hw1

Giang Vu

1/23/2021

Homework 1

a)

If the data didn't support the null, we would conclude that value of g differs through time. Because wa are already assuming that the scientists tried really hard to make sure the measurements are of the same location, the only thing that is different between different measurements is when they were taken. So maybe time could also affect the gravity acceleration g.

b)

Null hypothesis: the means of the 8 series are equal
Alternative hypothesis: the means of the 8 series are not equal
Assumptions: data are normally distributed with equal variances and the samples (series of measurements) are independent from one another.
An ANOVA F test was carried out and the F stat is 3.62, p value is small (0.00211) => we reject the null hypothesis.

```r
#read data
gravity <- read.delim("/Users/giangvu/Desktop/STAT 2132 - Applied Stat Method 2/HW/HW1/Gravity.txt")
# anova f test
gravity.aov <- aov(Value ~ as.character(Series), data = gravity)
# Summary of the analysis
summary(gravity.aov)
```

```
##                      Df Sum Sq Mean Sq F value  Pr(>F)
## as.character(Series)  7   2798   399.7    3.62 0.00211 **
## Residuals            73   8061   110.4
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

c)

The likelihood ratio statistic was obtained by comparing two models, one with the assumption of equal variances among series, and another with the assumption of different variances among series using ANOVA. The likelihood ratio statisic is 18.187, with a small p-value of 0.0112 so it is statistically significant.

```r
gravity$Series <- as.factor(gravity$Series)
library(nlme)
```

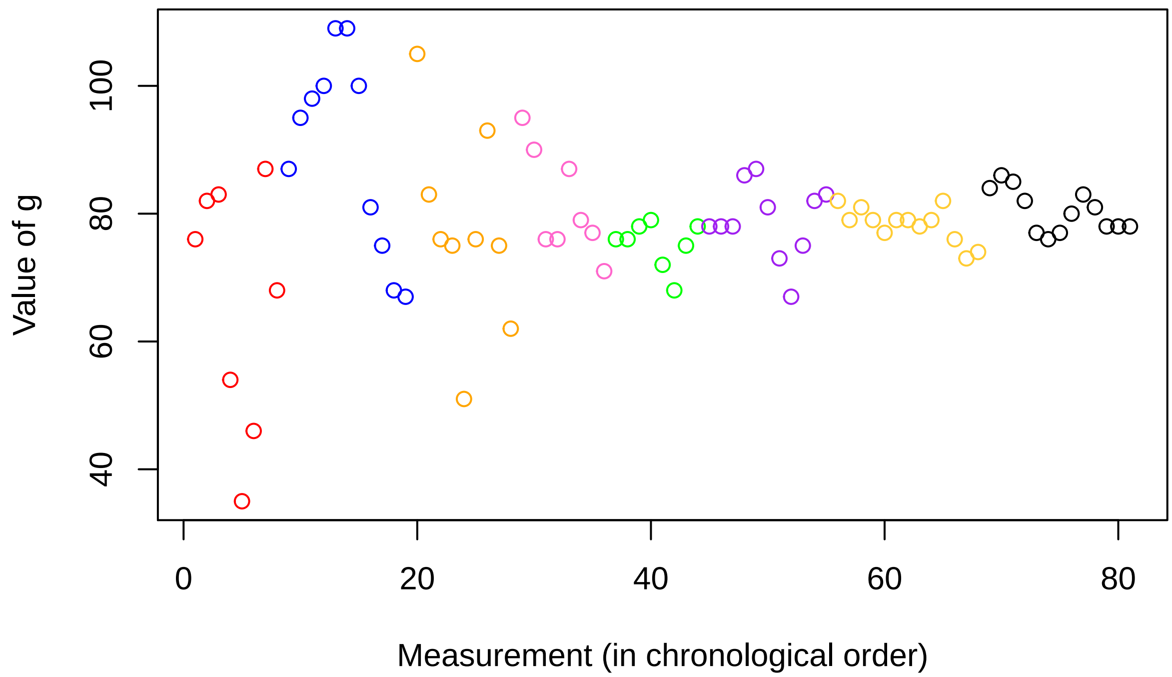
```
##
## Attaching package: 'nlme'

## The following object is masked from 'package:dplyr':
##
##     collapse
```

```r
#Model with different variances
model1 <- gls(Value~Series,data=gravity,
              method ="ML",
              weights = varIdent(form = ~1|Series))

#Model with equal variance assumption
model2 <- gls(Value~1,data=gravity,
              method ="ML",
              weights = varIdent(form = ~1|Series))

#likelihood ratio test to compare the two models
anova(model1,model2)
```

```
##        Model df      AIC      BIC    logLik   Test  L.Ratio p-value
## model1     1 16 561.5651 599.8763 -264.7826
## model2     2  9 565.7519 587.3019 -273.8759 1 vs 2 18.18672  0.0112
```

d)

In the plot below, data points represent measurements in a chronological order from left to right, colors signify different series, series are also ordered chronologically from left to right.
The more test the scientists carried out at that location, the more stable the result seemed to be. The variance of the first series of measurements are quite large compared to the later series.
This could mean that the samples (series) may not be independent from one another (the first measurements could have impacted the later measurements somehow) => our assumption might have been violated.



e)

First I tried removing outliers (using IQR rule) and repeat the previous analyses again. I failed to reject null hypothesis that the mean of different series are equal, so that is a better outcome than original data. The likelihood ratio statistics of 2 models, one with equal variance assumption, and one without it, is no longer statistically significant with the new data. This means with this new data we can safely assume equal variance, which can also be seen in the new plot. Then it would make sense to use the mean of the new data to get the estimate for g. Our estimate would then be 9.800785 m/s^2 with a standard error of 0.6442267. This shows us that our original data has some serial correlation, the samples are not independent, that's why our previous analysis had some violated assumption, leading to undesirable results.

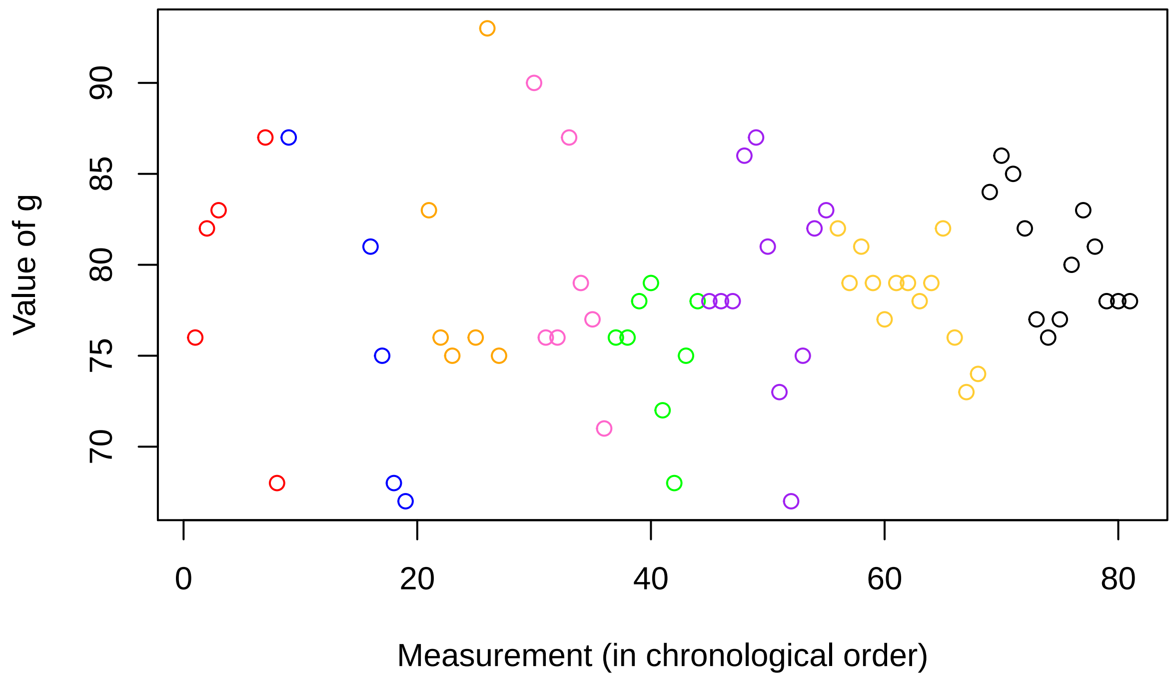
```
#remove outliers
iqr<-IQR(gravity$Value)
upbound <- quantile(gravity$Value,3/4,names = F)+iqr*1.5
lowbound <- quantile(gravity$Value,1/4,names = F)-iqr*1.5
gravity1 <- gravity[gravity$Value<=upbound & gravity$Value>=lowbound,]

#ANOVA
summary(aov(Value ~ Series, data = gravity1))
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## Series       7  191.6   27.37   0.966  0.464
## Residuals   60 1699.3   28.32
```

```r
#likelihood ratio
#model with different variance
model3 <- gls(Value~Series,data=gravity1,
              method ="ML",
              weights = varIdent(form = ~1|Series))

#Model with equal variance assumption
model4 <- gls(Value~1,data=gravity1,
              method ="ML",
              weights = varIdent(form = ~1|Series))

#likelihood ratio test to compare the two models
anova(model3,model4)
```

```
##        Model df      AIC      BIC    logLik   Test  L.Ratio p-value
## model3     1 16 425.4849 460.9971 -196.7425
## model4     2  9 421.5564 441.5320 -201.7782 1 vs 2 10.07144  0.1846
```

```r
#plot for new data
plot(gravity1$index,gravity1$Value, xlab = "Measurement (in chronological order)", ylab = "Value of g",
     col=ifelse(gravity1$Series==1, "red",
               ifelse(gravity1$Series==2,"blue",
                      ifelse(gravity1$Series==3,"orange",
                             ifelse(gravity1$Series==4,"#FF66CC",
                                    ifelse(gravity1$Series==5,"green",
                                           ifelse(gravity1$Series==6,"purple",
                                                  ifelse(gravity1$Series==7,"#FFCC33","black"))))))))
```




```
#mean of values
mean(gravity1$Value)
```

```
## [1] 78.54412
```

```
#standard error
sd(gravity1$Value)/sqrt(nrow(gravity1))
```

```
## [1] 0.6442267
```