Chapter 13

Violating Assumptions

This chapter introduces the detection and correction of problems due to the violation of the key assumptions necessary for the OLS model to work. Specifically, the issues of heteroscedasticity, multicollinearity, and autocorrelation are covered. The following R packages are needed for this chapter: car, lmtest, orcutt, prais, and sandwich. There are also slides and three videos associated with the chapter.

13.1 Heteroscedasticity

A key assumption of the OLS model is homoscedasticity error terms. That is, the error variance is constant:

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

With heteroscedasticity, the variance of the error term is not constant:

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

For a bivariate regression model with heteroscedastic data, it can be shown that

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

This is different from the variance of the coefficient estimate under homoscedasticity:

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

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. Homoscedasticity is needed to justify the t-test, F-test,

and confidence intervals. The F-statistic does no longer have an F-distribution. In short, hypothesis tests on the β-coefficients are no longer valid.

If σ_i^2 was known, the use of a Generalized Least Squares (GLS) model would be appropriate:

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

Dividing both sides by the known variance:

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

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

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

Under GLS model:

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

That is, GLS minimizes the weighted sum of the residual squares. Since in reality, the variance of σ^2 is not known, other techniques have to be employed to obtain so-called heteroscedasticity-consistent (HC) standard errors. But first, two tests are introduced to detect heteroscedasticity.

13.1.1 Detecting Heteroscedasticity

Two test are presented to detect heteroscedasticity:

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

The steps necessary for the **Goldfeld-Quandt Test** are as follows:

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, i.e., with the "lower" and "upper" part.
4. Compute
$$\lambda = \frac{RSS_2/df}{RSS_1/df}$$
5. λ follows an F-distribution.

The Goldfeld-Quandt Test can be illustrated with `gqtestdata`. In a first step, the data is separated into two groups with $C = 6$. In a second step, both groups are used to run a regression. And lastly, λ is calculated.

```
gqdata1   = gqdata[1:22,]
gqdata2   = gqdata[29:50,]
bhat1     = lm(price~sqft,data=gqdata1)
bhat2     = lm(price~sqft,data=gqdata2)
lambda    = sum(bhat2$residuals^2)/sum(bhat1$residuals^2)
```

Of course, there is also a function in R called gqtest which simplifies the procedure.

```
library(lmtest)
bhat      = lm(price~sqft,data=gqdata)
gqtest(bhat,fraction = 6)
```

```
##
##  Goldfeld-Quandt test
##
## data:  bhat
## GQ = 2.6467, df1 = 20, df2 = 20, p-value = 0.01745
## alternative hypothesis: variance increases from segment 1 to 2
```

In any case, the hypothesis of homoscedasticity is rejected for `gqtestdata`.

The **Breusch-Pagan-Godfrey Test** is an alternative and does not rely on choosing C as the number of central observations to be dropped. The steps include the following:

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
$$p_i = \alpha_0 + \alpha_1 \cdot x_{i1} + \alpha_2 \cdot x_{i2} + ...$$
5. Obtain the explained sum of squares (ESS) and define $\Theta = 0.5 \cdot ESS$. Then $\Theta \sim \chi^2_{m-1}$.

The much simpler procedure is to use the function `bptest()` in R.

```
library(lmtest)
bhat      = lm(price~sqft,data=gqdata)
bptest(bhat)
```

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

13.1.2 Correcting Heteroscedasticity

To correct for heteroscedasticity, robust standard errors must be obtained.

```
bhat      = lm(price~sqft,data=gqdata)
summary(bhat)
vcov      = vcovHC(bhat)
coeftest(bhat,vcov.=vcov)
```

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.

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

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

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

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

13.2 Multicollinearity

Multicollinearity describes the situation in which two or more independent variables are linearly related. Under perfect multicollinearity:

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

where λ_i are constants that are not all zero simultaneously. For example, consider $x_1 = \{8, 12, 15, 17\}$, $x_2 = \{24, 36, 45, 51\}$, and $x_3 = \{2, 3, 3.75, 4.25\}$. In this case, $\lambda_1 = 1$, $\lambda_2 = -1/5$, and $\lambda_3 = 2$. Note, multicollinearity refers to linear relationships! Including a squared or cubed term is not an issue of multicollinearity. It can be shown that the variance of the estimator increases in the presence of multicollinearity. There are various indications that the data suffers from multicollinearity:

- 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.

13.2.1 Variance Inflated Factors (VIF)

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

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

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. The 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.

13.2.2 Examples

To illustrate the concept of multicollinearity, the data set from `nfl` is used (Berri et al. (2011)). The first model includes total salary as the dependent variable and the following independent variables: prior season passing yards, pass attempts, experience (squared) in the league, draft round pick, veteran (more than 3 years in the league), pro bowl appearance, and facial symmetry.

```
bhat = lm(log(total)~yards+att+exp+exp2+draft1+draft2+veteran+
               changeteam+pbowlever+symm,data=nfl)
summary(bhat)
```

After estimating the results, the function `vif()` from the package `car` is used:

```
library(car)
round(vif(bhat),1)
```

```
##      yards        att        exp       exp2     draft1     draft2    veteran changeteam  pbowl
##       32.5       30.9       39.9       26.7        1.6        1.2        5.3        1.2
```

The results indicate multicollinearity for *yards*, *att*, and experience. Passings yards and attempts may be correlated and thus, one of them (*att*) is dropped.

```
bhat = lm(log(total)~yards+exp+exp2+draft1+draft2+veteran+
               changeteam+pbowlever+symm,data=nfl)
summary(bhat)
```

```
round(vif(bhat),1)
```

```
##      yards        exp       exp2     draft1     draft2    veteran changeteam  pbowle
##        1.5       39.3       26.2        1.6        1.2        5.3        1.1
```

This improves the estimation but experience (and its squared term) are still problematic. The last estimation removes experience and the VIF terms are now in the acceptable range.

```
bhat = lm(log(total)~yards+draft1+draft2+veteran+changeteam+pbowlever+symm,data=nfl)
summary(bhat)
```

```
round(vif(bhat),1)
```

```
##      yards     draft1     draft2    veteran changeteam  pbowlever       symm
##        1.4        1.7        1.2        2.0        1.1        1.4        1.0
```

The important part is that the conclusion of the paper has not changed with regard to facial symmetry.

13.3 Other Issues and Problems with Data

More serious problems than heteroscedasticity:

- Functional form mis-specification
- Measurement error
- Missing data: Estimating a 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, e.g., missing education data for people with lower education
- Non-random samples
- Outliers

13.4 Autocorrelation

The correlation of error terms is called autocorrelation. The issue usually arises if there is a time component in the data. Recall the main types of data available for 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)
- Panel data (same observations at different time points)

There is a distinction between serial correlation and autocorrelation:

- Serial correlation: Correlation between two series

- Autocorrelation: Correlation with lagged variables

The OLS estimator is still unbiased but there is no longer minimum variance since $E(\epsilon_i \epsilon_j) \neq 0$. Autocorrelation is unlikely for cross-sectional data except in the case of spatial auto-correlation. One cause of autocorrelation could be inertia in economic variables. For example, variables such as income, production, or employment increase after a recession. But there are a number of other reasons for autocorrelation.

Autocorrelation could be caused by specification bias due to excluded variables or incorrect functional forms. For example, assume that the correct equation is

$$q_{beef} = \beta_0 + \beta_1 \cdot p_{beef} + \beta_2 \cdot p_{income} + \beta_3 \cdot p_{pork} + \epsilon_t$$

The estimated equation is:

$$q_{beef} = \beta_0 + \beta_1 \cdot p_{beef} + \beta_2 \cdot p_{income} + \upsilon_t$$

The error terms in both equations are denoted ϵ_t and υ_t, respectively. This results in a systematic patters of υ_t:

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

Correlation between the error terms can also be caused by specifying an incorrect functional form. Assume that the correct equation is written as follows:

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

But the estimated equation is

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

Serial correlation is caused by lagged terms in the regression equation:

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

The issues of lagged terms will be covered in the part on dynamic regression and time series and this section serves only as an introduction to first-order autoregressive schemes. Consider the model:

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

Assume the following form of υ:

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

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

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

This can be illustrated with simulated data. Consider the following model:

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

and assume the following form of υ:

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

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

13.4.1 Durbin Watson d-Test

The test statistic of the Durbin-Watson test is written as:

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

Assumptions underlying the test are

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

13.4.2 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:

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

Then

$$(n - p) \cdot R^2 \sim \chi_p^2$$