In-class worksheet 2

Jan 21, 2016

1. t test

We will try the t test on the built-in data set `PlantGrowth`. However, first we need to reformat the data set, which we do with the function `unstack()`. We store the reformatted data set in a variable `plants`:

```
head(PlantGrowth)
```

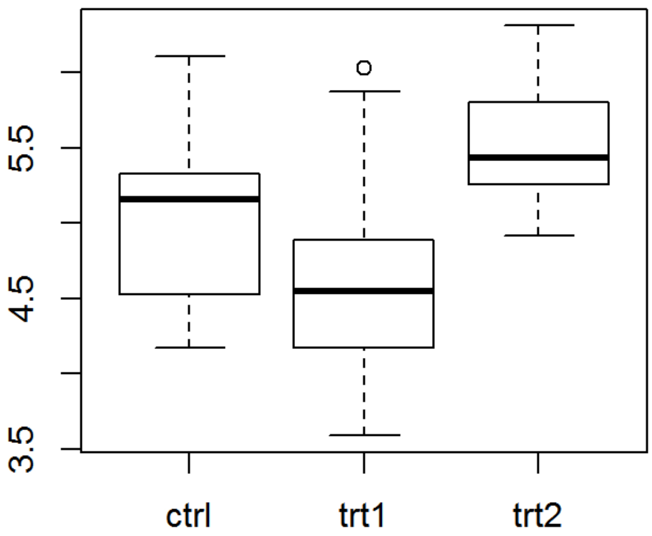
```
##   weight group
## 1   4.17  ctrl
## 2   5.58  ctrl
## 3   5.18  ctrl
## 4   6.11  ctrl
## 5   4.50  ctrl
## 6   4.61  ctrl
```

```
plants <- unstack(PlantGrowth)
head(plants)
```

```
##   ctrl trt1 trt2
## 1 4.17 4.81 6.31
## 2 5.58 4.17 5.12
## 3 5.18 4.41 5.54
## 4 6.11 3.59 5.50
## 5 4.50 5.87 5.37
## 6 4.61 3.83 5.29
```

The data set contains plant growth yield (dry weight) under one control and two treatment conditions:

```
boxplot(plants)
```



Question: Is the mean control weight significantly different from the mean weight under treatment 1? Is the mean weight under treatment 1 significantly different from the mean weight under treatment 2? Use the function `t.test()` to find out.

```
# R code goes here.
t.test(plants$ctrl, plants$trt1)
```

```
##
##  Welch Two Sample t-test
##
## data:  plants$ctrl and plants$trt1
## t = 1.1913, df = 16.524, p-value = 0.2504
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.2875162  1.0295162
## sample estimates:
## mean of x mean of y
##     5.032     4.661
```

```
t.test(plants$trt1, plants$trt2)
```

```
##
##  Welch Two Sample t-test
##
## data:  plants$trt1 and plants$trt2
## t = -3.0101, df = 14.104, p-value = 0.009298
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.4809144 -0.2490856
## sample estimates:
## mean of x mean of y
##     4.661     5.526
```

The first t.test has a p value if 0.25 which means we fail to reject the null hypothesis, and the control treatment can be said to be no different than treatment 1. The second treatment has a p-value of 0.009 meaning we can reject the null and treatment 1 and treatment 2 do have significantly different means.

2. Correlation

We will try the correlation test on the built-in data set `cars`. The data set contains the speed of cars and the distances taken to stop, measured in the 1920s:

```
head(cars)
```

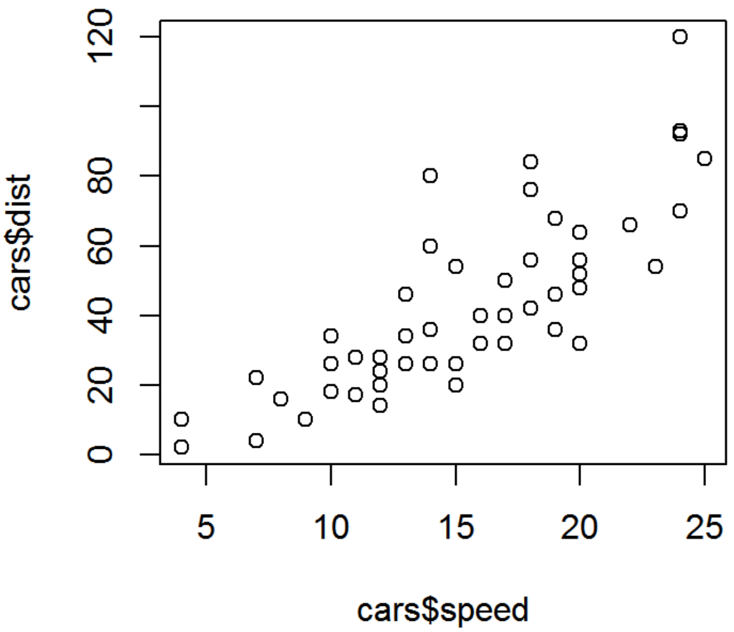
```
##   speed dist
## 1     4    2
## 2     4   10
## 3     7    4
## 4     7   22
## 5     8   16
## 6     9   10
```

Is there a relationship between speed and stopping distance? Use the function `cor.test()` to find out. Then make a scatterplot of speed vs. stopping distance, using the function `plot()`.

```
# R code goes here.
cor.test(cars$speed, cars$dist)
```

```
##
##  Pearson's product-moment correlation
##
## data:  cars$speed and cars$dist
## t = 9.464, df = 48, p-value = 1.49e-12
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.6816422 0.8862036
## sample estimates:
##       cor
## 0.8068949
```

```
plot(cars$speed, cars$dist)
```



There is a significant correalation between car speed and stopping distance, 0.81. This means that 66% of the cars variation in stopping distance can be attributed to speed (0.81^2 = 0.66)

3. Regression

We will do a regression analysis on the data set `cabbages` from the R package MASS. The data set contains the weight (`HeadWt`), vitamin C content (`VitC`), the cultivar (`Cult`), and the planting date (`Date`) for 60 cabbage heads:

```
library(MASS) # Load the MASS library to make the data set available
head(cabbages)
```

```
##   Cult Date HeadWt VitC
## 1  c39  d16    2.5   51
## 2  c39  d16    2.2   55
## 3  c39  d16    3.1   45
## 4  c39  d16    4.3   42
## 5  c39  d16    2.5   53
## 6  c39  d16    4.3   50
```

Use a multivariate regression to find out whether weight and cultivar have an effect on the vitamin C content. You will need to use the functions `lm()` and `summary()`.

```
# R code goes here.
fit <- lm(VitC~Cult+HeadWt, data = cabbages)
summary(fit)
```

```
## 
## Call:
## lm(formula = VitC ~ Cult + HeadWt, data = cabbages)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -12.233  -3.796  -1.064   4.542  14.061 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  67.9297     3.1159  21.801  < 2e-16 ***
## Cultc52       9.3578     1.7433   5.368 1.52e-06 ***
## HeadWt       -5.6524     0.9962  -5.674 4.88e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 6.304 on 57 degrees of freedom
## Multiple R-squared:  0.625,  Adjusted R-squared:  0.6119 
## F-statistic:  47.5 on 2 and 57 DF,  p-value: 7.234e-13
```

```
anova(fit)
```

```
## Analysis of Variance Table
##
## Response: VitC
##           Df Sum Sq Mean Sq F value    Pr(>F)
## Cult       1 2496.2 2496.15  62.811 9.145e-11 ***
## HeadWt     1 1279.5 1279.48  32.196 4.884e-07 ***
## Residuals 57 2265.2   39.74
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Both Cultivar and Head weight have significant impact on Vitamin C content. As head weight increases vitamin c content will decrease, as shown by the negative estimate value.

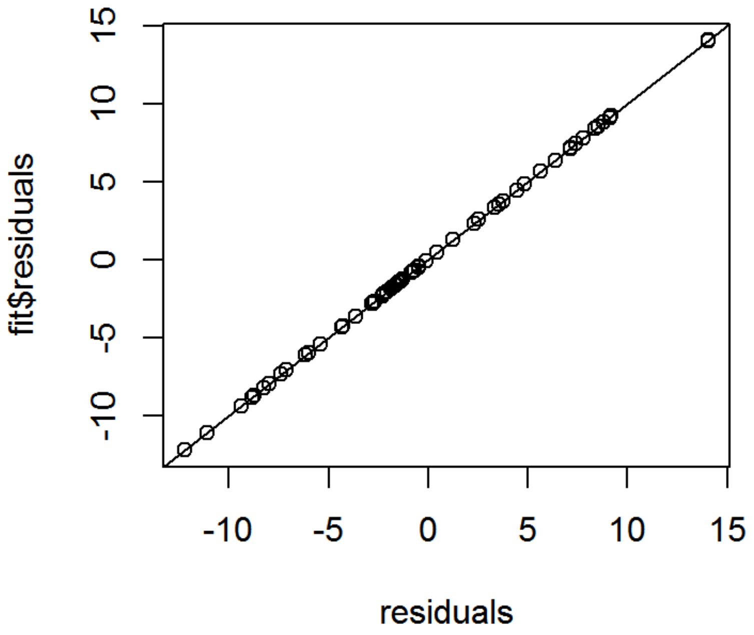
4. If this was easy

Look into the function `predict()`. Can you use it to estimate the vitamin C content of a c52 cultivar with a weight of 4? Can you use it to calculate the residuals of the regression model?

```
# R code goes here.
d <- data.frame(Cult="c52", HeadWt=4)
predict(fit,d)
```

```
##        1
## 54.67786
```

```
residuals <- cabbages$VitC - predict(fit, cabbages )
plot(residuals, fit$residuals)
abline(0,1)
```



We can predict that with cultivar C52 and a head weight of 4 we will have a vitamin c content of 54.68