Stats Exam 2016: Question 3

Sergii Gladchuk

December 26, 2016

Background

Lemmings are lousy runners. Professor Plupp wanted to know whether they actually do not have the muscles or if they are just lazy. To investigate this, he randomly collected 20 lemmings from different locations and took them into the lab. In the lab, he made them run on treadmills and measured their speed. Before the running experiment, some of the lemmings were treated with coffee. Prof. Plupp thought that the coffee would have an effect if lemmings were in general lazy, but otherwise not. While sampling the lemmings, the professor also noted at which elevation each lemming was caught. You find the data in the file "lemmingspeed" (.sav and .csv). The first variable is lemming id. Next is a variable coffee, indicating the treatment (coffee = 0 means no coffee; coffee = 1 means coffee was given). The variable elev gives the elevation (in meters) where each lemming was caught. Finally, the variable speed is the measured speed in m/s.

Analyses of co-variance (ANCOVA) should be conducted to answer questions based on both categorical variable *coffee* and numerical variable *elev*. There is no pseudoreplicaiton since all the lemmings are unique (have unique IDs).

The first thing is to read the data and convert integer *coffee* variable to factor.

```
speed.data <- read.csv('~/Documents/courses/stats/Exam/lemmingspeed.csv', sep=';')
str(speed.data)
```

```
## 'data.frame':    20 obs. of  4 variables:
##  $ id    : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ coffee: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ elev  : int  901 1132 1090 1250 1348 1203 1432 1298 1004 1227 ...
##  $ speed : num  0.3 0.35 0.37 0.41 0.42 0.4 0.46 0.36 0.26 0.43 ...
```

```
speed.data$coffee <- factor(speed.data$coffee)
str(speed.data)
```

```
## 'data.frame':    20 obs. of  4 variables:
##  $ id    : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ coffee: Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ elev  : int  901 1132 1090 1250 1348 1203 1432 1298 1004 1227 ...
##  $ speed : num  0.3 0.35 0.37 0.41 0.42 0.4 0.46 0.36 0.26 0.43 ...
```

Now data is fine, no missing values and factor for *coffee*.

Sub-question a)

Ignoring the elevation, is there an effect of the coffee treatment on the lemmings' running performance?

Since coffee is two-level factor, it is sufficient to use *t-test* to test the null-hypothesis that mean of speed for lemmings treated with coffee is the same as as mean of speed for lemmings which did not have coffee before running. Also box plot, error plot, and confidence intervals plot would be nice to confirm numerical test.

```
t.test(speed.data$speed~speed.data$coffee)
```

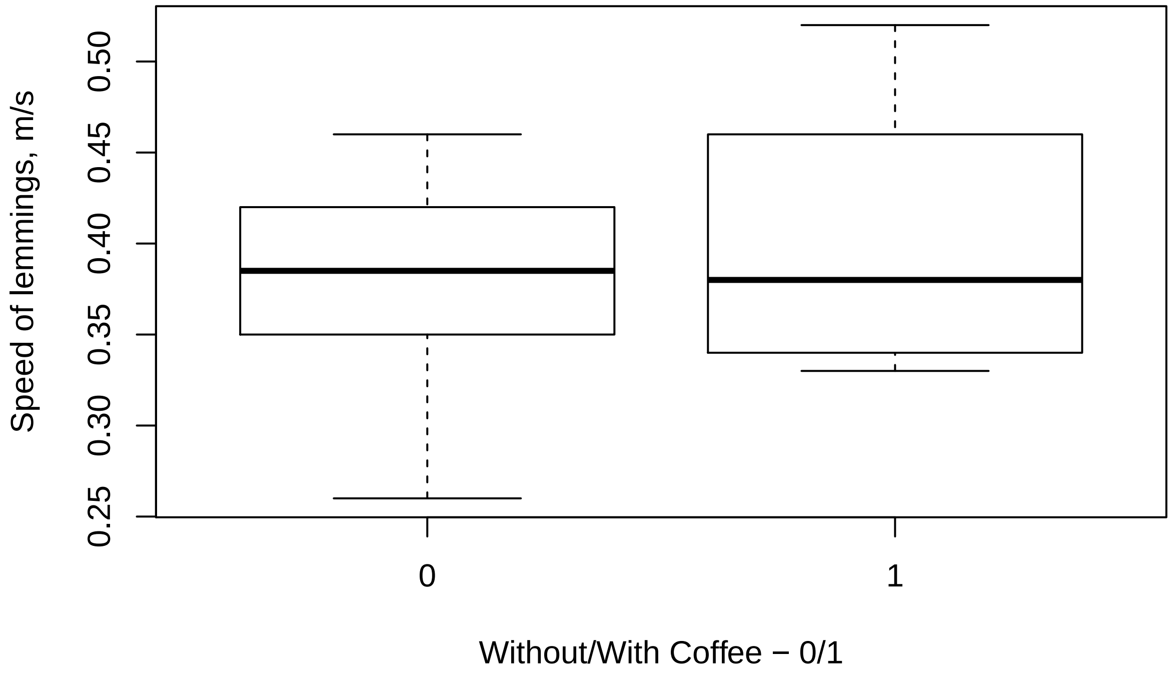
```
##
##  Welch Two Sample t-test
##
## data:  speed.data$speed by speed.data$coffee
## t = -0.83227, df = 17.837, p-value = 0.4163
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.08462344  0.03662344
## sample estimates:
## mean in group 0 mean in group 1
##           0.376           0.400
```

```
boxplot(speed.data$speed~speed.data$coffee,
        xlab='Without/With Coffee - 0/1',  ylab='Speed of lemmings, m/s')

library(Rmisc)
```

```
## Loading required package: lattice
```

```
## Loading required package: plyr
```



```r
mean.ci.se <- summarySE(speed.data, 'speed', 'coffee', na.rm = TRUE)
mean.ci.se
```

```
##   coffee  N speed         sd         se         ci
## 1      0 10 0.376 0.06131884 0.01939072 0.04386485
## 2      1 10 0.400 0.06749486 0.02134375 0.04828291
```

```r
library(Hmisc)
```

```
## Loading required package: survival
```

```
## Loading required package: Formula
```

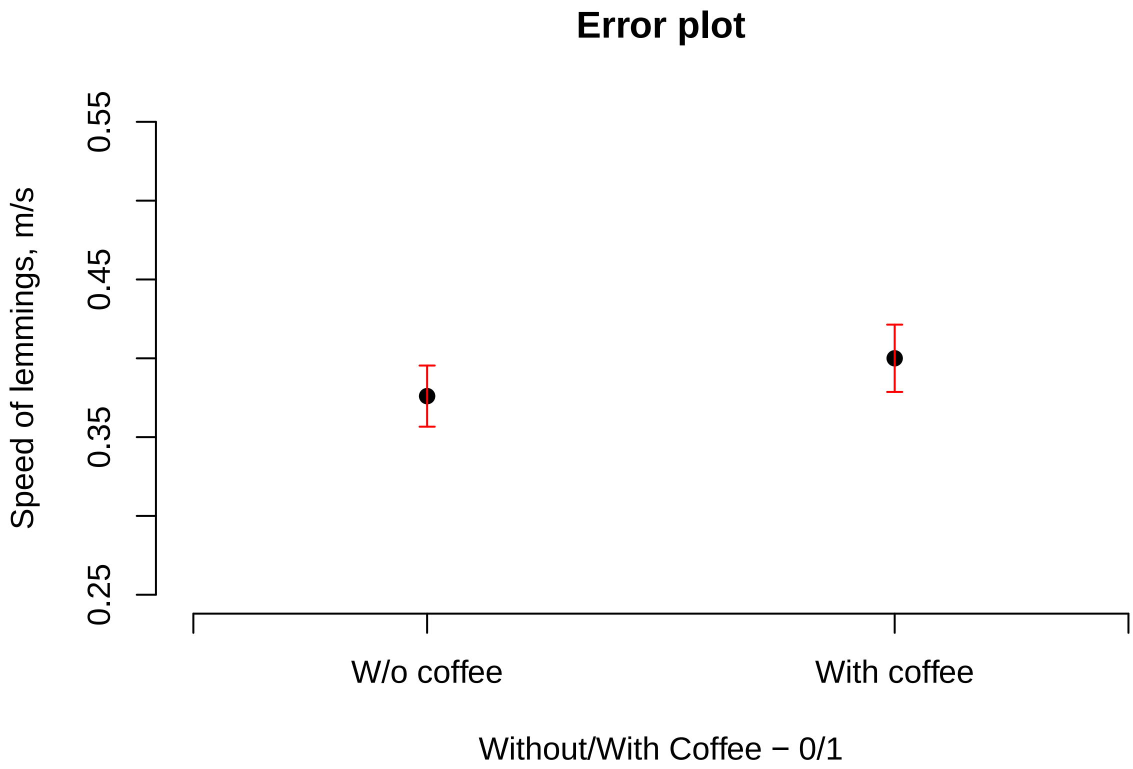
```
## Loading required package: ggplot2
```

```
## 
## Attaching package: 'Hmisc'
```

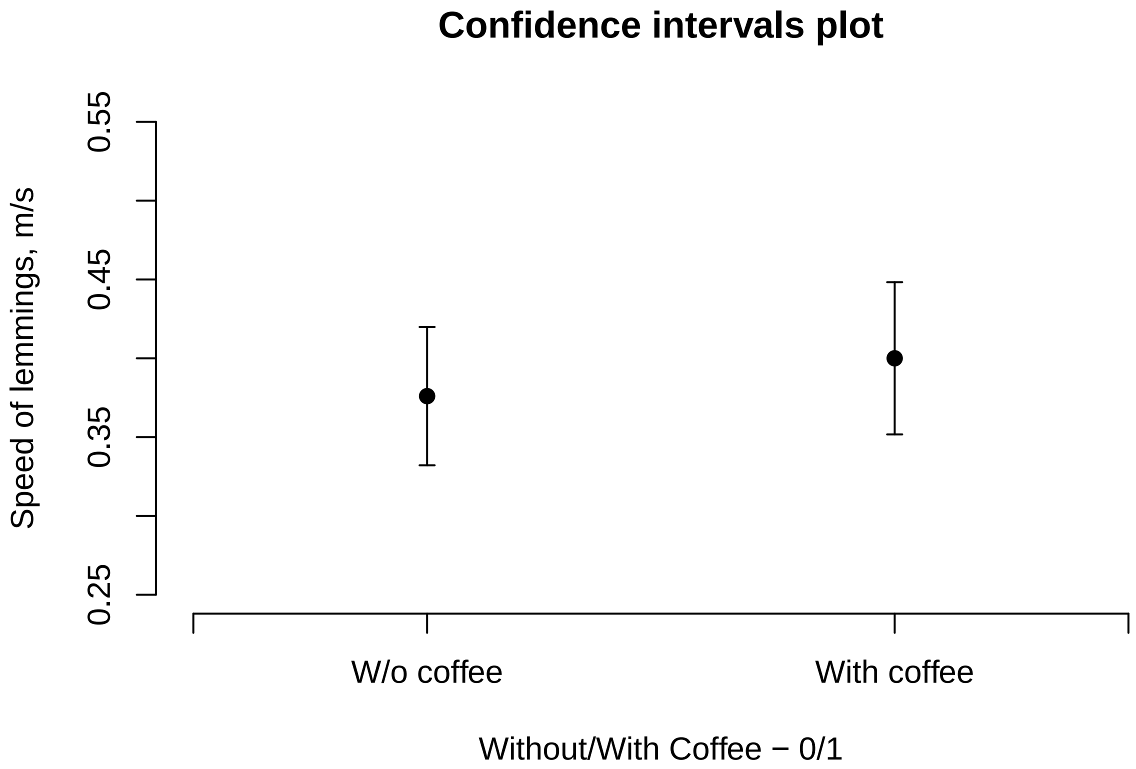
```
## The following objects are masked from 'package:plyr':
## 
##     is.discrete, summarize
```

```
## The following objects are masked from 'package:base':
## 
##     format.pval, round.POSIXt, trunc.POSIXt, units
```

```r
x <- c(1:2)
plot(x,mean.ci.se[,3], type='p', xlim=c(0.5,2.5),
     xlab='Without/With Coffee - 0/1', ylim=c(0.25,0.55), main='Error plot',
     ylab='Speed of lemmings, m/s', axes = FALSE)
axis(1, at=c(0.5,1,2,2.5), labels=c('','W/o coffee','With coffee',''))
axis(2, at=seq(0.25,0.55,0.05))
errbar(x, mean.ci.se[,3], mean.ci.se[,3] + mean.ci.se[,5],
       mean.ci.se[,3] - mean.ci.se[,5], add=TRUE,
       errbar.col='red')
```



```
plot(x,mean.ci.se[,3], type='p', xlim=c(0.5,2.5),
     xlab='Without/With Coffee - 0/1', ylim=c(0.25,0.55),
     main='Confidence intervals plot',
     ylab='Speed of lemmings, m/s', axes = FALSE)
axis(1, at=c(0.5,1,2,2.5), labels=c('','W/o coffee','With coffee',''))
axis(2, at=seq(0.25,0.55,0.05))
errbar(x, mean.ci.se[,3], mean.ci.se[,3] + mean.ci.se[,6],
       mean.ci.se[,3] - mean.ci.se[,6], add=TRUE)
```



The null-hypothesis is confirmed - pairwise, there is no significant difference in speed between lemming who drank coffee and who did not: p-value - 0.4163 > 0.05. Graphs confirm above statement. It is seen on the error plot that both means are located within error interval of other group.

Sub-question b)

If elevation is taken into account, is there a significant effect of the treatment?

Now ANCOVA time.

```
speed.aov <- aov(speed~elev*coffee, data=speed.data)
summary(speed.aov)
```

```
##             Df  Sum Sq Mean Sq F value   Pr(>F)
## elev         1 0.03537 0.03537  31.125 4.15e-05 ***
## coffee       1 0.02408 0.02408  21.191 0.000294 ***
## elev:coffee  1 0.00009 0.00009   0.082 0.778495
## Residuals   16 0.01818 0.00114
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Effects of the the separate *coffee* factor and *elevation* variable are significant. However, the interaction is not significant, which is logical since doctor did not treat lemmings with coffee specifically (e.g. lemmings from high elevations only). So similar slopes should be expected on the graph in the next sub-question.

The best fitted model will be without interaction. Since we reduce number of parameters (remove non-significant interaction).

```
speed.aov.best <- aov(speed~elev+coffee, data=speed.data)
summary(speed.aov.best)
```

```
##             Df  Sum Sq Mean Sq F value   Pr(>F)    
## elev         1 0.03537 0.03537    32.9 2.42e-05 ***
## coffee       1 0.02408 0.02408    22.4 0.000192 ***
## Residuals   17 0.01827 0.00107                     
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

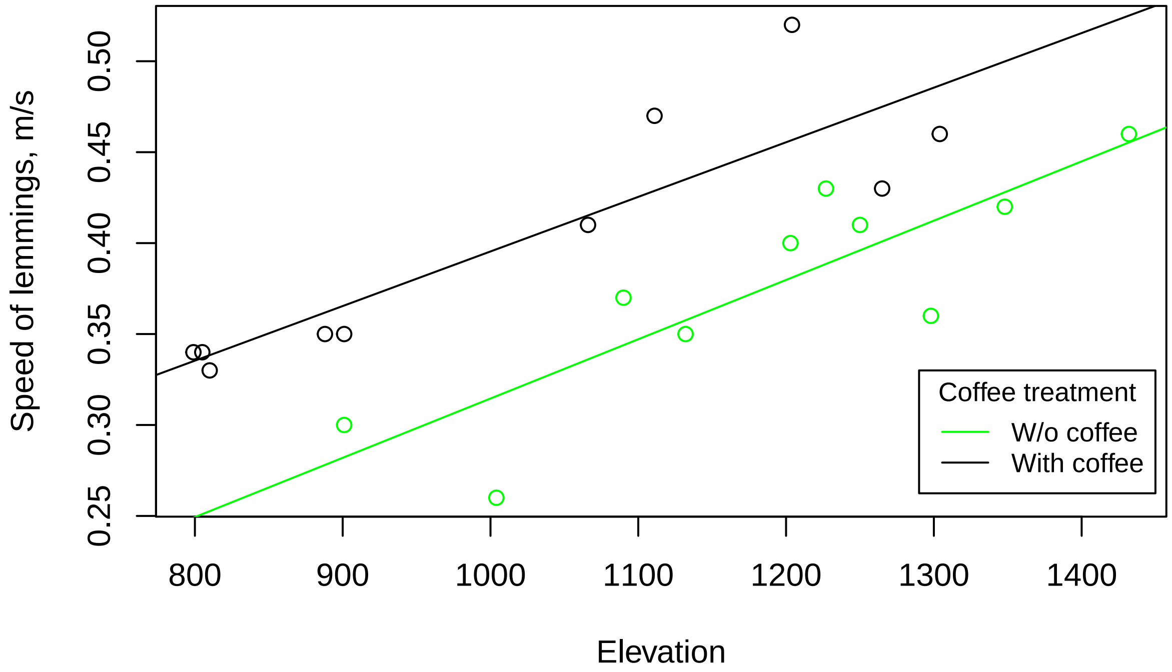
The p-values for each parameter got a little bit lower in comparison to model with interaction, which is sufficient for ANCOVA model selection process.

Sub-question c)

Make an illustration of your result in b)!

```
coffee.vec <- c('0','1')
coffee.col <- c('green','black')
plot (range(speed.data$elev), range(speed.data$speed),
      ylab='Speed of lemmings, m/s',
      xlab='Elevation', type='n')
for (i in 1:2){
  coffee.group <- subset(speed.data, coffee == coffee.vec[i])
  points(coffee.group$speed~coffee.group$elev, col=coffee.col[i])
  reg <- lm(coffee.group$speed~coffee.group$elev)
  abline(reg,col=coffee.col[i])

}
legend(1290,0.33, c('W/o coffee','With coffee'),
       cex=0.8, lty=1, col=coffee.col, title='Coffee treatment');
```



The above statement is supported by the graph - slopes are almost the same (no elevation dependence on coffee). So both coffee treatment and increase of elevation improve speed. In simple words lemmings who live on higher elevation run faster and/or lemmings, which are stimulated by coffee run faster too - so they are lazy indeed.

Sub-question d)

Express the model you found best in b), with speed as a function of (possibly) elevation and (possibly) coffee treatment. Make sure all parameter values are defined!

In order to express the best model found earlier, coefficients should be extracted

```
coef(speed.aov.best)
```

```
##  (Intercept)         elev      coffee1
## 0.0071365671 0.0003103605 0.0777544355
```

Since coffee is 2-level factor (0 or 1), then whole model can be expressed with single expression:

speed = 0.0071366 + 0.07775445*$\boldsymbol{coffee}$ + 0.00031036*$\boldsymbol{elev}$

where slope is 0.00031036, and intercept consist of two parts constant 0.0071366 and variable part 0.07775445 if *coffee* is 1 or 0 if *coffee* is 0.

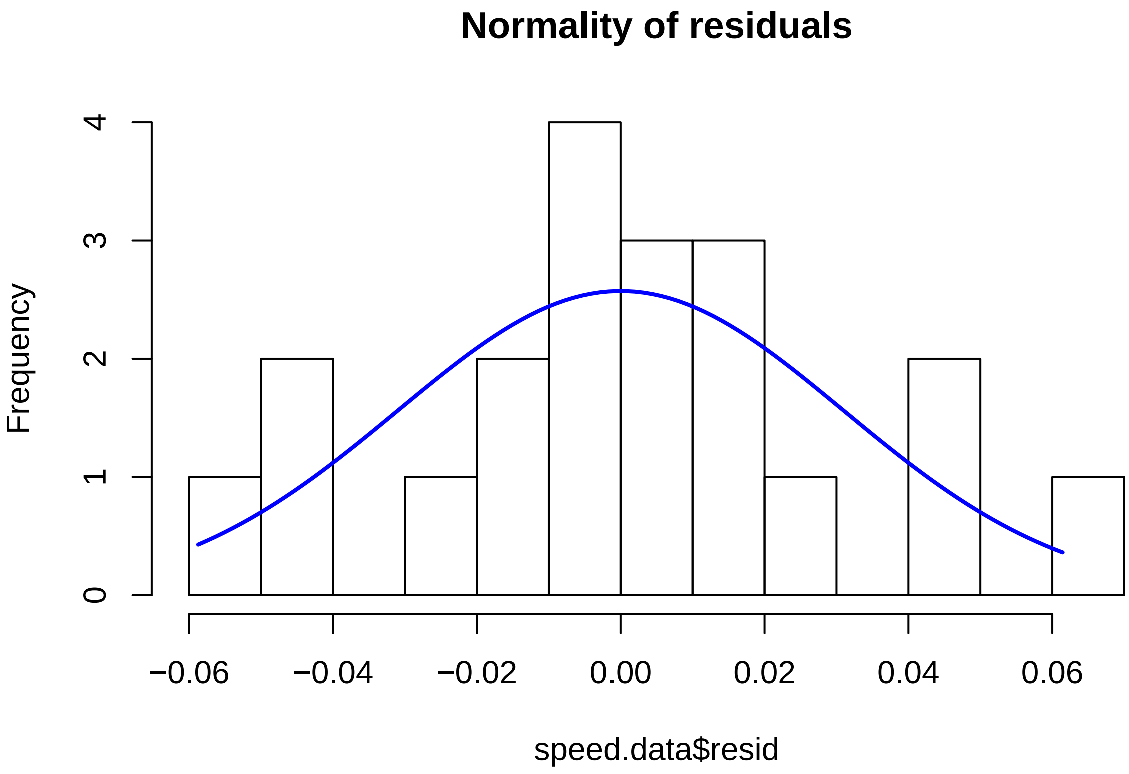
Sub-question e)

Test the assumptions of the test in b)!

List of the assumptions to be tested:

- **normality**. Test normality of residuals (histogram of residuals, Shapiro-Wilk normality test, Q-Q Plot)
- **independence** - plot residuals vs. elevation to confirm that there is no any trends.

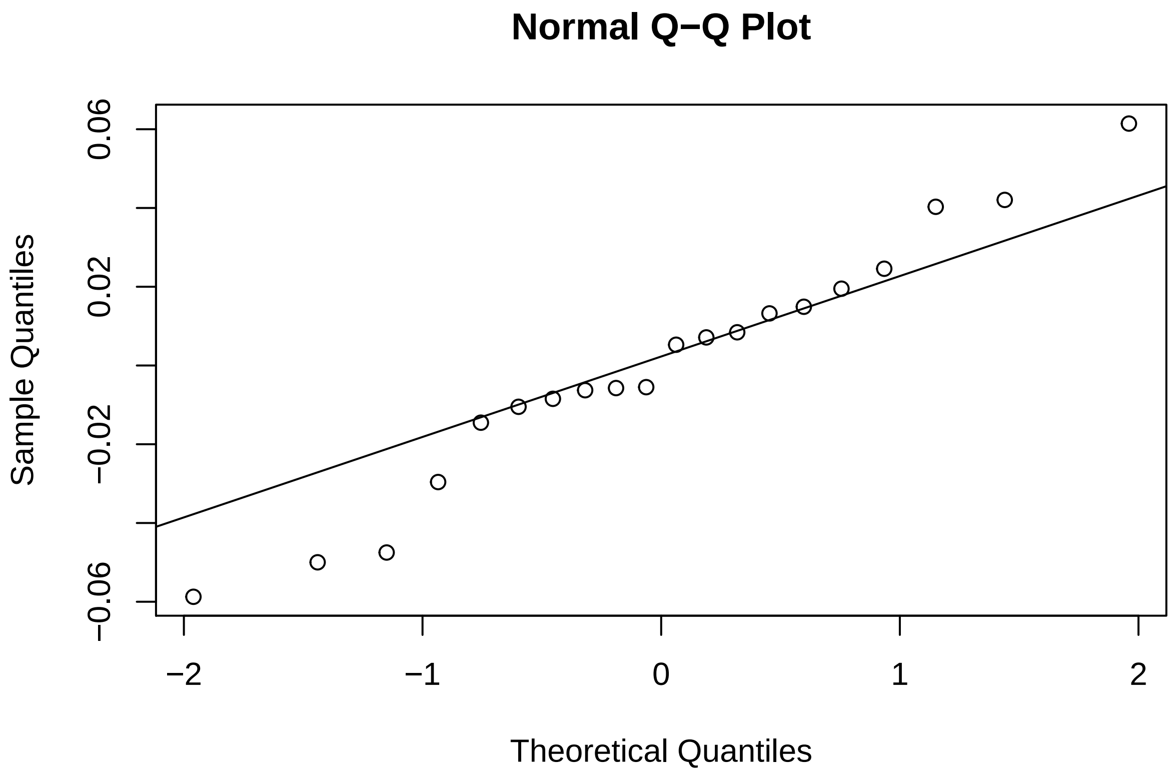
```
speed.data$resid <- residuals(speed.aov.best)
aveH <- hist(speed.data$resid, breaks=10, main = 'Normality of residuals')
xfit <- seq(min(speed.data$resid),max(speed.data$resid),length=100)
yfit <- dnorm(xfit,mean=mean(speed.data$resid),sd=sd(speed.data$resid))
yfit <- yfit*diff(aveH$mids[1:2])*length(speed.data$resid)
lines(xfit,yfit, col='blue', lwd=2)
```



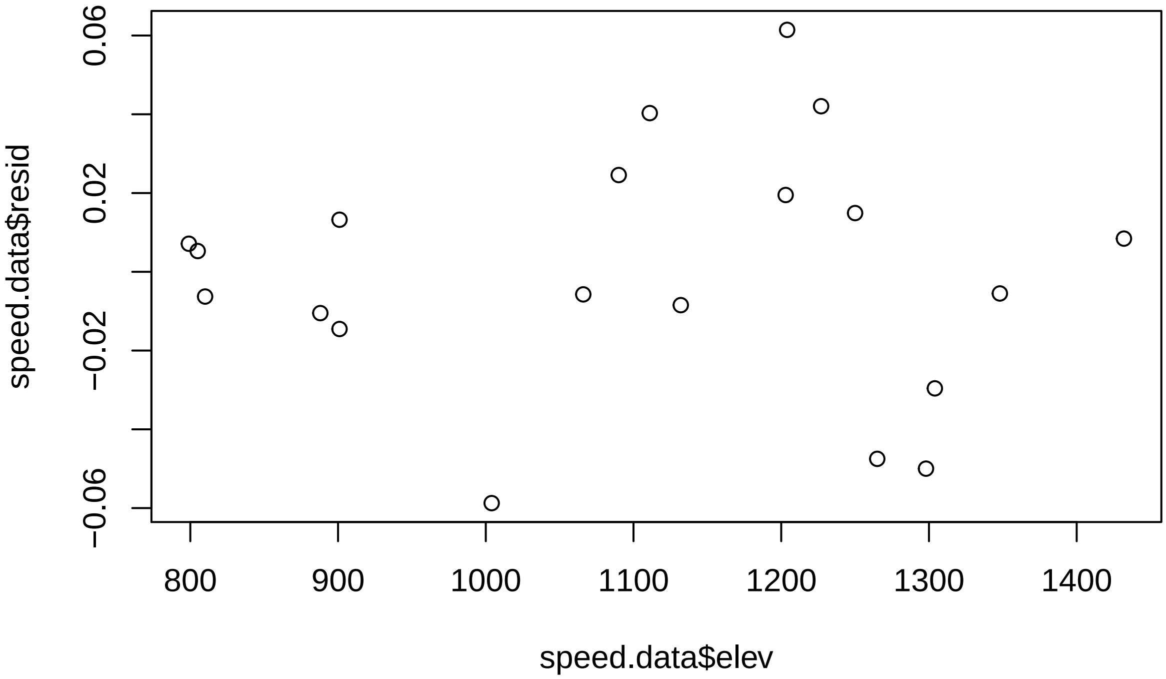
```
shapiro.test(speed.data$resid)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  speed.data$resid
## W = 0.97245, p-value = 0.8056
```

```r
qqnorm(speed.data$resid)
qqline(speed.data$resid)
```



```r
plot(speed.data$resid~speed.data$elev)
```



Conclusions:

- Residuals are normally distributed - the histogram and Q-Q plot could be better, but Shapiro-Wilk normality test is not significant (p-value > 20%) so distribution is close to normal.
- The independence assumption is also confirmed - plot does not show any trend in variance change.

ANCOVA assumptions were confirmed by the set of tests so as overall conclusion that speed of lemings increaseas with higher elevation and coffee treatment.