Assignment 2

Max Troeger

2025-02-21

1 Variable Distributions

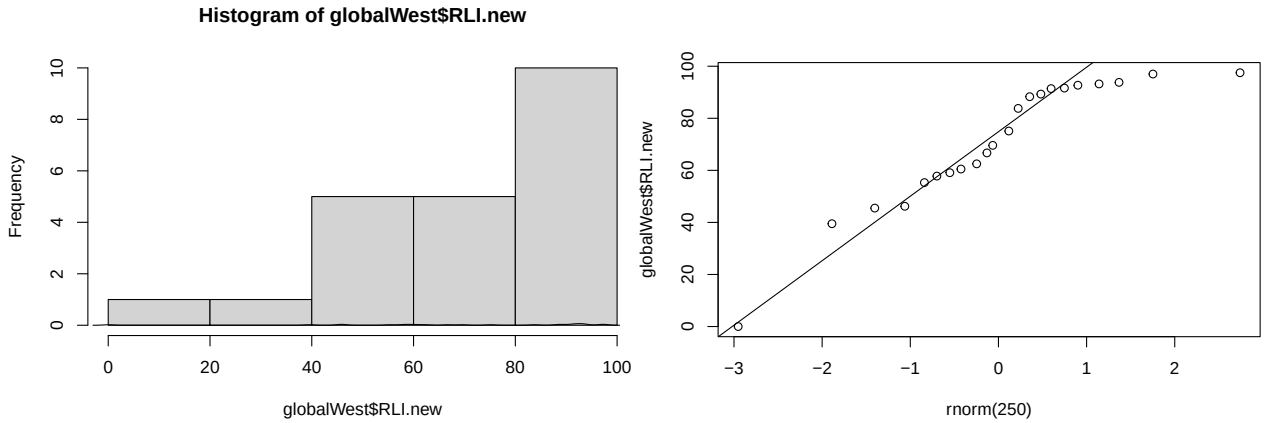
From the EPI dataset, we derive subsets for the Greater Middle East and the Global West

```
middleEast <- subset(epi_pop, region == 'Greater Middle East')
globalWest <- subset(epi_pop, region == 'Global West')
```

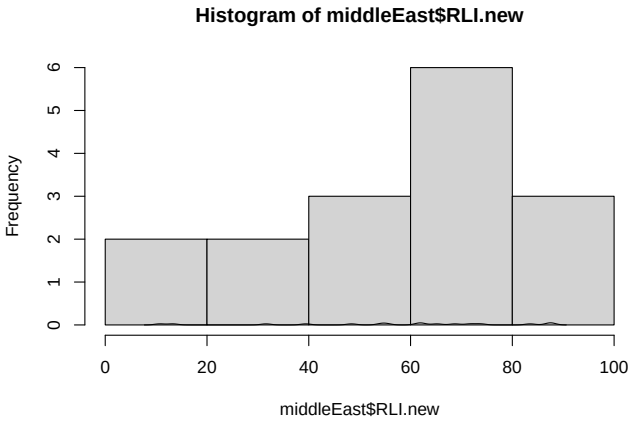
1.1 Histograms and Density Lines

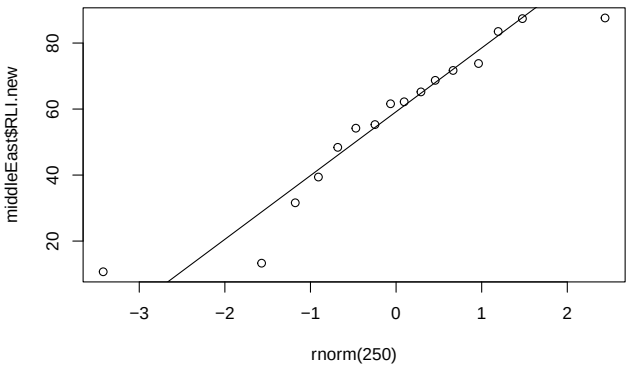
We consider the `RLI.new` variable

```
hist(globalWest$RLI.new)
lines(density(globalWest$RLI.new,bw=1.))
qqplot(rnorm(250), globalWest$RLI.new)
qqline(globalWest$RLI.new)
```




```
hist(middleEast$RLI.new)
lines(density(middleEast$RLI.new,bw=1.))
qqplot(rnorm(250), middleEast$RLI.new)
qqline(middleEast$RLI.new)
```





2 Linear Models

2.1 Initial Models

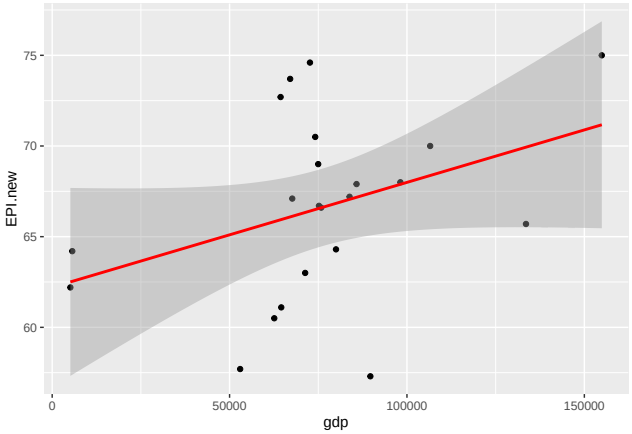
```
fit1 <- lm(EPI.new~gdp+population,data=globalWest)
coeftest(fit1)
```

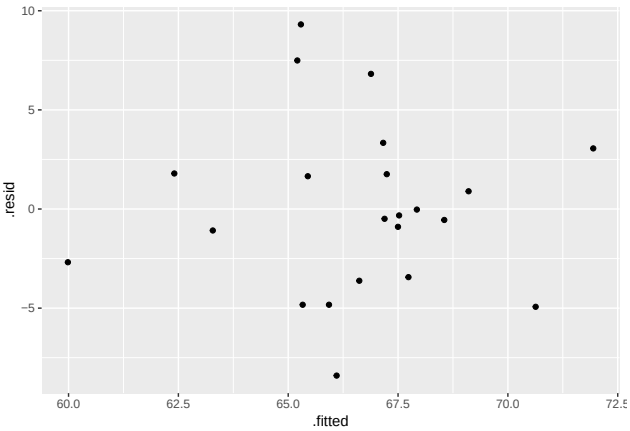
```
##
## t test of coefficients:
##
##                Estimate  Std. Error t value Pr(>|t|)
## (Intercept)  6.3247e+01  2.5725e+00 24.5860 7.25e-16 ***
## gdp          5.6236e-05  3.0477e-05  1.8452  0.08065 .
## population  -2.4182e-08  1.3664e-08 -1.7698  0.09281 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
ggplot(globalWest, aes(x = gdp, y = EPI.new)) +
  geom_point() +
  stat_smooth(method = "lm", col="red")
```

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

```
ggplot(fit1, aes(x = .fitted, y = .resid)) + geom_point()
```





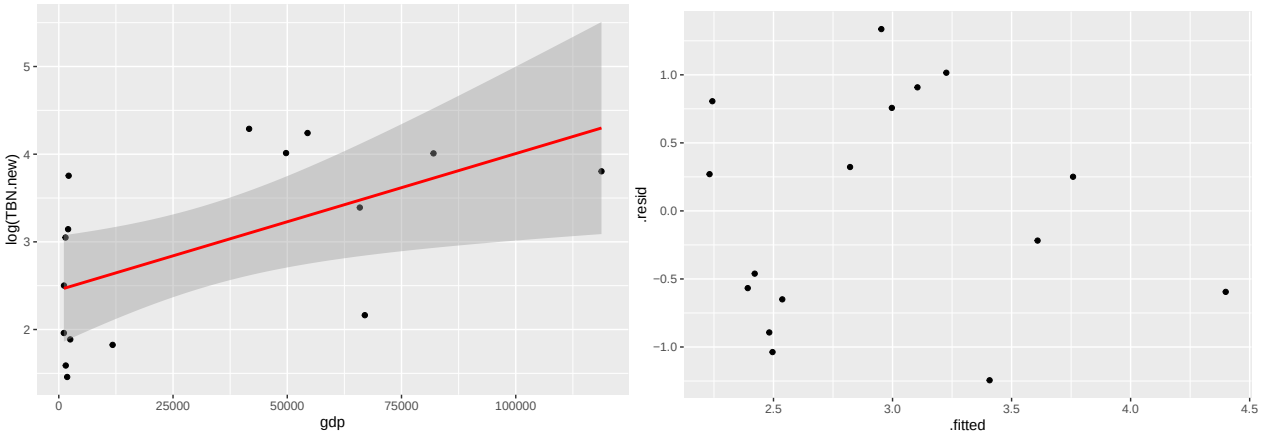

```
fit2 <- lm(log(TBN.new)~gdp+population,data=middleEast)
coeftest(fit2)
```

```
##
## t test of coefficients:
##
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2.1270e+00 4.5948e-01  4.6292 0.000472 ***
## gdp         1.8952e-05 7.0804e-06  2.6767 0.019018 *
## population  7.2393e-09 7.9675e-09  0.9086 0.380094
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
ggplot(middleEast, aes(x = gdp, y = log(TBN.new))) +
  geom_point() +
  stat_smooth(method = "lm", col="red")
```

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

```
ggplot(fit2, aes(x = .fitted, y = .resid)) + geom_point()
```



Now With Subsetting We take the subset of the "Global West" that is classified as "Western Europe":

```
westernEurope <- subset(globalWest, country == "Belgium" |
                                    country == "France" |
                                    country == "Ireland" |
                                    country == "Luxembourg" |
                                    country == "Netherlands" |
                                    country == "United Kingdom")
fit3 <- lm(EPI.new~gdp+population,data=westernEurope)
coeftest(fit3)
```

```
##
## t test of coefficients:
##
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 5.8286e+01 9.0122e+00  6.4675 0.007501 **
## gdp         8.5659e-05 7.0882e-05  1.2085 0.313432
## population  8.8070e-08 8.7533e-08  1.0061 0.388475
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The predictive accuracy of this model is actually worse than including all of the "Global West". This is likely because of the great variance in the size of population and gdp because of these countries; moreover, including all of the countries in the "Global West" makes the relationship between `EPI.new` and `gdp/population` clearer.

3 Classification (kNN)

3.1 Initial Model

```
twoRegions <- subset(epi_pop, region == "Global West" | region == "Greater Middle East")
twoRegions <- subset(twoRegions, select = c(region,EPI.new, RLI.new, APO.new))
knn.tR <- knn(train = twoRegions[,2:4], test = twoRegions[,2:4],
                                        cl = twoRegions$region, k = 5)
confuse.tR <- table(knn.tR, twoRegions$region, dnn=list('predicted','actual'))
confuse.tR
```

```
##                      actual
## predicted             Global West Greater Middle East
##   Global West                  21                   1
##   Greater Middle East           1                  15
```

We have (21+15)/38*100=94.74% accuracy.

3.2 Alternate Specification

```
twoRegions <- subset(epi_pop, region == "Global West" | region == "Greater Middle East")
twoRegions <- subset(twoRegions, select = c(region,SPI.new, TBN.new, NXA.new))
knn.tR <- knn(train = twoRegions[,2:4], test = twoRegions[,2:4],
                                        cl = twoRegions$region, k = 5)
confuse.tR <- table(knn.tR, twoRegions$region, dnn=list('predicted','actual'))
confuse.tR
```

```
##                      actual
## predicted             Global West Greater Middle East
##   Global West                  22                   1
##   Greater Middle East           0                  15
```

```
(22+15)/38*100
```

```
## [1] 97.36842
```

We have (22+15)/38*100=97.37% accuracy. This model is better because it correctly classifies one more country into the correct region compared to the original specification.