Let $Y = X\beta + \epsilon$ where the response vector $Y$ is $n \times 1$. The $n \times p$ design matrix $X$ has full rank $p < n$. The $p \times 1$ parameter vector is $\beta$. The $n \times 1$ disturbance vector $\epsilon$ is random. The OLS estimator is $\hat{\beta} = (X'X)^{-1}X'Y$. At the moment, no assumptions are imposed on $\epsilon$.

Lemma 1. $\hat{\beta} = \beta + (X'X)^{-1}X'\epsilon$.
Proof. Substitute the formula for $Y$ into the formula for $\hat{\beta}$:

$$\hat{\beta} = (X'X)^{-1}X'Y$$
$$= (X'X)^{-1}X'(X\beta + \epsilon)$$
$$= (X'X)^{-1}(X'X)\beta + (X'X)^{-1}X'\epsilon$$
$$= Ip_{p \times p} + (X'X)^{-1}X'\epsilon$$
$$= \beta + (X'X)^{-1}X'\epsilon$$

Theorem 1. $E(\hat{\beta}|X) = \beta + (X'X)^{-1}X'E(\epsilon|X)$.
Proof. Given $X$, related matrices like $(X'X)^{-1}X'$ are constant and factor out of the expectation. (This idea will be used several times below, without comment.) Lemma 1 completes the proof.

Corollary 1. If $E(\epsilon|X) = 0$ then $\hat{\beta}$ is conditionally unbiased.

Definition 1. If $U$ is random $p \times 1$, then

$$\text{cov}(U) = E\left\{[U - E(U)][U - E(U)]\right\} = E(UU') - E(U)E(U)'$$

Remark. You might want to check the equality, and the fact that $[E(U)]' = E(U')$.

Lemma 2. If $U$ is a random $p \times 1$ vector, while $A$ is a constant $p \times p$ matrix, and $B$ is a constant $p \times 1$ vector, then $\text{cov}(AU + B) = A\text{cov}(U)A'$.

Proof. The covariance does not depend on additive constants like $B$; these cancel. For simplicity, assume $E(U) = 0_{p \times 1}$. Then $E(AU) = AE(U) = 0_{p \times 1}$. Recall that $(CD)' = D'C'$. Now $\text{cov}(AU) = E((AU)(AU)') = E(AUU'A') = AE(UU')A' = A\text{cov}(U)A'$.

Theorem 2. $\text{cov}(\hat{\beta}|X) = (X'X)^{-1}X'\text{cov}(\epsilon|X)X(X'X)^{-1}$.
Proof. Use the lemmas and the fact that $X'X$ is symmetric.

Corollary 2. If $E(\epsilon|X) = 0$ and $\text{cov}(\epsilon|X) = \sigma^2 I_{n \times n}$ then $\hat{\beta}$ is conditionally unbiased and $\text{cov}(\hat{\beta}|X) = \sigma^2 (X'X)^{-1}$.

Proof. For the covariance, substitute into the theorem:

$$\text{cov}(\hat{\beta}|X) = (X'X)^{-1}X'\sigma^2 I_{n \times n}X(X'X)^{-1}$$
$$= \sigma^2 (X'X)^{-1}X'(I_{n \times n}X)(X'X)^{-1}$$
$$= \sigma^2 (X'X)^{-1}(X'X)(X'X)^{-1}$$
$$= \sigma^2 (X'X)^{-1}Ip_{p \times p}$$
$$= \sigma^2 (X'X)^{-1}$$
Definition 2. If $A$ is a square matrix, the “trace” of $A$ is the sum of the diagonal elements of $A$.

Lemma 3. (i) If $A$ is $m \times n$ and $B$ is $n \times m$, then $\text{trace}(AB) = \text{trace}(BA)$. (ii) If $C$ and $D$ are $m \times m$, then $\text{trace}(C + D) = \text{trace}(C) + \text{trace}(D)$; if $\alpha$ is a scalar constant, then $\text{trace}(\alpha C) = \alpha \text{trace}(C)$.

The “hat matrix” $H = X(X'X)^{-1}X'$ is symmetric and idempotent ($H^2 = H$); ditto for $I_{n \times n} - H$. The “fitted values” are $\hat{Y} = X\hat{\beta} = HY$. Confirm that $HX = X$. The “residuals” are $e = Y - \hat{Y} = (I_{n \times n} - H)Y = (I_{n \times n} - H)e$: substitute the formula for $Y$ into the formula for $e$, and check that $(I_{n \times n} - H)X = 0_{n \times p}$. The hat matrix projects onto the column space of $X$, and $I_{n \times n} - H$ projects onto the orthocomplement.

Lemma 4. $\text{trace}(H) = p$ and $\text{trace}(I_{n \times n} - H) = n - p$.

Proof. $\text{trace}[X(X'X)^{-1}X'] = \text{trace}[(X'X)^{-1}X'] = \text{trace}(I_{p \times p}) = p$: use lemma 3(i) to move $X$ from the left end of the product to the right end. Lemma 3(ii) completes the proof.

Theorem 3. $E(\|e\|^2 | X) = \text{trace}[(I_{n \times n} - H)E(\epsilon \epsilon' | X)]$.

Proof. $ee' = (I_{n \times n} - H)e\epsilon'(I_{n \times n} - H)$, because $(I_{n \times n} - H)' = (I_{n \times n} - H)$. Now $\|e\|^2 = e'e = \text{trace}(ee') = \text{trace}[(I_{n \times n} - H)e\epsilon'(I_{n \times n} - H)] = \text{trace}[(I_{n \times n} - H)e\epsilon']$.

Use lemma 3(i) to see that $e'e = \text{trace}(ee')$. Use lemma 3(ii) again to move $I_{n \times n} - H$ from right to left: keep in mind that $I_{n \times n} - H$ is idempotent. Finally, take the conditional expectation given $X$. The trace is linear by lemma 3(ii), and $H$ is conditionally a constant matrix, so $E\{\text{trace}[(I_{n \times n} - H)e\epsilon'] | X\} = \text{trace}[(I_{n \times n} - H)E\{\epsilon \epsilon' | X\}]$.

Corollary 3. If $E(\epsilon | X) = 0$ and $\text{cov}(\epsilon | X) = \sigma^2 I_{n \times n}$ then $E(\hat{\sigma}^2 | X) = \sigma^2$, where $\hat{\sigma}^2 = \|e\|^2/(n - p)$.

Proof. $E(\epsilon \epsilon' | X) = \sigma^2 I_{n \times n}$ and $\text{trace} (I_{n \times n} - H) = n - p$.

Corollary 4. Suppose $\epsilon$ is independent of $X$, the $\epsilon_i$ are IID, $E(\epsilon_i) = 0$, and $\text{var}(\epsilon_i) = \sigma^2$.

(i) $E(\hat{\beta} | X) = \beta$.

(ii) $\text{cov}(\hat{\beta} | X) = \sigma^2 (X'X)^{-1}$.

(iii) $E(\hat{\sigma}^2 | X) = \sigma^2$, where $\hat{\sigma}^2 = \|e\|^2/(n - p)$.