# yihan_shi_casestudy

Yihan Shi

10/31/2022

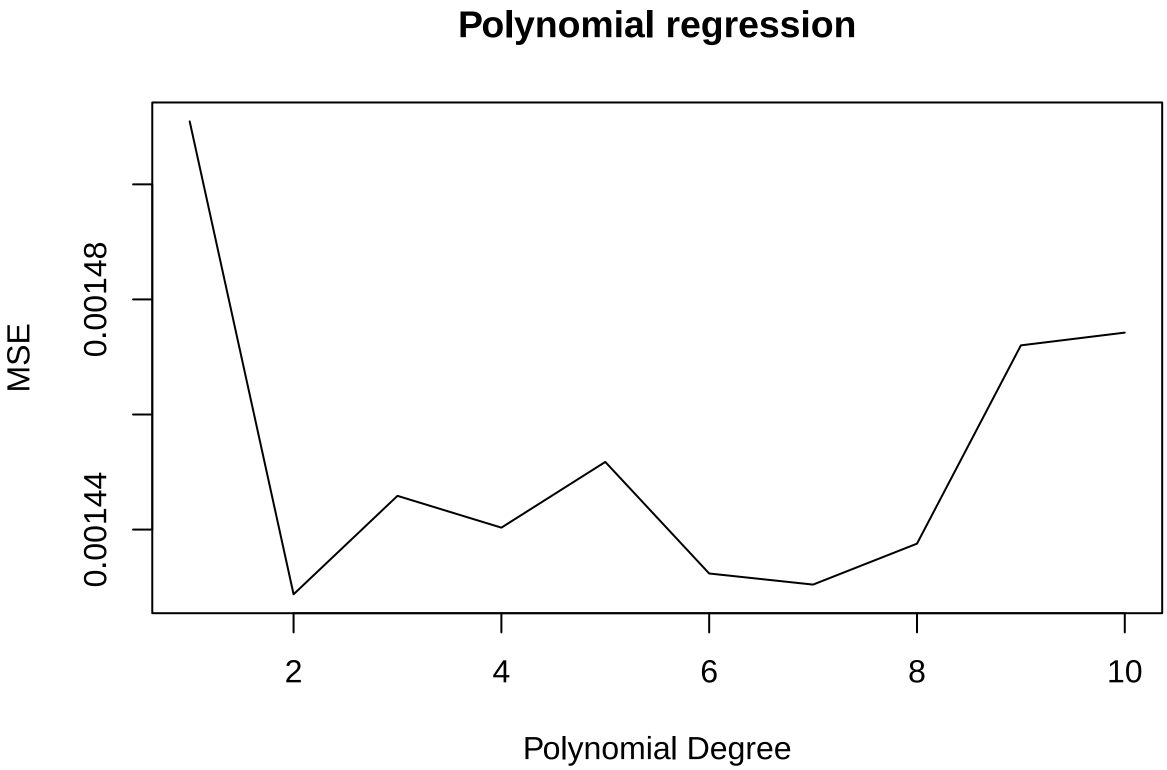
## Introduction

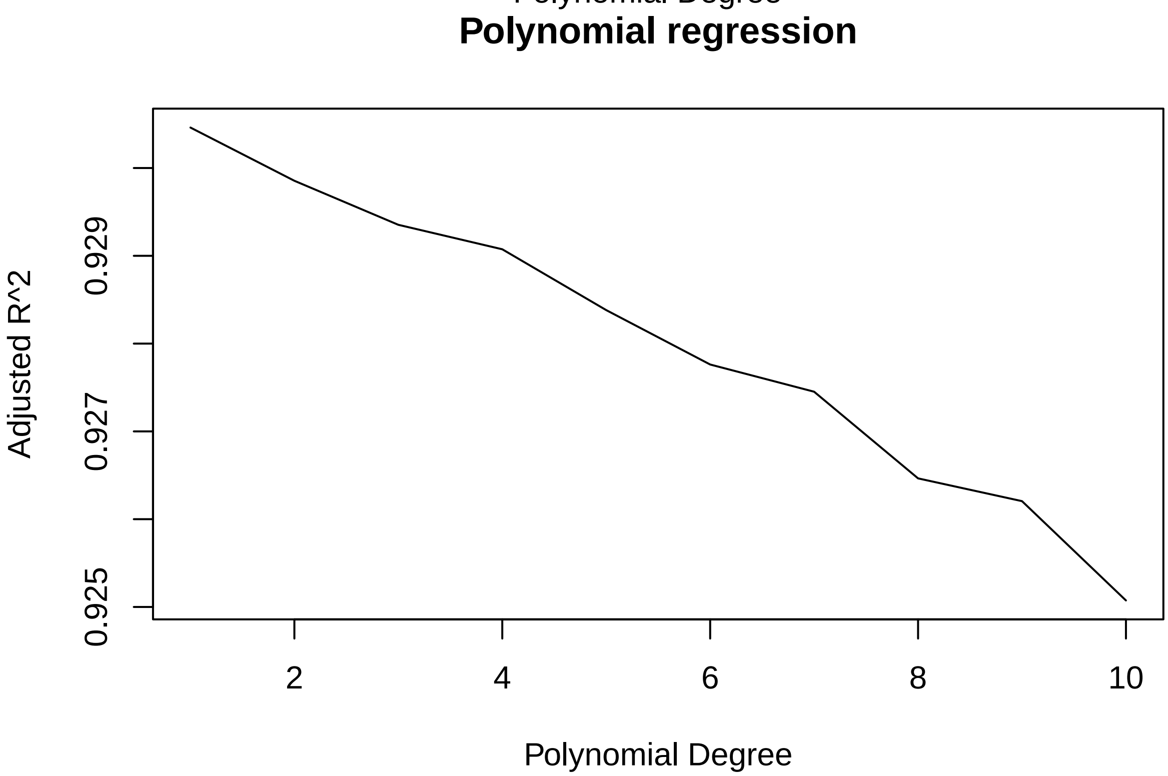
"We are experiencing some turbulence, please fasten your seat belt." Many of us might have heard this radio on the plane and felt bumpy. When we mix paint in water, we can also observe turbulence as the color dissipates. Turbulence is so common and easily observed in daily life, yet its causes and effects are hard to predict. In fluid dynamics, turbulence is "characterized by chaotic changes in pressure and flow velocity". With some knowledge and observation in parameters such as fluid density, flow speed, and the property of particles that cluster inside turbulent flows, we can gain insights into the distribution and clustering of particles in idealized turbulence. In this case study, we will investigate 3 observed features that might contribute to particle distribution in turbulence: Reynolds number (Re), which takes flow speed, viscosity, and density into account; Gravitational acceleration (Fr); Stokes number (St) that quantifies particle characteristics like size, relaxation time, and particle diameter. We hope to use these 3 features to explain changes in particle distribution as well as extrapolate beyond the scope of the known observations.

From the pair plot, there is somewhat linear or polynomial relationship between St and R_moment_1.

#### Linear regression

**Polynomial regression**



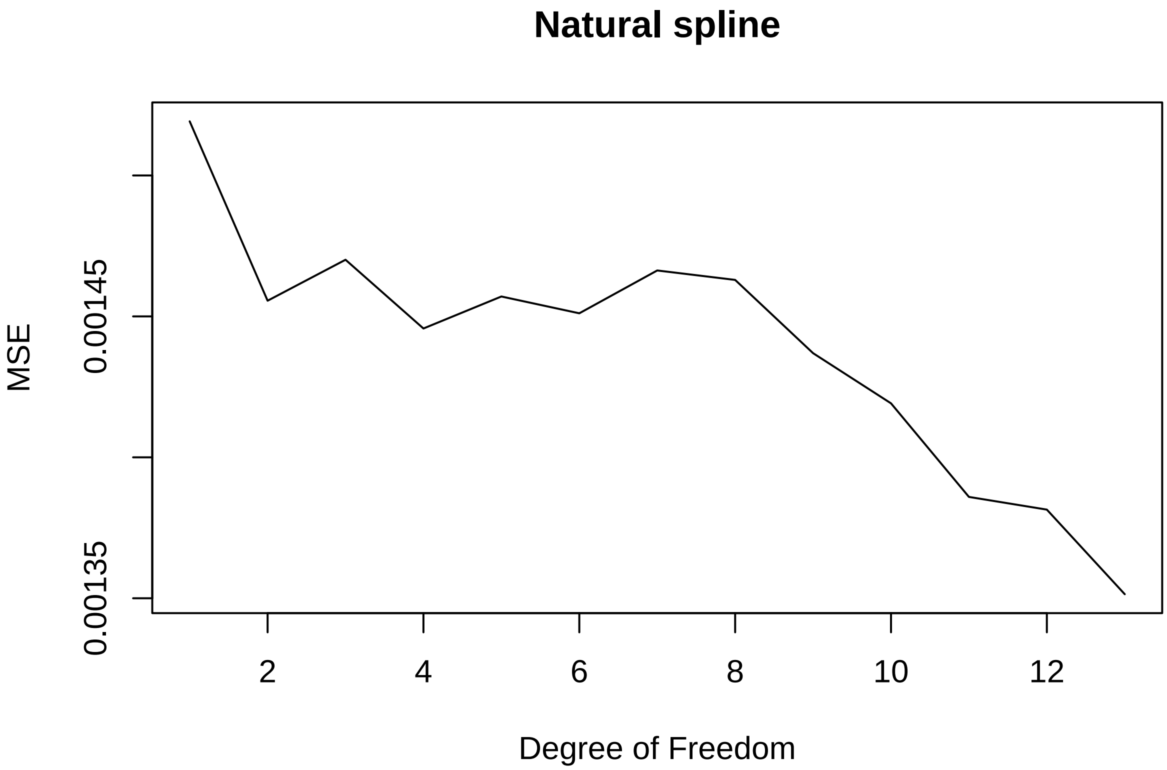



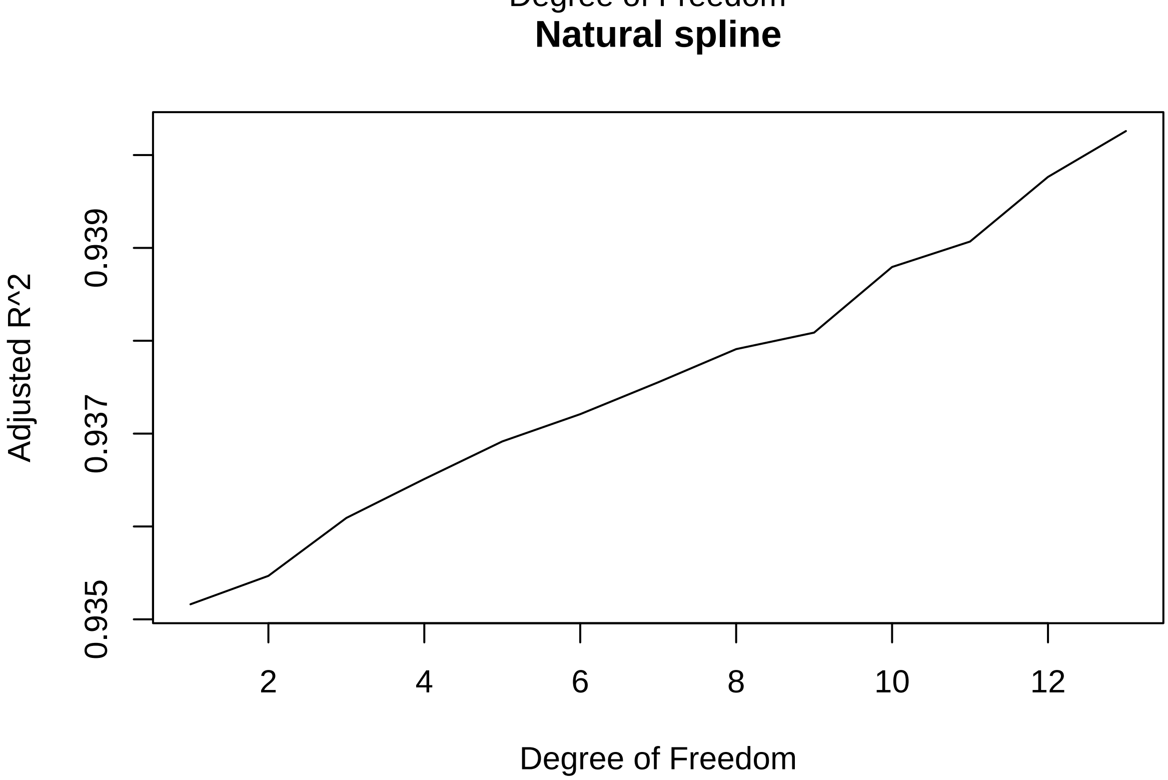

Using cross-validation with 5 folds, we see the adjusted $R^2$ decreases as polynomial degree increases, while the minimum MSE is achieved with order 2 polynomial on St. Linear regression is not suitable. Since adjusted $R^2$ is decent for any degree that we tested, we decided that the optimal polynomial order for St is 2.

We tried removing data of both high leverage and residual. However, this didn't change the MSE and adjusted

$R^2$ greatly. Since these observations take up 7% of the full training data, we decided that we don't want to exclude them.

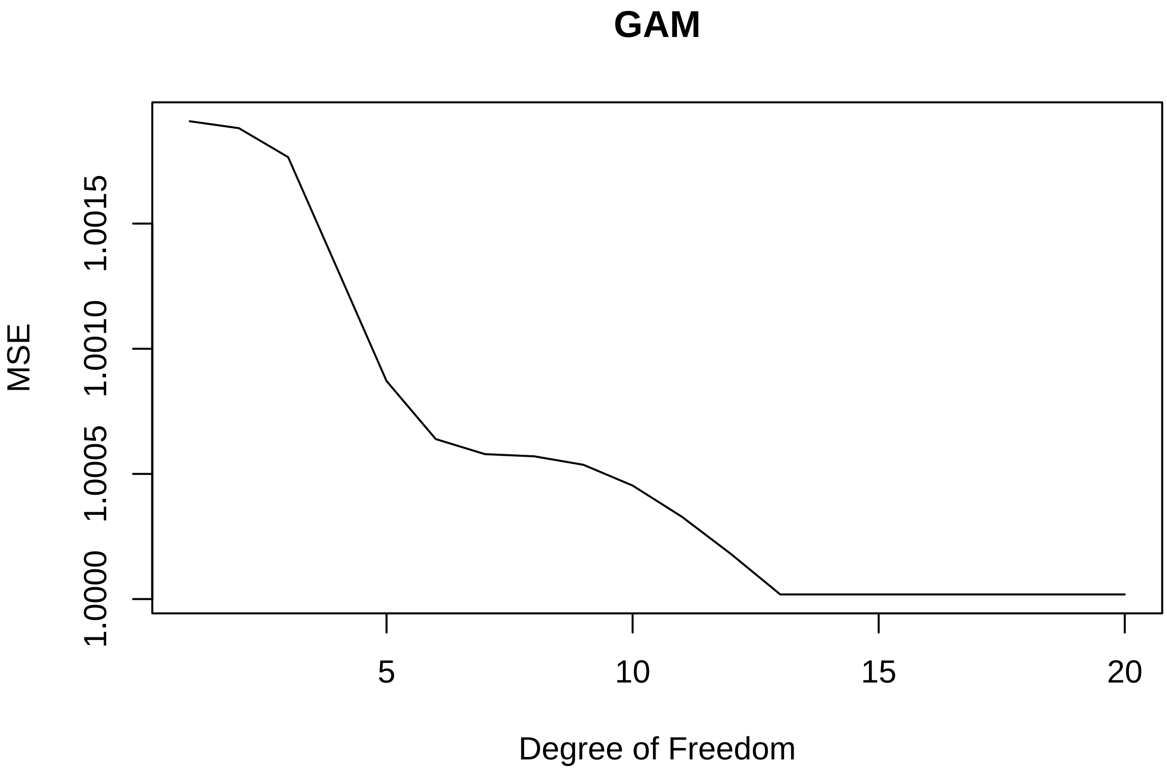
**Natural spline**

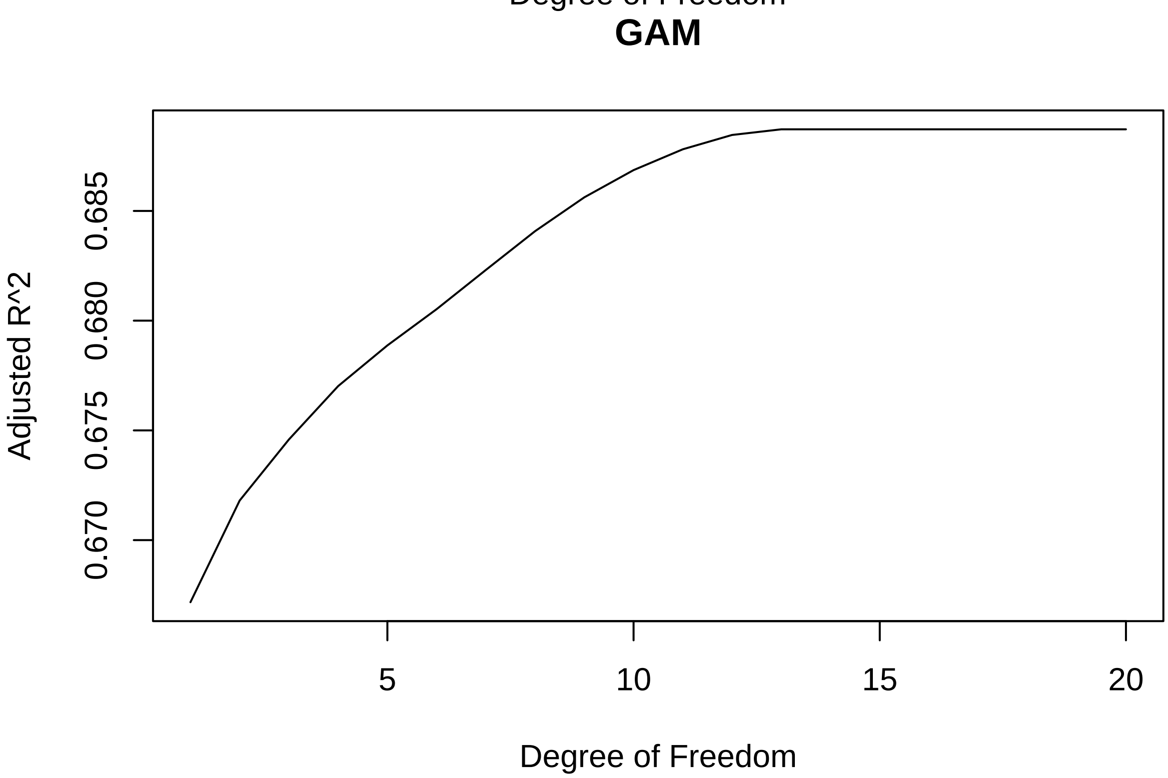



Using cross-validation with 5 folds, we see the adjusted $R^2$ increases as the degree of freedom increases, and the lowest MSE is achieved at degree of 13. The optimal polynomial degrees of freedom for St is 13, which achieves the lowest MSE and the highest adjusted $R^2$.

**GAM**




Using cross-validation with 5 folds, we see the adjusted $R^2$ increases as the degree of freedom increases to an extent

Table 1: Model MSE and Adjusted $R^2$

| models | formula | mse | adj.R |
|---|---|---:|---:|
| Linear regresion | R_moment_1 ~ Fr + Re + St + Fr * Re | 0.0015109 | 0.9304611 |
| Polynomial regression | R_moment_1 ~ Fr + Re + poly(St, 2) + Fr * Re | 0.0008570 | 0.9365328 |
| Natural spline | R_moment_1 ~ ns(St, df = 13) + Fr + Re + Fr * Re | 0.0007587 | 0.9429966 |
| Generalized additive model | R_moment_1 ~ s(St, 13) + Re + Fr + Fr * Re | 0.9588515 | 0.7163970 |

[*] Fr logit-transformed

(around 13). The optimal degree of freedom for GAM model to achieve the lowest MSE is also around 13. W might conclude that a degree of freedom around 13 is optimal.

Summary: Natural spline has the best results since it presents the highest adjusted $R^2$ and the lowest MSE. The second best model is polynomial regression, with only slight increase in MSE and decrease in adjusted $R^2$. Since natural splines are hard to interpret, we use it for prediction and polynomial regression for inference.