Assignment 2

Samuel Fraley
Eric Gutierrez
Corneel Moons

October 9, 2025

Contents

Question 1: Distances and Scale Transformations

(a) Compute the squared Euclidean distance $d_E(v_0, \bar{v})$ and the squared Mahalanobis distance $d_M(v_0, \bar{v})$ between

$$v_0 = \begin{bmatrix} 2 \\ 5 \\ 3 \end{bmatrix}, \qquad \bar{v} = \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix}, \qquad S = \begin{bmatrix} 5 & 1 & 2 \\ 1 & 6 & 1 \\ 2 & 1 & 7 \end{bmatrix}.$$

Include the functions (in `R`/`Python`) you used to compute both distances and report their values here.

(b) Suppose v_1 is rescaled so that each value is multiplied by 10. (i) Report the new $\bar{v}$, v_0, and S matrices. (ii) Recalculate $d_E(v_0, \bar{v})$ and $d_M(v_0, \bar{v})$ using your functions. Comment briefly on the effect of the rescaling on each distance measure.

Question 2: Effect of Units on OLS Estimation

We estimate the model:

$$y = \beta_1 + \beta_2 \ln(x_2) + \beta_3 x_3 + \varepsilon, \quad \hat{\beta} = (X'X)^{-1}X'y.$$

(a) Define matrix A such that $X^* = XA$ after rescaling x_2 and x_3 by constants a and b respectively.

(b) Derive the relationship between $\hat{\beta}^*$ and $\hat{\beta}$. Comment on how a change in the units of measurement affects each estimated parameter. You may use symbolic computation (`sympy` in Python or equivalent) to show the steps.

Question 3: Sample Verification of the Linear Model Equivalence

Consider the sample regression

$$\ln(wage) = \beta_1 + \beta_2\, educ + \varepsilon,$$

using data `dataFig311.csv` (Angrist & Pischke, 1980s U.S. sample).

(a) Estimate the regression using OLS and report $\hat{\beta}_1$ and $\hat{\beta}_2$.

(b) Compute conditional means of $\ln(wage)$ for each level of `educ`, and show that running the regression on these conditional means yields the same fitted line as using all individual observations. Include the weighted averages and plots as appropriate.

Question 4: Social Media and Corruption (Enikolopov et al., 2018)

We analyze the regression (R1):

$$Corruption = \beta_1 + \beta_2 \ln(gdp) + \beta_3\, smedia + \varepsilon,$$

using the dataset `corruption.csv` with $n = 35$ countries.

(a) **OLS estimation by formula.** Estimate parameters with the closed-form expression

$$\hat{\beta} = (X'X)^{-1}X'y,$$

and compute $SSE(\hat{\beta})$, $y'M_Xy$, and R^2.

(i) Calculate the associated sum of squared errors, $SSE(\hat{\beta})$, and R^2. Provide the values of $\hat{\beta}$, $SSE(\hat{\beta})$, $y'M_Xy$, and R^2 as your answer.

$\hat{\beta}_1$ (Intercept)	100.57
$\hat{\beta}_2$ ($\ln(gdp)$)	-0.38
$\hat{\beta}_3$ ($smedia$)	-0.00
$SSE(\hat{\beta})$	154.92
$y'M_Xy$	154.92
R^2	0.75

Table 1: Closed–form OLS results for regression (R1).

(ii) Are you surprised about the value of $SSE(\hat{\beta})$ compared to $y'M_Xy$?

No, we are not surprised that $SSE(\hat{\beta})$ is equal to $y'M_Xy$

(b) **Relative importance via Shapley value decomposition.**

Insert the printed Shapley values and include your generated plot below.

Q4b_shapley_plot.png

Comment briefly on which regressor contributes more to R^2 and interpret economically.

(c) **Frisch–Waugh–Lovell (FWL) regression (R2).**

Write explicitly your (R2) regression expression and verify that $\hat{\beta}_3$ here matches the $\hat{\beta}_3$ from (R1).

(d) **Interpretation of the added-variable plot.** Explain what variables are on each axis of Figure 1 (from the paper), why both axes can take positive and negative values, and what the plotted relationship represents in terms of residuals.

(e) **Interpretation of the slope.** State—in one rigorous sentence—what the slope of the added-variable plot says about the relationship between social media penetration and corruption.

References

- Angrist, J.D. & Pischke, J.S. (2009). *Mostly Harmless Econometrics: An Empiricist's Companion*.

- Enikolopov, R., Makarin, A., Petrova, M. (2018). “Social Media and Corruption.” *American Economic Journal: Applied Economics*.