Mini Project Linear Regression

Exercise: least squares regression

Use the /states.rds/ data set.

Fit a model predicting energy consumed per capita (energy) from the percentage of residents living in metropolitan areas (metro).

Be sure to:

1. Examine/plot the data before fitting the model
2. Print and interpret the model 'summary'
3. 'plot' the model to look for deviations from modeling assumptions

Load states.rds into a data frame and examine it's contents

```r
states <- readRDS("states.rds")
head(states)
```

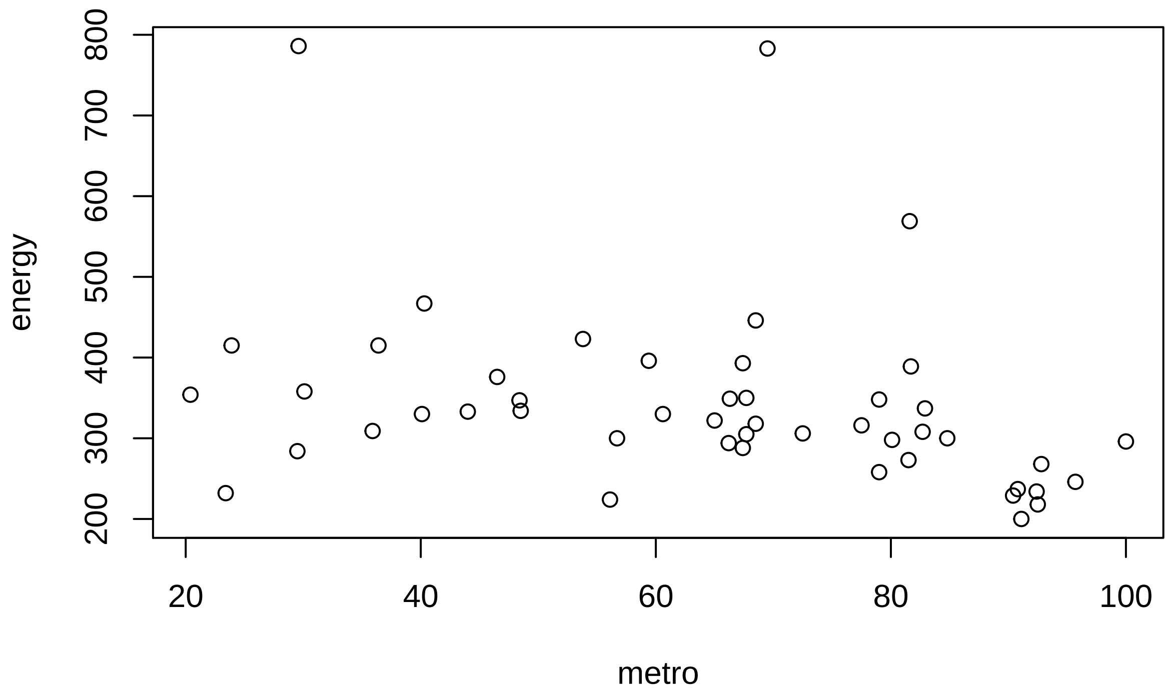
```
##        state region      pop   area density metro waste energy miles toxic
## 1    Alabama  South  4041000  52423   77.08  67.4  1.11    393  10.5 27.86
## 2     Alaska   West   550000 570374    0.96  41.1  0.91    991   7.2 37.41
## 3    Arizona   West  3665000 113642   32.25  79.0  0.79    258   9.7 19.65
## 4   Arkansas  South  2351000  52075   45.15  40.1  0.85    330   8.9 24.60
## 5 California   West 29760000 155973  190.80  95.7  1.51    246   8.7  3.26
## 6   Colorado   West  3294000 103730   31.76  81.5  0.73    273   8.3  2.25
##   green house senate csat vsat msat percent expense income high college
## 1 29.25    30     10  991  476  515       8    3627 27.498 66.9    15.7
## 2    NA     0     20  920  439  481      41    8330 48.254 86.6    23.0
## 3 18.37    13     33  932  442  490      26    4309 32.093 78.7    20.3
## 4 26.04    25     37 1005  482  523       6    3700 24.643 66.3    13.3
## 5 15.65    50     47  897  415  482      47    4491 41.716 76.2    23.4
## 6 21.89    36     58  959  453  506      29    5064 35.123 84.4    27.0
```

Examine and plot the data

```r
states.model1 <- subset(na.omit(states), select = c("metro", "energy"))
summary(states.model1)
```

```
##      metro            energy     
##  Min.   : 20.40   Min.   :200.0  
##  1st Qu.: 47.92   1st Qu.:287.0  
##  Median : 67.55   Median :320.0  
##  Mean   : 64.31   Mean   :343.6  
##  3rd Qu.: 81.62   3rd Qu.:362.5  
##  Max.   :100.00   Max.   :786.0  
```

```r
plot(states.model1)
```




```r
cor(states.model1)
```

```
##              metro     energy
## metro   1.0000000 -0.3116753
## energy -0.3116753  1.0000000
```

Metro and energy are not strongly correlated with a value of -.311

Print and interpret the model summary

```r
model1 <- lm(energy ~ metro, data = states.model1)
summary(model1)
```

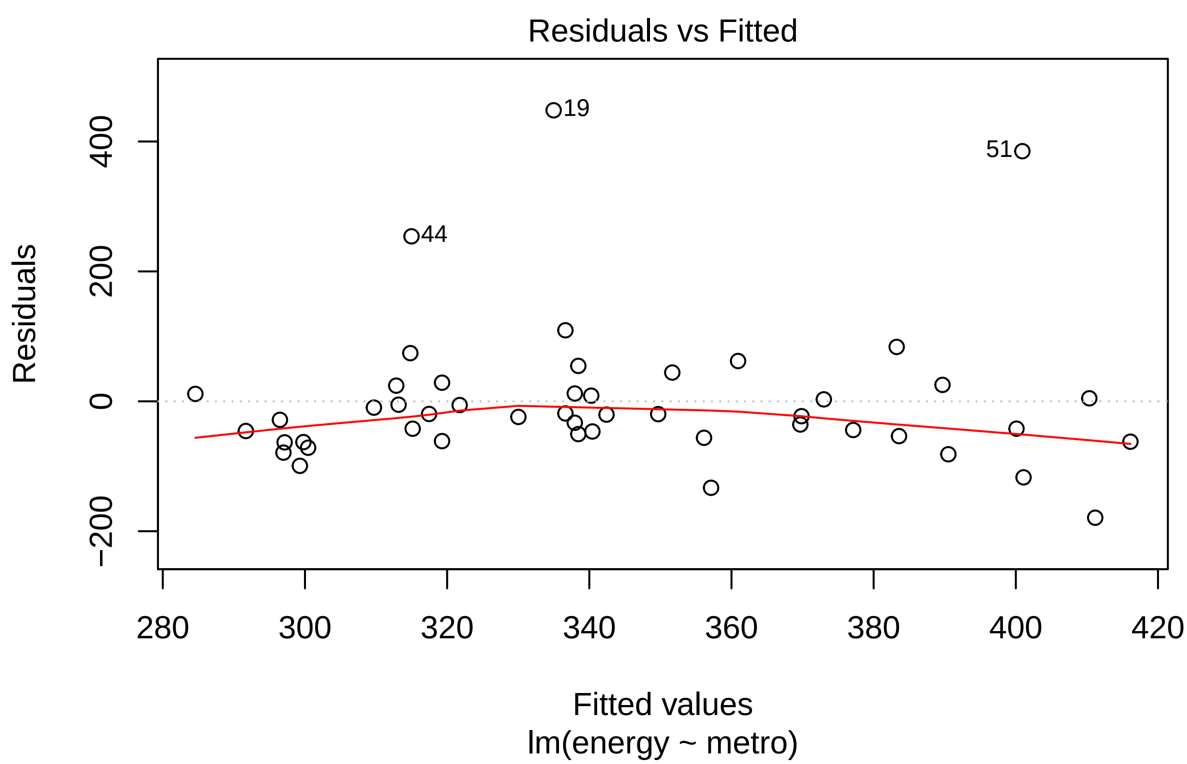
```
##
## Call:
## lm(formula = energy ~ metro, data = states.model1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -179.17  -54.21  -21.64   15.07  448.02
```

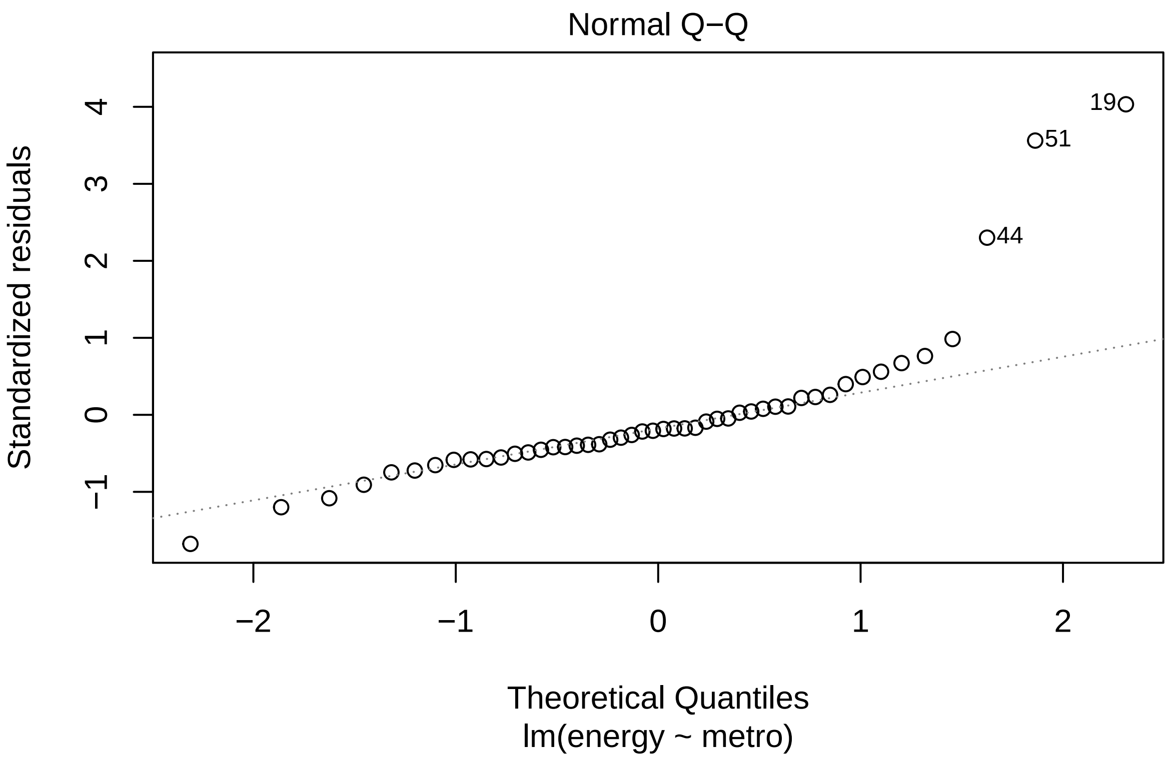
```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 449.8382    50.4472   8.917 1.37e-11 ***
## metro        -1.6526     0.7428  -2.225    0.031 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 112.3 on 46 degrees of freedom
## Multiple R-squared:  0.09714,    Adjusted R-squared:  0.07751
## F-statistic: 4.949 on 1 and 46 DF,  p-value: 0.03105
```

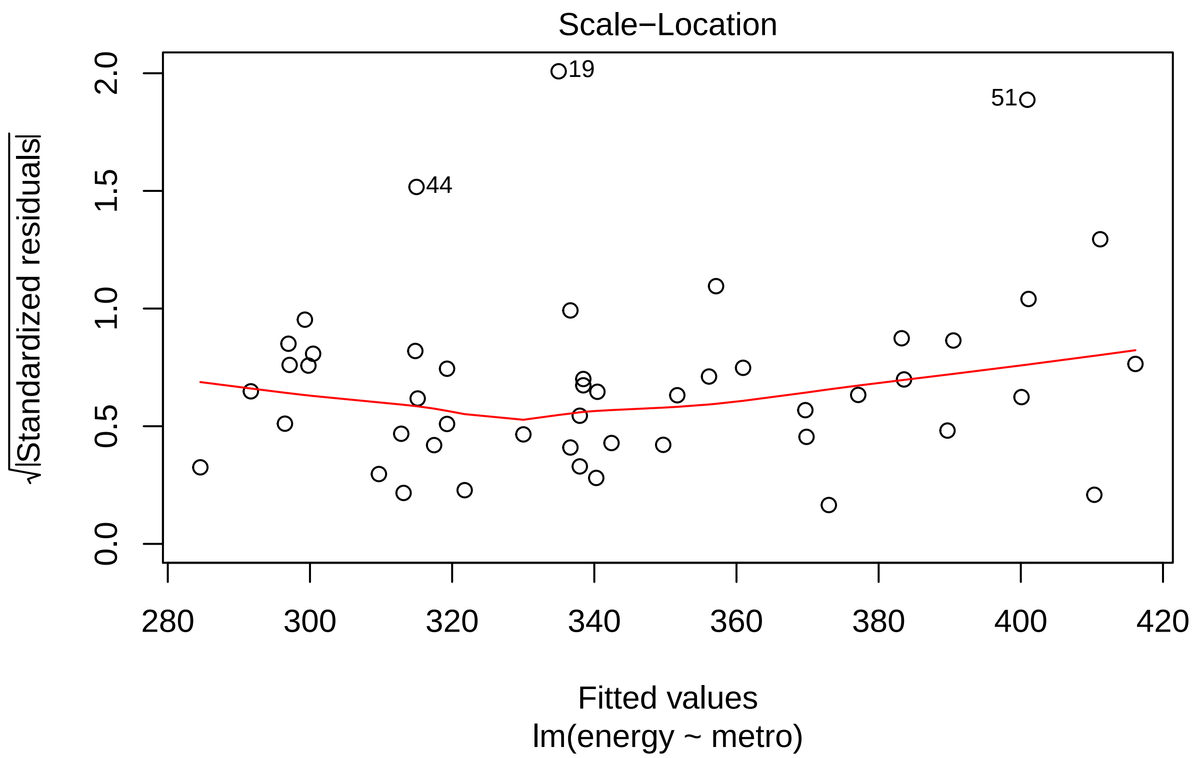
The P value for metro of .031 indicates it is a good predictor of energy, but with an R squared value of .078 there is a lot of error and not a good predictive model

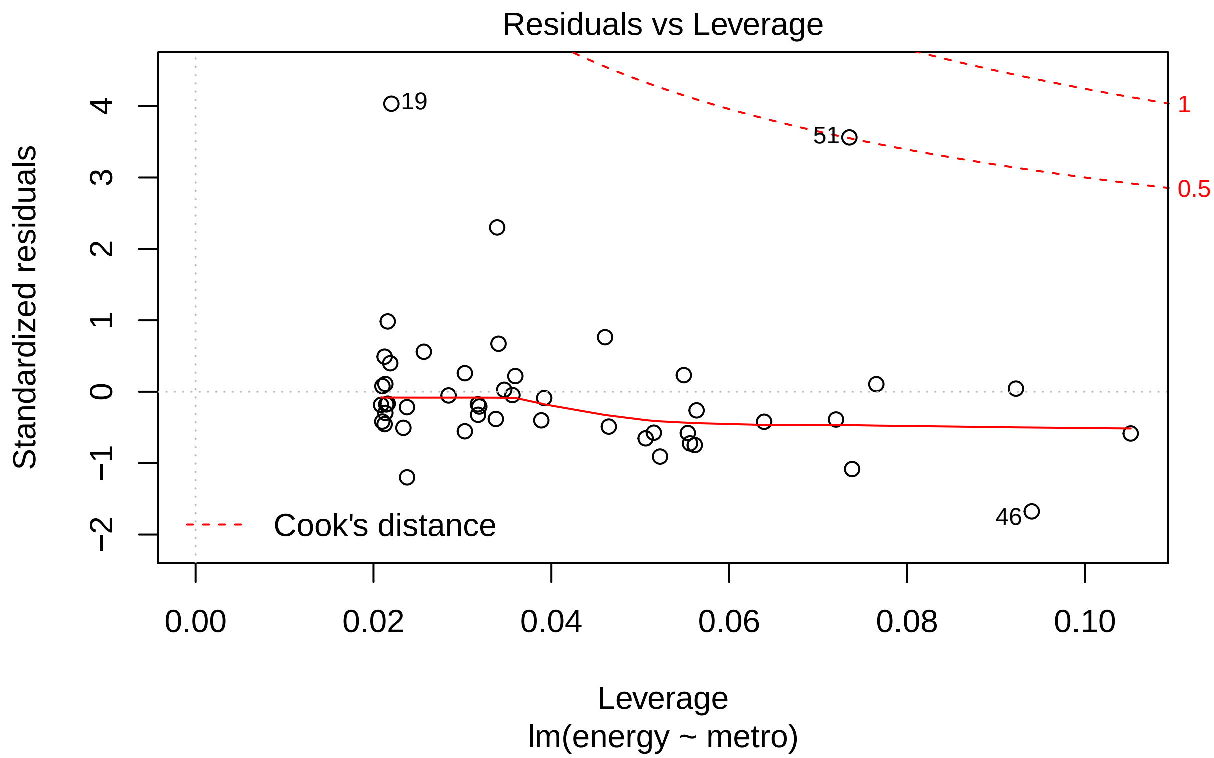
Plot the model

```r
plot(model1)
```




Normal Q−Q
Standardized residuals
4
3
2
1
0
−1
19
51
44
−2
−1
0
1
2
Theoretical Quantiles
lm(energy ~ metro)


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
19
51
44
280
300
320
340
360
380
400
420
Fitted values
lm(energy ~ metro)



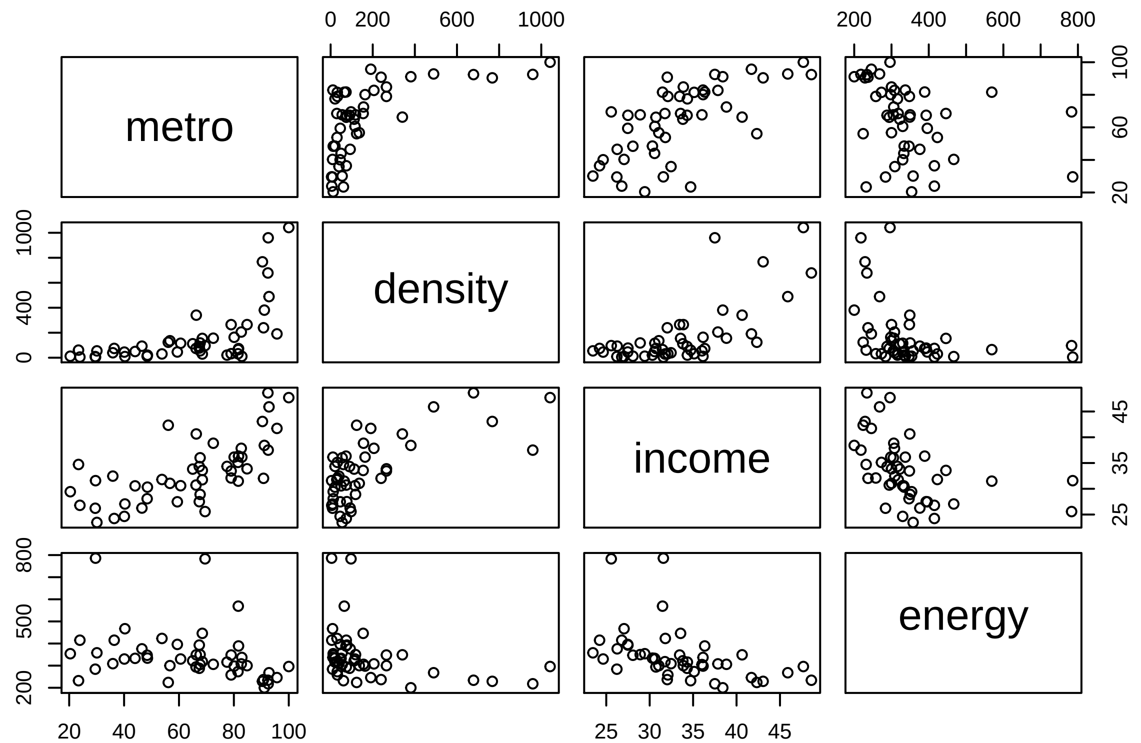
Select one or more additional predictors to add to your model and repeat steps 1-3.

Add density and income to the model

```r
states.model2 <- subset(na.omit(states), select = c("metro", "density", "income", "energy"))
summary(states.model2)
```

```
##      metro          density           income          energy     
##  Min.   : 20.40   Min.   :   4.68   Min.   :23.46   Min.   :200.0  
##  1st Qu.: 47.92   1st Qu.:  32.13   1st Qu.:29.30   1st Qu.:287.0  
##  Median : 67.55   Median :  75.76   Median :32.28   Median :320.0  
##  Mean   : 64.31   Mean   : 169.35   Mean   :33.38   Mean   :343.6  
##  3rd Qu.: 81.62   3rd Qu.: 170.41   3rd Qu.:36.20   3rd Qu.:362.5  
##  Max.   :100.00   Max.   :1041.92   Max.   :48.62   Max.   :786.0
```

```r
plot(states.model2)
```




```r
cor(states.model2)
```

```
##              metro    density     income     energy
## metro    1.0000000  0.5961558  0.6777118 -0.3116753
## density  0.5961558  1.0000000  0.6887342 -0.3432301
## income   0.6777118  0.6887342  1.0000000 -0.4483793
## energy  -0.3116753 -0.3432301 -0.4483793  1.0000000
```

None of the variables are highly correlated with energy

```r
model2 <- lm(energy ~ metro + density + income, data = states.model2)
summary(model2)
```

```
##
## Call:
## lm(formula = energy ~ metro + density + income, data = states.model2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -120.55  -50.91  -25.10   11.02  423.21
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 609.913918 111.150581   5.487  1.9e-06 ***
```

```
## metro         0.006966   1.000204   0.007   0.9945
## density      -0.032218   0.093680  -0.344   0.7326
## income       -7.828189   4.048116  -1.934   0.0596 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 107.9 on 44 degrees of freedom
## Multiple R-squared:  0.2033, Adjusted R-squared:  0.149
## F-statistic: 3.743 on 3 and 44 DF,  p-value: 0.01765
```

```
anova(model1, model2)
```

```
## Analysis of Variance Table
##
## Model 1: energy ~ metro
## Model 2: energy ~ metro + density + income
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1     46 580411
## 2     44 512168  2     68244 2.9314 0.06381 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

In this model, none of the variables are good predictors of energy. Although our error has improved (0.077 - 0.149) it is still weak, and this is not a good model either. Based on the ANOVA test results, the second model is not significantly better than the first

Exercise: interactions and factors

Use the states data set.

1. Add on to the regression equation that you created in exercise 1 by generating an interaction term and testing the interaction.

```
model3 <- lm(energy ~ metro * density, data = states)
coef(summary(model3))
```

```
##                    Estimate  Std. Error   t value     Pr(>|t|)
## (Intercept)    514.10424265 72.513405133  7.089782 6.683817e-09
## metro           -1.72111951  1.135554671 -1.515664 1.364465e-01
## density         -1.43817898  0.911292059 -1.578176 1.213782e-01
## metro:density    0.01386147  0.009534258  1.453859 1.527747e-01
```

None of these interactions appear to be significant

2. Try adding region to the model. Are there significant differences across the four regions?

```r
model4 <- lm(energy ~ metro * density + region, data = states)
anova(model4)
```

```
## Analysis of Variance Table
##
## Response: energy
##                Df Sum Sq Mean Sq F value  Pr(>F)
## metro           1 123064  123064  6.6011 0.01374 *
## density         1  25837   25837  1.3859 0.24557
## region          3  80605   26868  1.4412 0.24400
## metro:density  1  35018   35018  1.8783 0.17763
## Residuals      43 801642   18643
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

There do not appear to be significant differences across regions