MATH9952 Modern Applied Statistical Models
ASSIGNMENT MARKING SCHEME

Student No. _____ Name: Jerry Kiely

Assignment 1: SKIN CELLS

DATA ANALYSIS Ranking

R Programme [30]:	24

Comments:

V. good & code presented. Some code was not needed (and not discussed in your report). You might have calculated CI for the quadratic effect of time as well.

Q1: (DA) 38

Q2: 20

Statistical Modelling [50]:	35

Comments:

Good, but you needed to spend more time discussing radiation exposure — this was the focus of the experiment and should have been the focus of your report.

Q3: 30

total: 88

Presentation: [20]	16

Comments:

— good overall and reads well. —
You need to be more formal in structure in places and uses table/Figure numbers & titles..

Skin Cells

Jerry Kiely

26 February 2017

Introduction

Researchers in the Radiation and Environmental Science Centre are conducting research into the effect of solar radiation on the mortality of human skin cells. Colonies of human skin cells are placed in medium and transferred to wells on experimental plates. These plates are then exposed to radiation in a solar simulator for various amounts of time from 0 (control) to 3.5 minutes minutes. After exposure, the number of live cells in a sample are counted under a microscope and the total number of live cells in the colony is extrapolated from the sample result. The purpose of this experiment is to assess the effect of the varying times of exposure to radiation on cell death. Due to inherent variability in the response of cells, the researchers replicate their experiments a number of times. Their is also concern that environmental conditions within the lab from day to day may have an impact on the results. Therefore, the experiment is repeated over a number of days. Initially the experiment is designed with complete balance; i.e. the same number of replications at each exposure time on each day. But due to experimental loss, complete balance was not reflected in the final data (e.g. some plates were contaminated and had to be discarded).

The dataset is skincells.xls and is available on Webcourses. The variables are:

Variable	Description
day	The day (number code) that observation was recorded.
time	The amount of radiation exposure in minutes the colony was exposed to in the solar simulator.
logcells	The logarithm (base 2) of the number of live cells in the colony extrapolated from the sample result under the microscope.

No - it is time of exposure.

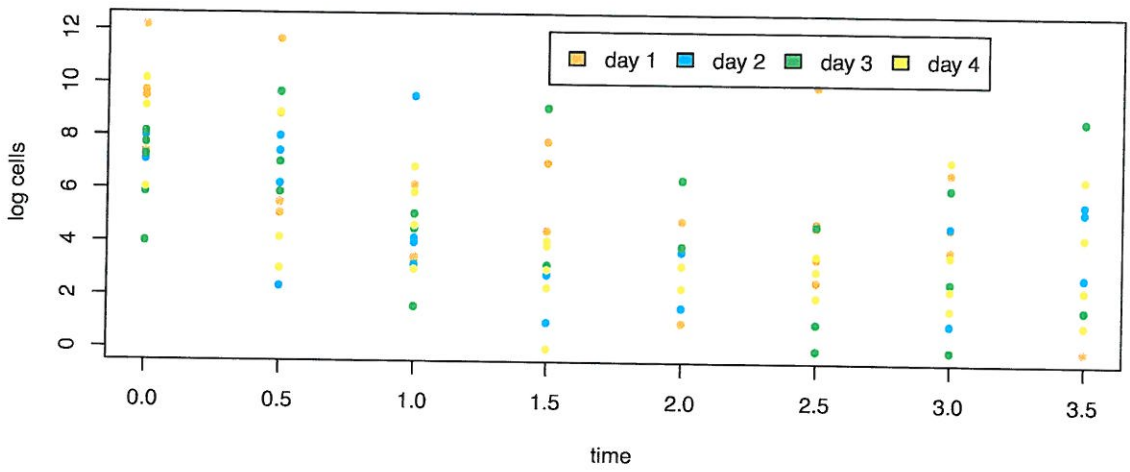
The central question is to establish whether the environmental conditions within the lab have an impact on the results. Below is a table of the means for each of the day, and the corresponding confidence intervals for each day:

Day	Mean	Lower 95%	Upper 95%
1	8.3201268	6.6060231	10.0342305
2	5.8858886	4.3942513	7.3775259
3	6.5115999	4.8604817	8.162718
4	6.8478244	5.1716733	8.5239755

Perhaps less decimal places to increase readability?

the above data would suggest that there is not much difference between the means, with maybe the means on day 1 and day 2 having the greatest difference.

use table numbers & titles?



colour choices?

The above scatterplot of the data with days having different colours would seem to confirm this.

use figure numbers.

First Model

Initially, a linear model was fitted with time as a continuous predictor, and day as a categorical predictor, with an interraction allowed between day and time.

what does this do? - explain.

$$y_i = \beta_0 + \beta_1(time_i) + \beta_2(\delta_{i2}) + \beta_3(\delta_{i3}) + \beta_4(\delta_{i4}) + \beta_5(\delta_{i2} \times time_i) + \beta_6(\delta_{i3} \times time_i) + \beta_7(\delta_{i4} \times time_i) + \epsilon_i$$

The null hypothesis may be stated as follows:

] of interaction

$$\begin{aligned} H_0 &: \quad \beta_5 = \beta_6 = \beta_7 = 0 \\ H_a &: \quad \beta_5, \beta_6, \text{ and } \beta_7 \text{ are different to } 0 \end{aligned}$$

The F-test statistic was found to be 0.4456585 on 3 and 110 degrees of freedom, yielding a p-value of 0.7208564. Therefore we fail to reject the null hypothesis and assume no interraction between the predictors (i.e. that the common slopes model is adequate).

Second Model

Secondly, a linear model was fitted with time as a continuous predictor, and day as a categorical predictor, with no interraction between day and time.

$$y_i = \beta_0 + \beta_1(time_i) + \beta_2(\delta_{i2}) + \beta_3(\delta_{i3}) + \beta_4(\delta_{i4}) + \epsilon_i$$

The null hypothesis may be stated as follows:

you should explain why this H_0 is relevant.

$$H_0: \quad \beta_2 = \beta_3 = \beta_4 = 0$$
$$H_a: \quad \beta_2, \beta_3, \text{ and } \beta_4 \text{ are different to } 0$$

The F-test statistic was found to be 2.4559368 on 3 and 113 degrees of freedom, yielding a p-value of 0.0667058. Therefore, although close, we fail to reject the null hypothesis and assume no significant difference between the days.

Third Model

Thirdly, a linear model was fitted with time as a continuous predictor, and a quadratic effect of time.

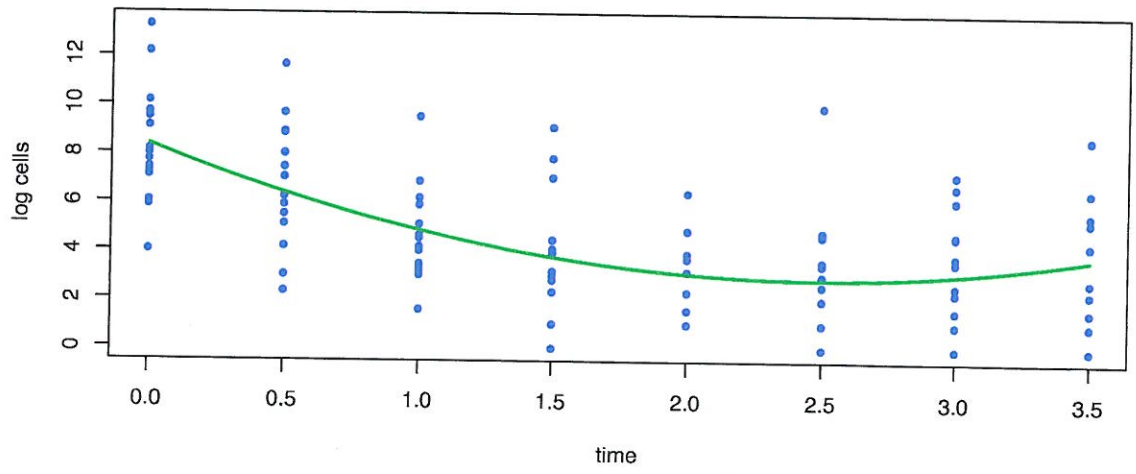
$$y_i = \beta_0 + \beta_1(time_i) + \beta_2(time_i{}^2) + \epsilon_i$$

The null hypothesis may be stated as follows:

$$H_0: \quad \beta_2 = 0$$
$$H_a: \quad \beta_2 \text{ is different to } 0$$

as per previous comment.

The F-test statistic was found to be 23.2011782 on 1 and 115 degrees of freedom, yielding a very small p-value of 0.00000449. Therefore we reject the null hypothesis and assume a quadratic effect.



The final plot of the data would seem to validate the research. ?

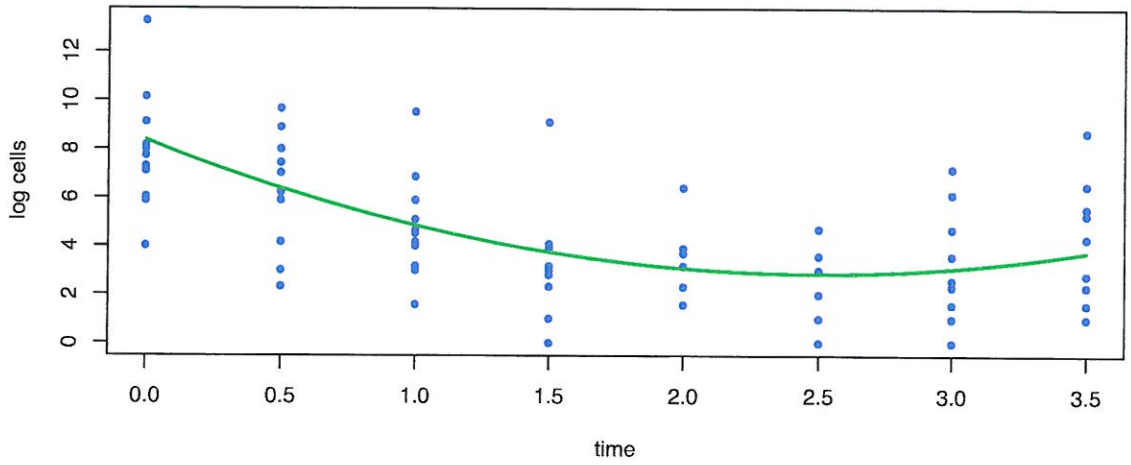
Comparison between days

– justify this ??

Although it is clear that day is not a predictor, there is grounds to exclude day 1 because the results on that day would seem to be significantly different from the results of the other days. Specifically, the differences between day2, day 3, and day 4 are not significant, but the difference between day 1 and day 2

may be statistically significant, with a p-value of 0.0248125. Therefore day 1 and day 2 may have significant differences, and perhaps day 1 should be excluded.

Below is a plot of the data excluding day 1, and with it's associated fitted line.



```r
# install and load required libraries

# install.packages("gdata")
# install.packages("multcomp")

library(gdata)
library(multcomp)


# set working directory

setwd("~/Workspace/College/DIT/MATH9952/Data")


# read in the data

data = read.xls("skincells.xls", header = T)
attach(data)


# split data by day and fit model for each day

fit_day1 = lm(logcells[day == 1] ~ time[day == 1], data = data)
fit_day2 = lm(logcells[day == 2] ~ time[day == 2], data = data)
fit_day3 = lm(logcells[day == 3] ~ time[day == 3], data = data)
fit_day4 = lm(logcells[day == 4] ~ time[day == 4], data = data)

coef(fit_day1)
coef(fit_day2)
coef(fit_day3)
coef(fit_day4)

confint(fit_day1)
confint(fit_day2)
confint(fit_day3)
confint(fit_day4)


# plot data by day

plot(time[day == 1], logcells[day == 1], pch = 20, col = 'orange', xlab = "time", ylab = "log cells")
points(time[day == 2], logcells[day == 2], pch = 20, col = 'cyan')
points(time[day == 3], logcells[day == 3], pch = 20, col = 'green')
points(time[day == 4], logcells[day == 4], pch = 20, col = 'yellow')
legend(1.5, 12, legend = c("day 1", "day 2", "day 3", "day 4"), fill = c("orange", "cyan", "green",
"yellow"), horiz = TRUE)


# fit full interraction model acroass all days and test

fit1 = lm(logcells ~ factor(day) * time, data = data)
summary(fit1)
anova(fit1)
drop1(fit1, test = 'F')


# drop interraction parameter and re-test

fit2 = update(fit1, ~. - factor(day):time)
summary(fit2)
anova(fit2)
drop1(fit2, test = 'F')


# drop day parameter and re-test

fit3 = update(fit2, ~. - factor(day))
summary(fit3)
anova(fit3)
drop1(fit3, test = 'F')


# fit empty model for test purposes

fit0 = lm(logcells ~ 1, data = data)
```

why ?

```
# test the fitted model by dropping parameters in all three directions - forward, backwards, and both

step(fit0, scope = list(lower = ~ 1, upper = ~ time * factor(day)), direction = "forward", trace = 1)
step(fit1, scope = list(lower = ~ 1, upper = ~ time * factor(day)), direction = "backward", trace = 1)
step(fit1, scope = list(lower = ~ 1, upper = ~ time * factor(day)), direction = "both", trace = 1)


# test days 2, 3, and 4 are equal to day 1

L = diag(5)[2:4,]
glh = glht(fit2, linfct = L)
summary(glh, test = Ftest())

# test day 2 is equal to day 3

L = matrix(c(0, 1, -1, 0, 0), nrow = 1)
glh = glht(fit2, linfct = L)
summary(glh, test = Ftest())

# test day 3 is equal to day 4

L = matrix(c(0, 0, 1, -1, 0), nrow = 1)
glh = glht(fit2, linfct = L)
summary(glh, test = Ftest())

# test day 2 is equal to day 4

L = matrix(c(0, 1, 0, -1, 0), nrow = 1)
glh = glht(fit2, linfct = L)
summary(glh, test = Ftest())

# test days 2 is equal to day 1

L = matrix(c(0, 1, 0, 0, 0), nrow = 1)
glh = glht(fit2, linfct = L)
summary(glh, test = Ftest())

# test days 3 is equal to day 1

L = matrix(c(0, 0, 1, 0, 0), nrow = 1)
glh = glht(fit2, linfct = L)
summary(glh, test = Ftest())

# test days 4 is equal to day 1

L = matrix(c(0, 0, 0, 1, 0), nrow = 1)
glh = glht(fit2, linfct = L)
summary(glh, test = Ftest())
```

hardly necessary here ?

also more easily available via summary(·) function.

Q2

$$Y = \left[X_A \mid X_B\right] \begin{bmatrix} \beta_A \\ \hline \beta_B \end{bmatrix} + \epsilon$$

$$= X_A \beta_A + X_B \beta_B + \epsilon$$

the underfitted model is

$$Y = X_A \beta_A + \epsilon$$

$$E[\hat{\beta}_A] = E\left[(X_A' X_A)^{-1} X_A' Y\right]$$

$$= (X_A' X_A)^{-1} X_A' E[Y]$$

$$= (X_A' X_A)^{-1} X_A' (X_A \beta_A + X_B \beta_B)$$

$$= (X_A' X_A)^{-1} X_A' X_A \beta_A + (X_A' X_A)^{-1} X_A' X_B \beta_B$$

$$\therefore E[\hat{\beta}_A] = \beta_A + A\beta_B$$

where $A = (X_A' X_A)^{-1} X_A' X_B$

Q2. $E[s^2] = E\left[\frac{SS_{error}}{n-p}\right]$

$$= \frac{1}{n-p} E\left[Y'(I-H_A)Y\right]$$

$$= \frac{1}{n-p}\left[trace\left[(I-H_A)\,VAR[Y]\right] + E[Y'](I-H_A)E[Y]\right]$$

$$= \frac{1}{n-p}\left[\sigma^2(n-p) + (X_A\beta_A + X_B\beta_B)'(I-H_A)'(I-H_A)(X_A\beta_A + X_B\beta_B)\right]$$

$$= \sigma^2 + \frac{((I-H_A)(X_A\beta_A + X_B\beta_B))'(I-H_A)(X_A\beta_A + X_B\beta_B)}{n-p}$$

$$= \sigma^2 + \frac{(X_B\beta_B - X_A(X_A'X_A)^{-1}X_A'X_B\beta_B)'}{n-p} \cdot \frac{(X_B\beta_B - X_A(X_A'X_A)^{-1}X_A'X_B\beta_B)}{n-p}$$

$$= \sigma^2 + \frac{((X_B - X_A(X_A'X_A)^{-1}X_A'X_B)\beta_B)'}{n-p} \cdot \frac{((X_B - X_A(X_A'X_A)^{-1}X_A'X_B)\beta_B)}{n-p}$$

$$\therefore E[s^2] = \sigma^2 + \frac{\beta_B'(X_B - X_A A)'(X_B - X_A A)\beta_B}{n-p}$$

20/20

Q3

$$Y = X_A \beta_A + \epsilon$$

The overfitted model is

$$Y = [X_A | X_B] \begin{bmatrix} \beta_A \\ \beta_B \end{bmatrix} + \epsilon$$

$$= X_A \beta_A + X_B \beta_B + \epsilon$$

$$E[\hat{\beta}] = E[(X'X)^{-1}X'Y] \quad \dots \quad X = [X_A | X_B]$$

$$= (X'X)^{-1}X' E[Y]$$

$$= (X'X)^{-1}X' (X_A \beta_A)$$

$$= (X'X)^{-1}X' \left(X \begin{bmatrix} \beta_A \\ 0 \end{bmatrix}\right)$$

$$E[\hat{\beta}] = \begin{bmatrix} \beta_A \\ 0 \end{bmatrix} = E\left[\begin{bmatrix} \hat{\beta}_A \\ \hat{\beta}_B \end{bmatrix}\right]$$

$$\therefore E[\hat{\beta}_A] = \beta_A$$

need to use Partitioned Matrix result.

yes, this is correct!

Q3 $$VAR[\hat{\beta}] = \sigma^2 \left(\begin{bmatrix} X_A' \\ X_B' \end{bmatrix} \begin{bmatrix} X_A & X_B \end{bmatrix} \right)^{-1}$$

$$= \sigma^2 \begin{pmatrix} X_A'X_A & X_A'X_B \\ X_B'X_A & X_B'X_B \end{pmatrix}^{-1}$$

from the identity provided, the top left

$$(X_A'X_A)^{-1}\left(I + X_A'X_B F X_B'X_A (X_A'X_A)^{-1}\right)$$

$$\begin{aligned} F &= \left[X_B'X_B - X_B'X_A (X_A'X_A)^{-1} X_A'X_B\right]^{-1} \\ &= \left[X_B'\left(X_B - X_A (X_A'X_A)^{-1} X_A'X_B\right)\right]^{-1} \\ &= \left[X_B'\left(I - X_A (X_A'X_A)^{-1} X_A'\right) X_B\right]^{-1} \\ &= \left[X_B'(I - H_A) X_B\right]^{-1} \end{aligned}$$

$\therefore$ expanding, we get

$$(X_A'X_A)^{-1} + (X_A'X_A)^{-1} X_A'X_B F X_B'X_A (X_A'X_A)^{-1}$$

$$(X_A'X_A)^{-1} + A\left[X_B'(I - H_A)X_B\right]^{-1} A'$$

$$\therefore VAR[\hat{\beta}_A] = \sigma^2 \left((X_A'X_A)^{-1} + A\left[X_B'(I - H_A)X_B\right]^{-1} A' \right)$$

X not asked for this - asked for $E[s^2]$??

(Q3) Compare with true regression model

$$VAR\left[\hat{\beta}_A\right] = \sigma^2 \left(X_A' X_A\right)^{-1}$$

$$\therefore A\left[X_B'\left(I - H_A\right)X_B\right]^{-1}A' = 0$$

$$\therefore X_A' X_B = 0 \quad \left(\text{and } X_B' X_A = 0\right)$$

NO
- still no, answer on orthogonality.

$$\therefore E\left[s^2\right] = E\left[\frac{SS_{error}}{n-p}\right]$$

$$= \sigma^2 + E\left[Y'(I-H)Y\right]$$

$$= \sigma^2 + E\left[Y'\right](I-H)E\left[Y\right]$$

$$= \sigma^2 + \left(X_A \beta_A\right)'(I-H)\left(X_A \beta_A\right)$$

$$= \sigma^2 + \left(X\begin{bmatrix}\beta_A \\ \hline 0\end{bmatrix}\right)'(I-H)\left(X\begin{bmatrix}\beta_A \\ \hline 0\end{bmatrix}\right)$$

$$= \sigma^2$$

$H = X(X'X)^{-1}X'$ yes

because $(I-H)\cdot X = 0$

$X = [X_A \mid X_B]$

(please see solutions.

10/10