R is also for Filipino Researchers

Joseph S. Tabadero, Jr.

October 28, 2017

What is R? (R Foundation 2017)

"R is a language and environment for statistical computing and graphics. It is a GNU project which is similar to the S language and environment which was developed at Bell Laboratories (formerly AT&T, now Lucent Technologies) by John Chambers and colleagues. R can be considered as a different implementation of S. There are some important differences, but much code written for S runs unaltered under R.

R provides a wide variety of statistical (linear and nonlinear modelling, classical statistical tests, time-series analysis, classification, clustering, …) and graphical techniques, and is highly extensible. The S language is often the vehicle of choice for research in statistical methodology, and R provides an Open Source route to participation in that activity."

(Some) Advantages of using R (from an R fanboy)

- R is the most comprehensive statistical analysis software available.
- R is a programming language developed by research and practicing statisticians for statisticians.
- The graphical capabilities and options in R far surpasses the available graphical capabilities in other statistical packages.
- R is free and open source licensed to The R Foundation for Statistical Computing under the GNU General Public License
- R has a large community of users, developers, and bug-fixers. You can contribute to the development of R, too, by becoming an active member of the community.
- R has over 10,000 packages available in CRAN and more available in bioconductor and Github repositories.
- There are a lot of free books, websites, and coursewares available for learning R.

(Some) Disadvantages of Using R

- R has a steep learning curve (?)
- Documentation is sometimes lacking
- The quality of some packages is sometimes questionable
- There is in general no one to complain to when something goes wrong
- R's memory management sucks (?)

Who Uses R? (Bhalla 2017)

- Facebook - For behavior analysis related to status updates and profile pictures.
- Google - For advertising effectiveness and economic forecasting.
- Twitter - For data visualization and semantic clustering
- Microsoft - Acquired Revolution R company and use it for a variety of purposes.
- Uber - For statistical analysis
- Airbnb - Scale data science.
- IBM - Joined R Consortium Group
- ANZ - For credit risk modeling

Why use R?

- It is free (and open source)!
- R is the most popular tool for analytics/data science (Piatetsky 2016).
- Ranked 5th in most popular software based on number of job offerings: SQL, Python, Java, Hadoop, R, C/C++/C#, SAS, Apache Spark, Tableau, Apache Hive (Muenchen 2017)
- R has surpassed SAS in scholarly use–but still way behind SPSS (Muenchen 2016)

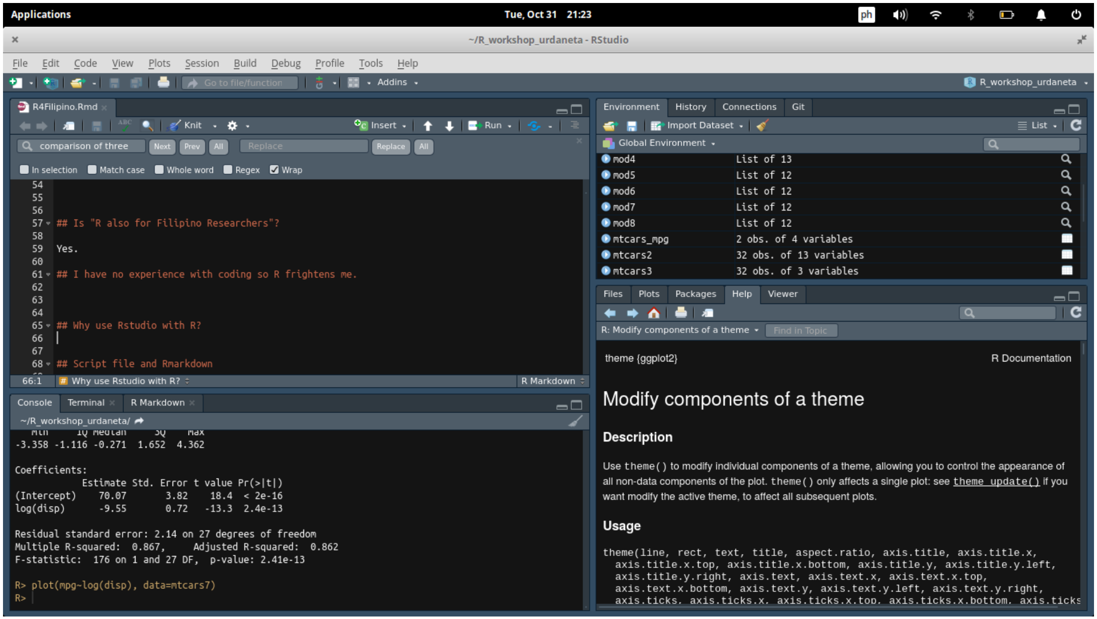
Is "R also for Filipino Researchers"?

Yes.

I have no experience with coding so R frightens me.

- You use Microsoft Excel, right?

Why use Rstudio with R?



Let's load the required packages first

```
library(tidyverse)
library(agricolae)
```

Introducing the `mtcars data set`

```
?mtcars
write.csv(mtcars, "mtcars.csv")
```

Using different ways to load data set into R

- Using the R console

```
mt1 <- read.table("mtcars.csv", sep = ",", header=TRUE)
mtcars2 <- read.csv("mtcars.csv")
mt3 <- read_csv("mtcars.csv")
mt4 <- data.table::fread("mtcars.csv")
```

- From Rstudio, click `File > Import Dataset`.

Linear Model for comparing the means of two groups

Means Model

$$y_{ij} = \mu_i + \varepsilon_{ij},\ i = 1, 2$$

- $H_0 : \mu_1 = \mu_2$
- $H_a : \mu_1 \neq \mu_2$

Linear Model for comparing the means of two groups

Effects Model

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij},\ i = 1, 2$$

where:

- $y_{ij} =$ is the jth value of `mpg` in the ith group of `am`
- $\mu =$ mean of Y
- $\alpha_i =$ effect of the ith group of `am` on `mpg`
- $\varepsilon =$ the random error due to the jth value of `mpg` in the ith value of `am`

Hypotheses

- $H_0 : \alpha_1 = \alpha_2$
- $H_a : \alpha_1 \neq \alpha_2$

A research question: Is there a difference in milleage for automatic and manual cars?

- $H_a : \alpha_1 = \alpha_2 = 0$ (or $\mu_1 = \mu_2$) The milleage per gallon differ based on transmission type of the car.
- $H_0 : \alpha_1 \neq \alpha_2$ (or $\mu_1 \neq \mu_2$) The milleage per gallon do not differ based on transmission type of the car.

A research question: Is there a difference in milleage for automatic and manual cars?

```
table(mtcars$am)
```

```
 0  1 
19 13 
```

```
aggregate(mtcars$mpg, by = list(mtcars$am), FUN="mean")
```

```
  Group.1     x
1       0 17.15
2       1 24.39
```

```
aggregate(mtcars$mpg, by = list(mtcars$am), FUN="var")
```

```
  Group.1     x
1       0 14.70
2       1 38.03
```

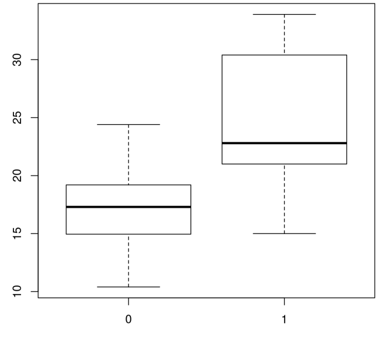
A research question: Is there a difference in milleage for automatic and manual cars?

```
mtcars_mpg <- mtcars %>%
  group_by(am) %>%
  summarise(
    mean_mpg = mean(mpg),
    var_mpg = var(mpg),
    n = n()
  )
mtcars_mpg
```

```
# A tibble: 2 x 4
     am mean_mpg var_mpg     n
  <dbl>    <dbl>   <dbl> <int>
1     0    17.15   14.70    19
2     1    24.39   38.03    13
```

Continuation of Eploratory Data Analysis

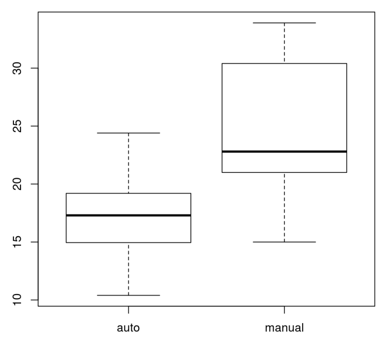
```
boxplot(mpg~am, data = mtcars)
```



Changing the labels of a plot; creating a new variable in a data set (`data.frame`)

Let us put some labels for the levels of `am`.

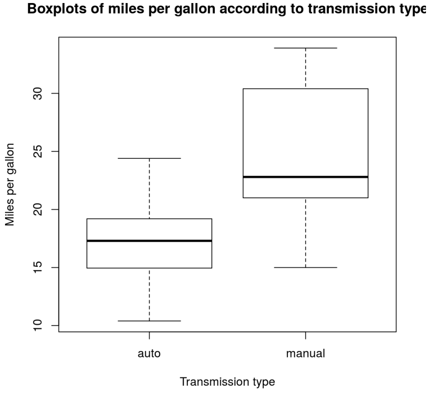
```
mtcars2$amf <- factor(mtcars2$am, levels = c(0,1), labels = c("auto", "manual"))
boxplot(mpg~amf, data=mtcars2)
```



Changing the x and y labels and putting a title

Let us put some labels for the levels of `am`.

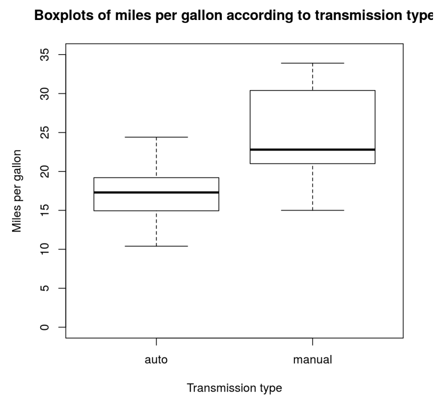
```
boxplot(mpg~amf, data=mtcars2,
        main = "Boxplots of miles per gallon according to transmission type",
        xlab = "Transmission type",
        ylab = "Miles per gallon")
```



Changing the range of values in the y-axis

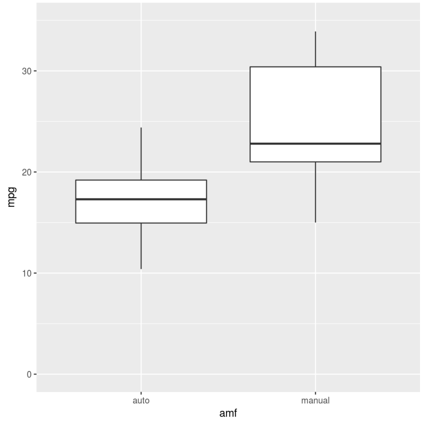
Let us start the boxplot at 0.

```
boxplot(mpg~amf, data=mtcars2,
        main = "Boxplots of miles per gallon according to transmission type",
        xlab = "Transmission type",
        ylab = "Miles per gallon",
        ylim = c(0,35))
```



Plotting with `ggplot2`

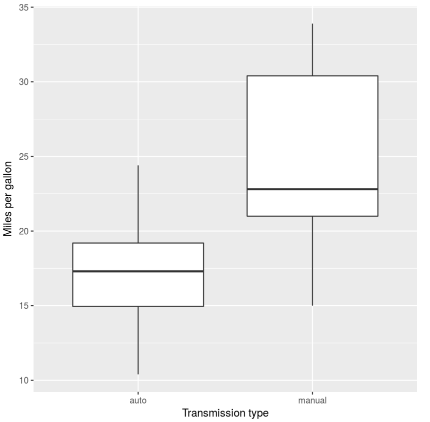
```
ggplot(mtcars2, aes(x = amf, y = mpg)) + geom_boxplot() + ylim(0,35)
```




```
# What is the difference?
ggplot(mtcars2, aes(x = amf, y = mpg)) + geom_boxplot() + coord_cartesian(ylim = c(0,35))
```

ggplot2 uses the language of graphics

```
p <- ggplot(mtcars2, aes(x = amf, y = mpg)) +
  geom_boxplot() +
  xlab("Transmission type") +
  ylab("Miles per gallon")
p
```



A review of t test

What are the assumptions of the independent samples t test?

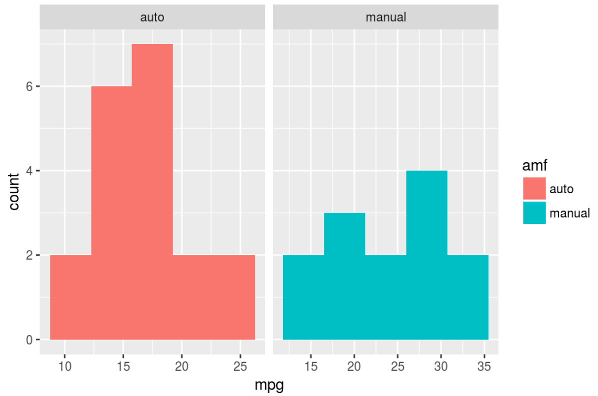
1. Dependent variable should be measured on a continuous scale (interval or ratio level)
2. Independent variable consist of two categorical, independet groups
3. Observations are independent of each other
4. No significant outliers
5. Dependent variable should be (approximately) normally distributed for each group of the independent variable
6. Variances should be homogenous

Applying the assumptions

1. What is the independent variable? What is the independent variable?
2. Is the dependent variable measured on a continuous scale?
3. Does the independent variable consist of two categorical, independent groups?
4. Are observations independent of each other?
5. Are there no significant outliers?
6. Is the dependent variable normally distributed for each group of the independent variable?
7. Are the variances homogenous?

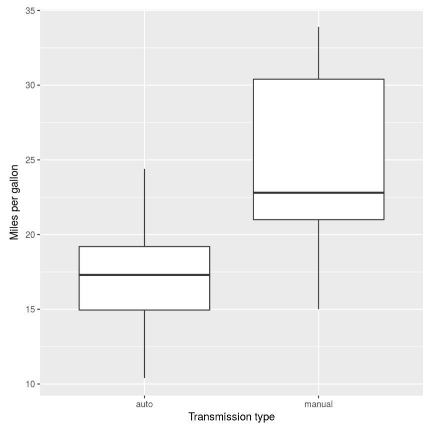
Normality for each group of the independent variable

```
ggplot(mtcars2, aes(x = mpg, fill = amf)) + geom_histogram(bins = 5) + facet_wrap(~amf, scales = "free_x")
```



Homogeneity of variance

p



Homogeneity of variance (cont…)

```
var.test(mpg~amf, data=mtcars2)
```

```
    F test to compare two variances

data:  mpg by amf
F = 0.39, num df = 18, denom df = 12, p-value = 0.07
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.1244 1.0703
sample estimates:
ratio of variances
            0.3866
```

t test results

```
(t1 <- t.test(mpg ~ amf, data = mtcars2))
```

```

    Welch Two Sample t-test

data:  mpg by amf
t = -3.8, df = 18, p-value = 0.001
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -11.28  -3.21
sample estimates:
  mean in group auto mean in group manual
               17.15                24.39
```

How to use `t.test`?

```
?t.test
```

How about independent samples t test?

```
(t2 <- t.test(mpg~amf, data=mtcars2, var.equal=TRUE))
```

```

    Two Sample t-test

data:  mpg by amf
t = -4.1, df = 30, p-value = 3e-04
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -10.848  -3.642
sample estimates:
  mean in group auto mean in group manual 
               17.15                24.39 
```

Non-parametric alternative

```
(w <- wilcox.test(mpg ~ amf, data = mtcars2, conf.int = TRUE))
```

```
Warning in wilcox.test.default(x = c(21.4, 18.7, 18.1, 14.3, 24.4, 22.8, :
cannot compute exact p-value with ties
```

```
Warning in wilcox.test.default(x = c(21.4, 18.7, 18.1, 14.3, 24.4, 22.8, :
cannot compute exact confidence intervals with ties
```

```

    Wilcoxon rank sum test with continuity correction

data:  mpg by amf
W = 42, p-value = 0.002
alternative hypothesis: true location shift is not equal to 0
95 percent confidence interval:
 -11.7  -2.9
sample estimates:
difference in location
                  -6.8
```

Confidence intervals

```
t1$conf.int
```

```
[1] -11.28  -3.21
attr(,"conf.level")
[1] 0.95
```

```
t2$conf.int
```

```
[1] -10.848  -3.642
attr(,"conf.level")
[1] 0.95
```

```
w$conf.int
```

```
[1] -11.7  -2.9
attr(,"conf.level")
[1] 0.95
```

Determining other values from the tests

```
names(t1)
```

```
[1] "statistic"   "parameter"   "p.value"     "conf.int"    "estimate"
[6] "null.value"  "alternative" "method"      "data.name"
```

Conclusion of comparison of milleage according to transmission type

The milleage per gallon differ between manual and automatic tramission-type vehicles by about 7.24 miles per gallon at 0.05 significance level (or 95% confidence level).

When to use one-tailed t test

```
?t.test
```

```
(t3 <- t.test(mpg ~ amf, data = mtcars2, alternative = "less"))
```

```

	Welch Two Sample t-test

data:  mpg by amf
t = -3.8, df = 18, p-value = 7e-04
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
   -Inf -3.913
sample estimates:
  mean in group auto mean in group manual
               17.15                24.39
```

```
t3 %>% broom::tidy()
```

```
  estimate estimate1 estimate2 statistic   p.value parameter conf.low
1   -7.245     17.15     24.39    -3.767 0.0006868     18.33     -Inf
  conf.high                  method alternative
1    -3.913 Welch Two Sample t-test        less
```

Comparison of the means of three groups

```
table(mtcars2$cyl)
```

```
 4  6  8 
11  7 14 
```

```
mtcars3 <- mtcars2 %>% select(mpg, cyl)
head(mtcars3)
```

```
   mpg cyl
1 21.0   6
2 21.0   6
3 22.8   4
4 21.4   6
5 18.7   8
6 18.1   6
```

Linear Model for the problem of comparing three means

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij},\ i = 1, 2, 3$$

where:

- $y_{ij} =$ is the jth value of `mpg` in the ith group of `cyl`
- $\mu =$ mean of Y
- $\alpha_i =$ effect of the ith group of `cyl` on `mpg`
- $\varepsilon =$ the random error due to the jth value of `mpg` in the ith value of `cyl`

Hypotheses

- $H_0 : \alpha_1 = \alpha_2 = \alpha_3 = 0$ (equivalently: $\mu_1 = \mu_2 = \mu_3$)
- $H_a : \alpha_i \neq 0$ for at least 1 i (equivalently: $\mu_a \neq \mu_b$ for at least one pair a and b)

Exploratory data analysis of mpg in terms of cyl

```
mtcars3 %>% group_by(cyl) %>% summarise(mean = mean(mpg), var = var(mpg), sd = sd(mpg), n = n())
```

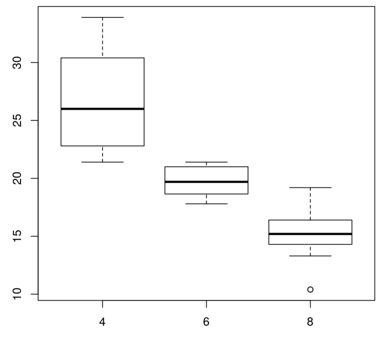
```
# A tibble: 3 x 5
    cyl  mean    var    sd     n
  <int> <dbl>  <dbl> <dbl> <int>
1     4 26.66 20.339 4.510    11
2     6 19.74  2.113 1.454     7
3     8 15.10  6.554 2.560    14
```

Hypotheses

- H_a : There are differences in mean milleage per gallon depending on number of the car's cylinders.
- H_o : There are no differences in mean milleage per gallong according to the number of the car's cylinders.

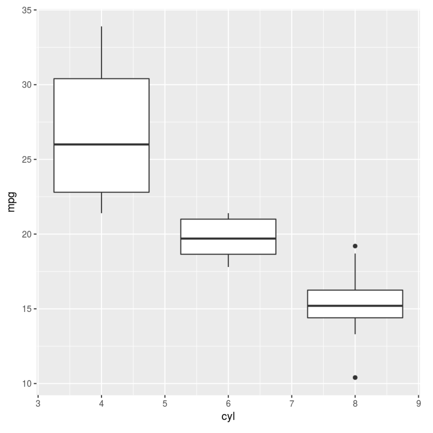
Investigation of Normality and equality of variances

```
boxplot(mpg ~ cyl, data = mtcars3)
```



Investigation of Normality and equality of variances

```
ggplot(mtcars3, aes(x=cyl, y=mpg, group=cyl)) + geom_boxplot()
```



Analysis of variance

```
mtcars3$cylf <- as.factor(mtcars3$cyl)
mod <- aov(mpg~cylf, data=mtcars3)
summary(mod)
```

```
            Df Sum Sq Mean Sq F value Pr(>F)
cylf         2    825     412    39.7  5e-09
Residuals   29    301      10
```

```
TukeyHSD(mod, "cylf")
```

```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = mpg ~ cylf, data = mtcars3)

$cylf
       diff     lwr     upr  p adj
6-4  -6.921 -10.769 -3.0722 0.0003
8-4 -11.564 -14.771 -8.3565 0.0000
8-6  -4.643  -8.328 -0.9581 0.0112
```

More on post-hoc analysis

```
with(mtcars3,
pairwise.t.test(mpg, cylf, p.adjust.method = "bonferroni")
)
```

```

    Pairwise comparisons using t tests with pooled SD

data:  mpg and cylf

  4     6
6 4e-04 -
8 3e-09 0.01

P value adjustment method: bonferroni
```

More on post-hoc analysis (cont…)

```
scheffe.test(mod, "cylf", console=TRUE)
```

```
Study: mod ~ "cylf"

Scheffe Test for mpg

Mean Square Error  : 10.39

cylf,  means

    mpg   std  r  Min  Max
4 26.66 4.510 11 21.4 33.9
6 19.74 1.454  7 17.8 21.4
8 15.10 2.560 14 10.4 19.2

Alpha: 0.05 ; DF Error: 29
Critical Value of F: 3.328

Groups according to probability of means differences and alpha level( 0.05 )

Means with the same letter are not significantly different.

    mpg groups
4 26.66      a
6 19.74      b
8 15.10      c
```

Non-parametric Kruskal-Wallis Test

```
kruskal.test(mpg~cyl, data=mtcars)
```

```

    Kruskal-Wallis rank sum test

data:  mpg by cyl
Kruskal-Wallis chi-squared = 26, df = 2, p-value = 3e-06
```

Non-parametric Kruskal-Wallis Test

```
with(mtcars, agricolae::kruskal(mpg, cyl, p.adj="BH", console = TRUE))
```

```
Study: mpg ~ cyl
Kruskal-Wallis test's
Ties or no Ties

Critical Value: 25.75
Degrees of freedom: 2
Pvalue Chisq  : 2.566e-06

cyl,  means of the ranks

     mpg  r
4 26.955 11
6 17.429  7
8  7.821 14

Post Hoc Analysis

P value adjustment method: BH
t-Student: 2.045
Alpha    : 0.05
Groups according to probability of treatment differences and alpha level.

Treatments with the same letter are not significantly different.

     mpg groups
4 26.955      a
6 17.429      b
8  7.821      c
```

Randomized Complete Block Design

Suppose we want to know the effect of the number of cylinders to mpg when we group the observations by type of transmission, which we know has an effect on mpg. That is, we want to isolate the effect of cyl on mpg when we group the observations by am.

The linear model is

$$y_{ij} = \mu + \alpha_i + \rho_j + \varepsilon_{ij}$$

where

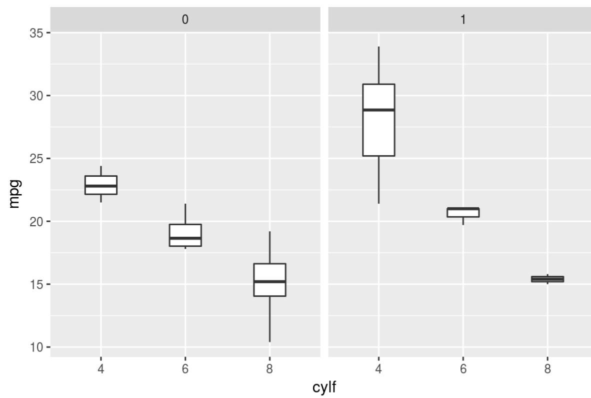
- $\mu =$ hypothesized mean
- $\rho_j =$ the effect of the jth blocking factor (`am`) to `mpg`
- $\alpha_i =$ effect of the the ith `cyl` to `mpg`
- $\varepsilon_{ij} =$ the random effect on the ij-th observation

Then the hypotheses are

- $H_0 : \alpha_i = 0$. The number of cylinders has no effect on milleage per gallon.
- $H_a : \alpha_i \neq 0$. The number of cylinders affect milleage per gallon.

Exploratory Data Analysis for RCBD

```
mtcars4 <- mtcars2 %>% select(mpg, am, cyl)
mtcars4$cylf <- as.factor(mtcars$cyl)
mtcars4$amf <- as.factor(mtcars$am)
ggplot(mtcars4, aes(x = cylf, y = mpg)) +
  geom_boxplot() +
  facet_wrap(~amf)
```



RCBD

```
mod2 <- aov(mpg ~ amf + cylf, data = mtcars4)
summary(mod2)
```

```
            Df Sum Sq Mean Sq F value  Pr(>F)
amf          1    405     405    42.9 4.2e-07
cylf         2    456     228    24.2 8.0e-07
Residuals   28    264       9
```

```
TukeyHSD(mod2, "cylf")
```

```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = mpg ~ amf + cylf, data = mtcars4)

$cylf
      diff     lwr     upr  p adj
6-4 -4.757  -8.434 -1.0798 0.0092
8-4 -7.330 -10.394 -4.2655 0.0000
8-6 -2.573  -6.093  0.9475 0.1853
```

Two-way ANOVA (Two-factor CBD)

What if prior to the experiment, we don't know the effect of any of `am` and `cyl` on `mpg`? We want to see how these factors affect `mpg` and whether they affect `mpg` independently or not.

- Model: $y_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ij}$

where

- $\mu =$ hypothesized mean
- $\alpha_i =$ the effect of the ith type of transmission (`am`) to `mpg`
- $\beta_j =$ effect of the the jth number of cylinders (`cyl`) to `mpg`
- $\gamma_{ij} =$ the interaction effect of the ij-th type of transmission and number of cylinders
- $\varepsilon_{ij} =$ the random effect on the ij-th observation

Three pairs of hypotheses for Two-way ANOVA

There are three pairs of hypotheses to be tested:

1. Interaction effects

- $H_0 : \gamma_{ij} = 0$. There is no interaction between `am` and `cyl`.
- $H_a : \gamma_{ij} \neq 0$. There is an interaction between `am` and `cyl`.

1. Effect of `am`

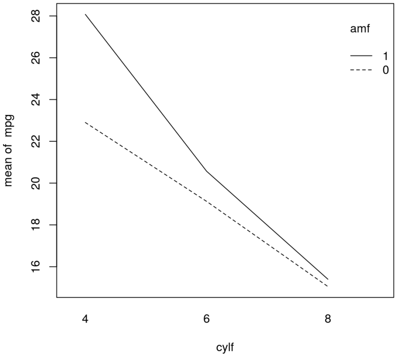
- $H_0 : \alpha_i = 0$. Controlling for other variables, `am` has no effect on `mpg`.
- $H_a : \alpha_i \neq 0$. Controlling for other variables, `am` has an effect on `mpg`.

1. Effect of `cyl`

- $H_0 : \beta_j = 0$. Controlling for other variables, `cyl` has no effect on `mpg`.
- $H_a : \beta_j \neq 0$. Controlling for other variables, `cyl` has an effect on `mpg`.

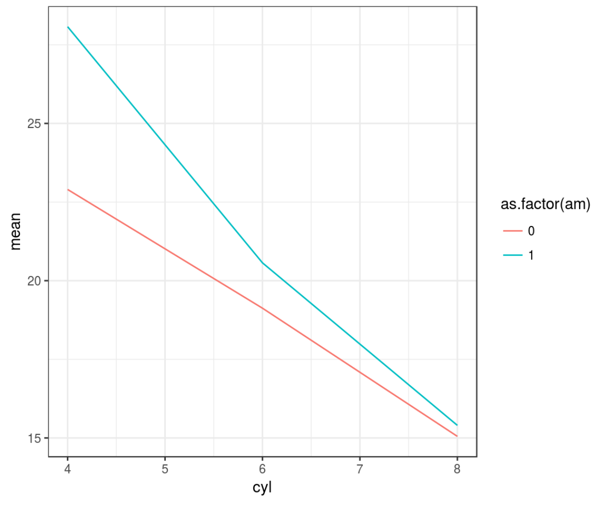
Interaction plot

```
with(mtcars4, interaction.plot(cylf, amf, mpg, fun = mean))
```



Interaction plot with `ggplot2`

Challenge: Create an interaction plot using `ggplot2`.



How to do two-way ANOVA in R

```
mod4 <- aov(mpg ~ amf * cylf, data = mtcars4)
summary(mod4)
```

```
            Df Sum Sq Mean Sq F value  Pr(>F)
amf          1    405     405   44.06 4.8e-07
cylf         2    456     228   24.82 9.4e-07
amf:cylf     2     25      13    1.38    0.27
Residuals   26    239       9
```

Conclusions from two-way ANOVA

1. There is no interaction between `am` and `cyl`.
2. `cyl` affects `mpg`.
3. `am` affects `mpg`.

Post hoc analyses for two-way ANOVA

```
TukeyHSD(mod4, "cylf")
```

```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = mpg ~ amf * cylf, data = mtcars4)

$cylf
      diff     lwr     upr  p adj
6-4 -4.757  -8.400 -1.1137 0.0088
8-4 -7.330 -10.365 -4.2937 0.0000
8-6 -2.573  -6.061  0.9151 0.1788
```

```
TukeyHSD(mod4, "amf")
```

```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = mpg ~ amf * cylf, data = mtcars4)

$amf
     diff   lwr   upr p adj
1-0 7.245 5.001 9.488     0
```

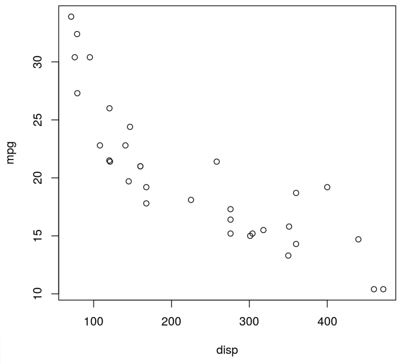
Finding relationships

Try any of the following codes to plot `mpg` against `disp` in the `mtcars` package.

```
plot(mpg~disp, data = mtcars)
```

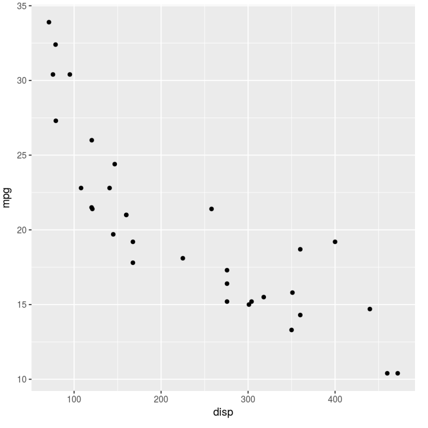
```
plot(mtcars$disp, mtcars$mpg)
```

```
with(mtcars, plot(disp, mpg))
```



Scatterplot with `ggplot2`

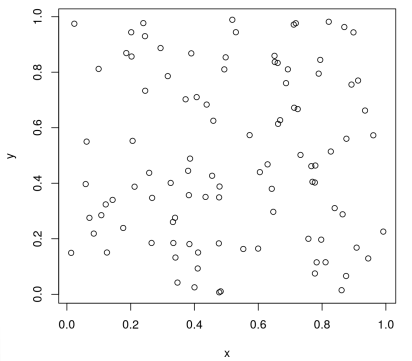
```
ggplot(mtcars, aes(x = disp, y = mpg)) + geom_point()
```



Correlation between **disp** and **mpg**

- Research problem: What is the relationship between `disp` and `mpg`?
- More specific research problem: Is there a linear relationship between `disp` and `mpg`?
- How does a scatterplot of no relationship between two variables look like?

```
set.seed(1); x = runif(100)
set.seed(2); y = runif(100)
plot(x,y)
```




```
with(mtcars, cor(disp, mpg))
```

Remember, correlation does not imply causation

But in controlled experiments where you test the variation in the dependent variable by manipulating the values of the independent variable, you can investigate causation.

Suppose we want to investigate whether `disp` has an effect on `mpg`.

The model is a linear regression of `mpg` on `disp`:

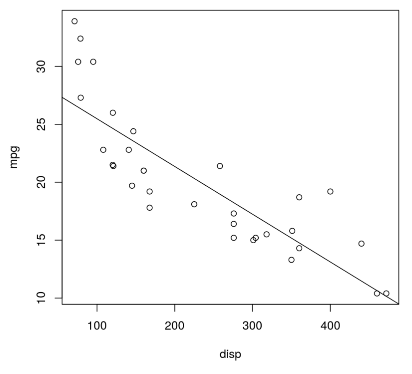
$$y = \beta_0 + \beta x + \varepsilon$$

where

- $y =$ `mpg`
- $x =$ `disp`
- $\beta_0 =$ intercept
- $\beta =$ the increase in `mpg` for every 1 unit increase in `disp`
- $\varepsilon =$ random error

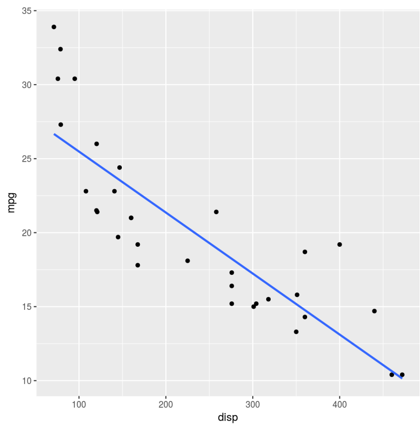
Plotting the line of best fit

```
mod5 <- lm(mpg~disp, data = mtcars)
with(mtcars, plot(disp, mpg))
abline(mod5)
```



Plotting the line of best fit with ggplot2

```
ggplot(mtcars, aes(disp, mpg)) + geom_point() + geom_smooth(method="lm", se=FALSE)
```



Testing the linear fit

```
summary(mod5)
```

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

Residuals:
   Min     1Q Median     3Q    Max 
-4.892 -2.202 -0.963  1.627  7.231 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 29.59985    1.22972   24.07  < 2e-16
disp        -0.04122    0.00471   -8.75  9.4e-10

Residual standard error: 3.25 on 30 degrees of freedom
Multiple R-squared:  0.718, Adjusted R-squared:  0.709 
F-statistic: 76.5 on 1 and 30 DF,  p-value: 9.38e-10
```

Results

We have the following results from this output:

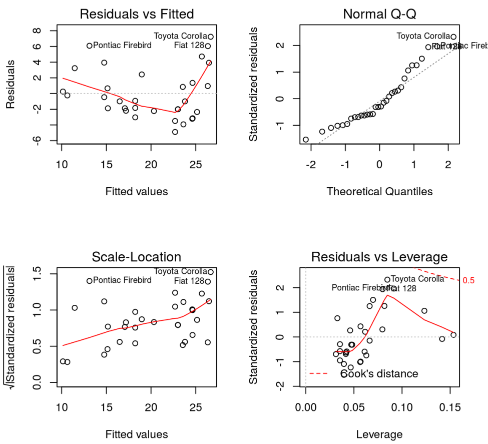
- The line of best fit has an equation: $y = 29.60 - 0.0412x$.
- `disp` has an affect on `mpg` at the .05 significance level ($p = 9.38 \times 10^{-10}$)
- `disp` explains about 72% of the variation in `mpg`

Four Principal Assumptions of linear regression

- **Linearity and additivity** of the relationship between dependent and independent variables / **Linearity of residuals**
- **Statistical independence** of the errors/residuals
- **Homoscedasticity** (equal variance) of the errors/residuals
- **Normality** of errors/residuals

Testing the linear fit

```
par(mfrow=c(2,2))
plot(mod5)
```

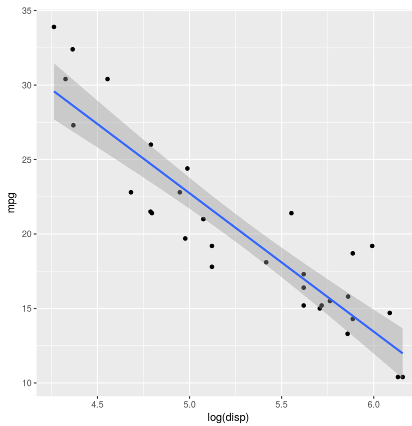


Interpreting the diagnostic plots

- The **residuals vs fitted** plot shows if residuals have non-linear patterns. This plot should show equally spread residuals around a horizontal line without distinct pattern.
- The **Normal Q-Q Plot** shows if the residuals are normally distributed. The residuals should follow a straight line well.
- The **Scale-Location Plot** shows if residuals are spread equally along the ranges of predictors. This plot can be used to check the assumption of equal variance (homoscedasticity). It should show a horizontal line with equally (randomly) spread points.
- The **Residuals vs Leverage Plot** helps us find influential cases (or subjects/observations) if any. There should be no points outside the dashed lines (or Cook's distance)

Transformations

```
ggplot(mtcars, aes(log(disp), mpg)) + geom_point() + geom_smooth(method="lm")
```



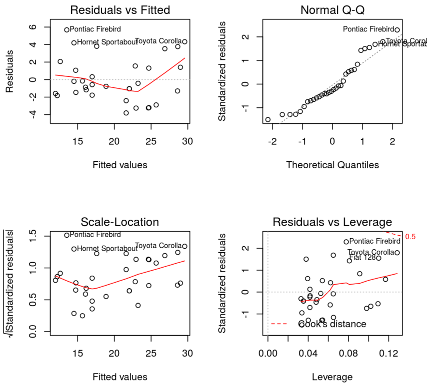
Regression with log transformation

```
mod6 <- lm(mpg~log(disp), data=mtcars)
summary(mod6)
```

```

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

Residuals:
   Min     1Q Median     3Q    Max 
-3.808 -1.634 -0.675  1.443  5.676 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   69.205      4.185    16.5  < 2e-16
log(disp)     -9.293      0.787   -11.8  8.4e-13

Residual standard error: 2.58 on 30 degrees of freedom
Multiple R-squared:  0.823, Adjusted R-squared:  0.817 
F-statistic:  139 on 1 and 30 DF,  p-value: 8.4e-13
```

Diagnostic plots of **mod6**

```
par(mfrow=c(2,2))
plot(mod6)
```

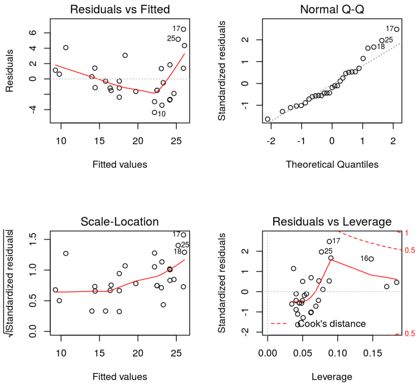


Interpretation of `mod6`

- The linear fit improved as the log of `disp` now explains about 82% of the variation in `mpg`.
- However, how do we now interpret the results?

Removing the influential observations

```
mtcars7 <- mtcars %>% filter(!rownames(.) %in% c("Pontiac Firebird",
"Toyota Corolla",
"Hornet Sportabout"))
mod7 <- lm(mpg~disp, data = mtcars7)
par(mfrow=c(2,2))
plot(mod7)
```



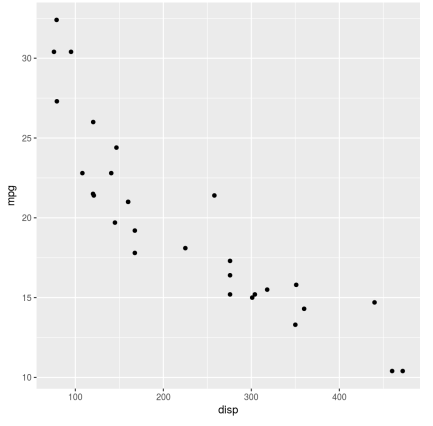
Effect of removing influential observations

```
summary(mod7)
```

```

Call:
lm(formula = mpg ~ disp, data = mtcars7)

Residuals:
   Min     1Q Median     3Q    Max 
 -4.37  -1.63  -0.50   1.38   6.47 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 29.26606    1.09492   26.73  < 2e-16
disp        -0.04237    0.00429   -9.87  1.9e-10

Residual standard error: 2.73 on 27 degrees of freedom
Multiple R-squared:  0.783, Adjusted R-squared:  0.775 
F-statistic: 97.3 on 1 and 27 DF,  p-value: 1.9e-10
```

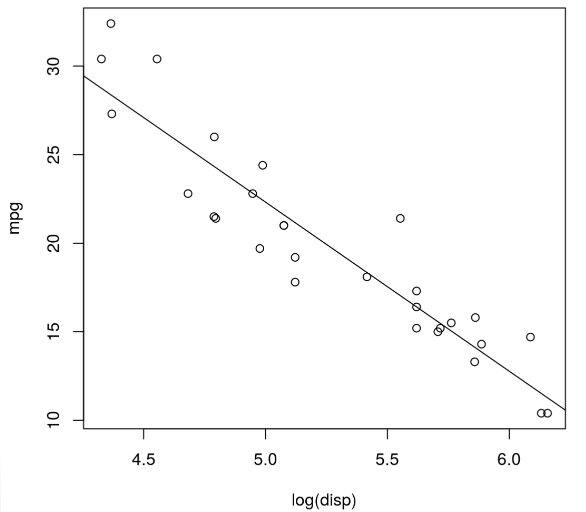
Plot without influential observations

```
ggplot(mtcars7, aes(disp, mpg)) + geom_point()
```



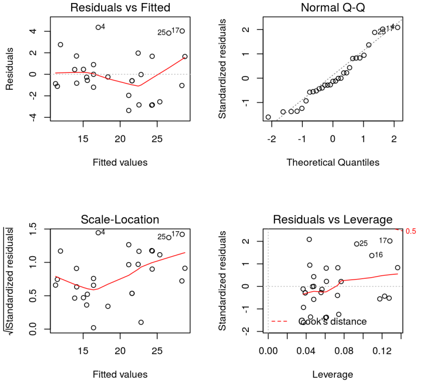
Trying transformations

```
mod8 <- lm(mpg~log(disp), data=mtcars7)
plot(mpg~log(disp), data=mtcars7)
abline(mod8)
```



Diagnostic plots of **mod8**

```
par(mfrow=c(2,2))
plot(mod8)
```



Fit of **mod8**

```
summary(mod8)
```

```

Call:
lm(formula = mpg ~ log(disp), data = mtcars7)

Residuals:
   Min     1Q Median     3Q    Max 
-3.358 -1.116 -0.271  1.652  4.362 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    70.07       3.82    18.4  < 2e-16
log(disp)      -9.55       0.72   -13.3  2.4e-13

Residual standard error: 2.14 on 27 degrees of freedom
Multiple R-squared:  0.867, Adjusted R-squared:  0.862 
F-statistic:  176 on 1 and 27 DF,  p-value: 2.41e-13
```

Multiple regression

```
head(mtcars)
```

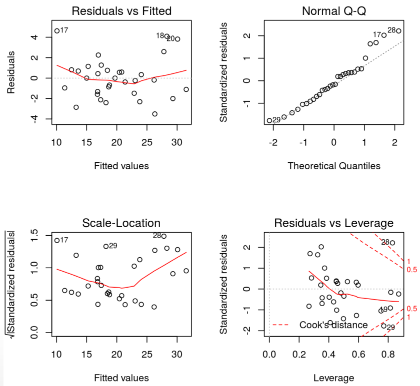
```
                   mpg cyl disp  hp drat    wt  qsec vs am gear carb
Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
Hornet 4 Drive    21.4   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
```

```
mtcars5 <- mtcars %>%
  mutate(
    cyl = as.factor(cyl),
    vs = as.factor(vs),
    am = as.factor(am),
    gear = as.factor(gear),
    carb = as.factor(carb)
  )
```

Continuation of Multiple Regression

```
mod7 <- lm(mpg~., data=mtcars5)
par(mfrow=c(2,2))
plot(mod7)
```

```
Warning: not plotting observations with leverage one:
  30, 31

Warning: not plotting observations with leverage one:
  30, 31
```



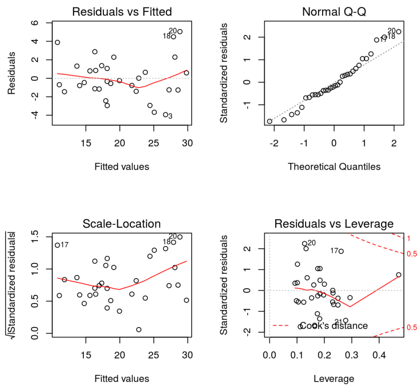
Step-wise regression

```
mod8 <- step(mod7, direction="both", trace=FALSE)
summary(mod8)
```

```
Call:
lm(formula = mpg ~ cyl + hp + wt + am, data = mtcars5)

Residuals:
   Min     1Q Median     3Q    Max 
-3.939 -1.256 -0.401  1.125  5.051 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  33.7083     2.6049   12.94  7.7e-13
cyl6         -3.0313     1.4073   -2.15   0.0407
cyl8         -2.1637     2.2843   -0.95   0.3523
hp           -0.0321     0.0137   -2.35   0.0269
wt           -2.4968     0.8856   -2.82   0.0091
am1           1.8092     1.3963    1.30   0.2065

Residual standard error: 2.41 on 26 degrees of freedom
Multiple R-squared:  0.866, Adjusted R-squared:  0.84 
F-statistic: 33.6 on 5 and 26 DF,  p-value: 1.51e-10
```

Diagnosic plots of result of step-wise regression

```
par(mfrow=c(2,2))
plot(mod8)
```



Prediction (for Demonstration Only)

https://stats.stackexchange.com/questions/244017/prediction-vs-inference

- Inference: Given a set of data you want to infer how the output is generated as a function of the data.
- Prediction: Given a new measurement, you want to use an existing data set to build a model that reliably chooses the correct identifier from a set of outcomes.
- Inference: You want to find out what the effect of Age, Passenger Class and, Gender has on surviving the Titanic Disaster. You can put up a logistic regression and infer the effect each passenger characteristic has on survival rates.
- Prediction: Given some information on a Titanic passenger, you want to choose from the set {lives,dies} and be correct as often as possible. (See bias-variance tradeoff for prediction in case you wonder how to be correct as often as possible.)

Let us use **mod5** for prediction

```
newdata <- data.frame(disp = sample(mtcars$disp,5) + rnorm(5))
predict(mod5, newdata, interval="prediction")
```

```
    fit    lwr   upr
1 14.75  7.895 21.61
2 18.24 11.480 24.99
3 23.67 16.875 30.47
4 13.09  6.149 20.03
5 23.79 16.991 30.59
```

For More on Prediction and Machine Learning

- https://www.datacamp.com/community/tutorials/machine-learning-in-r
- https://machinelearningmastery.com/machine-learning-in-r-step-by-step/
- https://www.coursera.org/learn/practical-machine-learning
- https://www.kaggle.com
- https://www.kdnuggets.com/2017/04/10-free-must-read-books-machine-learning-data-science.html

Challenge: multiple linear regression

Using the `diamonds` data set, create a model for pricing diamonds based on the other variables.

```
?diamonds
head(diamonds)
```

Jump start your self-learning of the R statistical package

```
install.packages("swirl")
library(swirl)
swirl()
```

Thank you!

```
library(ggplot2)
dat <- data.frame(x=seq(0, 2*pi, length.out=100))
shape <- function(x)2-2*sin(x) + sin(x)*(sqrt(abs(cos(x))))/(sin(x)+1.4)
ggplot(dat, aes(x=x)) + stat_function(fun=shape) + coord_polar(start=-pi/2)
```

References

Bhalla, Deepanshu. 2017. "List of Companies Using R." Data Science Central. https://www.datasciencecentral.com/profiles/blogs/list-of-companies-using-r.

Muenchen, Robert A. 2016. "R Passes SAS in Scholarly Use (finally)." http://r4stats.com/2016/06/08/r-passes-sas-in-scholarly-use-finally/.

———. 2017. "The Popularity of Data Science Software." Accessed January 1. http://r4stats.com/articles/popularity/.

Piatetsky, Gregory. 2016. "R, Python Duel As Top Analytics, Data Science software–KDnuggets 2016 Software Poll Results." https://www.kdnuggets.com/2016/06/r-python-top-analytics-data-mining-data-science-software.html.

R Foundation. 2017. "What Is R?" Accessed October 31. https://www.r-project.org/about.html.