

# STA610 Lab03

Hun Kang

2024-09-13

- Write down your answers in any blank sheet and submit your work in paper during the lab.
- Your work will not be graded. As long as you submit, you will get a full credit.
- For those who missed the lab today, you can submit it via email to me for half credit.

## Least Squares Review

How can we find  $c$  such that

$$\arg \min_{c \in \mathbb{R}} \sum_{i=1}^n (x_i - c)^2$$

One way is to start from

$$\sum_{i=1}^n (x_i - c)^2 = \sum_{i=1}^n (x_i - \bar{x} + \bar{x} - c)^2$$

Another approach is to use calculus

$$\frac{d}{dc} \sum_{i=1}^n (x_i - c)^2 = 0$$

**Q1-1: Compute  $c$  using both approaches.**

## Confidence Interval for Group Effects

```
radon<-readRDS(url("https://www2.stat.duke.edu/~pdh10/Teaching/610/Code/radonMN.rds"))
head(radon)
```

```
##      county    lon    lat    Uppm    radon
## 5081 AITKIN -93.415 46.608 0.502054 82.43790
## 5082 AITKIN -93.415 46.608 0.502054 82.43790
## 5083 AITKIN -93.415 46.608 0.502054 108.09157
## 5084 AITKIN -93.415 46.608 0.502054 39.18363
## 5085 ANOKA -93.246 45.273 0.428565 115.44118
## 5086 ANOKA -93.246 45.273 0.428565 93.41593
```

The following is the code to produce the first plot of the lecture slide s5GroupEstNP. We will modify this code to plot various confidence intervals of the group means.

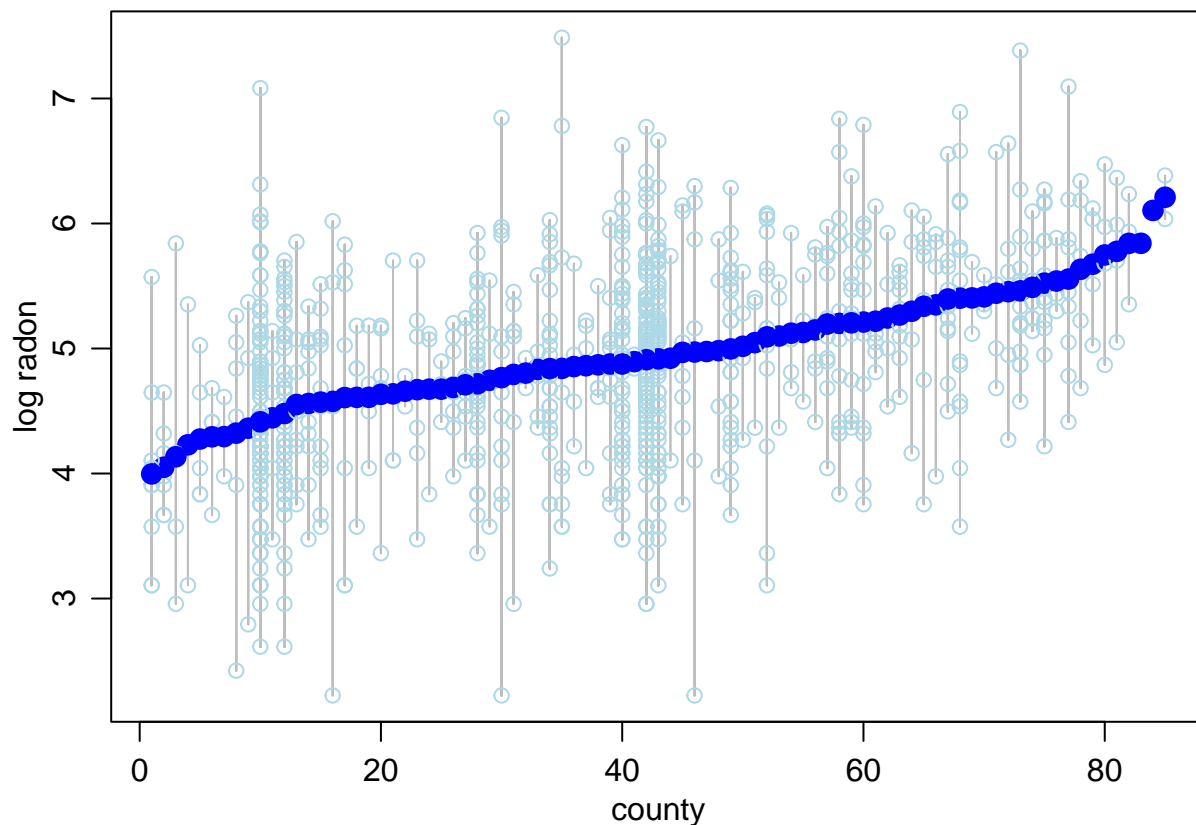
```

gdotplot<-function(y,g,xlab="group",ylab="response",mcol="blue",
                  ocol="lightblue",sortgroups=TRUE,...)
{
  m<-length(unique(g))
  rg<-rank( tapply(y,g,mean),ties.method="first")
  if(sortgroups==FALSE){ rg<-1:m ; names(rg)<-unique(g)}
  plot(c(1,m),range(y),type="n",xlab=xlab,ylab=ylab)

  for(j in unique(g))
  {
    yj<-y[g==j]
    rj<-rg[ match(as.character(j),names(rg)) ]
    nj<-length(yj)
    segments( rep(rj,nj) ,max(yj),rep(rj,nj),min(yj),col="gray")
    points( rep(rj,nj), yj,col=ocol,...)
    points(rj,mean(yj),pch=16,cex=1.5,col=mcol)
  }
}

par(mar=c(3,3,1,1), mgp=c(1.75,.75,0))
gdotplot(log(radon$radon),
        radon$county,
        xlab="county", ylab="log radon")

```



Q2-1: Identify the code lines that 1) compute the mean of each group, 2) order the

counties, 3) plot data points, 4) plot vertical lines.

**Q2-2:** Modify the plot so that counties are sorted according to the number of observations within county and draw a horizontal line marking the grand mean.

We first use ANOVA to evaluate statistical evidence of heterogeneity in across-county means.

```
anova(lm(log(radon) ~ county, data = radon))

## Analysis of Variance Table
##
## Response: log(radon)
##           Df Sum Sq Mean Sq F value    Pr(>F)
## county      84 126.93  1.51103   2.6406 2.981e-12 ***
## Residuals 834 477.23  0.57222
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Now we fit the hierarchical normal model to the data. Descriptions about the package with many useful examples can be found in <https://www.jstatsoft.org/article/view/v067i01>.

```
library(lme4)
```

**Q2-3:** Write down the model in the following code and identify the parameters and the random variables with corresponding distribution.

```
mod <- lmer(log(radon) ~ 1 + (1 | county), data = radon, REML = FALSE)
summary(mod)
```

```
## Linear mixed model fit by maximum likelihood ['lmerMod']
## Formula: log(radon) ~ 1 + (1 | county)
## Data: radon
##
##      AIC      BIC    logLik deviance df.resid
## 2164.1   2178.5  -1079.0   2158.1     916
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -3.6165 -0.6141  0.0292  0.6526  3.4932
##
## Random effects:
## Groups Name Variance Std.Dev.
## county (Intercept) 0.08804 0.2967
## Residual 0.57154 0.7560
## Number of obs: 919, groups: county, 85
##
## Fixed effects:
##              Estimate Std. Error t value
## (Intercept)  4.94656    0.04664   106.1
```

We can extract the estimates of  $\mu$ ,  $\tau$  and  $\sigma$  as below. Here, the estimate of  $\sigma$  is a pooled estimate.

```
fixef(mod)
```

```
## (Intercept)
##      4.946557
```

```
data.frame(VarCorr(mod))
```

```
##      grp      var1 var2      vcov      sdcor
## 1 county (Intercept) <NA> 0.08804027 0.2967158
## 2 Residual          <NA> <NA> 0.57153536 0.7559996
```

There are two different types of confidence interval we learned in class. One is the t-interval for group mean with pooled variance estimate:

$$\bar{y}_j \pm \frac{t_{1-\alpha/2}}{\sqrt{n_j/\hat{\sigma}^2}}$$

and the other is the Empirical Bayes interval

$$\left( \frac{\hat{\tau}^2}{\hat{\tau}^2 + \hat{\sigma}^2/n_j} \hat{\mu} + \frac{\hat{\sigma}^2/n_j}{\hat{\tau}^2 + \hat{\sigma}^2/n_j} \bar{y}_j \right) \pm \frac{t_{1-\alpha/2}}{\sqrt{n_j/\hat{\sigma}^2 + 1/\tau^2}}$$

**Q2-4:** Describe the differences between two intervals.

```
# pooled estimates
muhat = fixef(mod)
sighat = data.frame(VarCorr(mod))$sdcor[1]
tauhat = data.frame(VarCorr(mod))$sdcor[2]

# group-wise statistics
ybars = aggregate(log(radon) ~ county, data = radon, mean)[,2]
njs = aggregate(log(radon) ~ county, data = radon, length)[,2]
```

**Q2-5:** Plot two credible intervals for each group using the code above.