R for Research– A scientific approach

Yaseen

15 October, 2023

Descriptive analysis of data

Mean

The most basic estimate of location is the mean, or average value. The mean is the sum of all the values divided by the number of values.

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Trimmed mean

A variation of the mean is a *trimmed mean*, which you calculate by dropping a fixed number of sorted values at each end and then taking an average of the remaining values.

Example

```
library(knitr)
state=c('Kerala','Tamilnadu','Karnataka','Andra', 'Thelungana','Bihar','Bengal')
Population=c(5,6,4,35,22,15,79)
states=data.frame(state,Population)
kable(states)
```

state	Population
Kerala	5
Tamilnadu	6
Karnataka	4
Andra	35
Thelungana	22
Bihar	15
Bengal	79

#Finding basic statistics

```
summary(states)
```

```
##     state             Population
##  Length:7           Min.   : 4.00
##  Class :character   1st Qu.: 5.50
##  Mode  :character   Median :15.00
##                     Mean   :23.71
```

```
##                      3rd Qu.:28.50
##                      Max.   :79.00
```

```
mean(states$Population)
```

```
## [1] 23.71429
```

```
mean(states$Population, trim=0.15)
```

```
## [1] 16.6
```

```
median(states$Population)
```

```
## [1] 15
```

#Correlation and Regression

- Correlation determines if one variable varies systematically as another variable changes.

– The three forms of correlation presented here are Pearson, Kendall, and Spearman. The test determining the p-value for Pearson correlation is a parametric test that assumes that data are bi variate normal. Kendall and Spearman correlation use non parametric tests.

- Linear regression specifies one variable as the independent variable and another as the dependent variable. The resultant model relates the variables with a linear relationship.

– The tests associated with linear regression are parametric and assume normality, homoscedasticity, and independence of residuals, as well as a linear relationship between the two variables.

Immediate take away

- For Pearson correlation, two interval/ratio variables. Together the data in the variables are bi variate normal. The relationship between the two variables is linear. Outliers can detrimentally affect results.
- For Kendall correlation, two variables of interval/ratio or ordinal type.
- For Spearman correlation, two variables of interval/ratio or ordinal type.
- For linear regression, two interval/ratio variables. The relationship between the two variables is linear. Residuals are normal, independent, and homoscedastic. Outliers can affect the results unless robust methods are used.

Correlation Analysis

Null hypotheses

- For correlation, null hypothesis, H_0: The correlation coefficient (r, τ, or ρ) is zero. Or, *there is no correlation between the two variables.*
- For linear regression, null hypothesis, H_0: The slope of the fit line is zero. Or, *there is no linear relationship between the two variables.*

Concluding the test

If $p < 0.05$, then the null hypothesis is rejected with 95% confidence.

Correlation analysis in R

Packages required The packages used in this section include:

- psych

- Hmisc
- PerformanceAnalytics
- ggplot2
- rcompanion

Installation The following commands will install these packages if they are not already installed:

```
# library(devtools)
#
# install_github("cran/PerformanceAnalytics")
#
# if(!require(psych)){install.packages("psych")}


if(!require(ggplot2)){install.packages("ggplot2")}
```

```
## Loading required package: ggplot2
```

```
if(!require(rcompanion)){install.packages("rcompanion")}
```

```
## Loading required package: rcompanion
```

Examples for correlation

Brendon Small and company recorded several measurements for students in their classes related to their nutrition education program: Grade, Weight in kilograms, intake of Calories per day, daily Sodium intake in milligrams, and Score on the assessment of knowledge gain.

```
Input = ("
Instructor       Grade   Weight  Calories Sodium  Score
'Brendon Small'     6      43     2069    1287      77
'Brendon Small'     6      41     1990    1164      76
'Brendon Small'     6      40     1975    1177      76
'Brendon Small'     6      44     2116    1262      84
'Brendon Small'     6      45     2161    1271      86
'Brendon Small'     6      44     2091    1222      87
'Brendon Small'     6      48     2236    1377      90
'Brendon Small'     6      47     2198    1288      78
'Brendon Small'     6      46     2190    1284      89
'Jason Penopolis'   7      45     2134    1262      76
'Jason Penopolis'   7      45     2128    1281      80
'Jason Penopolis'   7      46     2190    1305      84
'Jason Penopolis'   7      43     2070    1199      68
'Jason Penopolis'   7      48     2266    1368      85
'Jason Penopolis'   7      47     2216    1340      76
'Jason Penopolis'   7      47     2203    1273      69
'Jason Penopolis'   7      43     2040    1277      86
'Jason Penopolis'   7      48     2248    1329      81
'Melissa Robins'    8      48     2265    1361      67
'Melissa Robins'    8      46     2184    1268      68
'Melissa Robins'    8      53     2441    1380      66
'Melissa Robins'    8      48     2234    1386      65
'Melissa Robins'    8      52     2403    1408      70
'Melissa Robins'    8      53     2438    1380      83
```

```
'Melissa Robins'    8      52     2360    1378      74
'Melissa Robins'    8      51     2344    1413      65
'Melissa Robins'    8      51     2351    1400      68
'Paula Small'       9      52     2390    1412      78
'Paula Small'       9      54     2470    1422      62
'Paula Small'       9      49     2280    1382      61
'Paula Small'       9      50     2308    1410      72
'Paula Small'       9      55     2505    1410      80
'Paula Small'       9      52     2409    1382      60
'Paula Small'       9      53     2431    1422      70
'Paula Small'       9      56     2523    1388      79
'Paula Small'       9      50     2315    1404      71
'Coach McGuirk'    10      52     2406    1420      68
'Coach McGuirk'    10      58     2699    1405      65
'Coach McGuirk'    10      57     2571    1400      64
'Coach McGuirk'    10      52     2394    1420      69
'Coach McGuirk'    10      55     2518    1379      70
'Coach McGuirk'    10      52     2379    1393      61
'Coach McGuirk'    10      59     2636    1417      70
'Coach McGuirk'    10      54     2465    1414      59
'Coach McGuirk'    10      54     2479    1383      61
")

Data = read.table(textConnection(Input),header=TRUE)


###  Order factors by the order in data frame
###  Otherwise, R will alphabetize them

Data$Instructor = factor(Data$Instructor,
                         levels=unique(Data$Instructor))


###  Check the data frame

library(psych)
```

```
##
## Attaching package: 'psych'

## The following object is masked from 'package:rcompanion':
##
##     phi

## The following objects are masked from 'package:ggplot2':
##
##     %+%, alpha
```

```
headTail(Data)
```

```
##        Instructor Grade Weight Calories Sodium Score
## 1   Brendon Small     6     43     2069   1287    77
## 2   Brendon Small     6     41     1990   1164    76
## 3   Brendon Small     6     40     1975   1177    76
## 4   Brendon Small     6     44     2116   1262    84
## ...          <NA>   ...    ...      ...    ...   ...
```

```
## 42  Coach McGuirk    10     52     2379   1393    61
## 43  Coach McGuirk    10     59     2636   1417    70
## 44  Coach McGuirk    10     54     2465   1414    59
## 45  Coach McGuirk    10     54     2479   1383    61
```

```
str(Data)
```

```
## 'data.frame':    45 obs. of  6 variables:
##  $ Instructor: Factor w/ 5 levels "Brendon Small",..: 1 1 1 1 1 1 1 1 2 ...
##  $ Grade     : int  6 6 6 6 6 6 6 6 7 ...
##  $ Weight    : int  43 41 40 44 45 44 48 47 46 45 ...
##  $ Calories  : int  2069 1990 1975 2116 2161 2091 2236 2198 2190 2134 ...
##  $ Sodium    : int  1287 1164 1177 1262 1271 1222 1377 1288 1284 1262 ...
##  $ Score     : int  77 76 76 84 86 87 90 78 89 76 ...
```

```
summary(Data)
```

```
##            Instructor     Grade        Weight         Calories
##  Brendon Small  :9    Min.   : 6   Min.   :40.00   Min.   :1975
##  Jason Penopolis:9    1st Qu.: 7   1st Qu.:46.00   1st Qu.:2190
##  Melissa Robins :9    Median : 8   Median :50.00   Median :2308
##  Paula Small    :9    Mean   : 8   Mean   :49.51   Mean   :2305
##  Coach McGuirk  :9    3rd Qu.: 9   3rd Qu.:53.00   3rd Qu.:2431
##                       Max.   :10   Max.   :59.00   Max.   :2699
##      Sodium         Score
##  Min.   :1164   Min.   :59.0
##  1st Qu.:1284   1st Qu.:67.0
##  Median :1380   Median :71.0
##  Mean   :1347   Mean   :73.2
##  3rd Qu.:1405   3rd Qu.:80.0
##  Max.   :1422   Max.   :90.0
```

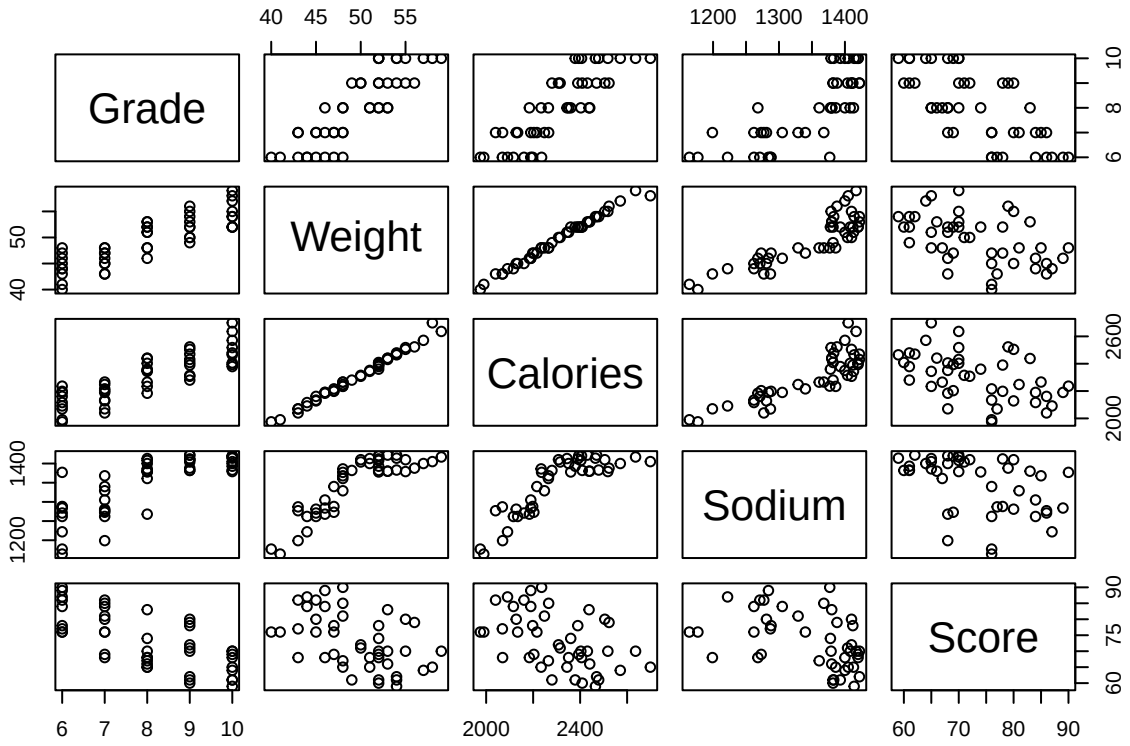
```
### Remove unnecessary objects

rm(Input)
```

Visualizing correlated variables

Multiple correlation

The pairs function can plot multiple numeric or integer variables on a single plot to look for correlations among the variables.

```
pairs(data=Data,
    ~ Grade + Weight + Calories + Sodium + Score)
```



Correlation Matrix

The corr.test function requires that the data frame contain only numeric or integer variables, so we will first create a new data frame called Data.num containing only the numeric and integer variables.

```
library(Hmisc)
```

```
##
## Attaching package: 'Hmisc'
```

```
## The following object is masked from 'package:psych':
##
##     describe
```

```
## The following objects are masked from 'package:base':
##
##     format.pval, units
```

```
Data.num = Data[,c("Grade", "Weight", "Calories", "Sodium", "Score")]

cm=rcorr(as.matrix(Data.num),type = "pearson")
cm$r
```

```
##                Grade     Weight   Calories     Sodium      Score
## Grade      1.0000000  0.8537015  0.8480573  0.7855545 -0.7032118
## Weight     0.8537015  1.0000000  0.9945259  0.8654492 -0.4840410
## Calories   0.8480573  0.9945259  1.0000000  0.8489548 -0.4846330
## Sodium     0.7855545  0.8654492  0.8489548  1.0000000 -0.4497510
## Score     -0.7032118 -0.4840410 -0.4846330 -0.4497510  1.0000000
```

```
cm$P
```

```
##                 Grade       Weight     Calories       Sodium        Score
```

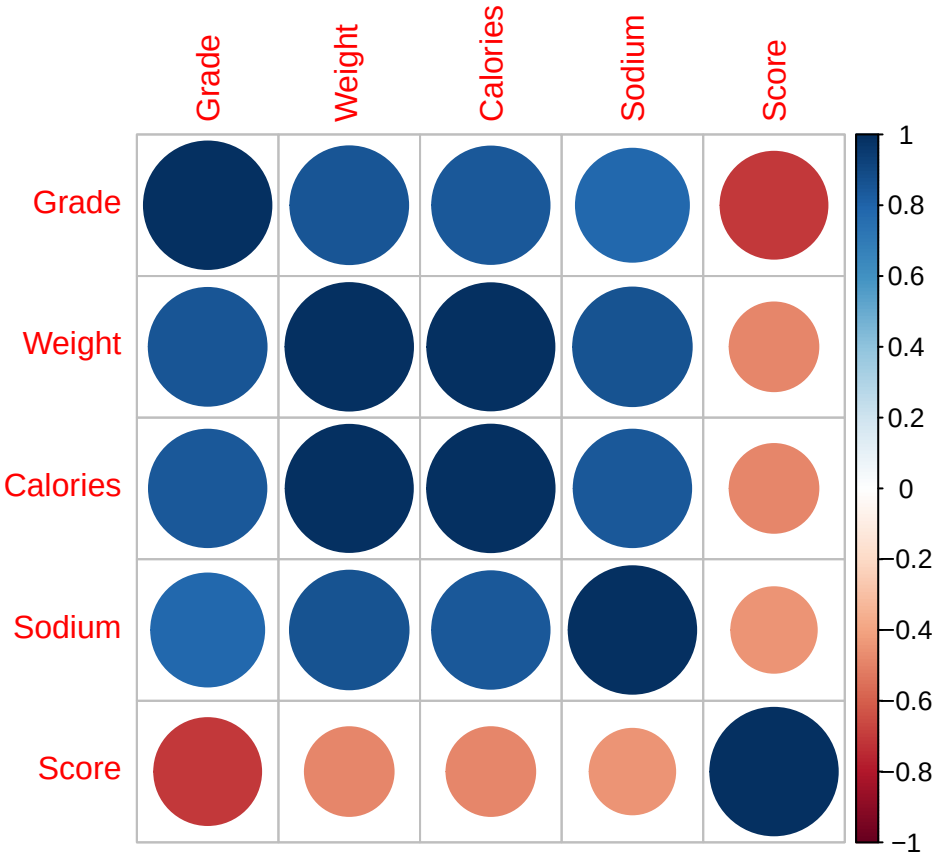
```
## Grade              NA 9.192647e-14 1.953993e-13 1.652476e-10 7.179472e-08
## Weight   9.192647e-14           NA 0.000000e+00 1.731948e-14 7.546753e-04
## Calories 1.953993e-13 0.000000e+00           NA 1.736389e-13 7.418414e-04
## Sodium   1.652476e-10 1.731948e-14 1.736389e-13           NA 1.937869e-03
## Score    7.179472e-08 7.546753e-04 7.418414e-04 1.937869e-03           NA
```

Correlation Plot

```
library(corrplot)
```

```
## corrplot 0.92 loaded
```

```
cp=Data[2:6]
corrplot(cor(cp),   method  =   "circle")
```

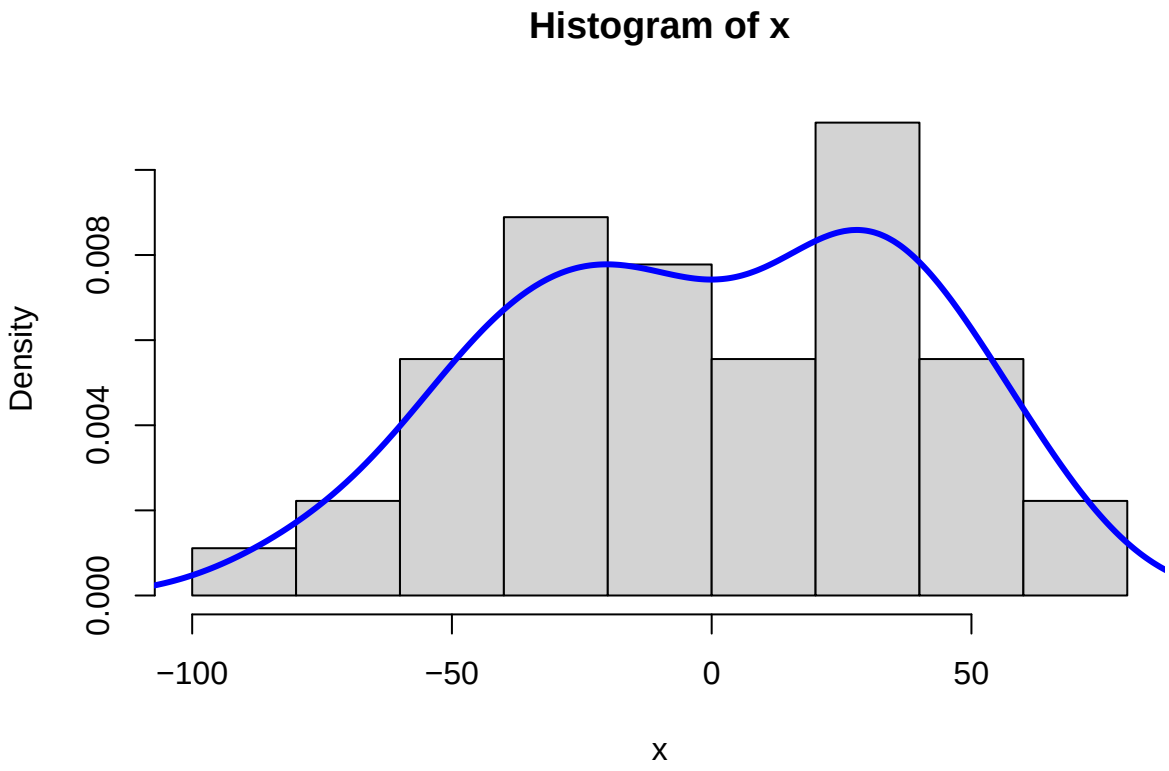


Plot residuals

It's not a bad idea to look at the residuals from Pearson correlation to be sure the data meet the assumption of bivariate normality. Unfortunately, the cor.test function doesn't supply residuals. One solution is to use the lm function, which actually redoes the analysis as a linear regression.

```
model = lm(Sodium ~ Calories,
           data = Data)

x = residuals(model)
```

```r
hist(x, freq=FALSE)
lines(density(model$residuals),lwd=3,col="blue")
```



Fitting a regression line

When we decide to consider one of the variables as as a response and the other as a predictor, we attempt to fit a line that best describes this relation. There are three types of lines we can fit, usually in this order:

1. Exploratory, non-parametric
2. Parametric
3. Robust

The first kind just gives a "smooth" impression of the relation. The second fits according to some optimality criterion; the classic least-squares estimate is in this class. The third is also parametric but optimizes some criterion that protects against a few unusual data values in favour of the majority of the data.

A common non-parametric fit is the LOWESS ("locally weighted regression and smoothing scatterplots") [35], computed by R method lowess. This has a useradjustable parameter, the smoother's "span", which is the proportion of points in the plot which influence the smooth at each value; larger values result in a smoother plot. This allows us to visualise the relation either up close (low value of parameter) or more generally (high). The default is 2/3.

#Linear regression

Dependent and Independent variables

When plotted, the dependent variable is usually placed on the y-axis, and the independent variable is usually placed in the x-axis.

Interpretation of coefficients The outcome of linear regression includes estimating the intercept and the slope of the linear model. Linear regression can then be used as a predictive model, whereby the model can be

used to predict a y value for any given x. In practice, the model shouldn't be used to predict values beyond the range of the x values used to develop the model.

Assumptions Linear regression assumes a linear relationship between the two variables, normality of the residuals, independence of the residuals, and homoscedasticity of residuals.

Note on writing r-squared

For bivariate linear regression, the r-squared value often uses a lower case r; however, some authors prefer to use a capital R. For multiple regression, the R in the R-squared value is usually capitalized. The name of the statistic may be written out as "r-squared" for convenience, or as r^2.

#Linear Regression using R

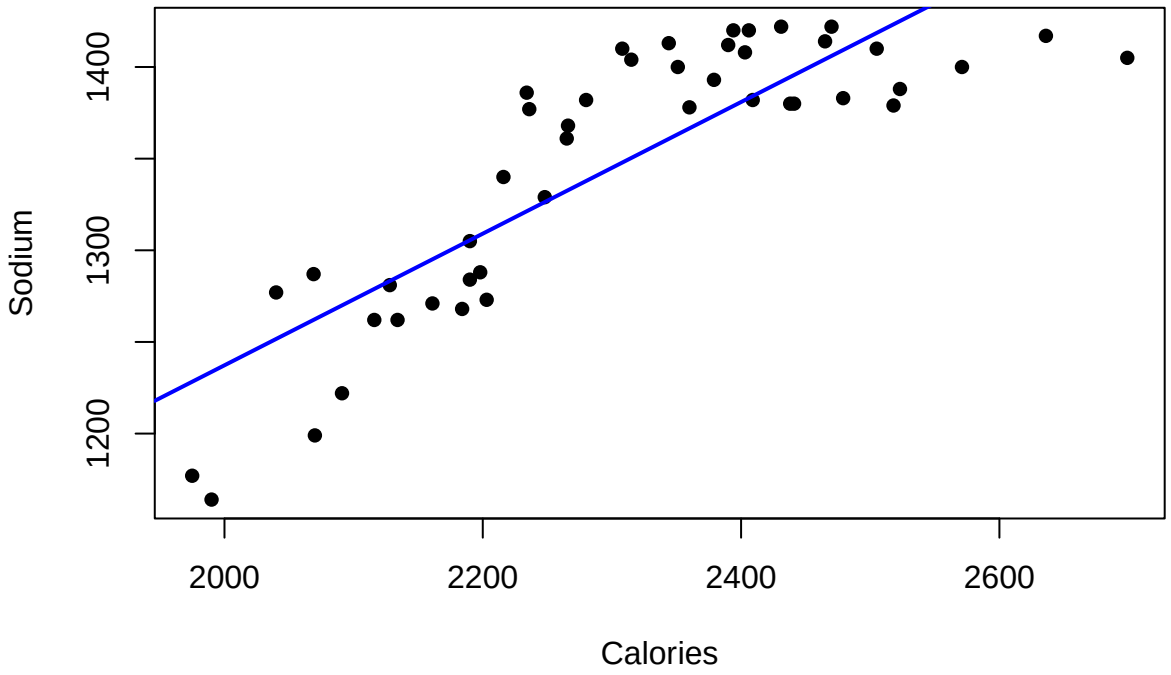
Linear regression can be performed with the *lm* function. The summary function for lm model objects includes estimates for model parameters (intercept and slope), as well as an r-squared value for the model and p-value for the model.

```
model = lm(Sodium ~ Calories,
           data = Data)

summary(model)
```

```
##
## Call:
## lm(formula = Sodium ~ Calories, data = Data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -83.263 -26.263  -0.486  29.973  64.714
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 519.07547   78.78211   6.589 5.09e-08 ***
## Calories      0.35909    0.03409  10.534 1.74e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 38.89 on 43 degrees of freedom
## Multiple R-squared:  0.7207, Adjusted R-squared:  0.7142
## F-statistic:   111 on 1 and 43 DF,  p-value: 1.737e-13
```

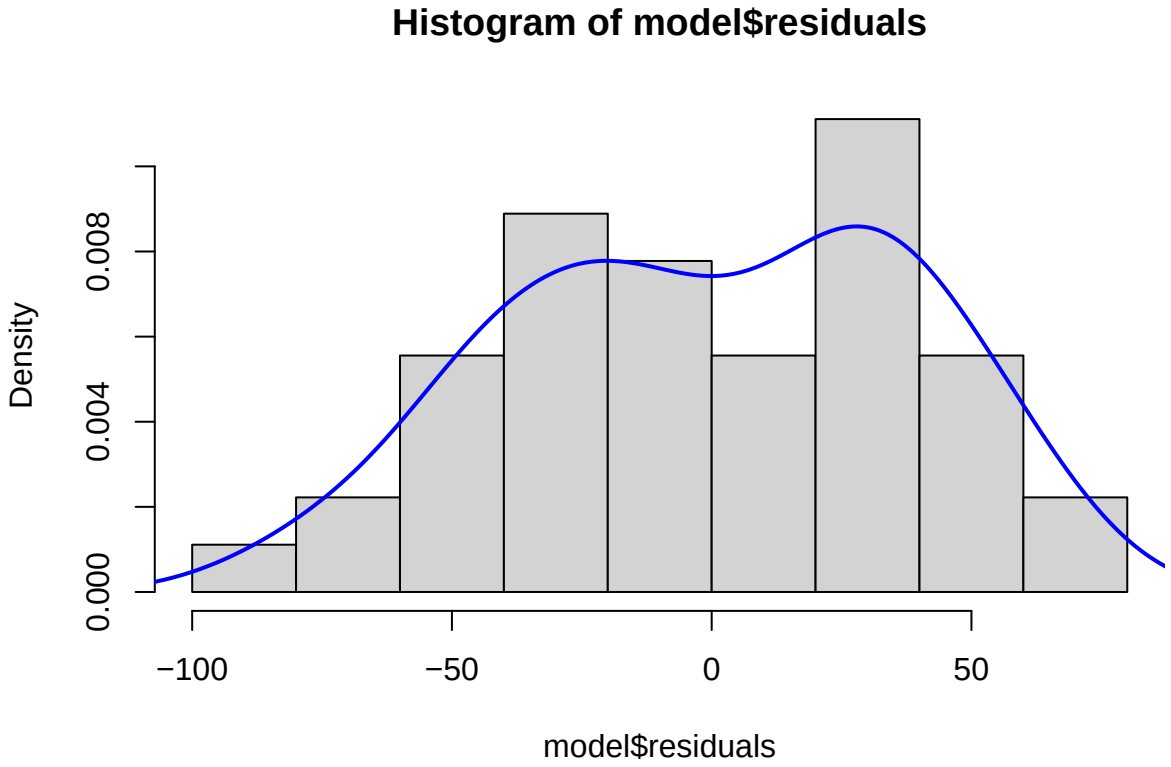
Plot data with best fit line

```
plot(Sodium ~ Calories,
     data=Data,
     pch=16,
     xlab = "Calories",
     ylab = "Sodium")

abline(model,
       col = "blue",
       lwd = 2)
```



Ploting Residuals in Regression

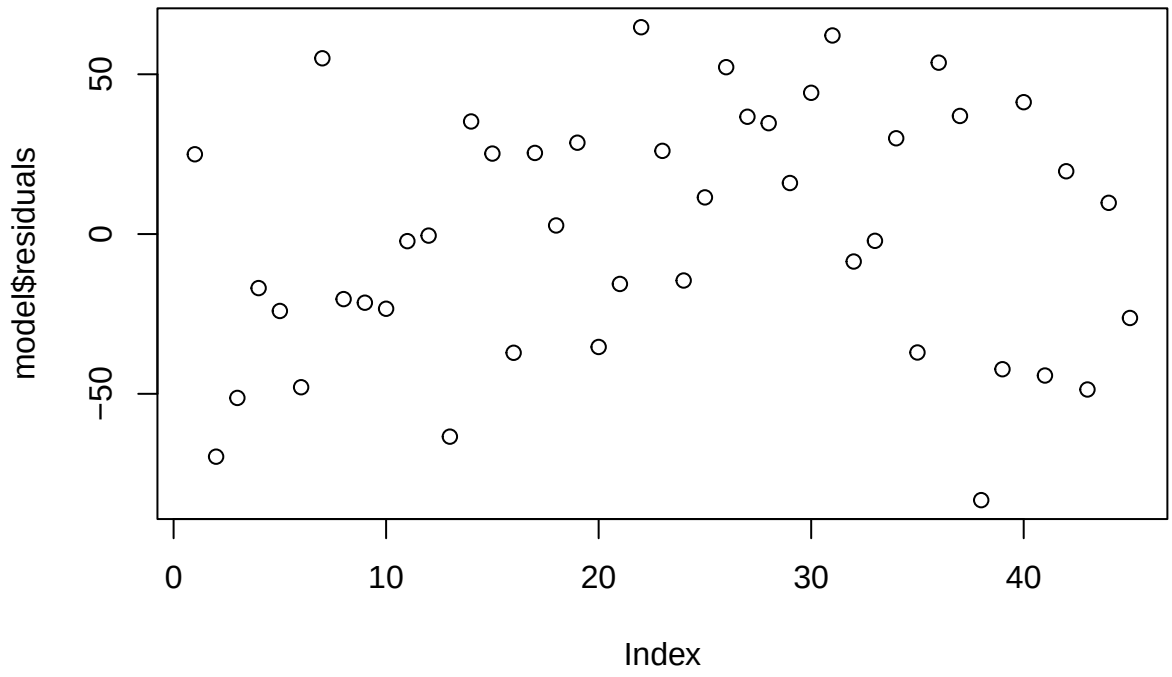
```
hist(model$residuals,freq = FALSE)
lines(density(model$residuals),lwd=2,col="blue")
```




```
#library(rcompanion)
#plotNormalHistogram(x)
```

Residual plot

```
plot( model$residuals)
```



#Polynomial Regression

Polynomial regression adds additional terms to the model, so that the terms include some set of the linear, quadratic, cubic, and quadratic, etc., forms of the independent variable.

```
model_1 = lm(Sodium ~ Calories,
             data = Data)

model_2 = lm(Sodium ~ I(Calories) + I(Calories^2),
             data = Data)

model_3 = lm(Sodium ~ I(Calories) + I(Calories^2) + I(Calories^3),
             data = Data)

model_4 = lm(Sodium ~ I(Calories) + I(Calories^2) + I(Calories^3) + I(Calories^4),
             data = Data)
```

Choosing Best Model

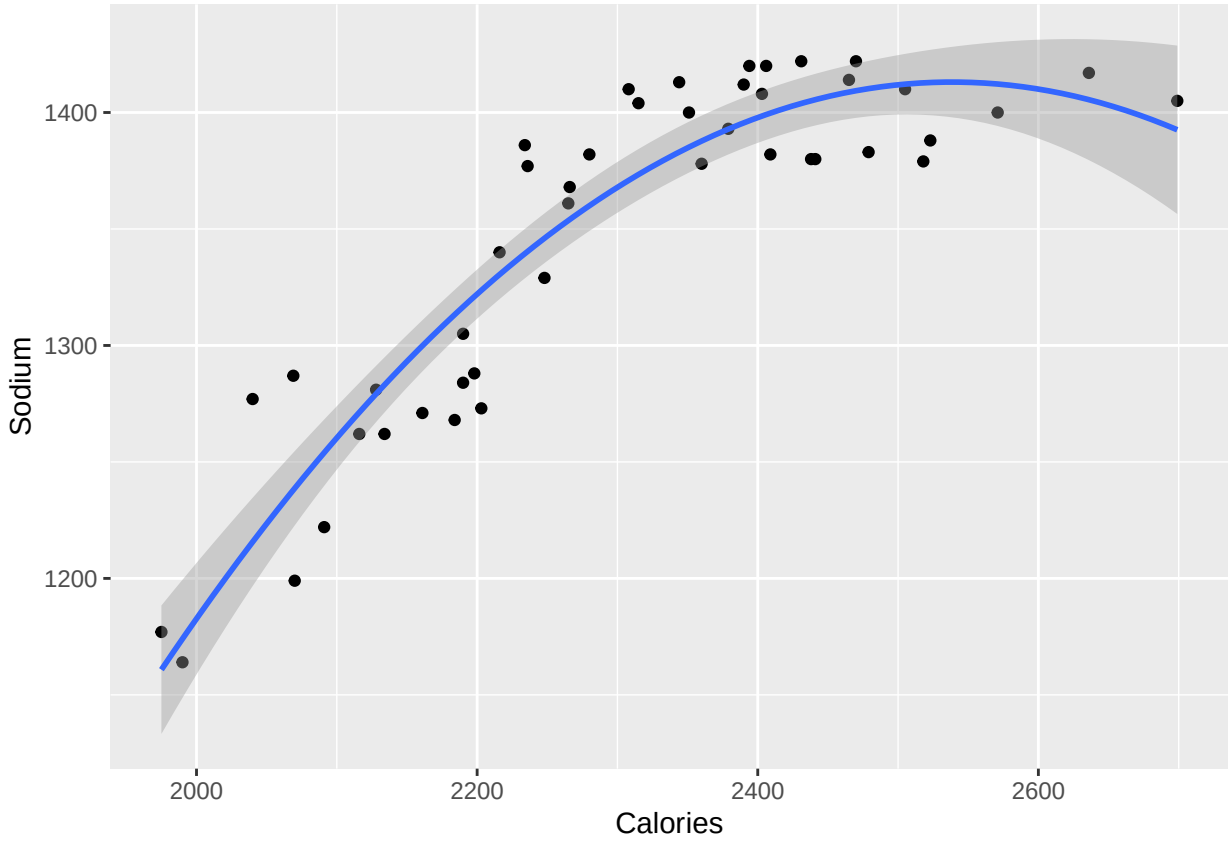
Chances are that we will not need all of the polynomial terms to adequately model our data. One approach to choosing the best model is to construct several models with increasing numbers of polynomial terms, and then use a model selection criterion like AIC, AICc, or BIC to choose the best one.

```
library(rcompanion)

cmdl=compareLM(model_1, model_2, model_3, model_4)
```

Plot of best fit line with confidence interval

Plot of best fit line with confidence interval

```
ggplot(data=Data,
               aes(x = Calories,
                   y = Sodium)) +
  geom_point() +
  geom_smooth(method  = "lm",
               formula = y ~ poly(x, 2, raw=TRUE), ### polynomial of order 2

               se      = TRUE)
```



Exercise-1

I. Consider the data from Brendon, Jason, Melissa, Paula, and McGuirk. Report for each answer, indicate how you know, when appropriate, by reporting the values of the statistic you are using or other information you used.

a. Which two variables are the most strongly correlated?

b. Which two variables are the least strongly correlated?

c. Are there any pairs of variables that are uncorrelated? Which?

d. Name a pair of variables that is positively correlated.

e. Name a pair of variables that is negatively correlated.

f. Is Sodium significantly correlated with Calories?

g. By linear regression, is there a significant linear relationship of Sodium vs. Calories?

h. Does the quadratic polynomial model fit the Sodium vs. Calories data better than the linear model? Consider the p-value, the r-squared value, the range of values for each of Sodium and Calories, and your practical conclusions.

Exercise-II

II. 2. As part of a professional skills program, a 4-H club tests its members for typing proficiency (Words.per.minute), Proofreading skill, proficiency with using a Spreadsheet, and acumen in Statistics.

```
input2=("Instructor  Grade Words.per.minute Proofreading Spreadsheet  Statistics
'Dr. Katz'  6     35               53           75           61
'Dr. Katz'  6     50               77           24           51
'Dr. Katz'  6     55               71           62           55
'Dr. Katz'  6     60               78           27           91
'Dr. Katz'  6     65               84           44           95
'Dr. Katz'  6     60               79           38           50
'Dr. Katz'  6     70               96           12           94
'Dr. Katz'  6     55               61           55           76
'Dr. Katz'  6     45               73           59           75
'Dr. Katz'  6     55               75           55           80
'Dr. Katz'  6     60               85           35           84
'Dr. Katz'  6     45               61           49           80
'Laura'     7     55               59           79           57
'Laura'     7     60               60           60           60
'Laura'     7     75               90           19           64
'Laura'     7     65               87           32           65
'Laura'     7     60               70           33           94
'Laura'     7     70               84           27           54
'Laura'     7     75               87           24           59
'Laura'     7     70               97           38           74
'Laura'     7     65               86           30           52
'Laura'     7     72               91           36           66
'Laura'     7     73               88           20           57
'Laura'     7     65               86           19           71
'Ben Katz'  8     55               84           20           76
'Ben Katz'  8     55               63           44           94
'Ben Katz'  8     70               95           31           88
'Ben Katz'  8     55               63           69           93
'Ben Katz'  8     65               65           47           70
'Ben Katz'  8     60               61           63           92
'Ben Katz'  8     70               80           35           60
'Ben Katz'  8     60               88           38           58
'Ben Katz'  8     60               71           65           99
'Ben Katz'  8     62               78           46           54
'Ben Katz'  8     63               89           17           60
'Ben Katz'  8     65               75           33           77")
```

Analysis part

For each of the following, answer the question, and show the output from the analyses you used to answer the question. Where relevant, indicate how you know.

a. Which two variables are the most strongly correlated?

b. Name a pair of variables that are uncorrelated.

c. Name a pair of variables that is positively correlated.

d. Name a pair of variables that is negatively correlated.

e. Consider the correlation between Spreadsheet and Proofreading.

   i. What is the value of the correlation coefficient r for this correlation?

   ii. What is the value of tau?

   iii. What is the value of ρ?

f. Conduct a linear regression of Proofreading vs. Words.per.minute.

   i. What is the p-value for this model?

   ii. What is the r-squared value?

   iii. Do the residuals suggest that the linear regression model is an appropriate model?

   iv. What can you conclude about the results of the linear regression? Consider the p-value, the r-squared value, the range of values for each of Proofreading and Words.per.minute, and your practical conclusions.