R lesson 3: Basic statistical functions

t-test

The `t.test` function can be used to perform all the t-tests whether one or two sample, and so on, with
the use of the functions parameters.

```
t.test(x, y = NULL,
    alternative = c("two.sided", "less", "greater"),
    mu = 0, paired = FALSE, var.equal = FALSE,
    conf.level = 0.95, ...)
```

> A=c(79.98, 80.04, 80.02, 80.03, 80.04, 79.97, 80.05, 80.03, 80.02, 80.00, 80.02)
> t.test(A,mu=80)

```
One Sample t-test
data: A
t = 3.1246, df = 12, p-value = 0.008779
alternative hypothesis: true mean is not equal to 80
95 percent confidence interval:
  80.00629 80.03525
sample estimates:
  mean of x
  80.02077
```

> t.test(A,mu=80,alternative="greater")

```
One Sample t-test
data: A
t = 3.1246, df = 12, p-value = 0.004389
alternative hypothesis: true mean is greater than 80
95 percent confidence interval:
 80.00892 Inf
sample estimates:
  mean of x
  80.02077
```

> B=c(80.02, 79.94, 79.98, 79.97, 79.97, 80.03, 79.95, 79.97)
> t.test(A,B,mu=0.05)

```
Welch Two Sample t-test
data: A and B
t = -0.6173, df = 12.027, p-value = 0.5486
alternative hypothesis: true difference in means is not equal to 0.05
95 percent confidence interval:
 0.01385526 0.07018320
sample estimates:
  mean of x mean of y
  80.02077  79.97875
```

If we can assume that the two variances are equal, i.e. \( \sigma^2_A = \sigma^2_B \) then we can use the pooled variance test.

> t.test(A,B,mu=0.05,var.equal.test=TRUE)
Welch Two Sample t-test

data:  A and B
t = -0.6173, df = 12.027, p-value = 0.5486
alternative hypothesis: true difference in means is not equal to 0.05
95 percent confidence interval:
  0.01385526 0.07018320
sample estimates:
  mean of x  mean of y
  80.02077  79.97875

Remark: We can actually test for equality of variances, but it usually not recommended to perform the F-test to determine whether to used the pooled variance t-test as we are in fact increasing the error rate $\alpha$ by performing two tests. Better to stick with the general unequal variance assumption

> var.test(A,B)

F test to compare two variances

data:  A and B
F = 0.5837, num df = 12, denom df = 7, p-value = 0.3938
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
  0.1251097 2.1052687
sample estimates:
  ratio of variances
  0.5837405

Linear models: Regression and ANOVA

One of the most basic functions for linear models is the `lm()` function used to fit linear models. It can be used to carry out regression, single stratum analysis of variance and analysis of covariance (although `anova` may provide a more convenient interface for these).

Regression

A simple linear regression example where we try to model the weight of a rock based on its porosity. Details are skipped.

> weight=c(99,101.1,102.7,103,105.4,107,108.7,110.8,112.1,112.4,113.6,113.8,115.1,115.4,120)
> porosity=c(28.8,27.9,27,25.2,22.8,21.5,20.9,19.6,17.1,18.9,16,16.7,13,13.6,10.8)
> plot(weight~porosity)
> reg=lm(weight~porosity)
> summary(reg) # Obtain estimates

Call:
  lm(formula = weight ~ porosity)

Residuals:
     Min      1Q  Median      3Q     Max
-1.7607 -0.7715 -0.0314  0.8403  1.8903

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 130.85443   1.01216 129.286  < 2e-16 ***
porosity    -1.07644   0.04889  -22.020  1.12e-11 ***

2
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1  1

Residual standard error: 1.023 on 13 degrees of freedom
Multiple R-squared: 0.9739,  Adjusted R-squared:  0.9719
F-statistic: 484.8 on 1 and 13 DF,  p-value: 1.125e-11

> anova(reg)  # ANOVA TABLE

Analysis of Variance Table

Response: weight
   Df Sum Sq Mean Sq F value Pr(>F)
porosity 1 507.59 507.59 484.84 1.125e-11 ***
Residuals 13 13.61  1.05

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

> abline(reg)  # Plot regression line

> anova(reg)

Analysis of Variance Table

Response: weight
   Df Sum Sq Mean Sq F value    Pr(>F)
porosity 1 507.59  507.59 484.84 1.125e-11 ***
Residuals 13 13.61   1.05

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

Inference on $\beta_1$ (slope)

> reg$coefficients[2]+c(1,-1)*qt(0.025,reg$df.residual)*sqrt(vcov(reg)[2,2])
[1] -1.1820518  -0.9708262
> reg$coefficients[2]/sqrt(vcov(reg)[2,2]) # test stat

porosity
-22.01916

Confidence and Prediction Intervals on response for a value of 27 for porosity. Since 27 is in the data the correct interval type is *confidence*

> which(porosity==27) # the third observation
[1] 3

> predict.lm(reg,newdata=data.frame(porosity=27),interval="confidence",level=0.95)

  fit  lwr  upr
1 101.7906 100.8555 102.7257

> predict.lm(reg,newdata=data.frame(porosity=27),interval="prediction",level=0.95)

  fit  lwr  upr
1 101.7906 99.39047 104.1907

Next we check the model assumptions

> re=rstandard(reg) #standardized residuals
> par(mfrow=c(2,2))
> # Normality
> hist(re)
> qqnorm(re,datay=TRUE)
> qqline(re,datay=TRUE)
> # Independence
> plot(re,type="o",pch=22,xlab="Order",main="Independence")
> abline(h=0)
> # Homogeneity of variance/Model Fit
> plot(re~fitted.values(reg),main="Homogeneity / Fit")
> abline(h=0)
Box-Cox Power transformation:

> # Method 1
> library(MASS)
> bc1=boxcox(reg,seq(-1,3,1/10))
> bc1$x[which.max(bc1$y)]

[1] 1.505051

> # Method 2
> library(car)
> bc2=powerTransform(weight~porosity)
> summary(bc2)

bcPower Transformation to Normality

<table>
<thead>
<tr>
<th>Est.Power</th>
<th>Std.Err.</th>
<th>Wald Lower Bound</th>
<th>Wald Upper Bound</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y1</td>
<td>1.5037</td>
<td>-1.4088</td>
<td>4.4162</td>
</tr>
</tbody>
</table>

Likelihood ratio tests about transformation parameters

<table>
<thead>
<tr>
<th>LRT</th>
<th>df</th>
<th>pval</th>
</tr>
</thead>
<tbody>
<tr>
<td>LR test, lambda = (0)</td>
<td>0.9846913</td>
<td>1</td>
</tr>
<tr>
<td>LR test, lambda = (1)</td>
<td>0.1141192</td>
<td>1</td>
</tr>
</tbody>
</table>

> yT=bcPower(weight,1.5037)

Multiple regression example

> dat=read.table("http://www.stat.ufl.edu/~dathien/STA6167/table8-4.csv",header=TRUE)
> dat1=data.frame(cbind(dat,dat$Speed*dat$Pause,dat$LCR^2,dat$Speed^2,dat$Pause^2))
> colnames(dat1)[5:8]=c("Speed_Pause","LCR2","Speed2","Pause2")
> attach(dat1)
> head(dat1)
> head(dat1)
<table>
<thead>
<tr>
<th></th>
<th>Speed</th>
<th>Pause</th>
<th>LCR</th>
<th>Overhead</th>
<th>Speed_Pause</th>
<th>LCR2</th>
<th>Speed2</th>
<th>Pause2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5</td>
<td>10</td>
<td>9.426</td>
<td>428.90</td>
<td>50</td>
<td>88.84948</td>
<td>25</td>
<td>100</td>
</tr>
<tr>
<td>2</td>
<td>5</td>
<td>20</td>
<td>8.318</td>
<td>443.68</td>
<td>100</td>
<td>69.18912</td>
<td>25</td>
<td>400</td>
</tr>
<tr>
<td>3</td>
<td>5</td>
<td>30</td>
<td>7.366</td>
<td>452.38</td>
<td>150</td>
<td>54.25796</td>
<td>25</td>
<td>900</td>
</tr>
<tr>
<td>4</td>
<td>5</td>
<td>40</td>
<td>6.744</td>
<td>461.24</td>
<td>200</td>
<td>45.48154</td>
<td>25</td>
<td>1600</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>50</td>
<td>6.059</td>
<td>475.07</td>
<td>250</td>
<td>36.71148</td>
<td>25</td>
<td>2500</td>
</tr>
<tr>
<td>6</td>
<td>10</td>
<td>10</td>
<td>16.456</td>
<td>446.06</td>
<td>100</td>
<td>270.79994</td>
<td>100</td>
<td>100</td>
</tr>
</tbody>
</table>

\[
\text{reg_full}=\text{lm(Overhead}\sim\text{LCR+Speed+Pause+Speed\_Pause+LCR2+Speed2+Pause2})
\]

\[
\text{summary(reg_full)}
\]

Call:
\[
\text{lm(formula = Overhead} \sim \text{LCR + Speed + Pause + Speed\_Pause + LCR2 + Speed2 + Pause2)}
\]

Residuals:

<table>
<thead>
<tr>
<th></th>
<th>Min</th>
<th>1Q</th>
<th>Median</th>
<th>3Q</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>-12.0242</td>
<td>-3.0847</td>
<td>0.2109</td>
<td>4.0988</td>
<td>8.6939</td>
</tr>
</tbody>
</table>

Coefficients:

|                         | Estimate | Std. Error | t value | Pr(>|t|) |
|-------------------------|----------|------------|---------|----------|
| (Intercept)             | 367.96413| 19.40264   | 18.965  | 7.12e-13 *** |
| LCR                     | 3.47669  | 2.12913    | 1.633   | 0.12087 |
| Speed                   | 3.04382  | 1.59133    | 1.913   | 0.07278 . |
| Pause                   | 2.29237  | 0.69838    | 3.282   | 0.00439 ** |
| Speed\_Pause            | -0.01222 | 0.01534    | -0.797  | 0.43663 |
| LCR2                    | -0.10412 | 0.03192    | -3.262  | 0.00459 ** |
| Speed2                  | -0.03131 | 0.01906    | -1.643  | 0.11885 |
| Pause2                  | -0.01318 | 0.01045    | -1.261  | 0.22442 |

---

Signif. codes:  
0 ** ** ** 0.001 ** 0.01 ** 0.05 . 0.1 1

Residual standard error: 5.723 on 17 degrees of freedom
Multiple R-squared: 0.9723, Adjusted R-squared: 0.9609
F-statistic: 85.33 on 7 and 17 DF, p-value: 5.409e-12

> \text{anova(reg_full)}

Analysis of Variance Table

Response: Overhead

<table>
<thead>
<tr>
<th></th>
<th>Df</th>
<th>Sum Sq</th>
<th>Mean Sq</th>
<th>F value</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>LCR</td>
<td>1</td>
<td>1150.1</td>
<td>1150.1</td>
<td>35.1107</td>
<td>1.662e-05 ***</td>
</tr>
<tr>
<td>Speed</td>
<td>1</td>
<td>14266.1</td>
<td>14266.1</td>
<td>435.5036</td>
<td>1.492e-13 ***</td>
</tr>
<tr>
<td>Pause</td>
<td>1</td>
<td>1688.1</td>
<td>1688.1</td>
<td>51.5316</td>
<td>1.544e-06 ***</td>
</tr>
<tr>
<td>Speed_Pause</td>
<td>1</td>
<td>0.1</td>
<td>0.1</td>
<td>0.0034</td>
<td>0.95427</td>
</tr>
<tr>
<td>LCR2</td>
<td>1</td>
<td>2246.6</td>
<td>2246.6</td>
<td>68.5815</td>
<td>2.273e-07 ***</td>
</tr>
<tr>
<td>Speed2</td>
<td>1</td>
<td>164.4</td>
<td>164.4</td>
<td>5.0178</td>
<td>0.03874 *</td>
</tr>
<tr>
<td>Pause2</td>
<td>1</td>
<td>52.1</td>
<td>52.1</td>
<td>1.5896</td>
<td>0.22442</td>
</tr>
<tr>
<td>Residuals</td>
<td>17</td>
<td>556.9</td>
<td>32.8</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

---

Signif. codes:  
0 ** ** ** 0.001 ** 0.01 ** 0.05 . 0.1 1

> \text{SSE.f=}\text{anova(reg_full)}["Residuals","Sum Sq"]

We note that with \text{F}-test statistic of 85.33 on 7 and 17 degrees of freedom, the p-value is approximately 0 and conclude that at least one predictor is significant.
Next we turn our attention to the individual t-tests and try to identify any variables that can be removed. We contemplate removing Speed-Pause interaction and Pause² due to the large p-values. Consequently, we fit a reduced model that does not contain these variables and test $H_0: \beta_4 = \beta_7 = 0$ by observing the difference in SSE from the full to the reduced model.

```r
> reg_red=lm(Overhead~LCR+Speed+Pause+LCR2+Speed2)
> summary(reg_red)

Call:
  lm(formula = Overhead ~ LCR + Speed + Pause + LCR2 + Speed2)

Residuals:
     Min       1Q   Median       3Q      Max
-15.6130  -3.9060    0.9894   4.5577   8.3459

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  372.59816   16.93351   22.004  5.58e-15 ***
   LCR       4.79907    1.93506    2.480  0.02267 *
     Speed    1.99340    1.02272    1.949  0.06620 .
     Pause   1.44790    0.25873    5.596  2.14e-05 ***
    LCR2   -0.12345    0.03400   -3.631  0.00178 **
    Speed2  -0.02120    0.01746   -1.214  0.23966

---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.368 on 19 degrees of freedom
Multiple R-squared: 0.9617,    Adjusted R-squared:  0.9516
F-statistic: 95.45 on 5 and 19 DF,  p-value: 8.694e-13
```

```r
> SSE.r=anova(reg_red)["Residuals","Sum Sq"]
> # manually
> T.f=((SSE.r-SSE.f)/(reg_red$df.residual-reg_full$df.residual))/
+ (SSE.f/reg_full$df.residual)
> T.f

[1] 3.260553
```

```r
> 1-pf(T.f,reg_red$df.residual-reg_full$df.residual,reg_full$df.residual)

[1] 0.06330311
```

```r
> # p-value the right tail probability (1-left probability)
> 
> # using the anova function
> anova(reg_red,reg_full)

Analysis of Variance Table

Model 1: Overhead ~ LCR + Speed + Pause + LCR2 + Speed2
Model 2: Overhead ~ LCR + Speed + Pause + Speed_Pause + LCR2 + Speed2 +
       Pause2
     Res.Df RSS Df Sum of Sq     F Pr(>F)
1     19 770.50   
2     17 556.88   2   213.62 3.2606 0.0633 .

---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
With a p-value of 0.0633 we fail to reject H₀ and can consequently remove the Speed-Pause and Pause² terms from the model. However, note that p-value is somewhat close to the default α = 0.05, so it may not be surprising for someone to keep those terms if they want a larger $R^2$.

**ANOVA**

A metal alloy that undergoes one of four possible strengthening procedures is tested for strength.

<table>
<thead>
<tr>
<th>Factor</th>
<th>Alloy Strength</th>
<th>Mean</th>
<th>St. Dev.</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>250 264 256 260 239</td>
<td>253.8</td>
<td>9.757</td>
</tr>
<tr>
<td>B</td>
<td>263 254 267 265 267</td>
<td>263.2</td>
<td>5.4037</td>
</tr>
<tr>
<td>C</td>
<td>257 279 269 273 277</td>
<td>271.0</td>
<td>8.7178</td>
</tr>
<tr>
<td>D</td>
<td>253 258 262 264 273</td>
<td>262.0</td>
<td>7.4498</td>
</tr>
</tbody>
</table>

> # Create data frame
> alloy=data.frame(strength=c(250,264,256,260,239,263,254,267,265,
+ 267,257,279,269,273,277,253,258,262,264,273),
+ treat=factor(rep(c("A","B","C","D"),each=5)))
> head(alloy)

<table>
<thead>
<tr>
<th>strength</th>
<th>treat</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 250</td>
<td>A</td>
</tr>
<tr>
<td>2 264</td>
<td>A</td>
</tr>
<tr>
<td>3 256</td>
<td>A</td>
</tr>
<tr>
<td>4 260</td>
<td>A</td>
</tr>
<tr>
<td>5 239</td>
<td>A</td>
</tr>
<tr>
<td>6 263</td>
<td>B</td>
</tr>
</tbody>
</table>

> means=tapply(alloy$strength,alloy$treat,mean) # obtain means by trt
> means

    A    B    C    D
253.8 263.2 271.0 262.0

> tapply(alloy$strength,alloy$treat,sd) # obtain st.dev. by trt

    A    B    C    D
9.757049 5.403702 8.717798 7.449832

> # Dot plot
> stripchart(strength~treat, data=alloy, method="stack", vertical=TRUE,
+ pch=1, cex=1.5, xlab="Factor", ylab="strength", main="Dotplots by Treatments")
> title(sub="pre-analysis plot", adj=0, cex=5/6)
> mtext("Example")
> points(c(1,2,3,4),tapply(alloy$strength,alloy$treat,mean),col=2,pch=8)
> abline(h=mean(alloy$strength),col=3)
> legend(3,250, c("Observations", " Trt Mean","Grand Mean"), col = c(1,2,3), text.col= "black",
+ lty=c(0,0,1),pch=c(1,8,NA),bg="gray90")
We conduct the test for $H_0: \alpha_1 = \cdots = \alpha_4 = 0$

```r
> attach(alloy)
> m1=aov(strength~treat)
> summary(m1)

<table>
<thead>
<tr>
<th>Df</th>
<th>Sum Sq</th>
<th>Mean Sq</th>
<th>F value</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>treat</td>
<td>3</td>
<td>743.4</td>
<td>247.80</td>
<td>3.873</td>
</tr>
<tr>
<td>Residuals</td>
<td>16</td>
<td>1023.6</td>
<td>63.98</td>
<td></td>
</tr>
</tbody>
</table>
```

Next we perform multiple 95% pairwise comparisons:

- **Bonferroni**

```r
> library(combinat)
> lett=combn(LETTERS[1:4],2);lett

[1,] "A"  "A"  "A"  "B"  "B"  "C"
[2,] "B"  "C"  "D"  "C"  "D"  "D"

> d=unname(-combn(means,2,diff));d
[1]  -9.4 -17.2  -8.2  -7.8  1.2  9.0

> # decide to do 95% CI overall on all pairwise differences so c=6
> # each CI must be made at 100(1-0.05/6)=99.1667% level
> ME=qt(0.05/(6*2),16,lower.tail=FALSE)*sqrt(MSE*(1/5+1/5));ME

[1] 15.21813
> for(i in 1:6)print(c(lett[,i],round(d[i]+c(-1,1)*ME,4)))

[1] "A" "B" "-24.6181" "5.8181"
[1] "A" "C" "-32.4181" "-1.9819"
[1] "A" "D" "-23.4181" "7.0181"
[1] "B" "C" "-23.0181" "7.4181"
[1] "B" "D" "-14.0181" "16.4181"
[1] "C" "D" "-6.2181" "24.2181"

> rbind(lett,abs(d)>=ME)

[1,] "A" "A" "A" "B" "B" "C"
[2,] "B" "C" "D" "C" "D" "D"
[3,] "FALSE" "TRUE" "FALSE" "FALSE" "FALSE" "FALSE"

- Tukey's

> alloy.Tukey=TukeyHSD(m1, "treat")
> print(alloy.Tukey)

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

Fit: aov(formula = strength ~ treat)

$treat
diff  lwr  upr   p adj
B-A  9.4  5.072915 23.872915 0.2839920
C-A 17.2  2.727085 31.672915 0.0172933
D-A  8.2  6.272915 22.672915 0.053011
C-B  7.8  6.672915 22.672915 0.0352295
D-B -1.2 15.672915 13.472915 0.995084
D-C -9.0 23.472915 5.472915 0.3185074

> plot(alloy.Tukey, sub="Alloy Data", adj=0)
> mtext("Tukey Honest Significant Differences", side=3, line=0.5)
• Fisher's LSD

\[
\text{ME}_F = q_t(0.975, 16) \times \sqrt{\text{MSE} \times (1/5 + 1/5)}; \text{ME}_F
\]

\[
[1] 10.72387
\]

\[
\text{rbind(lett, abs(d) >= ME}_F)
\]

\[
[1,] "A" "A" "A" "B" "B" "C"
[2,] "B" "C" "D" "C" "D" "D"
[3,] "FALSE" "TRUE" "FALSE" "FALSE" "FALSE" "FALSE"
\]

• Dunnet's: Assume that Factor A is the default group

\[
\text{library(multcomp)}
\]

\[
\text{alloy.Dunnett = glht(m1, linfct = mcp(treat = "Dunnett"))}
\]

\[
\text{summary(alloy.Dunnett)} \# \text{This gives the Hypothesis tests if trt means are equal}
\]

Multiple Comparisons of Means: Dunnett Contrasts

Fit: aov(formula = strength ~ treat)

Linear Hypotheses:

| Estimate | Std. Error | t value | Pr(>|t|) |
|----------|------------|---------|----------|
| B - A == 0 | 9.4000 | 5.059 | 1.858 | 0.19102 |
| C - A == 0 | 17.2000 | 5.059 | 3.400 | 0.00996 ** |
| D - A == 0 | 8.2000 | 5.059 | 1.621 | 0.28017 |

---

Signif. codes: 0 **** S 0.001 ** S 0.01 * S 0.05 S 0.1 S S 1

(Adjusted p values reported -- single-step method)

\[
\text{confint(alloy.Dunnett)} \# \text{This gives the CI}
\]

Multiple Comparisons of Means: Dunnett Contrasts

Fit: aov(formula = strength ~ treat)

Quantile = 2.5907

95% family-wise confidence level

Linear Hypotheses:

<table>
<thead>
<tr>
<th>Estimate</th>
<th>lwr</th>
<th>upr</th>
</tr>
</thead>
<tbody>
<tr>
<td>B - A == 0</td>
<td>9.4000</td>
<td>-3.7056</td>
</tr>
<tr>
<td>C - A == 0</td>
<td>17.2000</td>
<td>4.0944</td>
</tr>
<tr>
<td>D - A == 0</td>
<td>8.2000</td>
<td>-4.9056</td>
</tr>
</tbody>
</table>

> plot(alloy.Dunnett, sub="Alloy Data", adj=0)
> mtext("Dunnet Method", side=3, line=0.5)
95% family-wise confidence level
Dunnett Method

Alloy Data