OLS Assumptions_Detection and Solution

OLS assumptions

1. Correct specification of the model.
2. Model has to be linear in parameters.
3. Number of observations is greater than number of parameters.
4. The variance of each independent variables is not zero (not all observations have the same value for the independent variable).
5. Independent variables are deterministic.
6. No perfect multicollinearity
7. Homoskedasticity: Constant variance of the errors across value of independent variables.
8. No correlation between residuals. For two given and different values of the X the errors are not correlated.
9. The covariance between X and the error is zero.
10. The mean of the errors for a given X is zero.
11. The errors are normally distributed.

OLS properties

If the OLS assumptions hold, then the OLS-estimator will be the best, linear, unbiased one.

- **unbiased**, i.e. the estimates converge towards their true value with an increasing number of observations.
- **consistent**, i.e. the variance (standard error) decreases with an increasing number of observations.
- **efficient**, i.e. it has the lowest variance (standard error) among all estimators available.

Assumption violation: Detection & Solutions

Multicollinearity

Imagine we run a regression including revenue and total assets as explanatory variables in our model. The OLS will not be able to distinguish between the effects of the two correctly. We will then have high multicollinearity, which results to

less precise parameter estimates => increased standard errors of the estimates => lower t-values

Detection:

- High R-squared but few significant parameters
- High pairwise correlations between independent variables $|corr| > 0.8$
- High Variance inflation factor: $VIF >= 5$

Solutions:

- Usage of more and/or better data
- Exclusion of one or more variables (particularly in the case of perfect multicollinearity) but risk of specification errors
- Other methods such as factor analysis

Example

Let the data "MarketPower.xlsx"

```
##        id              year      FixedassetsthEUR      StockthEUR
##  Min.   :  2.0   Min.   :2009   Min.   :     0.01   Min.   :
0.64
##  1st Qu.: 99.0   1st Qu.:2011   1st Qu.:   369.81   1st Qu.:
159.22
##  Median :211.5   Median :2013   Median :  1094.67   Median :
570.29
##  Mean   :266.9   Mean   :2013   Mean   : 11037.69   Mean   :
4575.00
##  3rd Qu.:360.0   3rd Qu.:2015   3rd Qu.:  4157.77   3rd Qu.:
2445.29
##  Max.   :854.0   Max.   :2017   Max.   :307300.03   Max.
:115594.53
##
##  TotalassetsthEUR   ShareholdersfundsthEUR  RevenuethEUR
##  Min.   :    57.5   Min.   :     0.77      Min.   :     87.8
##  1st Qu.:  1347.7   1st Qu.:   380.46      1st Qu.:   2374.9
##  Median :  4023.6   Median :  1226.03      Median :   6471.1
##  Mean   : 25972.1   Mean   :  9742.84      Mean   :  60579.3
##  3rd Qu.: 15302.3   3rd Qu.:  5288.93      3rd Qu.:  38280.1
##  Max.   :637866.8   Max.   :221237.52      Max.   :1904158.2
##
##    SalesthEUR        PLbeforetaxthEUR    TaxationthEUR
NetincomethEUR
##  Min.   :     86.3   Min.   :-40711.22   Min.   :-3245.7   Min.   :-
74575.49
##  1st Qu.:   2213.1   1st Qu.:    13.79   1st Qu.:    0.0   1st Qu.:
9.99
##  Median :   6035.1   Median :    94.86   Median :    2.7   Median :
```

```
82.45
##  Mean   :  55727.4   Mean   :  1376.51   Mean   :  387.9   Mean   :
733.29
##  3rd Qu.:  35189.9   3rd Qu.:   426.05   3rd Qu.:   92.2   3rd Qu.:
315.87
##  Max.   :1678914.9   Max.   :138773.03   Max.   :36945.3   Max.   :
87741.72
##
##  CostsofemployeesthEUR InterestpaidthEUR   DebtorsthEUR
Numberofemployees
##  Min.   :    12.16     Min.   :  -0.976   Min.   :     0.11   Min.
:   1.0
##  1st Qu.:   280.30     1st Qu.:   4.236   1st Qu.:   310.76   1st
Qu.:   6.0
##  Median :   975.02     Median :  15.752   Median :   835.96   Median
:  21.0
##  Mean   :  5667.68     Mean   : 131.763   Mean   :  5488.00   Mean
: 114.3
##  3rd Qu.:  3407.14     3rd Qu.:  66.089   3rd Qu.:  3656.96   3rd
Qu.:  81.0
##  Max.   :195330.71     Max.   :5601.592   Max.   :186865.22   Max.
:4615.0
##
##  ExportrevenuethEUR MaterialcoststhEUR       age               FC
##  Min.   :    -3.8   Min.   :      2.7   Min.   :  1.00   Min.   :
14.8
##  1st Qu.:     0.0   1st Qu.:   1447.4   1st Qu.: 14.00   1st Qu.:
338.2
##  Median :     0.0   Median :   3797.9   Median : 25.00   Median :
1175.1
##  Mean   : 10783.0   Mean   :  40980.5   Mean   : 33.07   Mean   :
12422.9
##  3rd Qu.:  1246.5   3rd Qu.:  23534.2   3rd Qu.: 50.00   3rd Qu.:
5747.4
##  Max.   :366990.4   Max.   :1513689.7   Max.   :117.00   Max.
:695486.8
##  NA's   :1
##       FCR               ROA                eqshare             RevGR
##  Min.   : 0.2189   Min.   :-0.446323   Min.   : 0.01582   Min.   :-
77.983
##  1st Qu.:12.6032   1st Qu.: 0.005531   1st Qu.:27.58443   1st Qu.: -
3.026
##  Median :18.1771   Median : 0.038869   Median :40.34883   Median :
4.881
##  Mean   :19.9883   Mean   : 0.048962   Mean   :40.14749   Mean   :
6.194
##  3rd Qu.:25.4689   3rd Qu.: 0.080204   3rd Qu.:54.47316   3rd Qu.:
12.217
##  Max.   :82.0954   Max.   : 0.522538   Max.   :91.85318   Max.
:409.126
```

```
##
##      Markup
##  Min.   : 1.026
##  1st Qu.: 2.646
##  Median : 3.068
##  Mean   : 3.280
##  3rd Qu.: 3.526
##  Max.   :27.018
##
```

Lets us run an OLS regression

```
OLSbase = lm(Markup~RevGR+eqshare+FCR+age+TotalassetsthEUR, dat=mpow)
summary(OLSbase)
```

```
##
## Call:
## lm(formula = Markup ~ RevGR + eqshare + FCR + age +
TotalassetsthEUR,
##     data = mpow)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4257 -0.6224 -0.1528  0.3274 22.6193
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)       2.505e+00  1.144e-01  21.889  < 2e-16 ***
## RevGR             1.034e-03  1.548e-03   0.668   0.5043
## eqshare           4.802e-03  1.865e-03   2.574   0.0102 *
## FCR               2.621e-02  3.465e-03   7.565 7.03e-14 ***
## age               1.614e-03  1.506e-03   1.072   0.2840
## TotalassetsthEUR -5.210e-08  5.413e-07  -0.096   0.9233
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.346 on 1382 degrees of freedom
## Multiple R-squared:  0.05115,    Adjusted R-squared:  0.04772
## F-statistic:  14.9 on 5 and 1382 DF,  p-value: 2.881e-14
```

The phenomenon of "high-R-squared, few significant parameters rule", is not observed in our model as there are indeed only one or two significant parameters but the the $R^2 = 0.05 << 1$. Thus, we would conclude that there is no multicollinearity issue concerning the specific linear model?

As a second detection method let us calculate the pairwise correlation coefficients

```
indepvar = cbind(mpow[,5], mpow[,18], mpow[,20], mpow[,22:23])
cor(indepvar)
```

```
##                   TotalassetsthEUR          age         FCR
eqshare
## TotalassetsthEUR       1.00000000 -0.011653111  0.14379833 -
0.05080195
## age                   -0.01165311  1.000000000 -0.22079044
0.10496520
## FCR                    0.14379833 -0.220790438  1.00000000
0.14725112
## eqshare               -0.05080195  0.104965203  0.14725112
1.00000000
## RevGR                 -0.06390431 -0.001616312 -0.03174999 -
0.05254241
##                            RevGR
## TotalassetsthEUR -0.063904313
## age              -0.001616312
## FCR              -0.031749994
## eqshare          -0.052542407
## RevGR             1.000000000
```

Again we observe that no pairwise correlation coefficient has absolute value is greater than 0.8. According to this criterion there is no multicollinearity in our model

Finally, let us examine according to VIF detection method

```
library(car)

## Loading required package: carData

vif(OLSbase)

##            RevGR          eqshare              FCR              age
##         1.007462         1.052529         1.112074         1.074119
## TotalassetsthEUR
##         1.031896
```

As no VIF is equal to or higher than 5, we conclude again that there is no multicollinearity issue with the selected independent variables of this linear model

Heteroskedasticity

The variance of the errors varies across observations leading to distorted standard errors => t-tests become inaccurate (usually indicate higher significance). (Suppose you regress a common-product consumption on income => for large incomes, the errors will be larger, i.e. their variance increases)

Detection

- Goldfeld/Quandt test: equality of error variance cross subsamples tested (F-test)

- Method of Glesjer: Regress absolute values of residuals on independent variables.
- Breusch-Pagan-test: Regress squared residuals on independent variables. Conduct a Chi-squared test with k-1 degrees of freedom (k = number of explanatory variable) where: $BP = nR^2$
- White test: Regress squared residuals on independent variables and their products. Conduct a Chi-squared test with k-1 degrees of freedom (k = number of explanatory variables) where: $W = nR^2$

Solutions

- Solve specification errors (omitted variables)
- Use other estimators (such as weighted least squares)
- Transform the error term (Generalized least squares)
- Use heteroskedasticity robust standard errors (most popular)

Example

Let us consider again the data "MarketPower.xlsx" and the same linear model "OLSbase".

We start with the Breusch-Pagan test for Heteroskedasticity

```
mpow$sqres = OLSbase$residuals^2
BPreg = lm(sqres~RevGR+eqshare+FCR+age+TotalassetsthEUR, dat=mpow)
BP = nrow(mpow)*summary(BPreg)$r.squared
BP

## [1] 57.90686

BPpv = pchisq(BP, length(BPreg$coefficients)-1,lower.tail=FALSE)
BPpv

## [1] 3.287753e-11
```

As the $p - value << 1$ we conclude that under the H0 : 'there is no Heteroskedasticity' a value $BP = 57.907$ it is extremely unlikely to be measured. Thus we reject the null hypothesis in favor of heteroskedasticity.

Alternatively we can conduct a White-test

```
WTreg =
lm(sqres~RevGR+eqshare+FCR+age+TotalassetsthEUR+I(RevGR*RevGR)+RevGR*eq
share+RevGR*FCR

+RevGR*age+RevGR*TotalassetsthEUR+I(eqshare*eqshare)+eqshare*FCR+eqshar
e*age
```

```
+eqshare*TotalassetsthEUR+I(FCR*FCR)+FCR*age+FCR*TotalassetsthEUR

+I(age*age)+age*TotalassetsthEUR+I(TotalassetsthEUR*TotalassetsthEUR),
dat=mpow)
summary(WTreg)

##
## Call:
## lm(formula = sqres ~ RevGR + eqshare + FCR + age + TotalassetsthEUR
+
##     I(RevGR * RevGR) + RevGR * eqshare + RevGR * FCR + RevGR *
##     age + RevGR * TotalassetsthEUR + I(eqshare * eqshare) + eqshare
*
##     FCR + eqshare * age + eqshare * TotalassetsthEUR + I(FCR *
##     FCR) + FCR * age + FCR * TotalassetsthEUR + I(age * age) +
##     age * TotalassetsthEUR + I(TotalassetsthEUR * TotalassetsthEUR),
##     data = mpow)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -50.18  -1.43  -0.56   0.46 471.80
##
## Coefficients:
##                                          Estimate Std. Error t value
Pr(>|t|)
## (Intercept)                             6.574e+00  2.652e+00   2.479
0.01330
## RevGR                                  -1.777e-01  5.933e-02  -2.994
0.00280
## eqshare                                -5.087e-02  7.690e-02  -0.661
0.50841
## FCR                                    -4.561e-01  1.497e-01  -3.047
0.00236
## age                                    -7.879e-02  9.125e-02  -0.863
0.38805
## TotalassetsthEUR                       -1.874e-05  2.741e-05  -0.684
0.49427
## I(RevGR * RevGR)                       -2.793e-04  1.371e-04  -2.037
0.04183
## I(eqshare * eqshare)                   -5.639e-04  9.248e-04  -0.610
0.54213
## I(FCR * FCR)                            8.316e-03  1.858e-03   4.475
8.27e-06
## I(age * age)                           -5.552e-05  5.742e-04  -0.097
0.92298
## I(TotalassetsthEUR * TotalassetsthEUR)  1.327e-11  4.382e-11   0.303
0.76199
## RevGR:eqshare                          -1.188e-03  1.089e-03  -1.091
0.27562
## RevGR:FCR                               1.070e-02  1.902e-03   5.628
```

```
2.21e-08
## RevGR:age                                 2.553e-03  8.807e-04   2.899
0.00380
## RevGR:TotalassetsthEUR                   -6.665e-07  4.723e-07  -1.411
0.15836
## eqshare:FCR                               3.748e-03  1.916e-03   1.956
0.05070
## eqshare:age                               1.302e-03  1.058e-03   1.231
0.21846
## eqshare:TotalassetsthEUR                  3.562e-07  3.840e-07   0.928
0.35371
## FCR:age                                   1.334e-03  2.122e-03   0.629
0.52958
## FCR:TotalassetsthEUR                     -3.259e-07  5.485e-07  -0.594
0.55245
## age:TotalassetsthEUR                      3.031e-07  3.591e-07   0.844
0.39869
##
## (Intercept)                              *
## RevGR                                    **
## eqshare
## FCR                                      **
## age
## TotalassetsthEUR
## I(RevGR * RevGR)                         *
## I(eqshare * eqshare)
## I(FCR * FCR)                             ***
## I(age * age)
## I(TotalassetsthEUR * TotalassetsthEUR)
## RevGR:eqshare
## RevGR:FCR                                ***
## RevGR:age                                **
## RevGR:TotalassetsthEUR
## eqshare:FCR                              .
## eqshare:age
## eqshare:TotalassetsthEUR
## FCR:age
## FCR:TotalassetsthEUR
## age:TotalassetsthEUR
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.91 on 1367 degrees of freedom
## Multiple R-squared:  0.09113,    Adjusted R-squared:  0.07784
## F-statistic: 6.854 on 20 and 1367 DF,  p-value: < 2.2e-16
```

```
WT =  nrow(mpow)*summary(WTreg)$r.squared
WT
```

```
## [1] 126.4928
```

```
WTpv = pchisq(WT, length(WTreg$coefficients)-1,lower.tail=FALSE)
WTpv

## [1] 1.767624e-17
```

With that test we also have that $p - value << 1$ .We conclude that under the H0 : 'there is no Heteroskedasticity' a value $W = 126.49$ it is extremely unlikely to be measured. Thus we reject the null hypothesis in favor of heteroskedasticity.

A way to resolve the issue of heteroskedastic errors we use robust standard errors, which by construction are wider(and thus more realistic) than the ones from the simple model.

Autocorrelation

The residuals are correlated over time. Hence, autocorrelation is mostly relevant when working with times series or panel data. Autocorrelation typically leads to downward biased standard errors.

Detection

- Look at the scatterplot of the residuals from t against those from t-1.
- Durbin-Watson test: The test statistic is calculated as: d = $\sum([u(t) - u(t-1)]^2)/\sum([u(t)]^2)$ where

  d lies between 0 and 4 and we look up lower and upper d from the table.

Solutions

- Mostly changing the model specification.

Example

We consider the time-series data "solardat.csv"

```
library(readr)
soldat <- read.csv("D:/data/Empirical Research/solardat.csv",sep=";")
head(soldat)

##         Date  CV_daily PV_daily_MWh
## 1 06.01.2015 0.2706780     32887.25
## 2 07.01.2015 0.4032656     17114.75
## 3 08.01.2015 0.3745514      8598.25
## 4 09.01.2015 0.4617914      6823.75
## 5 10.01.2015 0.3316949     20475.00
## 6 11.01.2015 4.5214577     19811.25
```

Let us run a linear model

```
OLSsol = lm(CV_daily~PV_daily_MWh, dat=soldat)
summary(OLSsol)
```

```
##
## Call:
## lm(formula = CV_daily ~ PV_daily_MWh, data = soldat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.3277 -0.1751 -0.1205 -0.0423 22.0132
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   5.083e-01  4.833e-02   10.52   <2e-16 ***
## PV_daily_MWh -6.067e-07  4.019e-07   -1.51    0.131
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9392 on 1301 degrees of freedom
## Multiple R-squared:  0.001749,   Adjusted R-squared:  0.0009813
## F-statistic: 2.279 on 1 and 1301 DF,  p-value: 0.1314
```

Obtain residuals

```
library(Hmisc)
soldat$solres= OLSsol$residuals
soldat$lagsolres = Lag(soldat$solres)
```

Generate the test statistic

```
soldat$difres = (soldat$solres-soldat$lagsolres)^2
soldat$sqres = soldat$solres^2
dtest = sum(soldat$difres, na.rm=TRUE)/sum(soldat$sqres, na.rm=TRUE)
dtest

## [1] 1.851654
```

We use predefined function in R to perform the Durbin-Watson test

```
library(car)
durbinWatsonTest(OLSsol)

##  lag Autocorrelation D-W Statistic p-value
##    1      0.07414144      1.851654   0.046
##  Alternative hypothesis: rho != 0
```

A possible way to resolve the issue of autocorrelation is the following

```
soldat$lagcv = Lag(soldat$CV_daily)
autoco = lm(CV_daily~PV_daily_MWh+lagcv,dat=soldat)
summary(autoco)

##
## Call:
## lm(formula = CV_daily ~ PV_daily_MWh + lagcv, data = soldat)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.6544 -0.1671 -0.1179 -0.0435 22.0181
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   4.692e-01  5.039e-02   9.310  < 2e-16 ***
## PV_daily_MWh -5.524e-07  4.017e-07  -1.375  0.16939
## lagcv         7.562e-02  2.768e-02   2.732  0.00638 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9372 on 1299 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.007465,   Adjusted R-squared:  0.005937
## F-statistic: 4.885 on 2 and 1299 DF,  p-value: 0.007698
```

```
durbinWatsonTest(autoco)
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1    -0.003691424      2.007359   0.676
##  Alternative hypothesis: rho != 0
```

As we can observe the autocorrelation is now not statistical significant.