Introduction

The skin cell data consists of 118 observations of human skin cells colonies. Samples were 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

Day of Observation	Time Exposed to Radiation (minutes)	Number of Live Cells (log2)
1	0.0	7.39232
1	0.0	9.67948
1	0.0	9.50978
1	0.0	12.15893
1	0.0	9.47371
1	0.5	5.08746
1	0.5	5.49185
1	0.5	8.83605
1	0.5	11.64971
1	1.0	3.45943
1	1.0	6.16993
1	1.0	1.58496
1	1.0	3.32193

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. The lowest mean survival rate and upper confidence interval can be seen at the 2-minute interval with a mean of 2.91(log2) and upper bound of 3.93 (log2) meanwhile the lowest lower bound confidence interval can be seen at the 2.5 min interval at 1.77 (log2).

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

Time Interval (minutes)	Mean	Lower CI 95%	Upper CI 95%
0.0	8.34	7.17	9.52
0.5	6.69	5.18	8.20
1.0	4.47	3.32	5.61
1.5	3.75	2.55	4.95
2.0	2.91	1.88	3.93
2.5	3.20	1.77	4.62
3.0	3.25	2.18	4.31
3.5	3.50	1.86	5.14

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:

$$y_i = \beta_0 + \beta_1(radiation_i) + \beta_2(\delta_{i(0.5)}) + \ldots + \beta_9(\delta_{i(3.5)}) + \beta_{10}(\delta_{i(0.5)} * radiation_i) + \ldots + \beta_{17}(\delta_{i(3.5)} * radiation_i) + \epsilon_i$$

where is day ??

Where y_i is the live cell count for the sample i , $\delta_{i(o.5)}$ = 1 if skin cell sample is exposed to radiation for 0.5 minutes and 0 otherwise, repeat for all time intervals. These are the regression line equations for each time interval that skin cells were exposed to radiation:

0.5-minute interval = $\beta_0 + \beta_1 * (radiation)$

1-minute interval = $(\beta_0 + \beta_2) + (\beta_1 + \beta_{12}) * (radiation)$

1.5-minute interval = $(\beta_0 + \beta_3) + (\beta_1 + \beta_{13}) * (radiation)$

2-minute interval = $(\beta_0 + \beta_4) + (\beta_1 + \beta_{14}) * (radiation)$

2.5-minute interval =$(\beta_0 + \beta_5) + (\beta_1 + \beta_{15}) * (radiation)$

3-minute interval = $(\beta_0 + \beta_6) + (\beta_1 + \beta_{16}) * (radiation)$

3.5-minute interval = $(\beta_0 + \beta_7) + (\beta_1 + \beta_{17}) * (radiation)$

very confusing model Presentation → there are 4 days and 8 times ⇒ 26 parameters would be required ??

A test of the null hypothesis H_0: $\beta_{12} = \beta_{17} = 0$ can be interpreted as a test of the common slopes assumption where the slope for 8-time intervals is the same but their intercepts ✗

No.

might differ. This hypothesis was tested using the standard formulation for general linear hypothesis, results in Table 3 and Table 4.

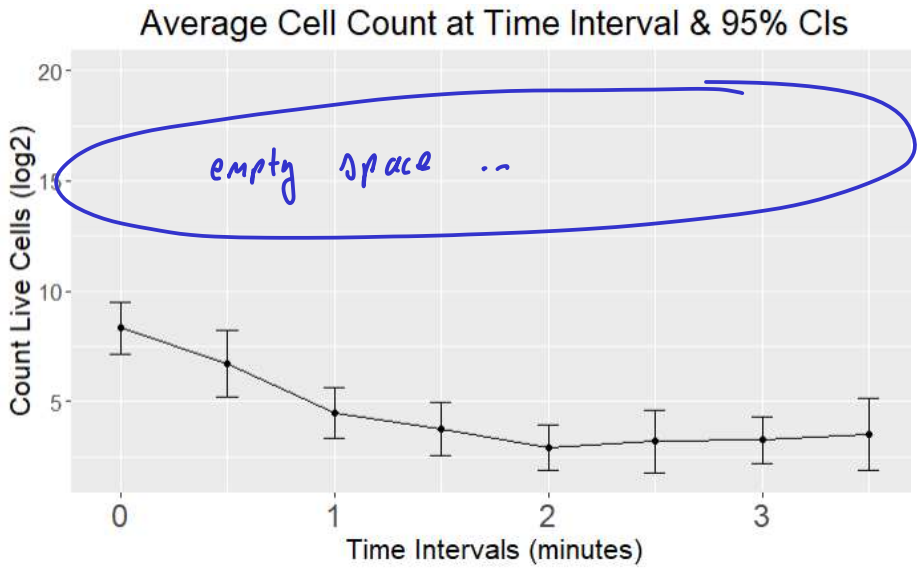


Figure 1: Mean difference by cell survival count with upper and lower 95% confidence

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 for a final model because there doesn't seem to be any positive interaction between day of observation and time exposed to radiation bar 1 outlier documented above. A quadratic effect of day of observation was also added which yielded a p-value of 0.0554 further proving that day of observation had no impact on the live cell count when paired with time exposed to radiation resulting in it being dropped from the model. The null hypothesis in this case in not rejected.

Model 2

Common slope model:

$$y_i = \beta_0 + \beta_1(radiation_i) + \beta_2(\delta_{i(0.5)}) + \ldots + \beta_9(\delta_{i(3.5)}) + \epsilon_i$$

Leads to the following regression line equations for each time interval skins cells are exposed to radiation:

0.5-minute interval = $\beta_0 + \beta_1 * (radiation)$

1-minute interval = $(\beta_0 + \beta_2) + \beta_1 * (radiation)$

1.5-minute interval = $(\beta_0 + \beta_3) + \beta_1 * (radiation)$

2-minute interval = $(\beta_0 + \beta_4) + \beta_1 * (radiation)$

2.5-minute interval = $(\beta_0 + \beta_5) + \beta_1 * (radiation)$

3-minute interval = $(\beta_0 + \beta_6) + \beta_1 * (radiation)$

3.5-minute interval = $(\beta_0 + \beta_7) + \beta_1 * (radiation)$

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. Then when comparing models by dropping 1 variable the best model yielded a p-value of 5.856e-11 on 7 and 110 degrees of freedom with an f-statistic of 11.584. The model summary indicates that exposing cell colonies to radiation for any amount of time shows a significant drop in live cell count.

Comparison among time intervals

The goal of this analysis was to find the effect of exposure to radiation at varying time intervals on skin cells. This analysis gives evidence to suggest that exposing cells to radiation for 3 minutes had the most significant effect with a p-value of 2.05e-09 and least significant at 0.5 minutes with a p-value of 0.046. However, when comparing differences between the baseline and other exposure times the 2-minute mark with a p-value of 5.20e-09 shows the biggest drop off in live cell count at 5.44 (log2). Meanwhile exposing skin cells to any form of radiation is harmful when compared to the baseline at 0 minutes as seen in Table 3.

Table 3: Most Lethal Exposure Time intervals

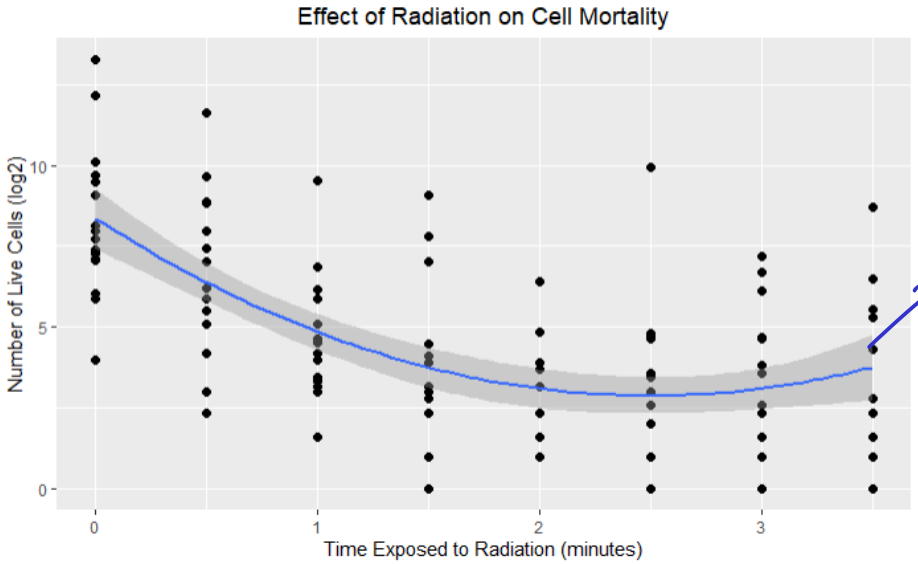
Comparison of time intervals	difference in live cell count (log2)	Standard Error	P-Value
0.5 to 0	1.66	0.82	0.473
1 to 0	3.88	0.81	< 0.001
1.5 to 0	4.59	0.78	< 0.001
2 to 0	5.44	0.86	< 0.001
2.5 to 0	5.15	0.82	< 0.001
3 to 0	5.1	0.78	< 0.001
3.5 to 0	4.85	0.86	< 0.001

The biggest differences of exposure lethality to skin cells where between the range of 1.5 minutes to 3.5-minute which can be seen in Table 4, other time intervals had p-values in the range of 0.6 to 1 leaving very little difference between them.

again - what model is being used?

Table 4: Biggest Difference in Lethal Exposure Time intervals

Comparison of time intervals	difference in live cell count (log2)	Standard Error	P-Value
1.5 to 0.5	-2.9	0.82	0.012
2 to 0.5	-3.8	0.89	0.001
2.5 to 0.5	-3.5	0.86	0.002
3 to 0.5	-3.4	0.82	0.001
3.5 to 0.5	-3.2	0.89	0.012



neither models 1 or 2 presented above include a quadratic effect of time?

Figure 2: Scatterplot of Cell Survival Count

There is little evidence to suggested that the variability of the lab environment had any impact on the result of this experiment. There seems to be a trend of diminishing returns once skin cells reach the 2.5-minute time interval as can be seen in Figure 2 above.

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
library(tidyverse) # data wrangling
library(gtsummary) # generate table for model results
library(multcomp) # needed for glht

setwd("~/GitHub/Masters-Classes-L-O/Applied Statistics/data_stats")
data = read.csv("skincells.csv")
data$logcells = round(data$logcells, digit=5)

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`)


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))


data = tmp_df
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']][5:4]),
                     unlist(group_means[['1.5']][5:4]),
                     unlist(group_means[['2']][5:4]),
                     unlist(group_means[['2.5']][5:4]),
                     unlist(group_means[['3']][5:4]),
                     unlist(group_means[['3.5']][5:4])),
                     nrow=8, ncol=3, byrow=T)
```

```
group_means
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=2)
group_means$X2 = round(group_means$X2, digit=2)
group_means$X3 = round(group_means$X3, digit=2)

colnames(group_means)=c('Mean','Lower CI 95%','Upper CI 95%','Time Interval (minutes)')

group_means

ggplot(group_means, aes(x=`Time Interval (minutes)`, y=group_means[,1])) +
  geom_errorbar(aes(ymin=`Lower CI 95%`, ymax=`Upper CI 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`)

plot(group_means$`Time Interval (minutes)`, group_means$Mean)

kbl(group_means) %>%
 kable_classic(full_width = F)


m1 = lm(data$`Number of Live Cells (log2)` ~ data$`Time Exposed to Radiation
(minutes)`+factor(data$`Day of Observation`)+data$`Time Exposed to Radiation
(minutes)`:factor(data$`Day of Observation`))

anovatab(m1)
drop1(m1,test='F')
summary(m1)
cov(m1)
confint(m1)

m1 %>% tbl_regression()

m1_a=lm(data$`Number of Live Cells (log2)` ~ factor(data$`Time Exposed to Radiation (minutes)`)
+ data$`Day of Observation`)
summary(m1_a)
m1_a = update(m1_a,.~.+I(data$`Day of Observation`^2))
drop1(m1_a, test='F')

m2=lm(`Number of Live Cells (log2)` ~ factor(`Time Exposed to Radiation (minutes)`), data =
data)
anovatab(m2)
drop1(m2,test='F')
summary(m2)
m2 %>% tbl_regression(label=c("factor(`Time Exposed to Radiation (minutes)`)" ~ "Time Exposed to
Radiation (minutes)"))
confint(m2)
glht1 = glht(m2, mcp("factor(`Time Exposed to Radiation (minutes)`)"="Tukey",
interaction_average=TRUE))
summary(glht1)
```

→ this would result in a model with 26 beta's - this is not presented in your report?

- how - day is continuous?

↓ you were explicitly instructed to model 'time' as a continuous predictor.

R Programme:	20
Statistical Modelling:	19
Presentation:	18
Total:	~~57%~~ exceeds page limit penalty → 47 %

Your coding and standard presentation is very good.

The statistical modelling presented is, at times, very confused.