Statistics 101B HW 1

Earle Aguilar
(804501476)
January 13, 2017

1. Fundamental Principles

The first fundamental principle, randomization, was applied by randomizing the run orders and by both researchers having used different random orders. The second fundamental principle, replication, was utilized when both researchers conducted the experiment independently and used the same cell culture.

2. Analysis

All 36 Trials

For the analysis of all trials I decided to create a preliminary model which includes all the drugs. The model suggests that a high level dosage of D is significant with a p-value of 8.61×10^{-9}, and both the diluted and high level dosages of E are significant with p-values 6.51×10^{-5} and 2.88×10^{-6} respectively. Examination of the diagnostic plots indicates that there is a strong trend line in the residual plot suggesting that the model is not fitting the data correctly.

In order to find the simplest model the backwards step-wise regression algorithm and Bayes Information Criteria (BIC) were used to find a good fit. The backwards step-wise regression algorithm shows that BIC is minimized when 4 variables are used and the algorithm suggests that the high dosage and diluted versions of drugs D and E are the most significant.

I used this information to create a new model which only used D and E. The new model shows that all the variables are significant. The diluted and high dose D have p-values 0.0286 and 1.17×10^{-9} respectively and the diluted and high dose E have p-values 1.49×10^{-5} and 1.05×10^{-6} respectively. This model states that if we hold all other variables constant then a 1 unit increase in $D = 0$ results in a 13.453% decrease in infected cells on average, $D = 1$ results in a 39.068% decrease in infected cells on average, $E = 1$ results in a 30.321% decrease in infected cells on average, and $E = 1$ results in a 27.379% decrease in infected cells on average.

Then I examined the diagnostics plots to understand how my model is incorrect. The residual plot has a trend line indicating that the data has not been fit by the model. The normal Q-Q plot indicates that the data is not following the normal distribution. The scale-location plot shows that the covaraince is constant. All the points in this model contain high leverage points but none of these points are 2 or more standardized residuals away, additionally there are no high influation points.

First 16 Trials

A preliminary model of these 16 trials was created using all 5 drugs. This preliminary model indicates that a high dose of D and E are significant with p-values 5.34×10^{-5} and 0.00278 respectively. This residual plot for this model has a quadratic trend, therefore this model is not a good fit.

The backwards step-wise regression algorithm and Bayes Information Criteria (BIC) were used to find a good fit. BIC is minimized when 2 variables are used, and the algorithm suggests that high doses of drugs D and E are significant.

A new model was created using D and E.The diluted and high dose D have p-values 0.0286 and 1.17×10^{-9} respectively and the diluted and high dose E have p-values 1.49×10^{-5} and 1.05×10^{-6} respectively. Therefore

these variables are significant. This model states that if we hold all other variables constant then a 1 unit increase in $D = 1$ results in a 41.956% decrease in infected cells on average, and $E = 1$ results in a 24.644% decrease in infected cells on average. The residual plot has a trend line again indicating that the data has not been fit by the model. The normal Q-Q plot shows that the data is not following the normal distribution, however we have fewer than 30 observations. The scale-location plot indicates that the variance is constant.

Last 18 Trials

A preliminary model of the last 18 trials which utilized all 5 drugs indicates that a high dosage level of D is significant (p-value 0.0001538) and a both a high and diluted doses of E are significant with p-values 0.002628 and 0.003087 respectively. This model's residual plot also has a quadratic trend line which is indication of lack of fit.

Bayes information criteria is minimized when 4 variables are used and the backwards step-wise algorithm suggest that the high dosage and diluted versions of drugs D and E are the most significant.

Using a new model consisting of D and E was created however this model indicates that a diluted version of drug D is not significant. The high level dose has p-value of 0.00313 and the high and low level doses of E have p-vales 0.000705 and 0.00897 respectively. This model states that if we hold all other variables constant then a 1 unit increase in $D = 1$ results in a 35.217% decrease in infected cells on average, $E = 0$ results in a 31.967% decrease in infected cells on average, and $E = 1$ results in a 31.025% decrease in infected cells on average. The residual plot of this model has a trend line indicating this model is not the best fit for the data. The normal Q-Q plot shows that the data is following the normal distribution. The scale-location plot show a slight trend indication potential non-constant variance. Points 17, 21, and 32 are close to 2 standardized residuals away and could be points of high leverage.

3. Conclusion

The three final models indicate the drugs D and E are significant in reducing the percentag of infected cells suggesting that they are effective for treating HSV-1. More specifically the diluted version of drug D is the least significant since it is missing from the model of the first 16 trials and it has a p-value greater than 0.05 in the model of the last 18 trials. The high doses of D and E have high significance and the diluted dose of E is present in the analysis of all 36 trials and the last 18 trials. This seems to suggest that a a combination of the high and low dose of E and a high dose of D are the best choice if the goal is to reduce the percent of infected cells. The models produced all showed a lack of fit with the data, however a transformation can assist in creating a linear relationship.