Answer to assignment8

December 13, 2015

Preparing data

```
setwd("~/Documents/R/DAE/Lab5/Lab5instruction/Ordination/")

# Import the data from CSV files
spe <- read.csv("DoubsSpe.csv", row.names = 1)
env <- read.csv("DoubsEnv.csv", row.names = 1)
data <- data.frame(LOC = spe$LOC, pho = env$pho, alt = env$alt)
```

ANalysis Of VAriance (ANOVA)

Q1: How does pho affect LOC's abundance?

We can use `aov` function to fit an analysis of variance model.

```
# one-way ANOVA
aov1 <- aov(LOC ~ pho, data = data)
summary(aov1)
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## pho          1  22.70  22.705   7.509 0.0106 *
## Residuals   28  84.66   3.024
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The F-test of `pho` ($p = 0.0106 < 0.05$) indicates that `pho` does affect `LOC`'s abundance and it can account for $22.70/(22.70 + 84.66) * 100 = 0.2114$ variance of `LOC`. However, the output of `aov` function can not quantitatively measure how `pho` affects `LOC`'s abundance. Hence, we can first fit the one-way ANOVA model using the `lm` function.

```
lm1 <- lm(LOC ~ pho, data = data)
summary(lm1)
```

```
##
## Call:
## lm(formula = LOC ~ pho, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.9862 -1.5092 -0.1738  1.2732  2.6094
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.9963     0.3782   7.924 1.25e-08 ***
## pho          -1.0096     0.3684  -2.740   0.0106 *
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.739 on 28 degrees of freedom
## Multiple R-squared:  0.2115, Adjusted R-squared:  0.1833
## F-statistic: 7.509 on 1 and 28 DF,  p-value: 0.01057
```

Since this model is a univariate linear model, the Multiple R-squared (0.2115) is actual the proportion of `LOC`'s variance that `pho` can explain.

Then an analysis of variance table for this model can be produced via the `anova` command.

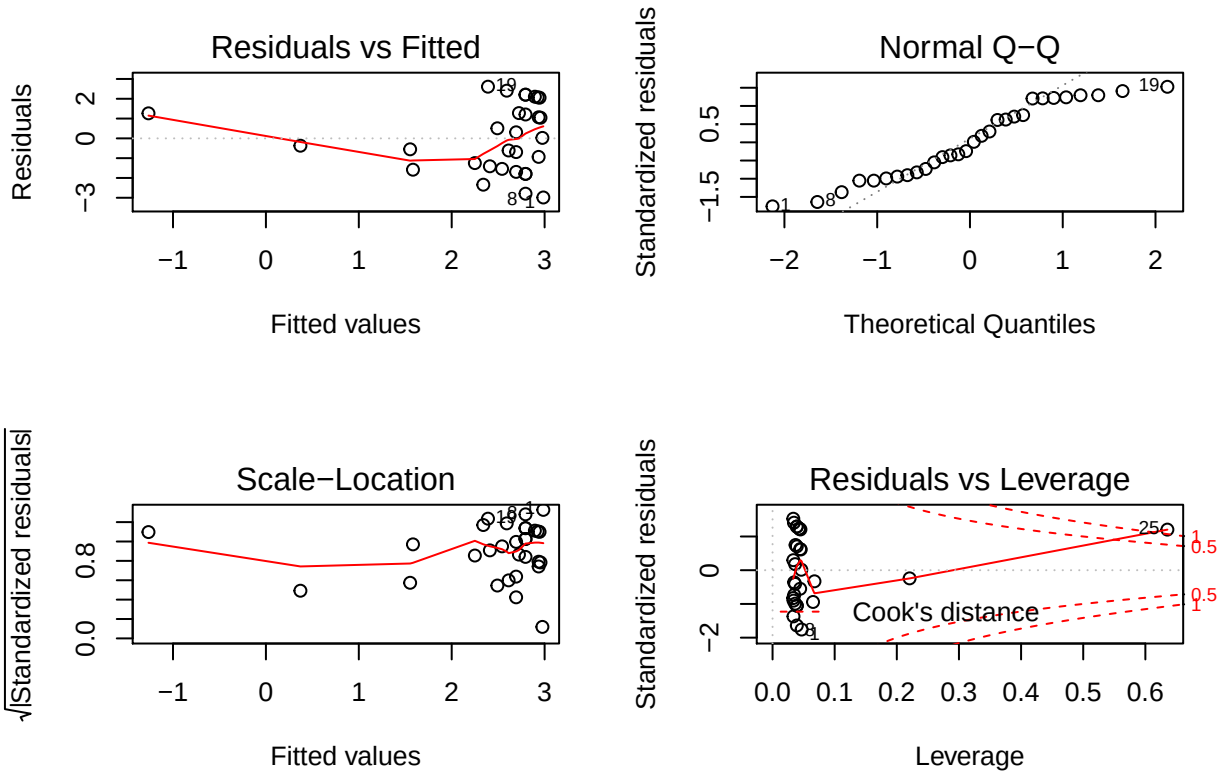
```
anova(lm1)
```

```
## Analysis of Variance Table
##
## Response: LOC
##           Df Sum Sq Mean Sq F value  Pr(>F)
## pho        1 22.705 22.7048  7.5091 0.01057 *
## Residuals 28 84.662  3.0236
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The output of `anova` is the same as that of `aov`, but the coefficient of `pho` (-1.0096) from `lm` suggests that there is a negative relation between `LOC` and `pho`.

At last, we need to investigate the model diagnostics to ensure that the various assumptions are broadly valid.

```
par(mfrow = c(2, 2))
plot(lm1)
```



Q2: How do pho and alt affect LOC's abundance?

To answer Q2, we can conduct a two-way ANOVA model using the methods described in Q1.

```
# two-way ANOVA
aov2 <- aov(LOC ~ pho + alt, data = data)
summary(aov2)
```

```
##             Df Sum Sq Mean Sq F value  Pr(>F)
## pho          1  22.70  22.705   7.741 0.00973 **
## alt          1   5.47   5.471   1.865 0.18330
## Residuals   27  79.19   2.933
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the analysis of variance table we can see `alt` can explain only 0.0510 variance of `LOC` and the p-value of F-test of `alt` is 0.18330, which is far greater than 0.05.

```
lm2 <- lm(LOC ~ pho + alt, data = data)
summary(lm2)
```

```
##
## Call:
## lm(formula = LOC ~ pho + alt, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.6598 -1.1645 -0.2851  1.4251  2.9051
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.001055   0.818416   2.445   0.0213 *
## pho         -0.765350   0.404524  -1.892   0.0693 .
## alt          0.001784   0.001306   1.366   0.1833
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.713 on 27 degrees of freedom
## Multiple R-squared:  0.2624, Adjusted R-squared:  0.2078
## F-statistic: 4.803 on 2 and 27 DF,  p-value: 0.01642
```

```
anova(lm2)
```

```
## Analysis of Variance Table
##
## Response: LOC
##           Df Sum Sq Mean Sq F value   Pr(>F)
## pho        1 22.705 22.7048  7.7411 0.009727 **
## alt        1  5.471  5.4706  1.8652 0.183297
## Residuals 27 79.191  2.9330
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Though including `alt` into predictors improves the Multiple R-squared by 0.0510, but the coefficients of both `pho` and `alt` are not statistically significant. Another usage of `anova` function is to compare models. To determine whether the inclusion of `alt` is appropriate, we can use `anova` to compare `lm2` with `lm1`.

```
anova(lm2, lm1)
```

```
## Analysis of Variance Table
##
## Model 1: LOC ~ pho + alt
## Model 2: LOC ~ pho
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     27 79.191
## 2     28 84.662 -1   -5.4706 1.8652 0.1833
```

Here `anova` performs the Chi-square test to compare `lm2` and `lm1`. It tests whether reduction in the residual sum of squares are statistically significant or not. From the output of `anova`, the reducation in the residual sum of squares is -5.4706 and the p-value of the test is 0.1833. It means that the fitted model `lm2` is not significantly different from `lm1` at the level of $\alpha = 0.05$. Therefore, for the simplicity of fitted model and avoiding overfitting, it's more appropriate not to include `alt` as a predictor.

In such a case, of course, it is unncessary to further consider the interaction of `pho` and `alt`.