Question 1. (40 points)

The dataset prostate comes from a study on 97 men with prostate cancer who were due to receive radical prostatectomy. The description of each variable can be found at https://rafalab.github.io/pages/649/prostate.html. Use data(prostate, package="faraway") to import this dataset and answer the following questions.

a. Fit a regression model with lpsa as the response and lcavol as the predictor. Show the R^2 of this model.

b. Add lweight, svi, lbph, age, lcp, pgg45 and gleason as predictors to the regression model. Show the R^2 of this model.

c. Compare the R^2 of these two models. Explain why you observe such a comparison result.

d. Use the method introduced in lecture slides to manually fit the model in b. Construct a matrix X, a response vector y, and then obtain the Least Squares estimator. Compare the manually estimated parameters with the result from the lm function.

e. Consider the model in part b. For each parameter associated with a predictor, conduct the following hypothesis test ($\alpha = 0.05$) using the manual method in lecture notes.

$$H_0: \beta_i = 0$$

$$H_1: \beta_i \neq 0$$

f. Use lm and summary in R to do the above test. Are the test statistics the same as those computed in e)?

g. Compute a 95% CI for the parameter associated with each predictor.

Question 2. (40 points)

Let $y \in \mathbb{R}^n$ and $X \in \mathbb{R}^{n \times p}$ with $n \geq p$. Consider the linear model

$$y = X\beta + \varepsilon,$$

where $\beta \in \mathbb{R}^p$ is unknown. Assume:

1. X is a fixed design matrix with full column rank p so $X^\top X$ is invertible.
2. $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$ with $\sigma^2 > 0$ (equivalently, ε_i are i.i.d. $\mathcal{N}(0, \sigma^2)$).

(a) Derive the OLS estimator.

Starting from $\min_\beta \|y - X\beta\|_2^2$, show that the unique minimizer is

$$\widehat{\beta} = (X^\top X)^{-1} X^\top y.$$

(b) Unbiasedness.

Show that $\mathbb{E}[\widehat{\beta}] = \beta$.

(c) Variance.

Show that $\mathrm{Var}(\widehat{\beta}) = \sigma^2 (X^\top X)^{-1}$.

(d) Distribution.

Using that linear transformations of multivariate normal vectors are normal, prove that

$$\widehat{\beta} \sim \mathcal{N}\big(\beta,\ \sigma^2 (X^\top X)^{-1}\big).$$

Hint:

You may use: if $Z \sim \mathcal{N}(\mu, \Sigma)$ and A is fixed, then $AZ \sim \mathcal{N}(A\mu,\ A\Sigma A^\top)$.

Question 3. (20 points)

You have regressed y on variables $x_1, x_2, \ldots, x_p$. Your friend, Bob, has regressed y on the variables $z_1, z_2, \ldots, z_p$, where

$$z_j = c_{j0} + \sum_{k=1}^{p} c_{jk} x_k$$

That is, Bob has applied a linear transformation to the predictors (but not to the response).

(a). Show that Bob's n × (p + 1) design matrix Z is related to yours via $Z = XC$ for some (p + 1) × (p + 1) matrix C; explain how the entries in C are related to coefficients c_{jk}.

(b). Show that the predictions from your model and Bob's model are exactly equal, if C is invertible.