MA500 Geometric Foundations of Data Analysis

Each homework should be submitted as a single .pdf document with an accompanying .py file to both Graham Ellis and Emil Sköldberg. The .pdf document should provide your answers, the methods used to obtain your answers, and an appendix listing any Python code used. The .py file should be a machine readable version of the appendix code.

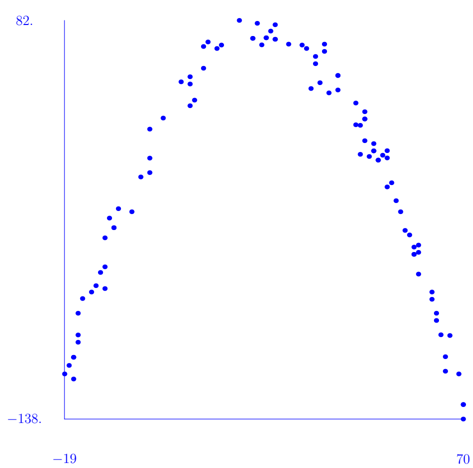
The homework will be graded according to a scheme in which *content* is weighted at 70% and *presentation* is weighted at 30%.

1 First Homework

Please try to submit this by 04.02.2019 as two files:
MA500_First_Homework_firstname_familyname.pdf
MA500_First_Homework_firstname_familyname.py

1.1

The scatter plot



represents a set of points $(x_1, y_1), (x_2, y_2), \ldots, (x_{100}, y_{100})$ produced using a model $y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \epsilon_i$ with independent random errors ϵ_i of mean 0 and finite variance. The numerical values of the points (x_i, y_i) are as follows:

```
x_1 = 70,   y_1 = -130
x_2 = 3,   y_2 = 28.1
x_3 = 67,   y_3 = -91.90000000000003
x_4 = 38,   y_4 = 47.59999999999999
x_5 = 46,   y_5 = 36.39999999999998
x_6 = -16,   y_6 = -91.59999999999999
x_7 = 64,   y_7 = -79.60000000000002
x_8 = 10,   y_8 = 38
x_9 = 55,   y_9 = -17.5
```

```
x_10 = -17,   y_10 = -115.9
x_11 = 51,   y_11 = 4.899999999999977
x_12 = 23,   y_12 = 72.09999999999999
x_13 = 26,   y_13 = 72.39999999999999
x_14 = 12,   y_14 = 67.59999999999999
x_15 = 34,   y_15 = 68.39999999999999
x_16 = 58,   y_16 = -36.40000000000003
x_17 = 0,   y_17 = 6
x_18 = -18,   y_18 = -108.4
x_19 = 9,   y_19 = 34.9
x_20 = -9,   y_20 = -27.1
x_21 = 50,   y_21 = 10
x_22 = 27,   y_22 = 76.09999999999999
x_23 = 50,   y_23 = 14
x_24 = 48,   y_24 = 31.59999999999999
x_25 = 9,   y_25 = 46.9
x_26 = 26,   y_26 = 72.39999999999999
x_27 = 63,   y_27 = -67.90000000000003
x_28 = 66,   y_28 = -111.6
x_29 = 47,   y_29 = 8.099999999999994
x_30 = 60,   y_30 = -42
x_31 = 37,   y_31 = 62.09999999999999
x_32 = -13,   y_32 = -67.90000000000001
x_33 = 48,   y_33 = 27.59999999999999
x_34 = -10,   y_34 = -38
x_35 = 70,   y_35 = -138
x_36 = 20,   y_36 = 82
x_37 = 24,   y_37 = 80.40000000000001
x_38 = 35,   y_38 = 66.5
x_39 = 28,   y_39 = 71.59999999999999
x_40 = 15,   y_40 = 66.5
x_41 = 60,   y_41 = -58
x_42 = 56,   y_42 = -23.60000000000002
x_43 = 59,   y_43 = -43.10000000000002
x_44 = 23,   y_44 = 72.09999999999999
x_45 = 9,   y_45 = 50.9
x_46 = 48,   y_46 = 15.59999999999999
x_47 = 13,   y_47 = 70.09999999999999
x_48 = 51,   y_48 = 4.899999999999977
x_49 = 49,   y_49 = 6.899999999999977
x_50 = 16,   y_50 = 68.40000000000001
x_51 = 36,   y_51 = 44.40000000000001
x_52 = 12,   y_52 = 55.6
x_53 = 42,   y_53 = 43.59999999999999
x_54 = -8,   y_54 = -32.4
x_55 = -15,   y_55 = -71.5
x_56 = 65,   y_56 = -91.5
x_57 = -19,   y_57 = -113.1
x_58 = 7,   y_58 = 48.1
x_59 = 25,   y_59 = 68.5
x_60 = -16,   y_60 = -79.59999999999999
x_61 = -10,   y_61 = -54
x_62 = 31,   y_62 = 68.89999999999999
x_63 = 39,   y_63 = 64.90000000000001
```

```
x_64 = 70,   y_64 = -130
x_65 = 42,   y_65 = 51.59999999999999
x_66 = 53,   y_66 = -9.900000000000034
x_67 = 59,   y_67 = -47.10000000000002
x_68 = -17,   y_68 = -103.9
x_69 = 54,   y_69 = -7.600000000000023
x_70 = -16,   y_70 = -95.59999999999999
x_71 = -17,   y_71 = -103.9
x_72 = 53,   y_72 = 6.099999999999966
x_73 = 42,   y_73 = 51.59999999999999
x_74 = -10,   y_74 = -66
x_75 = 37,   y_75 = 58.09999999999999
x_76 = 69,   y_76 = -113.1
x_77 = 48,   y_77 = 27.59999999999999
x_78 = -8,   y_78 = -32.4
x_79 = 59,   y_79 = -47.10000000000002
x_80 = 28,   y_80 = 71.59999999999999
x_81 = 63,   y_81 = -71.90000000000003
x_82 = 0,   y_82 = 22
x_83 = 64,   y_83 = -83.60000000000002
x_84 = 66,   y_84 = -103.6
x_85 = 50,   y_85 = 10
x_86 = -7,   y_86 = -21.9
x_87 = 39,   y_87 = 68.90000000000001
x_88 = 47,   y_88 = 24.09999999999999
x_89 = 46,   y_89 = 24.39999999999998
x_90 = 53,   y_90 = 10.09999999999997
x_91 = 40,   y_91 = 42
x_92 = -2,   y_92 = -4.4
x_93 = 60,   y_93 = -46
x_94 = -11,   y_94 = -57.1
x_95 = -4,   y_95 = -23.6
x_96 = 0,   y_96 = -2
x_97 = -12,   y_97 = -64.40000000000001
x_98 = 28,   y_98 = 79.59999999999999
x_99 = 57,   y_99 = -33.90000000000003
x_100 = 52,   y_100 = 7.599999999999966
```

1. Determine the values of b_0, b_1, b_2 for which

$$y = b_0 + b_1 x + b_2 x^2$$

   is the least squares estimator for the model $y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \epsilon_i$.

2. Exhibit a single plot of the data points (in say blue) and the curve $y = b_0 + b_1 x + b_2 x^2$ (in say red).

3. Determine the coefficient of determination $r^2 = 1 - (SSE/SSTO)$ for this least squares fit.

1.2

The observations below, taken on 10 incoming shipments of chemicals in drums arriving at a warehouse, show number of drums in shipment (x_1), total weight of shipment (x_2, in hundred pounds), and number of man-minutes required to handle the shipment (y_i):

i :	1	2	3	4	5	6	7	8	9	10
x_{i1} :	7	18	5	14	11	5	23	9	16	5
x_{i2} :	5.11	16.70	3.20	7.00	11.00	4.00	22.10	7.00	10.60	4.80
y_i :	58	152	41	93	101	38	203	78	117	44

1. Assume a model

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i \tag{1}$$

   in which errors are independent $N(0, \sigma^2)$.

   (a) Determine the least squares estimator $y = b_0 + b_1 x_1 + b_2 x_2$.

   (b) Test whether there is a regression equation, using a level of significance of 0.05.

   (c) Estimate β_1 and β_2 jointly, using a 95% family confidence coefficient.

   (d) Management desires simultaneous interval estimates of the mean handling times for five typical shipments specified to be as follows:

	1	2	3	4	5
x_1 :	5	6	10	14	20
x_2 :	3.20	4.80	7.00	10.00	18.00

   Obtain the family of estimates, using a 90 family confidence coefficient.

2. Obtain the residuals and make appropriate residual plots to ascertain whether model (1) with normal error terms is appropriate. Summarize your findings.

2 Second Homework

Please try to submit this by 25.02.2019 as two files:
MA500_Second_Homework_firstname_familyname.pdf
MA500_Second_Homework_firstname_familyname.py

2.1

The online article *Face Recognition with Python* by Philipp Wagner provides guidance for this assignment.

1. Downlaod the AT&T Facedatabase, details of which can be found in the online article. Import the images (as vectors) into Python and perform a principal component analysis. Let $P(n)$ denote the vector space generated by those eigenvectors corresponding to the n largest eigenvalues. For $n = 10, 50, 100$ and 300 determine how much of the variability of the database is captured by projecting onto $P(n)$?

2. Take an image of yourself and store it in the same format as the AT&T images. Display, as an image (rather than a vector), the projection of your original image onto $P(n)$ for $n = 10, 50, 100$ and 300.

3. Take an image of a friend and determine the distance between the projections of your own image and your friend's image onto $P(300)$. Specify which metric you are using to compute this distance.

3 Third Homework

Please try to submit this by 04.03.2019 as two files:
MA500_Third_Homework_firstname_familyname.pdf
MA500_Third_Homework_firstname_familyname.py

1. Implement an algorithm that applies single-linkage hierarchical clustering to an $n \times n$ matrix of distances (or dissimilarities) and returns the corresponding barcode.

2. Create a sample S of n points in $\mathbb{R}^2$ that are clearly partitioned into several distinct 'clusters'. Plot the points S.

3. For the Euclidean metric, and then the taxicab metric, construct the two $n \times n$ distance matrices for your set S of points.

4. Apply your implementation to the two matrices in (3) and display the resulting barcodes.