Case Study

Beck Addison and Marian Simo

10/12/2021

Introduction

The task of modeling turbulent fluid systems has consistently been one of the most difficult problems to solve in modern mathematics and physics; in fact, despite over two centuries of concerted research into the topic, no analytical method currently exists to accurately predict the evolution of turbulent systems over time.

Nevertheless, the successful simulation of turbulent systems remains uniquely essential to a diverse array of sciences and industries with common use cases including modeling airflow in aerodynamic design, analyzing blood flow in medicine (Sallam and Hwang 1984), and meteorological forecasting (Kim et al. 2015). To perform this simulation, high-resolution direct numerical simulation (DNS) is typically performed using the Navier-Stokes equation, but this is often very complex and computationally expensive to perform on any arbitrary set of inputs. In this case study, we circumvent these time-intensive simulation techniques by interpolating a set of results collected from Navier-Stokes DNS to produce a predictive model given a set of input parameters. Our objectives are therefore as follows:

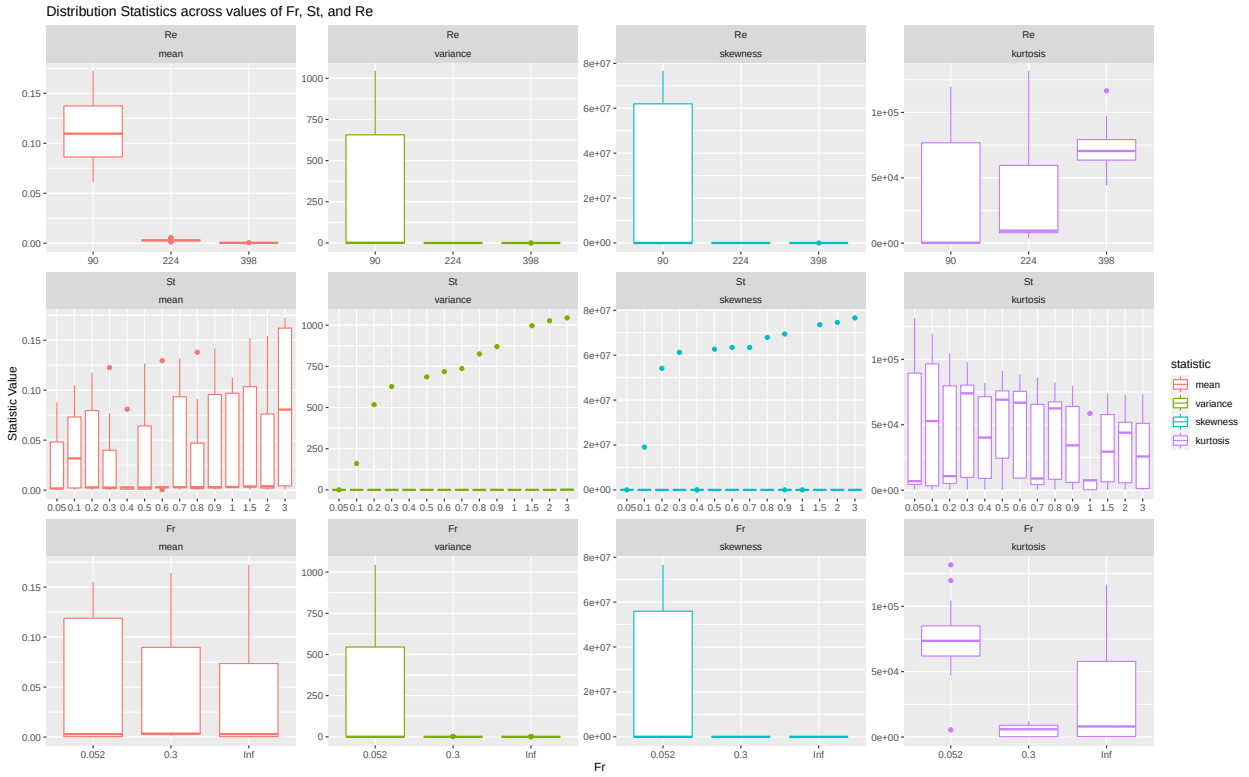
- Given a set of training and test data describing the inputs of a DNS simulation, namely the Reynolds Number Re, the gravitational acceleration Fr, and the Stokes number St, generate a model or set of models that can accurately predict the particle cluster volume distribution in terms of four moments used as dependent variables. We will call this our *prediction* objective.
- Given the set of input and output parameters, determine how each parameter the distribution of cluster volumes. We will call this our *inference* objective.

Methodology

To fulfill the objectives we pose in our introduction, we must collect more information about our modeling objectives. First, we have to define cluster volume in terms of the moments we are given. Since each simulation in our dataset represents a single row of three inputs $< Re, Fr, St >$ and four outputs $< E[X], E[X^2], E[X^3], E[X^4] >$ representing information about a *distribution* of several cluster volumes, we must consider that we are measuring the change in shape of a distribution of volumes rather than the volumes themselves. For this analysis, we defined key characteristics about these models in terms of these moments:

1. Mean: $E[X]$
2. Variance: $E[X^2] - E[X]^2$
3. Skewness: $(\frac{E[X^3]}{E[X^2]})^2$
4. Kurtosis: $\frac{E[X^4]}{E[X^2]^2}$

Our goal here is not only to make interpretations of our model predictions simpler but also to attempt to find some distribution that connects these outputs to our input parameters. To do this, we must first examine how these model characteristics are distributed in our dataset.



This table of graphs suggests the following key notes:

1. The mean, variance, and skewness of the distribution of cluster volumes drops significantly as `Re` increases, while kurtosis appears to increase.
2. There isn't a strong trend for `St` except to note that, as it increases, the maximum value of the mean, variance, and skewness increases fairly dramatically, but the kurtosis falls steadily.
3. The mean of the particle cluster volume remains mostly consistent for values of `Fr`, but the variance and skewness drop quickly. The kurtosis with respect to `Fr` doesn't seem to have a clear trend.

Other things to note include:

1. The variance and skewness of the distribution don't seem to be dependent on `St`. The mean and particularly the kurtosis of the distribution does appear to have some relationship with `St`, however, though kurtosis seems more closely related than the variance in what appears to be a roughly linear decline.
2. `Re` evokes a strong change in all of the distribution statistics as it increases, sharply reducing the variance in their distributions. I'd hypothesize that `Re` has a quadratic or otherwise polynomial effect on the distribution of the particle size, since it appears to change the variance and skewness of the set so sharply.
3. `Fr` evokes similar changes in the variance and skewness of the distribution as `Re`, but not in the kurtosis and mean. This could mean there is some interaction effect in the second moment of the data between `Re` and `St`.

References

Kim, Jung-Hoon, William N Chan, Banavar Sridhar, and Robert D Sharman. 2015. "Combined Winds and Turbulence Prediction System for Automated Air-Traffic Management Applications." *Journal of Applied Meteorology and Climatology* 54 (4): 766–84. https://doi.org/10.1175/JAMC-D-14-0216.1.

Sallam, Ahmed M, and Ned H C Hwang. 1984. "Human red blood cell hemolysis in a turbulent shear flow: Contribution of Reynolds shear stresses." *Biorheology* 21: 783–97. https://doi.org/10.3233/BIR-1984-21605.

Appendix A: Residual and Q-Q Plots

Raw Moments as Response Variables

```
## Warning: Unknown or uninitialised column: `moments`.
```

Linear Models, No Interactions

In the graphs above, we've plotted the residuals and Normal Q-Q plots for all of our linear regressions where each of the raw moments represents a response variable to the three predictors $< Re, Fr, St >$. The first row represents information with the first raw moment as the response, the second row represents the second raw moment, and so on. For this first assessment, we have no interaction variables.

These attempts of linear regression suggests the following key notes:

1. Using the Residual Standard Error and Adjusted R^2, they suggest a simple linear regression model does not fit the data well. Raw moments 2, 3, and 4 suffer from a large RSE and very low R^2 values. Raw moment 1 has a higher R^2 value of 0.6271 and a much lower RSE value of 0.0341; however, those values are indicative of weak linearity.
2. Looking at the residual plots, for all 4 raw moments it appears a linear model does fit well up until larger values of the response variable. This could be indicative of a needed transformation for the response values.
3. The QQ plots for raw moments 2, 3, and 4 also suggest linearity up until the larger values of the response variables. The QQ plot of raw moment 1 follows a "S" pattern that could suggest under-dispersed data.

Linear Models, 2-Way interactions with backward stepwise selection optimizing AIC

Important takeaways from backward selection:

1. In all three raw moments for which backward selection was done, St and Fr resulted in an interaction term as well as Re and Fr.
2. Both raw moment 3 and 4 contain all 3 interaction terms whereas raw moment 2 leaves out the interaction between St and Re.

Nonlinear Models, 2-degree polynomials

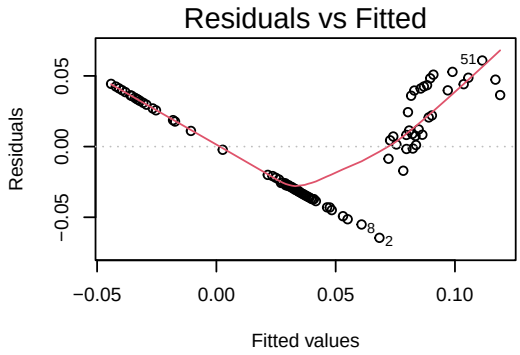
Here again, we now compare a set of 2-degree nonlinear polynomial regression models, again with each of the raw moments as response variables to the predictor set from before.

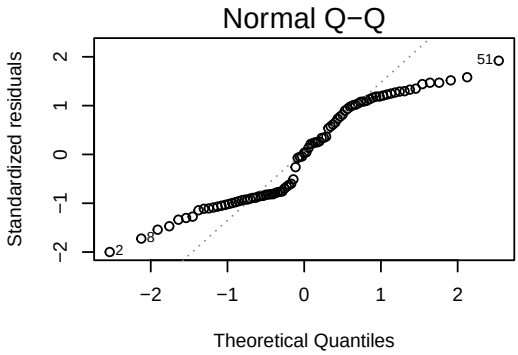
Attempting polynomial regression suggests the following key notes:

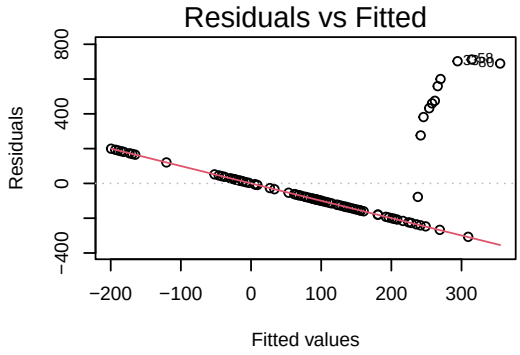
1. Polynomial regression only benefited raw moment 1, increasing its adjusted R^2 to 0.9646 and reducing its RSE to 0.01051. The QQ plot now follows a more linear pattern and the residual vs fitted plot also better follows the line of best fit. Thus, this suggests a polynomial model is better suited for raw moment 1.
2. Raw moments 2, 3, and 4 did not benefit from a polynomial model, each still suffered from a low R^2 and very high RSE.

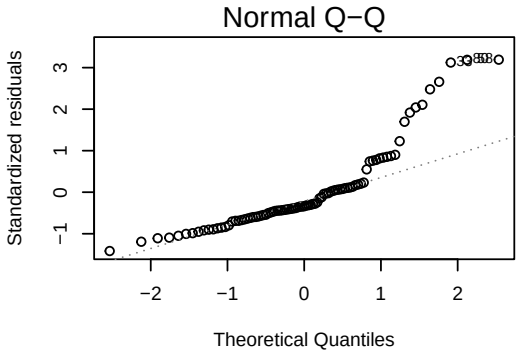
Mean/Variance/Skewness/Kurtosis as Response Variables

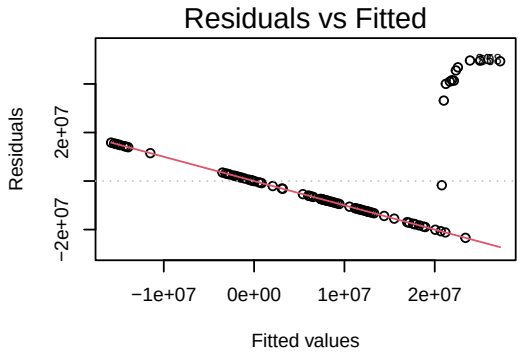
Linear Models, No Interactions

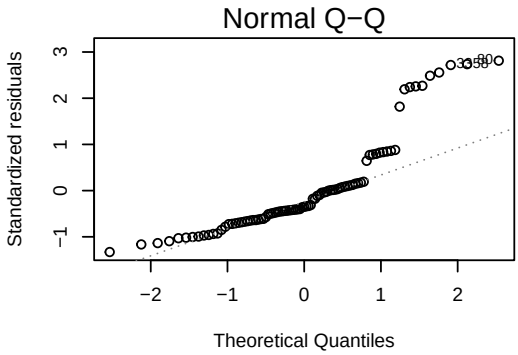


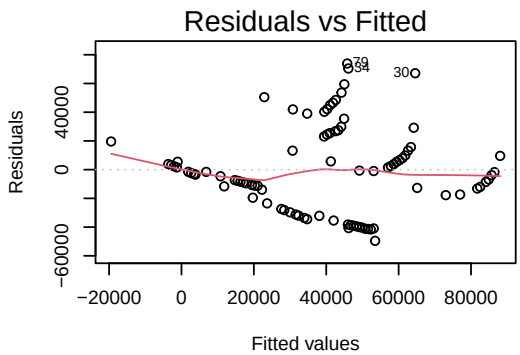


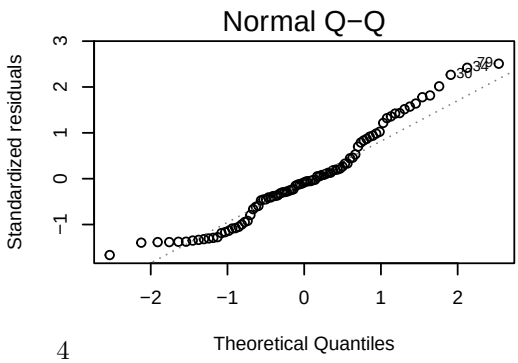




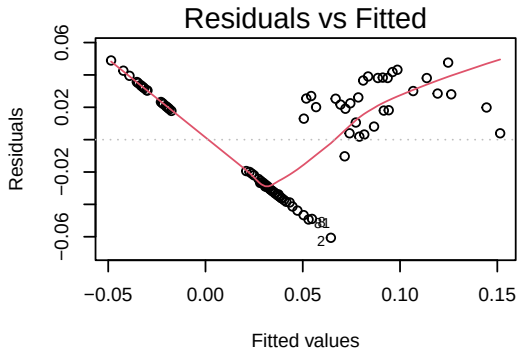


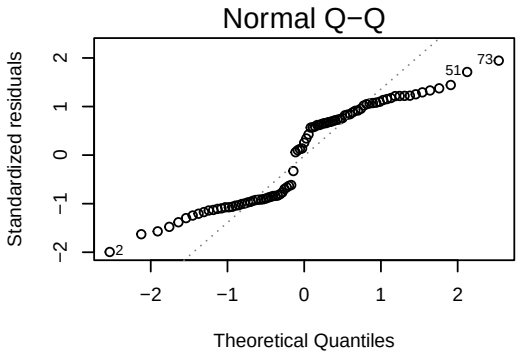


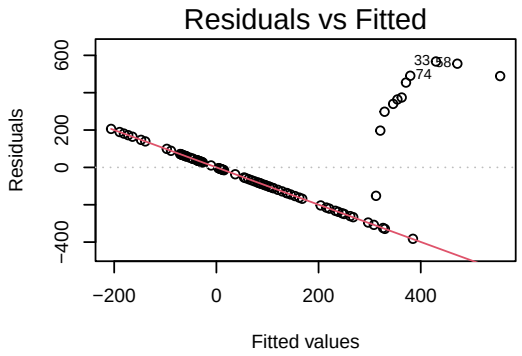


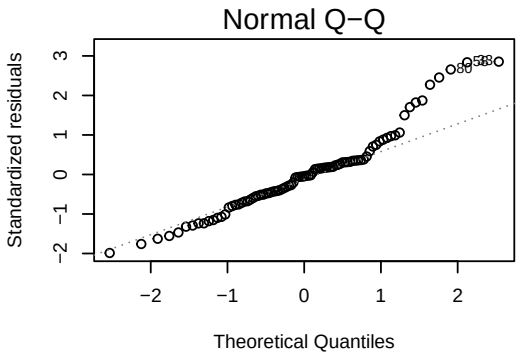


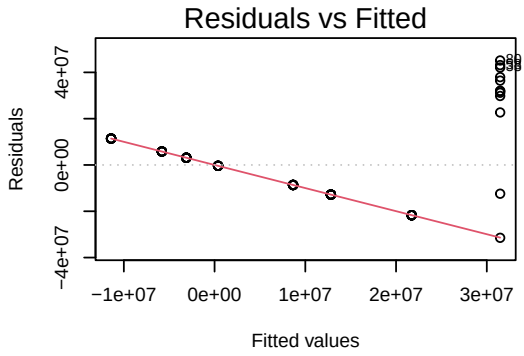
Linear Models, 2-Way interactions with backward stepwise selection optimizing AIC

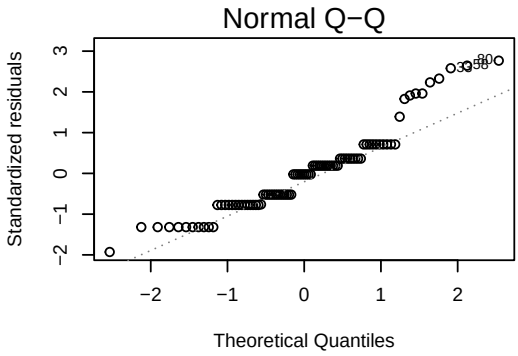


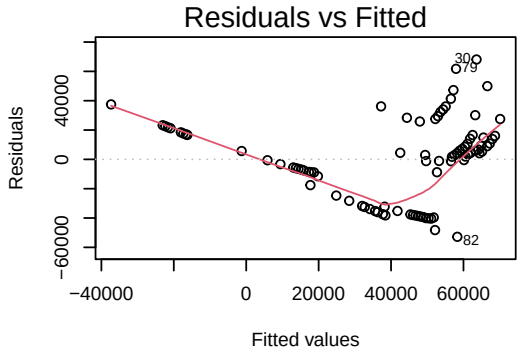


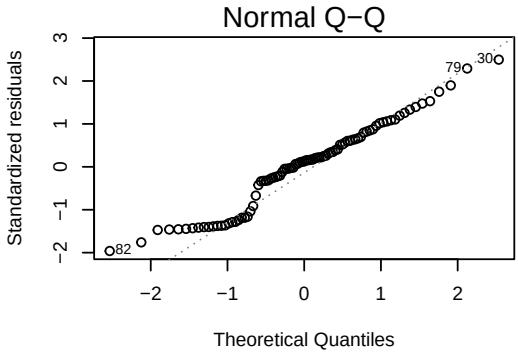












Nonlinear Models, 2-degree polynomials