Exercise 1

Exercise introduction

This is Exercise 1 in Part 3 of the course.

The purpose of the exercise is to cover correlations and regression.

Part 1: Blood pressure and Heart rate

Import the data using `Rcmdr`

```
library(Rcmdr)
```

```
## Warning: package 'Rcmdr' was built under R version 4.0.5
```

```
## Loading required package: splines
```

```
## Loading required package: RcmdrMisc
```

```
## Warning: package 'RcmdrMisc' was built under R version 4.0.5
```

```
## Loading required package: car
```

```
## Loading required package: carData
```

```
##
## Attaching package: 'car'
```

```
## The following object is masked _by_ '.GlobalEnv':
##
##     densityPlot
```

```
## The following object is masked from 'package:dplyr':
##
##     recode
```

```
## The following object is masked from 'package:purrr':
##
##     some
```

```
## Loading required package: sandwich
```

```
## Loading required package: effects
```

```
## Warning: package 'effects' was built under R version 4.0.5
```

```
## Registered S3 methods overwritten by 'lme4':
##   method                          from
##   cooks.distance.influence.merMod car
##   influence.merMod                car
##   dfbeta.influence.merMod         car
##   dfbetas.influence.merMod        car
```

```
## lattice theme set by effectsTheme()
## See ?effectsTheme for details.
```

```
## The Commander GUI is launched only in interactive sessions
```

```
library(car)
library(RcmdrMisc)
# include this code chunk as-is to enable 3D graphs
library(rgl)
```

```
## Warning: package 'rgl' was built under R version 4.0.5
```

```
knitr::knit_hooks$set(webgl = hook_webgl)
bp <-
  read.table("../p02_inputs/bp.txt",
   header=TRUE, stringsAsFactors=TRUE, sep="\t", na.strings="NA", dec=".",
  strip.white=TRUE)
```

Summarizing the data

First using `Rcmdr` Statistics -> Summaries -> Active data set

```
summary(bp)
```

```
##    Heart.rate    Blood.pressure
##  Min.   :50.00   Min.   :149.0
##  1st Qu.:64.25   1st Qu.:167.2
##  Median :69.50   Median :175.5
##  Mean   :70.60   Mean   :178.1
##  3rd Qu.:75.75   3rd Qu.:187.0
##  Max.   :98.00   Max.   :221.0
```

Then using `Rcmdr` Statistics -> Summaries -> Numerical summaries

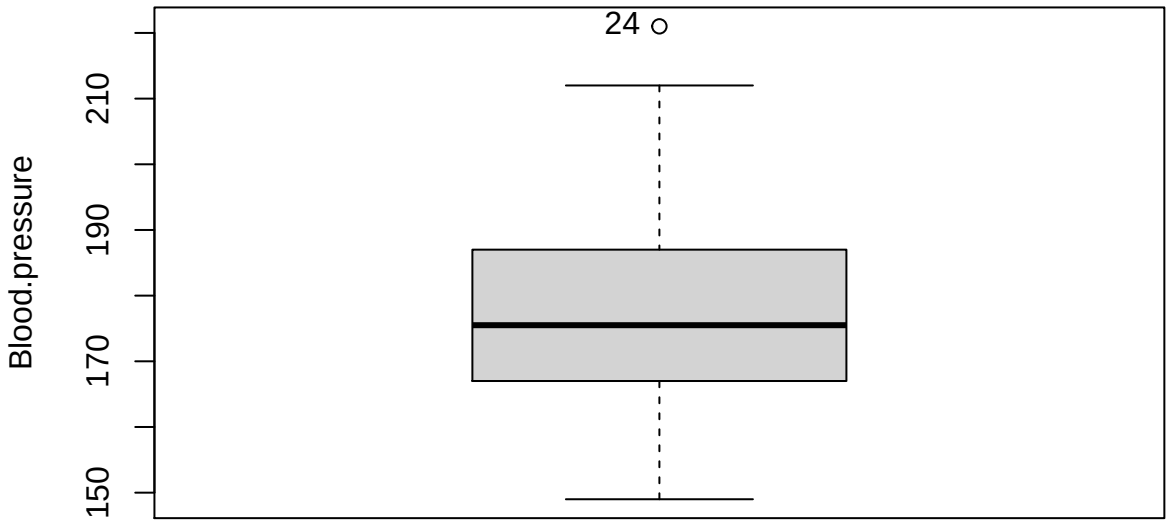
```
library(abind, pos = 65)
library(e1071, pos = 66)

numSummary(bp[,c("Blood.pressure", "Heart.rate"), drop=FALSE],
  statistics=c("mean", "sd", "IQR", "quantiles"), quantiles=c(0,.25,.5,.75,1))
```

```
##                     mean       sd   IQR  0%    25%   50%    75% 100%  n
## Blood.pressure 178.1333 17.01669 19.75 149 167.25 175.5 187.00  221 30
## Heart.rate      70.6000 10.31103 11.50  50  64.25  69.5  75.75   98 30
```

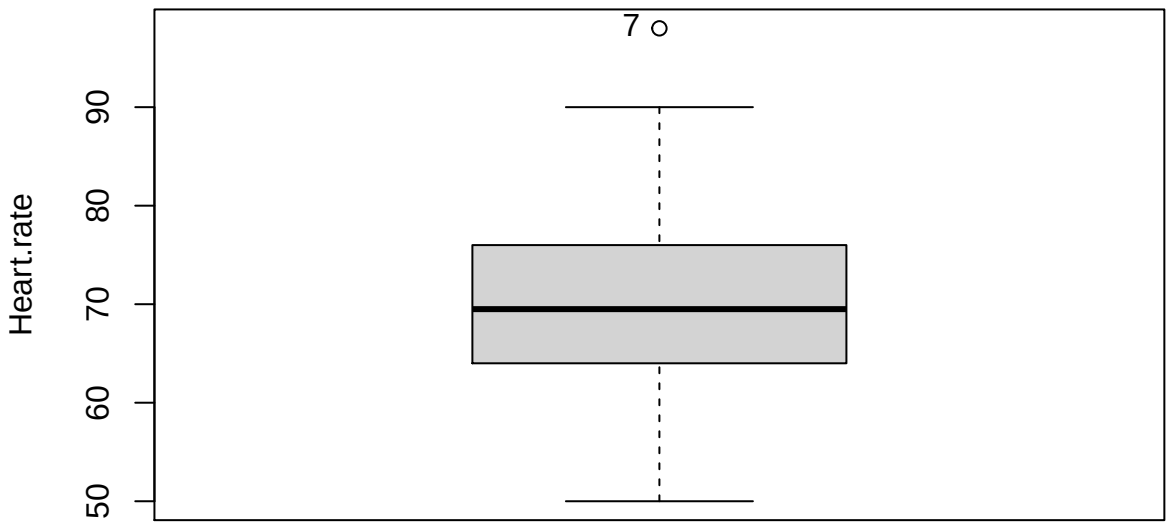
Box plots of the variables

```
Boxplot( ~ Blood.pressure, data=bp, id=list(method="y"))
```




```
## [1] "24"
```

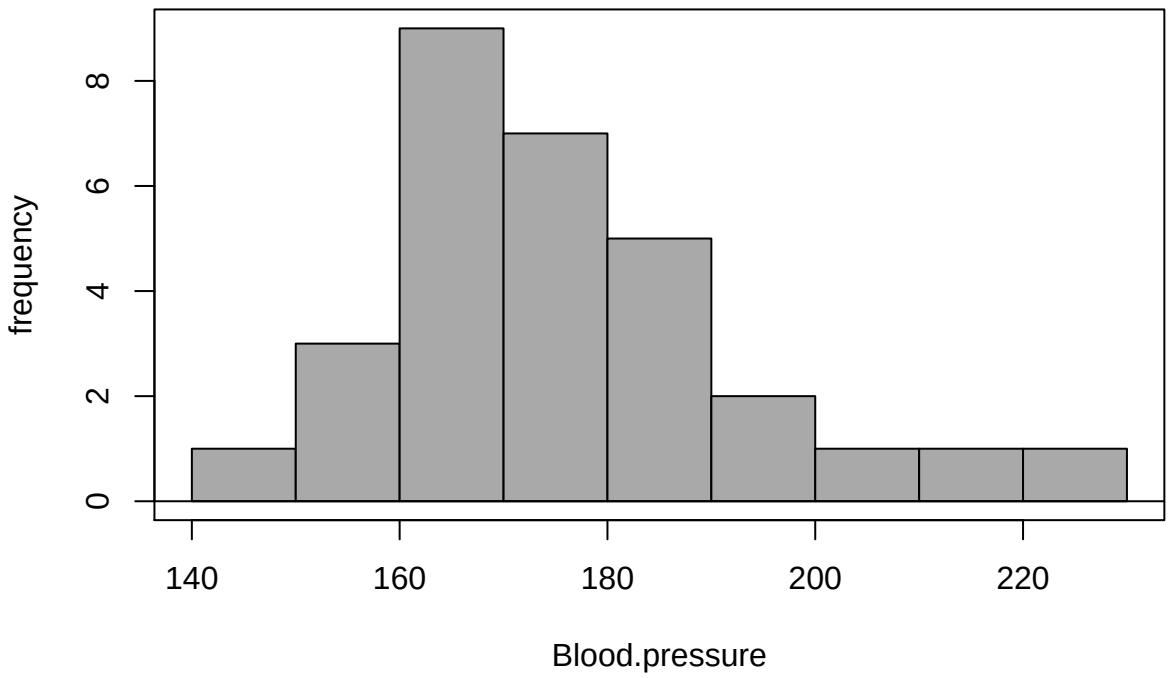
```
Boxplot( ~ Heart.rate, data=bp, id=list(method="y"))
```



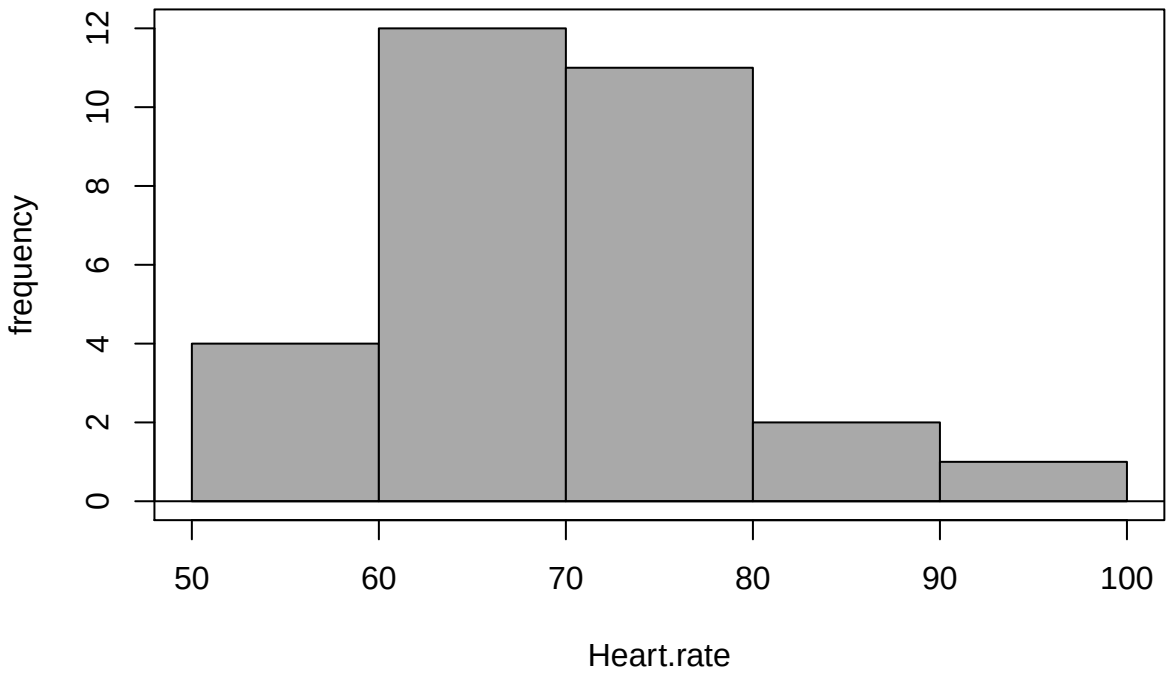

```
## [1] "7"
```

Histograms of variables

```
with(bp, Hist(Blood.pressure, scale="frequency", breaks="Sturges",
  col="darkgray"))
```




```
with(bp, Hist(Heart.rate, scale="frequency", breaks="Sturges",
  col="darkgray"))
```



Both have a slight right-skew.

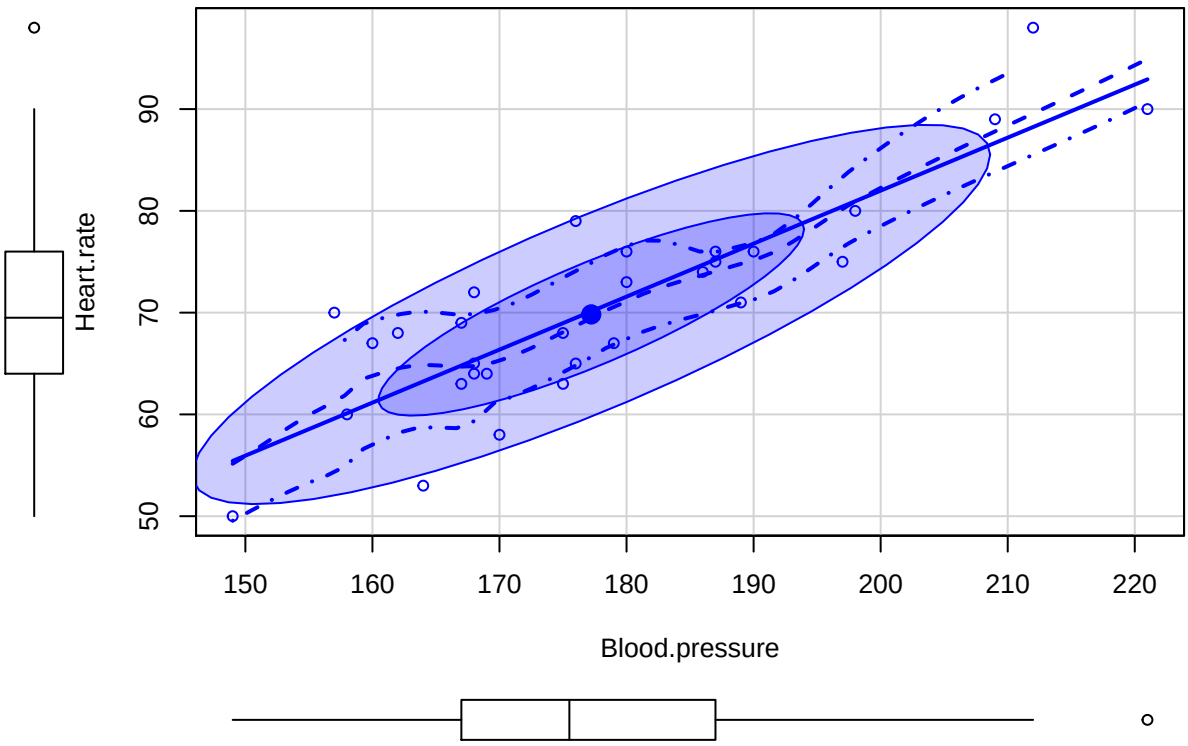
Correlation test

```
with(bp, cor.test(Blood.pressure, Heart.rate, alternative="two.sided",
  method="pearson"))
```

```
##
##  Pearson's product-moment correlation
##
## data:  Blood.pressure and Heart.rate
## t = 8.8993, df = 28, p-value = 0.000000001183
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.7232232 0.9313877
## sample estimates:
##      cor
## 0.859536
```

Scatter plot to support correlation analysis

```
scatterplot(Heart.rate~Blood.pressure, regLine=TRUE, smooth=list(span=0.5,
  spread=TRUE), id=list(method='identify'), boxplots='xy',
  ellipse=list(levels=c(.5, .9)), data=bp)
```



Part 2: CO_2 emissions, temperature and year

Data retrieval

Download the temperature data and CO_2 data

Then pre-process it in Excel and save as tab-delimited text files.

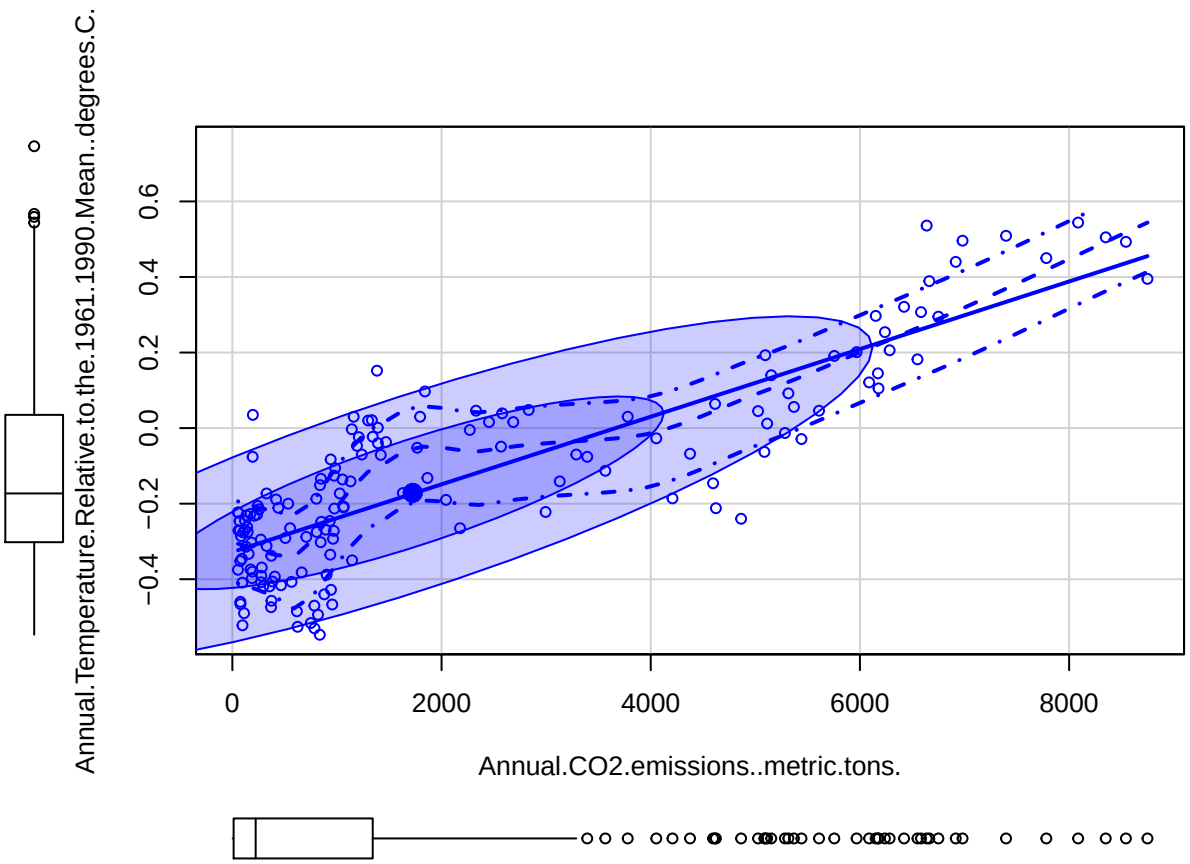
```
tempByCo2 <-
  read.table("D:/Coding/Projects/sh stat and viz R I/Part 3/p02_inputs/tempbyco2.txt",
   header=TRUE, stringsAsFactors=TRUE, sep="\t", na.strings="NA", dec=".",
  strip.white=TRUE)
```

Pearson's product-moment and Spearman's rank

Pearson's product-moment is for normally-distributed data with evenly distributed residuals (homoscedasticity). If either pre-requisite is violated, Spearman's rank correlation is an alternative.

Make a scatter plot first.

```
scatterplot(Annual.Temperature.Relative.to.the.1961.1990.Mean..degrees.C.~Annual.CO2.emissions..metric.t
   regLine=TRUE, smooth=list(span=0.5, spread=TRUE), boxplots='xy',
  ellipse=list(levels=c(.5, .9)), data=tempByCo2)
```



There appears to be a positive relationship between CO_2 emissions and temperature. The box plot indicates that the CO_2 emissions are not normally distributed (right-skew). Furthermore, the scatterplot with a simple line of best fit indicates that the data are heteroscedastic, with residuals tending to be negative when CO_2 emissions are under 1000, positive for emissions between 1000 and 2000, and perhaps homoscedastic otherwise. My intuition then is to use Spearman's rank correlation to determine of the positive relationship is statistically significant.

```
with(tempByCo2, cor.test(Annual.CO2.emissions..metric.tons.,
  Annual.Temperature.Relative.to.the.1961.1990.Mean..degrees.C.,
  alternative="two.sided", method="spearman"))
```

```
## Warning in cor.test.default(Annual.CO2.emissions..metric.tons.,
## Annual.Temperature.Relative.to.the.1961.1990.Mean..degrees.C., : Cannot compute
## exact p-value with ties

##
##  Spearman's rank correlation rho
##
## data:  Annual.CO2.emissions..metric.tons. and Annual.Temperature.Relative.to.the.1961.1990.Mean..degr
## S = 136488, p-value < 0.00000000000000022
## alternative hypothesis: true rho is not equal to 0
```

```
## sample estimates:
##       rho
## 0.7962626
```

The p-value is nearly 0, and thus the relationship is statistically significant.

Regression

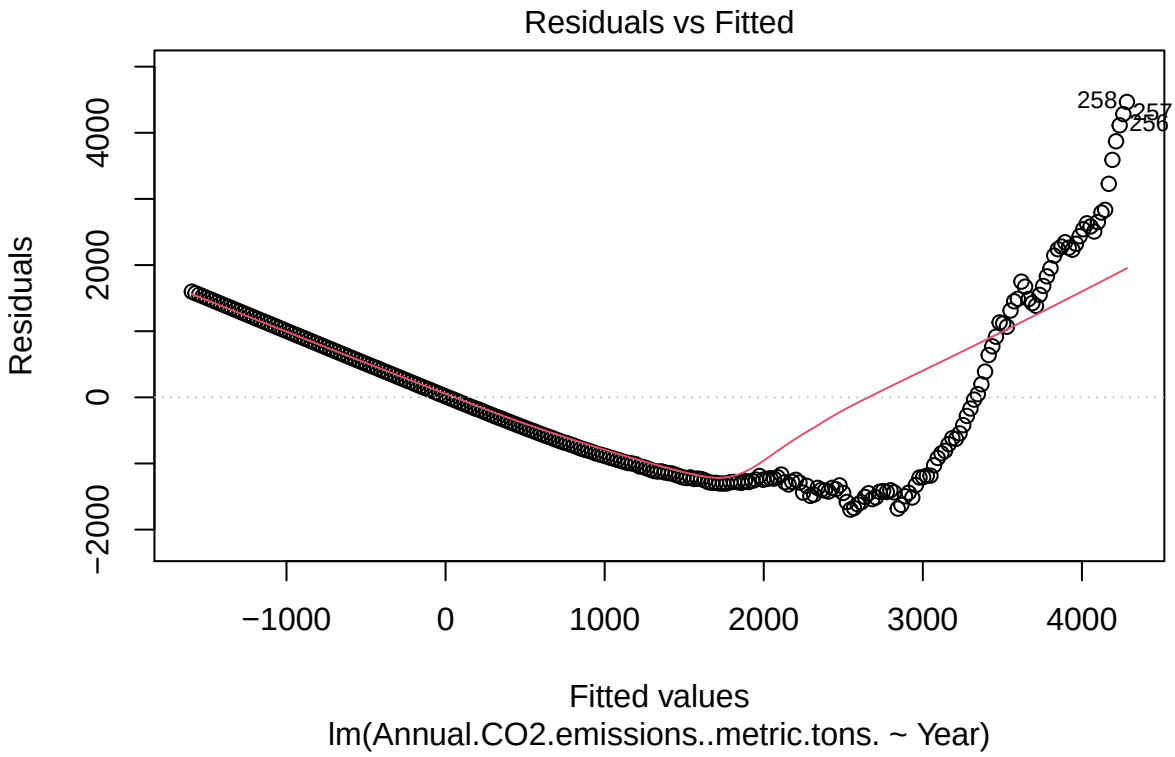
Building a model where CO_2 emissions are dependent on year.

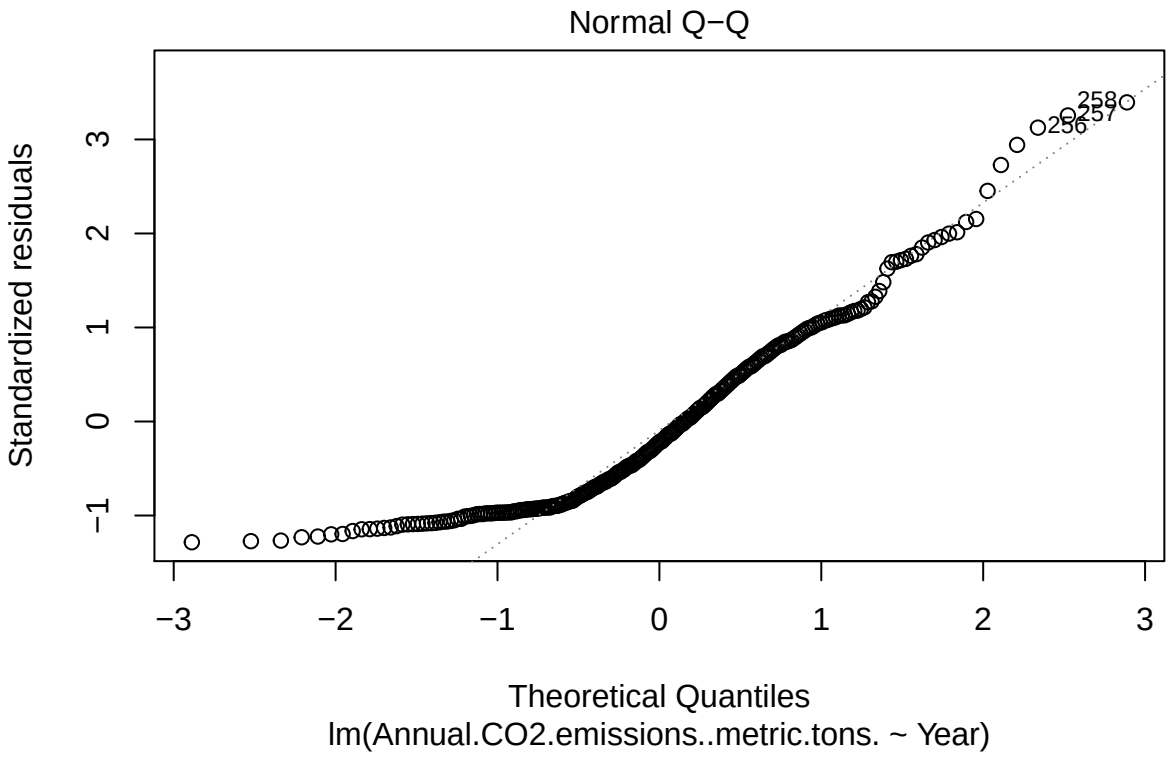
```
RegModel.1 <- lm(Annual.CO2.emissions..metric.tons.~Year, data=tempByCo2)
summary(RegModel.1)
```

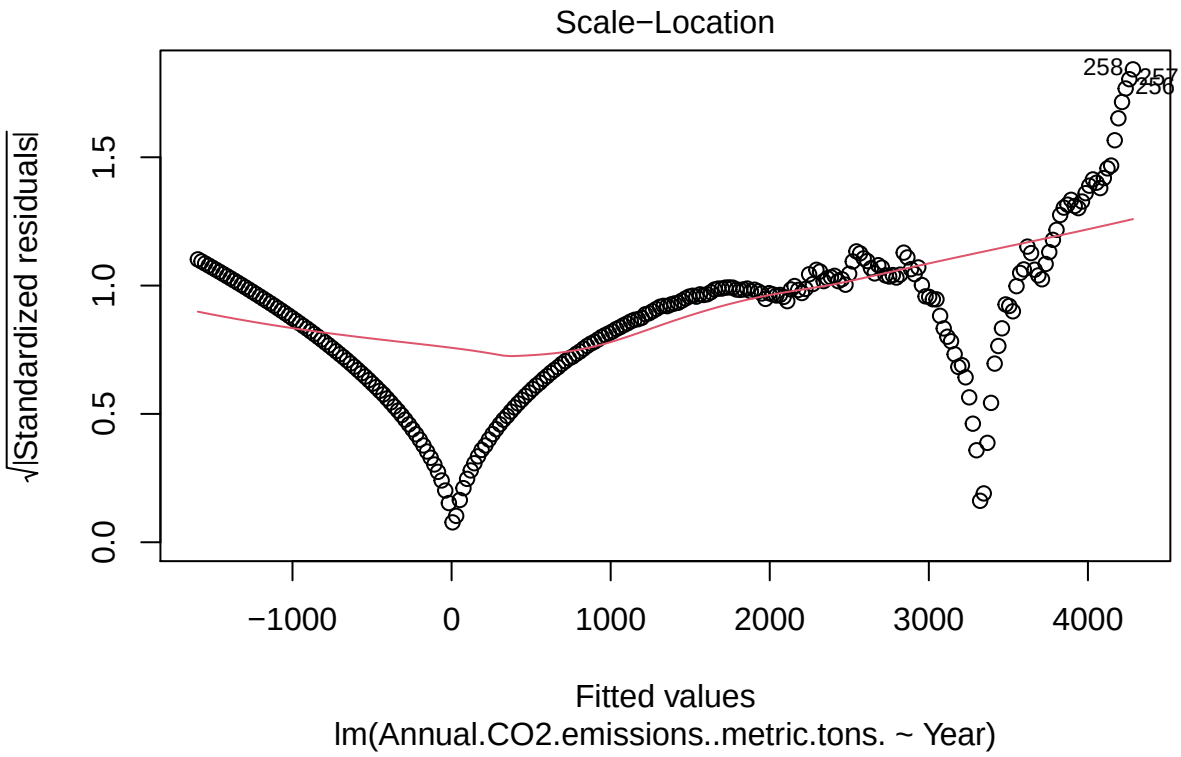
```
##
## Call:
## lm(formula = Annual.CO2.emissions..metric.tons. ~ Year, data = tempByCo2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1697.8 -1203.6  -291.4   952.8  4466.0
##
## Coefficients:
##               Estimate Std. Error t value            Pr(>|t|)
## (Intercept) -41642.179   2084.700  -19.98 <0.0000000000000002 ***
## Year            22.871      1.108   20.64 <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1326 on 256 degrees of freedom
##   (7 observations deleted due to missingness)
## Multiple R-squared:  0.6245, Adjusted R-squared:  0.6231
## F-statistic: 425.8 on 1 and 256 DF,  p-value: < 0.00000000000000022
```

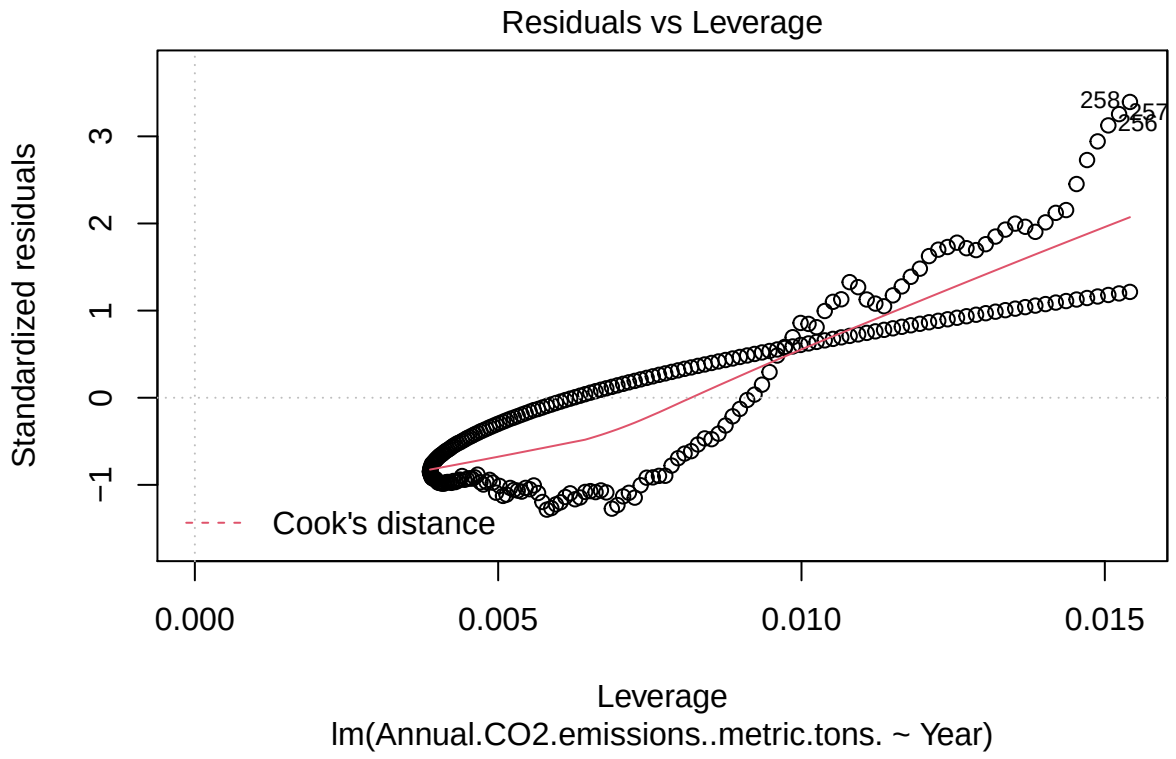
Using diagnostic plots to determine if a model is appropriate. Some resources regarding diagnostic plots: here here here

```
# Commenting out changes to par since want to save the graphs in full-size in the knitted document. Whe
# oldpar <- par(oma=c(0,0,3,0), mfrow=c(2,2))
plot(RegModel.1)
```


Residuals vs Fitted
Residuals
Fitted values
lm(Annual.CO2.emissions..metric.tons. ~ Year)
258
257
256
−2000
0
2000
4000
−1000
0
1000
2000
3000
4000


Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(Annual.CO2.emissions..metric.tons. ~ Year)
258
257
256
−3
−2
−1
0
1
2
3


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
258
257
256
−1000
0
1000
2000
3000
4000
Fitted values
lm(Annual.CO2.emissions..metric.tons. ~ Year)




```
# par(oldpar)
```

The Residuals vs Fitted plot should not have a distinctive pattern; this plot does have a v-shaped pattern, indicating that perhaps the model is not appropriate.

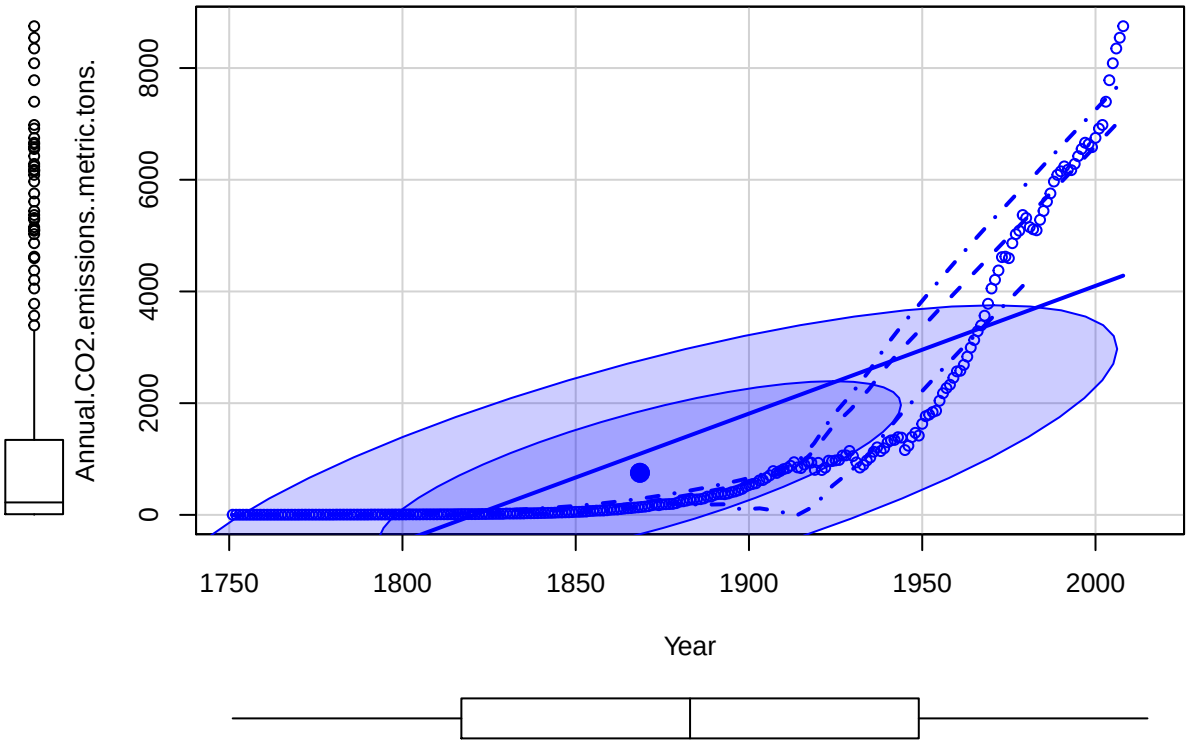
The Normal Q-Q plot should have the points falling close to the upward sloping line without a clear patterned deviation; this plot has many values in the lower-left quadrant that have residuals much lower than the theoretical residuals, indicating that perhaps the model is not appropriate.

The Scale-Location / Spread-Location figure is related to the Residuals vs Fitted plot, although the y-axis is the square root of the standardized residuals. When a model is a good fit then you should observe a horizontal line and no clear pattern to how the points are distributed around the line; in this case, there is a distinctive 'W' shape to the points, indicating the model is not a good fit.

The final diagnostic plot is the Residuals vs Leverage plot. This plot shows how the individual data points influence the overall regression model. Points that have a high Cook's distance score inordinately influence the model and might therefore be considered outliers. Every model will have Cook's distance plotted differently and the shape/pattern of the points in this figure are not specifically useful for interpretation, but rather if any points have high Cook's distance; in this case, no points have a high Cook's distance.

A scatterplot can also be useful for determining the appropriateness of a linear model.

```
scatterplot(Annual.CO2.emissions..metric.tons.~Year, regLine=TRUE,
  smooth=list(span=0.5, spread=TRUE), boxplots='xy', ellipse=list(levels=c(.5,
   .9)), data=tempByCo2)
```



This relationship is not well-characterized by a simple straight line. This, plus the interpretations from the diagnostic plots means that the a different model needs to be developed.

Alternative regression models

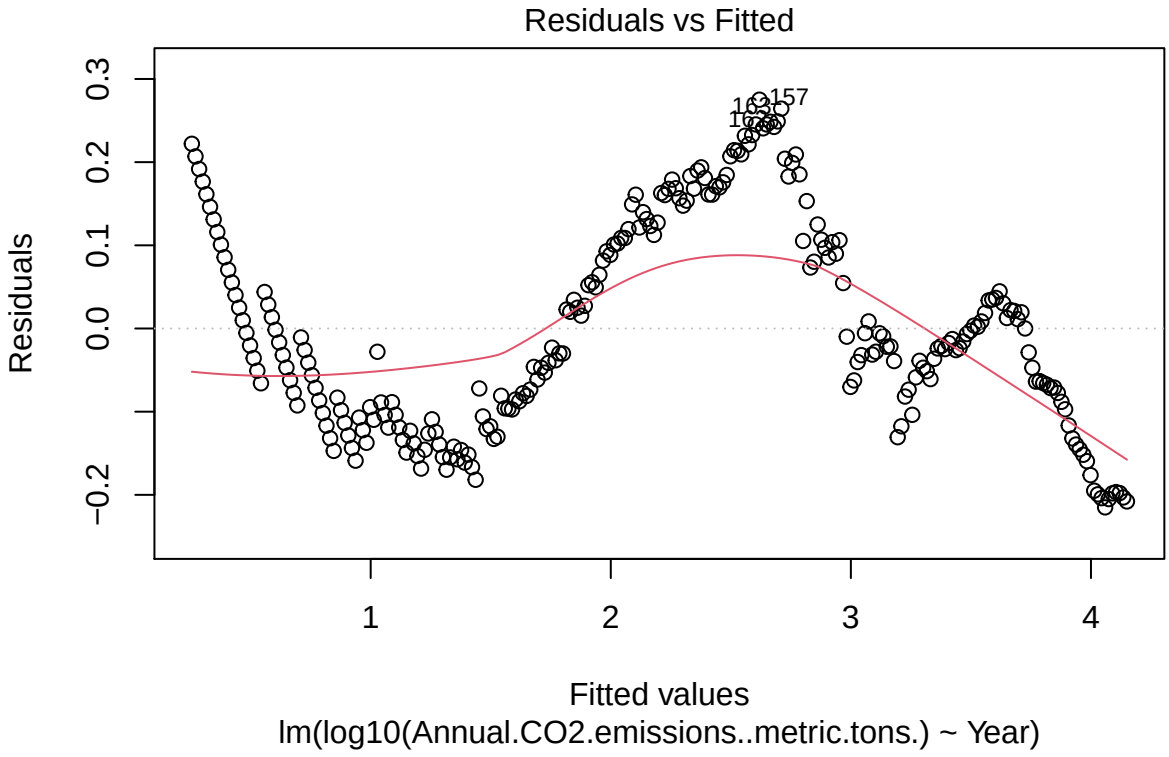
First will try log10 transformation of CO_2.

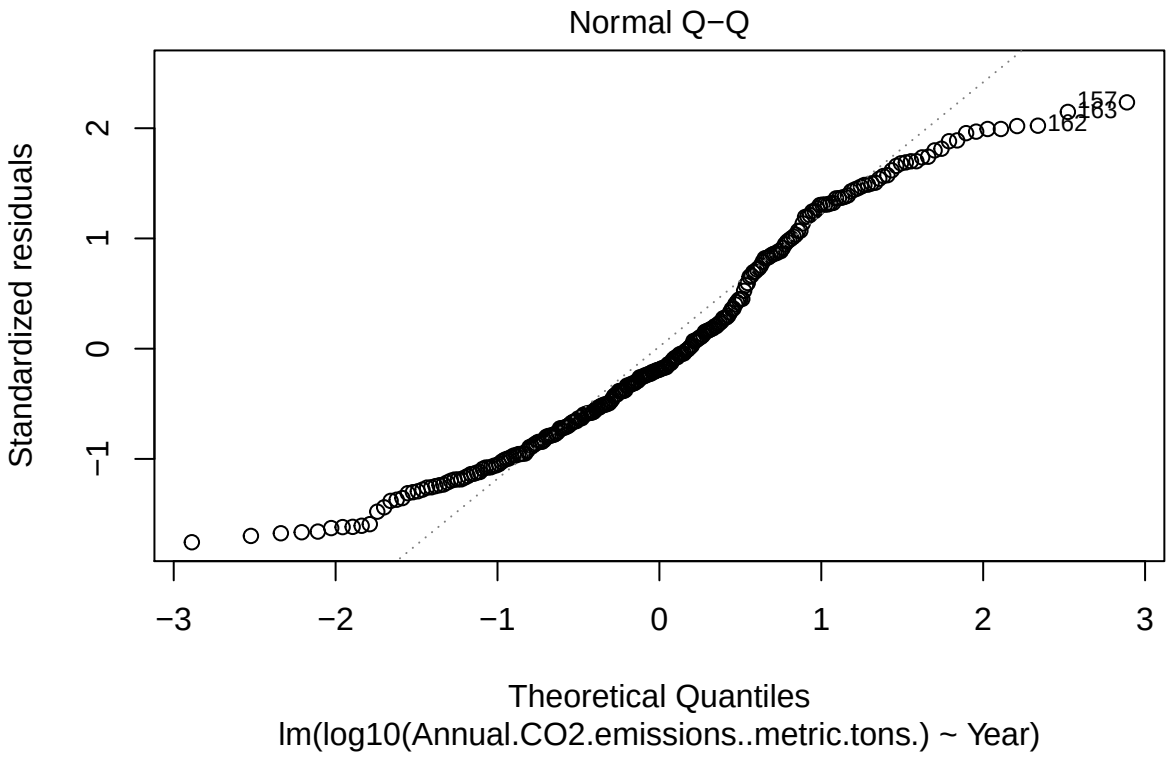
```
RegModel.2 <- lm(log10(Annual.CO2.emissions..metric.tons.)~Year, data=tempByCo2)
summary(RegModel.2)
```

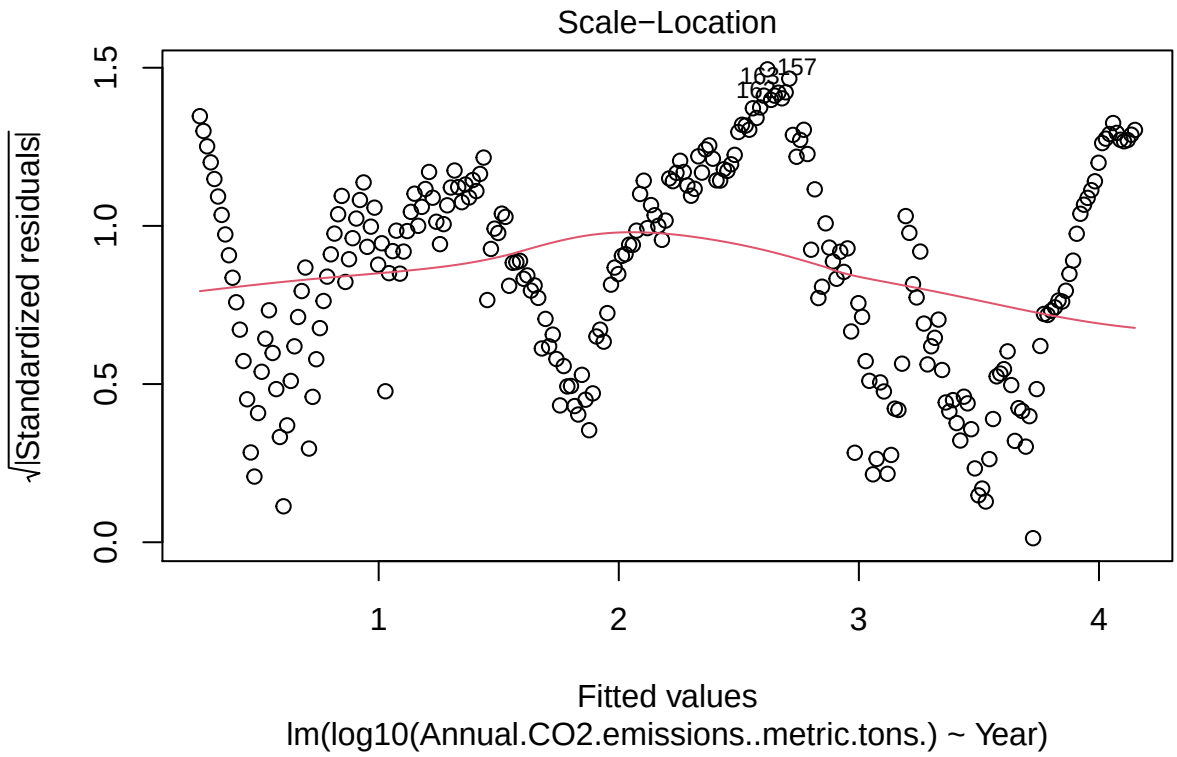
```
##
## Call:
## lm(formula = log10(Annual.CO2.emissions..metric.tons.) ~ Year,
##     data = tempByCo2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.21506 -0.09698 -0.02335  0.10182  0.27506
##
## Coefficients:
##                Estimate  Std. Error t value            Pr(>|t|)
## (Intercept) -26.2812570   0.1939634  -135.5 <0.0000000000000002 ***
## Year          0.0151550   0.0001031   147.0 <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

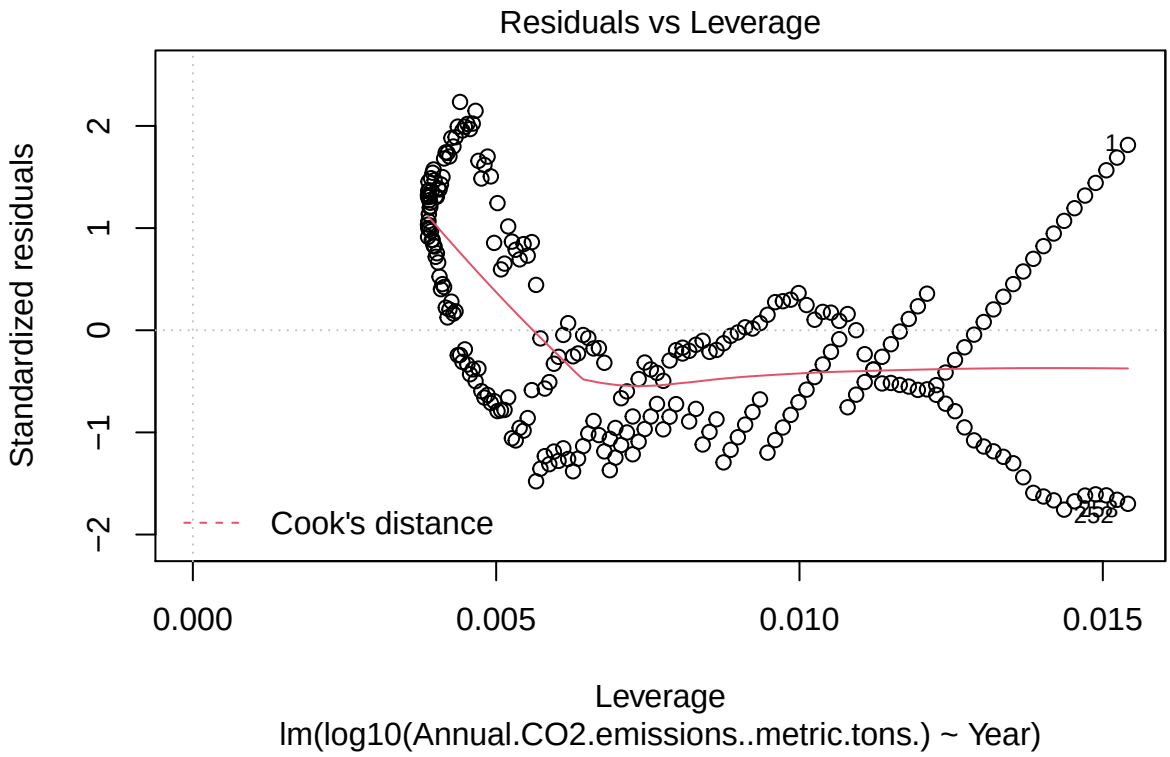
```
##
## Residual standard error: 0.1234 on 256 degrees of freedom
##   (7 observations deleted due to missingness)
## Multiple R-squared:  0.9883, Adjusted R-squared:  0.9882
## F-statistic: 2.16e+04 on 1 and 256 DF,  p-value: < 0.00000000000000022
```

```
plot(RegModel.2)
```

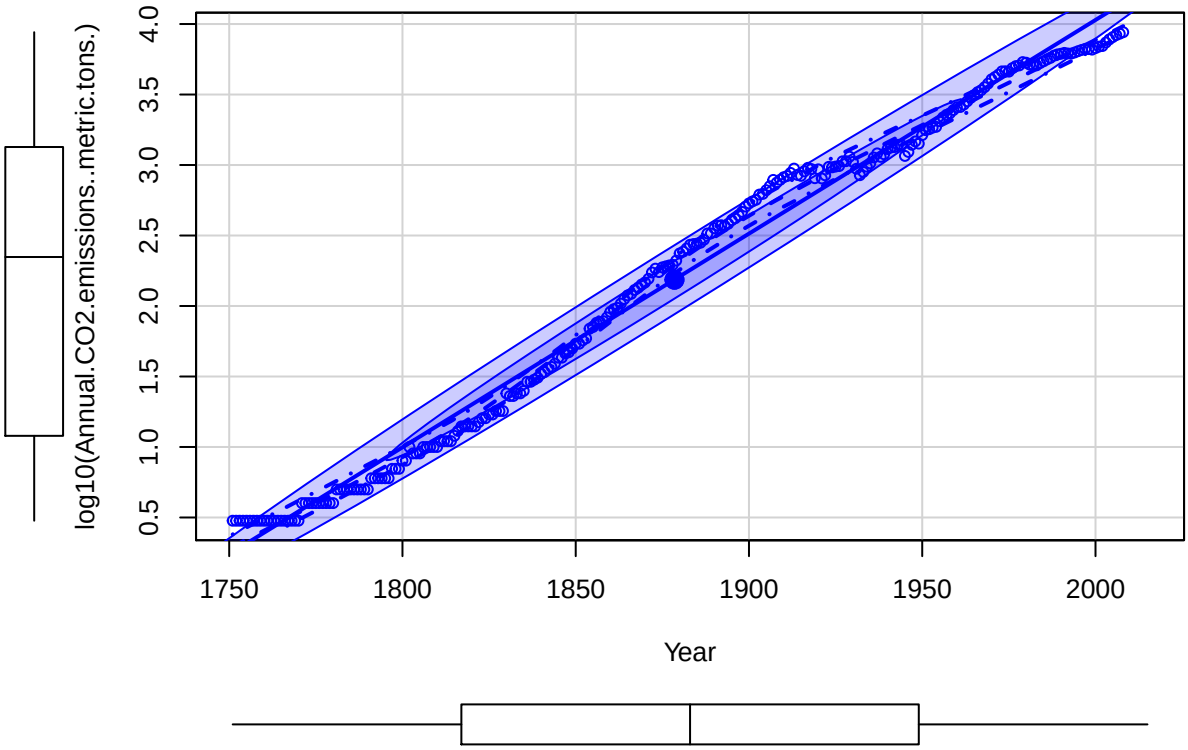



Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(log10(Annual.CO2.emissions..metric.tons.) ~ Year)
157
163
162
−3
−2
−1
0
1
2
3
−1
0
1
2


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
1
2
3
4
157
Fitted values
lm(log10(Annual.CO2.emissions..metric.tons.) ~ Year)




```
scatterplot(log10(Annual.CO2.emissions..metric.tons.)~Year, regLine=TRUE,
  smooth=list(span=0.5, spread=TRUE), boxplots='xy', ellipse=list(levels=c(.5,
    .9)), data=tempByCo2)
```



The log10-transformed model still has distinctive patterns in the Residual vs. Fitted and Scale-Location/Spread-Location plots. The Q-Q plot also clearly has points at the extremes that do not fall on the line of identity. However, the scatter plot looks much better, with the data aligning much better to a linear pattern.

Furthermore, the R^2 value for the log10-transformed model is much better than the untransformed model (0.9882 vs 0.6231).

Square-root transformation

Exploring how another transformation method may affect model performance.

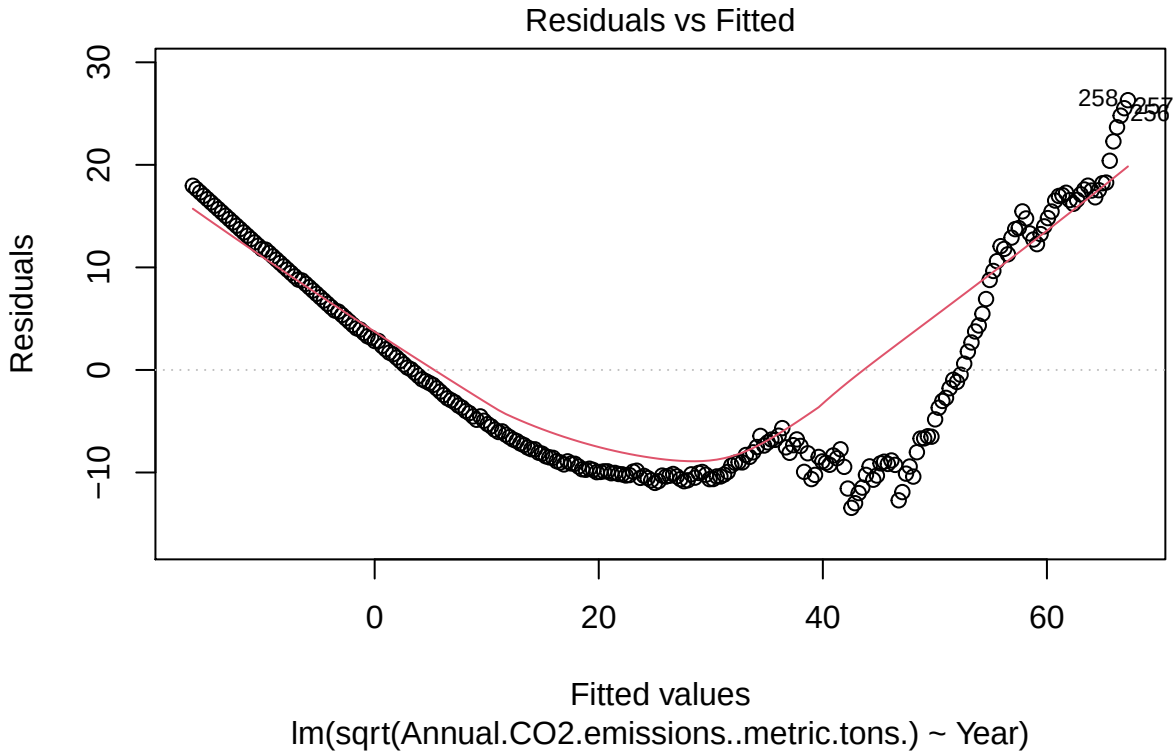
```
RegModel.3 <- lm(sqrt(Annual.CO2.emissions..metric.tons.)~Year, data=tempByCo2)
summary(RegModel.3)
```

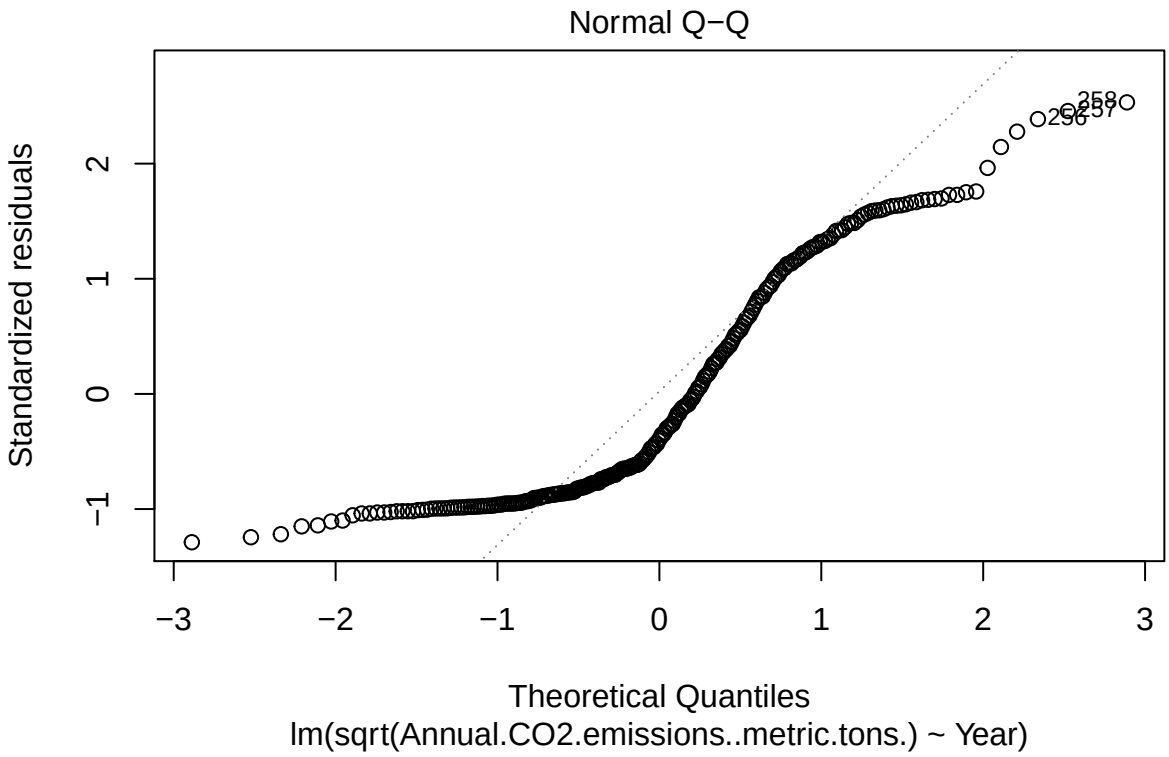
```
##
## Call:
## lm(formula = sqrt(Annual.CO2.emissions..metric.tons.) ~ Year,
##     data = tempByCo2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -13.447  -9.161  -4.107   9.605  26.306
##
## Coefficients:
##               Estimate  Std. Error t value           Pr(>|t|)
```

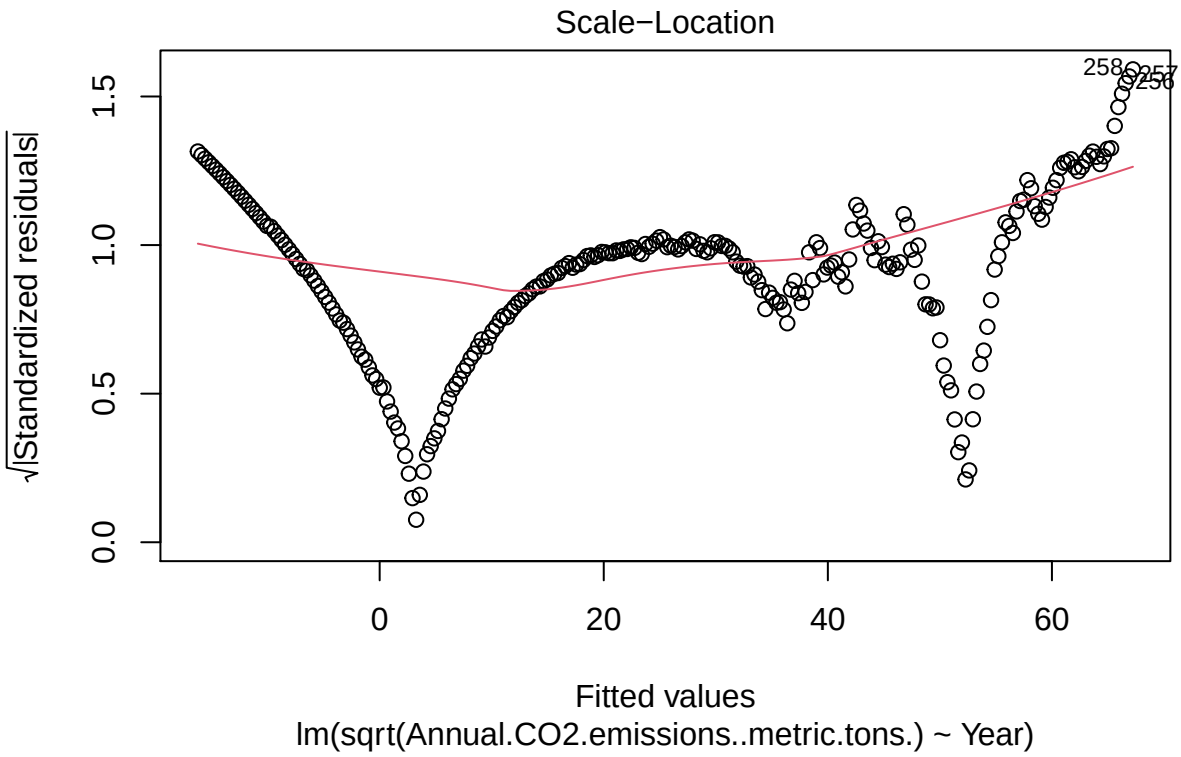
```
## (Intercept) -584.843727   16.460812  -35.53 <0.0000000000000002 ***
## Year           0.324738    0.008751   37.11 <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.47 on 256 degrees of freedom
##   (7 observations deleted due to missingness)
## Multiple R-squared:  0.8432, Adjusted R-squared:  0.8426
## F-statistic:  1377 on 1 and 256 DF,  p-value: < 0.00000000000000022
```

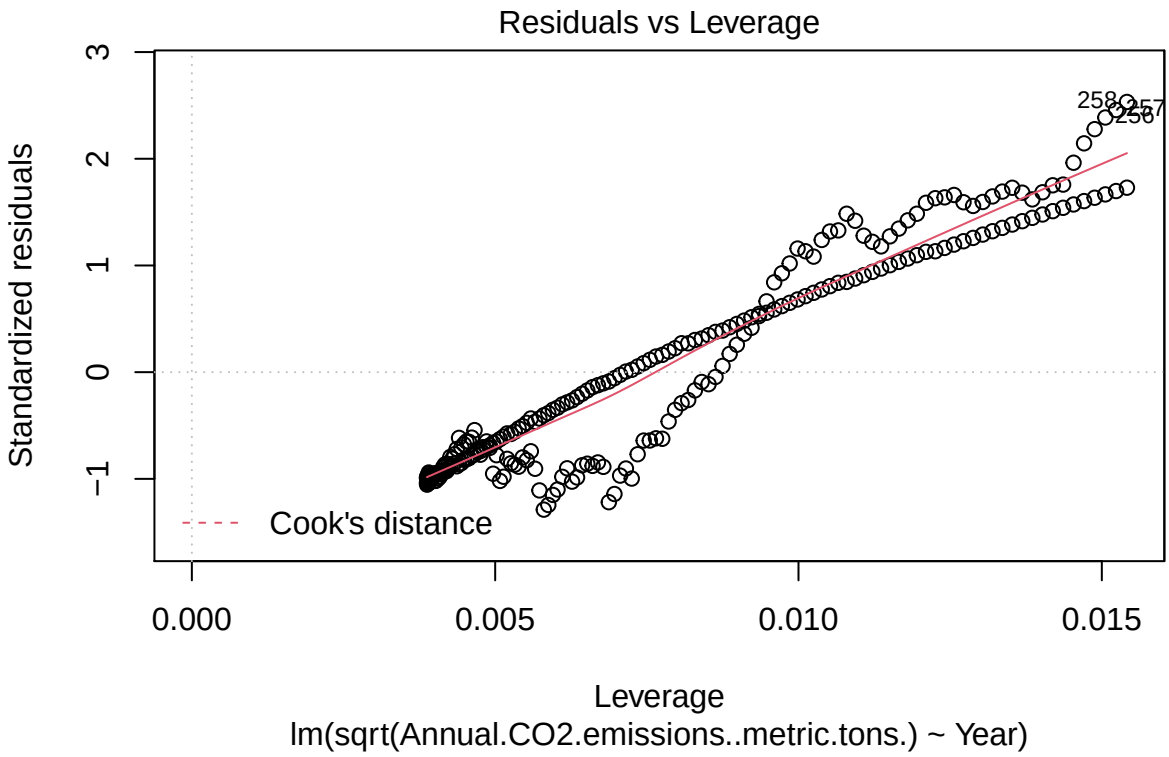
The R^2 is greater than the untransformed model, but not as good as the log10-transformed model. On to the diagnostic plots.

```
plot(RegModel.3)
```

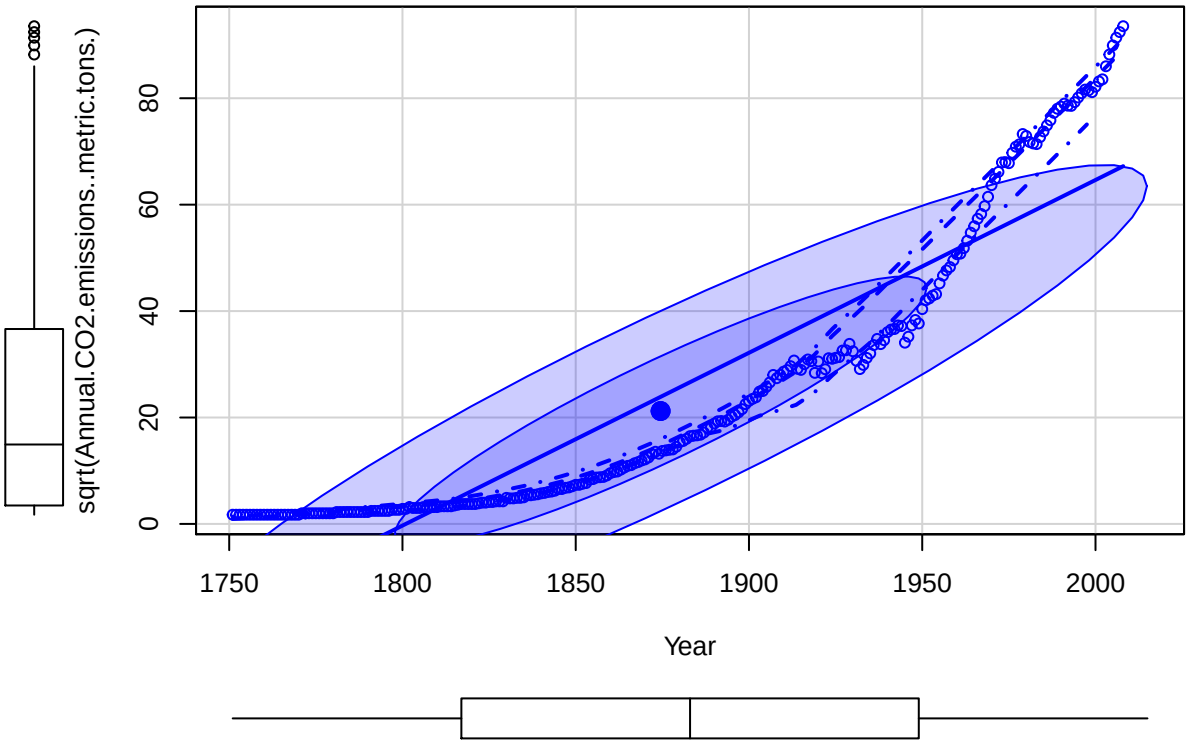



Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(sqrt(Annual.CO2.emissions..metric.tons.) ~ Year)
259
257
256
−3
−2
−1
0
1
2
3
−1
0
1
2


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
0
20
40
60
258
257
256
Fitted values
lm(sqrt(Annual.CO2.emissions..metric.tons.) ~ Year)




```
scatterplot(sqrt(Annual.CO2.emissions..metric.tons.)~Year, regLine=TRUE,
  smooth=list(span=0.5, spread=TRUE), boxplots='xy', ellipse=list(levels=c(.5,
    .9)), data=tempByCo2)
```



The scatter plot for the sqrt-transformed data looks better than the untransformed, but not as linear as the log10-transformed data. The other diagnostic plots appear similar; they do not provide much support for using this model.

Key learnings

- In `Rcmdr`, one must 'import' data for it to become part of the active datasets list; 'loading' the data is not sufficient since it is not saved as an object that can be re-accessed.
- Library `abind` is meant for helping combine multidimensional arrays
- Library `e1071` is a library with misc functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien
- The `pos` argument in the `library` function specifies the position on the search list at which to attach the loaded namespace.
- When you build a graph in `Rcmdr`, depending on the graph type and options, you may be able to use the mouse to click on individual points to add ID labels to them.
- The diagnostic plots for assessing the appropriateness of a linear model include: Residual vs Fitted, Q-Q, Scale-Location/Spread-Location, and Residuals vs. Leverage. In Residual vs Fitted and Scale-Location/Spread-Location, seeing no clear patterns in the data is ideal. In the Q-Q plot, seeing that the points align well with the upward-sloping line is ideal. In the Residuals vs. Leverage, having no points with a large Cook's distance is ideal.

Unresolved questions

- How does one enable the mouse-click labelling feature of `Rcmdr` figures outside of the context of using `Rcmdr`?

- At some point in my session, the code stopped being transcribed to the `R Markdown` tab in `Rcmdr`, although it was still transcribed to the `R Script` tab. What causes this error and how does one fix it?
- When would one favor Kendall's tau over Spearman's rank correlation?
- What are the formal methods for testing a dataset's normal distribution and homoscedasticity? Is it not Shapiro-Wilk for normality? What about the Breasch-Pagan test for heteroscedasticity?
- Are there *objective* ways to diagnose linear model appropriateness instead of 'eyeballing' graphs and making a judgment call?