SLR_superconductivity

Sowmya, Disha, Anushka, Pranjali, Ashmita

19/12/2021

Summary:

Aim:

Plotting Linear Regression for best correlated variable with "critical temperature" in superconductivity data.

Steps:

Finding most correlated variables

Exploratory Analysis for these variables to understand more about them

Finding trend of correlated variables with target variable "critical_temp" to find which variable is best for SLR

Checking assumptions for SLR and performing SLR

Checking if the model is a good fit

```
sconduct <- read.csv('C:/Users/disha/OneDrive/Desktop/501/final project/superconduct/train.csv')
ncol(sconduct)
```

```
## [1] 82
```

```
sconduct_x <- subset(sconduct, select = -critical_temp)
sconduct_y <- sconduct['critical_temp']
names_col = names(sconduct)
names_col
```

```
##  [1] "number_of_elements"            "mean_atomic_mass"
##  [3] "wtd_mean_atomic_mass"          "gmean_atomic_mass"
##  [5] "wtd_gmean_atomic_mass"         "entropy_atomic_mass"
##  [7] "wtd_entropy_atomic_mass"       "range_atomic_mass"
##  [9] "wtd_range_atomic_mass"         "std_atomic_mass"
## [11] "wtd_std_atomic_mass"           "mean_fie"
## [13] "wtd_mean_fie"                  "gmean_fie"
```

```
## [15] "wtd_gmean_fie"                    "entropy_fie"
## [17] "wtd_entropy_fie"                  "range_fie"
## [19] "wtd_range_fie"                    "std_fie"
## [21] "wtd_std_fie"                      "mean_atomic_radius"
## [23] "wtd_mean_atomic_radius"           "gmean_atomic_radius"
## [25] "wtd_gmean_atomic_radius"          "entropy_atomic_radius"
## [27] "wtd_entropy_atomic_radius"        "range_atomic_radius"
## [29] "wtd_range_atomic_radius"          "std_atomic_radius"
## [31] "wtd_std_atomic_radius"            "mean_Density"
## [33] "wtd_mean_Density"                 "gmean_Density"
## [35] "wtd_gmean_Density"                "entropy_Density"
## [37] "wtd_entropy_Density"              "range_Density"
## [39] "wtd_range_Density"                "std_Density"
## [41] "wtd_std_Density"                  "mean_ElectronAffinity"
## [43] "wtd_mean_ElectronAffinity"        "gmean_ElectronAffinity"
## [45] "wtd_gmean_ElectronAffinity"       "entropy_ElectronAffinity"
## [47] "wtd_entropy_ElectronAffinity"     "range_ElectronAffinity"
## [49] "wtd_range_ElectronAffinity"       "std_ElectronAffinity"
## [51] "wtd_std_ElectronAffinity"         "mean_FusionHeat"
## [53] "wtd_mean_FusionHeat"              "gmean_FusionHeat"
## [55] "wtd_gmean_FusionHeat"             "entropy_FusionHeat"
## [57] "wtd_entropy_FusionHeat"           "range_FusionHeat"
## [59] "wtd_range_FusionHeat"             "std_FusionHeat"
## [61] "wtd_std_FusionHeat"               "mean_ThermalConductivity"
## [63] "wtd_mean_ThermalConductivity"     "gmean_ThermalConductivity"
## [65] "wtd_gmean_ThermalConductivity"    "entropy_ThermalConductivity"
## [67] "wtd_entropy_ThermalConductivity" "range_ThermalConductivity"
## [69] "wtd_range_ThermalConductivity"    "std_ThermalConductivity"
## [71] "wtd_std_ThermalConductivity"      "mean_Valence"
## [73] "wtd_mean_Valence"                 "gmean_Valence"
## [75] "wtd_gmean_Valence"                "entropy_Valence"
## [77] "wtd_entropy_Valence"              "range_Valence"
## [79] "wtd_range_Valence"                "std_Valence"
## [81] "wtd_std_Valence"                  "critical_temp"
```

Finding most correlated variables

```
df2 = cor(sconduct)
df2 = df2[,82] # only select correlation with critical temp

df2 = sort(df2, decreasing = TRUE)
df2
```

```
##                   critical_temp     wtd_std_ThermalConductivity
##                      1.00000000                      0.72127108
##       range_ThermalConductivity             range_atomic_radius
##                      0.68765391                      0.65375904
##         std_ThermalConductivity         wtd_entropy_atomic_mass
##                      0.65363198                      0.62693040
##       wtd_entropy_atomic_radius              number_of_elements
##                      0.60349398                      0.60106857
##                       range_fie           wtd_std_atomic_radius
```

```
##                     0.60079038                     0.59919866
##                entropy_Valence            wtd_entropy_Valence
##                     0.59859091                     0.58966370
##                    wtd_std_fie                    entropy_fie
##                     0.58201326                     0.56781694
##         wtd_entropy_FusionHeat             std_atomic_radius
##                     0.56324427                     0.55962857
##          entropy_atomic_radius            entropy_FusionHeat
##                     0.55893744                     0.55270871
##            entropy_atomic_mass                        std_fie
##                     0.54361941                     0.54180381
##              range_atomic_mass  wtd_range_ThermalConductivity
##                     0.49196981                     0.46957158
##                entropy_Density       entropy_ElectronAffinity
##                     0.45716939                     0.43720681
##            wtd_entropy_Density                   wtd_mean_fie
##                     0.40019011                     0.39879637
##                wtd_entropy_fie   wtd_mean_ThermalConductivity
##                     0.38835905                     0.37933606
##                std_atomic_mass       mean_ThermalConductivity
##                     0.37876583                     0.37581286
##            wtd_std_atomic_mass                  wtd_gmean_fie
##                     0.35930616                     0.34374657
##       wtd_std_ElectronAffinity                  wtd_range_fie
##                     0.31514734                     0.30048245
##         range_ElectronAffinity           std_ElectronAffinity
##                     0.27970455                     0.26210348
##                  range_Density   wtd_entropy_ElectronAffinity
##                     0.26053562                     0.23764782
##                wtd_std_Density     wtd_range_ElectronAffinity
##                     0.20766315                     0.18534849
##                    std_Density      wtd_mean_ElectronAffinity
##                     0.11524263                     0.11151585
##             mean_atomic_radius                       mean_fie
##                     0.10527269                     0.10226805
##    entropy_ThermalConductivity                      gmean_fie
##                     0.08586207                    -0.02510325
##     wtd_gmean_ElectronAffinity               mean_atomic_mass
##                    -0.10735929                    -0.11352325
## wtd_entropy_ThermalConductivity               range_FusionHeat
##                    -0.11672759                    -0.14071369
##                   range_Valence            gmean_atomic_radius
##                    -0.14354639                    -0.14377022
##          mean_ElectronAffinity             wtd_std_FusionHeat
##                    -0.19355049                    -0.19557117
##                 std_FusionHeat                    std_Valence
##                    -0.20131006                    -0.20807162
##               gmean_atomic_mass              wtd_range_Density
##                    -0.23034537                    -0.28472937
##         wtd_mean_atomic_radius                wtd_std_Valence
##                    -0.29727212                    -0.30002801
##           wtd_mean_atomic_mass           wtd_range_FusionHeat
##                    -0.31227202                    -0.31417848
##           wtd_range_atomic_mass        wtd_range_atomic_radius
```

```
##                     -0.33713139                     -0.34409981
##                    mean_Density           wtd_gmean_atomic_mass
##                     -0.36826182                     -0.36985838
##   wtd_gmean_ThermalConductivity          gmean_ElectronAffinity
##                     -0.37160145                     -0.38056775
##                  mean_FusionHeat       gmean_ThermalConductivity
##                     -0.38550926                     -0.38719231
##             wtd_mean_FusionHeat         wtd_gmean_atomic_radius
##                     -0.39411694                     -0.40517561
##                 gmean_FusionHeat            wtd_gmean_FusionHeat
##                     -0.43179461                     -0.43236451
##                 wtd_mean_Density               wtd_range_Valence
##                     -0.43393963                     -0.43990078
##                wtd_gmean_Density                   gmean_Density
##                     -0.54004559                     -0.54168441
##                    gmean_Valence                    mean_Valence
##                     -0.57306806                     -0.60008486
##                wtd_gmean_Valence                wtd_mean_Valence
##                     -0.61565330                     -0.63240102
```

Picking the 5 variables that vary lineary (hit and trial gave these best variables varying linearly, even though correlation might be less)

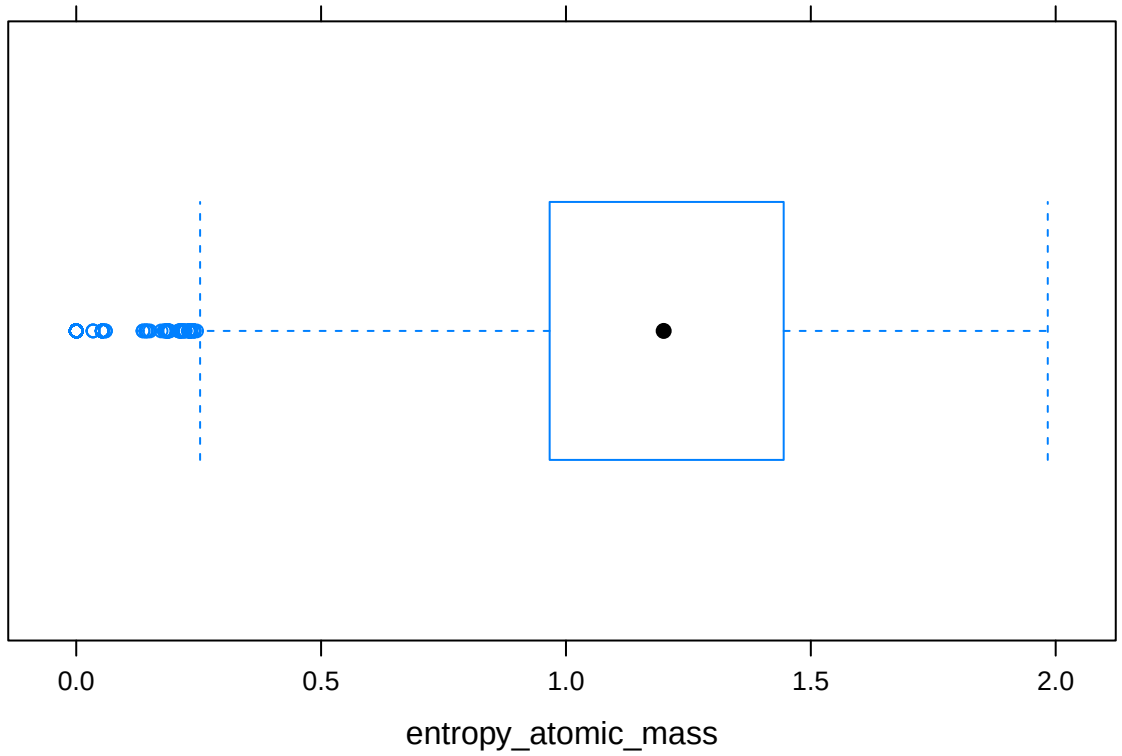
Exploratory Analysis for these variables to understand more about them

Index used :

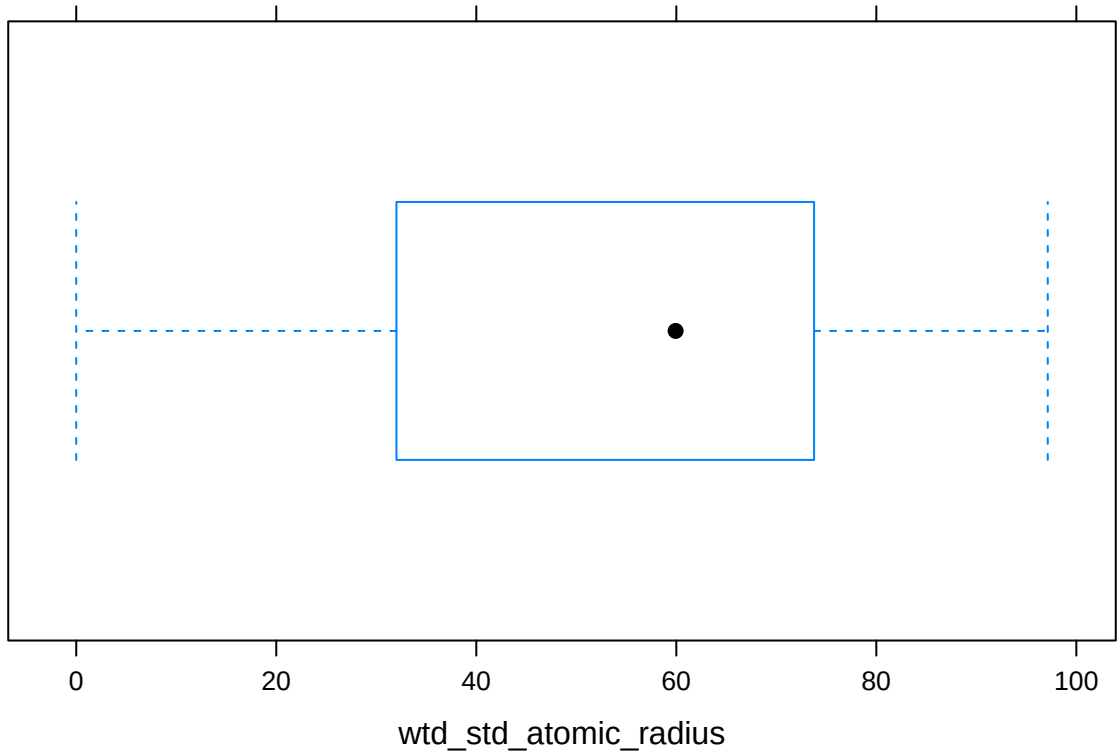
```
0 : entropy_atomic_mass
1 : wtd_std_atomic_radius
2 : wtd_entropy_FusionHeat
3 : wtd_entropy_Valence
4 : wtd_range_ThermalConductivity
```

All variables continuous, quantitative except number of elements

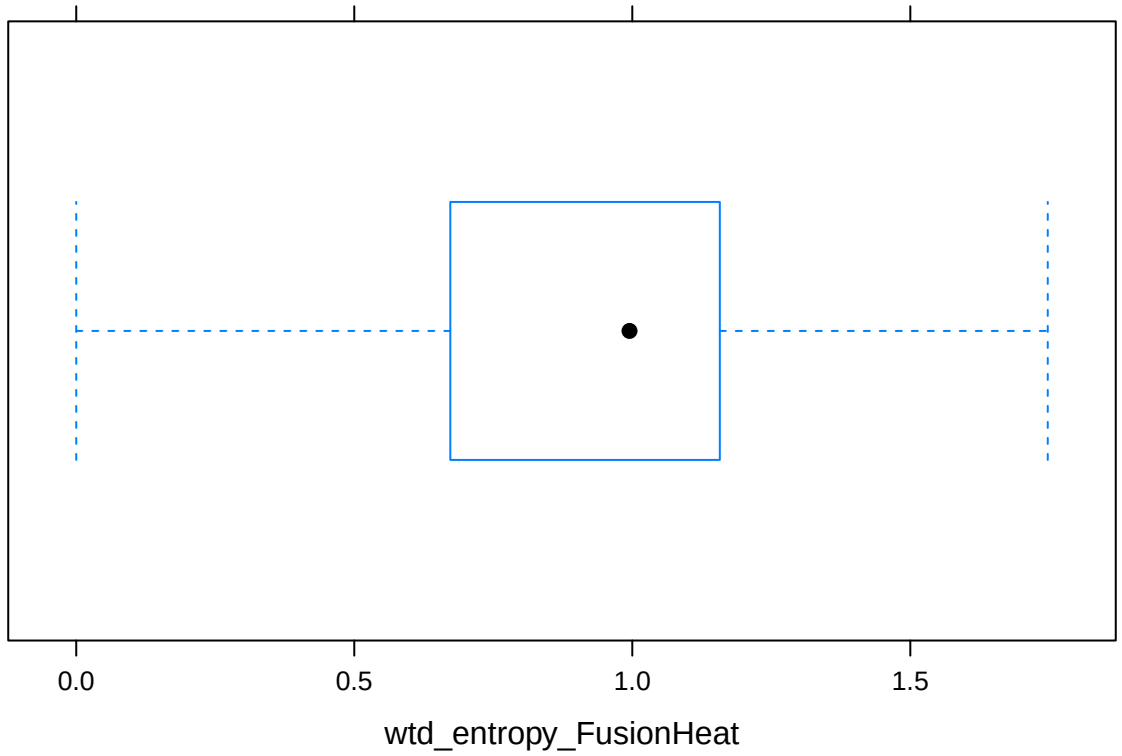
```
bwplot(~entropy_atomic_mass, data=sconduct)
```



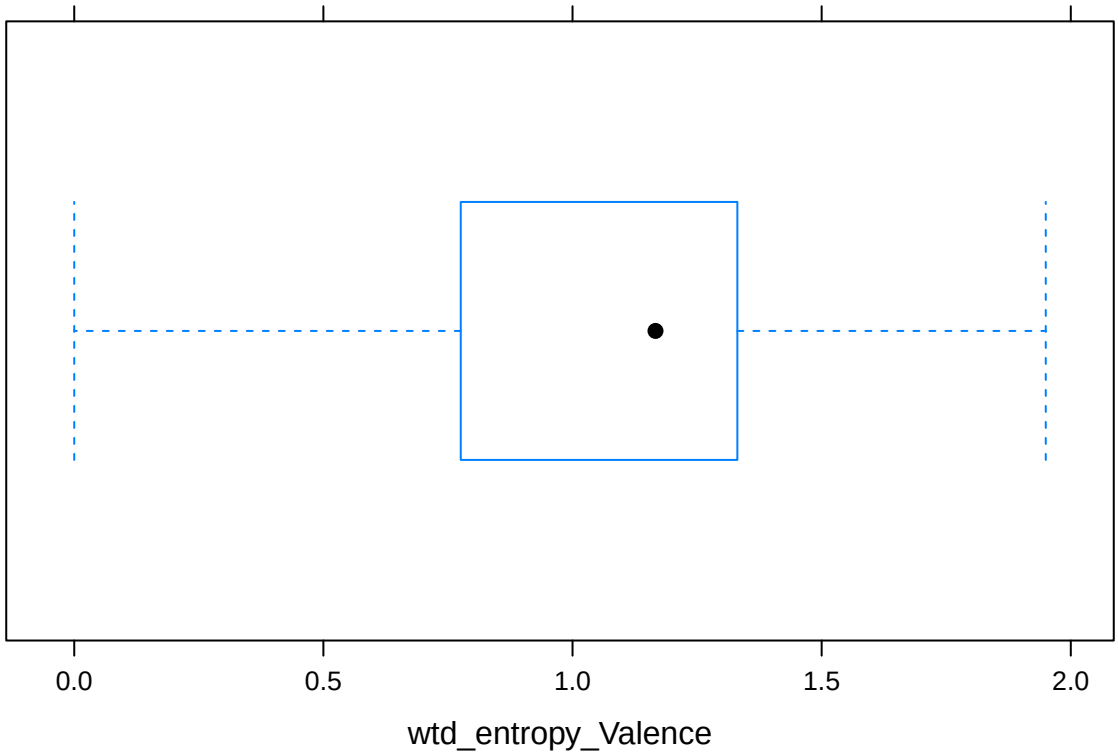

```
bwplot(~wtd_std_atomic_radius, data=sconduct)
```



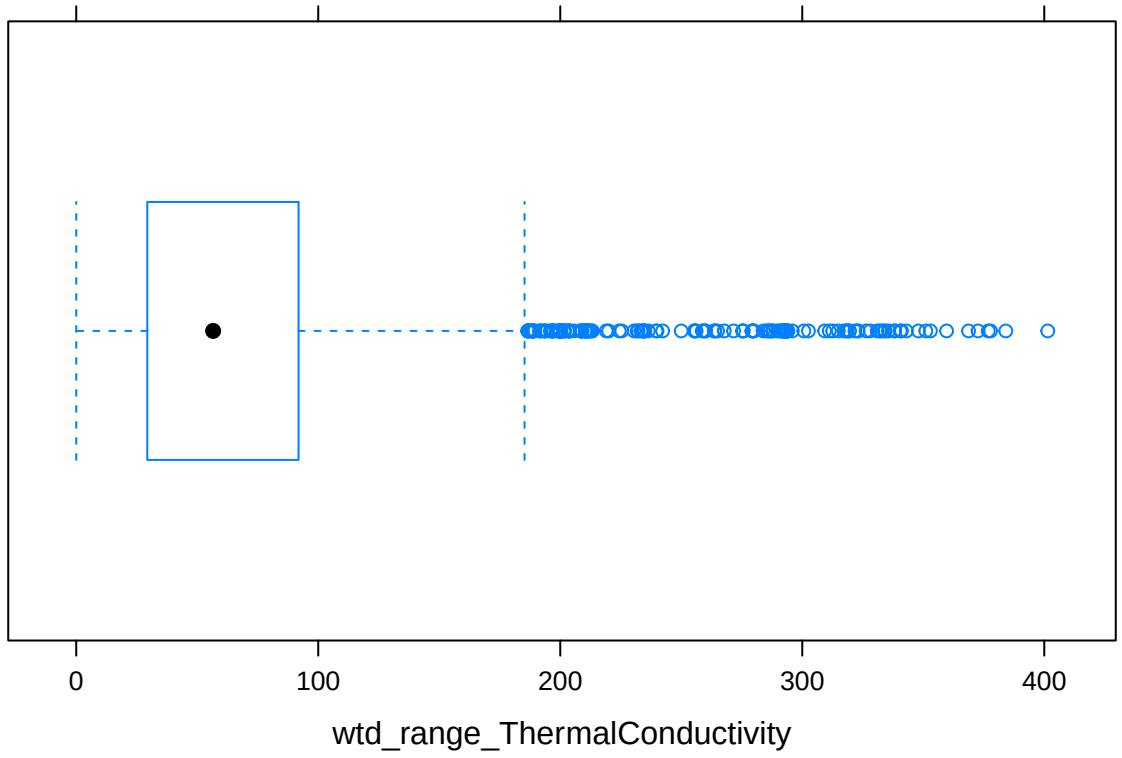

```
bwplot(~wtd_entropy_FusionHeat, data=sconduct)
```



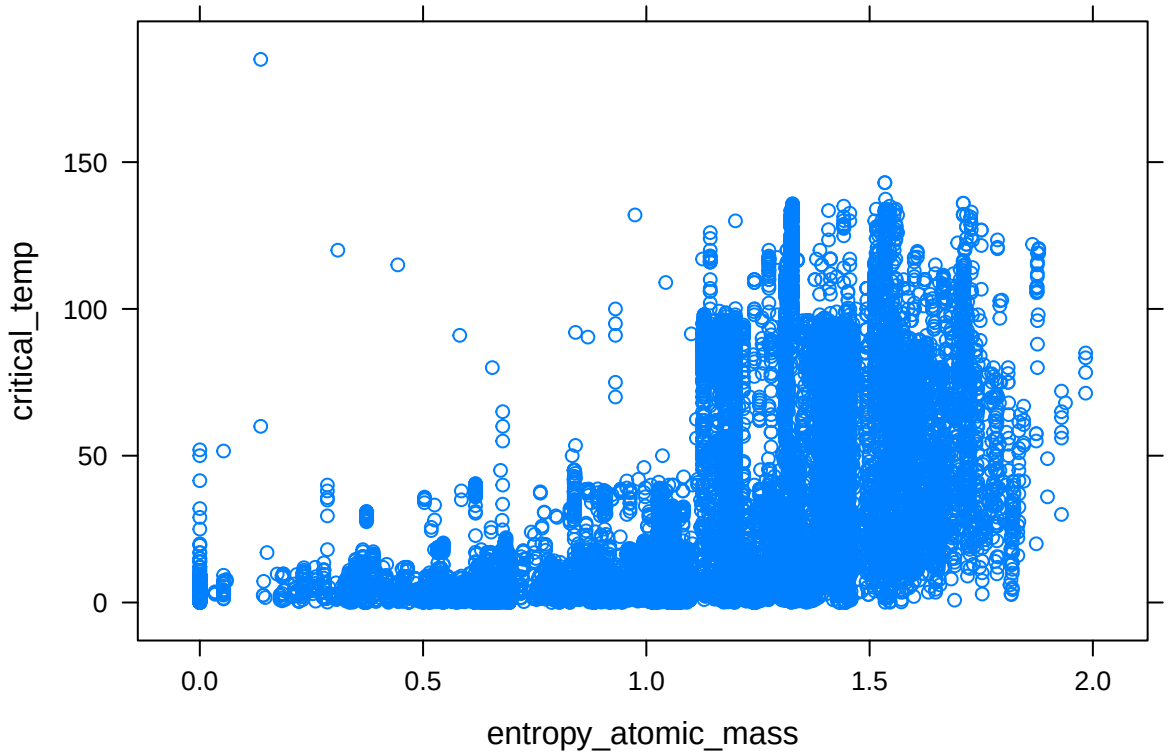

```
bwplot(~wtd_entropy_Valence, data=sconduct)
```



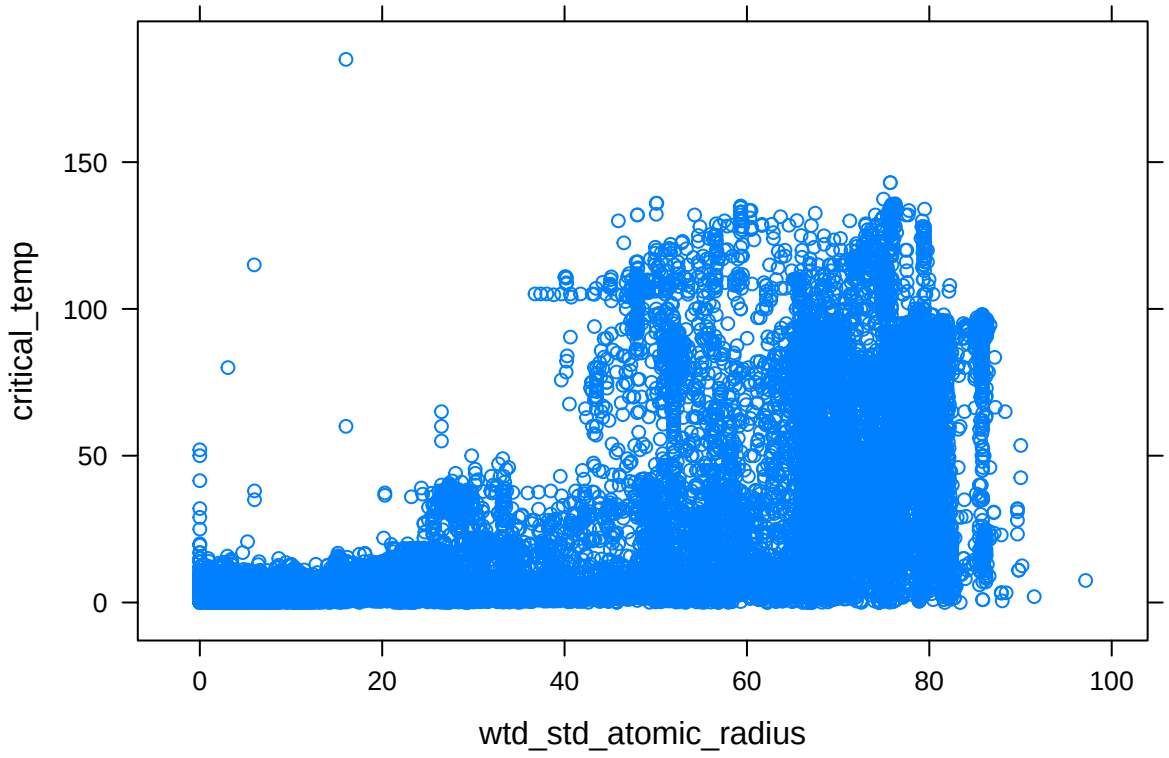

```
bwplot(~wtd_range_ThermalConductivity, data=sconduct)
```



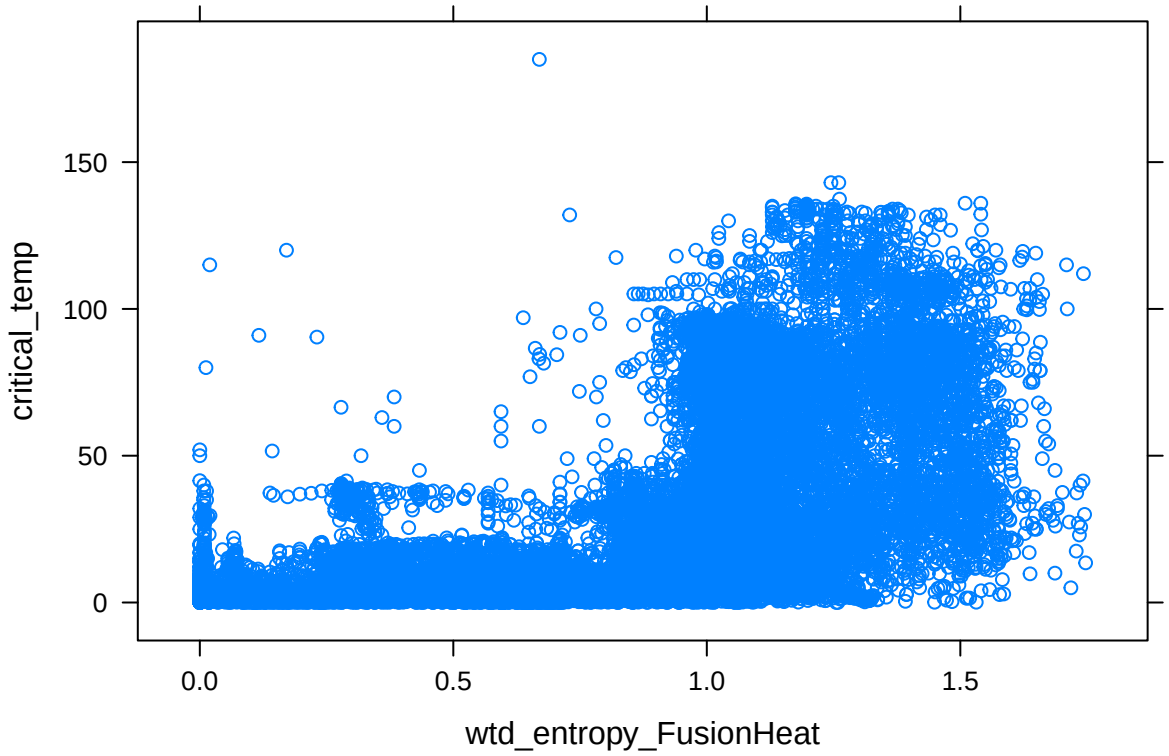

```
xyplot(critical_temp ~ entropy_atomic_mass, data=sconduct)
```



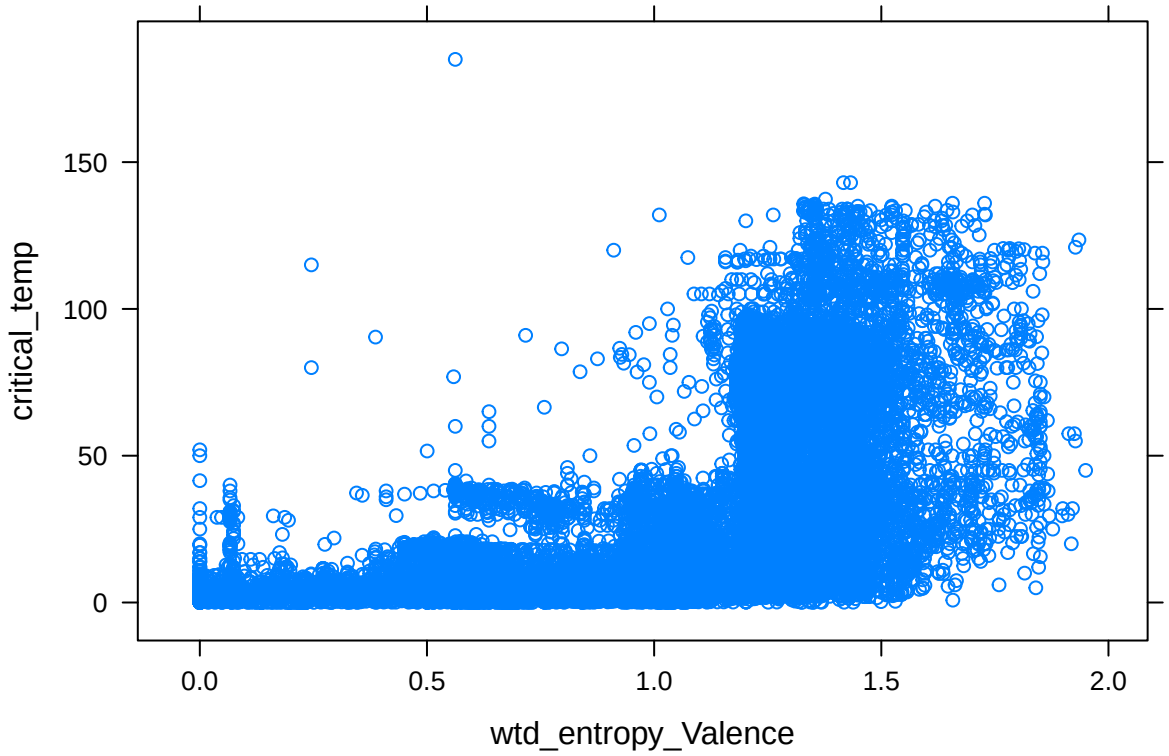

```
xyplot(critical_temp ~ wtd_std_atomic_radius, data=sconduct)
```



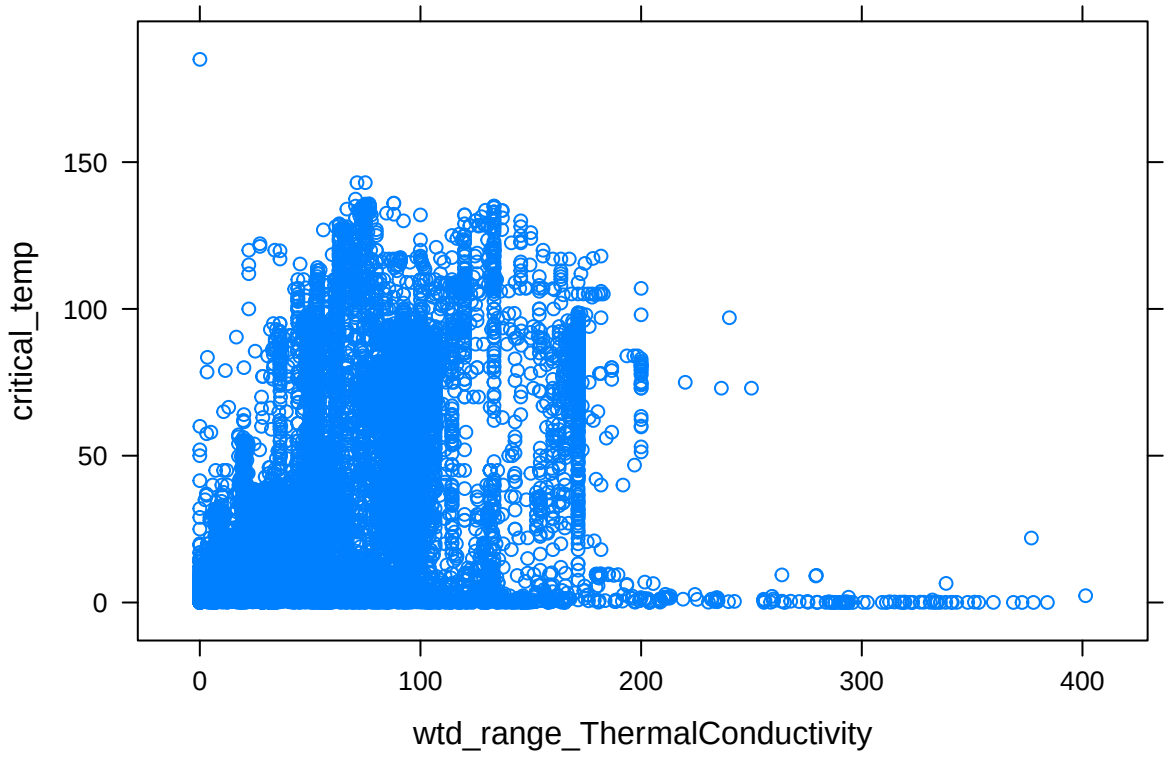

```
xyplot(critical_temp ~ wtd_entropy_FusionHeat, data=sconduct)
```




```
xyplot(critical_temp ~ wtd_entropy_Valence, data=sconduct)
```



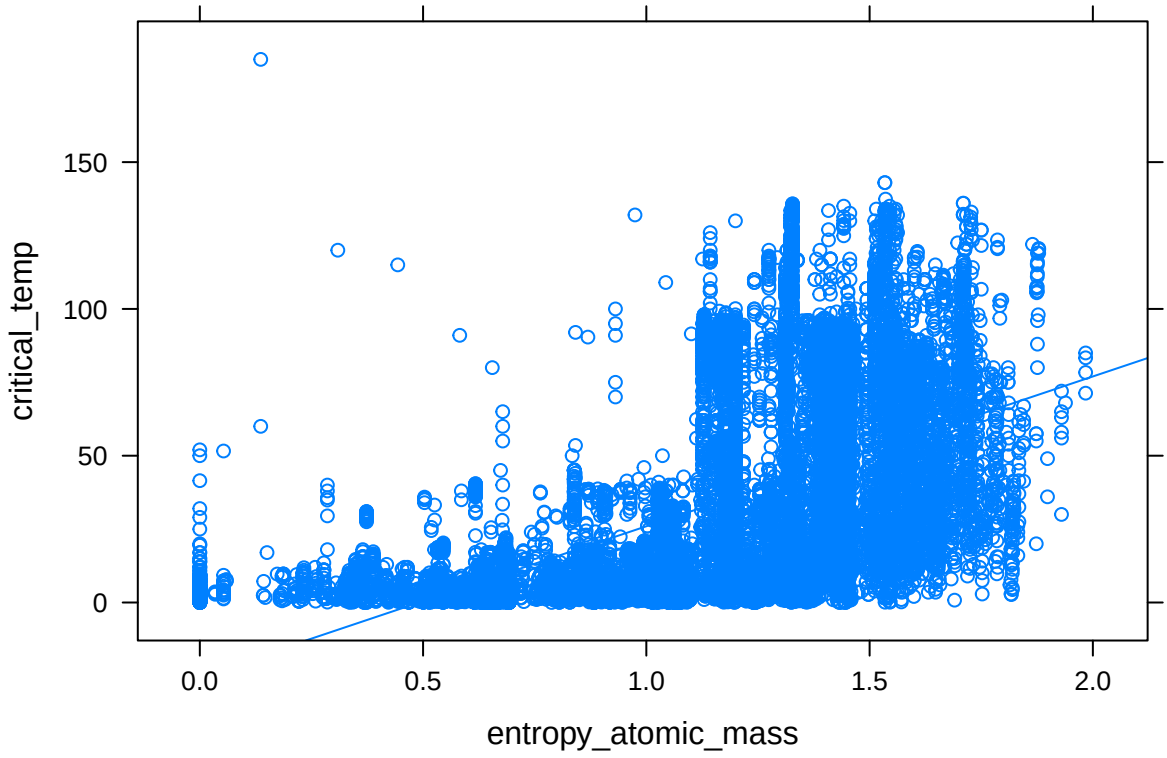

```
xyplot(critical_temp ~ wtd_range_ThermalConductivity, data=sconduct)
```



Fitting the linear model on these variables and Checking assumptions for these 5 variables

Fit

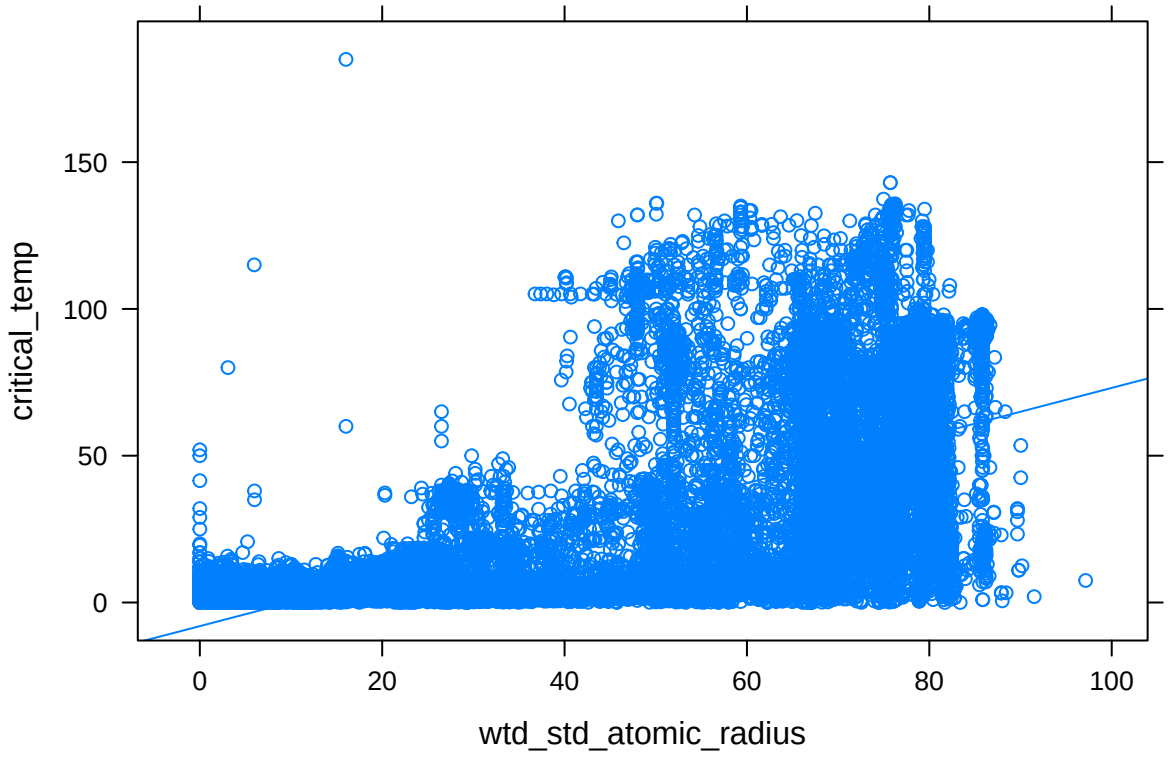
```
# entropy_atomic_mass
xyplot(critical_temp ~ entropy_atomic_mass, sconduct, type=c("p", "r"))
```




```
lm0 <- lm(critical_temp ~ entropy_atomic_mass, sconduct)
summary(lm0)
```

```
##
## Call:
## lm(formula = critical_temp ~ entropy_atomic_mass, data = sconduct)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -64.981 -22.201  -6.004  19.222 203.097
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)         -25.0564     0.6599  -37.97   <2e-16 ***
## entropy_atomic_mass  51.0271     0.5403   94.44   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 28.75 on 21261 degrees of freedom
## Multiple R-squared:  0.2955, Adjusted R-squared:  0.2955
## F-statistic:  8919 on 1 and 21261 DF,  p-value: < 2.2e-16
```

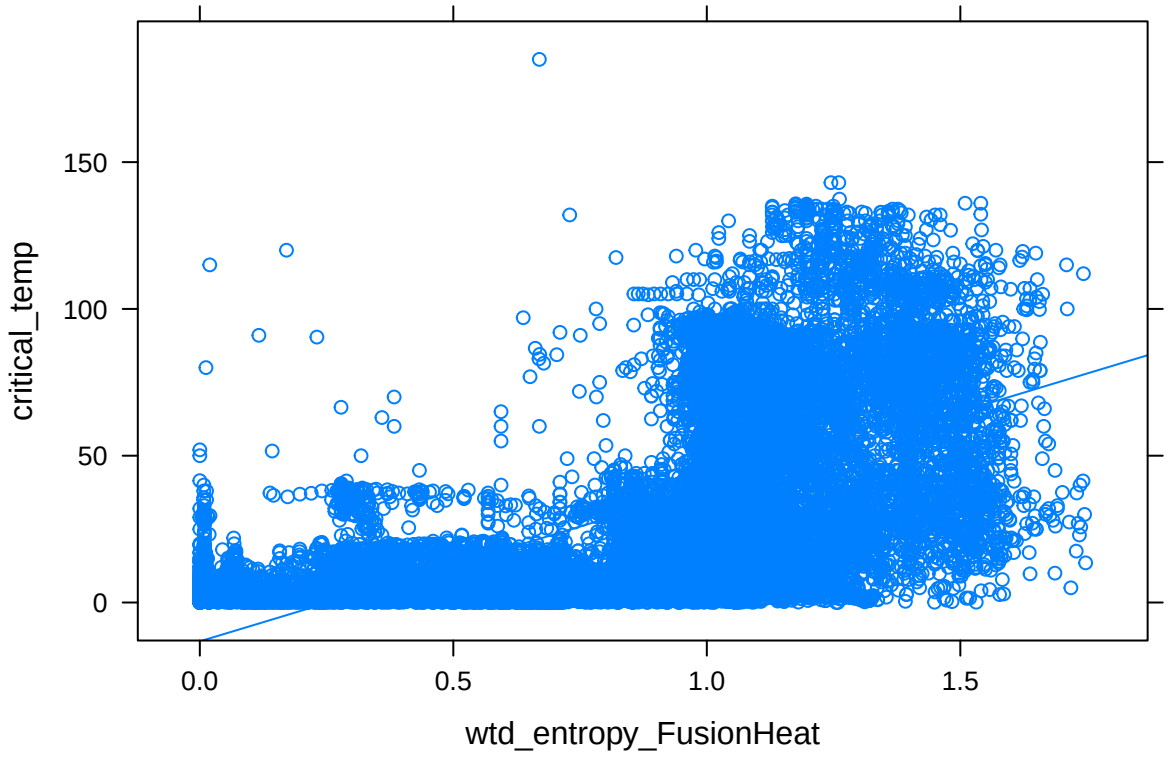
```
# wtd_std_atomic_radius
xyplot(critical_temp ~ wtd_std_atomic_radius, sconduct, type=c("p", "r"))
```




```
lm1 <- lm(critical_temp ~ wtd_std_atomic_radius, sconduct)
summary(lm1)
```

```
##
## Call:
## lm(formula = critical_temp ~ wtd_std_atomic_radius, data = sconduct)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -64.197 -20.737  -1.913  16.121 180.045
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)            -8.050299   0.432242  -18.62   <2e-16 ***
## wtd_std_atomic_radius   0.811447   0.007436  109.13   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 27.42 on 21261 degrees of freedom
## Multiple R-squared:  0.359,  Adjusted R-squared:  0.359
## F-statistic: 1.191e+04 on 1 and 21261 DF,  p-value: < 2.2e-16
```

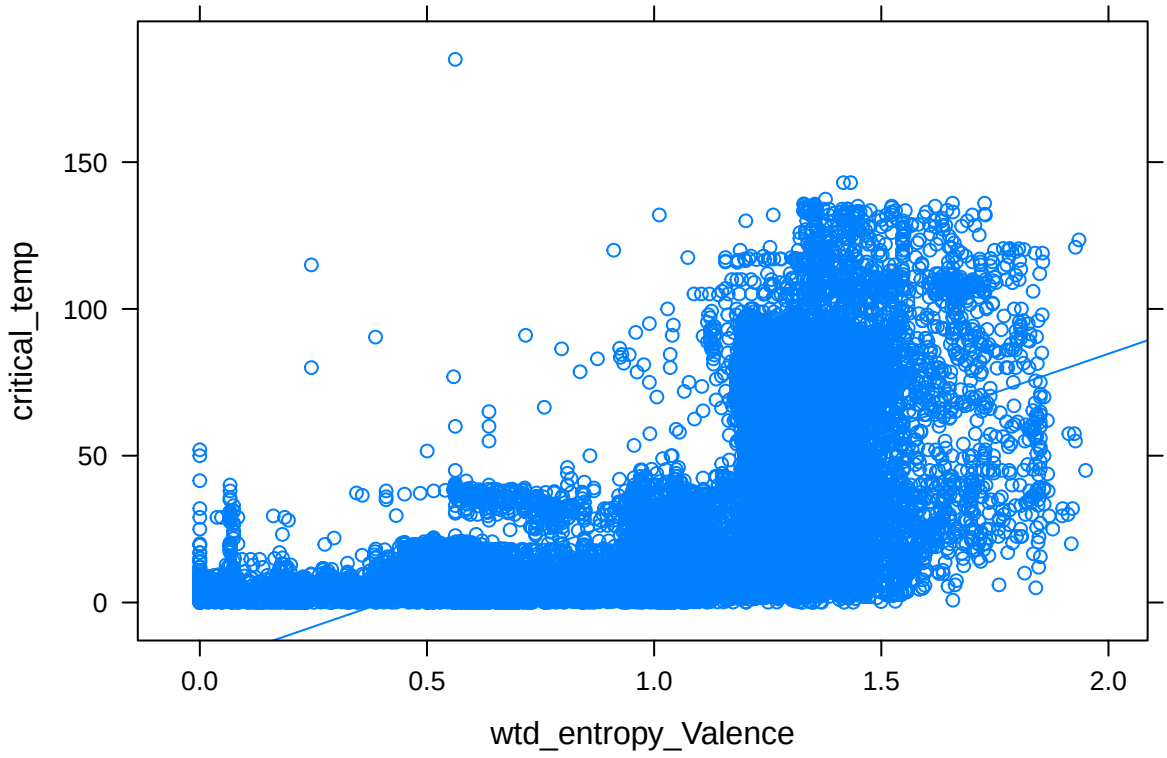
```
# wtd_entropy_FusionHeat
xyplot(critical_temp ~ wtd_entropy_FusionHeat, sconduct, type=c("p", "r"))
```




```
lm2 <- lm(critical_temp ~ wtd_entropy_FusionHeat, sconduct)
summary(lm2)
```

```
##
## Call:
## lm(formula = critical_temp ~ wtd_entropy_FusionHeat, data = sconduct)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -71.333 -21.149  -5.749  18.468 163.317
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)             -13.2264     0.5172  -25.57   <2e-16 ***
## wtd_entropy_FusionHeat  52.1272     0.5245   99.39   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 28.3 on 21261 degrees of freedom
## Multiple R-squared:  0.3172, Adjusted R-squared:  0.3172
## F-statistic:  9879 on 1 and 21261 DF,  p-value: < 2.2e-16
```

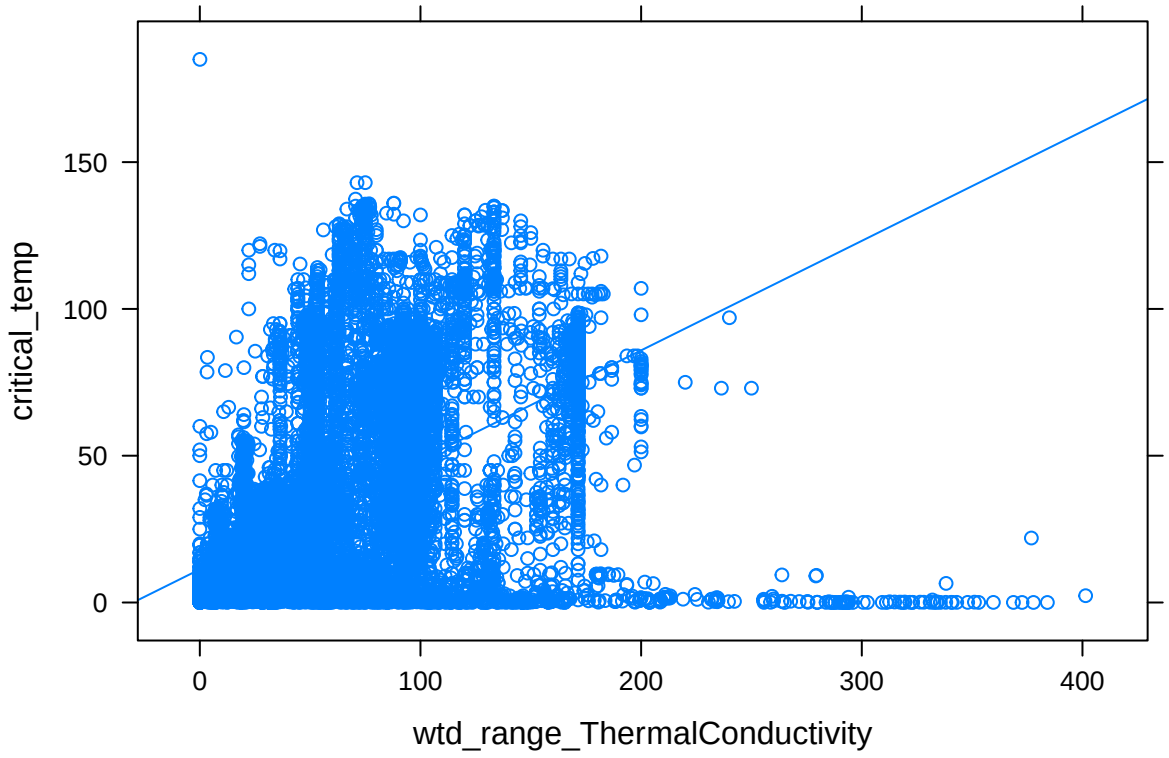
```
# wtd_entropy_Valence
xyplot(critical_temp ~ wtd_entropy_Valence, sconduct, type=c("p", "r"))
```




```
lm3 <- lm(critical_temp ~ wtd_entropy_Valence, sconduct)
summary(lm3)
```

```
##
## Call:
## lm(formula = critical_temp ~ wtd_entropy_Valence, data = sconduct)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -71.235 -21.088  -5.188  20.522 176.631
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)         -21.4989     0.5585  -38.49   <2e-16 ***
## wtd_entropy_Valence  53.1135     0.4989  106.46   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 27.67 on 21261 degrees of freedom
## Multiple R-squared:  0.3477, Adjusted R-squared:  0.3477
## F-statistic: 1.133e+04 on 1 and 21261 DF,  p-value: < 2.2e-16
```

```
# wtd_entropy_atomic_mass
xyplot(critical_temp ~ wtd_range_ThermalConductivity, sconduct, type=c("p", "r"))
```




```
lm4 <- lm(critical_temp ~ wtd_range_ThermalConductivity, sconduct)
summary(lm4)
```

```
##
## Call:
## lm(formula = critical_temp ~ wtd_range_ThermalConductivity, data = sconduct)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -158.699  -18.057   -7.692   17.292  173.698
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   11.28306    0.36337   31.05   <2e-16 ***
## wtd_range_ThermalConductivity  0.37300    0.00481   77.55   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 30.24 on 21261 degrees of freedom
## Multiple R-squared:  0.2205, Adjusted R-squared:  0.2205
## F-statistic:  6014 on 1 and 21261 DF,  p-value: < 2.2e-16
```

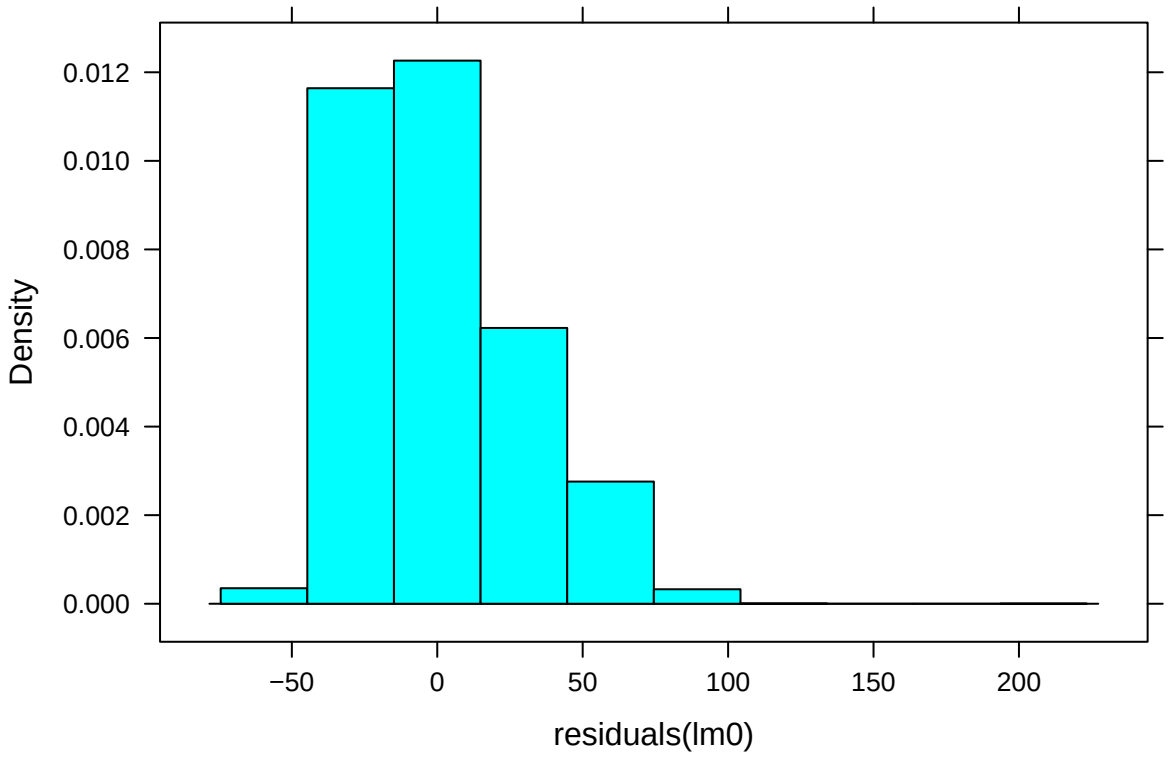
Check assumptions

1. Linearity : Verified (appears linear)

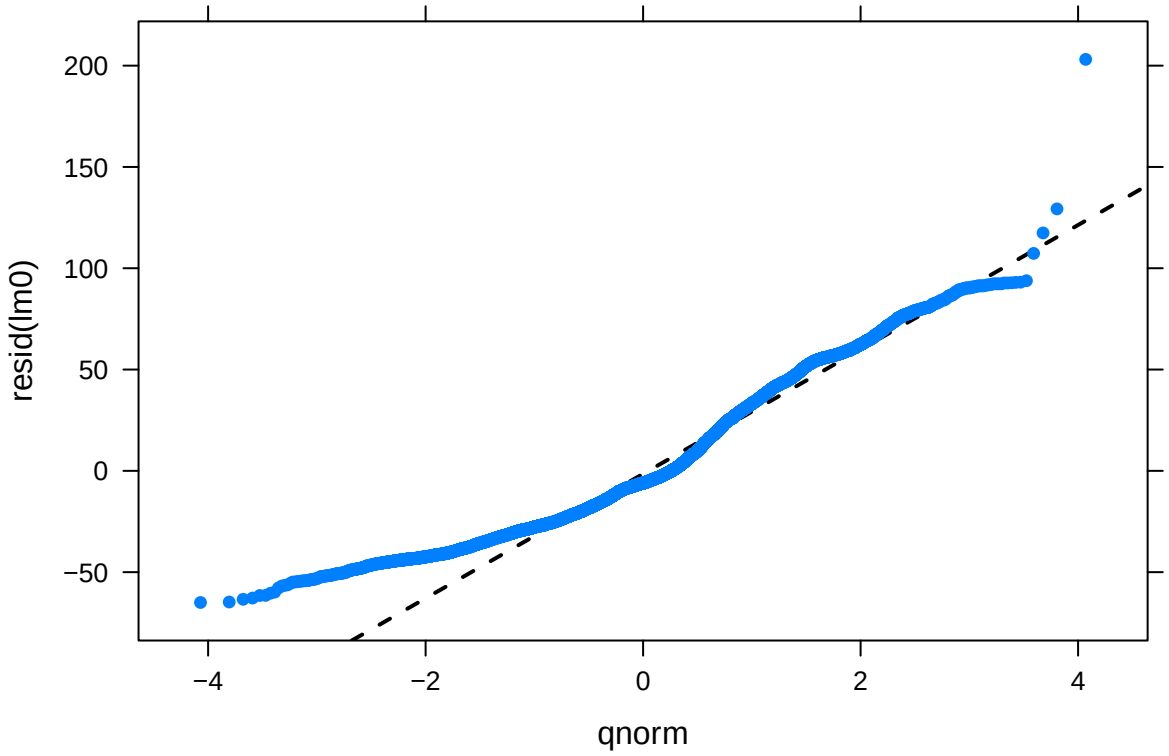
2. Independence : Assume true

3. Residuals Normal :

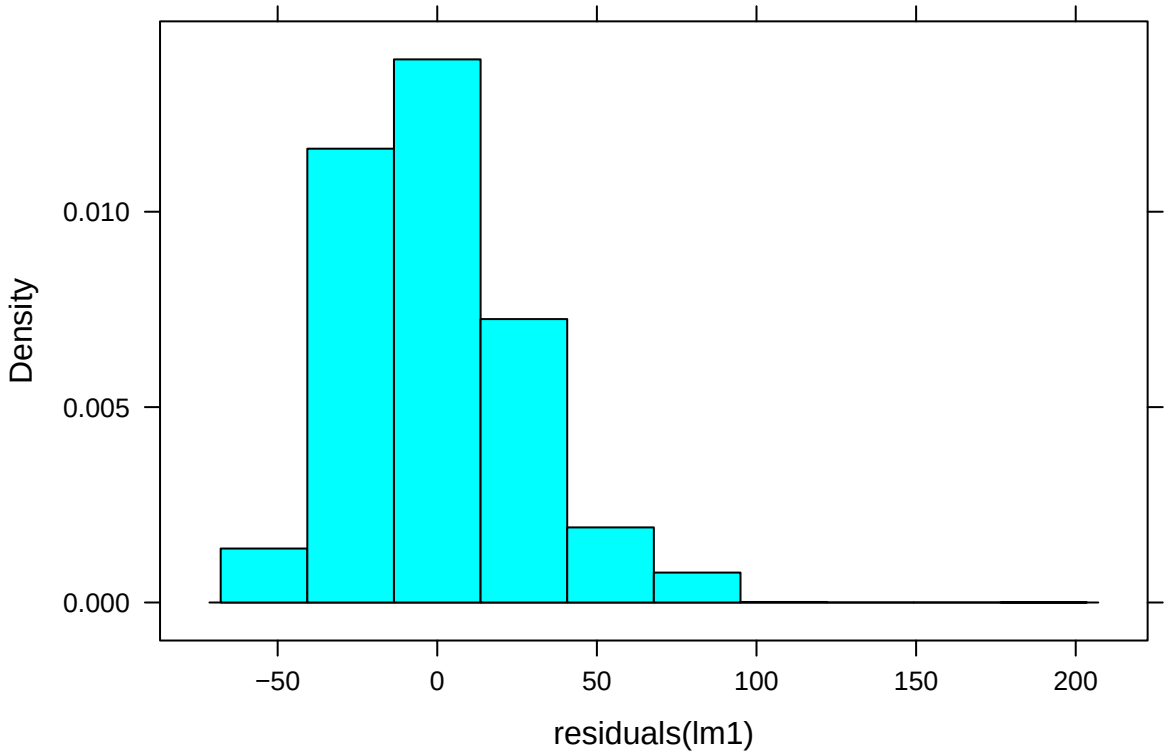
```
par(mfrow=c(5, 2))
histogram(~ residuals(lm0), nint=10)
```



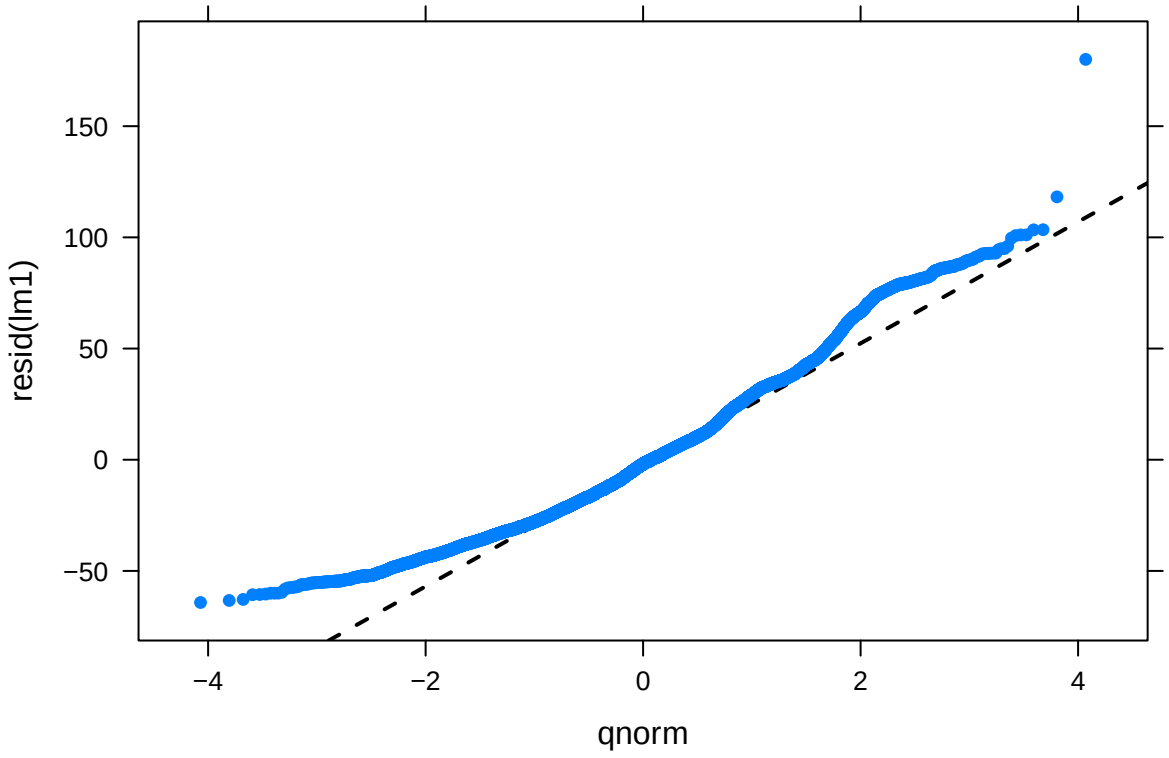

```
xqqmath(~resid(lm0))
```



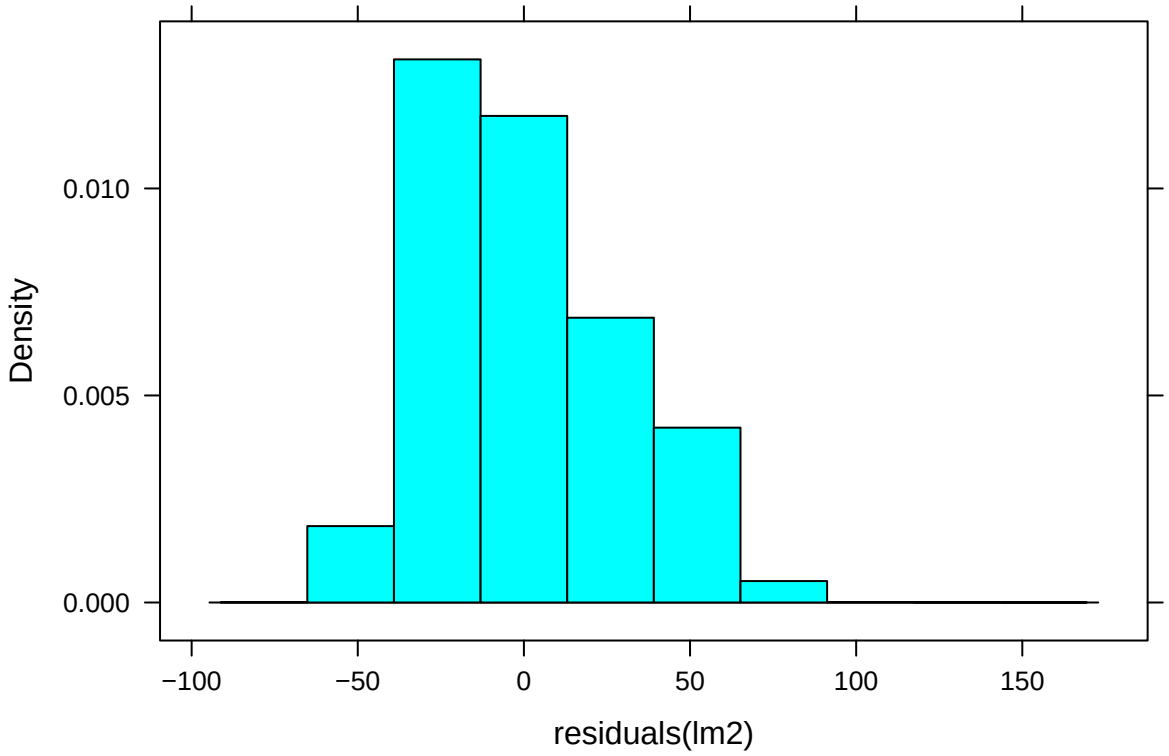

```
histogram(~ residuals(lm1), nint=10)
```



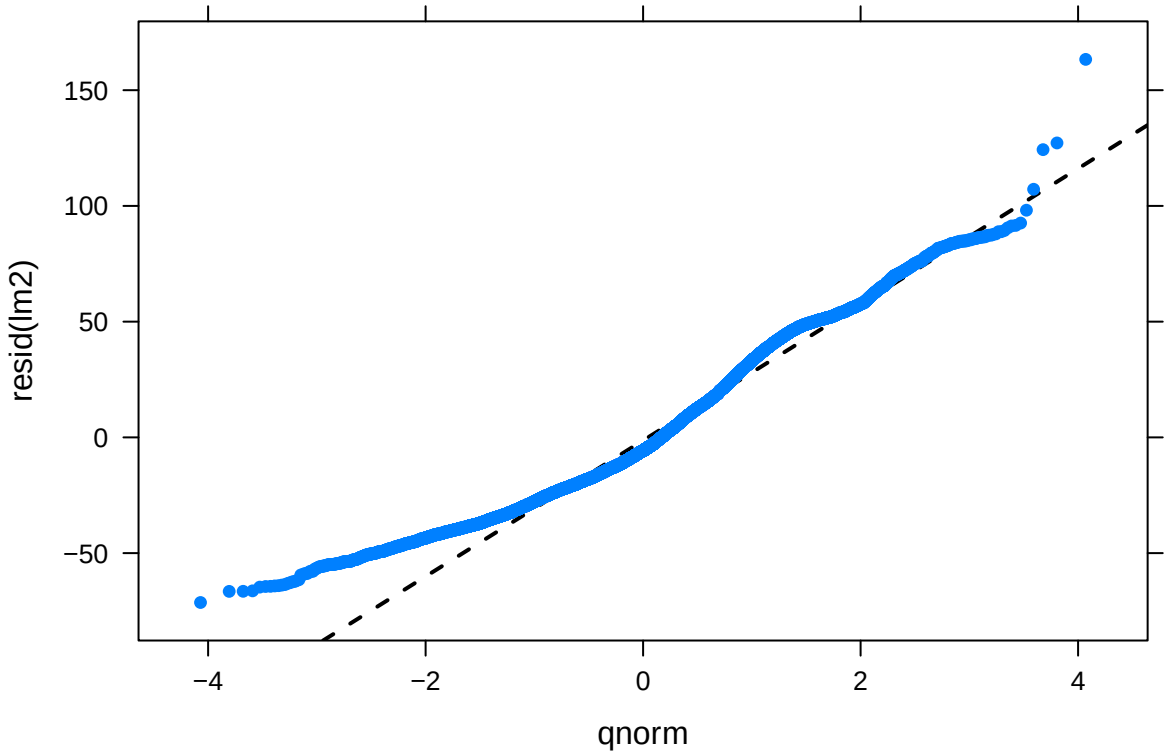

```
xqqmath(~resid(lm1))
```



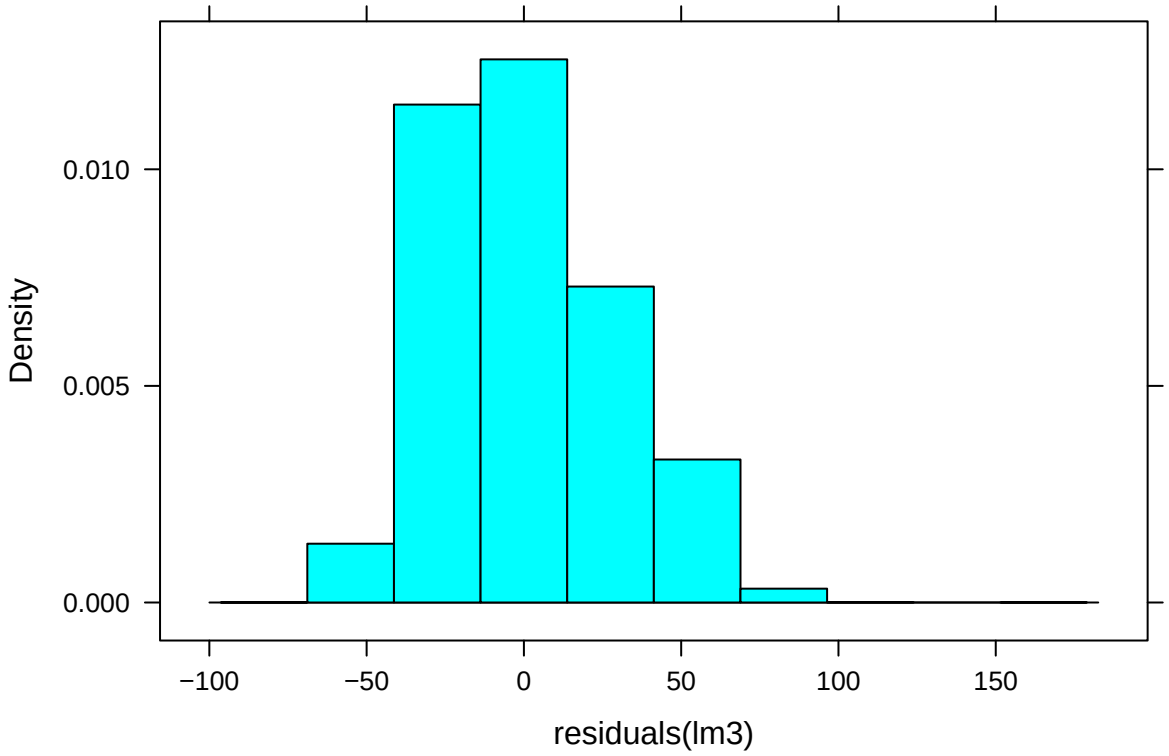

```
histogram(~ residuals(lm2), nint=10)
```



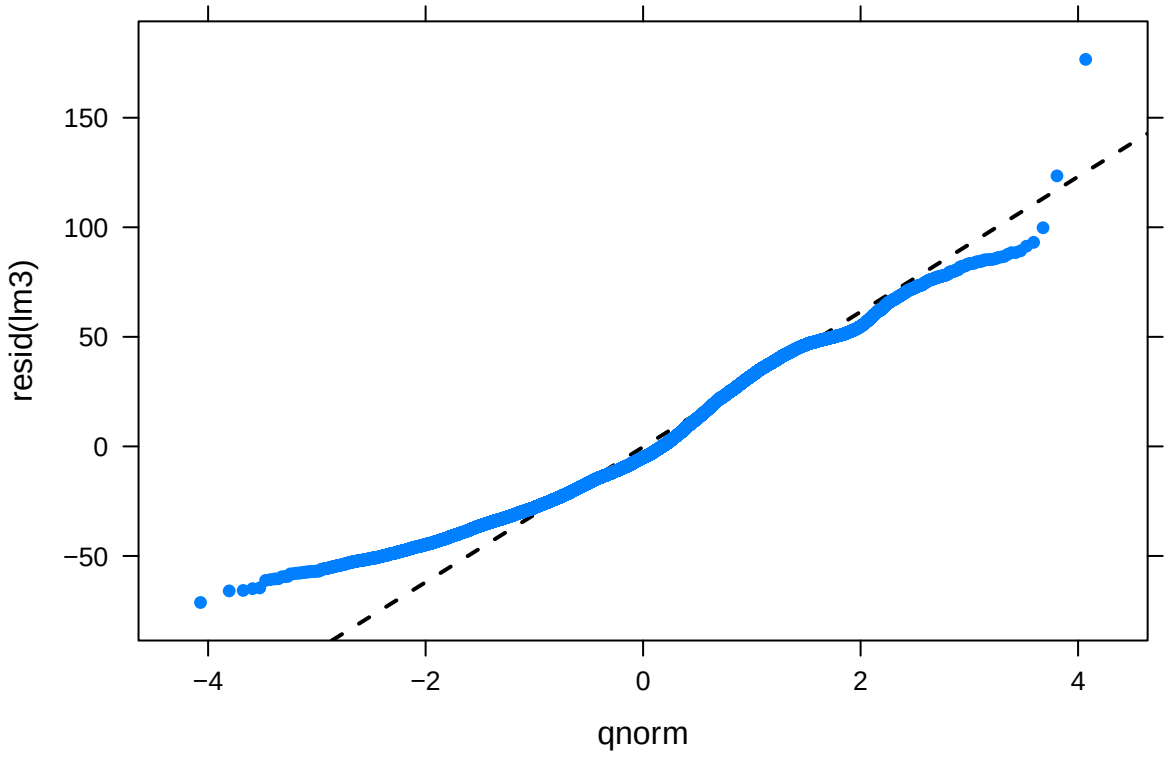

```
xqqmath(~resid(lm2))
```



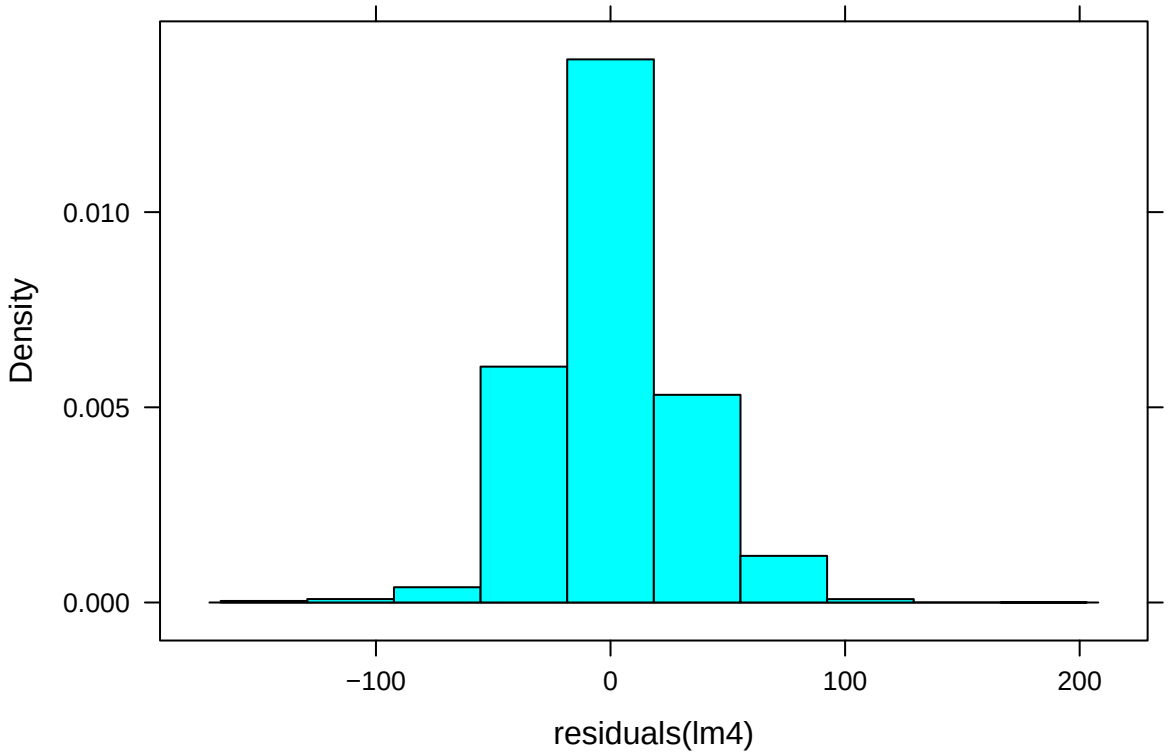

```
histogram(~ residuals(lm3), nint=10)
```



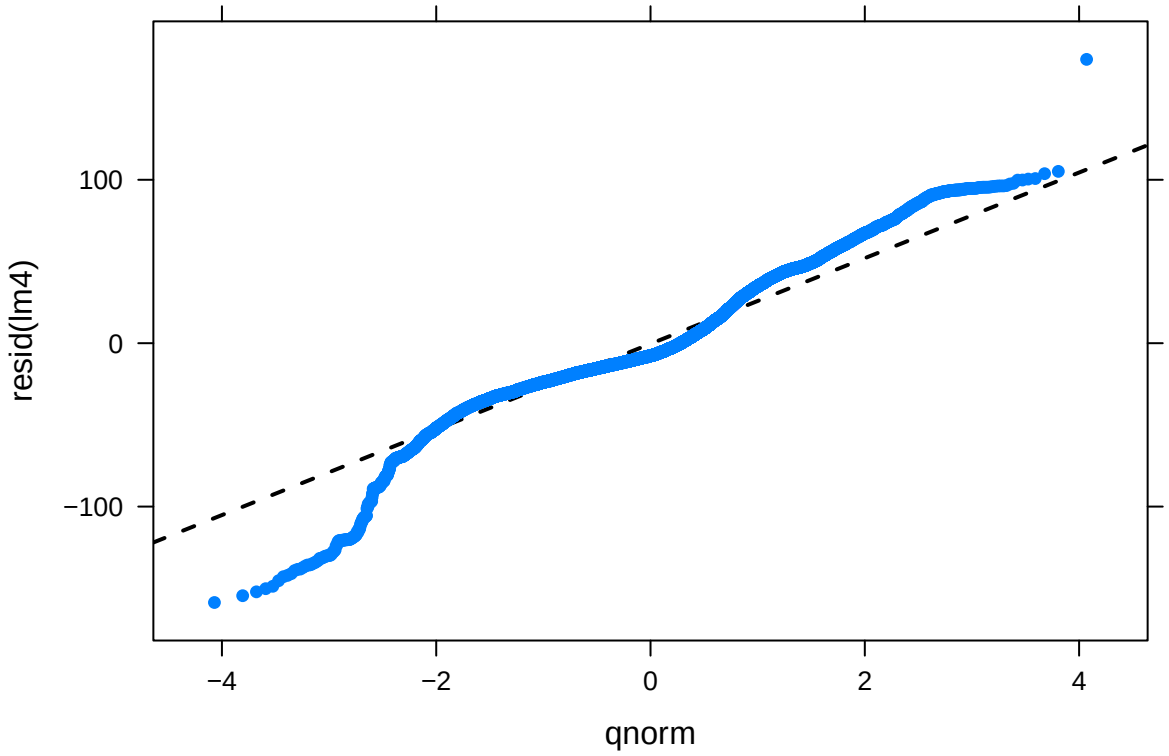

```
xqqmath(~resid(lm3))
```




```
histogram(~ residuals(lm4), nint=10)
```




```
xqqmath(~resid(lm4))
```

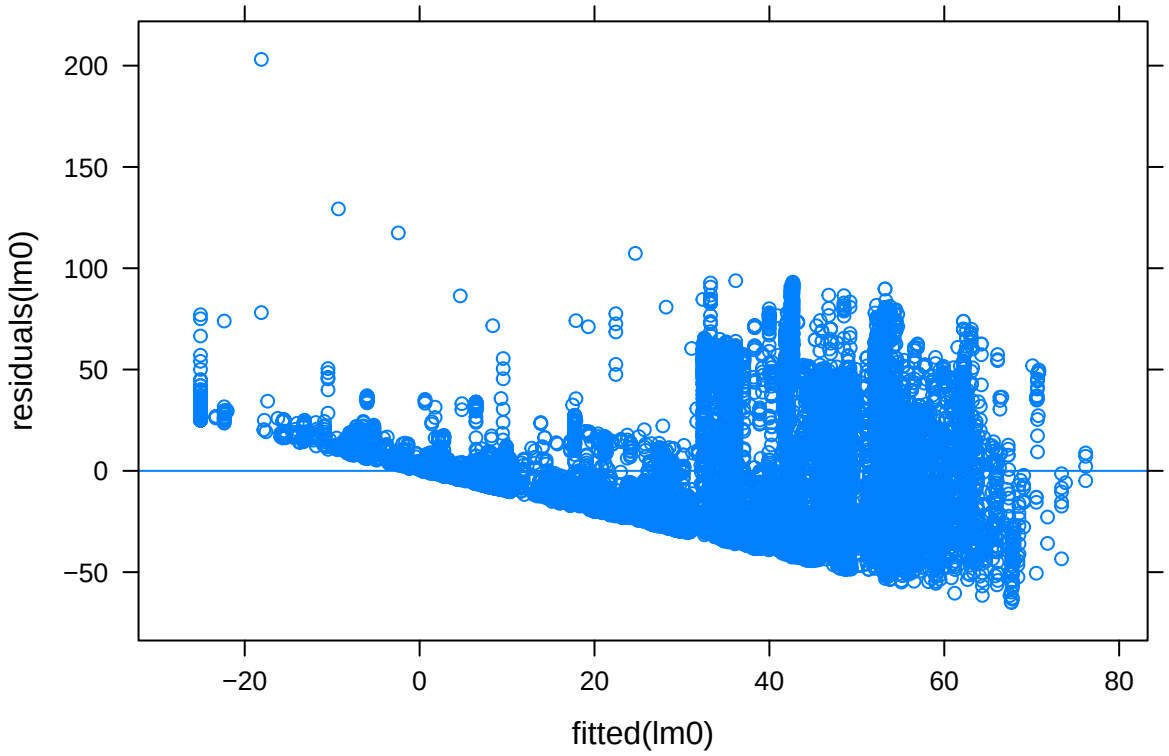


Histogram : Can assume Normal: 4 Not sure due to skewness: 0, 1, 2, 3

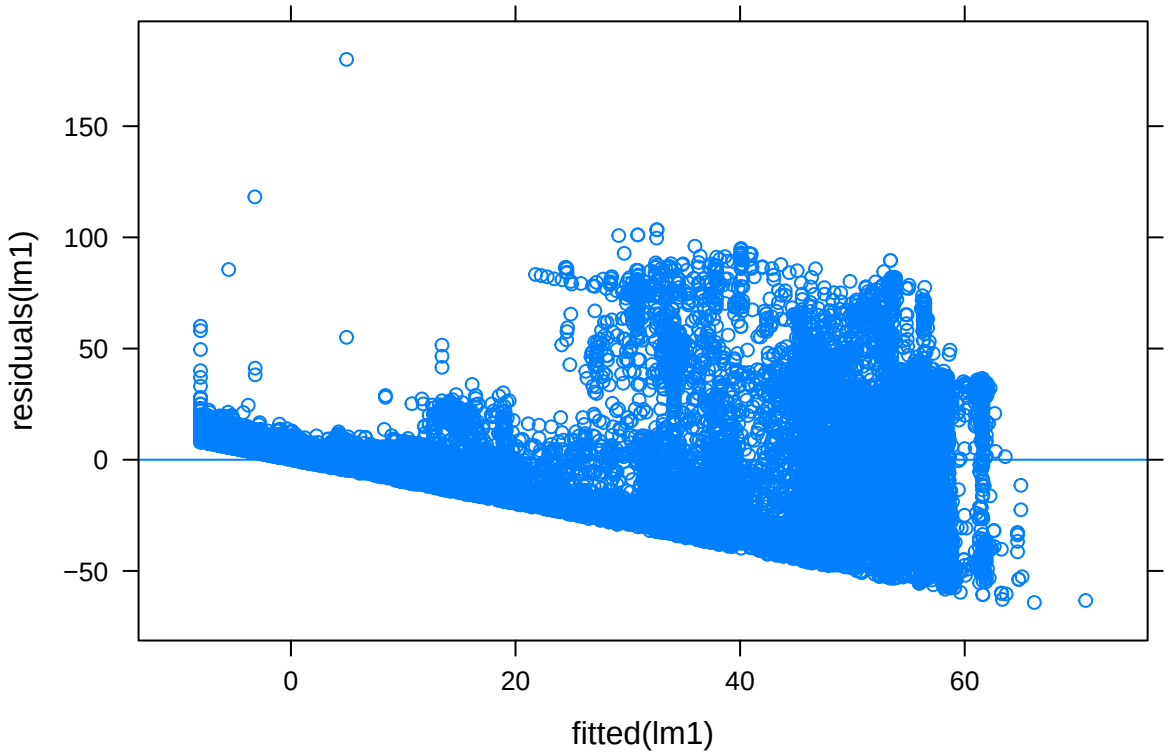
QQPlot : Perfect Normal: None Moderate Normal : All

4. Equal Variance Assumption:

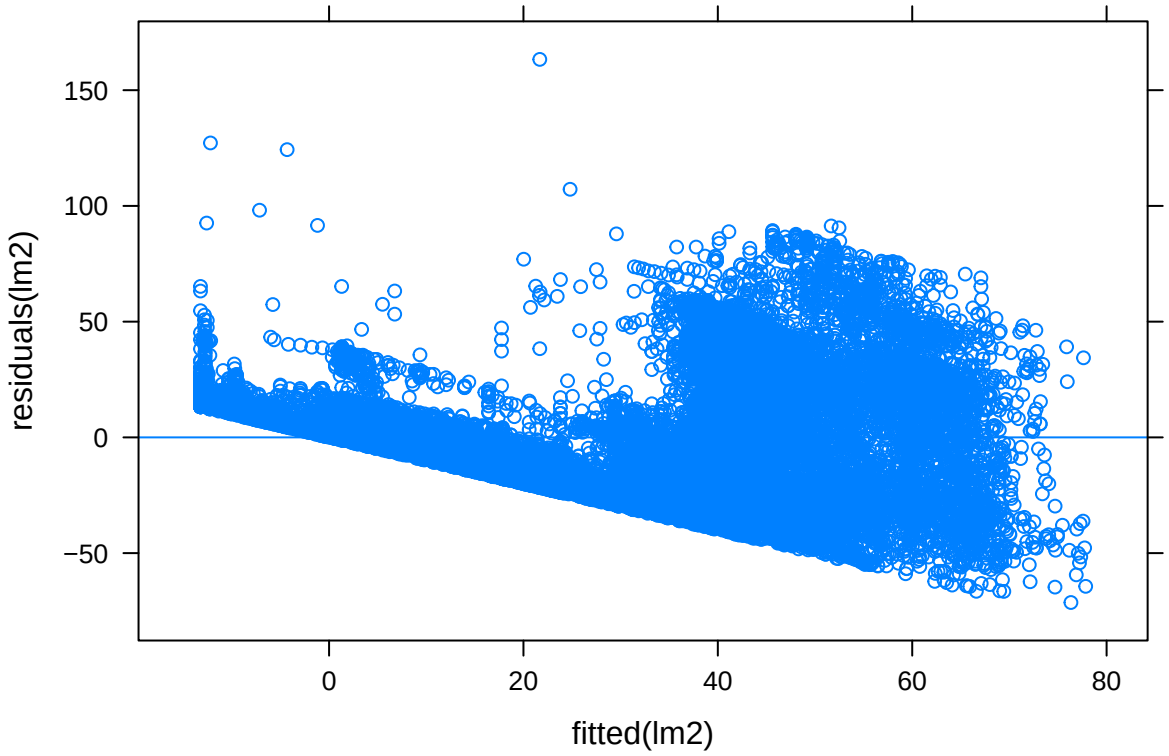
```
xyplot(residuals(lm0) ~ fitted(lm0), type=c("p", "r"))
```



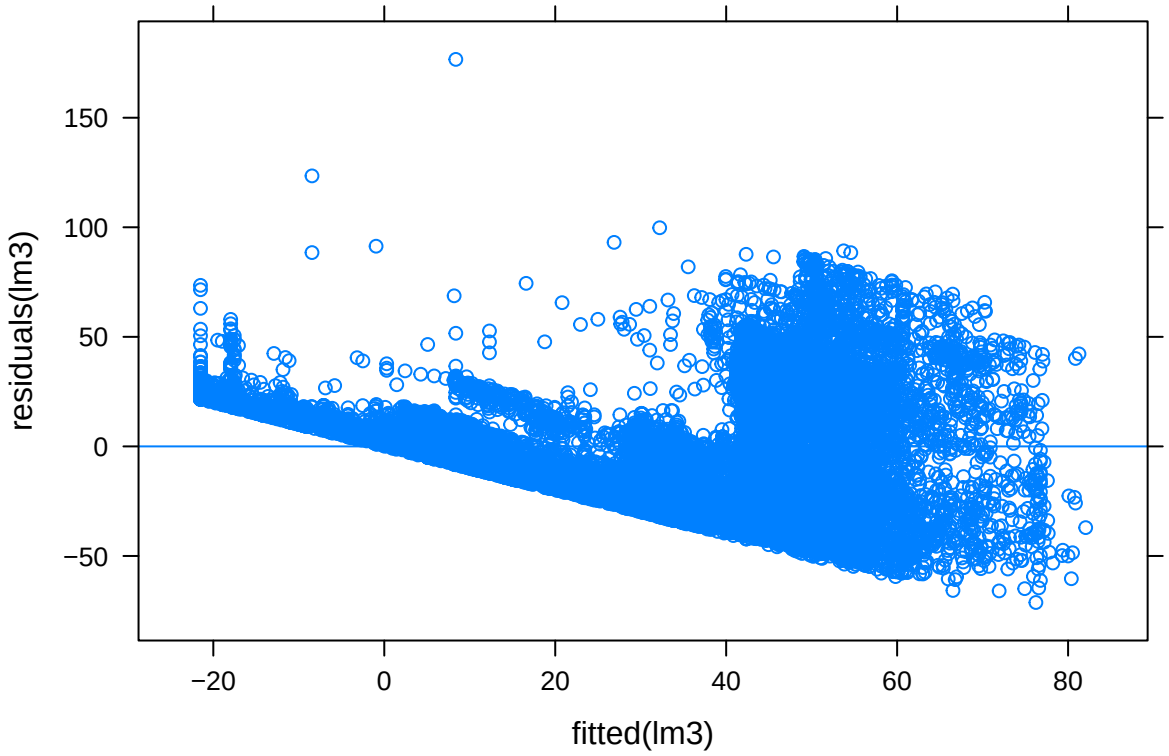

```
xyplot(residuals(lm1) ~ fitted(lm1), type=c("p", "r"))
```



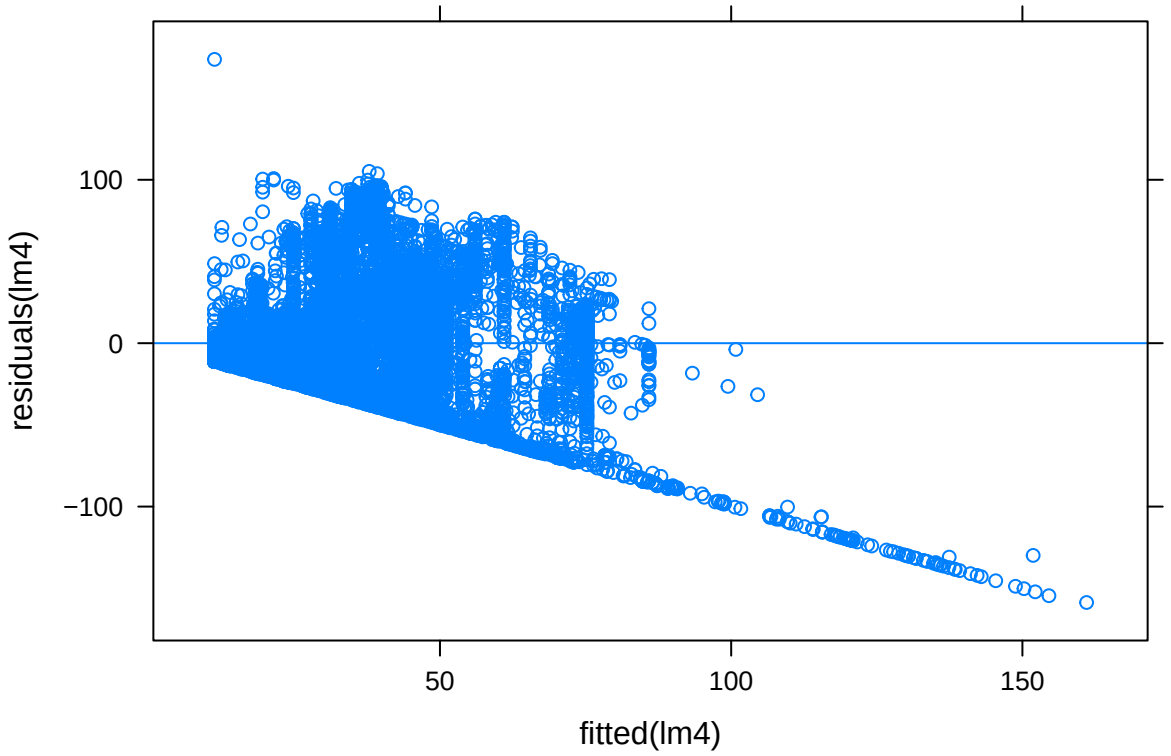

```
xyplot(residuals(lm2) ~ fitted(lm2), type=c("p", "r"))
```




```
xyplot(residuals(lm3) ~ fitted(lm3), type=c("p", "r"))
```




```
xyplot(residuals(lm4) ~ fitted(lm4), type=c("p", "r"))
```



Since the residuals do not look fairly evenly scattered about the horizontal line of 0 in any of the plots, are assumption is not getting justified for this dataset (tried many variables). Also the R-squared is not very great (~0.30-0.34 for all graphs).

We should not fit regression on this dataset.

F-test Look at the tables and write results. Mention formulae.

t-test Look at the table and write results. Mention formulae.