Linear Least Squares

This homework is about using the QR factorization to find the least squares fit of a polynomial to noisy data.

When $A \in \mathbf{R}^{m\times n}$ with $m > n$ the system of linear equations $Ax = b$ is said to be over determined. Such systems do not generally have a solution. Therefore, one looks for an x that minimizes the residual error $\|Ax - b\|$. Such an x can be obtained by solving $\widetilde{R}x = \widetilde{Q}^T b$ where $A = \widetilde{Q}\widetilde{R}$ is the reduced QR factorization of A. Recall,

- $\widetilde{Q} \in \mathbf{R}^{m\times n}$ with $\widetilde{Q}^T\widetilde{Q} = I$.
- $\widetilde{R} \in \mathbf{R}^{m\times m}$ with $\widetilde{R}$ upper triangular.
- If the columns of A are independent then $\widetilde{R}$ is invertible.

In Julia the left division operator `\` automatically uses the QR factorization to find the least-squares solution for over-determined systems. Thus, one can write `x=A\b` to find the x that minimizes $\|Ax - b\|$.

Fitting a Linear Model

One application of least squares is fitting a linear model

$$F_c(x) = c_1\phi_1(x) + c_2\phi_2(x) + \cdots + c_n\phi_n(x)$$

to noisy data. For example, given data for $i = 1, \ldots, N$ of the form

$$(x_i, y_i) \qquad \text{where} \qquad y_i = F_c(x_i) + noise$$

suppose the goal is to find the constants c_j such that the polynomial

$$p(x) = c_1 + c_2x + c_3x^2 + c_4x^3$$

is the one from which the data most likely came.

Under the assumption the noise process is independent, identically distributed and Gaussian, the desired c_j are exactly the values which minimize

$$J(c) = \sum_{i=1}^{N} \big|y_i - p(x_i)\big|^2.$$

Happily, minimizing $J(c)$ is the same as solving $Vc = y$ where V is the $N \times 4$ Vandermonde matrix and y the vector such that

$$V = \begin{bmatrix} 1 & x_1 & x_1^2 & x_1^3 \\ 1 & x_2 & x_2^2 & x_2^3 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_N & x_N^2 & x_N^3 \end{bmatrix} \qquad \text{and} \qquad y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}.$$

In this assignment you will fit a cubit polynomial to the points (x_i, y_i) provided in an individualized data file. These points are stored one per line with x_i the first value and y_i the second. There are $N = 100$ lines in the file. Click on the following link to retrieve your file:

`https://fractal.math.unr.edu/~ejolson/466-22/lsquare/lsdata.cgi`

Please do not use anyone else's data.

Although this is a homework assignment you are welcome to use the computer lab during the week when it's available. Be aware that other instructors may be holding their final exams in the room during finals week. Final exams have prioirty. Please let me know if you have any difficulty obtaining access to the lab. You may also use your own personal computer.

The data which appears when I click on the above link is different than what you will obtain when you click the same link. To finish this homework assignment please repeat these same steps but for own individualized data.

Upon clicking on the link, I obtained the example

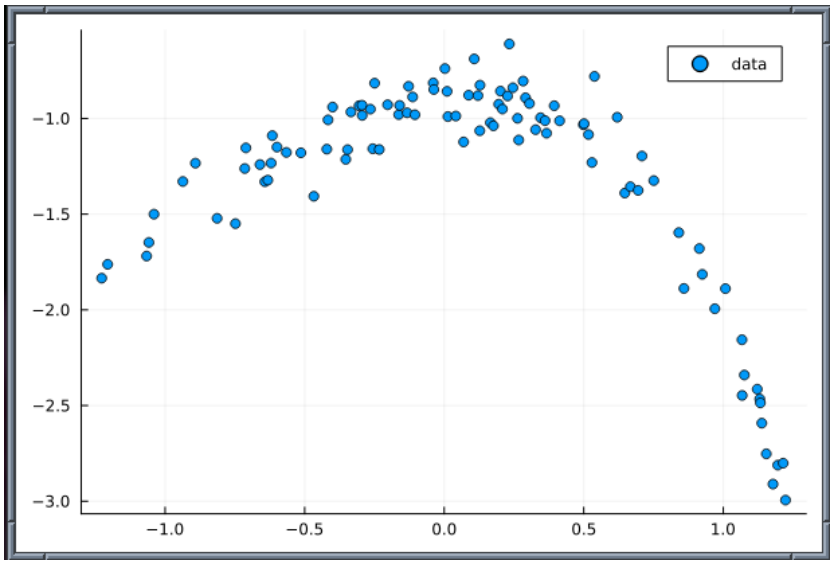
> The data was generated from the model
>
> $$y_i = c_1 + c_2 x_i + c_3 x_i^2 + c_4 x_i^3 + \text{noise}.$$
>
> Your data download is here.
>
> In Julia-compatible code the true values of the c_j are
>
> ```
> ctrue=[-0.9,0.32,-0.92,-0.58]
> ```

Make a working directory for this homework, for example `hw05`, then right click on the *here* link and save the file in that directory. In my case the name of the data file was `DE229EC9.dat`. It will be different for you.

It's always a good idea to visualize the data before performing any kind of numerical fit. Begin by reading in the data and plotting it. To do this use the `DelimitedFiles` and `Plots` libraries as follows:

```
1 using DelimitedFiles, Plots
2
3 data=readdlm("DE229EC9.dat")
4 x=data[:,1]
5 y=data[:,2]
6 display(scatter(x,y,label="data"))
```

At this point you should obtain a graph that looks similar to



The exact shape depends on your individualized data file.

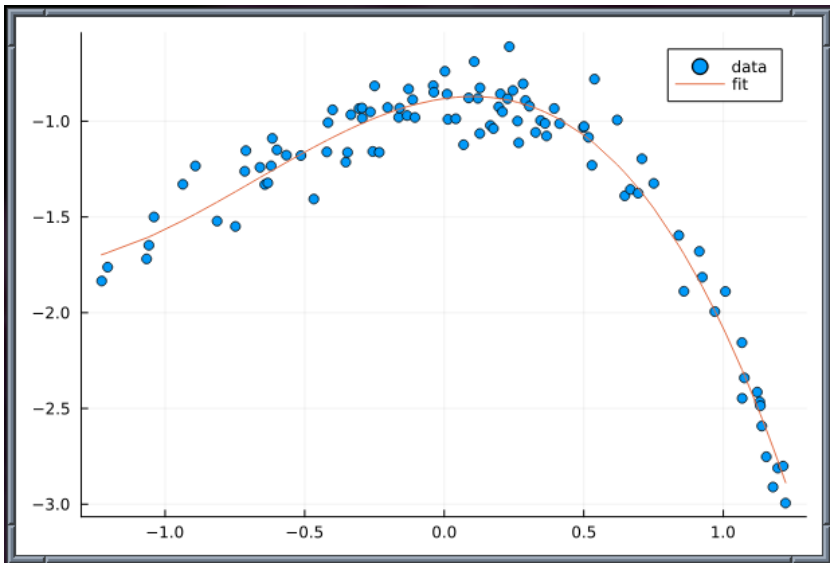
Next create the Vandermonde matrix A and solve the least squares problem $Vc = y$ to obtain the coefficients `cfit` that correspond to the polynomial from which the data most likely came.

```
8 phi=[x->1,x->x,x->x^2,x->x^3]
9 V=[phi[j](x[i]) for i=1:length(x), j=1:4]
10
11 cfit=V\y
12 yfit=V*cfit
13 display(plot!(x,yfit,label="fit"))
```

After executing line 11 the polynomial coefficients were

```
julia> cfit=V\y
4-element Vector{Float64}:
 -0.8818279900787308
  0.20464432548902906
 -0.941495558668323
 -0.46278525707704005
```

The resulting least-squares fit looks like



Again note your data and fit may have a different shape.

In applications the mechanism that generated the data is generally unknown. That's the whole point of finding the least-squares fit in the first place. In such cases after finding the least-squares fit one must subsequently ask whether the fit is statistically good enough.

In this assignment the coefficients for the exact polynomial are known. This provides a test case to check that our numerical codes work as expected. To this end, we compare the fitted polynomial with the coefficients on the web page for the polynomial that actually generated the data.

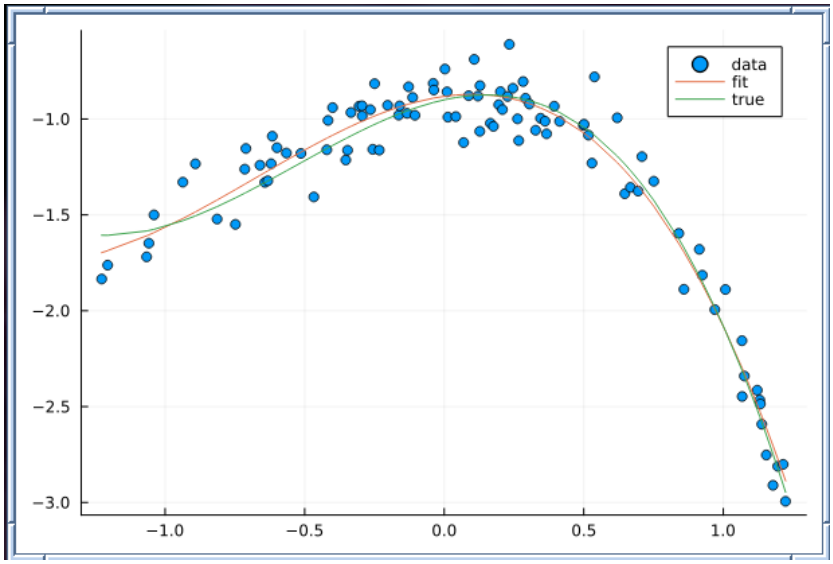
```
15 ctrue=[-0.9,0.32,-0.92,-0.58]
16 norm(cfit-ctrue)/norm(ctrue)
17 ytrue=V*ctrue
18 display(plot!(x,ytrue,label="true"))
```

Note that line 15 is pasted from the web page. The output

```
julia> norm(cfit-ctrue)/norm(ctrue)
0.11526839330427804
```

of line 16 indicates the fitted coefficients differ from the true coefficients by more than 10 percent. This is due to the noise in the data. Your fitted coefficients may be closer or further from the true coefficients.

For the data currently being analyzed the final graph comparing the polynomials looks like



Submitting Your Work

Two PDF files should be submitted for grading. The first file should contain two parts:

- A program that computes and then displays the fitted coefficients of the polynomial, the true coefficients and the norm of the difference between fitted and true coefficients.
- The output from running that program.

The second file should be graph showing the data, the fitted polynomial and the true polynomial. This can be obtained by placing the command `savefig("graph.pdf")` at the end of your program.

After debugging and making sure your program runs correctly, please prepare your submission by typing

```
$ julia hw05.jl >hw05.out
$ j2pdf -o hw05.pdf hw05.jl hw05.out
```

If you are using your personal computer, the `j2pdf` command will not be available. In that case a `hw05.pdf` file suitable for submission can be made by dropping your code and the corresponding output into a word processing document and using the print to PDF option.

Before uploading, check `hw05.pdf` and `graph.pdf` with

```
$ evince hw05.pdf graph.pdf &
```

to ensure the output and graph looks correct. If you did this assignment in the computer lab please reboot into Microsoft Windows before leaving.