## MATH-GA.2711 Machine Learning & Computational Statistics

# Homework 3

Prof. Ivailo Dimov
Due: February 16, 2026

# Instruction

This homework is to be done individually. No collaboration and/or code sharing permitted.

# Questions

1. The classical linear regression model.
   (a) Show that $\mathbf{P} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ is a symmetric and orthogonal projection from $\mathbb{R}^n \to \mathcal{R}(\mathbf{X})$.
   (b) Using NumPy generate a random full-rank design matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$ (e.g. $n = 50, p = 5$), form $\mathbf{P}$ as above. Compute and report: (i) $\|\mathbf{P} - \mathbf{P}'\|_F$ (ii) $\|\mathbf{P}^2 - \mathbf{P}\|_F$ (iii) for a random $\mathbf{y}$, verify $\mathbf{X}'(\mathbf{y} - \mathbf{P}\mathbf{y}) \approx 0$ Also report rank($\mathbf{P}$) via `np.linalg.matrix_rank` and compare to $p$.
   (c) Can the $R^2$ from a classical linear regression be negative? If yes, provide an example. If no, prove it is $\geq 0$.
   (d) Show that for a symmetric matrix $\mathbf{A}$ the following identity holds

$$\nabla_{\boldsymbol{\beta}} \boldsymbol{\beta}' \mathbf{A} \boldsymbol{\beta} = 2\mathbf{A}\boldsymbol{\beta}\,.$$

   (e) Explain the difference between fixed and random regressors.
   (f) Simulate in NumPy two scenarios for $\mathbf{X}$ with $n = 200, p = 3$: (i) *fixed* $\mathbf{X}$: draw once and hold it fixed across $B$ Monte Carlo repetitions, (ii) *random* $\mathbf{X}$: redraw $\mathbf{X}$ each repetition. Generate $\mathbf{y} = \mathbf{X}\beta + \varepsilon$ with $\varepsilon \sim \mathcal{N}(0, \sigma^2 I)$, compute $\widehat{\beta}$ each time, and compare: (a) the empirical $\mathrm{Var}(\widehat{\beta} \mid \mathbf{X})$ in case (i) to $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$, (b) the unconditional variance across repetitions in case (ii). Summarize what changes conceptually between conditioning on $\mathbf{X}$ and averaging over $\mathbf{X}$.

2. Finite-sample properties of OLS.

(a) List the assumptions under which $s^2$ is an unbiased estimator of $\sigma^2$. Then prove that $s^2$ is unbiased under these assumptions.

(b) Fix $n = 80, p = 5$ and in NumPy draw a random full-rank $\mathbf{X}$ and random $\beta$. Set $\sigma^2 = 2$. For $B = 5000$ repetitions, simulate $\varepsilon \sim \mathcal{N}(0, \sigma^2 I)$, set $\mathbf{y} = \mathbf{X}\beta + \varepsilon$, compute $\widehat{\beta}, \quad \hat{\varepsilon} = \mathbf{y} - \mathbf{X}\widehat{\beta}$ and $s^2$. Demonstrate $s^2$ is unbiased estimator of $\sigma^2$. What would need to change in the above setup to create a biased estimator? Demonstrate numerically.

(c) List the assumptions under which

$$\mathrm{var}[\widehat{\boldsymbol{\beta}}|\mathbf{X}] = \sigma^2(\mathbf{X}'\mathbf{X})^{-1} . \tag{1}$$

Then show this assertion.

(d) Assume the classical linear regressions assumptions (I)-(IV) hold. Show

$$\mathrm{cov}[\widehat{\boldsymbol{\beta}}, \mathbf{r}|\mathbf{X}] = 0 . \tag{2}$$

(e) For fixed $\mathbf{X}$, simulate $B$ datasets in NumPy as above, and for each compute $\widehat{\boldsymbol{\beta}}$, $\hat{\mathbf{y}}$, $\mathbf{r}$ Estimate the cross-covariance matrix between $\widehat{\boldsymbol{\beta}}$ and $\mathbf{r}$:

$$\widehat{C} = \frac{1}{B-1} \sum_{b=1}^{B} (\widehat{\boldsymbol{\beta}}^{(b)} - \overline{\widehat{\boldsymbol{\beta}}})(\mathbf{r}^{(b)} - \overline{\mathbf{r}})' \in \mathbb{R}^{p \times n}.$$

Report $\|\widehat{C}\|_F$ (or $\max |\widehat{C}_{ij}|$) and verify it shrinks as $B$ increases.

3. Hypothesis testing.

(a) (i) Describe what is a linear hypothesis in the context of the classical linear regression model

$$y = \mathbf{x}'\boldsymbol{\beta} + \varepsilon , \tag{3}$$

where $y \in \mathbb{R}$ and $\varepsilon \in \mathbb{R}$ are random variables, $\mathbf{x} \in \mathbb{R}^p$ is a random vector and $\boldsymbol{\beta} \in \mathbb{R}^p$ are the parameters. (ii) How do you test this hypothesis using the $F$-ratio?

(b) Using (a) describe how you would test

$$H_0 : \beta_2 = \ldots = \beta_p = 0 . \tag{4}$$

(c) List the assumptions under which the $F$-ratio

$$F|\mathbf{X}, H_0 \sim F_{(d,n-p)} . \tag{5}$$

Then show this assertion.