Stats 101A Homework 1
Lucy Shao
304575686
Lecture 1
Disc 1B

Report

1. How were the fundamental principles applied in the study?

The fundamentals principles applied are Randomization, Replication, and Blocking.

Randomization is applied by randomizing the experiment materials and the order of the individual run. To be more specifically, in this experiment, The first 16 runs form a half-fraction design and the last 18 runs form an orthogonal array both have the individual run randomized, or without intended order, ant those runs also have the antiviral drugs, or the experiment material randomized. Also, Randomization also assists balancing the extraneous factors such as variation of experimental environment, as mentioned in the experiment "the same cell culture". Randomization makes the assumption of independent randomly distribution of the observation valid.

Another principle that is critical to our experiment is Replication. In this experiment, the two researchers conducted the experiment independently using the same cell culture, yielding two independent repeated run of each factor combination. Replication allows the researchers to avoid some measurement error, and it allows people to check wether the results from the two trials are statistically different. In this experiment, the overall data difference is small, but for trial 14, the two replicates yields an obvious different results of 42.2 and 23.2 with large standard deviation. Replication also allow the sample mean to estimate the true mean. Since there are two replicates, the experimental error are expected to be reasonably smaller than only one replicate, but more replicates are needed for better accuracy.

Blocking is an other technique that is used to improve the precision. In this experiment, blocking is used to reduce the variability result from "nuisance factors", that is factors such as experimental environment controlled by the "same cell culture". The same cell culture ensures the researchers to have the same experimental temperature, in other words it allows the researchers to those factors could affect the experimental result that the researchers are not directly interested. Those three principles is crucial for the experiment's validity.

2. Analyze the data using all 34 runs. Analyze the data using the first 16 runs only. Analyze the data using the last 18 runs only.

One of the problems is how to handle the two replicates. The replicates are serving to reduce measurement variability in order to find the result that is closest to the true mean, so it is reasonable for us to find the standard deviation of the two replicates. If the standard deviation is not too large, then we can use the mean of the two replicates to generate results. In addition we

should test wether the two replicates have the same mean using pooled t-test assuming equal variances for the two replicates.

First we read the table in R. Check wether the two measurements from the two researchers has the same mean. To test the null hypothesis that the two replicates have the same mean with the alternative hypothesis that the two replicates does not have the same mean, we have got an pooled variance of 656.339 and a t-score of 0.006485216, and we fail to reject the null hypothesis and conclude that the two replicates have the same mean. Since there is only two replicates in this experiment, the distribution and variance of the two replicates for each trial does not contribute to the analyze significantly, so we simply take the average of the two replicates for each trial and use this data for linear regression analysis.

First we use all 34 runs for regression analysis to determine which drugs statistically significantly contributed to the study. First we test the correlation between the five drugs, and we get the result of zero correlation for every of them. Then we use the average of the two replicates as the dependent variable, and using the five drugs as the independent variables initially. It yields the model result=25.8515- 0.6679A- 3.9946B- 2.7054C- 19.5339D-13.6893E. From the summary of the model, we have some predictors with large p-value that does not seems significant for the dependent variable. Check wether multicollinearity exists by calculating the variance inflater factor, and the results are all smaller than 5 which indicates there are no multicollinearity in this regression analysis. Now, we try to delete the predictors that are not significant to our model by the method of backwardAIC. The method of backward AIC, or Akaike's information criterion with backward selection, eliminates insignificant predictors step by step and yields the model of result =25.851 -3.995B -19.534D -13.689E. However, we observed that the R-squared does not improve, so we use the old model. Then we check the linear assumptions for the old model by looking at the graphs. The Q-Q plot is attached to the normal line and it shows that there is not a big problem with normality of the data. The residual plot vs. fitted plot does not have a completely straight trend but it shows the errors are independent and relations are linear. Also the standardized residual plot tells there is a constant variance, ant it also shows a few points are outliers by cooks distance. The assumptions are better satisfied after eliminating the predictors A and B. The overall assumptions are somehow satisfied for a experiment with only 34 runs.

Then we use the first 16 runs for regression analysis to determine which drugs statistically significantly contributed to the study. First we subset the data into the first 16 runs. Then we use the average of the two replicates as the dependent variable, and using the five drugs as the independent variables initially. It yields the model result=27.8469- 2.9906A- 2.5469B- 0.8156C-20.9781D-12.3219E. From the summary of the model, we have some predictors with large p-value that does not seems significant for this linear model. Check wether multicollinearity exists by calculating the variance inflater factor, and the results are all smaller than 5 which indicates there are no multicollinearity in this regression analysis. Now, we try to delete the predictors that are not significant to our model by the method of backwardAIC. The method of backward AIC, eliminates insignificant predictors step by step and yields the model of result =27.85 -20.98B -

12.32E. This is the final model with all predictors to be significant. However, we observed that the R-squared does not improve, so we use the old model. Then we check the linear assumptions for the old model by looking at the graphs. The Q-Q plot is somehow attached to the normal line and it shows the normality. The residual plot vs. fitted plot does not have a completely straight trend but it still shows the errors are independent and relations are almost linear. Also the standardized residual plot tells there is a constant variance, and the residuals vs leverage plot shows a few points are outliers by cooks distance. The overall assumptions are basically satisfied for a experiment with only 16 runs.

Last we use last 18 runs for regression analysis to determine which drugs statistically significantly contributed to the study. We use the average of the two replicates as the dependent variable, and using the five drugs as the independent variables initially. It yields the model result=22.569+4.692A- 3.662B- 2.962C- 15.346D-13.250E. From the summary of the model, we have some predictors with large p-value that does not seems significant for the dependent variable. Check wether multicollinearity exists by calculating the variance inflater factor, and the results are all smaller than 5 which indicates there are no multicollinearity in this regression analysis. Now, we try to delete the predictors that are not significant to our model by the method of backwardAIC. The method of backward AIC eliminates insignificant predictors step by step and yields the model of result =22.039 +5.487A -14.551D -12.455 E. However, we observed that the R-squared does not improve, so we use the old model. Then we check the linear assumptions for the old model by looking at the graphs. The residual plot vs. fitted plot does not have a completely straight trend but it still shows the errors are independent and relations are almost linear. The QQ-plot with the dots slightly deviated from the normal line. Also the standardized residual plot tells there is not likely a constant variance, and the residuals vs leverage plot shows a few points are outliers by cooks distance. The overall assumptions are somehow satisfied for a experiment with only 18 runs.

3. Draw conclusions (and possible recommendations). Which drugs are effective for treating the HSV-1?

For all three analysis, after eliminate the predictors that does not seem significant, the R-squared dose not improve. This indicates that even those predictors does not seem significant still contribute to the result, but we can still tell the most effective drugs from the regression. From the three analysis, by looking at the p-value in the model summary, Drug D and Drug E are most effective drugs for treating the HSV-1.

However, the experiment can be improved by having more replicates. With only two replicates, if one of the researcher made an significant measurement error, the test result could be errant. I would also suggest the experiment to have more runs of different drug dosage composites to better improve the regression model.