"STA106 MIDTERM PROJECT II - Spring 2022"

Aditya Mittal & Mitchell Lawrence

Maxime Pouokam (Instructor)

Contents

PROBLEM I

Part A: Introduction

The data set Helicopter.csv contains two columns: "Shift" and "Count." Count is the response variable that records the number of times a helicopter was called to a sheriff's office during a one year span. Shift was the dependent categorical variable with four levels: I (between 2AM and 8AM), II (between 8AM and 2PM), III (between 2PM to 8PM), and IV (between 8PM to 2AM). The dataset contains 80 subjects.

With this information, we can ask the the question, "Does Shift have a statistically significant effect on the number of times a helicopter is called in to a sheriff's office?"

Part B: Model Fit and Assumptions

Since we are interested in testing for a single factor (Shift) effects on the number of times a helicopter is called, we can use the single factor ANOVA (SFA) model to fit our data. More specifically, the Single Factor ANOVA "cell means model" is a correct model fit for this data since we are interested in finding and comparing the group means of each group.

The cell means model is: $Yij = \mu_i + \epsilon_{ij}$ where:

Yij = jth value of Y observation in ith group

μ_i = unknown true population mean for ith group

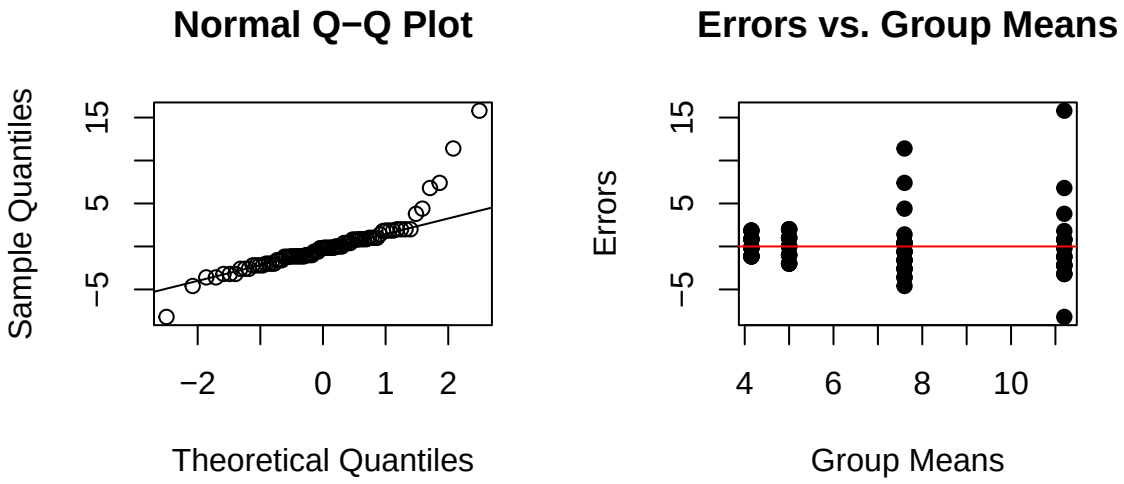
ϵ_{ij} = jth residual for ith group

To ensure that the ANOVA model is the best fit for our data, we must check if our data follows the needed assumptions.

A general ANOVA model must follow three basic assumptions: 1. All Yij's (each individual sample) were independently, randomly sampled 2. All groups are independent 3. Each error term, eij follows a normal distribution with mean = 0, equal population variance = $sigma^2$

The first two conditions are met as each subjects were independently, randomly sampled and each Shift group (I, II, III, IV) is also independent as the number of helicopters called in one shift does not does not provide information about the subjects in other groups. For our third assumption, we need to check if the random error terms follow normality and have equal population variances across groups.

Using plots, we create the Normal QQ plot to assess normality and the Errors vs. Group Means plot to test for homoscedasticity.



From the normal QQ plot, we can see that upper tail of the data deviates significantly from line, suggesting the distribution of the data is NOT normal. From the Errors vs Group Means plot, the "funnel" spread of the errors suggests heteroscedasticity due to unequal vertical spread based on grouping. However, this chart may be not be fully accurate due to difference in sample sizes among groups. The plots suggest both conditions are violated, but we conduct statistical tests to assess normality and homoscedasticity since using plots is subjective. We use common alpha level = 0.05 for our hypothesis testing.

1. Test for normality: Shapiro Wilkis test.

$H0$: The data is normally distributed vs. HA: The data is not normally distributed. Since our Shapiro-Wilkis test pvalue = 3.94535721771431e-09 < common alpha levels (0.01,0.05,0.1), we reject H0 in favor of Ha. We conclude that the the distribution of eij are non-normal. Thus, we can statistically conclude the condition of normality is violated.

2. Test for homoscedasticity: Brown-Forsythe Test.

$H_O : \sigma_I^2 = \sigma_{II}^2 = \sigma_{III}^2 = \sigma_{IV}^2$ vs. H_A: At least one group variance is unequal. Since our BF Test p-value = 0.0318596 < alpha = 0.05, we reject H0 in favor of Ha and conclude sample variance among groups are unequal.

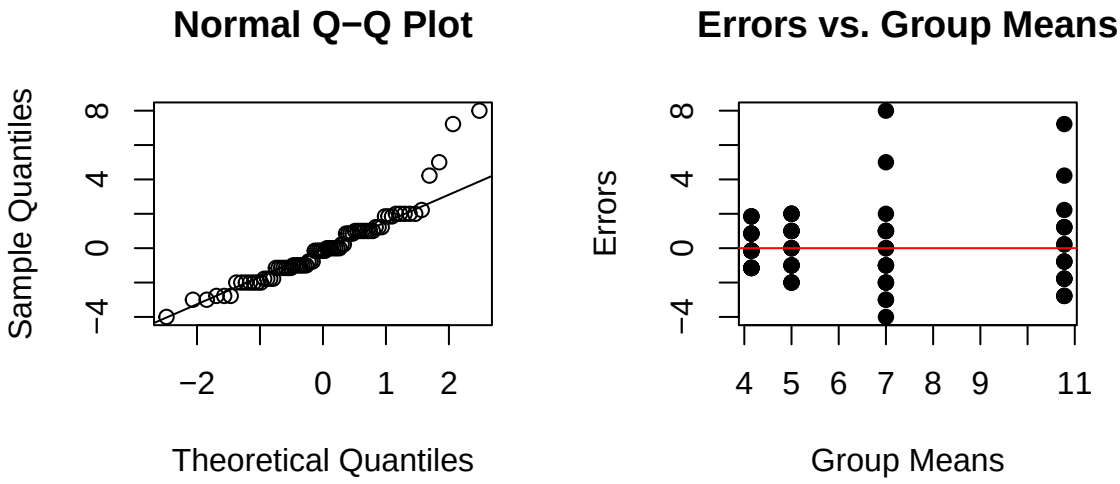
Thus, both parts of this condition are violated and we can move on to removing outliers/transformation of variables to try and alleviate these violations.

Part C: Outliers and Transformations

We first begin by removing outliers using the studentized/standardized residuals method as we have discovered heteroscedasticity of variance among the groups. The studentized residuals relaxes the assumption that all population variances are exactly equal. Furthermore, we are using the 99th percentile t-cutoff to determine if points are outliers.

$CO.rij = 1, 15, 67$: the following rows of the data set were identified as outliers, thus we can remove these rows and create the new dataset. Note, we removed approximately 3.75% of our data (3/80 observations). After removing our outliers, our new data set contains 77 entries.

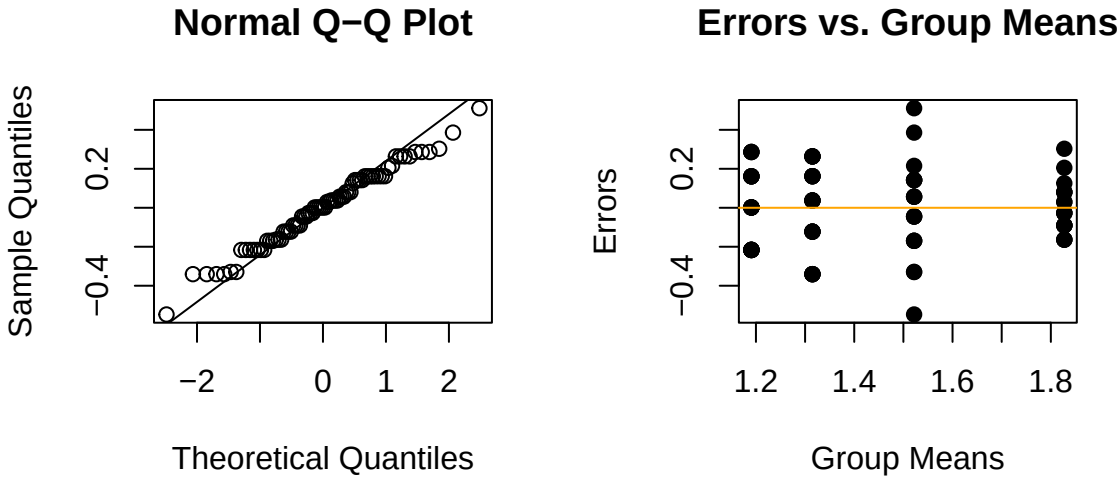
Here are the associated plots and diagnostic tests for the dataset without outliers:



Both plots still suggest possible non-normality and heteroscedasticity as the points deviate from the QQ line towards the upper tail (Normal QQ plot) and there is unequal vertical spread of residuals among groups (Errors vs. Group Means).

The P-value for Shapiro-Wilkis test: 2.0234568629526e-05. The P-value for the BF test: 0.0461419. Both p-value are higher than our original data, but they still smaller than alpha so we reject H0 and the assumptions are still not fully met. Since the data still does not meet the conditions, we can begin transformations of variables. More specifically, we are going to conduct three types of Box-Cox transformations to find different lambda values and select best one. Our lambda values based on: correlation coefficient of "best line" of QQ plot and transformed data, maximum p-value of the Shapiro-Wilkis test, the Log-Likelihood method.

Using the QQ PLOT method, our new transformed data has the following plots and tests:

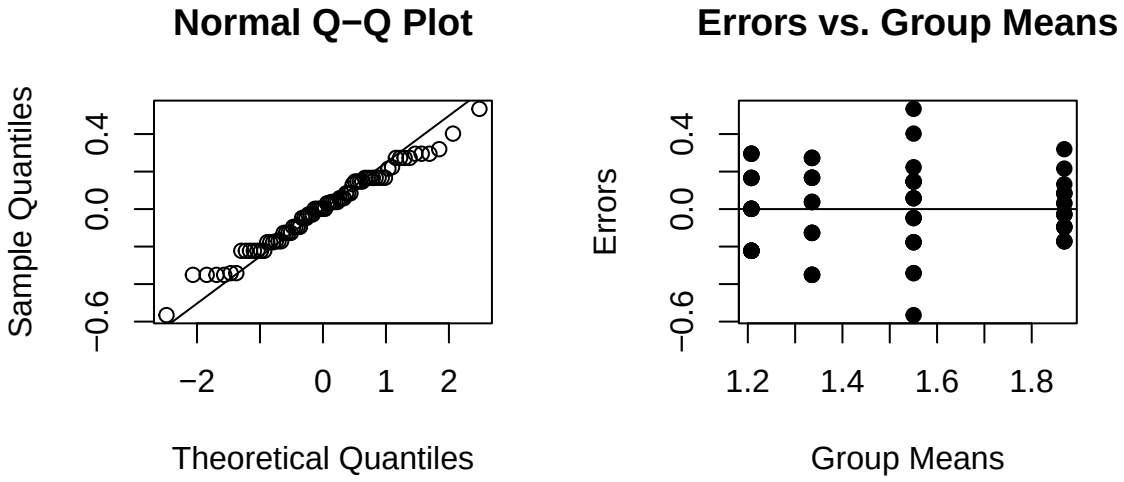


Using the QQ plot method, our calculated lambda was -0.2231027. Based on the normal QQ plot, the points are somewhat closer to the line compared to the original data set so the data may now approximately normal. However, it is important to note many points around both tails still deviate from the QQ line. The

Error vs. Group Means also suggests there is now equal variance as the vertical spread across groups is now approximately equal (with some deviations).

The P-value for Shapiro-Wilkis test: 0.620618843044404. The P-value for the BF test: 0.1433945. According to both tests, the data is now normal and follows homoscedasticity as our respective p-values are greater than alpha.

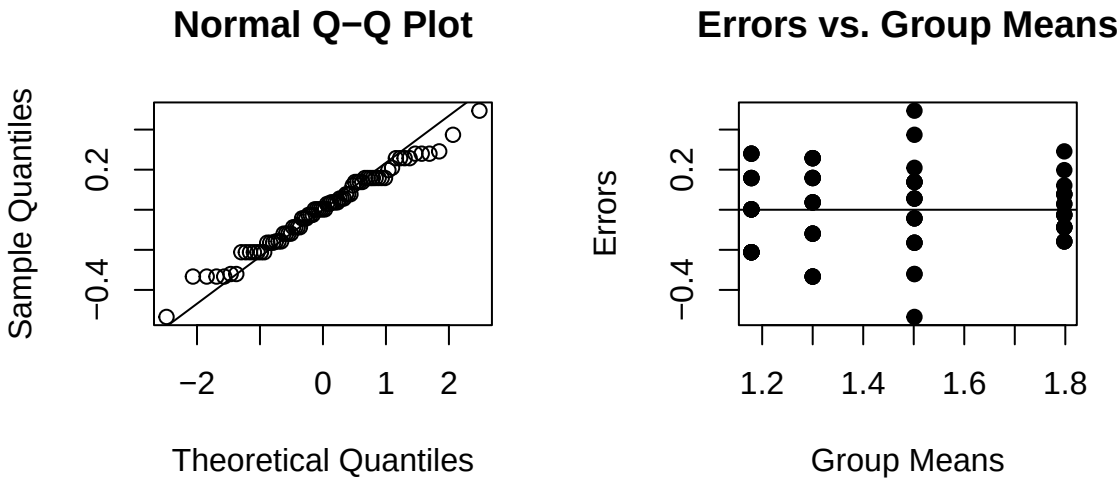
Using the Shapiro Wilkis p-value lambda, our new transformed data has plots such as:



Using the Shapiro-Wilkis pvalue method, our calculated lambda was -0.2024117. Based on the normal QQ plot, the points are somewhat closer to the line compared to the original data set so the data may now be normal. However, it is important to note many points around both tails still deviate from the QQ line. The Error vs. Group Means also suggests there is now equal variance as the vertical spread across groups is now approximately equal (with some deviations).

The P-value for Shapiro-Wilkis test: 0.623863913450328. The P-value for the BF test: 0.1514025. According to both tests, the data is now normal and follows homoscedasticity as our respective p-values are greater than alpha.

Using the Log Likelihood lambda, our new transformed data has plots such as:



Using the Log Likelihood pvalue method, our calculated lambda was -0.2381939. Based on the normal QQ plot, the points are somewhat closer to the line compared to the original data set so the data may now be normal. However, it is important to note many points around both tails still deviate from the QQ line. The Error vs. Group Means also suggests there is now equal variance as the vertical spread across groups is now approximately equal (with some deviations).

The P-value for Shapiro-Wilkis test: 0.614222969246058. The P-value for the BF test: 0.1376601. According to both tests, the data is now normal and follows homoscedasticity as our respective p-values are greater than alpha.

Since the plots of all three transformation suggests the conditions have been met, we use the statistical tests to find the highest p-values and make a conclusion on the best transformation. Here is a summary table of all test results of our datasets:

	Data with Outliers	Data without Outliers	Data Transformed using lambda = -0.2231027, (QQ plot)	Data Transformed using lambda = -0.2024117, (Shapiro-Wilkis)	Data Transformed using lambda = -0.2381939 (Log Likelihood)
Shapiro-Wilkis Test	3.945357e-09	2.023457e-05	0.6206188	0.6238639	0.614223
Brown-Forsythe Test	0.03185955	0.04614186	0.1433945	0.1514025	0.1376601

For our "best" combination, we should first use the studentized residuals method to find outliers, using the 99 percentile t-cutoff. Next, in terms of transformations, it appears all three transformations alleviate non-normality and heteroscedasticity. However, to pick one, we should go with the Box-Cox Transformation

Shapiro Wilkis P-value method on the dataset without outliers as both tests calculated in the highest p-value, indicating the highest probability of observing our data or more extreme given null hypothesis is true.

PART D: Conclusion

The transformation of variables helped alleviate both violations of non-normality and heteroscedasticity. Thus, conducting transformation of variables was ultimately beneficial and can be a statistically better fit to run the SFA Anova model. If the client wanted to use this dataset of ANOVA, I would recommend they may use it provided they account for outliers using studentized residual and use the Shapiro Wilkis Box-Cox transformation to ensure their data fits all assumptions.

With this, there are several downsides to transformations: accurate interpretation can be very difficult and reversing them can also become a very complicated process. Thus, it may be the case that the calculated CI and hypothesis testing does not provide the client the accurate information they might be looking for.

PROBLEM II

PART A: Introduction/Background

As the technology industry becomes an increasingly important factor driving the growth of many businesses, competition for hiring new talent has become extremely hard. As such, these companies are offering very high salaries to attract new employees over their competitors. With this, certain professions and regions have seen a growth in average salaries offered to employees. New employees considering potential job prospects can use this information to narrow possible professions and places they'd want to live where they can maximize their salaries.

More specifically, the data was recorded through an observation with two categorical factors profession and region, and a response variable annual salary. We are interested in finding if either or both factors have significant effects on the average salaries and if there's a possible interaction effect between both factors.

We will use two methods to test for differences in salary offers based on profession AND region. Since we have two possible factors, we will be using the Two Factor Anova Model.

1. We will use hypothesis testing, using F-tests, to find the most reduced model that still suggests statistically significant effects. More specifically, we will first test for interactions between the TFA interactions model and the TFA no interaction model. If we conclude there are no significant interactions, we can move to testing for individual factor effects. If both factors are significant, we use TFA no interactions as our final model or we can use Single Factor ANOVA model for the significant factor as our final model.

2. Next, we are interested in creating 6 separate confidence intervals. Our first three confidence intervals will pairwise, comparing level means of the factor "profession." More specifically, we will create pairwise CI comparing Data Scientists to Software Engineers, Data Scientists to Bioinformatics Engineer, Software Engineers to Bioinformatics Engineer. Our fourth CI will be a pairwise CI comparing factor B level means "region:" San Francisco and Seattle. Our final two CI will be contrasts, comparing the salaries of Data Scientists and Bioinformatics Engineer in Seattle to salaries of Data Scientists and Bioinformatics Engineer in SF. The second contrast compares salaries of Software Engineers and Bioinformatics Engineer in Seattle to Software Engineers and Bioinformatics Engineer in SF. The contrasts will help us test if there's interactions based on CI.

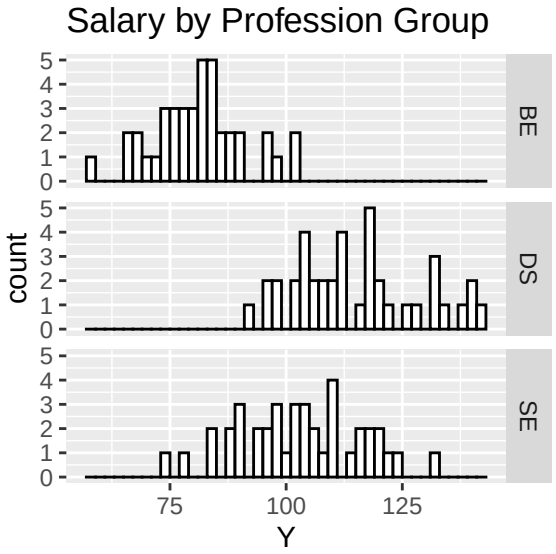
Ultimately, both of these methods will help us determine whether or not there is statistically significant in the average salaries based on both/ either factors, and if there are possible interactions between the factors. New employees can use this information to narrow their future job prospects.

PART B: Summary

The salary data set contains 120 subjects, with two factor levels: "Profession" (3 levels: Data Scientist, Software Engineer, Bioinformatics Engineer) and "Region" (2 levels: San Francisco and Seattle). To get a more accurate visualization of the data, we create a summary table to showcase the group means, standard deviations, and sample sizes.

First, we show summary statistics for Factor A: Profession.

```
##                BE        DS        SE
## Means    81.086996 115.14799 102.90644
## Std. Dev  9.662515  13.66819  13.24031
## Sizes    40.000000  40.00000  40.00000
```



From the summary table, it appears that the average salaries for Data Scientists is higher than Software Engineers and Bioinformatics Engineers. The standard deviations of Data Scientists and Software Engineers are approximately equal (around 13), while Bioinformatics Engineer has a lower standard deviation (approx 9). Based on the spread of our boxplot and histogram, the spread of the annual salary appears to vary based each factor level. This suggests there are statistically significant effects based on factor A. It is important to note each factor level has equal sample size, so our confidence intervals will follow equal weighting (this would affect the boundary of our CI).

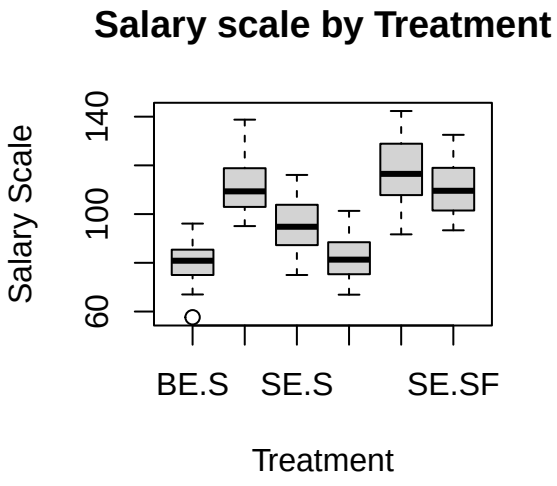
Here are the group means and standard deviations for factor B:

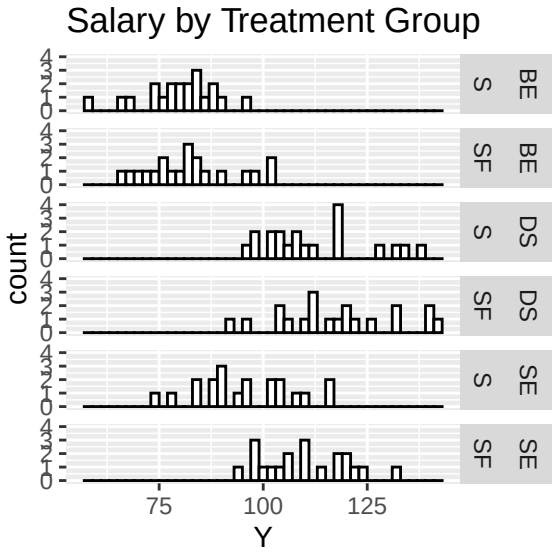
```
##                S        SF
## Means    95.94358 103.48403
## Std. Dev 17.41791  19.29842
## Sizes    60.00000  60.00000
```

The average salaries in San Francisco are higher than Seattle. However, both levels have high standard deviations which suggests there may actually not be any statistically significant effects based on region. The spread of data on both plots is also very similar, which suggests no effect based on region. In part 4, we will use F-tests to get more accurate results of factor B effects. Again, sample sizes based on factor levels are equal so the CI will have equal weighting.

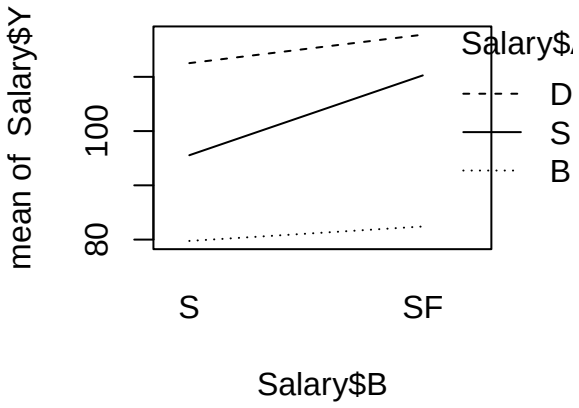
Here are the group means for treatment groups:

```
##            S        SF
## BE  79.75485  82.41914
## DS 112.52715 117.76883
## SE  95.54875 110.26412
```





Based on the summary table and the spread of the histogram/boxplot of treatment means by group, there appears to be some possible interactions based on group means. A more accurate visualization can be seen with interaction plot.



From the interaction plot, there seems to be some interactions between factors A and B as shows by the non-parallel lines. More specifically, treatment group of SF region and Software Engineer profession caused an additional increase in annual salary over either single treatment combined.

PART C: Assumptions and Diagnostics

In part D, we show the TFA no-interactions model as our final model. Thus, we will use the TFA no-interactions model to run our model diagnostics.

To ensure that the ANOVA model is the best fit for our data, we must check if our data follows the needed assumptions.

The TFA no interaction model is: $Yijk = \mu_{..} + \gamma_i + \delta_j + \epsilon_{ijk}$ where

$Yijk$ = kth value of Y observation in ith group

$\mu_{..}$ = overall mean

γ_i = effect of Factor A, i-th group

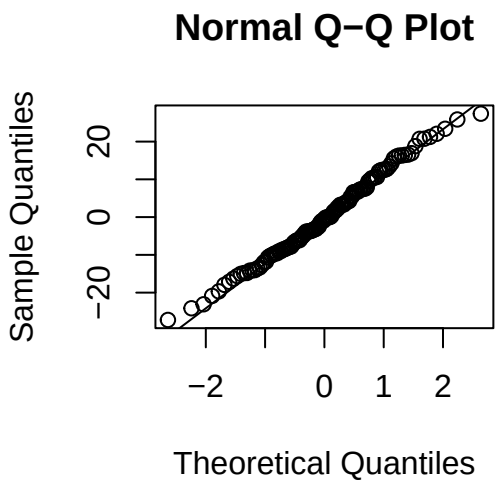
δ_j = effect of Factor B, j-th group

ϵ_{ijk} = individual error

This ANOVA model must follow these basic assumptions: 1. All Yij's (each individual sample) were independently, randomly sampled 2. All levels of factor A are independent 3. All levels of factor B are independent 4. Each error term, Eij follows a normal distributed with mean = 0, equal population variance = $sigma^2$

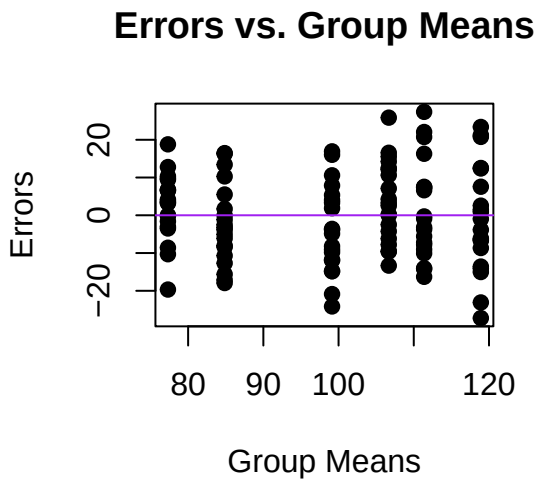
The condition of independence is the most important assumption for the ANOVA model's validity. The first and second condition are met as all salaries were independently, randomly sampled. Each profession group (Data Scientist, Software Engineer, Bioinformatic Engineer) is also independent as the salaries of employees in one profession does not provide information about the subjects in other groups. Same reasoning follows for Factor B region, SF and Seattle. For our third assumption, we need to check if the random error terms follow normality and have equal population variances across groups. We use alpha = 0.05 for our tests to control Type I error.

We can assess normality using the normal QQ plot and the Shapiro Wilkis test.



Based on the normal QQ plot, we can conclude that the data is approximately normal based on the points lying close to the line. Since our pvalue = 0.669780080844438 > alpha, we can conclude the data is normal from the Shapiro Wilkis test.

Now, we also need to test for homoscedasticity using Error vs Group Means Plot and the BF test:



Based on the plot, the errors seem equally spread across groups so we may conclude the data has approximately equal variances across groups. Based on the BF test, our pval: 0.3048319 > alpha, so we conclude data variances follow homoscedasticity.

Thus, our data meets all assumptions of the TFA no-interactions model and we can now move on to analysis.

PART D: Analysis

We want to try and find a more reduced model if possible, since it is easier to interpret than a model with more parameters. Thus, we will first test for interactions between the TFA interactions model and the TFA no interactions model:

```
## Analysis of Variance Table
##
## Model 1: Y ~ A + B
## Model 2: Y ~ A * B
##   Res.Df   RSS Df Sum of Sq      F  Pr(>F)
## 1    116 16058
## 2    114 15253  2    805.41 3.0098 0.05324 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Our null hypothesis is: $(\gamma\delta)_{ij} = 0$ for all ij, TFA no interactions model is a better fit. Our alternate hypothesis is: at least one $(\gamma\delta)_{ij}$!= 0, the TFA interactions model is a better fit.

The calculated F-statistic is: 3.0097976 at df(num) = 2, df(denom) = 114. Since our corresponding p-value 0.0532358 > alpha = 0.05, we fail to reject the H0 in favor of HA. We can conclude the TFA no interactions model is statistically a better fit than TFA interactions model between profession and region. Furthermore, we can calculate the corresponding condition R^2 to measure the % of reduction in error when adding interactions to a model with no interactions. The associated conditional $R^2 = 0.0501551$. When we add an interaction to a model with A and B effects, the reduction of error is 5.0155104. The small value of reduction error also suggests that interactions are not very significant.

Now, we can move on to testing for individual effects. First we can test for Factor A effects and compute the corresponding conditional R^2.

```
## Analysis of Variance Table
##
## Model 1: Y ~ B
## Model 2: Y ~ A + B
##   Res.Df   RSS Df Sum of Sq      F    Pr(>F)
## 1    118 39873
## 2    116 16058  2     23815 86.014 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Our null hypothesis is: $\gamma_i = 0$ for all i, Factor A effects are NOT significant. Our alternate hypothesis is: γ_i != 0 for at least one i,Factor A effects are significant.

The calculated F-statistic is: 86.0142968 at df(num) = 2, df(denom) = 118. Since our p-value $1.2339522 \times 10^{-23}$ < alpha = 0.05, we reject the H0 in favor of HA. We can conclude the there are significant Factor A effects. The conditional $R^2 = 0.5972622$. When we add an factor A to a model with B effects, the reduction of error is 59.7262208. The high value of reduction also suggests that factor A effects are significant.

Now, we can test for Factor B effects and compute the corresponding conditional R^2.

```
## Analysis of Variance Table
##
## Model 1: Y ~ A
## Model 2: Y ~ A + B
##   Res.Df   RSS Df Sum of Sq      F    Pr(>F)
## 1    117 17764
## 2    116 16058  1    1705.8 12.322 0.0006385 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Our null hypothesis is: $\delta_j = 0$ for all j, Factor B effects are NOT significant. Our alternate hypothesis is: δ_j != 0 for at least one j,Factor B effects are significant.

The calculated F-statistic is: 12.3217684 at df(num) = 1, df(denom) = 117. Since our p-value 6.3846549×10^{-4} < alpha = 0.05, we reject the H0 in favor of HA. We can conclude the there are significant Factor B effects. The conditional $R^2 = 0.0960224$. When we add an factor B to a model with A effects, the reduction of error is 9.6022433. The high value of reduction also suggests that factor B effects are significant. Thus, we use the TFA no interactions model as our final model as both individual factors are significant and interactions are not.

Now that we've establish each individual factor is significant, we create pairwise and contrast CI comparing factor level/treatment means to see how much salary difference there is between each factor level.

Here are following CI as mentioned in the introduction:

We are overall 95% confident that the difference in salaries of Bioinformatics Engineer is lower than Data Scientists by 26.6303 and 41.4917 dollars (in thousands). Ignoring interaction effects.

We are overall 95% confident that the difference in salaries of Bioinformatics Engineer is lower than Software Engineer by 14.3887 and 29.2502 dollars (in thousands). Ignoring interaction effects.

We are overall 95% confident that the difference in salaries of Data Scientists is higher than Software Engineer by 4.8108 and 19.6723 dollars (in thousands). Ignoring interaction effects.

We are overall 95% confident that the difference in salaries of employees in Seattle is lower than employees in San Francisco by 3.3931 and 11.6878 dollars (in thousands). Ignoring interaction effects.

We are overall 95% confident there is no statistical difference between the average salaries of Bioinformatics Engineer and Data Scientists in Seattle and average salaries of Bioinformatics Engineer and Data Scientists in San Francisco. 0 is contained within this interval.

We are overall 95% confident that the average salaries of Bioinformatics Engineer and Software Engineers in Seattle is less than the average salaries of Bioinformatics Engineer and Software Engineers in San Francisco by 0.0078355 and 17.3718299 dollars (in thousands).

PART D: Interpretation

Based on the hypothesis testing conduced in part C, we concluded that the two way no interactions model should statistically be our most reduced, final model. This indicates that there both individual factor A and B effects are significant. Thus, we created pairwise CI to compare factor level means to see how much salaries differ based on profession and how much salaries differ based on region while ignoring possible interaction effects.

Our first two confidence intervals indicated that Bioinformatic Engineers earn lesser salaries compared to both Software Engineer or Data Scientist, if we ignoring interaction effects from location. We then calculated the pairwise CI comparing Software Engineer and Data Scientists, concluding Data Scientists earn higher salaries. With this, we can statistically conclude Data Scientist earn the highest salaries of the three professions, followed by Software Engineers and then Bioinformatics Engineers. Note, this is ignoring interaction effects from location. Our fourth pairwise CI revealed that employees in San Francisco earn more than Seattle, ignoring interaction effects from profession. Thus, our pairwise helped us understand which individual factors had the highest salary offers, and we can conclude that data scientist in SF are likely to earn the highest salaries.

We also created contrasts, comparing the average salaries of Data Scientists and Bioinformatics Engineer in Seattle to Data Scientists and Bioinformatics Engineer in SF. The second contrast compares average salaries of Software Engineers and Bioinformatics Engineer in Seattle to Software Engineers and Bioinformatics Engineer in SF. With this, we concluded there is no statistical difference between the data Scientists and Bioinformatics Engineer in Seattle to Data Scientists and Bioinformatics Engineer in SF. However, there is difference in averages of Software Engineers and Bioinformatics Engineer in Seattle to Software Engineers and Bioinformatics Engineer in SF. This suggests there are possible minor interactions between profession and region as this can be corroborated due to the possible reduction in error when adding an interaction effect

to a model with factor A and B, effects although we decided the interaction was statistically insignificant using hypothesis testing at alpha = 0.05.

PART E: Conclusion

The goal of our tests and evaluation of the data was to determine the factor effects in our model for salaries across different job titles and regions. We have concluded that it the two factor anova model with no interaction effect would be the best model to fit our data. This was because when we tested for interaction effect we found it not to be significant, but we did find both the factor of job title and region both had a significant effect on salary. This led us to take into account both factors when modeling our data. This is a good result because we know the data we collected is relevant and we chose the right factors to observe in our data.

Thanks to our confidence intervals we constructed throughout the report, we were able to answer many questions about our data. We found that in fact it is more advantageous for employees to apply for jobs in San Francisco because, ignoring interaction effects, we found that San Francisco employees earn more. This increase of salary in San Francisco makes sense because this could also possibly be explained by a higher cost of living in the crowded expensive living areas of San Francisco. Similarly, using simultaneous confidence intervals we were able to determine that the highest paying salary was for the data scientists workers, while software engineers and bioinformatics engineers come thereafter. This information we extracted from our data also makes sense given that San Francisco is near the Silicon valley, at the heart of production from data scientists. Overall, this data set gave us a lot of information on job salaries across different job titles and in different locations.

Appendix of Code Used

```
knitr::opts_chunk$set(warning = FALSE, message = FALSE)
# PROBLEM I
# read dataset
Helicopter <- read.csv("~/Desktop/Year1/Spring Quarter 2022/STA106/Project II/Helicopter.csv")
# diagnostics
model.helicopter = lm(Count~Shift, data = Helicopter)
qqnorm(model.helicopter$residuals)
qqline(model.helicopter$residuals)
plot(model.helicopter$fitted.values, model.helicopter$residuals, main = "Errors vs. Group Means",xlab =
abline(h = 0,col = "red")

helicopter.original.ei = model.helicopter$residuals
helicopter.original.the.SWtest = shapiro.test(helicopter.original.ei)
library(car)
the.BFtest = leveneTest(helicopter.original.ei~ Shift, data=Helicopter, center=median)
helicopter.original.p.val = the.BFtest[[3]][1]
# Outliers
Helicopter$ei = model.helicopter$residuals
Helicopter.nt = nrow(Helicopter)
Helicopter.a = length(unique(Helicopter$Shift))
t.cutoff= qt(1-0.01,Helicopter.nt-Helicopter.a)
Helicopter.rij = rstandard(model.helicopter)
Helicopter.CO.rij = which(abs(Helicopter.rij) > t.cutoff)
Helicopter.outliers = Helicopter.CO.rij
Helicopter.new.data = Helicopter[-Helicopter.outliers,]
Helicopter.new.model = lm(Count ~ Shift,data = Helicopter.new.data)

qqnorm(Helicopter.new.model$residuals)
qqline(Helicopter.new.model$residuals)
plot(Helicopter.new.model$fitted.values, Helicopter.new.model$residuals, main = "Errors vs. Group Means
abline(h = 0,col = "red")
helicopter.outlier.ei = Helicopter.new.model$residuals
helicopter.outlier.the.SWtest = shapiro.test(helicopter.outlier.ei)
library(car)
helicopter.outlier.the.BFtest = leveneTest(helicopter.outlier.ei~ Shift, data=Helicopter.new.data, cent
helicopter.outlier.p.val = helicopter.outlier.the.BFtest[[3]][1]

# BOX COX QQ Plot
library(EnvStats)
L1 =boxcox(Helicopter.new.model ,objective.name = "PPCC",optimize = TRUE)$lambda
QQ.YT = (Helicopter.new.data$Count^(L1)-1)/L1
QQ.t.data = data.frame(Count = QQ.YT, Shift = Helicopter.new.data$Shift)
QQ.t.model = lm(Count ~ Shift,data = QQ.t.data)
qqnorm(QQ.t.model$residuals)
qqline(QQ.t.model$residuals)
plot(QQ.t.model$fitted.values, QQ.t.model$residuals, main = "Errors vs. Group Means",xlab = "Group Means
abline(h = 0,col = "orange")
QQ.ei = QQ.t.model$residuals
QQ.the.SWtest = shapiro.test(QQ.ei)
QQ.the.BFtest = leveneTest(QQ.ei~ Shift, data=Helicopter.new.data, center=median)
QQ.p.val = QQ.the.BFtest[[3]][1]
# BOX COX SHAPIRO
```

```
L2 =boxcox(Helicopter.new.model, objective.name = "Shapiro-Wilk" ,optimize = TRUE)$lambda
Shapiro.YT = (Helicopter.new.data$Count^(L2)-1)/L2
Shapiro.t.data = data.frame(Count = Shapiro.YT, Shift = Helicopter.new.data$Shift)
Shapiro.t.model = lm(Count ~ Shift,data = Shapiro.t.data)
qqnorm(Shapiro.t.model$residuals)
qqline(Shapiro.t.model$residuals)
plot(Shapiro.t.model$fitted.values, Shapiro.t.model$residuals, main = "Errors vs. Group Means",xlab = "
abline(h = 0,col = "black")
Shapiro.ei = Shapiro.t.model$residuals
Shapiro.the.SWtest = shapiro.test(Shapiro.ei)
Shapiro.the.BFtest = leveneTest(Shapiro.ei~ Shift, data=Helicopter.new.data, center=median)
Shapiro.p.val = Shapiro.the.BFtest[[3]][1]
# BOX COX LOG LIKELIHOOD
L3 = boxcox(Helicopter.new.data$Count,objective.name = "Log-Likelihood",optimize = TRUE)$lambda
Log.YT = (Helicopter.new.data$Count^(L3)-1)/L3
Log.t.data = data.frame(Count = Log.YT, Shift = Helicopter.new.data$Shift)
Log.t.model = lm(Count ~ Shift,data = Log.t.data)
qqnorm(Log.t.model$residuals)
qqline(Log.t.model$residuals)
plot(Log.t.model$fitted.values, Log.t.model$residuals, main = "Errors vs. Group Means",xlab = "Group Mea
abline(h = 0,col = "black")
Log.ei = Log.t.model$residuals
Log.the.SWtest = shapiro.test(Log.ei)
Log.the.BFtest = leveneTest(Log.ei~ Shift, data=Helicopter.new.data, center=median)
Log.p.val = Log.the.BFtest[[3]][1]
knitr::include_graphics("/Users/adityamittal/Desktop/Year1/Spring Quarter 2022/STA106/Project II/Screen
# setup basic function
find.means = function(the.data,fun.name = mean){
  a = length(unique(the.data[,2]))
  b = length(unique(the.data[,3]))
  means.A = by(the.data[,1], the.data[,2], fun.name)
  means.B = by(the.data[,1],the.data[,3],fun.name)
  means.AB = by(the.data[,1],list(the.data[,2],the.data[,3]),fun.name)
  MAB = matrix(means.AB,nrow = b, ncol = a, byrow = TRUE)
  colnames(MAB) = names(means.A)
  rownames(MAB) = names(means.B)
  MA = as.numeric(means.A)
  names(MA) = names(means.A)
  MB = as.numeric(means.B)
  names(MB) = names(means.B)
  MAB = t(MAB)
  results = list(A = MA, B = MB, AB = MAB)
  return(results)
}

Partial.R2 = function(small.model,big.model){
SSE1 = sum(small.model$residuals^2)
SSE2 = sum(big.model$residuals^2)
PR2 = (SSE1 - SSE2)/SSE1
return(PR2)
}

find.mult = function(alpha,a,b,dfSSE,g,group){
```

```
if(group == "A"){
Tuk = round(qtukey(1-alpha,a,dfSSE)/sqrt(2),3)
Bon = round(qt(1-alpha/(2*g), dfSSE ) ,3)
Sch = round(sqrt((a-1)*qf(1-alpha, a-1, dfSSE)),3)
}else if(group == "B"){
Tuk = round(qtukey(1-alpha,b,dfSSE)/sqrt(2),3)
Bon = round(qt(1-alpha/(2*g), dfSSE ) ,3)
Sch = round(sqrt((b-1)*qf(1-alpha, b-1, dfSSE)),3)
}else if(group == "AB"){
Tuk = round(qtukey(1-alpha,a*b,dfSSE)/sqrt(2),3)
Bon = round(qt(1-alpha/(2*g), dfSSE ) ,3)
Sch = round(sqrt((a*b-1)*qf(1-alpha, a*b-1, dfSSE)),3)
}
results = c(Bon, Tuk,Sch)
names(results) = c("Bonferroni","Tukey","Scheffe")
return(results)
}

scary.CI = function(the.data,MSE,equal.weights = TRUE,multiplier,group,cs){
   if(sum(cs) != 0 & sum(cs !=0 ) != 1){
    return("Error - you did not input a valid contrast")
  }else{
    the.means = find.means(the.data)
    the.ns =find.means(the.data,length)
    nt = nrow(the.data)
    a = length(unique(the.data[,2]))
    b = length(unique(the.data[,3]))
    if(group =="A"){
      if(equal.weights == TRUE){
        a.means = rowMeans(the.means$AB)
        est = sum(a.means*cs)
        mul = rowSums(1/the.ns$AB)
        SE = sqrt(MSE/b^2 * (sum(cs^2*mul)))
        N = names(a.means)[cs!=0]
        CS = paste("(",cs[cs!=0],")",sep = "")
        fancy = paste(paste(CS,N,sep =""),collapse = "+")
        names(est) = fancy
      } else{
        a.means = the.means$A
        est = sum(a.means*cs)
        SE = sqrt(MSE*sum(cs^2*(1/the.ns$A)))
        N = names(a.means)[cs!=0]
        CS = paste("(",cs[cs!=0],")",sep = "")
        fancy = paste(paste(CS,N,sep =""),collapse = "+")
        names(est) = fancy
      }
    }else if(group == "B"){
      if(equal.weights == TRUE){
        b.means = colMeans(the.means$AB)
        est = sum(b.means*cs)
        mul = colSums(1/the.ns$AB)
        SE = sqrt(MSE/a^2 * (sum(cs^2*mul)))
        N = names(b.means)[cs!=0]
```

```
        CS = paste("(",cs[cs!=0],")",sep = "")
        fancy = paste(paste(CS,N,sep =""),collapse = "+")
        names(est) = fancy
      } else{
        b.means = the.means$B
        est = sum(b.means*cs)
        SE = sqrt(MSE*sum(cs^2*(1/the.ns$B)))
        N = names(b.means)[cs!=0]
        CS = paste("(",cs[cs!=0],")",sep = "")
        fancy = paste(paste(CS,N,sep =""),collapse = "+")
        names(est) = fancy
      }
    } else if(group == "AB"){
      est = sum(cs*the.means$AB)
      SE = sqrt(MSE*sum(cs^2/the.ns$AB))
      names(est) = "someAB"
    }
    the.CI = est + c(-1,1)*multiplier*SE
    results = c(est,the.CI)
    names(results) = c(names(est),"lower bound","upper bound")
    return(results)
  }
}
# PROBLEM II
# read dataset Salary.csv
Salary <- read.csv("~/Desktop/Year1/Spring Quarter 2022/STA106/Project II/Salary.csv")
# Summary data
Salary.the.means =find.means(Salary,mean)
Salary.the.sizes=find.means(Salary,length)
Salary.the.sds=find.means(Salary,sd)
names(Salary) = c("Y","A","B")
the.summary.A = rbind(Salary.the.means$A, Salary.the.sds$A, Salary.the.sizes$A)
colnames(the.summary.A) = names(Salary.the.means$A)
rownames(the.summary.A) = c("Means","Std. Dev", "Sizes")
the.summary.A
boxplot(Salary$Y ~ Salary$A, main = "Salary scale by group",
        ylab = "Salary Scale")
library(ggplot2)
ggplot(Salary, aes(x = Y)) + geom_histogram(binwidth = 2,,color = "black",fill = "white") +
  facet_grid(A ~.) +ggtitle("Salary by Profession Group")
the.summary.B = rbind(Salary.the.means$B, Salary.the.sds$B, Salary.the.sizes$B)
colnames(the.summary.B) = names(Salary.the.means$B)
rownames(the.summary.B) = c("Means","Std. Dev", "Sizes")
the.summary.B
boxplot(Salary$Y ~ Salary$B, main = "Salary scale by group",
        ylab = "Salary Scale")
library(ggplot2)
ggplot(Salary, aes(x = Y)) + geom_histogram(binwidth = 2,,color = "black",fill = "white") +
  facet_grid(B ~.) +ggtitle("Salary by Treatment Group")

Salary.the.means$AB
boxplot(Y ~ A*B, main = "Salary scale by Treatment",
        ylab = "Salary Scale", xlab = "Treatment", data = Salary)
```

```
library(ggplot2)
ggplot(Salary, aes(x = Y)) + geom_histogram(binwidth = 2,,color = "black",fill = "white") +
  facet_grid(A*B ~.) +ggtitle("Salary by Treatment Group")

interaction.plot(Salary$B, Salary$A, Salary$Y)
# diagnostics
AB = lm(Y ~ A*B,Salary)
A.B = lm(Y ~ A + B,Salary)
A = lm(Y ~ A,Salary)
B = lm(Y ~ B,Salary)
N = lm(Y ~ 1, Salary)
all.models = list(AB,A.B,A,B,N)
SSE = t(as.matrix(sapply(all.models,function(M) sum(M$residuals^2))))
colnames(SSE) = c("AB","(A+B)","A","B","Empty/Null")
rownames(SSE) = "SSE"
qqnorm(A.B$residuals)
qqline(A.B$residuals)
ei = A.B$residuals
the.SWtest = shapiro.test(ei)
# equal variance
plot(A.B$fitted.values, A.B$residuals, main = "Errors vs. Group Means",xlab = "Group Means",ylab = "Erro
abline(h = 0,col = "purple")
library(car)
the.BFtest = leveneTest(ei~ A*B, data=Salary, center=median)
p.val = the.BFtest[[3]][1]
interaction.results = anova(A.B,AB)
interaction.results
R.squared.one = Partial.R2(A.B, AB)*100
factorA.results = anova(B,A.B)
factorA.results
R.squared.two = Partial.R2(B, A.B)*100
factorB.results = anova(A,A.B)
factorB.results
R.squared.three = Partial.R2(A, A.B)*100
a = length(unique(Salary$A))
b = length(unique(Salary$B))
nt = nrow(Salary)
SSE = round(sum(AB$residuals^2),2)
B = round(AB$coefficients,3)
MSE = round(SSE/(nt-a-b+1),4)

MULT = find.mult(0.05,a,b,nt-a-b+1,6,"AB")
Tuk = MULT[2]
CI1= round(scary.CI(Salary,MSE,equal.weights = TRUE,Tuk,"A",c(1,-1,0)),4)
CI2= round(scary.CI(Salary,MSE,equal.weights = TRUE,Tuk,"A",c(1,0,-1)),4)
CI3= round(scary.CI(Salary,MSE,equal.weights = TRUE,Tuk,"A",c(0,1,-1)),4)
Tuk.two= find.mult(0.05,a,b,nt-a-b+1,4,"B")[2]
CIb1= round(scary.CI(Salary,MSE,equal.weights = TRUE,Tuk.two,"B",c(1,-1)),4)
AB.cs.1 = matrix(0,nrow = a, ncol = b)
AB.cs.1[1,1] = 1/2
AB.cs.1[2,1] = 1/2
AB.cs.1[1,2] = -1/2
AB.cs.1[2,2] = -1/2
```

```
Sheffe= MULT[3]
CIAB1 = scary.CI(Salary,MSE,equal.weights = TRUE,Sheffe,"AB",AB.cs.1)

AB.cs.2 = matrix(0,nrow = a, ncol = b)
AB.cs.2[1,1] = 1/2
AB.cs.2[3,1] = 1/2
AB.cs.2[1,2] = -1/2
AB.cs.2[3,2] = -1/2
CIAB2 = scary.CI(Salary,MSE,equal.weights = TRUE,Sheffe,"AB",AB.cs.2)
```