lab02

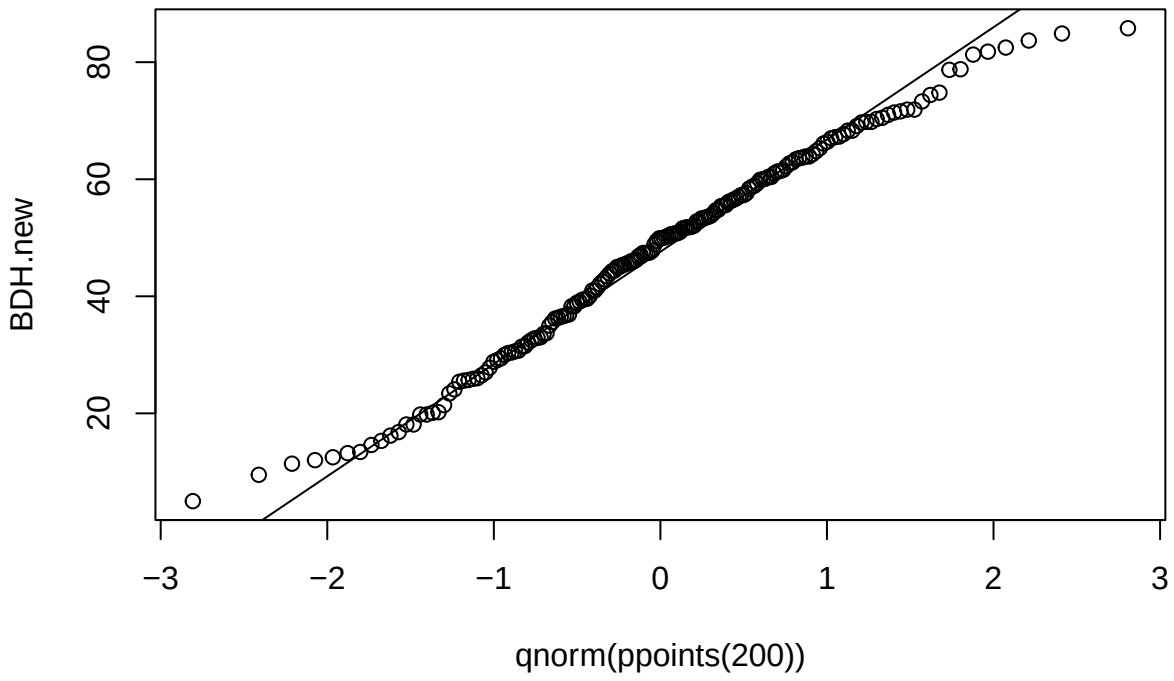
Sean Fitch

2024-09-20

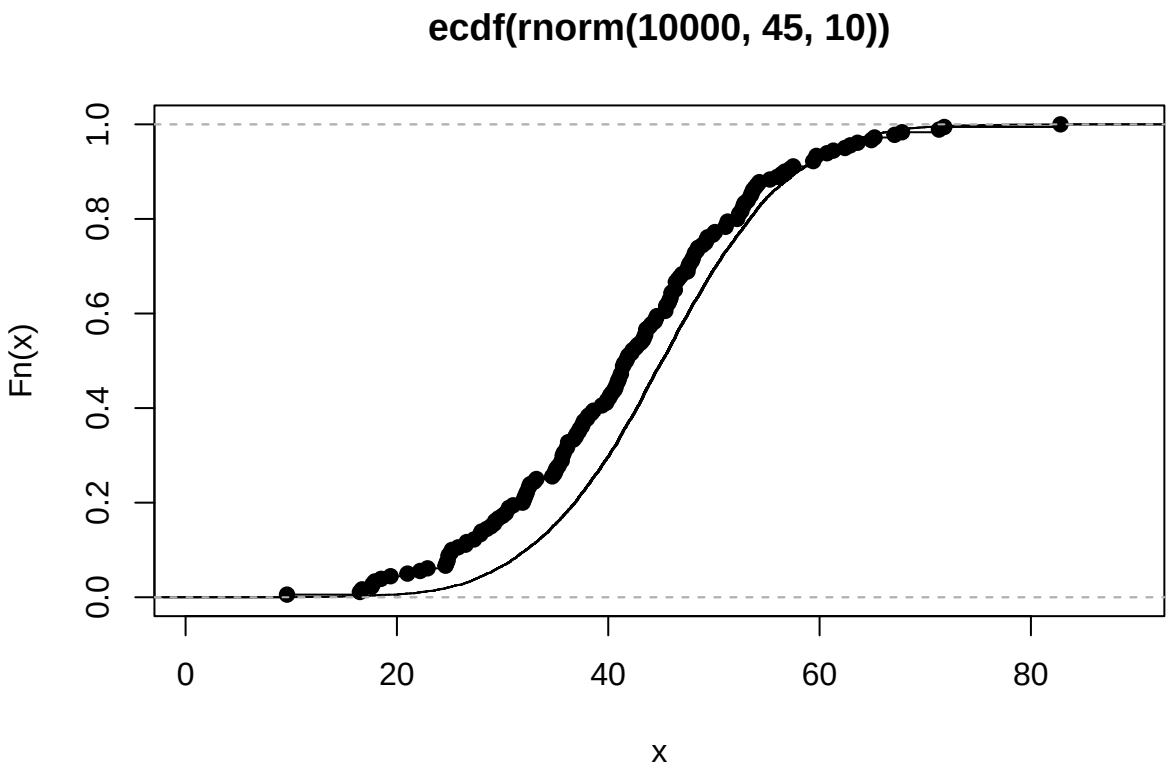
```
library(ggplot2)
```

```
EPI_data <- read.csv("../epi2024results06022024.csv")
epi.results <- read.csv("../epi2024results06022024.csv", header=TRUE)
epi.weights <- read.csv("../epi2024weights.csv")
attach(EPI_data)
```

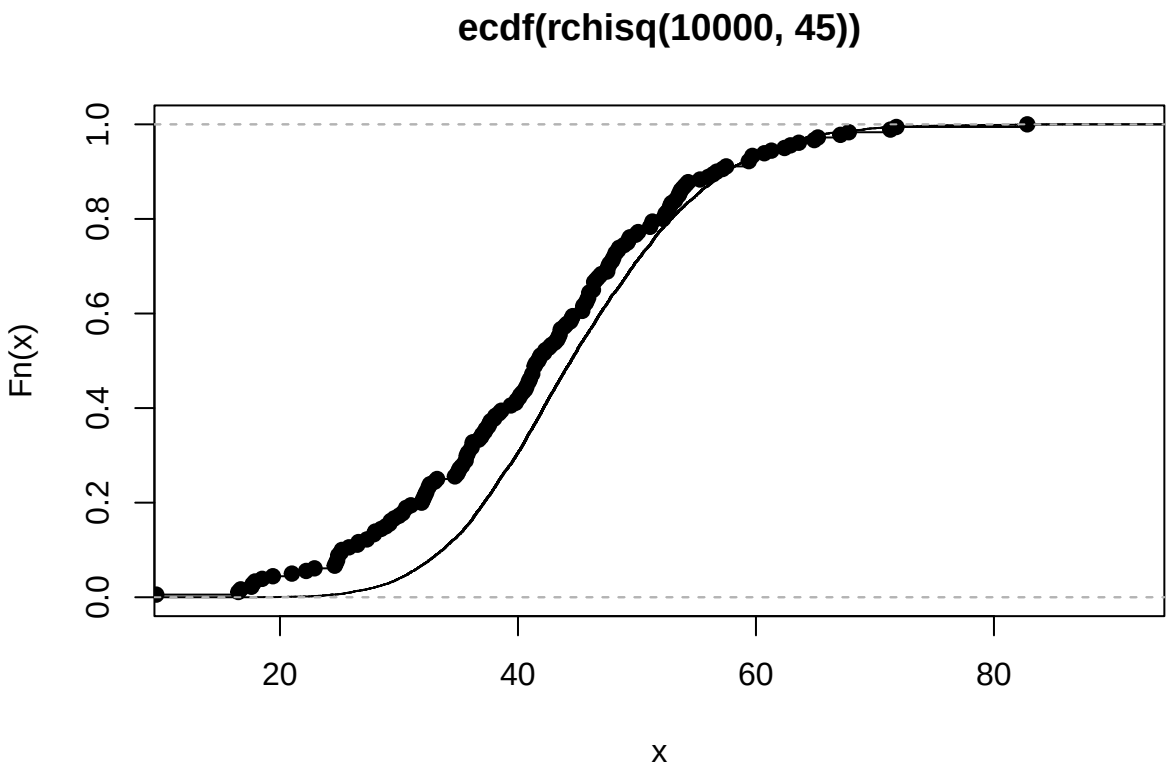
```
qqplot(qnorm(ppoints(200)),BDH.new)
qqline(BDH.new)
```




```
plot(ecdf(rnorm(10000, 45, 10)), do.points=FALSE)
lines(ecdf(CCH.new))
```




```
plot(ecdf(rchisq(10000, 45)), do.points=FALSE)
lines(ecdf(CCH.new))
```



Create population data set

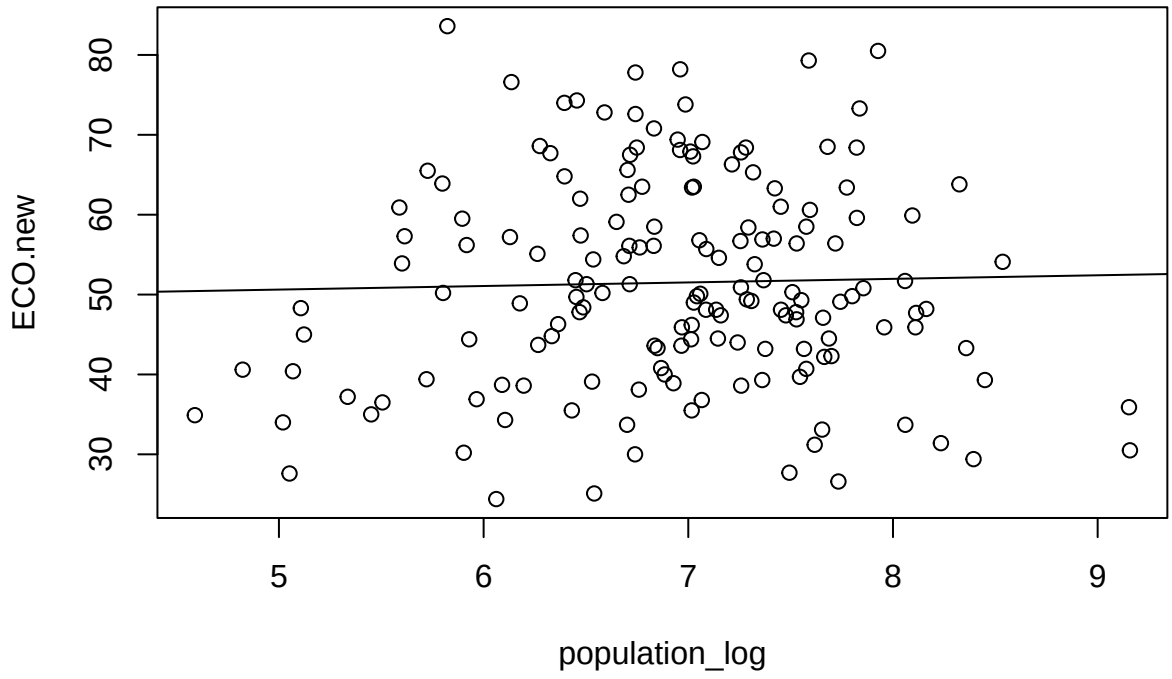
```
# read data
populations_2023 <- read.csv("../countries_populations_2023.csv")
# drop countries not in epi results
populations <- populations_2023[-which(!populations_2023$Country %in% epi.results$country),]
# sort populations by country
populations <- populations[order(populations$Country),]
# drop countries not in populations
epi.results.sub <- epi.results[-which(!epi.results$country %in% populations$Country),]
# sort epi results by country
epi.results.sub <- epi.results.sub[order(epi.results.sub$country),]
# only keep necessary columns
# epi.results.sub <- epi.results.sub[,c("country","EPI.old","EPI.new")]
# convert population to numeric
epi.results.sub$population <- as.numeric(populations$Population)
# compute population log base 10
epi.results.sub$population_log <- log10(epi.results.sub$population)
```

```
attach(epi.results.sub)
```

```
## The following objects are masked from EPI_data:
##
##     AGR.new, AGR.old, AIR.new, AIR.old, APO.new, APO.old, BCA.new,
##     BCA.old, BDH.new, BDH.old, BER.new, BER.old, BTO.new, BTO.old,
##     BTZ.new, BTZ.old, CBP.new, CBP.old, CCH.new, CCH.old, CDA.new,
```

```
##     CDA.old, CDF.new, CDF.old, CHA.new, CHA.old, code, COE.new,
##     COE.old, country, ECO.new, ECO.old, ECS.new, ECS.old, EPI.new,
##     EPI.old, FCD.new, FCD.old, FCL.new, FCL.old, FGA.new, FGA.old,
##     FLI.new, FLI.old, FSH.new, FSH.old, FSS.new, FSS.old, GHN.new,
##     GHN.old, GTI.new, GTI.old, GTP.new, GTP.old, H2O.new, H2O.old,
##     HFD.new, HFD.old, HLT.new, HLT.old, HMT.new, HMT.old, HPE.new,
##     HPE.old, IFL.new, IFL.old, iso, LED.new, LED.old, LUF.new, LUF.old,
##     MHP.new, MHP.old, MKP.new, MKP.old, MPE.new, MPE.old, NDA.new,
##     NDA.old, NOD.new, NOD.old, NXA.new, NXA.old, OEB.new, OEB.old,
##     OEC.new, OEC.old, OZD.new, OZD.old, PAE.new, PAE.old, PAR.new,
##     PAR.old, PCC.new, PCC.old, PFL.new, PFL.old, PHL.new, PHL.old,
##     PRS.new, PRS.old, PSU.new, PSU.old, RCY.new, RCY.old, RLI.new,
##     RLI.old, RMS.new, RMS.old, SDA.new, SDA.old, SHI.new, SHI.old,
##     SMW.new, SMW.old, SNM.new, SNM.old, SOE.new, SOE.old, SPI.new,
##     SPI.old, TBN.new, TBN.old, TCG.new, TCG.old, TKP.new, TKP.old,
##     USD.new, USD.old, UWD.new, UWD.old, VOE.new, VOE.old, WMG.new,
##     WMG.old, WPC.new, WPC.old, WRR.new, WRR.old, WRS.new, WRS.old,
##     WWC.new, WWC.old, WWG.new, WWG.old, WWR.new, WWR.old, WWT.new,
##     WWT.old
```

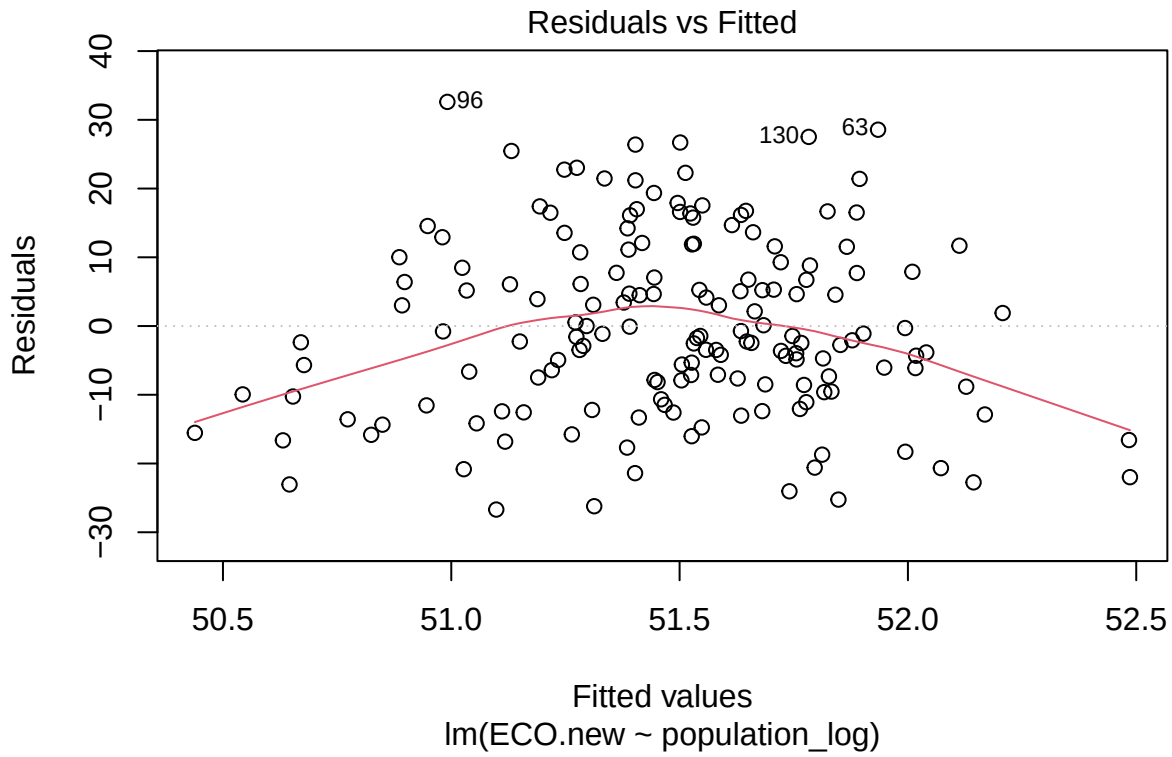
```
lin.mod.epinew <- lm(ECO.new~population_log,epi.results.sub)
plot(ECO.new~population_log)
abline(lin.mod.epinew)
```

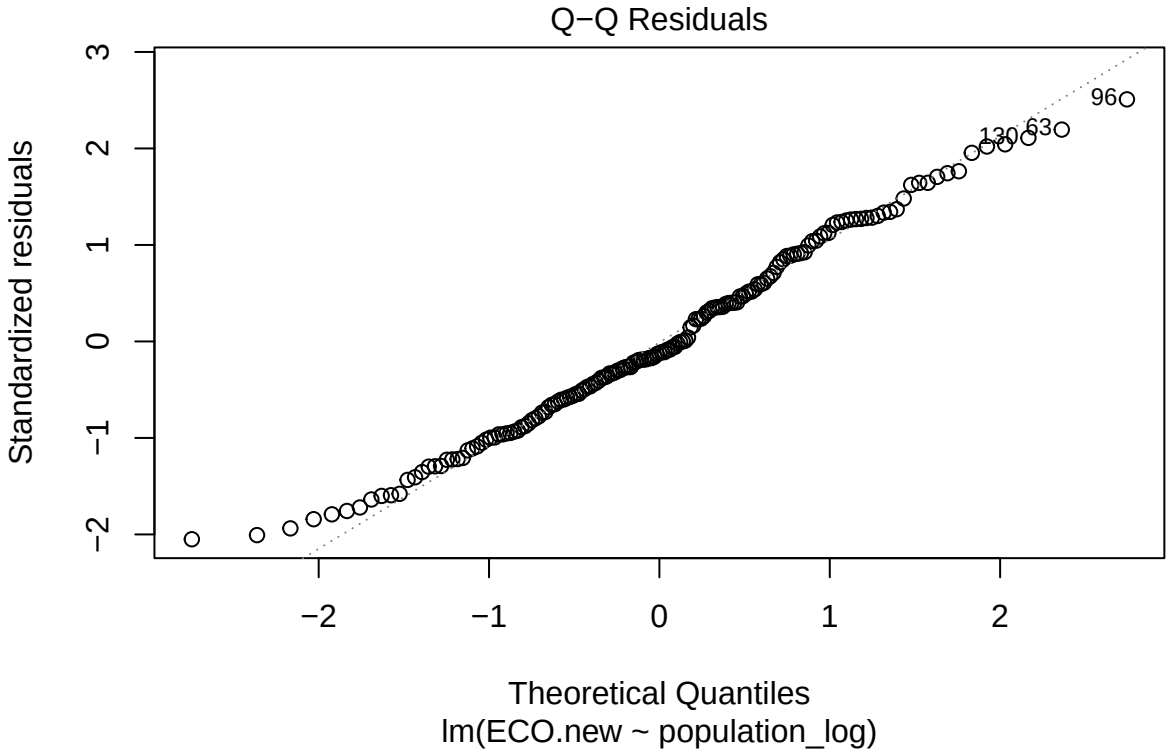


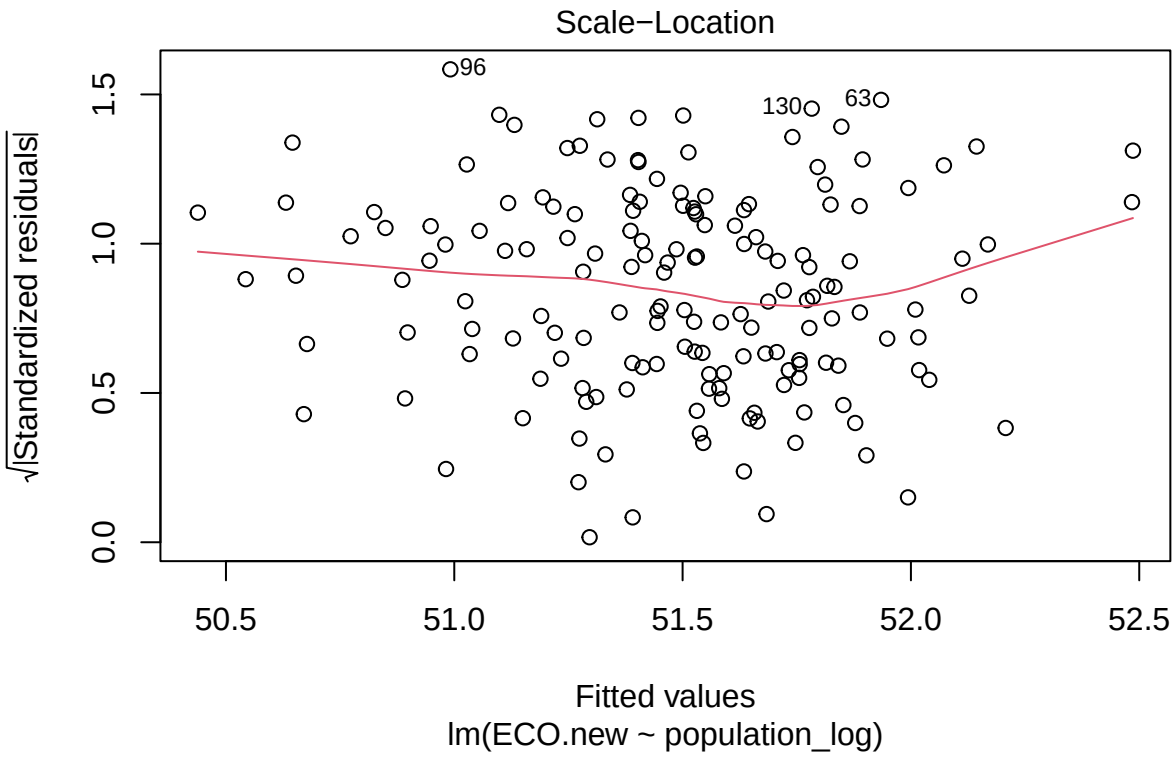
```
summary(lin.mod.epinew)
```

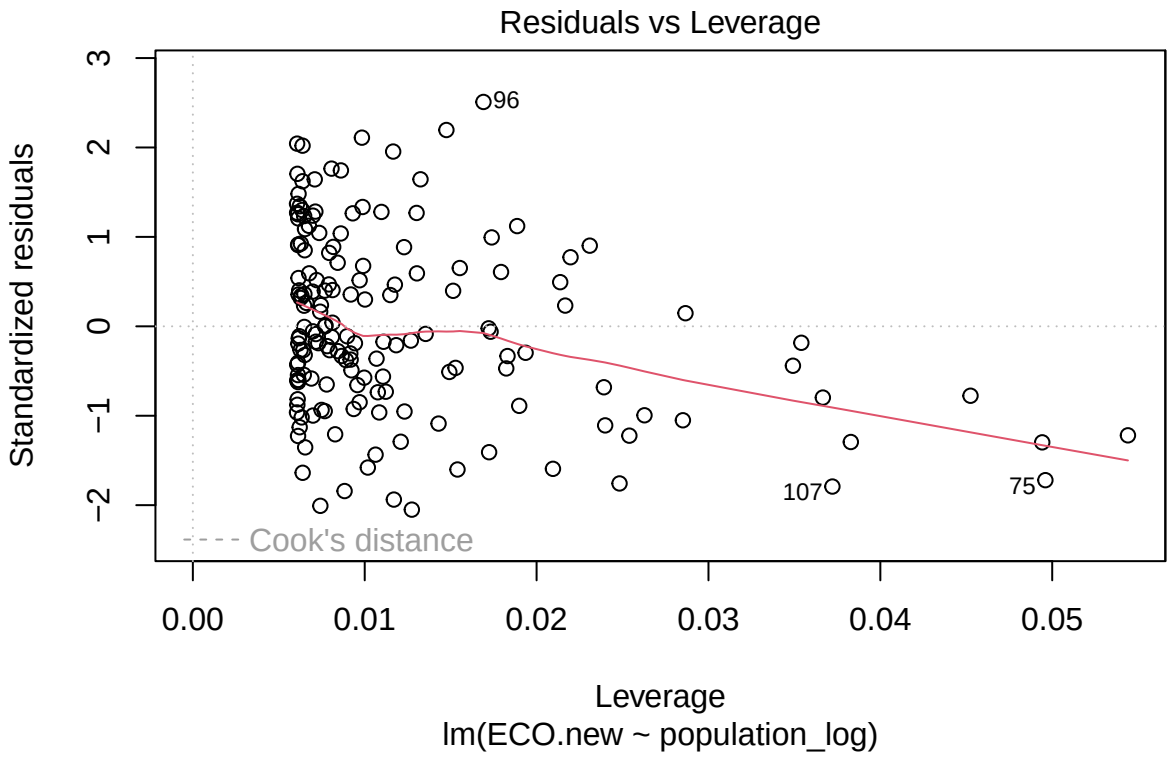
```
##
## Call:
## lm(formula = ECO.new ~ population_log, data = epi.results.sub)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -26.699  -9.532  -1.574   9.278  32.609
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     48.3814     8.5855   5.635 7.52e-08 ***
## population_log   0.4482     1.2295   0.365    0.716
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.11 on 163 degrees of freedom
## Multiple R-squared:  0.0008147,  Adjusted R-squared:  -0.005315
## F-statistic: 0.1329 on 1 and 163 DF,  p-value: 0.7159
```

```
plot(lin.mod.epinew)
```



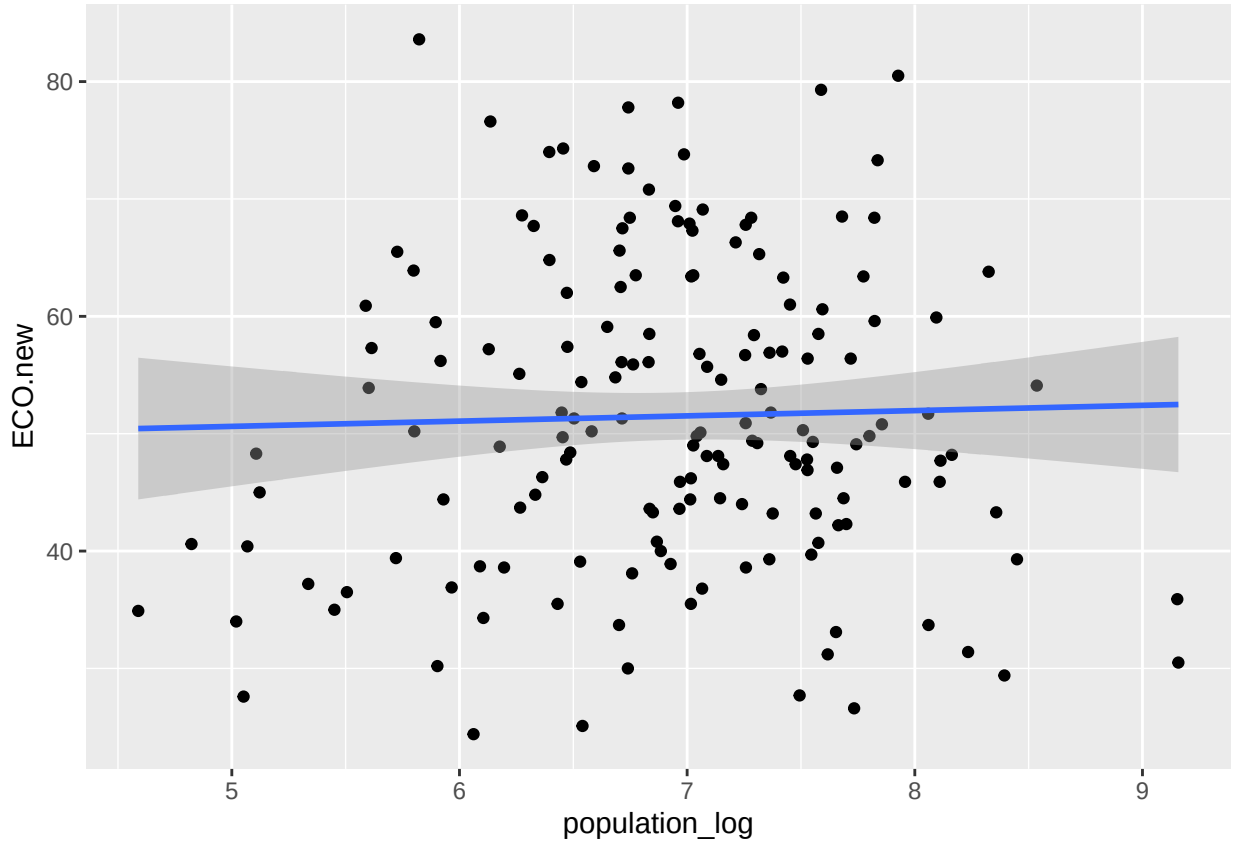

Q−Q Residuals
Standardized residuals
Theoretical Quantiles
lm(ECO.new ~ population_log)
96
130
63
−2
−1
0
1
2
3


Scale−Location
96
130
63
√|Standardized residuals|
0.0
0.5
1.0
1.5
50.5
51.0
51.5
52.0
52.5
Fitted values
lm(ECO.new ~ population_log)

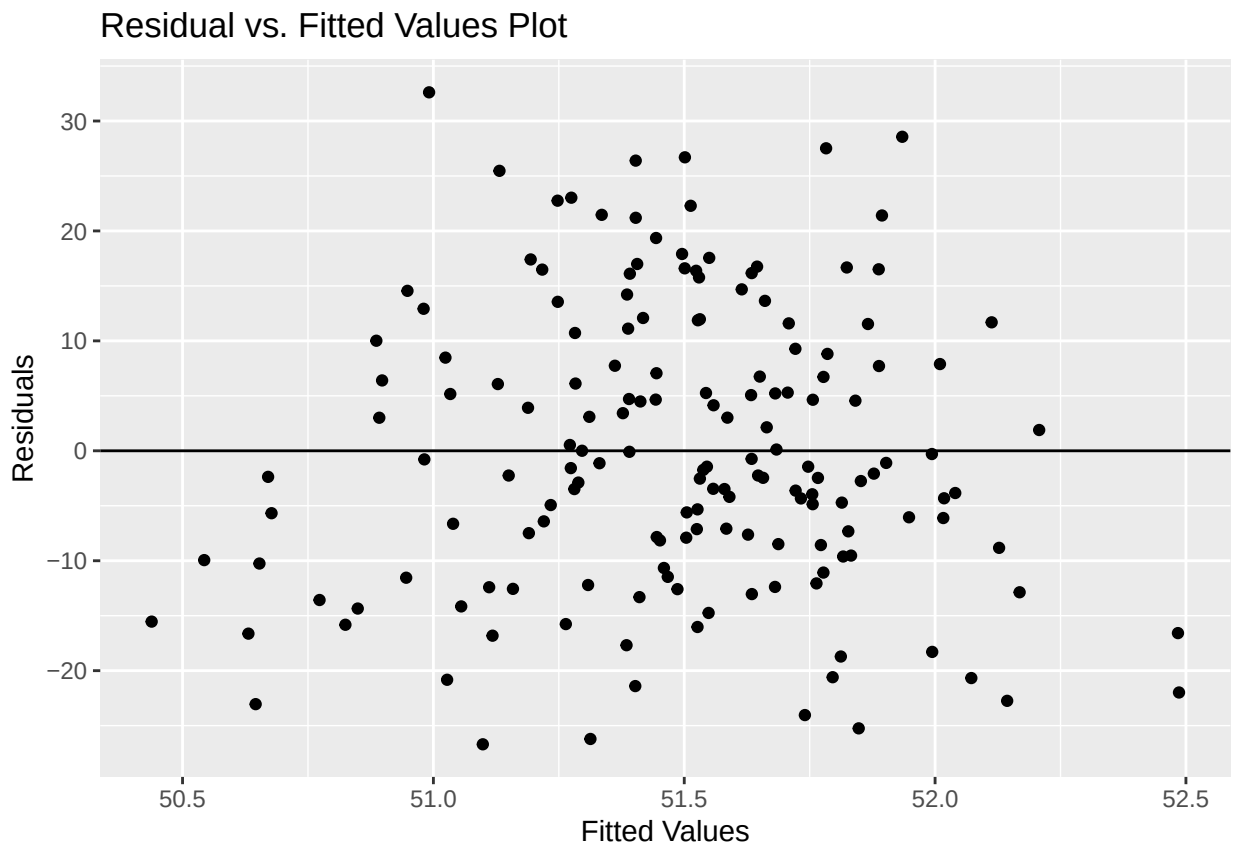



```
ggplot(epi.results.sub, aes(x = population_log, y = ECO.new)) +
geom_point() +
stat_smooth(method = "lm")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```




```
ggplot(lin.mod.epinew, aes(x = .fitted, y = .resid)) +
geom_point() +
geom_hline(yintercept = 0) +
labs(title='Residual vs. Fitted Values Plot', x='Fitted Values', y='Residuals')
```

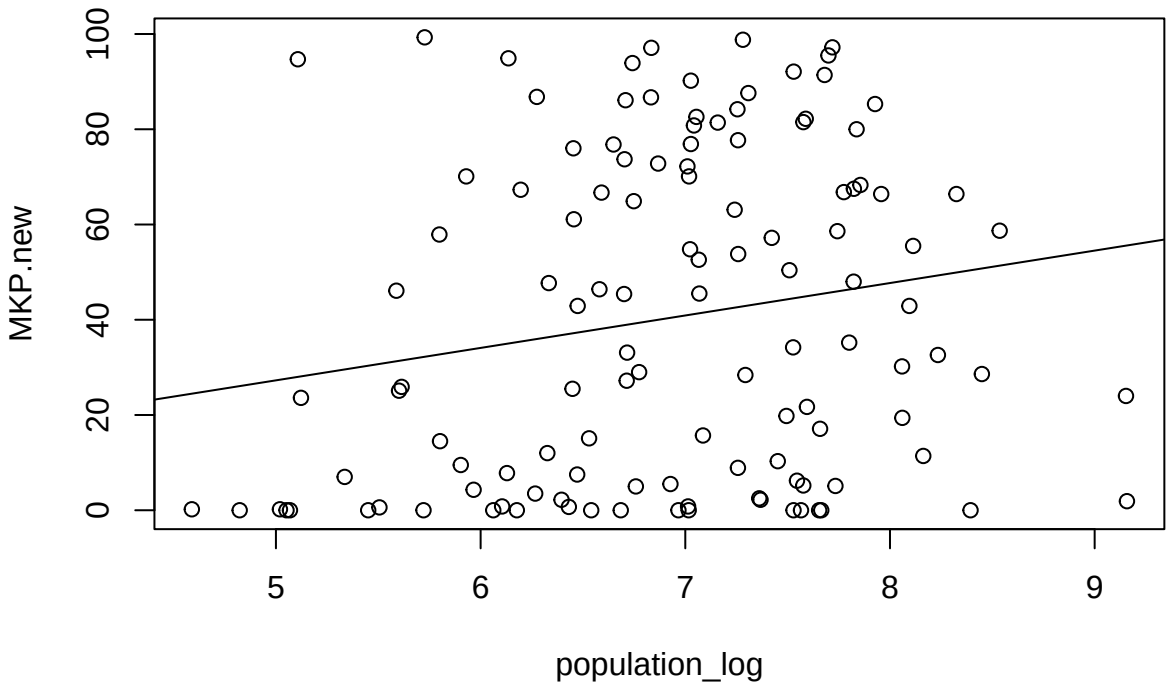



```
attach(epi.results.sub)
```

```
## The following objects are masked from epi.results.sub (pos = 3):
##
##     AGR.new, AGR.old, AIR.new, AIR.old, APO.new, APO.old, BCA.new,
##     BCA.old, BDH.new, BDH.old, BER.new, BER.old, BTO.new, BTO.old,
##     BTZ.new, BTZ.old, CBP.new, CBP.old, CCH.new, CCH.old, CDA.new,
##     CDA.old, CDF.new, CDF.old, CHA.new, CHA.old, code, COE.new,
##     COE.old, country, ECO.new, ECO.old, ECS.new, ECS.old, EPI.new,
##     EPI.old, FCD.new, FCD.old, FCL.new, FCL.old, FGA.new, FGA.old,
##     FLI.new, FLI.old, FSH.new, FSH.old, FSS.new, FSS.old, GHN.new,
##     GHN.old, GTI.new, GTI.old, GTP.new, GTP.old, H2O.new, H2O.old,
##     HFD.new, HFD.old, HLT.new, HLT.old, HMT.new, HMT.old, HPE.new,
##     HPE.old, IFL.new, IFL.old, iso, LED.new, LED.old, LUF.new, LUF.old,
##     MHP.new, MHP.old, MKP.new, MKP.old, MPE.new, MPE.old, NDA.new,
##     NDA.old, NOD.new, NOD.old, NXA.new, NXA.old, OEB.new, OEB.old,
##     OEC.new, OEC.old, OZD.new, OZD.old, PAE.new, PAE.old, PAR.new,
##     PAR.old, PCC.new, PCC.old, PFL.new, PFL.old, PHL.new, PHL.old,
##     population, population_log, PRS.new, PRS.old, PSU.new, PSU.old,
##     RCY.new, RCY.old, RLI.new, RLI.old, RMS.new, RMS.old, SDA.new,
##     SDA.old, SHI.new, SHI.old, SMW.new, SMW.old, SNM.new, SNM.old,
##     SOE.new, SOE.old, SPI.new, SPI.old, TBN.new, TBN.old, TCG.new,
##     TCG.old, TKP.new, TKP.old, USD.new, USD.old, UWD.new, UWD.old,
##     VOE.new, VOE.old, WMG.new, WMG.old, WPC.new, WPC.old, WRR.new,
##     WRR.old, WRS.new, WRS.old, WWC.new, WWC.old, WWG.new, WWG.old,
##     WWR.new, WWR.old, WWT.new, WWT.old
```

```
## The following objects are masked from EPI_data:
##
##     AGR.new, AGR.old, AIR.new, AIR.old, APO.new, APO.old, BCA.new,
##     BCA.old, BDH.new, BDH.old, BER.new, BER.old, BTO.new, BTO.old,
##     BTZ.new, BTZ.old, CBP.new, CBP.old, CCH.new, CCH.old, CDA.new,
##     CDA.old, CDF.new, CDF.old, CHA.new, CHA.old, code, COE.new,
##     COE.old, country, ECO.new, ECO.old, ECS.new, ECS.old, EPI.new,
##     EPI.old, FCD.new, FCD.old, FCL.new, FCL.old, FGA.new, FGA.old,
##     FLI.new, FLI.old, FSH.new, FSH.old, FSS.new, FSS.old, GHN.new,
##     GHN.old, GTI.new, GTI.old, GTP.new, GTP.old, H2O.new, H2O.old,
##     HFD.new, HFD.old, HLT.new, HLT.old, HMT.new, HMT.old, HPE.new,
##     HPE.old, IFL.new, IFL.old, iso, LED.new, LED.old, LUF.new, LUF.old,
##     MHP.new, MHP.old, MKP.new, MKP.old, MPE.new, MPE.old, NDA.new,
##     NDA.old, NOD.new, NOD.old, NXA.new, NXA.old, OEB.new, OEB.old,
##     OEC.new, OEC.old, OZD.new, OZD.old, PAE.new, PAE.old, PAR.new,
##     PAR.old, PCC.new, PCC.old, PFL.new, PFL.old, PHL.new, PHL.old,
##     PRS.new, PRS.old, PSU.new, PSU.old, RCY.new, RCY.old, RLI.new,
##     RLI.old, RMS.new, RMS.old, SDA.new, SDA.old, SHI.new, SHI.old,
##     SMW.new, SMW.old, SNM.new, SNM.old, SOE.new, SOE.old, SPI.new,
##     SPI.old, TBN.new, TBN.old, TCG.new, TCG.old, TKP.new, TKP.old,
##     USD.new, USD.old, UWD.new, UWD.old, VOE.new, VOE.old, WMG.new,
##     WMG.old, WPC.new, WPC.old, WRR.new, WRR.old, WRS.new, WRS.old,
##     WWC.new, WWC.old, WWG.new, WWG.old, WWR.new, WWR.old, WWT.new,
##     WWT.old
```

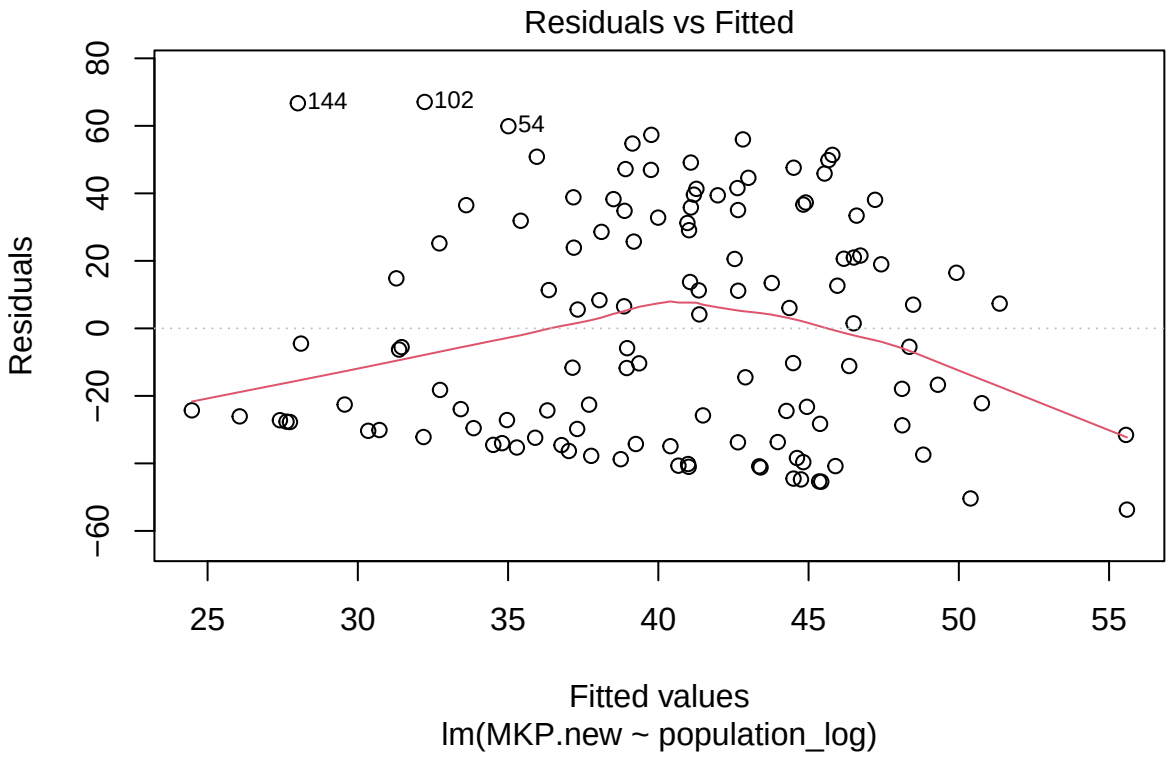
```
lin.mod.epinew <- lm(MKP.new~population_log,epi.results.sub)
plot(MKP.new~population_log)
abline(lin.mod.epinew)
```

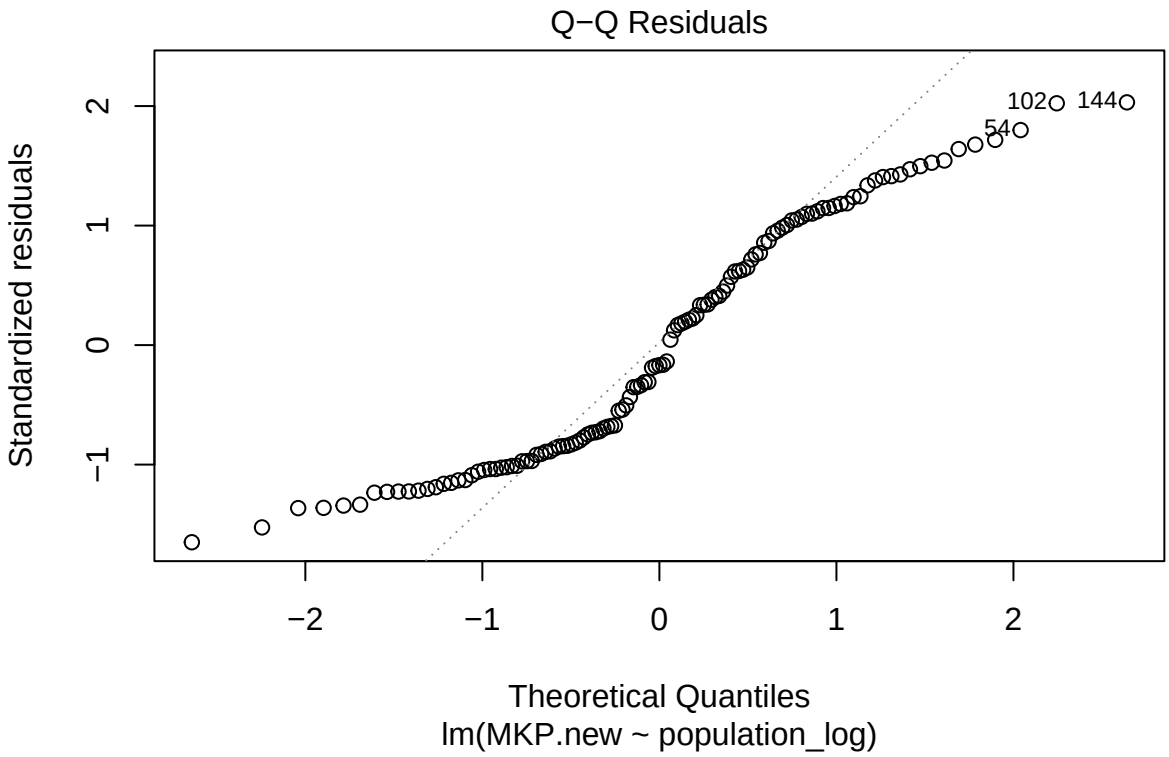


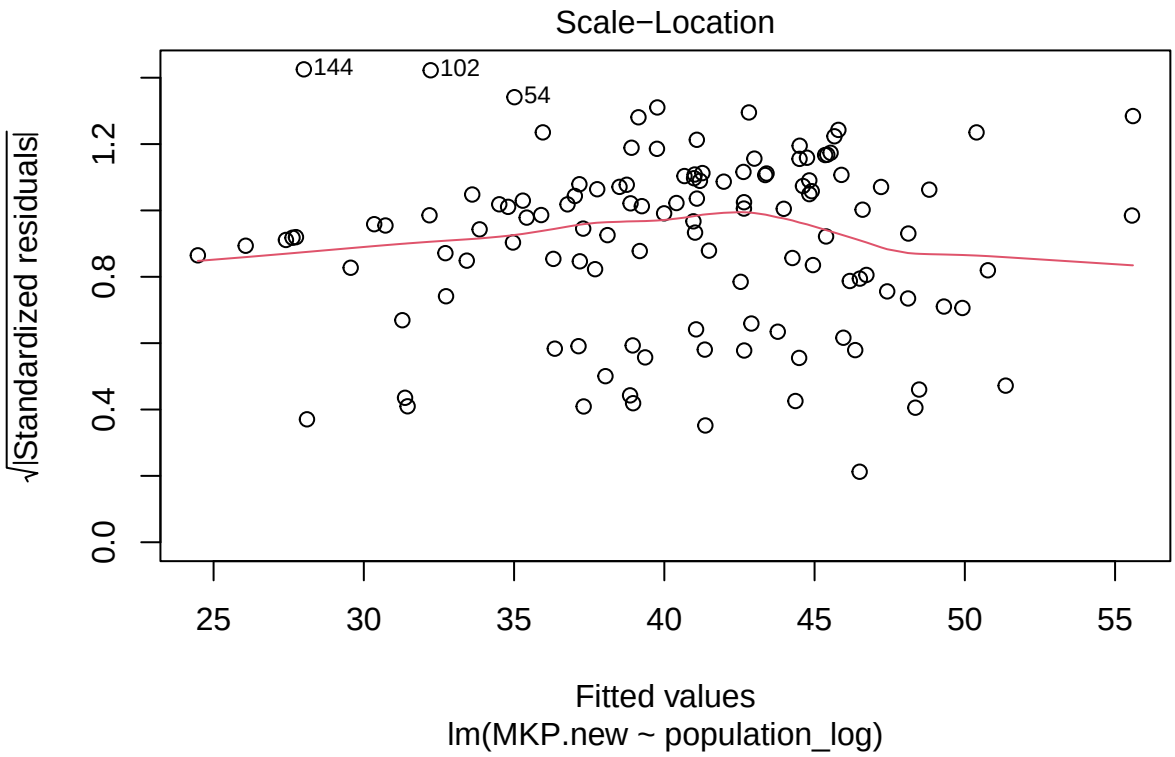

```
summary(lin.mod.epinew)
```

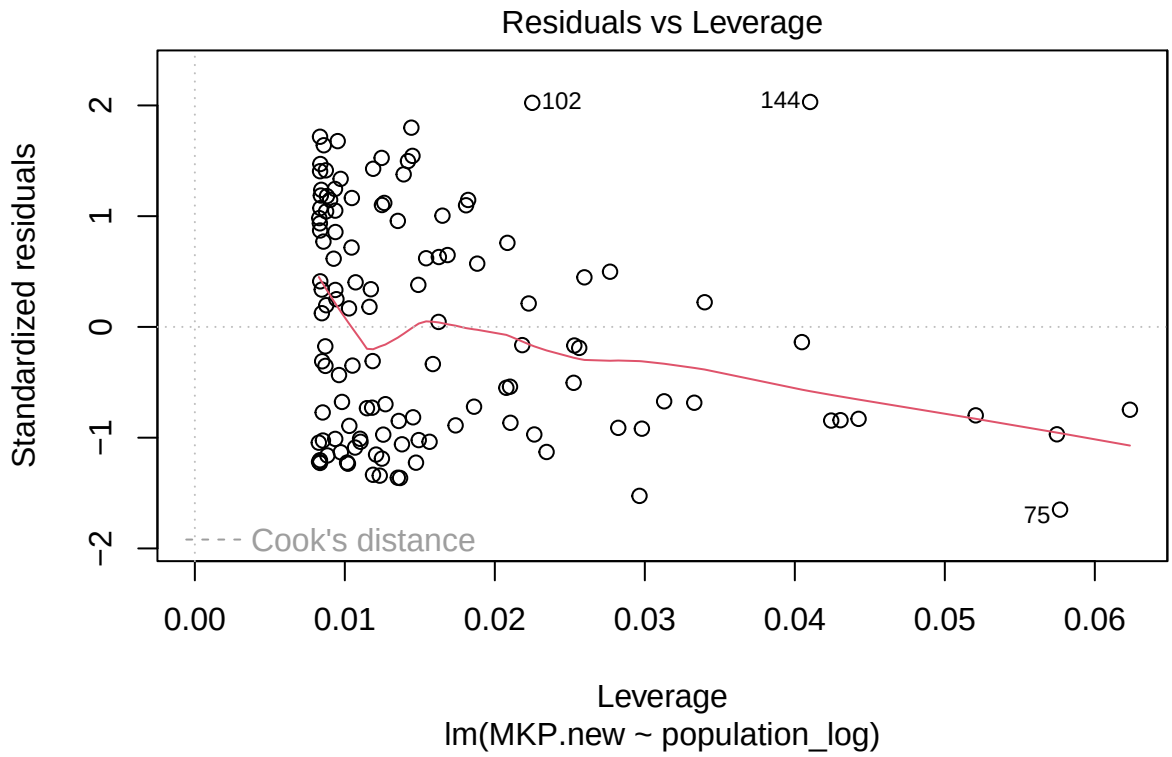
```
##
## Call:
## lm(formula = MKP.new ~ population_log, data = epi.results.sub)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -53.695 -30.119  -5.557  31.880  67.076
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -6.782     23.317  -0.291   0.7717
## population_log    6.811      3.338   2.040   0.0435 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 33.53 on 119 degrees of freedom
##   (44 observations deleted due to missingness)
## Multiple R-squared:  0.0338, Adjusted R-squared:  0.02568
## F-statistic: 4.163 on 1 and 119 DF,  p-value: 0.04352
```

```
plot(lin.mod.epinew)
```


Residuals vs Fitted
Residuals
Fitted values
lm(MKP.new ~ population_log)
144
102
54
80
60
40
20
0
−20
−60
25
30
35
40
45
50
55


Q−Q Residuals
Standardized residuals
Theoretical Quantiles
lm(MKP.new ~ population_log)
102
144
54
−2
−1
0
1
2


Scale−Location
144
102
54
√|Standardized residuals|
0.0
0.4
0.8
1.2
25
30
35
40
45
50
55
Fitted values
lm(MKP.new ~ population_log)

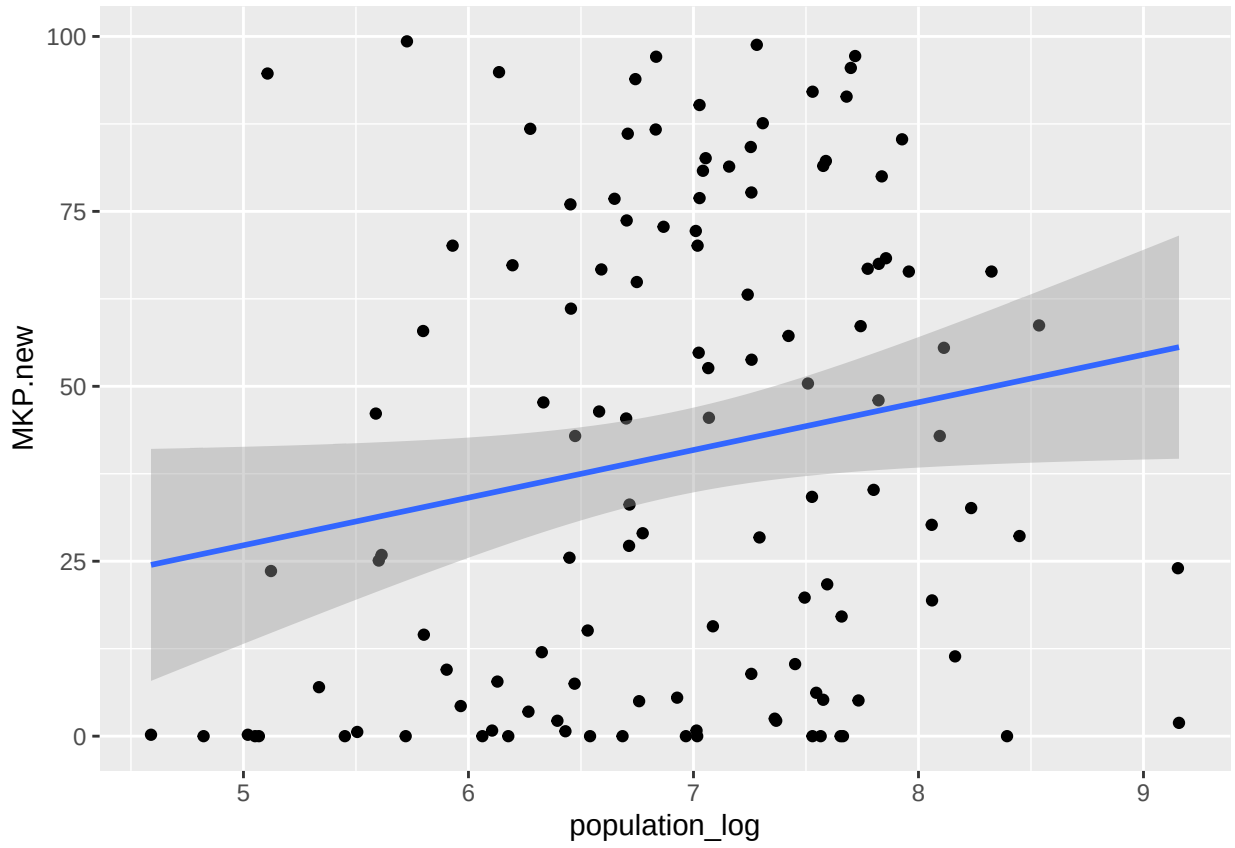



```
ggplot(epi.results.sub, aes(x = population_log, y = MKP.new)) +
geom_point() +
stat_smooth(method = "lm")
```

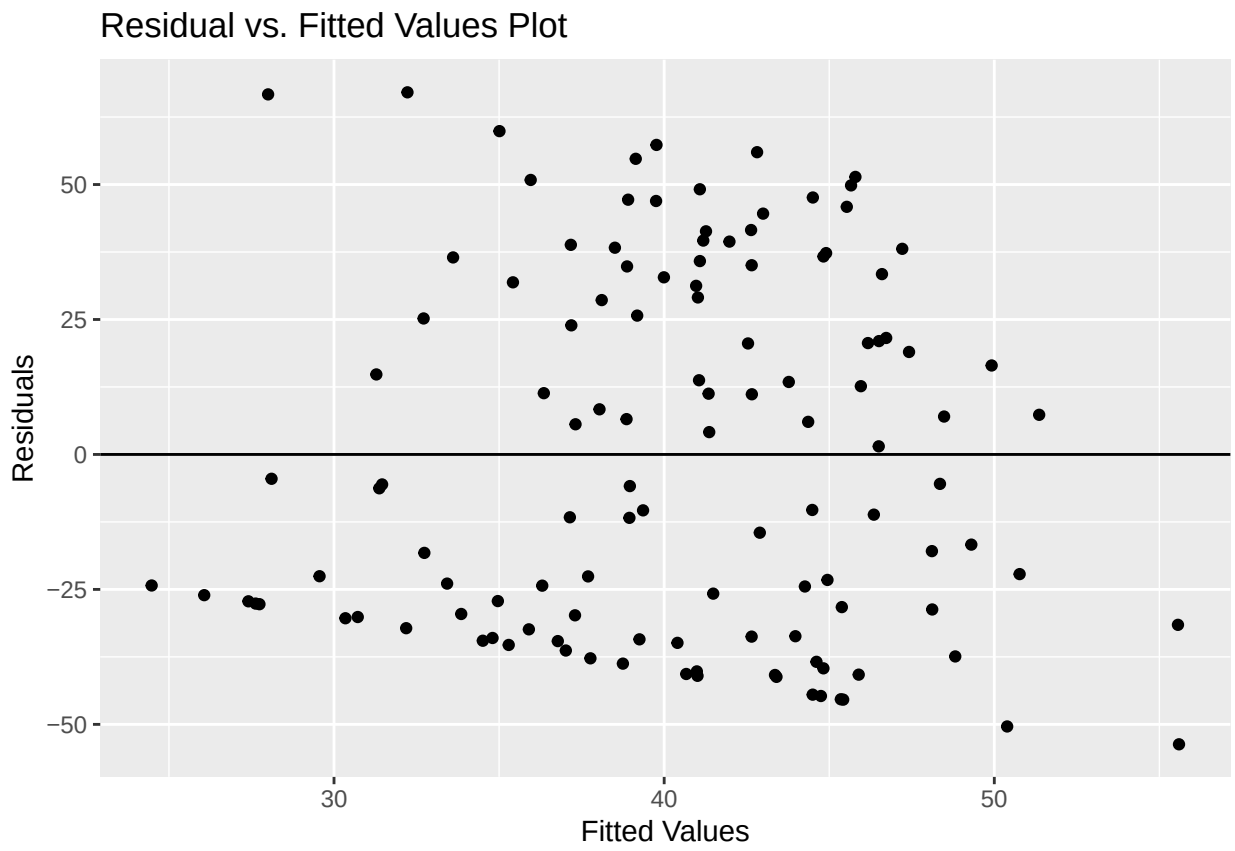
```
## `geom_smooth()` using formula = 'y ~ x'
```

```
## Warning: Removed 44 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```

```
## Warning: Removed 44 rows containing missing values or values outside the scale range
## (`geom_point()`).
```




```
ggplot(lin.mod.epinew, aes(x = .fitted, y = .resid)) +
geom_point() +
geom_hline(yintercept = 0) +
labs(title='Residual vs. Fitted Values Plot', x='Fitted Values', y='Residuals')
```


Residual vs. Fitted Values Plot
Residuals
50
25
0
−25
−50
30
40
50
Fitted Values