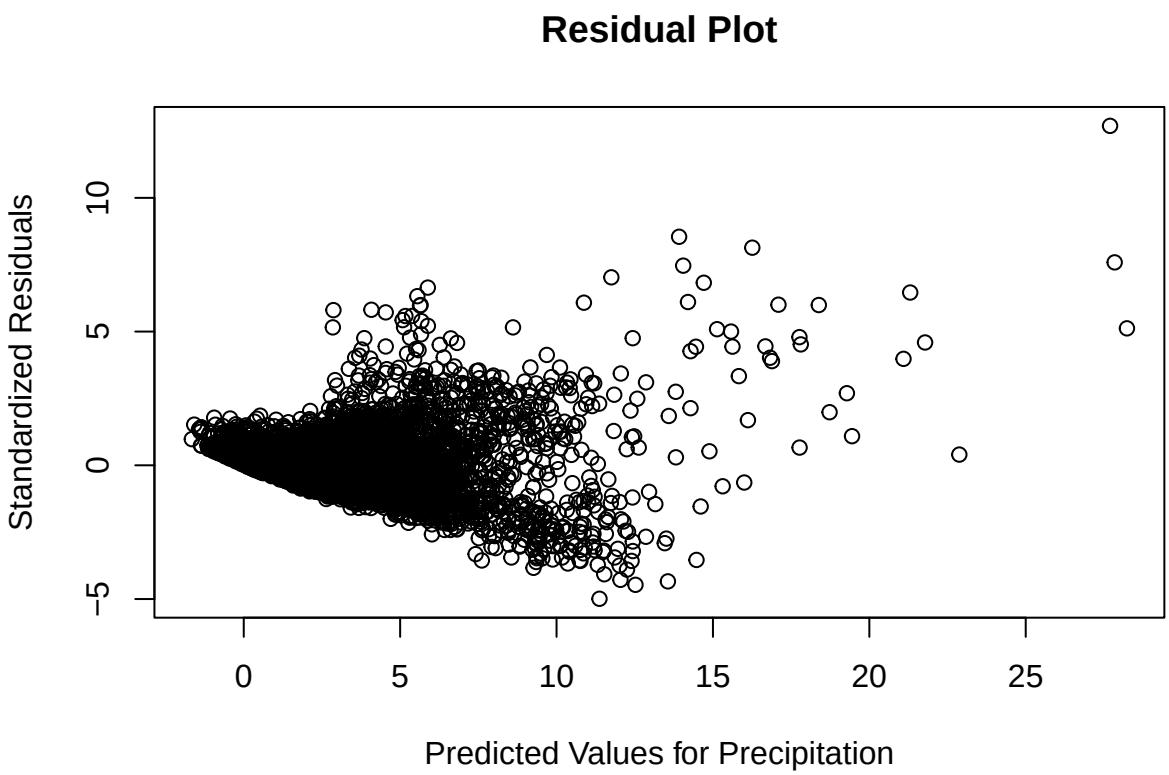
Weighted Least Squares Regression

unknownda

2023-03-26

Residual Analysis



- The residual plot is funnel shaped, hence indicating that heteroscedasticity is present in the data.

Heteroscedasticity Removal

1. Transforming the response variable

As a first measure, we transform the response variable by taking its logarithm.

```
##
## Call:
## lm(formula = log(Precipitation) ~ Year + Tmin + Tmax + I(Solar^-1) +
##     I(Wind^-1) + Relative_Humidity + Tmin:Tmax, data = df_cleaned)
##
## Residuals:
```

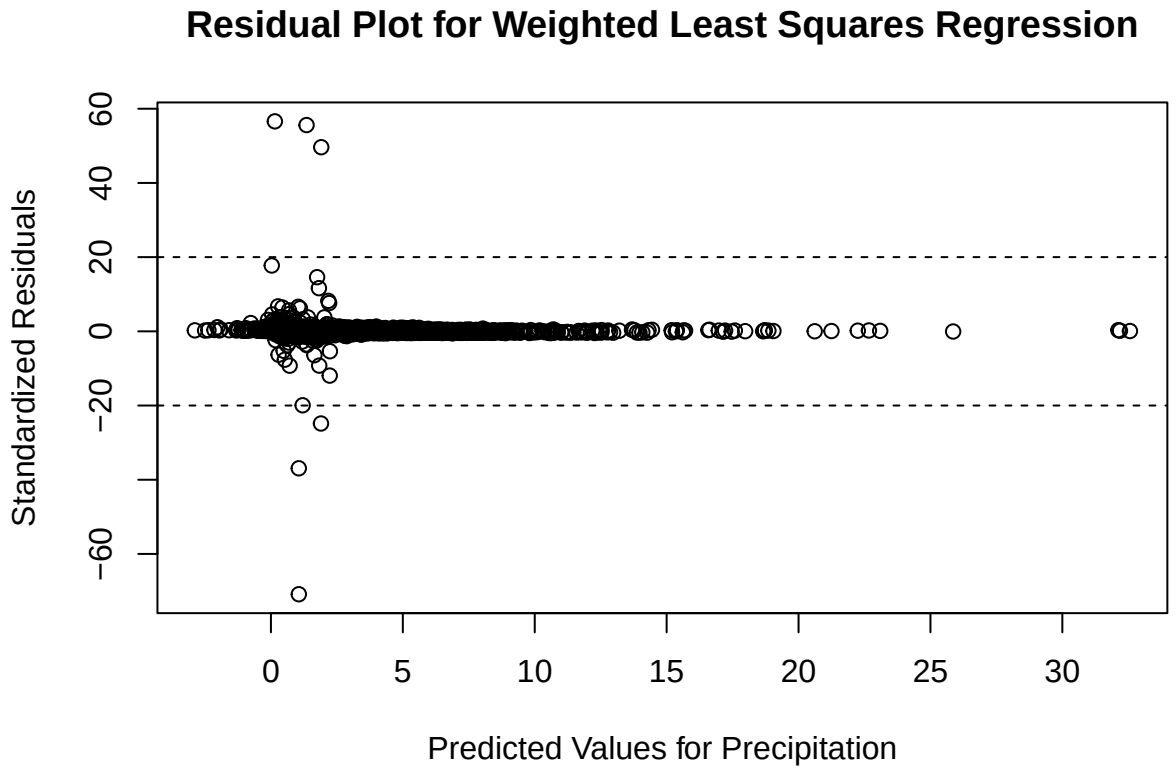
```
##     Min      1Q  Median      3Q     Max
## -5.0242 -0.2374  0.0690  0.3335  1.8563
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -1.019e+01  1.235e+00  -8.251  < 2e-16 ***
## Year               2.841e-03  6.088e-04   4.666 3.13e-06 ***
## Tmin               5.674e-03  5.192e-03   1.093    0.275
## Tmax               7.685e-03  5.853e-03   1.313    0.189
## I(Solar^-1)        5.420e+01  1.707e+00  31.749  < 2e-16 ***
## I(Wind^-1)         7.222e-01  7.363e-02   9.808  < 2e-16 ***
## Relative_Humidity  3.341e+00  9.562e-02  34.944  < 2e-16 ***
## Tmin:Tmax          3.636e-04  8.260e-05   4.403 1.08e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5488 on 7732 degrees of freedom
## Multiple R-squared:  0.5428, Adjusted R-squared:  0.5424
## F-statistic:  1311 on 7 and 7732 DF,  p-value: < 2.2e-16
```

- This measure was unsuccessful as the R-squared degraded further.

2. Weighted Least Squares Regression

```
##
## Call:
## lm(formula = Precipitation ~ Year + Tmin + Tmax + I(Solar^-1) +
##     I(Wind^-1) + Relative_Humidity + Tmin:Tmax, data = df_cleaned,
##     weights = w)
##
## Weighted Residuals:
##      Min       1Q   Median       3Q      Max
## -205.628   -1.165   -0.156    1.091  282.861
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -7.916e+01  2.314e+00 -34.214  < 2e-16 ***
## Year               3.036e-02  1.009e-03  30.079  < 2e-16 ***
## Tmin               3.418e-02  6.853e-03   4.987 6.25e-07 ***
## Tmax              -3.930e-02  8.026e-03  -4.896 9.96e-07 ***
## I(Solar^-1)        4.260e+02  8.238e+00  51.712  < 2e-16 ***
## I(Wind^-1)        -7.079e+00  2.174e-01 -32.564  < 2e-16 ***
## Relative_Humidity  4.428e+00  1.261e-01  35.127  < 2e-16 ***
## Tmin:Tmax          5.823e-04  1.069e-04   5.446 5.30e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.223 on 7732 degrees of freedom
## Multiple R-squared:  0.8593, Adjusted R-squared:  0.8591
## F-statistic:  6744 on 7 and 7732 DF,  p-value: < 2.2e-16
```

- R-squared value increased from 0.6065 to 0.8593 which is a huge improvement.
- In order to confirm the removal of heteroscedasticity, we need to analyze the residual plot again.




```
## integer(0)
```

- Barring a few outliers, the residual plot does not show any trend, hence implying removal of heteroscedasticity.

Removal of Outliers

1. First, we need to find the index of the data point which has the hisghest absolute residual.

```
which.max(abs(rstandard(wls)))
```

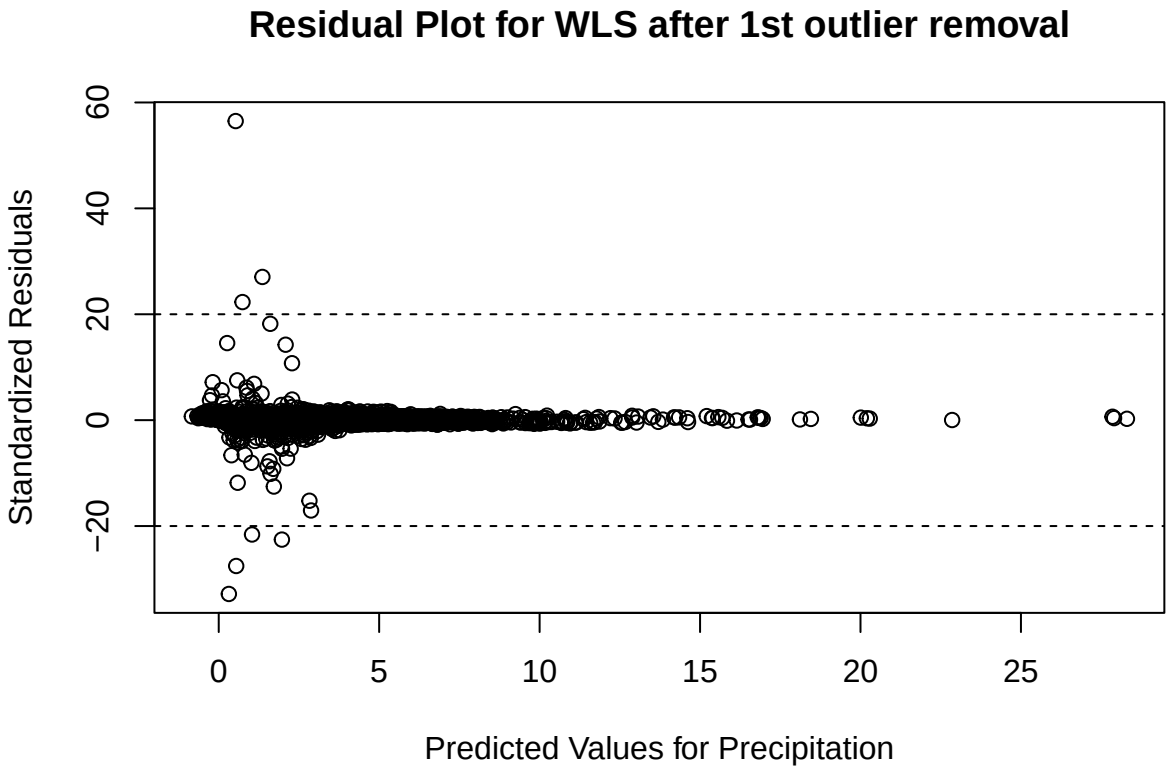
```
## 7332
## 7332
```

2. Next, we remove this specific data point from the dataset, retrain the model and plot the residual plot.

```
##
## Call:
## lm(formula = Precipitation ~ Year + Tmin + Tmax + I(Solar^-1) +
##     I(Wind^-1) + Relative_Humidity + Tmin:Tmax, data = df_o1,
##     weights = w_o1)
##
## Weighted Residuals:
##     Min      1Q  Median      3Q     Max
## -90.806  -1.035  -0.082   0.943 158.774
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -1.066e+02  1.403e+00  -75.97   <2e-16 ***
```

```
## Year               4.601e-02  6.160e-04   74.70   <2e-16 ***
## Tmin               5.962e-02  4.161e-03   14.33   <2e-16 ***
## Tmax              -5.407e-02  4.879e-03  -11.08   <2e-16 ***
## I(Solar^-1)        3.368e+02  5.087e+00   66.21   <2e-16 ***
## I(Wind^-1)        -2.988e+00  1.366e-01  -21.88   <2e-16 ***
## Relative_Humidity  4.540e+00  7.640e-02   59.42   <2e-16 ***
## Tmin:Tmax         -7.679e-04  6.689e-05  -11.48   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.821 on 7731 degrees of freedom
## Multiple R-squared:  0.9471, Adjusted R-squared:  0.947
## F-statistic: 1.976e+04 on 7 and 7731 DF,  p-value: < 2.2e-16
```

- Adjusted R-squared increases to 0.947 and Residual Standard Error decreases to 3.821



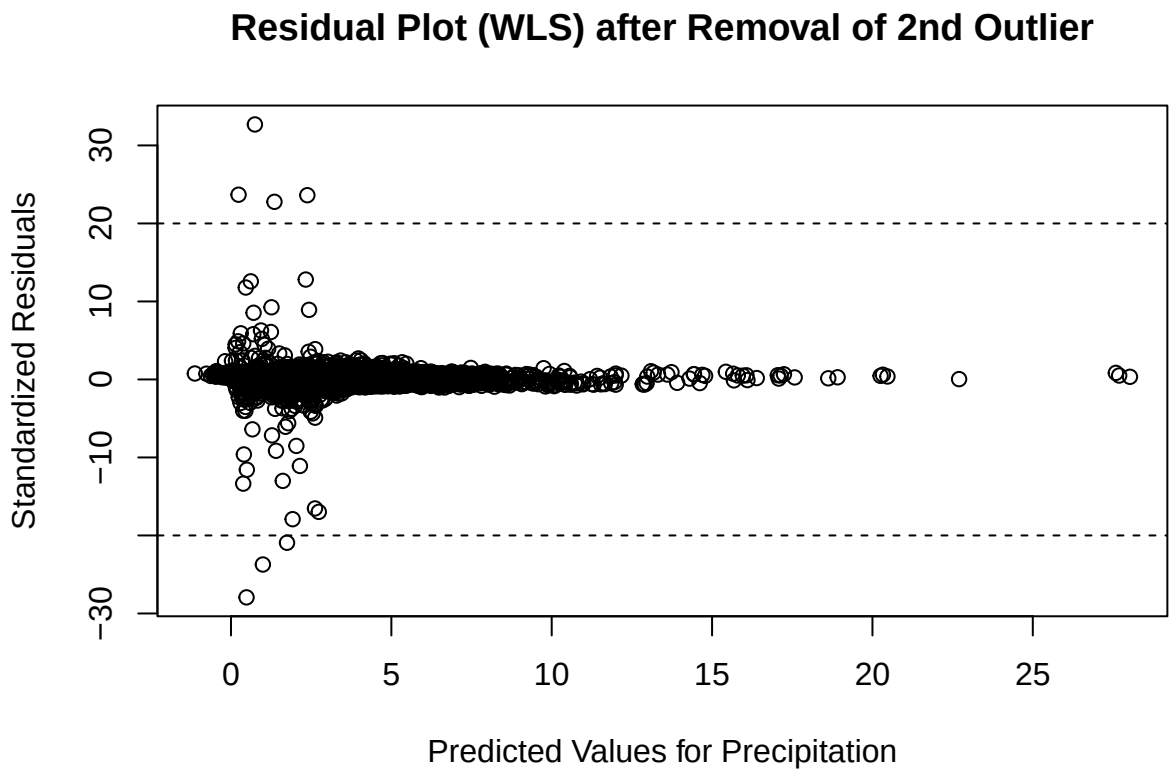

```
## integer(0)
```

3. Repeat Steps 1 and 2 till satisfactory results.

```
##
## Call:
## lm(formula = Precipitation ~ Year + Tmin + Tmax + I(Solar^-1) +
##     I(Wind^-1) + Relative_Humidity + Tmin:Tmax, data = df_o2,
##     weights = w_o2)
##
## Weighted Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -72.861  -0.956  -0.068   0.914  54.650
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -6.168e+01  1.234e+00 -49.970   <2e-16 ***
## Year               2.385e-02  5.581e-04  42.737   <2e-16 ***
## Tmin               9.401e-02  3.198e-03  29.396   <2e-16 ***
## Tmax              -8.858e-02  3.742e-03 -23.668   <2e-16 ***
## I(Solar^-1)        3.310e+02  3.880e+00  85.320   <2e-16 ***
## I(Wind^-1)        -1.037e+00  1.076e-01  -9.644   <2e-16 ***
## Relative_Humidity  3.770e+00  5.888e-02  64.030   <2e-16 ***
## Tmin:Tmax         -1.188e-03  5.203e-05 -22.840   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.948 on 7730 degrees of freedom
## Multiple R-squared:  0.9815, Adjusted R-squared:  0.9815
## F-statistic: 5.861e+04 on 7 and 7730 DF,  p-value: < 2.2e-16
```

- Adjusted R-squared increases to 0.9815 and Residual Standard Error decreases to 2.948.

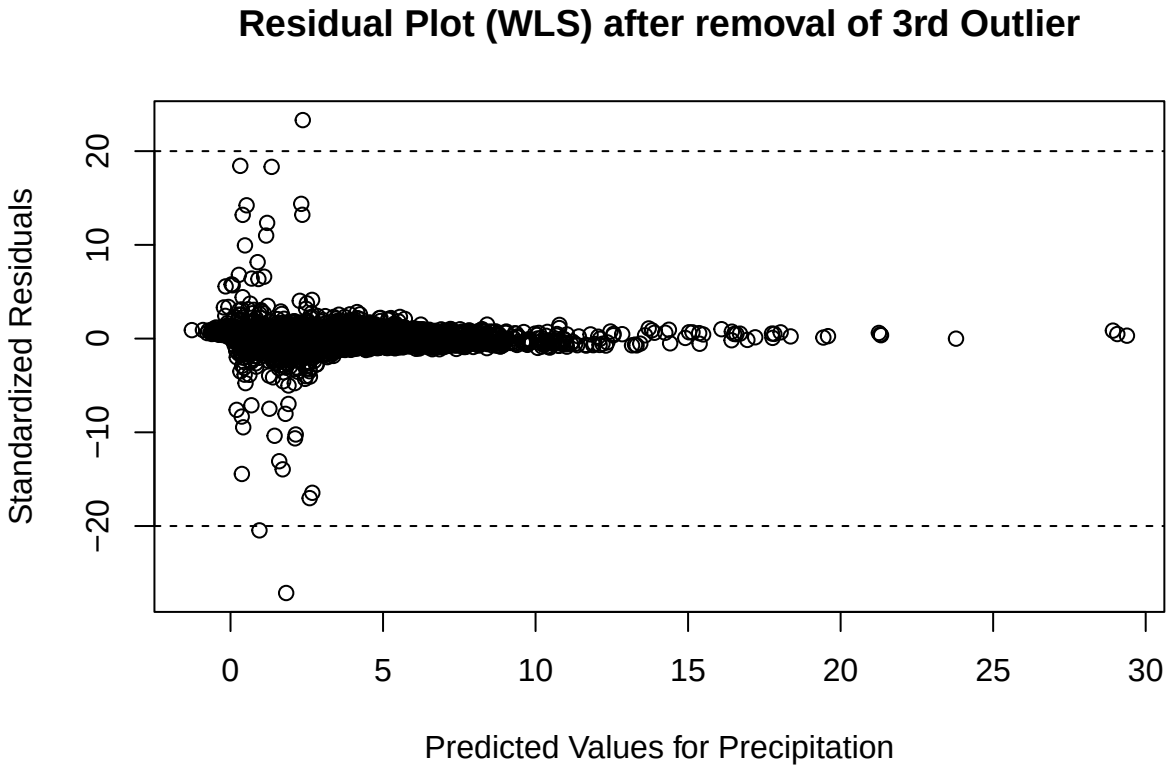



```
## integer(0)
```

```
##
## Call:
## lm(formula = Precipitation ~ Year + Tmin + Tmax + I(Solar^-1) +
##     I(Wind^-1) + Relative_Humidity + Tmin:Tmax, data = df_o3,
##     weights = w_o3)
```

```
##
## Weighted Residuals:
##     Min      1Q  Median      3Q     Max
## -71.744  -0.933  -0.038   0.919  44.418
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -6.781e+01  1.158e+00 -58.554  < 2e-16 ***
## Year               2.681e-02  5.248e-04  51.082  < 2e-16 ***
## Tmin               1.080e-01  3.009e-03  35.894  < 2e-16 ***
## Tmax              -1.233e-01  3.675e-03 -33.563  < 2e-16 ***
## I(Solar^-1)        3.471e+02  3.631e+00  95.607  < 2e-16 ***
## I(Wind^-1)         6.886e-01  1.137e-01   6.054 1.48e-09 ***
## Relative_Humidity  2.659e+00  6.399e-02  41.547  < 2e-16 ***
## Tmin:Tmax         -9.410e-04  4.930e-05 -19.088  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.76 on 7729 degrees of freedom
## Multiple R-squared:  0.9842, Adjusted R-squared:  0.9842
## F-statistic: 6.864e+04 on 7 and 7729 DF,  p-value: < 2.2e-16
```

- Adjusted R-squared increases to 0.9842 and Residual Standard Error reduces to 2.76.

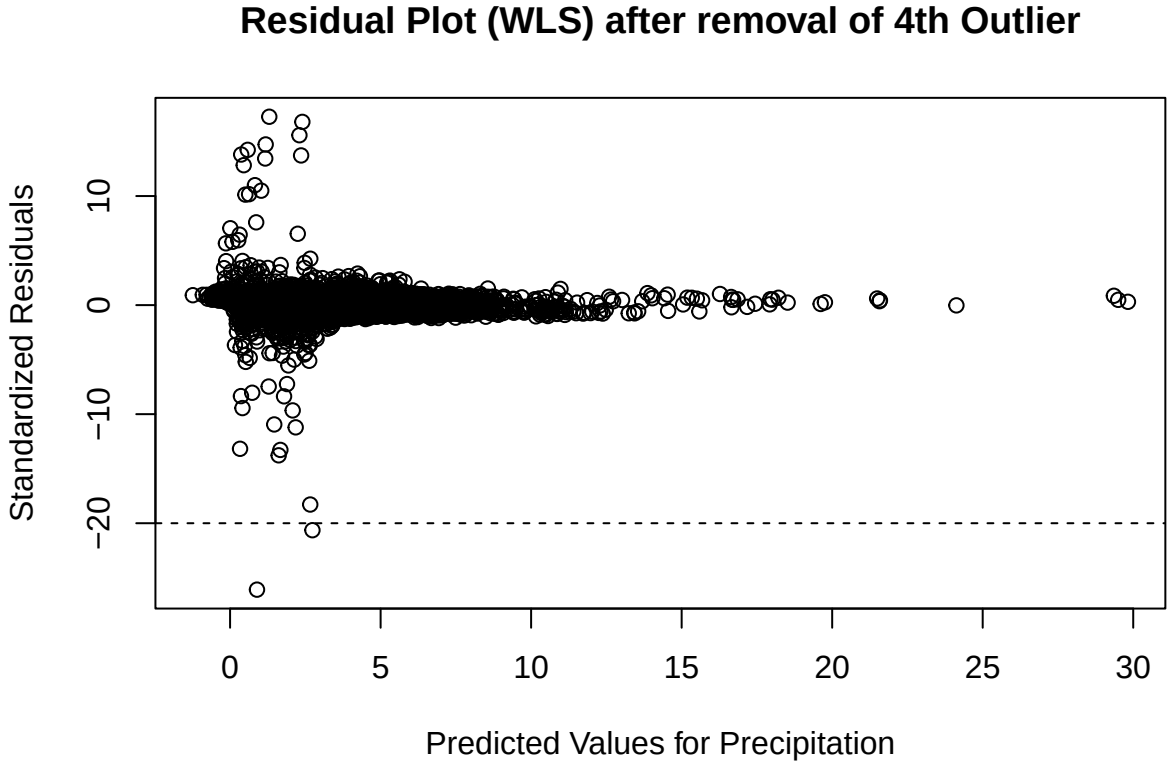



```
## integer(0)
```

```
##
## Call:
```

```
## lm(formula = Precipitation ~ Year + Tmin + Tmax + I(Solar^-1) +
##     I(Wind^-1) + Relative_Humidity + Tmin:Tmax, data = df_o4,
##     weights = w_o4)
##
## Weighted Residuals:
##     Min      1Q  Median      3Q     Max
## -50.309  -0.943  -0.050   0.906  40.789
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -7.467e+01  1.122e+00 -66.540  < 2e-16 ***
## Year               3.007e-02  5.098e-04  58.979  < 2e-16 ***
## Tmin               1.009e-01  2.873e-03  35.133  < 2e-16 ***
## Tmax              -1.166e-01  3.531e-03 -33.024  < 2e-16 ***
## I(Solar^-1)        3.529e+02  3.516e+00 100.376  < 2e-16 ***
## I(Wind^-1)         7.514e-01  1.098e-01   6.841 8.45e-12 ***
## Relative_Humidity  2.617e+00  6.249e-02  41.882  < 2e-16 ***
## Tmin:Tmax         -9.699e-04  4.822e-05 -20.112  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.682 on 7728 degrees of freedom
## Multiple R-squared:  0.9825, Adjusted R-squared:  0.9825
## F-statistic: 6.187e+04 on 7 and 7728 DF,  p-value: < 2.2e-16
```

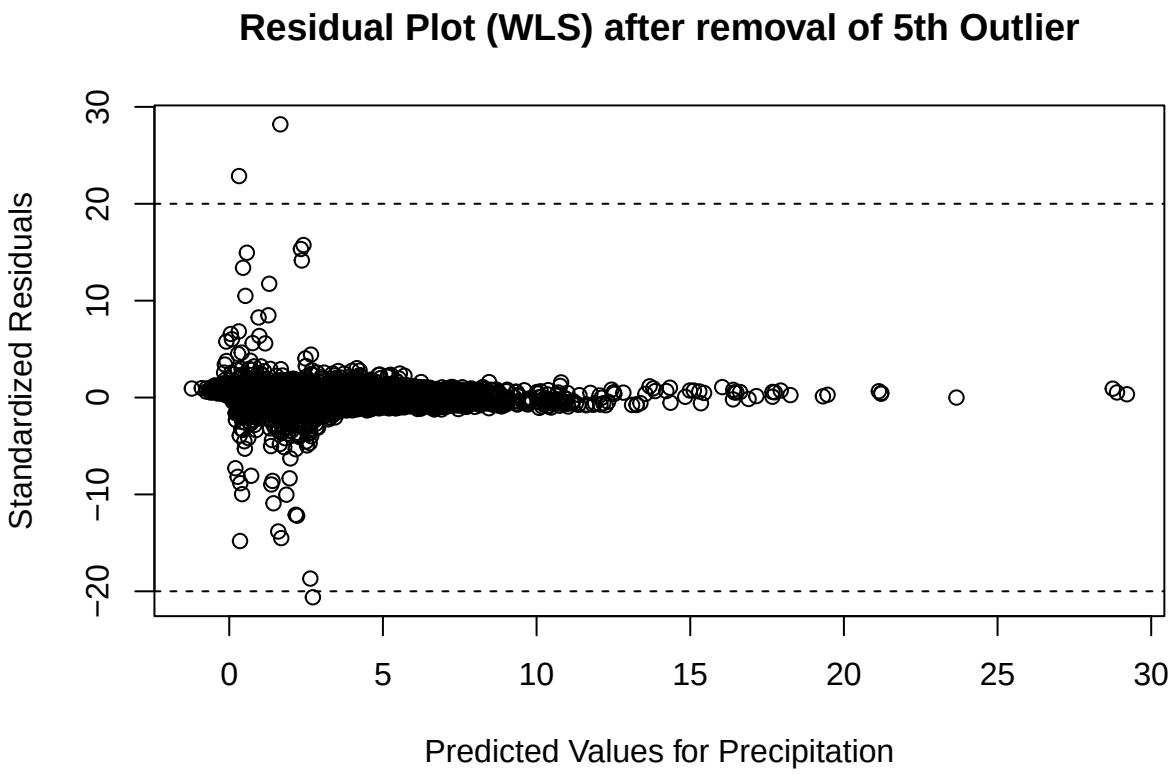
- Adjusted R-squared decreases to 0.9825 and Residual Standard Error reduces to 2.682.



```
## integer(0)
```

```
##
## Call:
## lm(formula = Precipitation ~ Year + Tmin + Tmax + I(Solar^-1) +
##     I(Wind^-1) + Relative_Humidity + Tmin:Tmax, data = df_o5,
##     weights = w_o5)
##
## Weighted Residuals:
##     Min      1Q  Median      3Q     Max
## -48.440  -0.943  -0.060   0.896  38.578
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -7.077e+01  1.068e+00 -66.290  < 2e-16 ***
## Year               2.836e-02  4.833e-04  58.685  < 2e-16 ***
## Tmin               1.073e-01  2.739e-03  39.160  < 2e-16 ***
## Tmax              -1.228e-01  3.390e-03 -36.228  < 2e-16 ***
## I(Solar^-1)        3.430e+02  3.395e+00 101.029  < 2e-16 ***
## I(Wind^-1)         6.958e-01  1.052e-01   6.617 3.91e-11 ***
## Relative_Humidity  2.813e+00  6.064e-02  46.385  < 2e-16 ***
## Tmin:Tmax         -9.492e-04  4.650e-05 -20.413  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.576 on 7727 degrees of freedom
## Multiple R-squared:  0.9797, Adjusted R-squared:  0.9797
## F-statistic: 5.339e+04 on 7 and 7727 DF,  p-value: < 2.2e-16
```

- Adjusted R-squared drops to 0.9797 and Residual Standard Error reduces to 2.576.

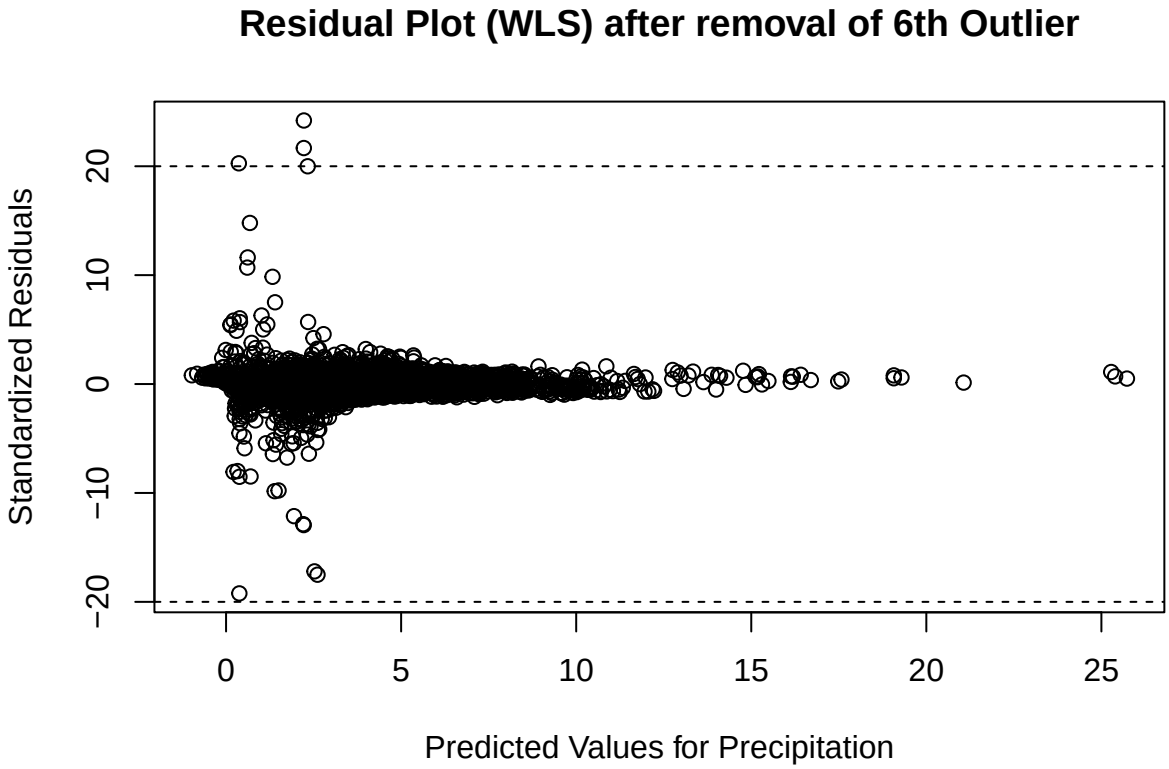



```
## integer(0)
```

```
##
## Call:
## lm(formula = Precipitation ~ Year + Tmin + Tmax + I(Solar^-1) +
##     I(Wind^-1) + Relative_Humidity + Tmin:Tmax, data = df_o6,
##     weights = w_o6)
##
## Weighted Residuals:
##     Min      1Q  Median      3Q     Max
## -44.092  -0.879  -0.059   0.847  55.267
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -6.262e+01  1.030e+00  -60.78   <2e-16 ***
## Year               2.519e-02  4.646e-04   54.22   <2e-16 ***
## Tmin               1.137e-01  2.626e-03   43.30   <2e-16 ***
## Tmax              -1.161e-01  3.274e-03  -35.46   <2e-16 ***
## I(Solar^-1)        2.837e+02  3.807e+00   74.50   <2e-16 ***
## I(Wind^-1)         2.196e+00  1.127e-01   19.49   <2e-16 ***
## Relative_Humidity  3.733e+00  6.718e-02   55.56   <2e-16 ***
## Tmin:Tmax         -1.118e-03  4.494e-05  -24.87   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.465 on 7726 degrees of freedom
## Multiple R-squared:  0.9457, Adjusted R-squared:  0.9457
```

```
## F-statistic: 1.923e+04 on 7 and 7726 DF,  p-value: < 2.2e-16
```

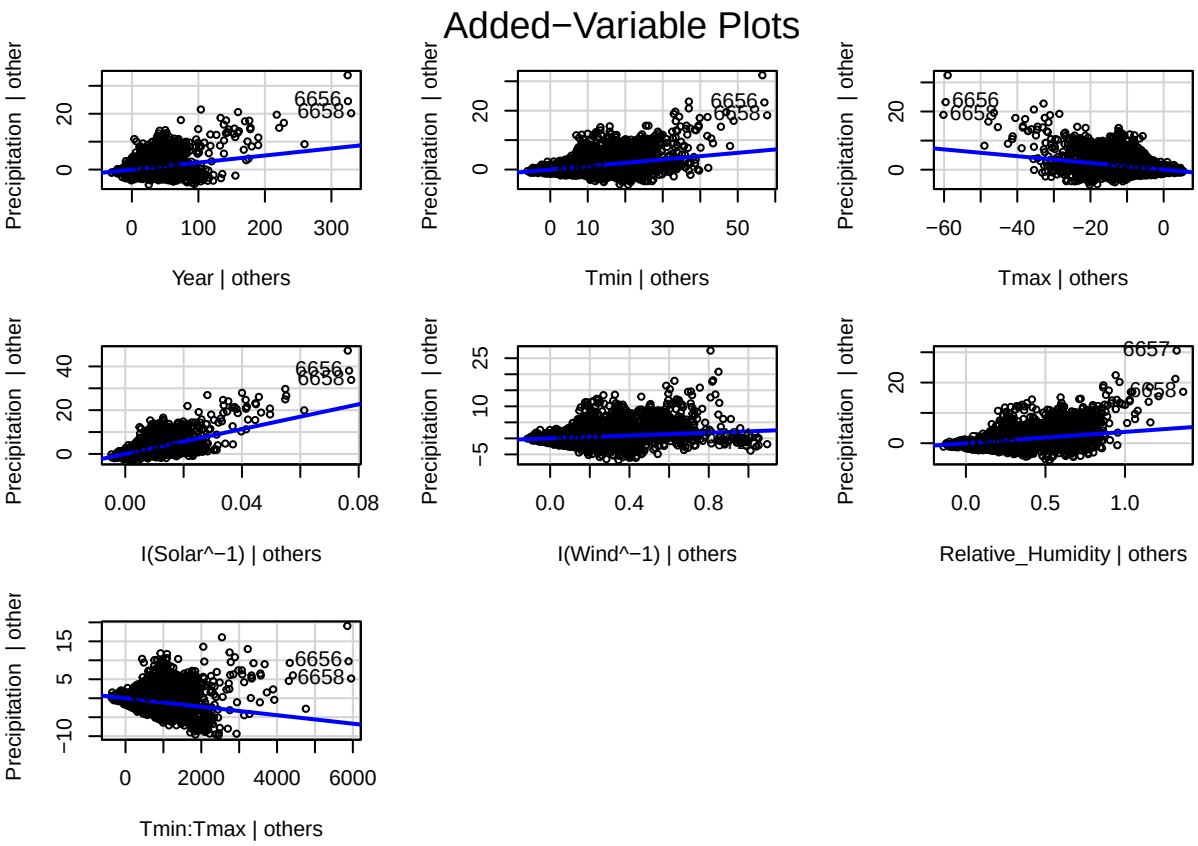
- Adjusted R-squared drops to 0.9457 and Residual Standard Error reduces to 2.465




```
## integer(0)
```

- On further iterations R-squared drops below acceptable range, so we stop here.

The final model plots are shown below:


Added−Variable Plots
Precipitation | others
Year | others
Tmin | others
Tmax | others
I(Solar^−1) | others
I(Wind^−1) | others
Relative_Humidity | others
Tmin:Tmax | others
6656
6658
6657