CSC 191B: Lab #8: Curve Fitting: Linear Regression

Learning Outcomes

- Modeling an experimental data set using regression
- Computing and evaluating best-fit polynomials in MATLAB

Background. Given a data set of paired values $\{(x_i, y_i)\}$, we often suspect that there is an underlying dependence of y_i on x_i. In this case, we would like to identify a simple function f such that $f(x_i) \approx y_i$ for all i, so that the dependence is fully specified by f. After identifying f, we can often draw further conclusions about our data set and also predict paired values that aren't in the original data, e.g., $(x', f(x'))$.

For (a famous) example, consider the history of the population of the United States. Every decade the census is performed, so we know with reasonable accuracy the population for each of those years. Here is the data (in millions of people):

Year	1900	1910	1920	1930	1940	1950
Population	75.995	91.972	105.711	123.203	131.669	150.697

Year	1960	1970	1980	1990	2000	2010
Population	179.323	203.212	226.505	249.633	281.422	308.746

1 Census GUI

Save `censusgui.m` to your MATLAB directory and run the script, which was written by Cleve Moler. You should see a scatter plot in blue of the data given in the tables above, as well as a green dot, which is an eyeball prediction of the US population in 2020. Play with the dropdown menu and other options to use regression models to predict the 2020 population.
Discussion questions:

1.1. What is the 2020 prediction using a polynomial of degree 2? of degree 7? using an exponential function? What general observations do you make?

1.2. Find an estimate for the US population this year (and cite your source). Based on this recent data, what model do you believe is most accurate for the 2020 prediction, and why?

2 Polynomial Curve Fitting

Find the `Pop_Data_China.csv` file, which comes from `http://databank.worldbank.org`, and enter the data into MATLAB. Write a MATLAB script called `pop_fit_china.m` that uses `polyfit` and `polyval` to compute a polynomial of best fit for this data with polynomial degrees varying from 1 to 6. Plot the data points along with the polynomials, and extend the range of years to 2020 to see what predictions each curve yields. Report the exact values of the 2020 prediction for each degree.
Discussion questions:

2.2. What prediction do you trust the most, and why?

2.3. (Bonus) How did you resolve the warning messages that MATLAB gave you when you used `polyfit`?

What to turn in.

- One MATLAB script called `pop_fit_china.m` that computes the polynomials of best fit, generates the plot, and includes the model predictions and answers to the discussion questions.
- One PDF file that contains the scatter plot of the data along with the polynomials of best fit, with clear legend and labels.

Grading rubric

- Code: 50 points for the script, which should be well organized and well documented
- Plot: 30 points for the plot, which should be well labeled
- Results: 5 points for the 2020 Chinese population prediction results
- Discussion: 5 points each for the discussion questions