**Problem Set 2 - Problem 2**

Nick Handelman
4/25/2018

# 1 Problem 2

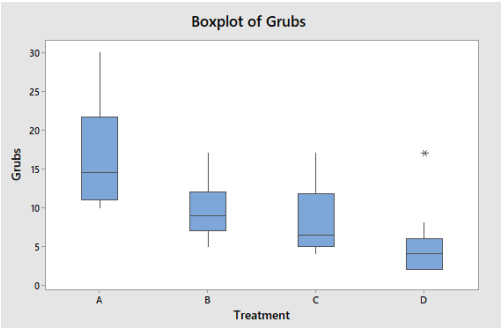
## 1.1 Question 1 - One-Way ANOVA on Treatment

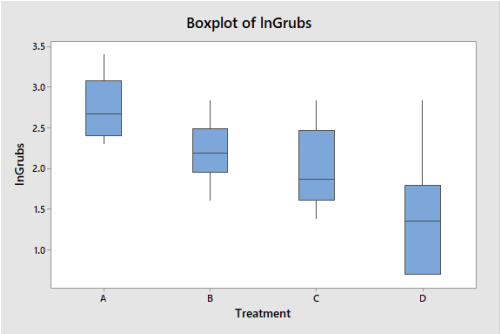
Is there a difference in the four treatments relative to their average counts?

### 1.1.1 Hypotheses

$H_0 : \mu_A = \mu_B = \mu_C = \mu_D$
$H_a$ : at least one of the means differs from the others

### 1.1.2 Boxplot and ANOVA Table




Analysis of Variance

| Source | DF | Adj SS | Adj MS | F-Value | P-Value |
|---|---|---|---|---|---|
| Treatment | 3 | 857.9 | 285.97 | 12.85 | 0.000 |
| Error | 44 | 978.9 | 22.25 | | |
| Total | 47 | 1836.8 | | | |

### 1.1.3 Choice of Statistical Technique

In the untransformed box plot (Boxplot of Grubs), all of the groups are approximately normal (though C is skewed more so), but A appears to have a larger variance. It appears that the variance increases somewhat with increasing mean so I performed a log transformation on the data. The transformed boxplot (Boxplot of lnGrubs) does appear to create more normality within the groups and it accounts for the outlier in group D in the untransformed data. However, it doesn't fix the unequal variance issue, so I ran the ANOVA on the untransformed data. From the problem description, I see no reason to think that independence within or between groups has been violated.
I ran a test for equal variances on the untransformed data and the p-values of the multiple comparisons test and Levene's test were 0.352 and 0.406, respectively. At a 95% confidence level, I did not reject the equal variances test null hypothesis that all variances are equal. So, I ran the One-Way ANOVA assuming equal variances.

### 1.1.4 Conclusions

With a 95% confidence level and a p-value of less than 0.001, I reject $H_0$ and accept that at least one of the group means differs from the others. The treatments are assigned randomly among the test sites in each plot, so it is acceptable to make the claim that the different treatments are the cause of rejecting $H_0$.

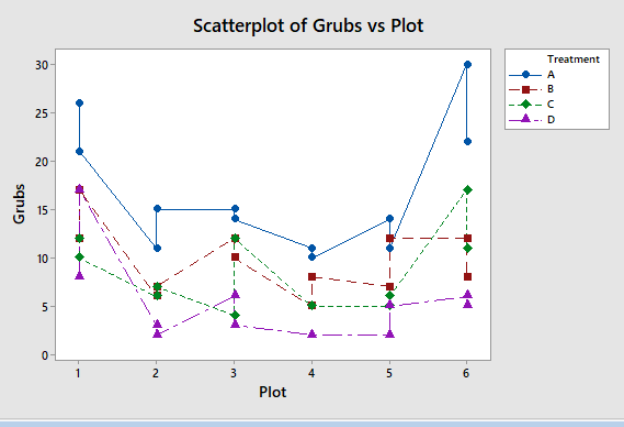
## 1.2 Question 1 - Multiple Regression

Is there a difference in the four treatments relative to their average counts?

### 1.2.1 Hypotheses

$H_0 : \mu_{log(A)} = \mu_{log(B)} = \mu_{log(C)} = \mu_{log(D)}$
$H_a$ : at least one of the means differs from the others

### 1.2.2 Visual Analysis

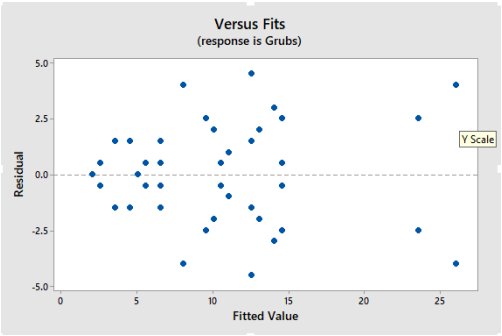



The coded scatterplot indicates that the test sites where treatment A was applied were almost always associated with larger grub counts. For treatments B and C the distinction isn't quite as clear, though the test sites on which treatment D was applied typically yielded the smallest amount of grubs. The group variances are all fairly close and only one point (plot 6, point at the top) gives any indication of possibly being an outlier. So, I decided on applying a multiple regression with two categorical variables (equivalently, a two-way ANOVA).

### 1.2.3 Multiple Regression and Evaluation of the Saturated Model

**Analysis of Variance**

| Source | DF | Seq SS | Contribution | Adj SS | Adj MS | F-Value | P-Value |
|---|---|---|---|---|---|---|---|
| Regression | 23 | 1635.3 | 89.03% | 1635.3 | 71.101 | 8.47 | 0.000 |
| Plot | 5 | 587.7 | 31.99% | 414.7 | 82.933 | 9.88 | 0.000 |
| Treatment | 3 | 857.9 | 46.71% | 188.4 | 62.792 | 7.48 | 0.001 |
| Plot*Treatment | 15 | 189.7 | 10.33% | 189.7 | 12.649 | 1.51 | 0.180 |
| Error | 24 | 201.5 | 10.97% | 201.5 | 8.396 | | |
| Total | 47 | 1836.8 | 100.00% | | | | |



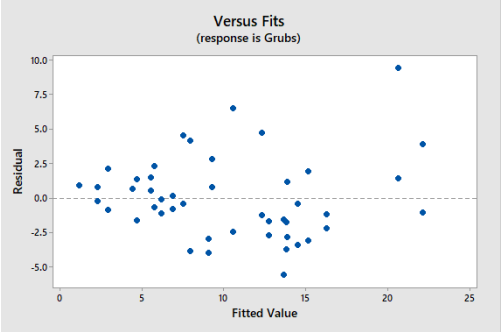

The p-value of 0.180 for the interaction term does not provide significant evidence that the coefficient is not 0. For each individual interaction level, Minitab performs a T-test to determine if there is enough evidence that each level's coefficient is not 0. Only Plot 6, Treatment D had a coefficient that the evidence suggests isn't 0.
Tests of the case influence statistics for each observation revealed two possible outliers. Both were marked as "unusual residuals" and were the residuals for both observations in Plot 1 Treatment D. However, their studentized residuals of 2.2 and -2.2 aren't particularly egregious and their Cook's Distance of 0.2 isn't close to 1. So, there don't appear to be any particularly influential observations.
The symmetric nature of the residual plot isn't important. It just indicates that the model was fit with 2

observations at each of the interaction levels. There does appear to be a slight funnel shape: the residuals are increasing with increasing as the observation value increases. This suggests a transform may be in order, but first I continued with the additive model.

### 1.2.4 Multiple Regression and Evaluation of the Additive Model

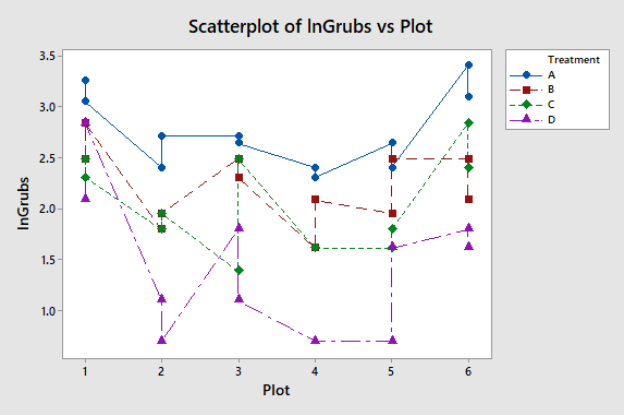
Analysis of Variance

| Source | DF | Seq SS | Contribution | Adj SS | Adj MS | F-Value | P-Value |
|---|---|---|---|---|---|---|---|
| Regression | 8 | 1445.6 | 78.70% | 1445.6 | 180.698 | 18.01 | 0.000 |
| Plot | 5 | 587.7 | 31.99% | 587.7 | 117.537 | 11.72 | 0.000 |
| Treatment | 3 | 857.9 | 46.71% | 857.9 | 285.965 | 28.51 | 0.000 |
| Error | 39 | 391.2 | 21.30% | 391.2 | 10.032 | | |
| Lack-of-Fit | 15 | 189.7 | 10.33% | 189.7 | 12.649 | 1.51 | 0.180 |
| Pure Error | 24 | 201.5 | 10.97% | 201.5 | 8.396 | | |
| Total | 47 | 1836.8 | 100.00% | | | | |




The p-values for Plot and Treatment indicate that their associated coefficients aren't 0. In this respect, the additive model looks adequate. One of the observations associated with Plot 6 Treatment A has a large studentized residual of 3.29, though its Cook's Distance of 0.28 is acceptable. The primary issue is the funnel shape of the residuals in the residual plot. Since the additive model did not correct the issue, a log transformation is in order.

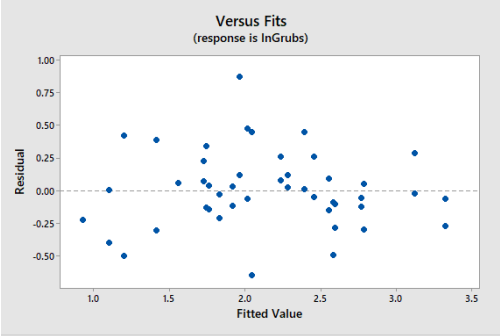
### 1.2.5 Multiple Regression and Evaluation of the Log Transformed Additive Model




While I hadn't initially considered the log transform necessary, it does appear to somewhat improve the equality of the variances among the groups. Thus, multiple regression will still provide valid conclusions on the transformed data and will hopefully provide a better fit than on the untransformed data.

**Analysis of Variance**

| Source | DF | Seq SS | Contribution | Adj SS | Adj MS | F-Value | P-Value |
|---|---|---|---|---|---|---|---|
| Regression | 8 | 17.788 | 82.24% | 17.788 | 2.22344 | 22.57 | 0.000 |
| Plot | 5 | 6.474 | 29.93% | 6.474 | 1.29490 | 13.14 | 0.000 |
| Treatment | 3 | 11.313 | 52.30% | 11.313 | 3.77102 | 38.28 | 0.000 |
| Error | 39 | 3.842 | 17.76% | 3.842 | 0.09852 | | |
| Lack-of-Fit | 15 | 1.474 | 6.81% | 1.474 | 0.09826 | 1.00 | 0.489 |
| Pure Error | 24 | 2.368 | 10.95% | 2.368 | 0.09868 | | |
| Total | 47 | 21.630 | 100.00% | | | | |




The ANOVA table indicates that Plot and Treatment are significant factors in estimating the mean of the natural log of grubs. More importantly, over the previous models, is that the residual plot lacks any apparent pattern. This is a strong indication that the regression of plot and treatment on the natural log of grubs provides a better fit than regression on grubs.

### 1.2.6 Conclusions

**Coefficients**

| Term | Coef | SE Coef | 95% CI | T-Value | P-Value | VIF |
|---|---|---|---|---|---|---|
| Constant | 3.321 | 0.136 | (3.046, 3.596) | 24.44 | 0.000 | |
| Plot | | | | | | |
| 2 | -0.868 | 0.157 | (-1.186, -0.551) | -5.53 | 0.000 | 1.67 |
| 3 | -0.553 | 0.157 | (-0.871, -0.236) | -3.52 | 0.001 | 1.67 |
| 4 | -1.041 | 0.157 | (-1.358, -0.723) | -6.63 | 0.000 | 1.67 |
| 5 | -0.769 | 0.157 | (-1.086, -0.451) | -4.90 | 0.000 | 1.67 |
| 6 | -0.204 | 0.157 | (-0.521, 0.113) | -1.30 | 0.201 | 1.67 |
| Treatment | | | | | | |
| B | -0.538 | 0.128 | (-0.797, -0.279) | -4.20 | 0.000 | 1.50 |
| C | -0.728 | 0.128 | (-0.987, -0.469) | -5.68 | 0.000 | 1.50 |
| D | -1.358 | 0.128 | (-1.618, -1.099) | -10.60 | 0.000 | 1.50 |

Since the regression is on log transformed response data, it is necessary to make inferences on the ratio of medians instead of on the change in means. Consider the coefficients of treatments B, C and D (A is the reference treatment). B's coefficient of -0.538 indicates it decreases the median number of grubs in a test site by about half, when compared to A. Similar arguments apply to the coefficients of C and D. Since the p-values associated with each of the coefficients are all less than 0.001, there is strong evidence that population median ratio isn't 0. Thus we can answer the original question: there is strong evidence that there is a difference in the four treatments relative to their average counts. Ok, so the conclusion relates to the median (not the average), but it still essentially answer the same question.

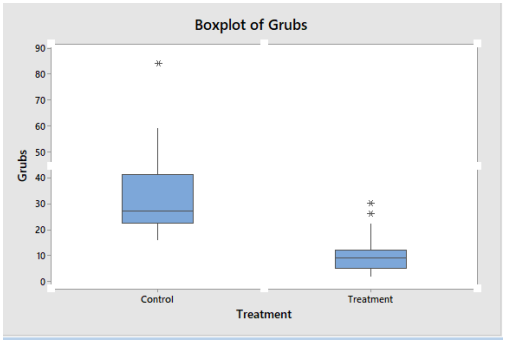
## 1.3 Question 2 - 2 Sample T-test on Treatment

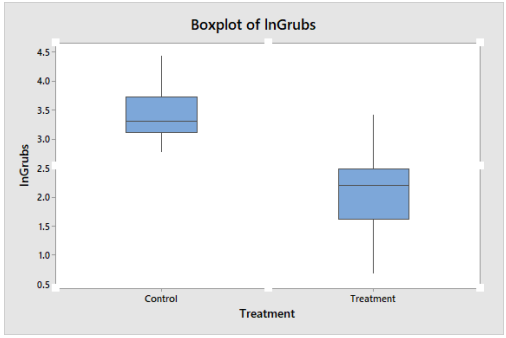
Is the average numbers of grubs on the test sites receiving the four treatments less than the average numbers on the control sites?

### 1.3.1 Hypotheses

Treatment consists of the 4 treatments: A,B,C,D.
$H_0 : \mu_{log(Control)} = \mu_{log(Treatment)}$
$H_a : \mu_{log(Control)} > \mu_{log(Treatment)}$

### 1.3.2 Boxplots and 2 Sample T-test Statistics




**Test**

| | |
|---|---|
| Null hypothesis | $H_0: \mu_1 - \mu_2 = 0$ |
| Alternative hypothesis | $H_1: \mu_1 - \mu_2 > 0$ |

| T-Value | DF | P-Value |
|---|---|---|
| 10.43 | 67 | 0.000 |

### 1.3.3 Choice of Statistical Technique

In the untransformed box plot (Boxplot of Grubs), the Control group is quite skewed but Treatment is approximately normal (minus the outliers). The outlier in Control group is over 4 standard deviations from the mean, and Treatment has multiple outliers. Visually, it appears that the group variances are different. Numerically this is also true with standard deviations of 15.42 (Control) and 6.251. So, I decided to consider the log transform. The transformed boxplot (Boxplot of lnGrubs) does appear to improve normality within the groups and it accounts for all of the outliers. It does appear to improve the equality of the group variances, confirmed numerically with standard deviations of 0.4025 (Control) and 0.6784. Thus, I performed the t-test on the transformed data. From the problem description, I see no reason to think that independence within or between groups has been violated.
I ran a test for equal variances on the transformed data and the p-values of the multiple comparisons test and Levene's test were 0.011 and 0.018, respectively. At a 95% confidence level, I reject the equal variances test null hypothesis that all variances are equal. So, I ran the T-test not assuming equal variances.

### 1.3.4 Conclusions

In the figure, $\mu_1 = \mu_{log(Control)}$ and $\mu_2 = \mu_{log(Treatment)}$. With a 95% confidence level and a p-value of less than 0.001, I reject $H_0$ and accept that $H_a$. Thus, the median of the Control group is greater than the median of the Treatment group. The treatments are assigned randomly among the test sites in each plot, so it is acceptable to make the claim that the different treatments are the cause of rejecting $H_0$.

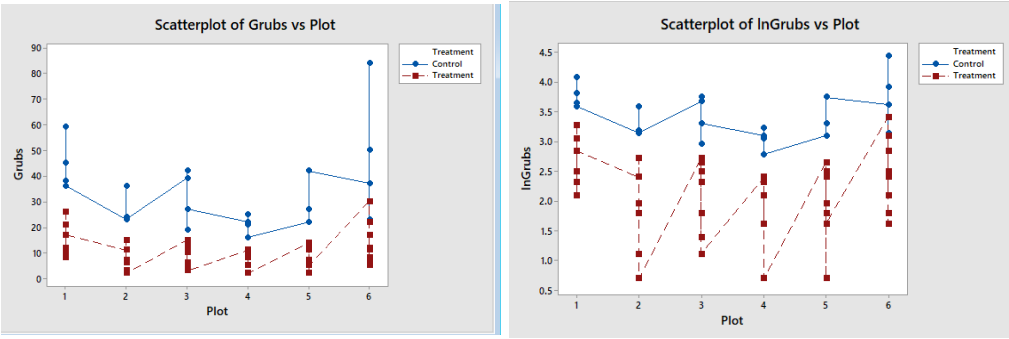
## 1.4 Question 2 - Multiple Regression

Is the average numbers of grubs on the test sites receiving the four treatments less than the average numbers on the control sites?

### 1.4.1 Hypotheses

Treatment consists of the 4 treatments: A,B,C,D.
$H_0 : \mu_{log(Control)} = \mu_{log(Treatment)}$
$H_a : \mu_{log(Control)} > \mu_{log(Treatment)}$

### 1.4.2 Visual Analysis

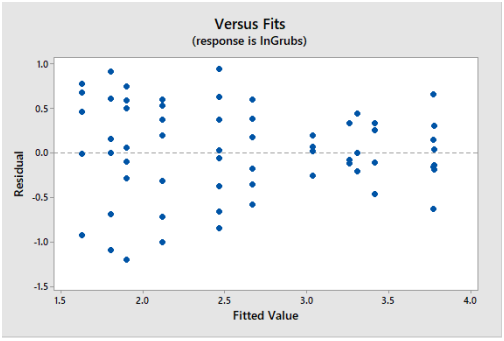



The coded scatterplot indicates that the test sites where no treatment (Control) was applied were almost always associated with larger grub counts. There appears to be nonconstant variance among the groups and a few outliers, so I considered a log transformation. The transformation appears to have largely corrected both problems, making the observations more amenable to regression tools. I decided again (as in question 1) on applying a multiple regression with two categorical variables (equivalently, a two-way ANOVA). Now, the Treatment factor only has 2 levels, instead of 4 previously.

### 1.4.3 Multiple Regression and Evaluation of the Log Transformed Saturated Model

**Analysis of Variance**

| Source | DF | Seq SS | Contribution | Adj SS | Adj MS | F-Value | P-Value |
|---|---|---|---|---|---|---|---|
| Regression | 11 | 36.6645 | 68.17% | 36.6645 | 3.33314 | 11.68 | 0.000 |
| Plot | 5 | 8.0309 | 14.93% | 1.7657 | 0.35313 | 1.24 | 0.303 |
| Treatment | 1 | 28.4244 | 52.85% | 3.2928 | 3.29280 | 11.54 | 0.001 |
| Plot*Treatment | 5 | 0.2092 | 0.39% | 0.2092 | 0.04185 | 0.15 | 0.980 |
| Error | 60 | 17.1166 | 31.83% | 17.1166 | 0.28528 | | |
| Total | 71 | 53.7811 | 100.00% | | | | |



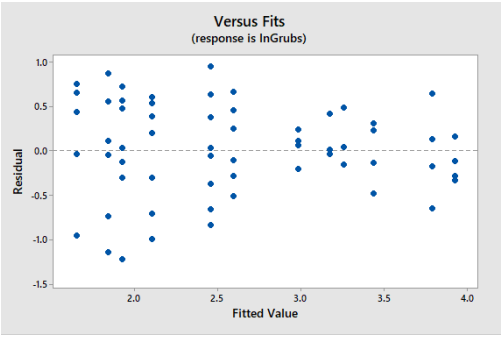

The p-value of 0.980 for the interaction term does not provide significant evidence that its coefficient is not 0. Interestingly, the p-value of Plot was 0.303. However, the extra sum of squares F-test that calculates that p-value is logically questionable, so I wouldn't place much importance on this result. Tests of the case influence statistics for each observation revealed three possible outliers. Their studentized residuals were no more(less) than 2.5(-2.5) and their Cook's Distances were all less than 0.1. So, there don't appear to be any particularly influential observations.
In the residual plot, there is a pattern: the residuals tend to be greater at the smallest and largest fitted values and decrease as the fitted values move further towards the middle. This suggests that something

needs to be changed in the model: possibly a transform or a change in the explanatory variables. Let's see how the additive model looks.

### 1.4.4 Multiple Regression and Evaluation of the Log Transformed Additive Model

Analysis of Variance

| Source | DF | Seq SS | Contribution | Adj SS | Adj MS | F-Value | P-Value |
|---|---|---|---|---|---|---|---|
| Regression | 6 | 36.4553 | 67.78% | 36.4553 | 6.0759 | 22.79 | 0.000 |
| Plot | 5 | 8.0309 | 14.93% | 8.0309 | 1.6062 | 6.03 | 0.000 |
| Treatment | 1 | 28.4244 | 52.85% | 28.4244 | 28.4244 | 106.64 | 0.000 |
| Error | 65 | 17.3258 | 32.22% | 17.3258 | 0.2666 | | |
| Lack-of-Fit | 5 | 0.2092 | 0.39% | 0.2092 | 0.0418 | 0.15 | 0.980 |
| Pure Error | 60 | 17.1166 | 31.83% | 17.1166 | 0.2853 | | |
| Total | 71 | 53.7811 | 100.00% | | | | |




The ANOVA table indicates that Plot and Treatment are significant factors in estimating the mean of the natural log of grubs. However, the residual plot has barely changed, due to the fact that the interaction term in the saturated model had such a small effect.

### 1.4.5 Conclusions

Coefficients

| Term | Coef | SE Coef | 95% CI | T-Value | P-Value | VIF |
|---|---|---|---|---|---|---|
| Constant | 3.924 | 0.172 | (3.580, 4.268) | 22.80 | 0.000 | |
| Plot | | | | | | |
| 2 | -0.752 | 0.211 | (-1.173, -0.331) | -3.57 | 0.001 | 1.67 |
| 3 | -0.491 | 0.211 | (-0.912, -0.070) | -2.33 | 0.023 | 1.67 |
| 4 | -0.942 | 0.211 | (-1.363, -0.521) | -4.47 | 0.000 | 1.67 |
| 5 | -0.670 | 0.211 | (-1.091, -0.249) | -3.18 | 0.002 | 1.67 |
| 6 | -0.137 | 0.211 | (-0.558, 0.284) | -0.65 | 0.517 | 1.67 |
| Treatment | | | | | | |
| Treatment | -1.333 | 0.129 | (-1.591, -1.075) | -10.33 | 0.000 | 1.00 |

Since the regression is on log transformed response data, it is necessary to make inferences on the ratio of medians instead of on the change in means. Consider the coefficient of Treatment (Control is the reference). Its coefficient of -1.333 indicates it decreases the median number of grubs in a test site by about 133%, when compared to Control. Since the p-value associated with the coefficient is less than 0.001, there is strong evidence that population median ratio isn't 0. Thus we can essentially answer the original question: there is strong evidence that, on average, there are fewer grubs at the test sites with the four treatments than at control sites.