

OLS: Specification, Estimation and Small Sample Results(1) Assumptions of the Model

- (A0) $y = X\beta + \epsilon$ ← Linear Functional Form
 (A1) $\text{rank}(X) = K$ ← Identification
 (A2) $E(\epsilon|X) = 0$ ← Orthogonality (of regressors and disturbance)
 (A3) $E(\epsilon\epsilon' | X) = \sigma^2 I$ ← Sphericity (Homoskedasticity, No Serial Correlation)
- (A2') $\epsilon \sim N(0, \sigma^2 I)$ conditional on X Normality

where y is an $n \times 1$ vector of observations on the left-hand-side variable or dependent variable, X is an $n \times K$ matrix of observations on the K right-hand-side variables or regressors, and ϵ is an $n \times 1$ vector of unobserved disturbances.

Linear Functional Form: This is not as restrictive an assumption as it may appear. To see why imagine a primitive set of data $\{w_i, z_i\}_{i=1}^n$ where w_i is the variable to be "explained" and z_i is an $L \times 1$ vector of explanatory variables. Then if the relationship between w_i and z_i can be written in the form:

$$\underbrace{g(w_i)}_{y_i} = \underbrace{[f_1(z_i) \quad \dots \quad f_K(z_i)]}_{X_i' \quad (1 \times K)} \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_K \end{bmatrix} + \epsilon_i$$

where $g(\cdot)$ and $f_k(\cdot)$ $k=1, \dots, K$ are observable functions then the regression model is linear in the transformed data $\{y_i, X_i\}$ even though it may not be linear in the original or untransformed data $\{w_i, z_i\}$. Also note that the dimension of z_i need not equal the dimension of X_i (ie $K \neq L$ potentially). Some examples follow:

Ex1: Quadratic Model

$$\frac{w_i}{y_i} = \beta_1 + \beta_2 \frac{z_i}{x_{i2}} + \beta_3 \frac{z_i^2}{x_{i3}} + \varepsilon_i$$

Ex2: Log-Linear Model

$$w_i = e^{\beta_1} \left(\prod_{k=2}^K z_{ik}^{\beta_k} \right) e^{\varepsilon_i} \Rightarrow \frac{\ln w_i}{y_i} = \beta_1 + \beta_2 \frac{\ln z_{i2}}{x_{i2}} + \dots + \beta_k \frac{\ln z_{ik}}{x_{ik}} + \varepsilon_i$$

Ex3: Logistic Model

$$w_i = \frac{1}{1 + e^{-(x_i' \beta + \varepsilon_i)}} \Rightarrow \frac{\ln \left(\frac{w_i}{1-w_i} \right)}{y_i} = x_i' \beta + \varepsilon_i$$

Identification: The reason that $\text{rank}(X) = k$ is called an identification assumption is because if $\text{rank}(X) < k$ (i.e. if the columns of X are linearly dependent) then there exists a $k \times 1$ vector $\gamma \neq 0$ such that $X\gamma = 0$. But then we could write the model as

$$Y = X \underbrace{(\beta + c\gamma)}_{\beta(c)} \quad \text{For any } c \in \mathbb{R}$$

Thus, there are an (uncountably) infinite number of models or parameter vectors $\beta(c)$ which are consistent with the data and so the true model (i.e. true β) cannot be identified from the data, $\{y_i, x_{i1}, \dots\}$.

Orthogonality (of the unobserved disturbance and the regressors): This is the most crucial assumption on the list. Clearly, identification is crucial also, but it is much less often the source of econometric angst than is the nonorthogonality problem. The orthogonality assumption can be given a number of economic interpretations. One interpretation that makes it easy to

see how things can go wrong if the assumption is violated is to view the assumption as stating that any potential regressors (sources of variation in y) which have been omitted from the model specification (ie are not in X) are uncorrelated (at every lead and lag) with the regressors that have been included in the specification of the model (X). To see why such a lack of correlation is important, consider the following example:

An environmental activist wants to encourage the City of Boston to undertake a tree-planting project in inner-city neighborhoods and so wants to estimate some measure of the value people place on green space to convince the mayor that this is a worthwhile project. To get such a measure he gathers data on the sale price of homes in Roxbury (an inner-city neighborhood) and Weston (a posh suburb of Boston) and regresses the sale price on a measure of the density of trees surrounding the house. Low and behold he finds that trees can explain much of the difference in housing prices between Weston and Roxbury. Does this empirical procedure pass the orthogonality assumption sniff test? Note quite. Some of the things that might be correlated with housing prices: size + quality of house, size of lot, quality of schools, and crime level to name a few would also tend to be highly correlated with the greenspace differential between Weston and Roxbury and so the estimated $\hat{\beta}$ on trees is picking up not just the tree effect but also the effect of these other factors. Thus, in this case, the violation of the orthogonality condition leads us to overestimate the value of trees.

Note: The above example in no way reflects my views on trees or environmental activists.

Sphericity: If one thinks of ϵ as noise, mucking up the relationship between y and X , then this assumption says two things about the nature of the noise:

Homoskedasticity: $E(\epsilon^2|X) = \sigma^2 \forall i \Rightarrow$ The amount of noise in an observation, as captured by the variance of the disturbance, is the same across

observations. If it were not ($E(\epsilon_i^2 | x) = \sigma_i^2 \neq \text{constant}$) then we would probably want to give less weight to observations with more noise (high σ_i^2). Since OLS assumes $\sigma_i^2 = \sigma^2 \forall i$, it treats all observations equally. If σ_i^2 varies across observations and we have an auxiliary model for σ_i^2 , then we will be able to do better than OLS by assigning observations weights that are inversely proportional to σ_i^2 .

No Serial Correlation: $E(\epsilon_i \epsilon_j | x) = 0 \quad i \neq j$. This assumption states that the disturbance from one observation tells you nothing about the disturbance for other observations. For example, suppose we have time series data and that $\epsilon_t = \rho \epsilon_{t-1} + u_t \quad |\rho| < 1$ so that $E(\epsilon_t \epsilon_{t+1} | x) = \rho \sigma_u^2 \neq 0$. The true noise in the t th observation is u_t not ϵ_t , and so we would be doing the observed data an injustice by treating ϵ_t as if it were all noise.

To make the discussion of nonsphericity more intuitive observe that

$$\hat{\beta}_{OLS} = \underset{\beta}{\operatorname{argmin}} \sum_i (y_i - x_i' \beta)^2 = \underset{\beta}{\operatorname{argmin}} \sum_i \epsilon_i^2$$

OLS chooses $\hat{\beta}$ to minimize the (equally weighted) sum of squared disturbances, or total noise. This makes intuitive sense if the ϵ_i are truly just white noise. In that case we want to choose a $\hat{\beta}$ which gives maximum explanatory power to the regressors, x_i , and hence minimum explanatory power to the ϵ_i .

Ex1: Heteroskedasticity

Suppose that $E(\epsilon_i^2 | x) = \sigma_i^2$. Then we would probably prefer an estimator that downweights observations with high σ_i^2 . Specifically we would want:

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \sum_i (\epsilon_i^2 / \sigma_i^2) = \underset{\beta}{\operatorname{argmin}} \sum_i \tilde{\epsilon}_i^2$$

where $\tilde{\epsilon}_i = \epsilon_i / \sigma_i$. By downweighting ϵ_i w/ big σ_i^2 we will improve the precision w/ which we estimate β .

Ex2. AR(1) autocorrelation $\epsilon_t = \rho \epsilon_{t-1} + u_t$ $E(u_t^2 | x) = \sigma^2$, $E(u_t u_s | x) = 0$
 In this case the true "disturbance" is u_t and since u_t is homoskedastic we would probably prefer an estimator that solves:

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \sum_t u_t^2$$

over $\hat{\beta}_{OLS}$.

Normality: This assumption allows us to place confidence intervals around our estimator and to conduct hypothesis tests. (It is not crucial for large samples since we know from our list of CLT's that everything of interest will be normally distributed as $n \rightarrow \infty$.) (We will see this in handouts #3 and #6).

(2) Derivation and Interpretation of $\hat{\beta}_{OLS}$

$$\begin{aligned} S_n(\beta) &= \sum_t \epsilon_t^2 = \epsilon' \epsilon = (Y - X\beta)'(Y - X\beta) = Y'Y - Y'X\beta - \beta'X'Y + \beta'X'X\beta \\ &= Y'Y - 2(X'Y)'\beta + \beta'(X'X)\beta \end{aligned}$$

$$\min_{\beta} S_n(\beta) = Y'Y - 2(X'Y)'\beta + \beta'(X'X)\beta$$

(1A) H01 \Rightarrow FOC: $\frac{\partial S_n(\beta)}{\partial \beta} = -2X'Y + 2(X'X)\beta = 0 \Rightarrow \boxed{(X'X)\beta = X'Y}$ Normal Equations
 (1) and (3)

Since $\operatorname{rank}(X'X) = \operatorname{rank}(X) = K$, $(X'X)$ is invertible. Thus, we have

$$\boxed{\hat{\beta}_{OLS} = (X'X)^{-1} X'Y}$$

(1A) H01 \Rightarrow SOC: $\frac{\partial^2 S_n(\beta)}{\partial \beta' \partial \beta} = (X'X) \leftarrow$ PD by (1B) H01 $\Rightarrow \hat{\beta}$ minimizes $S_n(\beta)$ as desired // (2)

An Intuitive (MM) Interpretation of $\hat{\beta}_{OLS}$: Assume that the data is iid and that $E X_i X_i'$ is invertible. Next, observe that $E X_i \epsilon_i = E E(X_i \epsilon_i | X_i) = E X_i E(\epsilon_i | X_i) = 0$ by (A2). Thus, we have

$$E X_i y_i = E X_i (X_i' \beta + \varepsilon_i) = (E X_i X_i') \beta + E X_i \varepsilon_i \Rightarrow \boxed{\beta = (E X_i X_i')^{-1} E X_i y_i}$$

Now notice that using our multiplication rules for partitioned matrices we can write:

$$\hat{\beta} = (X'X)^{-1} X'Y = \begin{bmatrix} X_1' \\ \vdots \\ X_n' \end{bmatrix}^{-1} \begin{bmatrix} X_1' \\ \vdots \\ X_n' \end{bmatrix} \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} X_1 \cdots X_n \\ \vdots \\ X_n \end{bmatrix}^{-1} \begin{bmatrix} X_1 \cdots X_n \\ \vdots \\ X_n \end{bmatrix} \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$$

$$= (\sum_i X_i X_i')^{-1} (\sum_i X_i y_i) = \boxed{(\frac{1}{n} \sum_i X_i X_i')^{-1} (\frac{1}{n} \sum_i X_i y_i)}$$

Thus, we can interpret $\hat{\beta}$ as a sample analog of the true β , that is, as a method of moments estimator of β .

Decomposing y : Let $\hat{y} = X\hat{\beta}$ be a linear prediction of y given $\hat{\beta}$ and let $\hat{\varepsilon} = y - \hat{y}$ the associated residual. Observe that

$\begin{pmatrix} P & Q \\ n \times n & n \times n \end{pmatrix}$

$$\hat{y} = X (X'X)^{-1} X' y = P y \text{ implying that } \hat{\varepsilon} = (I - P) y = Q y.$$

Recall from Section (5) of H01 that P and Q are symmetric, idempotent projection matrices w/ $PX = X$, $QX = 0$ and $PQ = 0$. Gathering up all our notation we can decompose y in all of the following equivalent ways:

$$y = X\hat{\beta} + \hat{\varepsilon} \quad y = \hat{y} + \hat{\varepsilon} \quad \boxed{y = Py + Qy} \quad y = Py + Q\varepsilon$$

where $Qy = Q\varepsilon$ since $y = X\beta + \varepsilon$ and $QX = 0$. Thus, OLS decomposes y into two orthogonal components, one of which is contained in the k -dimensional subspace of \mathbb{R}^n spanned by the columns of X (this component is $\hat{y} = X\hat{\beta} = Py$) and the other contained in the orthogonal complement to this subspace (this component is $\hat{\varepsilon} = Qy = Q\varepsilon$). Another interpretation of OLS, therefore, is that it takes the probabilistic orthogonality of X and ε ($E \varepsilon X_i = 0$) and translate it into the Euclidean orthogonality of X and $\hat{\varepsilon}$ ($X'\hat{\varepsilon} = 0$).

In Greene:
 $X(X'X)^{-1}X' = P$
 $I - X(X'X)^{-1}X' = M$
 $\hat{y} = PY$
 $\hat{\varepsilon} = MY$
 Steve calls Q Greene calls M
 what calls M

(3) Partitioned Regression:

Let X_1 and X_2 partition the columns of X so that $X = [X_1 \ X_2]$ and $Y = X_1 \beta_1 + X_2 \beta_2 + \varepsilon$. Further let $Q_1 = I - X_1(X_1'X_1)^{-1}X_1'$. Then we have the following theorem:

(T1) Theorem: All of the following estimators of β_2 are numerically identical.

- $\hat{\beta}_2$ obtained by doing OLS of Y on X_1 and X_2
- $\tilde{\beta}_2$ obtained by doing OLS of \tilde{Y} on \tilde{X}_2
- $\bar{\beta}_2$ obtained by doing OLS of Y on \hat{X}_2

where \tilde{Y} is the residual ^{vector} from the regression of Y on X_1 and similarly \tilde{X}_2 are the residual _{matrix} from the regressions of the columns of X_2 on X_1 .

Proof:

From the above discussion on decomposing Y we know that $\tilde{Y} = Q_1 Y$ and $\tilde{X}_2 = Q_1 X_2$. The equivalence of the estimators then follows from (E4) of H01 since

$$\hat{\beta}_2 = (X_2' Q_1 X_2)^{-1} X_2' Q_1 Y \quad (\text{From (E4) H01})$$

$$\tilde{\beta}_2 = (X_2' Q_1' Q_1 X_2)^{-1} X_2' Q_1' Q_1 Y = (X_2' Q_1 X_2)^{-1} X_2' Q_1 Y = \hat{\beta}_2$$

$$\bar{\beta}_2 = (X_2' Q_1' Q_1 X_2)^{-1} X_2' Q_1' Y = (X_2' Q_1 X_2)^{-1} X_2' Q_1 Y = \hat{\beta}_2$$

where we have used the symmetry and idempotency of Q_1 . ||

(C1) Corollary: Regressing Y on a constant and a set, X_2 , of $K-1$ other regressors produces the same slope coefficients as (1) regressing deviations of Y from its mean on deviations of X_2 from its mean or (2) regressing Y on deviations of X_2 from its mean.

Note: Only procedure (2) yields mean zero residuals

(T2) Theorem: IF $X_1'X_2 = 0$ (the columns of X_1 are orthogonal to the columns of X_2) then the estimators of β_2 from (T1) are all numerically equivalent to the estimator obtained by doing OLS of Y on X_2 :

Proof:

$$X_2'Q_1 = X_2'[I - X_1(X_1'X_1)^{-1}X_1'] = X_2' - \frac{X_2'X_1(X_1'X_1)^{-1}X_1'}{0} = X_2'$$

$$\Rightarrow \hat{\beta}_2 = (X_2'Q_1X_2)^{-1}X_2'Q_1Y = (X_2'X_2)^{-1}X_2'Y \quad ||$$

(4) Statistical Properties of OLS

(T3) Theorem: Under assumptions (A0) to (A3) we have:

(1) $E(\hat{\beta}|X) = \beta$ and $V(\hat{\beta}|X) = \sigma^2(X'X)^{-1}$

(2) $\hat{\beta}$ is the unique BLUE of $\beta \leftarrow$ Gauss-Markov Theorem (GMT)

(3) $S^2 \equiv \hat{\epsilon}'\hat{\epsilon}/(n-k)$ is an unbiased estimator of σ^2

Proof (1)

$$E(Y|X) = X\beta + E(\epsilon|X) = X\beta \quad \text{by (A2)}$$

$$\Rightarrow V(Y|X) = E(\epsilon\epsilon'|X) = \sigma^2 I \quad \text{by (A3)}$$

$$(1) E(\hat{\beta}|X) = (X'X)^{-1}X'E(Y|X) = (X'X)^{-1}X'X\beta = \beta$$

$$V(\hat{\beta}|X) = (X'X)^{-1}X' \frac{V(Y|X)}{\sigma^2 I} X(X'X)^{-1} = \sigma^2 (X'X)^{-1}X'X(X'X)^{-1} = \sigma^2 (X'X)^{-1}$$

Proof (2) (GMT):

Let $\tilde{\beta} = AY$ be another linear estimator of β .

$$E(\tilde{\beta}|X) = AE(Y|X) = AX\beta = \beta \quad \forall \beta \Leftrightarrow AX = I$$

$$V(\tilde{\beta}|X) = AV(Y|X)A' = \sigma^2 AA'$$

wlog assume $\sigma^2 = 1$. Now let $A^* = (X'X)^{-1}X'$ and observe that

$$V(\tilde{\beta} - \hat{\beta}|X) = V(AY - A^*Y|X) = V[(A - A^*)Y|X] = (A - A^*) \overbrace{V(Y|X)}^I (A - A^*)'$$

$$= (A - A^*)(A - A^*)' = AA' - AA^* - A^*A' + A^*A^*$$

$$= AA' - AX(X'X)^{-1} - (X'X)^{-1}X'A' + (X'X)^{-1}X'X(X'X)^{-1}$$

using $AX = I \rightarrow$

$$= AA' - (X'X)^{-1} - (X'X)^{-1} + (X'X)^{-1} = AA' - (X'X)^{-1}$$

$$= V(\tilde{\beta}|X) - V(\hat{\beta}|X)$$

Since $V(\hat{\beta} - \beta | x)$ is a variance matrix it is PSD. Thus, $V(\hat{\beta} | x) - V(\beta | x)$ is PSD also and so $\hat{\beta}$ is BLUE. For uniqueness suppose that \exists linear unbiased $\tilde{\beta}$ w/ $V(\tilde{\beta} | x) = V(\hat{\beta} | x)$. Then $\forall c \in \mathbb{R}^k, c \neq 0$

$$0 = c' [V(\tilde{\beta} | x) - V(\hat{\beta} | x)] c = c' V(\tilde{\beta} - \hat{\beta} | x) c = V(c'(\tilde{\beta} - \hat{\beta}) | x)$$

But $V(c'(\tilde{\beta} - \hat{\beta}) | x) = 0 \quad \forall c \neq 0 \Leftrightarrow c'(\tilde{\beta} - \hat{\beta})$ is constant $\forall c \neq 0$
 And $c'(\tilde{\beta} - \hat{\beta})$ is constant $\forall c \neq 0 \Leftrightarrow \tilde{\beta} - \hat{\beta} = m \in \mathbb{R}^k$, in which case.
 $m = E(m | x) = E(\tilde{\beta} | x) - E(\hat{\beta} | x) = \beta - \beta = 0 \quad //$

Proof (3)

$$\hat{\beta}' \hat{\beta} = (Qy)'(Qy) = (Q\varepsilon)'(Q\varepsilon) = \varepsilon' Q \varepsilon$$

scalar

$$\begin{aligned} \Rightarrow E(\hat{\beta}' \hat{\beta} | x) &= E(\varepsilon' Q \varepsilon | x) \\ &= E(\text{tr}(\varepsilon' Q \varepsilon) | x) && \text{since } \text{tr}(c) = c \quad \forall c \in \mathbb{R} \\ &= E(\text{tr}(\varepsilon \varepsilon' Q) | x) && \text{since } \text{tr}(AB) = \text{tr}(BA) \\ &= \text{tr}(E(\varepsilon \varepsilon' Q | x)) && \text{Since } \text{tr}(\cdot) \text{ is a linear operator (it is} \\ &= \text{tr}(E(\varepsilon \varepsilon' | x) Q) && \text{just a sum operator and can always} \\ &= \text{tr}(\sigma^2 Q) && \text{bring } E(\cdot) \text{ inside a sum)} \\ &= \sigma^2 \text{tr}(Q) && \text{Since } \text{tr}(cA) = c \text{tr}(A) \quad c \in \mathbb{R} \\ &= \sigma^2 \text{tr}[I_n - x(x'x)^{-1}x'] \\ &= \sigma^2 [\text{tr}(I_n) - \text{tr}(x(x'x)^{-1}x')] && \text{Since } \text{tr}(A+B) = \text{tr}(A) + \text{tr}(B) \\ &= \sigma^2 [n - \text{tr}(x'x(x'x)^{-1})] && \text{Since } \text{tr}(AB) = \text{tr}(BA) \\ &= \sigma^2 [n - \text{tr}(I_k)] \\ &= \sigma^2 (n - k) \quad // \end{aligned}$$

$$E S^2 = E \frac{\hat{\beta}' \hat{\beta}}{n-k} = \sigma^2$$

$$\text{tr}(Q) = n - k = \text{rank}(Q)$$

(for sym idempotent matrix rank = trace)

(T4) Theorem: Under assumptions (A0) to (A2') we can add to (T3) the following

- (1) $\hat{\beta} \sim N(\beta, \sigma^2(x'x)^{-1})$
- (2) $(n-k) S^2 / \sigma^2 \sim \chi^2_{n-k}$
- (3) $\hat{\beta}, S^2$ independent
- (4) $(\hat{\beta}_j - \beta_j) / SE_j \sim t_{n-k} \quad \forall j = 1, \dots, k$ where $SE_j = \sqrt{S^2 [(x'x)^{-1}]_{jj}}$
- (5) $\hat{\beta}_{OLS} = \hat{\beta}_{MLE}$ and achieves the Cramér-Rao Lower Bound so that $\hat{\beta}_{OLS}$ is BLUE. (best unbiased, not only linear)

Proof

(1) Follows from (T4) HOZ and result (1) of (T3) on this handout since $\hat{\beta}$ is a linear transformation of y and hence also of ϵ .

(2) $(n-k)S^2/\sigma^2 = \hat{\epsilon}'\hat{\epsilon}/\sigma^2 = \epsilon'Q\epsilon/\sigma^2 = (\epsilon/\sigma)'Q(\epsilon/\sigma) \sim \chi^2_{n-k}$ by (T5) of HOZ since $\epsilon/\sigma \sim N(0, I)$ and since Q is symmetric, idempotent w/ $\text{rank}(Q) = n-k$ ← derived in proof of (T3) of this handout

(3) Since $\hat{\beta}$ and $\hat{\epsilon}$ are both normally distributed, we can show that they are independent by showing that $\text{Cov}(\hat{\beta}, \hat{\epsilon}|X) = 0$. Further, since S^2 is a function of $\hat{\epsilon}$ ($S^2 = \hat{\epsilon}'\hat{\epsilon}/(n-k)$) it would also follow that $\hat{\beta}$ and S^2 are independent (since: X, Y independent $\Rightarrow f(x), g(y)$ independent $\forall f(\cdot), g(\cdot)$)
 Now observe that $E(\hat{\epsilon}|X) = E(y - X\hat{\beta}|X) = E(y|X) - X E(\hat{\beta}|X) = X\beta - X\beta = 0$ and that $\hat{\beta} - \beta = (X'X)^{-1}X'y - \beta = (X'X)^{-1}X'(X\beta + \epsilon) - \beta = (X'X)^{-1}X'\epsilon$. Thus, we have:

$$\begin{aligned} \text{Cov}(\hat{\beta}, \hat{\epsilon}|X) &= E[(\hat{\beta} - \beta)\hat{\epsilon}'|X] = E[(X'X)^{-1}X'\epsilon(Q\epsilon)'|X] \\ &= (X'X)^{-1}X'E(\epsilon\epsilon'|X)Q = \sigma^2(X'X)^{-1}X'Q = 0 \end{aligned}$$

(4) Observe that since $\hat{\beta} \sim N(\beta, \sigma^2(X'X)^{-1})$, we have

$$\begin{aligned} Z_j &= \frac{\hat{\beta}_j - \beta_j}{\sqrt{\sigma^2[(X'X)^{-1}]_{jj}}} \sim N(0,1) \quad \forall j=1, \dots, k \text{ and from (2) } W = \frac{(n-k)S^2}{\sigma^2} \sim \chi^2_{n-k} \\ \Rightarrow t &\equiv \frac{\hat{\beta}_j - \beta_j}{\sqrt{S^2[(X'X)^{-1}]_{jj}}} = \frac{\hat{\beta}_j - \beta_j}{\sqrt{\sigma^2[(X'X)^{-1}]_{jj}}} / \sqrt{\frac{S^2}{\sigma^2}} = \frac{Z_j}{\sqrt{W/(n-k)}} \stackrel{\substack{N(0,1) \\ \sqrt{\chi^2_{n-k}/(n-k)} \\ \sim t_{n-k}}}{\text{by (T1) HOZ}} \end{aligned}$$

since Z_j, W independent by (3)

(5) Observe that the likelihood and log-likelihood functions for (β, σ^2) are

$$L(\beta, \sigma^2) = (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left[-(y - X\beta)'(y - X\beta)/2\sigma^2\right]$$

$$\ell(\beta, \sigma^2) = k - \frac{n}{2} \ln \sigma^2 - \frac{1}{2\sigma^2} \underbrace{(y - X\beta)'(y - X\beta)}_{S_n(\beta)}$$

$$(\hat{\beta}, \hat{\sigma}^2)_{MLE} = \underset{(\beta, \sigma^2)}{\text{argmax}} \ell(\beta, \sigma^2) \quad S_n(\beta)$$

FOC's:

$$\frac{\partial \ell}{\partial \beta} = \frac{-1}{2\sigma^2} \frac{\partial S_n(\beta)}{\partial \beta} = \frac{-2(X'Y - X'X\beta)}{-2\sigma^2} = 0 \Rightarrow \hat{\beta}_{MLE} = (X'X)^{-1}X'Y = \hat{\beta}_{OLS}$$

$$\frac{\partial \ell}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{(Y - X\hat{\beta})(Y - X\hat{\beta})'}{2\sigma^4} = 0 \Rightarrow \hat{\sigma}_{MLE}^2 = \frac{(Y - X\hat{\beta}_{MLE})'(Y - X\hat{\beta}_{MLE})}{n} = \frac{\hat{\varepsilon}'\hat{\varepsilon}}{n}$$

$$\hat{\sigma}_{MLE}^2 = \frac{\hat{\varepsilon}'\hat{\varepsilon}}{n} = \frac{n-k}{n} S^2 \neq S^2$$

Solving for the CRLB:

$$\frac{\partial^2 \ell}{\partial \beta \partial \beta} = -\frac{(X'X)}{\sigma^2} \Rightarrow E\left[\frac{\partial^2 \ell}{\partial \beta \partial \beta} \mid X\right] = -\frac{(X'X)}{\sigma^2}$$

$$\frac{\partial^2 \ell}{\partial (\sigma^2)^2} = \frac{n}{2\sigma^4} - \frac{\varepsilon'\varepsilon}{\sigma^6} \Rightarrow E\left[\frac{\partial^2 \ell}{\partial (\sigma^2)^2} \mid X\right] = \frac{n}{2\sigma^4} - \frac{n\sigma^2}{\sigma^6} = -\frac{n}{2\sigma^4}$$

$$\frac{\partial^2 \ell}{\partial \sigma^2 \partial \beta} = -\frac{(X'Y - X'X\beta)}{\sigma^4} \Rightarrow E\left[\frac{\partial^2 \ell}{\partial \sigma^2 \partial \beta} \mid X\right] = -\frac{X'(E(Y|X) - X\beta)}{\sigma^4} = 0$$

$$\begin{aligned} \text{CRLB}\left(\begin{matrix} \beta \\ \sigma^2 \end{matrix}\right) &= [I\left(\begin{matrix} \beta \\ \sigma^2 \end{matrix}\right)]^{-1} = -\left[\begin{array}{cc} E\left[\frac{\partial^2 \ell}{\partial \beta \partial \beta} \mid X\right] & E\left[\frac{\partial^2 \ell}{\partial \sigma^2 \partial \beta} \mid X\right] \\ E\left[\frac{\partial^2 \ell}{\partial \sigma^2 \partial \beta} \mid X\right] & E\left[\frac{\partial^2 \ell}{\partial (\sigma^2)^2} \mid X\right] \end{array} \right]^{-1} \\ &= -\left[\begin{array}{cc} -\frac{(X'X)}{\sigma^2} & 0 \\ 0 & -\frac{n}{2\sigma^4} \end{array} \right]^{-1} = \left[\begin{array}{cc} \sigma^2(X'X)^{-1} & 0 \\ 0 & \frac{2\sigma^4}{n} \end{array} \right] \end{aligned}$$

$$\Rightarrow \text{CRLB}(\beta) = \sigma^2(X'X)^{-1} = V(\hat{\beta}_{OLS} \mid X) \quad \parallel$$

$$E\hat{\sigma}_{MLE}^2 = \frac{n-k}{n}\sigma^2 \neq \sigma^2 \text{ biased}$$

$$\hat{\sigma}_{MLE}^2 \sim \frac{\sigma^2 \chi^2_{n-k}}{n}$$

$$V(\hat{\sigma}_{MLE}^2) = \frac{2(n-k)\sigma^4}{n^2}$$

does not achieve CRLB
which implies that MLE is not efficient