Linear regression: residuals and model diagnostics

2025-12-08

Table of contents

```
library(ggplot2)
library(ggtext)
```

1 Understanding Residuals: The Foundation of Diagnostics

Before diving into assumptions and diagnostic tests, we need to understand residuals—the cornerstone of all regression diagnostics. Many students confuse residuals with error terms, but understanding the distinction is crucial for interpreting diagnostic plots and tests.

1.1 What Are Residuals?

Residuals are the observed differences between actual values and predicted values from your fitted model:

$$e_i = y_i - \hat{y}_i$$

where:

- e_i is the residual for observation i
- y_i is the actual observed value
- $\hat{y}_i$ is the predicted value from your regression model

In plain language: A residual tells you how far off your prediction was for each observation. If your model predicts a customer will spend 100 EUR but they actually spent 120 EUR, the residual is 20 EUR (see also Figure 1).

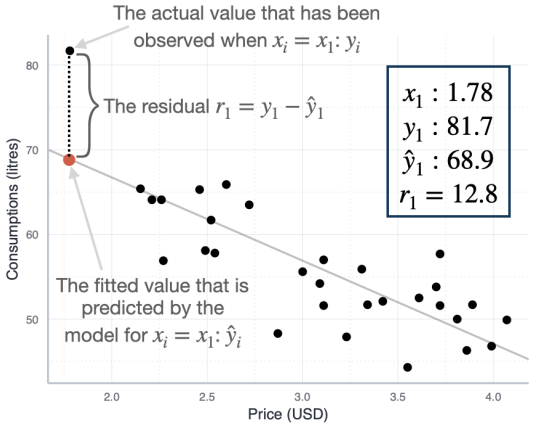

Figure 1: Illustration of a residual.

1.2 Residuals vs. Error Terms: A Critical Distinction

Students often use "residual" and "error" interchangeably, but they are fundamentally different concepts:

Concept	Symbol	What It Is	Can We Observe It?
Error term	ϵ_i	True, unknown deviation from the population regression line	**No** - it's a theoretical concept
Residual	e_i	Observed deviation from our fitted sample regression line	**Yes** - we calculate it from our data

The error term (ϵ_i) is the true, unobservable deviation in the population model:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

This represents the "true" relationship in the population. We never observe ϵ_i because we don't know the true population parameters (β_0 and β_1).

The residual (e_i) is what we actually observe from our fitted sample model:

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + e_i$$

This is our estimate based on sample data. We calculate e_i because we have estimates $\hat{\beta}_0$ and $\hat{\beta}_1$.

1.3 Why This Distinction Matters

1. **Assumptions are about error terms, not residuals:** When we say "errors are normally distributed" or "errors have constant variance," we're making statements about the unobservable ϵ_i in the population.
2. **Diagnostics use residuals to learn about errors:** Since we can't observe ϵ_i, we use residuals e_i as our best approximation. We examine residuals hoping they reveal the properties of the true errors.
3. **Residuals are estimates of errors:** Under the regression assumptions, residuals should behave similarly to errors. If they don't (showing patterns, non-constant variance, etc.), it suggests the assumptions are violated.

1.4 Types of Residuals in Diagnostics

You'll encounter several types of residuals in diagnostic output:

1. **Raw residuals** (e_i):

$$e_i = y_i - \hat{y}_i$$

These are what we've been discussing—the basic difference between observed and predicted values.

2. **Standardized residuals** (e_i^*):

$$e_i^* = \frac{e_i}{\hat{\sigma}\sqrt{1 - h_i}}$$

where $\hat{\sigma}$ is the residual standard error and h_i is the leverage. Standardized residuals have approximately unit variance, making them comparable across observations. Values beyond ± 2 or ± 3 suggest potential outliers.

Understanding Leverage

Leverage (denoted as h_i or "hat value") measures how unusual or extreme an observation's predictor values (X values) are compared to the rest of the data. It quantifies how far an observation is from the center of the predictor space.
Mathematical definition:
In simple linear regression, leverage for observation i is:

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{j=1}^{n}(x_j - \bar{x})^2}$$

In multiple regression, leverage values come from the diagonal of the "hat matrix" $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$, which "puts the hat" on $\mathbf{y}$ to get $\hat{\mathbf{y}}$.
Key properties:

- Leverage ranges from $\frac{1}{n}$ to 1
- Average leverage is $\frac{p}{n}$ where p is the number of parameters (including intercept)
- Sum of all leverage values equals p: $\sum_{i=1}^{n} h_i = p$

Interpretation:

- **Low leverage** ($h_i \approx \frac{p}{n}$): The observation's X values are typical, close to the center of the data
- **High leverage** ($h_i > \frac{2p}{n}$ or $\frac{3p}{n}$): The observation's X values are unusual or extreme

Why leverage matters:
High leverage observations have greater potential to influence the regression line because they're far from the center of the data. Think of a lever: observations far from the fulcrum (center) have more power to move the line.
However, **high leverage alone doesn't make a point influential**:

- If a high leverage point fits the model well (small residual), it reinforces the pattern and isn't problematic

- If a high leverage point doesn't fit the model (large residual), it can dramatically pull the regression line toward itself

Example:
Imagine studying the relationship between study hours and exam scores for students who studied 10-20 hours. A student who studied 40 hours would have high leverage—they're far from the typical range. If they score as predicted by the pattern, they confirm it. If they score much lower or higher than expected, they could dramatically change your estimated relationship.
Why leverage appears in standardized residuals:
The formula $e_i^* = \frac{e_i}{\hat{\sigma}\sqrt{1-h_i}}$ adjusts for the fact that observations with high leverage naturally have smaller residuals because the regression line is "pulled" toward them. Without this adjustment, high leverage points would appear to fit better than they actually do, masking potential problems.

3. Studentized residuals:

Similar to standardized but use a different estimate of variance that excludes observation i. More robust for identifying outliers.

Why standardize? Raw residuals naturally have different variances depending on the leverage of each observation. Standardization accounts for this, putting all residuals on the same scale for fair comparison.

1.5 Why Residuals Are Central to Diagnostics

Residuals are the empirical manifestation of all our modeling assumptions. Here's why they're so important:

1. They reveal assumption violations:

- **Linearity:** If the relationship isn't linear, residuals will show systematic patterns (curves, waves)
- **Homoscedasticity:** If variance isn't constant, residuals will spread out (or contract) across fitted values
- **Normality:** If errors aren't normal, the distribution of residuals will deviate from normality
- **Independence:** If observations aren't independent, residuals will show autocorrelation or clustering

2. They're model-free diagnostics:

Residuals don't require you to know the "true" model. They simply show you what's left unexplained after fitting your model. Large, systematic patterns in residuals mean your model is missing something important.

3. They provide a common framework:

Almost every diagnostic tool examines residuals in some way:

- Plotting residuals to check for patterns

- Testing the distribution of residuals
- Examining residual variance across groups
- Identifying observations with unusual residuals

4. They quantify model performance:

The magnitude of residuals tells you about prediction accuracy. Ideally, residuals should be:

- Small in magnitude (good predictions)
- Random in pattern (no systematic errors)
- Homogeneous in variance (consistent precision)
- Approximately normal (valid inference)

1.6 The Diagnostic Philosophy

When you examine diagnostic plots, you're asking: **"Do these residuals behave like random noise, or do they show systematic patterns that suggest model problems?"**

If residuals are truly random (as they should be when assumptions hold), they should:

- Scatter randomly around zero with no patterns
- Have roughly constant spread across all fitted values
- Follow a normal distribution (for the population errors)
- Show no relationship to predictor variables, fitted values, or observation order

Any deviation from this ideal suggests where your model or assumptions may be failing.

1.7 A Practical Example

Imagine you're modeling house prices based on square footage:

- **Model:** Price = \$50,000 + \$200 × SqFt
- **Observation:** A 1,500 sq ft house sells for \$350,000
- **Prediction:** $\hat{y}$ = \$50,000 + \$200 × 1,500 = \$350,000
- **Residual:** e = \$350,000 - \$350,000 = \$0 (perfect prediction!)

But for another house:

- **Observation:** A 2,000 sq ft house sells for \$500,000
- **Prediction:** $\hat{y}$ = \$50,000 + \$200 × 2,000 = \$450,000
- **Residual:** e = \$500,000 - \$450,000 = \$50,000 (underpredicted by \$50,000)

If most large houses have large positive residuals, this pattern suggests your linear model is inadequate—perhaps you need a quadratic term or the relationship changes at higher square footage.

---

2 Understanding the Assumptions

Linear regression relies on four key assumptions about the data and the relationship between variables. Understanding these assumptions is crucial because violations can lead to:

- **Biased parameter estimates** (incorrect coefficients)
- **Invalid standard errors** (unreliable confidence intervals)
- **Incorrect hypothesis tests** (wrong p-values and conclusions)
- **Poor predictions** (inaccurate forecasts)

2.1 1. Linearity

What it means: The relationship between each predictor variable and the outcome variable is linear. In mathematical terms, the conditional expectation of Y given X follows a linear function: $E[Y|X] = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p$

Why it matters: Linear regression models can only capture linear relationships. If the true relationship is curved or more complex, the model will systematically mispredict values, leading to biased estimates and poor fit.

What happens if violated: The model will show systematic patterns in residuals (curves, waves), predictions will be systematically wrong in certain ranges, and R^2 will underestimate the true strength of the relationship.

2.2 2. Independence

What it means: Each observation is independent of all other observations. The residual (error) for one observation does not depend on the residual for any other observation: $Cov(\epsilon_i, \epsilon_j) = 0$ for $i \neq j$.

Why it matters: Independence is required for valid standard errors and hypothesis tests. When observations are correlated (e.g., repeated measurements from the same person, or time series data), standard errors are typically underestimated, leading to overconfident conclusions.

What happens if violated: Standard errors are incorrect (usually too small), confidence intervals are too narrow, p-values are too small (Type I error inflation), and you may falsely conclude that relationships are significant.

Common violations: Time series data (autocorrelation), clustered data (students within schools), repeated measures (multiple observations per subject), spatial data (nearby locations are similar).

2.3 3. Homoscedasticity

What it means: The variance of the residuals is constant across all levels of the predictor variables: $Var(\epsilon_i) = \sigma^2$ for all i. The "spread" of residuals should be the same whether you're predicting low or high values.

Why it matters: Heteroscedasticity (non-constant variance) leads to inefficient estimates and incorrect standard errors. While coefficient estimates remain unbiased, hypothesis tests and confidence intervals become unreliable.

What happens if violated: Standard errors are incorrect, confidence intervals are wrong (may be too wide or too narrow), hypothesis tests are invalid, and the model is inefficient (there are better ways to estimate the relationships).

Common causes: The variance of Y naturally increases with X (e.g., spending variance increases with income), measurement error that varies, or misspecified functional form (wrong model).

2.4 4. Normality

What it means: The residuals (errors) follow a normal distribution: $\epsilon_i \sim N(0, \sigma^2)$. Note that this assumption is about the residuals, not the variables themselves.

Why it matters: Normality is primarily required for valid hypothesis tests and confidence intervals, especially in small samples. The Central Limit Theorem helps here: with large samples (n > 30-50), inference remains approximately valid even with moderate departures from normality.

What happens if violated: Confidence intervals and hypothesis tests may be inaccurate, especially in small samples. Predictions and coefficient estimates remain unbiased, but uncertainty quantification becomes unreliable.

Important note: This is the least critical assumption for large samples due to the Central Limit Theorem. Focus more on linearity, independence, and homoscedasticity.

---

3 Summary Table of Assumptions and Tests

Table 2: Summary of linear regression assumptions and diagnostic approaches

Assumption	Description	Visual Test	Quantitative Test
Linearity	The relationship between predictors and outcome is linear	Residuals vs Fitted plot (Tukey-Anscombe)	RESET test (Ramsey)

Assumption	Description	Visual Test	Quantitative Test
Independence	Observations are independent of each other	Residuals vs Order plot; ACF plot	Durbin-Watson test
Homoscedasticity	Constant variance of residuals across all levels of predictors	Residuals vs Fitted; Scale-Location plot	Breusch-Pagan test; White test
Normality	Residuals follow a normal distribution	QQ plot; Histogram of residuals	Shapiro-Wilk test; Kolmogorov-Smirnov test

---

4 Essential Visual Diagnostic Tests

Visual diagnostics are your first and most important line of defense. R automatically generates four diagnostic plots with `plot(lm_model)` that provide a comprehensive visual assessment. Always examine these plots before moving to quantitative tests.

4.1 The Four Standard Diagnostic Plots

4.1.1 Plot 1: Residuals vs Fitted (Tukey-Anscombe Plot)

Purpose: Checks linearity and homoscedasticity simultaneously.

How it works: Plots residuals ($e_i = y_i - \hat{y}_i$) against fitted values ($\hat{y}_i$). If the linearity and homoscedasticity assumptions hold, residuals should be randomly scattered around zero with no patterns or trends.

What to look for:

- **Good:** Random scatter around the horizontal line at zero, with no discernible pattern (see Figure 2)
- **Problem - Non-linearity:** Curved pattern (U-shape, inverted U, or other systematic curves) indicates the relationship is not linear (see Figure 3)
- **Problem - Heteroscedasticity:** Funnel shape (expanding or contracting spread) indicates non-constant variance (see Figure 4)

Interpretation: The smoothed line (blue) should be approximately horizontal at zero. Any systematic deviation indicates assumption violations.

Figure 2: Residuals vs Fitted plot showing a good model fit. Residuals are randomly scattered around zero with no patterns, indicating linearity and homoscedasticity are satisfied.

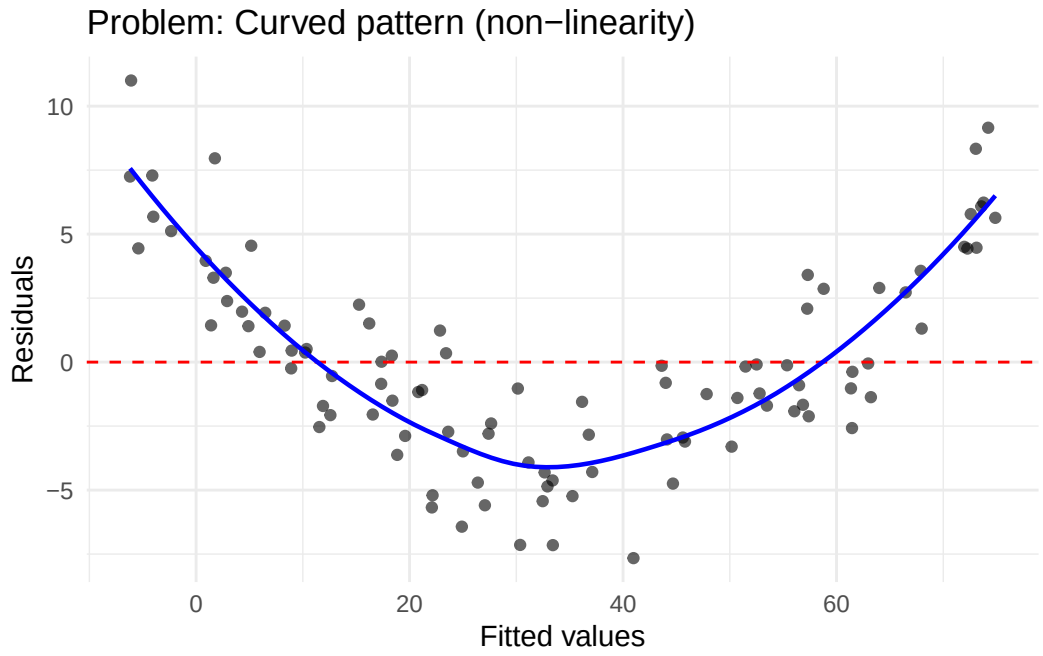

Figure 3: Residuals vs Fitted plot showing a non-linear relationship. The curved pattern in the residuals indicates that a linear model is inadequate and transformations or polynomial terms may be needed.

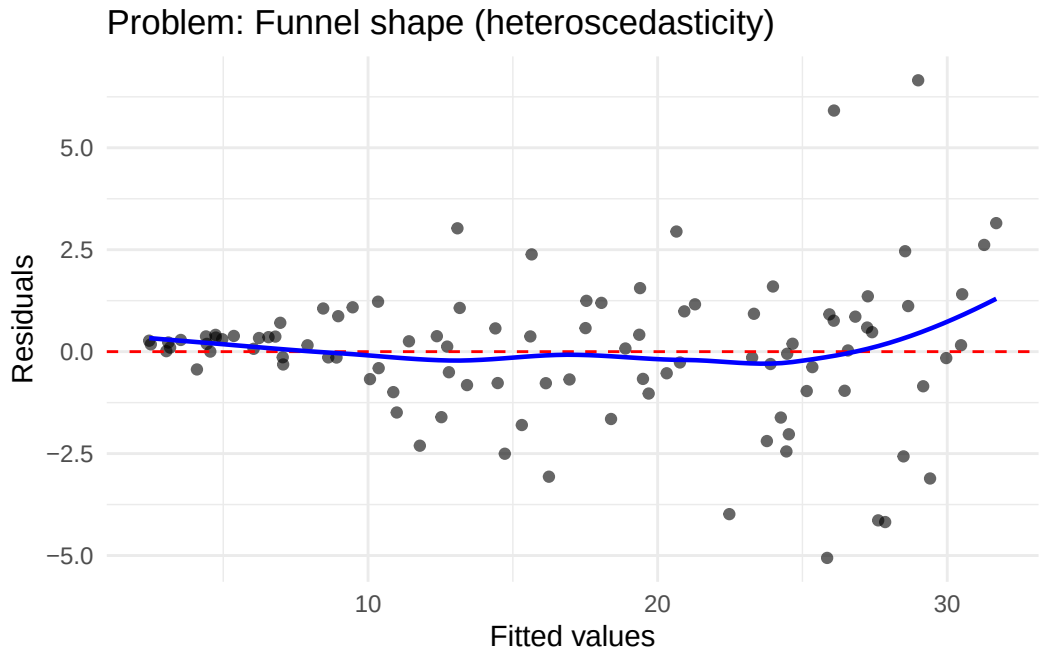

Figure 4: Residuals vs Fitted plot showing heteroscedasticity. The funnel shape (increasing spread of residuals) indicates that variance is not constant across the range of fitted values.

4.1.2 Plot 2: Normal Q-Q Plot

Purpose: Checks whether residuals follow a normal distribution.

How it works: Plots the quantiles of standardized residuals against theoretical quantiles from a standard normal distribution. If residuals are normally distributed, points should fall along the diagonal reference line.

What to look for:

- **Good:** Points closely follow the diagonal line from lower-left to upper-right (see Figure 5)
- **Problem - Heavy tails:** Points deviate from the line at both extremes, curving away (see Figure 6)
- **Problem - Skewness:** Points deviate systematically in one direction, showing an S-curve
- **Problem - Light tails:** Points deviate toward the line at extremes

Interpretation: Minor deviations are often acceptable, especially with larger sample sizes (n > 30-50) where the Central Limit Theorem provides robustness. Focus on gross departures from normality rather than minor wiggles.

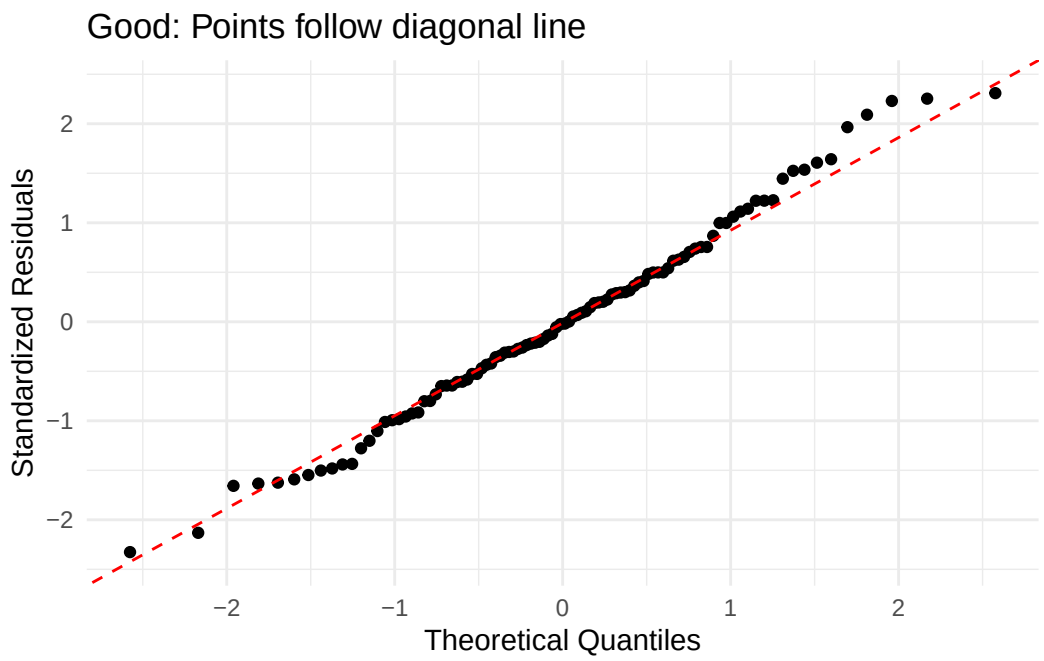

Figure 5: Normal Q-Q plot showing normally distributed residuals. Points fall closely along the diagonal reference line, indicating the normality assumption is satisfied.

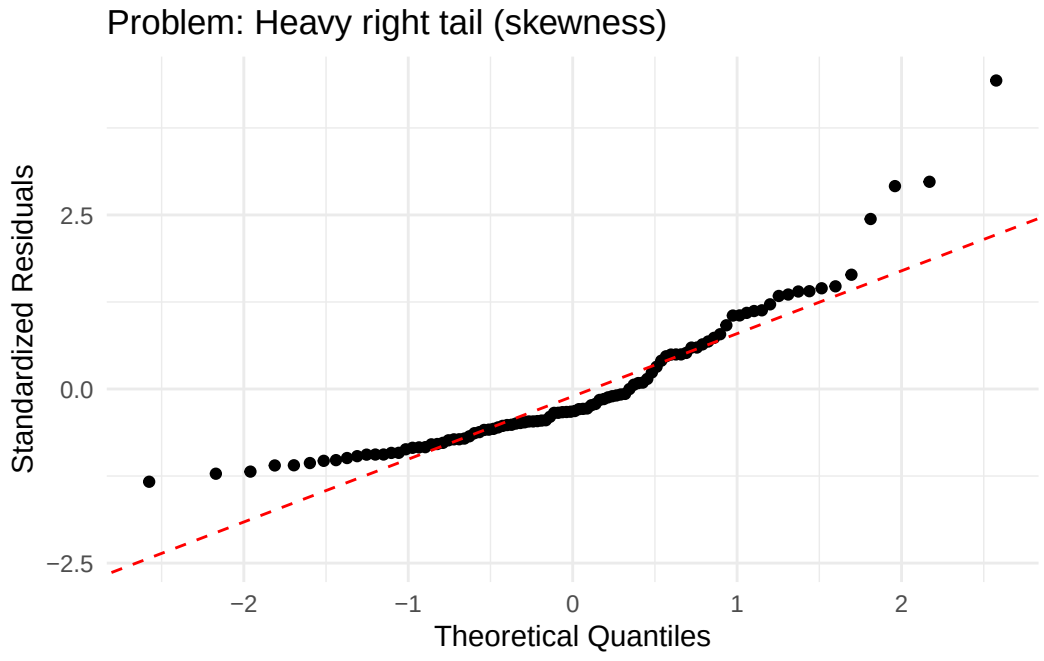

Figure 6: Normal Q-Q plot showing non-normally distributed residuals. The systematic deviation from the diagonal line (particularly in the tails) indicates a violation of the normality assumption.

4.1.3 Plot 3: Scale-Location Plot

Purpose: Checks homoscedasticity more clearly than the Residuals vs Fitted plot.

How it works: Plots the square root of standardized residuals ($\sqrt{|e_i^*|}$) against fitted values. The square root transformation stabilizes variance and makes patterns of heteroscedasticity easier to detect.

What to look for:

- **Good:** Roughly horizontal smoothed line with points evenly spread around it (see Figure 7)
- **Problem:** Smoothed line has a clear trend (upward or downward slope), or the spread of points increases/decreases systematically (see Figure 8)

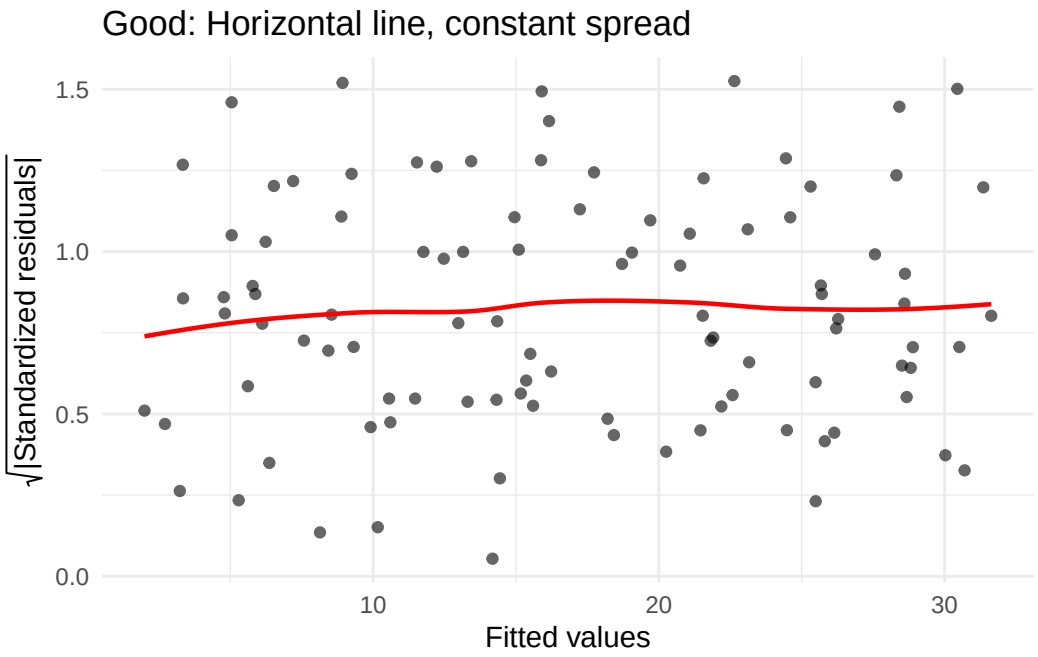

Figure 7: Scale-Location plot showing homoscedasticity. The smoothed line is approximately horizontal and the spread of points is roughly constant, indicating constant variance of residuals.

Interpretation: This plot makes heteroscedasticity violations more apparent than the standard Residuals vs Fitted plot because the square root transformation amplifies patterns in variance.

4.1.4 Plot 4: Residuals vs Leverage

Purpose: Identifies influential observations that disproportionately affect the regression model.

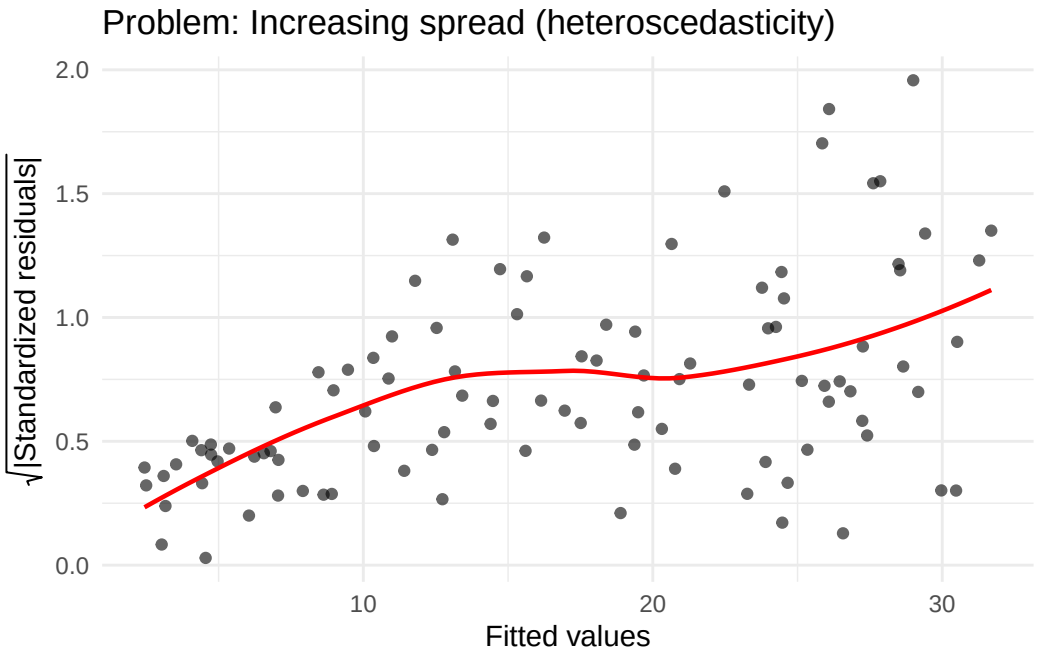

Figure 8: Scale-Location plot showing heteroscedasticity. The upward trend in the smoothed line indicates increasing variance of residuals as fitted values increase.

How it works: Plots standardized residuals against leverage (hat values). Includes Cook's distance contours to identify problematic points that are both unusual in their predictor values (high leverage) and poorly predicted by the model (large residuals).

Understanding Cook's Distance and Contours

Cook's Distance measures the influence of each observation on the fitted values. It quantifies how much all fitted values change when observation i is deleted from the analysis:

$$D_i = \frac{\sum_{j=1}^{n} (\hat{y}_j - \hat{y}_{j(i)})^2}{p \cdot MSE}$$

where $\hat{y}_{j(i)}$ is the predicted value for observation j when observation i is excluded, p is the number of parameters, and MSE is the mean squared error.
Alternatively, Cook's distance combines leverage and residual size:

$$D_i = \frac{(e_i^*)^2}{p} \times \frac{h_i}{1 - h_i}$$

where e_i^* is the standardized residual and h_i is the leverage (hat value).
Interpretation:

- $D_i > 0.5$: Investigate this observation carefully
- $D_i > 1.0$: This observation is highly influential and seriously affects your results

Contours on the plot:

The dashed curves on the Residuals vs Leverage plot show Cook's distance contours (typically at 0.5 and 1.0). Points that fall outside these contours have high influence.
What makes a point influential?
Two factors contribute to influence:

1. **Leverage** (h_i): How unusual the predictor values are. High leverage means the X values are far from the mean.
2. **Residual size**: How poorly the model predicts this observation.

A point needs BOTH high leverage AND a large residual to be truly influential. High leverage alone (if well-predicted) or large residuals alone (if typical X values) are less concerning.
What to do about influential points:

1. Check for data entry errors
2. Determine if the point represents a different population
3. Investigate whether the point provides important information or is an anomaly
4. Consider reporting results both with and without influential observations
5. Never automatically delete influential points - understand why they're influential first

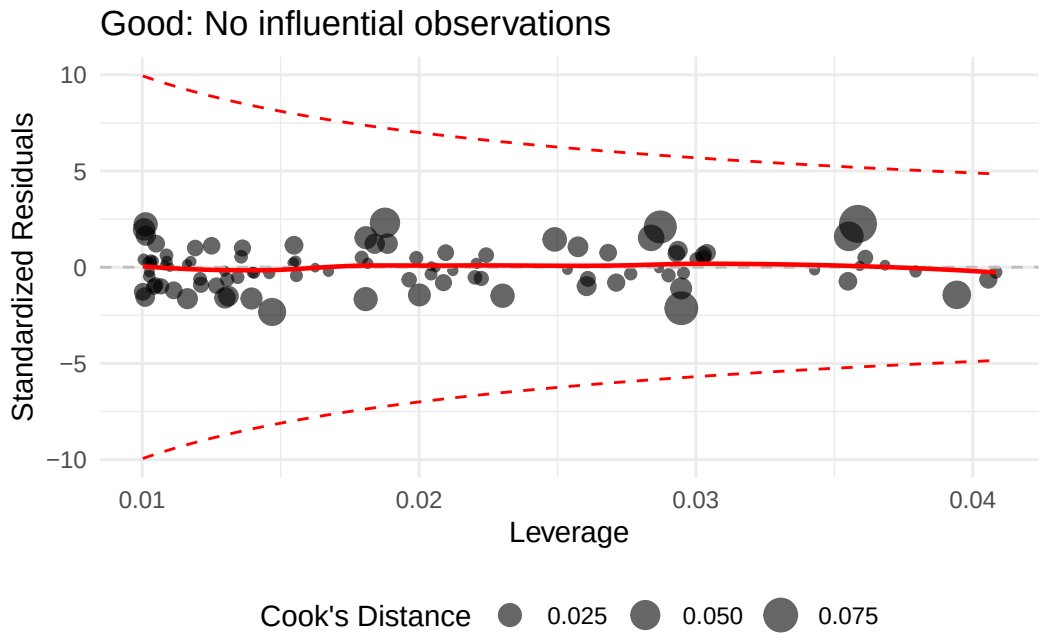

Figure 9: Residuals vs Leverage plot for a well-behaved model. All points fall well within Cook's distance contours (dashed red lines at 0.5 and 1.0), with no concerning leverage or Cook's distance values.

What to look for:

- **Good:** All points have moderate leverage and standardized residuals, with small Cook's distance values (see Figure 9)

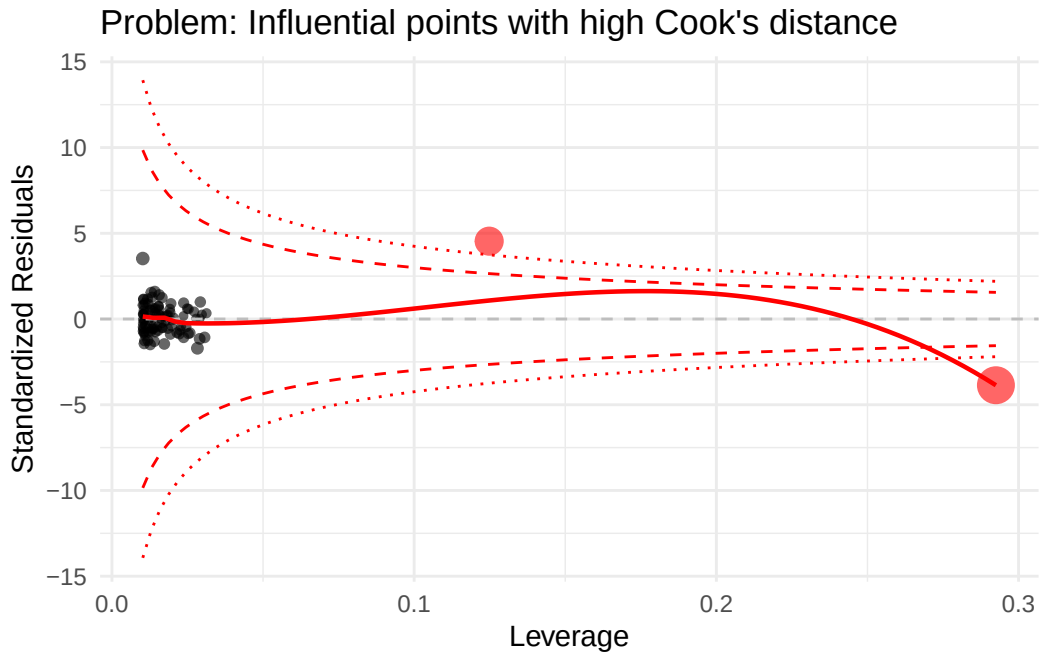

Figure 10: Residuals vs Leverage plot showing influential observations. Several points (shown with larger symbols) have high Cook's distance values and fall outside or near the contours. The dashed red lines show Cook's distance contours at 0.5 and 1.0. Points outside these contours, especially in the corners (high leverage AND large residuals), are particularly concerning.

- **Problem:** Points in the corners (high leverage AND large residuals) with large Cook's distance values (see Figure 10)

Key concepts:

- **Leverage (h_i):** Measures how far an observation's predictor values are from the mean of the predictors. Average leverage is p/n where p is the number of parameters.
- **High leverage:** Points with $h_i > 2p/n$ or $h_i > 3p/n$
- **Influential:** High leverage AND large residual, indicated by large Cook's distance

Important note on outliers vs. influential points:

- **Outliers** have unusual Y values (large residuals) but may not be influential if they have typical X values
- **High leverage points** have unusual X values but may not be influential if well-predicted
- **Influential points** combine both: unusual X values (high leverage) AND large residuals

This fourth diagnostic plot specifically helps identify influential observations that warrant investigation, even though "no influential outliers" is not a formal statistical assumption of linear regression.

---

5 Quantitative Diagnostic Tests (Optional)

While visual diagnostics should always be your primary tool, quantitative tests can provide additional confirmation and are useful when you need to report formal test results. However, be aware that these tests can be overly sensitive in large samples, detecting trivial violations that don't meaningfully affect your results.

5.1 Tests for Normality

5.1.1 Shapiro-Wilk Test

Purpose: Formal test of whether residuals follow a normal distribution.

How it works: Tests the null hypothesis H_0: the data come from a normal distribution. The test statistic (W) measures how well the ordered residuals match the expected pattern from a normal distribution, with W ranging from 0 to 1 (higher values indicate better normality).

R code:

```
shapiro.test(residuals(model))
```

Decision rule:

- If p-value < 0.05: Reject normality assumption (residuals are not normally distributed)
- If p-value 0.05: Cannot reject normality assumption (consistent with normality)

Important notes:

- Very sensitive to sample size (may detect trivial deviations in large samples)
- Should be used in conjunction with QQ plots, not as a replacement
- Moderate violations are often acceptable, especially with n $>$ 30-50 (Central Limit Theorem)
- Maximum sample size is 5000; use Kolmogorov-Smirnov for larger samples

5.1.2 Alternative Normality Tests

Kolmogorov-Smirnov Test:

- More appropriate for larger samples (n $>$ 50)
- Less powerful than Shapiro-Wilk but works with any sample size
- Code: `ks.test(residuals(model), "pnorm", mean=0, sd=sd(residuals(model)))`

Anderson-Darling Test:

- Available in `nortest` package
- Gives more weight to tails than Shapiro-Wilk
- Good for detecting departures in the extremes
- Code: `library(nortest); ad.test(residuals(model))`

5.2 Tests for Homoscedasticity

5.2.1 Breusch-Pagan Test

Purpose: Formal test for heteroscedasticity (non-constant variance).

How it works: Regresses the squared residuals on the predictors to test whether variance is related to predictor values. Under H_0 (homoscedasticity), squared residuals should not be systematically related to predictors.

The test statistic follows a χ^2 distribution with degrees of freedom equal to the number of predictors.

R code:

```
library(lmtest)
bptest(model)
```

Decision rule:

- If p-value < 0.05: Reject homoscedasticity (variance is not constant)
- If p-value 0.05: Cannot reject homoscedasticity (consistent with constant variance)

Solutions if violated:

- Weighted least squares (WLS)
- Robust standard errors (heteroscedasticity-consistent standard errors)
- Variance-stabilizing transformations (log, square root)

5.2.2 White Test

Purpose: More general test for heteroscedasticity than Breusch-Pagan.

How it works: Similar to Breusch-Pagan but includes cross-products and squared terms of predictors when regressing squared residuals. Tests for any form of heteroscedasticity, not just linear relationships between variance and predictors.

R code:

```
library(skedastic)
white_test(model)
# Or using lmtest:
library(lmtest)
bptest(model, ~ fitted(model) + I(fitted(model)^2))
```

When to use: When you suspect complex patterns of heteroscedasticity that Breusch-Pagan might miss, such as variance that changes non-linearly with predictors.

5.3 Test for Independence

5.3.1 Durbin-Watson Test

When to use: Primarily for time series data or any data with a natural ordering where adjacent observations might be correlated.

Purpose: Tests for autocorrelation in residuals (correlation between consecutive residuals).

How it works: Computes a test statistic based on the differences between consecutive residuals:

$$DW = \frac{\sum_{t=2}^{n}(e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2}$$

The DW statistic ranges from 0 to 4:

- **DW 2**: No autocorrelation (good)

- **DW < 2**: Positive autocorrelation (consecutive residuals tend to have the same sign)
- **DW > 2**: Negative autocorrelation (consecutive residuals tend to alternate signs)

R code:

```
library(lmtest)
dwtest(model)
```

Interpretation:

- DW 2: No autocorrelation (ideal)
- DW < 1.5 or > 2.5: Potential concern
- Check p-value for formal test of H_0: no autocorrelation

Solutions if violated:

- Add lagged variables to the model
- Use time series methods (ARIMA, GLS)
- Use robust standard errors that account for autocorrelation
- Consider whether independence assumption is appropriate for your data

Important note: Only tests for first-order autocorrelation. For higher-order autocorrelation, examine ACF (autocorrelation function) plots or use Ljung-Box test.

5.4 Test for Linearity

5.4.1 RESET Test (Ramsey's Regression Equation Specification Error Test)

Purpose: Tests for functional form misspecification - whether the linear form is adequate or if you need transformations/polynomial terms.

How it works: Adds powers of fitted values (e.g., $\hat{y}^2$, $\hat{y}^3$, $\hat{y}^4$) to the original model and tests whether they are jointly significant. If they are, the linear form is inadequate because higher-order terms improve the fit.

The logic: if the relationship is truly linear, powers of fitted values shouldn't add predictive power.

R code:

```
library(lmtest)
resettest(model, power = 2:3)  # Tests squared and cubed terms
```

Decision rule:

- If p-value < 0.05: Linear form is inadequate; consider transformations or polynomial terms

- If p-value 0.05: Linear form appears adequate

Solutions if violated:

- Transform the outcome variable (log, square root, Box-Cox)
- Transform predictor variables
- Add polynomial terms (X^2, X^3)
- Consider non-linear modeling approaches

5.5 Tests for Multicollinearity

5.5.1 Variance Inflation Factor (VIF)

When to use: Multiple regression with 2 or more predictors.

Purpose: Detects multicollinearity (high correlation among predictors), which inflates standard errors and makes it difficult to isolate individual predictor effects.

How it works: For each predictor, VIF measures how much the variance of its coefficient estimate is inflated due to correlation with other predictors:

$$VIF_j = \frac{1}{1 - R_j^2}$$

where R_j^2 is the R^2 from regressing predictor j on all other predictors.

Interpretation: VIF quantifies how much more variance the coefficient has compared to if predictors were uncorrelated. $VIF = 1$ means no correlation; $VIF = 5$ means variance is 5 times larger than if uncorrelated.

R code:

```
library(car)
vif(model)
```

Decision rules:

- **VIF $<$ 5**: Generally acceptable
- **VIF 5-10**: Moderate multicollinearity; investigate
- **VIF $>$ 10**: Serious multicollinearity; action needed

Solutions:

- Remove redundant predictors (keep the most important/interpretable)
- Combine correlated predictors into a composite variable or index
- Use ridge regression or principal components regression
- Collect more data (multicollinearity is a data problem, not a model problem)

Important note: High VIF doesn't bias coefficient estimates, but it inflates standard errors, making it harder to detect significant effects. Your model can still make good predictions with high VIF.

5.6 Tests for Influential Observations

5.6.1 Cook's Distance Analysis

Purpose: Systematically identify influential observations using a quantitative threshold.

How it works: Cook's distance measures how much all fitted values change when observation i is deleted:

$$D_i = \frac{(e_i^*)^2}{p} \times \frac{h_i}{1 - h_i}$$

where e_i^* is the standardized residual, h_i is leverage, and p is the number of parameters.

R code:

```
# Cook's distance plot (automatically labeled)
plot(model, which = 4)

# Extract values for custom analysis
cooks_d <- cooks.distance(model)
influential <- which(cooks_d > 0.5)
highly_influential <- which(cooks_d > 1.0)

# View most influential observations
head(sort(cooks_d, decreasing = TRUE), 10)
```

Decision rules:

- **D $<$ 0.5**: Not influential
- **D $>$ 0.5**: Investigate this observation
- **D $>$ 1.0**: Highly influential; serious concern

What to do with influential points:

1. **Check for errors:** Data entry mistakes, measurement errors, coding errors
2. **Understand why:** Does this observation represent a different population or process?
3. **Investigate context:** Is this a meaningful outlier (e.g., major event) or noise?
4. **Report transparently:** Show results with and without influential observations
5. **Never automatically delete:** Deletion requires substantive justification, not just statistical criteria

Alternative influence measures:

- **DFFITS**: Similar to Cook's D but focuses on change in individual fitted value
- **DFBETAS**: Change in each coefficient when observation is deleted
- **Leverage**: Hat values (h_i); values $> 2p/n$ or $3p/n$ indicate high leverage

---

6 Practical Workflow for Diagnostics

Follow this systematic approach for every regression analysis:

6.1 1. Always Start with Visual Diagnostics

```
# Generate all four diagnostic plots at once
par(mfrow = c(2, 2))
plot(model)
par(mfrow = c(1, 1))
```

Examine each plot carefully:

- Plot 1: Check for patterns in residuals (linearity, homoscedasticity)
- Plot 2: Check for deviations from the diagonal (normality)
- Plot 3: Check for trends in spread (homoscedasticity)
- Plot 4: Check for influential points (Cook's distance)

6.2 2. Run Key Quantitative Tests (Optional)

```
# Normality
shapiro.test(residuals(model))

# Homoscedasticity
library(lmtest)
bptest(model)

# Multicollinearity (if multiple predictors)
library(car)
vif(model)
```

6.3 3. Investigate Issues Identified

If linearity is violated:

```
# Try RESET test
resettest(model)
# Consider transformations or polynomial terms
```

If heteroscedasticity is detected:

```
# Use robust standard errors
library(sandwich)
library(lmtest)
coeftest(model, vcov = vcovHC(model, type = "HC3"))
```

If influential points are found:

```
# Extract Cook's distance
cooks_d <- cooks.distance(model)
# Identify influential observations
influential_obs <- which(cooks_d > 0.5)
# Examine these observations
data[influential_obs, ]
# Refit model without them and compare
model_no_influential <- lm(y ~ x, data = data[-influential_obs, ])
```

If time series data:

```
# Add Durbin-Watson test
dwtest(model)
# Examine residual plot over time
plot(residuals(model), type = "l")
```

6.4 4. Additional Tests When Needed

Use these only when initial diagnostics suggest specific problems:

- **RESET test** if Residuals vs Fitted shows curves
- **White test** if heteroscedasticity pattern is complex
- **Alternative normality tests** if Shapiro-Wilk fails in large samples
- **Cook's distance analysis** if Plot 4 suggests potential influential observations

---

7 Key Principles to Remember

1. **Visual diagnostics come first:** Plots are more informative than p-values, especially in large samples where tests become overly sensitive.
2. **Statistical tests are supplements:** Use quantitative tests to confirm what you see visually, not as a replacement for visual inspection.
3. **Context matters:** Not all violations are equally serious. Linearity and independence are typically most critical; moderate violations of normality are often acceptable with reasonable sample sizes.
4. **Large samples are robust:** With $n > 100$, minor violations of normality and homoscedasticity have minimal practical impact due to the Central Limit Theorem and robustness of OLS.
5. **Investigate influential points, don't automatically delete:** Understand why they're influential first. They may contain important information or indicate model misspecification.
6. **Report transparently:** If you identify violations, acknowledge them and discuss their potential impact. If you exclude influential observations, report results both ways.
7. **Multiple diagnostics:** Don't rely on a single test or plot. A comprehensive assessment examines multiple perspectives on each assumption.
8. **Four assumptions, not five:** Linear regression has four core distributional assumptions (linearity, independence, homoscedasticity, normality). Checking for influential observations is an important diagnostic practice, but it's not a formal assumption of the model.

Remember: Perfect adherence to all assumptions is rare in real data. The goal is to understand where and how assumptions are violated, assess the severity of violations, and make informed decisions about how to proceed with your analysis.