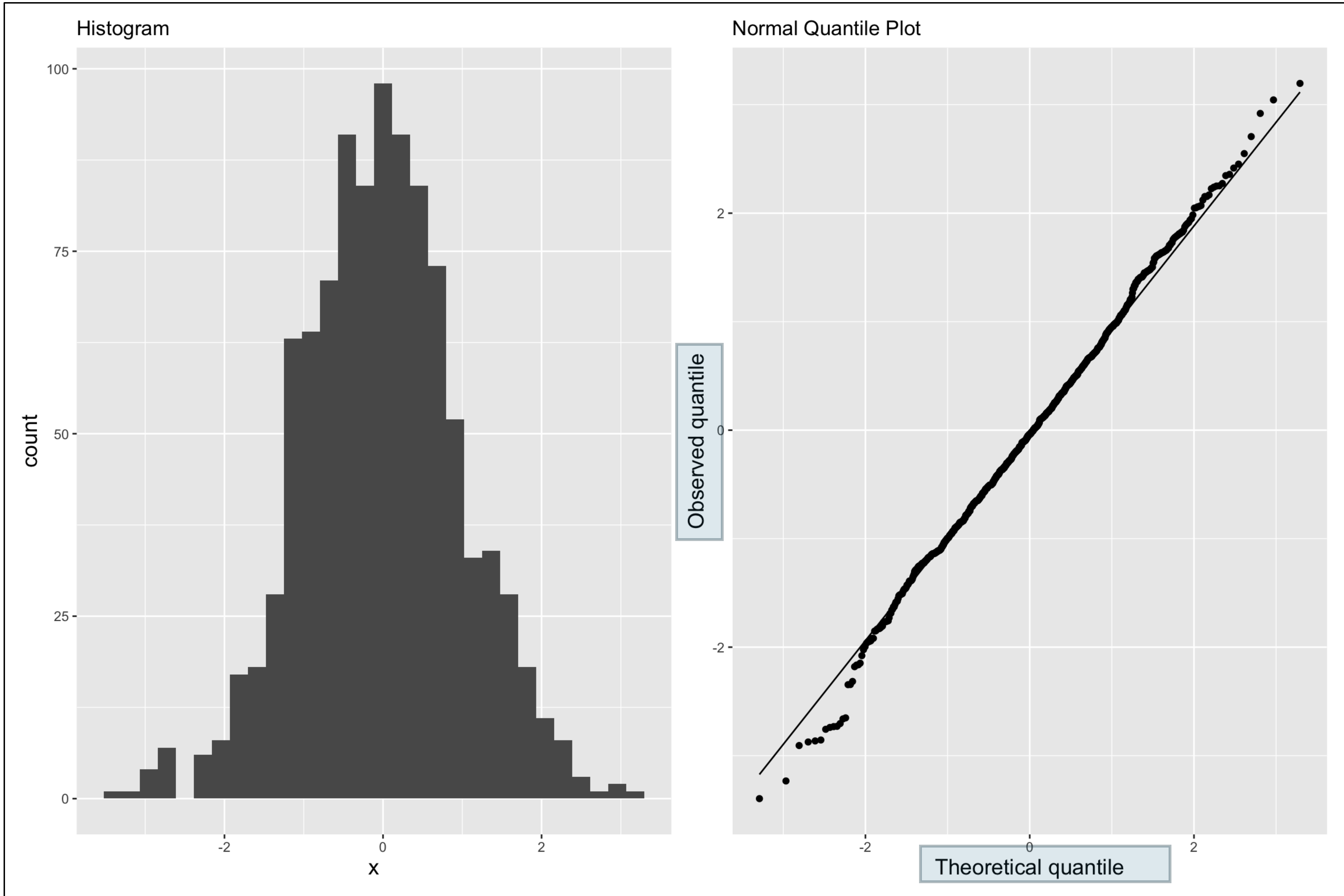
- Problem: If errors are not Gaussian, we get
  - Invalid p-values
  - Incorrect confidence intervals
  - Unreliable hypothesis tests

- Why is this a common assumption?
  - The central limit theorem suggests normality of sampling distributions with large enough samples

✅ : Points fall along a straight diagonal line on the normal quantile plot.

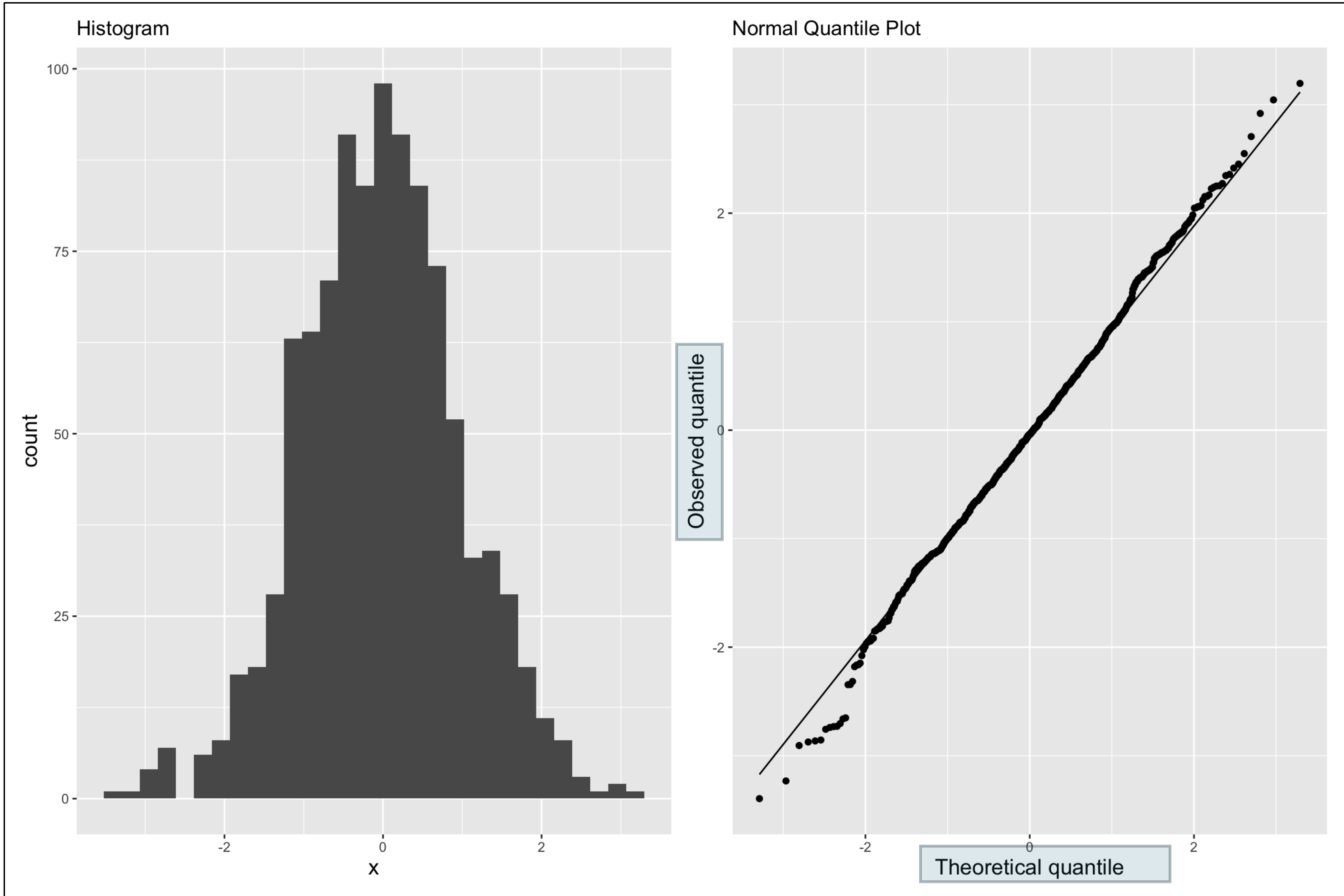


✅ : Points fall along a straight diagonal line on the normal quantile plot.



✅ : Points fall along a straight diagonal line on the normal quantile plot.

- Perfect normal? Straight line.
- Not normal?
  - Curves = skewed
  - S-shape = heavy tails
  - Zigzag = outliers



3. Homoscedasticity

- $y = \beta_0 + \beta_1 x + \varepsilon$

  Note here that we assume $\varepsilon \sim N(0, \sigma^2)$

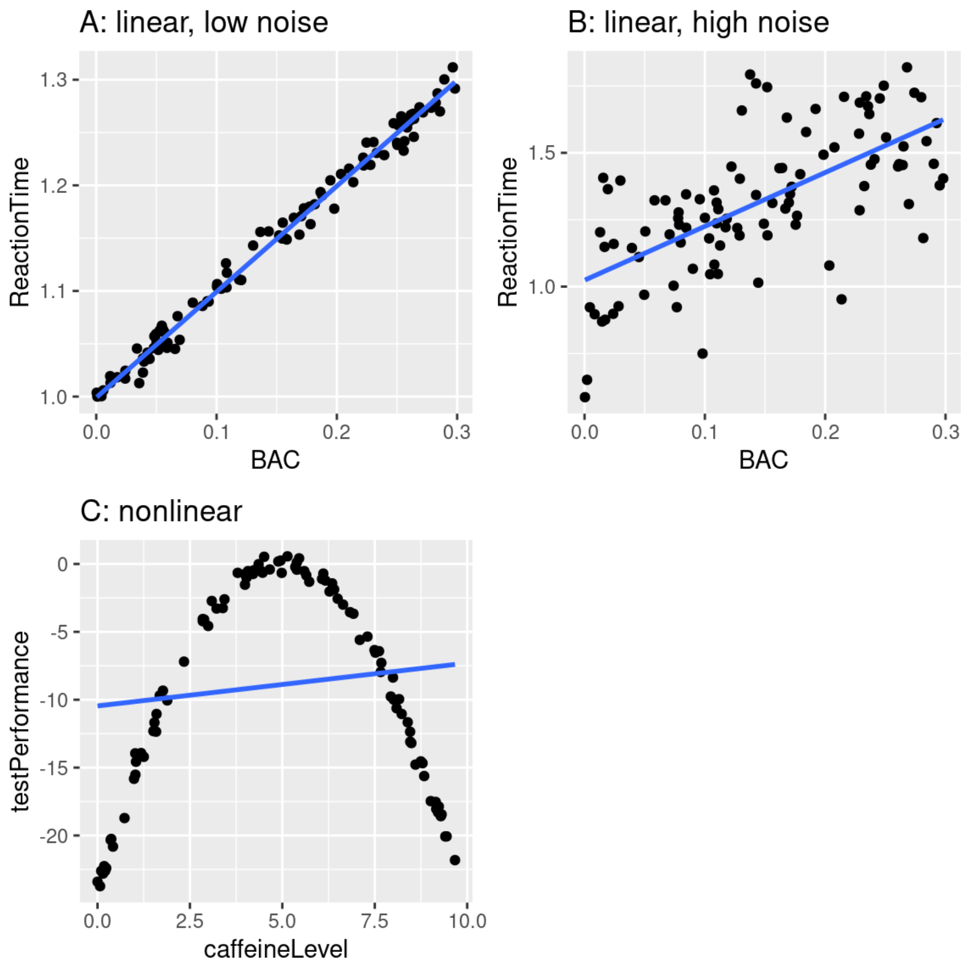
  And *σ^2 is a constant value*

3. Homoscedasticity

- $y = \beta_0 + \beta_1 x + \varepsilon$

  Note here that we assume $\varepsilon \sim N(0, \sigma^2)$

  And σ^2 is a constant value



3. Homoscedasticity

- $y = \beta_0 + \beta_1 x + \varepsilon$

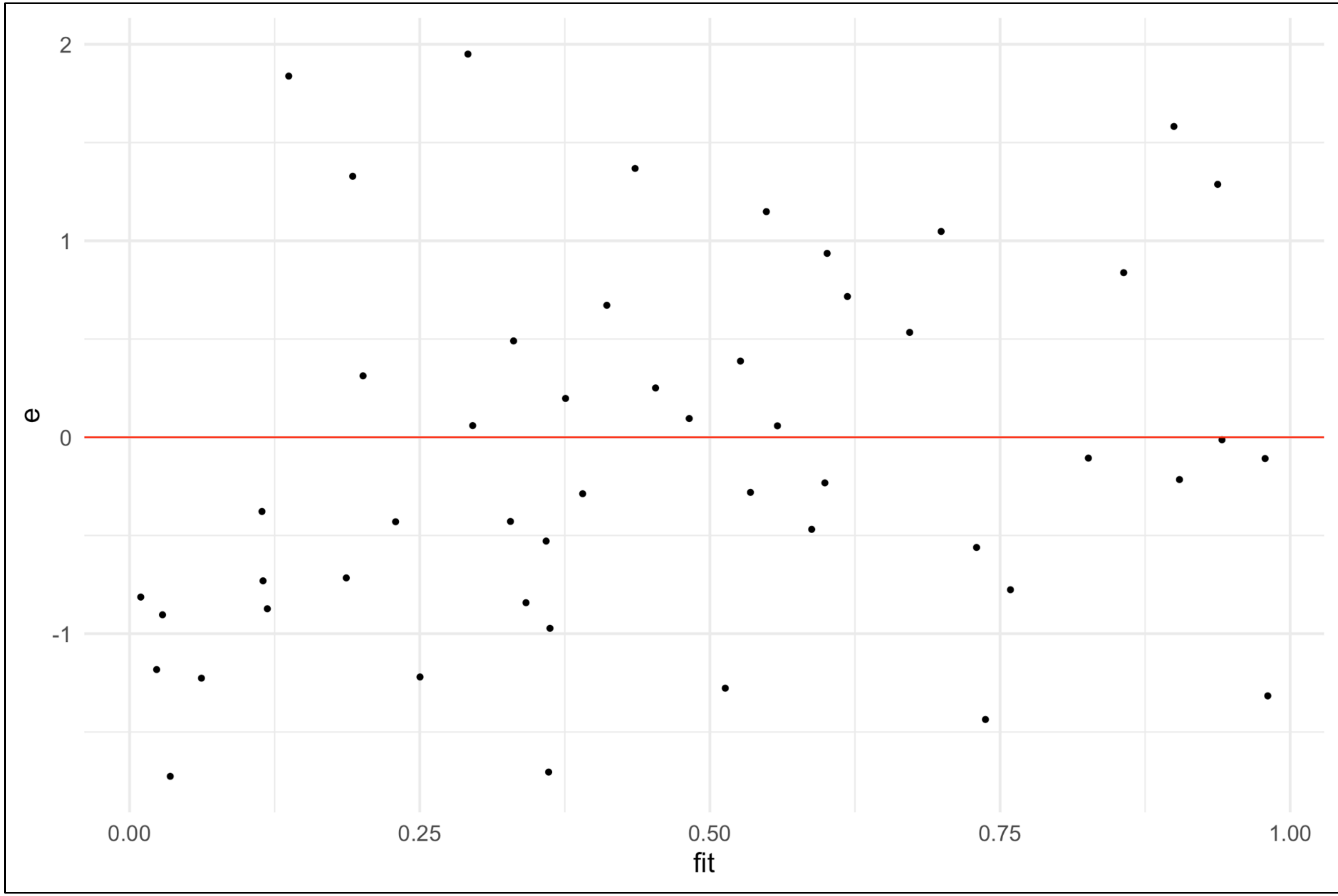
  Note here that we assume $\varepsilon \sim N(0, \sigma^2)$

  And *σ^2 is a constant value*

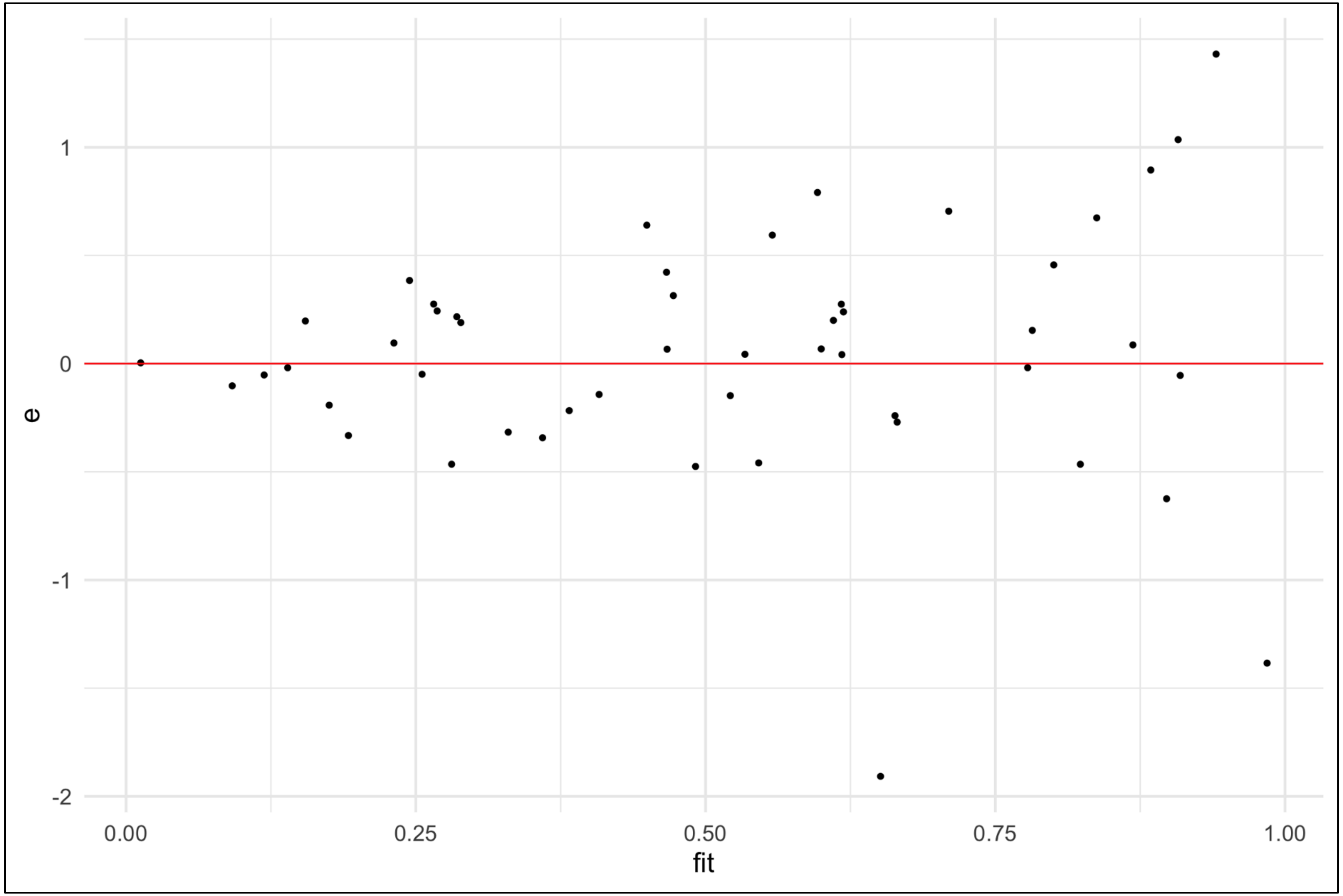
- Scatterplot of fitted values vs residuals
  - Constant error => No correlation between predictor and residuals
  - What are we looking for?
    - Random variation above and below 0
    - No patterns
    - Width of the band of points is constant

- Good

✅ : There is no distinguishable pattern or structure. The residuals are randomly scattered.



- Not so good
- There is a distinguishable pattern or structure.



4. Independence

- **Independence:** The errors are independent from each other

4. Independence

- **Independence:** The errors are independent from each other

- Common violations:
  - Time Series:
  - Yesterday affects today
  - Stock prices
  - Issues in experiment design!

- Checks:
  - Similarly via scatterplots

4. Independence

- **Independence:** The errors are independent from each other

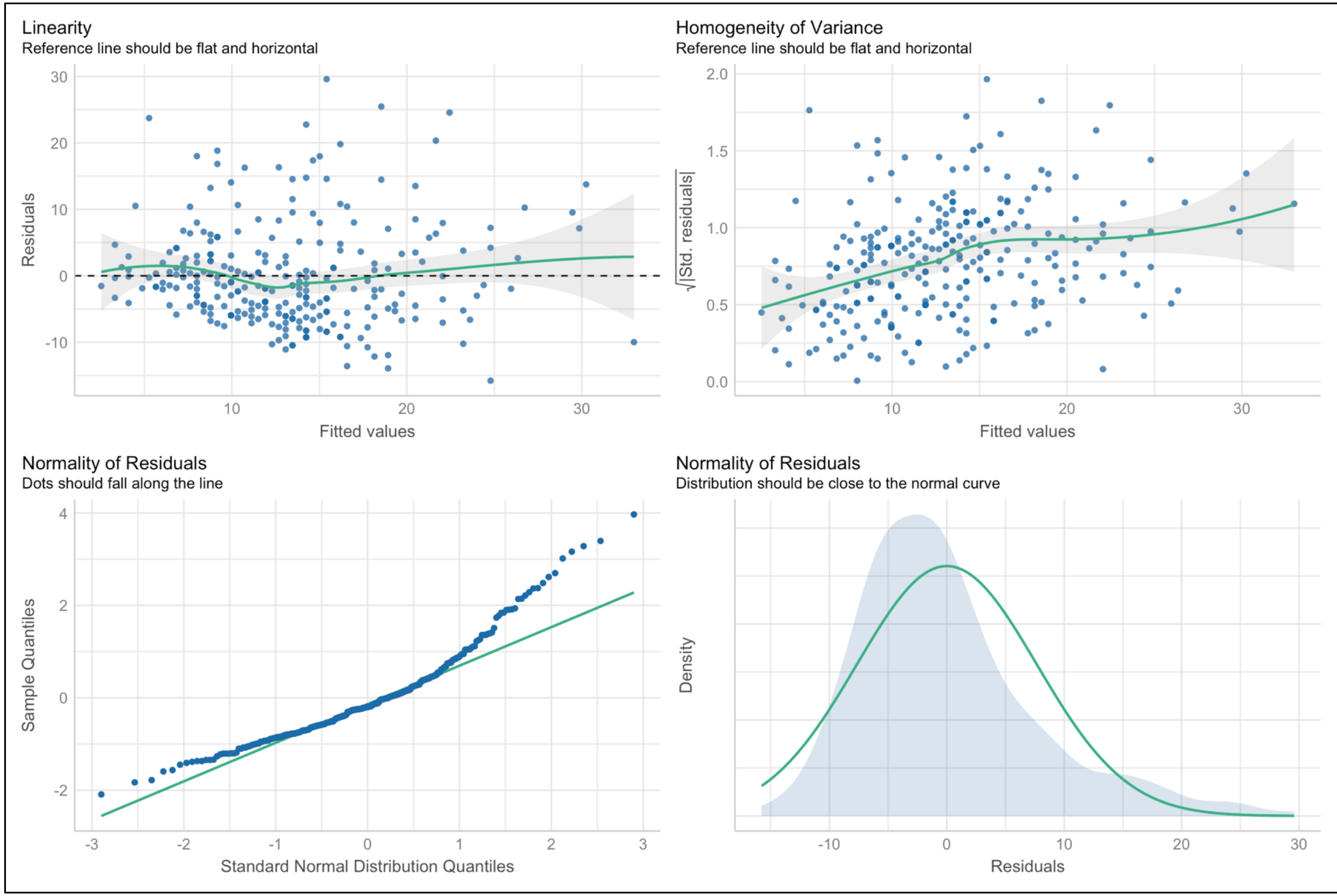
- Common violations:
  - Time Series:
  - Yesterday affects today
  - Stock prices
  - Issues in experiment design!

- Checks:
  - Similarly via scatterplots
- Let’s assume this is met

easystats:Performance

```{r}
performance::check_model(model1, check=c("normality", "linearity", "homogeneity", "qq"))
```

- Visual Model Checks



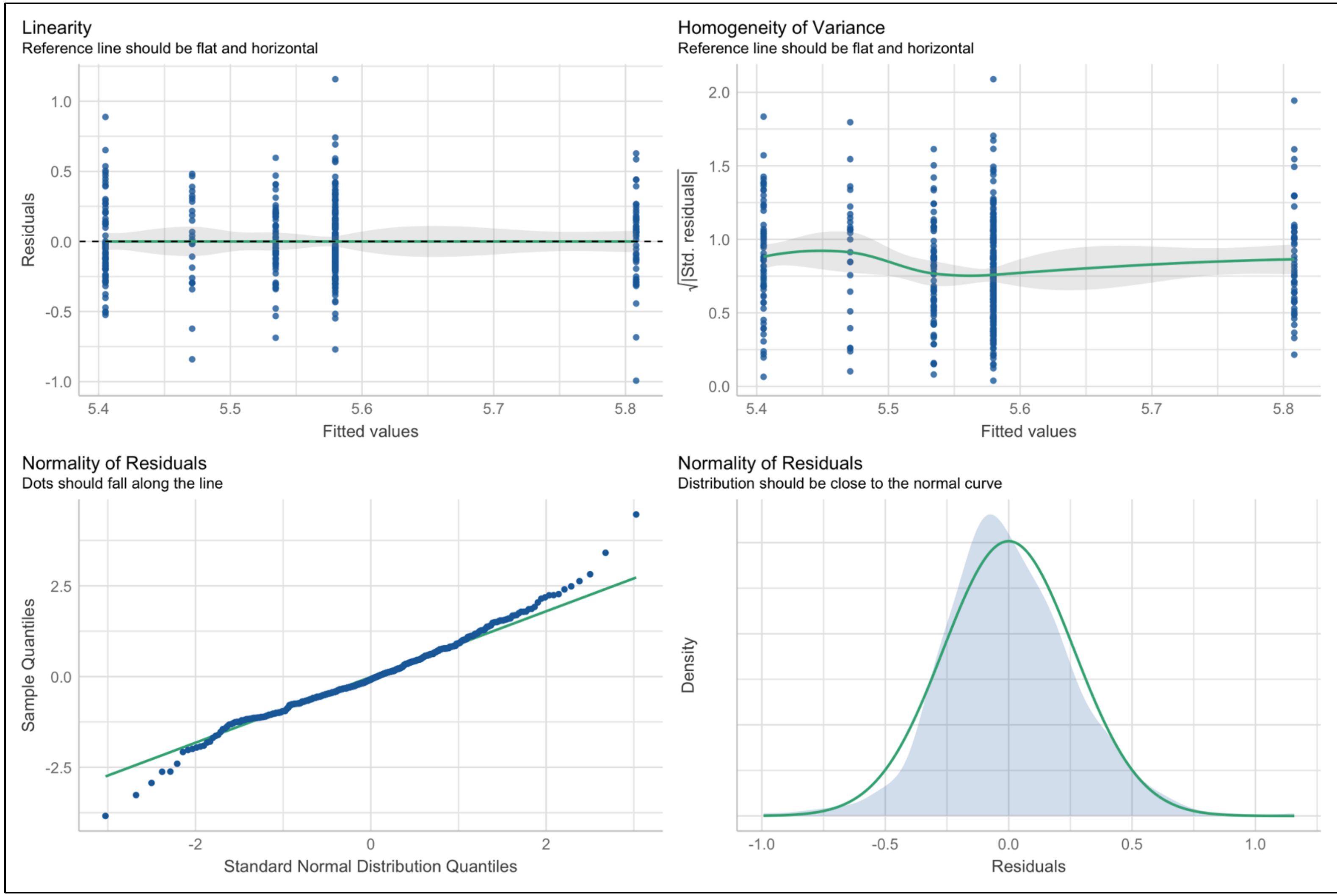
Assumptions for categorical predictors

- Linear models can be easily extended to categorical predictors
  - Interpretation of intercept and slope are a bit different
- Interpretation of test statistics and statistical significance are the same
  - So are assumptions checks!

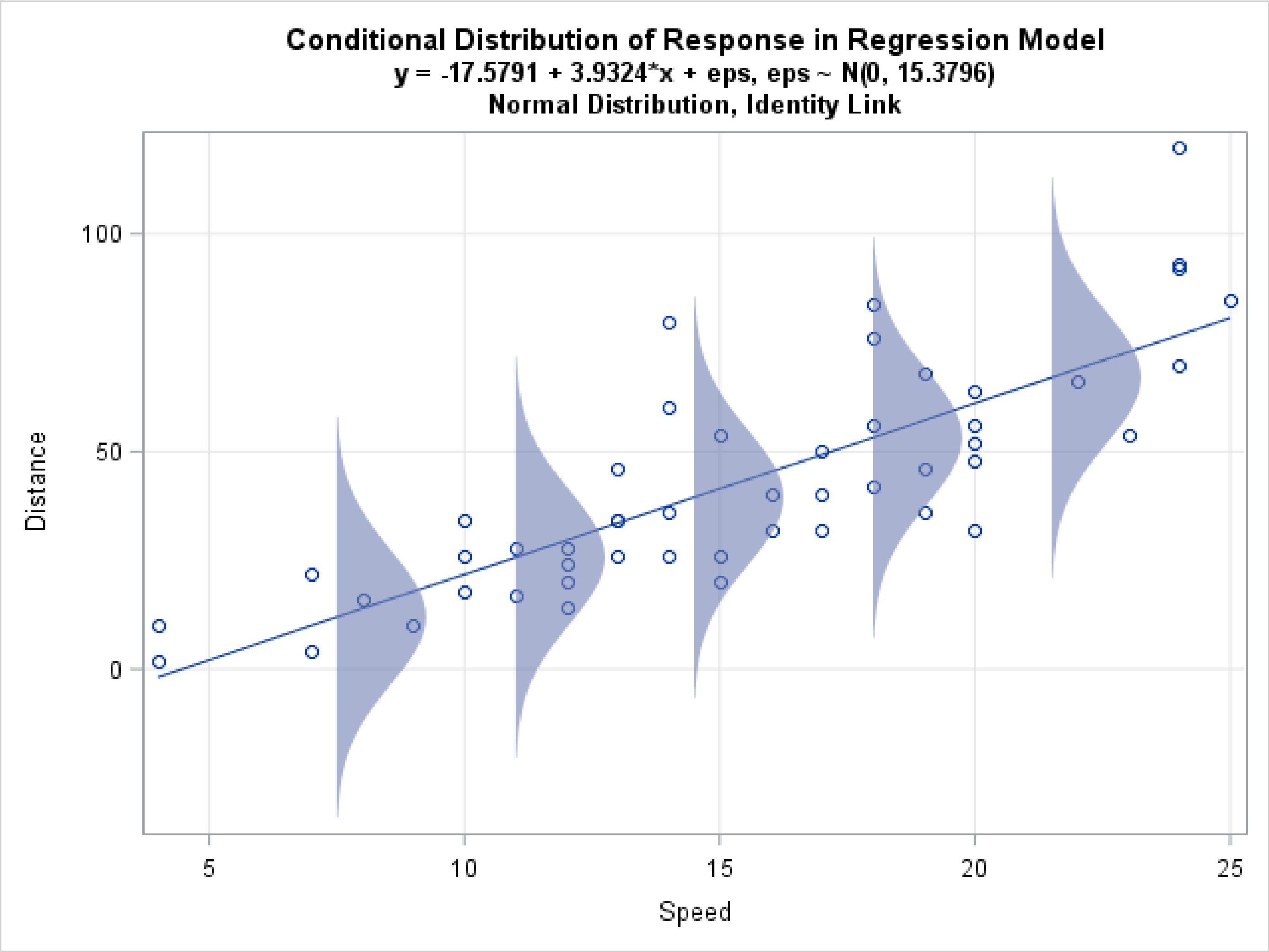
easystats:Performance

````
```{r}
performance::check_model(model1, check=c("normality", "linearity", "homogeneity", "qq"))
```
````

- Visual Model Checks

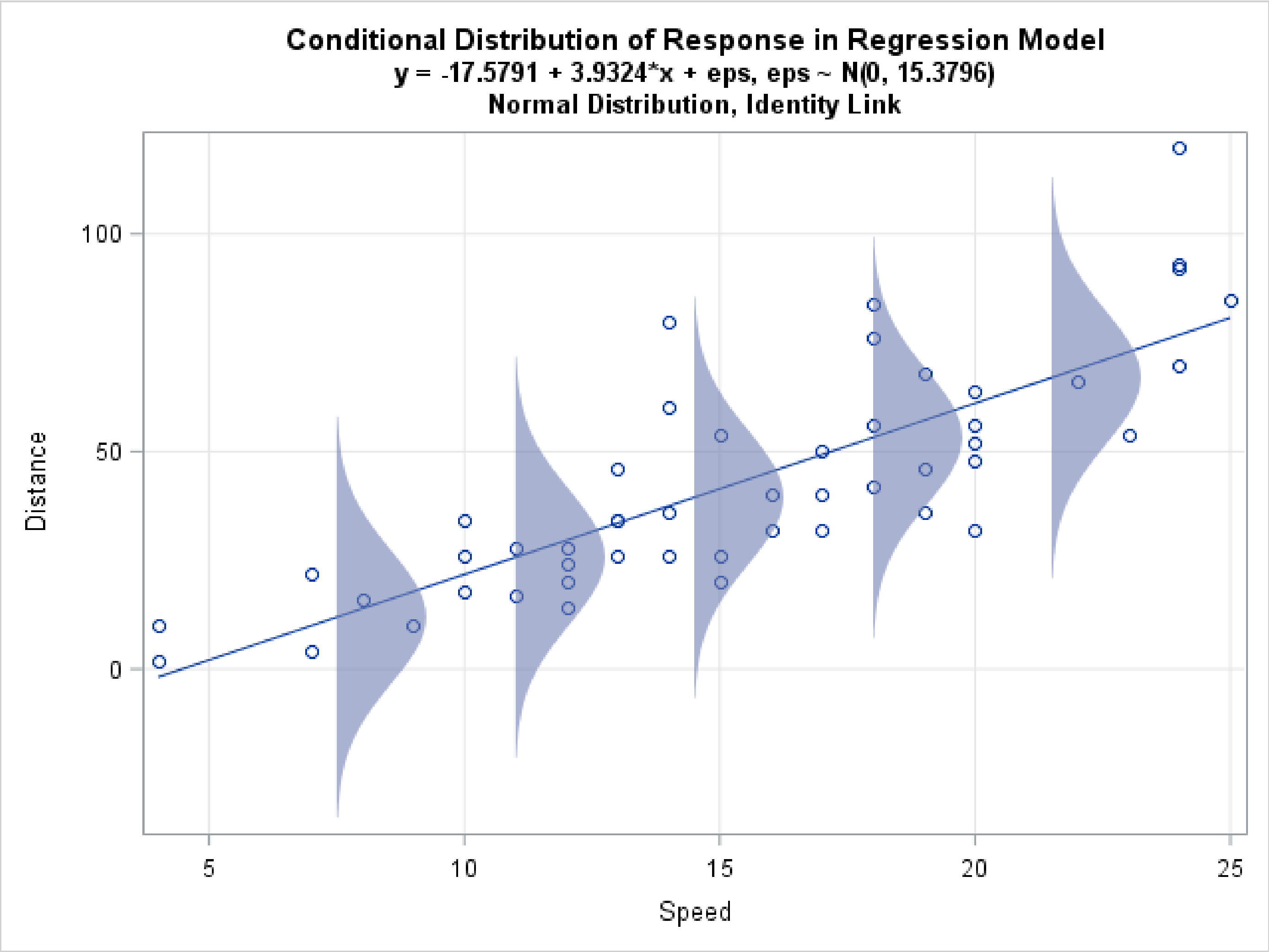


What these look like

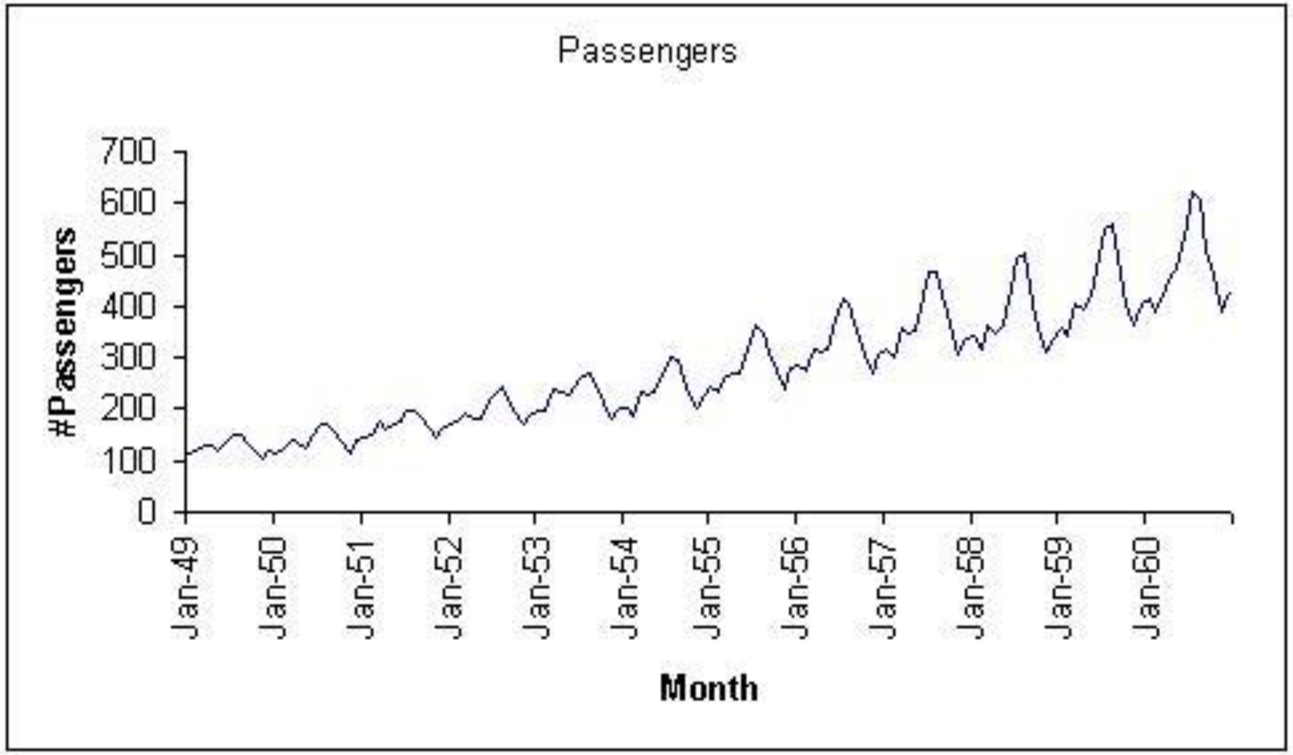


What these look like

- ***Linearity*** ⇒ means of the distributions lie on the line.
- ***Constant variance*** ⇒ the widths don’t change
- ***Normality*** ⇒ each distribution looks gaussian



Violation of independence



Framing in terms of observations vs residuals

Assumption	t-test	t-test as a linear model
Normality	Sampling distribution of differences must be normal	Residuals must be normal
Variance	Equal variance in both groups (Student's t-test)	Equal variance across groups
Independence	Scores in different conditions are independent	Each residual is independent

Assumption	t-test	ANOVA	Simple Regression	Multiple Regression
Normality	Sampling distribution of differences must be normal	Residuals must be normal	Residuals must be normal (mean = 0)	Residuals must be normal (mean = 0)
Variance	Equal variance in both groups (Student's t-test)	Equal variance across groups	Constant variance at all predictor levels (homoscedasticity)	Constant variance at all predictor levels (homoscedasticity)
Independence	Scores in different conditions are independent	Observations are independent	Residuals are independent (uncorrelated)	Errors are independent
Linearity	t-test is a special case of regression/linear model	ANOVA is a special case of linear model	Relationship between variables is linear	Relationship is linear (often assumes additive effects)
Data Types	At least interval level	At least interval scale	Outcome: quantitative, continuous, unbounded. Predictors: at least interval	Outcome: quantitative, continuous, unbounded. Predictors: quantitative or categorical

Beyond Visual Tests

- A variety of Hypothesis Tests
    - Null Hypothesis: an assumption is not violated
    - Run to see
        - If NH needs to be rejected

Beyond Visual Tests

- **Normality:** Shapiro-Wilk test
- **Homogeneity of variance:** Levene’s test, Bartlett test, NCV test
- **Linearity:** RESET test
- **Independence:** Durbin-Watson

When to use formal tests

- Use formal tests when:
  - Borderline visual cases
  - Need to document/justify decisions
  - Comparing models
  - Automated screening
- Don't use them as sole decision maker. Combine with visual checks and effect sizes.

Problem with Tests

- Dependent on sample size
  - n = 50: Tests have no power. Miss real violations.
  - n = 5000: Tests reject everything. Even harmless violations significant.

- IMO, residual plots tell you more than p-values.

Problem with Tests

- Dependent on sample size
  - n = 50: Tests have no power. Miss real violations.
  - n = 5000: Tests reject everything. Even harmless violations significant.

- IMO, residual plots tell you more than p-values.

R-packages

- Base R can run ANOVA, but these packages solve practical problems:
  - **afex**: Simplifies ANOVA syntax, handles complex designs
  - **emmeans**: Post-hoc tests and marginal means –
  - **performance**: Quick assumption checks
  - **pwr**: Power analysis for study design

afex

- **Why use it?**
  - One function for different ANOVA types
  - Handles unbalanced designs correctly (Type III SS)
  - Cleaner output than `aov()`

````
```{r}
library(afex)
model <- aov_ez(id = "subject_id",
                dv = "response_time",
                between = "condition",
                data = my_data)

```
````

emmeans

- **Why use it?**
  - Post-hoc pairwise comparisons
  - Marginal means adjusted for design
  - Multiple comparison corrections built-in

````
```{r}
library(emmeans)

# Get marginal means
emm <- emmeans(model, ~ condition)

# Pairwise comparisons
pairs(emm, adjust = "bonferroni")
```
````

performance

- **Why use it?**
  - One function checks all assumptions
  - Visual + statistical tests
  - Works with many model types

````
```{r}
library(performance)

# Check all assumptions at once
check_model(model)

# Or specific checks
check_normality(model)
check_homogeneity(model)
```
````