Violating Assumptions

Jerome Dumortier

17 August 2023

Overview

Required Packages

For this lecture, we will need the following packages:

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

Assumptions

Important assumptions for unbiasedness:

- A1: Linear regression model, i.e., linear in terms of coefficients
- A2: Zero mean value of disturbances ϵ, i.e., $E(\epsilon_i|x_i) = 0$
- A3: Homoscedasticity or equal variance of ϵ_i, i.e., $Var(\epsilon_i) = \sigma^2$
- A4: No autocorrelation between the disturbance terms, i.e., $Cov(\epsilon_i, \epsilon_j) = 0$
- A5: No covariance between ϵ_i and x_i
- A6: Number of observations is greater than number of parameters to be estimated
- A7: No multicollinearity

Topics Covered

Violation of the following assumptions

- Non-constant error variance, i.e., heteroscedasticity (A3)
  - Detection of heteroscedasticity
  - Heteroskedasticity-consistent (robust) standard errors
- Multicollinearity (A7)
  - Variance Inflation Factor
- Autocorrelation (A4)

Non-Constant Error Variance

Overview

Homoscedasticity:

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

Heteroscedasticity:

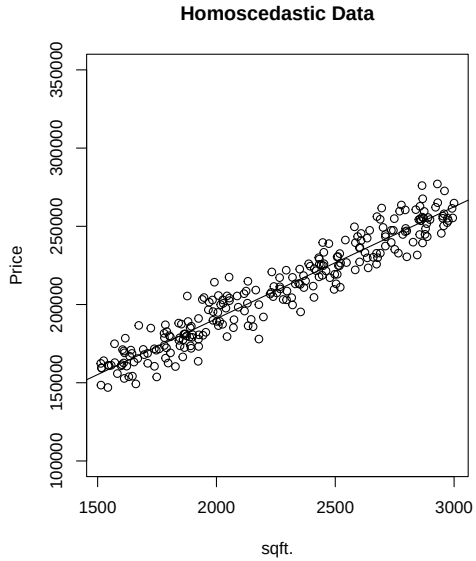
$$Var(\epsilon_i) = \sigma_i^2$$

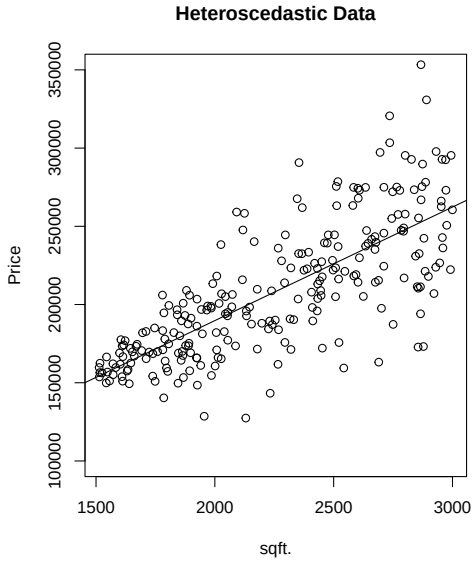
It can be shown that

$$Var(\hat{\beta}_1) = \frac{\sum x_i^2 \cdot \sigma_i^2}{(\sum x_i^2)^2}$$

Unbiasedness of the OLS estimator is not affected but the variance of β_1 will be larger compared to other estimators. Note that the measure of R^2 is unaffected by heteroscedasticity.

Homoscedastic vs. Heteroscedastic Data





Causes and Examples of Heteroscedasticity

Examples:

- Variance of the unobserved factors affecting savings increases with income.
- As the hours of typing practice increases, the variance in typing errors decreases. Practice has the effect of both decreasing the average number of typing errors and shrinking the range of the typing errors sampling distribution.
- Companies with larger profits show more variability in dividend payments.

Homoscedasticity is needed to justify t-test, F-test and confidence intervals:

- F-statistics no longer have the F-distribution.

In short, 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

Usual OLS:

$$\sum_{i=1}^{N} e_i^2 = \sum_{i=1}^{N} \left(y_i - \hat{\beta}_0 + \hat{\beta}_1 \cdot x_i \right)^2$$

GLS:

$$\sum_{i=1}^{N} w_i \cdot e_i^2 = \sum_{i=1}^{N} w_i \cdot \left(y_i - \hat{\beta}_0 + \hat{\beta}_1 \cdot x_i \right)^2$$

GLS minimizes the weighted sum of the residual squares.

Standard Errors of the Coefficients

Note that there are multiple variations to calculate the standard error and thus, it is possible for slight variations among the results from different packages:

$$Var(\hat{\beta}_1) = \frac{\sum_{i=1}^{N}(x_i - \bar{x})^2 \cdot e_i^2}{\sum_{i=1}^{N}(x_i - \bar{x})^2}$$

The square root of the following equation is called heteroscedastic-robust standard error:

$$\widehat{Var}(\hat{\beta}_j) = \frac{\sum_{i=1}^{N} \hat{r}_{ij}^2 \cdot e_i^2}{\sum_{i=1}^{N}(x_i - \bar{x})^2}$$

Standard errors can be either larger or smaller. Note that in this example, we do not know whether heteroscedasticity is present or not.

Tests for Heteroscedasticity

Methods to detect heteroscedasticity:

- Goldfeld-Quandt Test (1965)
- Breusch-Pagan-Godfrey Test (1979)

Goldfeld-Quandt Test

Steps for the GQ-Test:

1. Sort observations by ascending order of the dependent variable.
2. Pick C as the number of central observations to drop in the middle of the dependent variable.
3. Run two separate regression equations.
4. Compute
$$\lambda = \frac{RSS_2/df}{RSS_1/df}$$
5. λ follows an F-distribution and a hypothesis test can be conducted.

Following slides: Example using `gqtestdata`, which is already sorted in ascending order, using $C = 4$.

Goldfeld-Quandt Test in R I

```
gqtestdata1     = gqtestdata[1:13,]
gqtestdata2     = gqtestdata[18:30,]
bhat            = lm(price~sqft,data=gqtestdata)
bhat1           = lm(price~sqft,data=gqtestdata1)
bhat2           = lm(price~sqft,data=gqtestdata2)
lambda          = sum(bhat2$residuals^2)/sum(bhat1$residuals^2)
```

Goldfeld-Quandt Test in R II

```
gqtest(bhat,fraction=4)
```

```
##
##  Goldfeld-Quandt test
##
## data:  bhat
## GQ = 2.7805, df1 = 21, df2 = 21, p-value = 0.01165
## alternative hypothesis: variance increases from segment 1 to 2
```

Breusch-Pagan-Godfrey Test

Steps for the BPG-Test

1. Run a regular OLS model and obtain the residuals.
2. Calculate
$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{N} e_i^2}{N}$$
3. Construct the variable p_i as follows: $p_i = e_i^2/\hat{\sigma}^2$
4. Regress p_i on the X's as follows
$$p_i = \alpha_0 + \alpha_1 \cdot x_{i1} + \alpha_2 \cdot x_{i2} + \ldots$$
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

```
bptest(bhat)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  bhat
## BP = 7.9195, df = 1, p-value = 0.00489
```

Correcting for Heteroscedasticity in R I

Robust standard errors

- Correction of standard errors to account for conditional heteroscedasticity

Run regular OLS model

```
bhat = lm(price~sqft,data=gqtestdata)
```

Correcting for Heteroscedasticity in R II

```
summary(bhat)
```

```
##
## Call:
## lm(formula = price ~ sqft, data = gqtestdata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -107626  -13190    4920   15834   68901
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 64384.92   20848.33   3.088  0.00334 **
## sqft           60.11       9.32   6.449 5.14e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 29680 on 48 degrees of freedom
## Multiple R-squared:  0.4642, Adjusted R-squared:  0.4531
## F-statistic: 41.59 on 1 and 48 DF,  p-value: 5.144e-08
```

Correcting for Heteroscedasticity in R III

```
coeftest(bhat,vcov.=vcovHC)
```

```
##
## t test of coefficients:
##
##               Estimate Std. Error t value  Pr(>|t|)
## (Intercept) 64384.923  21076.873  3.0548   0.00367 **
## sqft           60.110     11.017  5.4561 1.679e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Other Issues and Problems with Data

More serious problems than heteroscedasticity:

- Functional form misspecification
- Measurement error
- Missing data, non-random samples, and outliers

Missing Data and Non-Random Samples

Consequences and remedies:

- Standard regression model is not possible with missing values
- All statistical software packages ignore missing data

Missing data is a minor problem if it is due to random error. Missing data can be problematic if it is systematically missing

- Missing education data for people with lower education
- Missing IQ scores from people with higher IQ's

Examples of exogenous sample selection or sample selection based on the independent variable

Multicollinearity

Introduction

Perfect multicollinearity

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

where λ_i are constants that are not all zero simultaneously. Consider the following example:

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

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

In this case, $\lambda_1 = 1$ and $\lambda_2 = -1/3$, i.e., $x_1 - 1/3 \cdot x_2 = 0$. Note, multicollinearity refers to linear relationships! Including a squared or cubed term is not an issue of multicollinearity.

Problem

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.*"

Possible Reasons and Examples

Model construction

- Measuring electricity or natural gas consumption based on income and house size
- Income and house size are correlated

Over-determined model

- Number of variables k larger than number of observations n

Additional reason later discussed in the semester:

- Common trend, i.e., variables increase or decrease over time. Mostly an issue with data on population, income, wealth, etc.

Detection

It can be shown that the variance of the estimator increases in the presence of multicollinearity. Signs that indicate that some multicollinearity is present:

- High R^2 but few significant variables
- Fail to reject the hypothesis for H_0: $\beta_i = 0$ based on t-values but rejection all slopes being simultaneously zero based on F-test.
- High correlation among explanatory variables
- Variation of statistically significant variables between models.

Variance Inflated Factors (VIF)

- Identifies possible correlation among multiple independent variables and not just two as in the case of a simple correlation coefficient.

Variance Inflated Factors (VIF)

Overview

Consider the model:

$$y_i = \beta_0 + \beta_k \cdot x_{ik} + \epsilon_i$$

The estimated variances of the coefficient β_k is written as

$$Var(\beta_k)^* = \frac{\sigma^2}{\sum_{i=1}^{N}(x_{ik} - \bar{x}_k)^2}$$

Without any multicollinearity, this variance is minimized. If some some independent variables are correlated with the independent variable k, then

$$Var(\beta_k) = \frac{\sigma^2}{\sum_{i=1}^{N}(x_{ik} - \bar{x}_k)^2} \cdot \frac{1}{1 - R_k^2}$$

where R_k^2 is the R^2 if variable x_k is taken as the dependent variable.

Overview II

VIF can be written as

$$\frac{Var(\beta_k)}{Var(\beta_k)^*} = \frac{1}{1 - R_k^2}$$

If $VIF = 1$, then there is no relationship between the variable x_k and the remaining independent variables. Otherwise, $VIF > 1$. In general, the interpretation is as follows:

- VIF of 4 warrants attention
- VIF of 10 indicates a serious problem.

Example on the following slides: `bloodpressure`

- Blood pressure (BP), body surface area (BSA), and duration of hypertension (Dur).

VIF Example using Blood Pressure Data I

```
round(cor(bloodpressure),3)
```

```
##              pt    bp   age weight    bsa   dur pulse stress
## pt        1.000 0.031 0.043  0.025 -0.031 0.176 0.112  0.343
## bp        0.031 1.000 0.659  0.950  0.866 0.293 0.721  0.164
## age       0.043 0.659 1.000  0.407  0.378 0.344 0.619  0.368
## weight    0.025 0.950 0.407  1.000  0.875 0.201 0.659  0.034
## bsa      -0.031 0.866 0.378  0.875  1.000 0.131 0.465  0.018
## dur       0.176 0.293 0.344  0.201  0.131 1.000 0.402  0.312
## pulse     0.112 0.721 0.619  0.659  0.465 0.402 1.000  0.506
## stress    0.343 0.164 0.368  0.034  0.018 0.312 0.506  1.000
```

VIF Example using Blood Pressure Data II

Output of the following regression is omitted due to space constraints:

```
bhat=lm(bp~age+weight+bsa+dur+pulse+stress,data=bloodpressure)
summary(bhat)
```

VIF Example using Blood Pressure Data III

Using the function `vif` from the package "cars":

```
vif(bhat)
```

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

Variance Inflated Factors Example using Blood Pressure Data III

Remove `Weight` from the regression equation:

```
bhat = lm(weight~age+bsa+dur+pulse+stress,data=bloodpressure)
summary(bhat)
```

Manual Calculation of VIF

The results indicate that $R^2 = 0.8812$ then

$$VIF = \frac{1}{1 - 0.8812} = 8.417508$$

Solution:

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

```
bhat=lm(bp~age+weight+dur+pulse+stress,data=bloodpressure)
```

Autocorrelation

Correlated Error Terms

Data available in research:

- Cross-sectional data (multiple observations at same time point)
- Time series data (one variable observed over time)
- Pooled data (multiple observations at different time points) and panel data (same observations at different time points)

Serial correlation versus autocorrelation:

- Serial correlation: Correlation between two series
- Autocorrelation: Correlation with lagged variables

Consequences

- OLS estimator is still unbiased but there is no longer minimum variance since $E(\epsilon_i \epsilon_j) \neq 0$.

Autocorrelation unlikely for cross-sectional data except for spatial auto-correlation

Causes of Autocorrelation I

Inertia

- Upward trend after recession of variables such as income, production, employment, prices.

Specification bias: Excluded variable

- Correct equation

$$Q_{beef} = \beta_0 + \beta_1 \cdot P_{beef} + \beta_2 \cdot P_{income} + \beta_3 \cdot P_{pork} + \epsilon_t$$

- Estimated equation

$$Q_{beef} = \beta_0 + \beta_1 \cdot P_{beef} + \beta_2 \cdot P_{income} + \upsilon_t$$

- Systematic pattern since

$$\upsilon_t = \beta_3 \cdot P_{pork} + \epsilon_t$$

Causes of Autocorrelation II

Specification bias: Incorrect functional form

- Correct equation

$$y_i = \beta_0 + \beta_1 \cdot x_i + \beta_2 \cdot x_i^2 + \epsilon_i$$

- Estimated equation

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

Cobweb phenomenon (e.g., production decision before prices are observed such as in agriculture):

$$supply_t = \beta_0 + \beta_1 \cdot p_{t-1}$$

Lags:

$$consumption_t = \beta_0 + \beta_1 \cdot income_t + \beta_3 \cdot consumption_{t-1} + \epsilon_t$$

First-order Autoregressive Scheme

Consider the model:

$$y_t = \beta_0 + \beta_1 \cdot x_t + \upsilon_t$$

Assume the following form of υ:

$$\upsilon_t = \rho \cdot \upsilon_{t-1} + \epsilon_t$$

This is called a first-order autoregressive AR(1) scheme. An AR(2) would be written as

$$\upsilon_t = \rho_1 \cdot \upsilon_{t-1} + \rho_2 \cdot \upsilon_{t-2} + \epsilon_t$$

Example with Simulated Data

Consider the model:

$$y_t = 1 + 0.8 \cdot x_t + \upsilon_t$$

Assume the following form of υ:

$$\upsilon_t = 0.7 \cdot \upsilon_{t-1} + \epsilon_t$$

Procedure

- Simulate the above model 100 times
- Compare variance of coefficients under different two different methods: (1) OLS and (2) Cochrane-Orcutt

Durbin Watson d Test

Test statistic:

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

Assumptions

- No intercept
- AR(1) process, i.e., $\upsilon_t = \rho \cdot \upsilon_{t-1} + \epsilon_t$
- No lagged independent variables

Original papers derive lower (d_L) and upper (d_U) bounds, i.e., critical values, that depend on N and k only.

- $d \approx 2 \cdot (1 - \rho)$ and since $-1 \leq \rho \leq 1$, we have $0 \leq d \leq 4$.

Rule of thumb indicates that $d = 2$ signals no problems.

Breusch-Godfrey Test

Consider the following model $y_t = \beta_0 + \beta_1 x_t + \upsilon_t$ with the following error term structure:

$$\upsilon_t = \rho_1 \upsilon_{t-1} + \rho_2 \upsilon_{t-2} + \cdots + \rho_p \upsilon_{t-p} + \epsilon_t$$

The null hypothesis for the test is expressed as follows:

$$H_0 : \rho_1 = \rho_2 = \cdots = \rho_p = 0$$

When the following regression is executed:

$$\hat{\upsilon}_t = \alpha_0 + \alpha_1 x_t + \hat{\rho}_1 \hat{\upsilon}_{t-1} + \hat{\rho}_2 \hat{\upsilon}_{t-2} + \cdots + \hat{\rho}_p \hat{\upsilon}_{t-p} + \epsilon_t$$

Then

$$(n-p)R^2 \sim \chi^2_p$$

Wages and Productivity in the United States 1959-1998

Consider the data in *business.csv* and do the following:

- Plot the data in a scatter plot
- Run the regression in level form as well as log format
- Plot the diagnostic plots.
- Run the Durbin-Watson test and the Breusch-Godfrey test. What do you conclude?
- Run the regression by (1) including a trend variable and (2) a squared term but no trend.