Regression

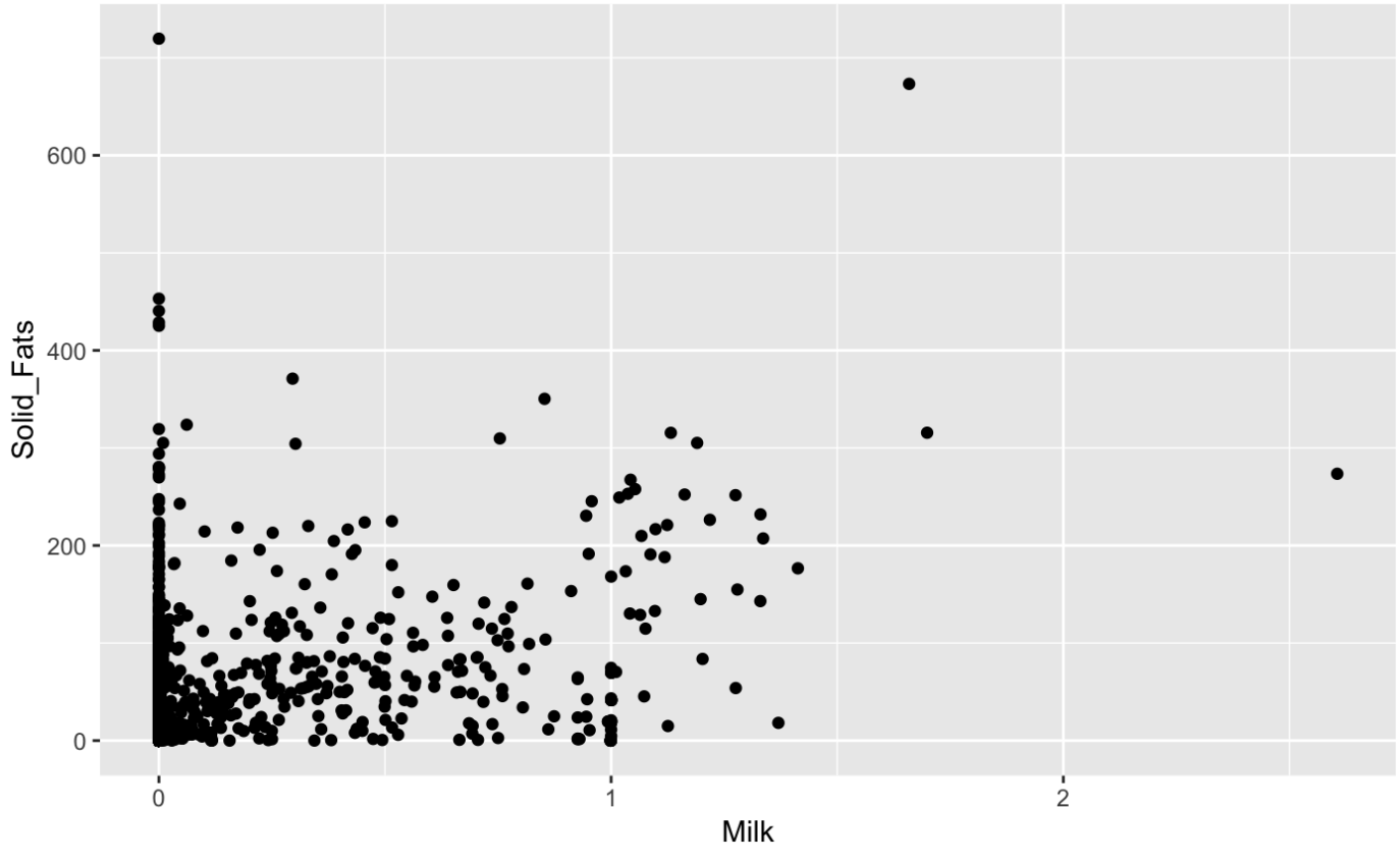
Code

Part One

I am checking to see if there is a correlation between the number of servings of dairy (this is denoted in the data as milk, so I am making an assumption), and calories from solid fat.

Hide

```
ggplot(nutrition, aes(x=Milk, y=Solid_Fats)) + geom_point()
```



Hide

```
cor(nutrition$Milk, nutrition$Solid_Fats)
```

```
[1] 0.406807
```

There seems to be a modest positive correlation between servings of dairy and calories from solid fat. Now, I will build a linear regression model. The independent variable is servings of dairy, and the dependent variable are calories from fat. In another words, the amount of calories from fat is somehow affected by the number of servings of dairy.

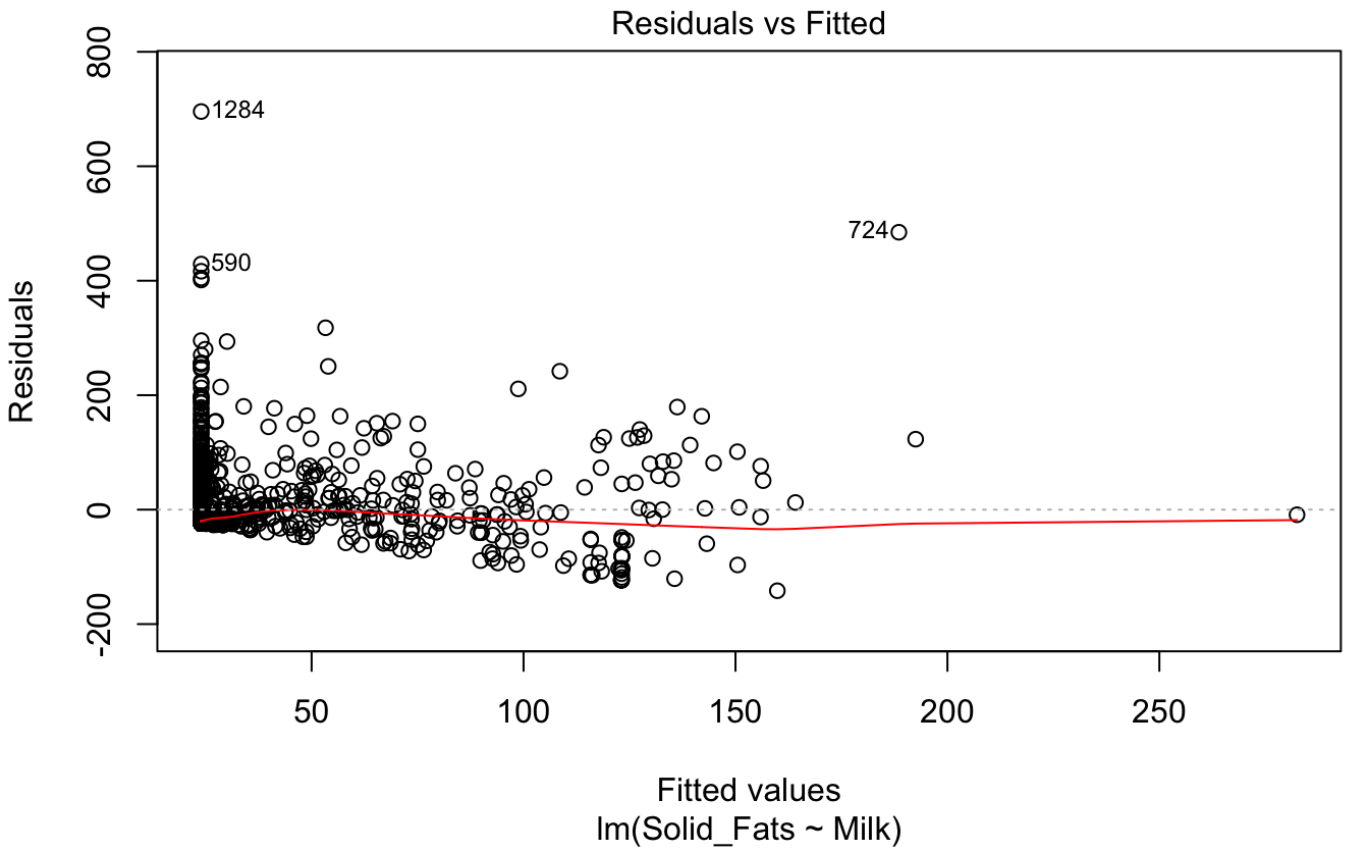
Hide

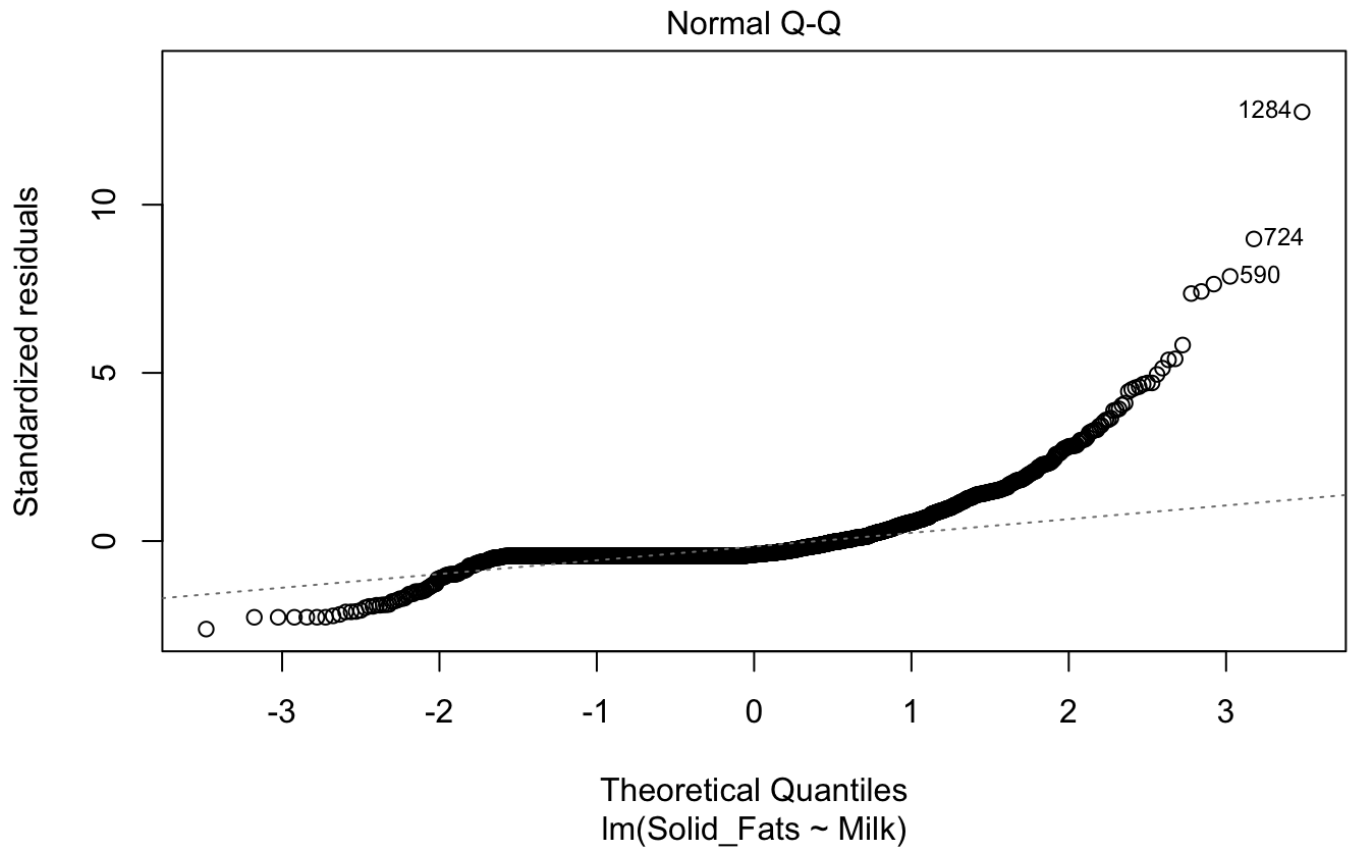
```
m1 <- lm(data=nutrition, Solid_Fats ~ Milk)
summary(m1)
```

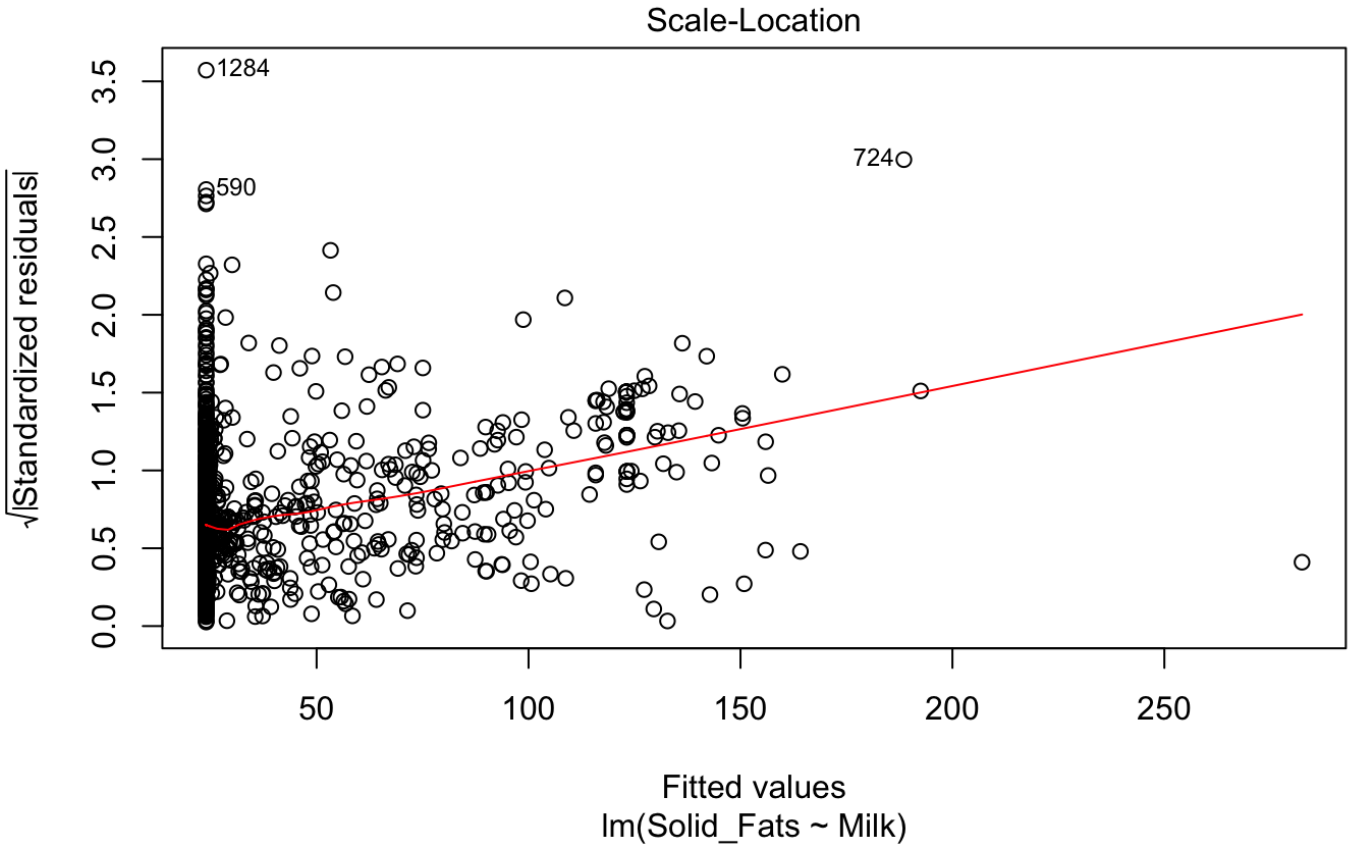
```

Call:
lm(formula = Solid_Fats ~ Milk, data = nutrition)

Residuals:
    Min      1Q  Median      3Q     Max
-141.71  -23.95  -21.58    6.05  695.65

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   23.953      1.285   18.64   <2e-16 ***
Milk          99.215      4.967   19.98   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 54.55 on 2012 degrees of freedom
Multiple R-squared:  0.1655,    Adjusted R-squared:  0.1651
F-statistic:   399 on 1 and 2012 DF,  p-value: < 2.2e-16
```

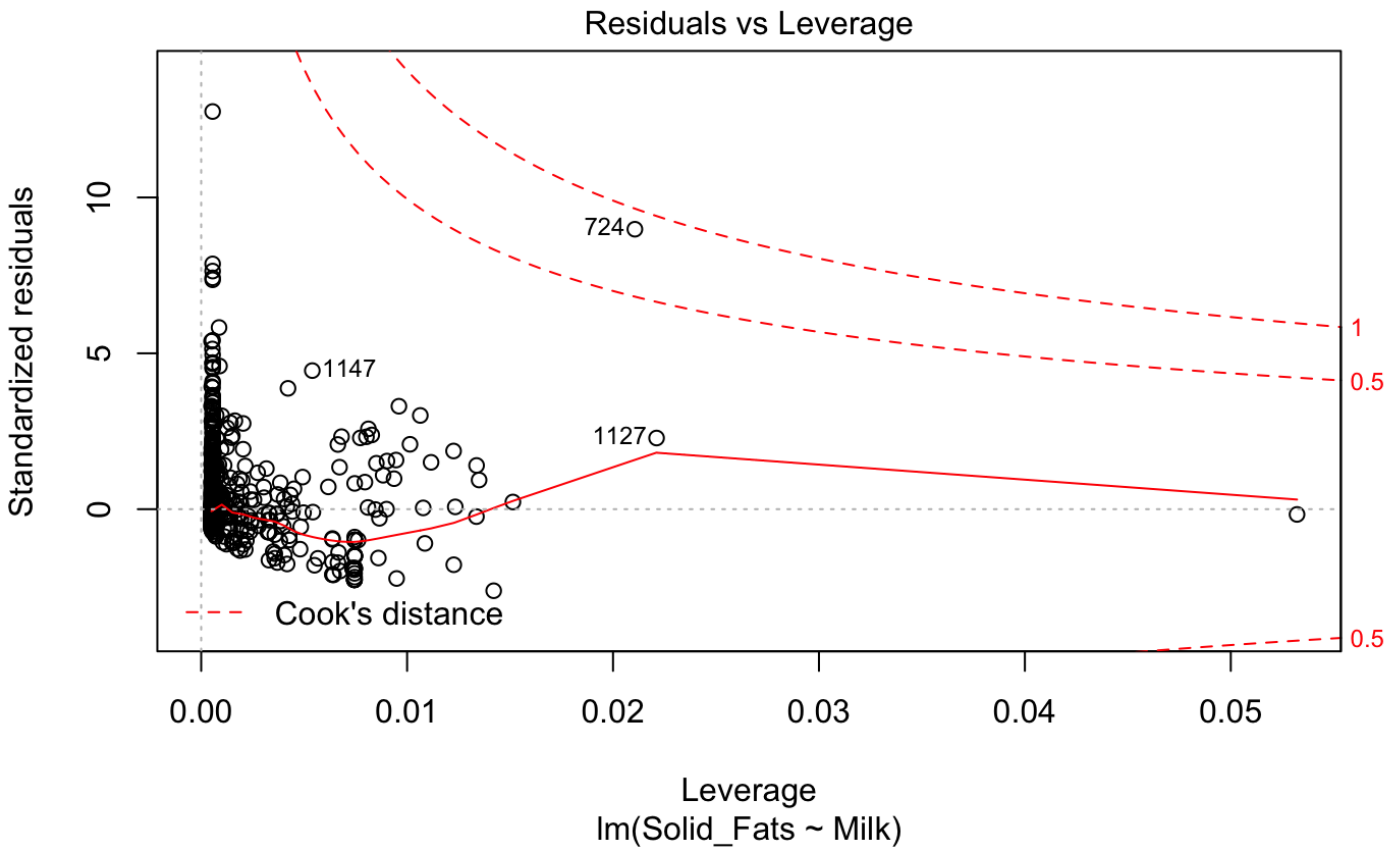
Hide

```
plot(m1)
```




Normal Q-Q
Standardized residuals
1284
724
590
10
5
0
-3
-2
-1
0
1
2
3
Theoretical Quantiles
lm(Solid_Fats ~ Milk)


Scale-Location
√|Standardized residuals|
1284
724
590
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
50
100
150
200
250
Fitted values
lm(Solid_Fats ~ Milk)



I'm not really sure how to interpret these results. It doesn't seem like there are any patterns.

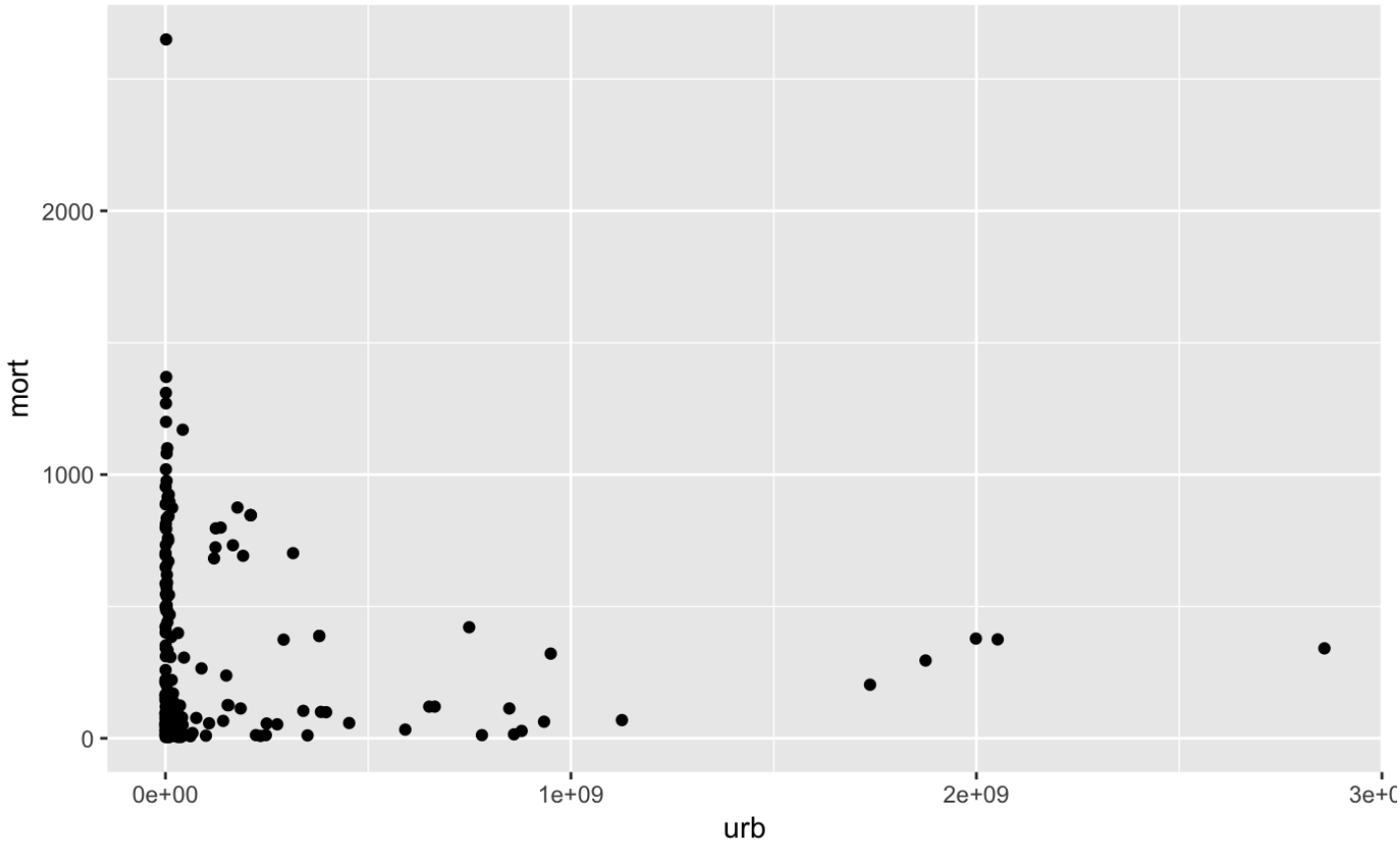
Part Two

I am comparing maternal mortality ratio with urban population.

My alternative hypothesis is that mortality is inversely correlated with urban population. In other words, countries with higher urban populations would have lower numbers of infant mortality.

The null hypothesis is that there is no correlation between mortality and urban population.

A major limitation of this dataset is that these numbers are absolute, not percentages. I don't have the overall population of the countries, so I can't normalize the data in terms of percentage of the entire population. This will probably skew the results and make them less meaningful.



This distribution suggests that most countries absolute mortality is between 0 and 1000. There are several countries with large urban populations and relatively lower mortality, which may represent richer, developed nations.

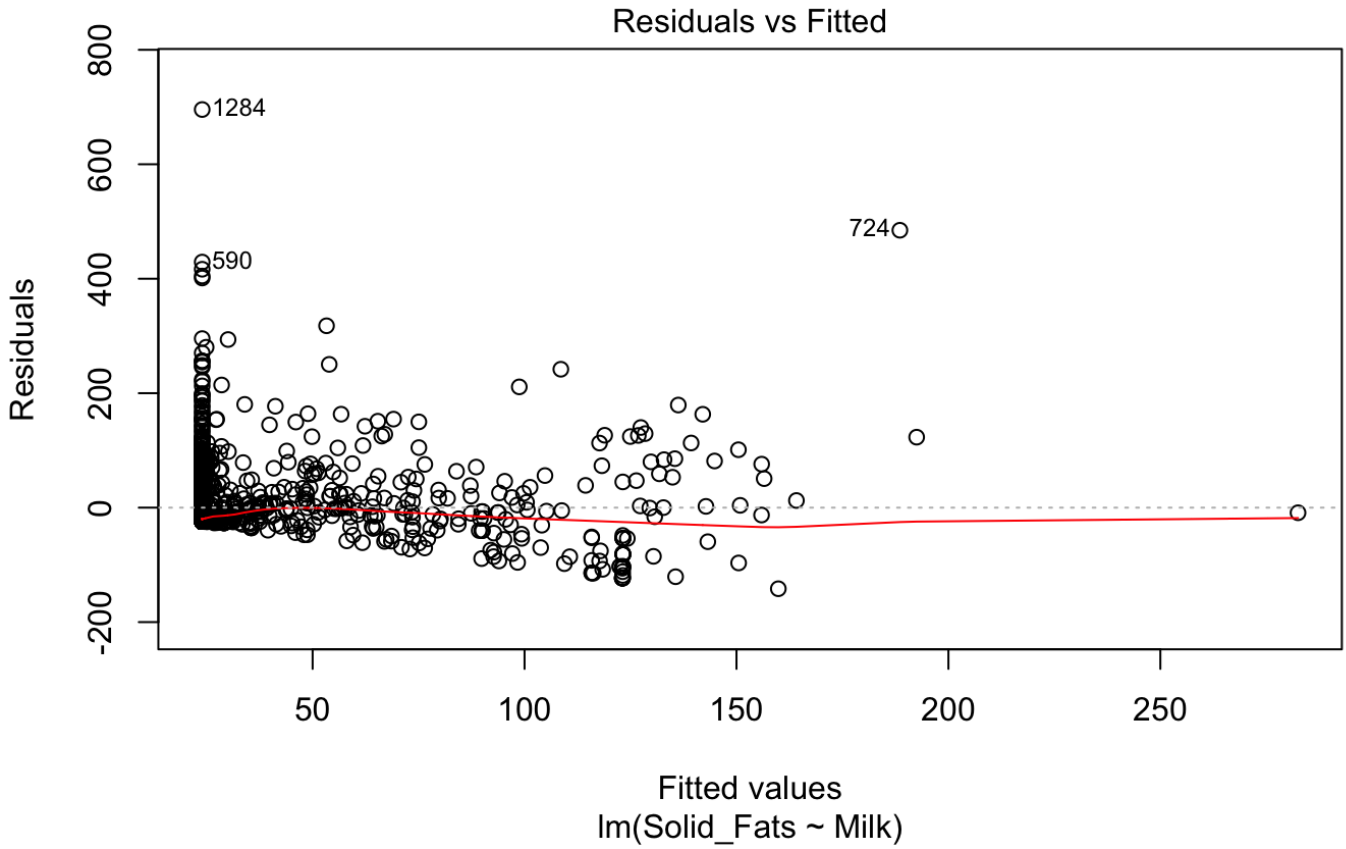
Hide

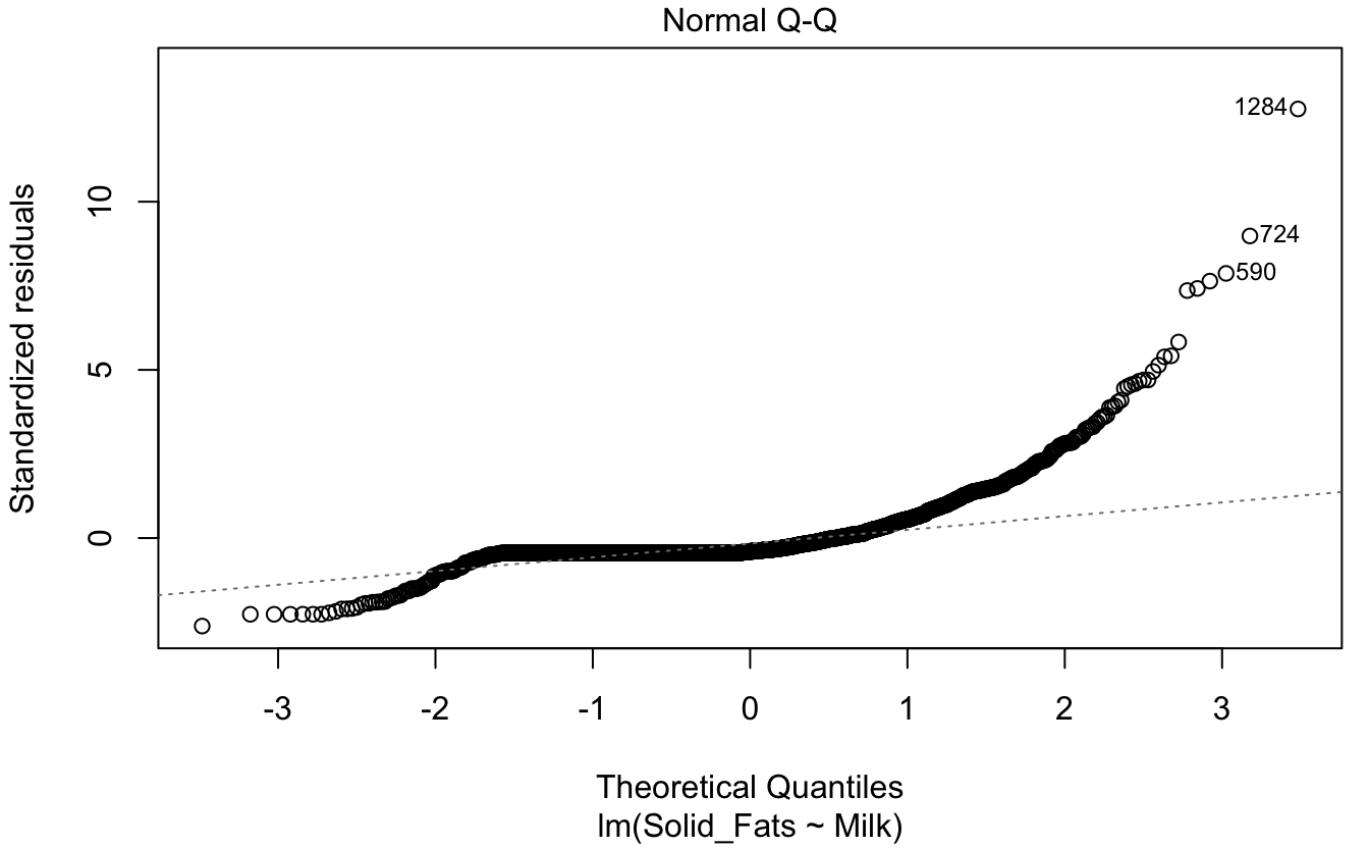
```
m2 <- lm(data=combined, mort ~ urb)
summary(m2)
```

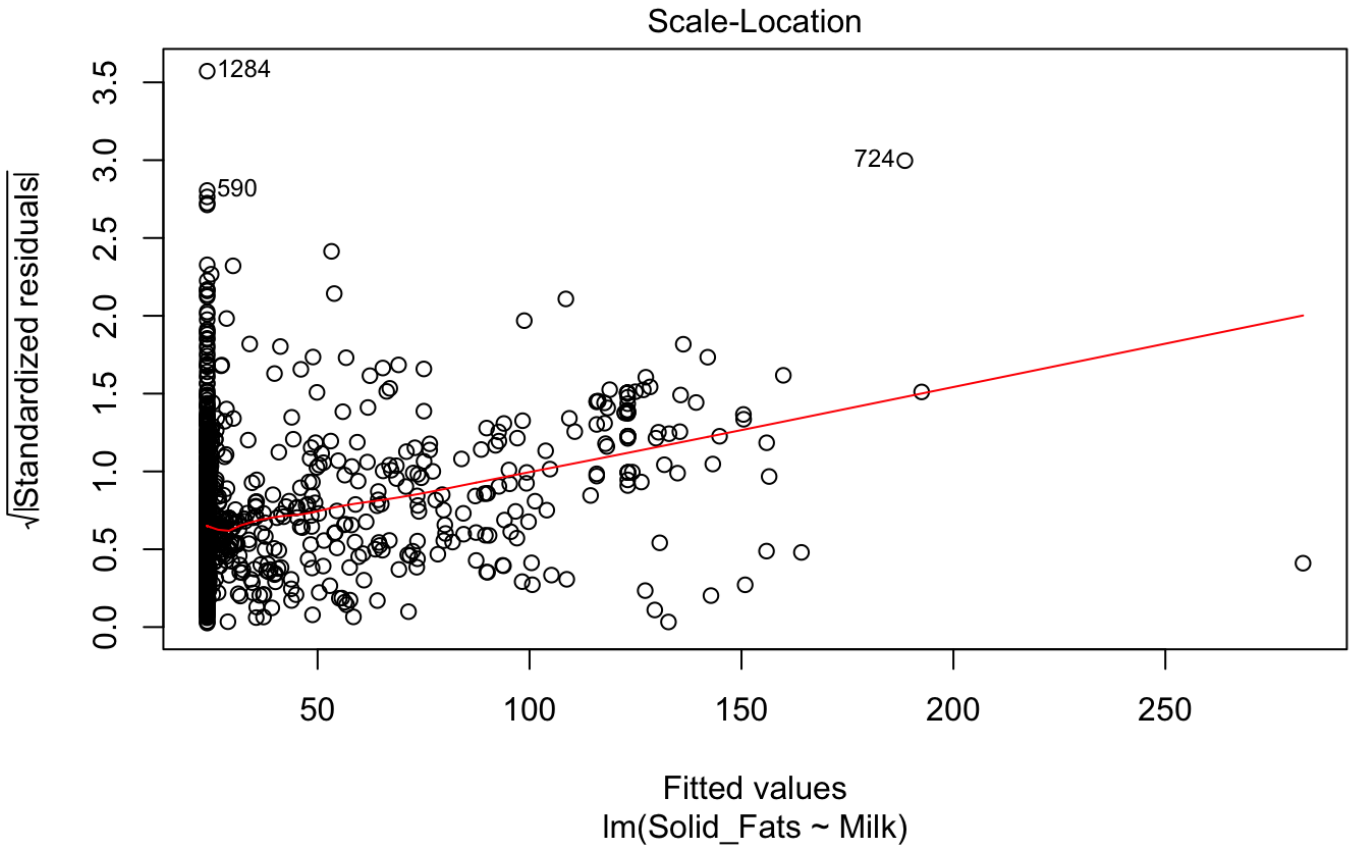
```
Call:
lm(formula = mort ~ urb, data = combined)

Residuals:
   Min     1Q Median     3Q    Max
-278.8 -252.0 -186.0  184.2 2367.0

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.830e+02  2.546e+01  11.117   <2e-16 ***
urb         -3.086e-08  6.750e-08  -0.457    0.648
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 363.9 on 227 degrees of freedom
Multiple R-squared:  0.00092,   Adjusted R-squared:  -0.003481
F-statistic: 0.209 on 1 and 227 DF,  p-value: 0.648
```

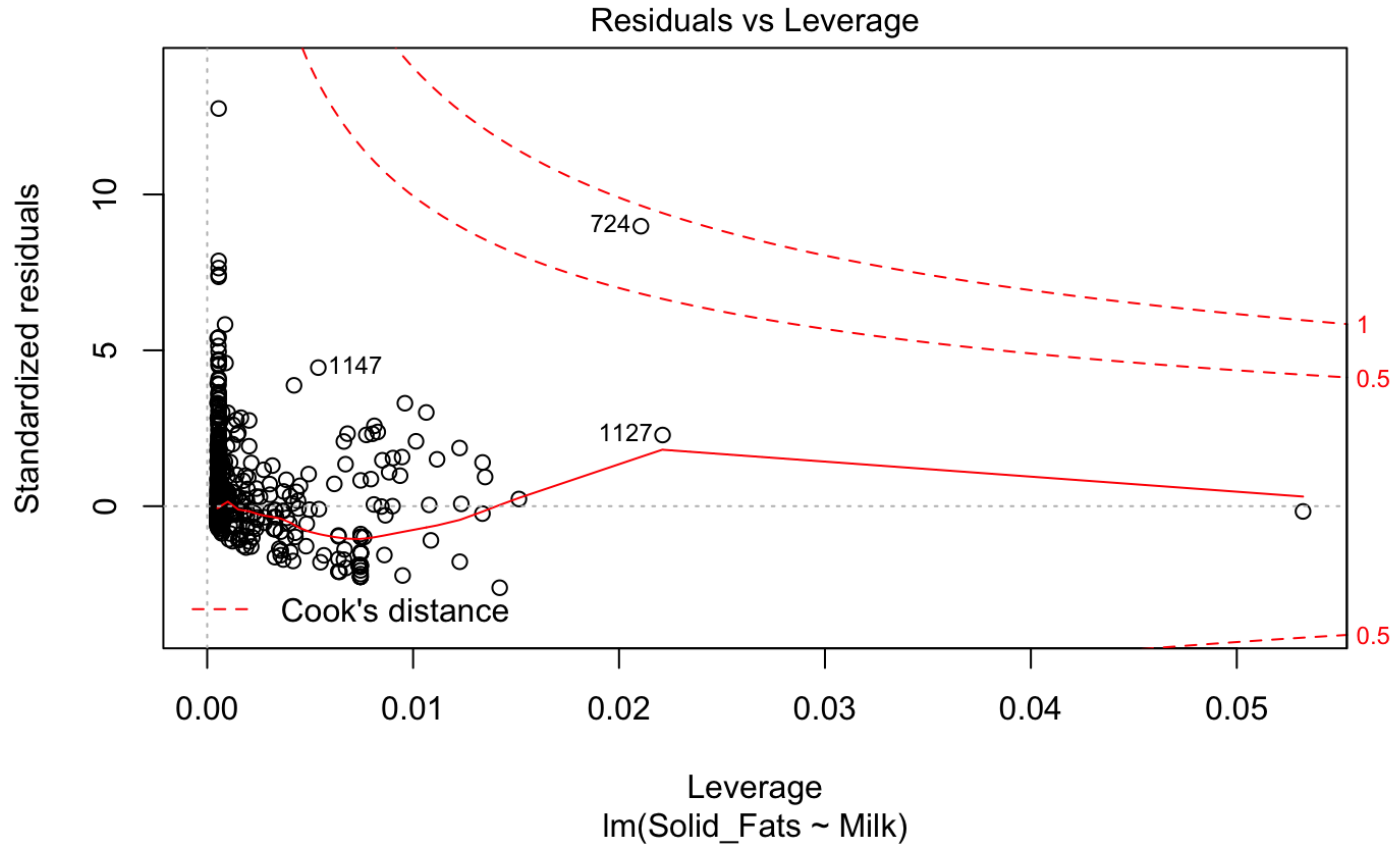
Hide

```
plot(m1)
```


Residuals vs Fitted
Residuals
Fitted values
lm(Solid_Fats ~ Milk)
1284
724
590
-200
0
200
400
600
800
50
100
150
200
250


Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(Solid_Fats ~ Milk)
1284
724
590
0
5
10
-3
-2
-1
0
1
2
3





There doesn't seem to be any correlations in the data, and thus the null hypothesis is not rejected. It would be helpful to normalize the data by percentage of the entire population and then rerun the analysis to see if there are any meaningful correlations.