

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

Lecture – 97
Lecture 42B - Estimation Methods 1 -4

So, very quickly let us go through some standard stuff and then come to the most important question how well does a least square method perform in terms of estimating parameters? At the moment we have not asked at until now we have not asked that question. So, now, that you have an optimal estimate you also want to now predict using this model, the first thing that we always do is when we want to see how good the model how good a model we have obtain.

We apply the same model on to the same training data set that we have used or it can be used in a fresh data set this, these equations fair fairly generate.

(Refer Slide Time: 00:52)

Model and LS estimates

Predictions and Residuals

The optimal prediction of y and the associated residuals are:

$$\hat{y}_{LS} = \Phi \hat{\theta}_{LS} = \Phi (\Phi^T \Phi)^{-1} \Phi^T y = P y \quad (12)$$
$$e_{LS} = y - \hat{y}_{LS} = (I - \Phi (\Phi^T \Phi)^{-1} \Phi^T) y = P^\perp y \quad (13)$$

Arun K. Tangirala Applied TSA October 26, 2018

So, given any new phi or the old phi does not matter, the predictions are simple phi times theta hat. Now I have used a hat earlier when I solving the theoretical least square problem I did not use the hat, this theta hat is an estimate of the true theta least squares that you would have obtain, had you looked at the entire vector y and the entire matrix phi ok.

So, \hat{y} is simply $\Phi \hat{\theta}$ very simple plug in the estimate get your prediction. It turns out that you can write this \hat{y} as you see in equation 12, you can write \hat{y} as simple $P y$; P is a matrix purely constructed from the regressors, what this tells us is in order to construct a prediction of y I do not need θ , I just need Φ , but we have to go through the problem to arrive at optimal prediction, I do not have to compute θ necessarily I can directly use the regressor to the regressor that I used for my estimation remember there is that is very very important; here we are assuming that we are using the same training data set.

(Refer Slide Time: 02:16)

Suppose let me put it this way suppose I am given a new set of regressor, there is a fresh data set that I have obtain and there is an old data set from where I have estimated θ . So, let us actually abbreviated let us call this Φ_n as Φ_n and Φ_{old} as just simply the Φ_o . So, this is the one that I have used for estimating θ , this is the one that perhaps is coming out of it fresh data set.

Now, if I want to predict y on this fresh data set that I have, then the first equation that I write is $\Phi_n \hat{\theta}$, but remember $\hat{\theta}$ is constructed from Φ_o which is your training data, it is like you have as an instructor you have a question bank your partition them into one for homework and one for exam other one for exam. So, the Φ_o comes from your training data set from your homework data set and that is what is used in construct estimating θ .

Now, you have to be careful here where I am just pointing out that the phi that is being written there is being written without the subscript, but you have to be careful. Generally in r what happens is this theta is told. So, none of this is probably not so relevant in general when you fit a model using l m, the object that is being written stores the coefficients and then you apply that to a fresh data it simply do this calculation for you, but there is this there are this equations which bypass the use of theta.

So, you have to be carefully when you are using this at the moment assume that we are predicting on the training data set itself so that the subscript do not matter. What you can see is that y hat and epsilon; y hat is being written as p times y and epsilon is written as p perpendicular y, y is what is this p perpendicular? It is a orthogonal compliment of p.

(Refer Slide Time: 04:54)

World and LS estimates

Predictions and Residuals

The optimal prediction of y and the associated residuals are:

$$\hat{y}_{LS} = \Phi \hat{\theta}_{LS} = \Phi(\Phi^T \Phi)^{-1} \Phi^T y = P y \quad (12)$$

$$\epsilon_{LS} = y - \hat{y}_{LS} = (I - \Phi(\Phi^T \Phi)^{-1} \Phi^T) y = P^\perp y \quad (13)$$

where

$$P = \Phi(\Phi^T \Phi)^{-1} \Phi^T \quad \text{and} \quad P^\perp = I - P \quad (14)$$

are said to be the *projection matrix* and its *orthogonal complement* respectively. The latter name is due to the fact that

$$P P^\perp = P^\perp P = 0 \quad (15)$$

Amn K. Tongolo Applied TSA October 26, 2018

Remember our least square solution is such that epsilons are orthogonal to the regressor, and you should be able to see that when I take the inner product of epsilon with y hat I will get 0. So, that is what it is conforming, it is just re affirming what we already known about the least square solution.

So, these equations that you are seeing are being provided for two reasons: one is to compute the predictions given a set of observations and the other one is of course, computing the residuals and to reaffirm the fact that epsilons are indeed orthogonal to the regressor that is all. So, this p is called the projection matrix and p perpendicular is simply the orthogonal compliment anyway I mean this is just for information, let us

move on we have talked about pseudo inverse, so I am going to skip. I have also talked about the equivalence of OLS with method of moments right last time in the last class when we started off with the theoretical least squares problem, we wrote this equation that you seen equation 20 $\Sigma \theta = \Sigma y$.

(Refer Slide Time: 06:00)

MOM and LS estimators

Equivalence of OLS with MoM

From the projection theorem, it is somewhat evident that the OLS estimate of θ in the linear regression is equivalent to the method of moments estimate.

The equivalence comes from the theoretical LS estimator:

$$\Sigma_{\psi\psi}\theta = \Sigma_{\psi y} \implies \theta_{LS}^* = \Sigma_{\psi\psi}^{-1}\Sigma_{\psi y} \quad (20)$$

where the Σ stands for covariance matrix.

Replacing the theoretical covariances by the sample versions gives rise to the sample LS estimator in (11) as well as the MoM estimator.

Arav K. Tongolo Applied TSA October 26, 2018

We had use the symbol $\Sigma \theta = \Sigma y$.

So, from a method of moments prospective, what is idea and method of moments you write the theoretical relations between the parameters and the moments and then replace the theoretical moments with time averages. So, if I were to replace here of course, we are assuming y and ψ to be random, if I were to replace $\Sigma \psi \psi$ or Σy with its respective estimate, what would be the respective estimate this (Refer Time: 06:36). So, this would be $\hat{\Sigma} \psi \psi$ and likewise the cross covariance with this term.

So, this would give me an estimate of the cross covariance then you get method of moments, but you see method of moments says you can replace it with any other estimate also; if you replace the theoretical once with these estimates then you recover the OLS. So, method of moments and OLS give the same solution; however, when we talked of OLS solution, we did not have to invoke any randomness that part you should remember whereas, with method of moments right from step one both the regressor and the data are assume to be random in nature.

So, including a constant term, until now we have not included the intercept so called intercept term, but it is very easy.

(Refer Slide Time: 07:44)

Multi and LS estimators

Including a constant term

A constant term (intercept term, or a non-zero mean) can be accommodated in the regression model by simply appending the regressor with a vector of ones as seen below.

$$y[k] = \varphi^T[k]\theta + \beta = \begin{bmatrix} \varphi[k] & 1 \end{bmatrix}^T \begin{bmatrix} \theta \\ \beta \end{bmatrix} \quad (21)$$

Interestingly, the LS estimate of the constant term β can be obtained sequentially by first obtaining $\hat{\theta}_{LS}$ followed by,

$$\hat{\beta}_{LS} = \bar{y} - \bar{\varphi}^T \hat{\theta}_{LS} \quad (22)$$

where \bar{y} and $\bar{\varphi}$ are the sample means of $y[k]$ and the regressors respectively.

Alan N. Sargante Applied TSA October 26, 2018

All you have to do is you can think of this intercept term as beta suppose the intercept term is beta, then think of this intercept term as being beta times 1 so that now you are including one in the regressor matrix, so that is why I have written this equation; if you look at equation 21 what have we done? We have converted the non zero intercept problem into a once in the regressors problem, this is a very standard trick that is used in linear regression where if there is an intercept term all it does is it includes a vector 1 in the regressor vector and it says there is a new regressor whose value is 1 and finds the corresponding parameters that is all. So, in additional parameters is estimator.

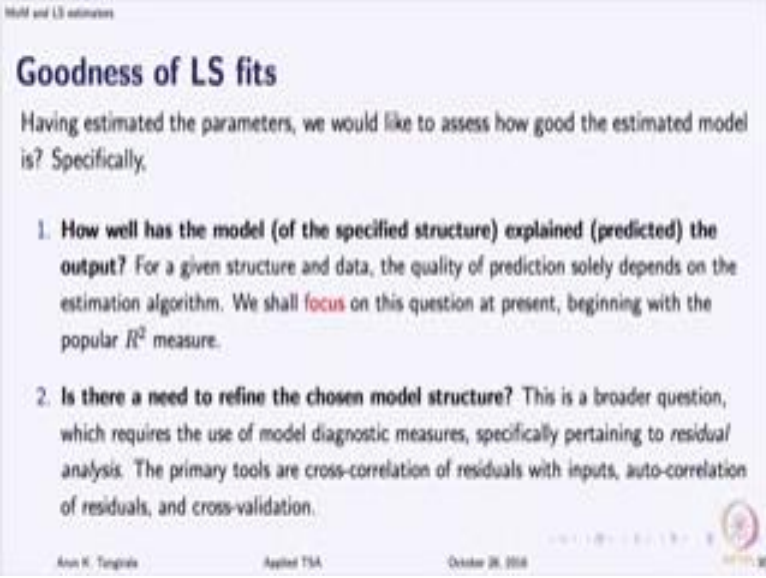
It turns out interestingly that the optimal estimate of this intercept term is nothing, but \bar{y} which is the sample mean of the simple average minus the psi transpose theta hat. So, it is $\bar{\varphi}^T \theta_{LS}$ sorry; $\bar{\varphi}$ is the sample means of the regressors. So, it is a very simple solution when you work out everything it turns out that the optimal estimate of beta is simply the difference between the sample mean of y and the sample mean of the regressors times theta hat that is all. So, which means you can effort to decouple this problem; that means, you can estimate the intercept term later on you do not have to estimate it simultaneously with the other parameters. Estimate your theta

calculates the time averages of y and the regressors and use equation 22 to get your beta hat.

So, therefore, most of the least square problems are presented without the intercept term because you do not have to include that in your least square formulation. In R when you use `lm` by default it assumes that you are estimating an intercept term. Now we come to the practical aspects of this least squares, where we ask two different questions; for the given model whatever model I have given what kind of fit has been obtain; that means, how good is the fit for this given model whenever and in fact, this is not the question only pertaining to least squares, any estimation method any modelling exercise you should ask two questions.

One how well has the given model explained, two is there any scope for improvement; both are related, but they are different kind of questions one is telling about the goodness of the fit obtained by the model and the estimation method and the other is talking about if there is anything if there is a scope for improvement, if there is if there are further refinements that are necessary. So, here we are asking the first question how well has the model explained or predicted the given y ?

(Refer Slide Time: 11:02)



Model and LS estimates

Goodness of LS fits

Having estimated the parameters, we would like to assess how good the estimated model is? Specifically,

1. **How well has the model (of the specified structure) explained (predicted) the output?** For a given structure and data, the quality of prediction solely depends on the estimation algorithm. We shall **focus** on this question at present, beginning with the popular R^2 measure.
2. **Is there a need to refine the chosen model structure?** This is a broader question, which requires the use of model diagnostic measures, specifically pertaining to residual analysis. The primary tools are cross-correlation of residuals with inputs, auto-correlation of residuals, and cross-validation.

Amir K. Taparia Applied TSA October 26, 2018

And usually you will come across measures like R square, you must have heard of R square or adjusted R square; we will just briefly talk about that and come to the second question.

(Refer Slide Time: 11:17)

Model and LS estimates

R^2 measure

The R^2 measure is a goodness-of-fit index. It gives a bird's eye view of how well the model has explained the variations in the data. Its definition is based on an important feature of LS estimation:

$$\underbrace{\sum_{k=0}^{N-1} (y[k] - \bar{y})^2}_{\text{sum square total (SST)}} = \underbrace{\sum_{k=0}^{N-1} (\hat{y}[k] - \bar{y})^2}_{\text{sum square predictions (SSP)}} + \underbrace{\sum_{k=0}^{N-1} e^2[k]}_{\text{sum square errors (SSE)}} \quad (23)$$

Ann K. Tongolo Applied TSA October 26, 2018

So, this R square measures the famous one that is being presented in all linear regression exercises, is a measure of how much your model has managed to capture the total variation in y ; that means, y is changing because of some reason and you believe it is primarily because of the regressor, how well has your model now manage to capture that fluctuation and you can show first of all for the least squares when you use the least squares method to estimate parameters, what it is essentially doing? This is another prospective; we have talked about a few other prospective before.

This is another prospective of least square what it is doing essentially is? It is breaking up the sum squares of $y[k] - \bar{y}$; for now assume that \bar{y} is zero just for the sake of discussion \bar{y} is simply the time averages, then what you have on the left hand side of this equation? Is simply the square two norm of y and what you have on the right hand side what do you have?

Student :(Refer Time: 12:28).

The squared 2 norms of \hat{y} and epsilons; so essentially what it is doing is its minimizing this, but what it is doing is it is breaking up.

(Refer Slide Time: 12:41)



Assume \bar{y} is 0 it is essentially breaking up in this way, it has to be right because if you go back to this equation this is not surprising, let me write the equation here and show you that this is not a big deal because you have already seen this before in a different form; what we had said is find θ such that ϵ is orthogonal to \hat{y} that is what least square is doing.

So, when I take here inner product, what is the square two norm of y ? It is essentially the inner product of y vector with itself. So, I have to take inner product of \hat{y} plus ϵ with \hat{y} plus ϵ and when you do the expansion you will get this result; using the fact that what is what is the fact that we use?

Student: (Refer Time: 13:35).

So, least squares give you solution as that ϵ is orthogonal to \hat{y} . So, this breaking up of you can say some square y or you can say square two norm, some people would call it energy, some other would call it variation whatever you call this is what is happening by virtue of least squares. Other methods do not necessarily do this, it is based on this expression that R square measure is derived. The R square measure essentially looks at what fraction of this variation or energy has gone unexplained by this model and by least squares remember least squares is doing it.

So, there are two factors playing the roles here in this decomposition, we are doing a signal decomposition we said earlier, but what we are also doing is a sum square decomposition have we seen this kind of a relation before?

Student: (Refer Time: 14:38).

Anything else in Fourier analysis, we have talked about signal decomposition followed by parsevals relation, but their of course, in the parsevals relation we said completely the sum square in time is preserved by sum square in the frequency, we do not say decomposition, but your right it is a kind of a Pythagoras theorem, essentially if you look at this schematic that we have drawn before, you have y , \hat{y} which is optimal projection and then you have epsilons so you can see here as well.

So, only least squares gives you this kind of a break up and R square measure essentially is defined based on this, it is sum square prediction by sum square total.

(Refer Slide Time: 15:25)

The slide is titled "R² measure ... contd." and contains the following text: "The total variance of the output is broken up into two additive terms - the variance explained by the model and the variance of the residuals." Below this, a red header reads "Coefficient of determination R²". The main content is a mathematical equation (24) defining R² as the ratio of the Sum of Squares of Predictions (SSP) to the Sum of Squares of Total (SST), which is equal to 1 minus the ratio of the sum of squared residuals to the total sum of squares.

$$R^2 \triangleq \frac{SSP}{SST} = 1 - \frac{\sum_{k=0}^N \epsilon^2[k]}{\sum_{k=0}^{N-1} (y[k] - \hat{y})^2} = 1 - \frac{\|\hat{y} - y\|_2^2}{\|y - \hat{y}\|_2^2} \quad (24)$$

At the bottom of the slide, there is a footer with the text: "Arun K. Tongida Applied TSA October 26, 2018".

That means the square 2 norm of \hat{y} by square 2 norm of y . Higher the R square better is the fit all right. So, ideally people want R square 1. In fact, getting R square 1 is not possible because always y will contain something that the regressors cannot explain, which means epsilon is always going to be non zero. But how much we have missed out is given by R square and because see the other point that you should remember is this R square measured that we are defining, can be defined for other estimators as well; I may

~~am I use m l e, I mayam-I~~ use some other method does not matter. I can always define R square, but only when I use R square with least squares, I can guarantee that R square will be between 0 and 1 why is that?

Student: (Refer Time: 16:29).

Correct because of this property that least square enjoys, only least squares guarantees that \hat{y} .

(Refer Slide Time: 16:41)



This sum square at the square two norm of \hat{y} is less than or equal to the sum square two norm of y and R square is simply a ratio of that right. So, straight away you can see it is always going to be between 0 and 1. So, if you try to compute R square which you can for some other method, for the same model if you use some other method, it is possible that R square can be greater than 1; do not panic, but then that means, it is not a good measure to use all right.

Now, but there is a problem with R square high values of R square indicate good fits very good, but it does not tell you whether the model is having fixed, that is whether the model has been over parameterized. If you have burden it with too many parameters then there is an issue, why I say this why do I why do you think R square cannot detect what happens if you over parameterized?

Student :(Refer Time: 17:57).

What becomes very R square? Because you want R square high and high you want very good fits. So, you keep increasing the number of regressors in you are model; obviously, I mean numerically as you keep increasing more and more regressor, more and more fits can be obtained better and better fit can be obtained; that means, you will be shrinking epsilons. So, R square is being taken close to 1; what we would be forgetting there is that by including more and more parameters, which will become quite obvious soon we would be also increasing the error in the estimate of theta, as a result it becomes unsuitable for use on fresh data sets.

Its cross validation abilities will be poor, it becomes more and more specialised to the data right it is like may be double or triple PhD seriously if you take the PhD, if you ask anything outside the area of specialization answer would be measurable (Refer Time: 18:58) are kind of good least square fits on the subject about to be depends all right. So, that is a problem with R square and it is what with that reason that adjusted R square was introduced.

(Refer Slide Time: 19:14).

Wald and LS estimates

Adjusted R^2

- R^2 has a poor sensitivity with respect to inclusion (or exclusion) of additional regressors. Thus, it cannot be used to determine overfits.
- An **adjusted R^2** that is based on the mean square,

$$\hat{R}^2 = 1 - \frac{SSE/(N-p)}{SST/N-1} = 1 - \frac{N-1}{N-p}(1-R^2) \quad (26)$$

is useful for determining overfits. The factors $(N-1)$ and $(N-p)$ denote the degrees of freedom associated with the SST and SSE respectively.

- The modified measure can assume negative values unlike the classical R^2 .
- It measures the balance between prediction bias and the variability of estimates.
- In practice, sophisticated measures based on information theory such as AIC and SIC are employed.

Applied T&A October 26, 2008

Adjusted R square is introduced so as to take into account over fitting. So, that the user is being told that there is a penalty for including more and more parameters.

How is the definition modified? Now the definition is modified such that you do not base it on sum square errors and sum square total, you take the average. Now you calculate the sum square error not the sum square error, but what is known as mean square error

how is the mean square error calculated? So, if you look at this equation here we have 1 minus SSE by N minus p in the numerator, why do we have N minus P, why is the mean square error, how many terms do we have in sum square error? N right, epsilons is an N by 1 component ideally if I have to calculate the mean square error, I should have simple said 1 over N sum square error, but we are calculating mean square error by 1 over N as 1 over N minus p why are we doing that any idea?

So, all that we are doing is in going from R square to adjusted R square, we are replacing the sum square error that we have in equation 24 with mean square error and sum square total by mean square total. When I am calculating sum square total I have N minus 1 as the divider and when I am calculating sorry mean square total I am using SST by n minus 1 and when I am calculating mean square error I am using N minus p why is that? Remember your sum square total is actually this term here on the left.

(Refer Slide Time: 21:08)

R^2 measure

The R^2 measure is a goodness-of-fit index. It gives a bird's eye view of how well the model has explained the variations in the data. Its definition is based on an important feature of LS estimation:

$$\underbrace{\sum_{k=0}^{N-1} (y[k]^k - \bar{y})^2}_{\text{sum square total (SST)}} = \underbrace{\sum_{k=0}^{N-1} (\hat{y}[k] - \bar{y})^2}_{\text{sum square predictions (SSP)}} + \underbrace{\sum_{k=0}^{N-1} \epsilon^2[k]}_{\text{sum square errors (SSE)}} \quad (23)$$

Atan K. Tongidis Applied TSA October 26, 2018

In calculating y bar when I begin I have n observations, but when I calculate y bar I have lost one degree of freedom.

So, I have effectively N minus 1 degrees of freedom, we have talked about is degrees of freedom before. So, the correct way of calculating mean square total is to divide this by N minus 1. Now you can also explain why we are using N minus p for calculating mean square error. To begin with I have n observation and n errors, but before I have computed, before computing those n errors I have estimated p parameters. So, I have lost

p degrees of freedom and therefore, in calculating mean square error we use N minus p that is a reason why the mean square error and mean square total are calculated with N minus p and N minus 1 respectively.

(Refer Slide Time: 22:09)

Multiple Regression

Adjusted R^2

- R^2 has a poor sensitivity with respect to inclusion (or exclusion) of additional regressors. Thus, it cannot be used to determine overfits.
- An **adjusted R^2** that is based on the mean square,

$$\bar{R}^2 = 1 - \frac{SSE/(N-p)}{SST/N-1} = 1 - \frac{N-1}{N-p}(1-R^2) \quad (26)$$

is useful for determining overfits. The factors $(N-1)$ and $(N-p)$ denote the degrees of freedom associated with the SST and SSE respectively.

- The modified measure can assume negative values unlike the classical R^2 .
- It measures the balance between prediction bias and the variability of estimates.
- In practice, sophisticated measures based on information theory such as AIC and SIC are employed.

Applied TSA October 26, 2018

So, that is how the adjusted R square is all you can see now is the number of parameters comes into play. As you increase p what happens to R square? R square is independent of p, R square is in fact, getting better as you increase p, but as you increase p the denominator in the adjusted R square is coming down as a result of which the adjusted R square what happens? Decreases because you have 1 minus that; so the adjusted R square is gives you some compromise between the fit that you have on the test training data and the performance that you may expect to see on a fresh data, but we have come across another measured we have talked about another measured that offers is trade of recall aka ike information criteria AIC also does this right. So, today people use AIC with l m for example, if you want to run l m, if you recall it reports r it reports adjusted R square, it reports many other things but now hopefully you have an understanding of what it is.