Case Study: Turbulence

Julia Rosner, Emily Mittleman, Tracy , *Ashley*

2022-11-3

Introduction

Understanding and predicting turbulence in fluid motion is incredibly important to a vast range of problems, such as air pollution, population dynamics, and weather. However, turbulence looks random, irregular, unpredictable, making it difficult to understand. Thus, our goals are as follows: For a new parameter setting of (Re, Fr, St), predict its particle cluster volume distribution in terms of its four raw moments. Inference: Investigate and interpret how each parameter affects the probability distribution for particle cluster volumes.

Methodology

Inference Our univariate exploratory data analysis of (Re, Fr, St) and each moment revealed that the R moments are heavily right skewed, which poses a problem to linear regression. We applied log transformations on each moment to obtain more normally distributed variables. We converted raw moments to central moments for inference so that our results are more interpretable. In addition, despite their numerical natures, Fr and Re only contain three levels ($Fr \in \{0.052, 0.3, \infty\}$ and $Re \in \{90, 224, 398\}$) in the training data. We decided to make them categorical variables for our inference model for 2 main reasons. Firstly, treating them as numeric variables puts our model at risk for extrapolation due to lack of data at many levels of Re and Fr, making our model unable to learn the trends around such regions; and secondly, we believe that these categories could carry real life significance. For instance, Fr = 0.3 is representative of cumulonimbus clouds and 0.052 is representative of cumulus clouds. Focusing on observations collected at such specific levels may lend unique insights into practical problems. A closer examination of the data revealed interesting interactive patterns among the independent and dependent variables. Specifically, St appears to assume a strong, non-linear relationship with each of the moments. This relationship between St and R moments appeared logarithmic, so we log-transformed St. Also, while treating the Fr variable as categorical, we observed that there is a linear and decreasing relationship between Fr (gravitational acceleration) and R moments. Lastly, there is also a linear and decreasing relationship between Re (Reynold's number) and R moments. We also explored multicollinearity through VIFs for each model, which were very low. DISCUSS SPLINE MODEL WITH INTERACTIONS

Prediction Despite certain advantages in modeling Re and Fr as categorical variables, we recognize the need to extrapolate beyond the three levels of Re and Fr to model real life circumstances and enable a wider range of prediction. To address the limitations of our current models, we provide the following related models using splines that can be used for extrapolation. In these models, Re is nuemric, and we transformed Fr to get rid of the infinity value that it takes on.

One of our goals for this case study was to make a model that is specialized in making accurate predictions, meaning that we are not concerned with inference for this particular model. Because we are not concerned with interpreting this model, we explored more complex models because they tend to have better predictive performance than simple models.

In order to predict values for new observations between and/or outside the ranges of values we have seen, we had to make a model that could both interpolate and extrapolate. Instead of using categorical values for our

predictor variables Fr and Re, we used numerical values so that our model is more generalizable to accept numerical values between or outside the ranges of Fr and Re that were given in the training data. In order to make Fr a numerical variable, we needed to perform a transformation on it to make Fr take on only finite values for modeling. To do this, we transformed Fr so that the new value is equal to $1 - 2^{-Fr}$, so that the new Fr takes on values between 0 and 1.

DISCUSS PREDICTION SPLINE MODEL W INTERACTIONS

OLDER STUFF DELETE ALL BELOW?

Our univariate exploratory data analysis of (Re, Fr, St) and each moment revealed that the R moments are heavily right skewed, which poses a problem to linear regression. We applied log transformations on each moment to obtain more normally distributed variables. Additionally, we used raw moments for prediction, but converted raw moments to central moments for inference so that our results are more interpretable. We also plotted each response variable (four moments) against each predictor variable to discern any relationships between them. From this, we saw that the relationship between St and R moments appear logarithmic, so we also log-transformed St. We also noticed that Fr has an infinity value, so transformed Fr to get rid of the infinity value transformation to get rid of infinity# From what we can see, treating the Fr variable as categorical, there is also a linear and decreasing relationship between Fr (gravitational acceleration) and R moments. Lastly, there is also a linear and decreasing relationship between Re (Reynold's number) and R moments. We also explored multicollinearity through VIFs for each model, which were very low.

Accordingly, we fit a basic linear model onto each log-transformed response variable. Linear models were evaluated based on adjusted R^2 values and P-values for coefficient estimates. We chose to treat Re and Fr as factors or categorical variables, as Re only takes on the values of 90, 224, and 398; Fr only takes on the values 0.052, 0.3, and infinity. While the adjusted R^2 value for the first raw moment was very high at 0.9949, subsequent moments exhibited decreasing adjusted R^2 values, with the fourth raw moment having an adjusted R^2 value of 0.6518. We also explored the addition of interaction terms to the model. The only interaction term which was significant for all raw moments was the interaction between Re and Fr. St and Re only have significant interaction for the first moment. Therefore, we constructed the linear models with an interaction between Re and Fr:

$$\begin{aligned} log(R_{moment1}) = -2.73 + 0.25(st) - 3.816(Re_{224}) - 5.988(Re_{398}) - 0.263(Fr_{0.3}) - 0.329(Fr_{\infty}) \\ +0.221(Re_{224} * Fr_{0.3}) + 0.402(Re_{224} * Fr_{\infty}) + 0.502(Re_{398} * Fr_{\infty}) \end{aligned}$$

$$\begin{aligned} log(R_{moment2}) = 5.187 + 0.834(st) - 7.434(Re_{224}) - 11.384(Re_{398}) - 6.416(Fr_{0.3}) - 6.652(Fr_{\infty}) \\ +4.387(Re_{224} * Fr_{0.3}) + 4.718(Re_{224} * Fr_{\infty}) + 7.076(Re_{398} * Fr_{\infty}) \end{aligned}$$

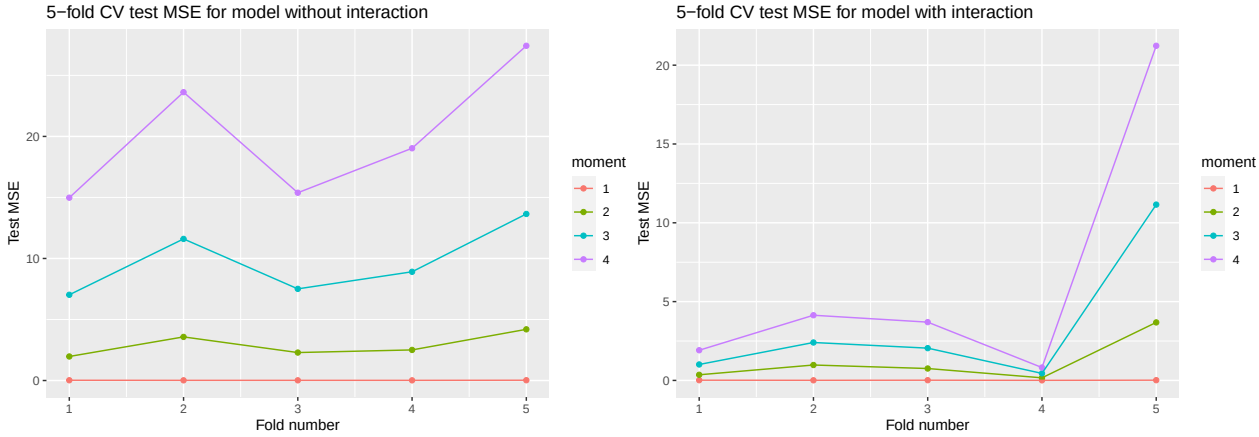
$$\begin{aligned} log(R_{moment2}) = 13.399 + 1.174(st) - 11.164(Re_{224}) - 17.032(Re_{398}) - 12.478(Fr_{0.3}) - 12.772(Fr_{\infty}) \\ +8.3648(Re_{224} * Fr_{0.3}) + 8.7718(Re_{224} * Fr_{\infty}) + 13.3707(Re_{398} * Fr_{\infty}) \end{aligned}$$

$$\begin{aligned} log(R_{moment2}) = 21.695 + 1.469(st) - 14.906(Re_{224}) - 22.715(Re_{398}) - 18.471(Fr_{0.3}) - 18.811(Fr_{\infty}) \\ +12.276(Re_{224} * Fr_{0.3}) + 12.756(Re_{224} * Fr_{\infty}) + 19.568(Re_{398} * Fr_{\infty}) \end{aligned}$$

Adding the interaction term between Re and Fr improved the fit of the model according to the adjusted R^2 values, which are much higher for every moment. With this new interaction term included, the adjusted R^2 value for R_moment_1 was slightly higher than before at 0.9966, and increased for moment 2 at 0.8909, moment 3 at 0.8770, and moment 4 0.8809 respectively.

To analyze predictive performance of our models, we applied an 80/20 train-test split and performed 5-fold cross validation. The linear models with the interaction term for Re and Fr outperformed any other linear model, producing lower test MSEs for every moment of R. Having this interaction term significantly improved the test MSEs of the linear model. # ADD PHYSICAL EXPLANATION; Reynolds number:t the ratio of

inertial forces to viscous forces within a fluid which is subjected to relative internal movement due to different fluid velocities. Whenever the Reynolds number is less than about 2,000, flow in a pipe is generally laminar, whereas, at values greater than 2,000, flow is usually turbulent. Once again, should we use central moments? Much easier to interpret: first central moment is related to mean, is always zero. Second central moment is the variance (spread) of the distribution. Third central moment measures symmetry, fourth central moment measures kurtosis (skewedness) of the distribution.



Although the linear model with interaction between Re and Fr performs much better than the linear model without interaction, it does not perform nearly as well in predicting higher raw moments. We can see this in the plot above: in the fifth fold, the test MSE for the first raw moment was 0.052, but the test MSE for the fourth raw moment was 28.83. The drastic difference between the two shows us that fitting a linear model for all four raw moments may be too simple.

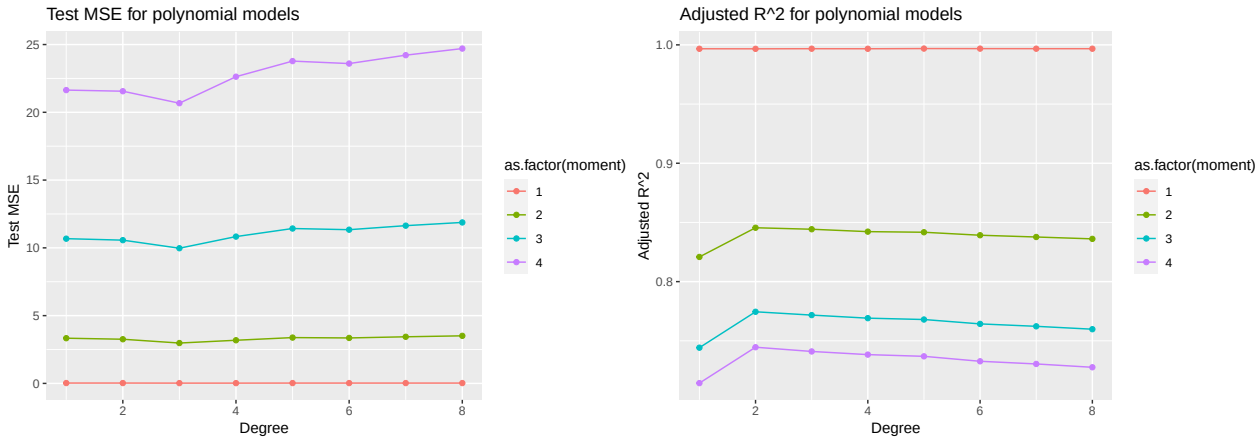
Polynomial Regression and Splines We also explore other linear models such as polynomial regression. Similarly we treat Fr and Re as categorical values and performed a log transformation on each raw moment as illustrated in the following function. In order to decide the optimal polynomial degree for each raw moment, we apply a simple validation approach to select the best among models with D different polynomial orders. However, after comparing the validation MSEs between those polynomial models and the above linear regression model with interactions, both training and testing accuracy are much smaller.

$$log(R_{moment_i}) = \beta_0 + \beta_{1...d} poly(St, d) + \beta_{d+1} Fr + \beta_{d+2} Re$$

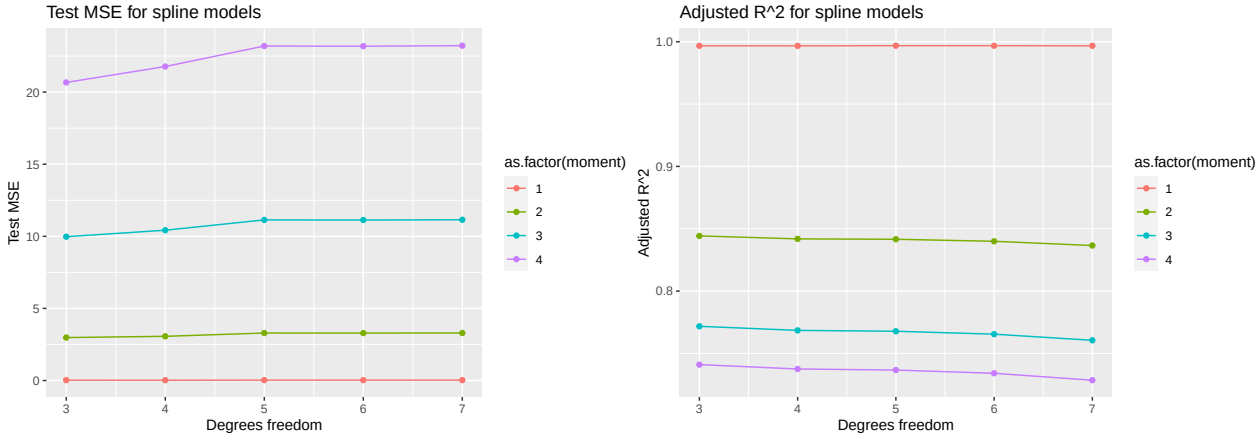
We tried to tackle this by adding interaction terms to the initial polynomial model. And this significantly improved the regression model's performance both on training adjusted R^2 and testing MSEs. The regression function is formulated as below:

$$log(R_{moment_i}) = \beta_0 + \beta_{1...d} poly(St, d) + \beta_{d+1} Fr + \beta_{d+2} Re + \beta_* Fr * Re$$

After training and selecting the optimal polynomial degrees for each raw moment, the adjusted training R^2 reaches 0.9983 at 1st raw moment, and 2nd raw moment at 0.9677, 3nd raw moment at 0.9585, 4th raw moment at 0.9572, respectively. Also, the optimal polynomial models' predictive performances are much better than the above linear model with interaction according to their testing MSEs for all moments.

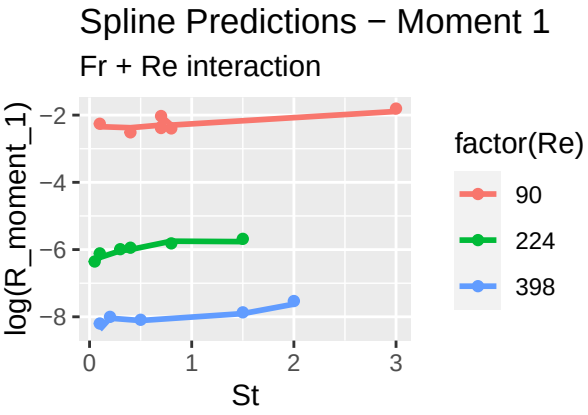


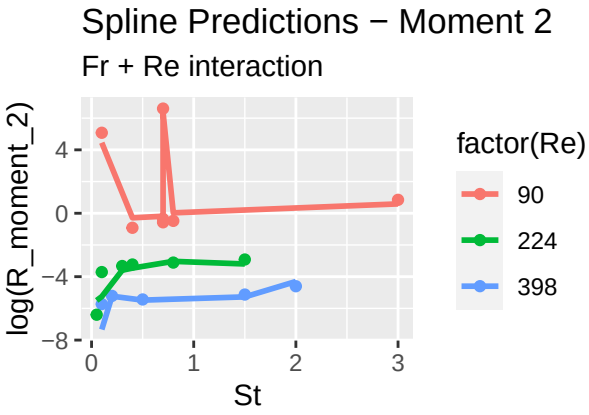
ADD INTERPRETATION OF SPLINES

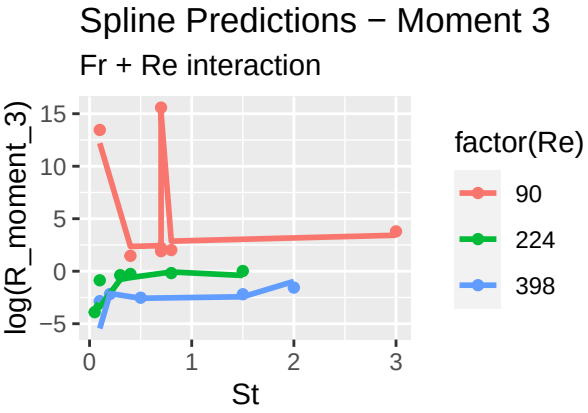


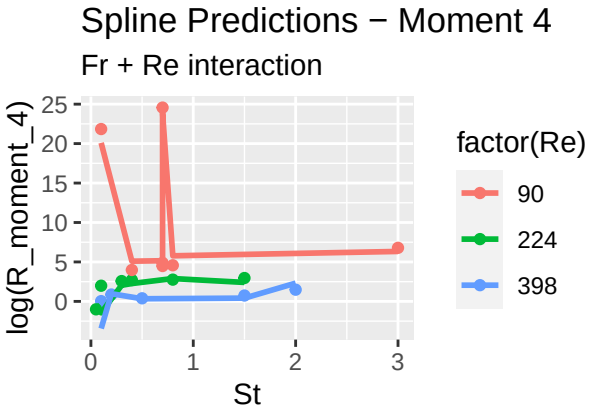
- interpret interaction between Re and Fr physically
- Are the effects identified above similar over all central moments (i.e., over all response variables), or are there effects which differ between, say, the mean and the variance? Try to interpret the latter effects using the three parameters mean physically
- central vs raw moments?

Since the performance of the cubic polynomial model still showed much room for improvement, we further explored splines with three degrees of freedom. (cubic splines?) We expected the splines to perform better compared to polynomial regression because instead of fitting a single polynomical function to the entire dataset, spline interpolation fits a piecewise continuous function composed of many polynomials to model the data set.





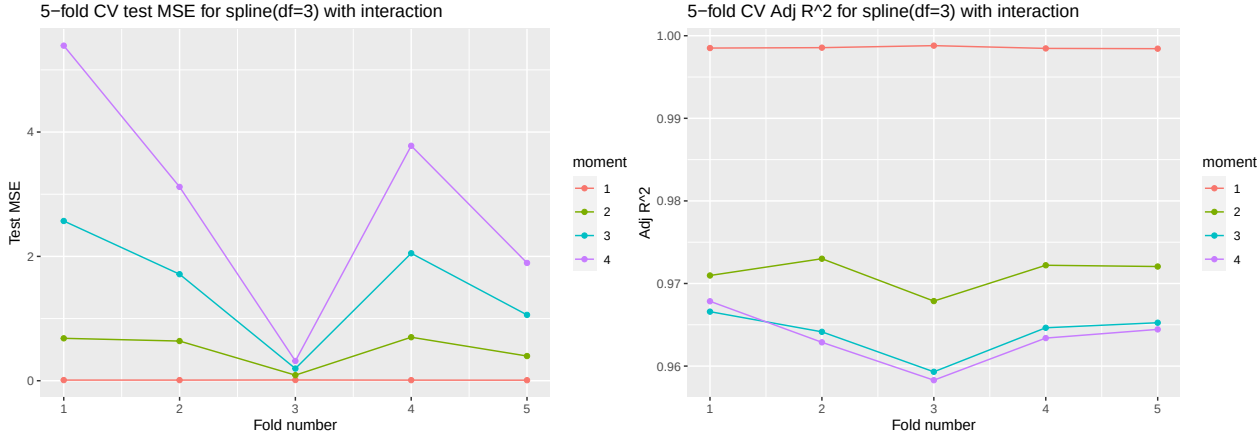




	Moment 1	Moment 2	Moment 3	Moment 4
MSE	0.007	0.424	1.077	1.895
Ajd R^2	0.998	0.970	0.962	0.961

Cross Validation again

RMSE	0.098	0.740	1.827	0.098
Rsquared	0.998	0.968	0.940	0.998
MAE	0.069	0.450	1.026	0.069



Note: try cross validation for spline with interaction? To assess whether it's overfitting?

- base linear model: St doesn't even seem to be a significant predictor in some of these models. Linear regression might not work super well here.
- interaction terms: Re and g(Fr) appear to be significant, but St doesn't have any significant interactions.
- overall linear regression both with and without interaction terms: MSE comparable for each moment, Really bad results, interaction terms aren't helping much. Let's try some other methods other than simple linear regression.
- splines without interaction: better performance than linear model, but still room for improvement
- splines with interaction: Adding the interaction term to the splines makes the performance of this model superior to anything we've used previously (when considering numerical datapoints) based on Adj R^2 and MSE in test validation.

Formatting here needs to be fixed

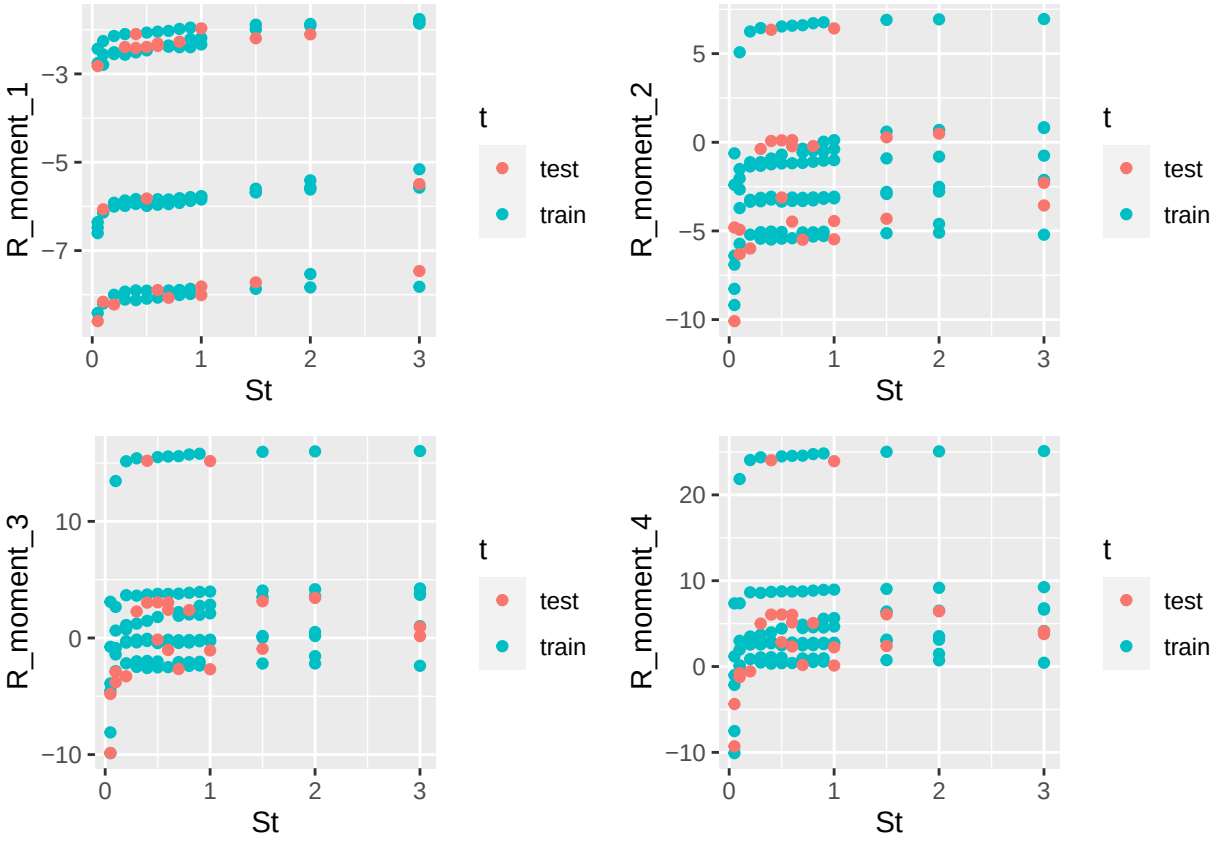
Prediction

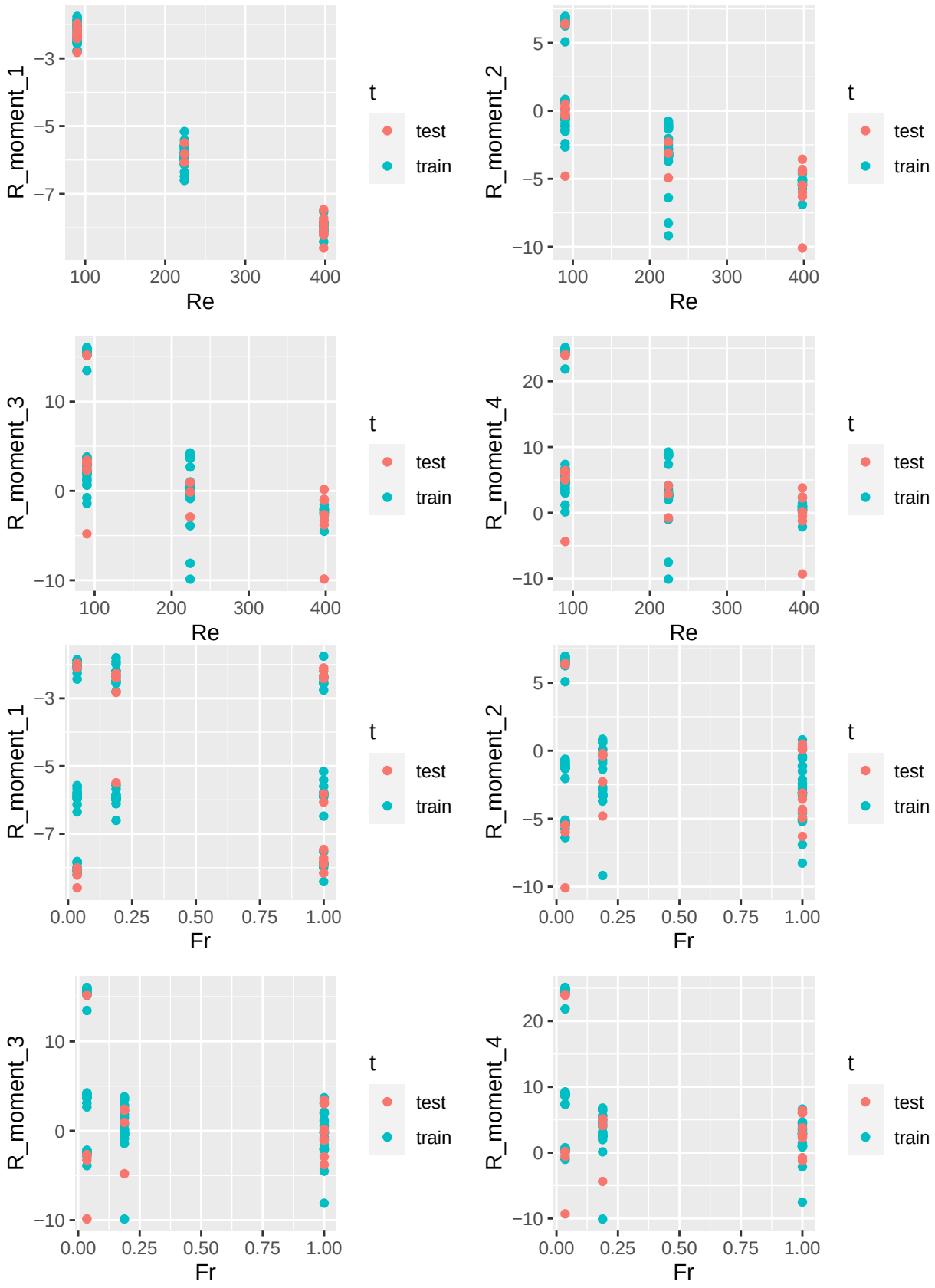
```
## Warning in predict.lm(spline1, data_test): prediction from a rank-deficient fit
## may be misleading
```

```
## Warning in predict.lm(spline2, data_test): prediction from a rank-deficient fit
## may be misleading
```

```
## Warning in predict.lm(spline3, data_test): prediction from a rank-deficient fit
## may be misleading
```

```
## Warning in predict.lm(spline4, data_test): prediction from a rank-deficient fit
## may be misleading
```




R_moment_1
Re
R_moment_2
Re
R_moment_3
Re
R_moment_4
Re
R_moment_1
Fr
R_moment_2
Fr
R_moment_3
Fr
R_moment_4
Fr
t
test
train

Results

Conclusion