Lecture 3: Ordinary Least Squares

Zhentao Shi

September 20, 2018

Notation: y_i is a scalar, and x_i is a $K \times 1$ vector. Y is an $n \times 1$ vector, and X is an $n \times K$ matrix.

1 Algebra of Least Squares

1.1 OLS estimator

As we have learned from the linear project model, the parameter β

$$\begin{aligned} y_i &= x_i'\beta + e_i \\ E[x_i e_i] &= 0 \end{aligned}$$

can be written as $\beta = \left(E\left[x_i x_i'\right]\right)^{-1} E\left[x_i y_i\right]$.

While population is something imaginary, in reality we possess a sample of n observations. We thus replace the population mean $E\left[\cdot\right]$ by the sample mean, and the resulting estimator is

$$\widehat{\beta} = \left(\frac{1}{n}\sum_{i=1}^{n} x_i x_i'\right)^{-1} \frac{1}{n}\sum_{i=1}^{n} x_i y_i = \left(X'X\right)^{-1} X'y.$$

This is one way to motivate the OLS estimator.

```
In [1]: n = 100
        beta0 = c(1.0, 1.0, 0.0)
        X = cbind(rnorm(n), rpois(n, 3) )
        e = rlogis(n) # the error term does not have to be normally distributed

        y = cbind(1, X ) %*% beta0 + e # generate data
        # in reality, we observe y and X but not e and beta0
```

Alternatively, we can derive the OLS estimator from minimizing the sum of squared residuals

$$Q\left(\beta\right) = \sum_{i=1}^{n}\left(y_i - x_i'\beta\right)^2 = \left(Y - X\beta\right)'\left(Y - X\beta\right).$$

By the first-order condition

$$\frac{\partial}{\partial\beta} Q\left(\beta\right) = -2X'\left(Y - X\beta\right),$$

the optimality condition gives exactly the same $\widehat{\beta}$. Moreover, the second-order condition

$$\frac{\partial^2}{\partial\beta\partial\beta'}Q(\beta) = 2X'X$$

shows that $Q(\beta)$ is convex in β. ($Q(\beta)$ is strictly convex in β if $X'X$ is positive definite.)

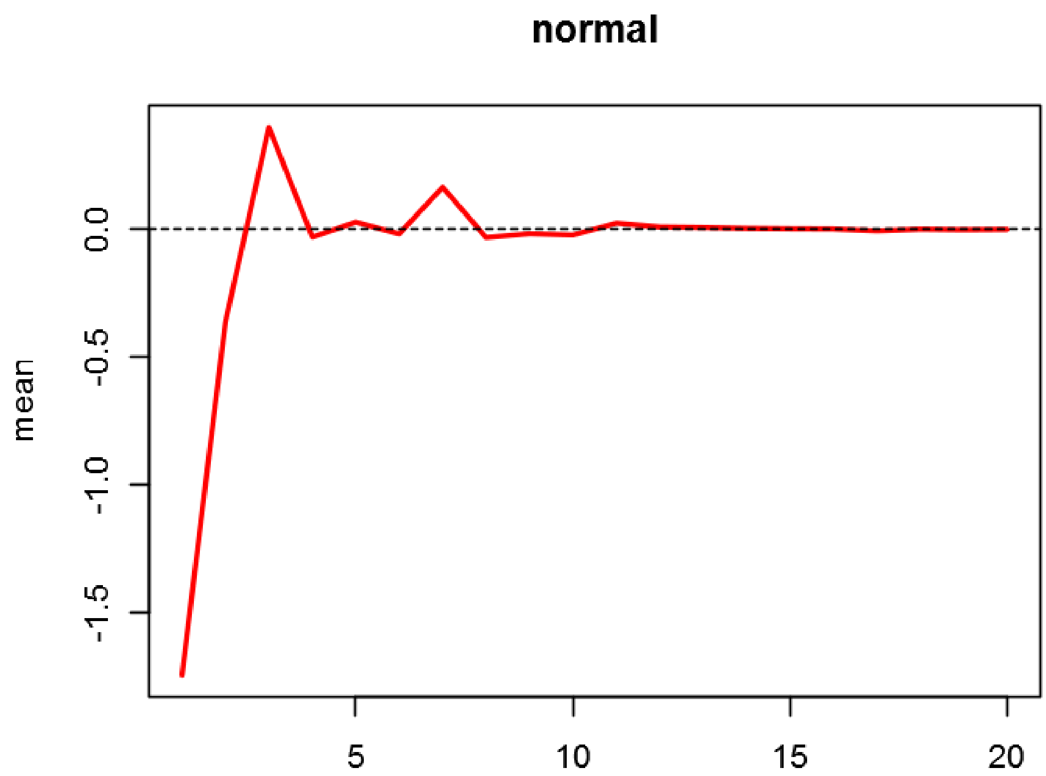
```
In [2]: reg1 = lm( y ~ X ) # OLS regression
        print(reg1)

        X1 = cbind(1, X) # the first column of X is a constant
        bhat = solve(t(X1)%*%X1, t(X1) %*% y )
        print(bhat)
```

```
Call:
lm(formula = y ~ X)

Coefficients:
(Intercept)           X1           X2
    0.84500      0.77722     -0.01699

            [,1]
[1,]  0.84500343
[2,]  0.77722133
[3,] -0.01698568
```

Here we introduce some definitions and properties in OLS estimation.

- Fitted value: $\widehat{Y} = X\widehat{\beta}$.
- Projector: $P_X = X(X'X)^{-1}X$; Annihilator: $M_X = I_n - P_X$.
- $P_X M_X = M_X P_X = 0$.
- If $AA = A$, we call it an idempotent matrix. Both P_X and M_X are idempotent.
- Residual: $\widehat{e} = Y - \widehat{Y} = Y - X\widehat{\beta} = M_X Y = M_X(X\beta + e) = M_X e$.
- $X'\widehat{e} = XM_X e = 0$.
- $\frac{1}{n}\sum_{i=1}^{n}\widehat{e}_i = 0$ if x_i contains a constant.

```
In [3]: yhat = predict( reg1, data  = X ) # predicted value from the OLS regression
        matplot( x = X[,1], y = cbind(y, yhat), pch = 1:2, xlab = "x", ylab = "y") # a graph bet

        library(repr)
        options(repr.plot.width=6, repr.plot.height=5)
        legend(x = 1.2, y = -2, pch = 1:2, col = 1:2, legend = c("y", "predicted"))
```




```
In [4]: # check the orthogonality of ehat and X1
        ehat = y - X1 %*% bhat
        print( t(X1) %*% ehat )

             [,1]
[1,] 4.218847e-15
[2,] 2.337195e-14
[3,] 0.000000e+00
```

```
In [5]: cat("The mean of the residual is ", mean(ehat), "and the sum is", sum(ehat), "\nBut the

The mean of the residual is  3.008878e-17 and the sum is 3.01148e-15
But the mean of the true error term is -0.18831
```

Real Data Example

We check the relationship between *health status* and three control variables: *the number of doctor visits*, *the number of children in the household*, and *access to health care*.

```
In [6]: library(Ecdat, quietly = TRUE, warn.conflicts = FALSE)

        data(Doctor)
        head(Doctor) # display the data structure
```

```
Attaching package: 'Ecfun'

The following object is masked from 'package:base':

    sign
```

doctor	children	access	health
0	1	0.50	0.495
1	3	0.17	0.520
0	4	0.42	-1.227
0	2	0.33	-1.524
11	1	0.67	0.173
3	1	0.25	-0.905

```
In [7]: reg = lm(health ~ doctor + children + access, data = Doctor)
        print(reg)
```

```
Call:
lm(formula = health ~ doctor + children + access, data = Doctor)

Coefficients:
(Intercept)       doctor     children       access
   -0.02810      0.12059      0.03323     -0.63320
```

```
In [8]: summary(reg)
```

```
Call:
lm(formula = health ~ doctor + children + access, data = Doctor)

Residuals:
    Min      1Q  Median      3Q     Max
-3.3370 -1.0085 -0.3261  0.6938  6.1266

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.02810    0.18281  -0.154    0.878
doctor       0.12059    0.01884   6.399 3.71e-10 ***
```

```
children      0.03323    0.04771   0.697    0.486
access       -0.63320    0.33724  -1.878    0.061 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.378 on 481 degrees of freedom
Multiple R-squared:  0.08221,Adjusted R-squared:  0.07649
F-statistic: 14.36 on 3 and 481 DF,  p-value: 5.628e-09
```

1.2 Goodness of Fit

The so-called R-square is the most popular measure of goodness-of-fit in the linear regression. R-square is well defined only when a constant is included in the regressors. Let $M_\iota = I_n - \frac{1}{n}\iota\iota'$, where ι is an $n \times 1$ vector of 1's. M_ι is the *demeaner*, in the sense that $M_\iota (z_1, \ldots, z_n)' = (z_1 - \bar{z}, \ldots, z_n - \bar{z})'$, where $\bar{z} = \frac{1}{n}\sum_{i=1}^n z_i$. For any X, we can decompose $Y = P_X Y + M_X Y = \widehat{Y} + \widehat{e}$. The total variation is

$$Y' M_\iota Y = \left(\widehat{Y} + \widehat{e}\right)' M_\iota \left(\widehat{Y} + \widehat{e}\right) = \widehat{Y}' M_\iota \widehat{Y} + 2\widehat{Y}' M_\iota \widehat{e} + \widehat{e}' M_\iota \widehat{e} = \widehat{Y}' M_\iota \widehat{Y} + \widehat{e}'\widehat{e}$$

where the last equality follows by $M_\iota \widehat{e} = \widehat{e}$ as $\frac{1}{n}\sum_{i=1}^n \widehat{e}_i = 0$, and $\widehat{Y}'\widehat{e} = Y' P_X M_X e = 0$. R-square is defined as $\widehat{Y}' M_\iota \widehat{Y} / Y' M_\iota Y$.

1.3 Frish-Waugh-Lovell Theorem

The FWL theorem is an algebraic fact about the formula of a subvector of the OLS estimator. To derive the FWL theorem We need to use the inverse of partitioned matrix. For a positive definite symmetric matrix $A = \begin{pmatrix} A_{11} & A_{12} \\ A_{12}' & A_{22} \end{pmatrix}$, the inverse can be written as

$$A^{-1} = \begin{pmatrix} \left(A_{11} - A_{12}A_{22}^{-1}A_{12}'\right)^{-1} & -\left(A_{11} - A_{12}A_{22}^{-1}A_{12}'\right)^{-1} A_{12}A_{22}^{-1} \\ -A_{22}^{-1}A_{12}'\left(A_{11} - A_{12}A_{22}^{-1}A_{12}'\right)^{-1} & \left(A_{22} - A_{12}'A_{11}^{-1}A_{12}\right)^{-1} \end{pmatrix}.$$

In our context of OLS estimator,

$$\begin{aligned} \widehat{\beta} = \begin{pmatrix} \widehat{\beta}_1 \\ \widehat{\beta}_2 \end{pmatrix} &= \begin{pmatrix} X_1'X_1 & X_1'X_2 \\ X_2'X_1 & X_2'X_2 \end{pmatrix}^{-1} \begin{pmatrix} X_1 y \\ X_2 y \end{pmatrix} \\ &= \begin{pmatrix} \left(X_1 M_{X_2}' X_1\right)^{-1} & -\left(X_1 M_{X_2}' X_1\right)^{-1} X_1' X_2 \left(X_2'X_2\right)^{-1} \\ \cdot & \cdot \end{pmatrix} \begin{pmatrix} X_1 y \\ X_2 y \end{pmatrix}. \end{aligned}$$

The subvector

$$\begin{aligned} \widehat{\beta}_1 &= \left(X_1 M_{X_2}' X_1\right)^{-1} X_1 y - \left(X_1 M_{X_2}' X_1\right)^{-1} X_1' X_2 \left(X_2' X_2\right)^{-1} X_2 y \\ &= \left(X_1 M_{X_2}' X_1\right)^{-1} \left(X_1 y - X_1' P_{X_2} y\right) \\ &= \left(X_1 M_{X_2}' X_1\right)^{-1} X_1 M_{X_2} y. \end{aligned}$$

Similar derivation can also be carried out in the population linear projection. See Hansen's Chapter 2.21-23.

```
In [9]: X2 = X1[,1:2]
        PX2 = X2 %*% solve( t(X2) %*% X2) %*% t(X2)
        MX2 = diag(rep(1,n)) - PX2

        X3 = X1[,3]

        bhat3 =  solve(t(X3)%*% MX2 %*% X3, t(X3) %*% MX2 %*% y )
        print(bhat3)

            [,1]
[1,] -0.01698568
```

2 Statistical Properties of Least Squares

To talk about the statistical properties in finite sample, we impose the following assumptions.

1. The data $(y_i, x_i)_{i=1}^n$ is a random sample from the same data generating process $y_i = x_i'\beta + e_i$.
2. $e_i|x_i \sim N\left(0, \sigma^2\right)$.

2.1 Maximum Likelihood Estimation

Under the normality assumption, $y_i|x_i \sim N\left(x_i'\beta, \gamma\right)$, where $\gamma = \sigma^2$. The *conditional* likelihood of observing a sample $(y_i, x_i)_{i=1}^n$ is

$$\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\gamma}} \exp\left(-\frac{1}{2\gamma}\left(y_i - x_i'\beta\right)^2\right),$$

and the (conditional) log-likelihood function is

$$L\left(\beta, \gamma\right) = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log\gamma - \frac{1}{2\gamma}\sum_{i=1}^{n}\left(y_i - x_i'\beta\right)^2.$$

Therefore, the maximum likelihood estimator (MLE) coincides with the OLS estimator, and $\widehat{\gamma}_{\mathrm{MLE}} = \widehat{e}'\widehat{e}/n$.

2.2 Finite Sample Distribution

We can show the finite-sample exact distribution of $\widehat{\beta}$. *Finite sample distribution* means that the distribution holds for any n; it is in contrast to *asymptotic distribution*, which is a large sample approximation to the finite sample distribution.

Since

$$\widehat{\beta} = \left(X'X\right)^{-1} X'y = \left(X'X\right)^{-1} X'\left(X'\beta + e\right) = \beta + \left(X'X\right)^{-1} X'e,$$

we have the estimator $\widehat{\beta}|X \sim N\left(\beta, \sigma^2 \left(X'X\right)^{-1}\right)$, and

$$\widehat{\beta}_k|X \sim N\left(\beta_k, \sigma^2 \eta_k' \left(X'X\right)^{-1} \eta_k\right) \sim N\left(\beta_k, \sigma^2 \left(X'X\right)^{-1}_{kk}\right),$$

where $\eta_k = \left(1\left\{l = k\right\}\right)_{l=1,\ldots,K}$ is the selector of the k-th element.

In reality, σ^2 is an unknown parameter, and

$$s^2 = \widehat{e}'\widehat{e}/\left(n-K\right) = e'M_X e/\left(n-K\right)$$

is an unbiased estimator of σ^2. Consider the T-statistic

$$T_k = \frac{\widehat{\beta}_k - \beta_k}{\sqrt{s^2\left[\left(X'X\right)^{-1}\right]_{kk}}} = \frac{\left(\widehat{\beta}_k - \beta_k\right)/\sqrt{\sigma^2\left[\left(X'X\right)^{-1}\right]_{kk}}}{\sqrt{\frac{e'}{\sigma}M_X\frac{e}{\sigma}/\left(n-K\right)}}.$$

The numerator follows a standard normal, and the denominator follows $\frac{1}{n-K}\chi^2\left(n-K\right)$. Moreover, the numerator and the denominator are independent. As a result, $T_k \sim t\left(n-K\right)$.

2.3 Mean and Variance

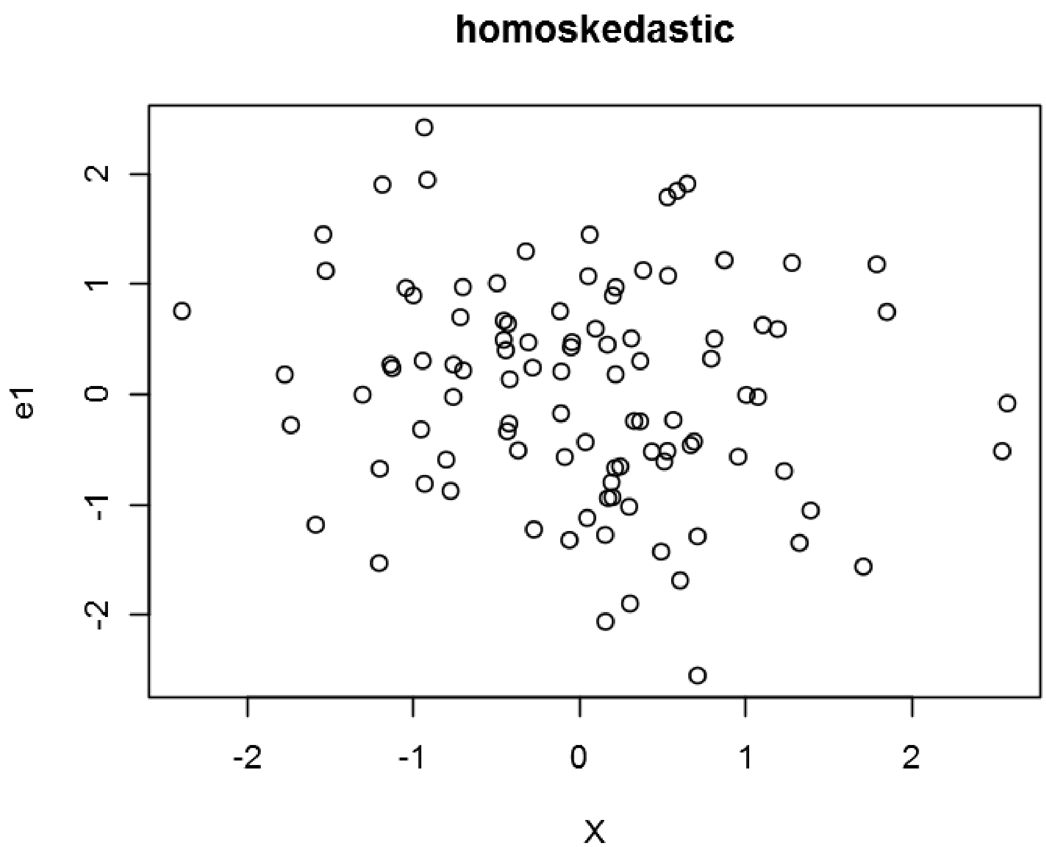
Now we relax the normality assumption and statistical independence. Instead, we assume a regression model $y_i = x_i'\beta + e_i$ and

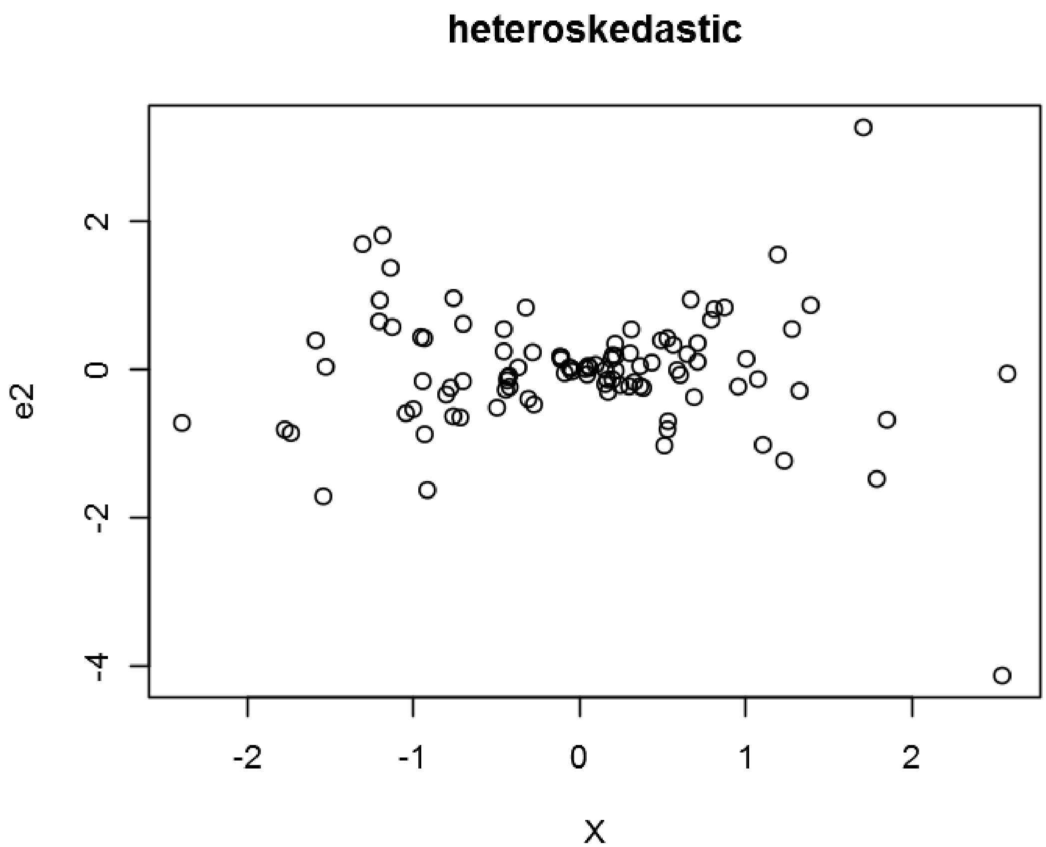
$$E[e_i|x_i] = 0$$
$$E\left[e_i^2|x_i\right] = \sigma^2.$$

where the first condition is the *mean independence* assumption, and the second condition is the *homoskedasticity* assumption.

Example (Heteroskedasticity) If $e_i = x_i u_i$, where x_i is a scalar random variable, u_i is independent of x_i, $E\left[u_i\right] = 0$ and $E\left[u_i^2\right] = \sigma^2$. Then $E\left[e_i|x_i\right] = 0$ but $E\left[e_i^2|x_i\right] = \sigma_i^2 x_i^2$ is a function of x_i. We say e_i^2 is a heteroskedastic error.

```
In [10]: n = 100
         X = rnorm(n)

         e1 = rnorm(n)
         plot( y = e1, x = X, main = "homoskedastic")

         e2 = X * rnorm(n)  # the source of heteroskedasticity
         plot( y = e2, x = X,  main = "heteroskedastic")
```


homoskedastic
e1
2
1
0
-1
-2
-2
-1
0
1
2
X



These assumptions are about the first and second moment of e_i conditional on x_i. Unlike the normality assumption, they do not restrict the entire distribution of e_i.

- Unbiasedness:

$$E\left[\widehat{\beta}|X\right] = E\left[\left(X'X\right)^{-1} XY|X\right] = E\left[\left(X'X\right)^{-1} X\left(X'\beta + e\right)|X\right] = \beta.$$

  Unbiasedness does not rely on homoskedasticity.

- Variance:

$$\begin{aligned}
\text{var}\left(\widehat{\beta}|X\right) &= E\left[\left(\widehat{\beta} - E\widehat{\beta}\right)\left(\widehat{\beta} - E\widehat{\beta}\right)'|X\right] \\
&= E\left[\left(\widehat{\beta} - \beta\right)\left(\widehat{\beta} - \beta\right)'|X\right] \\
&= E\left[\left(X'X\right)^{-1} X'ee'X\left(X'X\right)^{-1}|X\right] \\
&= \left(X'X\right)^{-1} X'E\left[ee'|X\right] X\left(X'X\right)^{-1} \\
&= \left(X'X\right)^{-1} X'\left(\sigma^2 I_n\right) X\left(X'X\right)^{-1} \\
&= \sigma^2\left(X'X\right)^{-1}.
\end{aligned}$$

2.4 Gauss-Markov Theorem

Gauss-Markov theorem justifies the OLS estimator as the efficient estimator among all linear unbiased ones. *Efficient* here means that it enjoys the smallest variance in a family of estimators.

There are numerous linearly unbiased estimators. For example, $(Z'X)^{-1} Z'y$ for $z_i = x_i^2$ is unbiased because $E\left[(Z'X)^{-1} Z'y\right] = E\left[(Z'X)^{-1} Z'(X\beta + e)\right] = \beta$.

Let $\tilde{\beta} = A'y$ be a generic linear estimator, where A is any $n \times K$ functions of X. As

$$E\left[A'y|X\right] = E\left[A'(X\beta + e)|X\right] = A'X\beta.$$

So the linearity and unbiasedness of $\tilde{\beta}$ implies $A'X = I_n$. Moreover, the variance

$$\mathrm{var}\left(A'y|X\right) = E\left[(A'y - \beta)(A'y - \beta)'|X\right] = E\left[A'ee'A|X\right] = \sigma^2 A'A.$$

Let $C = A - X(X'X)^{-1}$.

$$\begin{aligned}
A'A - (X'X)^{-1} &= \left(C + X(X'X)^{-1}\right)'\left(C + X(X'X)^{-1}\right) - (X'X)^{-1} \\
&= C'C + (X'X)^{-1} X'C + C'X(X'X)^{-1} \\
&= C'C,
\end{aligned}$$

where the last equality follows as

$$(X'X)^{-1} X'C = (X'X)^{-1} X'\left(A - X(X'X)^{-1}\right) = (X'X)^{-1} - (X'X)^{-1} = 0.$$

Therefore $A'A - (X'X)^{-1}$ is a positive semi-definite matrix. The variance of any $\tilde{\beta}$ is no smaller than the OLS estimator $\widehat{\beta}$.

Homoskedasticity is a restrictive assumption. Under homoskedasticity, $\mathrm{var}\left(\widehat{\beta}\right) = \sigma^2 (X'X)^{-1}$. Popular estimator of σ^2 is the sample mean of the residuals $\widehat{\sigma}^2 = \frac{1}{n}\widehat{e}'\widehat{e}$ or the unbiased one $s^2 = \frac{1}{n-K}\widehat{e}'\widehat{e}$. Under heteroskedasticity, Gauss-Markov theorem does not apply.