Assignment 3: (Generalized) Linear Regression Models

Chalkiopoulos Georgios | *p*3352124

December 24, 2021

Exercise 1. *On November 23 2021, the European Commission made the twit shown on Figure 1:*

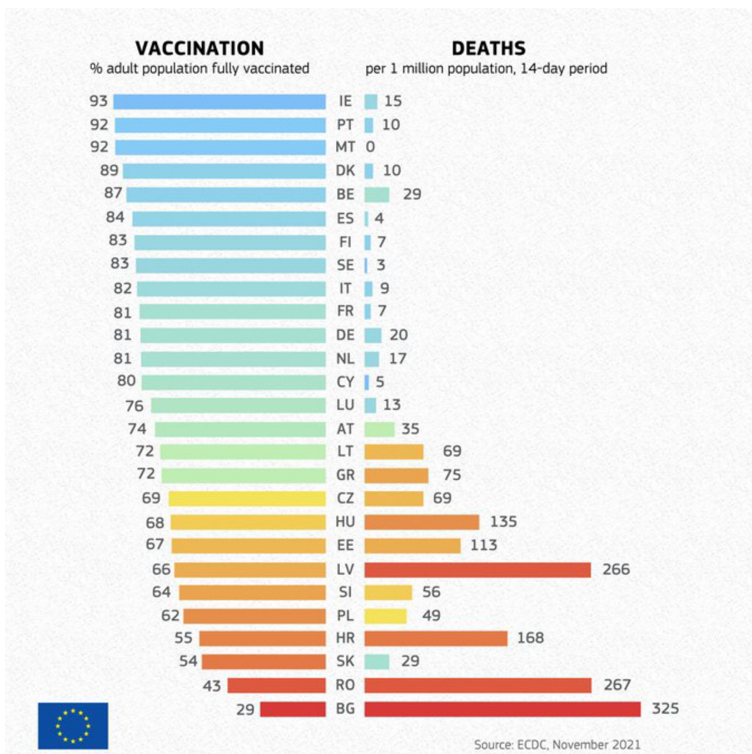

Figure 1: The twit of the European Commision on 23/11/2021. https://twitter.com/EU_Commission/status/1463119478099693571

We are interested to describe the number of deaths (per 1 M, 14 day period) due to Covid19 based on the percentage of adult population which are fully vaccinated, according to the data shown in Figure 1. Propose, estimate, describe and compare sensible statistical model(s) in order to analyze the dataset. Explain the model(s) to nonexperts. Discuss any limitations that may apply.

As the Exercise suggested, we would like to describe the number of deaths (per 1 million population in a 14-day period) based on the percentage of adult population which are fully vaccinated. Having the data available from the graph, we will use R to import the data, describe it and then run various models to see if the data could be explained by any of these.
The R code will be provided for each question along with explanations for each step of the process.

First we will import the data and report basic descriptive statistics. Since we have two variables, and we are interested in finding a relationship between them, we will also provide a basic scatter-plot.

```
> vac = c(93, 92, 92, 89, 87, 84, 83, 83, 82, 81, 81, 81,
+ 80, 76, 74, 72, 72, 69, 68, 67, 66, 64, 62, 55,
+ 54, 43, 29)
> deaths = c(15, 10, 0, 10, 29, 4, 7, 3, 9, 7, 20, 17, 5, 13,
+ 35, 69, 75, 69, 135, 113, 266, 56, 49, 168,
+ 29, 267, 325)
> # parse the data into a dataframe
> df <- data.frame(vac, deaths)
> colnames(df) <- c("vac", "deaths")
```

The datat has been imported. We may now run basic statistics.

```
> # Describe and summary
> summary(df)
      vac           deaths      
 Min.   :29.0   Min.   :  0.00  
 1st Qu.:66.5   1st Qu.:  9.50  
 Median :76.0   Median : 29.00  
 Mean   :73.3   Mean   : 66.85  
 3rd Qu.:83.0   3rd Qu.: 72.00  
 Max.   :93.0   Max.   :325.00  
> describe(df)
       vars  n  mean    sd median trimmed   mad min
vac       1 27 73.30 15.29     76   74.87 11.86  29
deaths    2 27 66.85 90.15     29   52.61 35.58   0

       max range  skew kurtosis    se
vac     93    64 -1.02     0.74  2.94
deaths 325   325  1.61     1.43 17.35
```

We also run basic plots (hist and boxplot) along with a scatterplot with a lm line.

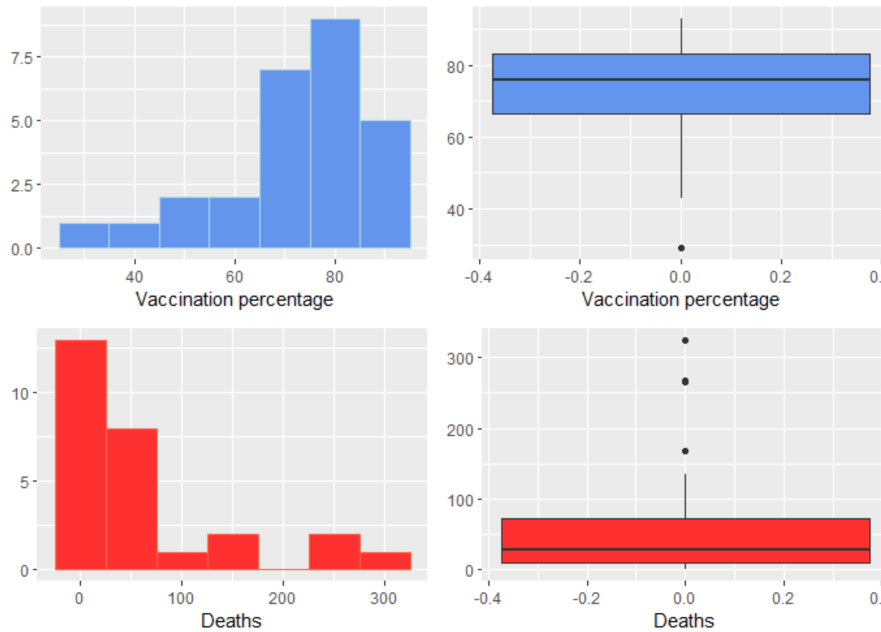

Figure 2: Histogram and Boxplot for Vaccination and Deaths.

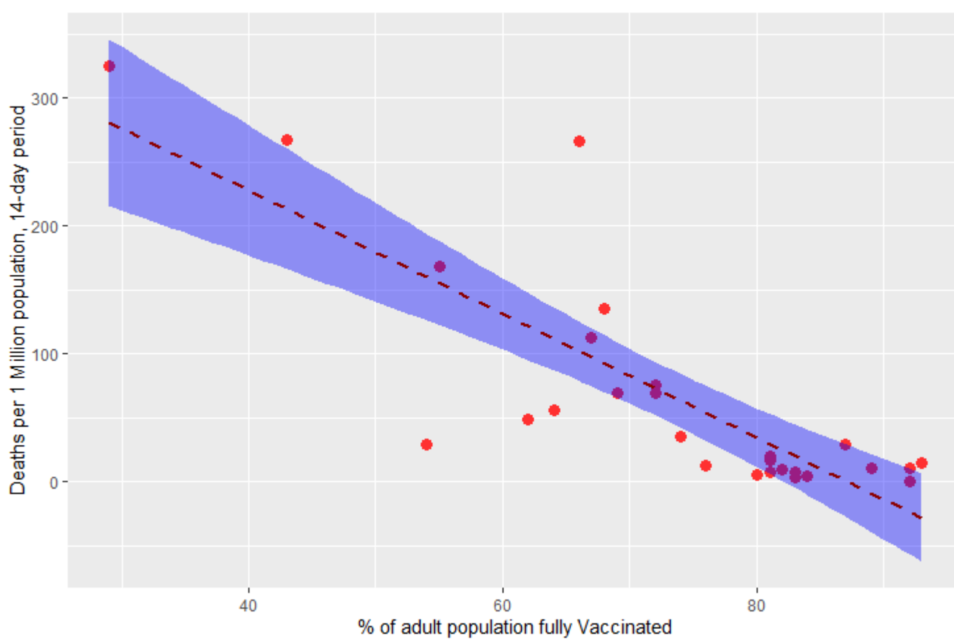

Figure 3: Scatter plot of Vaccinations Deaths.

At first glance we observe that there is some kind of linear relation between the two variables, but we also see that two values might be affecting the results,being outliers. The outliers can also be seen in Figure 2. We will run a detailed Linear Regression model to get accurate values as well as examine the significance of each variable. For each model, the

- Simple Linear model (Including outliers)

```
> df_fit <- run_reg(df)

Call:
lm(formula = deaths ~ vac, data = f_df)

Residuals:
    Min      1Q  Median      3Q     Max 
-131.09  -20.62   -9.63   25.94  163.89 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 420.9999    50.6021   8.320 1.14e-08 ***
vac          -4.8317     0.6764  -7.144 1.74e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 52.72 on 25 degrees of freedom
Multiple R-squared:  0.6712,Adjusted R-squared:  0.658 
F-statistic: 51.03 on 1 and 25 DF,  p-value: 1.735e-07
```

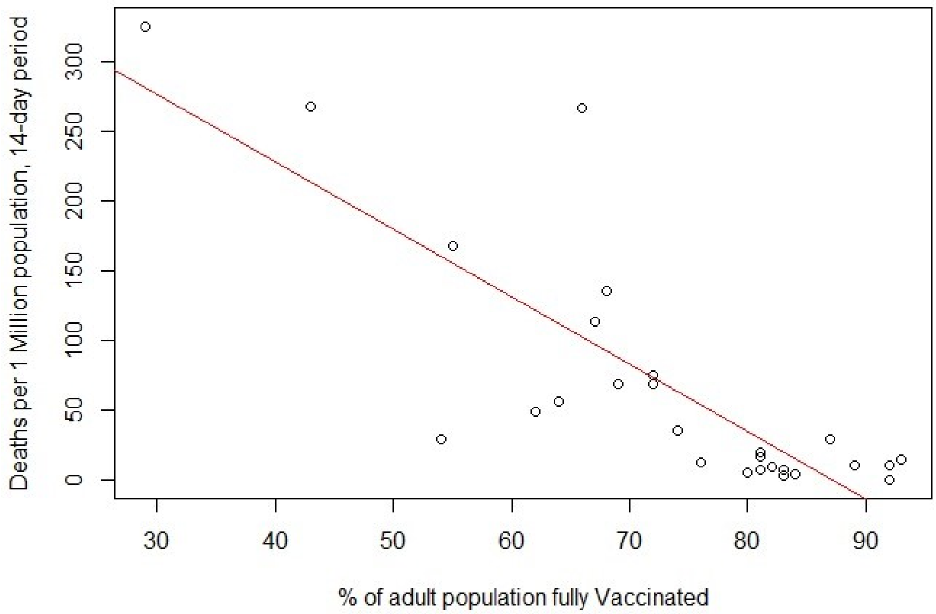

Figure 4: Deaths vs Vaccinations along with the OLS line.

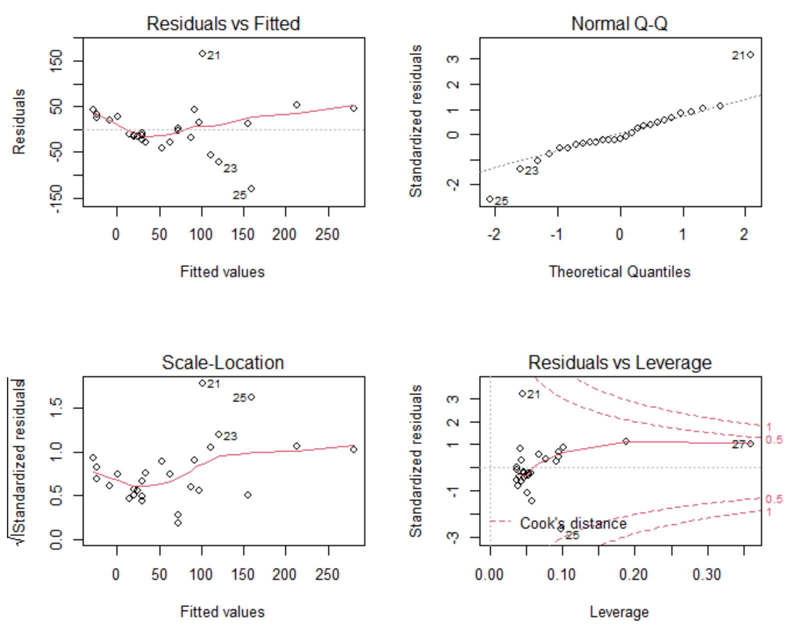

Figure 5: Diagnostic plots for the linear model.

- Linear Model with Quadratic term

```
> df_fit2 <- run_quard_reg(df)

Call:
lm(formula = deaths ~ vac + vac2, data = f_df_q)

Residuals:
     Min       1Q   Median       3Q      Max
-122.973  -18.142   -5.749   16.574  178.101

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 614.80398  136.65847   4.499 0.000149 ***
vac         -11.21487    4.24905  -2.639 0.014363 *
vac2          0.04896    0.03220   1.521 0.141410
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 51.39 on 24 degrees of freedom
Multiple R-squared:  0.7001,Adjusted R-squared:  0.6751
F-statistic: 28.01 on 2 and 24 DF,  p-value: 5.295e-07

Analysis of Variance Table

Response: deaths
          Df Sum Sq Mean Sq F value    Pr(>F)
vac        1 141839  141839 53.7122 1.435e-07 ***
vac2       1   6106    6106  2.3124    0.1414
Residuals 24  63377    2641
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

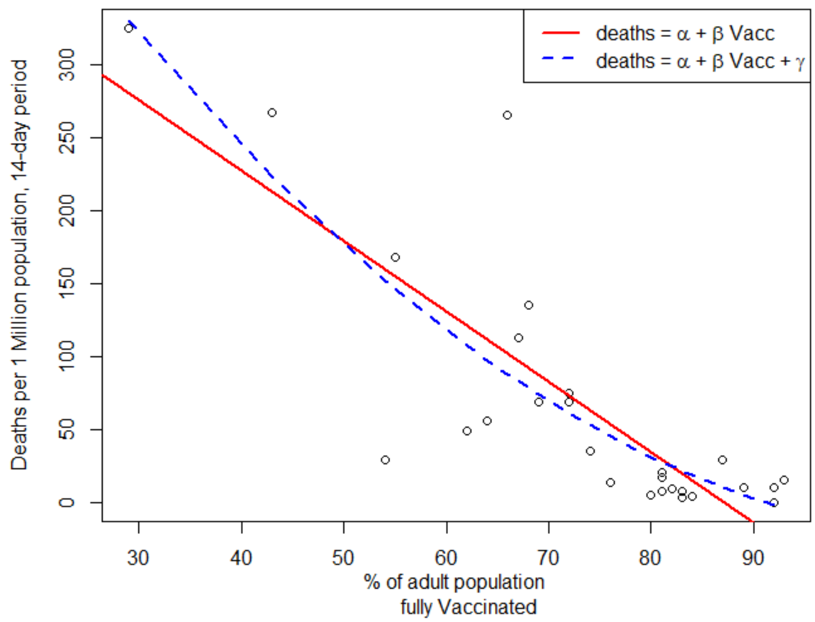

Figure 6: Deaths vs Vaccinations along with the OLS linear and quadratic fitted lines.

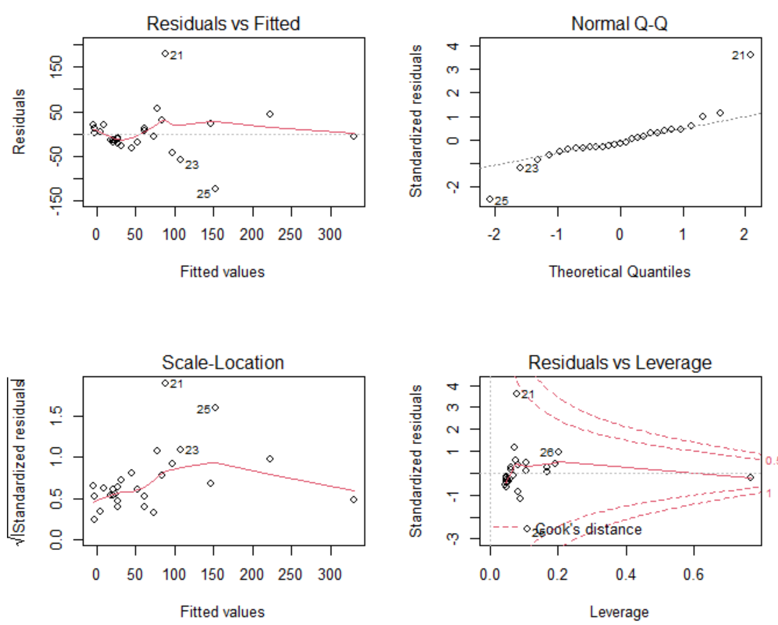

Figure 7: Diagnostic plots for the quadratic model.

- Log transformed model (for # of deaths)

```
> df_log_fit <- run_reg(df_log)

Call:
lm(formula = deaths ~ vac, data = f_df)

Residuals:
    Min      1Q  Median      3Q     Max
-1.9441 -0.6135  0.1226  0.6945  1.6963

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  8.83271    0.87308  10.117 2.54e-10 ***
vac         -0.07488    0.01167  -6.416 1.02e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9096 on 25 degrees of freedom
Multiple R-squared:  0.6222,Adjusted R-squared:  0.6071
F-statistic: 41.17 on 1 and 25 DF,  p-value: 1.02e-06
```

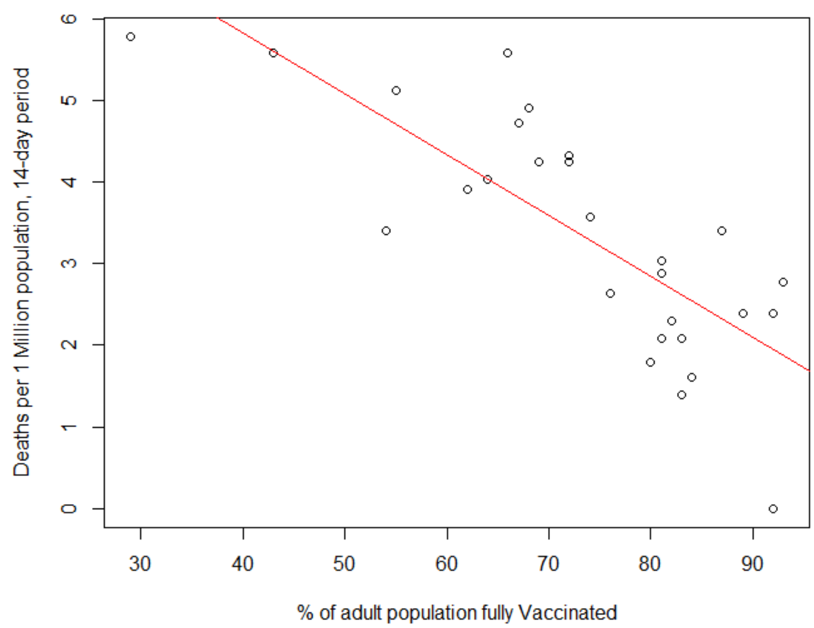

Figure 8: log(Deaths) vs Vaccinations along with the OLS.

By observing the original histogram inf Figure 2, we may observe that the number

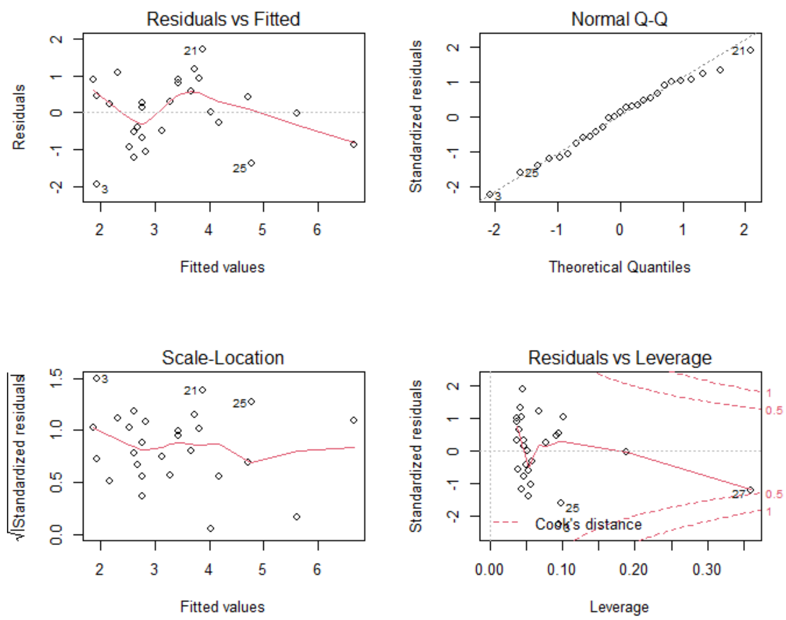

Figure 9: Diagnostic plots with log transformed data.

of deaths is right skewed. This is the first indication that the model with the log transformation (for the number of deaths) would perform well. The initial Linear model has p-values that would indicate a relation between the two variables (for the given significance levels) but the diagnostic plots make us reject this model. Even after removing outliers (provided in the `*.R` file) the results did not improve. Next we tried to introduce a quadratic term to the model, but the p-value of that indicated that the term was not significant. Moreover the diagnostic plots indicate that the assumptions made earlier (in order to use linear regression) are not confirmed. Finally, a log-transformation was test, in the number of deaths. This, of course, comes with a transformation in the original number of deaths, in which a constant (of one) was added in the original number of deaths. We will make the assumption that this transformation will not heavily impact the results (which could be tested individually as well). This model fetched the best results and the diagnostic plots look somewhat acceptable. The model could further be improved by removing outliers, but we will not remove unnecessary data if we don't have to.
Model interpretation: Having log-transformed the dependent variable we have the following model:

$$log(y) = \beta_0 + \beta_1 x \Rightarrow y = exp(\beta_0 + \beta_1 x) \Rightarrow y = exp(\beta_0)exp(\beta_1 x)$$

By using the coefficients we found the model becomes:

$$y = exp(8.833)exp(-0.075x)$$

The meaning of the model can be simplified in that a 1% increase in the vaccination percentage decreases the number of deaths by 0.075/100. For example, by increasing the vaccination percentage by 10% the total number of deaths will decrease by $0.075\dot{l}og(1.10) = 0.041$, which is around 4.1%.

Exercise 2. This exercise refers to the example in page 61 of the "Multiple Linear Regression" set of slides. Use the logarithm with base 2 of the body weight and the categorical variable D as explanatory variables in order to describe the outcome variable TS.

1. *Give a concise description of the estimated model.*

In the paper published in 1976, Allison and Cicchetti presented data on sleep patterns of 62 mammal species along with several other possible predictors of sleep. This study was used in the book "Applied Linear Regression" by S. Weisberg, where different methods were explored in order to investigate whether there were any variables that could explain the hours of sleep of mammals. Detailed explanation of how the final data-set was created may be found in the book.

We are interested in the following variables:

- **TS**: Total sleep, hrs/day
- **BodyWt**: Body weight in kg (transformed to log2(BodyWt))
- **D**: Danger index, 1 = least danger,. . . , 5 = most danger

More specifically we are interested in predicting the total sleep (TS) based on the predictors Body Weight (BodyWt) and Danger index (D) using Multiple Linear Regression. Before doing so, we need to transform our predictors. Starting with the Dangers, which is a categorical value, we will use Dummy Variables to transform it to a discrete one. The first category (D1) will be dropped and will be used as "control".
Furthermore, regarding the body weight we will be using the logarithm with base two. The reason behind this is explained in detail in section 7.1. of the mentioned book, but a brief explanation is that we may use transformations when the initial values don't have a linear relation. In this case a log-transformation can be used in which the relation will be linear. A comparison between the original Weight and the log-transformed Weight may be found in Figure 10. Finally the usual assumptions needed for the linear regression model will be taken as granted. That is:

- The underline relationship is linear.
- The random errors ε are independent of the predictors $X_1, X_2, \ldots, X_k$.
- For the random errors we have: $\varepsilon \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$.

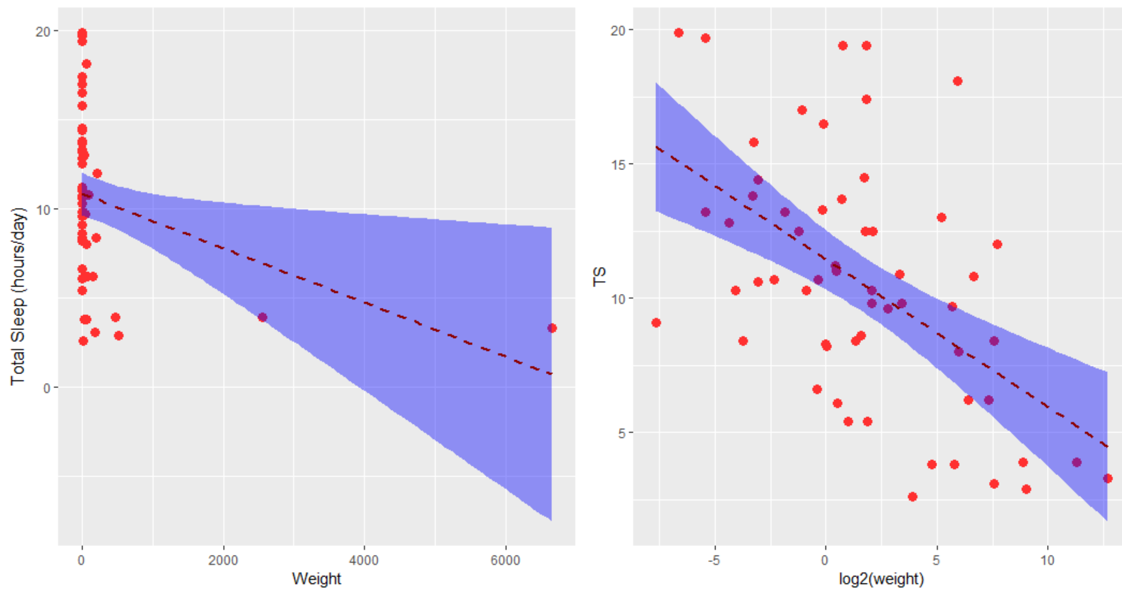

Figure 10: Weight and Log transformed Weight ∼ Hours of sleep

2. *What is the effect of an animal belonging to category 5 of the danger index when compared to animals in danger level equal to 1? Is this effect significant?*

In order to investigate the differences between the various Danger groups (more specifically Danger group 1 and 5) we will run a Multiple Linear regression model to get more insight.

```
> summary(fit2)
Call:
lm(formula = TS ~ logb(BodyWt, 2) + D, data = sleep1,
    na.action = na.omit)

Residuals:
    Min      1Q  Median      3Q     Max
-7.6165 -1.8447 -0.0214  1.9043  6.7414

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      13.9325     0.8173  17.046  < 2e-16 ***
logb(BodyWt, 2)  -0.4357     0.1113  -3.914 0.000266 ***
D2               -2.4287     1.2238  -1.985 0.052479 .
D3               -3.5836     1.3347  -2.685 0.009714 **
D4               -3.8535     1.3691  -2.815 0.006879 **
D5               -7.2945     1.5525  -4.699 1.96e-05 ***
```

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.343 on 52 degrees of freedom
  (4 observations deleted due to missingness)
Multiple R-squared:  0.5195,Adjusted R-squared:  0.4733
F-statistic: 11.25 on 5 and 52 DF,  p-value: 2.23e-07
```

We are interested in the effect of an animal belonging to category 5, of the danger index, compared to the ones in the first. This is represented in the `R` code results `(Intercept)` and `D5`.
First of all we observe that D5 has a coefficient estimate of -7.3 when compared to the Intercept (D1). That means that an animal, with the same body weight, that belongs to D5 (higher danger) will get around 7.3 less hours of sleep. Moreover, the effect is quite significant, for typical significance levels and then some, since the p-value is very close to zero (1.96e-05).
Looking at Figure 11 (the figure of question 3) we may better understand the above conclusion, since the regression lines for groups 1 and 5 are parallel, and the distance between the, corresponds to the coefficient found.

3. *Visualize the estimated model by superimposing the (5) estimated regression lines on top of the Figure in page 62.*

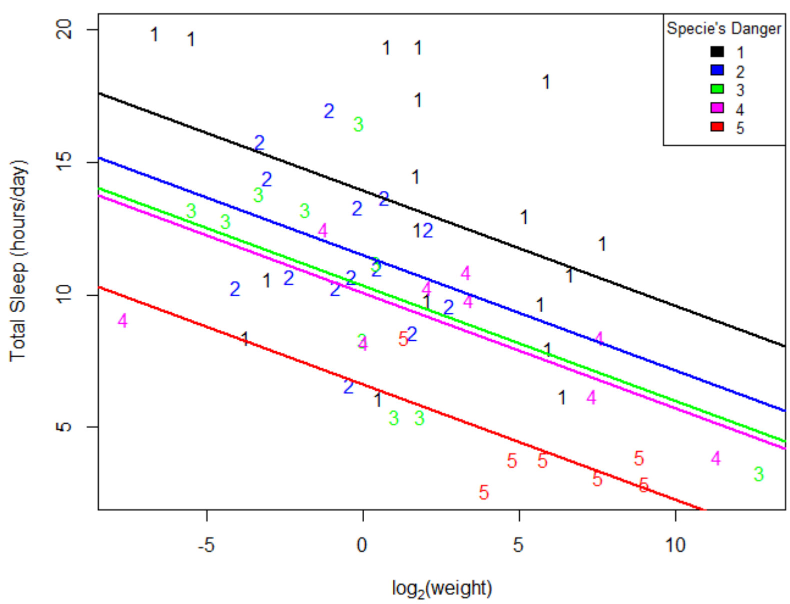

Figure 11: Estimated model with regression lines per Danger level.

4. *Explain the figure of the previous question into a nonexpert in Statistics.*

When examining the total hours of sleep of animals, two variables were taken into consideration. The weight, as well as the spice's danger. After testing the various relation between the variables, we concluded that the model that better explains this relation is the one visualized in Figure 11. This model implies that for an animal with the same weight, the hours of sleep are indeed affected by the danger level. More specifically, the less dangerous the category, the more hours of sleep the animal gets, and vice versa. For example, an animal that belong the the Danger Group 5 will sleep, on average, 7.3 hours less, given that the weight is the same. Finally, the heavier the animal, the less the hours of sleep, which is indicated by the negative slope in the graph.

Exercise 3. An automobile magazine is interested in identifying the factors influencing fuel consumption.

1. *Load the data into* R *using*

```
> require(starts)
> data(mtcars)
```

Using the library

```
> data(mtcars)
> head(mtcars)
                   mpg cyl disp  hp drat    wt  qsec vs am gear carb
Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
Valiant           18.1   6  225 105 2.76 3.460 20.22  1  0    3    1
```

2. *Define* ***am****,* ***vs*** *and* ***cyl*** *as factor variables. The rest of them should be treated as numeric variables.*

```
> cols <- c("am", "vs", "cyl")
> mtcars[cols] <- lapply(mtcars[cols], factor)
```

3. *Report basic descriptive statistics for each variable and illustrate them on suitable diagrams.*

We will run the summary for all variables, including the categorical variables. Then we will produce density plots for numerical variables and barplots for the three categorical variables. Moreover, a pairwise plot including all pairs will be provided just to get a high level overview.

```
> describe(mtcars)[c(1:9)]
     vars  n   mean     sd median trimmed    mad   min    max
mpg     1 32  20.09   6.03  19.20   19.70   5.41 10.40  33.90
cyl*    2 32   2.09   0.89   2.00    2.12   1.48  1.00   3.00
disp    3 32 230.72 123.94 196.30  222.52 140.48 71.10 472.00
hp      4 32 146.69  68.56 123.00  141.19  77.10 52.00 335.00
```

```
drat     5 32   3.60   0.53   3.70    3.58   0.70  2.76   4.93
wt       6 32   3.22   0.98   3.33    3.15   0.77  1.51   5.42
qsec     7 32  17.85   1.79  17.71   17.83   1.42 14.50  22.90
vs*      8 32   1.44   0.50   1.00    1.42   0.00  1.00   2.00
am*      9 32   1.41   0.50   1.00    1.38   0.00  1.00   2.00
gear    10 32   3.69   0.74   4.00    3.62   1.48  3.00   5.00
carb    11 32   2.81   1.62   2.00    2.65   1.48  1.00   8.00
>
> describe(mtcars)[c(10:13)]
      range  skew kurtosis    se
mpg   23.50  0.61    -0.37  1.07
cyl*   2.00 -0.17    -1.76  0.16
disp 400.90  0.38    -1.21 21.91
hp   283.00  0.73    -0.14 12.12
drat   2.17  0.27    -0.71  0.09
wt     3.91  0.42    -0.02  0.17
qsec   8.40  0.37     0.34  0.32
vs*    1.00  0.24    -2.00  0.09
am*    1.00  0.36    -1.92  0.09
gear   2.00  0.53    -1.07  0.13
carb   7.00  1.05     1.26  0.29
```

Density plots for Numerical variables:

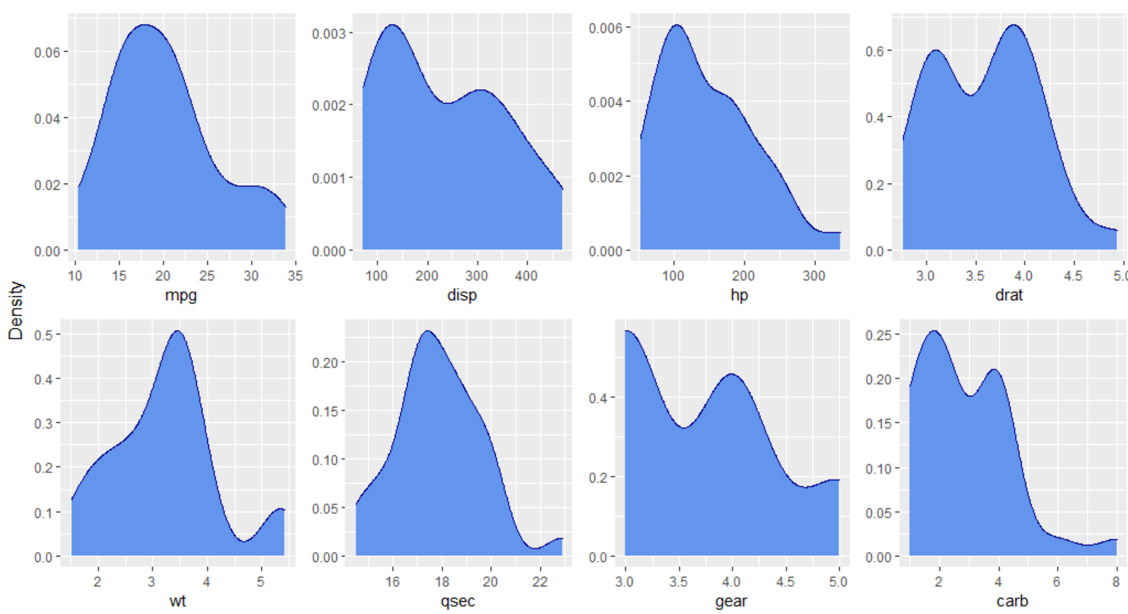

Figure 12: Density plots for Numerical Variables.

Bar plots for categorical variables:

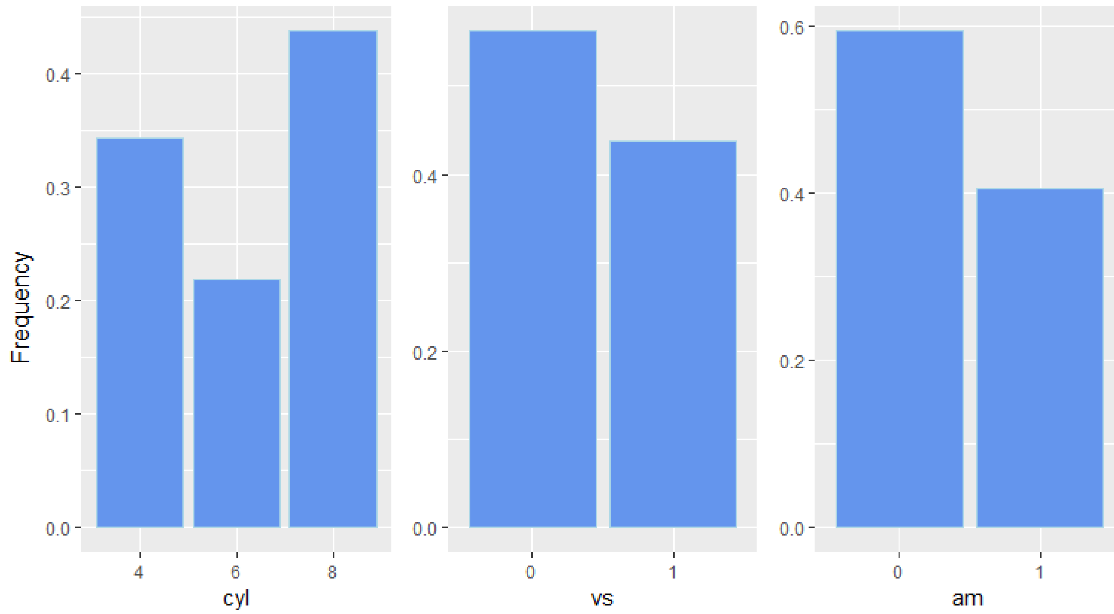

Figure 13: Barplots for categorical variables

4. *Produce all pairwise scatter plots for the numeric variables and compute the corresponding correlation coefficients.*

The requested Fugire may be found in page 18 (Figure 14).

5. *Is there any difference in consumption between automatic and manual cars?*

We will answer the question using Linear Regression, although we could answer that by performing other tests.
First we will provide a series of Figures, to visualize the data, and then we will run the regression model and investigate the p-values.

```
> summary(fit_consumption)

Call:
lm(formula = mpg ~ am, data = mtcars)

Residuals:
    Min      1Q  Median      3Q     Max
-9.3923 -3.0923 -0.2974  3.2439  9.5077
```

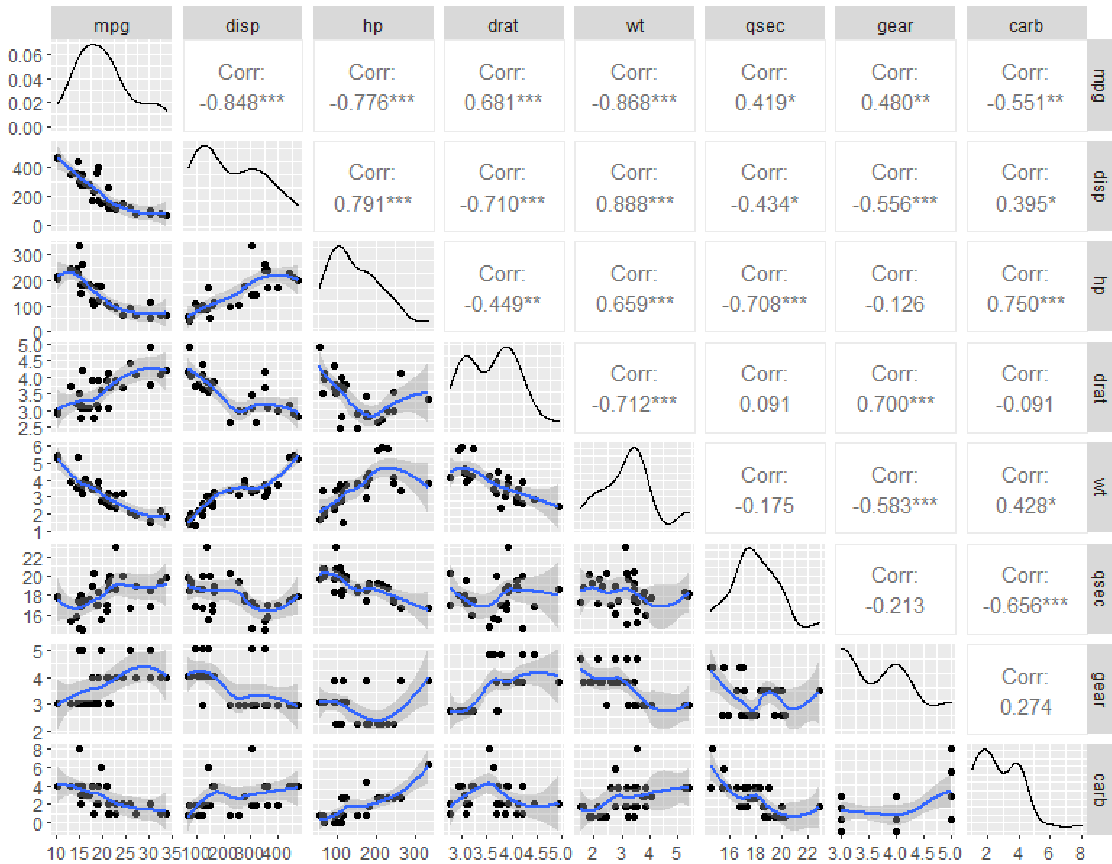

Figure 14: Pairplots for all variables.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   17.147      1.125  15.247 1.13e-15 ***
am1            7.245      1.764   4.106 0.000285 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.902 on 30 degrees of freedom
Multiple R-squared:  0.3598,Adjusted R-squared:  0.3385
F-statistic: 16.86 on 1 and 30 DF,  p-value: 0.000285
```

The p-value indicates that the transmission is a significant value if the mpg, thus there is a difference in the value of different groups.

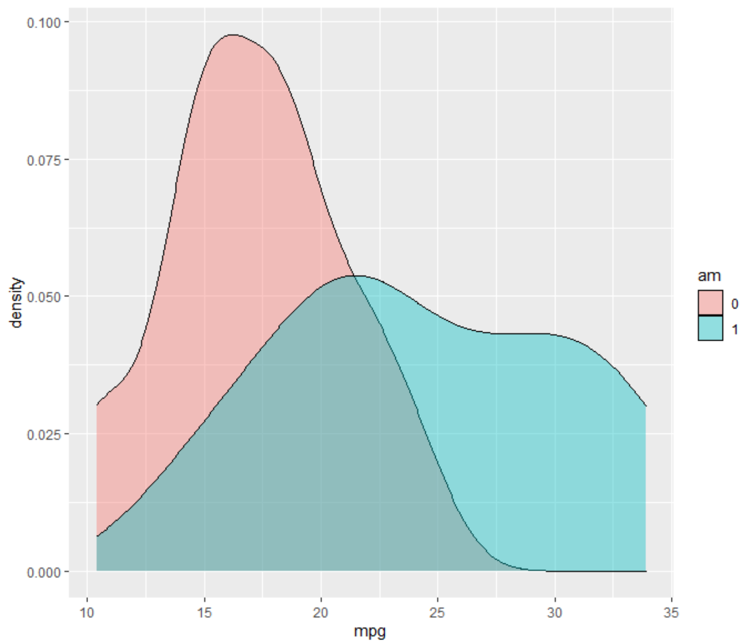

Figure 15: mpg per am Density plot

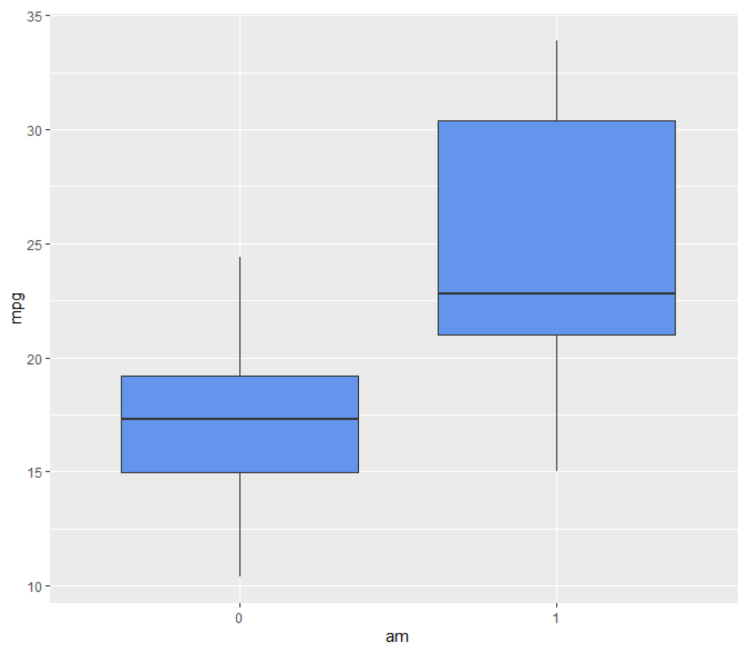

Figure 16: mpg per am box plot

The plots give a visual representation of the means.

6. *Is there any difference in consumption among cars with different number of cylinders?*

The results are similar for the number of cylinders as well.

```
> summary(fit_consumption)

Call:
lm(formula = mpg ~ cyl, data = mtcars)

Residuals:
    Min      1Q  Median      3Q     Max
-5.2636 -1.8357  0.0286  1.3893  7.2364

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  26.6636     0.9718  27.437  < 2e-16 ***
cyl6         -6.9208     1.5583  -4.441 0.000119 ***
cyl8        -11.5636     1.2986  -8.905 8.57e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.223 on 29 degrees of freedom
Multiple R-squared:  0.7325,Adjusted R-squared:  0.714
F-statistic:  39.7 on 2 and 29 DF,  p-value: 4.979e-09
```

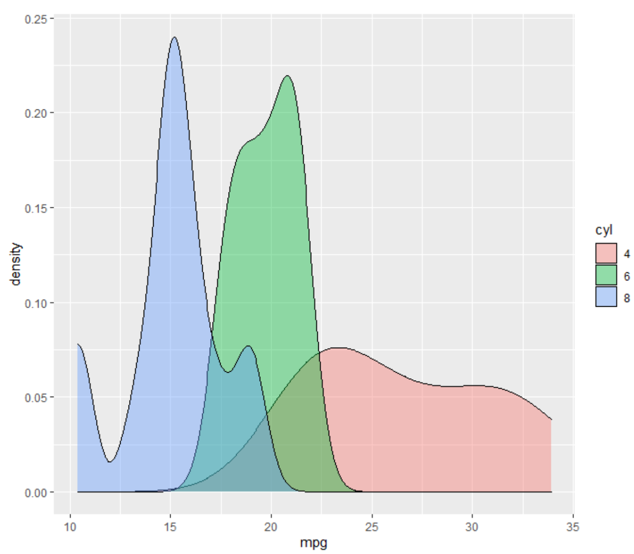

Figure 17: mpg per cyl Density plot

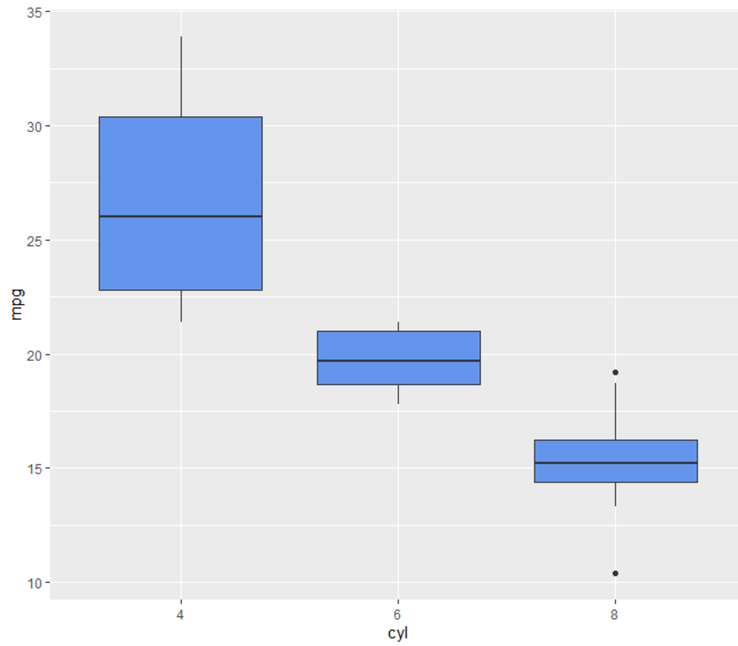

Figure 18: mpg per cyl box plot

7. *Fit and interpret the full regression model using all available explanatory variables. According to the Bayesian Information Criterion, which variables mostly effect consumption? Check all modelling assumptions and interpret the selected model.*

```
> n <- dim(mtcars)[1]
> stepBE<-step(fitall_log, scope=list(lower = ~ 1,
+ upper= ~ log2(mpg) + log2(disp) + log2(hp) + log2(drat)+
+ +log2(wt) + log2(qsec) + cyl + vs + am + gear + carb),
+ direction="backward", data=mtcars,  k = log(n))
Start:  AIC=-86.99
log2(mpg) ~ log2(disp) + log2(hp) + log2(drat) + +log2(wt) +
    log2(qsec) + cyl + vs + am + gear + carb

             Df Sum of Sq     RSS     AIC
- cyl         2  0.019572 0.59509 -92.855
- log2(drat)  1  0.000474 0.57600 -90.433
- vs          1  0.002679 0.57820 -90.311
- am          1  0.003370 0.57889 -90.273
- log2(qsec)  1  0.007634 0.58316 -90.038
- log2(disp)  1  0.016660 0.59218 -89.547
- log2(hp)    1  0.023269 0.59879 -89.191
```

```
- carb          1   0.033708 0.60923 -88.638
- gear          1   0.063531 0.63905 -87.109
<none>                       0.57552 -86.994
- log2(wt)      1   0.068853 0.64438 -86.844

Step:  AIC=-92.86
log2(mpg) ~ log2(disp) + log2(hp) + log2(drat) + log2(wt) +
        log2(qsec) + vs + am + gear + carb

             Df Sum of Sq     RSS     AIC
- log2(drat)  1  0.000264 0.59536 -96.307
- log2(qsec)  1  0.002622 0.59772 -96.180
- am          1  0.004231 0.59933 -96.094
- vs          1  0.005002 0.60010 -96.053
- log2(disp)  1  0.007880 0.60297 -95.900
- carb        1  0.021542 0.61664 -95.183
- log2(hp)    1  0.030802 0.62590 -94.706
- gear        1  0.048725 0.64382 -93.803
<none>                    0.59509 -92.855
- log2(wt)    1  0.074239 0.66933 -92.559

Step:  AIC=-96.31
log2(mpg) ~ log2(disp) + log2(hp) + log2(wt) +
        log2(qsec) + vs + am + gear + carb

             Df Sum of Sq     RSS     AIC
- log2(qsec)  1  0.002992 0.59835 -99.612
- am          1  0.004353 0.59971 -99.539
- vs          1  0.005263 0.60062 -99.491
- log2(disp)  1  0.007619 0.60298 -99.366
- carb        1  0.022074 0.61743 -98.608
- log2(hp)    1  0.030747 0.62611 -98.161
- gear        1  0.049917 0.64528 -97.196
<none>                    0.59536 -96.307
- log2(wt)    1  0.075127 0.67048 -95.970

Step:  AIC=-99.61
log2(mpg) ~ log2(disp) + log2(hp) + log2(wt) + vs + am + gear +
    carb

             Df Sum of Sq     RSS      AIC
- vs          1  0.002878 0.60123 -102.924
```

```
- am          1  0.007627 0.60598 -102.673
- log2(disp)  1  0.016078 0.61443 -102.229
- carb        1  0.035308 0.63366 -101.243
- log2(hp)    1  0.040433 0.63878 -100.986
- gear        1  0.049800 0.64815 -100.520
<none>                    0.59835  -99.612
- log2(wt)    1  0.099353 0.69770  -98.162

Step:  AIC=-102.92
log2(mpg) ~ log2(disp) + log2(hp) + log2(wt) + am + gear + carb

             Df Sum of Sq     RSS     AIC
- am          1  0.005108 0.60634 -106.12
- log2(disp)  1  0.013270 0.61450 -105.69
- carb        1  0.032447 0.63368 -104.71
- log2(hp)    1  0.041329 0.64256 -104.26
- gear        1  0.047013 0.64824 -103.98
<none>                    0.60123 -102.92
- log2(wt)    1  0.112822 0.71405 -100.89

Step:  AIC=-106.12
log2(mpg) ~ log2(disp) + log2(hp) + log2(wt) + gear + carb

             Df Sum of Sq     RSS     AIC
- log2(disp)  1  0.012190 0.61853 -108.95
- carb        1  0.036074 0.64241 -107.74
- gear        1  0.042245 0.64858 -107.43
- log2(hp)    1  0.045119 0.65146 -107.29
<none>                    0.60634 -106.12
- log2(wt)    1  0.110717 0.71705 -104.22

Step:  AIC=-108.95
log2(mpg) ~ log2(hp) + log2(wt) + gear + carb

           Df Sum of Sq     RSS     AIC
- carb      1  0.026209 0.64474 -111.09
- gear      1  0.051572 0.67010 -109.85
<none>                  0.61853 -108.95
- log2(hp)  1  0.230270 0.84880 -102.29
- log2(wt)  1  0.287558 0.90608 -100.20

Step:  AIC=-111.09
```

```
log2(mpg) ~ log2(hp) + log2(wt) + gear

           Df Sum of Sq     RSS      AIC
- gear      1   0.02548 0.67022 -113.311
<none>                  0.64474 -111.086
- log2(wt)  1   0.42510 1.06984  -98.346
- log2(hp)  1   0.45675 1.10148  -97.413

Step:  AIC=-113.31
log2(mpg) ~ log2(hp) + log2(wt)

           Df Sum of Sq     RSS      AIC
<none>                  0.67022 -113.311
- log2(hp)  1   0.44169 1.11191 -100.578
- log2(wt)  1   0.95617 1.62639  -88.409
> stepBE$anova
          Step Df     Deviance Resid. Df Resid. Dev        AIC
1              NA           NA        20  0.5755229  -86.99396
2        - cyl  2 0.0195717515        22  0.5950946  -92.85530
3 - log2(drat)  1 0.0002635495        23  0.5953582  -96.30687
4 - log2(qsec)  1 0.0029920073        24  0.5983502  -99.61219
5         - vs  1 0.0028782994        25  0.6012285 -102.92437
6         - am  1 0.0051079679        26  0.6063365 -106.11938
7 - log2(disp)  1 0.0121895382        27  0.6185260 -108.94818
8       - carb  1 0.0262093696        28  0.6447354 -111.08590
9       - gear  1 0.0254806509        29  0.6702160 -113.31131
```

Looking at the results, we finally have a model that can be explained using only two variables. The weight and horsepower of the car. This is reasonable and could be expected.

Exercise 4. *Consider the crosssection data on the Home Mortgage Disclosure Act (HMDA), which is available on the AER package in R*

```
> library("AER")
> data(HMDA)
```

and use the `?HMDA` *command to retrieve useful information about the variables in the dataset. The aim is to describe whether the mortgage is denied or not based on the remaining variables in the dataset. Propose a statistical model and explain your findings. Use AIC and BIC to select the relevant explanatory variables. According to the model selected by AIC, how the odds of mortgage denial are affected when comparing singles versus married (but otherwise similar) persons?*

After loading the data, we run run the usual statistics and then plot numerical and categorical values.

```
> data(HMDA)
> # ?HMDA
> head(HMDA)

{not shown}

> summary(HMDA)
  deny          pirat            hirat            lvrat       
 no :2095   Min.   :0.0000   Min.   :0.0000   Min.   :0.0200  
 yes: 285   1st Qu.:0.2800   1st Qu.:0.2140   1st Qu.:0.6527  
            Median :0.3300   Median :0.2600   Median :0.7795  
            Mean   :0.3308   Mean   :0.2553   Mean   :0.7378  
            3rd Qu.:0.3700   3rd Qu.:0.2988   3rd Qu.:0.8685  
            Max.   :3.0000   Max.   :3.0000   Max.   :1.9500  
 chist    mhist    phist          unemp        selfemp   
 1:1353   1: 747   no :2205   Min.   : 1.800   no :2103  
 2: 441   2:1571   yes: 175   1st Qu.: 3.100   yes: 277  
 3: 126   3:  41              Median : 3.200             
 4:  77   4:  21              Mean   : 3.774             
 5: 182                       3rd Qu.: 3.900             
 6: 201                       Max.   :10.600             
 insurance  condomin    afam       single     hschool   
 no :2332   no :1694   no :2041   no :1444   no :  39  
 yes:  48   yes: 686   yes: 339   yes: 936   yes:234
```

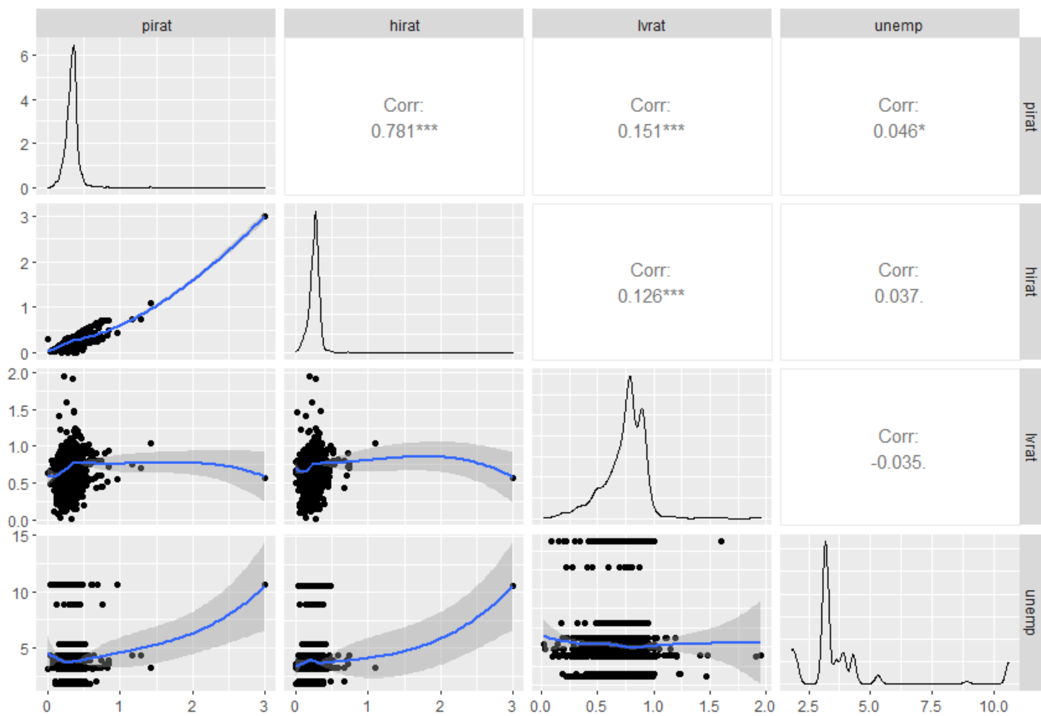

Figure 19: scatter plots and histogram

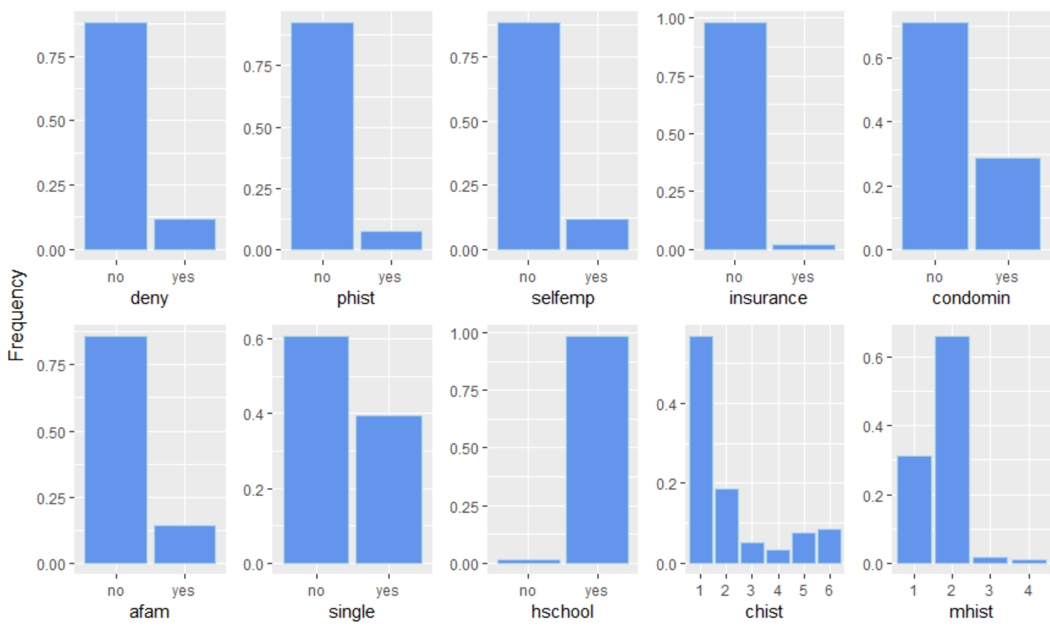

Figure 20: Barplot for categorical values

We will use Logistic regression and more specifically Poisson regression.

```
> n <- dim(HMDA)[1]
> HMDA_glm <- glm(deny~.,data = HMDA, family =poisson)
> summary(HMDA_glm)

Call:
glm(formula = deny ~ ., family = poisson, data = HMDA)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.4887  -0.4247  -0.3337  -0.2721   2.3174

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  -3.52566    0.44391  -7.942 1.99e-15 ***
pirat         2.89941    0.70773   4.097 4.19e-05 ***
hirat        -1.99976    0.74198  -2.695  0.00704 **
lvratmedium   0.43014    0.13356   3.221  0.00128 **
lvrathigh     1.02257    0.21193   4.825 1.40e-06 ***
chist         0.21369    0.03321   6.434 1.24e-10 ***
mhist         0.14892    0.11122   1.339  0.18057
phistyes      0.61440    0.15241   4.031 5.55e-05 ***
unemp         0.03745    0.02730   1.371  0.17024
selfempyes    0.43640    0.17294   2.523  0.01162 *
insuranceyes  1.57069    0.17352   9.052  < 2e-16 ***
condominyes  -0.10073    0.13385  -0.753  0.45173
afamyes       0.42428    0.14123   3.004  0.00266 **
singleyes     0.37512    0.12474   3.007  0.00264 **
hschoolyes   -0.84224    0.34128  -2.468  0.01359 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 1209.75  on 2379  degrees of freedom
Residual deviance:  881.59  on 2365  degrees of freedom
AIC: 1481.6

Number of Fisher Scoring iterations: 6
```

We then use AIC and BIC to obtain the best models using both direction. The results are as follows: AIC:

```
Step:  AIC=1480
deny ~ pirat + hirat + lvrat + chist + phist + unemp + selfemp +
    insurance + afam + single + hschool
```

BIC:

```
Step:  AIC=1541.6
deny ~ pirat + lvrat + chist + phist + insurance + single

> # GOF tests
> with(m1, pchisq(deviance, df.residual, lower.tail = FALSE))
[1] 1
> with(m2, pchisq(deviance, df.residual, lower.tail = FALSE))
[1] 1
```

In order to test the mortage denial we could run the model again and plot the fitted model to see the difference or run a test to check the difference, as has been done in previous exercises (wish I had the time to do it..).