Part I

1) A biologist wants to test if there is any linear relationship between the amount of fertilizer supplied to aubergine plants and the subsequent yield of aubergines. She selected 2 aubergine plants of the same variety and treated them weekly with a solution in which *x* grams of fertilizer was dissolved in a fixed quantity of water. The yield, *y* kilograms, of aubergines was recorded. (10)

Aubergine	x	y
Plant A	1	1.5
Plant B	3.9	4.4

a) Form an estimated regression model.
b) Find the estimates using OLS.

2) Carina obtains cash from an ATM (cash machine). She suspects that the rate at which she spends money is affected by the amount of cash she withdrew at her previous visit to an ATM. To investigate this she deliberately varies the amount she withdraws. She records, for each visit to an ATM, the amount, *$x*, withdrawn, and the number of hours, *y*, until her next visit to an ATM. (10)

Withdrawal	1	2	3	4	5	6	7	8	9	10
x	40	10	100	110	120	150	20	90	80	130
y	56	62	195	330	94	270	48	196	214	286

a) What is the correlation between the amount of cash withdrawn and number of hours until next visit to the ATM?
b) Interpret your results.
c) What is the difference between correlation and covariance?

Part II

3) Using wage data on 250 randomly selected male workers and 280 female workers, a researcher estimates the OLS regression, where wage is measured in $/hour and Male is a binary variable that is equal to 1 if the person is a male and 0 if the person is a female. (10)

$$\widehat{wage} = 12.52 - 2.12 \times Male \qquad R^2 = 0.06$$

a) What do you understand by binary variable?
b) What is the slope coefficient in this equation?
c) What is the estimated gender gap?
d) What is the mean wage of men and women?

4) For a sample of 306 students in a basic business communications course, the sample regression line (10)

$\hat{y} = 58.813 + 0.2875x$ was obtained. Here

y = final student score at the end of the course

x = score on a diagnostic writing skills test given at the beginning of the course

The coefficient of determination was 0.1158 and the estimated standard deviation of the estimator of the slope of population regression line was 0.04566.

a) Define and interpret the estimates of the regression equation.
b) Interpret the coefficient of determination.
c) What will happen to the final student score if his diagnostic writing test score increases by 2?
d) Test the null hypothesis that the slope of the population regression line is 0 against the alternative that the true slope is positive. Interpret your result. (Use 5% significance level)
e) How can you improve this model?

5) Research topic (10)
'Relationship between population density and violence in the US'
Explain how you would conduct an empirical analysis on the above topic using the 'Guns' data set and data description PDF posted with this assignment. Your answer must include
- Population regression equation

- Variables you need to use
- Detailed description of the steps you need to carry out in STATA
- Estimated regression equation
- **You do not need to show STATA results**