# University of Rome Tor Vergata
# Department of Economics and Finance

## Static Regression - Fall 2018 - Prof. Cubadda

## Problem Set 1 - Solutions

TA Claudia De Palo
claudia.depalo@alumni.uniroma2.eu

---

## Exercise 1

In econometrics, a model for a conditional expectation is often specified to depend on a vector of parameters $\beta \in B$, which gives a parametric model of $E(y|x)$ of the form $E(y|x) = m(x, \beta)$, where

$$\underset{K\times 1}{x} = (x_1, x_2, ..., x_k, ...x_K)' \quad \text{vector of explanatory variables}$$
$$\underset{K\times 1}{\beta} = (\beta_1, \beta_2, ..., \beta_k, ...\beta_K)' \quad \text{vector of parameters}$$

Often, the first explanatory variable is equal to 1. This gives you a model with the intercept.

- Let us assume that the conditional expectation $m(x, \beta)$ is linear in the components of the vectors $\beta$ and $x$, i.e., $E(y|x) = x'\beta$

a) Write down the regression curve and define the error term. How can you write the population regression equation?

b) Show the implications on the error term.

c) Using your previous results, derive the following moment condition: $E(xx')\beta = E(xy)$. What assumption do you need for the true parameter $\beta$ to exist?

**Solution**

a) The regression curve of $y$ on $x_1...x_K$ is defined to be the conditional expectation $E(y|x_1...x_K)$. Given the linearity assumption made above, the regression curve is

$$E(y|x) = x'\beta = \beta_1 x_1 + ...\beta_K x_K \tag{1}$$

Define the **error term** or **disturbance term** as:

$$\varepsilon = y - E(y|x) = y - x'\beta \tag{2}$$

It follows from (1) and (2) that we can always write $y$ as its conditional expectation, $E(y|x)$, plus an error term that has conditional mean zero. Therefore, the population regression equation can be written as:

$$y = E(y|x) + \varepsilon = x'\beta + \varepsilon \tag{3}$$

where

$$E(\varepsilon|x) = 0 \tag{4}$$

b) The zero conditional mean in (4) implies that:
- $E(\varepsilon) = E_X[E(\varepsilon|x)] = E_X[0] = 0$
- $E(x\varepsilon) = E_X[E(x\varepsilon|x)] = E_X[xE(\varepsilon|x)] = 0$ **(orthogonality condition)**

c) Replace the definition given in (2) in the orthogonality condition:

$$\begin{aligned} 0 = E(x\varepsilon) &= E[x(y - x'\beta)] \\ &= E(xy - xx'\beta) \\ &= E(xy) - E(xx')\beta \end{aligned}$$

Hence,

$$E(xx')\beta = E(xy)$$

This implies that, for $\beta$ to exists, we need to assume that $E(xx')$ is nonsingular or equivalently that rank $E(xx') = K$

# Exercise 2

Assume that the conditional variance of $y|x$ is a constant function of $x$: $Var(y|x) = \sigma^2$.
Show that:

$$\begin{aligned} Var(\varepsilon) &= Var(y) - \beta^{'} Var(x)\beta \\ &= Var(y) - Cov(y,x)Var(x)^{-1}Cov(x,y) \\ &= \sigma^2 \end{aligned}$$

**Solution**
Notice that:

$$\sigma^2 = Var(y|x) = Var(\beta'x + \varepsilon|x) = Var(\varepsilon|x)$$

In addition, by the Law of Iterated Expectation (LIE):

$$Var(\varepsilon) = E(\varepsilon^2) = E_X[E(\varepsilon^2|x)] = E_X[Var(\varepsilon|x)] = \sigma^2$$

Recall now the low of total variance:

$$Var(y) = E[Var(y|x)] + Var[E(y|x)]$$

Hence,

$$\begin{aligned} Var(y) - Var[E(y|x)] &= E[Var(y|x)] \\ Var(y) - Var[(x'\beta)] &= E[Var(y|x)] \\ Var(y) - \beta' Var(x)\beta &= E[Var(y|x)] \\ &= E[Var(\varepsilon|x)] \\ &= Var(\varepsilon) \\ &= \sigma^2 \end{aligned}$$

Recall now the following definitions:

$$\begin{aligned} \varepsilon &= y - \beta' x \\ \beta &= Var(x)^{-1} Cov(x, y) \\ \beta' &= \underbrace{Cov(y, x)}_{=Cov(x,y)'} Var(x)^{-1} \end{aligned}$$

Hence,

$$\begin{aligned} Var(\varepsilon) &= Var(y) + \beta' Var(x)\beta - 2\beta' Cov(x, y) \\ &= Var(y) + Cov(y, x)Var(x)^{-1}Var(x)Var(x)^{-1}Cov(x, y) - 2Cov(y, x)Var(x)^{-1}Cov(x, y) \\ &= Var(y) - Cov(y, x)Var(x)^{-1}Cov(x, y) \end{aligned}$$

# Exercise 3

Consider a model where the population regression function is equal to

$$E(y \mid x) = x_1'\beta_1 + x_2'\beta_2$$

where $x_1$ is a $K_1 \times 1$ vector and $x_2$ is a $K_2 \times 1$ vector.

Use the properties of the conditional expectation to show that

$$E\{[y - E(y \mid x_2)] \mid [x_1 - E(x_1 \mid x_2)]\} = [x_1 - E(x_1 \mid x_2)]'\beta_1$$

*Hint*
*You can write* $E(y \mid x)$ *as* $E(y \mid x_1, x_2) = x_1'\beta_1 + x_2'\beta_2$ *, which is the same as*

$$y = x_1'\beta_1 + x_2'\beta_2 + \varepsilon$$

*where,*

$$E(\varepsilon \mid x_1, x_2) = 0$$

*(implication: $\varepsilon$ is orthogonal to $x_1$, $x_2$ and to any function of $x_1$ and $x_2$).*

**Solution**
Using the property of linearity of the conditional expectation, we have that

$$\begin{aligned} E\{[y - E(y \mid x_2)] \mid [x_1 - E(x_1 \mid x_2)]\} &= E\{y \mid [x_1 - E(x_1 x_2)]\} - E\{E(y \mid x_2) \mid [x_1 - E(x_1 x_2)]\} \\ &= E\{x_1'\beta_1 + x_2'\beta_2 + \varepsilon \mid [x_1 - E(x_1 x_2)]\} - E\{E(y \mid x_2) \mid [x_1 - E(x_1 x_2)]\} \end{aligned}$$

Recall that: $E(y \mid x_2) = E(x_1'\beta_1 + x_2'\beta_2 + \varepsilon \mid x_2) = E(x_1 \mid x_2)'\beta_1 + x_2'\beta_2$.

Hence,

$$= E\{x_1'\beta_1 + x_2'\beta_2 + \varepsilon \mid [x_1 - E(x_1 x_2)]\} - E\{E(x_1 \mid x_2)'\beta_1 + x_2'\beta_2 \mid x_2) \mid [x_1 - E(x_1 \mid x_2)]\}$$

Applying again the linearity of the conditional expectation and since $E\{\varepsilon \mid [x_1 - E(x_1 \mid x_2)]\} = 0$, we get

$$\begin{aligned} &= E\{x_1'\beta_1 \mid [x_1 - E(x_1 \mid x_2)]\} + E\{x_2'\beta_2 \mid [x_1 - E(x_1 \mid x_2)]\} - E\{E(x_1 \mid x_2)'\beta_1 + x_2'\beta_2 \mid x_2) \mid [x_1 - E(x_1 \mid x_2)]\} \\ &= E\{x_1'\beta_1 \mid [x_1 - E(x_1 \mid x_2)]\} + E\{x_2'\beta_2 \mid [x_1 - E(x_1 \mid x_2)]\} - E\{E(x_1 \mid x_2)'\beta_1 \mid [x_1 - E(x_1 \mid x_2)]\} - \\ &- E\{x_2'\beta_2 \mid [x_1 - E(x_1 \mid x_2)]\} \\ &= \beta_1 E\{x_1' \mid [x_1 - E(x_1 \mid x_2)]\} - \beta_1 E\{E(x_1 \mid x_2)' \mid [x_1 - E(x_1 \mid x_2)]\} \\ &= \beta_1\{E[x_1' - E(x_1 \mid x_2)'] \mid [x_1 - E(x_1 \mid x_2)]\} \\ &= \beta_1\{E[x_1 - E(x_1 \mid x_2)]' \mid [x_1 - E(x_1 \mid x_2)]\} \\ &= \beta_1[x_1 - E(x_1 \mid x_2)]' \end{aligned}$$

# Exercise 4

Write down the OLS problem and

a) derive $\hat{\beta}_{OLS}$

b) compute $E(\hat{\beta}_{OLS})$ and $V(\hat{\beta}_{OLS})$

c) state the Gauss Markov theorem, recalling the assumptions you need to take

**Solution**
The ordinary least-squares (OLS) problem is

$$\min_{\beta} Q_n(\beta) = n^{-1}\sum_{n=1}^{N}(y_n - x_n'\beta)^2 = (y - X\beta)'(y - X\beta)$$

The method of OLS consists in finding the value of $\beta$ which minimizes the function $Q_n$, called OLS the criterion.

a) Differentiating $Q_n$ with respect to $\beta$ gives

$$\frac{\partial Q_n(\beta)}{\partial \beta} = -2X'y + 2X'X\beta$$

which set equal to zero gives the **normal equations**

$$X'y - X'X\hat{\beta}_{OLS} = 0$$

If the $K \times K$ matrix $n^{-1}\sum_n x_n x_n' = X'X$ is nonsingular, solving the OLS normal equations gives the OLS estimate

$$\hat{\beta}_{OLS} = \left(\sum\nolimits_n x_n x_n'\right)^{-1} \left(\sum\nolimits_n x_n y_n\right)^{-1} = (X'X)^{-1}X'y = 0$$

b) Notice that

$$\begin{aligned}
\hat{\beta}_{OLS} &= (X'X)^{-1}X'y \\
&= (X'X)^{-1}X'(X\beta + \varepsilon) \\
&= \underbrace{(X'X)^{-1}X'X}_{=I_k}\beta + (X'X)^{-1}X'\varepsilon \\
&= \beta + (X'X)^{-1}X'\varepsilon
\end{aligned}$$

Taking expectation on both sides, we get

$$\begin{aligned}
E(\hat{\beta}_{OLS}) &= \beta + (X'X)^{-1}X'\underbrace{E(\varepsilon)}_{=0} \\
&= \beta
\end{aligned}$$

Notice now that

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

The variance of $\hat{\beta}_{OLS}$ is therefore

$$\begin{aligned}
E[(\hat{\beta}_{OLS} - \beta)'(\hat{\beta}_{OLS} - \beta)] &= (X'X)^{-1}X'\underbrace{E(\varepsilon'\varepsilon)}_{\sigma^2}X(X'X)^{-1} \\
&= \sigma^2(X'X)^{-1}\underbrace{X'X(X'X)^{-1}}_{=I_k} \\
&= \sigma^2(X'X)^{-1}
\end{aligned}$$

c) **Assumption A1: Linearity in the parameters**.
The conditional expectation of Y $E(Y|X)$ is a linear function of the parameters, the $\beta$'s. It may or

may not be linear in the variable X.

**Assumption A2: Random sample of n observations**.
This assumption is composed of three related sub-assumptions.
- Assumption A2.1: The sample consists of n-paired observations that are drawn randomly from the population.
- Assumption A2.2: The number of observations is greater than the number of parameters to be estimated, usually written $n > k$.
- Assumption A2.3: X is a non-stochastic $(n \times k)$ matrix, with $k < n$, which has full (column) rank $(rank(X) = k)$.
Two conditions are necessary to ensure this assumption:
- the number of observations cannot be smaller than the number of explanatory variables in the model. So, $n \geq k$.
- there cannot be an exact linear relationship between two explanatory variables. This means that it is impossible to include one variable twice or include a variable which is a linear combination of another variable as this would lead to perfect collinearity.
If A2.3 fails, then we have $(X'X)$ not invertible and cannot compute $\hat{\beta}$.

**Assumption A3: Strict exogeneity**.

$$E(\varepsilon|X) = 0$$

This statement indicates there is no relationship between the error terms and the explanatory variables.
Implications:
- $E(\varepsilon) = 0$ as $E(\varepsilon) = E(\varepsilon|X) = E(0) = 0$
- $E(Y|X) = X\beta$. If it fails, we have misspecification in the regression function (e.g. omitted variables).

**Assumption A4: Independent and identically distributed error terms**.
The error terms of the population $\varepsilon_i$ are independent and identically distributed with zero expected value and constant variance $\sigma^2$:

$$\varepsilon_i \sim (0, \sigma^2)$$

This implies:
- A4.1: $E(\varepsilon_i) = 0$
This is less strong than $E(\varepsilon_i|X) = 0$. If $E(\varepsilon_i|X) = 0$ is fulfilled, it implies also $E(\varepsilon_i) = 0$.
$E(\varepsilon_i) = 0$ does not imply $E(\varepsilon_i|X) = 0$.
- A4.2 **Homoschedasticity**: $Var(\varepsilon_i) = \sigma^2$
- A4.3 **No autocorrelation**: $Cov(\varepsilon_i, \varepsilon_j) = 0$ for $i \neq j$
A4.2 and A4.3 can be summarized as $Var(\varepsilon) = \sigma^2 I_n$.

The following is not a Gauss Markov assumption so we cannot list among the other assumptions: $\varepsilon_i$

are normally distributed in the population.
OLS estimator will still be BLUE even if $\varepsilon_i$ are not normally distributed in the population. We don't worry too much because we have normality asymptotically.

Under these assumptions, the **Gauss Markov Theorem** states that $\hat{\beta}_{OLS}$ is the best linear unbiased estimator (BLUE)of $\beta$, i.e., $Var(\hat{\beta}_{OLS}) \leq Var(\bar{\beta})$ $\forall$ linear and unbiased estimators $\bar{\beta}$.

# Exercise 5

Consider the projection matrix $M = I_n - X(X'X)^{-1}X'$
Show that:

a) $M' = M$; $MM = M$ ($M$ is symmetric and idempotent)

b) $rank(M) = trace(M) = N - K$

Which is the relationship between $M$, the fitted values and the residuals?

**Solution**

a)

$$M' = [I_n - X(X'X)^{-1}X']' = I_n - [X(X'X)^{-1}X'] = I_n - X(X'X)^{-1}X' = M$$

$$\begin{aligned} MM &= [I_n - X(X'X)^{-1}X'][I_n - X(X'X)^{-1}X'] \\ &= I_n - X(X'X)^{-1}X' - X(X'X)^{-1}X' + [X(X'X)^{-1}X'][X(X'X)^{-1}X'] \\ &= I_n - X(X'X)^{-1}X' - X(X'X)^{-1}X' + X(X'X)^{-1}\underbrace{X'X(X'X)^{-1}}_{=I_k}X' \\ &= I_n - X(X'X)^{-1}X' = M \end{aligned}$$

b) In general, notice that if A is a symmetric matrix, then $A = V\Delta V'$, where $V$ is the orthogonal matrix containing the eigenvectors of $A$ and $\Delta$ is the diagonal matrix of eigenvalues. The $i-th$ diagonal element of the matrix $\Delta$ is indicated as $d_i$.
Now,

$$\Sigma a_{ij} = tr(A) = tr(A = V\Delta V') = tr(\Delta \underbrace{VV'}_{=I}) = tr(\Delta) = \Sigma d_i.$$

We know that:
- if A symmetric, then $rank(A) = \#\text{eigenvalues}(A) \neq 0$

- idempotent matrices have eigenvalues equal to zero or one
Our matrix $M$ is symmetric and idempotent, therefore

$$\begin{aligned} rank(M) &= \#eigenvalues(M) \neq 0 \\ &= \Sigma d_i \\ &= tr(M) \end{aligned}$$

Now,

$$\begin{aligned} tr(M) &= tr(I_n - X(X'X)^{-1}X') \\ &= tr(I_n) - tr(X(X'X)^{-1}X') \\ &= tr(I_n) - tr((X'X)^{-1}X'X) \\ &= tr(I_n) - tr(I_k) \\ &= N - K \end{aligned}$$

Concerning the relationship between $M$ and residuals:

$$\begin{aligned} e &= y - X\hat{\beta} \\ &= y - X(X'X)^{-1}X'y \\ &= [I - X(X'X)^{-1}X']y \\ &= My \end{aligned}$$

With respect to fitted values, we have that

$$\begin{aligned} \hat{y} &= X\hat{\beta} \\ &= y - e \\ &= y - My \\ &= (I - M)y \\ &= X(X'X)^{-1}X'y \end{aligned}$$

# Exercise 6

Prove the Gauss Markov Theorem, i.e., $Var(\hat{\beta}_{OLS}) \leq Var(\bar{\beta})$ $\forall$ linear and unbiased estimators $\bar{\beta}$.

**Solution**
Essentially, we have to show that $Var(\bar{\beta}) - Var(\hat{\beta})$ is positive semi-definite.
To prove the theorem, we need to introduce another linear and unbiased estimator of $\beta$. Let us consider

$\bar{\beta} = C'y$, where $C$ is a $N \times K$ matrix.

$$\begin{aligned} E(\bar{\beta}) &= C'E(y) \\ &= C'E(X\beta + \varepsilon) \\ &= C'X\beta + \underbrace{E(\varepsilon)}_{=0} \\ &= C'X\beta \\ &= \beta \end{aligned}$$

which is true only if $C'X = I_k$. Then

$$\begin{aligned} Var(\bar{\beta}) &= Var(C'y) \\ &= C'Var(y)C \\ &= \sigma^2 C'C \end{aligned}$$

Notice that, under unbiasedness

$$\begin{aligned} Var(\hat{\beta}_{OLS}) &= \sigma^2 (X'X)^{-1} \\ &= \sigma^2 \underbrace{C'X}_{=I_k} (X'X)^{-1} \underbrace{X'C}_{=I_k} \end{aligned}$$

Hence,

$$\begin{aligned} Var(\bar{\beta}) - Var(\hat{\beta}_{OLS}) &= \sigma^2 C'C - \sigma^2 C'X(X'X)^{-1}X'C \\ &= \sigma^2 C'[I_n - X(X'X)^{-1}X']C \\ &= \sigma^2 C'M'MC \\ &= \sigma^2 C'MC \qquad \text{which is positive semi-definite} \end{aligned}$$

# Exercise 7

Consider the classical Gaussian linear model $Y \sim \mathcal{N}_n(X\beta, \sigma^2 I_n)$

a) Show that maximizing the log-likelihood with respect to $\beta$ is equivalent to minimizing the OLS criterion.

b) What can you say about the estimation of $\sigma^2$? Discuss.

**Solution**

a) Our linear regression model is:

$$Y = X\beta + \varepsilon$$

where $\varepsilon \sim N(0, \sigma^2)$ with first and second moments $E(Y|X) = X\beta$ and $Var(Y|X) = Var(\varepsilon|X) = \sigma^2$.

We know that $Y = \underbrace{\begin{pmatrix} y_1 \\ \vdots \\ y_N \end{pmatrix}}_{N\times 1}$.

If we express our model abandoning the matrix notation, we have as well-known:

$$y_i = \beta' x_i + \varepsilon_i$$

where $\varepsilon_i \sim i.i.d.N(0, \sigma^2)$, so our y's are distributed as $y_i \sim N(\beta' x_i, \sigma^2)$ with $i = 1, ..., N$.
The probability density function, conditioning on the parameters $\beta$ and $\sigma^2$, is denoted by:

$$f(y_i|x_i; \beta, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\left[-\frac{1}{2\sigma^2}(y_i - \beta' x_i)^2\right]$$

The joint density function is, by independence, equal to the product of the marginal densities:

$$f(y_1, ..., y_N|X; \beta, \sigma^2) = f(y_1|X; \beta, \sigma^2)...f(y_N|X; \beta, \sigma^2) = \prod_{i=1}^{N} f(y_i|x_i; \beta, \sigma^2)$$

The likelihood function is defined as the joint density treated as a function of the parameters:

$$\begin{aligned} L(\beta, \sigma^2|y_1, ..., y_N; X) &= \prod_{i=1}^{N} f(y_i|x_i; \beta, \sigma^2) \\ &= \prod_{i=1}^{N} \left\{ \frac{1}{\sqrt{2\pi\sigma^2}} exp\left[-\frac{1}{2\sigma^2}(y_i - \beta' x_i)^2\right] \right\} \\ &= \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{N}{2}} \prod_{i=1}^{N} exp\Big[-\frac{1}{2\sigma^2}(y_i - \beta' x_i)^2\Big] \end{aligned}$$

It is usually simpler to work with the log of the likelihood function:

$$\begin{aligned} \ell(\beta, \sigma^2|y_1, ..., y_N; X) = logL(\beta, \sigma^2|y_1, ..., y_N; X) &= \sum_{i=1}^{N} logf(y_i|x_i; \beta, \sigma^2) \\ &= -\frac{N}{2} log2\pi - \frac{N}{2} log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i}^{N}(y_i - \beta' x_i)^2 \end{aligned}$$

In matrix notation:

$$\ell(\beta, \sigma^2|y_1, ..., y_N; X) = logL(\beta, \sigma^2|y_1, ..., y_N; X) = -\frac{N}{2} log2\pi - \frac{N}{2} log\sigma^2 - \frac{1}{2\sigma^2}(y - X\beta)'(y - X\beta)$$

A maximum likelihood estimator of $(\beta, \sigma^2)$ is a solution to the maximization problem:

$$\max_{\beta,\sigma^2 \in \ominus} \ell(\beta, \sigma^2)$$

Note that the solution to an optimization problem is invariant to a strictly monotone increasing transformation of the objective function, a MLE can be obtained as a solution to the following problem:

$$\max_{\beta,\sigma^2 \in \ominus} \ell(\beta, \sigma^2) = \max_{\beta,\sigma^2 \in \ominus} L(\beta, \sigma^2)$$

According to the OLS criterion, the OLS problem is equivalent to the problem:

$$\min_{\beta \in \ominus} \quad S(\beta) = (y - X\beta)'(y - X\beta)$$

A solution $\hat{\beta}$ must necessarily satisfy the system of linear equations called normal equations.

$$X'X\beta - X'y = 0$$

If you look at the log-likelihood function as:

$$\ell(\beta, \sigma^2) = -\frac{N}{2} log 2\pi - \frac{N}{2} log \sigma^2 - \frac{1}{2\sigma^2} S(\beta)$$

You see that maximizing the log-likelihood function with respect to $\beta$ is equivalent to minimizing the OLS criterion. Then, the Gaussian ML and the OLS estimates of $\beta$ coincides.

Taking the derivative of the log likelihood function with respect to $\beta$:

$$\frac{\partial \ell}{\partial \beta} = -\frac{1}{2\sigma^2} 2(-X')(y - X\beta)$$

$$\frac{\partial \ell}{\partial \beta} = -\frac{1}{2\sigma^2}(-2X'y + 2X'X\beta)$$

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

Thus, $\hat{\beta}_{ML} = \hat{\beta}_{OLS}$ is unbiased.

(b) Taking the derivative of the log likelihood function with respect to the parameter $\sigma^2$:

$$\frac{\partial \ell}{\partial \sigma^2} = -\frac{N}{2}\frac{1}{\sigma^2} - \frac{1}{2}\Big(-\frac{1}{\sigma^4}\Big)(y - X\beta)'(y - X\beta)$$

$$\hat{\sigma}^2_{ML} = \frac{(y - X\hat{\beta}_{ML})'(y - X\hat{\beta}_{ML})}{N} = \frac{e'e}{N}$$

The ML estimator $\hat{\sigma^2_{ML}}$ is biased for $\sigma^2$ but has a lower sample variance than the unbiased estimator.

In fact:

$$
\begin{aligned}
E(\hat{\sigma}^2_{ML}) &= E\Big(\frac{e'e}{N}\Big) = \frac{1}{N}E(e'e) \\
&= \frac{1}{N}\Big[E(\varepsilon' \underbrace{M'M}_{M}\varepsilon))\Big] = \frac{1}{N}E\Big[tr(\varepsilon' M\varepsilon)\Big] && \text{since } \varepsilon' M\varepsilon \text{ is a scalar} \\
&= \frac{1}{N}E\Big[tr(M\varepsilon\varepsilon')\Big] && \text{since tr(ABC)=tr(BCA)} \\
&= \frac{1}{N}tr\Big[E(M\varepsilon\varepsilon')\Big] && \text{since expectation is a linear operator} \\
&= \frac{1}{N}tr\Big[ME(\varepsilon\varepsilon')\Big] && \text{since M is non-stochastic} \\
&= \sigma^2\frac{1}{N}\underbrace{tr(M)}_{N-k} && \text{since tr(aA)=atr(A) where a is a scalar} \\
&= \sigma^2\frac{N-k}{N} < \sigma^2
\end{aligned}
$$

Therefore, the ML estimator of the variance is downward biased.