Regression Tutorial

Mingrui Duan

0. Before we start: How to load data and check whether we have done it right?

To run “Regression”, we must load the dataset first to start our analysis. To achieve that, we can use the following function to load .csv file(Or the appropriate one, depends on your dataset type):

```
# Load the ".csv" file
data <- read.csv("Class7-WHR20_DataForFigure2.1.csv")
```

After loaded successfully, the very first thing we do is to do the following checks to make sure we could use the data properly afterwards: - The overall data size(How many points you have) - The attribute list(The columns) - The exact number of each attribute(The rows) You can run the following function to check quickly:

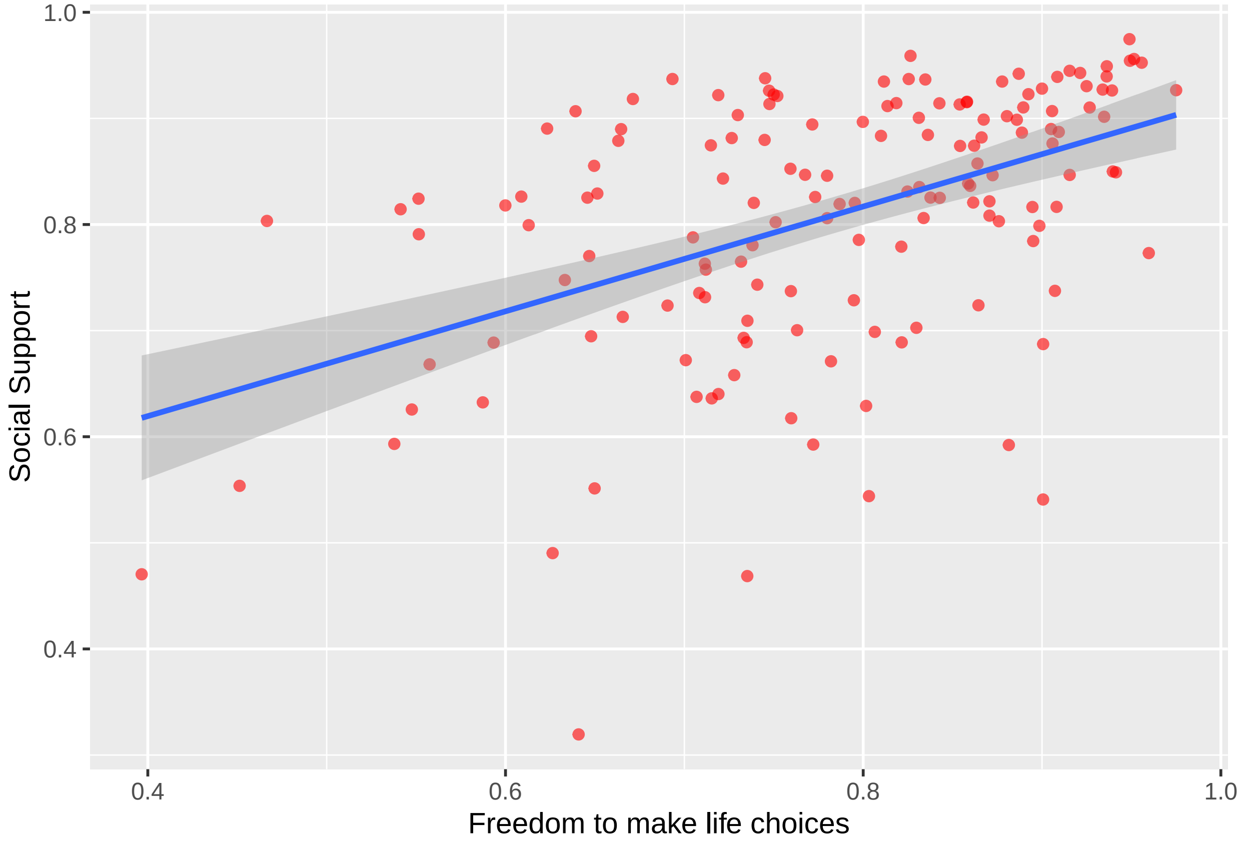
```
# Check the data quickly
summary(data)
```

```
##  Country.name       Regional.indicator  Ladder.score
##  Length:153         Length:153         Min.   :2.567
##  Class :character   Class :character   1st Qu.:4.724
##  Mode  :character   Mode  :character   Median :5.515
##                                        Mean   :5.473
##                                        3rd Qu.:6.229
##                                        Max.   :7.809
##  Standard.error.of.ladder.score  upperwhisker    lowerwhisker
##  Min.   :0.02590                Min.   :2.628   Min.   :2.506
##  1st Qu.:0.04070                1st Qu.:4.826   1st Qu.:4.603
##  Median :0.05061                Median :5.608   Median :5.431
##  Mean   :0.05354                Mean   :5.578   Mean   :5.368
##  3rd Qu.:0.06068                3rd Qu.:6.364   3rd Qu.:6.139
##  Max.   :0.12059                Max.   :7.870   Max.   :7.748
##  Logged.GDP.per.capita Social.support   Healthy.life.expectancy
##  Min.   : 6.493        Min.   :0.3195   Min.   :45.20
##  1st Qu.: 8.351        1st Qu.:0.7372   1st Qu.:58.96
##  Median : 9.456        Median :0.8292   Median :66.31
##  Mean   : 9.296        Mean   :0.8087   Mean   :64.45
##  3rd Qu.:10.265        3rd Qu.:0.9067   3rd Qu.:69.29
##  Max.   :11.451        Max.   :0.9747   Max.   :76.80
##  Freedom.to.make.life.choices   Generosity       Perceptions.of.corruption
##  Min.   :0.3966              Min.   :-0.30091   Min.   :0.1098
##  1st Qu.:0.7148              1st Qu.:-0.12701   1st Qu.:0.6830
##  Median :0.7998              Median :-0.03366   Median :0.7831
##  Mean   :0.7834              Mean   :-0.01457   Mean   :0.7331
##  3rd Qu.:0.8777              3rd Qu.: 0.08543   3rd Qu.:0.8492
##  Max.   :0.9750               Max.   : 0.56066   Max.   :0.9356
##  Ladder.score.in.Dystopia Explained.by..Log.GDP.per.capita
##  Min.   :1.972            Min.   :0.0000
##  1st Qu.:1.972            1st Qu.:0.5759
##  Median :1.972            Median :0.9185
##  Mean   :1.972            Mean   :0.8688
##  3rd Qu.:1.972            3rd Qu.:1.1692
##  Max.   :1.972            Max.   :1.5367
##  Explained.by..Social.support Explained.by..Healthy.life.expectancy
##  Min.   :0.0000               Min.   :0.0000
##  1st Qu.:0.9867               1st Qu.:0.4954
##  Median :1.2040               Median :0.7598
##  Mean   :1.1556               Mean   :0.6929
##  3rd Qu.:1.3871               3rd Qu.:0.8672
##  Max.   :1.5476               Max.   :1.1378
##  Explained.by..Freedom.to.make.life.choices Explained.by..Generosity
##  Min.   :0.0000                             Min.   :0.0000
##  1st Qu.:0.3815                             1st Qu.:0.1150
##  Median :0.4833                             Median :0.1767
##  Mean   :0.4636                             Mean   :0.1894
##  3rd Qu.:0.5767                             3rd Qu.:0.2555
##  Max.   :0.6933                             Max.   :0.5698
##  Explained.by..Perceptions.of.corruption Dystopia...residual
##  Min.   :0.00000                         Min.   :0.2572
##  1st Qu.:0.05580                         1st Qu.:1.6299
##  Median :0.09844                         Median :2.0463
##  Mean   :0.13072                         Mean   :1.9723
##  3rd Qu.:0.16306                         3rd Qu.:2.3503
##  Max.   :0.53316                         Max.   :3.4408
```

1. What is "Linear Regression" and the technique to achieve that

"Linear Regression" can be split into two words: Linear & Regression, which "Linear" means a straight line-like and "Regression" basically means that you want to simulate or predict the output with some input given the history data you have. So "Linear Regression" means you want to fit all the data points into some straight line. Here's an example(The data we have and how we fit that into a line):

```r
library("ggplot2") # Load the lib we're going to use
plt <- ggplot(data, aes(x=Freedom.to.make.life.choices, y=Social.support)) +
  geom_point(color = "red", alpha = 0.6) +
  stat_smooth(method = "lm", formula = y~x) +
  xlab("Freedom to make life choices") +
  ylab("Social Support")
plt
```



To run the best fit, which means this line are the "closest" to all the current points, we must use something to measure the "distance" each point to the line: That's the meaning of "Residual"(The mathematical error between prediction and true value). We will choose the line which has the minimal value after summing all points' residual.

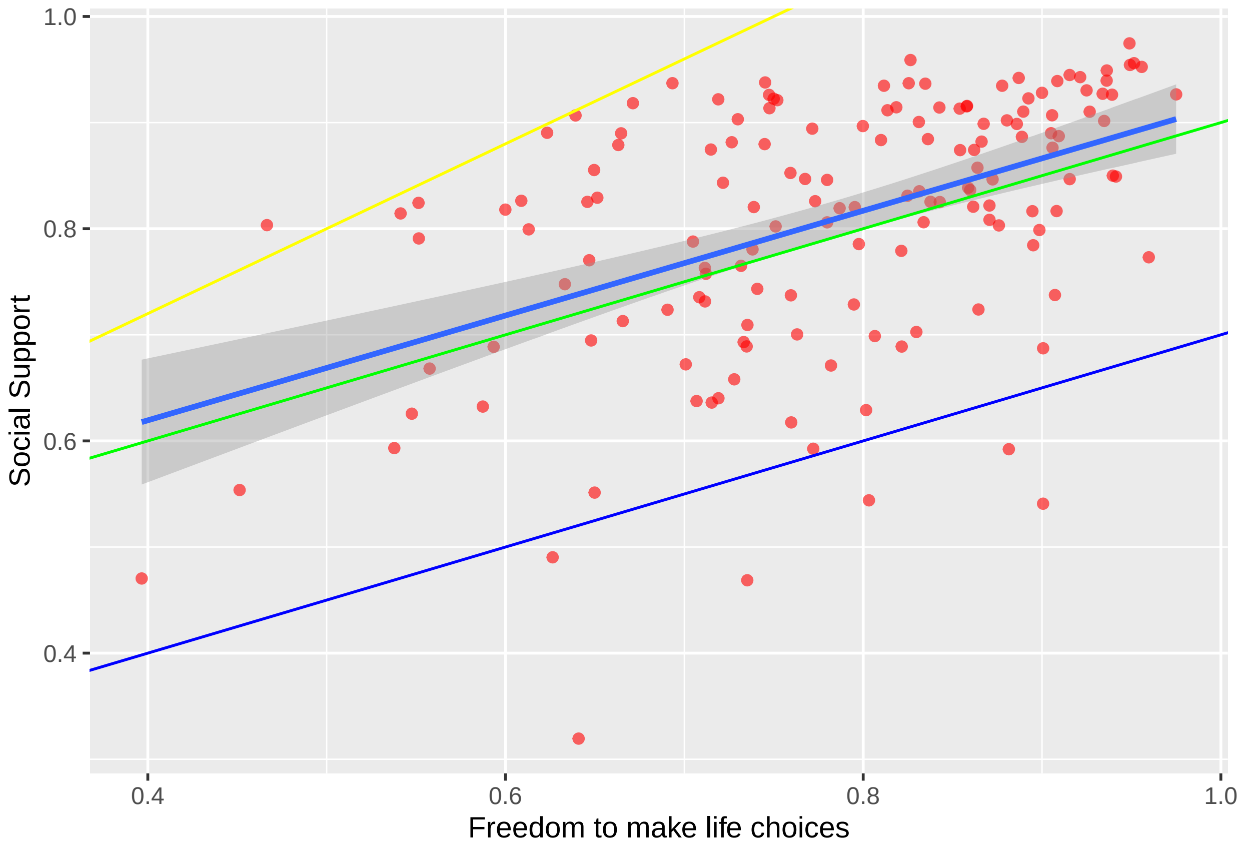
2. How can we conduct linear Regression in R?

To run linear regression in R, you should use the “lm” function and input the “predictors” in a proper order related to the attribute you want to “regression”, you can then get the linear regression you want.

```
# Run the linear regression
l <- lm(Social.support ~ Freedom.to.make.life.choices, data = data)
# Check the result of what we got
summary(l)
```

```
##
## Call:
## lm(formula = Social.support ~ Freedom.to.make.life.choices, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.41891 -0.04835  0.01808  0.07092  0.17287
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   0.42192    0.05835   7.231 2.24e-11 ***
## Freedom.to.make.life.choices  0.49377    0.07367   6.703 3.82e-10 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.107 on 151 degrees of freedom
## Multiple R-squared:  0.2293, Adjusted R-squared:  0.2242
## F-statistic: 44.93 on 1 and 151 DF,  p-value: 3.818e-10
```

```
plt + geom_abline(intercept = 0.4, slope = 0.5, color = "green") +
  geom_abline(intercept = 0.4, slope = 0.8, color = "yellow") +
  geom_abline(intercept = 0.2, slope = 0.5, color = "blue")
```

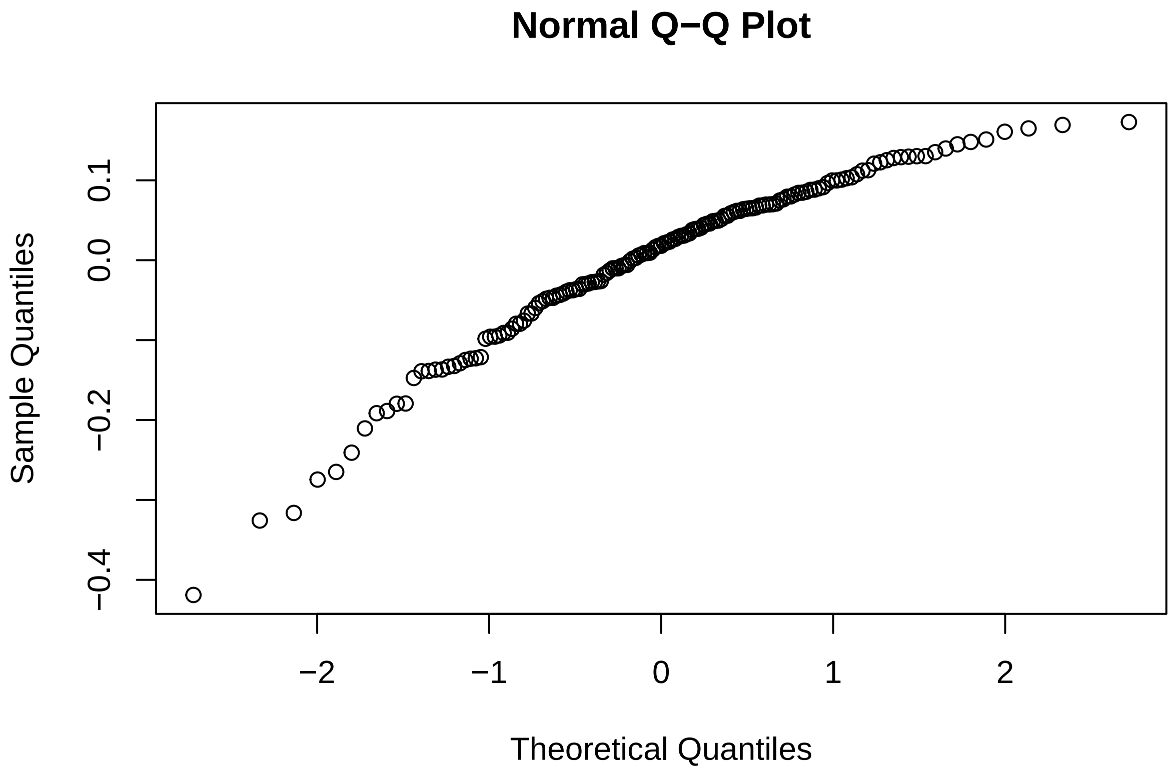


Here "slope" means how much your predictor can affect the output(graphically means how steer your line is); "Intercept" means when your predictors are "0", what will the output be(graphically means the value your line conjoint with the y-axis). p-value here indicates the probability that the observed relationship between the variables occurred by chance.

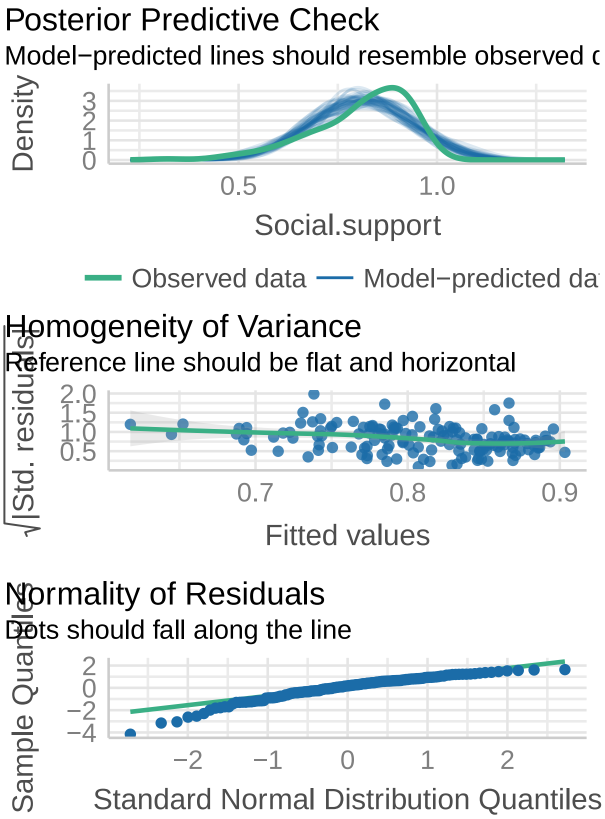
3. Key assumptions before doing linear regression

However, we cannot use linear regression without following these assumptions, otherwise our output won't be convincing: - The residuals must be normal for the data we use - The effects are linear - The standard deviation of the residuals is constant To check the assumption of the normality of residuals, we should use QQ plot; To check the effect are linear, we could function "check_model";

```
# Check whether the residuals are normal
qqnorm(residuals(l))
```



```
# Check whether the effect is linear
library(performance)
check_model(l)
```

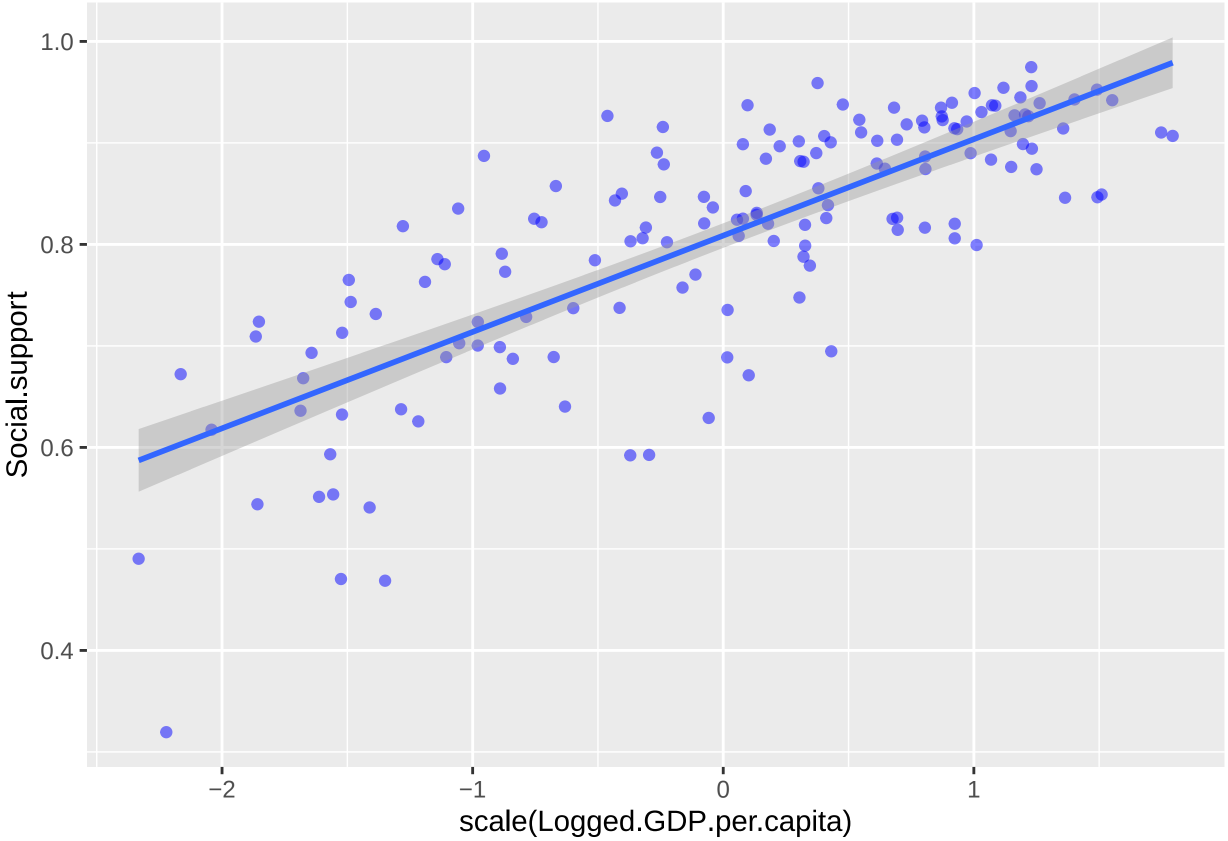


If the plot we got have the points fit perfectly with the central line, it illustrates that the residual distribution of out data is normal, which means then we can conduct liner regression on them. If the output of “cor()” is close to 1/-1, then it’s linear.

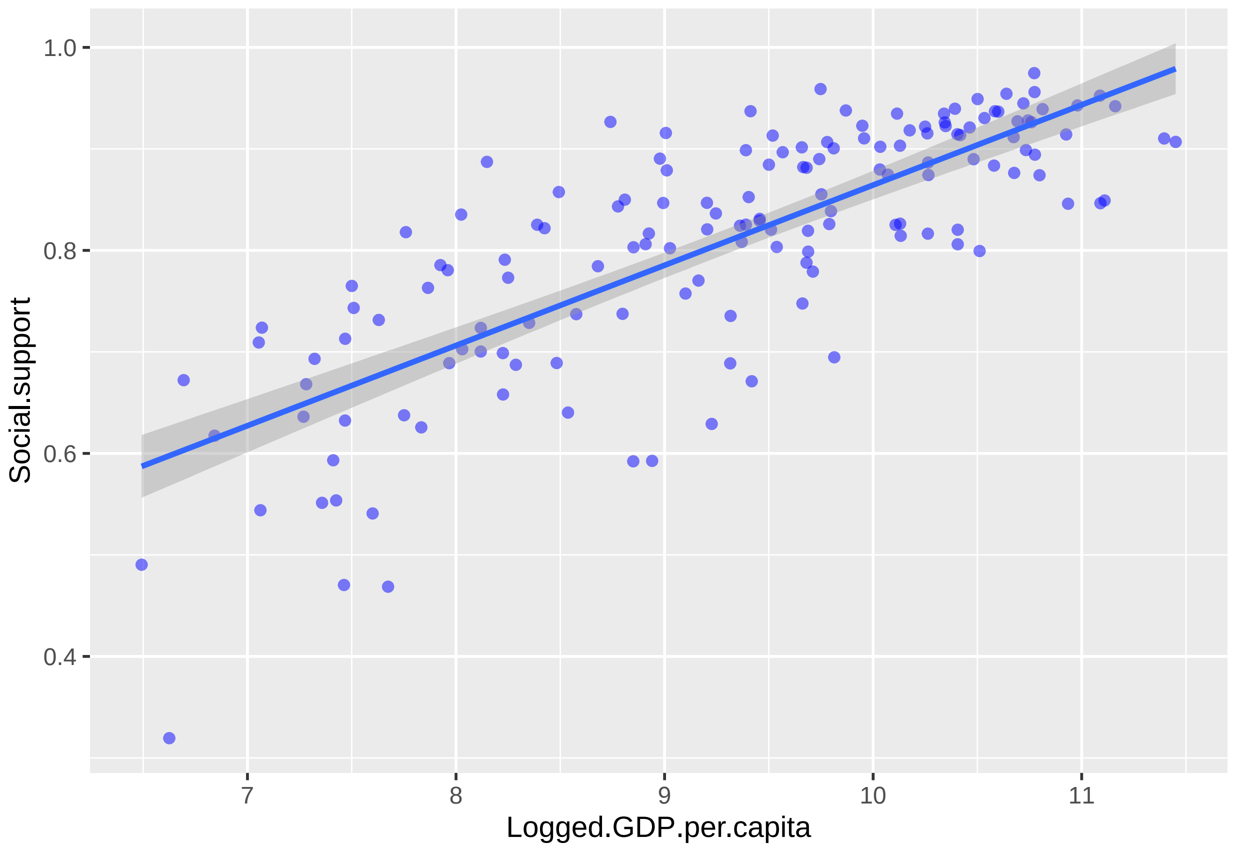
4. “Standardize” in Regression, Why & How

“Standardize” is really common in doing regressions, the reason is that in some situations our predictors’ scale can be extremely incomparable. Running the standardization process can help us avoid the bias due to different scale, reduce the impact of outliers and easier to interpret(coefficient means the change in the variable with one standard deviation change in the predictor variable).

```r
# scatter plots of standardized
plt.standardized <- ggplot(data, aes(x=scale(Logged.GDP.per.capita),y=Social.support)) +
  geom_point(color="blue", alpha=0.5) +
  stat_smooth(method="lm", formula=y~x)
plt.standardized
```



```r
# scatter plot of un-standardized
plt.unstandardized <- ggplot(data, aes(x=Logged.GDP.per.capita,y=Social.support)) +
  geom_point(color="blue", alpha=0.5) +
  stat_smooth(method="lm", formula=y~x)
plt.unstandardized
```



Graphically, after standardized, the predictors' distribution won't change but the middle point will be 0; the regression(let's say, the line) will have the same scale if every predictor are standardized.

5. Example of Continuous X with continuous interaction

```
# Run the linear regression
conti <- lm(Healthy.life.expectancy ~ scale(Logged.GDP.per.capita) +
  scale(Freedom.to.make.life.choices) +
  scale(Logged.GDP.per.capita)*scale(Freedom.to.make.life.choices),
            data = data)
summary(conti)
```

```
##
## Call:
## lm(formula = Healthy.life.expectancy ~ scale(Logged.GDP.per.capita) +
##     scale(Freedom.to.make.life.choices) + scale(Logged.GDP.per.capita) *
##     scale(Freedom.to.make.life.choices), data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.0011  -2.0530   0.8404   2.8241   7.0602
##
## Coefficients:
##                                                                   Estimate
## (Intercept)                                                        64.6909
## scale(Logged.GDP.per.capita)                                        5.6222
## scale(Freedom.to.make.life.choices)                                 0.6931
## scale(Logged.GDP.per.capita):scale(Freedom.to.make.life.choices)  -0.5894
##                                                                   Std. Error
## (Intercept)                                                           0.3194
## scale(Logged.GDP.per.capita)                                          0.3266
## scale(Freedom.to.make.life.choices)                                   0.3304
## scale(Logged.GDP.per.capita):scale(Freedom.to.make.life.choices)     0.2928
##                                                                   t value
## (Intercept)                                                       202.548
## scale(Logged.GDP.per.capita)                                       17.215
## scale(Freedom.to.make.life.choices)                                 2.098
## scale(Logged.GDP.per.capita):scale(Freedom.to.make.life.choices)  -2.013
##                                                                   Pr(>|t|)
## (Intercept)                                                         <2e-16 ***
## scale(Logged.GDP.per.capita)                                        <2e-16 ***
## scale(Freedom.to.make.life.choices)                                 0.0376 *
## scale(Logged.GDP.per.capita):scale(Freedom.to.make.life.choices)   0.0459 *
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 3.652 on 149 degrees of freedom
## Multiple R-squared:  0.7376, Adjusted R-squared:  0.7323
## F-statistic: 139.6 on 3 and 149 DF,  p-value: < 2.2e-16
```

Here both predictors are continuous. Interaction here means value of one coefficient depends on the value of another. The intercept value means when both predictors are mean, what y value would be; As for the other coefficients about the two predictors, it means how much can each predictor effect Healthy.life.expectancy; the last coefficient measures the level of one predictor effect the other one. To recover the predicted Healthy.life.expectancy value, you just need to compute the following formula with the coefficient you got from the summary: y = intercept + coefficient of predictor_1 * predictor_1 + coefficient of predictor_2 * predictor_2 + last coefficient * predictor_1 * predictor_2. To “summary”, you just need to call the “summary” function. p-value here means whether the model we simulated is fit; t-value here means the significance of each coefficient.

6. Example of discrete & continuous X with interacting with a slop and intercept.

```
# Add an indicator for whether you are in Europe
data$InEurope <- is.element(data$Regional.indicator,
                         c("Central and Eastern Europe", "Western Europe"))
dis <- lm(Healthy.life.expectancy ~ InEurope * scale(Freedom.to.make.life.choices) +
  scale(Logged.GDP.per.capita),
          data=data)
summary(dis)
```

```
##
## Call:
## lm(formula = Healthy.life.expectancy ~ InEurope * scale(Freedom.to.make.life.choices) +
##     scale(Logged.GDP.per.capita), data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.9107  -2.1247   0.3782   2.3535   6.1865
##
## Coefficients:
##                                                  Estimate Std. Error t value
## (Intercept)                                       63.9999     0.3508 182.416
## InEuropeTRUE                                       2.2126     0.8071   2.741
## scale(Freedom.to.make.life.choices)                1.1449     0.3561   3.215
## scale(Logged.GDP.per.capita)                       5.1835     0.3713  13.961
## InEuropeTRUE:scale(Freedom.to.make.life.choices)  -1.4613     0.7477  -1.954
##                                                  Pr(>|t|)
## (Intercept)                                       < 2e-16 ***
## InEuropeTRUE                                      0.00687 **
## scale(Freedom.to.make.life.choices)               0.00160 **
## scale(Logged.GDP.per.capita)                      < 2e-16 ***
## InEuropeTRUE:scale(Freedom.to.make.life.choices)  0.05255 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.6 on 148 degrees of freedom
## Multiple R-squared:  0.7466, Adjusted R-squared:  0.7398
## F-statistic:   109 on 4 and 148 DF,  p-value: < 2.2e-16
```

Here we got a discrete factor: "InEurope", and we assume it has interaction with "Freedom to make life choices". The intercept value means when the region is "not Europe" and the Logged.GDP.per.capita are the mean value, what would be the predicted Healthy.life.expectancy; As for the other coefficients about the two predictors, it means how much can each predictor effect y; The coefficient of the term "InEurope * scale(Freedom.to.make.life.choices)" means when the region belongs to Europe, how much the Freedom.to.make.life.choices affects the Healthy.life.expectancy. To recover, you should know whether the current region is in Europe. If it's in Europe, then you need to multiply the coefficient of the term "InEurope * scale(Freedom.to.make.life.choices)" with InEurope is TRUE and add the other predictor; However if the current region isn't in Europe, then you just only need to use another predictor. p-value here means whether the model we simulated is fit; t-value here means the significance of each coefficient. For the last t-value, it means how much significance InEurope do to the "scale(Freedom.to.make.life.choices)".