Lecture 33 Worksheet

Chrysafis Vogiatzis

Every worksheet will work as follows.

1. You will be entered into a Zoom breakout session with other students in the class.
2. Read through the worksheet, discussing any questions with the other participants in your breakout session.
   - You can call me using the "Ask for help" button.
   - Keep in mind that I will be going through all rooms during the session so it might take me a while to get to you.
3. Answer each question (preferably in the order provided) to the best of your knowledge.
4. While collaboration between students in a breakout session is highly encouraged and expected, each student has to submit their own version.
5. You will have 24 hours (see gradescope) to submit your work. This is no longer valid. All worksheets are now flexibly due 2 days before the corresponding exam.

Worksheet 1: Respiratory function

In this first set of problems, we tackle linear regression before finally seeing how a quadratic version works.

The following table contains information about respiratory function (as measured by forced expiratory volume) and smoking. The dataset contains information about age (x_1, note that all subjects are between 13 and 19 years old), height (x_2), and whether they smoke (1) or not (0) (x_3). The last column called FEV measures forced expiratory volume and is our dependent variable (y).

ID	Age (x_1)	Height (x_2)	Smoking (x_3)	FEV (y)
1	13	67	1	3.994
2	13	61	0	3.208
3	14	64.5	0	2.997
4	14	72.5	1	4.271
5	16	72	1	4.872
6	16	63	0	2.795
7	19	72	1	5.102
8	19	66	0	3.519
9	18	60	0	2.853
10	17	70.5	1	4.724
11	16	69.5	1	4.070

Problem 1: Simple linear regression

First perform a linear regression between height (x_2) and FEV (y). What is the adjusted R^2 score?

Answer to Problem 1.

Problem 2: Simple quadratic regression

Does the adjusted R^2 score improve if we perform a regression on height *squared* (x_2^2) and FEV (y)? [1]

Answer to Problem 2.

[1] Recall what we saw in the notes: create a new column with **only** x_2^2 values and use that one!

Problem 3: A full quadratic regression

Finally, do a regression between age, height squared, smoking (x_1, x_2^2, x_3) and FEV (y). What is R^2_{adj} now?

Answer to Problem 3.

Worksheet 2: Blood pressure

In this worksheet, we check another dataset, one on blood pressure. We have collected the following data:

1. age (in years);
2. time spent in an urban environment (in years).

We want to relate those two factors to the systolic blood pressure. The data is presented in the following table.

ID	Age (x_1)	Years in urban area (x_2)	Systolic pressure (y)
1	22	6	120
2	24	5 5	125
3	28	5	120
4	33	10	114
5	34	15	130
6	35	18	118
7	41	32	128
8	47	1	116
9	50	43	132
10	54	40	152

Problem 4: A ratio regression

Let's not spend time doing a multiple factor regression again. Instead, assume you want to find a regression between **the ratio of years in urban area and age**. That is, you want a regression $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 \frac{x_2}{x_1}$. What is the regression line? [2]

[2] Use simple linear regression!

Answer to Problem 4.

Worksheet 3: Model selection

We have collected $n = 16$ house sales in an urban setting. We are interested in what increases the price of a house per square meter, so we tried to keep track of some sale details. More specifically, we collect:

1. the age of the house (Age, in years);
2. the distance to the closest subway station (Distance, in meters);
3. the number of convenience and grocery stores in walking distance (Stores);
4. the latitude (Latitude);
5. and the longitude (Longitude).

The goal is to use some or all of these factors to predict y, the price per square meter. Here is the table of the data:

#	Age	Distance	Stores	Latitude	Longitude	Price
1	14.7	1717.19	2	24.96	121.52	23
2	12.7	170.13	1	24.97	121.53	37.3
3	26.8	482.76	5	24.97	121.54	35.5
4	7.6	2175.03	3	24.96	121.51	27.7
5	12.7	187.48	1	24.97	121.53	28.5
6	30.9	161.94	9	24.98	121.54	39.7
7	16.4	289.32	5	24.98	121.54	41.2
8	23	130.99	6	24.96	121.54	37.2
9	1.9	372.14	7	24.97	121.54	40.5
10	5.2	2408.99	0	24.96	121.56	22.3
11	18.5	2175.74	3	24.96	121.51	28.1
12	13.7	4082.02	0	24.94	121.5	15.4
13	5.6	90.46	9	24.97	121.54	50
14	18.8	390.97	7	24.98	121.54	40.6
15	8.1	104.81	5	24.97	121.54	52.5
16	6.5	90.46	9	24.97	121.54	63.9

Problem 5: All subsets selection

How many different subsets should you consider before declaring the absolute best combination of factors? [3]

[3] There are 5 factors – how many subsets can we create with 5 factors?

Answer to Problem 5.

Problem 6: Comparing subsets

Clearly enumerating all of the previous subsets is a computationally expensive task. However, comparing two or three of them is much easier. Say we agree that Age, Distance, and Stores are the three most important factors, which one of the three models is a more accurate predictor of price?

1. Age and Distance.
2. Age and Stores.
3. Distance and Stores.

Answer to Problem 6.

Problem 7: Backwards selection

Starting from a full model, do at least one iteration of the backwards selection heuristic. Which one is the least significant factor (that is, which one is the first factor to be removed)?

Answer to Problem 7.

Problem 8: Forwards selection

Starting from an empty model, do at least two iterations of the forwards selection heuristic. Which one is the most significant factor (that is, which one is the first factor to be added to the regression)? Which one is the second one to be added to the regression?

Answer to Problem 8.