Violating Assumptions
Jerome Dumortier
Overview
Non-Constant Error Variance
Theoretical Concepts
Testing for Heteroscedasticity
Correcting for Heteroscedasticity
Multicollinearity
Theoretical Concepts
Detection and Variance Inflation Factors (VIF)
VIF Example
Autocorrelation
Causes
Omitted Variables

Violating Assumptions

Jerome Dumortier

18 February 2025

Required Packages

The following packages are needed for the material presented in the slides

- car
- lmtest
- MASS
- nlme
- orcutt
- prais
- sandwich

Assumptions

Key assumptions underlying the ordinary least square (OLS) model

1. **Linear in coefficients**: Linear relationship between y and $x_1, \ldots, x_k$
2. **Zero mean of the error terms**: $E(\epsilon | x_1, \ldots, x_k) = 0$ and normally distributed error terms
3. **Homoscedasticity**: $Var(\epsilon_i) = \sigma^2$
4. **No autocorrelation between error terms**: $Cov(\epsilon_i, \epsilon_j) = 0$ for all $i \neq j$
5. **Exogeneity of independent variables**: $E(\epsilon_i | x_1, \ldots, x_k)$, i.e., independent variables contain no information to predict error terms
6. **Full rank** (linear independence of all columns) of X (matrix of independent variables): Perfect multicollinearity (i.e., one independent variable being perfectly predicted from a linear combination of one or more other independent variables) leads to a rank deficiency of X

Topics Covered

Non-constant error (ϵ_i) variance, i.e., heteroscedasticity

- Testing for heteroscedasticity using the Goldfeld-Quandt Test (1965) and the Breusch-Pagan-Godfrey Test (1979)
- Correcting for heteroscedasticity by using heteroskedasticity-consistent (robust) standard errors

Multicollinearity

- Detecting multicollinearity with Variance Inflation Factors (VIF)

Autocorrelation

Overview

Homoscedasticity

$$Var(\epsilon_i) = \sigma^2$$

Heteroscedasticity

$$Var(\epsilon_i) = \sigma_i^2$$

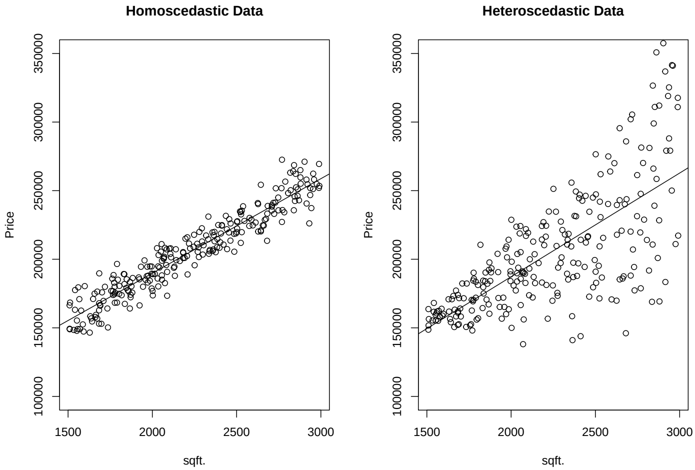
It can be shown that

$$Var(\hat{\beta}_1) = \underbrace{\frac{\sigma_i^2}{\sum x_i^2}}_{Hetero.} \neq \underbrace{\frac{\sigma^2}{\sum x_i^2}}_{Homo.}$$

Notes

- Coefficient estimates and R^2 are unaffected by heteroscedasticity
- Variance of β_1 is larger

Homoscedastic vs. Heteroscedastic Data



Examples and Effects of Heteroscedasticity I

Examples

- *Income, savings, and consumption*: People with higher incomes tend to have more variability in their savings and expenditures whereas low-income individuals spend close to their income
- *Firms and dividends*: Companies with larger profits show more variability in dividend payments
- *Education and income*: Wages may be more predictable for lower education levels while higher education degrees introduce greater variability due to differences in occupation, industry, and experience
- *House price and square footage*: Small price variations for smaller homes compared to larger homes

Examples and Effects of Heteroscedasticity II

Examples

- *Municipal budget variability and city size*: Larger cities experience greater fluctuations in budget expenditures due to the complexity and unpredictability of managing diverse public services
- *Public program effectiveness and demographics*: Policy interventions show more variable outcomes in diverse populations (e.g., in terms of socio-economics) compared to more homogeneous communities, leading to inconsistent program effectiveness

Effects of heteroscedasticity

- Requirement of homoscedasticity for t-test, F-test, and confidence intervals
- F-statistics no longer have the F-distribution
- Bottom line: Hypothesis tests on the β coefficients are no longer valid

Generalized Least Squares (GLS) I

If σ_i^2 was known:

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

Dividing both sides by the known variance:

$$\frac{y_i}{\sigma_i} = \beta_0 \cdot \frac{1}{\sigma_i} + \beta_1 \cdot \frac{x_i}{\sigma_i} + \frac{\epsilon_i}{\sigma_i}$$

If $\epsilon_i^* = \epsilon_i / \sigma_i$, then it can be shown that $Var(\epsilon_i^*) = 1$, i.e., constant.

Generalized Least Squares (GLS) II

Regular OLS

$$\sum_{i=1}^{N} \epsilon_i^2 = \sum_{i=1}^{N} (y_i - \beta_0 + \beta_1 \cdot x_i)^2$$

GLS with $w_i = 1/\sigma_i$

$$\sum_{i=1}^{N} w_i \cdot \epsilon_i^2 = \sum_{i=1}^{N} w_i \cdot (y_i - \beta_0 + \beta_1 \cdot x_i)^2$$

GLS: Minimization of the weighted sum of squared residuals

Generalized Least Squares (GLS) III

Implementation of GLS

1. Estimate the heteroscedasticity structure, e.g., using a White test or Breusch-Pagan test
2. Model the variance function σ_i^2, e.g., as a function of explanatory variables
3. Compute weights $w_i = 1/\hat{\sigma}_i$.
4. Transform the dependent and independent variables using those weights
5. Perform weighted least squares (WLS) regression on the transformed data

Goldfeld-Quandt Test: Steps

Steps for the Goldfeld-Quandt Test

1. Sorting observations in ascending order of an independent variable likely introducing heteroscedasticity
2. Choosing c as the number of central observations to drop resulting in sample sizes $n_1 = n_2 = (n-c)/2$
3. Running two separate regression equations
4. Compute λ with k as the number of coefficients to be estimated including the intercept
$$\lambda = \frac{RSS_2/(n_2-k)}{RSS_1/(n_1-k)}$$
5. λ follows F-distribution and a hypothesis test can be conducted

Goldfeld-Quandt Test: Manual Implementation

Setup

- Example using gqdata with *sqft* being sorted in ascending order,
- $C = 4$

```
gqdata1          = gqdata[1:20,]
gqdata2          = gqdata[31:50,]
bhat             = lm(price~sqft,data=gqdata)
bhat1            = lm(price~sqft,data=gqdata1)
bhat2            = lm(price~sqft,data=gqdata2)
sum(bhat2$residuals^2)/sum(bhat1$residuals^2)
```

```
## [1] 2.826607
```

Goldfeld-Quandt Test: R Function

```
gqtest(bhat,fraction=10)
```

```
##
##  Goldfeld-Quandt test
##
## data:  bhat
## GQ = 2.8266, df1 = 18, df2 = 18, p-value = 0.01665
## alternative hypothesis: variance increases from segment 1 to 2
```

Breusch-Pagan-Godfrey Test: Setps

Steps for the Breusch-Pagan-Godfrey Test

1. Run a regular OLS model and obtain the residuals
2. Calculate
$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{N} \epsilon_i^2}{N}$$
3. Construct the variable $p_i = \epsilon_i^2 / \hat{\sigma}^2$
4. Run a regression as follows with x_i as the independent variables from the original regression
$$p = \alpha_0 + \alpha_1 \cdot x_1 + \alpha_2 \cdot x_2 + \dots$$
5. Obtain the explained sum of squares (ESS) and define $\Theta = 0.5 \cdot ESS$. Then $\Theta \sim \chi^2_{m-1}$.

Or simply use `bptest(bhat)` in R

Breusch-Pagan-Godfrey Test: R Function

```r
bptest(bhat)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  bhat
## BP = 3.8751, df = 1, p-value = 0.04901
```

Robust Standard Errors: Steps

Robust standard (heteroscedasticity-consistent) errors

- Estimation of a covariance matrix (usually denoted Ω in books)

Steps in R

1. Estimation of a regular OLS model
2. Estimation of a covariance matrix using `vcovHC()` from the sandwich package
3. Applying the function `coeftest()` from the nlme package

Simultaneous execution of steps 2 and 3

Robust Standard Errors: Methods

HC0: Default heteroscedasticity-consistent (HC) standard error estimator

- Uses squared residuals without any adjustment
- Suitable for large samples

HC1: Adjusts HC0 for small sample bias by scaling the residuals

- Equivalent to HC0 multiplied by $n/(n-k)$ where k is the number of independent variables

HC2: Corrects for leverage effects in small samples

- Division of squared residuals by $1-h_i$ where $0 \leq h_i \leq 1$ is the leverage of observation i (i.e., influence of i on regression coefficients)

HC3: Additional adjustment compared to HC2 for small sample size

- Division of squared residuals by $(1-h_i)^2$

Robust Standard Errors: Implementation

```
bhat = lm(price~sqft,data=gqdata)
b1   = coeftest(bhat,vcov=vcovHC(bhat,type="HC0"))
b2   = coeftest(bhat,vcov=vcovHC(bhat,type="HC1"))
b3   = coeftest(bhat,vcov=vcovHC(bhat,type="HC2"))
b4   = coeftest(bhat,vcov=vcovHC(bhat,type="HC3"))
```

Robust Standard Errors: Implementation

```
##
## ===================================================================================
##                                          Dependent variable:
##                    ---------------------------------------------------------------
##                            price
##                             OLS                         coefficient
##                                                             test
##                             (1)           (2)       (3)       (4)       (5)
## -----------------------------------------------------------------------------------
## sqft                     73.911***     73.911*** 73.911*** 73.911*** 73.911***
##                           (8.062)       (6.894)   (7.036)   (7.088)   (7.289)
## -----------------------------------------------------------------------------------
## Observations                50
## R2                         0.636
## Adjusted R2                0.629
## Residual Std. Error  24,873.760 (df = 48)
## F Statistic         84.045*** (df = 1; 48)
## ===================================================================================
## Note:                                              *p<0.1; **p<0.05; ***p<0.01
```

Problem with Multicollinearity

From the book Basic Econometrics by Gujarati:

"*If multicollinearity is perfect [. . .], the regression coefficients of the X variables are indeterminate and their standard errors are infinite. If multicollinearity is less than perfect [. . .], the regression coefficients, although determinate, possess large standard errors (in relation to the coefficients themselves), which means the coefficients cannot be estimated with great precision or accuracy.*"

Overview

Non-Constant Error Variance

Theoretical Concepts

Testing for Heteroscedasticity

Correcting for Heteroscedasticity

Multicollinearity

Theoretical Concepts

Detection and Variance Inflation Factors (VIF)

VIF Example

Autocorrelation

Causes

Omitted Variables

Overview

Perfect multicollinearity with λ_i representing constants that are not all zero simultaneously

$$\lambda_1 \cdot x_1 + \lambda_2 \cdot x_2 + \cdots + \lambda_k \cdot x_k = 0$$

Example

$$x_1 = \{8, 12, 15, 45\}$$

$$x_2 = \{24, 36, 15, 51\}$$

$\lambda_1 = 1$ and $\lambda_2 = -1/3$ are such that $x_1 - 1/3 \cdot x_2 = 0$. Multicollinearity refers to linear relationships and including a squared or cubed term does not represent multicollinearity

Examples

Estimation of energy consumption based on income and home size

- Likely high correlation between income and house size

Estimation of education quality (e.g., test scores, graduation rates) based on public spending (e.g., per-capita education budget, teacher salary, and number of schools)

- Correlation between education budget and teacher salary as well as education budget and number of schools

Estimation of crime based on crime prevention policies and public safety expenditures

- Likely correlation of public safety expenditures and the ability to fund crime prevention policies

Over-determined model

- Number of variables k larger than number of observations n

Indications of Multicollinearity

Signs of multicollinearity

- High R^2 but few significant variables
- Failure to reject H_0 (i.e.,$\beta_i = 0$) based on t-values but rejection of F-test (i.e., all slopes being simultaneously zero)
- High correlation among explanatory variables
- Variation of statistically significant variables between models that include different sets of independent variables

Consequences of multicollinearity

- Increase in variances of coefficients

Overview

Non-Constant Error Variance

Theoretical Concepts

Testing for Heteroscedasticity

Correcting for Heteroscedasticity

Multicollinearity

Theoretical Concepts

Detection and Variance Inflation Factors (VIF)

VIF Example

Autocorrelation

Causes

Omitted Variables

VIF: Overview

Purpose

- Identification of possible correlation among multiple independent variables and not just two as in the case of a correlation coefficient
- Detect inflated variance based on multicollinearity

Theoretical aspects

- Existence of a VIF for each independent variable in the model

Regressing each independent variable on all other independent variables

VIF: Calculation and Interpretation

Calculation

- VIF for variable k

$$VIF_k = \frac{1}{1 - R_k^2}$$

Interpretation

- $VIF = 1$: No relationship between the variable x_k and the remaining independent variables
- $VIF > 1$: Some degree of multicollinearity
- $VIF > 4$: Warrants attention
- $VIF > 10$: Indication of serious problem

The latter two are rules of thumb

Setup

Data used: `bloodpressure`

- Patient ID (*pt*), blood pressure (*bp*), body surface area (*bsa*), and duration of hypertension (*dur*)

Correlation matrix

```
##          bp  age weight  bsa  dur pulse stress
## bp     1.00 0.66   0.95 0.87 0.29  0.72   0.16
## age    0.66 1.00   0.41 0.38 0.34  0.62   0.37
## weight 0.95 0.41   1.00 0.88 0.20  0.66   0.03
## bsa    0.87 0.38   0.88 1.00 0.13  0.46   0.02
## dur    0.29 0.34   0.20 0.13 1.00  0.40   0.31
## pulse  0.72 0.62   0.66 0.46 0.40  1.00   0.51
## stress 0.16 0.37   0.03 0.02 0.31  0.51   1.00
```

Regular OLS Regression Results

```
##
## ===========================================
##                    Dependent variable:
##                ----------------------------
##                            bp
## -------------------------------------------
## age                  0.703*** (0.050)
## weight               0.970*** (0.063)
## bsa                   3.776** (1.580)
## dur                     0.068 (0.048)
## pulse                  -0.084 (0.052)
## stress                  0.006 (0.003)
## -------------------------------------------
## Observations              20
## R2                       0.996
## F Statistic    560.641*** (df = 6; 13)
## ===========================================
## Note:          *p<0.1; **p<0.05; ***p<0.01
```

Overview
Non-Constant Error Variance
Theoretical Concepts
Testing for Heteroscedasticity
Correcting for Heteroscedasticity
Multicollinearity
Theoretical Concepts
Detection and Variance Inflation Factors (VIF)
VIF Example
Autocorrelation
Causes
Omitted Variables

VIF Calculation

Using the function `vif` from the package car:

```
vif(bhat1)
```

```
##      age   weight      bsa      dur    pulse   stress
## 1.762807 8.417035 5.328751 1.237309 4.413575 1.834845
```

VIF: Calculation of VIF for `weight`

```
##
## ========================================
##                    Dependent variable:
##              ---------------------------
##                          weight
## ----------------------------------------
## age                   -0.145 (0.206)
## bsa                  21.422*** (3.465)
## dur                    0.009 (0.205)
## pulse                 0.558*** (0.160)
## stress                -0.023 (0.013)
## ----------------------------------------
## Observations                20
## R2                         0.881
## F Statistic     20.768*** (df = 5; 14)
## ========================================
## Note:        *p<0.1; **p<0.05; ***p<0.01
```

Overview

Non-Constant Error Variance

Theoretical Concepts

Testing for Heteroscedasticity

Correcting for Heteroscedasticity

Multicollinearity

Theoretical Concepts

Detection and Variance Inflation Factors (VIF)

VIF Example

Autocorrelation

Causes

Omitted Variables

VIF: Manual Calculation

The results indicate that $R^2 = 0.881$ then

$$VIF = \frac{1}{1 - 0.881} = 8.403361$$

Solution:

- Eliminate `BSA` because `weight` is easier to obtain.
- `Pulse` may be an issue as well.

VIF: Final Regression

```
##
## ============================================================
##                           Dependent variable:
##              -----------------------------------------------
##                                   bp
##                         (1)                     (2)
## ------------------------------------------------------------
## age               0.703*** (0.050)       0.732*** (0.056)
## weight            0.970*** (0.063)       1.099*** (0.038)
## bsa                3.776** (1.580)
## dur                 0.068 (0.048)          0.064 (0.056)
## pulse              -0.084 (0.052)       -0.137** (0.054)
## stress              0.006 (0.003)         0.007* (0.004)
## ------------------------------------------------------------
## Observations             20                      20
## R2                      0.996                   0.994
## F Statistic  560.641*** (df = 6; 13) 502.503*** (df = 5; 14)
## ============================================================
## Note:                            *p<0.1; **p<0.05; ***p<0.01
```