Case Study

Abdel Shehata

2022-10-26

Introduction and Data

Introduction

Data Introduction

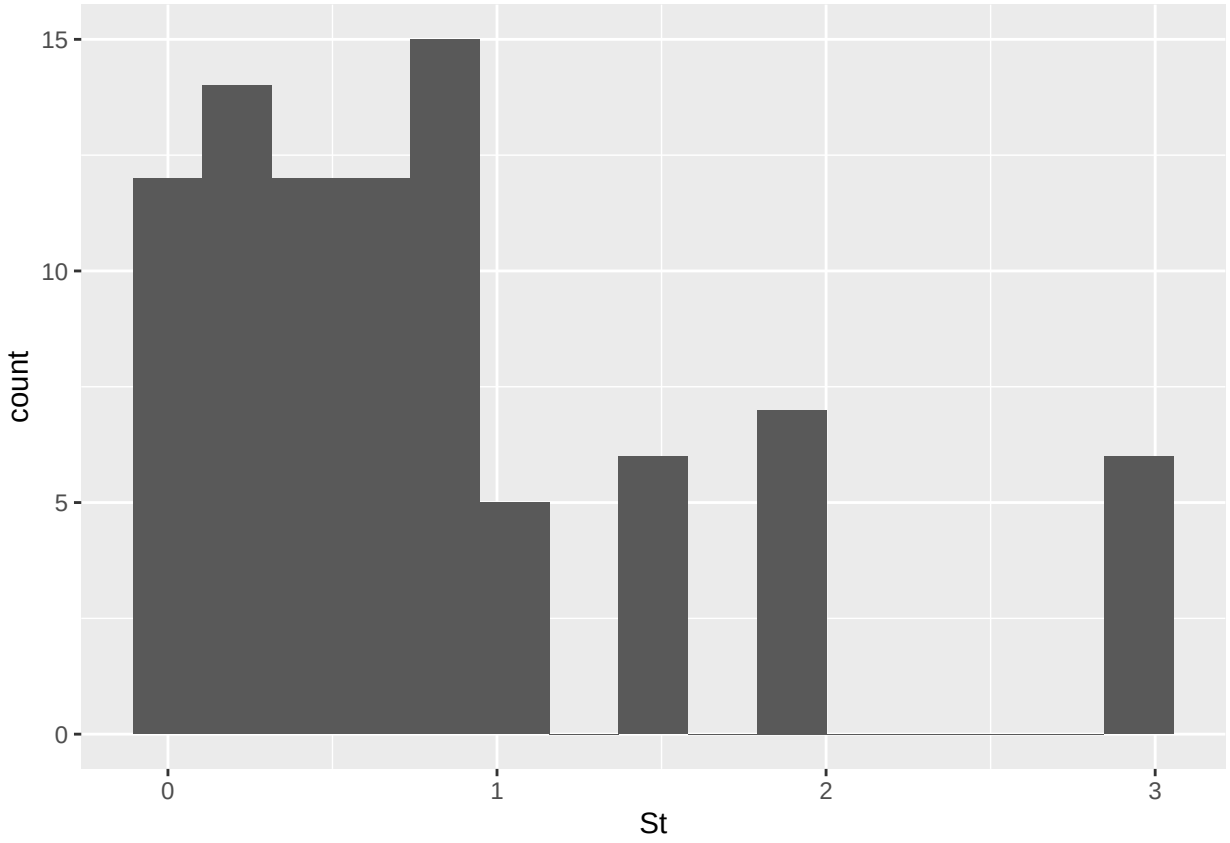
Exploratory Data Analysis

```
library(readr)
data_train <- read_csv("data-train.csv")
```

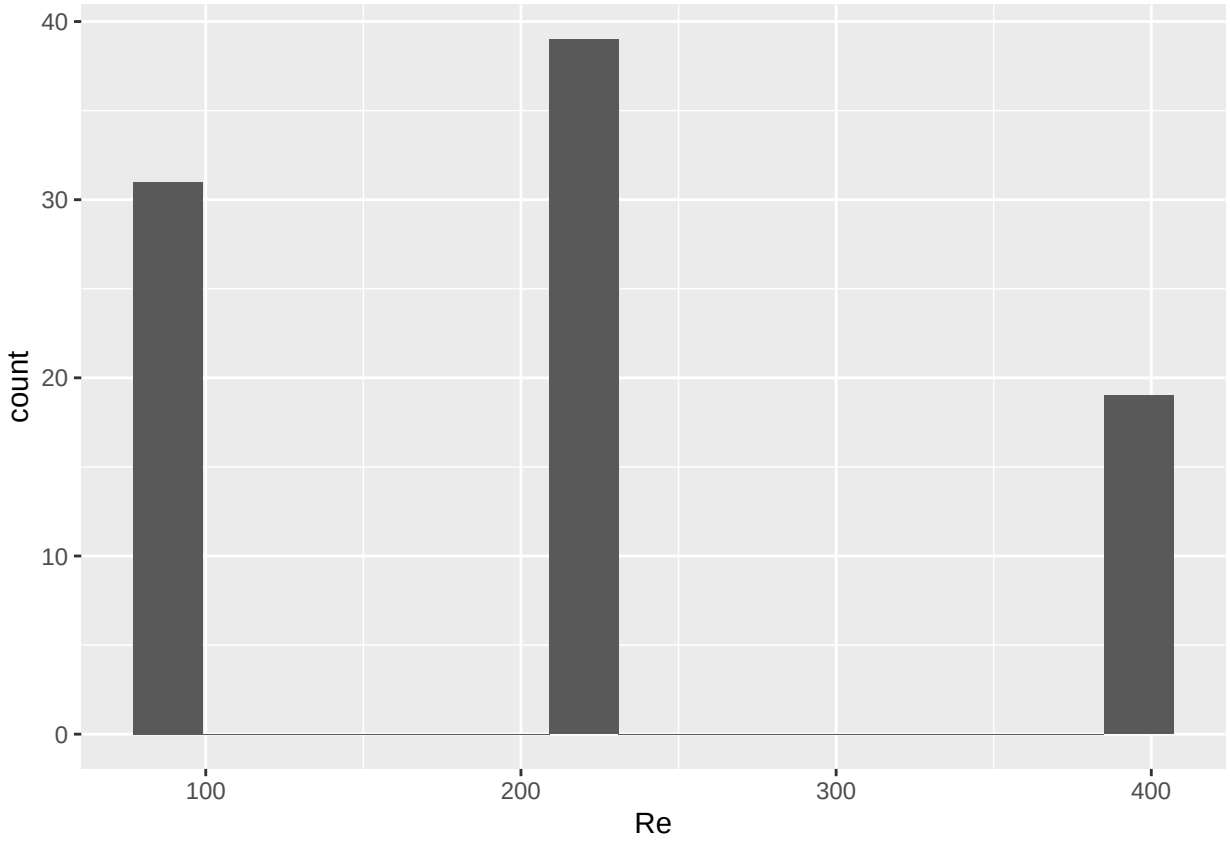
```
## Rows: 89 Columns: 7
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## dbl (7): St, Re, Fr, R_moment_1, R_moment_2, R_moment_3, R_moment_4
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
attach(data_train)
data_train <- data_train%>% mutate(TFr = case_when(Fr>1~ .99999, Fr<1~Fr))
data_train<-data_train%>%mutate(TFr=logit(TFr))

ggplot(data_train) +
  geom_histogram(aes(x = St), bins = 15)
```

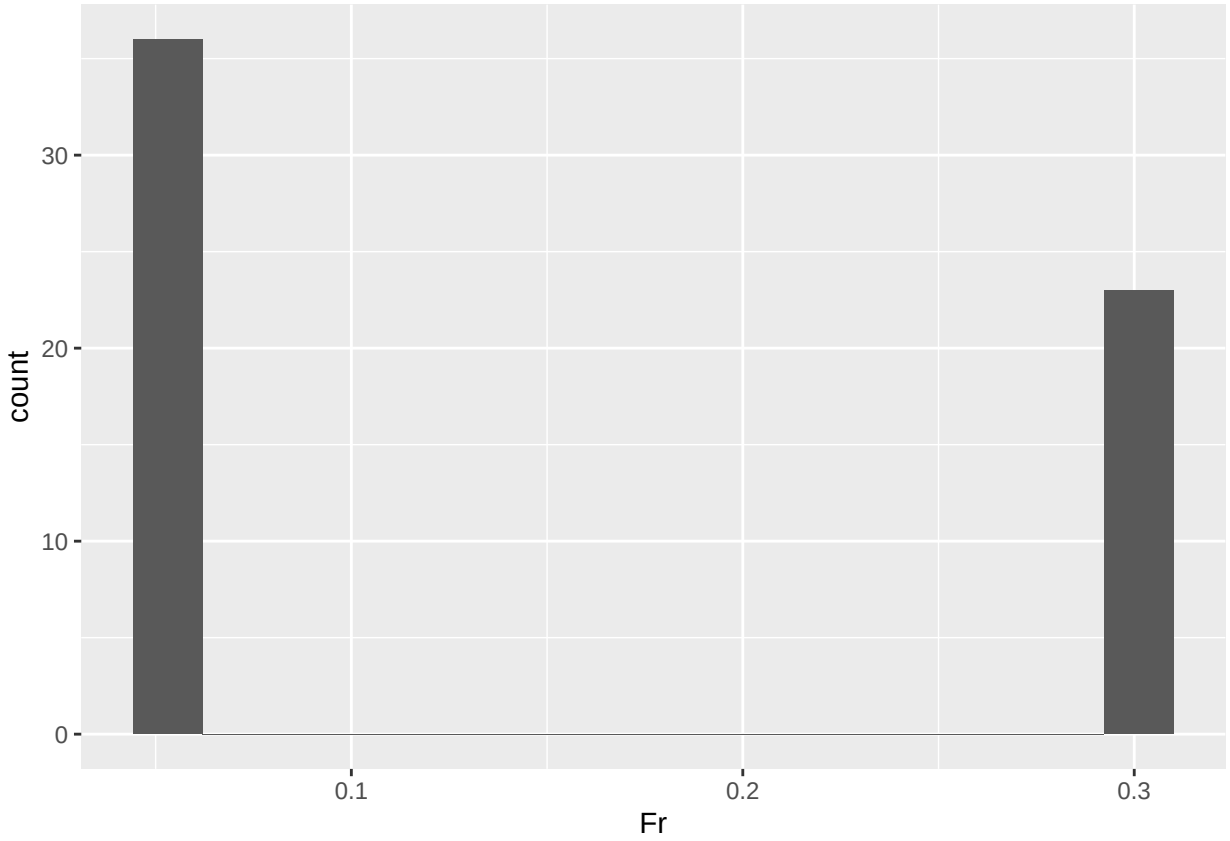



```
ggplot(data_train) +
  geom_histogram(aes(x = Re), bins = 15)
```

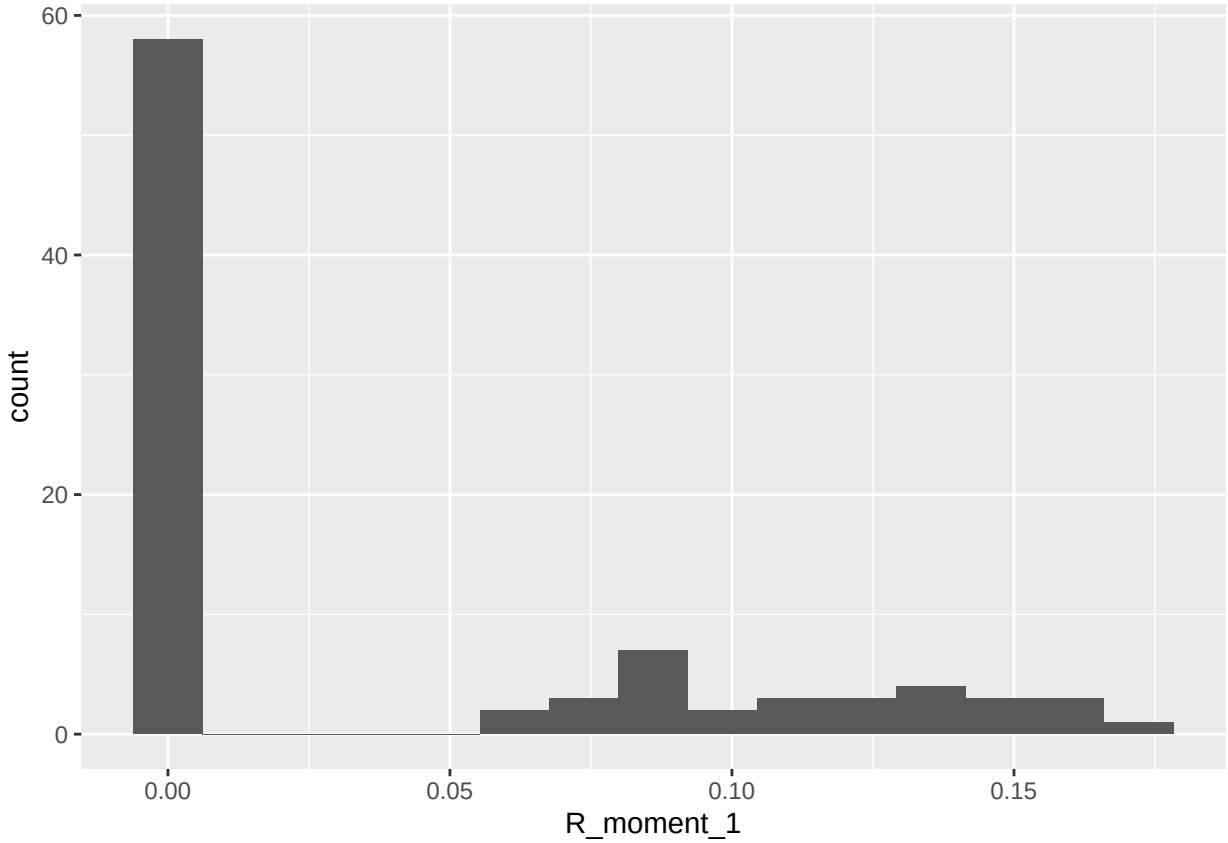



```
ggplot(data_train) +
  geom_histogram(aes(x = Fr), bins = 15)
```

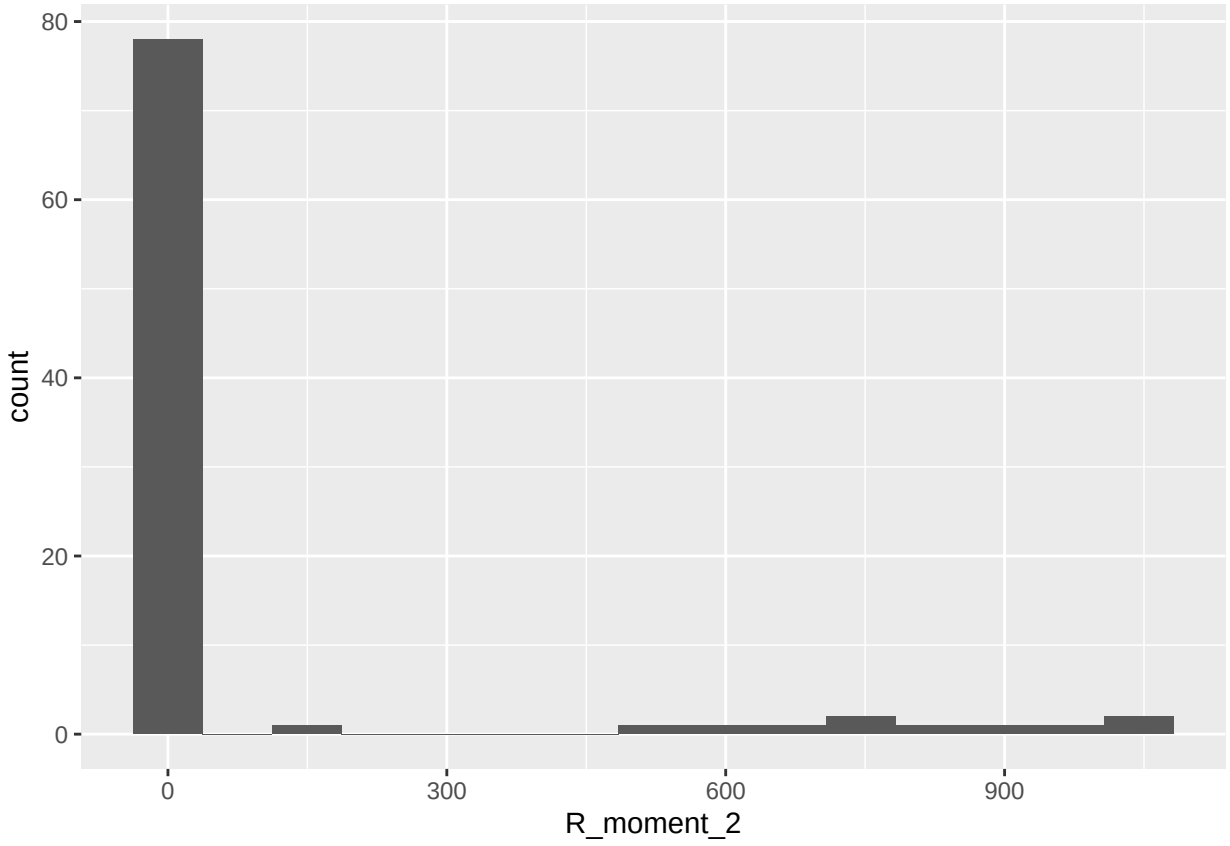
```
## Warning: Removed 30 rows containing non-finite values (stat_bin).
```



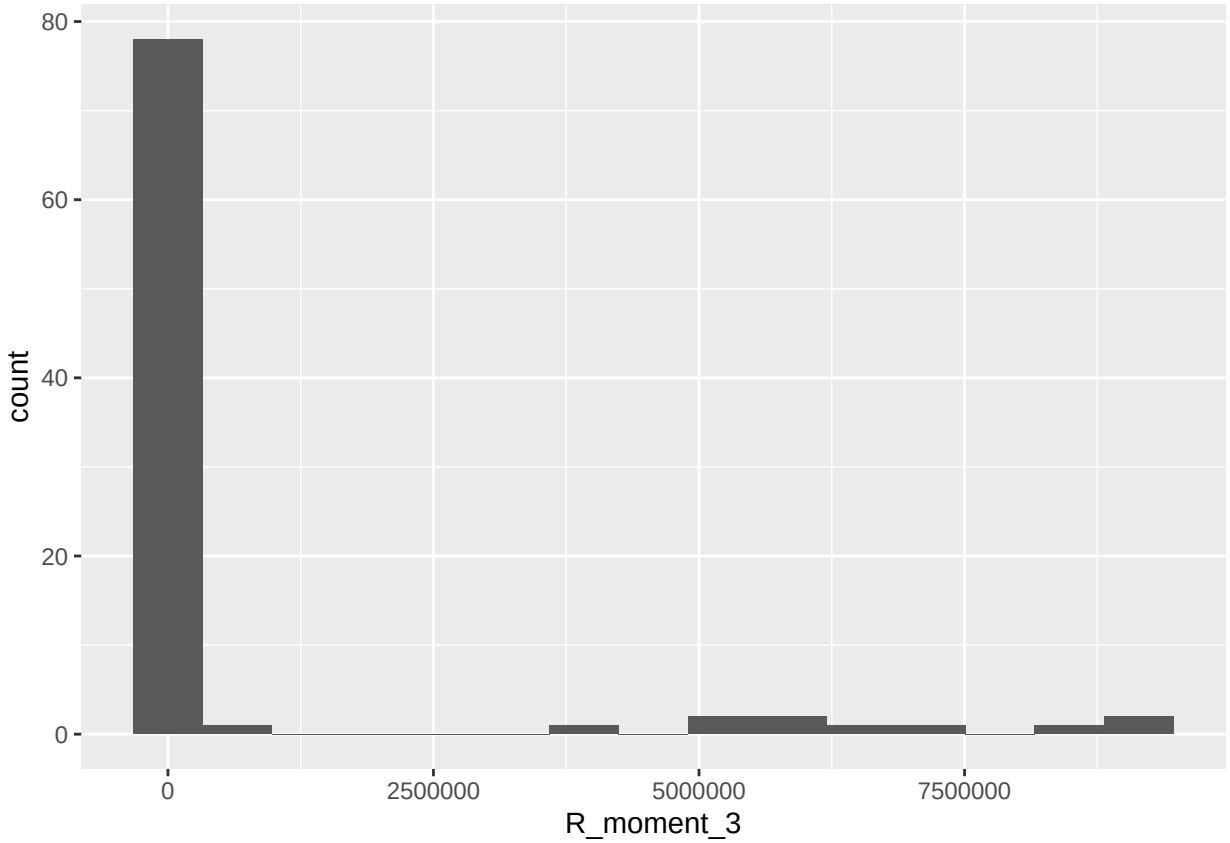

```
ggplot(data_train) +
  geom_histogram(aes(x = R_moment_1), bins = 15)
```



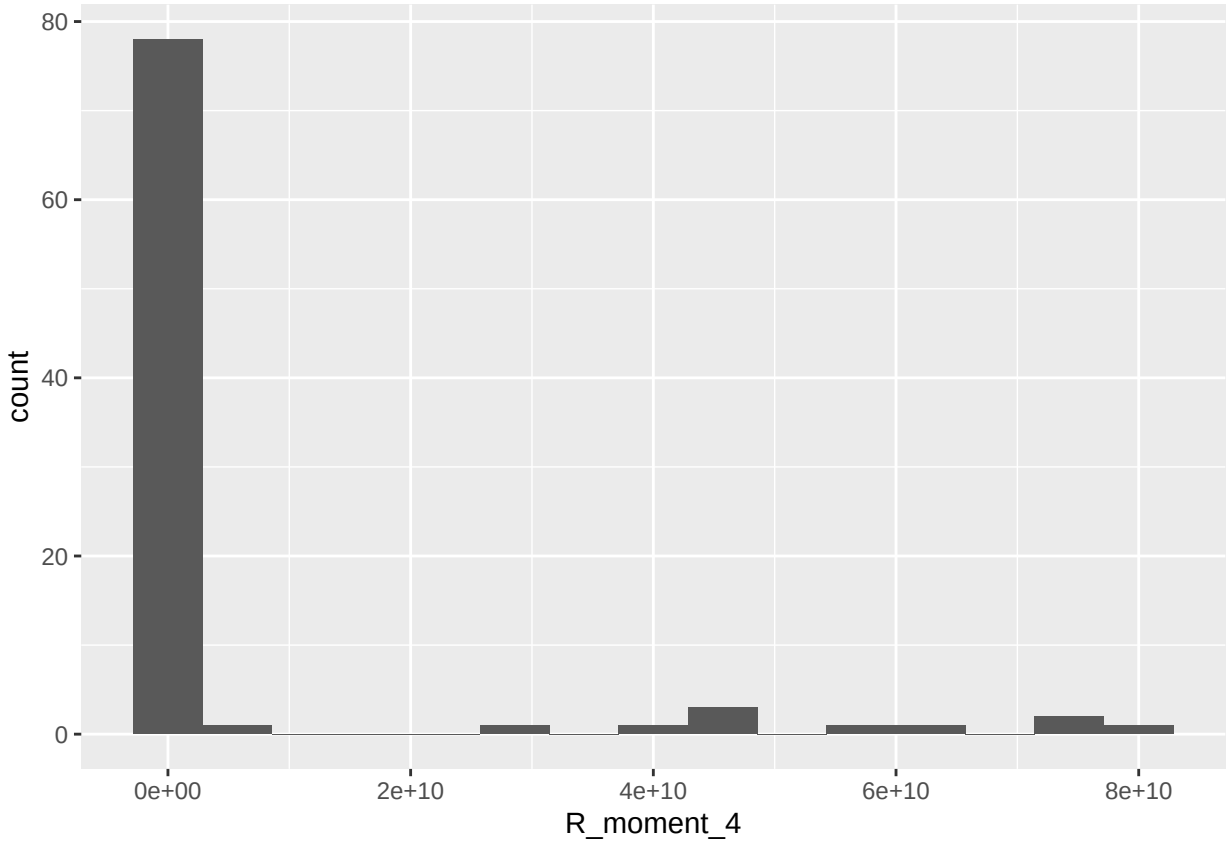

```
ggplot(data_train) +
  geom_histogram(aes(x = R_moment_2), bins = 15)
```




```
ggplot(data_train) +
  geom_histogram(aes(x = R_moment_3), bins = 15)
```




```
ggplot(data_train) +
  geom_histogram(aes(x = R_moment_4), bins = 15)
```



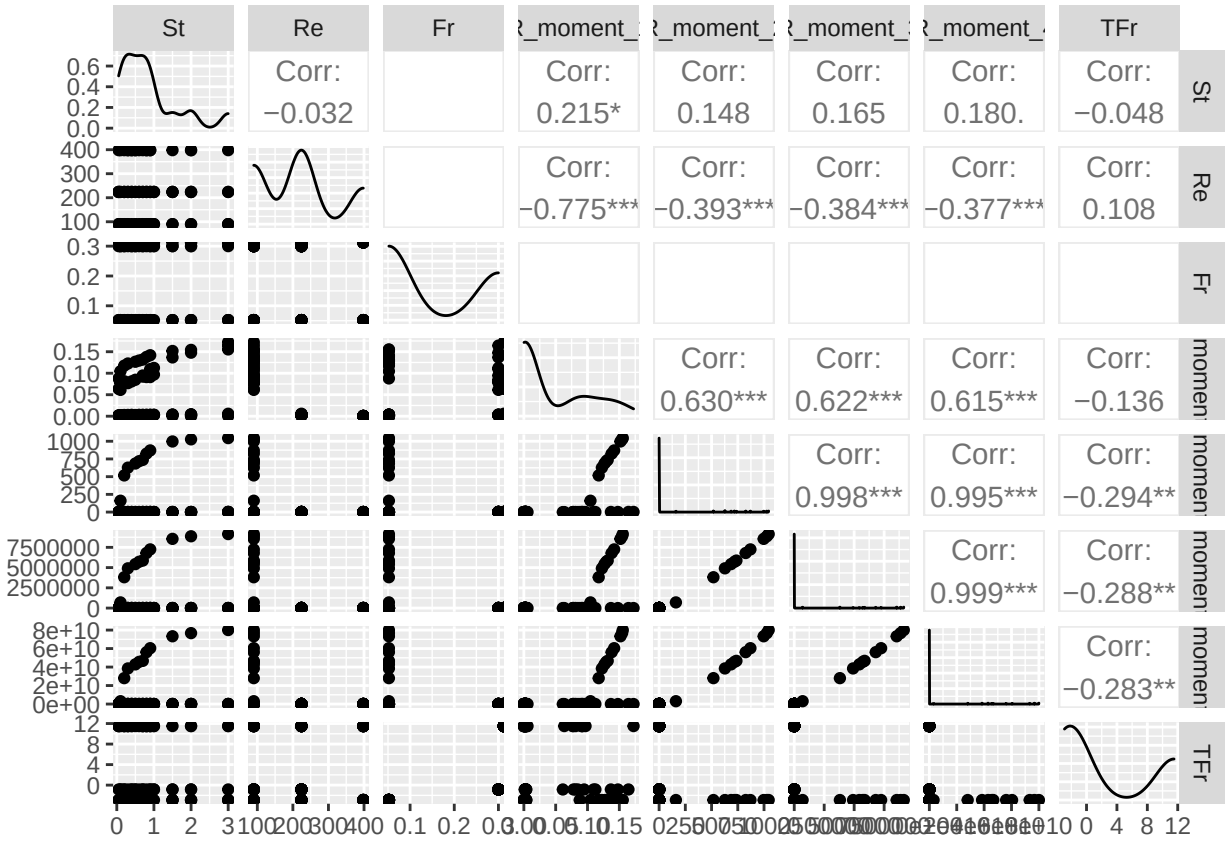

```
ggpairs(data_train)
```

```
## Warning: Removed 1 rows containing missing values (geom_text).

## Warning: Removed 1 rows containing missing values (geom_text).

## Warning: Removed 30 rows containing non-finite values (stat_density).

## Warning: Removed 1 rows containing missing values (geom_text).
## Removed 1 rows containing missing values (geom_text).
## Removed 1 rows containing missing values (geom_text).
## Removed 1 rows containing missing values (geom_text).
## Removed 1 rows containing missing values (geom_text).
```



Some brief notes:

Observations on predictors: St (size) seems to be mostly small particles with some trials with larger particles. Re (turbulence) seems to be in three groups: low (90), medium (224), and high (398). Perhaps it could be considered a categorical variable? Fr (gravitational acceleration) seems to also be in three groups: low (.052), medium (.3), and high (infinite). Could this also become a categorical variable? We have decided to do a logistic transformation on Fr in order to approximate the effects of infinity.

Also, I believe we should centralize the second through fourth moments. The first raw moment is actually helpful because it tells us about the average amount of turbulence. However, when it comes to the shape of the distribution (variance, skewness, and kurtosis) we need to centralize the moments in order to interpret them.

The code for transforming the variables is below:

```
data_train <- data_train %>% mutate(R_moment_1_central = 0)
data_train <- data_train %>% mutate(R_moment_2_central = R_moment_2 - (R_moment_1)^2)
data_train <- data_train %>% mutate(R_moment_3_central = R_moment_3 - 3*R_moment_1*R_moment_2 + 2*(R_mo
data_train <- data_train %>% mutate(R_moment_4_central = R_moment_4 - 4*R_moment_1*R_moment_3 + 6*((R_m
```

Correlations: Reynolds number is negatively correlated with all moments, which is surprising but I believe it is due to the fact that almost all of the observations of all the moments are mostly around 0 with only some exceptions. The 2nd, 3rd, and 4th moments are pretty correlated but this makes sense because they are all various measures of the width and shape of the tails.

Plots: St and the first moment seem to have a linear or quadratic relationship. I would not be surprised if it is true that bigger particles cluster more on average. St and the second, third, and fourth moments seem to have a linear or quadratic relationship. Perhaps bigger particles behave more unpredictably.

Methodology

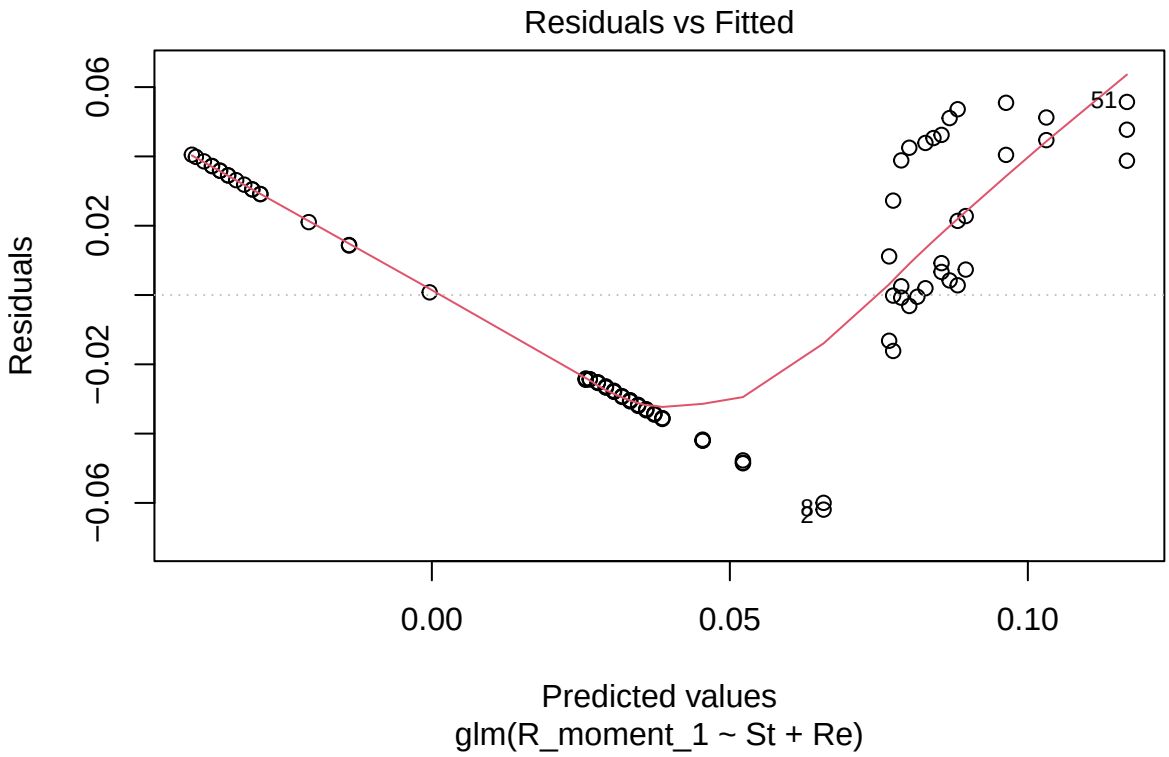
Linear

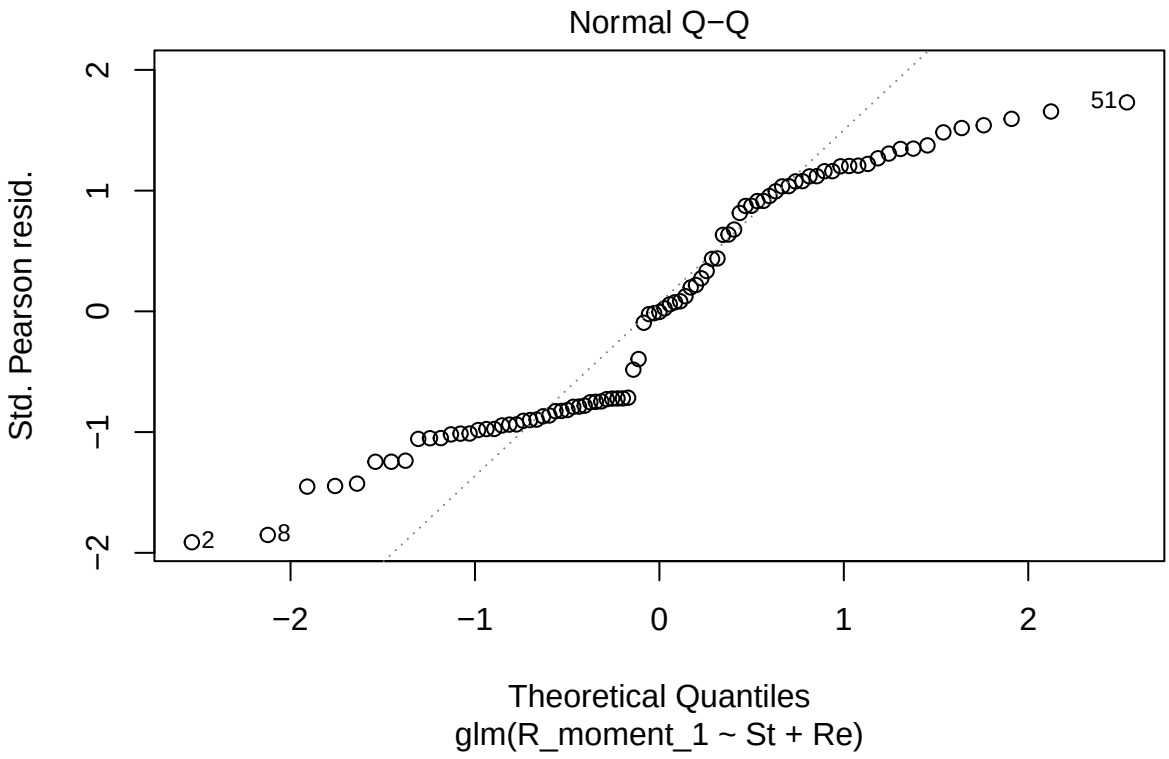
```
full_linear_E1 <- glm(R_moment_1 ~ St + TFr + Re, data = data_train)
step_full_linear_E1 <- stepAIC(full_linear_E1, direction = "both", trace = FALSE)
summary(step_full_linear_E1)
```

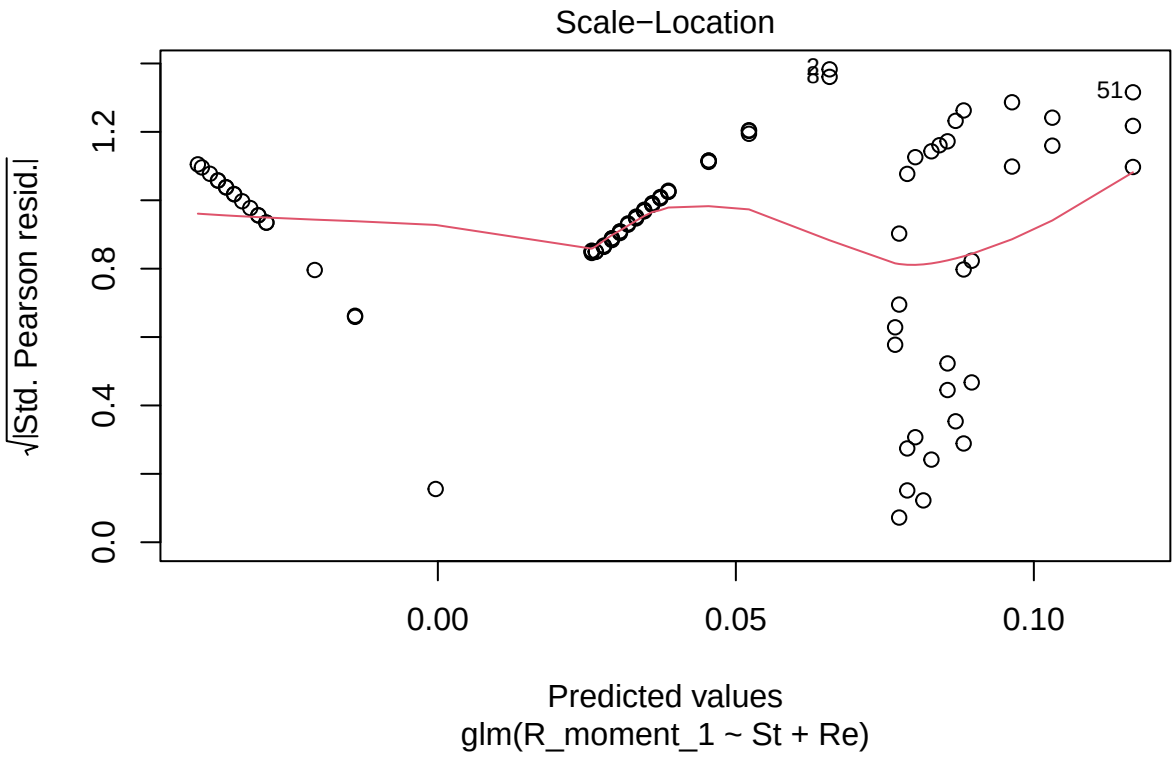
Linear Fitting

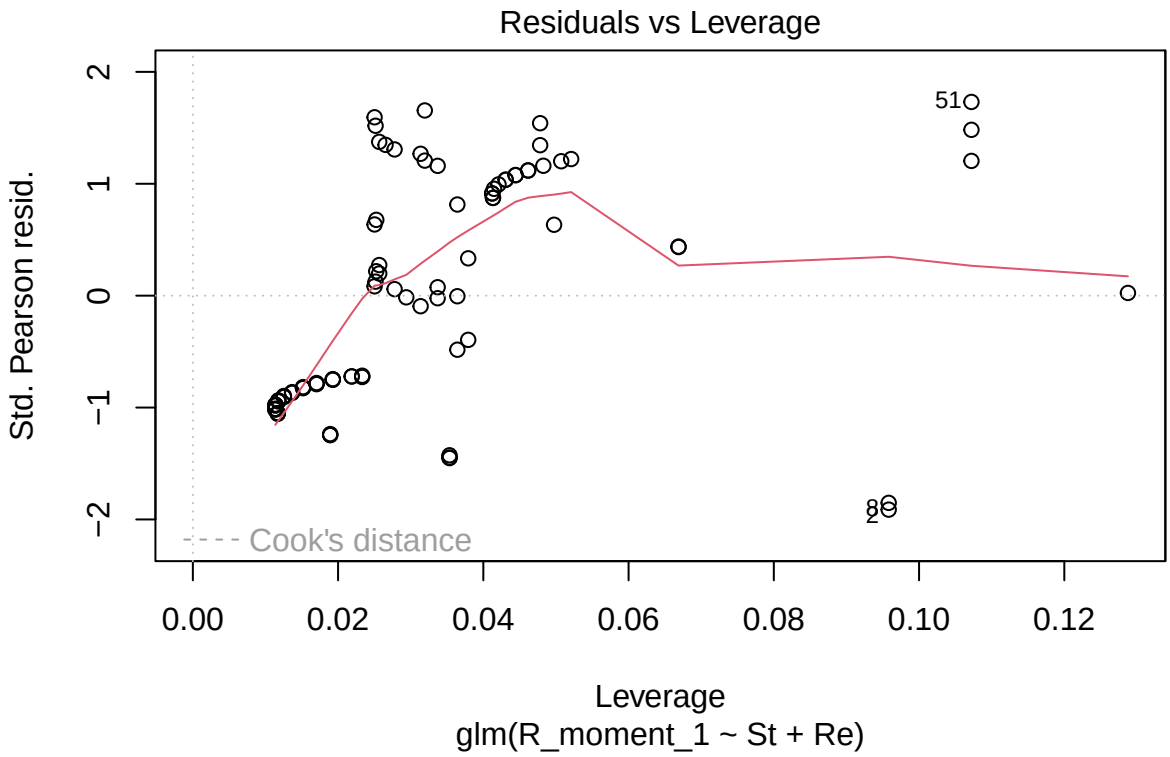
```
##
## Call:
## glm(formula = R_moment_1 ~ St + Re, data = data_train)
##
## Deviance Residuals:
##       Min         1Q     Median         3Q        Max
## -0.061936  -0.030347  -0.000174   0.034491   0.055714
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.102e-01  8.837e-03  12.475  < 2e-16 ***
## St           1.353e-02  4.621e-03   2.927  0.00438 **
## Re          -3.798e-04  3.215e-05 -11.816  < 2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for gaussian family taken to be 0.001160225)
##
##     Null deviance: 0.274427  on 88  degrees of freedom
## Residual deviance: 0.099779  on 86  degrees of freedom
## AIC: -344.04
##
## Number of Fisher Scoring iterations: 2
```

```
plot(step_full_linear_E1)
```


Residuals vs Fitted
Residuals
0.06
0.02
−0.02
−0.06
0.00
0.05
0.10
51
8
2
Predicted values
glm(R_moment_1 ~ St + Re)


Normal Q−Q
Std. Pearson resid.
Theoretical Quantiles
glm(R_moment_1 ~ St + Re)
2
8
51
−2
−1
0
1
2


Scale−Location
√|Std. Pearson resid.|
0.0
0.4
0.8
1.2
0.00
0.05
0.10
Predicted values
glm(R_moment_1 ~ St + Re)
30
51

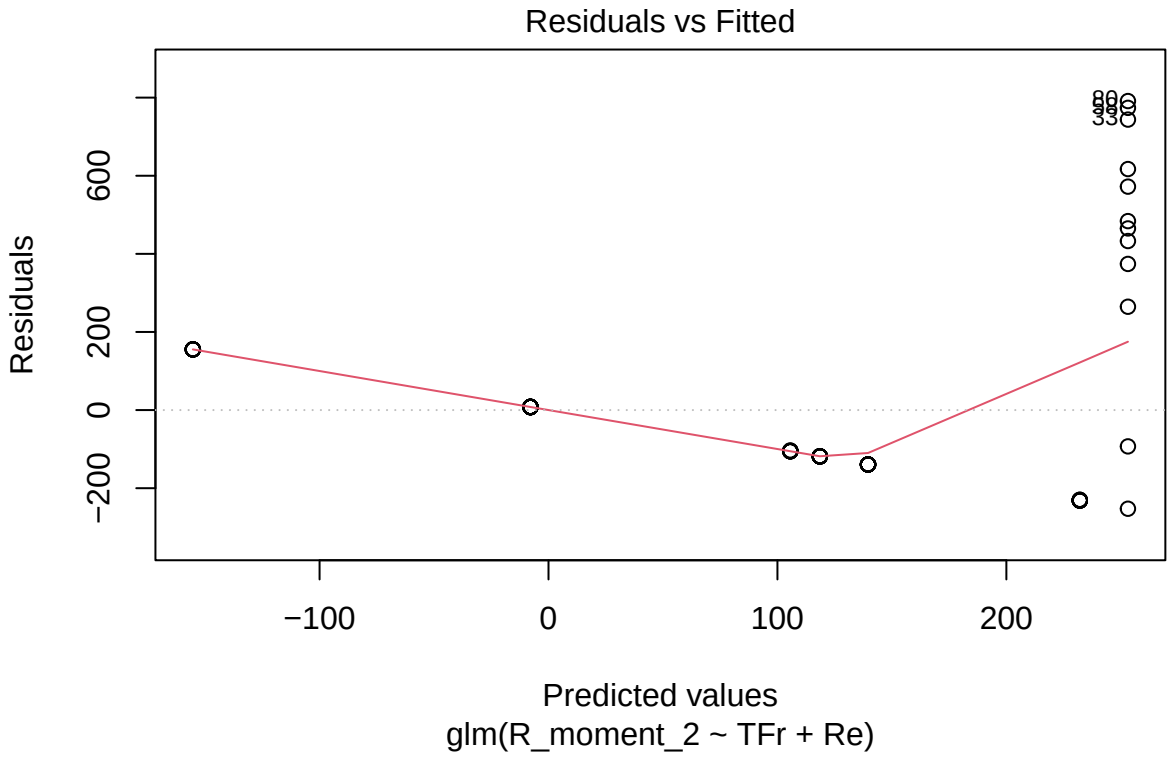


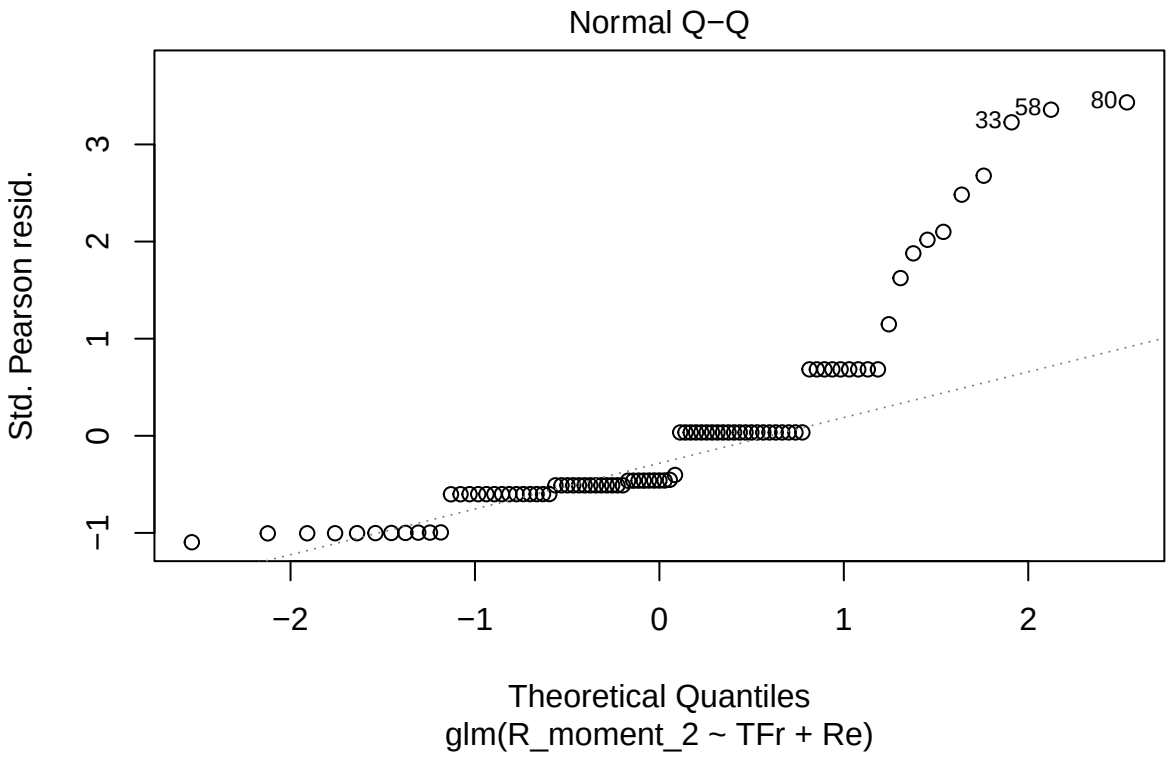

```
full_linear_E2 <- glm(R_moment_2 ~ St + TFr + Re, data = data_train)
step_full_linear_E2 <- stepAIC(full_linear_E2, direction = "both", trace = FALSE)
summary(step_full_linear_E2)
```

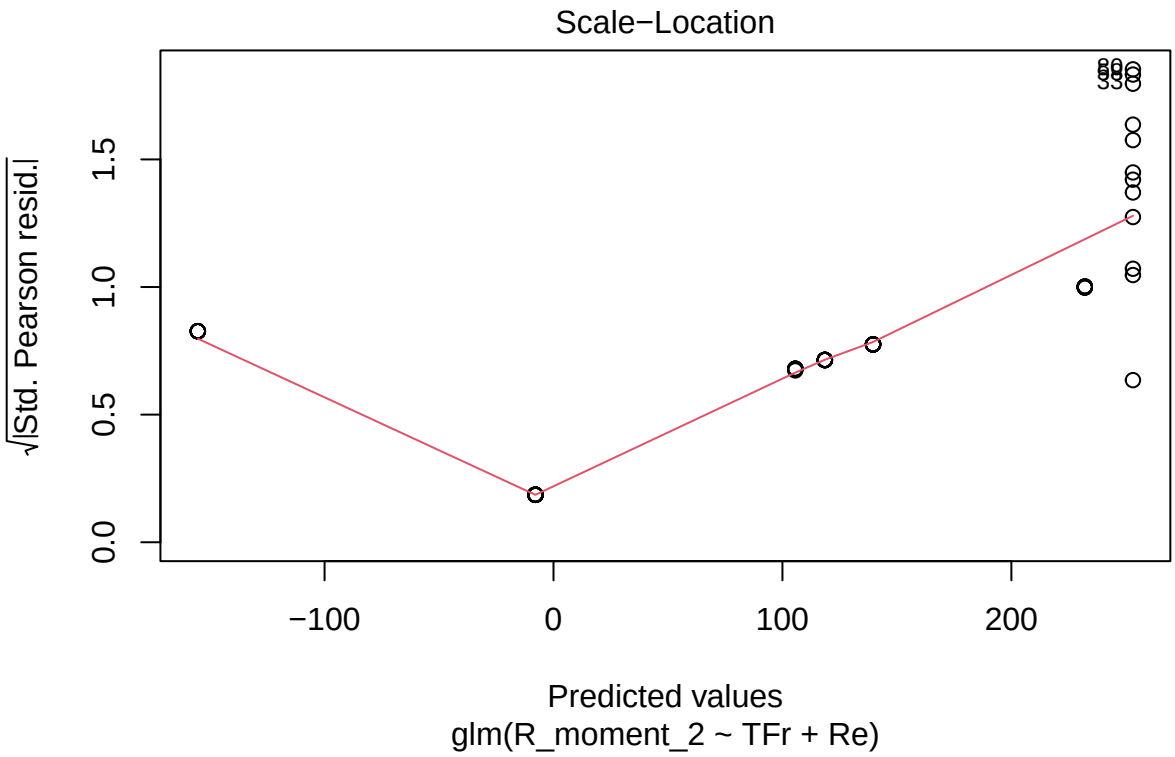
```
##
## Call:
## glm(formula = R_moment_2 ~ TFr + Re, data = data_train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -252.57  -139.17  -104.99     7.97   791.19
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 299.6593    53.6457   5.586 2.67e-07 ***
## TFr         -10.2317     3.8471  -2.660 0.009332 **
## Re           -0.8473     0.2221  -3.815 0.000256 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 54790.94)
##
##     Null deviance: 6032373  on 88  degrees of freedom
## Residual deviance: 4712021  on 86  degrees of freedom
## AIC: 1228.6
```

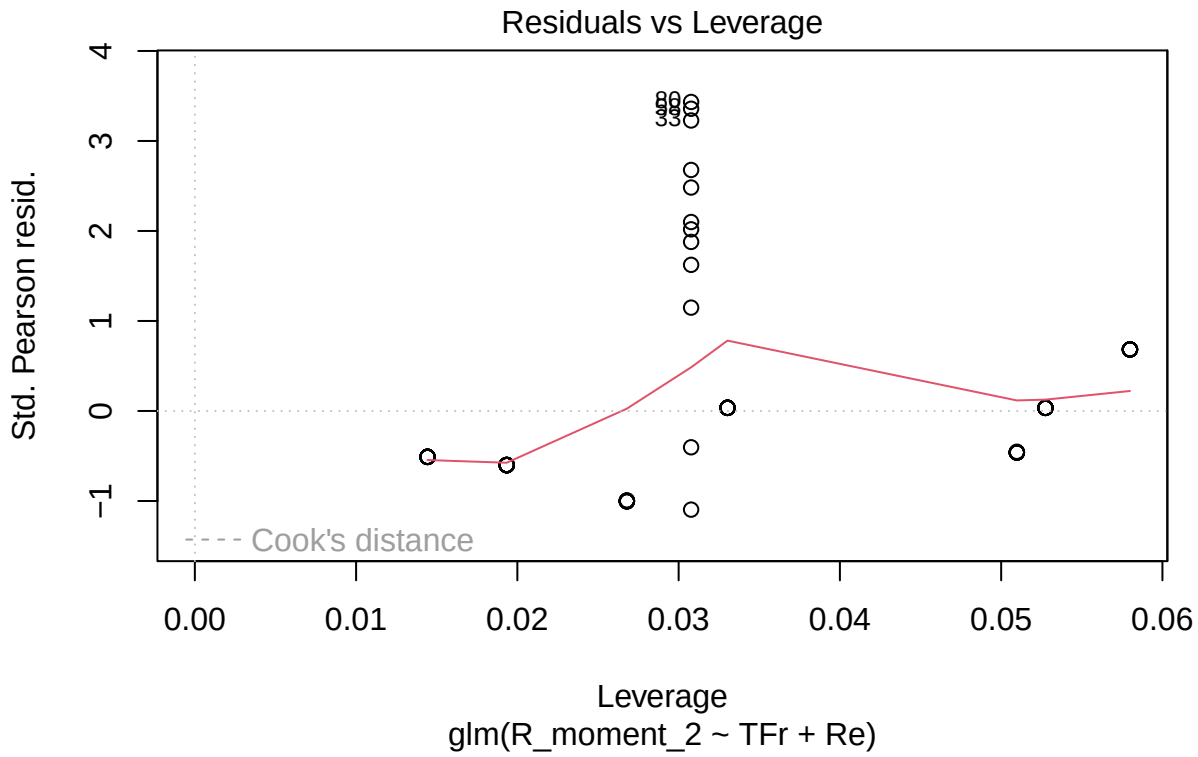
```
##
## Number of Fisher Scoring iterations: 2
```

```
plot(step_full_linear_E2)
```




Normal Q−Q
Std. Pearson resid.
−1
0
1
2
3
33
58
80
−2
−1
0
1
2
Theoretical Quantiles
glm(R_moment_2 ~ TFr + Re)


Scale−Location
√|Std. Pearson resid.|
0.0
0.5
1.0
1.5
−100
0
100
200
Predicted values
glm(R_moment_2 ~ TFr + Re)

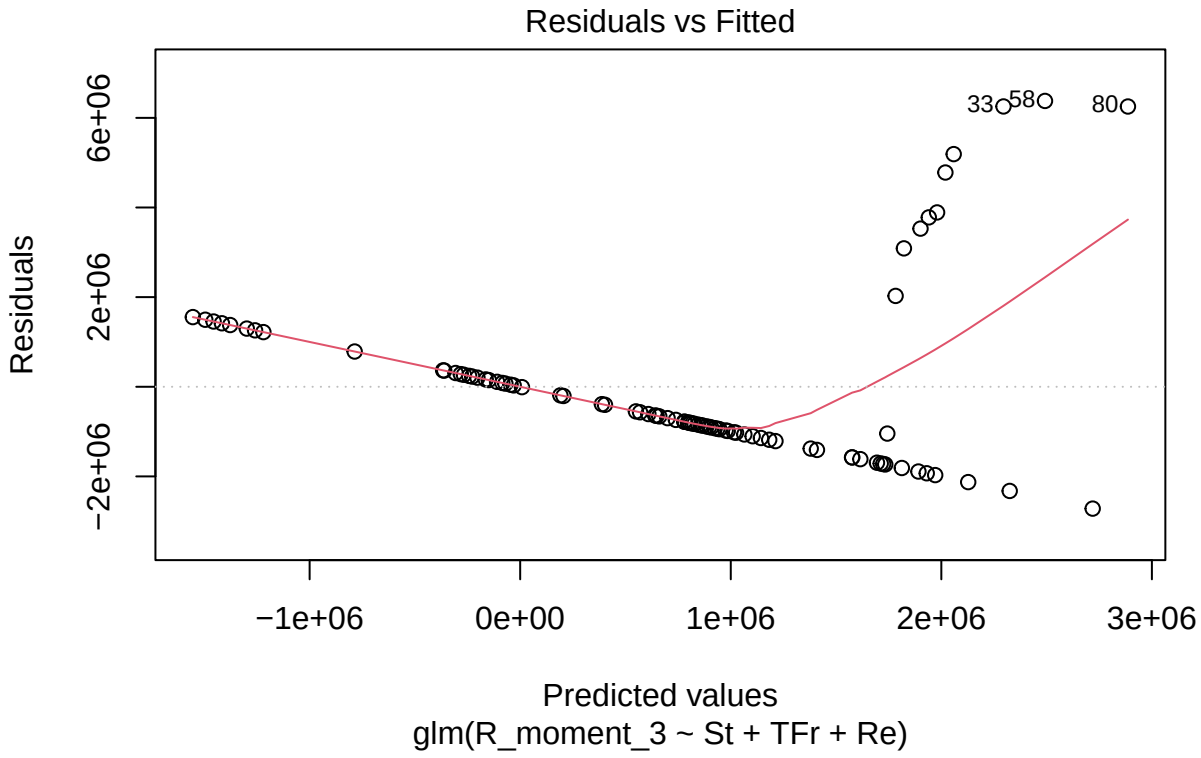


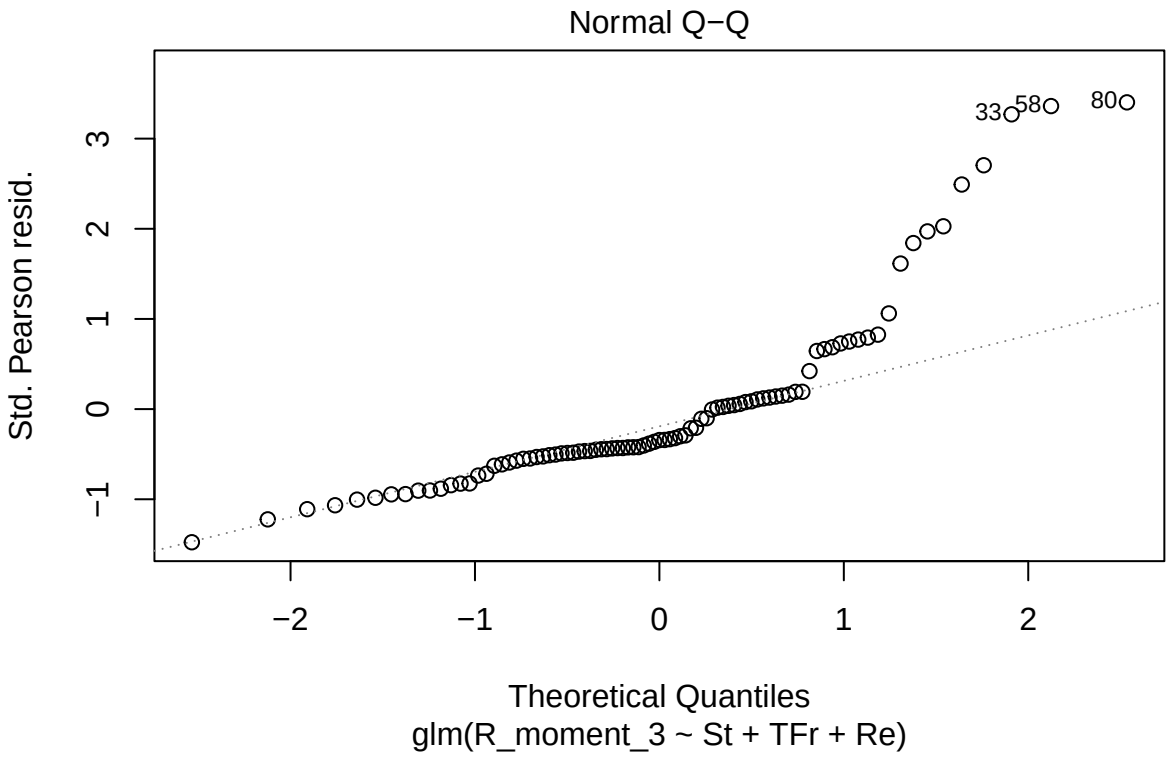

```
full_linear_E3 <- glm(R_moment_3 ~ St + TFr + Re, data = data_train)
step_full_linear_E3 <- stepAIC(full_linear_E3, direction = "both", trace = FALSE)
summary(step_full_linear_E3)
```

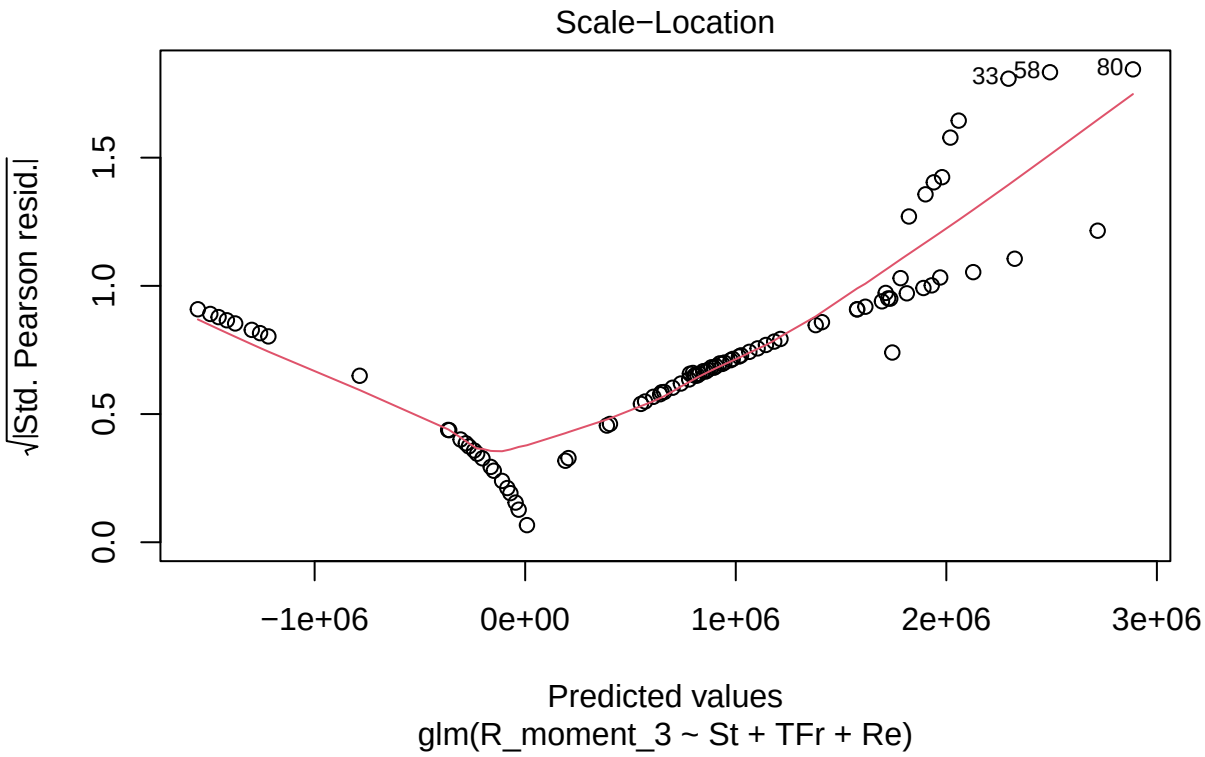
```
##
## Call:
## glm(formula = R_moment_3 ~ St + TFr + Re, data = data_train)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2718782  -1025004   -660570    281888   6377805
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2082451     508059   4.099 9.46e-05 ***
## St            394073     264795   1.488 0.140394
## TFr           -81448      32077  -2.539 0.012933 *
## Re             -6832       1851  -3.691 0.000393 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 3.801448e+12)
##
##     Null deviance: 4.1938e+14  on 88  degrees of freedom
## Residual deviance: 3.2312e+14  on 85  degrees of freedom
```

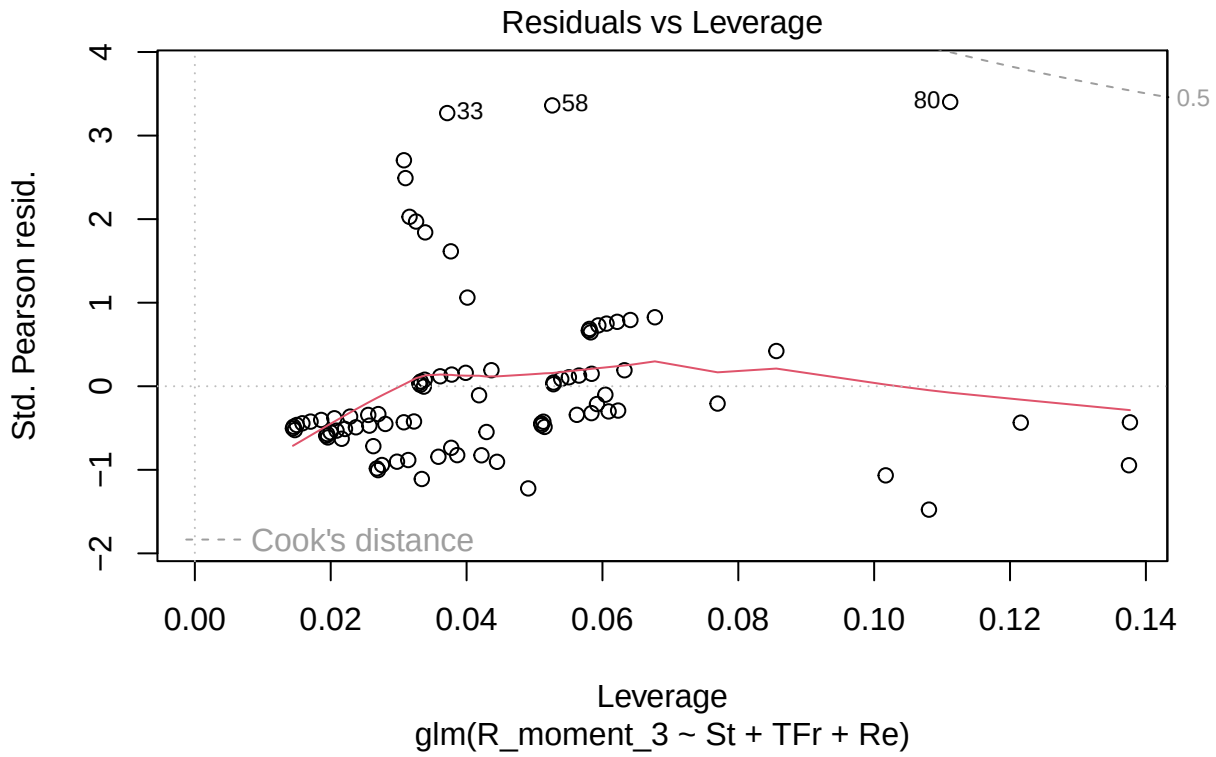
```
## AIC: 2836.5
##
## Number of Fisher Scoring iterations: 2
```

```
plot(step_full_linear_E3)
```




Normal Q−Q
Std. Pearson resid.
−1
0
1
2
3
33
58
80
−2
−1
0
1
2
Theoretical Quantiles
glm(R_moment_3 ~ St + TFr + Re)


Scale−Location
√|Std. Pearson resid.|
0.0
0.5
1.0
1.5
33
58
80
−1e+06
0e+00
1e+06
2e+06
3e+06
Predicted values
glm(R_moment_3 ~ St + TFr + Re)

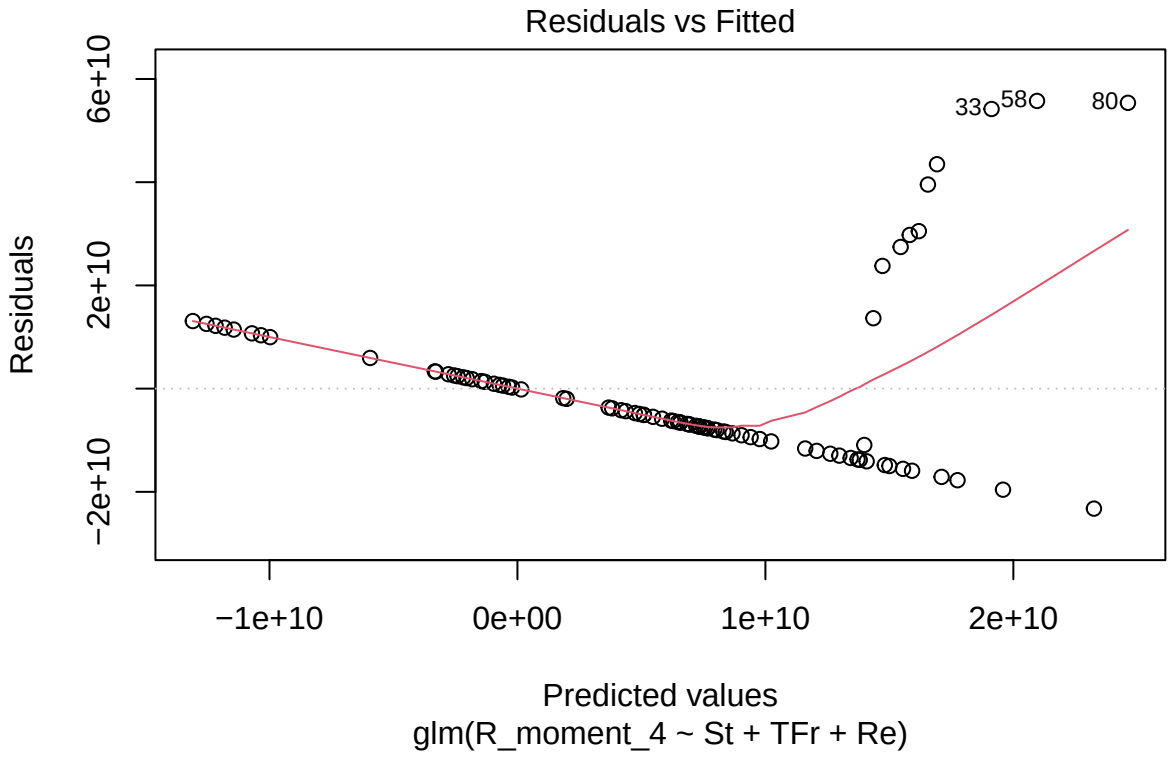


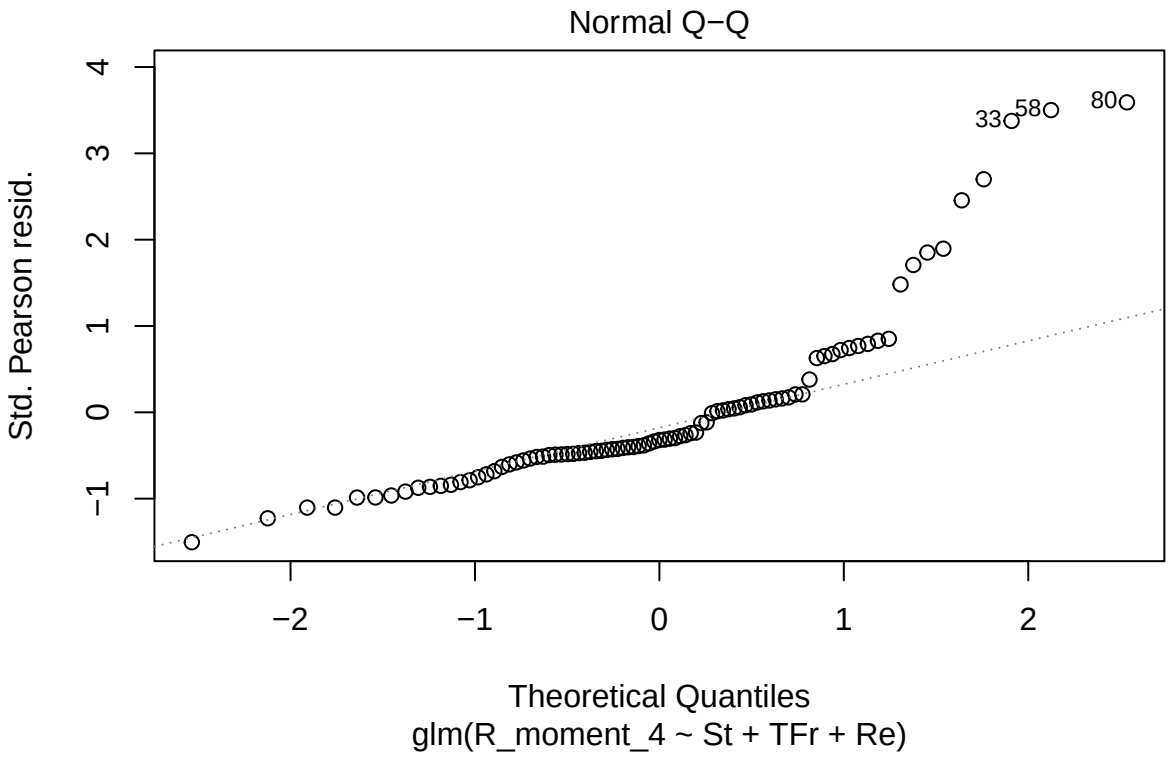

```
full_linear_E4 <- glm(R_moment_4 ~ St + TFr + Re, data = data_train)
step_full_linear_E4 <- stepAIC(full_linear_E4, direction = "both", trace = FALSE)
summary(step_full_linear_E4)
```

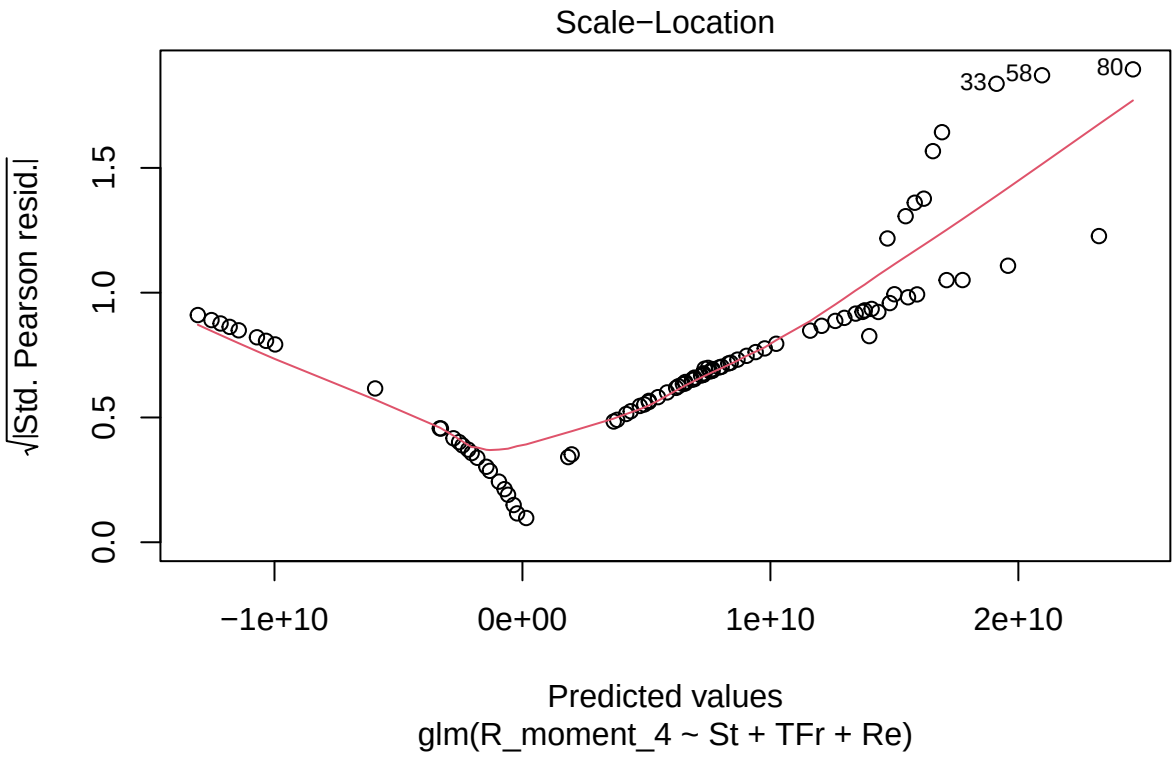
```
##
## Call:
## glm(formula = R_moment_4 ~ St + TFr + Re, data = data_train)
##
## Deviance Residuals:
##        Min          1Q      Median          3Q         Max
## -2.325e+10  -8.400e+09  -5.101e+09   2.554e+09   5.575e+10
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.673e+10  4.263e+09   3.925 0.000175 ***
## St           3.667e+09  2.222e+09   1.651 0.102520
## TFr         -6.673e+08  2.691e+08  -2.480 0.015126 *
## Re          -5.609e+07  1.553e+07  -3.612 0.000513 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 2.676008e+20)
##
##     Null deviance: 2.9413e+22  on 88  degrees of freedom
## Residual deviance: 2.2746e+22  on 85  degrees of freedom
```

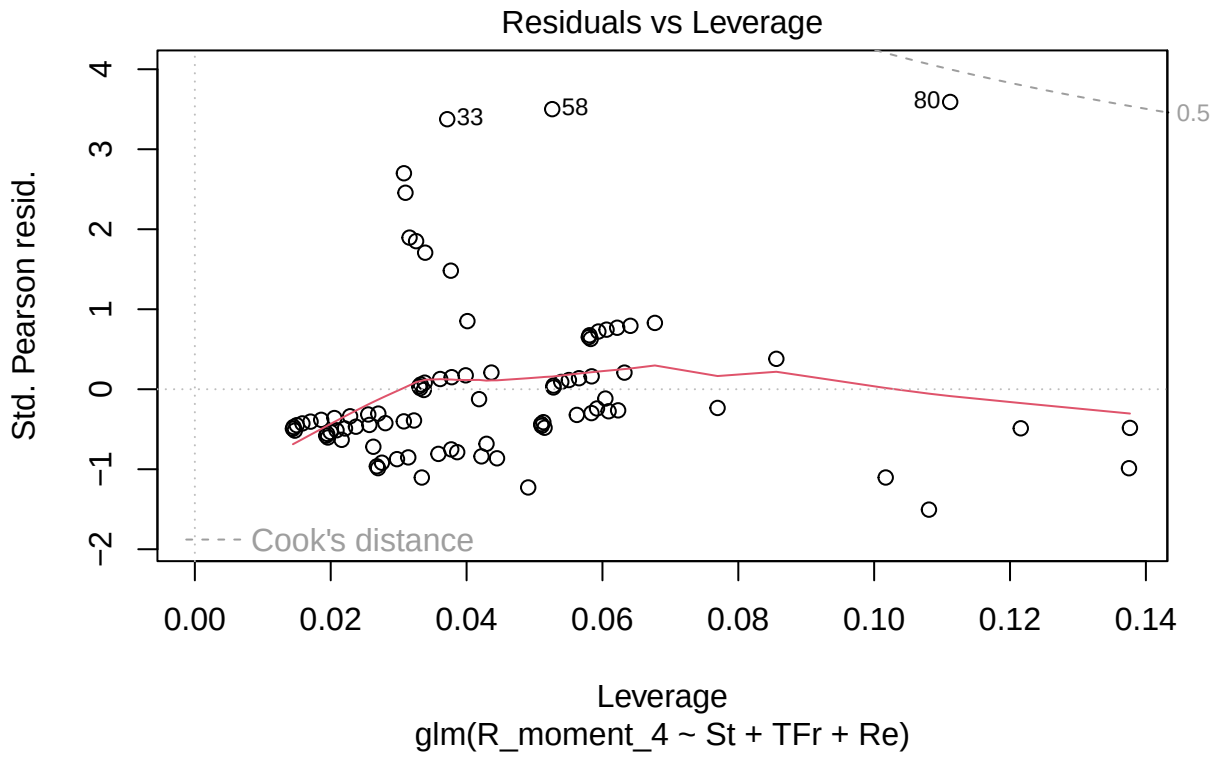
```
## AIC: 4444.7
##
## Number of Fisher Scoring iterations: 2
```

```
plot(step_full_linear_E4)
```




Normal Q−Q
Std. Pearson resid.
Theoretical Quantiles
glm(R_moment_4 ~ St + TFr + Re)
33
58
80


Scale−Location
√|Std. Pearson resid.|
0.0
0.5
1.0
1.5
33
58
80
−1e+10
0e+00
1e+10
2e+10
Predicted values
glm(R_moment_4 ~ St + TFr + Re)



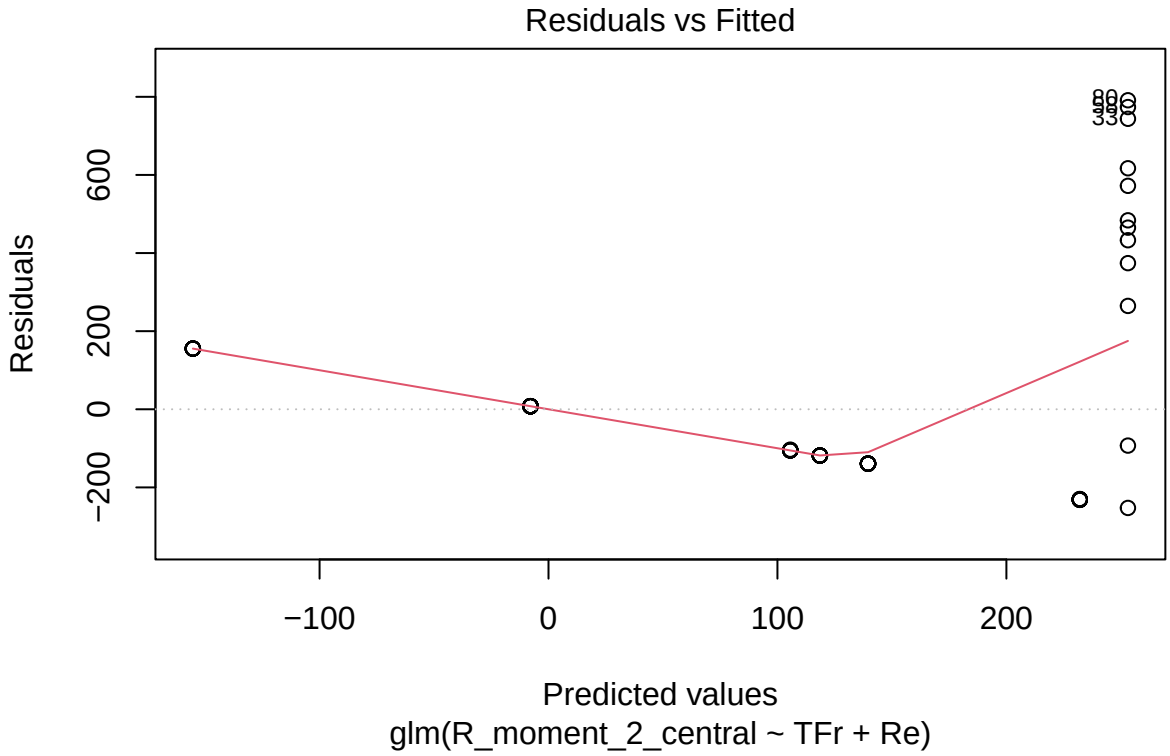

```
full_linear_E2_central <- glm(R_moment_2_central ~ St + TFr + Re, data = data_train)
step_full_linear_E2_central <- stepAIC(full_linear_E2_central, direction = "both", trace = FALSE)
summary(step_full_linear_E2_central)
```

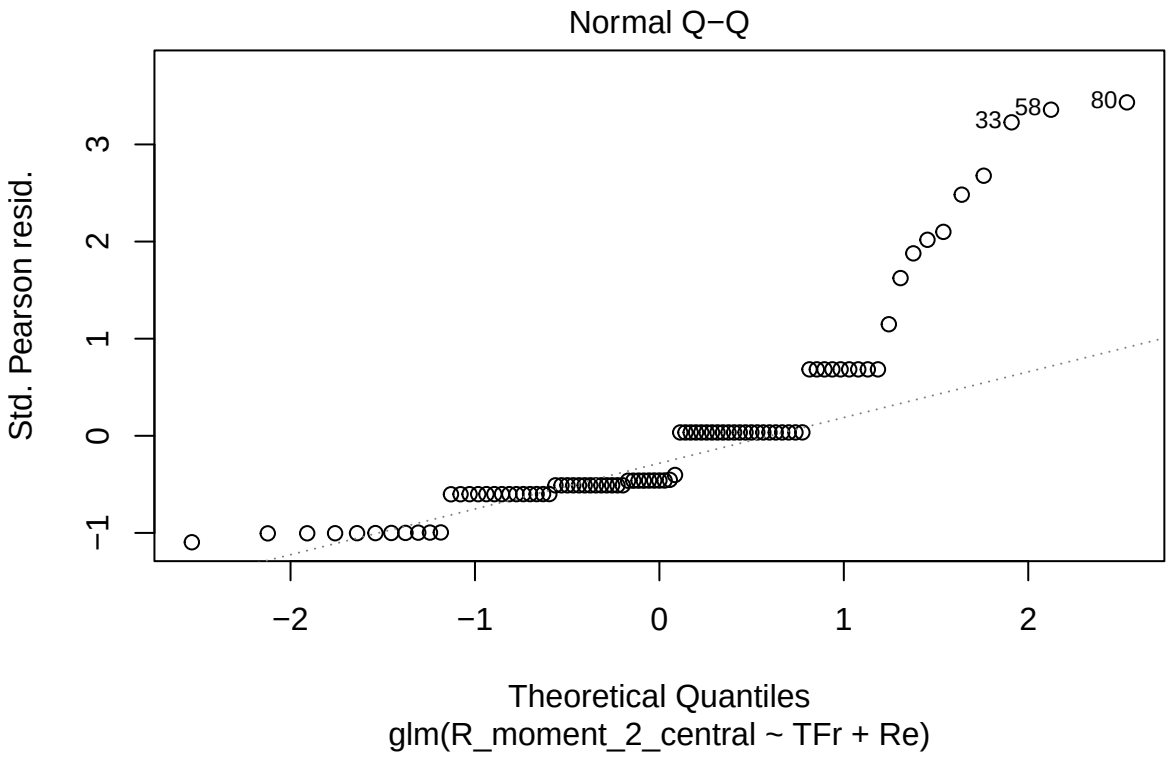
Linear fitting on central moments 2 through 4

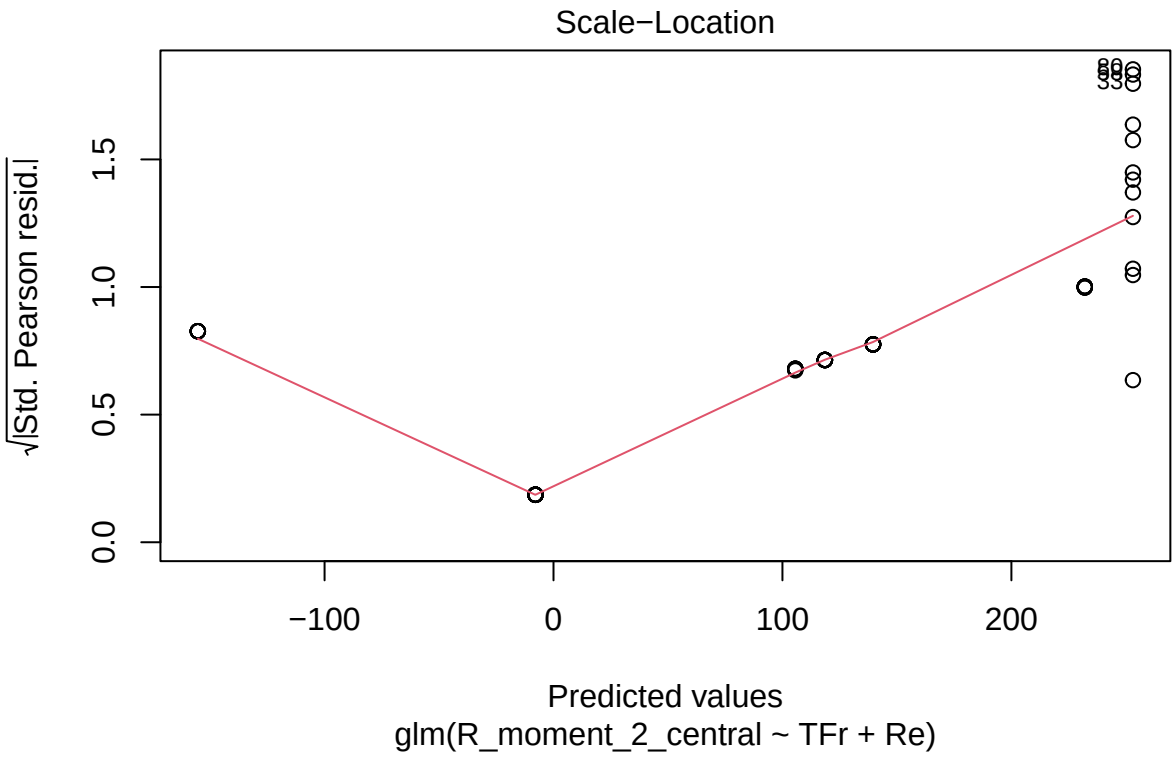
```
##
## Call:
## glm(formula = R_moment_2_central ~ TFr + Re, data = data_train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -252.57  -139.16  -104.99     7.98   791.18
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 299.6445    53.6449   5.586 2.68e-07 ***
## TFr         -10.2315     3.8471  -2.660 0.009332 **
## Re           -0.8472     0.2221  -3.815 0.000256 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 54789.33)
```

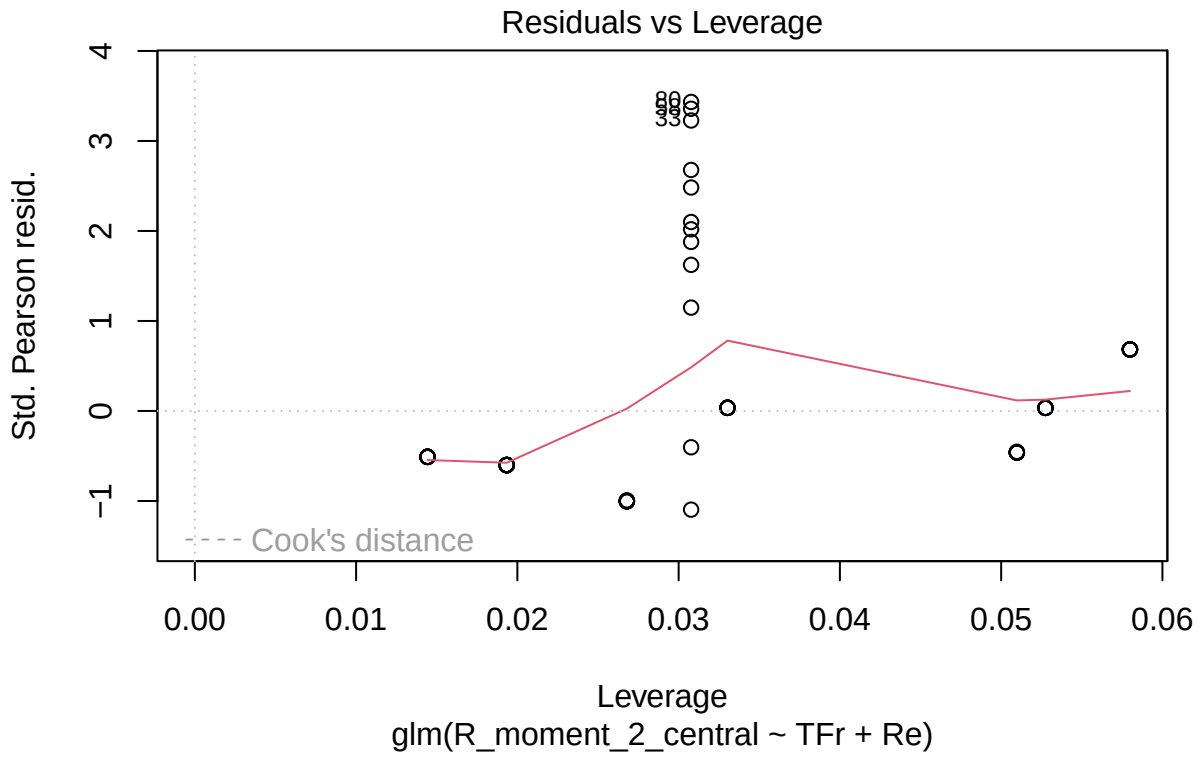
```
##
##     Null deviance: 6032130  on 88  degrees of freedom
## Residual deviance: 4711882  on 86  degrees of freedom
## AIC: 1228.6
##
## Number of Fisher Scoring iterations: 2
```

```
plot(step_full_linear_E2_central)
```




Normal Q−Q
Std. Pearson resid.
−1
0
1
2
3
33
58
80
−2
−1
0
1
2
Theoretical Quantiles
glm(R_moment_2_central ~ TFr + Re)


Scale−Location
√|Std. Pearson resid.|
0.0
0.5
1.0
1.5
−100
0
100
200
Predicted values
glm(R_moment_2_central ~ TFr + Re)

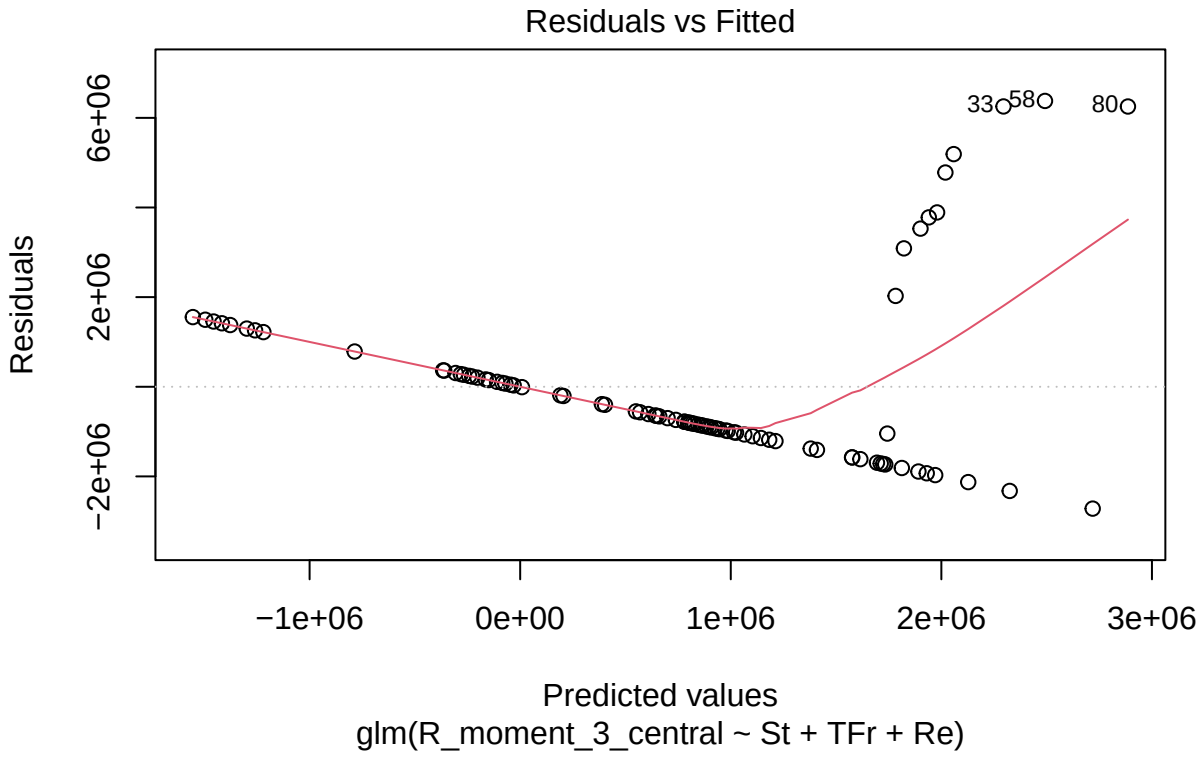


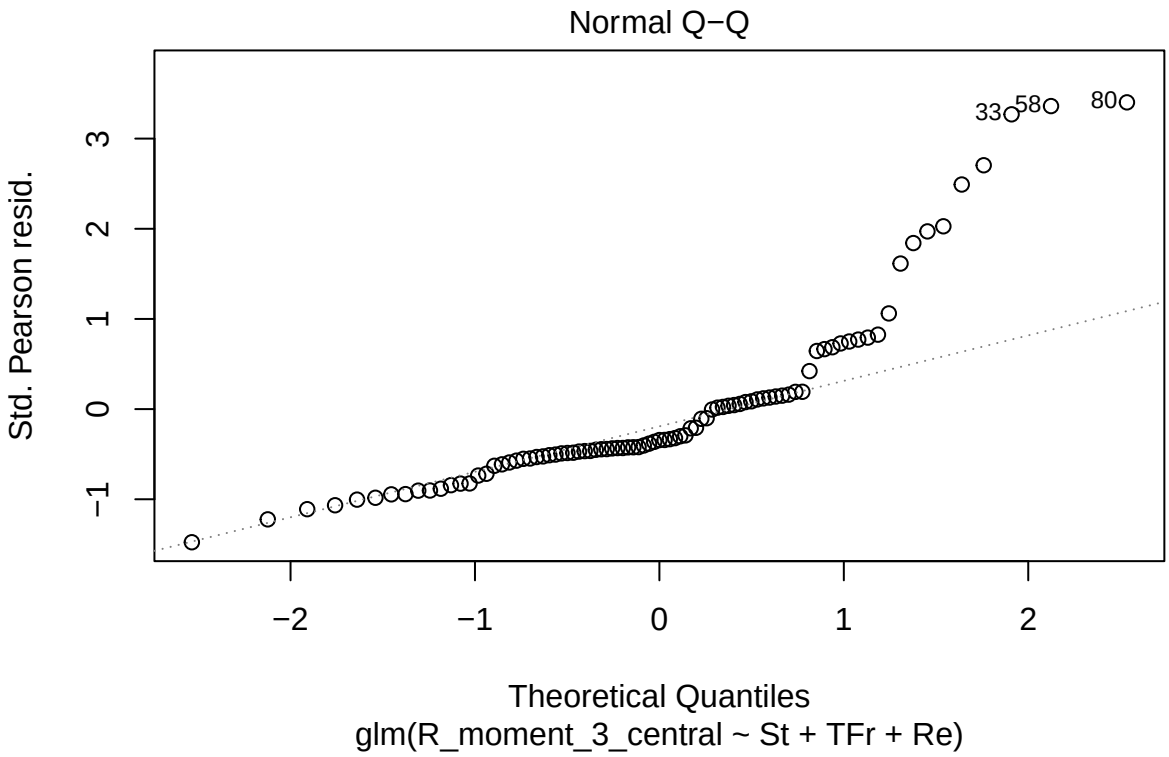

```r
full_linear_E3_central <- glm(R_moment_3_central ~ St + TFr + Re, data = data_train)
step_full_linear_E3_central <- stepAIC(full_linear_E3_central, direction = "both", trace = FALSE)
summary(step_full_linear_E3_central)
```

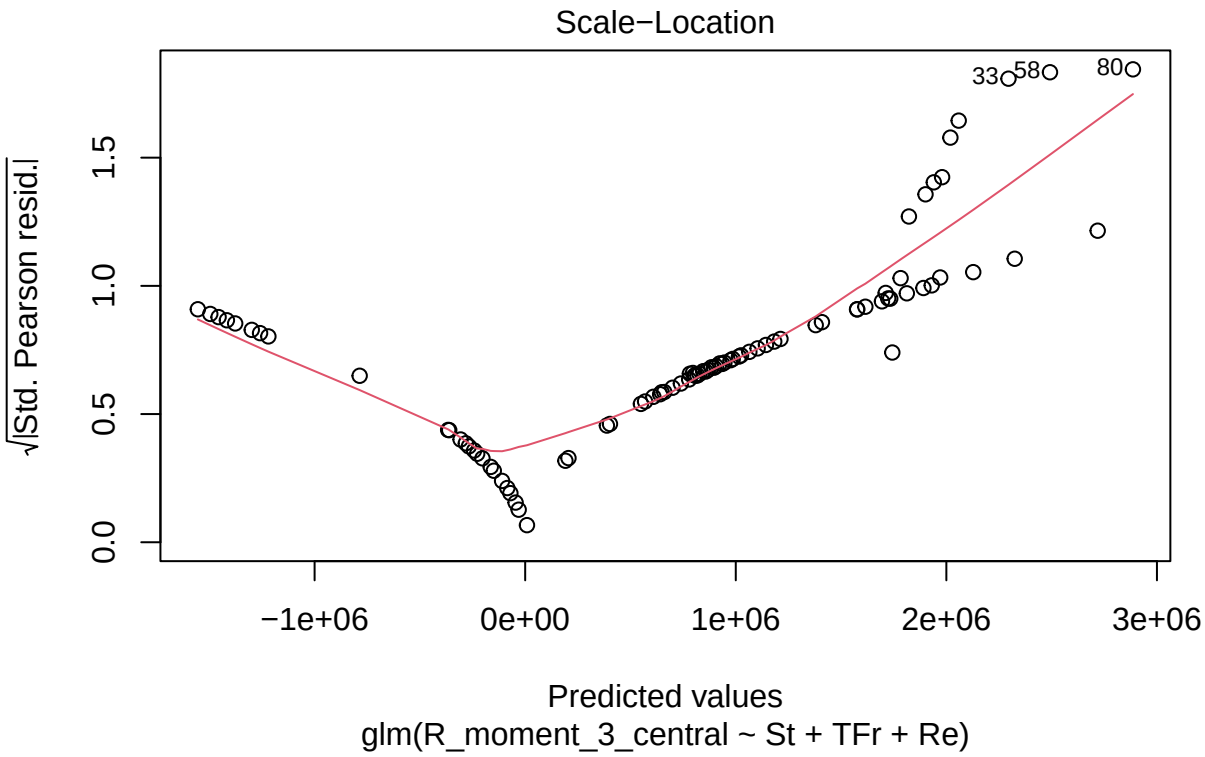
```
##
## Call:
## glm(formula = R_moment_3_central ~ St + TFr + Re, data = data_train)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2718640  -1024952   -660538    281872   6377459
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2082346     508033   4.099 9.46e-05 ***
## St            394050     264781   1.488 0.140396
## TFr           -81444      32075  -2.539 0.012933 *
## Re             -6831       1851  -3.691 0.000393 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 3.801057e+12)
##
##     Null deviance: 4.1934e+14  on 88  degrees of freedom
## Residual deviance: 3.2309e+14  on 85  degrees of freedom
```

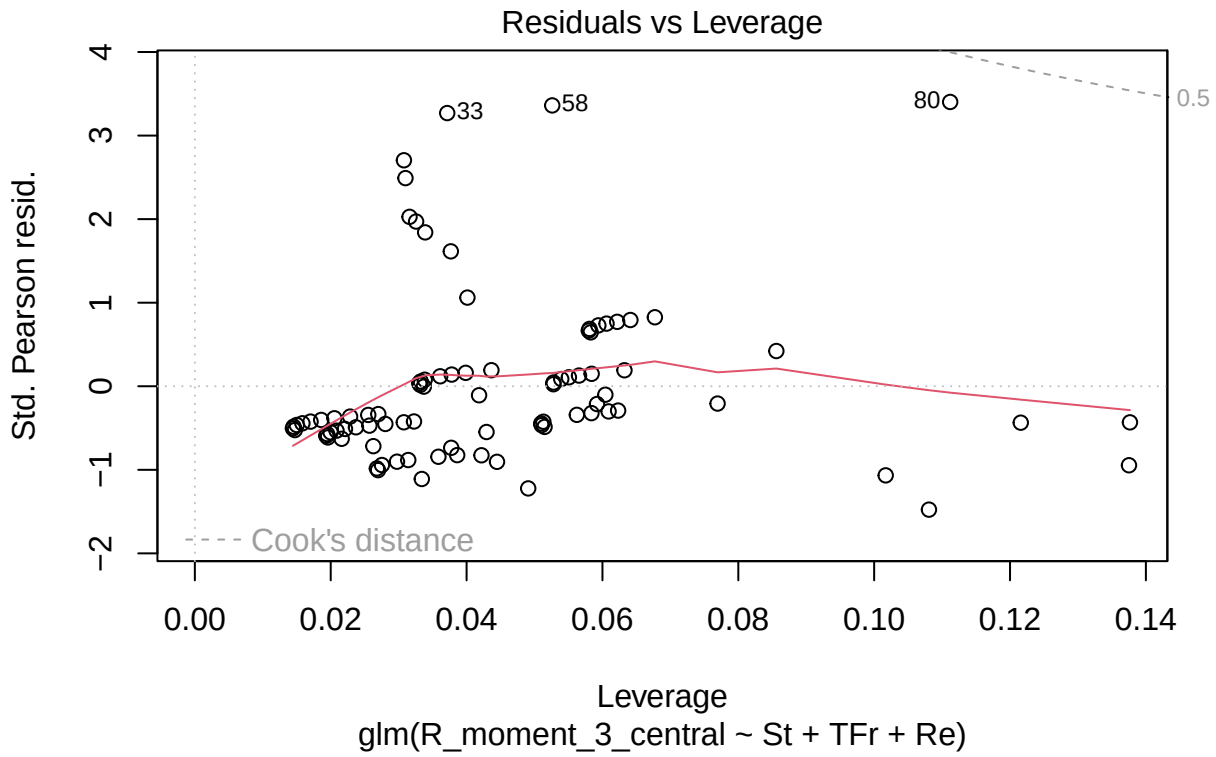
```
## AIC: 2836.5
##
## Number of Fisher Scoring iterations: 2
```

```
plot(step_full_linear_E3_central)
```




Normal Q−Q
Std. Pearson resid.
−1
0
1
2
3
33
58
80
−2
−1
0
1
2
Theoretical Quantiles
glm(R_moment_3_central ~ St + TFr + Re)


Scale−Location
√|Std. Pearson resid.|
0.0
0.5
1.0
1.5
33
58
80
−1e+06
0e+00
1e+06
2e+06
3e+06
Predicted values
glm(R_moment_3_central ~ St + TFr + Re)

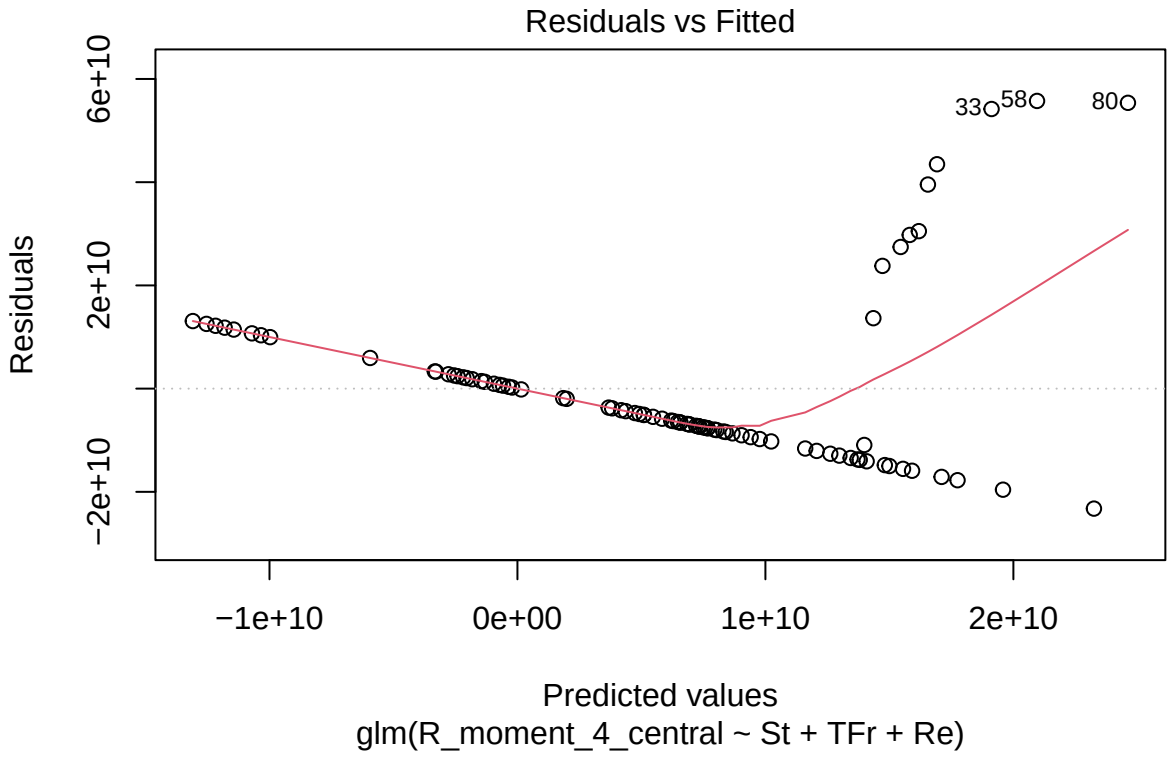


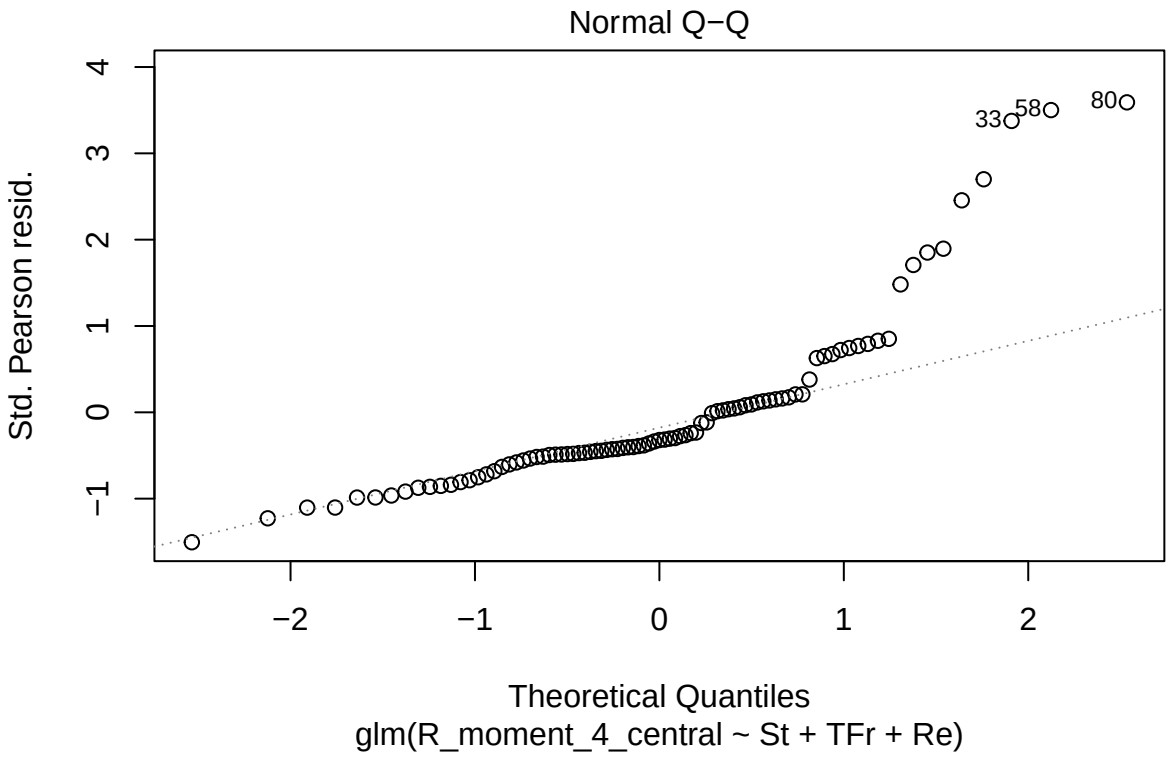

```
full_linear_E4_central <- glm(R_moment_4_central ~ St + TFr + Re, data = data_train)
step_full_linear_E4_central <- stepAIC(full_linear_E4_central, direction = "both", trace = FALSE)
summary(step_full_linear_E4_central)
```

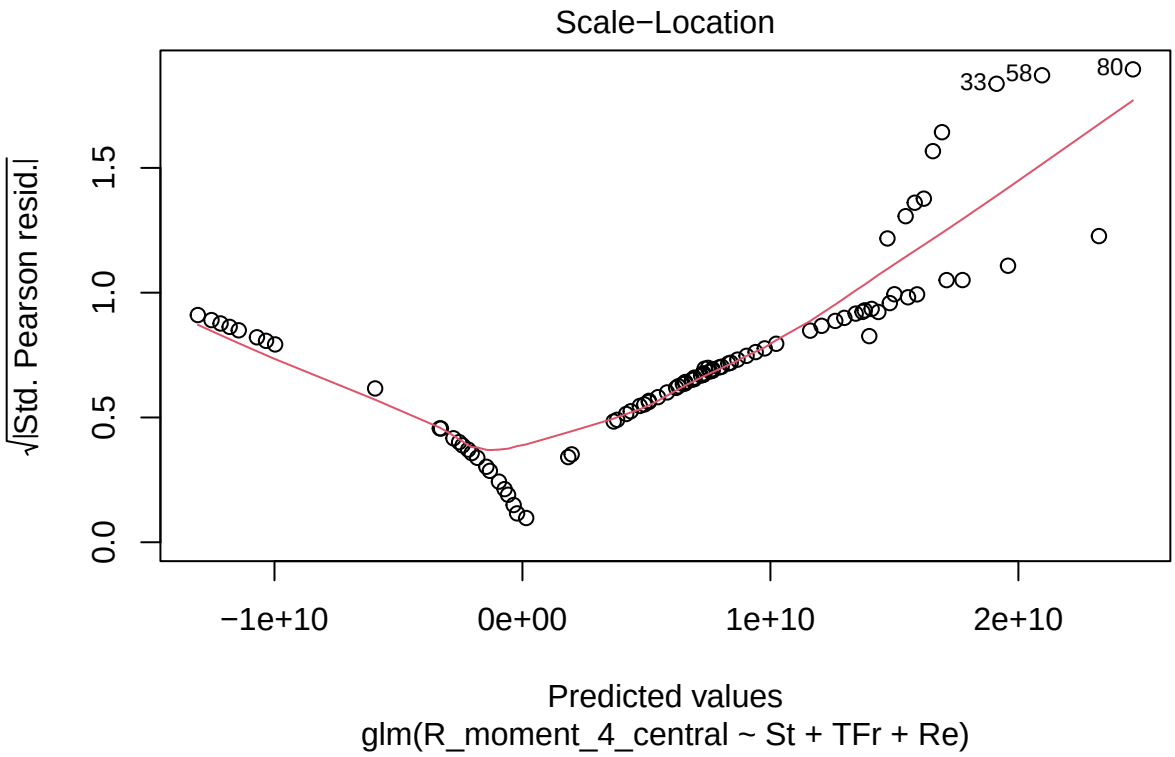
```
##
## Call:
## glm(formula = R_moment_4_central ~ St + TFr + Re, data = data_train)
##
## Deviance Residuals:
##        Min          1Q      Median          3Q         Max
## -2.325e+10  -8.400e+09  -5.100e+09   2.554e+09   5.574e+10
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.673e+10  4.262e+09   3.925 0.000175 ***
## St           3.667e+09  2.222e+09   1.651 0.102522
## TFr         -6.673e+08  2.691e+08  -2.480 0.015126 *
## Re          -5.609e+07  1.553e+07  -3.612 0.000513 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 2.675639e+20)
##
##     Null deviance: 2.9409e+22  on 88  degrees of freedom
## Residual deviance: 2.2743e+22  on 85  degrees of freedom
```

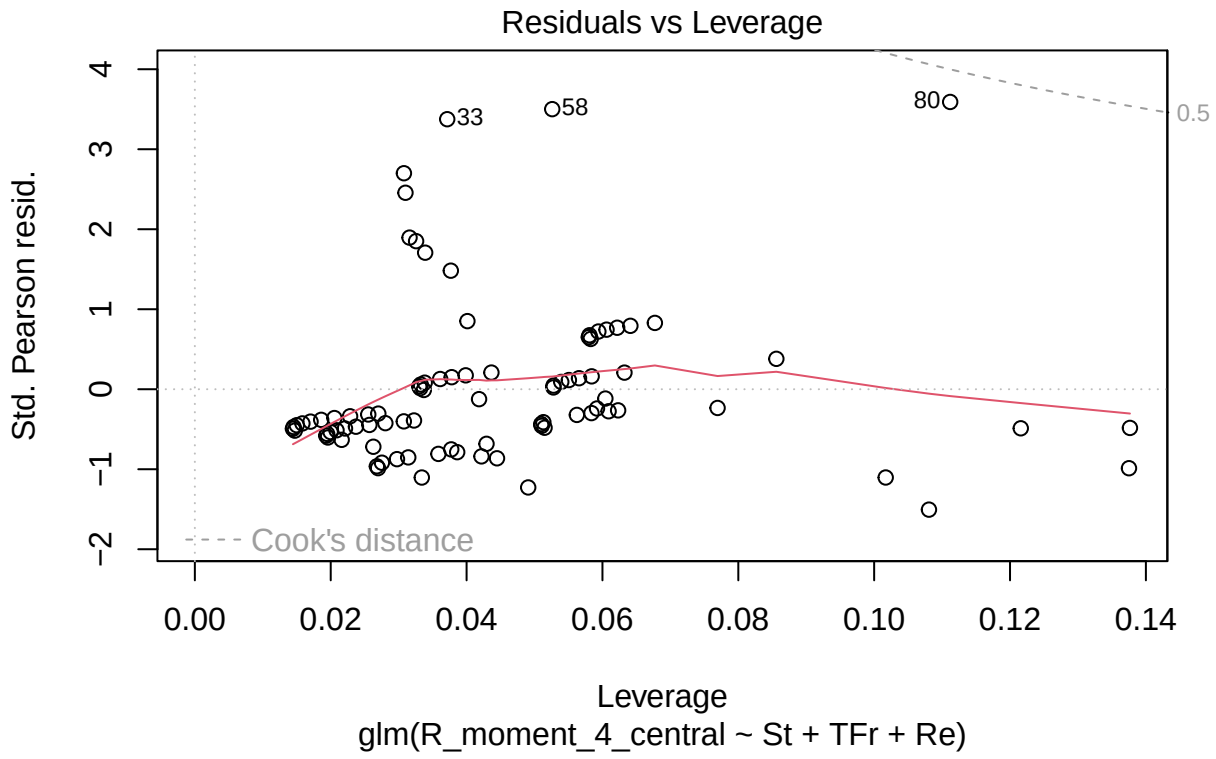
```
## AIC: 4444.7
##
## Number of Fisher Scoring iterations: 2
```

```
plot(step_full_linear_E4_central)
```




Normal Q−Q
Std. Pearson resid.
Theoretical Quantiles
glm(R_moment_4_central ~ St + TFr + Re)
33
58
80


Scale−Location
√|Std. Pearson resid.|
0.0
0.5
1.0
1.5
33
58
80
−1e+10
0e+00
1e+10
2e+10
Predicted values
glm(R_moment_4_central ~ St + TFr + Re)



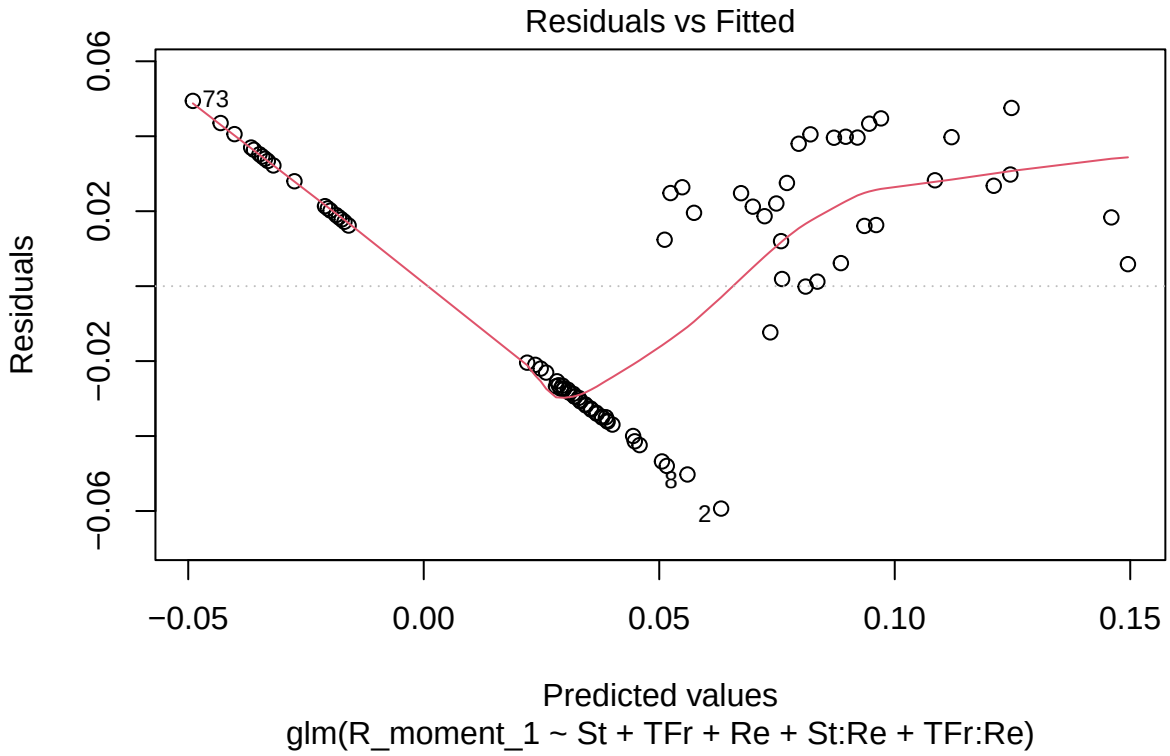

```
full_linear_interactions_E1 <- glm(R_moment_1 ~ St*TFr + St*Re + TFr*Re, data = data_train)
step_full_linear_interactions_E1 <- stepAIC(full_linear_interactions_E1, direction = "both", trace = FA
summary(step_full_linear_interactions_E1)
```

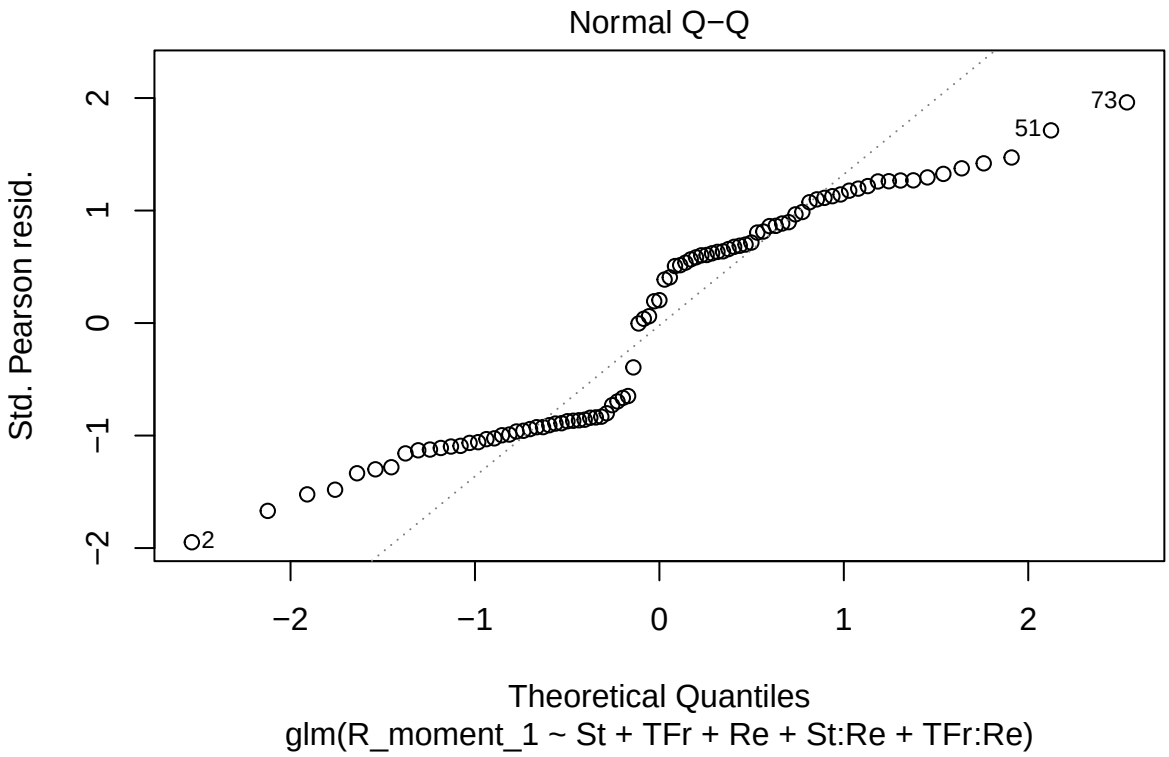
Best AiC model with interactions

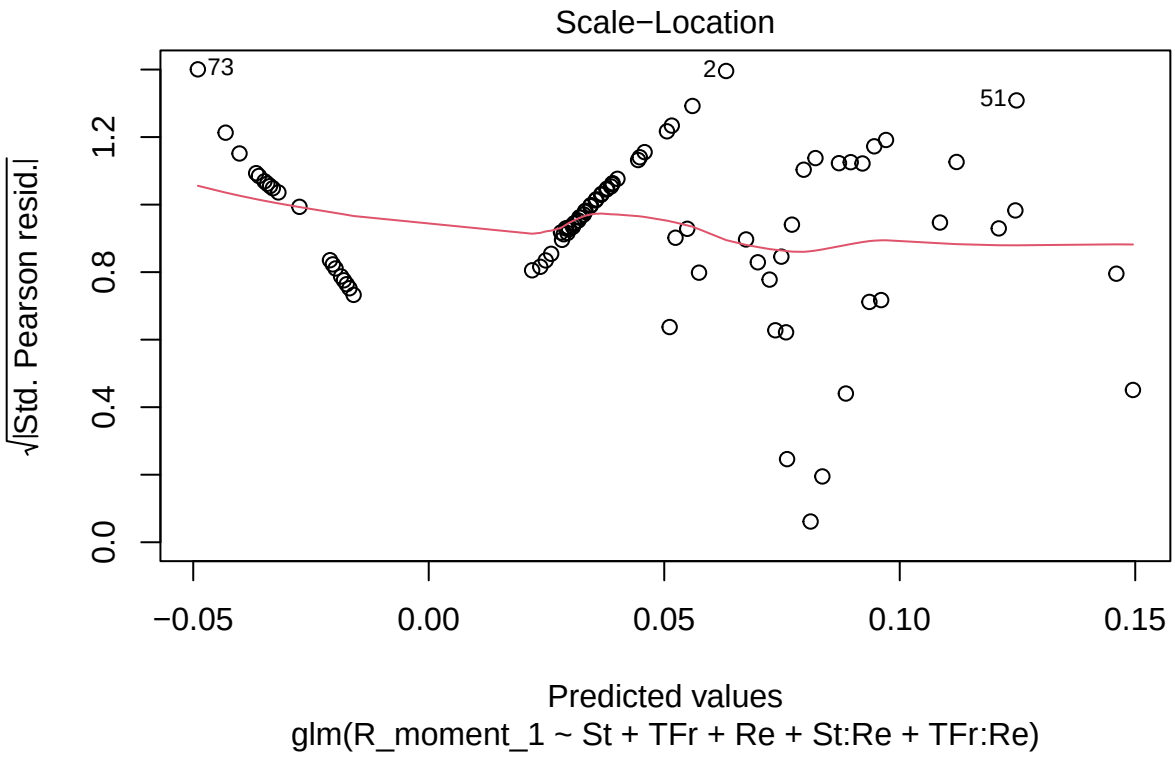
```
##
## Call:
## glm(formula = R_moment_1 ~ St + TFr + Re + St:Re + TFr:Re, data = data_train)
##
## Deviance Residuals:
##       Min         1Q     Median         3Q        Max
## -0.059348  -0.029496   0.006145   0.027529   0.049423
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  9.822e-02  1.140e-02   8.615 3.85e-13 ***
## St           3.398e-02  8.969e-03   3.789 0.000286 ***
## TFr         -2.534e-03  1.161e-03  -2.182 0.031927 *
## Re          -3.176e-04  4.925e-05  -6.448 7.08e-09 ***
## St:Re       -1.002e-04  3.899e-05  -2.570 0.011953 *
## TFr:Re       9.098e-06  4.559e-06   1.995 0.049275 *
## ---
```

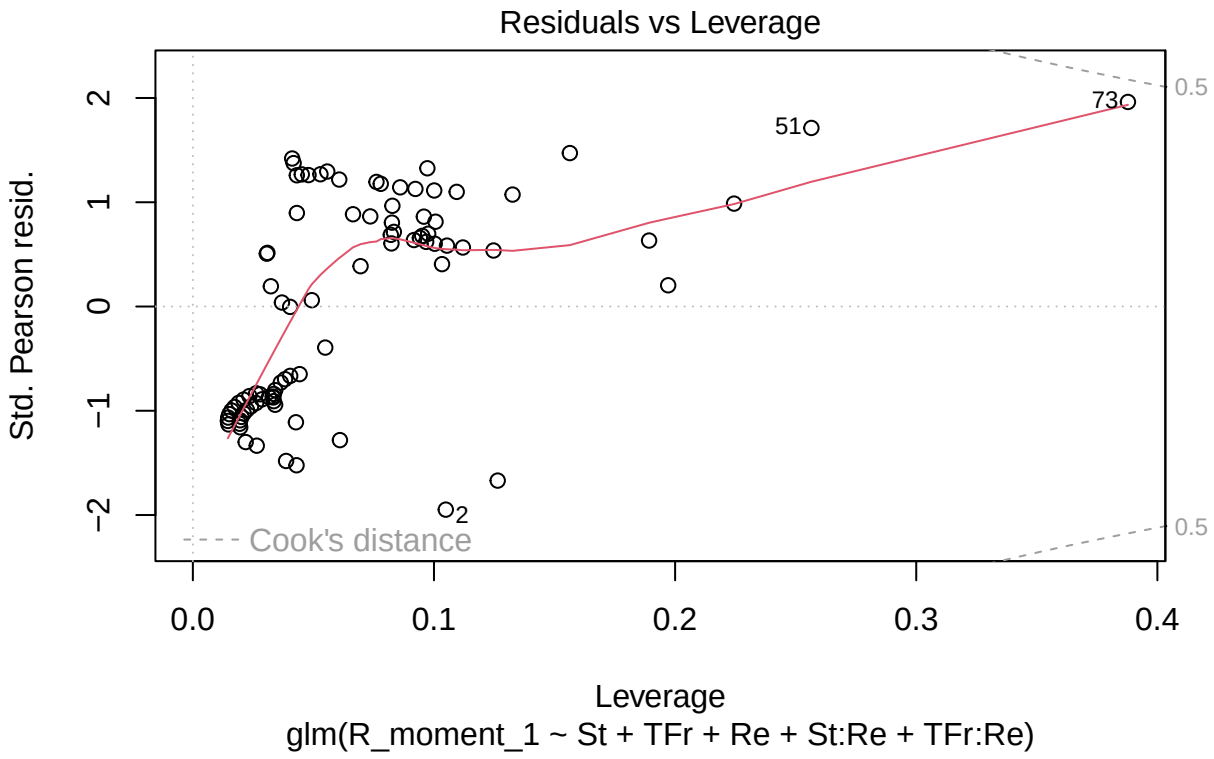
```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.001036755)
##
##     Null deviance: 0.274427  on 88  degrees of freedom
## Residual deviance: 0.086051  on 83  degrees of freedom
## AIC: -351.22
##
## Number of Fisher Scoring iterations: 2
```

```
plot(step_full_linear_interactions_E1)
```




Normal Q−Q
Std. Pearson resid.
Theoretical Quantiles
glm(R_moment_1 ~ St + TFr + Re + St:Re + TFr:Re)
73
51
2
−2
−1
0
1
2


Scale−Location
73
2
51
√|Std. Pearson resid.|
0.0
0.4
0.8
1.2
−0.05
0.00
0.05
0.10
0.15
Predicted values
glm(R_moment_1 ~ St + TFr + Re + St:Re + TFr:Re)

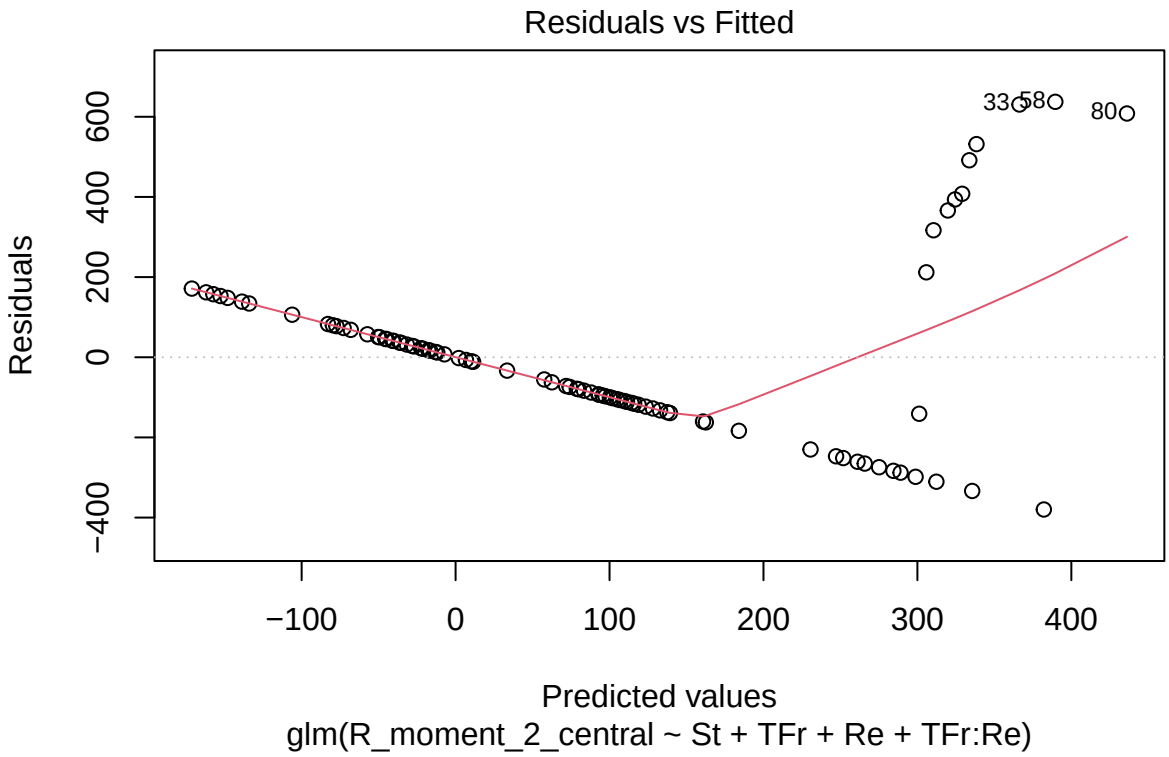


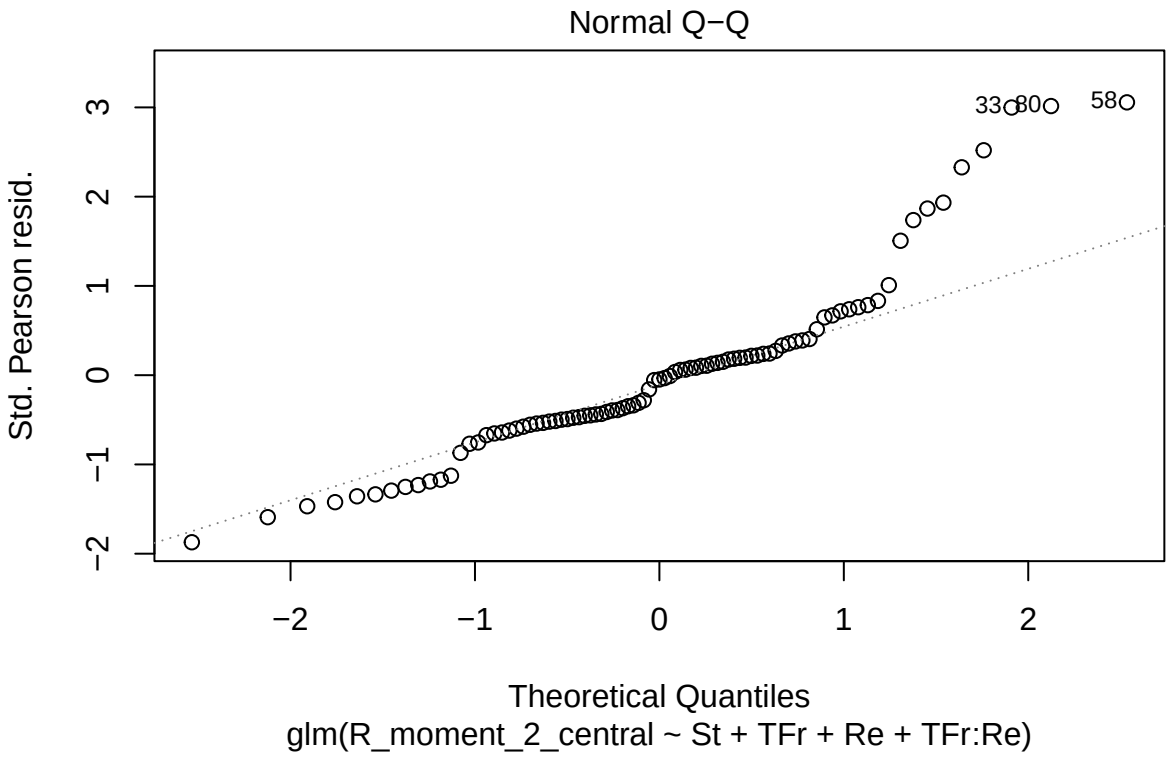

```
full_linear_interactions_E2 <- glm(R_moment_2_central ~ St*TFr + St*Re + TFr*Re, data = data_train)
step_full_linear_interactions_E2 <- stepAIC(full_linear_interactions_E2, direction = "both", trace = FA
summary(step_full_linear_interactions_E2)
```

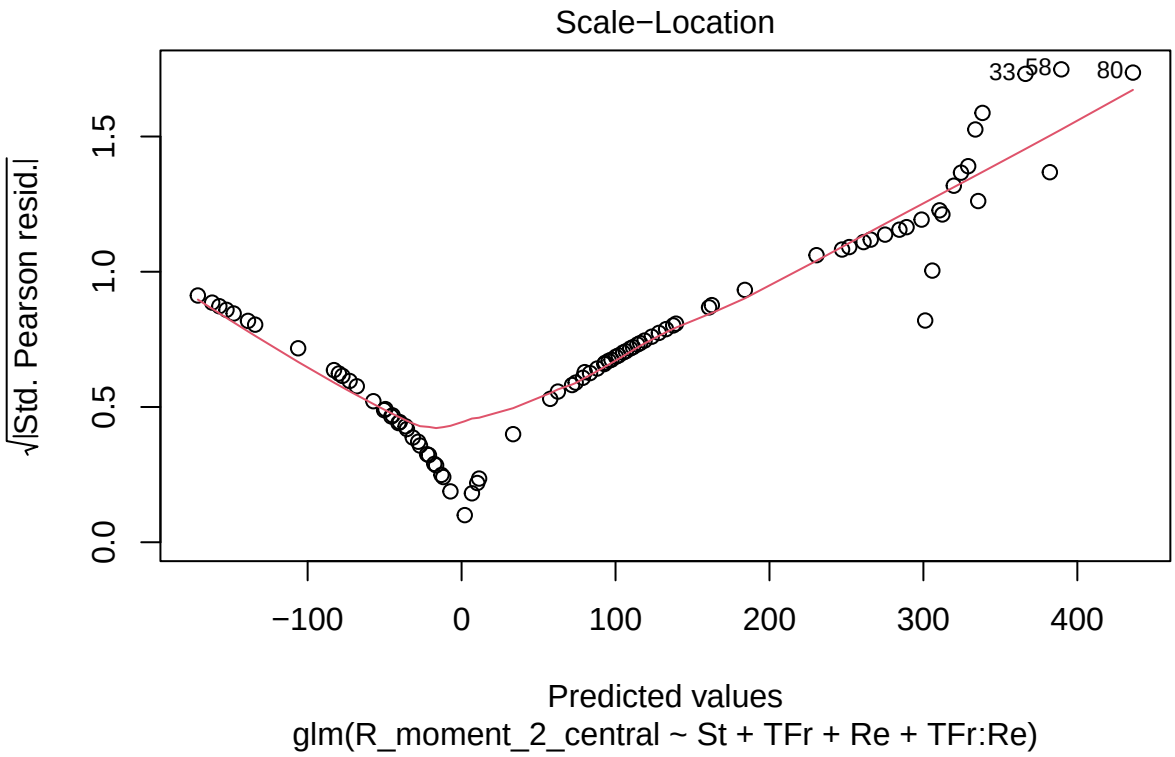
```
##
## Call:
## glm(formula = R_moment_2_central ~ St + TFr + Re + TFr:Re, data = data_train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -379.83  -115.96   -10.12    68.41   637.30
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 327.48973   58.60963   5.588 2.78e-07 ***
## St           46.54204   29.30260   1.588 0.115970
## TFr         -36.88006    7.71023  -4.783 7.29e-06 ***
## Re           -1.19141    0.22338  -5.333 8.00e-07 ***
## TFr:Re        0.11802    0.03007   3.925 0.000177 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 46461.98)
##
##     Null deviance: 6032130  on 88  degrees of freedom
```

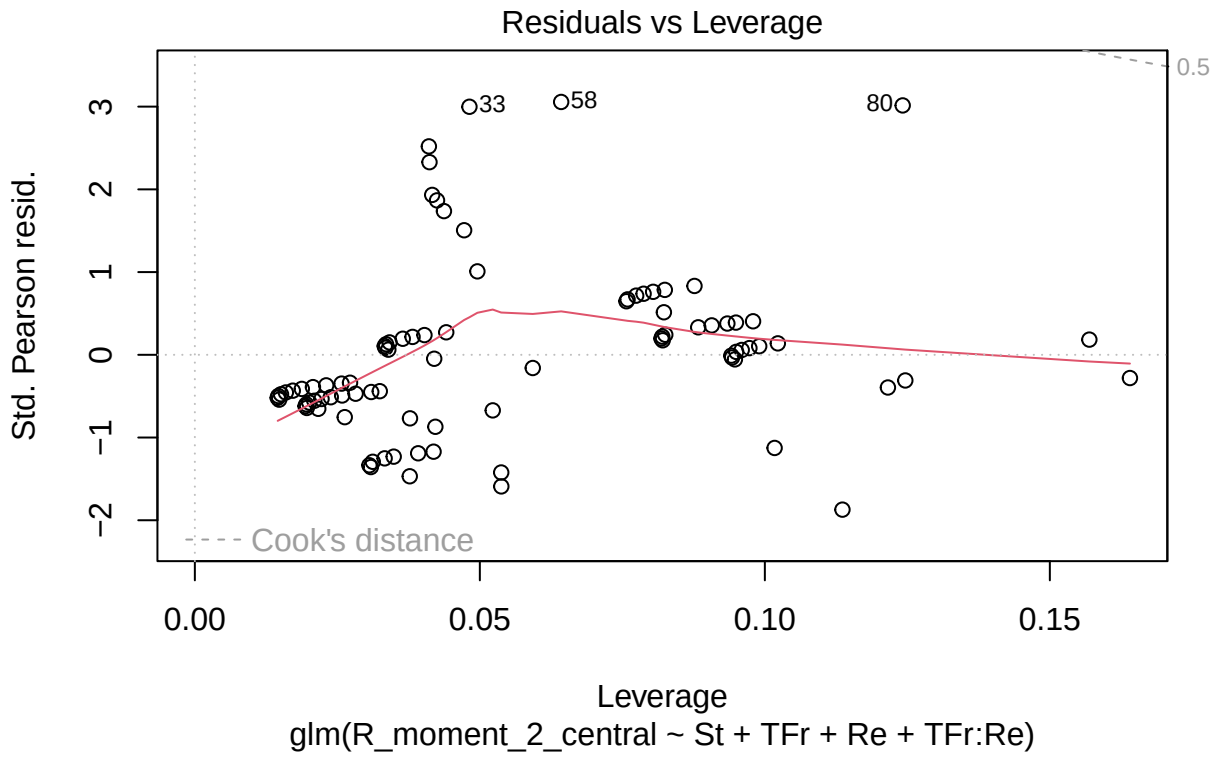
```
## Residual deviance: 3902807  on 84  degrees of freedom
## AIC: 1215.9
##
## Number of Fisher Scoring iterations: 2
```

```
plot(step_full_linear_interactions_E2)
```




Normal Q−Q
Std. Pearson resid.
Theoretical Quantiles
glm(R_moment_2_central ~ St + TFr + Re + TFr:Re)
33
80
58
−2
−1
0
1
2
3


Scale−Location
√|Std. Pearson resid.|
0.0
0.5
1.0
1.5
−100
0
100
200
300
400
33
58
80
Predicted values
glm(R_moment_2_central ~ St + TFr + Re + TFr:Re)

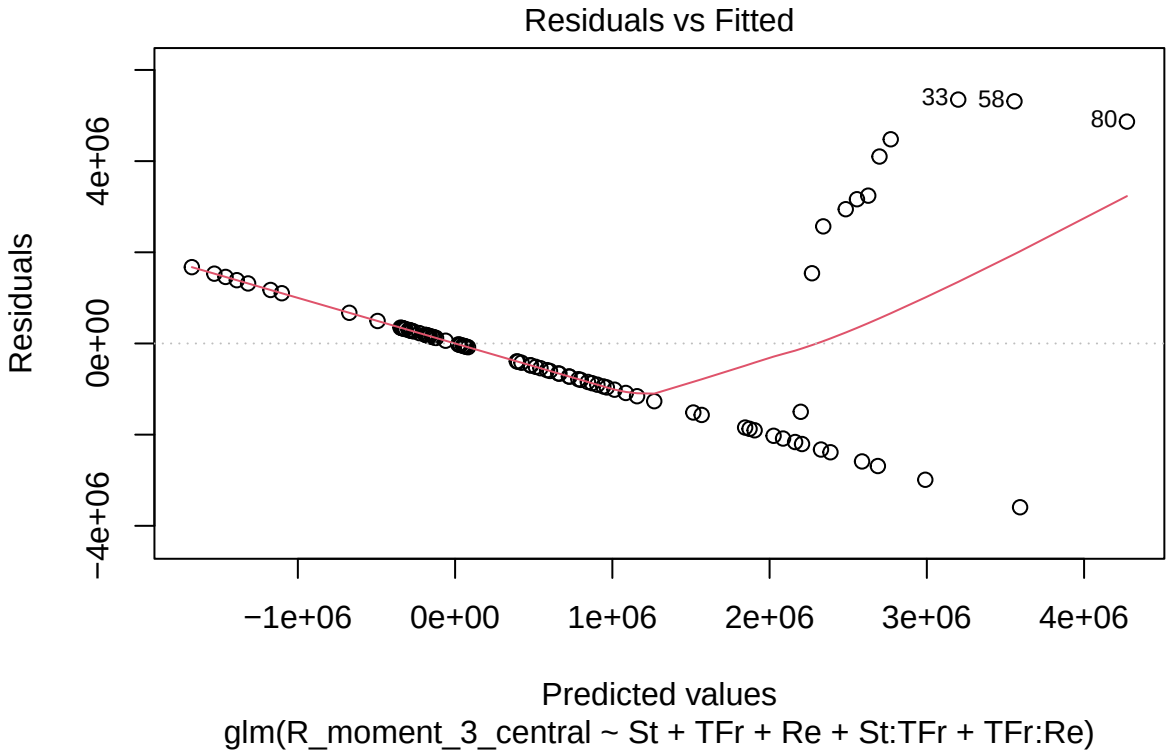


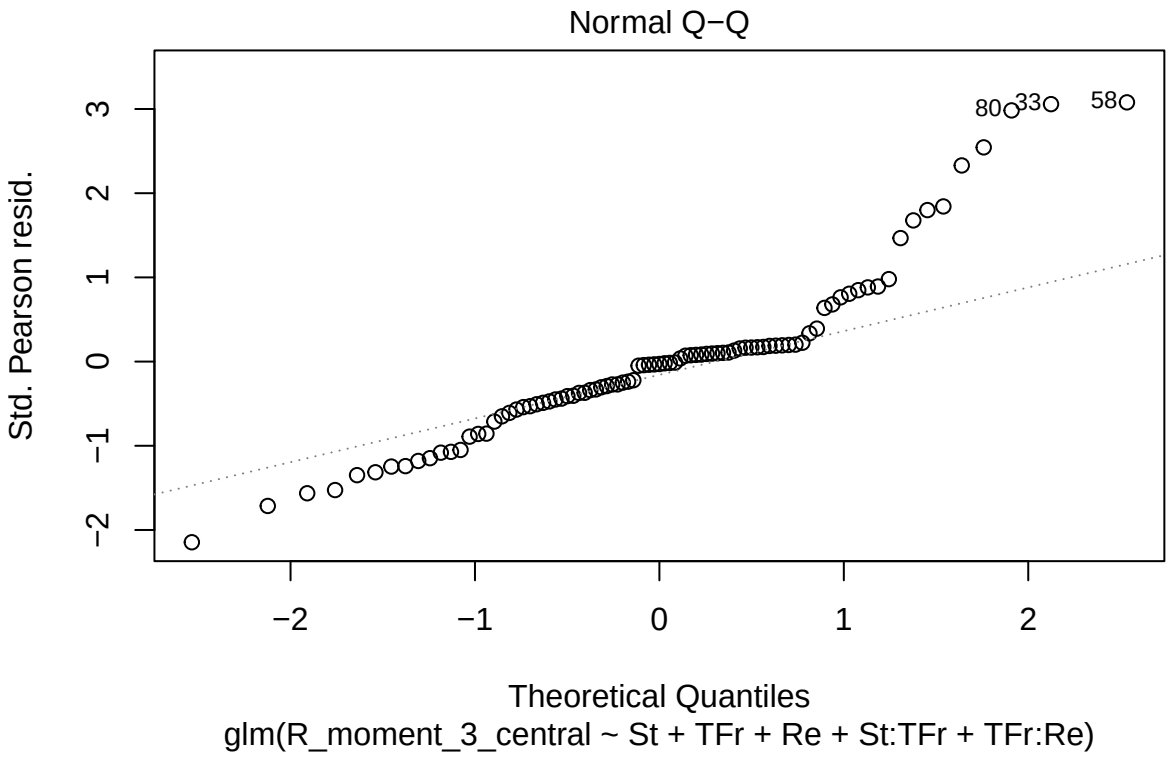

```
full_linear_interactions_E3 <- glm(R_moment_3_central ~ St*TFr + St*Re + TFr*Re, data = data_train)
step_full_linear_interactions_E3 <- stepAIC(full_linear_interactions_E3, direction = "both", trace = FA
summary(step_full_linear_interactions_E3)
```

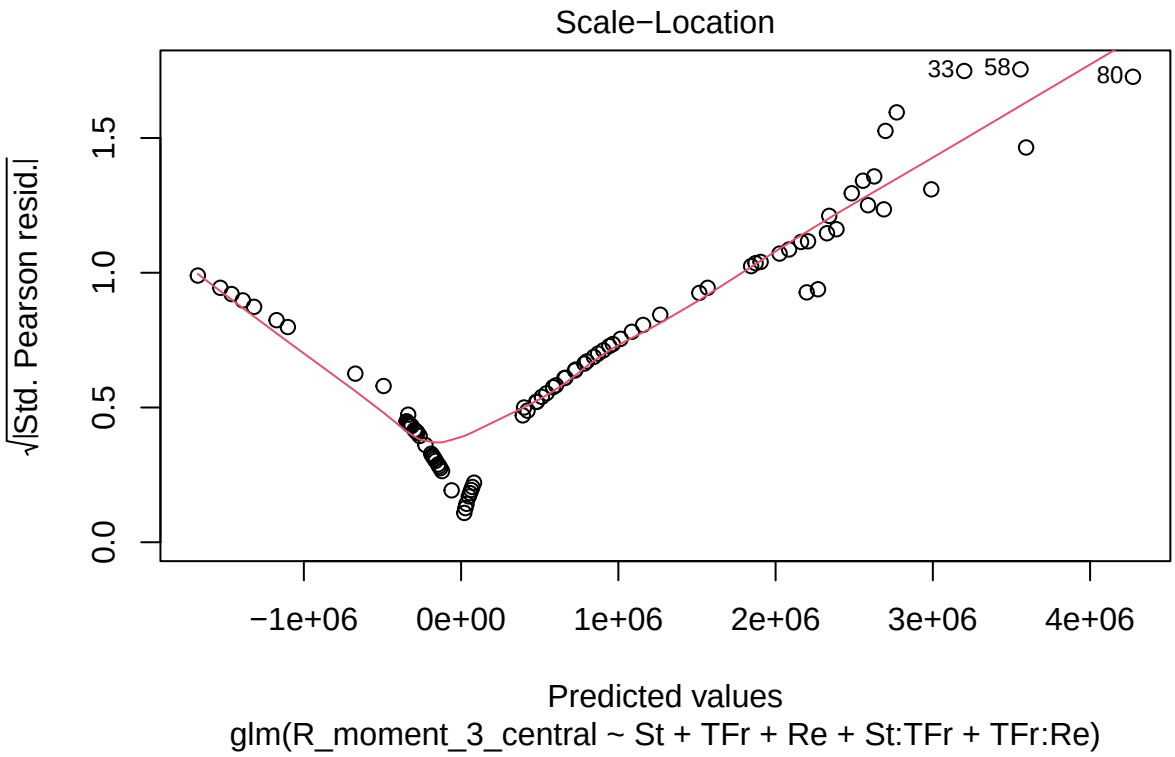
```
##
## Call:
## glm(formula = R_moment_3_central ~ St + TFr + Re + St:TFr + TFr:Re,
##     data = data_train)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -3592944   -904717    -49411    332389   5349989
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2525572.0   493187.5   5.121 1.94e-06 ***
## St           556593.3   258219.6   2.156 0.034018 *
## TFr         -252297.4    72716.5  -3.470 0.000829 ***
## Re            -9805.2     1864.1  -5.260 1.10e-06 ***
## St:TFr       -54689.6    37680.2  -1.451 0.150434
## TFr:Re           953.2      250.9   3.798 0.000276 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 3.231922e+12)
```

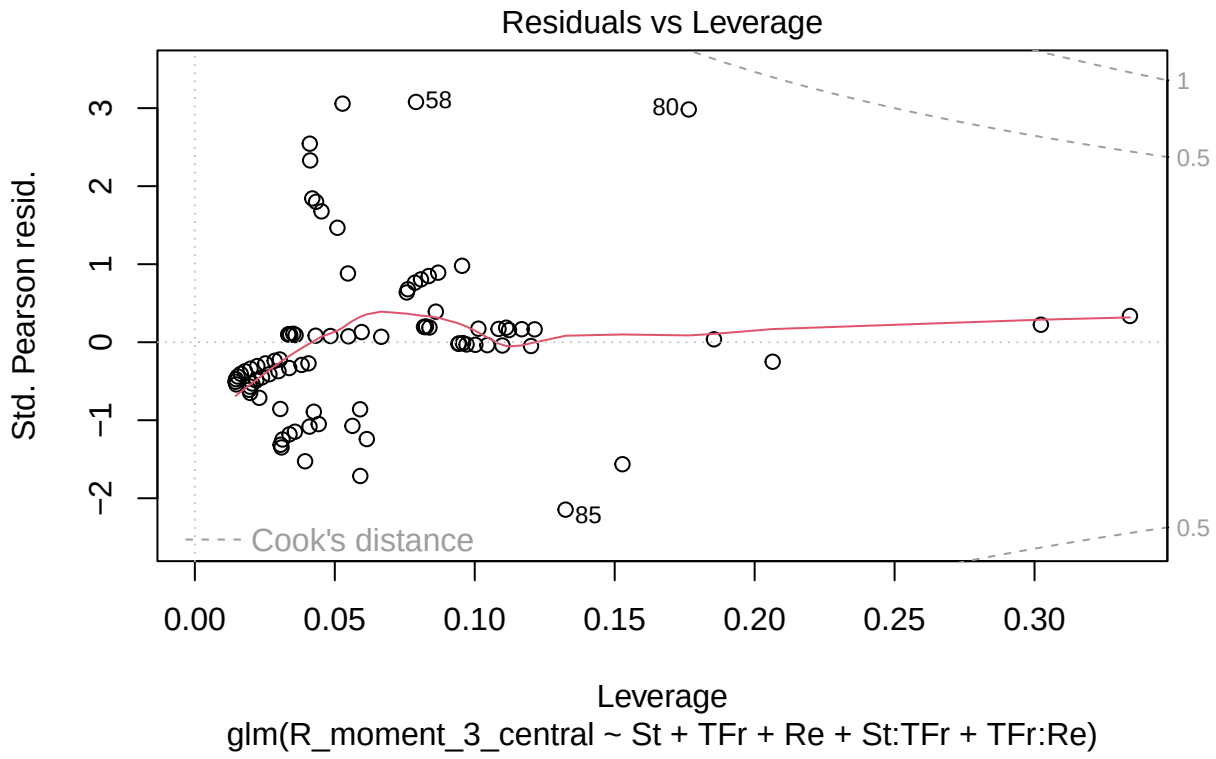
```
##
##     Null deviance: 4.1934e+14  on 88  degrees of freedom
## Residual deviance: 2.6825e+14  on 83  degrees of freedom
## AIC: 2823.9
##
## Number of Fisher Scoring iterations: 2
```

```
plot(step_full_linear_interactions_E3)
```




Normal Q−Q
Std. Pearson resid.
−2
−1
0
1
2
3
80
33
58
−2
−1
0
1
2
Theoretical Quantiles
glm(R_moment_3_central ~ St + TFr + Re + St:TFr + TFr:Re)


Scale−Location
√|Std. Pearson resid.|
0.0
0.5
1.0
1.5
33
58
80
−1e+06
0e+00
1e+06
2e+06
3e+06
4e+06
Predicted values
glm(R_moment_3_central ~ St + TFr + Re + St:TFr + TFr:Re)

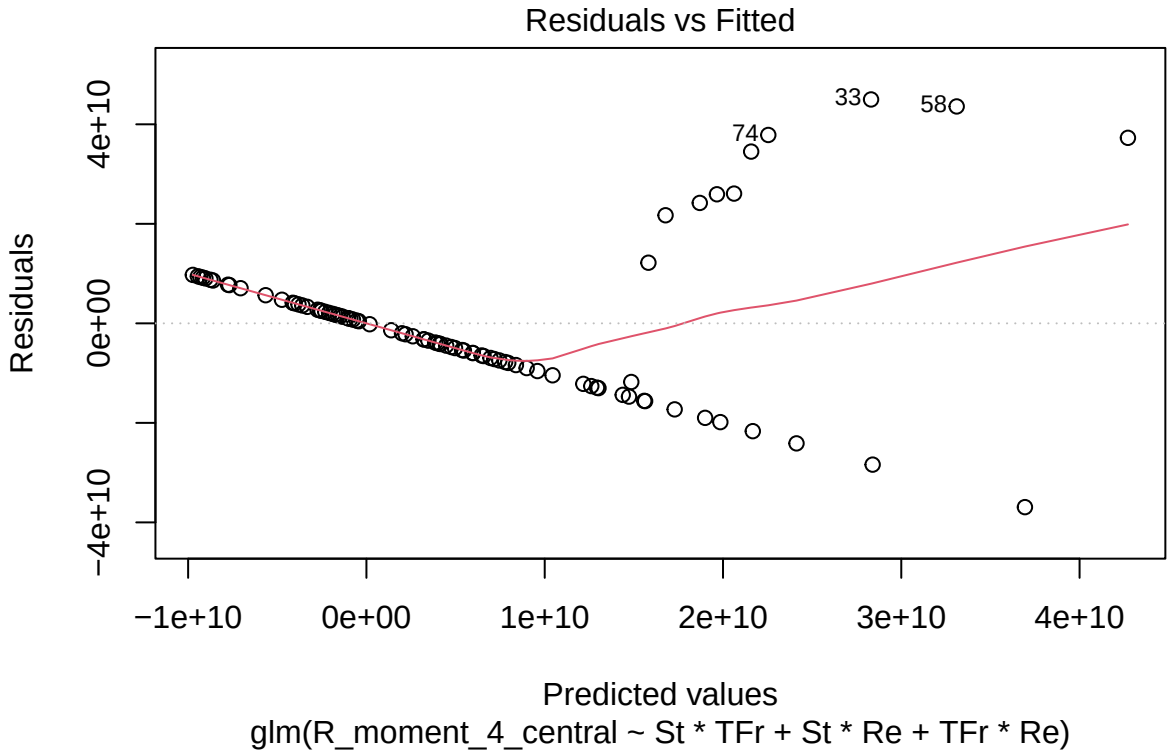


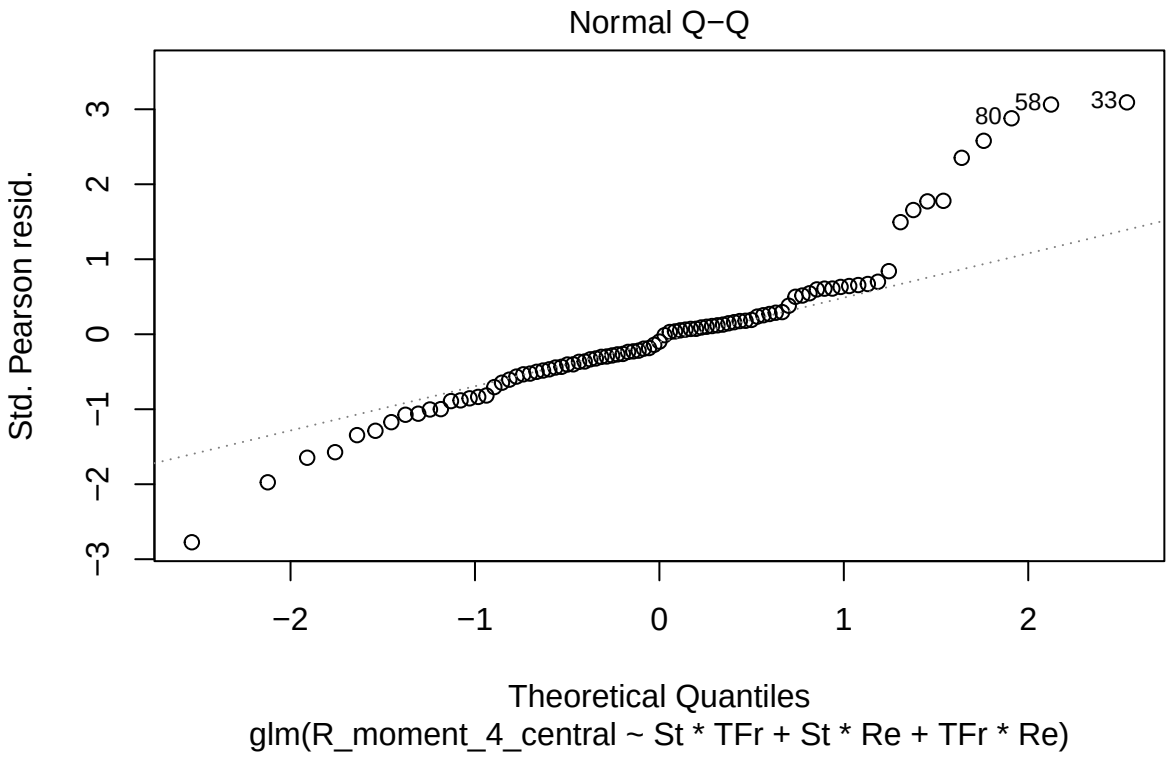

```
full_linear_interactions_E4 <- glm(R_moment_4_central ~ St*TFr + St*Re + TFr*Re, data = data_train)
step_full_linear_interactions_E4 <- stepAIC(full_linear_interactions_E4, direction = "both", trace = FA
summary(step_full_linear_interactions_E4)
```

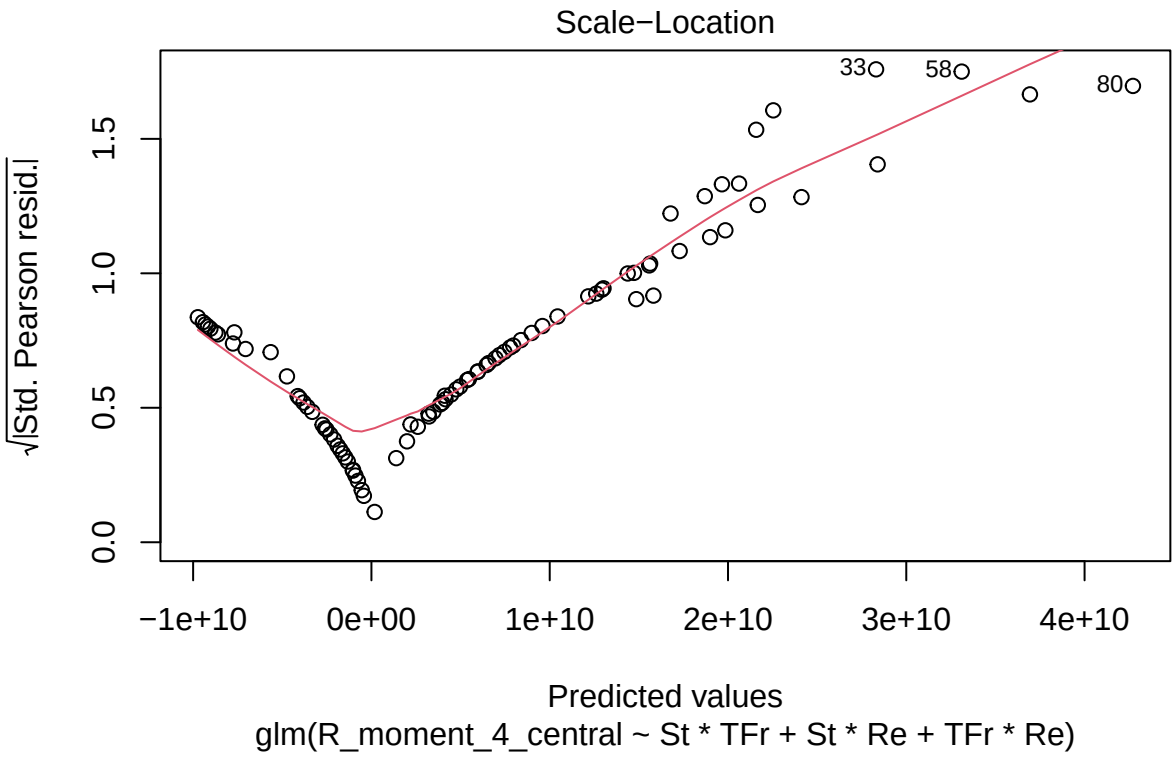
```
##
## Call:
## glm(formula = R_moment_4_central ~ St * TFr + St * Re + TFr *
##     Re, data = data_train)
##
## Deviance Residuals:
##        Min          1Q      Median          3Q         Max
## -3.694e+10  -7.457e+09  -1.392e+09   4.131e+09   4.499e+10
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.528e+10  5.349e+09   2.857 0.005418 **
## St           1.050e+10  4.256e+09   2.467 0.015707 *
## TFr         -1.915e+09  6.113e+08  -3.133 0.002401 **
## Re          -5.598e+07  2.293e+07  -2.442 0.016773 *
## St:TFr      -5.176e+08  3.144e+08  -1.646 0.103499
## St:Re       -2.662e+07  1.816e+07  -1.466 0.146598
## TFr:Re       7.303e+06  2.125e+06   3.438 0.000924 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

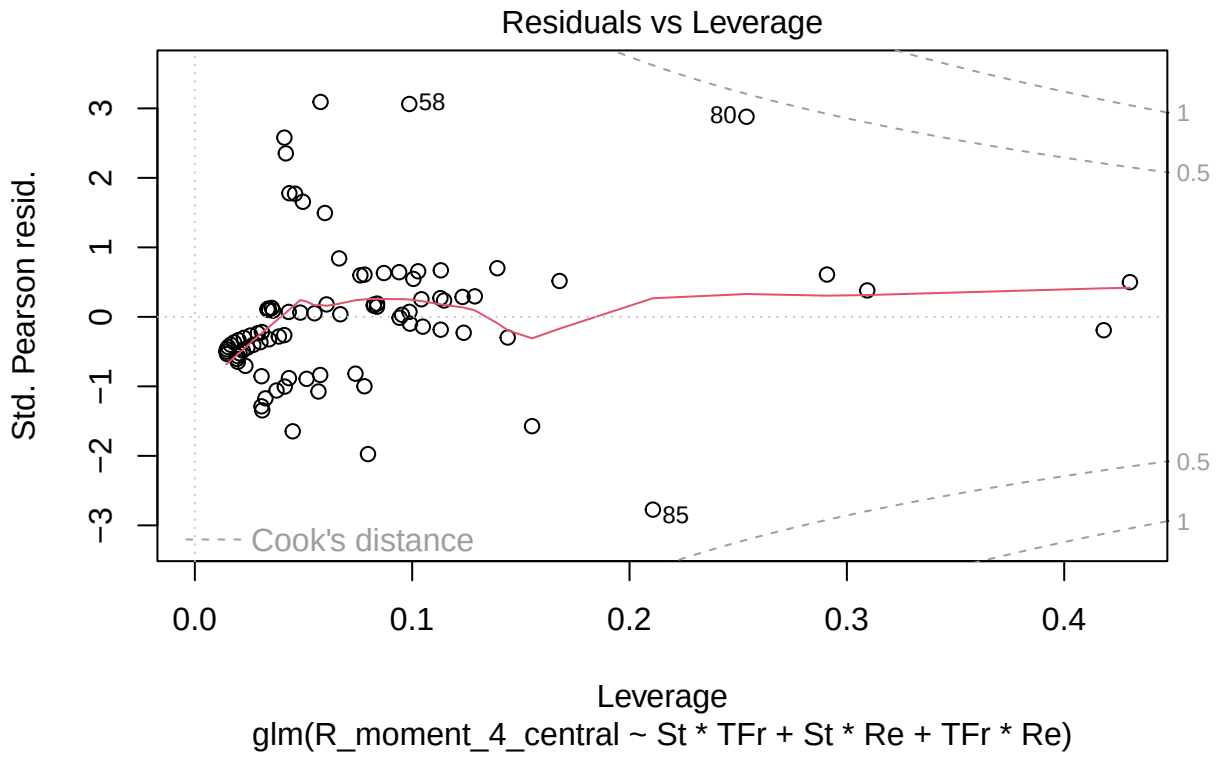
```
## (Dispersion parameter for gaussian family taken to be 2.246926e+20)
##
##     Null deviance: 2.9409e+22  on 88  degrees of freedom
## Residual deviance: 1.8425e+22  on 82  degrees of freedom
## AIC: 4431.9
##
## Number of Fisher Scoring iterations: 2
```

```
plot(step_full_linear_interactions_E4)
```




Normal Q−Q
Std. Pearson resid.
−3
−2
−1
0
1
2
3
80
58
33
−2
−1
0
1
2
Theoretical Quantiles
glm(R_moment_4_central ~ St * TFr + St * Re + TFr * Re)


Scale−Location
33
58
80
√|Std. Pearson resid.|
0.0
0.5
1.0
1.5
−1e+10
0e+00
1e+10
2e+10
3e+10
4e+10
Predicted values
glm(R_moment_4_central ~ St * TFr + St * Re + TFr * Re)




```
library(boot)
```

Model Evaluation (Linear)

```
##
## Attaching package: 'boot'

## The following object is masked from 'package:arm':
##
##     logit

## The following object is masked from 'package:lattice':
##
##     melanoma
```

```
cve_linear_E1 <- cv.glm(data_train, step_full_linear_E1, K=10)
cve_linear_E1$delta
```

```
## [1] 0.001229868 0.001224073
```

```
cve_linear_interactions_E1 <- cv.glm(data_train, step_full_linear_interactions_E1, K = 10)
cve_linear_interactions_E1$delta
```

```
## [1] 0.001090773 0.001083930
```

```
cve_linear_E2 <- cv.glm(data_train, step_full_linear_E2_central, K=10)
cve_linear_E2$delta
```

```
## [1] 57330.80 57097.37
```

```
cve_linear_interactions_E2 <- cv.glm(data_train, step_full_linear_E2_central, K = 10)
cve_linear_interactions_E2$delta
```

```
## [1] 56683.85 56489.28
```

```
cve_linear_E3 <- cv.glm(data_train, step_full_linear_E3_central, K=10)
cve_linear_E3$delta
```

```
## [1] 3.826997e+12 3.816443e+12
```

```
cve_linear_interactions_E3 <- cv.glm(data_train, step_full_linear_interactions_E3, K = 10)
cve_linear_interactions_E3$delta
```

```
## [1] 3.705014e+12 3.666260e+12
```

```
cve_linear_E4 <- cv.glm(data_train, step_full_linear_E4_central, K=10)
cve_linear_E4$delta
```

```
## [1] 2.883521e+20 2.865726e+20
```

```
cve_linear_interactions_E4 <- cv.glm(data_train, step_full_linear_interactions_E4, K = 10)
cve_linear_interactions_E4$delta
```

```
## [1] 2.466960e+20 2.443778e+20
```

It seems that the interactions are increasingly important. They are less important for the first and second moments. In fact, cross validation error increases for the second moment when interactions are added into the model. However, for the third through fourth moments, there is a pretty significant decrease in cross validation error when comparing the strictly linear models versus the ones with interactions.

This I think the linear models worth sharing with our physicist colleagues are the following:

```
summary(step_full_linear_E1)
```

```
##
## Call:
## glm(formula = R_moment_1 ~ St + Re, data = data_train)
##
```

```
## Deviance Residuals:
##       Min         1Q     Median         3Q        Max
## -0.061936  -0.030347  -0.000174   0.034491   0.055714
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.102e-01  8.837e-03  12.475  < 2e-16 ***
## St           1.353e-02  4.621e-03   2.927  0.00438 **
## Re          -3.798e-04  3.215e-05 -11.816  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.001160225)
##
##     Null deviance: 0.274427  on 88  degrees of freedom
## Residual deviance: 0.099779  on 86  degrees of freedom
## AIC: -344.04
##
## Number of Fisher Scoring iterations: 2
```

```
summary(step_full_linear_E2_central)
```

```
##
## Call:
## glm(formula = R_moment_2_central ~ TFr + Re, data = data_train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -252.57  -139.16  -104.99     7.98   791.18
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 299.6445    53.6449   5.586 2.68e-07 ***
## TFr         -10.2315     3.8471  -2.660 0.009332 **
## Re           -0.8472     0.2221  -3.815 0.000256 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 54789.33)
##
##     Null deviance: 6032130  on 88  degrees of freedom
## Residual deviance: 4711882  on 86  degrees of freedom
## AIC: 1228.6
##
## Number of Fisher Scoring iterations: 2
```

```
summary(step_full_linear_interactions_E3)
```

```
##
## Call:
## glm(formula = R_moment_3_central ~ St + TFr + Re + St:TFr + TFr:Re,
##     data = data_train)
##
```

```
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -3592944   -904717    -49411    332389   5349989
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2525572.0   493187.5   5.121 1.94e-06 ***
## St           556593.3   258219.6   2.156 0.034018 *
## TFr         -252297.4    72716.5  -3.470 0.000829 ***
## Re            -9805.2     1864.1  -5.260 1.10e-06 ***
## St:TFr       -54689.6    37680.2  -1.451 0.150434
## TFr:Re          953.2      250.9   3.798 0.000276 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 3.231922e+12)
##
##     Null deviance: 4.1934e+14  on 88  degrees of freedom
## Residual deviance: 2.6825e+14  on 83  degrees of freedom
## AIC: 2823.9
##
## Number of Fisher Scoring iterations: 2
```

```
summary(step_full_linear_interactions_E4)
```

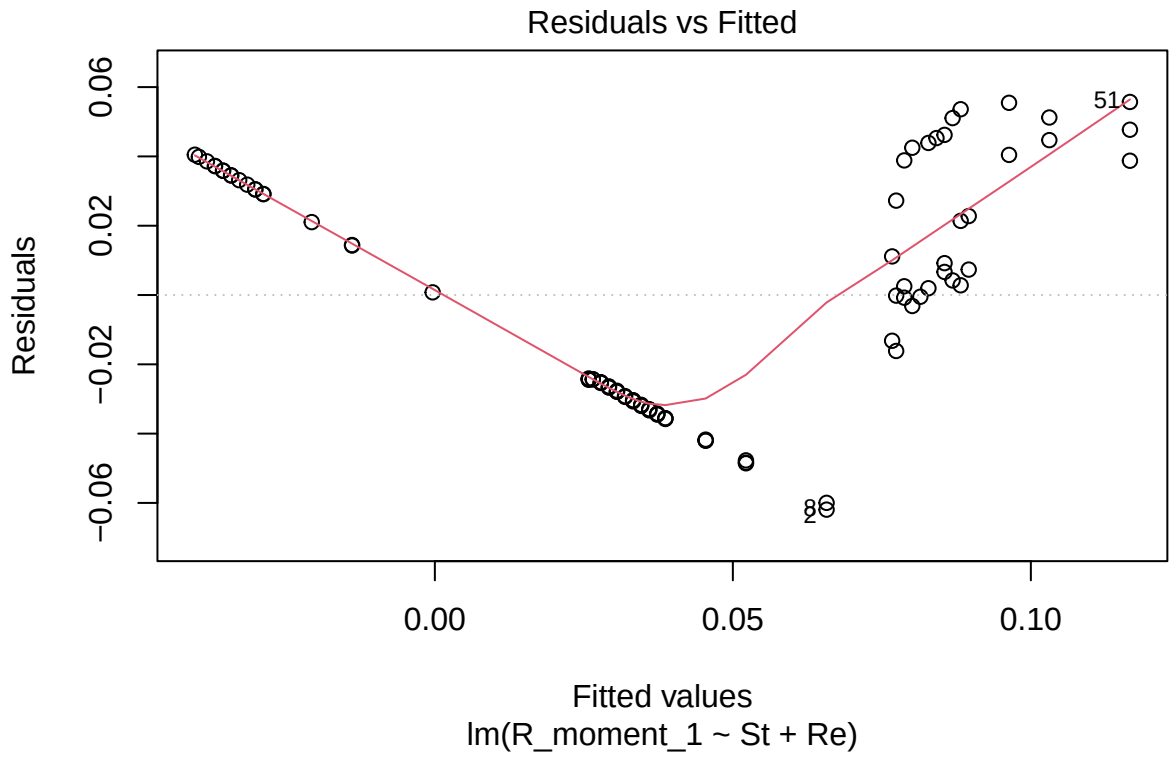
```
##
## Call:
## glm(formula = R_moment_4_central ~ St * TFr + St * Re + TFr *
##     Re, data = data_train)
##
## Deviance Residuals:
##        Min          1Q      Median          3Q         Max
## -3.694e+10  -7.457e+09  -1.392e+09   4.131e+09   4.499e+10
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.528e+10  5.349e+09   2.857 0.005418 **
## St           1.050e+10  4.256e+09   2.467 0.015707 *
## TFr         -1.915e+09  6.113e+08  -3.133 0.002401 **
## Re          -5.598e+07  2.293e+07  -2.442 0.016773 *
## St:TFr      -5.176e+08  3.144e+08  -1.646 0.103499
## St:Re       -2.662e+07  1.816e+07  -1.466 0.146598
## TFr:Re       7.303e+06  2.125e+06   3.438 0.000924 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 2.246926e+20)
##
##     Null deviance: 2.9409e+22  on 88  degrees of freedom
## Residual deviance: 1.8425e+22  on 82  degrees of freedom
## AIC: 4431.9
##
## Number of Fisher Scoring iterations: 2
```

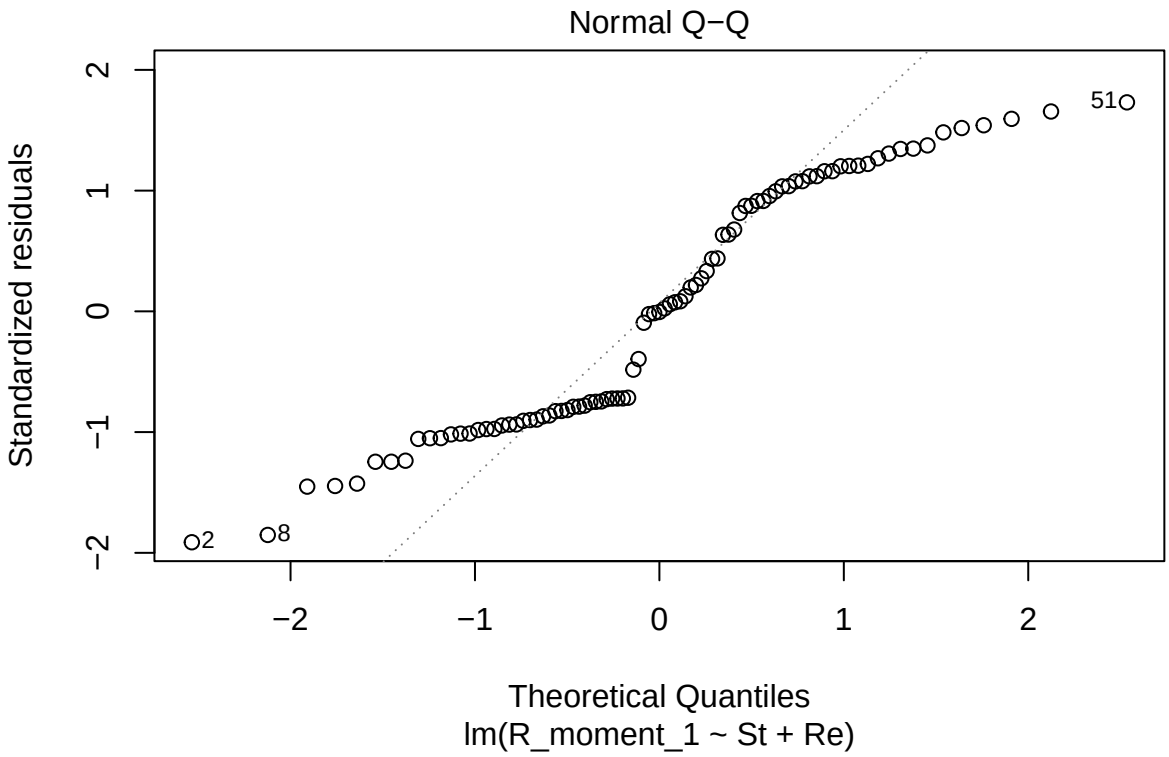
OR (by calling lm version of the functions)

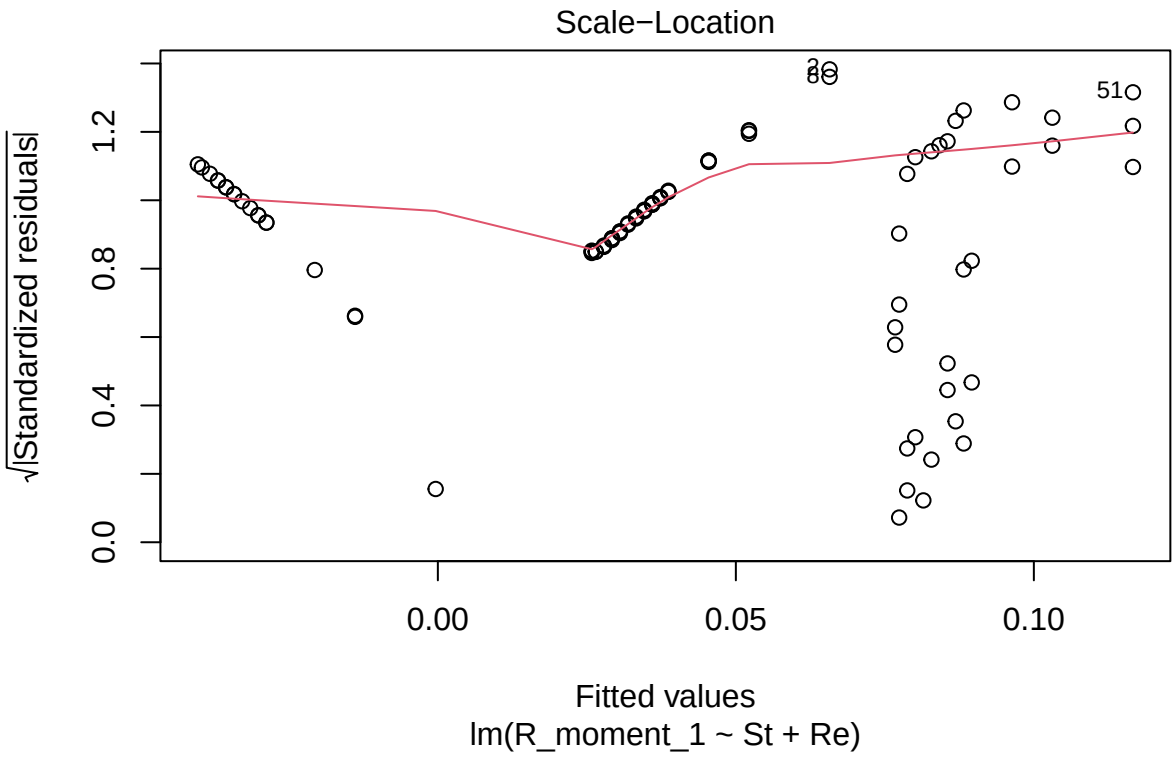
```
lm_fit_E1 <- lm(R_moment_1 ~ St + Re, data = data_train)
summary(lm_fit_E1)
```

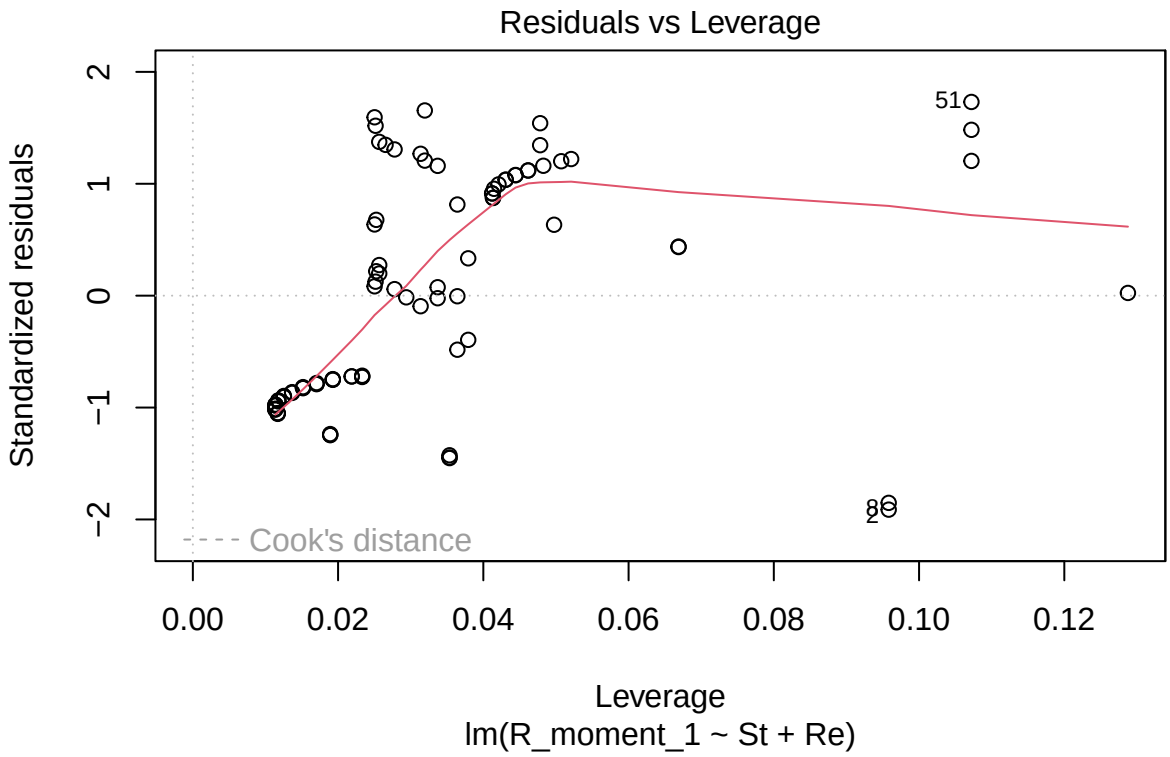
```
##
## Call:
## lm(formula = R_moment_1 ~ St + Re, data = data_train)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.061936 -0.030347 -0.000174  0.034491  0.055714
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.102e-01  8.837e-03  12.475  < 2e-16 ***
## St           1.353e-02  4.621e-03   2.927  0.00438 **
## Re          -3.798e-04  3.215e-05 -11.816  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.03406 on 86 degrees of freedom
## Multiple R-squared:  0.6364, Adjusted R-squared:  0.628
## F-statistic: 75.26 on 2 and 86 DF,  p-value: < 2.2e-16
```

```
plot(lm_fit_E1)
```




Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(R_moment_1 ~ St + Re)
2
8
51
−2
−1
0
1
2


Scale−Location
√|Standardized residuals|
0.0
0.4
0.8
1.2
0.00
0.05
0.10
Fitted values
lm(R_moment_1 ~ St + Re)
30
51




```
cve_linear_E1$delta
```

```
## [1] 0.001229868 0.001224073
```

Surprisingly, a 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. 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. 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.

```
lm_fit_E2 <- lm(R_moment_2_central ~ TFr + Re, data = data_train)
summary(lm_fit_E2)
```

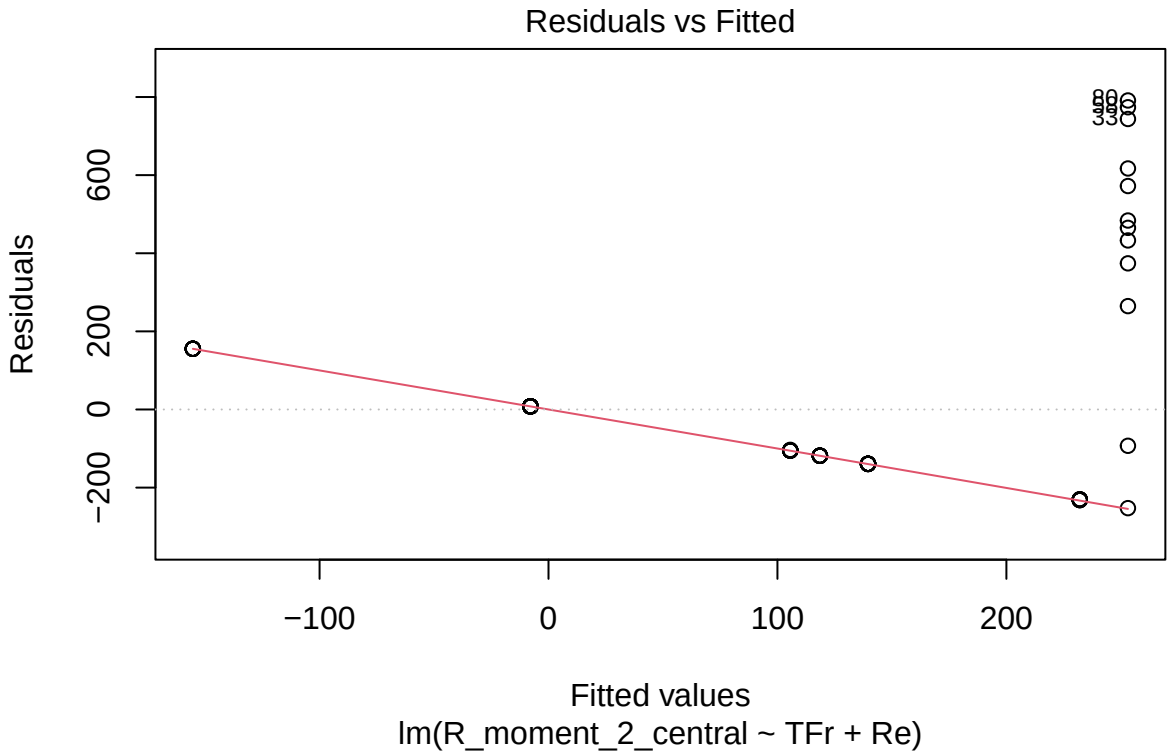
```
##
## Call:
## lm(formula = R_moment_2_central ~ TFr + Re, data = data_train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

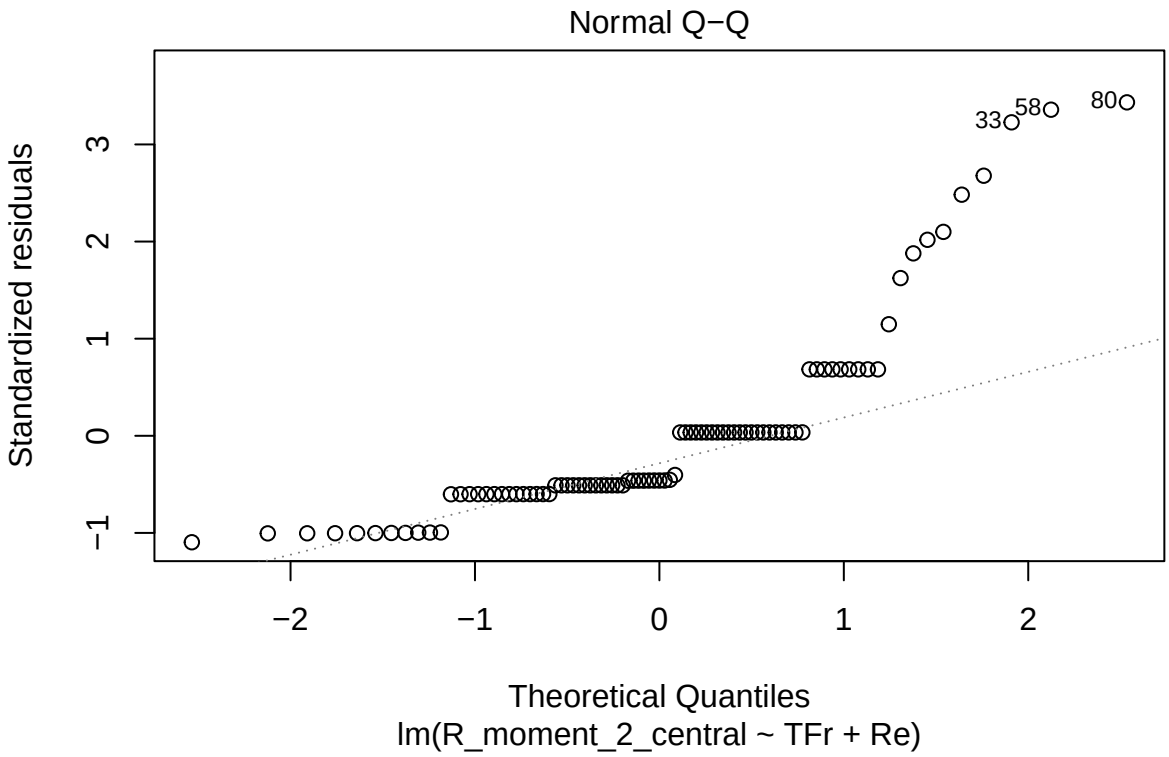
```
## -252.57 -139.16 -104.99    7.98  791.18
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 299.6445    53.6449   5.586 2.68e-07 ***
## TFr         -10.2315     3.8471  -2.660 0.009332 **
## Re           -0.8472     0.2221  -3.815 0.000256 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 234.1 on 86 degrees of freedom
## Multiple R-squared:  0.2189, Adjusted R-squared:  0.2007
## F-statistic: 12.05 on 2 and 86 DF,  p-value: 2.438e-05
```

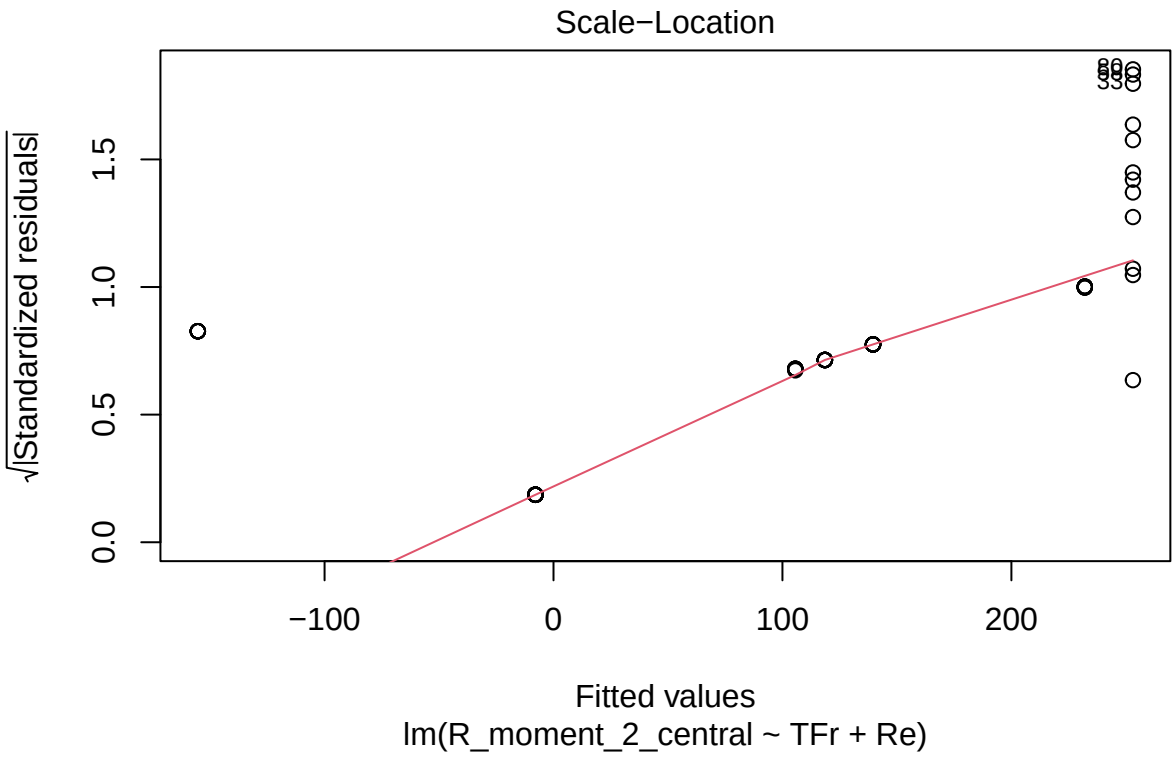
```
cve_linear_E2$delta
```

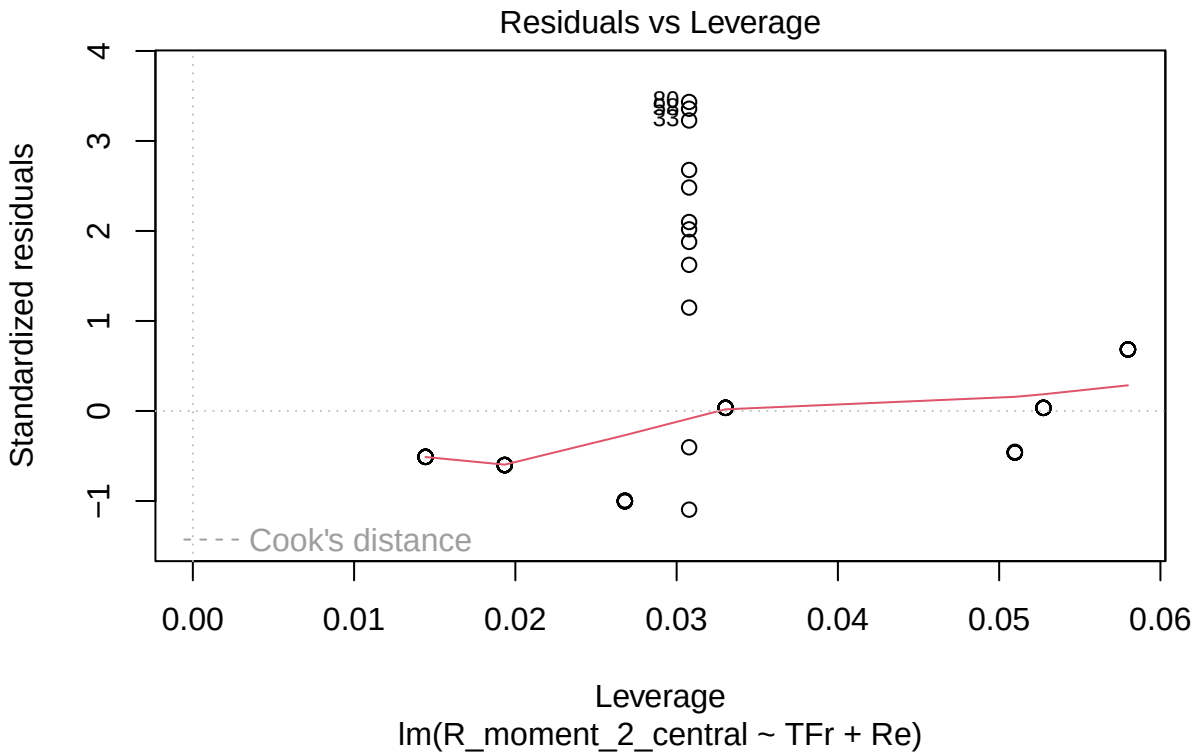
```
## [1] 57330.80 57097.37
```

```
plot(lm_fit_E2)
```




Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(R_moment_2_central ~ TFr + Re)
33
58
80


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
−100
0
100
200
Fitted values
lm(R_moment_2_central ~ TFr + Re)



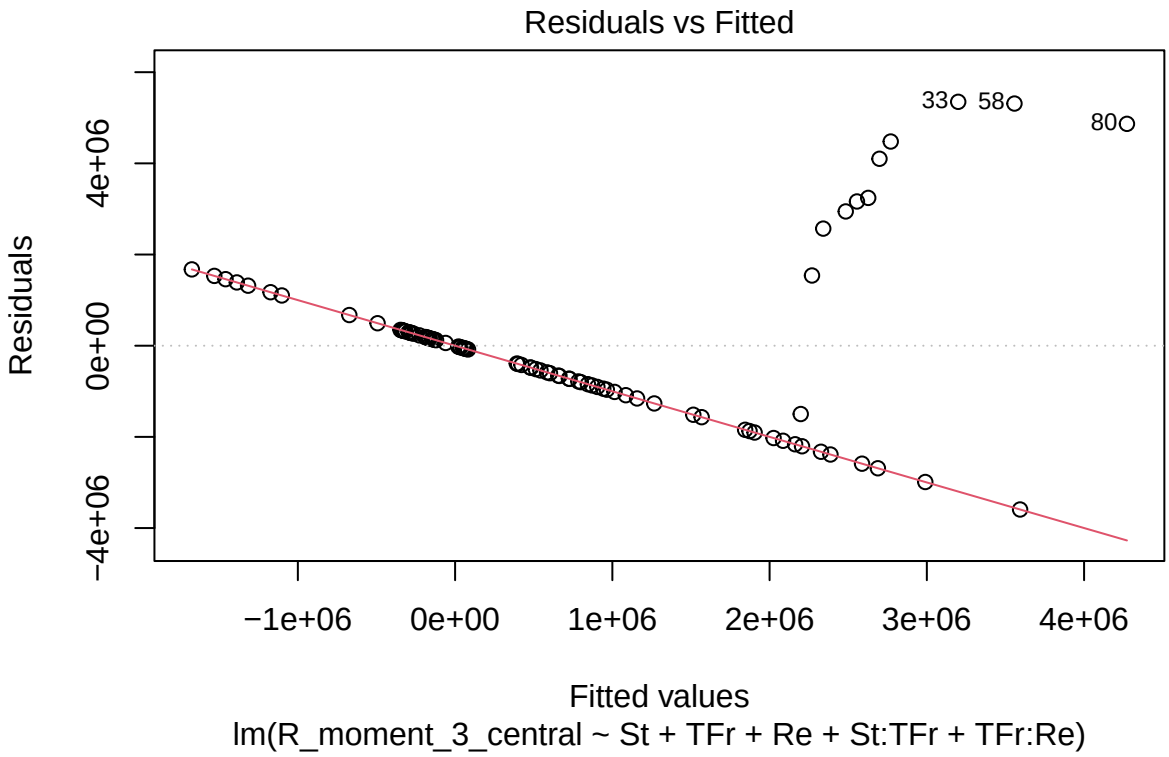
The R^2 is quite low at only about 20%. Generally, I believe this linear model is not very helpful. There is another clear pattern in the residuals plot and the linear fit consistently underestimates when the second moment is large. The truth is definitely closer to a nonlinear relationship.

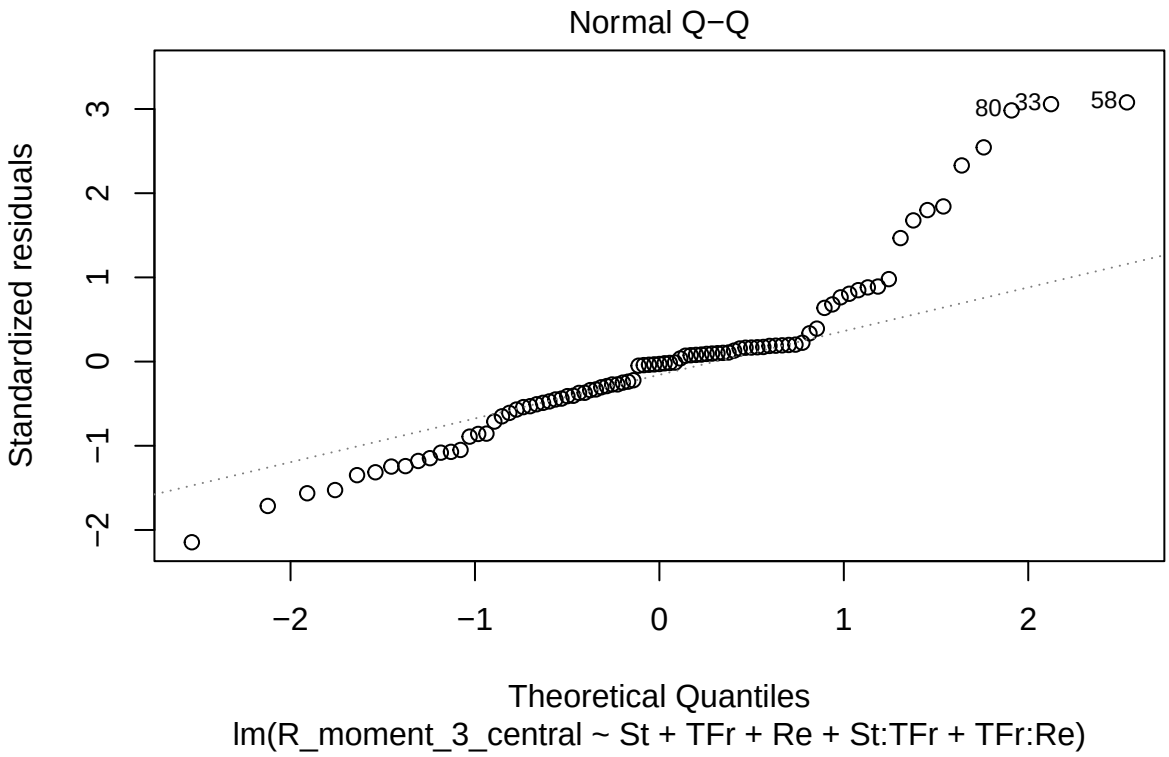
```
lm_fit_E3 <- lm(R_moment_3_central ~ St + TFr + Re + St:TFr + TFr:Re, data = data_train)
summary(lm_fit_E3)
```

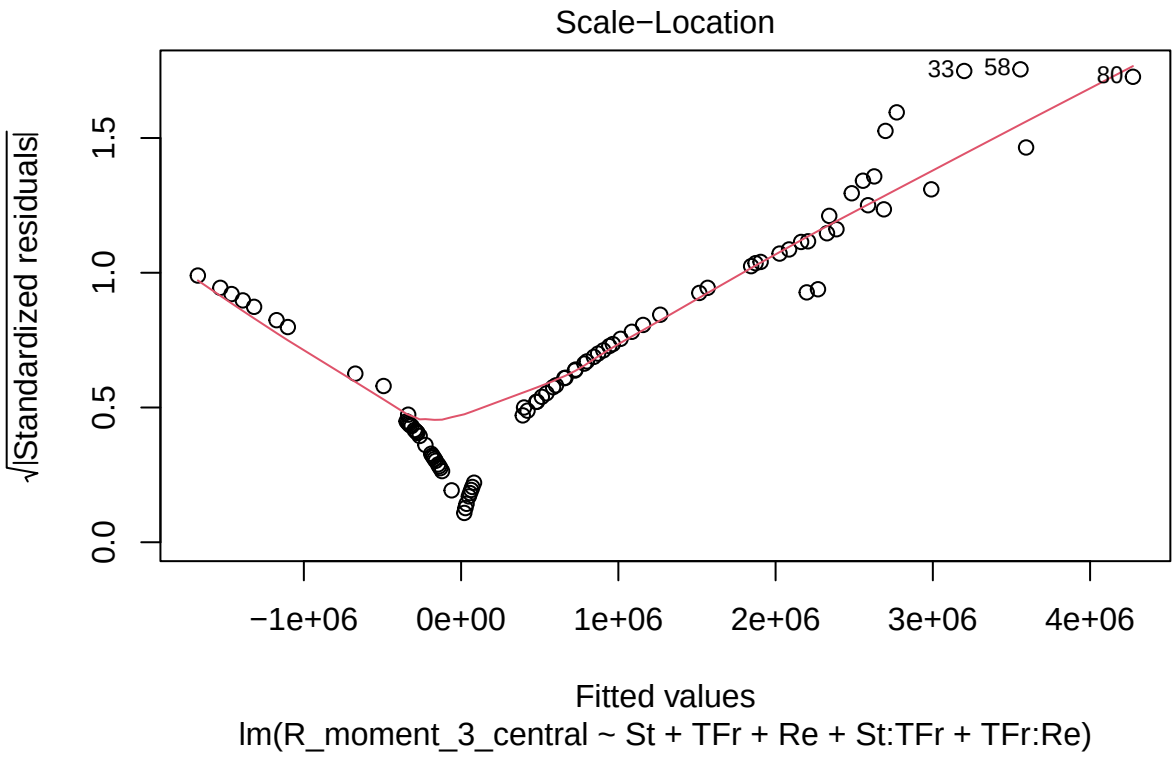
```
##
## Call:
## lm(formula = R_moment_3_central ~ St + TFr + Re + St:TFr + TFr:Re,
##     data = data_train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3592944  -904717   -49411   332389  5349989
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2525572.0   493187.5   5.121 1.94e-06 ***
## St           556593.3   258219.6   2.156 0.034018 *
## TFr         -252297.4    72716.5  -3.470 0.000829 ***
## Re            -9805.2     1864.1  -5.260 1.10e-06 ***
## St:TFr       -54689.6    37680.2  -1.451 0.150434
## TFr:Re          953.2      250.9   3.798 0.000276 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

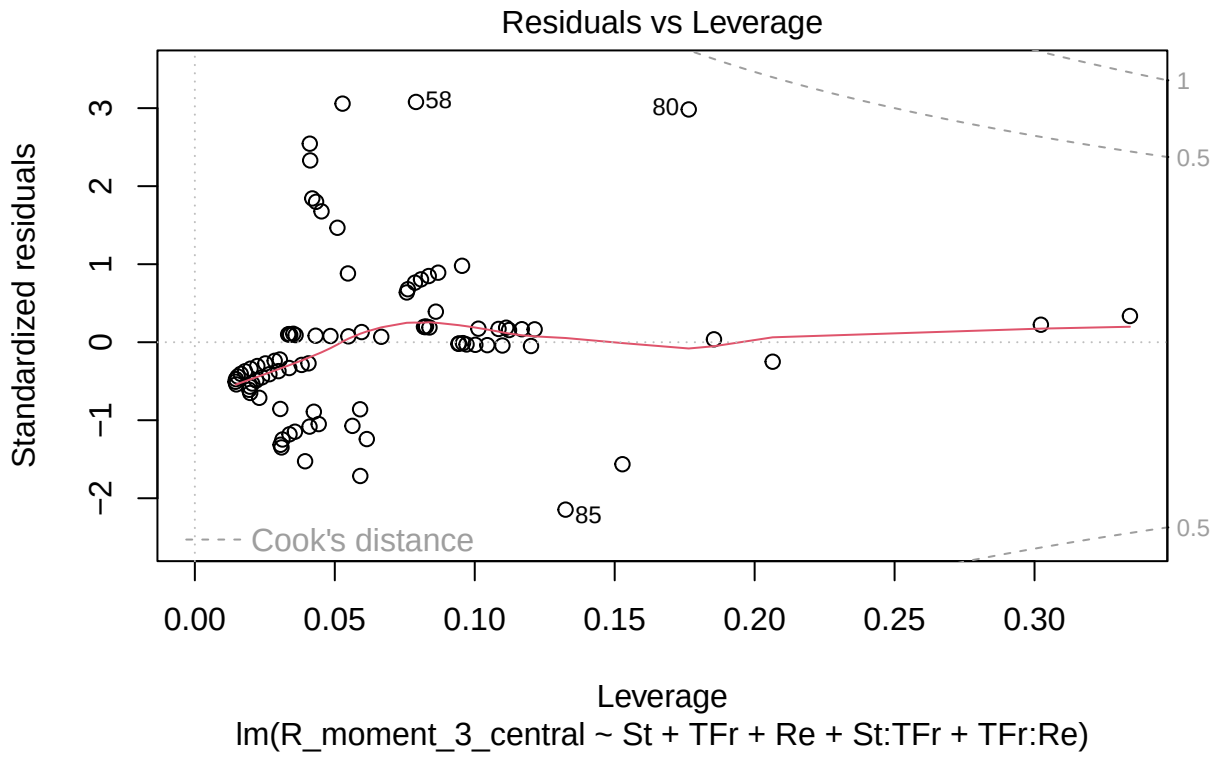
```
##
## Residual standard error: 1798000 on 83 degrees of freedom
## Multiple R-squared:  0.3603, Adjusted R-squared:  0.3218
## F-statistic:  9.35 on 5 and 83 DF,  p-value: 4.292e-07
```

```
plot(lm_fit_E3)
```




Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(R_moment_3_central ~ St + TFr + Re + St:TFr + TFr:Re)
80
33
58


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
33
58
80
−1e+06
0e+00
1e+06
2e+06
3e+06
4e+06
Fitted values
lm(R_moment_3_central ~ St + TFr + Re + St:TFr + TFr:Re)




```
cve_linear_interactions_E3$delta
```

```
## [1] 3.705014e+12 3.666260e+12
```

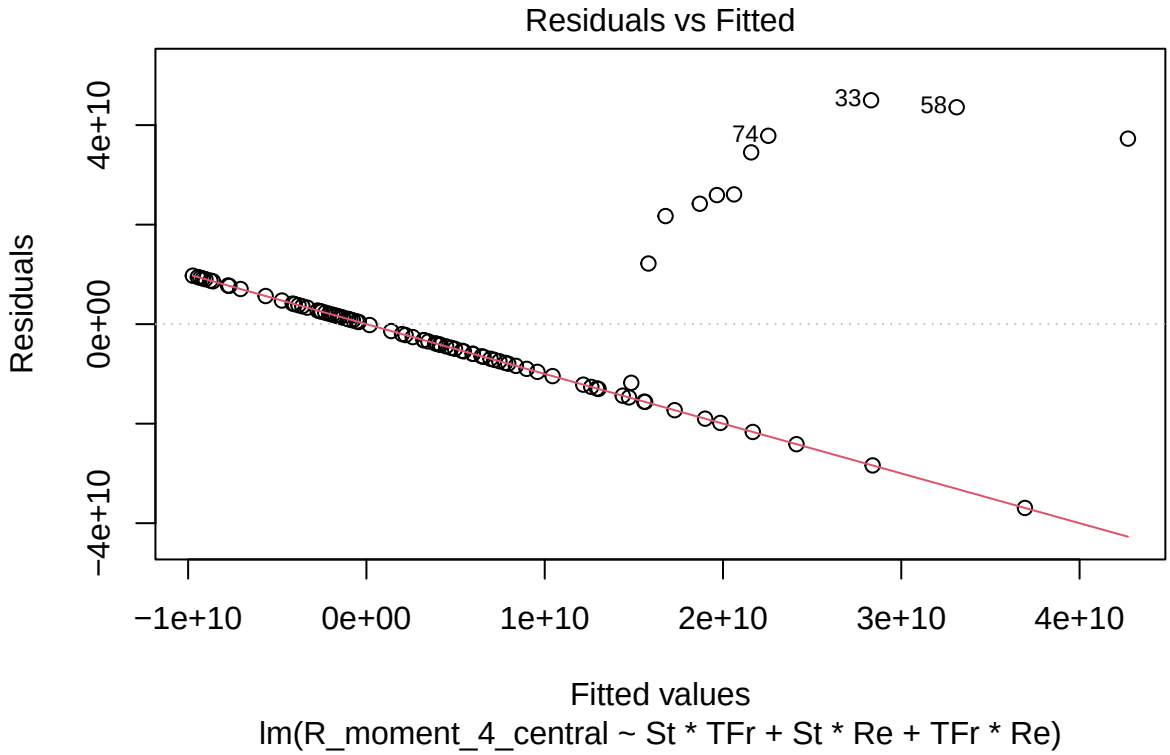
The R^2 is still not great at only about 32%. However, the interaction terms give important theoretical insights that are more in line with the limited theory that we know. The coefficient on TFr:Re is positive, which means that even though the coefficients on Re and TFr alone are negative, we can infer that at high enough levels of TFr, the effect of Re will actually be positive (since the interaction term means that for a given TFr, the coefficient on Re is (-9805.2 + 953.2 * TFr)). Similarly, the effect of TFr will be positive at high enough levels of Re. Thus, increasing rightward skewness of the probability density functions with Re or TFr seems to occur only for a combination of high values of TFr and Re. Otherwise, the main positive coefficient is St. Again, we see that the size of the particles has a particularly straightforward effect on turbulence. It increases the first moment and the right skewness of the PDF.

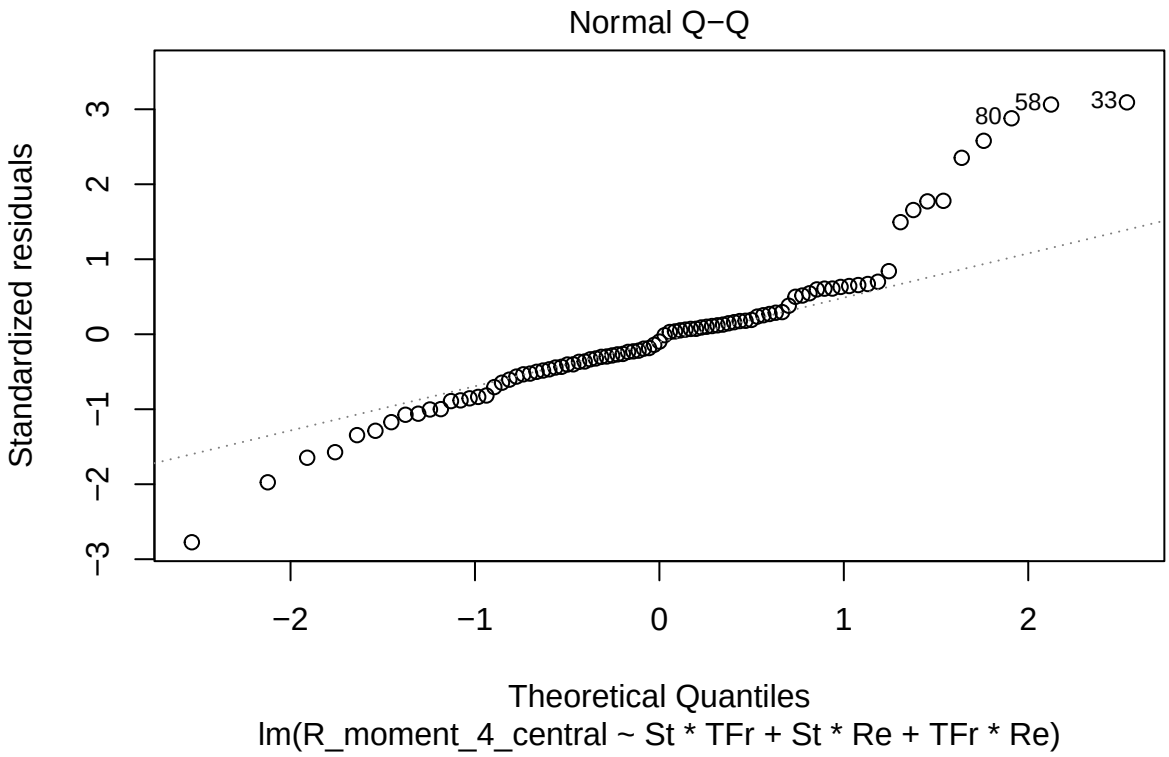
```
lm_fit_E4 <- lm(R_moment_4_central ~ St * TFr + St * Re + TFr * Re, data = data_train)
summary(lm_fit_E4)
```

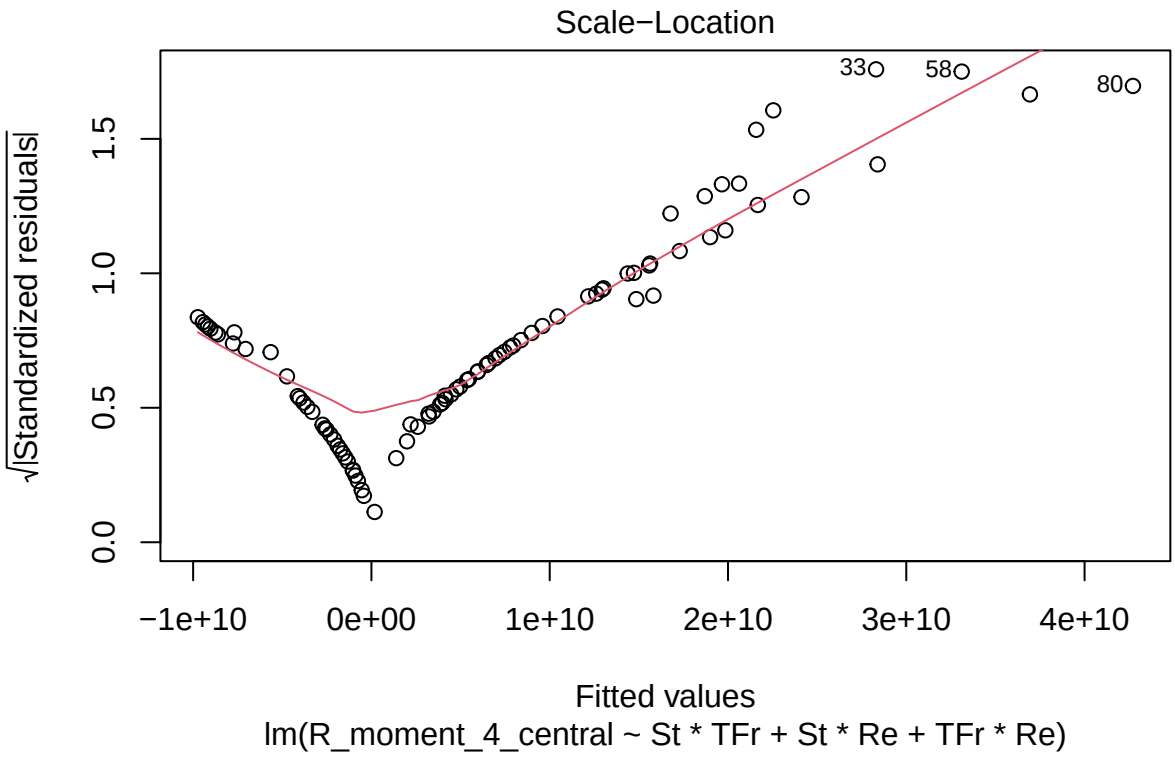
```
##
## Call:
## lm(formula = R_moment_4_central ~ St * TFr + St * Re + TFr *
##     Re, data = data_train)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
```

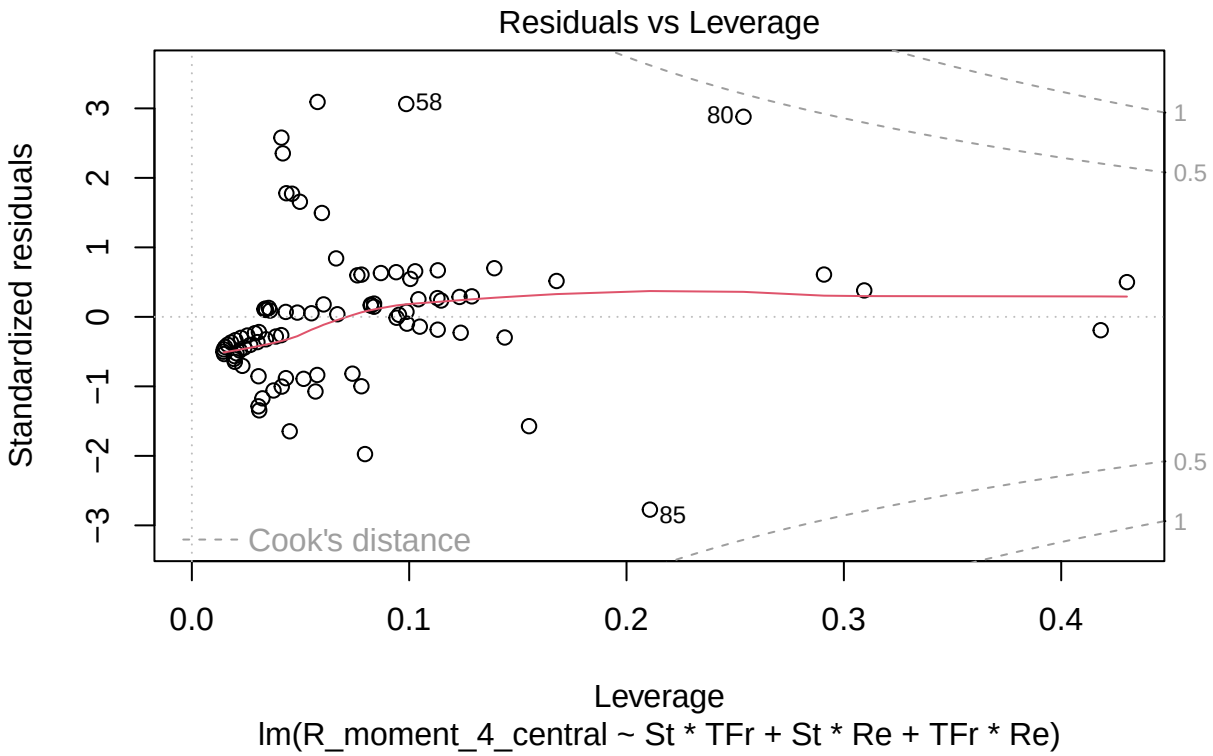
```
## -3.694e+10 -7.457e+09 -1.392e+09  4.131e+09  4.499e+10
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.528e+10  5.349e+09   2.857 0.005418 **
## St           1.050e+10  4.256e+09   2.467 0.015707 *
## TFr         -1.915e+09  6.113e+08  -3.133 0.002401 **
## Re          -5.598e+07  2.293e+07  -2.442 0.016773 *
## St:TFr      -5.176e+08  3.144e+08  -1.646 0.103499
## St:Re       -2.662e+07  1.816e+07  -1.466 0.146598
## TFr:Re       7.303e+06  2.125e+06   3.438 0.000924 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.499e+10 on 82 degrees of freedom
## Multiple R-squared:  0.3735, Adjusted R-squared:  0.3277
## F-statistic: 8.147 on 6 and 82 DF,  p-value: 6.439e-07
```

```
plot(lm_fit_E4)
```




Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(R_moment_4_central ~ St * TFr + St * Re + TFr * Re)
80
58
33


Scale−Location
33
58
80
√|Standardized residuals|
0.0
0.5
1.0
1.5
−1e+10
0e+00
1e+10
2e+10
3e+10
4e+10
Fitted values
lm(R_moment_4_central ~ St * TFr + St * Re + TFr * Re)



Our analysis of the best linear model for the fourth central moment is quite similar to that of the third central moment. St is the biggest positive driver of kurtosis in the PDF. TFr and Re individually are negative, but they have a positive interaction coefficient.

Complex Model

Best Degree Model

Model Evaluation?

Results (Final Models)

Discussion & Conclusion

References