MTH 420/520 Homework 3. **Yesh Godse**
Date: May 3, 2019
Prof. M. Peszynska Due: May 4, 2019

MTH 420-520 students turn in 1-2.
Extra credit is marked.
Show enough work and enough of `MATLAB code` to demonstrate this is your own hard work, but be concise.
Sloppy incomplete work, or lengthy algebraic calculations, or core dump of MATLAB output with errors will not receive much credit.

Problem 1. Work with real data (Better Life Index), `https://stats.oecd.org/index.aspx?DataSetCode=BLI`.

Pick at least n countries that appear both in BLI listings and at `https://en.wikipedia.org/wiki/Lists_of_countries_by_GDP_per_capita`. Use the data for the first n-2 countries as your training data. Develop a linear model for the dependence of Life Satisfaction l_j (data on BLI list) on some environmental or on some other data d_j reported in BLI list (your choice). Separately, develop a linear model for dependence on GDP, g_j.
Extra: develop a multilinear model $l_j \approx Ad_j + Bg_j + C$, $j = 1, \dots n-2$.

Now test your prediction of BLI for the countries not included in your training set. Report on what you found. Are you satisfied?
MTH 420 students should choose $n \geq 6$, MTH 520 at least $n \geq 10$.

Solution 1. The three models we will create are given by the following equations:

$$l_j = \theta_1 d_j + b_1 \tag{1}$$

$$l_j = \theta_2 g_j + b_2 \tag{2}$$

$$l_j = \theta_3 d_j + \theta_4 g_j + b_3 \tag{3}$$

We use 11 countries in the training data set and 2 countries (Germany and Estonia) for testing.

For models 1 and 2, we find the minimizer $u(t) = \theta * t + b$ of the following:

$$\phi(u) = \sum_i (\theta * t_i + b - l_i)^2 \tag{4}$$

Here, t_i is either the GDP g_j as in (2) above, or the years in education d_j as in (1) above.

To do this we set $\nabla\phi(u) = 0$ and get the following system of equations from simplifying the partial derivatives $\frac{\partial\phi}{\partial\theta}$ and $\frac{\partial\phi}{\partial b}$:

$$(\sum_i t_i)\theta + (\sum_i 1)b = (\sum_i l_i)1$$
$$(\sum_i t_i^2)\theta + (\sum_i t)b = (\sum_i l_i)1$$

From the class notes, we can represent this result with matrix algebra as $A^T A\hat{u} = A^T b$. To know that a minimum u for $\phi(u)$ exists, we can check if A is full rank. Because A is simply a column of 1s and a column of each input data point t_i, we can observe that A^TA is spd because not all t_i values are the same. This let's us conclude that A is full rank and a minimum exists.

Model 3 follows the same approach as models 1 and 2, but has an additional column in A and row in $\hat{u}$ to accommodate for the additional input variable.

We use MATLAB to solve the matrix equations and determine $\hat{u}$ for (1), (2) and (3):

```
%% Model 1 - years in education on life satisfaction
A = [ones(1,11)' data(:,1)];
f = data(:,3);
u_hat = A'*A \ A'*f;

my_scatter(data(:,1),data(:,3),u_hat);

% Test on test data
test_model_2d(testData(:,1),testData(:,3),u_hat);


%% Model 2 - gross domestic product on life satisfaction
A = [ones(1,11)' data(:,2)];
f = data(:,3);
u_hat = A'*A \ A'*f;

my_scatter(data(:,2),data(:,3),u_hat);

% Test on test data.
test_model_2d(testData(:,2),testData(:,3), u_hat);

%% Model 3 - years in education and gdp on life satisfaction

% Use normal equations to find u_hat from A and f
A = [ones(1,11)' data(:,1) data(:,2)];
f = data(:,3);
u_hat = A'*A \ A'*f;

my_3d_scatter(data(:,1),data(:,2),data(:,3),u_hat)

% Test on test data.
test_model_3d(testData(:,1),testData(:,2),testData(:,3), u_hat);
```

The above code listing generates the following best fit lines for model 1 and 2, and best fit plane for multi-linear model 3.

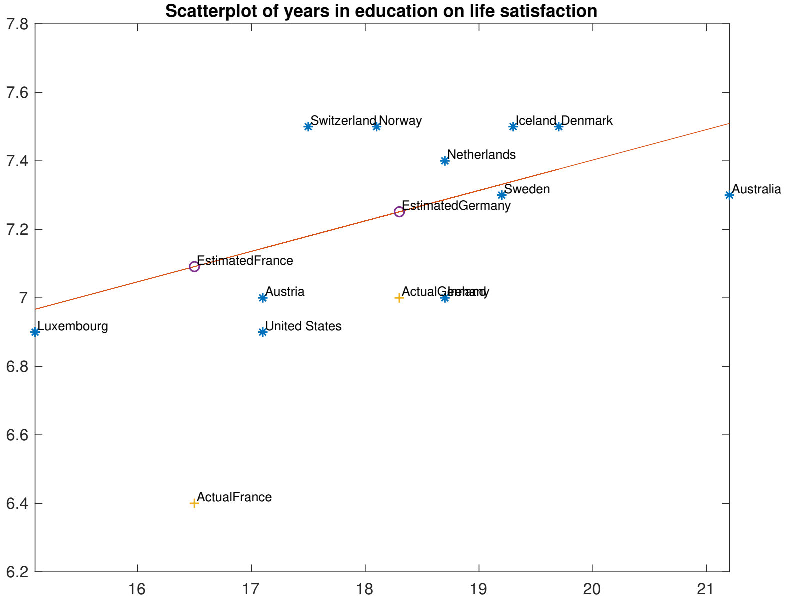

Figure 1. Model 1 - Years in Education on Life Satisfaction. Each point is labeled. Prediction for Germany is alright, prediction for France is very bad

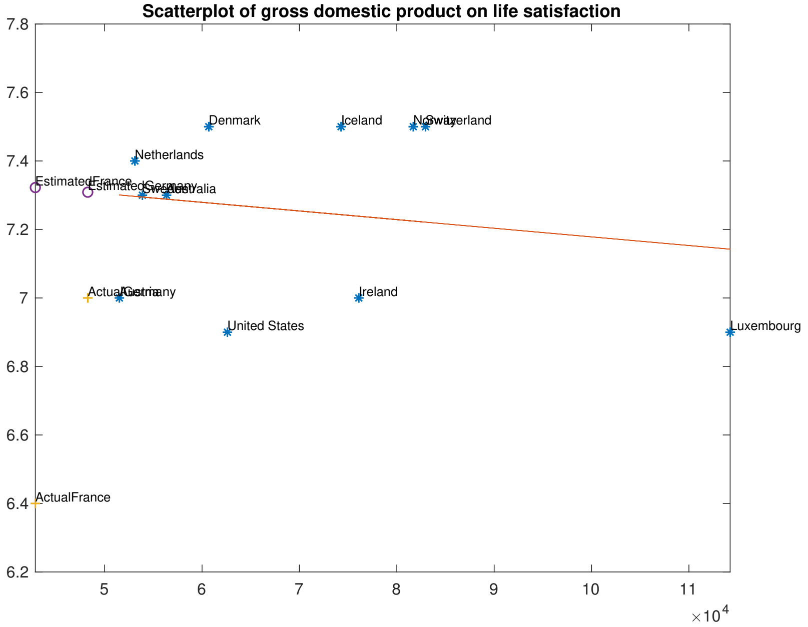

Figure 2. Model 2 - GDP on Life Satisfaction. Each point is labeled. Once again, prediction for Germany is alright, prediction for France is very bad

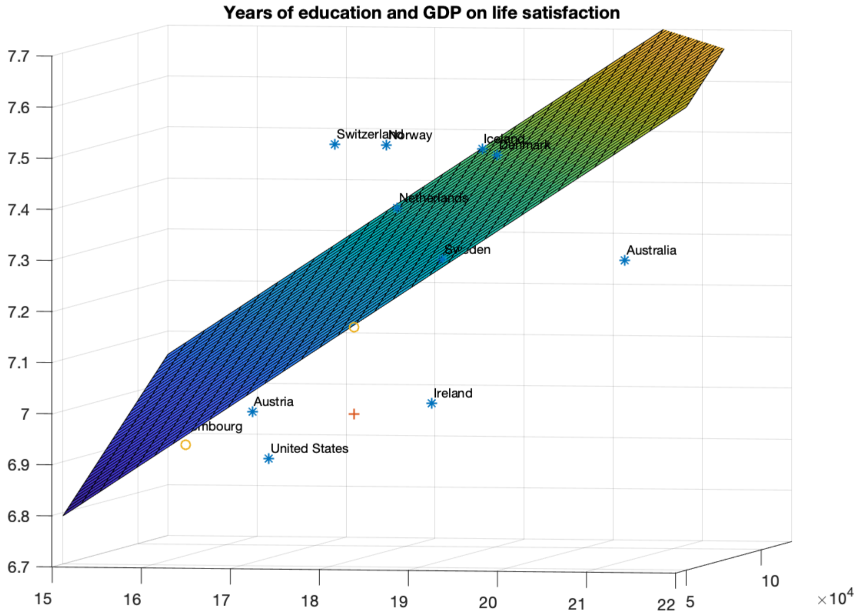

FIGURE 3. Model 3 - GDP and Years in Education on Life Satisfaction. Each point is labeled. Best fit plane goes through cloud of data and again makes pretty inaccurate predictions for Germany and Estonia

In all three models, Germany and France were not predicted with great accuracy. If France was included in the training data-set it could be an outlier in every single model. This suggests that much more variability exists in the real distribution of countries' years in education, GDP, and life satisfaction. To make each model more accurate, more data points are necessary. Based on the scatter-plot of the small data set that we used, moving to a non-linear model would not result in significant improvements as there simply isn't enough data to conclude whether a non-linear model would fit the data better.

Problem 2. Derive the equation that finds the best continuous least squares fit of a linear polynomial (an affine function) $u(x) = \theta x + b$ to a given function $q(x)$. This best fit minimizes

$$J(u) = \int_0^1 (u(x) - q(x))^2 dx. \tag{5}$$

Calculate this solution when $q(x) = x^2$. Plot both $q(x)$ and $u(x)$, and at least two other linear approximations $u_1(x), u_2(x)$ for $q(\cdot)$ that you can think of. Measure the quality of these approximations by calculating $\| u - q \|$.

Solution 2. We minimize $J(u)$ defined below to determine the best-fit line $u(x) = \theta x + b$ for $q(x) = x^2$.

$$J(u) = \int_0^1 (\theta x + b - x^2)^2 dx$$

We set $\nabla J(u) = 0$ and get the following equations.

$$\frac{\partial J}{\partial \theta} = \int_0^1 (2(\theta * x + b - x^2) * x)dx = 0$$
$$\frac{\partial J}{\partial b} = \int_0^1 (2(\theta * x + b - x^2) * (1))dx = 0$$

These equations can be further simplified by splitting up the integrals

$$\left(\int_0^1 x^2\right)\theta + \left(\int_0^1 x\right)b = \left(\int_0^1 q(x)\right)x$$
$$\left(\int_0^1 x\right)\theta + \left(\int_0^1 (1)\right)b = \left(\int_0^1 q(x)\right)(1)$$

By evaluating the definite integrals, we get the following equations. Note that the coefficients of x in the system of equations below form the same matrices $A^T A$ and $A^T b$ used for finding $\hat{u}$ in the discrete problems.

$$\frac{\theta}{3} + \frac{b}{2} = \frac{1}{4}, \quad \frac{\theta}{2} + b = \frac{1}{3}$$
$$\implies u(x) = x - \frac{1}{6}$$

The following MATLAB code plots $q(x)$ with the best fit line $u(x)$. In addition, two arbitrary lines $u_1(x) = (3/4) * x$ and $u_2(x) = (4/3)x - (1/4)$ are plotted. To evaluate the result of the best-fit calculation and make sure it is a closer fit than these arbitrary lines, the norm $\|u - q\|$ for each $u(x)$ is calculated.

```
x = linspace(0,1);

figure();
plot(x,x.^2,x, x - (1/6),x,(3/4)*x,x,(4/3)*x-(1/4));
legend('q(x)','u(x)','u_1(x)','u_2(x)');

% calculate norm of minimized u
norm_fun = @(t) ((t - 1/6) - t.^2).^2;
norm_u = sqrt(integral(norm_fun,0,1));

% calculate norm of arbitrary u_1
norm_fun = @(t) (((3/4)*t) - t.^2).^2;
norm_u2 = sqrt(integral(norm_fun,0,1));

% calculate norm of arbitrary u_2
norm_fun = @(t) (((4/3)*t-(1/4)) - t.^2).^2;
norm_u1 = sqrt(integral(norm_fun,0,1));
```

Here is the plot generated by the MATLAB code.

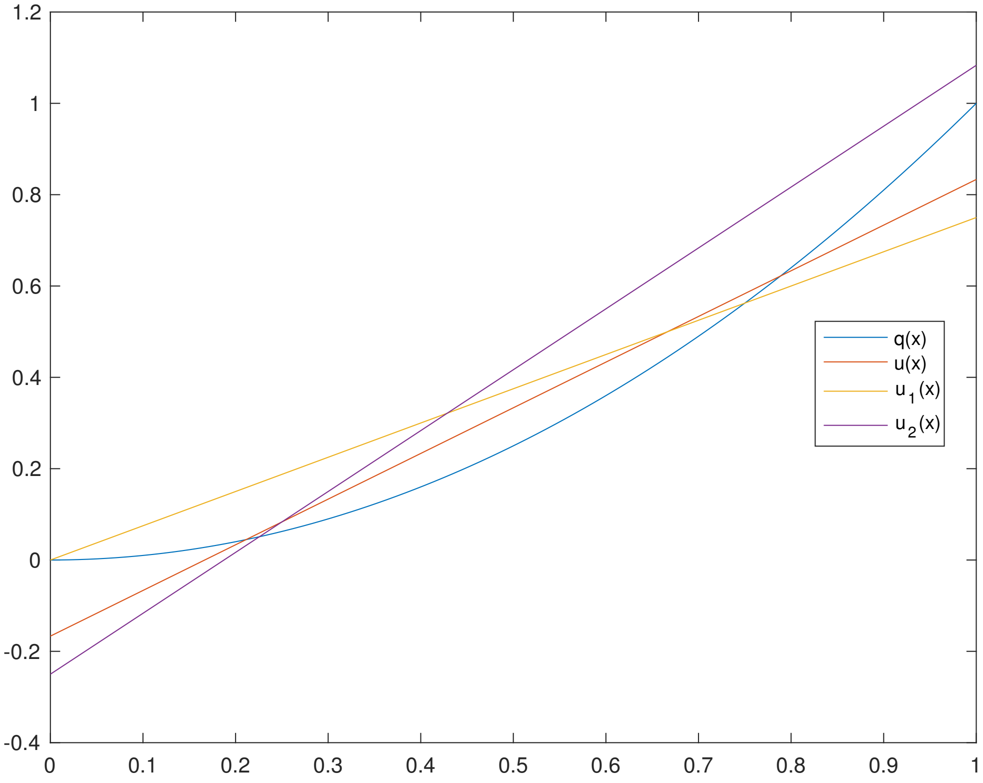

Figure 4. Plot of $q(x)$, $u(x)$ and arbitrary lines $u_1(x)$, $u_2(x)$

The following console output shows that the best-fit line $u(x)$ is better than the arbitrarily chosen lines $u_1(x)$ and $u_2(x)$

```
>> norm_u = 0.0745      >> norm_u1 = 0.1475      >> norm_u2 = 0.1118
```