Workshop5. Homoscedasticity

UNAM – FE Econometrics I

Esp. Humberto Acevedo
Assistant. Emilio Sandoval

Sept 26, 2022

Benford's law

Rethinking Neural Networks With Benford's Law

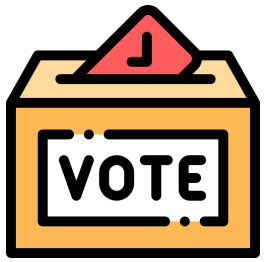
Surya Kant Sahu,[1] **Abhinav Java,** [21] **Arshad Shaikh** [1] **and Yannic Kilcher** [3]

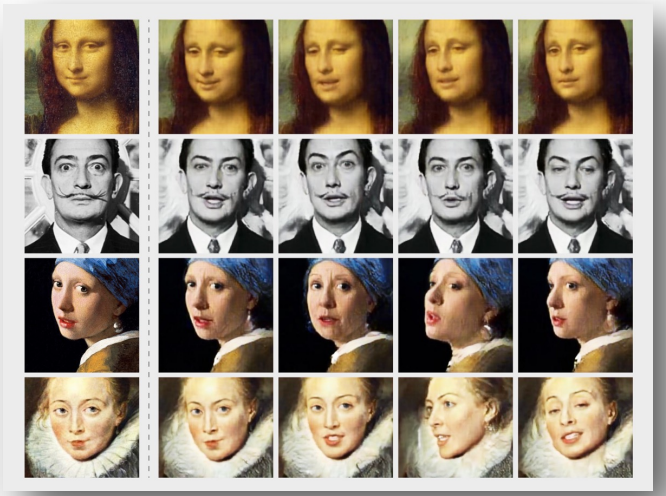
[1] The Learning Machines
[2] Delhi Technological University
[3] ETH Zürich
surya.oju@pm.me, java.abhinav99@gmail.com

Abstract

Benford's Law (BL) or the Significant Digit Law defines the probability distribution of the first digit of numerical values in a data sample. This Law is observed in many naturally occurring datasets. It can be seen as a measure of naturalness of a given distribution and finds its application in areas like anomaly and fraud detection. In this work, we address the following question: Is the distribution of the Neural Network parameters related to the network's generalization capability? To that end, we first define a metric, MLH (Model Enthalpy), that measures the closeness of a set of numbers to Benford's Law and we show empirically that it is a strong predictor of Validation Accuracy. Second, we use MLH as an alternative to Validation Accuracy for Early Stopping, removing the need for a Validation set. We provide experimental evidence that even if the optimal size of the validation set is known beforehand, the peak test accuracy attained is lower than not using a validation set at all. Finally, we investigate the connection of BL to Free Energy Principle and First Law of Thermodynamics, showing that MLH is a component of the internal energy of the learning system and optimization as an analogy to minimizing the total energy to attain equilibrium.

$$P(d) = log_{10}(d+1) - log_{10}(d) = log_{10}\left(\frac{d+1}{d}\right) = log_{10}\left(1+\frac{1}{d}\right)$$



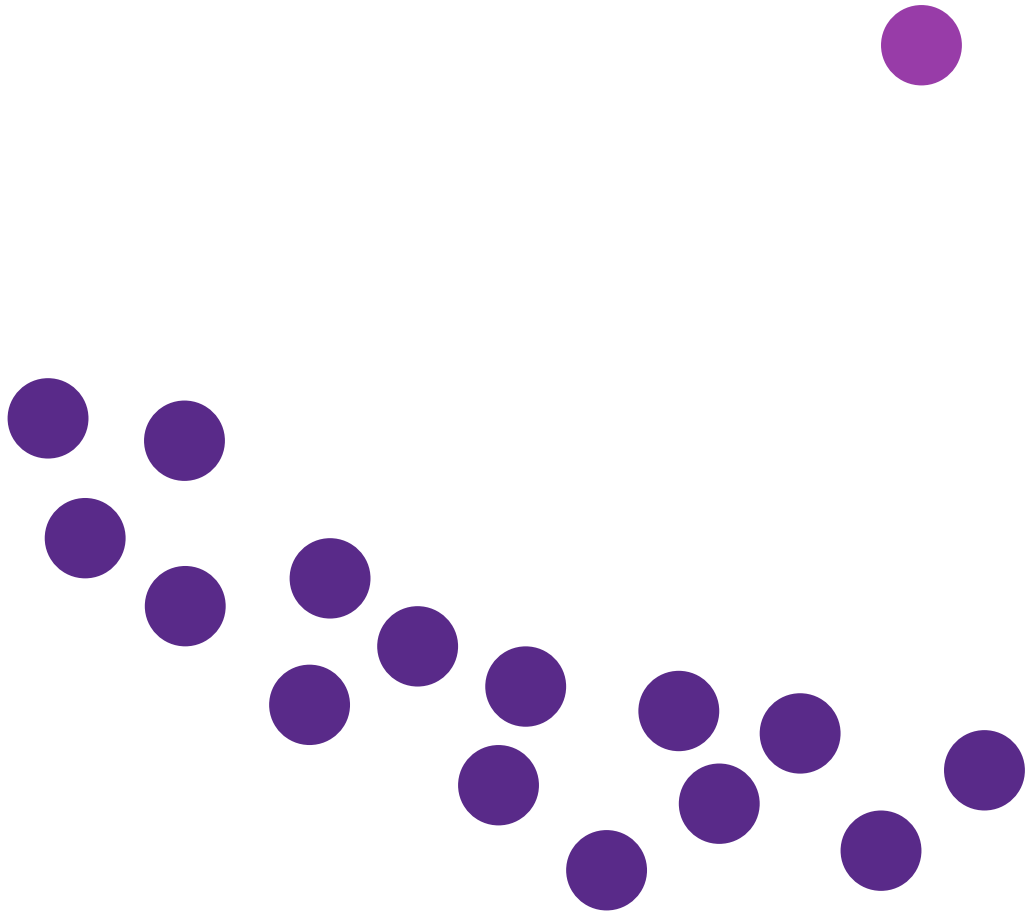


First of all…Why is it important the assumption of normality?

Normality in residuals ensure that OLS estimator is consistent and efficient.

Several tests such as t and F are calculated from the normal distribution assumption.

However, in practice we find some distribution shocks…



If there is an outlier caused by a special situation that is outside from the model, it can provoke a disruption in error distribution.

A datapoint is considered an *outlier* if the value for that point for any variable substantially differs from the rest of the observation's pattern.

Now, the question is…What can we do with these anormal datapoints?

If you have paid enough attention to classes, you must remember that one of the assumptions under linear regression was:

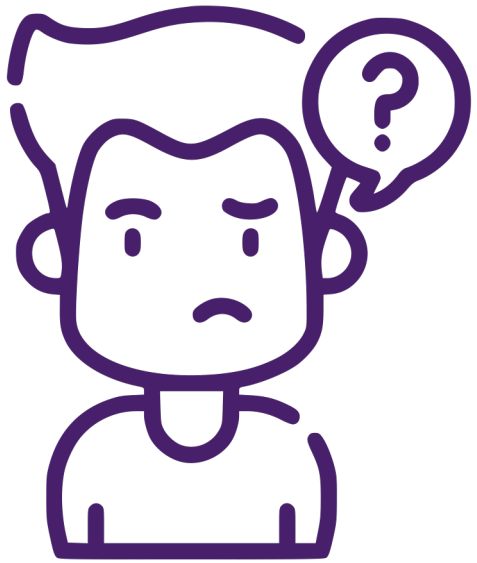
Homogeneity in residuals' variance

If variance in residuals is not constant, then variance in residuals expose heteroscedasticity

It is quite usual in cross-sectional data

We work with members of a population in a specific moment, such as families or industries, which can have different sizes.

We've got two cases:



Pure Heteroscedasticity

We specify the correct model and yet we observe non-constant variance in residuals

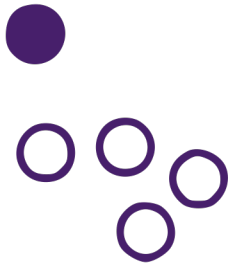
Impure Heteroscedasticity

We incorrectly specify the model, causing the non-constant variace

Causes of **homoscedasticity**

Explanatory variables with an asymetric distribution.

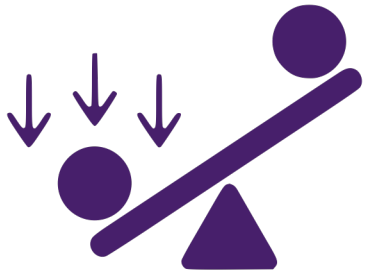
When we omit a variable, it will rely in stochastic term, perhaps causing its own variation.



Outliers necessarily imply an imbalance in disturbance variance.

An outlier can be considered as a sampling element belonging to another distribution (different variance.)

Consequences of **homoscedasticity**



Low heteroscedasticity (Estimator errors are biased)

In presence of heteroscedasticity usual statistics in hypothesis testing under the Gauss-Markov assumptions are not applicable.

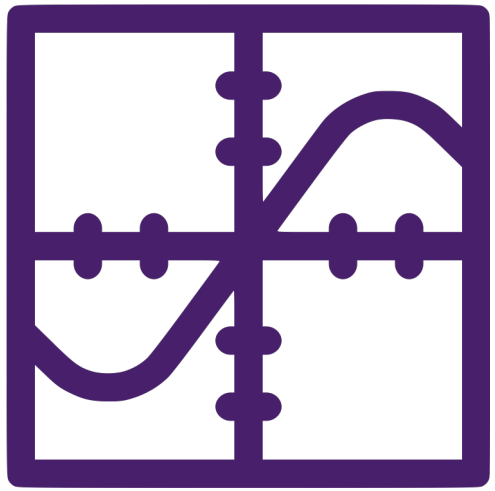
Due to $var(u|X)$ is not longer constant, OLS estimator is not BLUE and not asymptotically efficient.

It is possible to find estimators that are more efficient than the OLS estimator, but it is necessary to kwow the shape of heteroscedasticity

Effects of heteroscedasticity

Defective estimate of parameters

OLS estimator keeps linear, unbiasedness and consistent but is not longer efficient. Homoscedasticity of error term does not play any relevant role in biasedness or consistency



Incorrect computation of variances and inefficient parameters

Variances of OLS estimator despite the fact of not being minimal it cannot be calculated with the expression used in presence of homoscedasticity

We use elemapi2.dta which contains information about basic academic performance in USA

Let's prove that (`api00`) depends on free meals percentage given to students (`meals`), students currently learning English (`ell`) and percentage professors with recent accreditations (`emer`)

We use `rvfplot`

STATA COMMANDS

```
1. regress api00 meals ell emer
2. rvfplot, yline(0)
```

First, let's rename variables

Also, we changed their label

STATA COMMANDS

```
3. rename varivieja varnueva
4.  label variable nombrevariable nombreetiqueta
```

Breusch-Pagan test to identify heteroscedasticity **in residuals**

Null hypothesis assumes variance in errors is constant (homoscedastic)

STATA COMMANDS

```
5. regress api00 meals ell emer

6. estat hettest
```

Effects of heteroscedasticity

Advantages

- Easy to apply
- Does not require to know the functional form of heteroscedasticity

Disadvantages

- Relies on the error normality assumption
- Auxiliary equation is not exempted of specification errors from any regression

White test to identify heteroscedasticity **in residuals**

Null hypothesis assumes variance in errors is constant (homoscedastic)

STATA COMMANDS

```
7. regress api00 meals ell emer

8. estat imtest, white
```

Effects of heteroscedasticity

Advantages

- It is a general test
- Easy to apply

Disadvantages

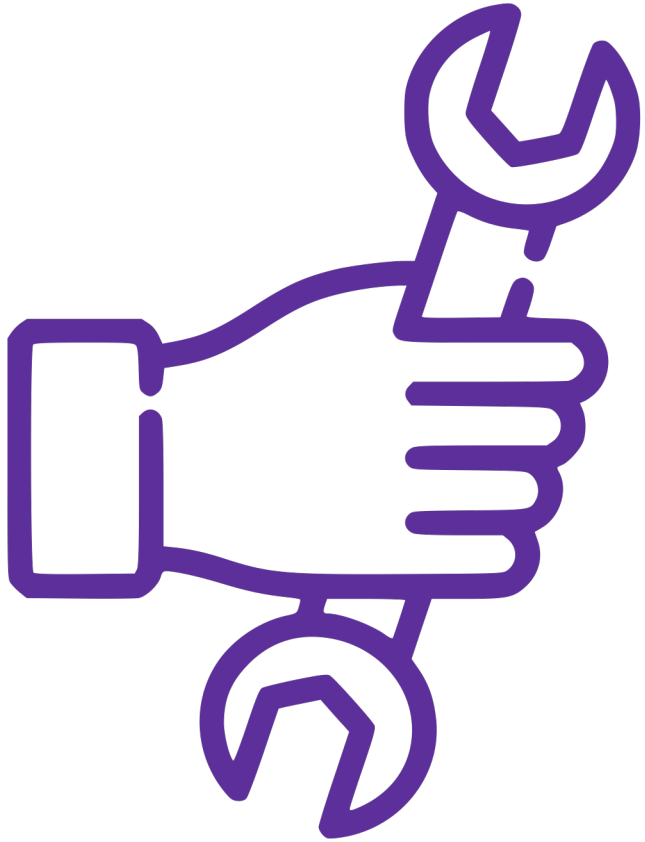
- Auxiliary equation may include to many independent variables
- Auxiliary equation is not exempted of specification errors from any regression

How to fix it?

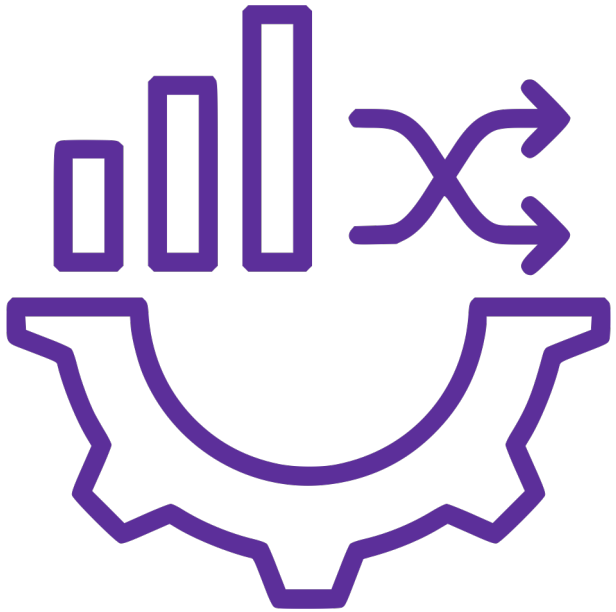
In presence of heteroscedasticity, the OLS estimator is linear, unbiasedness and consistent, but not efficient.

We can add the omitted variable, or to modify the structural form, to model with robust errors, or even to change the estimation method.

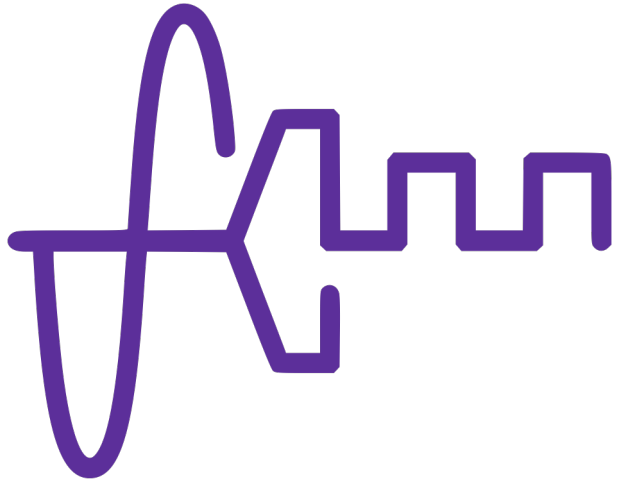
In the case of pure heteroscedasticity, the remedy may be complex…



Dealing with heteroscedasticity



Specify the model again / transform variables



Use Standard Robust Errors

To prove heteroscedasticity in a simple equation over the price of housing, we use the HPRICEI.RAW dataset

An advantage to use logarithmic form of dependent variable is that it may reduce heteroscedasticity.

STATA COMMANDS

```
9.   reg price lotsize sqrft bdrms. + heteroscedasticity price
```

References

- **Salvatore, D., & Sarmiento, J. C.** (1983). *Econometría* (No. HB141 S39). McGraw-Hill.
- **Gujarati, D. N.** (2009). Basic econometrics. Tata McGraw-Hill Education.
- **Wooldridge, J.M.** (2016). *Introductory Econometrics*, Cengage Learning, 6th edition.