QSAR Fish Toxicity Data Set] (https://archive.ics.uci.edu/ml/datasets/QSAR+fish+toxic (https://archive.ics.uci.edu/ml/datasets/QSAR+fish+toxic

Data

This dataset was used to develop quantitative regression QSAR models to predict acute aquatic toxicity towards the fish Pimephales promelas (fathead minnow) on a set of 908 chemicals. LC50 data, which is the concentration that causes death in 50% of test fish over a test duration of 96 hours, was used as model response. The model comprised 6 molecular descriptors: MLOGP (molecular properties), CIC0 (information indices), GATS1i (2D autocorrelations), NdssC (atom-type counts), NdsCH ((atom-type counts), SM1_Dz(Z) (2D matrix-based descriptors). Details can be found in the quoted reference: M. Cassotti, D. Ballabio, R. Todeschini, V. Consonni. A similarity-based QSAR model for predicting acute toxicity towards the fathead minnow (Pimephales promelas), SAR and QSAR in Environmental Research (2015), 26, 217-243; doi: 10.1080/1062936X.2015.1018938

Attribute Information

6 molecular descriptors and 1 quantitative experimental response:

1. CIC0
2. SM1_Dz(Z)
3. GATS1i
4. NdsCH
5. NdssC
6. MLOGP
7. quantitative response, LC50 [-LOG(mol/L)]

Datatable

Raw Data | Normalized Data | Statistics Summary of Normalized Data

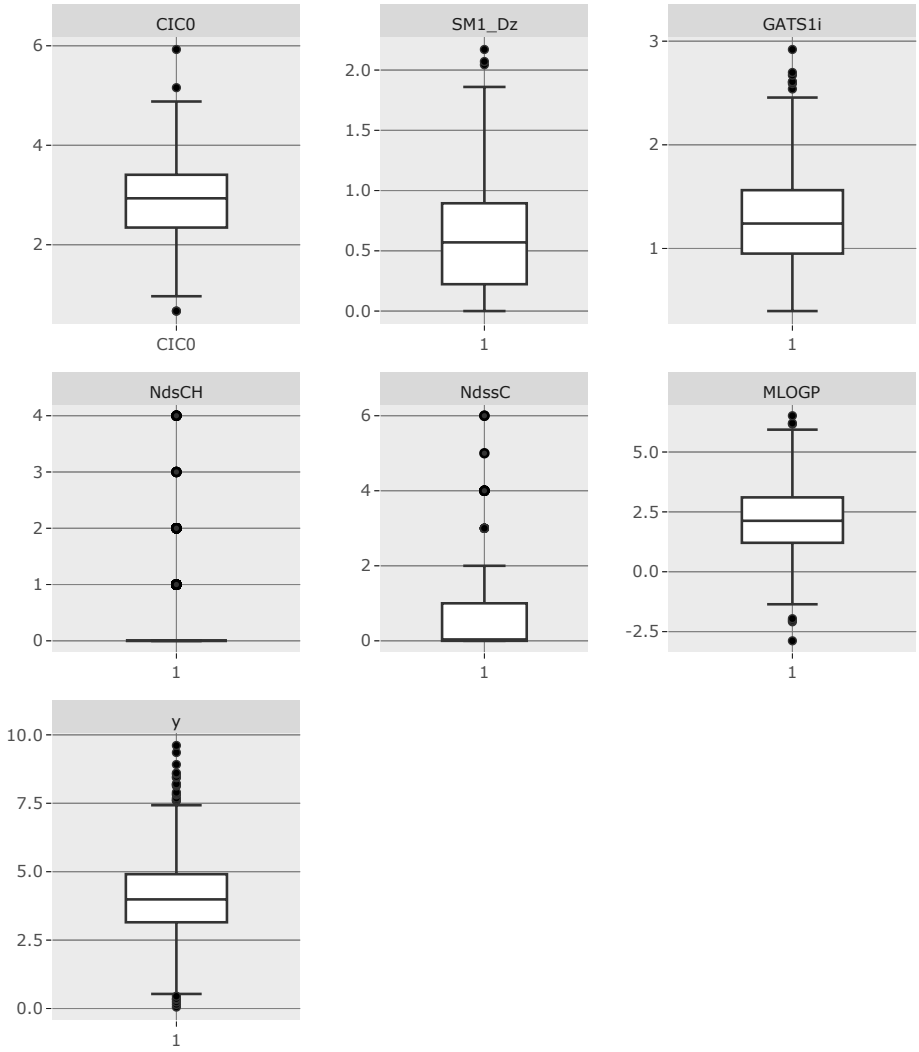
Show 10 **entries** **Search:**

Data

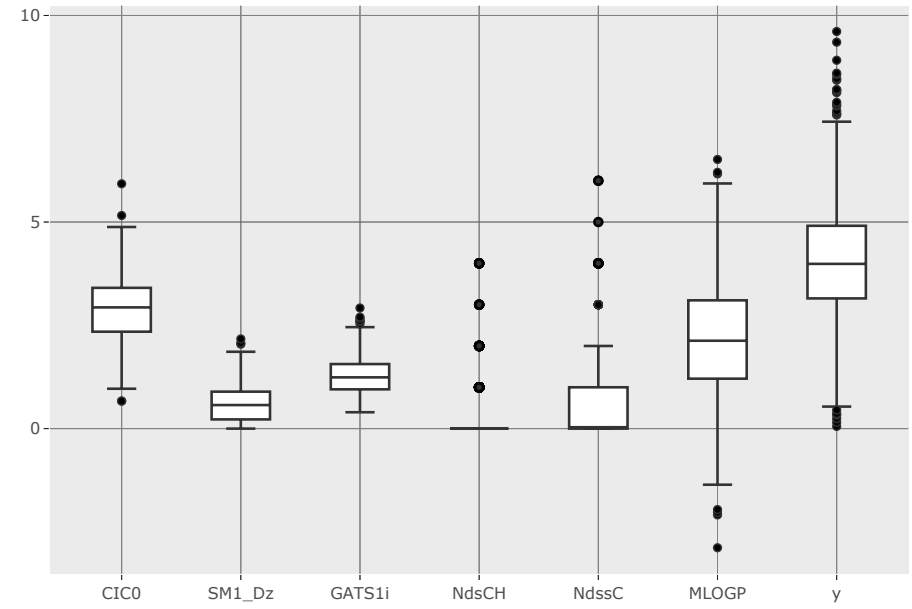
	CIC0	SM1_Dz	GATS1i	NdsCH	NdssC	MLOGP	y
1	3.26	0.829	1.676	0	1	1.453	3.77
2	2.189	0.58	0.863	0	0	1.348	3.115
3	2.125	0.638	0.831	0	0	1.348	3.531
4	3.027	0.331	1.472	1	0	1.807	3.51
5	2.094	0.827	0.86	0	0	1.886	5.39
6	3.222	0.331	2.177	0	0	0.706	1.819
7	3.179	0	1.063	0	0	2.942	3.947
8	3	0	0.938	1	0	2.851	3.513
9	2.62	0.499	0.99	0	0	2.942	4.402
10	2.834	0.134	0.95	0	0	1.591	3.021

Showing 1 to 10 of 908 entries Previous 1 2 3 4 5 … 91 Next

Boxplots

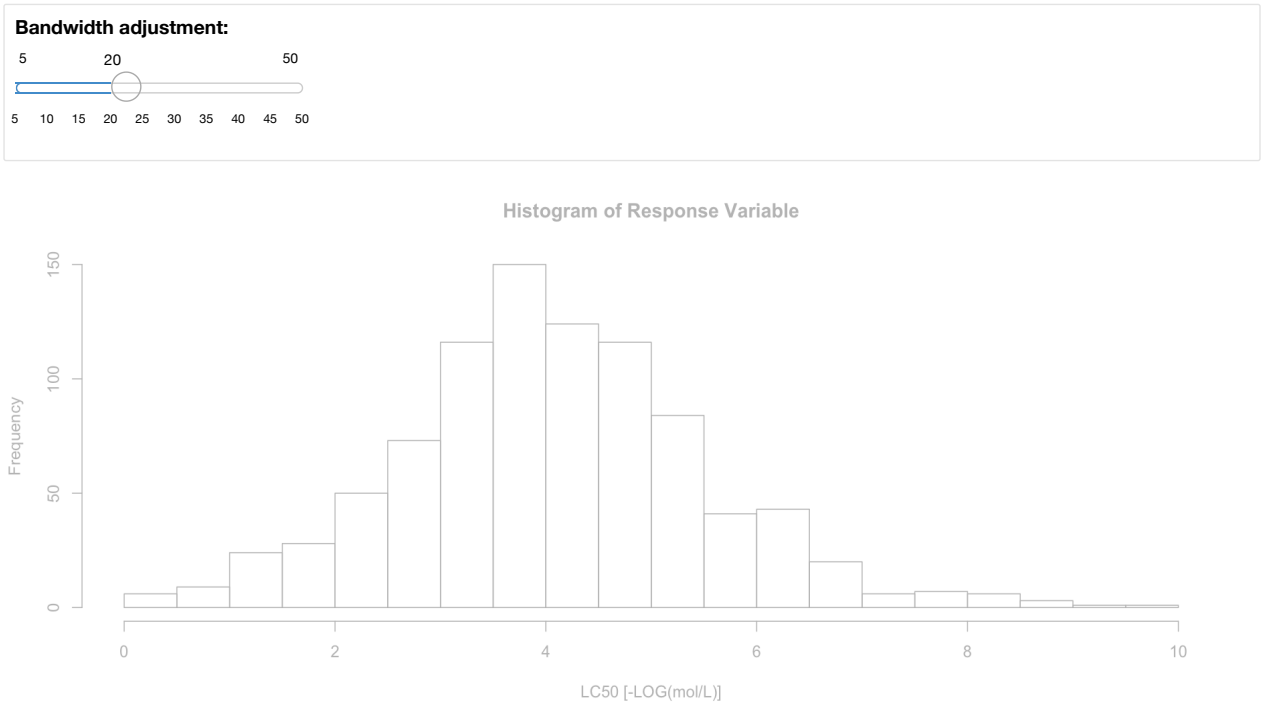



```
p <- ggplot(data = meltdt, aes(factor(variable),value)) + geom_boxplot() + xlab("") + ylab("")
p %>% ggplotly(.)
```

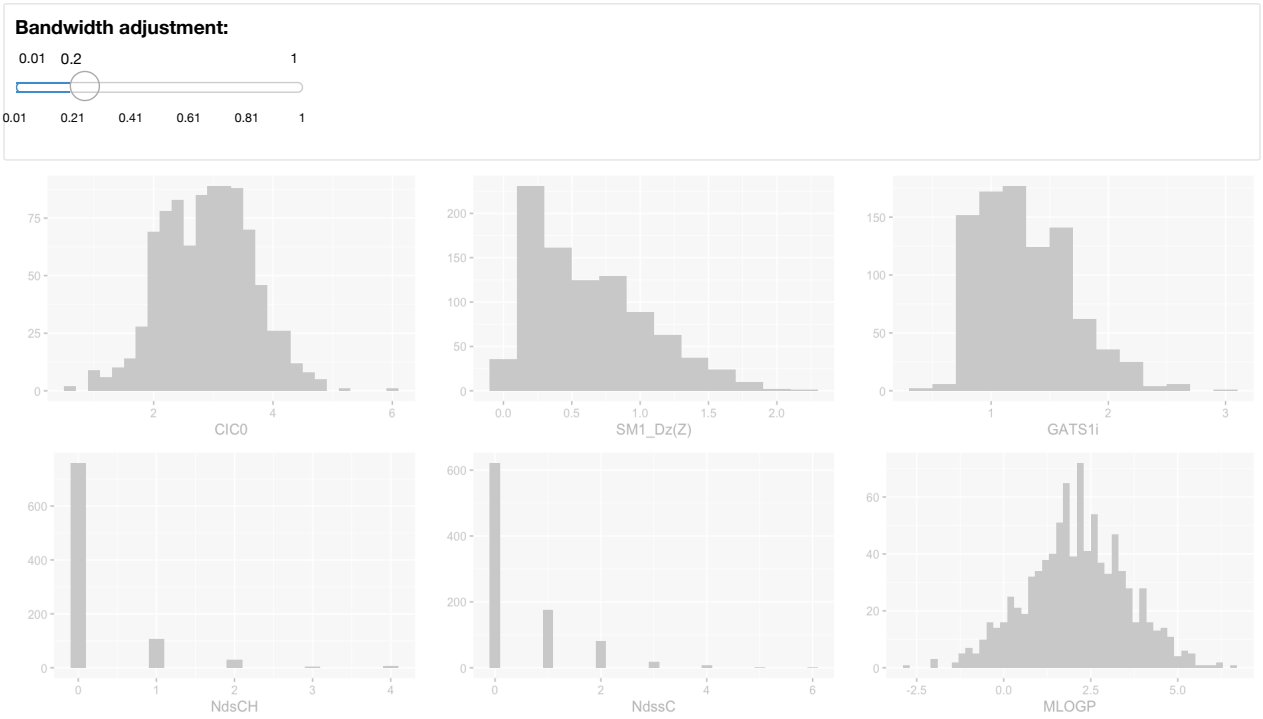


Histograms

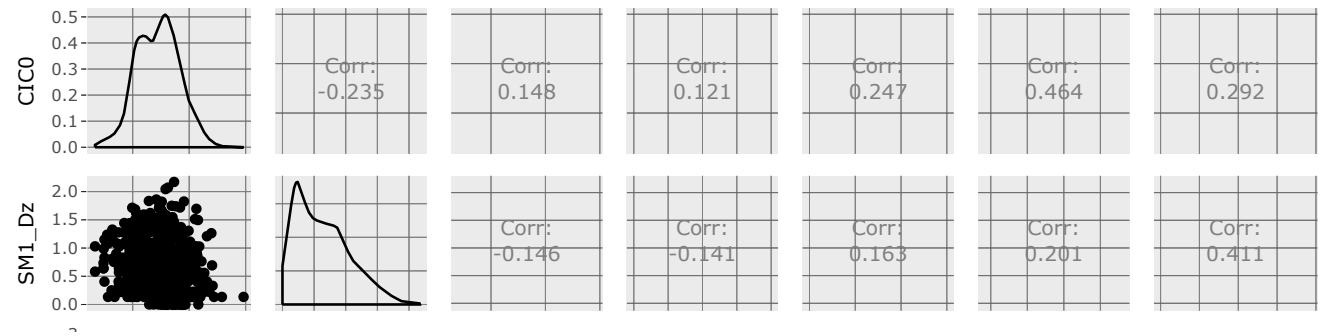
Histogram of Response Variable

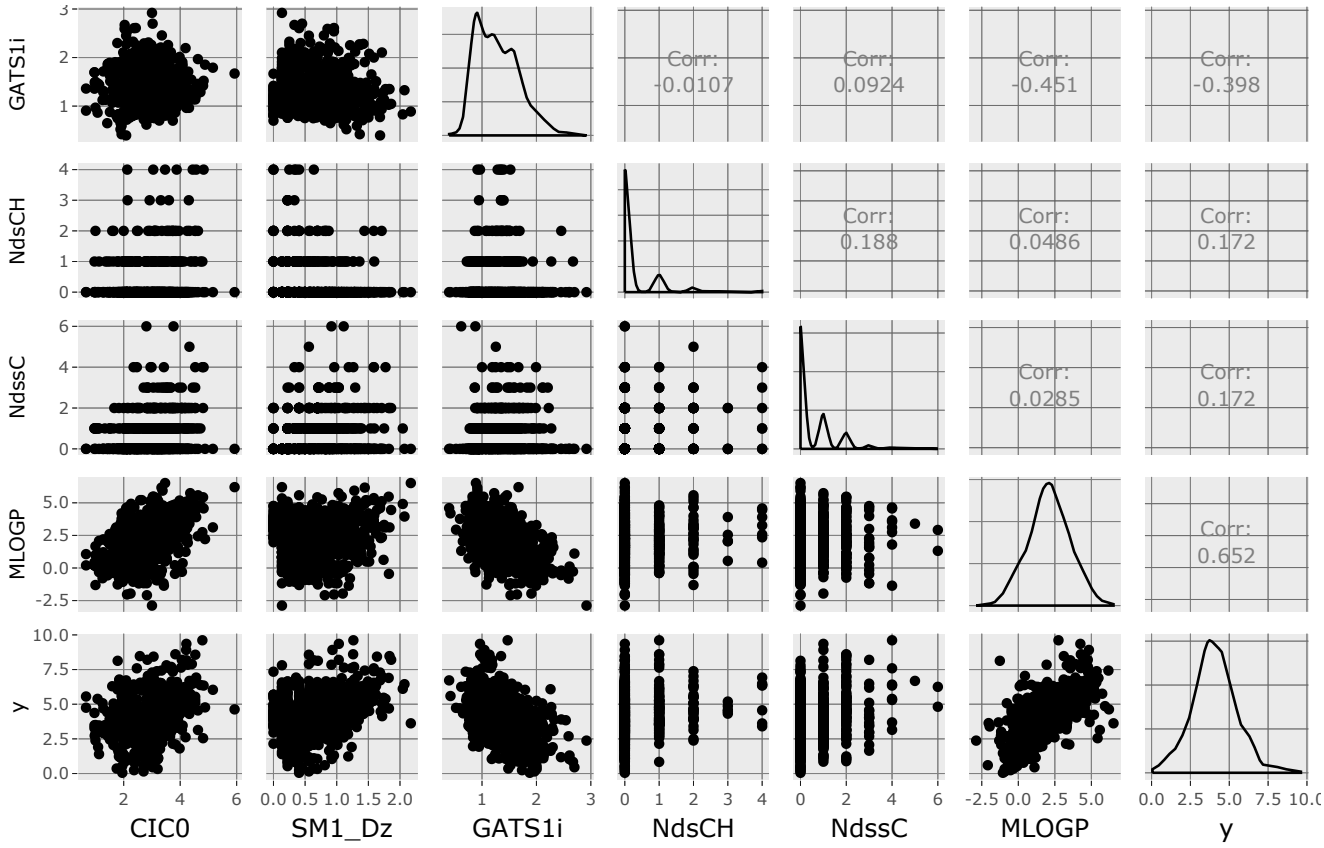


Histograms of Explanatory Variables

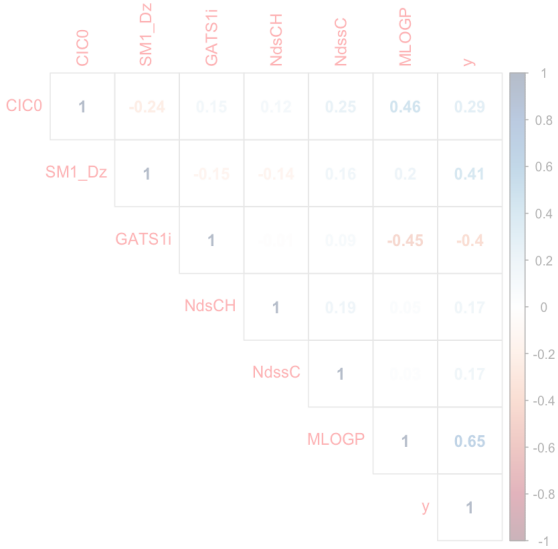


Scatter Plots & Correlations





Correlation Heat Map



Models (TODO)

Using `regsubsets()` to find best-fitted model

```
models <- regsubsets(dt$y ~., data = dt, method = "exhaustive")
models_summary <- summary(models); models_summary
```

```
## Subset selection object
## Call: eval(expr, envir, enclos)
## 6 Variables  (and intercept)
##        Forced in Forced out
## CIC0       FALSE      FALSE
## SM1_Dz     FALSE      FALSE
## GATS1i     FALSE      FALSE
## NdsCH      FALSE      FALSE
## NdssC      FALSE      FALSE
## MLOGP      FALSE      FALSE
## 1 subsets of each size up to 6
## Selection Algorithm: exhaustive
##          CIC0 SM1_Dz GATS1i NdsCH NdssC MLOGP
## 1  ( 1 ) " "  " "    " "    " "   " "   "*"
## 2  ( 1 ) " "  "*"    " "    " "   " "   "*"
## 3  ( 1 ) " "  "*"    " "    "*"   " "   "*"
## 4  ( 1 ) " "  "*"    "*"    "*"   " "   "*"
## 5  ( 1 ) "*"  "*"    "*"    "*"   " "   "*"
## 6  ( 1 ) "*"  "*"    "*"    "*"   "*"   "*"
```

```
models_res <- data.frame(
  Adj.R2 = which.max(models_summary$adjr2),
  CP = which.min(models_summary$cp)
); models_res # observation: model 6
```

```
##   Adj.R2 CP
## 1      6  6
```

```
models_summary$adjr2
```

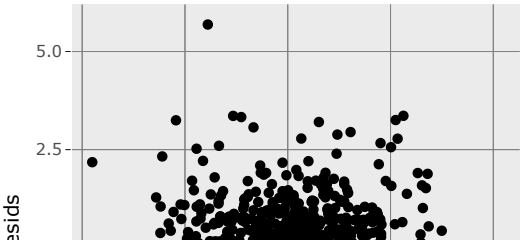
```
## [1] 0.4240310 0.5053395 0.5398406 0.5492448 0.5736370 0.5743478
```

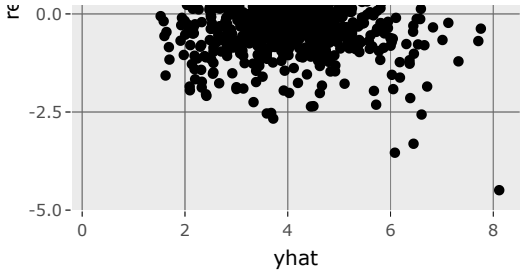
Best-fitted model summary

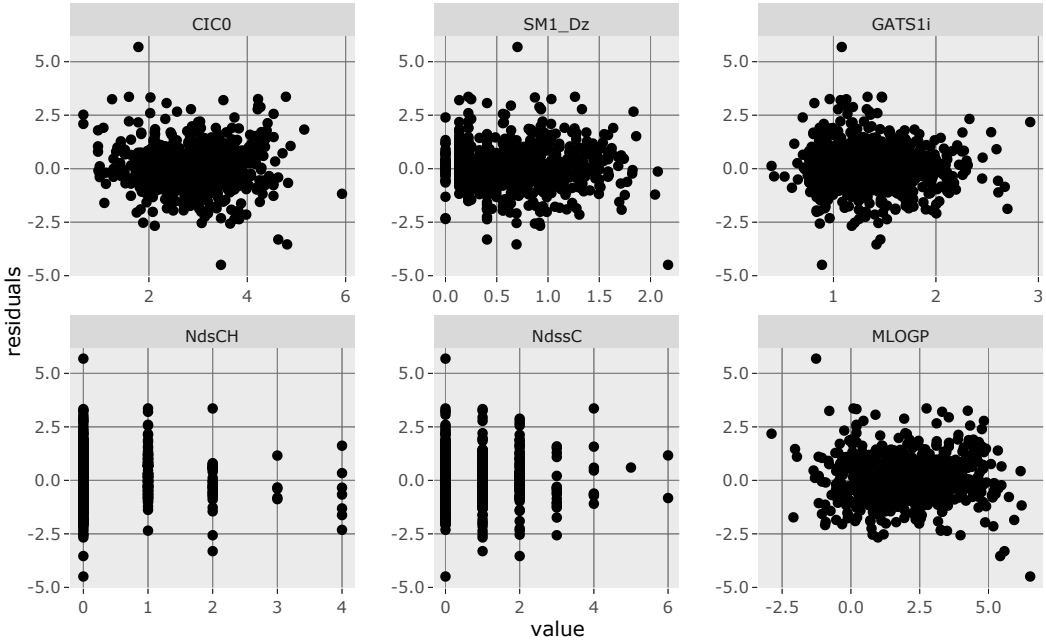
```
# Using the best model (model 6, selecting all variables) (& generate residual plot):
reg <- lm(dt$y ~ ., data = dt)
summary(reg)
```

```
##
## Call:
## lm(formula = dt$y ~ ., data = dt)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.4921 -0.5287 -0.0712  0.4861  5.6876
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.17456    0.18122  12.000  < 2e-16 ***
## CIC0         0.38563    0.06089   6.333 3.79e-10 ***
## SM1_Dz       1.25562    0.08702  14.430  < 2e-16 ***
## GATS1i      -0.74641    0.10135  -7.365 4.00e-13 ***
## NdsCH        0.41355    0.05410   7.644 5.41e-14 ***
## NdssC        0.06433    0.04064   1.583    0.114
## MLOGP        0.39005    0.03376  11.555  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9497 on 901 degrees of freedom
## Multiple R-squared:  0.5772, Adjusted R-squared:  0.5743
## F-statistic:   205 on 6 and 901 DF,  p-value: < 2.2e-16
```

Residual Plots




re
0.0
-2.5
-5.0
0
2
4
6
8
yhat


CIC0
SM1_Dz
GATS1i
NdsCH
NdssC
MLOGP
residuals
value