Lecture 13: Heteroskedastic data

Module 4: part 1

Spring 2024

Logistics

- Starting Module 4: What to do when our assumptions are violated
- Module 3 assessment posted, due Mar 16

Model Assumption Violations

Hypothesis Testing & Model Assumptions

Key Points

- Statistical inference relies on distributional assumptions.
- Hypothesis tests compare test statistics to theoretical null distributions.
- These null distributions are derived from model assumptions.
- When assumptions fail, our statistical conclusions may be invalid.

Common Model Violations

Consequences:

- Incorrect p-values and confidence intervals.
- Inflated Type I error rates (false positives).
- Reduced statistical power.
- Biased parameter estimates.

Solutions:

- Robust standard errors.
- Transformation of variables.
- Alternative testing procedures.
- Resampling methods (bootstrap).

Today's focus: Dealing with heteroskedasticity

Linear Model Framework

$$Y_i = \beta_0 + \sum_{j=1}^{p} \beta_j X_{i,j} + \varepsilon_i \tag{1}$$

$$\tag{2}$$

Core Assumptions

- **Linearity:** $\mathbb{E}(Y_i \mid \mathbf{X_i} = \mathbf{x}) = \beta_0 + \sum_j \beta_j x_j$
- **Independence:** $\varepsilon_i \perp\!\!\!\perp \varepsilon_j$ for $i \neq j$
- **Homoskedasticity:** $\mathrm{Var}(\varepsilon_i \mid \mathbf{X_i}) = \sigma^2$ (constant variance)
- **Exogeneity:** $\mathbb{E}(\varepsilon_i \mid \mathbf{X_i}) = 0$ (errors independent of predictors)

Distributional Assumption (for exact tests)

- **Normality:** $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$
- Less critical with large sample sizes (Central Limit Theorem)

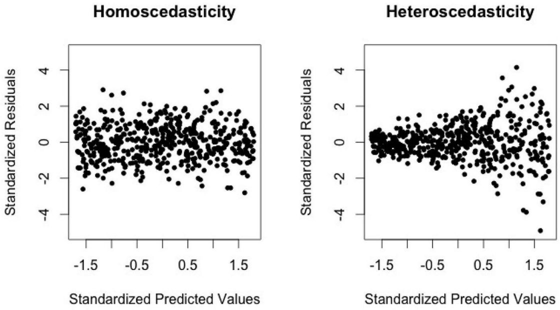
Understanding Heteroskedasticity

Homoskedastic

$$\text{Var}(\varepsilon_i \mid X_i) = \sigma^2$$
(Constant error variance)

Heteroskedastic

$$\text{Var}(\varepsilon_i \mid X_i) = \sigma_i^2$$
(Error variance depends on X)



Common Patterns

- Fan-shaped residuals (variance increases with mean)
- Grouped heteroskedasticity (different groups have different variances)
- Temporal heteroskedasticity (variance changes over time)

Consequences of Heteroskedasticity

Impact on Statistical Inference

- OLS estimators remain **unbiased** and **consistent**.
- Standard errors become **biased** (typically underestimated).
- Hypothesis tests no longer valid (incorrect Type I error rates).
- Confidence intervals have incorrect coverage probabilities.
- OLS no longer the most efficient estimator (BLUE property violated).

Key Insight

When heteroskedasticity is present, the sampling distribution of test statistics differs from what standard theory predicts, invalidating traditional inference.

Detecting Heteroskedasticity

Visual Methods:

- Residual plots (vs. fitted values).
- Residual plots (vs. predictors).
- Scale-location plots.
- Q-Q plots of squared residuals.

Formal Tests:

- Breusch-Pagan test
- White test
- Goldfeld-Quandt test
- NCV test (non-constant variance)

Breusch-Pagan Test (focus of today)

Tests whether estimated variance of residuals depends on the values of independent variables

1. Regress Y on X to obtain residuals $\hat{\varepsilon}_i$
2. Regress $\hat{\varepsilon}_i^2$ on X variables
3. Test if any coefficients in second regression are significant

Breusch-Pagan Test

Main Idea

We can test for whether the data is heteroskedastic using the Breusch–Pagan test.
Objective: Test if the variability of the residuals is associated with the covariates.

1. Fit model and get residuals $\hat{\varepsilon}_i = y_i - \hat{y}_i$
2. Let u_i be a standardized version of $\hat{\varepsilon}_i^2$ so that u_i has mean $= 1$

$$u_i = \frac{\hat{\varepsilon}_i^2}{\frac{1}{n}\sum_i \hat{\varepsilon}_i^2} \Rightarrow \frac{1}{n}\sum_i u_i = 1$$

3. Fit an auxiliary regression model:

$$u_i = \hat{\gamma}_0 + \sum_{k=1}^{p} \hat{\gamma}_k x_{i,k}$$

Intuition

If covariates can predict the squared residuals, then error variance depends on X values (heteroskedasticity).

Breusch-Pagan Test

Hypothesis Testing Framework

Under the null hypothesis that the data generating process is homoskedastic:

$$H_0 \ : \ \gamma_1 = \gamma_2 = \ldots = \gamma_p = 0 \tag{3}$$

$$H_A \ : \ \text{at least one coefficient is non-zero} \tag{4}$$

Breusch-Pagan Test

Hypothesis Testing Framework

Under the null hypothesis that the data generating process is homoskedastic:

$$H_0 \; : \; \gamma_1 = \gamma_2 = \ldots = \gamma_p = 0 \tag{3}$$

$$H_A \; : \; \text{at least one coefficient is non-zero} \tag{4}$$

Test Statistic

Calculate test statistic:

$$L = \frac{1}{2}\left(RSS_0 - RSS(\hat{\gamma})\right)$$

where $RSS_0 = \sum_i (u_i - 1)^2$ and $RSS(\hat{\gamma}) = \sum_i \left(u_i - \hat{\gamma}_0 - \sum_{k=1}^{p} \hat{\gamma}_k x_{i,k}\right)^2$.

Breusch-Pagan Test

Hypothesis Testing Framework

Under the null hypothesis that the data generating process is homoskedastic:

$$H_0 \; : \; \gamma_1 = \gamma_2 = \ldots = \gamma_p = 0 \tag{3}$$
$$H_A \; : \; \text{at least one coefficient is non-zero} \tag{4}$$

Test Statistic

Calculate test statistic:

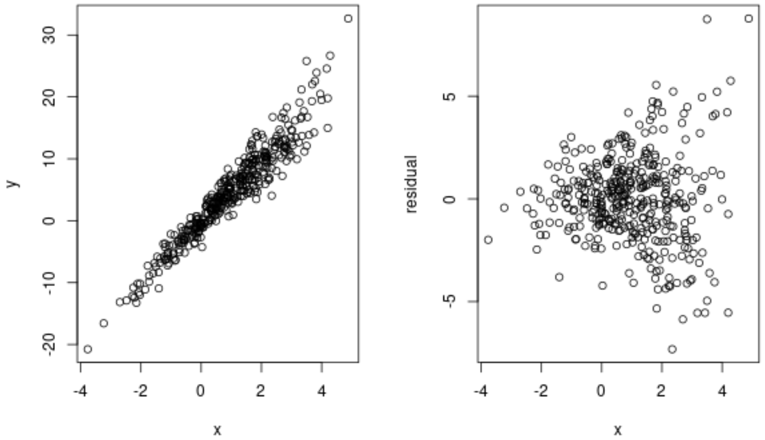
$$L = \frac{1}{2}\left(RSS_0 - RSS(\hat{\gamma})\right)$$

where $RSS_0 = \sum_i (u_i - 1)^2$ and $RSS(\hat{\gamma}) = \sum_i \left(u_i - \hat{\gamma}_0 - \sum_{k=1}^{p} \hat{\gamma}_k x_{i,k}\right)^2$.

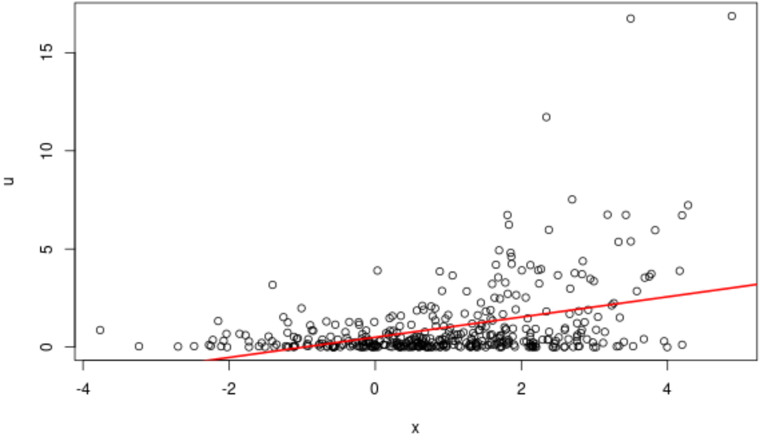
Statistical Decision

Under H_0, $L \sim \chi^2_p$ (chi-squared with p degrees of freedom).
Reject H_0 if $L > \chi^2_{p,\alpha}$ at significance level α.

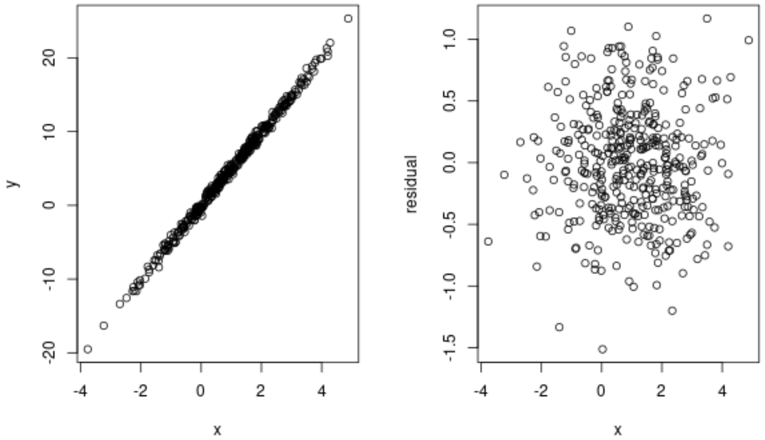
Breusch Pagan Test



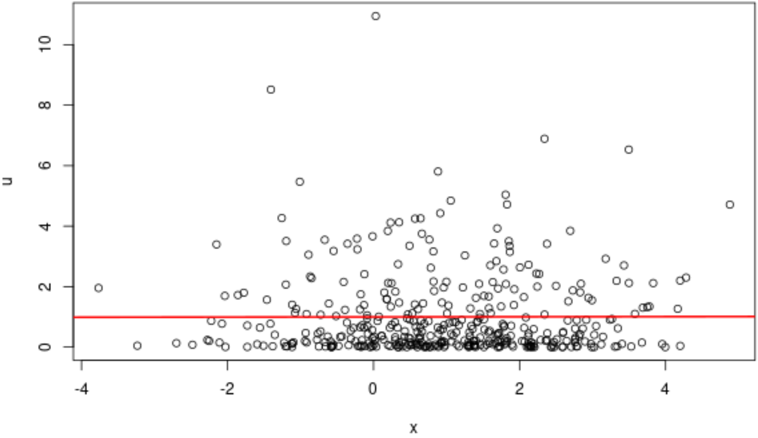
Breusch Pagan Test



Breusch Pagan Test



Breusch Pagan Test



Addressing Heteroskedasticity

Transformation Approaches:

- Log transformation: $\log(Y_i)$
- Power transformations: Y_i^λ
- Box-Cox transformations
- Variance-stabilizing transformations

Robust Inference:

- Heteroskedasticity-Consistent (HC) estimators
- Weighted Least Squares (WLS).

Sandwich Estimator (First HC estimator)

$$\widehat{\text{Var}}(\hat{\beta}) = (X'X)^{-1}X'\text{diag}(\hat{\varepsilon}_i^2)X(X'X)^{-1}$$

Sandwich Estimator for Heteroskedasticity

General Variance Formula for OLS Estimators

Under heteroskedasticity, the true variance of the OLS estimators is:

$$\mathrm{Var}(\hat{\beta} \mid \mathbf{X}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{W}\mathbf{X})(\mathbf{X}'\mathbf{X})^{-1}$$

- $\mathbf{W}$ is a diagonal matrix with $w_{ii} = \mathrm{Var}(\varepsilon_i)/\sigma^2$.
- When homoskedastic, $\mathbf{W} = \mathbf{I}$ and formula simplifies to $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$.
- When heteroskedastic, diagonal elements of $\mathbf{W}$ differ, preventing simplification.

Sandwich Estimator for Heteroskedasticity

General Variance Formula for OLS Estimators

Under heteroskedasticity, the true variance of the OLS estimators is:

$$\mathrm{Var}(\hat{\beta} \mid \mathbf{X}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{W}\mathbf{X})(\mathbf{X}'\mathbf{X})^{-1}$$

- $\mathbf{W}$ is a diagonal matrix with $w_{ii} = \mathrm{Var}(\varepsilon_i)/\sigma^2$.
- When homoskedastic, $\mathbf{W} = \mathbf{I}$ and formula simplifies to $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$.
- When heteroskedastic, diagonal elements of $\mathbf{W}$ differ, preventing simplification.

Sandwich Estimator (White, 1980)

When we substitute estimates for σ^2 and $\mathbf{W}$, we get the "sandwich" estimator:

$$\widehat{\mathrm{Var}(\hat{\beta} \mid \mathbf{X})} = \hat{\sigma}_\varepsilon^2(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\hat{\mathbf{W}}\mathbf{X})(\mathbf{X}'\mathbf{X})^{-1}$$

Robust Inference with Sandwich Estimator

Estimated Variance Matrix

With the sandwich estimator:

$$\widehat{\text{Var}(\hat{\beta} \mid \mathbf{X})} = \hat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\hat{\mathbf{W}}\mathbf{X})(\mathbf{X}'\mathbf{X})^{-1}$$

Hypothesis Testing

$$t = \frac{\hat{\beta}_j - \beta_j^0}{\sqrt{[\widehat{\text{Var}(\hat{\beta} \mid \mathbf{X})}]_{jj}}} \tag{5}$$

- β_j^0 is the hypothesized value.
- $[\cdot]_{jj}$ indicates the j-th diagonal element.

Confidence Intervals

$$\hat{\beta}_j \pm t_{1-\alpha/2,n-p-1}\sqrt{[\widehat{\text{Var}(\hat{\beta} \mid \mathbf{X})}]_{jj}} \tag{6}$$

- Use same formula as standard CI.
- Only the variance estimate changes.

Robust Inference with Sandwich Estimator

Estimated Variance Matrix

With the sandwich estimator:

$$\widehat{\text{Var}(\hat{\beta} \mid \mathbf{X})} = \hat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\hat{\mathbf{W}}\mathbf{X})(\mathbf{X}'\mathbf{X})^{-1}$$

Hypothesis Testing

$$t = \frac{\hat{\beta}_j - \beta_j^0}{\sqrt{[\widehat{\text{Var}(\hat{\beta} \mid \mathbf{X})}]_{jj}}} \tag{5}$$

- β_j^0 is the hypothesized value.
- $[\cdot]_{jj}$ indicates the j-th diagonal element.

Confidence Intervals

$$\hat{\beta}_j \pm t_{1-\alpha/2,n-p-1}\sqrt{[\widehat{\text{Var}(\hat{\beta} \mid \mathbf{X})}]_{jj}} \tag{6}$$

- Use same formula as standard CI.
- Only the variance estimate changes.

Large Sample Properties

As n (the sample size) increases, the hypothesis tests and confidence intervals become valid regardless of whether the data is homoskedastic or heteroskedastic.

Simulation Study: Comparing Methods

Data Generating Process

For fixed covariates $\mathbf{X} = (X_1, X_2, \ldots, X_n)$:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

Model 1 (Homoskedastic):

$$\varepsilon_i \mid X_i \sim N(0, (\bar{x} + 1)/3)$$

Model 2 (Heteroskedastic):

$$\varepsilon_i \mid X_i \sim N(0, (X_i + 1)/3)$$

Simulation Design (TO DO)

- Compare standard OLS inference vs. sandwich estimator inference.
- Sample sizes: $n = 25, 50, 100, 200, 400$.
- Measure Type I error: when $\beta_1 = 1$ and testing $H_0 : \beta_1 = 1$.
- Measure Type II error: when $\beta_1 = 1.1$ and testing $H_0 : \beta_1 = 1$.

Key Questions (TO ANSWER)

- Does the S.E. maintain correct Type I error rates under heteroskedasticity?
- How much power is lost when using the S.E. under homoskedasticity?
- At what sample size does the S.E. perform adequately?

Heteroskedasticity-Consistent (HC) Standard Errors

HC Variants

- **HC0:** Original White estimator ($\hat{\varepsilon}_i^2$).
- **HC1:** Small sample correction ($\frac{n}{n-k}\hat{\varepsilon}_i^2$).
- **HC2:** Leverage adjustment ($\frac{\hat{\varepsilon}_i^2}{1-h_{ii}}$).
- **HC3:** Jackknife-inspired ($\frac{\hat{\varepsilon}_i^2}{(1-h_{ii})^2}$).

where h_{ii} are the diagonal elements of the hat matrix $H = X(X'X)^{-1}X'$.

Weighted Least Squares (WLS)

Approach

If we know the structure of heteroskedasticity: $\text{Var}(\varepsilon_i) = \sigma^2 w_i$

1. Transform the model: $\frac{Y_i}{\sqrt{w_i}} = \frac{\beta_0}{\sqrt{w_i}} + \sum_j \beta_j \frac{X_{ij}}{\sqrt{w_i}} + \frac{\varepsilon_i}{\sqrt{w_i}}$.
2. Apply OLS to transformed model.
3. Result: efficient estimates with correct standard errors.

Challenge

The true weights w_i are typically unknown and must be estimated.

Common Weight Functions

- $w_i = |X_i|$ (variance proportional to predictor).
- $w_i = \hat{Y}_i$ (variance proportional to mean).
- $w_i = \hat{Y}_i^2$ (standard deviation proportional to mean).

Practical Recommendations

When to worry:

- Small samples with clear heteroskedasticity.
- Inference is primary goal (p-values, CIs).
- Prediction intervals needed.
- Efficiency of estimates matters.
- Financial or economic data.

Best practices:

- Always check for heteroskedasticity
- Use HC standard errors as default approach.
- Consider transformations for severe cases.
- Report results with and without corrections.
- Use bootstrap for complex situations.

Summary

- Heteroskedasticity affects inference but not point estimates.
- HC standard errors provide simple, robust solution in most cases.
- Transformations can address both heteroskedasticity and non-linearity.
- WLS is most efficient when variance structure is known.