

Applied Time-Series Analysis
Prof. Arun. K. Tangirala
Department of Chemical Engineering
Indian Institute of Technology, Madras

Lecture - 104
Lecture 45A - Estimation Methods 2 -2

Very good morning, so let us actually conclude on the weighted least squares problem today, briefly talk about non-linear least squares and then I will do a quick round up of MLE and if time permits will just get started on Bayesian. The idea in Bayesian as far as Bayesian estimation is concerned just to give you a preview; may not applied to estimating time series models is going to be a bit tough doing that in this course.

But certainly we will learn in the last week of this course which is next week how to now take this estimation methods and apply them to estimating two things, signal properties mainly ACF and the power spectral densities and in that process we will also learn what are whiteness test and then we will learn how to apply this estimation methods to estimating time series models. Once that is done then we will close the course with the brief discussion on forecasting on prediction. After all we have come all along only to forecast and if a pretty if you do not know how to predict once I have estimated the model.

We have visited those concepts in bits and pieces all along, it is just a matter of putting them together and formulizing certain concepts, but otherwise the basic result on prediction has already being given which is the conditional expectation is the best prediction, but then some more discussion on that is required.

So, let us get going on by weighted least square, I have already introduced to you the concept of weighted least squares and given you a few compelling reasons as to why one would be interested in formulating and solving a weighted least squares problem. Mainly for whatever reason you would like to attach different importance to different observations, it could be motivated by heteroskedasticity or may be correlation in the errors or many a times, there is this need for updating the model. So, when you are updating your model online then the brute forcing is to take the entire data from the time you started modeling to the present and rebuild your model; that means, re-estimate your theta. Obviously, common sense tells us that we may have to give more importance to

the most reason data and down play the data that is in the past and for this reason there is a concept called forgetting factor. We all have forgetting factors in build into us and it comes to learning remembering concepts, but when it comes to modeling unless you include that forgetting factor it will not forget the past.

So, this forgetting factor can be thought of again as a weighting. So, there are several scenarios in which we may want to formulate and solve a weighted least squares problem and yesterday I said that the weighted least squares problem can be solved very easily by recasting it as an ordinary least squares problem. And that is exactly what you see on the slide because W is positive definite and I have already told you why the waiting matrix should be positive definite, that is to preserve the convexity of the objective function.

(Refer Slide Time: 03:36)

MoM and LS estimators References

Solution to the WLS problem

Since W is positive-definite, we can perform a Cholesky factorization:

$$W = C^T C \quad (49)$$

Then, the objective function in (48) can be re-written as

$$(\mathbf{y} - \Phi\boldsymbol{\theta})^T C^T C (\mathbf{y} - \Phi\boldsymbol{\theta}) \quad (50)$$

Now, introduce scaled observations and regressors,

$$\mathbf{y}_S = C\mathbf{y}; \quad \Phi_S = C\Phi \quad (51)$$

Arun K. Tangirala Applied TSA November 3, 2016 NPTEL 61

We can perform a Cholesky factorization and scale the both the observations and the regressors accordingly, so the weighting factor determines this scaling.

(Refer Slide Time: 03:57)

MoM and LS estimators References

The WLS solution

The WLS problem can be then cast into an OLS formulation

$$\min_{\theta} (\mathbf{y}_S - \Phi_S \theta)^T (\mathbf{y}_S - \Phi_S \theta) \quad (52)$$

From the OLS solution, we thus have the WLS estimator

$$\hat{\theta}_{\text{WLS}} = (\Phi_S^T \Phi_S)^{-1} \Phi_S^T \mathbf{y}_S = (\Phi^T \mathbf{W} \Phi)^{-1} \Phi^T \mathbf{W} \mathbf{y} \quad (53)$$

► Scaling the data amounts to scaling the observation errors as well,

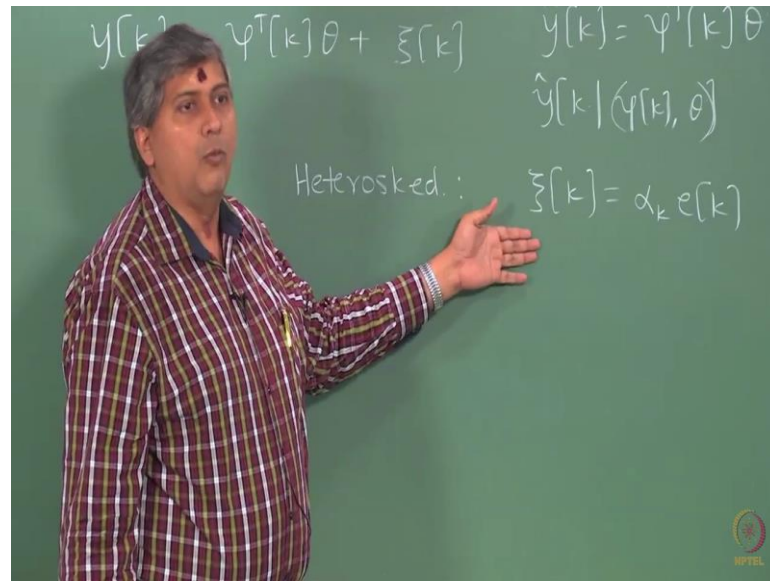
$$\xi_S = \mathbf{C} \xi \quad (54)$$

Arun K. Tangirala Applied TSA November 3, 2016 NPTEL 64

The nice thing is now in terms of the scaled data, we have an OLS. So, the WLS is nothing, but an OLS on scaled data, but the scaling factor being governed by \mathbf{W} . Since we already know the solution to the OLS, we can straight away write the solution to the weighted least squares; all one has to do is instead of Φ , use Φ_S and instead of \mathbf{y} use \mathbf{y}_S and then plug them back in. So, if you see this expression here instead of Φ_S ; we would substitute here \mathbf{C} times Φ .

So as to be able to write the solution in terms of the original data, not the scaled data, so that is it. So, you have here $\Phi^T \mathbf{W} \Phi^{-1} \Phi^T \mathbf{W} \mathbf{y}$ which is not a difficult solution to remember and straight away you should verify that when \mathbf{W} is a identity, you recover the OLS solution and what we will do is; we will try to now ask what governs the choice of \mathbf{W} and before we do that let me also tell you that as we have arrived the solution by scaling the data, we have also implicitly scaled the errors.

(Refer Slide Time: 05:27)



So, what we are saying is here is the original model linear regression model. So, let us assume I am got everything right what we are doing by scaling y and ψ we are also actually scaling the errors and that is the trick and this is the standard trick that you will see prevailing in parameter estimation, whenever the errors do not meet the criterion of ordinary least squares then one way out is to scale the errors, but of course the big issue is knowing the weights are prior.

So, here we are saying that these errors do not satisfy the conditions that are right for ordinary least squares, what are the conditions that are right for ordinary least squares they should be white. Whatever you have left out in your regression model should be white right that is when only your prediction we have been all along using the predictor as this. So this makes sense only from this prospective as well using this predictor makes sense only when what you are left out is white.

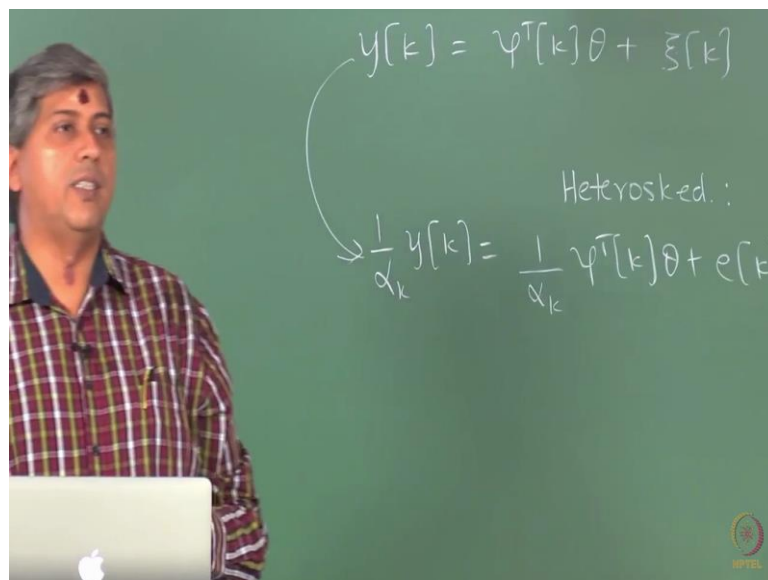
If what you are left out is not white; obviously, you should factor that into your prediction, remember all them writing here \hat{y} of k implicitly you should understand that this \hat{y} of k is given ψ k and θ . Normally we do not write this, but it is to be understood that I am given the regressors, I am given the θ and hence the prediction.

In time series models your regressors are past data and therefore, your \hat{y} becomes typically once step ahead prediction alright. Now if this was not white then the one step ahead prediction is not going to be 0 and again from that perspective you want to make

sure that what you are left out is white. So, coming back to the point, if you know for some reason that what you have left out is white and you do not want to really model this part and you are only bent on estimating this theta.

So, as an example suppose these are white, these are coloured but they are not necessarily white in the definition sense remember white noise process is a stationary uncorrelated process. The moment z has heteroskedasticity, the stationarity properties gone, it is still uncorrelated that is temporally, but it does not have the stationarity property. In such cases what you are saying is I know that this z_k can be written, so when I have heteroskedasticity maybe we could write this as $\alpha_k \epsilon_k$ where ϵ_k is now a stationary, your classical white noise process and α_k is a scaling factor that keeps changing with the observation or with the index and what we are doing by scaling is essentially; remember what as you will see shortly for the heteroskedastic case the weighting matrix turns out to be the inverse of this noise covariance matrix. And you will see essentially that your W 's would be inverses of alpha; that means, your W would be a diagonal matrix of $1/\alpha_k$. If you did not know the weighted least squares problem let us say you did not know the formulation, all you knew that OLS works best for the pure white noise case then if you are given that z has this kind of a structure.

(Refer Slide Time: 09:35)



Then what you do is, if you are given alpha you would rewrite this in this way $1/\alpha_k y_k = 1/\alpha_k \psi^T[k] \theta + e_k$. So, you would rewrite the same equation

given alpha k in this fashion. Now you defined a new y and a new regressor that new phi and new y and new regressor is nothing but your scaled data, but what this point also says is that in that process you are rescaled your errors such that now in the scale domain, the errors meet the requirements for the OLS to work for you. OLS will work, but it works best; that means, it will give you the most efficient estimates when the residuals are white or your noise is white. So, that is the basic ideas you should keep in mind, as you as scaling the data; we are also scaling the errors in a particular fashion.

So, the question is now how to choose the weights mat weighting matrix W who gives me this do I know alpha k for example, this in the heteroskedastic case there are as I have already said many other situations in which we run into W formulation. In all such formulations, how do I choose W?

(Refer Slide Time: 11:09)

MoM and LS estimators References



Choice of weights

How to choose the weights matrix W?

The answer depends on the application / criterion.

- i. **Model updation:** W is a diagonal matrix of *forgetting factors*
- ii. **Efficient estimates:** The goal here is to achieve minimum $\text{var}(\hat{\theta}_{\text{WLS}})$.

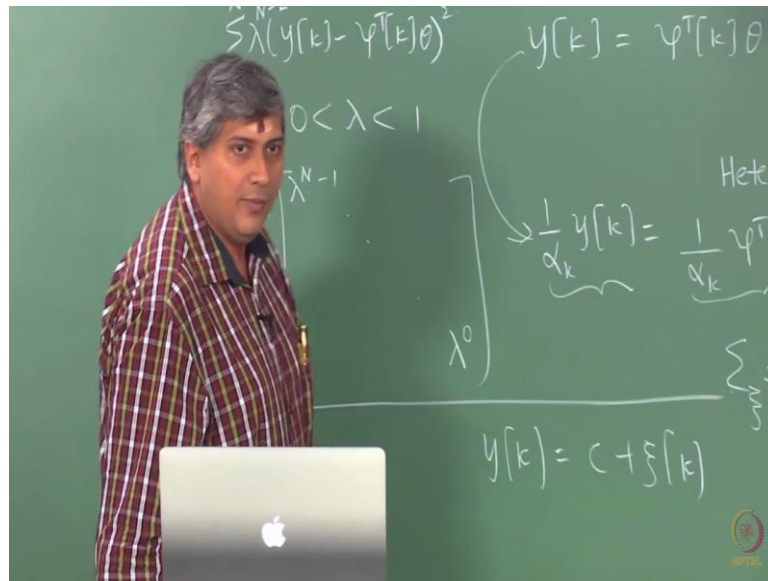
$$\Sigma_{\hat{\theta}}(\mathbf{W}) = \text{var}(\hat{\theta}_{\text{WLS}}) = (\Phi^T \mathbf{W} \Phi)^{-1} \Phi^T \mathbf{W} \Sigma_{\xi} \mathbf{W} \Phi (\Phi^T \mathbf{W} \Phi)^{-1} \quad (55)$$

Arun K. Tangirala Applied TSA November 3, 2016 66

Now there are many considerations that can going to it, but primarily there are two; one is model updation where W is a diagonal matrix of forgetting factors.

(Refer Slide Time: 11:22)



What we mean by forgetting factors is you would have your objective function as follows instead of simple $y[k] - \psi^T[k]\theta$ square you would have a λ^{n-k} ; that is k runs from 1 to n or 0 you can also modified this accordingly.

So, the idea is here λ is a forgetting factor it is a positive value between 0 and 1. So, that the most recent data that is when k equals n ; λ is λ^{n-k} is 1. So, the most recent data is given the maximum importance and the last one; that is the first observation in k equals 1, if you are standing at n then the past one which is really remotely in the past is k equals 1; that is given the least importance because λ is a factor between 0 and 1.

So this is a concept of forgetting factor; you can straight away now recognize what is W . What would be the W here, it would be a diagonal matrix and what would be the entries here; λ the first one would have $n-1$ up to λ^0 . So, when you are looking at model updation using the concept of forgetting factors W becomes a diagonal matrix of forgetting factors and λ is a users choice, there is nothing much there you would decided how much importance has to be given, but a far more important and prominent consideration is to obtain efficient estimates because that is how we started off. Remember we said that ordinary least squares does not work when z is not white

when the equation error is not white and then we set on to the weighted least square formulation.

So obviously, that consideration has to be now taken into account and the question to be asked is what is W , what is a weighting matrix now that will guarantee efficient estimates. When you are looking at a heteroskedastic case, it is very obvious what is W ; if I knew the variances of z , if I knew the variance or if I knew α_k ; all I have to do is what would be W diagonal matrix of 1 by α_k very good right; stating that if α_k is high then that observation should be given lower importance.

Obviously, if one observation if α_k is high for a particular k then that observation has a larger error. Suppose α_k is a highest for that particular observation then that observation has a largest error among all the observations and obviously, you want to down play that observation; W will be inversely proportion to α_k .

(Refer Slide Time: 14:45)

$(y[k] - \psi^T[k]\theta)^2$
 $0 < \lambda < 1$
 λ^{N-1}
 λ^0
 $y[k] = \psi^T[k]\theta + \xi[k]$
 $\hat{y}[k]$
 Heterosked.: $\xi[k] =$
 $\frac{1}{\alpha_k} y[k] = \frac{1}{\alpha_k} \psi^T[k]\theta + e[k]$
 $\sum = \begin{bmatrix} \alpha_1^2 & & \\ & \ddots & \\ & & \alpha_n^2 \end{bmatrix}$

So, you get some idea now that for obtaining efficient estimates we need to have W as the inverse of the noise covariance matrix; that is if I were to look at the noise covariance matrix of z in this case, it would be simply at the diagonal matrix, but let us assume $\sigma^2 e$ is 1 ; then I would have α_1^2 up to α_n^2 and W is going to be simply the inverse of this.

Remember I am scaling the data with 1 over alpha and when I move on to the objective function, when I rewrite the objective function now in terms of this scaled observations and regressors, it amounts to writing a objective function for the original data using a weighting of 1 over alphas square because there is a square. So this leads us actually to the answer, but theoretically the way you would arrive at this solution that is to the question what should be the waiting matrix W if I want the most efficient estimate and efficiency has got to do with the variance. So, you can show that in a similar way that we have shown earlier for the OLS, the sigma theta hat is nothing but that is for the weighted least squares is the simple looking expression here.

So, as you can see here it is a very simple expression, I do not expected to you remembers this at all even, I have tried to remembering it many times, but failed tot many W's and phis and so on, but what you should quickly check is; if W is identity and z is white, it simplifies to the OLS expression that we have seen sigma square e times phi transpose phi inverse. Now what people have done is taken this simple expression in days where there was a lot of time and asked what gets me the minimum value of this; this is a matrix. So, does not make sense to talk of minimum of sigma theta hat people are looked at minimize in the trace what is trace some of diagonal elements.

(Refer Slide Time: 16:59)

MoM and LS estimators References

Choice of weights for efficient estimates

The **optimal weighting matrix is the inverse of the covariance of equation errors,**

$$\mathbf{W}_{\text{opt}} = \Sigma_{\xi}^{-1} \quad (56)$$

With this choice, the variance of the WLS estimator is

$$\Sigma_{\hat{\theta}}(\mathbf{W}_{\text{opt}}) = (\Phi^T \mathbf{W}_{\text{opt}} \Phi)^{-1} = (\Phi^T \Sigma_{\xi} \Phi)^{-1} \quad (57)$$

The result is nicely understood in systems with heteroskedastic errors, as we shall observe next.

Arun K. Tangirala Applied TSA November 3, 2016 NPTEL 68

So, they have looked at that and essentially come up with this answer the optimal W is nothing, but sigma z inverse which is what intuitively we expected. Our arguments

where based on the heteroskedastic example, but this is a there is a formal proof available just look up the literature it is not of interest was at this moment, we argue without any rebate that this is a correct solution we agree.

So, intuitively we know that the solution makes sense of course, the big question is who gives me σ_z , it is a very very rare commodity you cannot find it online no where you can have to get it from data and that is the typical approach that is used in all weighted least squares or any weighted approach, where you are weighting meeting matrix depends on a noise covariance matrix, but look at how important understanding this entire theory is; this one equation should help you appreciate a formal the need for a formal study of the subject.

If you did not know any of this you would be performing ordinary linear regression on data which has probably heteroskedasticity or errors are not white and so on and assume that your fit is best that the model is good, but the fact is you should have worked with scaled data. In fact, there is a whole lot of literature and we are continuing to work on it in a straightly different frame work, using these ideas we have some really break through solutions very recently in our group and now we are trying to communicate that to the world.

So, this simple idea of weighting and incorporating that in the regression and that the fact that this weighting depends on the noise covariance matrix of the errors should really you know help us appreciate the beauty of this formulization. In fact, I told you the noise covariance matrix is perhaps or just the covariance matrix you been seeing it coming time and again in the course is perhaps the most ubiquitous quantity that you will keep encountering all through your life in data analysis either, there no escape to it even when you are non-linear world the covariance matrix will come and haunt you or right to help you with your data analysis in fact.

So, let us look at an example to understand the beauty of this result by the way. So, when you choose this weighting matrix it turns out that the σ_{θ} in the weighted least squares case turns out to be this expression here; $\phi^T \sigma_e \phi$ $\sigma_z \phi^T \phi^{-1}$. So, that is what easy answer and once again you should check when σ_z is white; what happens σ_z will be a diagonal matrix of σ^2 is on the diagonal and you recover the OLS solution and that is anywhere

we know is the most efficient. So, it just a corroboration of what we already know, but it is a good practice to keep checking so that you know the expression that you working with are correct.

(Refer Slide Time: 20:21)

MoM and LS estimators References

Example: Heteroskedastic errors

Sensor fusion: Steady-state estimation

Temperature measurements of a reactor at steady-state from 10 thermocouples that have different, but known error characteristics (variability).

Sensor	1	2	3	4	5	6	7	8	9	10
Meas. ($^{\circ}C$)	61.2	64.3	59.1	64.1	63.8	62.9	58.2	60.7	61.5	63.7
Variance	0.36	2.25	1.69	0.25	0.49	2.89	3.2	1.4	1.2	2.7

where the readings have already been adjusted for calibration.

Estimate the steady-state temperature from these ten different measurements.

Arun K. Tangirala Applied TSA November 3, 2016 NPTEL 71

Let us look at an example we have already talked about this, so this is an example as I pertaining to what I discussed earlier; there is a reactor and temperature measurements are being made by ten sensors. Let me tell you also this is a syntactic data, but representative of a realistic scenario. So, I definitely I have cocked up the data, but this is not something very rare you will see this kind of situations quite often and what I have given you in this case is the sensor variances. So, here the k for us is not time the k for us is sensor.

So, I have 10 sensors and I am suppose to estimate the average, if I did not know anything as a lay man I would take simple mean or sample mean of this 10 sensor readings and get the reactor temperature assumption is that the reaction temperature is constant during this experiments. But suppose I were to give the variances of each of this sensor, it is possible to obtain those in this kind of experiments. For example, say if I know the process at study state then I can observe the sensor noise in each sensor and estimate the variance and report them.

So, imagine that that is how I done it and the variance have been given. So, now the goal is to estimate is steady state temperature as I said the simple solution is the least squares

solution which is sample mean, but now that I am familiar with the weighted last squares concepts, I would rather construct the weighted average that is something that you should remember that when you are having heteroskedastic errors and you are estimating mean, what you would be working with for an estimator of mean is a weighted average and how does that come out to be because W, we know optimal W is 1 over sigma square along the diagonals and all you have to do is plug in to the solution here; phi transpose W phi inverse phi transpose W y. In every problem until you get use to it you should make it a habit to recognize what is the regressor, what is the data. The data y is the data given to us the regressor, in this case because I am estimating an average remember the problem of interest to me is that y our familiar problems c plus z k instead of e k I have z k.

So, this c is the unknown that I am estimating that is our theta, but there is no regressor; however, always this there are this two numbers in math 1 and 0 which are invisibly present in some form of other in every number. So, 1 is the invisible regressor are there and that is what makes up my phi there is only one regressor, so what would be phi transpose or let us a phi itself.

Student: (Refer Time: 23:21).

It will be a column matrix of once correct, as a result phi trans and W; we any anyway knows the diagonal matrix. So, the essentially it amounts to this expression here.

(Refer Slide Time: 23:38)

MoM and LS estimators References

Example: Heteroskedastic errors ... contd.

The WLS estimate of the average and its error variance are then given by,

$$\hat{\mu}_{WLS} = (\Phi^T \mathbf{W} \Phi)^{-1} \Phi^T \mathbf{W} \mathbf{y} = \frac{\sum_{k=1}^N \frac{y[k]}{\sigma_k^2}}{\sum_{k=1}^N \frac{1}{\sigma_k^2}} = 62.6086;$$

$$\text{var}(\hat{\mu}) = (\Phi^T \mathbf{W} \Phi)^{-1} = \frac{1}{\sum_{k=1}^N \frac{1}{\sigma_k^2}} = 0.0805$$

where k indicates the sensor index for this example.

Arun K. Tangirala Applied TSA November 3, 2016 NPTEL 73

So, $\Phi^T W \Phi$ would be your denominator because W is a diagonal matrix and $\Phi^T W y$ would be the numerator again because W is a diagonal matrix Φ^T would be a row vector of ones once you get it and that is all. So, the Φ^T here is the one that is responsible for the summation that you see and this y by σ_k^2 comes about by this multiplication W times y ; all I straight away recognize I am constructing a weighted average.

Now the value turns out to be 62 point something and the variance can also be computed, remember we said when we choose W to be the inverse of noise covariance matrix then the variance is simply $\Phi^T W \Phi$ inverse and once again compute that to be 0.0805. Now why did we go through this weighted least square approach because we thought this gives me better results in the least square approach better in what sense?

Student: Efficiency.

Efficiency what does efficiency mean.

Student: Variance should be lower than another one.

Variance should be lower than another one. So, we should expect WLS to be more efficient than the least squares the question is if it is.

(Refer Slide Time: 25:17)

MoM and LS estimators References

Example: Heteroskedastic errors ... contd.

Compare corresponding results from OLS

$$\hat{\mu}_{OLS} = \frac{1}{N} \sum_{k=1}^N y[k] = 61.95; \quad \text{var}(\hat{\mu}) = (\Phi^T \Phi)^{-1} \Phi^T \Sigma_v \Phi (\Phi^T \Phi)^{-1} = \frac{\sum_{k=1}^N \sigma_k^2}{N^2} = 0.1643 \quad (59)$$

The widths of confidence intervals for the average correspondingly would be proportional to $2\sigma_{\hat{\mu}}$, i.e., 0.2835 and 0.4053 respectively.

Arun K. Tangirala Applied TSA November 3, 2016 NPTEL 74

So let us look at this here; first of all look at the point estimate. Now this is where we need for computing the variance also is highlighted. So, you look at the point estimate is it significantly different from the point estimate that we computed with the weighted least squares; not much this gives a 61.95 something that gives me 62 point something. So, if you look at the difference in point estimates it is not much, but if you look at the variability in the estimate; what do you notice here you have 0.16 right and whereas, you have 0.08. So, the variance of the ordinary least squares is twice the variability that you seen weighted least squares.

If you translate that in terms of standard errors it is 1.41 raffle right that can be quite a lot in many applications, but also let me tell you that in this example the factor turns out to be 1.41 that is the in terms of errors, in some examples it may be much higher than this in other examples it may be lower than this, but you are always guaranteed that weighted least squares will not perform was an OLS; at best it will be OLS.

So, you do not lose much by working with WLS, but what you can lose out on is the computational time because it is an iterative typically W is not known, you have to iterate and therefore, you have to put in some effort. The question is whether the effect is worth it, you may not know that a priori sometimes you may know; sometimes you know clearly there is a need for the weighted least squares. So, you say it is worth the iteration sometimes it is ok there is a heteroskedasticity, but the fluctuation is very very low more or less I will treat it as a homo heteroskedastic case, I will scarifies bit of a efficiency in the in the estimates for computation efficiency. So, there is conflict between computational efficiency and estimate efficiency, so those are the standard things that you will come across in estimation. So, hopefully this example gives you a better picture of how weighted least squares works and the inner workings of it.