Problem Set 2

Due on Mar 18

Please note

- Submit your answers (in Microsoft Word or PDF format) and your code to canvas. Your answer shall be well written. Graph and Table shall be well-formatted. Your code shall be easy for TA to run and check. Your grade will be affected if your code does not provide proper outputs, or it is confusing so that TA cannot run it.
- Use proper regression model to analyze the empirical questions. Report the result both by words and regression results. Use Table and Figures to report the result wherever necessary.

1. Consider the dataset "SAT.csv"

a. Load this dataset into R. Compile a summary statistics table for all numerical and dummy variables for this data set.

b. Make a plot with 2×2 diagrams that show histograms of GPA, SAT, APMath, and APEng.

c. Make a scatter plot of SAT and GPA. Add a linear regression line and a non-linear (lowess) regression line on the plot. [Hint: see https://www.statmethods.net/graphs/scatterplot.html]

d. Use matrix algebra to compute a least square estimator for regression model:

$$\text{GPA} = \beta_1 + \beta_2 \text{SAT} + \beta_3 \text{APMath} + \beta_4 \text{APEng} + \beta_5 \text{ESL} + \beta_6 \text{gender} + \beta_7 \text{race} + \varepsilon.$$

You may follow these steps:

(i) Generate a vector with 65 one's as the constant regressor.

(ii) Make a X matrix that each regressor constitute a column. The matrix shall be 65×7.

(iii) Make a column vector y using GPA.

(iv) Compute $\hat{\beta}_{OLS} = (X'X)^{-1}X'y$.

(v) Compare with the result from build-in least-square function "lm".

2. Consider the dataset drug_price.csv. This is a data set about the price information of a medicine produced by a pharmaceutical company. This drug is sold in 32 countries and the Competition Commission is investigating whether the company practice international price discrimination. Here are the variables:

p.r	price of the drug in country *i* relative to U.S. price
cv	consumption volume of the drug in country *i*
cv.r	overall consumption volume of drugs in country *i* relative to U.S
GDP.r	per capita GDP of country *i* relative to U.S.
p.control	dummy for price control in country *i*
p.comp	dummy for price competition is encouraged in country *i*
patent	dummy for whether the drug is protected by patent in country *i*

Investigate whether the pharmaceutical company try to sell the drug for higher price in countries with higher per capita GDP. Use several regression results and determine the best linear regression model. Use a scatter plot to show the result.

3. In 2014, Hong Kong government launched a funding program to support innovative startups founded by university students, alumni, and professors. The program is called Technology Start-up Support Scheme for Universities (TSSSU). See `www.itf.gov.hk/en/funding-programmes/supporting-start-ups/tsssu/tsssu-directory/index.html`.

The dataset TSSSU.tex covers all startups supported by the program from 2014 to 2020. The variable *amount* is the funding amount in million HKD; *area* is the technological area; *survive* is the survival status from company registry record; *social_media* and *phone_call* are proxies indicating whether the startup is still active; *No_member, No_alumnus, No_professors*, *No_postgrad*, and *No_undergrad* are the composition of the founding team; *Employee* is the number of employees reported by some startups when they applied for COVID-19 Employment Support Scheme (`www.ess.gov.hk/en`); *private_fund* indicates whether the startup receives follow-up private investment; and *Sciencepark* indicates whether the startup receives follow-up support from Hong Kong Science Park (`www.hkstp.org`).

a. Produce a summary statistics table for all numerical variables.

b. Produce two figures. The first one is a histogram of the funding amount. The second one is a scatter plot showing the relationship between funding amount and number of employees.

c. Consider the following regression

$$y_i = \beta_1 + \beta_2 amount_i + \beta_3 No.undergrad_i + \beta_4 No.postgrad_i + \beta_5 No.professor_i + Univ.FE + year.FE + area.FE + \varepsilon_i,$$

where $Univ.FE$, $year.FE$, and $area.FE$ are fixed effects for university, year, and technological area, respectively.

Run six regressions with different dependent variables: *survive, Employee, social_media, phone_call, private_fund,* and *Sciencepark.* Report the results in one table. In the table, please report the results of *Univ.FE,* but do not report the coefficient estimates of *year.FE* and *area.FE.*

d. Replicate the following figure that illustrates the funding distribution across different areas.

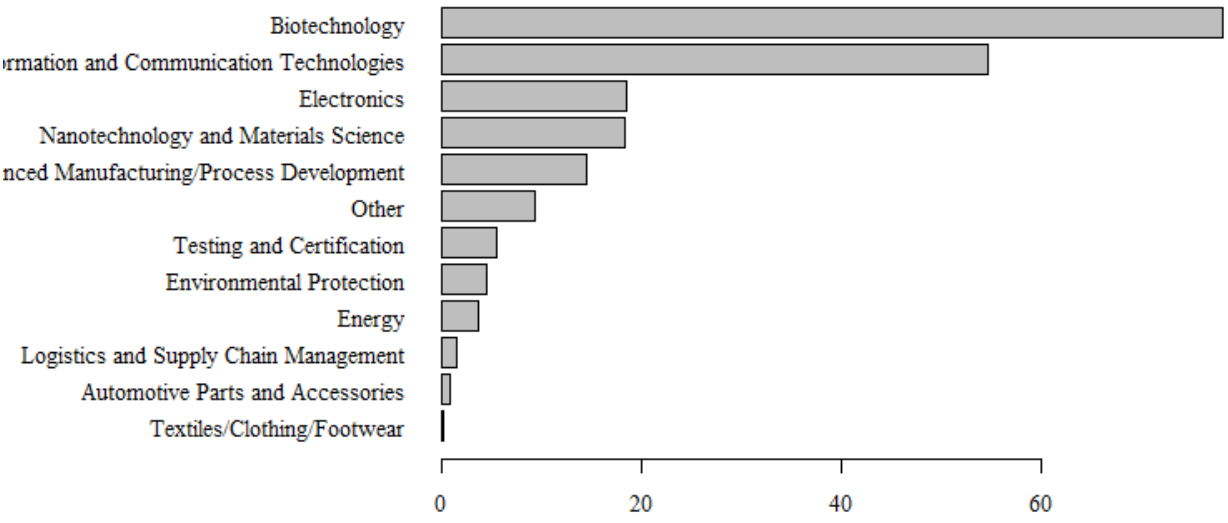

Figure 1: TSSSU Funding Distribution Acorss Areas