Case Study

Abdel Shehata

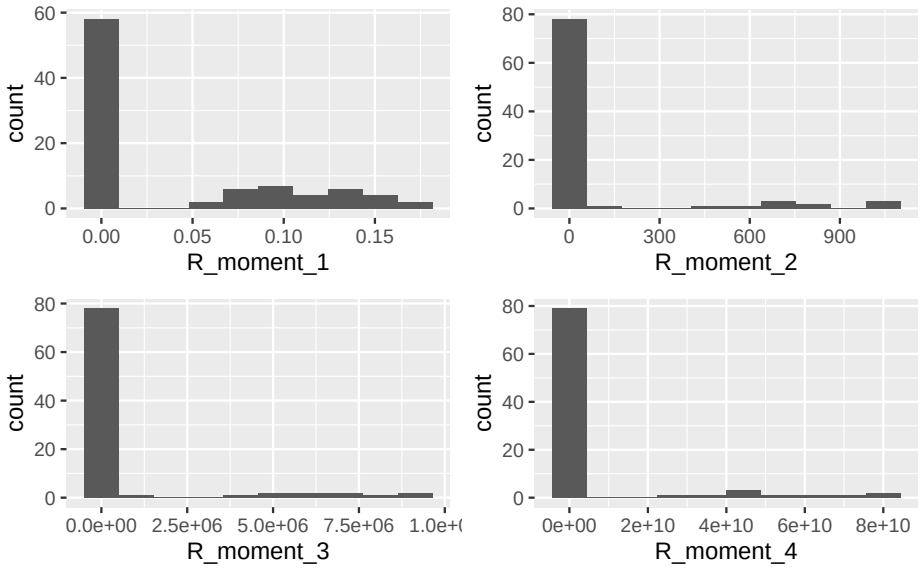
2022-10-26

Introduction and Data

Introduction

Data Introduction

Exploratory Data Analysis



Upon viewing the data, we encountered several challenges. First the Re and Fr variables only have three unique observed values. For Re we see low (90), medium (224), and high (398) and for Fr we see low (.052), medium (.3), and high (infinite). We decided against treating these variables as categorical because we want our models to be used for extrapolation to minimize the need for expensive mathematical modeling in the future for new values of these predictors. To support this goal, Fr needed to be transformed so it can be machine-readable. We decided to do a logistic transformation on Fr (creating a new variable called TFr) in order to approximate the effects of infinity. Values approaching -3 are roughly zero and values approaching 11.5 are roughly infinite. Additionally, the moment data observations are all raw, so we decided to centralize the second through fourth moments. The first raw moment is useful for interpretation because it tells us about the average amount of turbulence we can expect. However, when it comes to the shape of the probabilistic distribution of turbulence (variance, skewness, and kurtosis) we need to centralize the moments in order to interpret them because the raw moments include locational information (e.g. the first moment). Lastly, we observed that the large majority of the observations of all moments are essentially zero. This means that in most simulations, little clustering was observed. This influenced many models by making

coefficients on certain predictors negative. Also, because of the lack of dispersion in response variables, there is probably nonlinearities in the relationships between the three predictors and the responses, and there may be interactions between them where are specific thresholds of values, clustering increases dramatically.

Methodology

We began by fitting linear models to help us form intuitions to guide more complex and accurate models that will be more useful for prediction and more detailed interpretation. We fit one set of linear models based off of all the given predictors and another set that included all possible two-way interactions. We used bidirectional selection to narrow down variables to only include the most effective ones. Then, we compared the plain linear models to the ones with interactions.

Linear R-Squared	No Interactions	Interactions
R__1	0.6054832	0.6282726
R__2	0.1716913	0.2754936
R__3	0.161867	0.2650636
R__4	0.1539926	0.2466392

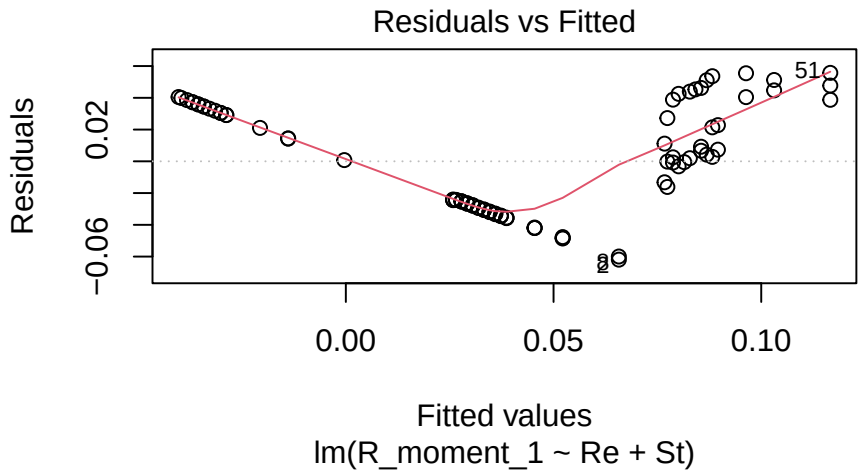
It seems that the interactions are increasingly helpful when explaining variation (by the R^2 value). Adding interactions to the first moment model does not help improve fit much. However, for the third through fourth moments, there is a pretty significant increase in R^2 when comparing the strictly linear models versus the ones with interactions. Thus, to improve model fit we should pay attention to interactions and nonlinear relationships between the predictors and response variables.

Result for Linear Model

$$\hat{R_1} = 0.0102 + 0.01353 * St - 0.0003798 * Re$$

This very simple linear model with only two out of the three predictors explains about 62% of the variation of the first moment. The Reynolds number coefficient is small and negative, which contradicts physics theory (Britannica). I believe this is due to the fact that the overwhelming majority of observations had small mean turbulence, so the regression fit a line with negative slope. On average, we just do not often observe turbulence no matter what predictors are used. However, the coefficient on St is slightly larger and positive. I believe this shows that perhaps the most important contributor to increases in the first moment is the size of the particles. In fact, adding interactions or the Fr predictor did not change the R^2 very much, so I believe that St is very important for increasing average turbulence and does so in a linear way. Intuitively, this appears to make sense. Larger particles have a greater chance of bumping into each other and clustering. This regression tells us that the increase in size perhaps increases the chance of bumping and clustering with a constant, linear effect.

Model Evaluation (Linear)



Nonetheless, there is a clear pattern to the residuals plot. First we underestimate, then overestimate, then underestimate again. This is evidence of a potential nonlinear relationship between the variables and the predictors which we will explore next.

[FOR ABDEL: make a 1x2 plot with the fitted vs residuals side by side with QQ plot]

Methodology for Complex Models

We started with a "full model" that includes the maximum number of two- and three-way interactions between TFr, Re, and St (including up to degree 2), fitted using least squares regression. Because two of the predictors have few unique observations and St already seems to have a linear relationship to the moments (as discovered through our linear regressions), we used only a 2nd degree polynomial. 2nd degree polynomials only need a minimum of 3 unique values to solve. Because the polynomial is relatively low degree, we are not worried about erratic behavior of the model and did not use a spline or smoothing method (when we did attempt to use these, they did not offer much better performance in terms of R^2 and error reduction). To increase our flexibility and capture nonlinear behavior between interactions, we considered many interactions, up to degree two. Nonetheless, from the "full model", we utilized a sequential forward stepwise selection method to select only the variables that decrease mean squared error (estimated by LOOCV) until we believe any added variables overfit the model.

Results for Complex Models

Model Evaluation