EXPLAINED GAUSS-MARKOV PROOF: ORDINARY LEAST SQUARES AND B.L.U.E [1]

This document aims to provide a concise and clear proof that the ordinary least squares model is BLUE. BLUE stands for Best, Linear, Unbiased, Estimator. In this example, we will start from back to front. The first thing to do is list the OLS estimator in functional form.

$$y = \beta x + \epsilon$$

This equation represents the population values of x, y, beta, and the error term. Since we can almost never find true population statistics, OLS serves to create an estimator dependent on a sample of that population. Estimate values are denoted with hat symbols on any estimated parameter.

$$\hat{y} = \hat{\beta} x + e$$

Since we know nothing about the error terms, an alternative measure is used.[2] Upon creating predicted values of y and β, we can look at the difference between the predicted y value and the actual y value for that observattion. This difference is known as a residual which denote with an e.

The OLS form can be expressed in matrix notation which will be used throughout the proof **where all matrices are denoted by boldface.**

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{e}$$

ESTIMATOR

This is the simplist part of determining whether OLS is blue. For OLS to be an estimator, it must predict an outcome based on the sample. In this case, $\hat{\boldsymbol{y}}$ and $\hat{\boldsymbol{\beta}}$ are estimator as the represent the predicted values of y and beta given the sample data x.

UNBIASED

In order to prove that OLS in matrix form is unbiased, we want to show that the expected value of $\hat{\boldsymbol{\beta}}$ is equal to the population coefficient of $\boldsymbol{\beta}$. First, we must find what $\hat{\boldsymbol{\beta}}$ is.

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}$$

$$\boldsymbol{\epsilon} = \boldsymbol{y} - \boldsymbol{X\beta}$$

Then if we want to derive OLS we must find the beta value that minimizes the squared residuals (e).

$$\boldsymbol{\epsilon}'\boldsymbol{\epsilon} = (\boldsymbol{y} - \boldsymbol{X\beta})'(\boldsymbol{y} - \boldsymbol{X\beta})$$

Note that the square of a matrix is denoted by the multiplication of the matrix transpose by itself. Our next step is to simply distribute the terms.

$$\boldsymbol{\epsilon}'\boldsymbol{\epsilon} = \boldsymbol{y}'\boldsymbol{y} - \boldsymbol{y}'(\boldsymbol{X\beta}) - (\boldsymbol{X\beta})'\boldsymbol{y} - (\boldsymbol{X\beta})'(\boldsymbol{X\beta})$$

An important simplification that we will use is that $\boldsymbol{y}'(\boldsymbol{X\beta}) = (\boldsymbol{X\beta})'\boldsymbol{y}$. This is found by Taking the transpose of each term and finding that it equals the other (e.g. $(\boldsymbol{y}'(\boldsymbol{XB}))' = (\boldsymbol{XB})'\boldsymbol{y}$ and vis versa). Our equation then simplifies.

[1] Jesse I. Kaczmarski | University of New Mexico, Department of Economics | jikaczmarski@unm.edu

[2] Error terms represent all the unexplained variation in the derivation of y. Since we cannot collect and regress all variables that might influence an observation's y value, we have to capture it for completeness.

$$\boldsymbol{\epsilon}'\boldsymbol{\epsilon} = \boldsymbol{y}'\boldsymbol{y} - 2(\boldsymbol{X}\boldsymbol{\beta})'\boldsymbol{y} - \boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta}$$

Now in order to find the beta that minimizes our subject, we want to take the derivative in respect to beta and set it equal to zero. This will find the point in the function where our slope is equal to zero, also known as a minimum point.

$$\frac{\partial \boldsymbol{\epsilon}'\boldsymbol{\epsilon}}{\partial \boldsymbol{\beta}} = -2\boldsymbol{X}'\boldsymbol{y} - 2\boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}} = 0$$

Since this is equal to zero, we can move terms around and set one part of the equation equal to the other. Note that the 2's cancelled out, and the $\boldsymbol{\beta}$ is on the right side of the $\boldsymbol{X}'\boldsymbol{X}$. This is because we are taking it in respect to $\boldsymbol{\beta}$, not $\boldsymbol{\beta}'$. In addition, $\boldsymbol{\beta}'\boldsymbol{\beta}$ is the same as a squared term, so the 2 appears by taking the derivative of a squared term.

$$\boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}'\boldsymbol{y}$$

Now in order to move one matrix to the other side of the equation, we must take the matrix equivalent of division, which is the inverse of a matrix.

$$(X'X)^{-1}X'X\hat{\beta} = (X'X)^{-1}X'y$$

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$$

This concludes the matrix derivation of $\hat{\boldsymbol{\beta}}$. Now in order to prove that $\hat{\boldsymbol{\beta}}$ is an unbiased estimator, we want to show that the expected value of $\hat{\boldsymbol{\beta}}$ is equal to $\boldsymbol{\beta}$. We can take what we know from above and substitute y into the most recent equation.

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon})$$

We can then distribute the terms accordingly where $(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ is a single term.

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\epsilon}$$

Now in the fist term, we find there is a matrix time it's inverse, which is known to be the identity matrix (a matrix notation for multiplying by 1) which we will denote with **I**.

$$\hat{\boldsymbol{\beta}} = \boldsymbol{I}\boldsymbol{\beta} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\epsilon}$$

When we take the expectation of both side, we find that the expected value of a number (in this case $\boldsymbol{\beta}$) is itself, and we must invoke OLS assumption number 3. This is the zero conditional mean assumption which states that the expected value of an error term dependent on X will be zero; $E(\epsilon_i|x_i) = 0$. This is required for OLS since we know nothing about the error terms. To say there was a value on this conditional would be to say that we know something about the error term. Therefore this entire second term goes to zero. This proves that the estimator for our OLS is unbiased.

$$E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$$

LINEAR

Now that we have proved that our estimator is ubiased, we also proved it was linear. The fact that we can write $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + \boldsymbol{A}\boldsymbol{\epsilon}$ where $\boldsymbol{A} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ proves that it is linear. This is because the **A** term is a liner combination of matrices. Matrix algebra only works in the presence of linearity. Therefore this assumption is proven.

BEST

This final part of the proof is the most difficult, and require many assumptions to be enacted. We must first define our objective. In order to prove that our estimator is the best, we must prove that our estimator is either equal to or less than the variance of all other unbiased estimators. The first step of this process is to find the variance of our current estimator. The variance of an estimator can be found by squaring the error terms. Similar to what was done before, but this time we will use what we learned in the unbiasedness proof to do some substitution. Note that our residuals can be defined as the difference between our predicted and actual values of $\boldsymbol{\beta}$ and the square of this is the variance.

$$var(\hat{\boldsymbol{\beta}}) = (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'$$

Now in order to find what $(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ is, we use the $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\epsilon$ equation we found before, and subtract the $\boldsymbol{\beta}$ onto the left side. Note that we do not let $\boldsymbol{X}'\boldsymbol{\epsilon}$ go to zero here since we are not taking the expectation of it. We now substitute this into the equation.

$$var(\hat{\boldsymbol{\beta}}) = ((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\epsilon})((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\epsilon})'$$

Now the transpose is distributed into the second term. Note that the transpose of an invesre matrix will end up being the same.

$$var(\hat{\boldsymbol{\beta}}) = ((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\epsilon})(\boldsymbol{\epsilon}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1})$$

$$var(\hat{\boldsymbol{\beta}}) = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\epsilon}\boldsymbol{\epsilon}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}$$

Note that we have a $\boldsymbol{\epsilon}\boldsymbol{\epsilon}'$ term in the middle *that is dependent on X*. This allows us to envoke OLS spherical assumption that states $E(\boldsymbol{\epsilon}\boldsymbol{\epsilon}'|\boldsymbol{x}) = \sigma^2\boldsymbol{I}$. By taking this expected value we are left with the following. Also note that the inverse of a matrix times itself is the identity matrix, in this case $(\boldsymbol{X}'\boldsymbol{X})^{-1}(\boldsymbol{X}'\boldsymbol{X}) = \boldsymbol{I}$

$$var(\hat{\boldsymbol{\beta}}) = \sigma^2\boldsymbol{I}[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}]$$

$$var(\hat{\boldsymbol{\beta}}) = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$$

Note that we still do not know the value of σ^2 as it is derived directly from our error term. In this case it will suffice for comparing best estimators. A derivation of the true sample variance can be found in the appendix. Now that we have found the variance for our estimator, we must find the variance for all other linear unbiased estimators. In order to do this we must create a new matrix "C" which is linear (as seen by $\boldsymbol{CX} = \boldsymbol{I}$). In addition, **C** can be expressed as $\boldsymbol{C} = \boldsymbol{D} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ where $\boldsymbol{DX} = 0$. This **C** matrix is unbiased in that **C** can be expressed as $\boldsymbol{D} = \boldsymbol{C} - (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}$ where $\boldsymbol{\beta}_0 = \boldsymbol{CY}$. This states that $\boldsymbol{\beta}_0$ is linear and it is assumed to be unbiased through the same proof we have done before. Now we must derive the variance of this estimator and see how it compares to the variance of $\hat{\boldsymbol{\beta}}$. The set up for this derivation is similar to that above.

$$var(\hat{\boldsymbol{\beta}_0}) = (\boldsymbol{\beta}_0 - \boldsymbol{\beta})(\boldsymbol{\beta}_0 - \boldsymbol{\beta})'$$

Note that if $\boldsymbol{\beta}_0 = \boldsymbol{CY}$ and $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}$.

$$\boldsymbol{\beta}_0 = \boldsymbol{C}(\boldsymbol{X\beta} + \boldsymbol{\epsilon})$$

$$\boldsymbol{\beta}_0 = \boldsymbol{CX\beta} + \boldsymbol{C\epsilon}$$

Now we invoke the fact that $\boldsymbol{CX} = \boldsymbol{I}$.

$$\boldsymbol{\beta}_0 = \boldsymbol{\beta} + \boldsymbol{C\epsilon}$$

$$\boldsymbol{\beta}_0 - \boldsymbol{\beta} = \boldsymbol{C\epsilon}$$

Now that we have found this subject, we can substitute it back into our variance formula and distribute the transpose in the second term.

$$var(\hat{\boldsymbol{\beta}_0}) = (\boldsymbol{C\epsilon})(\boldsymbol{C\epsilon})'$$

$$var(\hat{\boldsymbol{\beta}_0}) = (\boldsymbol{C\epsilon})(\boldsymbol{\epsilon}'\boldsymbol{C}')$$

Remeber from OLS assumption 4, that the expected value of the square of the residuals dependent on X's is $\sigma^2\boldsymbol{I}$. Also note that $\boldsymbol{C} = \boldsymbol{D} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ as defined at the start of this derivation. Our expression gets simplified and substituted.

$$var(\hat{\boldsymbol{\beta}_0}) = \sigma^2\boldsymbol{I}(\boldsymbol{CC}')$$

$$var(\hat{\boldsymbol{\beta}_0}) = \sigma^2(\boldsymbol{D} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}')(\boldsymbol{D} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}')'$$

After distributing the transpose we find the following expression

$$var(\hat{\boldsymbol{\beta}}_0) = \sigma^2(\boldsymbol{D} + (\boldsymbol{X'X})^{-1}\boldsymbol{X'X}(\boldsymbol{X'X})^{-1}\boldsymbol{D'})$$

$$var(\hat{\boldsymbol{\beta}}_0) = \sigma^2(\boldsymbol{DD'} + (\boldsymbol{X'X})^{-1})$$

$$var(\hat{\boldsymbol{\beta}}_0) = \sigma^2(\boldsymbol{DD'}) + \sigma^2(\boldsymbol{X'X})^{-1}$$

Since we know that the variance of $\hat{\boldsymbol{\beta}}$ is $\sigma^2(\boldsymbol{X'X})^{-1}$, then our equation simplifies further.

$$var(\hat{\boldsymbol{\beta}}_0) = \sigma^2(\boldsymbol{DD'}) + var(\hat{\boldsymbol{\beta}})$$

Since a squared term (DD') cannot be negative, we can conclude that the variance of $\hat{\boldsymbol{\beta}}_0$ will be equal to or greater to the variance of $\hat{\boldsymbol{\beta}}$ by a value of $\sigma^2(\boldsymbol{DD'})$. Since it must be non-negative, but it may be zero. This concludes that there are no linear unbiased estimators that are smaller in variance than the OLS estimator. While they may be the same, our OLS will remain BLUE.

APPENDIX A - DERIVING THE SAMPLE VARIANCE

In terms of comparing the OLS estimator we found originally to that of other unbiased linear estimators, it is acceptable to leave the variance in terms of $\sigma^2\boldsymbol{I}$ as we are looking for a magnitude change between $var(\hat{\boldsymbol{\beta}})$ and $var(\hat{\boldsymbol{\beta}}_0)$. It is actually impossible to know this value of σ^2 as it is derived directly from out error terms (remember that $E(\boldsymbol{\epsilon}\boldsymbol{\epsilon}'|\boldsymbol{x}) = \sigma^2\boldsymbol{I}$ via OLS spherical assumption) and we are not supposed to know anything about our error terms. This appendix will derive the sample variance using terms found throughout this document.

Let us assume a new matrix, **M**, that is defined as $\boldsymbol{M} = \boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ where **M** represents a symmetric and idempotent matrix.[3] We can then rewrite our OLS model as follows.

$$\boldsymbol{e} = \boldsymbol{M}\boldsymbol{y}$$

$$\boldsymbol{M}\boldsymbol{y} = (\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon})\boldsymbol{M} = \boldsymbol{M}\boldsymbol{\epsilon}$$

This relationship can seem slightly confusing without the proper explaination. This relation can be seen by substituting M into the original equation for the residuals. The mechanism where $\boldsymbol{M}(\boldsymbol{X}\boldsymbol{\beta})$ goes to 0 is the same mechanism described below.

$$\boldsymbol{e} = (\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}')(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})$$

$$\boldsymbol{e} = (\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} - (\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta})$$

$$\boldsymbol{e} = \boldsymbol{M}\boldsymbol{y} - (\boldsymbol{I} - \boldsymbol{X}\boldsymbol{I}\boldsymbol{\beta})$$

$$\boldsymbol{e} = \boldsymbol{M}\boldsymbol{y} = \boldsymbol{M}\boldsymbol{\epsilon}$$

This allows us to establish a relationship between the error terms and the residuals. Based off of this relationship, we find that $\boldsymbol{e} = \boldsymbol{M}\boldsymbol{\epsilon}$. Therefore the expected value of the squared residuals can be described as follows.

$$E[\boldsymbol{e}'\boldsymbol{e}] = E[\boldsymbol{\epsilon}'\boldsymbol{M}\boldsymbol{\epsilon}]$$

In order to find this expected value, we must take the trace of the righthand side.

$$E[\boldsymbol{e}'\boldsymbol{e}] = E[trace(\boldsymbol{\epsilon}'\boldsymbol{M}\boldsymbol{\epsilon})]$$

The trace of a matrix is the sum of the eigenvalues (i.e. these are known as the diagonal values of a matrix). The trace is only defined for a square matrix. We can expand our previous equation as follows.

$$E[\boldsymbol{e}'\boldsymbol{e}] = E[trace(\boldsymbol{\epsilon}'\boldsymbol{M}\boldsymbol{\epsilon})] = \sigma^2 trace[\boldsymbol{M}]$$

Expanding M, we find that.

$$E[\boldsymbol{e}'\boldsymbol{e}] = \sigma^2 trace[\boldsymbol{I}] - trace[\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}']$$

The resulting value is akin to σ^2 time $(n-k)$ where n is the number of observations and k is th number of estimators being used. Therefore we find that.

$$E[\boldsymbol{e}'\boldsymbol{e}] = \sigma^2(n-k)$$

Now in order to find what our σ^2 parameter equals, the $(n-k)$ term is simply divided to the other side. Since the expected value of a number is itself (note that e'e is a number), then we find our sample variance as.

$$S^2 = \frac{\boldsymbol{e}'\boldsymbol{e}}{(n-k)}$$

This would then convert our variance of our estimator to be the following.

$$var(\hat{\boldsymbol{\beta}}) = S^2(\boldsymbol{X}'\boldsymbol{X})^{-1}, \quad where\, S^2 = \frac{\boldsymbol{e}'\boldsymbol{e}}{(n-k)}$$

[3] An idempotent matrix states that M*M=M. Therefore, you wanted to do the math, you would find that $(\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}')(\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}) = \boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$.