Further Statistical Analysis using R

Mark Dunning, Matthew Eldridge and Sarah Vowler

Last modified: 20 Oct 2015

Contents

1 Introduction

In this practical, we will use several 'real-life' datasets to demonstrate some of the concepts you have seen in the lectures. We will guide you through how to analyse these datasets in R and the kinds of questions you should be asking yourself when faced with similar data. As this is not a course in using R, we will provide the R code that you will need. However, it is up to you to think carefully about the assumptions of a statistics tests and interpret the results that R gives.

The data you need for this practical are available as a `https://rawgit.com/bioinformatics-core-shared-training/intermediate-stats/master/CourseData.zip`. Please download this zip file and extract to a directory on your laptop. You will then need to tell RStudio where to find these files by setting your *working directory*

Session -> Set Working Directory -> Choose Directory

2 One-Way ANOVA

The data for this exercise are to be found in `amess.csv`. The data are the red cell folate levels in three groups of cardiac bypass patients given different levels of nitrous oxide (N_2O) and oxygen (O_2) ventilation. [There is a reference to the source of this data in Altman, Practical Statistics for Medical Research, p. 208.]

The treatments are

i) 50% N2O and 50% O2 continuously for 24 hours
ii) 50% N2O and 50% O2 during the operation
iii) No N2O but 35-50% O2 continuously for 24 hours

1. Import the file `amess.csv` into R. Verify that the dimensions of the object that R creates are correct.

```
amess <- read.csv("amess.csv")
dim(amess)
```

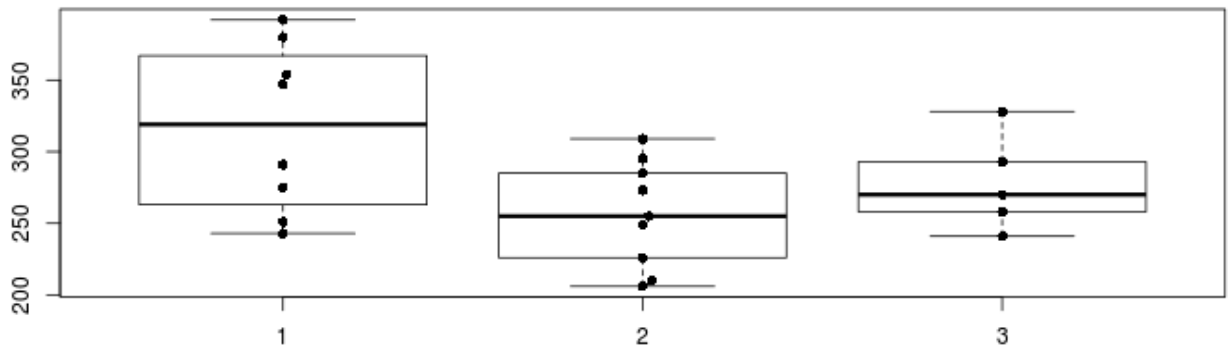
```
## [1] 22  2
```

NOTE: You can use the `View` function in RStudio to bring-up a display of the object you have created

2. Draw boxplots of the data. Does it look as though the assumptions for ANOVA are satisfied?

NOTE: It is often useful to overlay individual points on the boxplot

```
library(beeswarm)
boxplot(folate~treatmnt,data=amess)
beeswarm(folate~treatmnt,data=amess,add=TRUE,pch=16)
```



3. Perform t-tests for each of the three pair-wise comparisons. Make note of the t-test statistics obtained. Which groups, if any, differ from one another?

NOTE:If you have not performed a t test in R before, you may wish to consult the help page for `t.test`; `?t.test`.

```
t.test(folate~ treatmnt,data=amess[amess$treatmnt !=3,],var.equal=TRUE)
```

```
##
##  Two Sample t-test
##
## data:  folate by treatmnt
## t = 2.5582, df = 15, p-value = 0.02184
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   10.03871 110.32240
## sample estimates:
## mean in group 1 mean in group 2
##        316.6250        256.4444
```

```
t.test(folate~treatmnt,data=amess[amess$treatmnt !=2,],var.equal=TRUE)
```

```
##
##  Two Sample t-test
##
## data:  folate by treatmnt
## t = 1.3266, df = 11, p-value = 0.2115
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -25.45786 102.70786
## sample estimates:
## mean in group 1 mean in group 3
##         316.625         278.000
```

```
t.test(folate~treatmnt,data=amess[amess$treatmnt !=1,],var.equal=TRUE)
```

```
##
##  Two Sample t-test
##
## data:  folate by treatmnt
## t = -1.0725, df = 12, p-value = 0.3046
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -65.34826  22.23715
## sample estimates:
## mean in group 2 mean in group 3
##        256.4444        278.0000
```

4. Perform an analysis of variance on the data. Why is this a more valid analysis? Interpret the output.

```
mod <- aov(folate~factor(treatmnt),data=amess)
mod
```

```
## Call:
##    aov(formula = folate ~ factor(treatmnt), data = amess)
##
## Terms:
##                 factor(treatmnt) Residuals
## Sum of Squares          15515.77  39716.10
## Deg. of Freedom                2        19
##
## Residual standard error: 45.72003
## Estimated effects may be unbalanced
```

```
summary(aov(mod))
```

```
##                  Df Sum Sq Mean Sq F value Pr(>F)
## factor(treatmnt)  2  15516    7758   3.711 0.0436 *
## Residuals        19  39716    2090
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

5. Choose any pair of treatments and repeat the ANOVA for just these two treatments. Compare the test statistic value to the results of the t-test from question 3. What do you notice?

```
subset <- amess[amess$treatmnt != 3,]
summary(aov(folate~treatmnt,data=subset))
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## treatmnt     1  15339   15339   6.544 0.0218 *
## Residuals   15  35158    2344
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

6. Use a *post-hoc* test to compute p-value for all pairwise contrasts. Unlike the t-tests performed in Question 3, the p-values returned here are adjusted for multiple testing. Two functions for doing such tests in R are `TukeyHSD` and `pairwise.t.test`. Help for both these functions is available through RStudio; `?TukeyHSD`, `?pairwise.t.test`

```
TukeyHSD(mod)
```

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

```
##
## Fit: aov(formula = folate ~ factor(treatmnt), data = amess)
##
## $`factor(treatmnt)`
##          diff        lwr      upr     p adj
## 2-1 -60.18056 -116.61904 -3.74207 0.0354792
## 3-1 -38.62500 -104.84037 27.59037 0.3214767
## 3-2  21.55556  -43.22951 86.34062 0.6802018
```

```
pairwise.t.test(amess$folate,amess$treatmnt)
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  amess$folate and amess$treatmnt
##
##   1     2
## 2 0.042 -
## 3 0.310 0.408
##
## P value adjustment method: holm
```

3 Non-Parametric tests

```
amess <- read.csv("amess.csv")
```

1. If the assumptions for ANOVA don't hold, then there are non-parametric alternatives available.
   - If the assumption of normality doesn't hold then we might look at the *Kruskal-Wallis* test.
   - If the assumption of equal variances doesn't hold then we might use the *median test*.

Perform these tests on the `amess` dataset. Do they give the same answer as ANOVA? If not, why not? How important were the assumptions?

```
kruskal.test(folate~factor(treatmnt),data=amess)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  folate by factor(treatmnt)
## Kruskal-Wallis chi-squared = 4.1852, df = 2, p-value = 0.1234
```

```
library(RVAideMemoire)
mood.medtest(folate~factor(treatmnt),data=amess)
```

```
##
##  Mood's median test
##
## data:  folate by factor(treatmnt)
## p-value = 0.2332
```

The file `treatments.txt` records measurements from five subjects undergoing a treatment course. The data were taken from an example in *GraphPad Prism*.

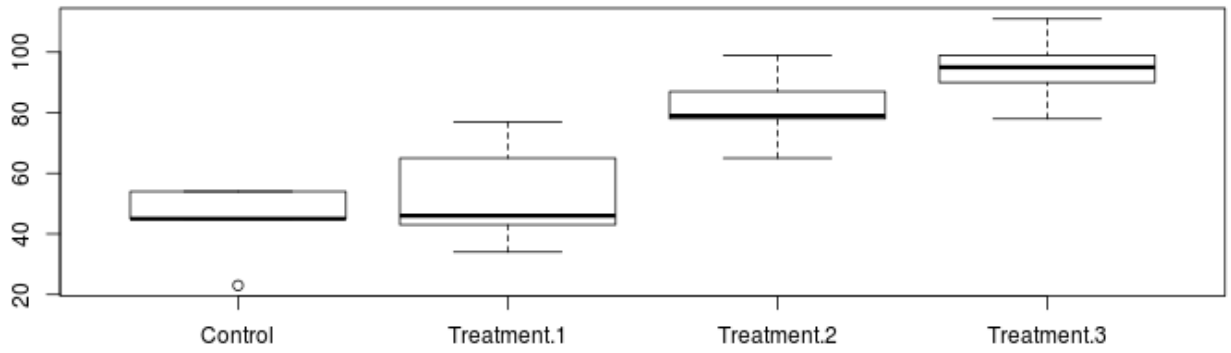
2. Import these data into R and visualise the distribution of the data

```
treatments <- read.delim("treatments.txt")
head(treatments)
```

```
##    X Control Treatment.1 Treatment.2 Treatment.3
## 1 GS      54          43          78         111
## 2 JM      23          34          65          99
## 3 HM      45          65          99          78
## 4 DR      54          77          79          90
## 5 PS      45          46          87          95
```

3. Use the Friedman test to assess the effect of treatment and interpret the result. Can you justify why this test was chosen?

```
boxplot(treatments[,-1])
```




```
friedman.test(as.matrix(treatments[,-1]))
```

```
##
##  Friedman rank sum test
##
## data:  as.matrix(treatments[, -1])
## Friedman chi-squared = 13.08, df = 3, p-value = 0.004467
```

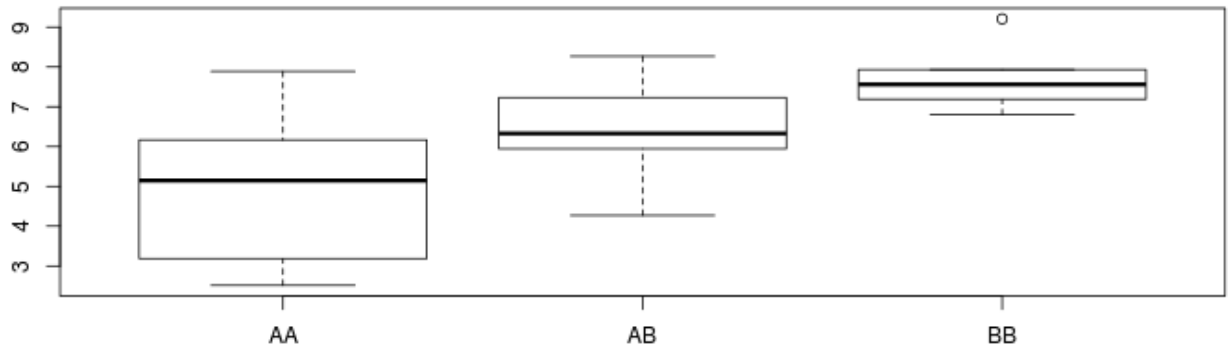
The file genotypes.txt contains measurements from patients in a gene expression study. For each patient in the study, their genotype for a particular gene was recorded.

4. Import the file genotypes.txt into R and visualise the data using a boxplot

```
gtypes <- read.delim("genotypes.txt")
gtypes
```

```
##         AA       AB       BB
## 1 2.513047 6.328862       NA
## 2 6.168767 5.607576 7.639488
## 3 3.184589 8.269598 6.795799
## 4 7.889960 4.271390 7.188640
## 5 5.146395 6.282917 7.482058
## 6       NA 7.274772 7.934725
## 7       NA 7.184816 9.208339
```

```
boxplot(gtypes)
```



NOTE: The boxplot function is also able to accept a data frame as input. In this case, it will plot the distribution of data in each column separately

5. Transform the data into two columns; one to indicate the genotype of the individual and the second for the gene expression level.

```
library(tidyr)
gtypes <- gather(gtypes)
gtypes$key <- rep(c(1,2,3),each=7)
```

6. Use an appropriate method to test the association between gene expression and genotype

```
library(clinfun)
jonckheere.test(gtypes$value,gtypes$key)
```

```
##
##  Jonckheere-Terpstra test
##
## data:
## JT = 86, p-value = 0.007535
## alternative hypothesis: two.sided
```

4 Breast Cancer Incidence

The file `globalBreastCancerRisk.csv` gives the number of new cases of Breast Cancer (per population of 10,000) in various countries around the world, along with various health and lifestyle risk factors. These data were collected from the gapminder resource [1]. Let's suppose we are initially interested in whether the number of breast cancer cases is significantly different in different regions of the world.

1. Read these data into R

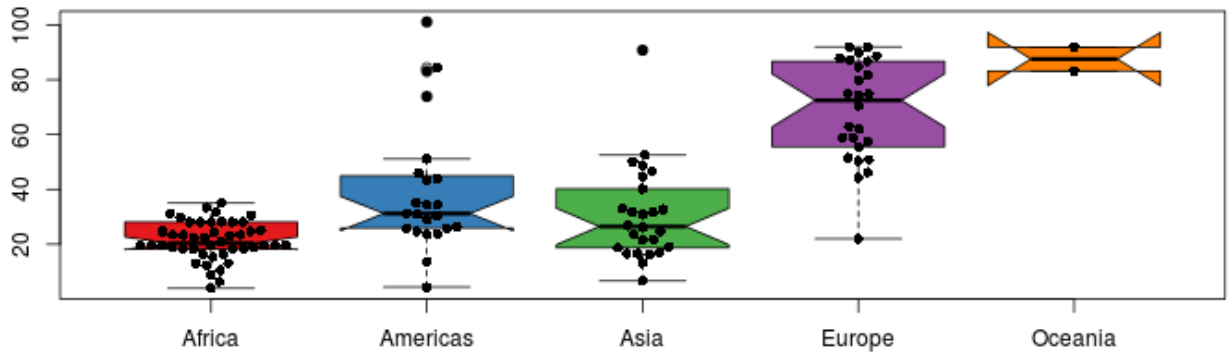
```
bcStats <- read.csv("globalBreastCancerRisk.csv")
head(bcStats)
```

[1] Similar data are also detailed in a blog post at: http://www.analyticsforfun.com/2014/06/performing-anova-test-in-r-results-and.html

```
##       country continent year lifeExp      pop  gdpPercap
## 1 Afghanistan      Asia 2002  42.129 25268405   726.7341
## 2     Albania    Europe 2002  75.651  3508512  4604.2117
## 3     Algeria    Africa 2002  70.994 31287142  5288.0404
## 4      Angola    Africa 2002  41.003 10866106  2773.2873
## 5   Argentina  Americas 2002  74.340 38331121  8797.6407
## 6   Australia   Oceania 2002  80.370 19546792 30687.7547
##   NewCasesOfBreastCancerIn2002 AlcoholComsumption BloodPressure
## 1                         26.8               0.02      124.2085
## 2                         57.4               6.68      129.0609
## 3                         23.5               0.96      130.4024
## 4                         23.1               5.40      129.9282
## 5                         73.9              10.00      119.6538
## 6                         83.2              10.02      120.5113
##   BodyMassIndex Cholestorol Smoking
## 1      20.65274    4.295170      NA
## 2      25.27082    4.918646     4.0
## 3      25.69948    4.848951     0.3
## 4      22.26093    4.499115      NA
## 5      26.70460    5.143871    25.4
## 6      26.25957    5.326858    21.8
```

2. Visualise the distribution of breast cancer incidence in each continent. Include a way of identifying how many observations belong to each group (continent). Are there any groups that you would consider removing?

```
library(RColorBrewer)
library(beeswarm)
boxplot(NewCasesOfBreastCancerIn2002~continent,data=bcStats,col=brewer.pal(5,"Set1"),notch=TRUE)
beeswarm(NewCasesOfBreastCancerIn2002~continent,data=bcStats,add=TRUE,pch=16)
```




```
bcStats.clean <- bcStats[bcStats$continent != "Oceania",]
```

3. Would a parameteric, or non-parametric, approach be suitable for this analysis? Use the result from a `bartlett.test` to support you answer. Proceed with your chosen approach to test the hypothesis that breast cancer incidence is different across the globe

```
bartlett.test(NewCasesOfBreastCancerIn2002~continent,data=bcStats.clean)
```

```
##
```

```
##  Bartlett test of homogeneity of variances
##
## data:  NewCasesOfBreastCancerIn2002 by continent
## Bartlett's K-squared = 48.26, df = 3, p-value = 1.875e-10
```

```
kruskal.test(NewCasesOfBreastCancerIn2002~continent,data=bcStats.clean)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  NewCasesOfBreastCancerIn2002 by continent
## Kruskal-Wallis chi-squared = 59.896, df = 3, p-value = 6.185e-13
```

4. In a manner similar to the *One-Way ANOVA* example, we can apply multiple-testing correction to all pairwise contrasts and assess which individual contrasts are significant. Which pairs of continents have significantly different breast cancer incidence rates to each other?

```
pairwise.wilcox.test(bcStats.clean$NewCasesOfBreastCancerIn2002,bcStats.clean$continent)
```

```
##
##  Pairwise comparisons using Wilcoxon rank sum test
##
## data:  bcStats.clean$NewCasesOfBreastCancerIn2002 and bcStats.clean$continent
##
##          Africa  Americas Asia
## Americas 5.4e-05 -        -
## Asia     0.04232 0.14359  -
## Europe   8.8e-11 0.00016  7.9e-07
##
## P value adjustment method: holm
```

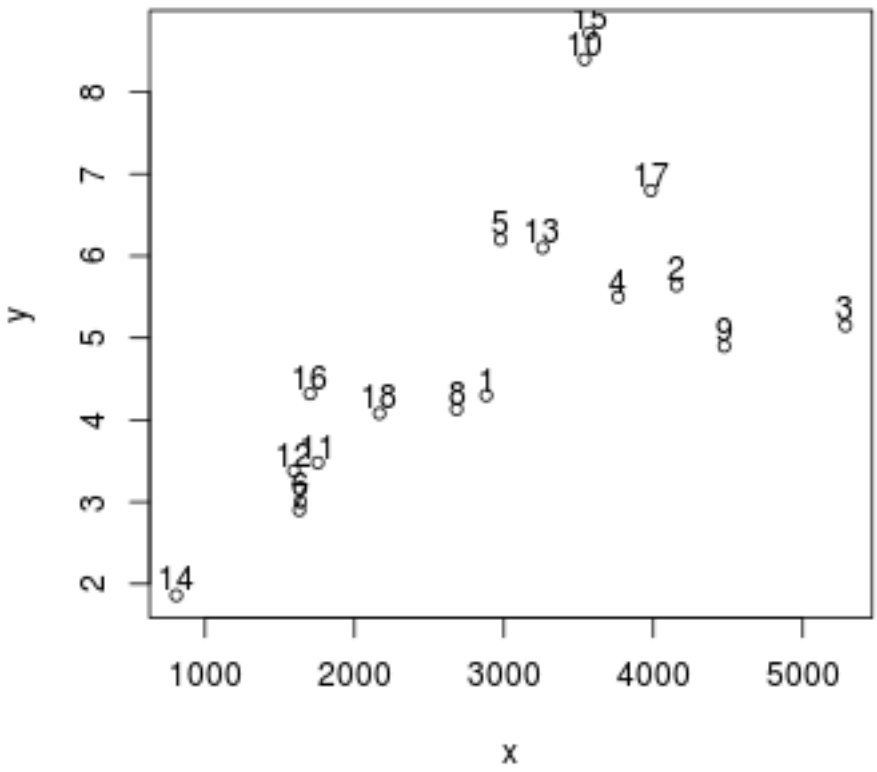
5 Linear Regression

1. Clinical Trial Data: A researcher wished to study the effect of a particular drug X on the immune system. Thus, the concentration of the drug was measured on various days, along with the white blood cell counts.
   a) Import the clinical data from the file `ClinicalTrials.txt`
   b) Make a scatter plot
   c) Label each point according to the date that the measurement was made
   d) Fit a linear model to the data and produce the diagnostic plots. Which observations fit the model least-well? Compare your answers to the scatter plot
   e) Use the output of the linear model to express the equation of the line of best-fit in the form $y = ax + b$
   f) Overlay the line-of-best fit on the scatter plot. Indicate the differences between each fitted and observed value.
   g) Extract the R^2 value from the model summary, and print this value on the plot

```
trials <- read.delim("clinicalTrials.txt")

x <- trials$Drug.concentration
y <- trials$Cell.Count

plot(x,y,cex=0.8)
text(x,y+0.2,labels=1:length(x))
```



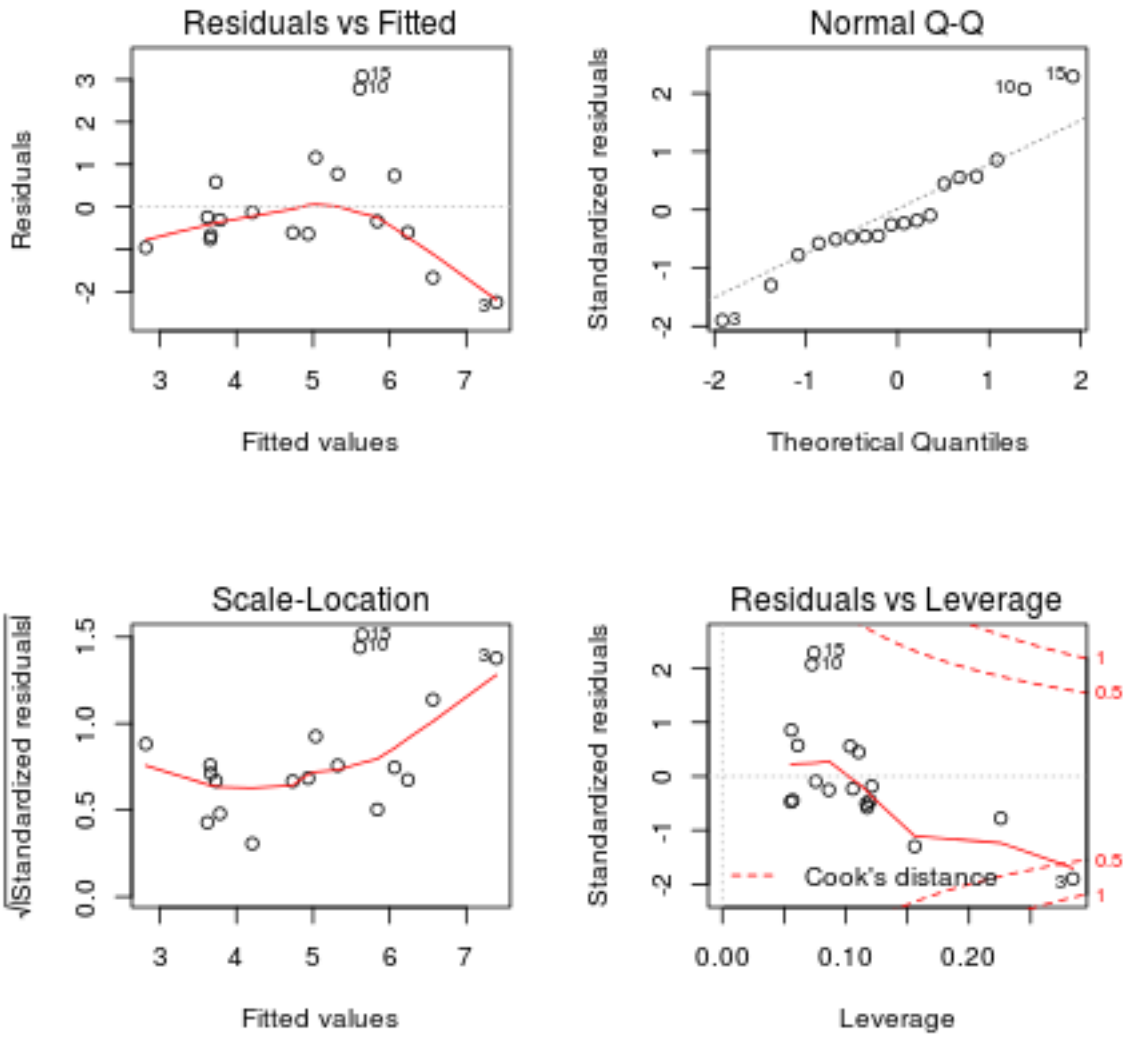
NOTE: We use the `text` function to overlay the number of each observation on-top of the scatter plot that we create initially. `text` requires a set of x and y coordinates. Here we add an offset of 0.2 to make sure the labels are not obscured by the points

```
model <- lm(y ~x )
summary(model)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.2437 -0.6580 -0.3219  0.6999  3.0792
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.9910174  0.8666011   2.298  0.03542 *
## x           0.0010221  0.0002781   3.675  0.00205 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.398 on 16 degrees of freedom
## Multiple R-squared:  0.4577, Adjusted R-squared:  0.4239
```

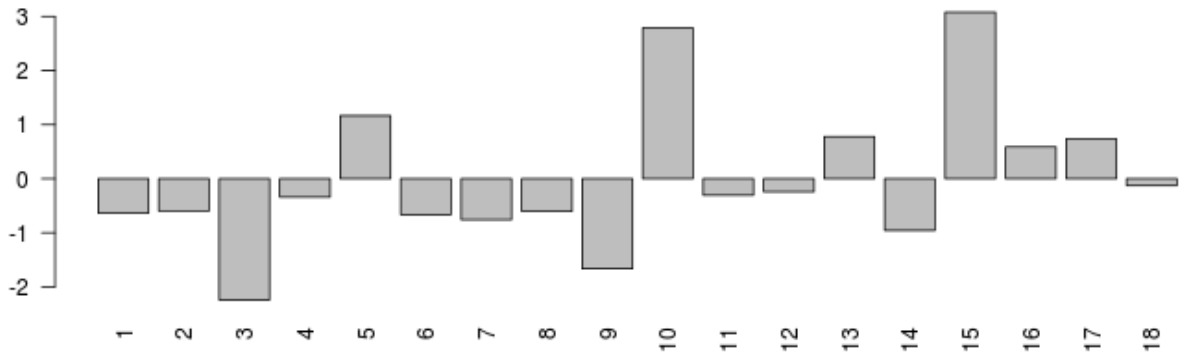
```
## F-statistic: 13.51 on 1 and 16 DF,  p-value: 0.002047
```

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



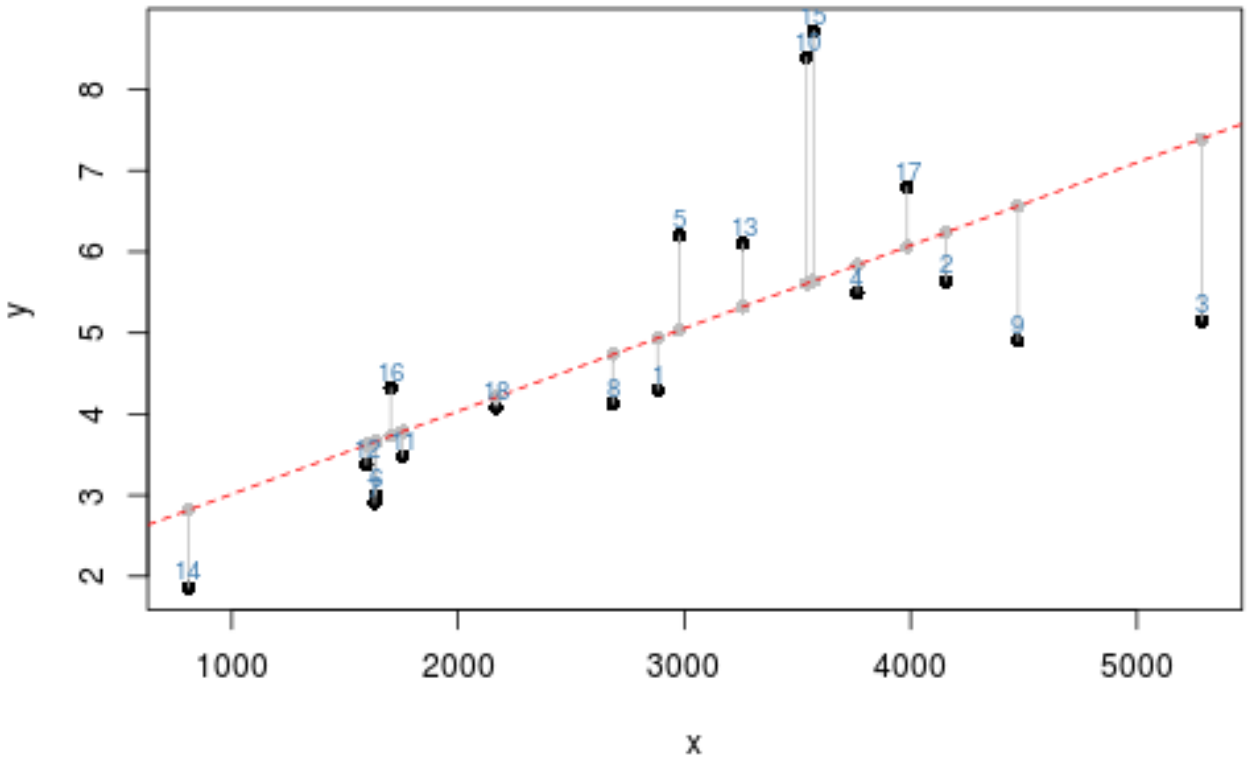
NOTE: We modify the plot layout by changing the mfrow parameter to allow for all the diagnostic plots to be shown in the same plotting device

```
par(mfrow=c(1,1))
barplot(model$residuals,names=1:length(x),las=2)
```



NOTE: The setting of the `las` parameter to 2 arranges the labels so they are perpendicular to the x-axis

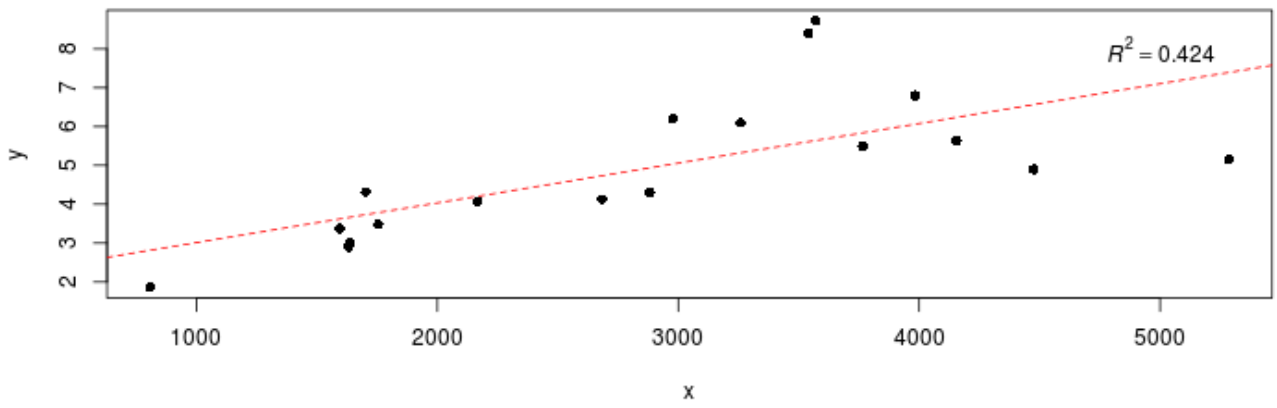
```
par(mfrow=c(1,1))
plot(x,y,pch=16)
points(x,model$fitted.values,pch=16,col="grey")
arrows(x,model$fitted.values,x,y,code=0,col="grey")
abline(model,col="red",lty=2)
text(x,y+0.2,labels=1:length(x),cex=0.8,col="steelblue")
```



***NOTE:** The `arrows` function was used to draw vertical lines on the plot. Each line is defined by a set of starting and ending x and y positions. In this case, we want the start and end positions to be the same and to vary the y coordinates*

```
plot(x,y,pch=16)
abline(model,col="red",lty=2)
result <- summary(model)
r2 <- round(result$adj.r.squared,3)
my.p = result$coefficients[2,4]

mylabel = bquote(italic(R)^2 == .(format(r2, digits = 3)))
text(x = 5000, y = 8, labels = mylabel)
```



2. Interpolating Unknown values

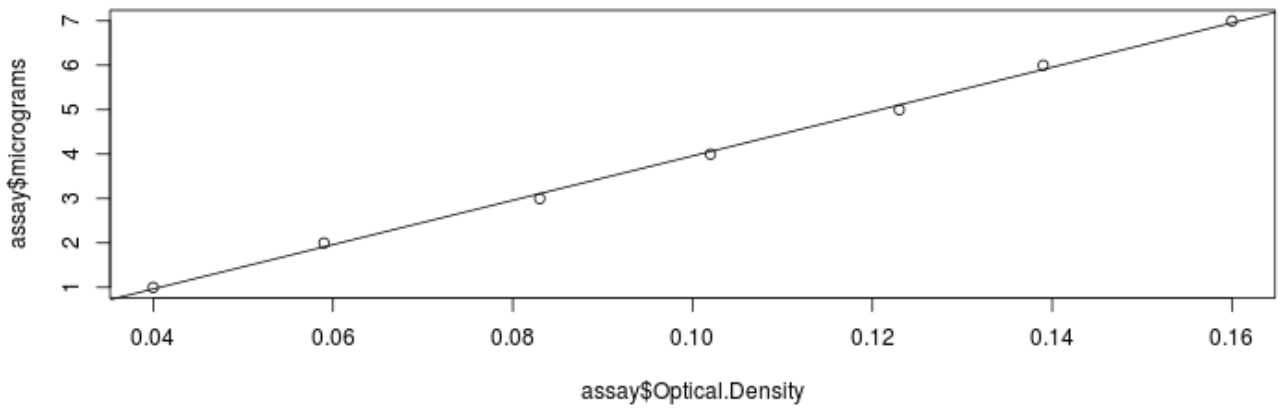
A standard curve is a graph relating a measured quantity (Optical Density in this example) to concentration of the substance of interest in "known" samples. A researcher has prepared and assayed "known" samples containing the substance in amounts chosen to span the range of concentrations that you expect to find in the "unknown" samples. These data are in the file `Assay.txt`.

a) Read the file `Assay.txt` and make a scatter plot of Concentration ('y') versus Optical Density ('x')
b) Fit a linear model to the data and show the results on the scatter plot
c) Use this linear model to predict values for the 'unknown' concentrations
d) Overlay your newly-predicted points on the scatter plot from a)

```
assay <- read.delim("Assay.txt")
model <- lm(micrograms~Optical.Density,data=assay)
summary(model)
```

```
##
## Call:
## lm(formula = micrograms ~ Optical.Density, data = assay)
##
## Residuals:
##       1       2       3       4       5       6       7
##  0.03774  0.08934 -0.10864 -0.05705 -0.10528  0.09606  0.04783
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -1.03438    0.09939  -10.41 0.000141 ***
## Optical.Density 49.91596    0.91597   54.49 3.94e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.09702 on 5 degrees of freedom
##   (3 observations deleted due to missingness)
## Multiple R-squared:  0.9983, Adjusted R-squared:  0.998
## F-statistic:  2970 on 1 and 5 DF,  p-value: 3.935e-08
```

```
plot(assay$Optical.Density, assay$micrograms)
abline(model)
```




```
y <- assay$micrograms
x <- assay$Optical.Density

topredict <- which(is.na(y))
new <- data.frame(x = x[which(is.na(y))])
y[which(is.na(y))] <- predict(lm(y~x), new)
```

```
cols <- rep("black",length(x))
cols[topredict] <- "red"
plot(x,y,pch=16,col=cols)
```