MPulleyM2Q3

Melissa Pulley

2024-04-23

```r
cell = read.csv("protocelldata.csv")
cell$temp = as.factor(cell$temp)
cell$nutrients = as.factor(cell$nutrients)

summary(cell)
```

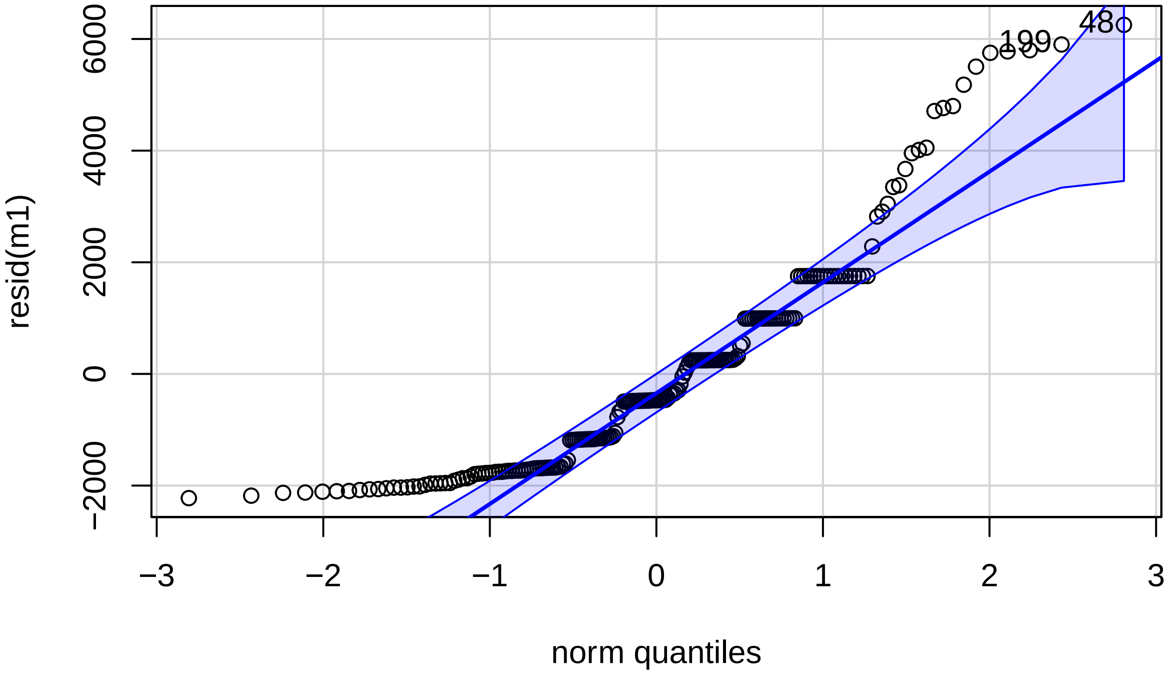
```
##       time         cellcount            temp     nutrients
##  Min.   : 1.0   Min.   :    5.042   cold:100   A:100
##  1st Qu.: 3.0   1st Qu.:   31.806   warm:100   B:100
##  Median : 5.5   Median :  248.060
##  Mean   : 5.5   Mean   : 1699.307
##  3rd Qu.: 8.0   3rd Qu.: 1848.318
##  Max.   :10.0   Max.   :11397.575
```

```r
library(car)
```

1. Test the assumptions of a linear model. Take into account that there might be an effect of the categorical variables on the response variable!

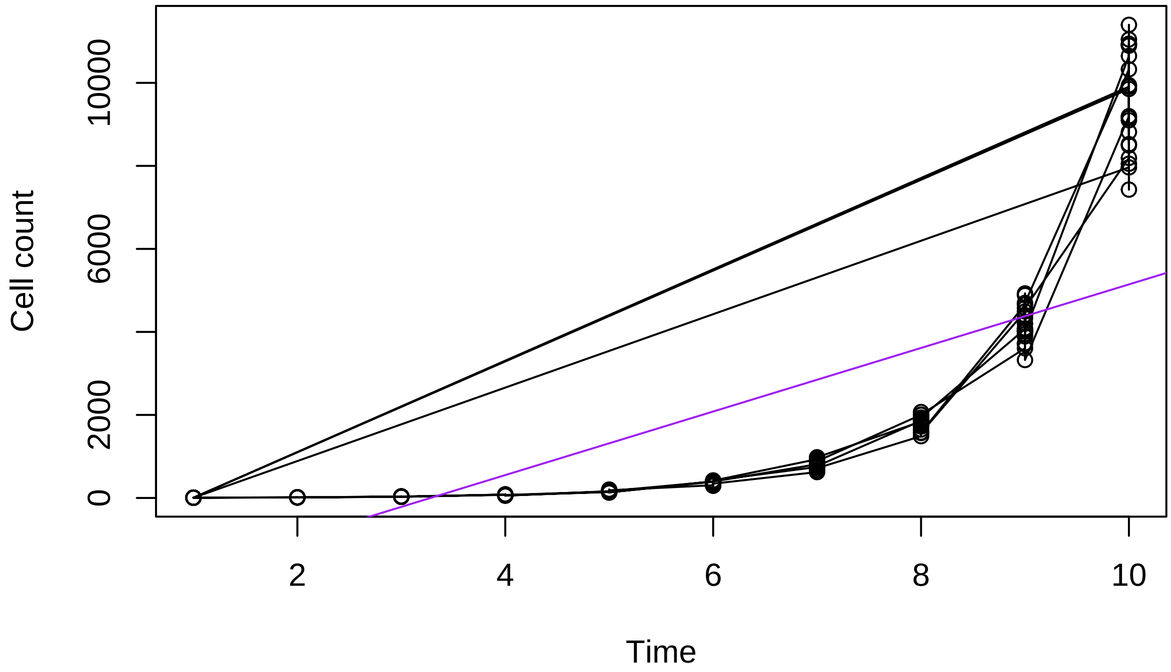
    a. The residuals are largely gathered close to the line near the center of the plot. Even though some of the residuals lie outside of the shaded region, we do have support that the errors are normally distributed.
    b. Since the data are not well-gathered along the linear model, we do not have significant support for linearity.
    c. The residual plot does not appear evenly distributed or symmetric around 0, so we do not have signicant support for the assumption of equal variance.
    d. Since the data is time related, the data cannot be independent. In the ACF plot, since many of the lag "spikes" have height that extend outside of the dotted line, they are statistically significant. So, we have support that cell count is highly correlated with time.

```r
m1 = lm(cellcount~time, data = cell)

#a1 - normality
qqPlot(resid(m1))
```

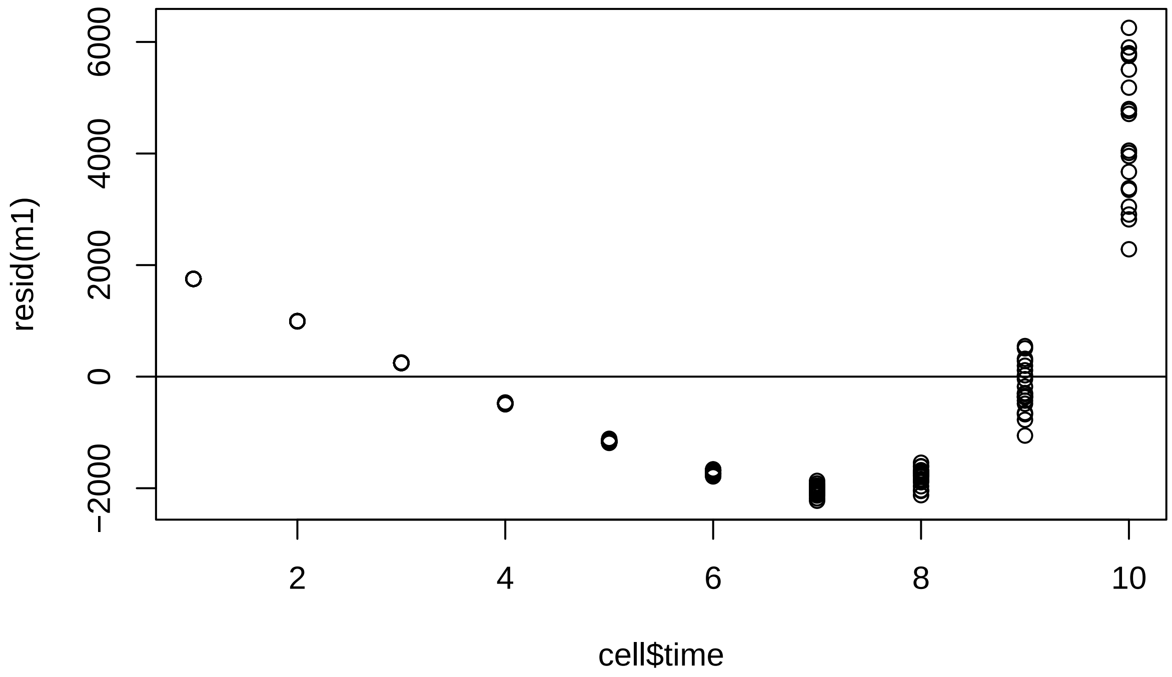


```
## [1]  48 199
```

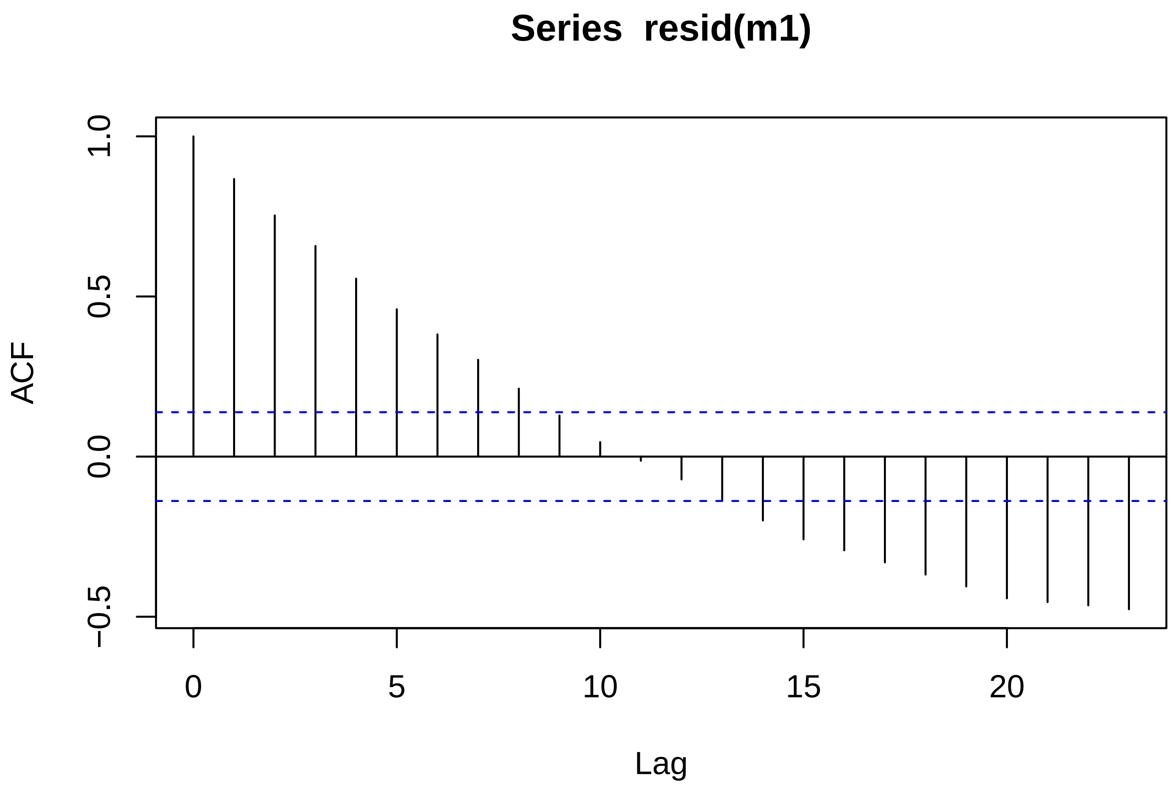
```
#a2 - linearity
plot(cellcount~time, ylab='Cell count', xlab= 'Time', cell)
lines(cellcount~time, data=cell)
abline(m1, col='purple')
```



```r
#a3 - equal variance
plot(resid(m1)~cell$time) # resid(model object) generates the residual values
abline(h=0)
```



```
#a4 - independence
acf(resid(m1))
```

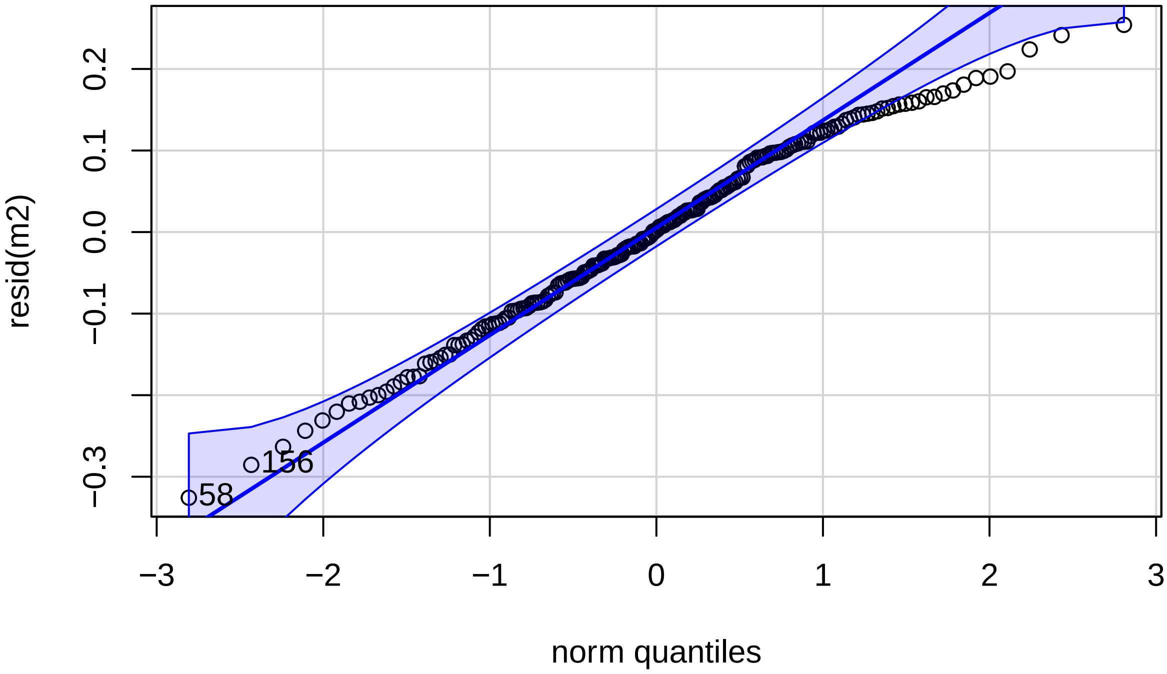


2. Make a decision based on the assumptions tests: You can make any modifications to the dataset, decide to do a GLM instead, or conclude that you are good to run with an multiple linear model. Whatever your decision is, state it, explain why, and take an action (or not, depending on the decision).

   In our evaluation of the linearity assumption, it appears that the data may be exponentially related. So, I have decided to apply a log transformation to the cell count data. Revisiting the assumptions with this new model, I am able to affirm my decision:

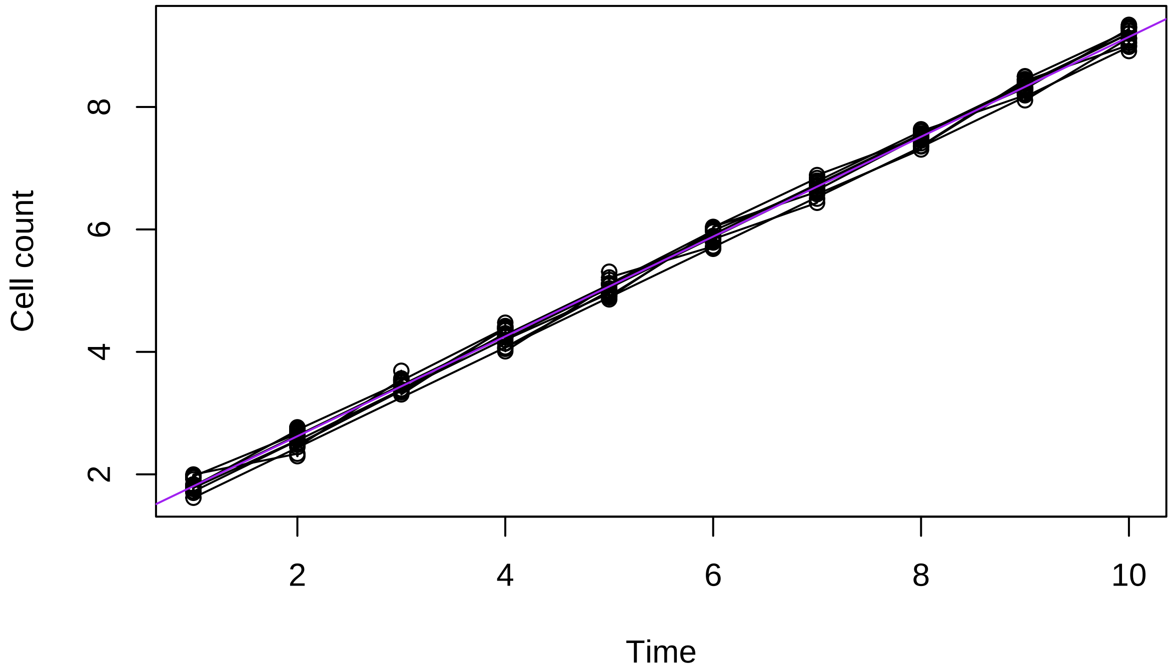
   a. The residuals are largely gathered close to the line near the center of the plot. Fewer residuals lie outside of the shaded region, so we do have support that the errors are normally distributed.
   b. Since the log-transformed data are well-gathered along the linear model, so we have significant support for linearity for the transformed model.
   c. The residual plot is evenly distributed and symmetric around 0, so we have support for the assumption of equal variance.
   d. Since the data is time related, the data cannot be independent. In the ACF plot, since many of the lag "spikes" have height that do not extend outside of the dotted lines, they are not statistically significant. So, we do not have support that transformed cell count data is highly correlated with time.

```r
cell$log_cellcount = log(cell$cellcount)

m2 = lm(log_cellcount~time, data = cell)

#a1 - normality
qqPlot(resid(m2))
```

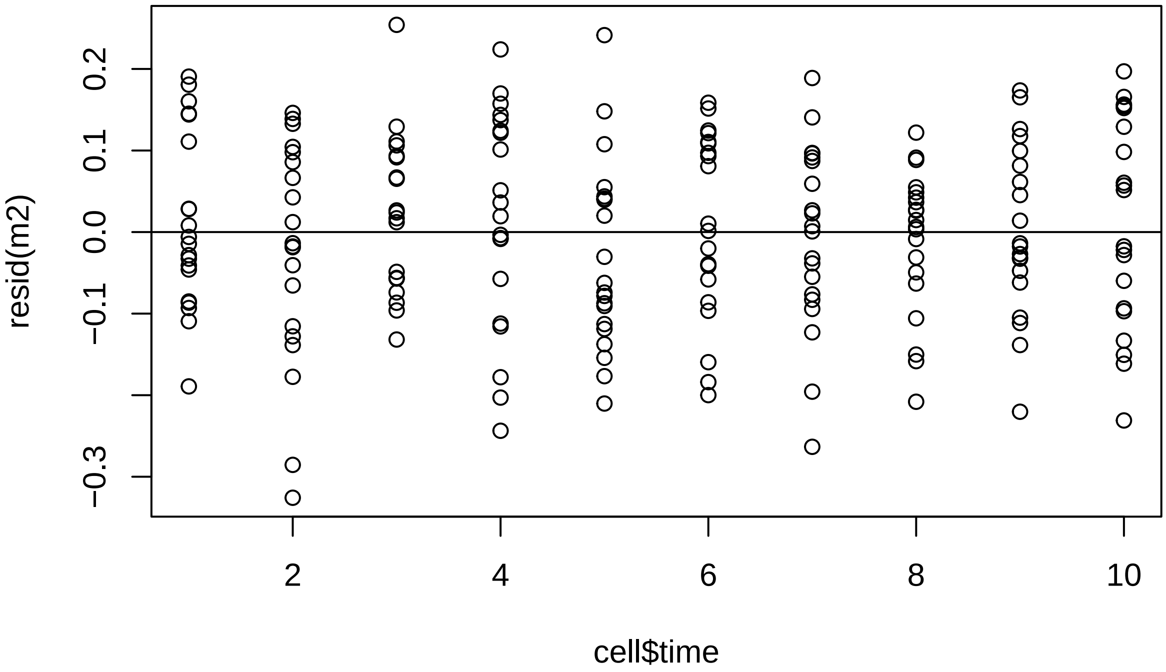


```
## [1]  58 156
```

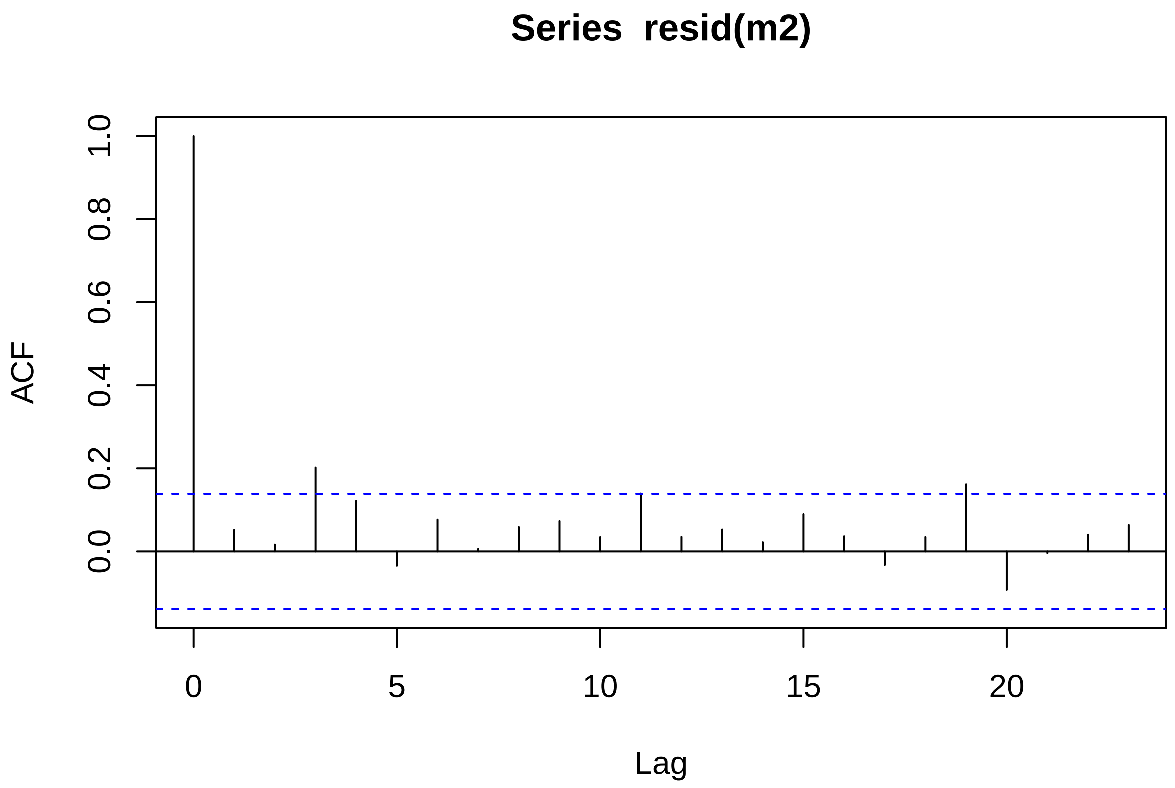
```
#a2 - linearity
plot(log_cellcount~time, ylab='Cell count', xlab= 'Time', cell)
lines(log_cellcount~time, data=cell)
abline(m2, col='purple')
```



```
#a3 - equal variance
plot(resid(m2)~cell$time)
abline(h=0)
```



```
#a4 - independence
acf(resid(m2))
```



3. Run at least 4 models looking at the potential effects of temperature and nutrients. Remember, at this point, you don't really know if anything has any effect! (upload the code)

```
m2 = lm(log_cellcount~time + temp, data = cell)
m3 = lm(log_cellcount~time + nutrients, data = cell)
m4 = lm(log_cellcount~time + temp + nutrients, data = cell)
m5 = lm(log_cellcount~time + temp*nutrients, data = cell)
m6 = lm(log_cellcount~time*temp*nutrients, data = cell)
```

4. Produce an AIC table, and define what the best model is

```
test = AIC(m2,m3,m4,m5,m6)
test
```

```
##    df       AIC
## m2  4 -296.3942
## m3  4 -309.2129
## m4  5 -309.1496
## m5  6 -307.2101
## m6  9 -304.4443
```

```
min(test$AIC)
```

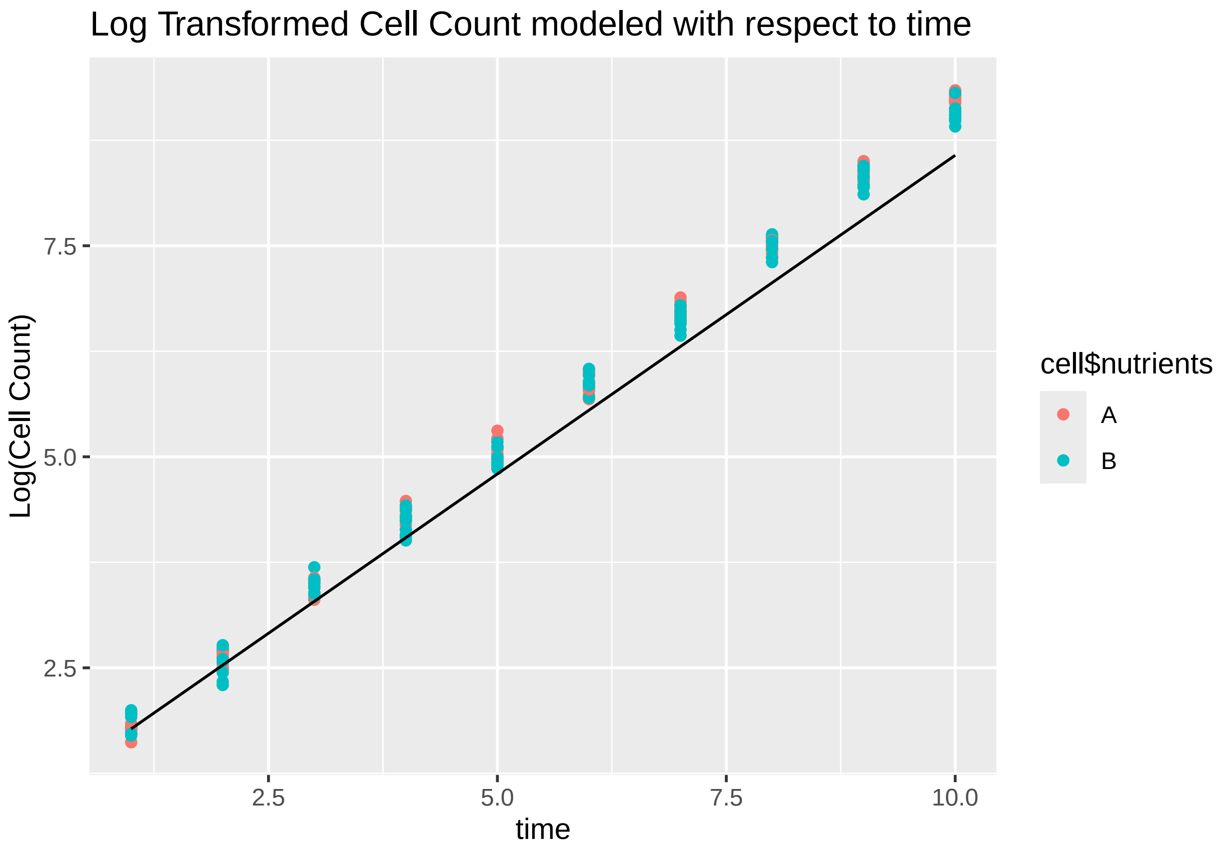
```
## [1] -309.2129
```

Model 3 has the lowest AIC of -309.2129.

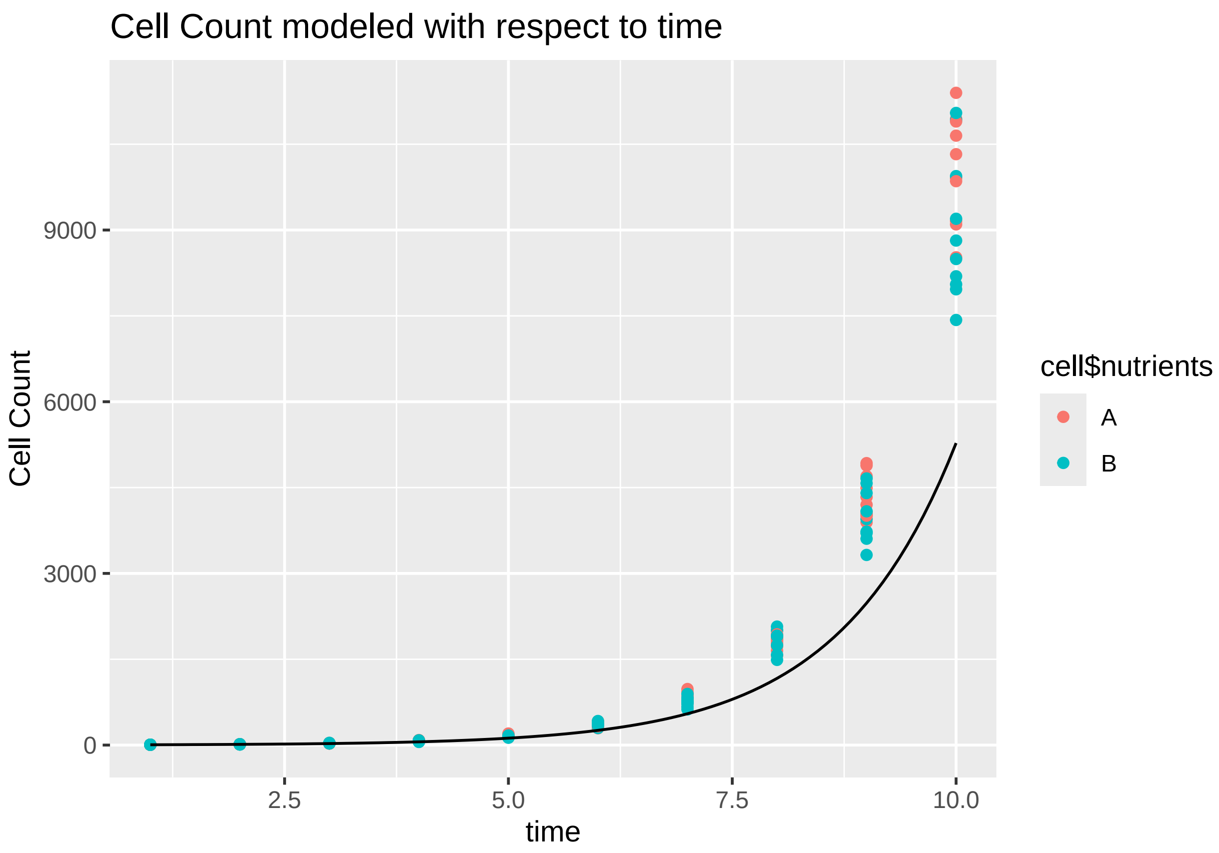
5. Describe and plot the best model.

The third model describes the relationship between the log transformed cell data with explanatory variables time and nutrient type with no interaction terms. Two plot are given. The first uses the log transformation, the second returns to the original values.

```
library(ggplot2)
library(dplyr)

#a2 - linearity

times = seq(min(cell$time), max(cell$time),length.out=100)

beta0 = m3$coefficients[1]
beta1 = m3$coefficients[2]
beta2 = m3$coefficients[3]

y= beta0 + (beta1 + beta2)*times


ggplot() + geom_point(aes(cell$time, cell$log_cellcount, colour=cell$nutrients)) +
  geom_line(aes(times,y)) + labs(title = "Log Transformed Cell Count modeled with respect to time") + xl
```



```
backy = exp(y)

ggplot() + geom_point(aes(cell$time, cell$cellcount, colour=cell$nutrients)) +
  geom_line(aes(times,backy)) + labs(title = "Cell Count modeled with respect to time") + xlab("time") +
```


Cell Count modeled with respect to time
Cell Count
9000
6000
3000
0
2.5
5.0
7.5
10.0
time
cell$nutrients
A
B