Example: Log transformations in regression

10/12/2022

Data source: Albyn Jones http://people.reed.edu/~jones/141/Guns.html.

Variables: Socioeconomic data from 1990/1991 -

- pop = population of state (in 1000s of people)
- area = area of state (in 1000 square miles)
- urban = percent urban population
- poverty = percent below poverty line
- gunreg = whether there are gun registration laws or not
- homicides = number of homicides in the past year

Research question: Are gun registration laws associated with the rate of homicides in a state?

Data import

```
"GunReg" <-
structure(.Data = list(
"pop" = c(4089, 2372, 30380, 3291, 598, 13277, 1135, 2795, 11543, 5996,
   4860, 9368, 4432, 5158, 6737, 635, 7760, 18058, 10939, 11961, 1004,
   3560, 4953, 17349, 1770, 5018, 570, 3750, 3377, 680, 6623,
   1039, 5610, 2495, 3713, 4252, 1235, 2592, 808, 1593, 1105,
   1548, 1284, 3175, 2922, 703, 6286, 567, 4955, 1801, 460.),
"area" = c(52.4, 53.2, 163.7, 5.5, 0.1, 65.8, 10.9, 56.3, 57.9, 10.6,
   12.4, 96.8, 86.9, 69.7, 53.8, 70.7, 8.7, 54.5, 44.8, 46.1, 1.5, 32,
   42.1, 268.6, 84.9, 71.3, 656.4, 114, 104.1, 2.5, 59.4, 83.6, 36.4,
   82.3, 40.4, 51.8, 35.4, 48.4, 147, 77.4, 9.4, 121.6, 110.6, 69.9,
   98.4, 77.1, 42.8, 9.6, 65.5, 24.2, 97.8),
"urban" = c(60, 54, 93, 79, 100, 85, 89, 61, 85, 84, 81, 70, 71, 53,
   50, 53, 89, 84, 74, 69, 86, 55, 61, 80, 87, 76, 68, 88, 82,
   73, 63, 57, 65, 69, 52, 68, 45, 47, 53, 66, 51, 73, 88,
   68, 71, 50, 69, 32, 66, 36, 65.),
"poverty" = c(19, 18.4, 14.2, 5.8, 19.2, 14.1, 10, 10.1, 13.3, 10.2,
   9.3, 13.9, 12, 13.6, 13.2, 13.5, 9, 14.1, 11.8, 10.8, 8.2, 16.5,
   16.9, 16.8, 9.8, 26.2, 11.2, 14.2, 12.1, 8.1, 16, 13.7, 14.1, 11.1,
   17.4, 22, 12.5, 23.8, 15.8, 10.9, 7.1, 20.9, 10.7, 15.8, 11.3, 13.5,
   10.6, 7.1, 9.2, 17.2, 10.6),
"gunreg" = c(1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
   1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0,
   0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0.),
"homicides" = c(410, 240, 3710, 170, 489, 1300, 44, 62, 1270, 200, 540,
   1020, 100, 550, 730, 11, 350, 2550, 760, 740, 38, 350, 470, 2660,
   43, 220, 56, 290, 155, 32, 720, 21, 380, 150, 260, 760,
   23, 370, 29, 43, 32, 160, 135, 220, 120, 9, 550, 24, 240,
   135, 20.)),
names = c("pop", "area", "urban", "poverty", "gunreg", "homicides"),
row.names = c("AL", "AR", "CA", "CT", "DC", "FL", "HI", "IA", "IL", "MA",
```

```
   "MD", "MI", "MN", "MO", "NC", "ND", "NJ", "NY", "OH", "PA", "RI", "SC",
   "TN", "TX", "UT", "WA", "AK", "AZ", "CO", "DE", "GA", "ID", "IN", "KS",
   "KY", "LA", "ME", "MS", "MT", "NE", "NH", "NM", "NV", "OK", "OR", "SD",
   "VA", "VT", "WI", "WV", "WY"), class = "data.frame")
```

Exploratory data analysis

Examine data set

```
names(GunReg)
```

```
## [1] "pop"       "area"      "urban"     "poverty"   "gunreg"    "homicides"
```

```
dim(GunReg)
```

```
## [1] 51  6
```

```
summary(GunReg)
```

```
##       pop             area            urban           poverty
##  Min.   :  460   Min.   :  0.10   Min.   : 32.00   Min.   : 5.80
##  1st Qu.: 1260   1st Qu.: 35.90   1st Qu.: 56.00   1st Qu.:10.60
##  Median : 3377   Median : 56.30   Median : 69.00   Median :13.30
##  Mean   : 4945   Mean   : 74.26   Mean   : 68.51   Mean   :13.47
##  3rd Qu.: 5803   3rd Qu.: 84.25   3rd Qu.: 81.50   3rd Qu.:15.90
##  Max.   :30380   Max.   :656.40   Max.   :100.00   Max.   :26.20
##      gunreg         homicides
##  Min.   :0.0000   Min.   :   9.0
##  1st Qu.:0.0000   1st Qu.:  50.0
##  Median :1.0000   Median : 220.0
##  Mean   :0.5098   Mean   : 469.8
##  3rd Qu.:1.0000   3rd Qu.: 545.0
##  Max.   :1.0000   Max.   :3710.0
```

```
str(GunReg)
```

```
## 'data.frame':    51 obs. of  6 variables:
##  $ pop      : num  4089 2372 30380 3291 598 ...
##  $ area     : num  52.4 53.2 163.7 5.5 0.1 ...
##  $ urban    : num  60 54 93 79 100 85 89 61 85 84 ...
##  $ poverty  : num  19 18.4 14.2 5.8 19.2 14.1 10 10.1 13.3 10.2 ...
##  $ gunreg   : num  1 1 1 1 1 1 1 1 1 1 ...
##  $ homicides: num  410 240 3710 170 489 1300 44 62 1270 200 ...
```

```
glimpse(GunReg)
```

```
## Rows: 51
## Columns: 6
## $ pop       <dbl> 4089, 2372, 30380, 3291, 598, 13277, 1135, 2795, 11543, 5996~
## $ area      <dbl> 52.4, 53.2, 163.7, 5.5, 0.1, 65.8, 10.9, 56.3, 57.9, 10.6, 1~
## $ urban     <dbl> 60, 54, 93, 79, 100, 85, 89, 61, 85, 84, 81, 70, 71, 53, 50,~
## $ poverty   <dbl> 19.0, 18.4, 14.2, 5.8, 19.2, 14.1, 10.0, 10.1, 13.3, 10.2, 9~
## $ gunreg    <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ~
## $ homicides <dbl> 410, 240, 3710, 170, 489, 1300, 44, 62, 1270, 200, 540, 1020~
```

```
# Top 10 population
GunReg %>% arrange(desc(pop)) %>% slice_head(n = 10)
```

```
##      pop  area urban poverty gunreg homicides
## CA 30380 163.7    93    14.2      1      3710
## NY 18058  54.5    84    14.1      1      2550
## TX 17349 268.6    80    16.8      1      2660
## FL 13277  65.8    85    14.1      1      1300
## PA 11961  46.1    69    10.8      1       740
## IL 11543  57.9    85    13.3      1      1270
## OH 10939  44.8    74    11.8      1       760
## MI  9368  96.8    70    13.9      1      1020
## NJ  7760   8.7    89     9.0      1       350
## NC  6737  53.8    50    13.2      1       730
```

```
# Top 10 number of homicides
GunReg %>% arrange(desc(pop)) %>% slice_head(n = 10)
```

```
##      pop  area urban poverty gunreg homicides
## CA 30380 163.7    93    14.2      1      3710
## NY 18058  54.5    84    14.1      1      2550
## TX 17349 268.6    80    16.8      1      2660
## FL 13277  65.8    85    14.1      1      1300
## PA 11961  46.1    69    10.8      1       740
## IL 11543  57.9    85    13.3      1      1270
## OH 10939  44.8    74    11.8      1       760
## MI  9368  96.8    70    13.9      1      1020
## NJ  7760   8.7    89     9.0      1       350
## NC  6737  53.8    50    13.2      1       730
```

```
# Bottom 10 number of homicides
GunReg %>% arrange(desc(pop)) %>% slice_tail(n = 10)
```

```
##     pop  area urban poverty gunreg homicides
## ID 1039  83.6    57    13.7      0        21
## RI 1004   1.5    86     8.2      1        38
## MT  808 147.0    53    15.8      0        29
## SD  703  77.1    50    13.5      0         9
## DE  680   2.5    73     8.1      0        32
## ND  635  70.7    53    13.5      1        11
## DC  598   0.1   100    19.2      1       489
## AK  570 656.4    68    11.2      0        56
## VT  567   9.6    32     7.1      0        24
## WY  460  97.8    65    10.6      0        20
```

Create new variables

```
# Gun registration character vector
# Homicide rates (per 1,000 person-years):
GunReg <- GunReg %>% mutate(
  gunreg_ind = gunreg,
  gunreg = case_when(
    gunreg_ind == 0 ~ "No",
    gunreg_ind == 1 ~ "Yes"
  ),
  rate = homicides/pop
  )
```

```r
# Top 10 homicide rates
GunReg %>% arrange(desc(rate)) %>% slice_head(n = 10)
```

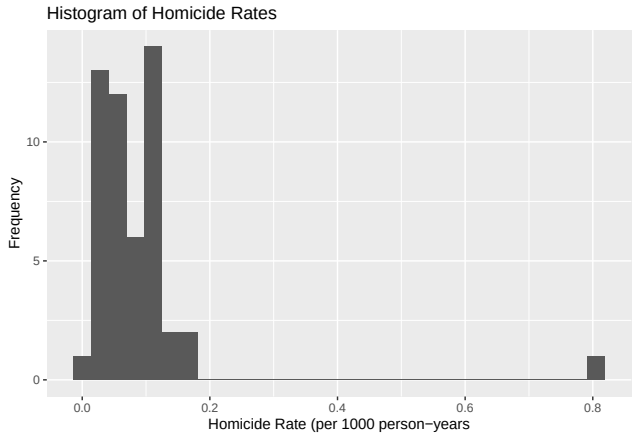
```
##      pop  area urban poverty gunreg homicides gunreg_ind      rate
## DC   598   0.1   100    19.2    Yes       489          1 0.8177258
## LA  4252  51.8    68    22.0     No       760          0 0.1787394
## TX 17349 268.6    80    16.8    Yes      2660          1 0.1533230
## MS  2592  48.4    47    23.8     No       370          0 0.1427469
## NY 18058  54.5    84    14.1    Yes      2550          1 0.1412117
## CA 30380 163.7    93    14.2    Yes      3710          1 0.1221198
## MD  4860  12.4    81     9.3    Yes       540          1 0.1111111
## IL 11543  57.9    85    13.3    Yes      1270          1 0.1100234
## MI  9368  96.8    70    13.9    Yes      1020          1 0.1088813
## GA  6623  59.4    63    16.0     No       720          0 0.1087121
```

```r
# Homicide rate by gun registration law status
GunReg %>% group_by(gunreg) %>%
  summarize(
    tot_pop = sum(pop),
    tot_hom = sum(homicides),
    overall_rate = sum(homicides)/sum(pop)
    )
```

```
## # A tibble: 2 x 4
##   gunreg tot_pop tot_hom overall_rate
##   <chr>    <dbl>   <dbl>        <dbl>
## 1 No       63143    4934       0.0781
## 2 Yes     189038   19027       0.101
```

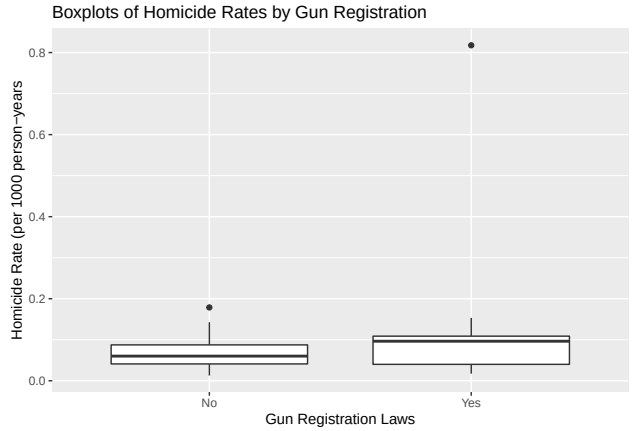
Data visualization

```r
# Distribution of homicide rates
GunReg %>% ggplot(aes(x = rate)) +
  geom_histogram(bins = 30) +
  labs(x = "Homicide Rate (per 1000 person-years",
       y = "Frequency",
       title = "Histogram of Homicide Rates")
```

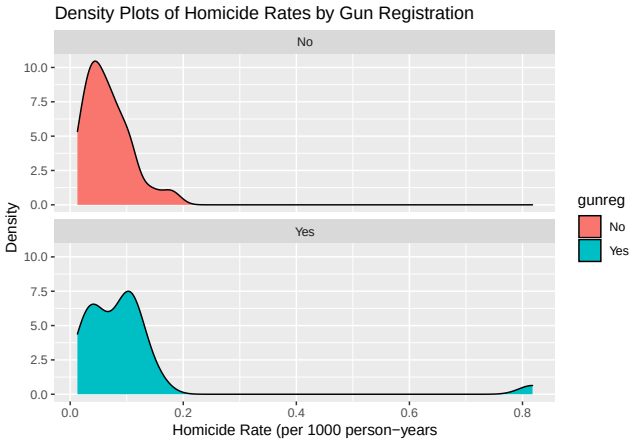



```r
# Homicide rates by gun registration status
GunReg %>% ggplot(aes(x = gunreg, y = rate)) +
  geom_boxplot() +
```

```
  labs(x = "Gun Registration Laws",
       y = "Homicide Rate (per 1000 person-years",
       title = "Boxplots of Homicide Rates by Gun Registration")
```




```
GunReg %>% ggplot(aes(x = rate, fill = gunreg)) +
  geom_density() + facet_wrap(vars(gunreg), nrow = 2) +
  labs(x = "Homicide Rate (per 1000 person-years",
       y = "Density",
       title = "Density Plots of Homicide Rates by Gun Registration")
```



Regression Modeling

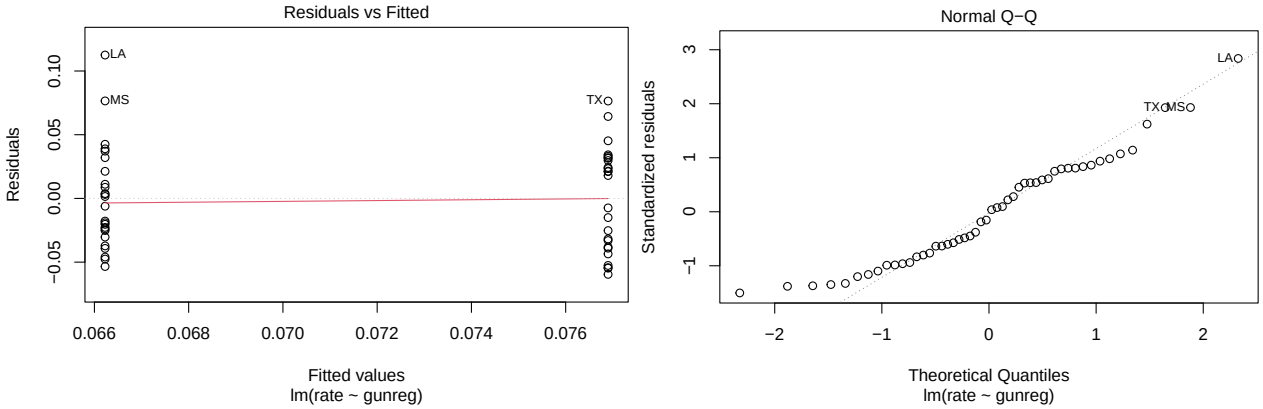
Model without the DC outlier:

```
mod1 <- lm(rate ~ gunreg, data = GunReg, subset = GunReg$rate < 0.2)
summary(mod1)
```

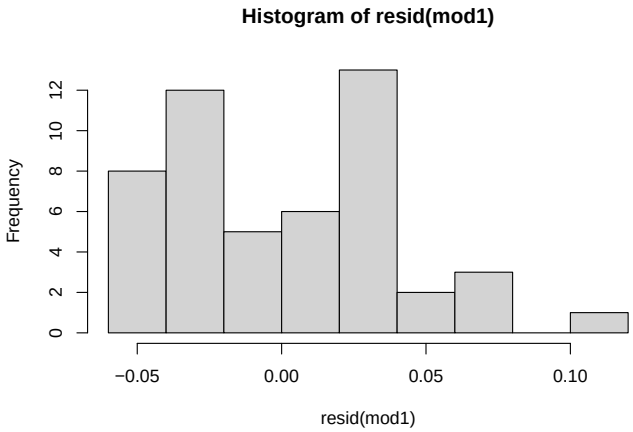
```
##
## Call:
## lm(formula = rate ~ gunreg, data = GunReg, subset = GunReg$rate <
##     0.2)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.059577 -0.032743 -0.002298  0.031025  0.112513
##
```

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.066226   0.008091   8.185 1.16e-10 ***
## gunregYes   0.010674   0.011443   0.933    0.356
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.04046 on 48 degrees of freedom
## Multiple R-squared:  0.01781,    Adjusted R-squared:  -0.002657
## F-statistic: 0.8702 on 1 and 48 DF,  p-value: 0.3556
```

```
plot(mod1, c(1,2))
```




```
hist(resid(mod1))
```



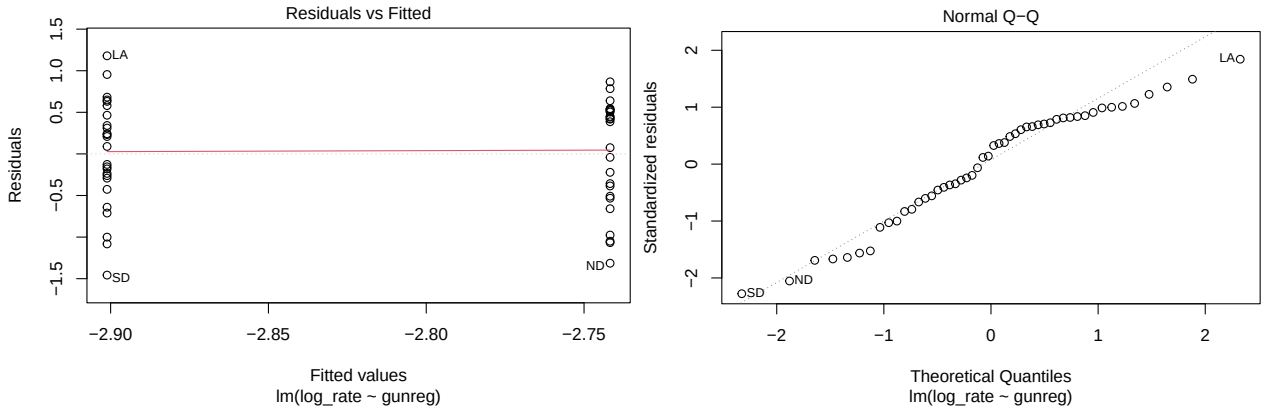

```
# Create log-transformed rate variable
GunReg <- GunReg %>% mutate(log_rate = log(rate))

mod2 <- lm(log_rate ~ gunreg, data = GunReg, subset = GunReg$rate < 0.2)
summary(mod2)
```

```
##
## Call:
## lm(formula = log_rate ~ gunreg, data = GunReg, subset = GunReg$rate <
##     0.2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -1.4570 -0.4160  0.1494  0.5154  1.1793
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -2.9011     0.1306 -22.211   <2e-16 ***
## gunregYes     0.1594     0.1847   0.863    0.393
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6531 on 48 degrees of freedom
## Multiple R-squared:  0.01527,    Adjusted R-squared:  -0.005242
## F-statistic: 0.7445 on 1 and 48 DF,  p-value: 0.3925
```

```
plot(mod2, c(1,2))
```




```
hist(resid(mod2))
```