Final Writeup

Kadriye Sude Almus

10/25/2020

The Data

```
data <- read.csv("data/data-train.csv")
```

Final Model and Predictions

```
attach(data)
set.seed(3)
train_ind <- sample(x = nrow(data), size = 0.8 * nrow(data))
test_ind_neg <- -train_ind
training <- data[train_ind, ]
testing <- data[test_ind_neg, ]
training
```

```
##      St  Re    Fr R_moment_1 R_moment_2 R_moment_3 R_moment_4
## 5  0.70 398   Inf 0.00036945 6.2242e-03 1.2649e-01 2.5714e+00
## 58 2.00  90 0.052 0.15433000 1.0269e+03 8.8700e+06 7.6700e+10
## 12 0.30 398   Inf 0.00036022 6.2830e-03 1.3546e-01 2.9211e+00
## 36 0.90 224 0.300 0.00280490 4.1143e-02 7.5132e-01 1.3729e+01
## 88 0.60 224   Inf 0.00291710 4.4317e-02 8.5282e-01 1.6431e+01
## 8  3.00 224   Inf 0.00574730 1.1966e-01 2.7480e+00 6.3159e+01
## 20 0.20  90 0.052 0.11760000 5.1774e+02 3.8100e+06 2.8000e+10
## 74 0.90  90 0.052 0.14184000 8.7019e+02 7.2500e+06 6.0400e+10
## 55 0.40 398 0.052 0.00029691 4.1375e-03 7.6124e-02 1.4014e+00
## 40 0.20 224 0.052 0.00257870 2.6830e-01 3.9080e+01 5.6959e+03
## 48 0.30 224 0.052 0.00256750 2.6547e-01 3.7665e+01 5.3451e+03
## 84 0.90 398 0.052 0.00034145 5.0555e-03 9.4083e-02 1.7522e+00
## 37 0.40 224   Inf 0.00292630 4.6261e-02 9.2914e-01 1.8681e+01
## 66 2.00 398 0.052 0.00039644 6.1040e-03 1.1209e-01 2.0593e+00
## 29 0.30  90 0.052 0.12261000 6.2727e+02 4.9100e+06 3.8500e+10
## 89 1.50 224   Inf 0.00370310 6.0910e-02 1.1829e+00 2.2990e+01
## 77 0.60  90 0.052 0.12946000 7.1816e+02 5.7200e+06 4.5600e+10
## 9  0.90 224   Inf 0.00302150 4.5244e-02 8.4530e-01 1.5809e+01
## 72 0.20 224 0.300 0.00246950 3.4818e-02 6.7088e-01 1.2939e+01
## 18 1.00  90 0.300 0.11236000 1.1261e+00 1.7335e+01 2.8261e+02
## 16 0.60 224 0.300 0.00257400 3.6621e-02 6.7102e-01 1.2309e+01
## 25 0.90 398   Inf 0.00038344 6.4432e-03 1.2925e-01 2.5935e+00
## 87 0.05 224 0.300 0.00135380 1.0303e-04 5.1400e-05 4.1600e-05
## 80 3.00  90 0.052 0.15538000 1.0443e+03 9.1400e+06 8.0000e+10
## 22 0.50 398   Inf 0.00036800 6.3559e-03 1.3341e-01 2.8013e+00
## 60 0.30  90   Inf 0.07694500 3.2652e-01 3.4052e+00 4.1042e+01
## 51 3.00  90   Inf 0.17234000 2.2386e+00 4.0454e+01 7.6198e+02
```

```
## 6  2.00  90 0.300 0.14780000 2.0068e+00 3.6249e+01 6.7167e+02
## 76 0.20  90 0.300 0.07798500 2.5598e-01 2.0965e+00 2.0849e+01
## 15 0.50  90 0.052 0.12670000 6.8596e+02 5.4300e+06 4.2900e+10
## 10 0.60 398 0.052 0.00031431 4.4672e-03 8.2060e-02 1.5077e+00
## 11 0.90  90   Inf 0.09102700 5.9539e-01 7.2454e+00 9.5166e+01
## 75 0.40 398   Inf 0.00036977 6.4986e-03 1.3933e-01 2.9880e+00
## 23 0.20 224   Inf 0.00269160 3.9016e-02 7.6384e-01 1.4978e+01
## 3  0.70 224   Inf 0.00290540 4.3499e-02 8.2200e-01 1.5551e+01
## 73 3.00 398 0.052 0.00040188 5.4492e-03 9.1871e-02 1.5565e+00
## 78 1.50  90 0.300 0.13678000 1.8254e+00 3.2833e+01 6.0903e+02
## 59 0.90  90 0.300 0.10962000 1.0319e+00 1.5797e+01 2.6136e+02
## 4  0.05  90   Inf 0.06352800 9.0653e-02 4.6746e-01 3.2696e+00
## 33 1.50  90 0.052 0.15181000 9.9690e+02 8.5500e+06 7.3300e+10
## 19 0.70 224 0.052 0.00285610 3.1273e-01 4.4529e+01 6.3423e+03
## 38 0.80 398 0.052 0.00033341 4.9036e-03 9.1143e-02 1.6948e+00
## 26 0.50 224 0.052 0.00274240 3.0355e-01 4.3911e+01 6.3530e+03
## 45 0.80 224 0.052 0.00295750 3.3361e-01 4.8161e+01 6.9539e+03
## 7  0.20  90   Inf 0.08127300 3.2450e-01 3.0363e+00 3.2976e+01
## 1  0.10 224 0.052 0.00215700 1.3035e-01 1.4374e+01 1.5865e+03
## 71 0.80 224 0.300 0.00268160 3.7714e-02 6.7549e-01 1.2112e+01
## 41 0.30 224   Inf 0.00283770 4.3589e-02 8.6962e-01 1.7373e+01
## 82 0.05  90 0.052 0.08786800 5.3449e-01 2.2205e+01 1.5679e+03
## 62 0.50  90 0.300 0.08477300 4.9728e-01 6.0317e+00 8.3287e+01
## 31 1.50 224 0.052 0.00341630 4.0300e-01 5.8417e+01 8.4710e+03
## 86 0.05 224   Inf 0.00153380 2.5653e-04 3.0407e-04 5.4466e-04
## 14 1.00 224 0.052 0.00312380 3.6478e-01 5.3322e+01 7.7958e+03
## 43 2.00 224 0.300 0.00381230 6.1927e-02 1.1844e+00 2.2705e+01
## 81 2.00 224 0.052 0.00363470 4.4512e-01 6.5387e+01 9.6105e+03
## 63 0.80 398   Inf 0.00037399 6.2457e-03 1.2542e-01 2.5193e+00
## 67 0.70 224 0.300 0.00260870 3.6438e-02 6.5445e-01 1.1765e+01
## 70 0.90 224 0.052 0.00305410 3.5419e-01 5.1795e+01 7.5758e+03
## 17 0.10  90   Inf 0.07722700 2.2120e-01 1.8833e+00 2.0190e+01
## 69 0.30 398 0.052 0.00030066 4.3488e-03 8.3446e-02 1.6023e+00
## 34 0.05 398   Inf 0.00022202 1.0055e-03 1.0857e-02 1.1782e-01
## 61 0.50 224 0.300 0.00250710 3.5152e-02 6.4378e-01 1.1801e+01
## 44 1.00  90   Inf 0.09691800 6.7696e-01 8.2384e+00 1.0602e+02
## 13 2.00 224   Inf 0.00447250 8.0804e-02 1.6668e+00 3.4408e+01
## 2  3.00 224 0.052 0.00379030 4.7042e-01 6.9940e+01 1.0404e+04
## 21 0.10  90 0.300 0.06125200 6.9867e-02 2.4338e-01 1.1379e+00
## 68 0.60 224 0.052 0.00279390 3.0594e-01 4.3745e+01 6.2554e+03
## 46 1.00 224 0.300 0.00289530 4.2300e-02 7.6755e-01 1.3941e+01
## 52 1.00 224   Inf 0.00309680 4.6454e-02 8.6381e-01 1.6077e+01
## 28 0.40 224 0.052 0.00268090 2.8897e-01 4.1585e+01 5.9861e+03
## 54 0.80  90 0.052 0.13793000 8.2524e+02 6.8000e+06 5.6100e+10
testing
```

```
##      St  Re    Fr R_moment_1 R_moment_2 R_moment_3 R_moment_4
## 24 1.50 398 0.052 0.00038321 5.9338e-03 1.1156e-01 2.1004e+00
## 27 0.10 398 0.052 0.00027479 3.2549e-03 5.8006e-02 1.0344e+00
## 30 0.05 224 0.052 0.00173740 1.6633e-03 2.0228e-02 3.6438e-01
## 32 0.80  90   Inf 0.09107400 6.1825e-01 7.4973e+00 9.7048e+01
## 35 0.80 224   Inf 0.00298090 4.4580e-02 8.3764e-01 1.5759e+01
## 39 1.50 224 0.300 0.00341050 5.4101e-02 1.0222e+00 1.9340e+01
## 42 0.30 224 0.300 0.00250630 3.5881e-02 6.8596e-01 1.3132e+01
```

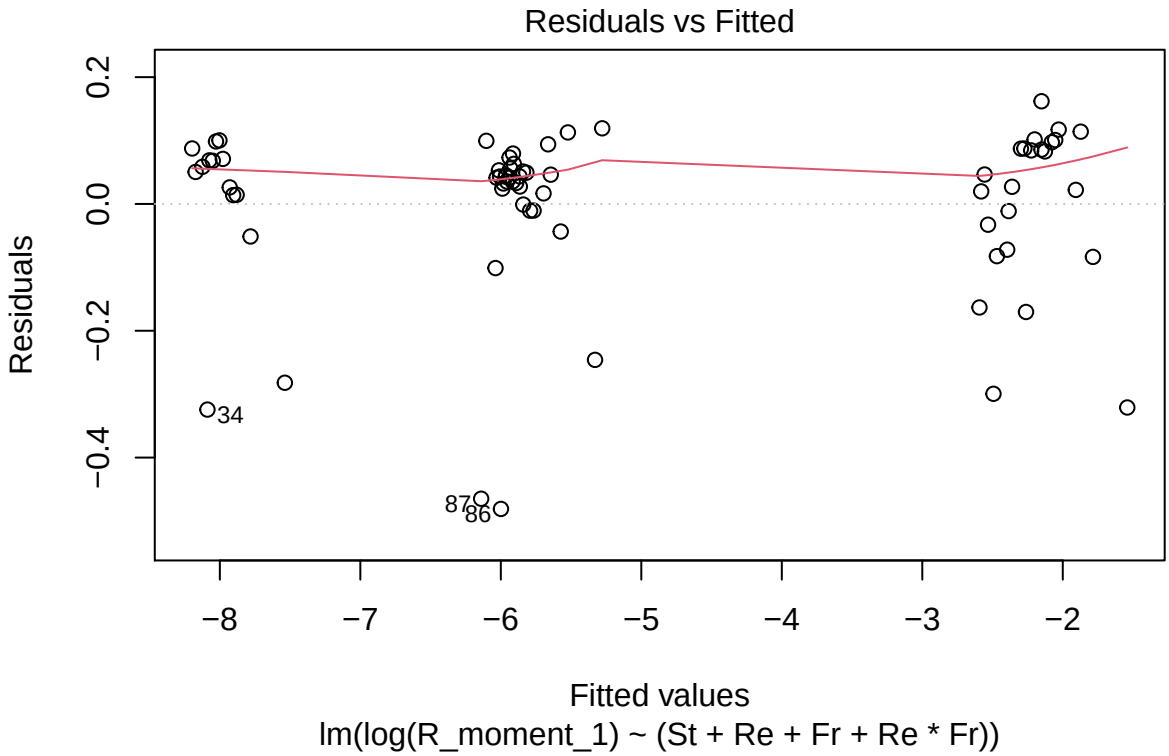
```
## 47 0.70  90 0.300 0.09471100 6.9751e-01 9.1793e+00 1.3187e+02
## 49 0.40 224 0.300 0.00262070 3.9502e-02 7.6851e-01 1.4966e+01
## 50 0.10 224 0.300 0.00221530 2.4475e-02 4.2167e-01 7.2842e+00
## 53 2.00 398   Inf 0.00053647 1.0022e-02 2.1023e-01 4.4109e+00
## 56 0.50 398 0.052 0.00030716 4.3494e-03 8.0143e-02 1.4770e+00
## 57 0.70  90   Inf 0.09217600 5.6482e-01 6.7191e+00 8.8723e+01
## 64 0.20 398   Inf 0.00033521 5.4505e-03 1.1408e-01 2.3884e+00
## 65 0.70  90 0.052 0.13173000 7.3694e+02 5.8700e+06 4.6700e+10
## 79 0.10  90 0.052 0.10464000 1.6015e+02 6.9900e+05 3.0700e+09
## 83 0.40  90 0.300 0.08095700 3.9996e-01 4.3303e+00 5.3618e+01
## 85 3.00  90 0.300 0.16433000 2.3317e+00 4.4516e+01 8.8779e+02
```

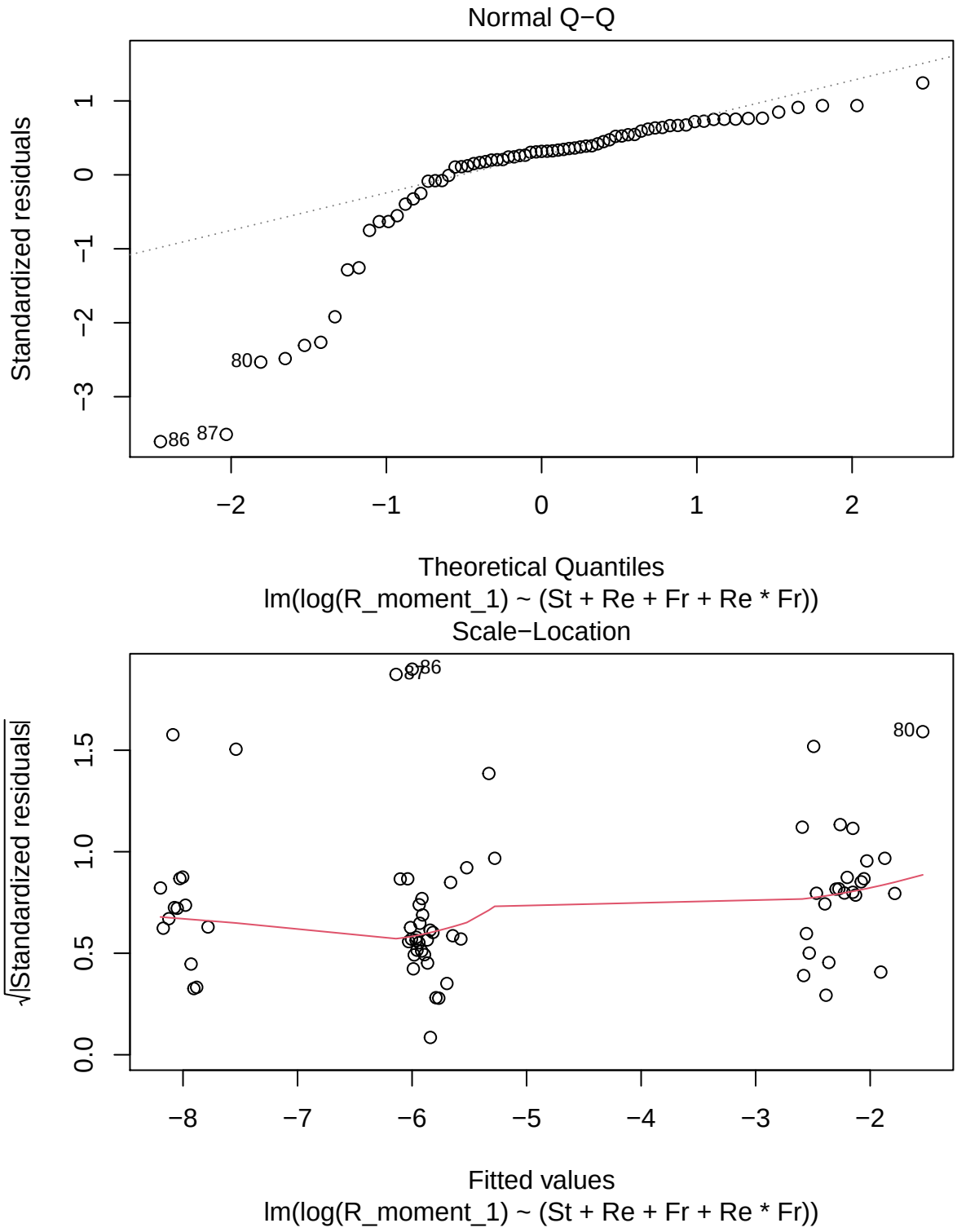
```
ftraining <- training
ftesting <- testing
ftraining$Fr <- factor(ftraining$Fr, levels = c(0.052, 0.300, Inf))
ftraining$Re <- factor(ftraining$Re, levels = c(90, 224, 398))
ftesting$Fr <- factor(ftesting$Fr, levels = c(0.052, 0.300, Inf))
ftesting$Re <- factor(ftesting$Re, levels = c(90, 224, 398))
```

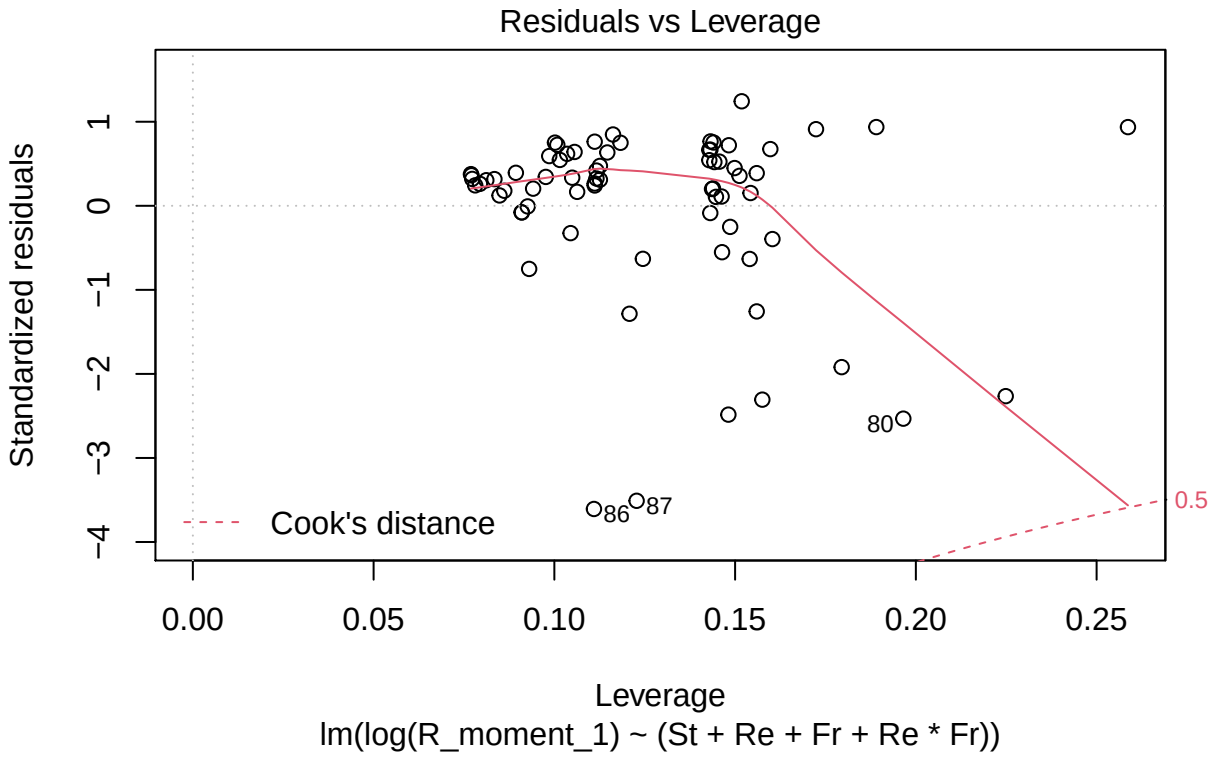
```
fit.lm1 <- lm(log(R_moment_1) ~ (St + Re + Fr + Re*Fr), data = ftraining)
pred.lm1 <- predict(fit.lm1, ftesting)
```

```
## Warning in predict.lm(fit.lm1, ftesting): prediction from a rank-deficient fit
## may be misleading
```

```
plot(fit.lm1)
```




Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(log(R_moment_1) ~ (St + Re + Fr + Re * Fr))
Scale−Location
√|Standardized residuals|
Fitted values
lm(log(R_moment_1) ~ (St + Re + Fr + Re * Fr))



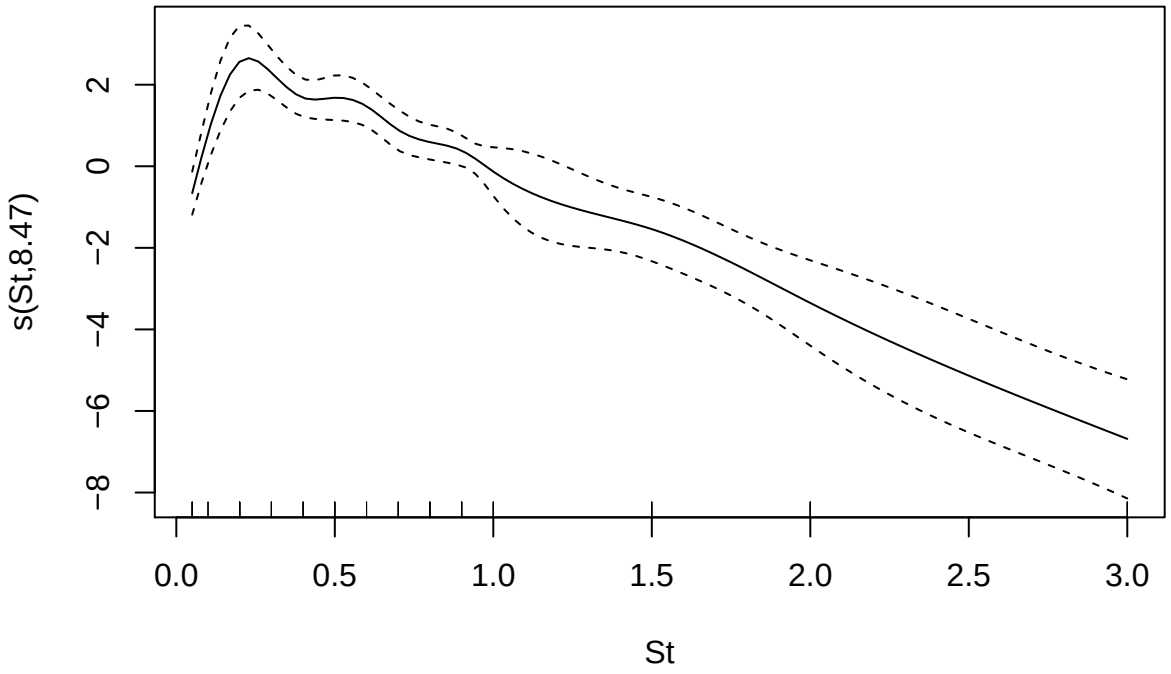

```
summary(fit.lm1)
```

```
##
## Call:
## lm(formula = log(R_moment_1) ~ (St + Re + Fr + Re * Fr), data = ftraining)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.48100 -0.01056  0.04300  0.08114  0.16195
##
## Coefficients: (1 not defined because of singularities)
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.27376    0.04961 -45.828  < 2e-16 ***
## St           0.24431    0.02181  11.203  < 2e-16 ***
## Re224       -3.78857    0.05950 -63.669  < 2e-16 ***
## Re398       -5.99652    0.06978 -85.932  < 2e-16 ***
## Fr0.3       -0.24403    0.06973  -3.500 0.000869 ***
## FrInf       -0.33146    0.06982  -4.747 1.26e-05 ***
## Re224:Fr0.3  0.15420    0.09285   1.661 0.101811
## Re398:Fr0.3       NA         NA      NA       NA
## Re224:FrInf  0.38257    0.09078   4.214 8.28e-05 ***
## Re398:FrInf  0.50112    0.10325   4.853 8.55e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1414 on 62 degrees of freedom
## Multiple R-squared:  0.9963, Adjusted R-squared:  0.9959
## F-statistic:  2106 on 8 and 62 DF,  p-value: < 2.2e-16
```

```
mse_test1 <- mean((pred.lm1 - log(testing$R_moment_1))^2)
mse_test1
```

```
## [1] 0.008822464
```

```
gam.m2 = gam(log(R_moment_2) ~ s(St) + Re + Fr + St:Re + St:Fr + Re:Fr, data = ftraining)
plot(gam.m2)
```

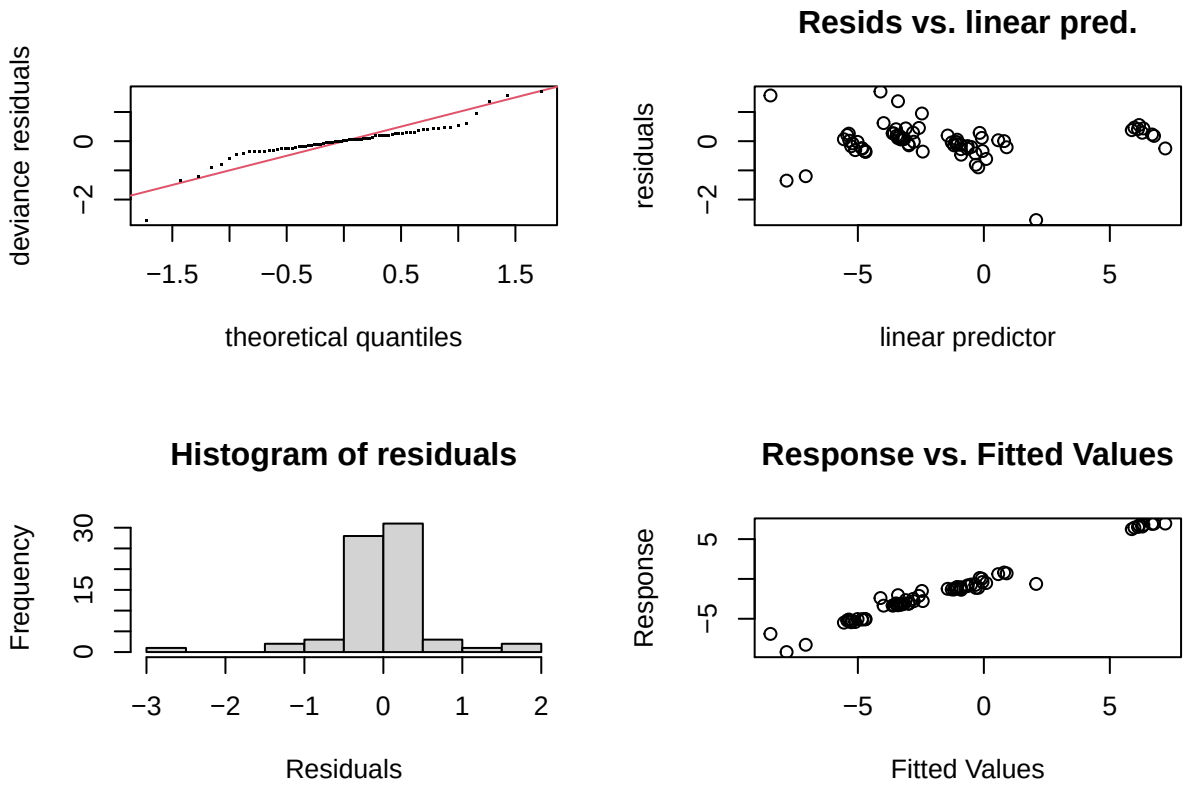



```
summary(gam.m2)
```

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## log(R_moment_2) ~ s(St) + Re + Fr + St:Re + St:Fr + Re:Fr
##
## Parametric coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.54657    0.37518   6.788 1.24e-08 ***
## Re224        -7.11591    0.37585 -18.933  < 2e-16 ***
## Re398       -11.08527    0.52216 -21.230  < 2e-16 ***
## Fr0.3        -6.87397    0.49181 -13.977  < 2e-16 ***
## FrInf        -6.17375    0.42687 -14.463  < 2e-16 ***
## Re90:St       3.77928    0.35437  10.665 1.60e-14 ***
## Re224:St      3.63936    0.36826   9.883 2.19e-13 ***
## Re398:St      3.26669    0.38455   8.495 2.71e-11 ***
## Fr0.3:St      0.51261    0.38567   1.329    0.190
## FrInf:St     -0.07419    0.25302  -0.293    0.771
## Re224:Fr0.3   4.06642    0.47277   8.601 1.86e-11 ***
## Re398:Fr0.3   0.00000    0.00000      NA       NA
## Re224:FrInf   4.15555    0.46108   9.013 4.39e-12 ***
## Re398:FrInf   6.74397    0.55523  12.146  < 2e-16 ***
```

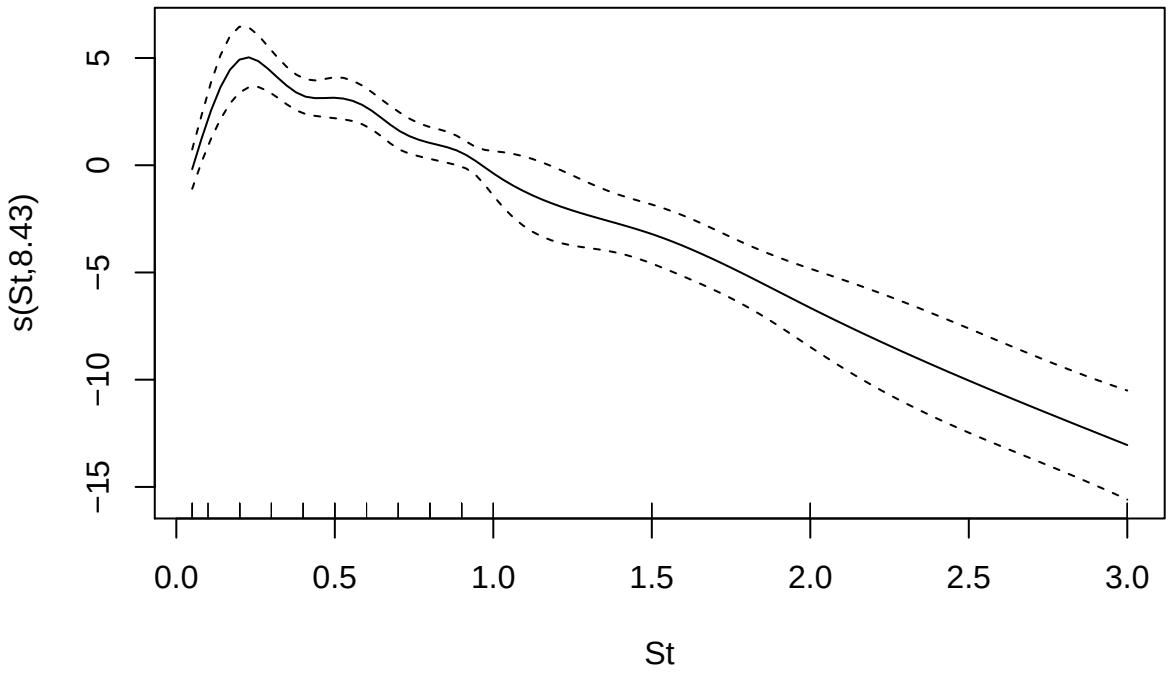
```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##         edf Ref.df     F p-value
## s(St) 8.467  8.799 17.03  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Rank: 21/23
## R-sq.(adj) =  0.965   Deviance explained = 97.5%
## GCV = 0.69646  Scale est. = 0.49431   n = 71
```

```
gam.check(gam.m2)
```




```
##
## Method: GCV   Optimizer: magic
## Smoothing parameter selection converged after 8 iterations.
## The RMS GCV score gradient at convergence was 3.715048e-07 .
## The Hessian was positive definite.
## Model rank =  21 / 23
##
## Basis dimension (k) checking results. Low p-value (k-index<1) may
## indicate that k is too low, especially if edf is close to k'.
##
##         k'  edf k-index p-value
## s(St) 9.00 8.47    1.08    0.71
```

```
gam.m3 = gam(log(R_moment_3) ~ s(St) + Re + Fr + St:Re + St:Fr + Re:Fr, data = ftraining)
plot(gam.m3)
```

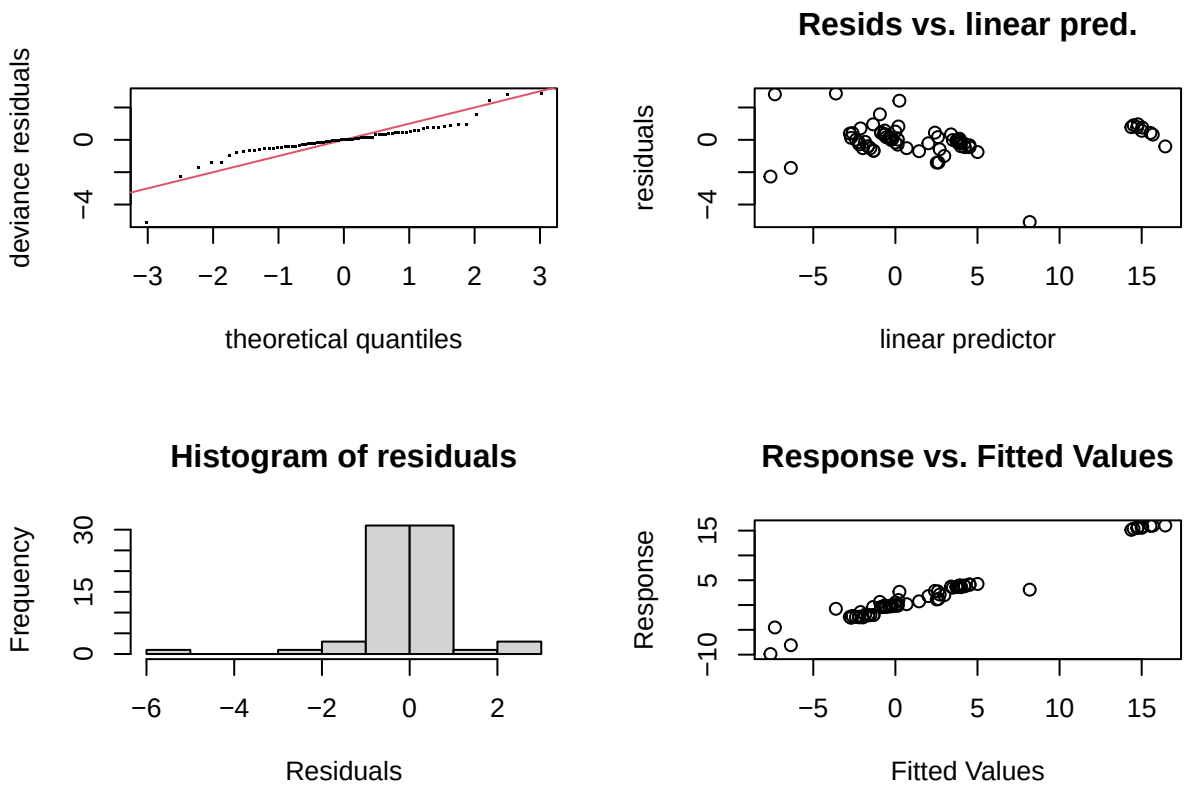



```
summary(gam.m3)
```

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## log(R_moment_3) ~ s(St) + Re + Fr + St:Re + St:Fr + Re:Fr
##
## Parametric coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   8.0050     0.6534  12.251  < 2e-16 ***
## Re224       -10.6005     0.6575 -16.124  < 2e-16 ***
## Re398       -16.4306     0.9132 -17.993  < 2e-16 ***
## Fr0.3       -13.0638     0.8601 -15.188  < 2e-16 ***
## FrInf       -11.7892     0.7467 -15.788  < 2e-16 ***
## Re90:St       7.1596     0.6164  11.614 7.29e-16 ***
## Re224:St      6.8820     0.6412  10.734 1.26e-14 ***
## Re398:St      6.2334     0.6705   9.296 1.63e-12 ***
## Fr0.3:St      0.6865     0.6746   1.018    0.314
## FrInf:St     -0.3003     0.4426  -0.678    0.501
## Re224:Fr0.3   7.8560     0.8270   9.499 8.08e-13 ***
## Re398:Fr0.3   0.0000     0.0000      NA       NA
## Re224:FrInf   7.8729     0.8066   9.761 3.30e-13 ***
## Re398:FrInf  12.7630     0.9712  13.141  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##         edf Ref.df     F p-value
## s(St) 8.435  8.788 19.87  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

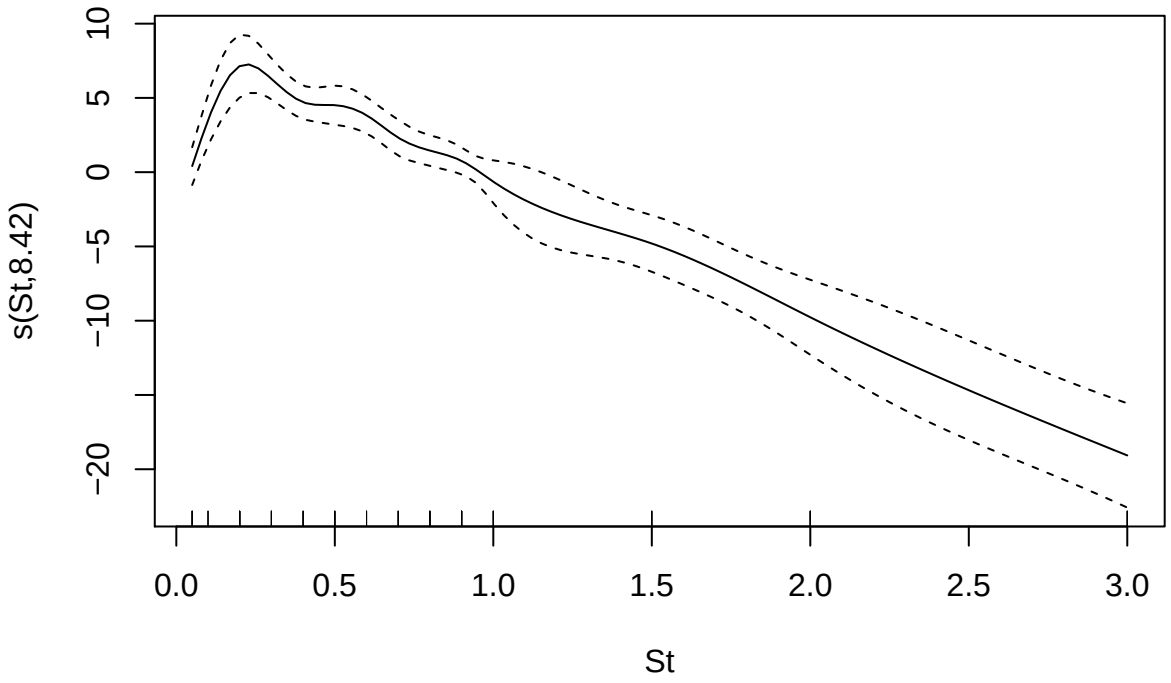
```
##
## Rank: 21/23
## R-sq.(adj) =  0.955   Deviance explained = 96.8%
## GCV =   2.13  Scale est. = 1.5127    n = 71
```

```
gam.check(gam.m3)
```




```
##
## Method: GCV   Optimizer: magic
## Smoothing parameter selection converged after 7 iterations.
## The RMS GCV score gradient at convergence was 8.319095e-05 .
## The Hessian was positive definite.
## Model rank =  21 / 23
##
## Basis dimension (k) checking results. Low p-value (k-index<1) may
## indicate that k is too low, especially if edf is close to k'.
##
##         k'  edf k-index p-value
## s(St) 9.00 8.43    1.08    0.75
```

```
gam.m4 = gam(log(R_moment_4) ~ s(St) + Re + Fr + St:Re + St:Fr + Re:Fr, data = ftraining)
plot(gam.m4)
```

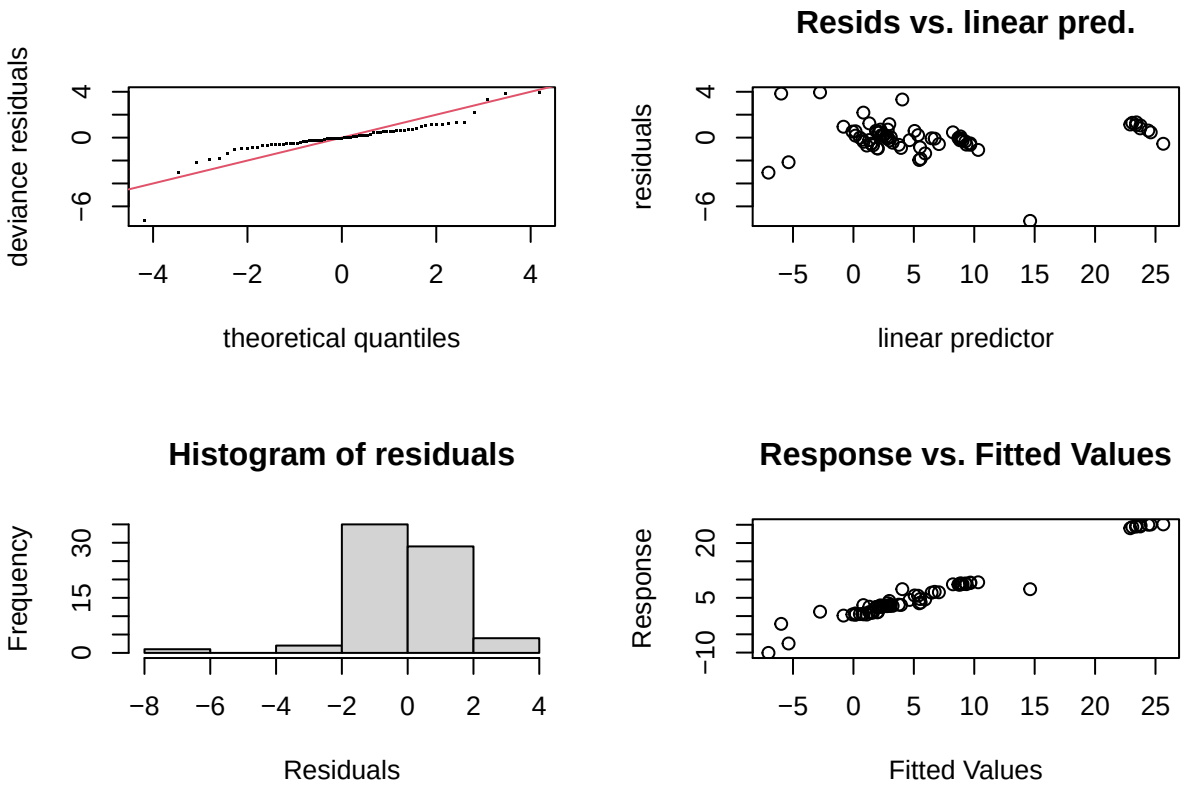



```
summary(gam.m4)
```

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## log(R_moment_4) ~ s(St) + Re + Fr + St:Re + St:Fr + Re:Fr
##
## Parametric coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  13.6971     0.9028  15.171  < 2e-16 ***
## Re224       -14.1498     0.9106 -15.538  < 2e-16 ***
## Re398       -21.8583     1.2647 -17.284  < 2e-16 ***
## Fr0.3       -19.1353     1.1913 -16.063  < 2e-16 ***
## FrInf       -17.3592     1.0342 -16.784  < 2e-16 ***
## Re90:St      10.3389     0.8511  12.148  < 2e-16 ***
## Re224:St      9.9495     0.8857  11.234 2.46e-15 ***
## Re398:St      9.0499     0.9271   9.762 3.27e-13 ***
## Fr0.3:St      0.8019     0.9343   0.858    0.395
## FrInf:St     -0.5199     0.6131  -0.848    0.400
## Re224:Fr0.3  11.6078     1.1455  10.134 9.28e-14 ***
## Re398:Fr0.3   0.0000     0.0000      NA       NA
## Re224:FrInf  11.5630     1.1172  10.350 4.49e-14 ***
## Re398:FrInf  18.7047     1.3453  13.904  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##         edf Ref.df     F p-value
## s(St) 8.416  8.782 22.52  <2e-16 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Rank: 21/23
## R-sq.(adj) =  0.954   Deviance explained = 96.7%
## GCV = 4.0851  Scale est. = 2.9023    n = 71
```

```
gam.check(gam.m4)
```




```
##
## Method: GCV   Optimizer: magic
## Smoothing parameter selection converged after 7 iterations.
## The RMS GCV score gradient at convergence was 0.0001068624 .
## The Hessian was positive definite.
## Model rank =  21 / 23
##
## Basis dimension (k) checking results. Low p-value (k-index<1) may
## indicate that k is too low, especially if edf is close to k'.
##
##         k'  edf k-index p-value
## s(St) 9.00 8.42    1.08    0.73
```

```
pred.gam2 <- predict(gam.m2, ftesting)
pred.gam3 <- predict(gam.m3, ftesting)
pred.gam4 <- predict(gam.m4, ftesting)

mse_gam2 <- mean((pred.gam2 - log(ftesting$R_moment_2))^2)
mse_gam3 <- mean((pred.gam3 - log(ftesting$R_moment_3))^2)
mse_gam4 <- mean((pred.gam4 - log(ftesting$R_moment_4))^2)

mse_gam2
```

```
## [1] 0.8611115
```

```
mse_gam3
```

```
## [1] 2.292613
```

```
mse_gam4
```

```
## [1] 4.100509
```

All predictions:

```
pred.lm1
```

```
##        24        27        30        32        35        39        42        47
## -7.903824 -8.245852 -6.050113 -2.409778 -5.815775 -5.785695 -6.078863 -2.346776
##        49        50        53        56        57        64        65        79
## -6.054432 -6.127724 -7.612019 -8.148130 -2.434209 -8.051770 -2.102746 -2.249329
##        83        85
## -2.420068 -1.784872
```

```
pred.gam2
```

```
##          24          27          30          32          35          39
## -5.17948165 -7.41314809 -5.04696725 -0.07130746 -3.14361085 -2.68974480
##          42          47          49          50          53          56
## -3.83754976 -0.43162283 -4.03771492 -6.16281375 -4.93557152 -5.22475170
##          57          64          65          79          83          85
## -0.14215862 -4.76176497  6.08351988  3.72338044 -0.93225528  1.86366209
```

```
pred.gam3
```

```
##         24         27         30         32         35         39         42
## -2.2839560 -5.6483304 -2.4335525  2.7423449 -0.2073580  0.3409741 -1.1813545
##         47         49         50         53         56         57         64
##  2.0664236 -1.5422310 -4.8924792 -2.2347144 -2.1631538  2.6502384 -1.3324056
##         65         79         83         85
## 14.6496699 10.8748850  1.3132769  5.4320688
```

```
pred.gam4
```

```
##         24         27         30         32         35         39         42
##  0.6109892 -3.7525810  0.4553173  5.6465167  2.7481803  3.3443108  1.5286265
##         47         49         50         53         56         57         64
##  4.6841043  1.0282650 -3.4013317  0.4754834  0.8783565  5.5349113  2.0377117
##         65         79         83         85
## 23.2580204 18.2346274  3.7259894  8.9135197
```

Creating a csv of predictions:

```
dataframe_all <- bind_rows(pred.lm1, pred.gam2, pred.gam3, pred.gam4)
dataframe_all
```

```
## # A tibble: 4 x 18
##     `24`  `27`   `30`    `32`   `35`   `39`  `42`   `47`  `49`  `50`   `53`
##    <dbl> <dbl>  <dbl>   <dbl>  <dbl>  <dbl> <dbl>  <dbl> <dbl> <dbl>  <dbl>
## 1 -7.90  -8.25 -6.05  -2.41   -5.82  -5.79  -6.08 -2.35  -6.05 -6.13 -7.61
## 2 -5.18  -7.41 -5.05  -0.0713 -3.14  -2.69  -3.84 -0.432 -4.04 -6.16 -4.94
## 3 -2.28  -5.65 -2.43   2.74   -0.207  0.341 -1.18  2.07  -1.54 -4.89 -2.23
## 4  0.611 -3.75  0.455  5.65    2.75   3.34   1.53  4.68   1.03 -3.40  0.475
## # ... with 7 more variables: `56` <dbl>, `57` <dbl>, `64` <dbl>, `65` <dbl>,
```

```
## #   `79` <dbl>, `83` <dbl>, `85` <dbl>
```

```
write.csv(dataframe_all,"data/predictions.csv", row.names = FALSE)
```