Assignment #2

Paper Design, Model Description, Justification of Parameters, Evaluation, Evidence, & Reflection

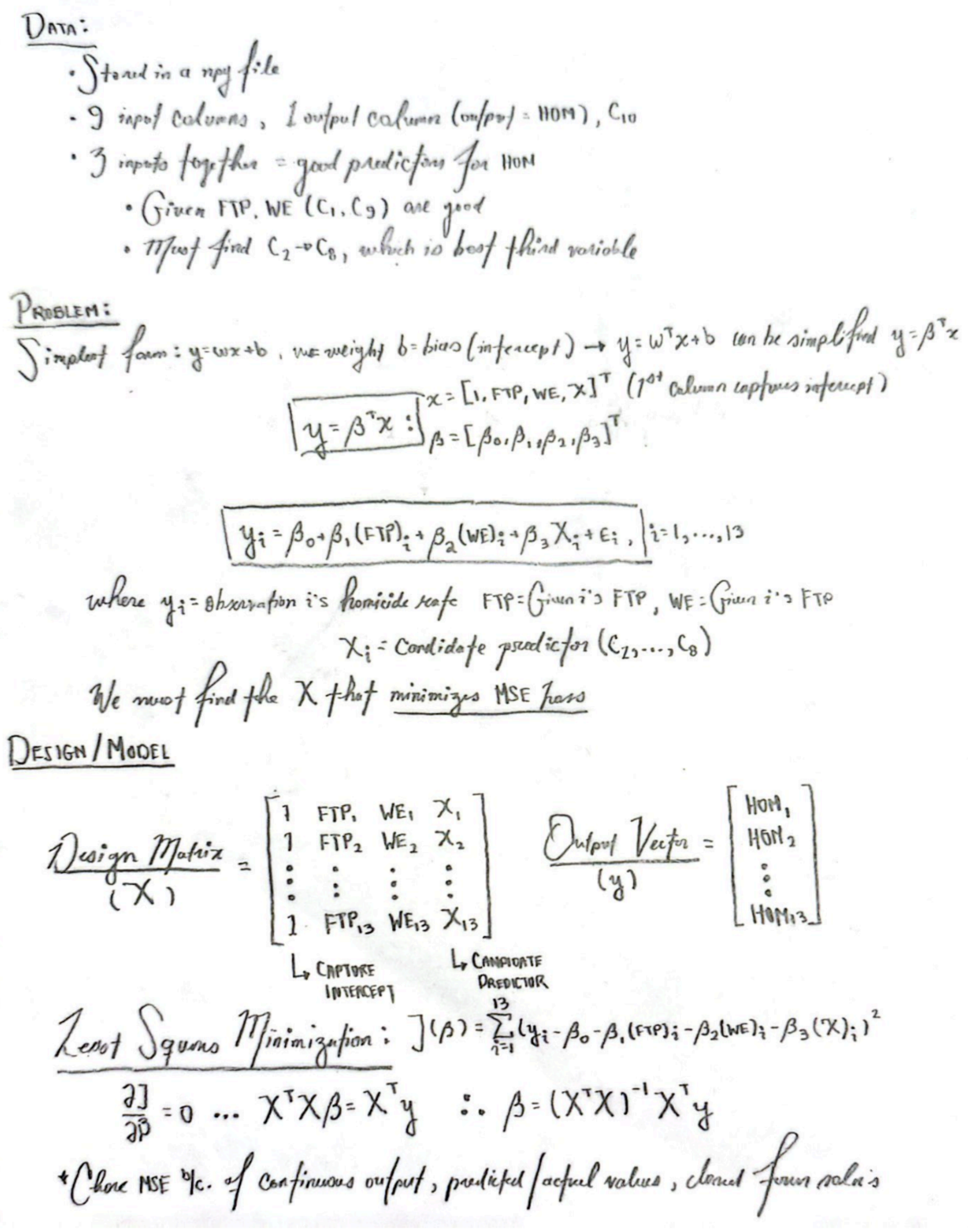
ENGS 106

Taka Khoo

Due Monday, Feb. 3: 8:00pm

ENGS 106: Assignment 2 (Coding) Written Portion

Part I: Paper Design (Scanned)



DATA:

- Stored in a npy file
- 9 input columns, 1 output column (output = HOM), C_{10}
- 3 inputs together = good predictions for HOM
  - Given FTP, WE (C_1, C_9) are good
  - Must find $C_2 \to C_8$, which is best third variable

PROBLEM:

Simplest form: $y = wx + b$, w= weight b= bias (intercept) → $y = w^T x + b$ can be simplified $y = \beta^T x$

$$\boxed{y = \beta^T x :} \quad \begin{aligned} x &= [1, FTP, WE, X]^T \text{ (1st column captures intercept)} \\ \beta &= [\beta_0, \beta_1, \beta_2, \beta_3]^T \end{aligned}$$

$$\boxed{y_i = \beta_0 + \beta_1 (FTP)_i + \beta_2 (WE)_i + \beta_3 X_i + \epsilon_i ,} \; i = 1, \dots, 13$$

where y_i = observation i's homicide rate FTP = Given i's FTP, WE = Given i's FTP

X_i = Candidate predictor ($C_2, \dots, C_8$)

We must find the X that <u>minimizes MSE loss</u>

DESIGN / MODEL

$$\frac{\text{Design Matrix}}{(X)} = \begin{bmatrix} 1 & FTP_1 & WE_1 & X_1 \\ 1 & FTP_2 & WE_2 & X_2 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & FTP_{13} & WE_{13} & X_{13} \end{bmatrix} \qquad \frac{\text{Output Vector}}{(y)} = \begin{bmatrix} HOM_1 \\ HOM_2 \\ \vdots \\ HOM_{13} \end{bmatrix}$$

↳ CAPTURE INTERCEPT

↳ CANDIDATE PREDICTOR

Least Squares Minimization: $J(\beta) = \sum_{i=1}^{13} (y_i - \beta_0 - \beta_1 (FTP)_i - \beta_2 (WE)_i - \beta_3 (X)_i)^2$

$$\frac{\partial J}{\partial \beta} = 0 \; \dots \; X^T X \beta = X^T y \quad \therefore \beta = (X^T X)^{-1} X^T y$$

*Chose MSE b/c. of continuous output, predicted/actual values, closed form solu's

Prediction / Error:

Predicted Output $(\hat{y}) = X\beta$

Mean Squared Error $(\text{MSE}) = \frac{1}{13}\sum_{i=1}^{13}(y_i - \hat{y}_i)^2$

- measures quality of fit
- repeat for each candidate predictor (UFMP, MAN, LIC, GR, NMAN, GOV, HE),
- select candidate w/ smallest MSE

Code Planning / Notes:

1. Load Data: np.load('detroit.npy')
   - Extract FTP, WE, HOM columns (C_0, C_8, C_9)
   - Candidate Predictors $C_1 \rightarrow C_7$

   } Making 1st Column → C_0 for easier code

2. Helper Functions:
   - Compute_Normal_Equation : $\beta = (X^T X)^{-1} X^T y$
   - Compute_MSE : $\frac{1}{13}\sum_{i=1}^{13}(y_i - \hat{y}_i)^2$

3. Design Matrix & β:
   - Each Row = $[1_i, \text{FTP}_i, \text{WE}_i, X_i]$ Corresp. to model
   - Compute β using helper
   - W/ β's predict $\hat{\text{HOM}}$ and calculate MSE

4.) Model Selection:
   - Candidate w/ smallest MSE = best 3rd variable to include in model
   - Code Plots of actual homicide rates vs. predicted model rates using the best candidate predictor to verify fit
   - Possibly display other candidate models to show they are worse

Part II: Model Description & Justification of Parameters

- I used a linear regression model that would predict the homicide rate, or HOM, in Detrois using three predictors at any given time simultaneously:
    - FTP: Full Time Police per 100,000 population
        - This was the first given indicator that is a “good” predictor for homicide rate
    - WE: Average Weekly Earnings
        - This was the second given indicator that is a “good” predictor for homicide rate
    - X: Candidate Predictor
        - 1 of the 7 remaining potential predictors (UEMP, MAN, LIC, GR, NMAN, GOV, or HE)

As shown in the above design the model was formulated:

$$HOM = \beta_o + \beta_1 FTP + \beta_2 WE + \beta_3 X + \epsilon$$

$$\beta_o = intercept\ (homicide\ rate);\ \beta_1 = effect\ of\ FTP;\ \beta_2 = effect\ of\ WE$$

$$\beta_3 = additional\ influence\ of\ X;\ \epsilon = residual\ error$$

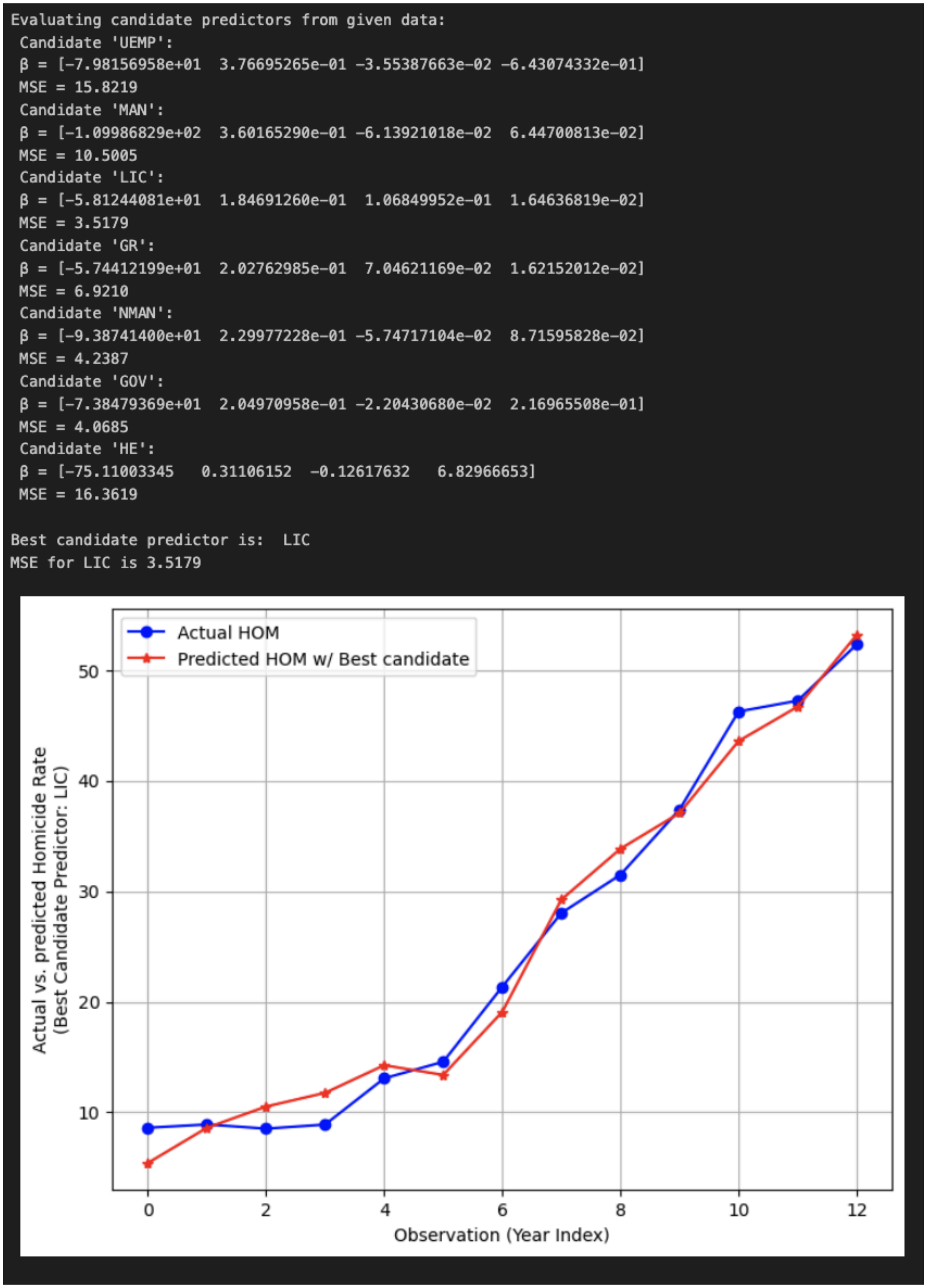
- Parameter Estimation
    - Methodology: I computed the coefficients using the derived normal equation for β which was derived and shown in Part I during my design and mathematical foundation; the normal equation provided a closed-form solution of:

$$\beta = (X^T X)^{-1} X^T y$$

-
    -
        - X was the design matrix, which was, as shown previously, constructed by stacking a columns of 1’s for β_0, the intercept, FTP, WE, and 1 candidate predictor at a time

- Justification
  - β_o adjusts the baseline prediction, as it is the intercept
  - $\beta_1(FTP)$ & $\beta_2(WE)$ are given and known “good” and hence significant predictors from prior research, which is why they appear in the design matrix for all potentials
  - β_3 is the candidate predictor, which is ultimately selected based on its ability to reduce the prediction error when added to the model
- This all works because the normal equation for β minimizes the Mean-Squared-Error (MSE)
  - This ensures that the parameter our model ultimately decide upon yield the best possible fit (in terms of least-squares)

Part III: Model Evaluation & Evidence



```
Evaluating candidate predictors from given data:
 Candidate 'UEMP':
 β = [-7.98156958e+01  3.76695265e-01 -3.55387663e-02 -6.43074332e-01]
 MSE = 15.8219
 Candidate 'MAN':
 β = [-1.09986829e+02  3.60165290e-01 -6.13921018e-02  6.44700813e-02]
 MSE = 10.5005
 Candidate 'LIC':
 β = [-5.81244081e+01  1.84691260e-01  1.06849952e-01  1.64636819e-02]
 MSE = 3.5179
 Candidate 'GR':
 β = [-5.74412199e+01  2.02762985e-01  7.04621169e-02  1.62152012e-02]
 MSE = 6.9210
 Candidate 'NMAN':
 β = [-9.38741400e+01  2.29977228e-01 -5.74717104e-02  8.71595828e-02]
 MSE = 4.2387
 Candidate 'GOV':
 β = [-7.38479369e+01  2.04970958e-01 -2.20430680e-02  2.16965508e-01]
 MSE = 4.0685
 Candidate 'HE':
 β = [-75.11003345   0.31106152  -0.12617632   6.82966653]
 MSE = 16.3619

Best candidate predictor is:  LIC
MSE for LIC is 3.5179
```

Actual HOM
Predicted HOM w/ Best candidate
Actual vs. predicted Homicide Rate
(Best Candidate Predictor: LIC)
10
20
30
40
50
0
2
4
6
8
10
12
Observation (Year Index)

- Evidence of Validity/Quality
    - LIC (# of Handgun licenses per 100,000 population) stands out with the lowest MSE of 3.5179
    - The small MSE implies that the model incorporating LIC as its third predictor in combination with FTP and WE produces predictions that are very close to the actual homicide rates
    - While there are other metrics that exist that can very much so be correlated to homicide rate, there is a subjective sense of validation that comes from the model outputting handgun licenses as the metric. Handguns are the instrument by which much homicide can be attributed to
    - The graph of the HOM produced by my model shows an extremely close correlation to the homicide rates

Part IV: Reflection

Implementing regression from scratch really deepened my understanding of this. Perhaps it is being required to draw out my own design matrices with real data (and a limited amount of it) brought this concept from being a theoretical to a real thing. I initially struggled with designing this, as given the basic linear regression for one variable, I couldn't find a way to implement the intercept for every single variable and with multiple. I was proud to figure out that implementing a row of 1s allowed this to happen, it was actually pretty straight forward looking back retroactively.

It was most interesting to me how I eventually carried this out versus how I initially expected to. I was thinking that I was going to generate some entity for each of the two given metrics, and try to figure out the third based off of that. But jut taking a look at the way regressions are supposed to be implemented, it was surprisingly much more simple to do it altogether in one design matrix. It was really nice to be able to implement the normalization equation, and seeing it really work in action.

Ultimately, I am beginning to see the importance of an 'evidence-based' approach, for this is the best way to build robust predictive models. Overall, I felt that my design and steps helped not only with the mathematical foundation needed for this project, but even with designing the code itself. Making the functions such as MSE beforehand as helpers guided me in the right direction early on and made the coding process much more efficient and organized than it would have been.