1. (a) This is just a special case of (b), in which $p = 1$.

   (b) Let $X$ denote the design matrix. The least square estimate of $\boldsymbol{\beta}$ is given by $\hat{\boldsymbol{\beta}} = (X^\top X)^{-1} X^\top Y$, and the fitted regression line is $\hat{y} = (1, x_1, \cdots, x_p)\hat{\boldsymbol{\beta}}$. Let $\mathbf{1}$ be a $1 \times n$ vector whose entries are all 1, then the datum $(1, \bar{x}_{\cdot 1}, \cdots, \bar{x}_{\cdot p})$ can be written as $\frac{1}{n}\mathbf{1}^\top X$ whose fitted value must be

$$\begin{aligned}
\hat{y} &= (1, \bar{x}_{\cdot 1}, \cdots, \bar{x}_{\cdot p})\hat{\boldsymbol{\beta}} \\
&= \frac{1}{n}\mathbf{1}^\top X(X^\top X)^{-1}X^\top Y \\
&= \frac{1}{n}\mathbf{1}^\top PY = \frac{1}{n}(P\mathbf{1})^\top Y \\
&= \frac{1}{n}\mathbf{1}^\top Y = \bar{y},
\end{aligned}$$

   which concludes the desired result. To elaborate $P\mathbf{1} = \mathbf{1}$ in the second-last equality, notice the presence of the intercept in the regression model, which indicates that the design matrix $X$ must have a column of all 1s, so it is guaranteed that $\mathbf{1}$ is in the column space of $X$, and hence the projection of $\mathbf{1}$ to the column space of $X$ is still $\mathbf{1}$.

   *Remark.* Alternatively, one can observe $\hat{\beta}_0 n + \hat{\beta}_1 \sum_{i=1}^n x_{i1} + \cdots + \hat{\beta}_p \sum_{i=1}^n x_{ip} = \sum_{i=1}^n y_i$ from the first entry of the normal equation. Dividing both sides by $n$ gives the desired result.

   (c) If an intercept column is not included in the design matrix $X$, then the column space of $X$ is not guaranteed to include $\mathbf{1}$. In this case, $P\mathbf{1} = \mathbf{1}$ is not true in general, so the results in (b) no longer work.

2. (a) Denote the orthogonal complement of the hat matrix of the $i$-th model as $Q_i = I_n - P_i = I_n - X_i(X_i^\top X_i)^{-1}X_i^\top$, whose trace is $n - p_i$. Under the assumption that $\mathbf{Y} = X_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}$, the residual sum of squares of the $i$-th model would be

$$\begin{aligned}
\|\hat{\boldsymbol{\varepsilon}}_i\|^2 &= \|\mathbf{Y} - \hat{\mathbf{Y}}_i\|^2 = \|Q_i\mathbf{Y}\|^2 \\
&= \|Q_i(X_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon})\|^2 = \|Q_iX_2\boldsymbol{\beta}_2 + Q_i\boldsymbol{\varepsilon}\|^2 \\
&= \|Q_iX_2\boldsymbol{\beta}_2\|^2 + \|Q_i\boldsymbol{\varepsilon}\|^2 + 2(Q_iX_2\boldsymbol{\beta}_2)^\top(Q_i\boldsymbol{\varepsilon}) \\
&= \|Q_iX_2\boldsymbol{\beta}_2\|^2 + \boldsymbol{\varepsilon}^\top Q_i\boldsymbol{\varepsilon} + 2(X_2\boldsymbol{\beta}_2)^\top Q_i\boldsymbol{\varepsilon}.
\end{aligned}$$

   Taking expectation on both sides, with the assumption that $\mathbb{E}[\boldsymbol{\varepsilon}] = \mathbf{0}$, yields

$$\begin{aligned}
\mathbb{E}\|\hat{\boldsymbol{\varepsilon}}_i\|^2 &= \mathbb{E}\|Q_iX_2\boldsymbol{\beta}_2\|^2 + \mathbb{E}[\boldsymbol{\varepsilon}^\top Q_i\boldsymbol{\varepsilon}] + 2(X_2\boldsymbol{\beta}_2)^\top Q_i\mathbb{E}[\boldsymbol{\varepsilon}] \\
&= \|Q_iX_2\boldsymbol{\beta}_2\|^2 + \sigma^2(n - p_i).
\end{aligned}$$

   Note that $\mathbf{v} \in \mathcal{C}(X_i) \iff P_i\mathbf{v} = \mathbf{v} \iff Q_i\mathbf{v} = \mathbf{0} \iff \|Q_i\mathbf{v}\|^2 = 0$ (geometrically, it means that if the projection of a vector onto a plane is the vector itself, then the vector itself must lie on the projection plane, and vice versa). Since $X_2\boldsymbol{\beta}_2 \in \mathcal{C}(X_2) \subsetneq \mathcal{C}(X_3)$, we must have $\|Q_iX_2\boldsymbol{\beta}_2\|^2 = 0$ for $i = 2, 3$, so $\mathbb{E}[\hat{\sigma}_i^2 - \sigma^2] = 0$ when $i = 2, 3$. On the other hand,

$$\mathbb{E}[\hat{\sigma}_1^2 - \sigma^2] = \frac{\|Q_1X_2\boldsymbol{\beta}_2\|^2}{n - p_1} = \frac{\|Q_1(X_2\boldsymbol{\beta}_2 - X_1\boldsymbol{\beta}_1)\|^2}{n - p_1}$$

   is non-zero in general, since $X_2\boldsymbol{\beta}_2 - X_1\boldsymbol{\beta}_1$ is the linear combination of $\mathbf{x}_{p_1+1}, \cdots, \mathbf{x}_{p_2}$, which cannot be in $\mathcal{C}(X_1)\backslash\{\mathbf{0}\}$ (otherwise it will violate the assumption of linear independence of $\mathbf{x}_i$). So $\hat{\sigma}_1^2$ is biased, whereas $\hat{\sigma}_2^2$ and $\hat{\sigma}_3^2$ are unbiased.

(b) Use the assumption $\boldsymbol{\varepsilon} \sim \mathbf{N}(\mathbf{0}, \sigma^2 I_n)$. From the lemma on page 43 of Lecture 2, we have $\boldsymbol{\varepsilon}^\top Q_i \boldsymbol{\varepsilon}/\sigma^2 \sim \chi^2_{n-p_i}$, and hence $\mathrm{Var}(\boldsymbol{\varepsilon}^\top Q_i \boldsymbol{\varepsilon}) = 2\sigma^4(n-p_i)$. Then for each $i$,

$$\begin{aligned}\mathrm{Var}(\|\hat{\boldsymbol{\varepsilon}}_i\|^2) &= \mathrm{Var}(\|Q_i X_2 \boldsymbol{\beta}_2\|^2 + \boldsymbol{\varepsilon}^\top Q_i \boldsymbol{\varepsilon} + 2(X_2\boldsymbol{\beta}_2)^\top Q_i \boldsymbol{\varepsilon}) \\ &= \mathrm{Var}(\boldsymbol{\varepsilon}^\top Q_i \boldsymbol{\varepsilon}) + 4\mathrm{Var}((X_2\boldsymbol{\beta}_2)^\top Q_i \boldsymbol{\varepsilon})) + 4\mathrm{Cov}(\boldsymbol{\varepsilon}^\top Q_i \boldsymbol{\varepsilon}, (X_2\boldsymbol{\beta}_2)^\top Q_i \boldsymbol{\varepsilon}) \\ &= 2\sigma^4(n-p_i) + 4\sigma^2\|Q_i X_2 \boldsymbol{\beta}_2\|^2 + 0,\end{aligned}$$

where the covariance vanishes because $\boldsymbol{\varepsilon}$ is symmetric about $\mathbf{0}$ (i.e. $-\boldsymbol{\varepsilon} \sim \mathbf{N}(\mathbf{0}, \sigma^2 I_n)$) and matrix operations are linear, so $\mathbb{E}[\boldsymbol{\varepsilon}^\top Q_i \boldsymbol{\varepsilon}]\mathbb{E}[(X_2\boldsymbol{\beta}_2)^\top Q_i \boldsymbol{\varepsilon}] = 0$ and

$$\begin{aligned}\mathrm{Cov}(\boldsymbol{\varepsilon}^\top Q_i \boldsymbol{\varepsilon}, (X_2\boldsymbol{\beta}_2)^\top Q_i \boldsymbol{\varepsilon}) &= \mathbb{E}[(\boldsymbol{\varepsilon}^\top Q_i \boldsymbol{\varepsilon})((X_2\boldsymbol{\beta}_2)^\top Q_i \boldsymbol{\varepsilon})] - 0 \\ &= \mathbb{E}[(-\boldsymbol{\varepsilon}^\top Q_i (-\boldsymbol{\varepsilon}))((X_2\boldsymbol{\beta}_2)^\top Q_i (-\boldsymbol{\varepsilon}))] \\ &= -\mathbb{E}[(\boldsymbol{\varepsilon}^\top Q_i \boldsymbol{\varepsilon})((X_2\boldsymbol{\beta}_2)^\top Q_i \boldsymbol{\varepsilon})] \\ &= -\mathrm{Cov}(\boldsymbol{\varepsilon}^\top Q_i \boldsymbol{\varepsilon}, (X_2\boldsymbol{\beta}_2)^\top Q_i \boldsymbol{\varepsilon}) \quad (= 0 \text{ by rearrangement}).\end{aligned}$$

The results above indicate that

$$\mathrm{Var}(\hat{\sigma}_i^2) = \frac{2\sigma^4}{n-p_i} + \frac{4\sigma^2\|Q_i X_2 \boldsymbol{\beta}_2\|^2}{(n-p_i)^2}.$$

From (a) we know that $Q_i X_2 \boldsymbol{\beta}_2 = \mathbf{0}$ for $i = 2, 3$, and thus

$$\mathrm{Var}(\hat{\sigma}_3^2) = \frac{2\sigma^4}{n-p_3} > \frac{2\sigma^4}{n-p_2} = \mathrm{Var}(\hat{\sigma}_2^2).$$

3. (a) Given a general $\mathbf{Y}$ and $X$, choose $\tilde{\mathbf{Y}} = (I_n - \frac{1}{n}J_n)\mathbf{Y}$ and $\tilde{X} = (I_n - \frac{1}{n}J_n)X$, where $J_n$ is the $n \times n$ matrix whose entries are all 1, and it can be checked that $(I_n - \frac{1}{n}J_n) = (I_n - \frac{1}{n}J_n)^\top = (I_n - \frac{1}{n}J_n)^2$. Then $\tilde{\mathbf{Y}} = \tilde{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ is in the deviation form. Since the first column of $\tilde{X}$ becomes 0 after transformation, the intercept is discarded as well.

Now, let $\hat{\boldsymbol{\beta}}$ be the OLS estimate from the original model, and from the arguments in Question 1(b) we know that the OLS estimate satisfies $J_n X\hat{\boldsymbol{\beta}} = J_n\mathbf{Y}$. Then we must have $(\tilde{X}^\top\tilde{X})\hat{\boldsymbol{\beta}} = (X^\top X - \frac{1}{n}X^\top J_n X)\hat{\boldsymbol{\beta}} = X^\top Y - \frac{1}{n}X^\top J_n Y = \tilde{X}^\top\tilde{Y}$, where the first row and first column of $\tilde{X}^\top\tilde{X}$ are zeros. Hence, except for $\beta_0$, all other OLS estimates from the original model will match with the OLS from the transformed model.

(b) Given a general $X$ with rank $p$, $X^\top X$ is invertible and thus we choose $\tilde{\mathbf{Y}} = \mathbf{Y}$ and $\tilde{X} = X(X^\top X)^{-\frac{1}{2}}$, where $(X^\top X)^{-\frac{1}{2}}$ is a square root matrix of $(X^\top X)^{-1}$. Then the new model $\tilde{\mathbf{Y}} = \tilde{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ satisfies $\tilde{X}^\top\tilde{X} = I_p$.

Although the estimates of $\hat{\boldsymbol{\beta}}$ in the new model will be different from the original model, it turns out that their LRT statistics in this context are the same. If we assume the errors are iid from $N(0, \sigma^2)$, then the MLE for $\sigma^2$ under $H_0$ is $\mathbf{Y}^\top\mathbf{Y}$ and the unconstrained MLE for $(\boldsymbol{\beta}, \sigma^2)$ is $((X^\top X)^{-1}X^\top\mathbf{Y}, \mathbf{Y}^\top(I_n - X(X^\top X)^{-1}X^\top)\mathbf{Y})$, so the likelihood ratio is

$$\begin{aligned}\lambda(\mathbf{Y}) &= \frac{(\mathbf{Y}^\top\mathbf{Y})^{-n/2}\exp\{-\frac{\mathbf{Y}^\top\mathbf{Y}}{2\mathbf{Y}^\top\mathbf{Y}}\}}{(\mathbf{Y}^\top(I_n - X(X^\top X)^{-1}X^\top)\mathbf{Y})^{-n/2}\exp\left\{-\frac{\mathbf{Y}^\top(I_n - X(X^\top X)^{-1}X^\top)\mathbf{Y}}{2\mathbf{Y}^\top(I_n - X(X^\top X)^{-1}X^\top)\mathbf{Y}}\right\}} \\ &= \frac{(\mathbf{Y}^\top\mathbf{Y})^{-n/2}}{(\mathbf{Y}^\top(I_n - X(X^\top X)^{-1}X^\top)\mathbf{Y})^{-n/2}} = \frac{(\mathbf{Y}^\top\mathbf{Y})^{-n/2}}{(\mathbf{Y}^\top(I_n - \tilde{X}\tilde{X}^\top)\mathbf{Y})^{-n/2}} \\ &= \frac{(\tilde{\mathbf{Y}}^\top\tilde{\mathbf{Y}})^{-n/2}}{(\tilde{\mathbf{Y}}^\top(I_n - \tilde{X}(\tilde{X}^\top\tilde{X})^{-1}\tilde{X}^\top)\tilde{\mathbf{Y}})^{-n/2}} = \lambda(\tilde{\mathbf{Y}}).\end{aligned}$$

Thus, it suffices to analyze $\tilde{\mathbf{Y}} = \tilde{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, in which the assumption $\tilde{X}^\top \tilde{X} = I_p$ is made.

(c) Given a general positive definite $V$, left-multiply both sides of $\mathbf{Y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ by $V^{-\frac{1}{2}}$ to obtain an equivalent (1-to-1) model $\tilde{\mathbf{Y}} = \tilde{X}\boldsymbol{\beta} + \tilde{\boldsymbol{\varepsilon}}$, where $\mathrm{Var}(\tilde{\boldsymbol{\varepsilon}}) = V^{-\frac{1}{2}}\sigma^2 V V^{-\frac{1}{2}} = \sigma^2 I_n$.

4. (a) It can be checked from induction that $u_i = \sum_{k=1}^{i} \rho^{i-k}\varepsilon_k$ for each $i$, so

$$\mathbf{u} = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix} = \begin{bmatrix} 1 & & & & & \\ \rho & 1 & & & & \\ \rho^2 & \rho & \ddots & & & \\ \vdots & \vdots & & \ddots & & \\ \rho^{n-2} & \rho^{n-3} & \cdots & \rho & 1 & \\ \rho^{n-1} & \rho^{n-2} & \cdots & \rho^2 & \rho & 1 \end{bmatrix} \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix} =: R\boldsymbol{\varepsilon},$$

where it is given that $\mathbb{E}[\boldsymbol{\varepsilon}] = \mathbf{0}$ and $\mathrm{Var}(\boldsymbol{\varepsilon}) = \sigma^2 I_n$. Therefore,

$$\mathbb{E}[\mathbf{u}] = R\mathbb{E}[\boldsymbol{\varepsilon}] = \mathbf{0}, \quad \mathrm{Var}(\mathbf{u}) = R\,\mathrm{Var}(\boldsymbol{\varepsilon})R^\top = \sigma^2 RR^\top,$$

where $[RR^\top]_{i,i+h} = [RR^\top]_{i+h,i} = \dfrac{\rho^h(1-\rho^{2i})}{1-\rho^2}$ for $i \geq 1, h \geq 0$.

(b) The unbiasedness of $\hat{\boldsymbol{\beta}}$ is due to $\mathbb{E}[\hat{\boldsymbol{\beta}}] = (X^\top X)^{-1}X^\top(X\boldsymbol{\beta} + \mathbb{E}[\mathbf{u}]) = \boldsymbol{\beta}$. On the other hand, $\mathrm{Var}(\hat{\boldsymbol{\beta}}) = (X^\top X)^{-1}X^\top \mathrm{Var}(\mathbf{u})X(X^\top X)^{-1} = \sigma^2(X^\top X)^{-1}X^\top RR^\top X(X^\top X)^{-1}$.

(c) Treat $u_i = \rho u_{i-1} + e_i$ as another linear regression model and we fit the model with $\hat{u}_i$, $i = 0, \cdots, n$ (for convention set $\hat{u}_0 = 0$). That is, we observed

$$\begin{bmatrix} \hat{u}_1 \\ \hat{u}_2 \\ \vdots \\ \hat{u}_n \end{bmatrix} = \rho \begin{bmatrix} \hat{u}_0 \\ \hat{u}_1 \\ \vdots \\ \hat{u}_{n-1} \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}$$

and hence an estimate for $\rho$ can be obtained through the OLS: $\hat{\rho} = \dfrac{\sum_{i=1}^{n} \hat{u}_i \hat{u}_{i-1}}{\sum_{i=1}^{n} \hat{u}_{i-1}^2}$.

(d) Rearranging terms will give us $\varepsilon_i = u_i - \rho u_{i-1}$ for each $i$. In matrix form, this is saying that

$$\boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix} = \begin{bmatrix} 1 & & & & & \\ -\rho & 1 & & & & \\ 0 & -\rho & 1 & & & \\ \vdots & \vdots & & \ddots & & \\ 0 & 0 & \cdots & -\rho & 1 & \\ 0 & 0 & \cdots & 0 & -\rho & 1 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix} =: D\mathbf{u}\,(= R^{-1}\mathbf{u}).$$

Therefore $(RR^\top)^{-1} = D^\top D$, and we can obtain another estimator of $\boldsymbol{\beta}$ through the GLS, in which

$$\tilde{\boldsymbol{\beta}} = (X^\top \hat{\mathbf{D}}^\top \hat{\mathbf{D}} X)^{-1} X^\top \hat{\mathbf{D}}^\top \hat{\mathbf{D}} \mathbf{Y},$$

where $\hat{\mathbf{D}} = D|_{\rho=\hat{\rho}}$ is the plug-in estimator of $D$ using the estimator in (c).