

School of Computer Science

Linear and General Regression Models

in Fulfilment of

MATH9903

Maksymilian Drzezdzon

C15311966

Degree: TU060/2

Module Coordinator: Joe Condon

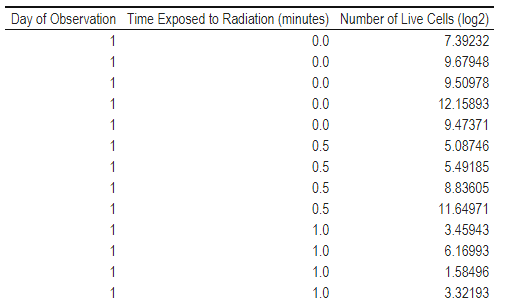
Declaration of Ownership: I declare that the attached work is entirely my own and that all sources have been acknowledged: 🗹

**Date: 04/03/2022**

Introduction

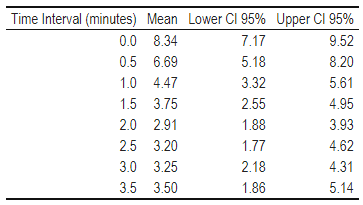
The skin cell data consists of 118 observations of human skin cells colonies. Samples where exposed to solar radiation from 0 (control) to 3.5 minutes in intervals of 0.5 minutes, this experiment was replicated over a ~28 days several times each day per interval of radiation exposure. Albeit some samples were contaminated and had to be discarded. The response variable for this analysis is the number of live cells remaining counted under a microscope after exposure to radiation documented in log2 form. The data are displayed in Table 1 in raw form.

Table 1: Skin Cell Data - first 13 observations



The question at hand is whether there is a statistically significant effect of solar radiation on the mortality of human skin cells. Table 2 presents the mean difference by cell survival count with upper and lower 95% confidence limits based on the student t distribution. There is a downward mean trend that levels off at the 2-minute interval showing an interesting rebound in both mean, lower and upper confidence intervals indicating variable survival rate of cells at longer exposure time. Further supported by Figure 1, group-wise confidence intervals overlap which results in the population mean difference in cell survivorship cannot be assumed statistically different.

Table 2: Cell survival count & 95% CI by time interval



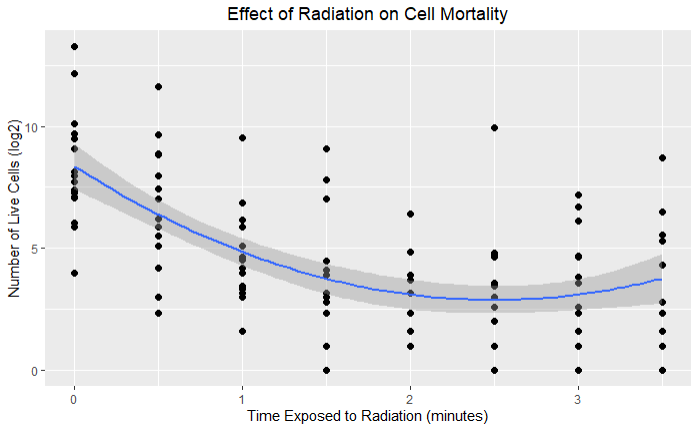


Figure 2: Scatterplot of Cell Survival Count

Model 1

A linear regression model was fitted to these data with *day of observation* as a categorical predictor and *time exposed to radiation (minutes)* as a continuous predictor. To assess interaction between *time exposed to radiation (minutes)*, *day of observation* was used as a factor to measure any variability imposed on *time exposed to radiation (minutes)*.

The model was: (fix the equation later)

Where is the live cell count for the sample , …….

This hypothesis was tested using the standard formulation for g

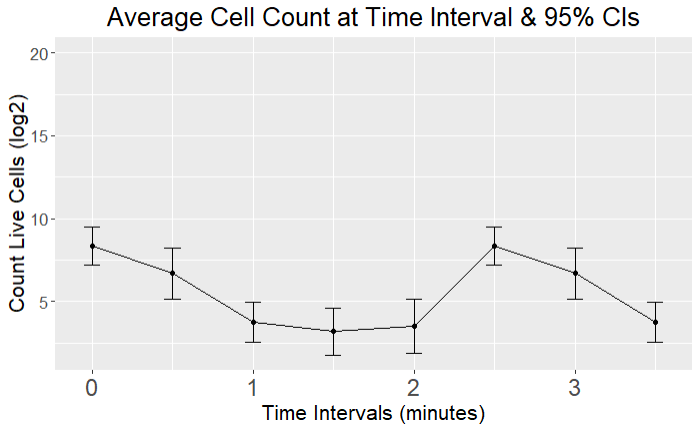
****

Figure 1: Mean difference by cell survival count with upper and lower 95% confidence

The ANOVA table for model 1 yields a p-value of 6.19e-08 on 7 degrees of freedom, indicating that there is a difference between means. As illustrated in Figure 1.

Then using the drop1() function in R using the F-statistic the best model yielded a p-value of 0.72 and f-statistic of 0.445 after iteratively dropping a single term. This model does not seem to be a good fit the p-value is astronomically high.

Upon examining the summary of the model which yielded an F-statistic of 8.1 on 7 and 110 degrees of freedom and a p-value of 6.193e-08. Few parameters were found statistically significant besides factor 2 of day of observation with a p-value of 0.034 and time exposed to radiation with a p-value of 3.17e-05, the intercept for this model is < 2e-16. Given the above output model 1 does not seem to be of good fit because there doesn’t seem to be any positive interaction between day of observation and time exposed to radiation bar 1 outlier documented above. The null hypothesis in this case in not rejected.

Model 2

Common slope model:

Xxxxx

Leads to the following regression line

After removing day of observation as a factor from the model and rerunning the model the ANOVA table shows that there is a difference between means with a p-value of 5.86e-11 on 7 degrees of freedom, much lower than observed in model 1.

on time exposed to radiation as a factor all time intervals have been found to be statistically significant. The effect of time exposed to radiation was statistically significant with an F-test statistic of 49.78 on 1 and 116 degrees of freedom, yielding a p-value of 1.339e-10. The slope for x was 6.13 at 2.5% and

Comparison among time intervals

Code sample

library(pastecs) #For creating descriptive statistic summaries

library(ggplot2) #For creating histograms with more detail than plot

library(psych) # Some useful descriptive functions

library(semTools) #For skewness and kurtosis

library(FSA) # For percentage

library(car) # For Levene's test for homogeneity of variance

library(effectsize) # To calculate effect size for t-test

library(kableExtra) # Used to generate report ready tables

# A copy paste of libraries I use in R projects

# Load data and setup project

setwd("~/GitHub/Masters-Classes-L-O/Applied Statistics/data\_stats")

getwd()

data = read.csv("skincells.csv")

anovatab <-

function(mod){

tab=as.matrix(anova(mod))

rows=dim(tab)[1]

moddf=sum(tab[,1])-tab[rows,1]

ssmodel=sum(tab[,2])-tab[rows,2]

msmodel=ssmodel/moddf

f=msmodel/tab[rows,3]

p=1-pf(f,moddf,tab[rows,1])

tab2=tab[(rows-1):rows,]

tab2[1,1:5]=c(moddf,ssmodel,msmodel,f,p)

tab2=rbind(tab2,c(moddf+tab2[2,1],ssmodel+tab2[2,2],rep(NA,3)))

rownames(tab2)=c('Model','Error','Total')

colnames(tab2)[1]='df'

return(print(tab2,na.print = "" , quote = FALSE,digits=3))

}

colnames(data)[1] = "Day of Observation"

colnames(data)[2] = "Time Exposed to Radiation (minutes)"

colnames(data)[3] = "Number of Live Cells (log2)"

tmp\_df = data # store reference

data$`Day of Observation` = as.factor(data$`Day of Observation`)

# Plots and table generation

scatter\_plot = ggplot(data, aes(x=`Time Exposed to Radiation (minutes)`,

y=`Number of Live Cells (log2)`,

show.legend = T)) +

geom\_point(size=2) +

ggtitle("Effect of Radiation on Cell Mortality") +

stat\_smooth(method = "lm", formula = y ~ x + I(x^2), size = 1) +

theme(plot.title = element\_text(hjust = 0.5))

scatter\_plot

data = tmp\_df # reset reference

kbl(data) %>%

kable\_classic(full\_width = F)

group\_means=by(data$`Number of Live Cells (log2)`, data$`Time Exposed to Radiation (minutes)`, t.test)

group\_means=matrix(c(unlist(group\_means[['0']][5:4]),

unlist(group\_means[['0.5']][5:4]),

unlist(group\_means[['1.0']][5:4]),

unlist(group\_means[['1.5']][5:4]),

unlist(group\_means[['2.0']][5:4]),

unlist(group\_means[['2.5']][5:4]),

unlist(group\_means[['3.0']][5:4]),

unlist(group\_means[['3.5']][5:4])),

nrow=8, ncol=3, byrow=T)

group\_means = data.frame(cbind(group\_means,c(0, 0.5, 1.0, 1.5, 2.0, 2.5, 3.0, 3.5)))

group\_means$X1 = round(group\_means$X1, digit=3)

group\_means$X2 = round(group\_means$X2, digit=3)

group\_means$X3 = round(group\_means$X3, digit=3)

colnames(group\_means)=c('Mean','Lower CL 95%','Upper CL 95%','Time Interval (minutes)')

ggplot(group\_means, aes(x=`Time Interval (minutes)`, y=group\_means[,1])) +

geom\_errorbar(aes(ymin=`Lower CL 95%`, ymax=`Upper CL 95%`), width=.1) +

geom\_line() +

geom\_point() +

expand\_limits(y=c(5, 20)) +

ylab("Count Live Cells (log2)") +

xlab('Time Intervals (minutes)') +

labs(title="Average Cell Count at Time Interval & 95% CIs") +

theme(text = element\_text(size=16), axis.text.x=element\_text(size=18), plot.title = element\_text(hjust = 0.5))

group\_means = group\_means %>% relocate(`Time Interval (minutes)`, .before = `Mean`)

kbl(group\_means) %>%

kable\_classic(full\_width = F)