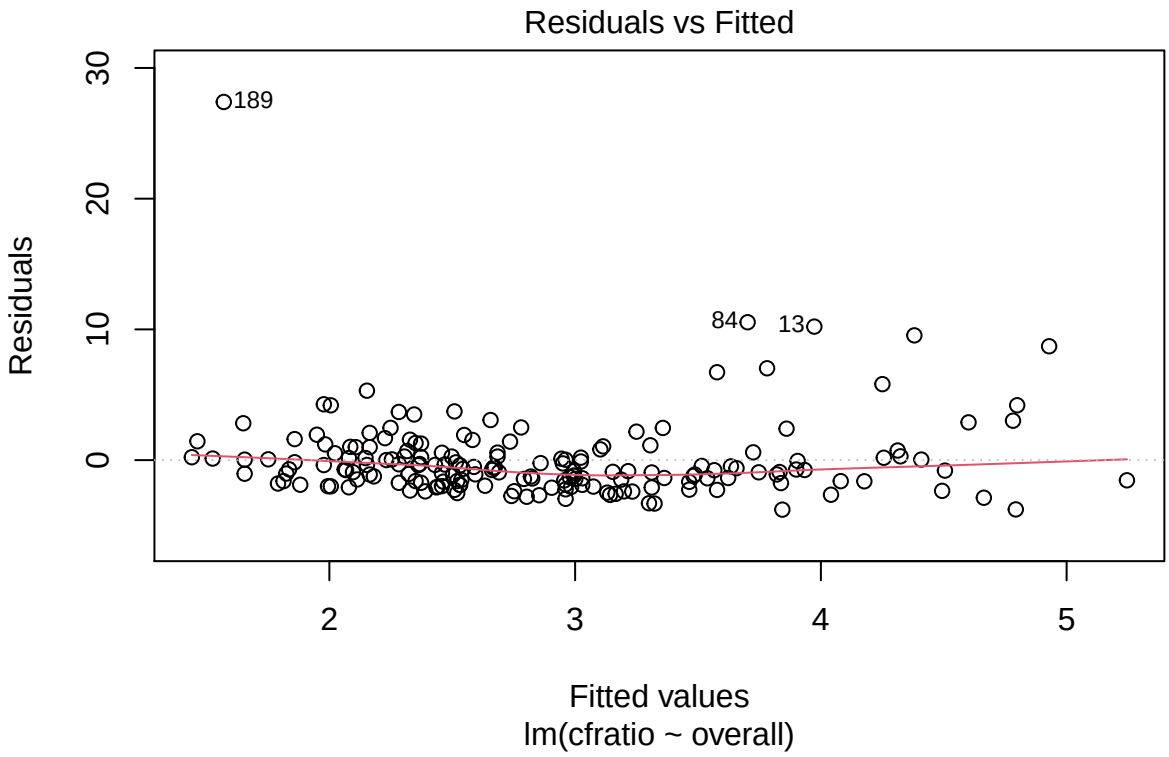
Outlier Screening

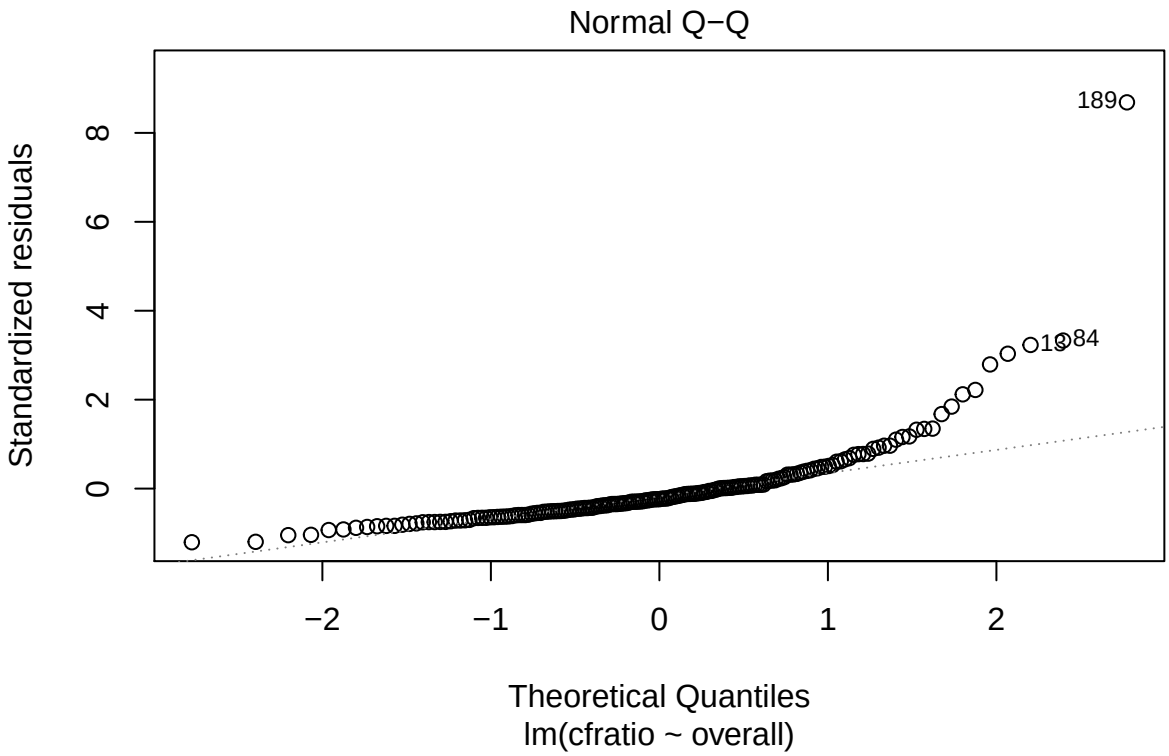
Emily Linebarger

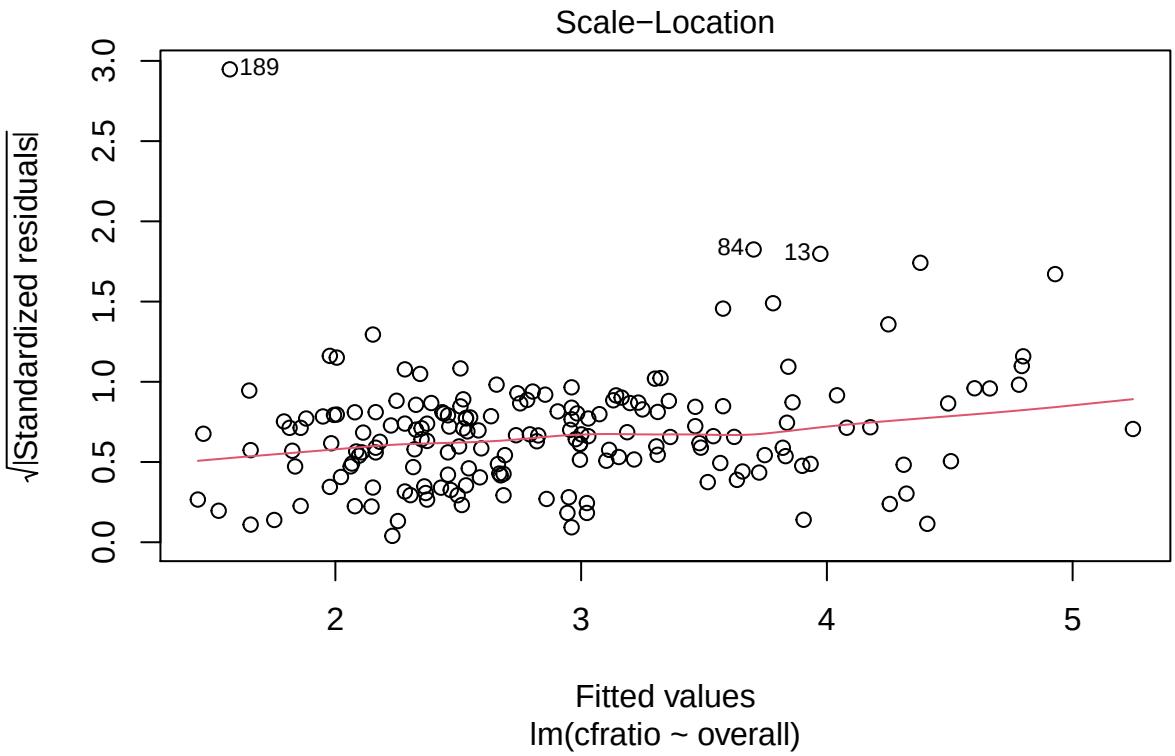
2/17/2021

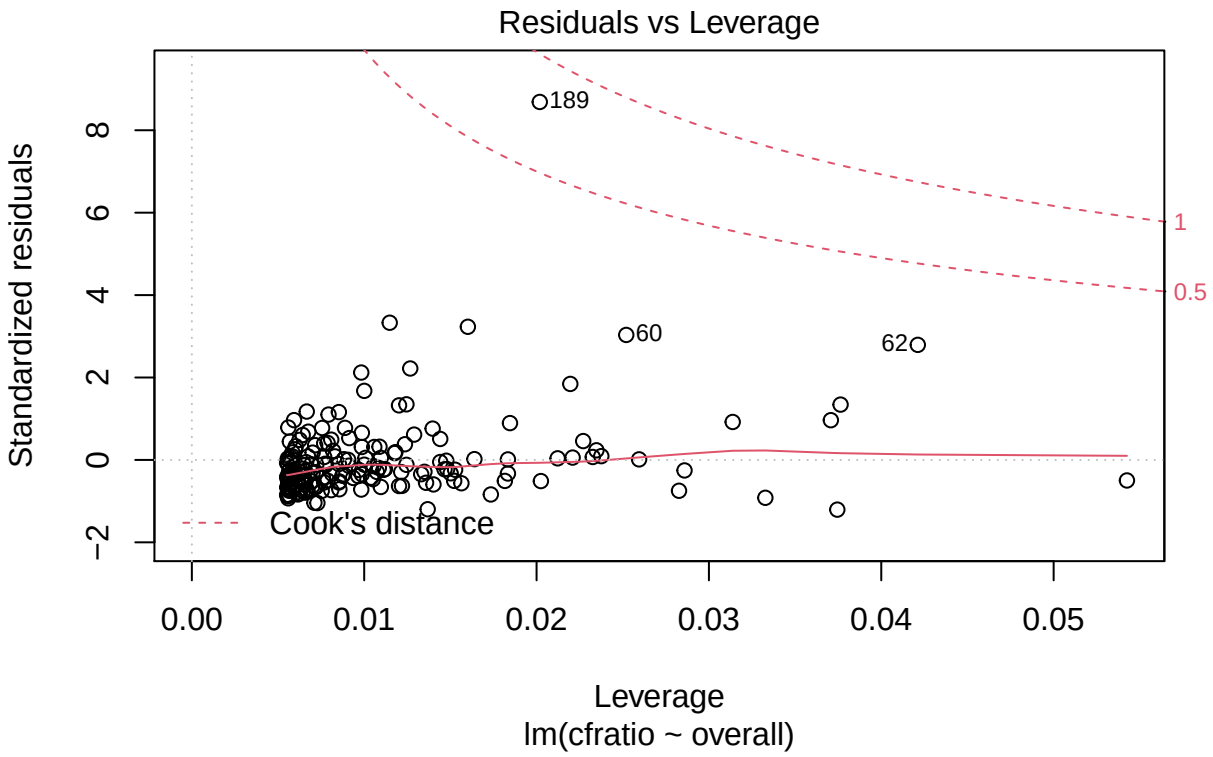
Run first regression, of just cases-per-capita to overall score.

```
##
## Call:
## lm(formula = cfratio ~ overall, data = dt)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.7871 -1.6501 -0.7537  0.5766 27.3988
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.52444    0.73506   0.713 0.476482
## overall      0.05654    0.01676   3.374 0.000908 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.186 on 179 degrees of freedom
##   (11 observations deleted due to missingness)
## Multiple R-squared:  0.05979,    Adjusted R-squared:  0.05454
## F-statistic: 11.38 on 1 and 179 DF,  p-value: 0.0009083
```


Residuals vs Fitted
Residuals
30
20
10
0
189
84
13
2
3
4
5
Fitted values
lm(cfratio ~ overall)


Normal Q−Q
Standardized residuals
0
2
4
6
8
189
184
−2
−1
0
1
2
Theoretical Quantiles
lm(cfratio ~ overall)


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
2.5
3.0
189
84
13
2
3
4
5
Fitted values
lm(cfratio ~ overall)




```
## pdf
##   2
```

There seems to be constant variance in plots 1 and 3. However, we have an issue with non-normality in our qq-plot. Examining this further with a histogram, we can see two issues: a large cluster of zeros, and a few high outliers with a case fatality ratio higher than 25.

```
## pdf
##   2
```

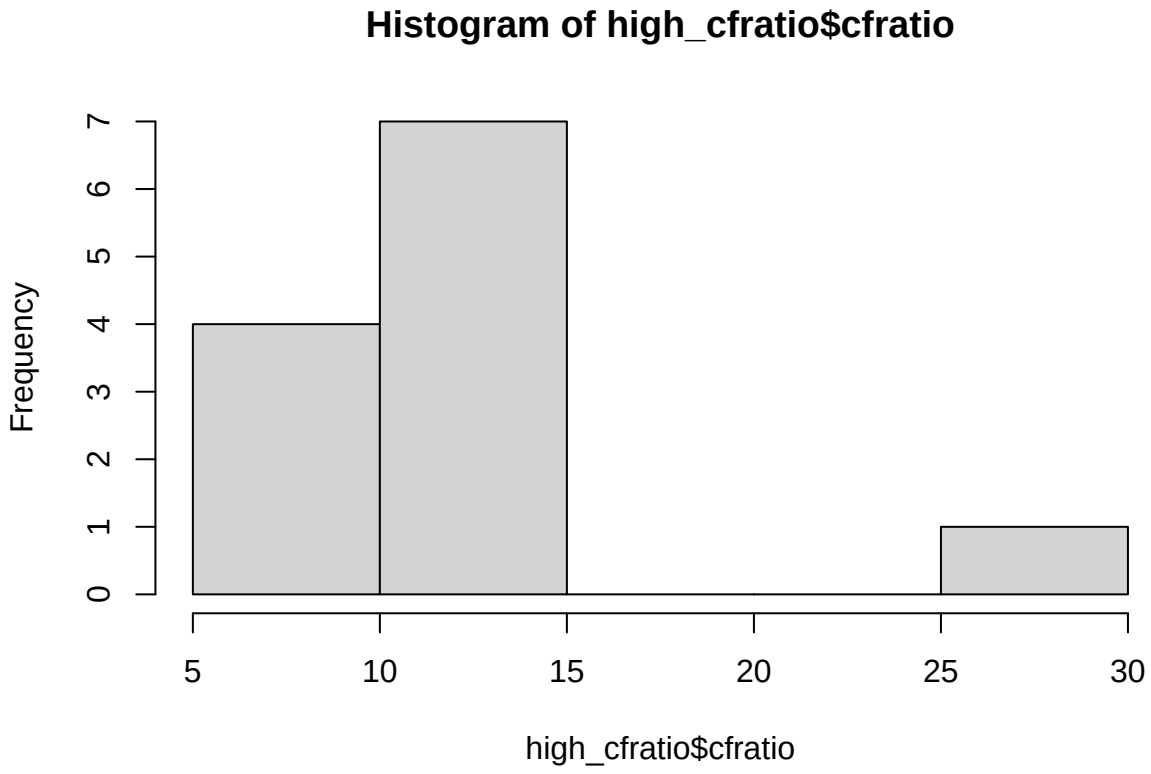
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   0.000   1.080   2.121   2.849   3.237  28.969       9
```

Review upper values

The upper whisker extends 1.5 times past the interquartile range, to 6.6975. The values that lie beyond that are:

country_code	Cases	Deaths	casepc	deathpc	cfratio
TCD	1085	81	0.0680384	0.0050794	7.465438
SWE	76516	5730	7.4392445	0.5570975	7.488630
CAN	114398	8919	3.0433691	0.2372752	7.796465
NLD	69224	6227	3.9938037	0.3592600	8.995435
ESP	282641	28441	6.0038302	0.6041407	10.062588

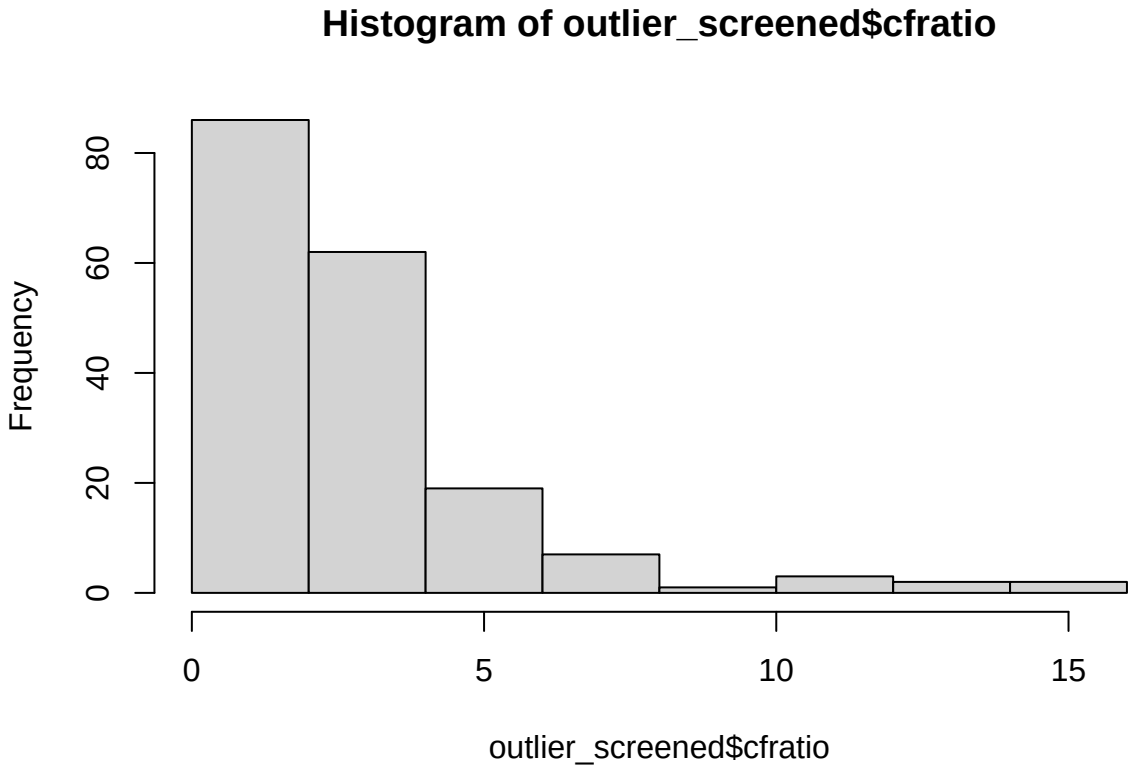
country_code	Cases	Deaths	casepc	deathpc	cfratio
HUN	5961	614	0.6101362	0.0628458	10.300285
MEX	568621	61450	4.4571322	0.4816754	10.806847
GBR	302261	41220	4.5225360	0.6167482	13.637221
FRA	216684	30169	3.2312014	0.4498815	13.923040
BEL	69402	9845	6.0433357	0.8572756	14.185470
ITA	246488	35123	4.0878714	0.5824961	14.249375
YEM	2047	593	0.0701943	0.0203347	28.969223

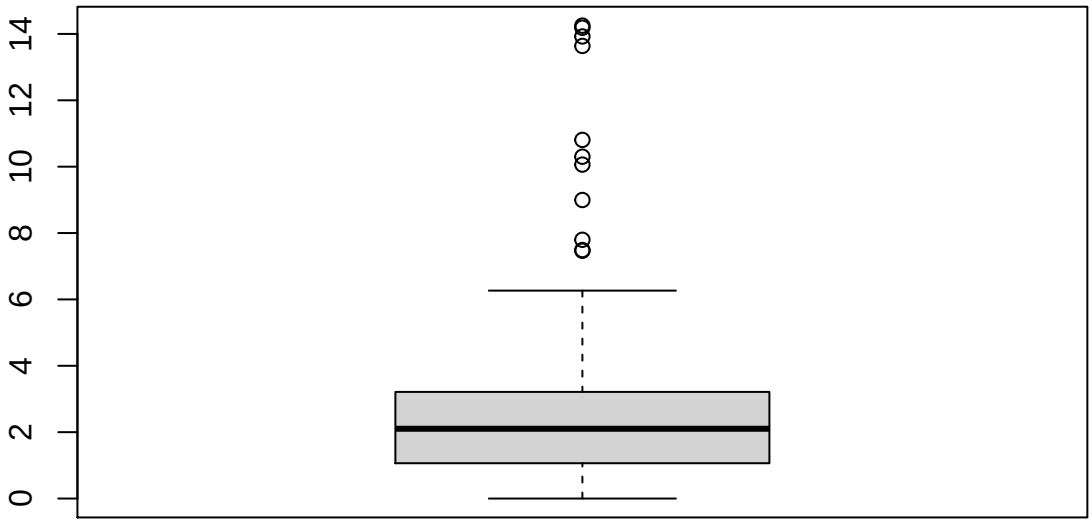


Although these do represent high fatality ratios, many of them represent reasonable values. For example, the second-highest, Italy, had one of the worst early waves in the pandemic and this fatality ratio is very likely six months after they started recording cases/deaths.

The only value that I would treat like an outlier here is Yemen, which has a case-fatality ratio of 28.9. It looks highly likely that Yemen has undercounted cases in these reported numbers, because over 1 in 4 people have died after catching COVID-19 in their reported numbers (593 out of 2047). Because of this, we will drop this data point from our analysis.

Review zeros/very low values



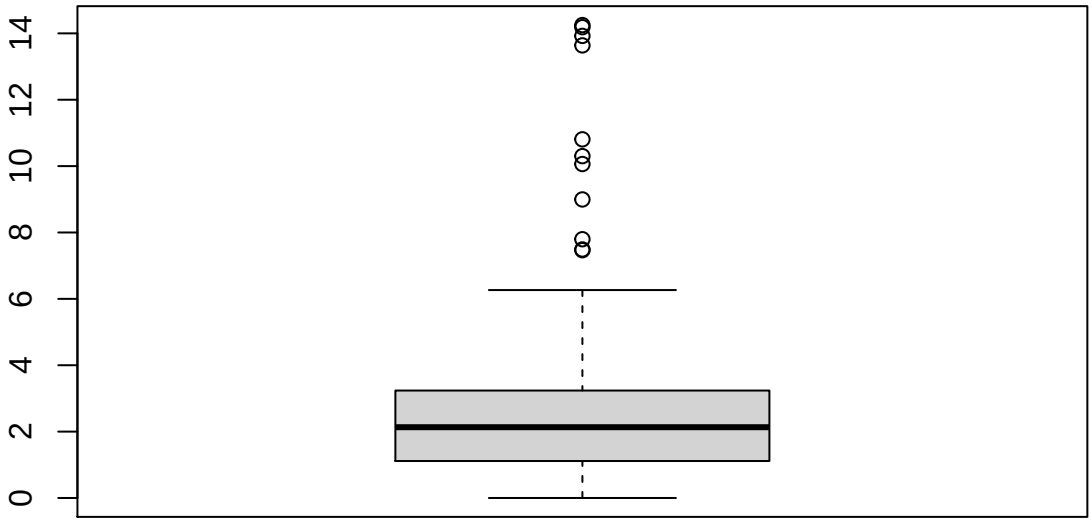


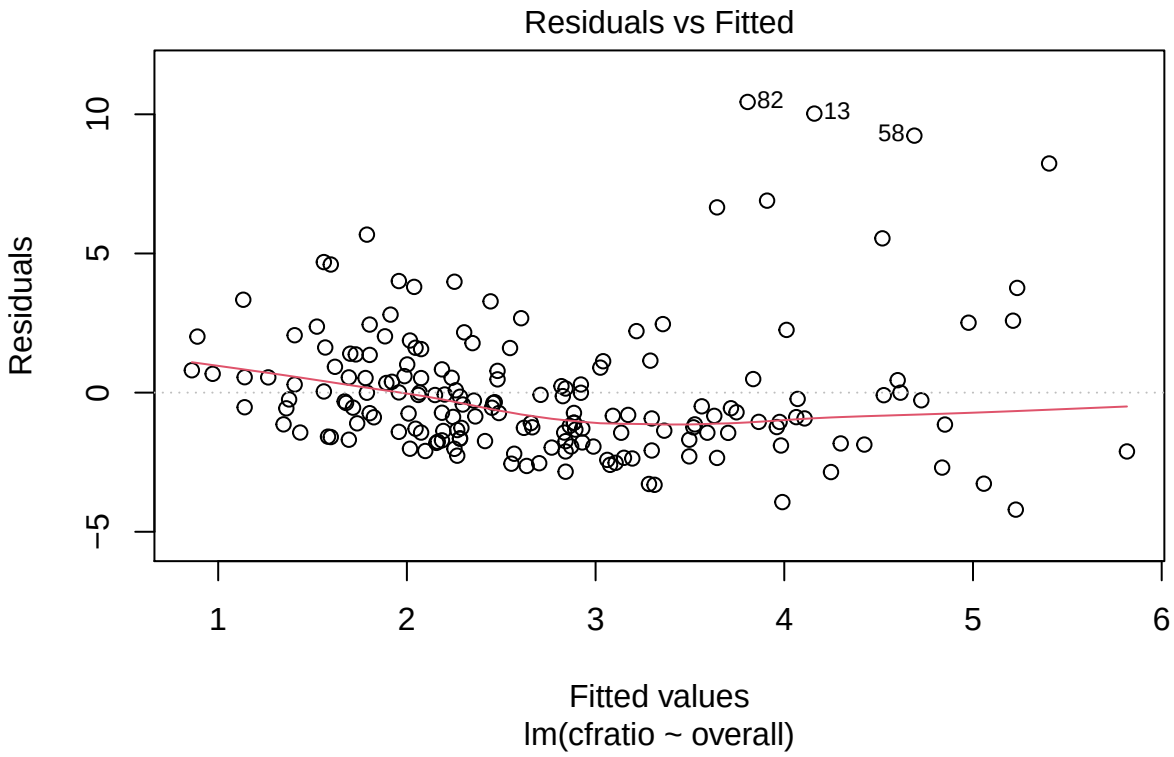
There are also a large cluster of countries with values clustered at zero (N=14), which would signify zero deaths. Review these to make sure these values are believable.

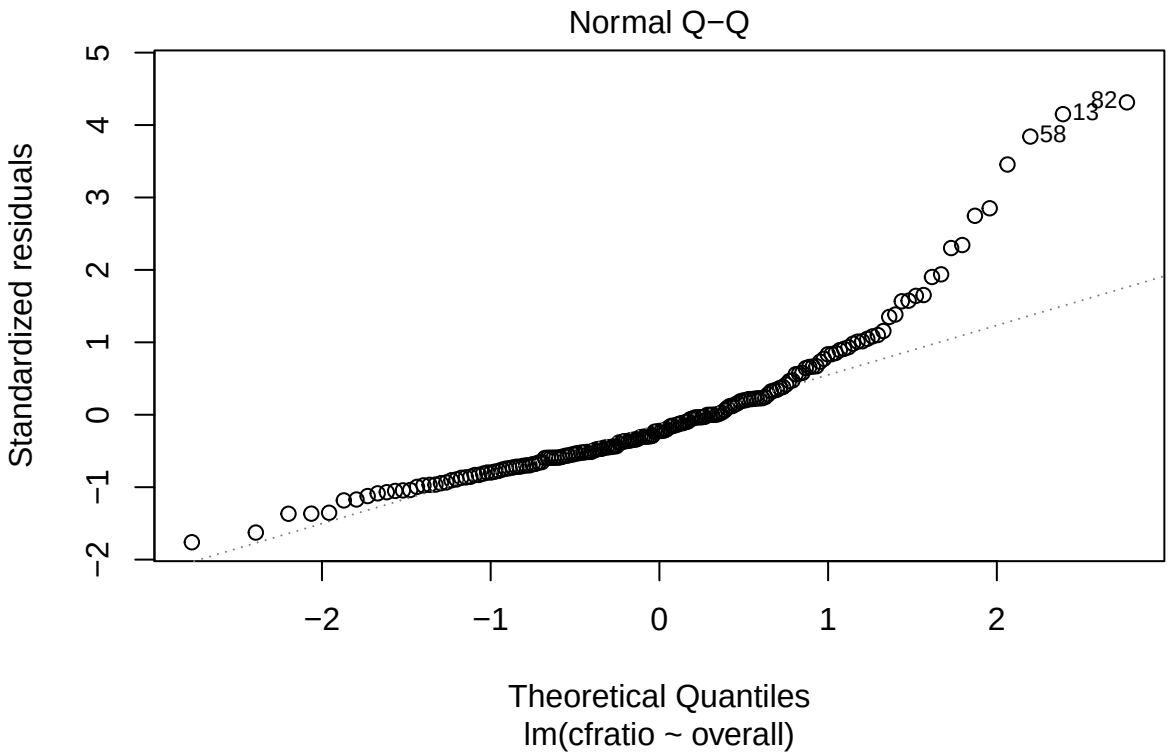
country_code	Cases	Deaths	deathpc	casepc	pop_2019	clean_date	cfratio
BTN	227	0	0	0.2974740	763092	2020-09-01	0
DMA	24	0	0	0.3342246	71808	2020-09-17	0
ERI	364	0	NA	NA	NA	2020-09-16	0
GRD	24	0	0	0.2142800	112003	2020-09-17	0
KHM	202	0	0	0.0122524	16486542	2020-07-24	0
KNA	17	0	0	0.3217625	52834	2020-09-20	0
LAO	23	0	0	0.0032081	7169455	2020-09-19	0
LCA	27	0	0	0.1477105	182790	2020-09-09	0
MNG	310	0	0	0.0961191	3225167	2020-09-05	0
SYC	137	0	0	1.4033291	97625	2020-09-09	0
TLS	27	0	0	0.0208797	1293119	2020-09-17	0
VAT	12	0	NA	NA	NA	2020-09-01	0
VCT	62	0	0	0.5606344	110589	2020-09-09	0
VNM	384	0	0	0.0039808	96462106	2020-07-20	0

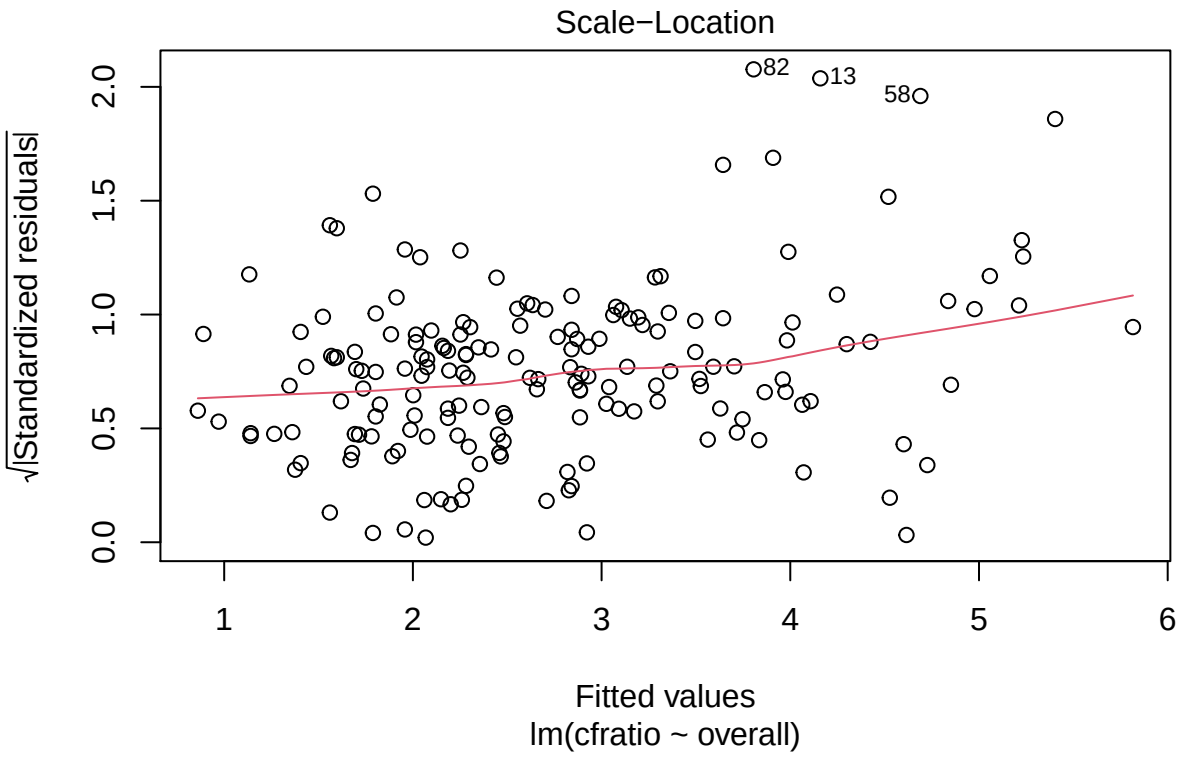
There are a few NA values here where population was not available. We will remove these from our analysis. However, for the rest of the values, they all have very low reported case numbers and may have truly recorded zero deaths. Without more substantial evidence to signify a reporting failure, we will leave these zeros in for now.

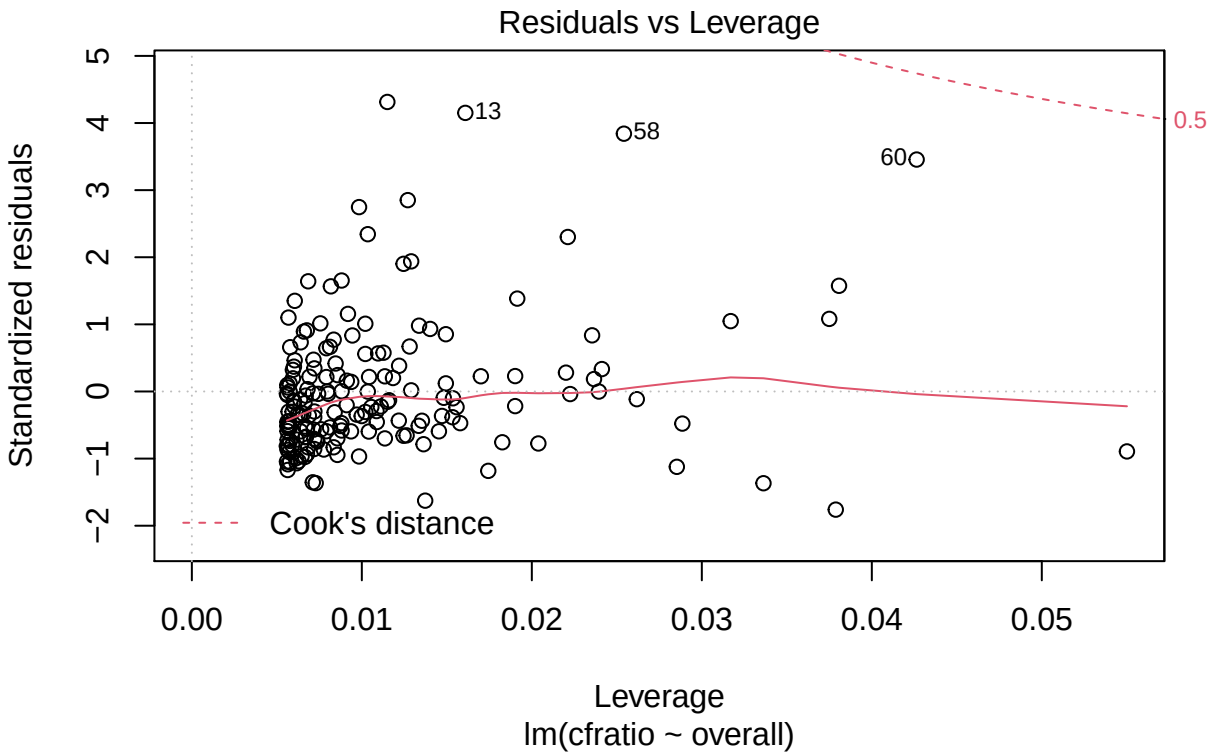
Finally, run initial analysis again and see if non-normality problem persists.




Residuals vs Fitted
82
13
58
Residuals
10
5
0
−5
1
2
3
4
5
6
Fitted values
lm(cfratio ~ overall)


Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(cfratio ~ overall)
82
13
58
−2
−1
0
1
2
3
4
5


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
82
13
58
1
2
3
4
5
6
Fitted values
lm(cfratio ~ overall)



After this cleaning, we do see some evidence of non-constant variance, and our non-normality problem persists. We will need to turn to data transformations to address these issues.

The outlier screening process has taken our total sample size from N=192 to N=179.