

Violations of Linear Regression Classic Assumptions and Their Remedies

Babak Rezaee Daryakenari
The Institute of Political Science
Leiden University
s.rezaeedaryakenari@fsw.leidenuniv.nl
For more research methods handouts:
<https://babakrezaee.github.io>

Spring 2020

Today's plan

- ▶ Reviewing the Gauss-Markov Assumptions (Classic OLS assumptions)
- ▶ OLS diagnostics
- ▶ Omitted variable problem
- ▶ Serial correlation vs. auto correlation
- ▶ Clustering and robust standard errors
- ▶ Does Multicollinearity violate the classic assumptions?

The Gauss-Markov Assumptions

1. **Linearity assumption:** $y = \beta_0 + \beta_1 x + \epsilon$
2. **X is a full rank matrix.**
3. $E(\epsilon|X) = 0$
4. $E(\epsilon\epsilon'|X) = \sigma^2 I$
5. **X and ϵ are orthogonal $X \perp \epsilon$**
6. $\epsilon|X \sim N(0, \sigma^2 I)$

Th assumption 6 is not actually required for the Gauss-Markov Theorem. However, we often assume it to make hypothesis testing easier.

Therefore, if assumptions 1-5 are violated, they mess up OLS *BLUEness*:

$$\hat{\beta} = (X^T X)^{-1} X^T y = \beta + (X^T X)^{-1} X^T \epsilon \quad (1)$$

OLS diagnostics

- ▶ How can we test if our estimated model does not violate the OLS assumptions?
- ▶ \mathcal{R} offer nice visualization tools to check the possibility of OLS assumptions violations.

Let's get back to our Acemoglu et al. (2001) example:

```
## MODEL INFO:
## Observations: 57
## Dependent Variable: logpgp95
## Type: OLS linear regression
##
## MODEL FIT:
## F(1,55) = 72.32, p = 0.00
## R2 = 0.57
## Adj. R2 = 0.56
##
## Standard errors: OLS
## -----
##              Est.    S.E.    t val.    p
## -----
## (Intercept)    4.89    0.38    12.79    0.00
## avexpr          0.49    0.06     8.50    0.00
## -----
```

Fitted values and residuals

- ▶ We need fitted values and residuals values of our estimated model to diagnose possible issues: $e = y - \hat{y}$.
- ▶ There are different methods to do this; here I use broom package.

```
library(broom)
```

```
OLS1_diag_metrics <- augment(OLS1_results)
```

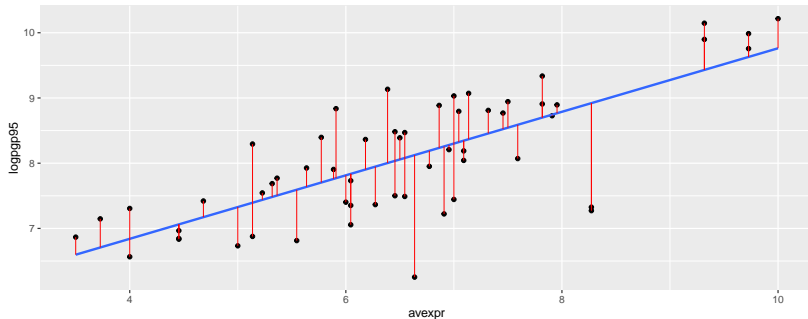
Diagnostcs metrics

```
head(OLS1_diag_metrics)
```

```
## # A tibble: 6 x 10
##   .rownames logpgp95 avexpr .fitted .se.fit .resid   .hat .sigma .cooksd
##   <chr>      <dbl>  <dbl>   <dbl>   <dbl> <dbl>   <dbl> <dbl>   <dbl>
## 1 372        6.87   3.5     6.60   0.193  0.270  0.0882  0.653  0.00916
## 2 219        7.15   3.73    6.71   0.181  0.439  0.0780  0.652  0.0210
## 3 271        6.57   4       6.84   0.168 -0.276  0.0667  0.653  0.00691
## 4 323        7.31   4       6.84   0.168  0.466  0.0667  0.651  0.0197
## 5 141        6.85   4.45    7.06   0.146 -0.216  0.0505  0.654  0.00311
## 6 264        6.84   4.45    7.06   0.146 -0.227  0.0505  0.654  0.00343
## # ... with 1 more variable: .std.resid <dbl>
```

Marking errors in your scatter-fit plot

```
library(ggplot2)
ggplot(OLS1_diag_metrics, aes(avexpr, logpgp95)) +
  geom_point() +
  stat_smooth(method = lm, se = FALSE) +
  geom_segment(aes(xend = avexpr, yend = .fitted), color = "red", size = 0.3)
```

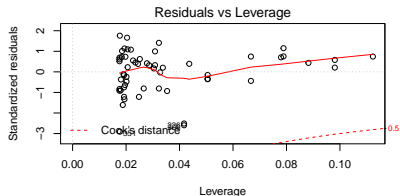
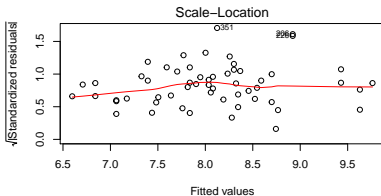
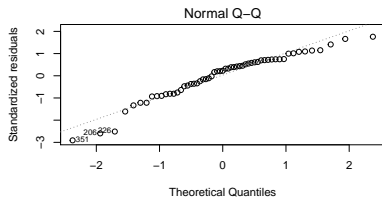
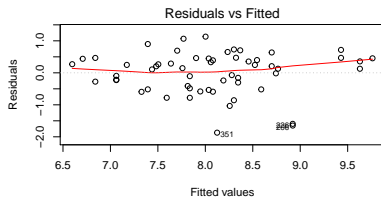


- If you switch to a multivariate regression model, then the above plot would be a little bit different because OLS should fit the data across more than two variables!

Regression diagnostics plots

```
par(mfrow = c(2, 2))
```

```
plot(OLS1_results)
```

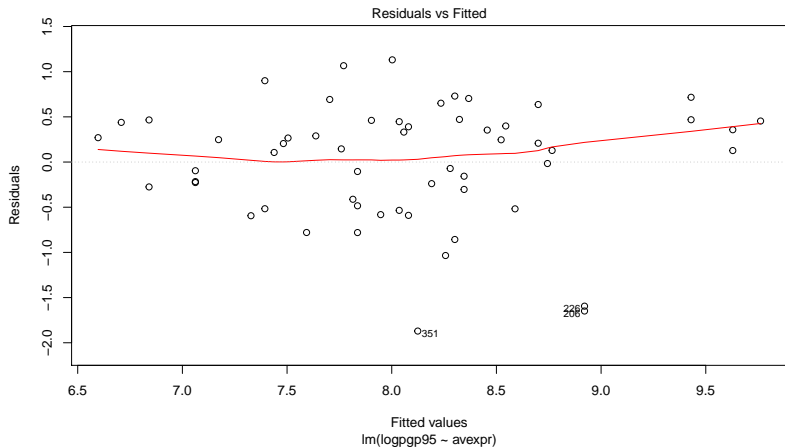


Regression diagnostics plots (2)

- 1. Residuals vs Fitted:** Checking the linear relationship assumptions. The trend line should be close to a horizontal line to show a linear relationship.
- 2. Normal Q-Q:** Checking if the residuals are normally distributed. The trend of residuals points around the 45° (dashed) line is a good sign.
- 3. Scale-Location (or Spread-Location):** Checking the homogeneity of variance of the residuals (homoscedasticity). Equal spread of the points around a horizontal line is a good indication of homoscedasticity. In our example, there is a small sign of heteroskedasticity.
- 4. Residuals vs Leverage:** Checkig if there are extreme values (outliers) that might influence the regression results.

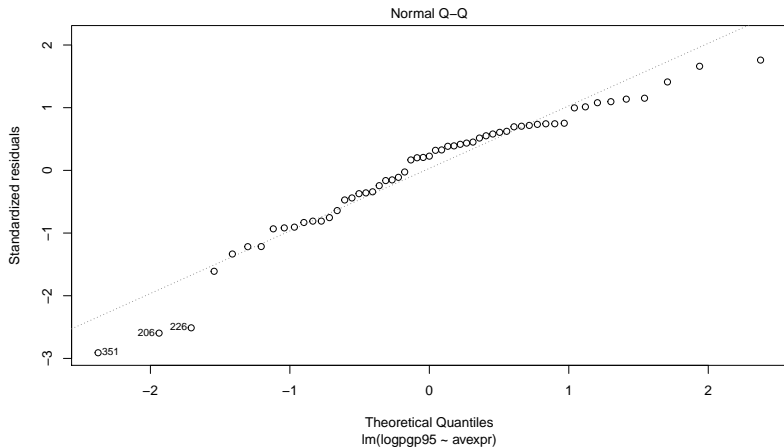
Regression diagnostics individual plots

```
plot(OLS1_results, 1)
```



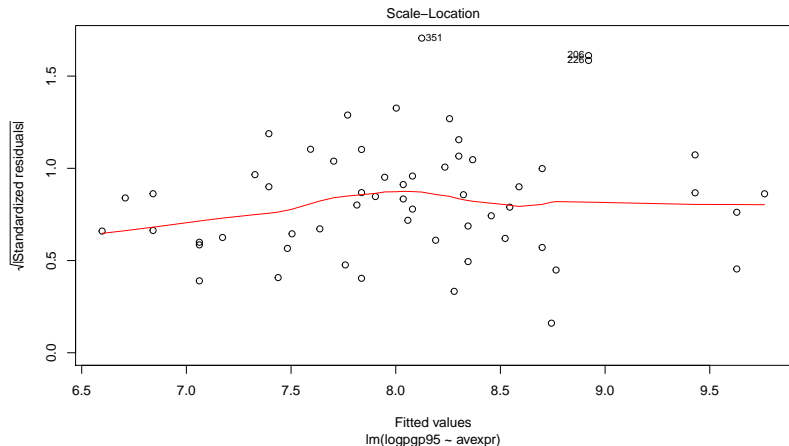
Regression diagnostics individual plots

```
plot(OLS1_results, 2)
```



Regression diagnostics individual plots

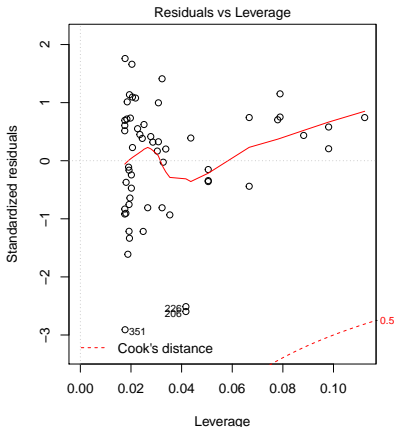
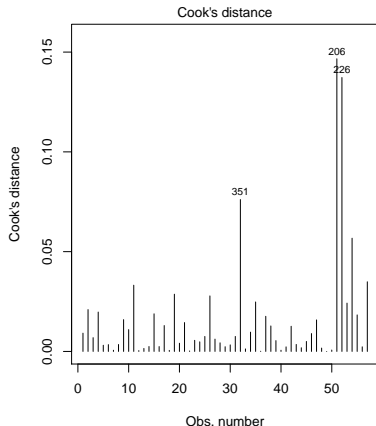
```
plot(OLS1_results, 3)
```



Identifying outliers/influential variables

There variables are called influential because adding or removing to the sample can change the results substantially. To measure the influence of observations, we can use the *Cook's distance*.

```
par(mfrow = c(1, 2))  
plot(OLS1_results, 4)  
plot(OLS1_results, 5)
```

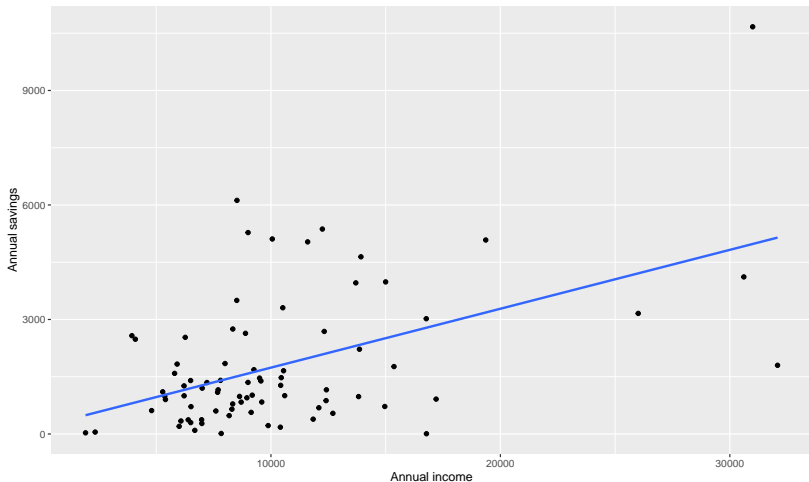


Robust and clustered standard errors

- ▶ *heteroskedasticity* does not lead to a biased coefficient, but it can lead to a biased estimation of the variance-covariance matrix.
- ▶ This can lead to incorrect t-statistics and confidence intervals. That is rejecting null hypothesis and finding support for our theory incorrectly.
- ▶ We learned how to identify this issue using *Scale-Location* (or *Spread-Location*), but how can we solve the problem?
- ▶ Serial correlation and auto correlation are among the main causes of heterogenous.
- ▶ Often, estimating robust and clustered standard errors can help to ease this problem.

A new example: the association between income and saving

$$\text{Saving}_i = \beta_0 + \beta_1 \text{Income}_i \quad (2)$$



Estimation: non-robust

```
model1 <- lm(sav ~ inc, data = saving)
```

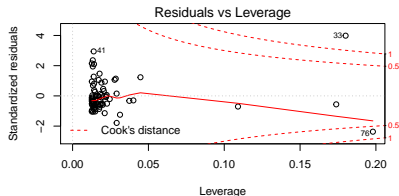
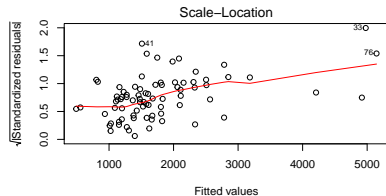
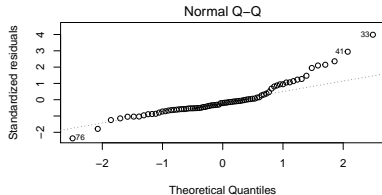
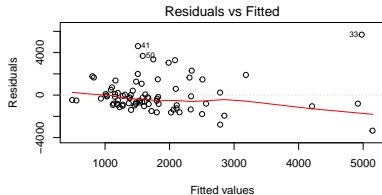
```
summary(model1)
```

```
##
## Call:
## lm(formula = sav ~ inc, data = saving)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3345   -900   -323    457   5690
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 195.1883   366.6416   0.53     0.6
## inc          0.1543     0.0312    4.95 0.0000042 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1580 on 77 degrees of freedom
## Multiple R-squared:  0.242, Adjusted R-squared:  0.232
## F-statistic: 24.5 on 1 and 77 DF, p-value: 0.00000424
```

diagnostics plots

```
par(mfrow = c(2, 2))
```

```
plot(model1)
```



Estimation: robust

- As always there are different ways to cluster standard errors; I here use `lmtest` and `sandwich` package.

```
library(lmtest)
library(sandwich)
```

```
coeftest(model1, vcov = sandwich)
```

```
##
```

```
## t test of coefficients:
```

```
##
```

```
##           Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept) 195.1883   524.2427    0.37  0.7107
```

```
## inc           0.1543    0.0572    2.70  0.0086 **
```

```
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
```

```
coeftest(model1, vcov = vcovHC(model1, type = "HCO"))
```

```
##
```

Cluster standard errors

- ▶ Again there are different ways to cluster standard errors; I here use multiwayvcov package.

```
# FE regression with SE clustered by firm
```

```
library(multiwayvcov)
```

```
model2<-miceadds::lm.cluster(sav ~ inc, data = saving, clus
```

```
model2
```

```
## $lm_res
```

```
##
```

```
## Call:
```

```
## stats::lm(formula = formula, data = data, weights = wgt,
```

```
##
```

```
## Coefficients:
```

```
## (Intercept)          inc
```

```
##      195.188        0.154
```

```
##
```

```
##
```

```
## $vcov
```

Omitted variable problem

$$y = \beta_0 + \beta_1 x + \epsilon \quad (3)$$

- ▶ When we estimate a regression model, how do we know that what variables should be included in the model?
- ▶ And, what is the potential problem caused by omitting an important variable?
- ▶ Let's check our Acemoglu et al. (2001) example again:

Model 1:

$$\log GDP = \beta_0 + \beta_1 \textit{Expropriation} + \epsilon \quad (4)$$

Model 2:

$$\log GDP = \beta_0 + \beta_1 \textit{Expropriation} + \beta_2 \textit{Democracy1900} + \epsilon \quad (5)$$

Results of Model 1

```
library(jtools)
# OLS model
OLS1_results <- lm(logpgp95~avexpr, data=myData)

summ(OLS1_results)
```

```
## MODEL INFO:
## Observations: 57
## Dependent Variable: logpgp95
## Type: OLS linear regression
##
## MODEL FIT:
## F(1,55) = 72.32, p = 0.00
## R2 = 0.57
## Adj. R2 = 0.56
##
## Standard errors: OLS
## -----
##               Est.   S.E.   t val.   p
## -----
## (Intercept)    4.89   0.38   12.79   0.00
## avexpr          0.49   0.06    8.50   0.00
## -----
```

Results of Model 2

```
## MODEL INFO:
## Observations: 57
## Dependent Variable: logpgp95
## Type: OLS linear regression
##
## MODEL FIT:
## F(2,54) = 44.65, p = 0.00
## R2 = 0.62
## Adj. R2 = 0.61
##
## Standard errors: OLS
## -----
##               Est.   S.E.   t val.   p
## -----
## (Intercept)    5.33   0.39   13.58   0.00
## avexpr         0.40   0.06    6.37   0.00
## democ00a       0.09   0.03    2.81   0.01
## -----
```

Confounding effect

- ▶ A situation in which the effect or association between an independent variable and outcome is distorted by the presence of another variable.
- ▶ Assume that we estimated $y = \beta_0 + \beta_1 x$. There is a variable z that is argued to be added to this model. Not including z in the model can cause a problem, only this *omitted variable* causes a confounding bias.
- ▶ An omitted variable problem can lead to a confounding bias if:
 1. x is correlated with the omitted variable z .
 2. The omitted variable has a causal association with the dependent variable Y .

Confounding effect and biased estimation

- ▶ If z is correlated with y , but not included in the model, its effect will show up as part of the residuals. Therefore, residuals will be associated with x (because z is correlated with x).
- ▶ This is a violation of 3rd classic assumption: $E(\epsilon|X) = 0$, leading to a biased estimation.

Solutions to omitted variable problem

- ▶ This is one of the most challenging problems to address
- ▶ Depends on whether the omitted variable is time-variant or invariant
- ▶ Adding the lag of dependent variable
- ▶ Adding fixed-effects and time effects