Homework 6: χ^2-fitting to polynomial forms

Caden Gobat

March 31, 2021

1 Introduction

In this assignment, address the ubiquitous problem of fitting an analytical function to discrete/numerical data. To do this, we will attempt to minimize a χ^2 (chi-squared) goodness-of-fit measure to optimize a set of coefficients on a polynomial. The χ^2 statistic is defined as

$$\chi^2 \equiv \sum_i \frac{(P(x_i) - y_i)^2}{\sigma_i^2}$$

Where $P(x)$ is a fitted polynomial function and x, y, and σ are the independent quantity, dependent quantity, and standard error in y, respectively. We will use this value to determine whether certain models appropriately describe our data.

2 Results

Question 1

Fitting the first 6 seconds of the data ($0 \leq t < 6$) yields a fit of

$$y(t) = -5.0535t^2 + 2.46t + 16381.7$$

With a χ^2 value of approximately 6.35. Because we know the equation of motion should be of the form $\frac{1}{2}gt^2 + v_0t + y_0$, we can interpret these parameters to find that the fit has returned a value for gravitational acceleration of $g \approx -10.11$ m/s^2.

Given that this subset of the data consists of 12 data points (two samples per second; 6 seconds) and we are fitting 3 parameters (coefficients on the quadratic, linear, and constant terms), the total degrees of freedom is 9. Thus the reduced chi-squared value (i.e. $\frac{\chi^2}{\text{dof}}$) for this fit is 0.706, which (because it is less than 1) indicates that this routine has *over*fit the data.

The percent confidence α for this fit can be calculated by

$$\int_{\chi^2_\nu}^{\infty} f(\chi^2) d\chi^2 = \alpha$$

where $f(\chi^2)$ is the chi-squared distribution.[1] This α is then the probability that any deviation in χ^2_ν from its expected value (based on the known χ^2 distribution) is due to a probabilistic fluke and not an actual flaw in the model/fit parameters. In `Mathematica`, this calculation can be accomplished by running `1-CDF[ChiSquareDistribution[`ν`], `χ^2`]` (with $\chi^2 \cong 6.35$ and $\nu = N - 3 = 9$), which yields $\alpha \cong 0.704$. This means we can be approximately 70% confident that this model is appropriate for these data.

[1] $f(\chi^2) = \dfrac{e^{-\chi^2/2}}{2^{\nu/2}\,\Gamma(\nu/2)} \cdot (\chi^2)^{\frac{\nu}{2}-1}$

	Value	Expectation
g	-10.11	-9.81
v_0	2.46	~ -15.62
y_0	16381.7	16381.567
χ^2	6.35	9
χ^2_ν	0.706	1
conf	0.704	

Table 1: Summary of second-order polynomial fit results for the first 6 seconds of `cannonball.dat` data.

Question 2

Using all 118 data points in the `cannonball.dat` file yields a best-fit of

$$y(t) = -4.6924t^2 - 3.4245t + 1.6395.64$$

With $\chi^2 \cong 140.24$. This fit has found $g \approx 9.385$ m/s^2, which is now an *under*estimate compared to its actual known value of ~ 9.81 m/s^2. Given that the number of data points for this fit is now $N = 118$, we now have $N - 3 = 115$ degrees of freedom, making our reduced chi-squared $\frac{\chi^2}{\text{dof}} = \frac{140.24}{115} \cong 1.22$. This can now be said to be an appropriate fit, as it is greater than but on the order of 1 ($\chi^2_\nu = 1$ indicates a perfect fit).

The confidence (as described previously) is $\alpha = 0.055$, meaning we can only be 5.5% sure that our model represents a good fit to these data. This could be due to the effects of air resistance, which introduces another term into the underlying differential equation and makes the acceleration non-constant. In the first 6 seconds, the effect would not be noticeable, but over the entire trajectory, it becomes noticeable and thus affects the goodness of the fit.

	Value	Expectation
g	-9.385	-9.81
v_0	-3.42	~ -15.62
y_0	16395.6	16381.567
χ^2	140.24	115
χ^2_ν	1.22	1
conf	0.055	

Table 2: Summary of second-order polynomial fit results for the entirety of the `cannonball.dat` data.

Question 3

An alternative formula for calculating a χ^2 statistic is

$$\chi^2 = \sum_{i=1}^{N} \frac{(\bar{x_i} - x_i)^2}{\bar{x_i}}$$

which does not rely on the data coming with any error σ. However, being able to calculate a chi-squared value for these data does not mean that we will necessarily get a good fit. Radioactive decay behaves exponentially, not as a polynomial, so we have more fundamental issues with our methodology. Second- and third- order polynomials do not even come close visually to fitting the data. A fourth-order polynomial with coefficients

$$0.0193t^4 - 3.931t^3 + 295.364t^2 - 9923.637t + 131394.55$$

comes within the realm of feasibility, providing the fit shown in Fig. 1. This yields $\chi^2 = 1606.5$ and we have 20 data points in `decay.dat`, so there are $20 - 5 = 15$ degrees of freedom in this fit. Therefore the reduced χ^2 is $\chi^2_\nu \cong 107.1$. This is still very poor. There is a chance that moving up to higher order polynomials may improve this, but a much better option would be to change the underlying model itself. For example, using an exponential model with fitted parameters

$$133565e^{-0.086t}$$

gives $\chi^2 = 12.2$ with $\chi^2_\nu = 0.677$. This is an improvement of two orders of magnitude, and requires fewer parameters to attain a much better fit. Therefore we can conclude that this is a better fundamental model.

We can determine the half life by solving $0.5 = e^{-0.086t}$. This yields $\ln(0.5) = -0.086t \implies t = 8.06$ days, or around 696000 seconds. This matches most closely with iodine-131, which has a half life of approximately 693 ks.

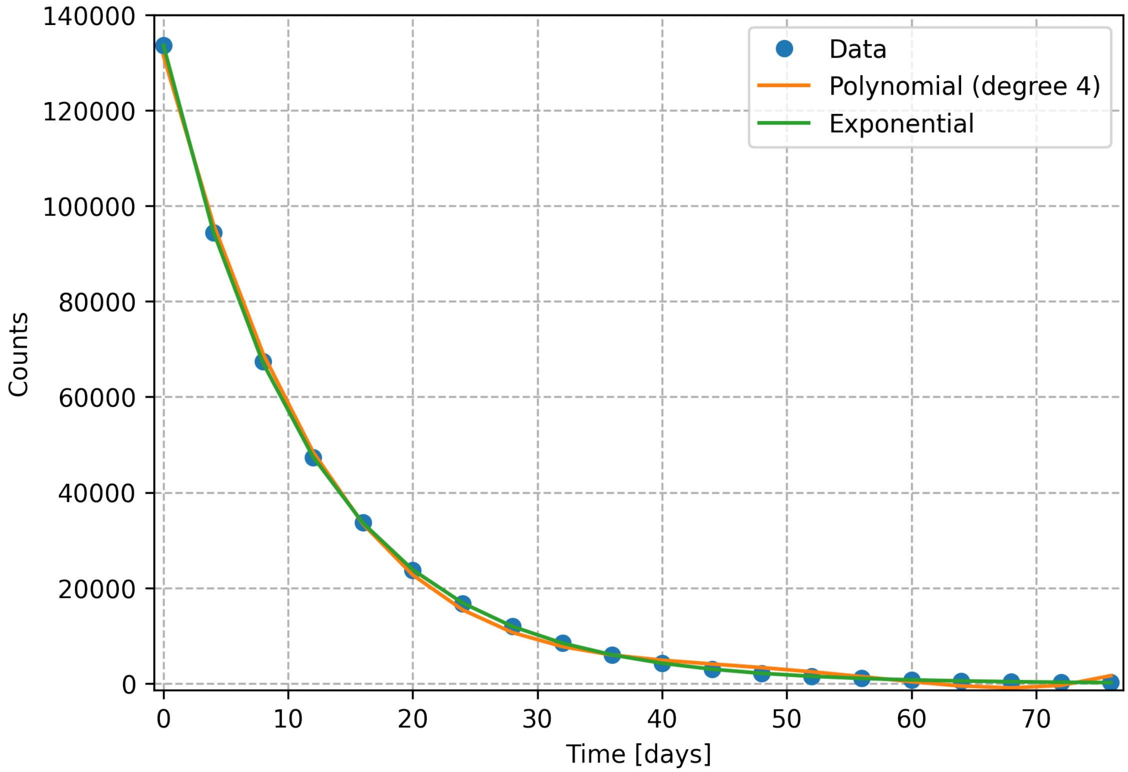

Figure 1: Radioactive decay data with two models fitted. Parameters for each model are those given above, with the exponential fit being decidedly better.

3 Conclusions

I greatly enjoyed this assignment, and I have a much better understanding of the mechanism behind χ^2 analysis now, having built a fitting engine from the ground up. I also appreciated learning the underlying statistical theory of confidence levels and distributions.

I am curious about how one might deal with errors in the horizontal direction (i.e. an error on the x-values as well), and am looking forward to continuing to work on these kinds of problems.