University of California, Davis

**Take Home Project 2**

Topic 1 Question 1: Helicopter Count & Shift

Topic 2 Question 1: Annual Salary & Profession, Region

Ana Boeriu & Julia Tien

STA 106

Dr. Erin Melcon

March 9, 2018

**PROJECT 1 TOPIC 1 - Helicopter**

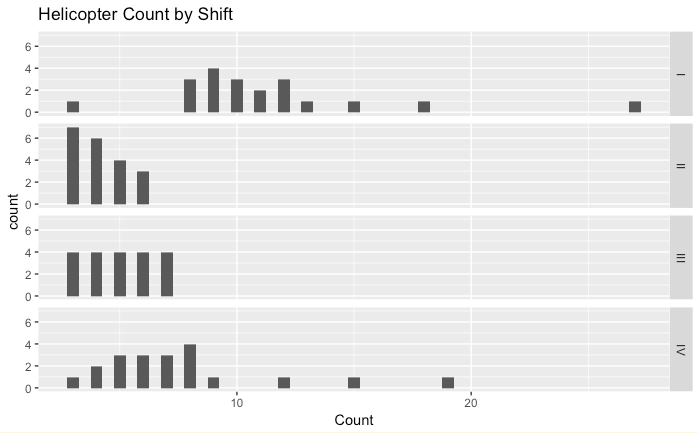
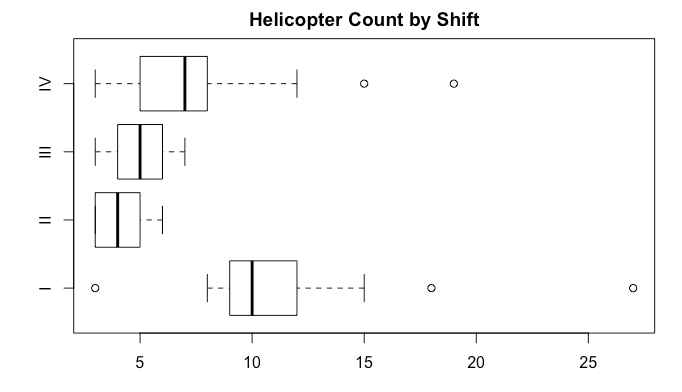
**Transformation of Variables**

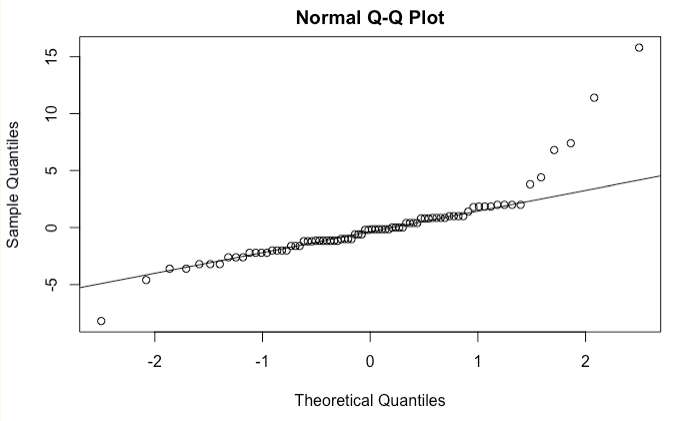
**a. Introduction**

Given data on the amount of helicopters requested over different times of the day for a sheriff’s office, the goal for this report is to transform the data so it best fits the ANOVA model. We will be looking at assumptions of normality and constant variance, the best transformation and if outliers must be removed using a Single Factor ANOVA model with variables Count (number of helicopters called) and Shift (I, II, III, IV).

**b. Diagnostics**

In this section we perform diagnostics to ensure that the data meets the assumptions of ANOVA which are that all Yijk are randomly sampled, all group levels of both factors are independent, and errors are distributed normally with mean of 0 and variance of σ2ɛ. Here we look at normality and constant variance.

1. **Boxplot & Histogram**
2. **Testing for Normality**
3. **QQ Plot**



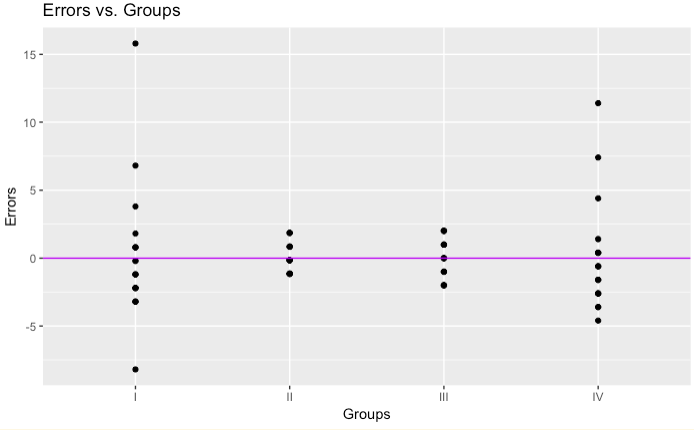
There are clear outliers on this QQ plot as many values leave the perfect distribution Y=X line especially on the upper right side of the plot.

1. **Shapiro-Wilkes Test**

|  |
| --- |
| P-value: 3.945e-09 |

Because this p-value is less than alpha of 0.05, we reject the null and conclude there is evidence to suggest that the errors of the original dataset are non-normal.

1. **Testing for Constant Variance**
2. **Errors Vs. Groups Plot**



This plot shows a rather unequal spread of data.

**b. Brown-Forsythe Test**

|  |
| --- |
| P-value: 0.03185955 |

Because this p-value is less than an alpha value of 0.05, we reject the null hypothesis and there is evidence to conclude this data does not have equal variance.

1. **Model Fit**

Group Means Model: Y=μi+εij

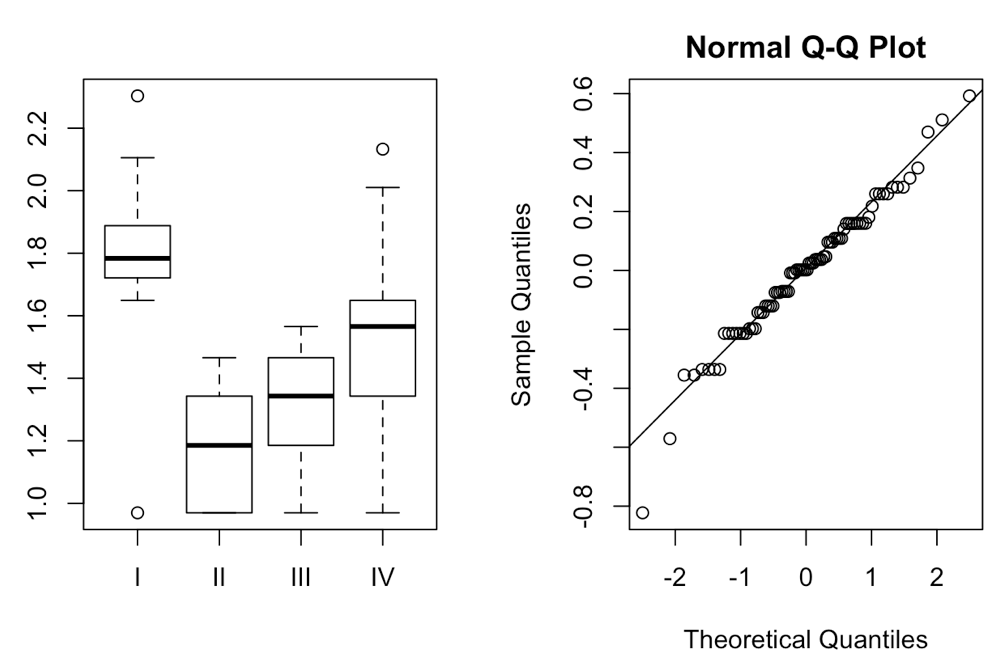
We are using MSE to estimate the variance of the errors.

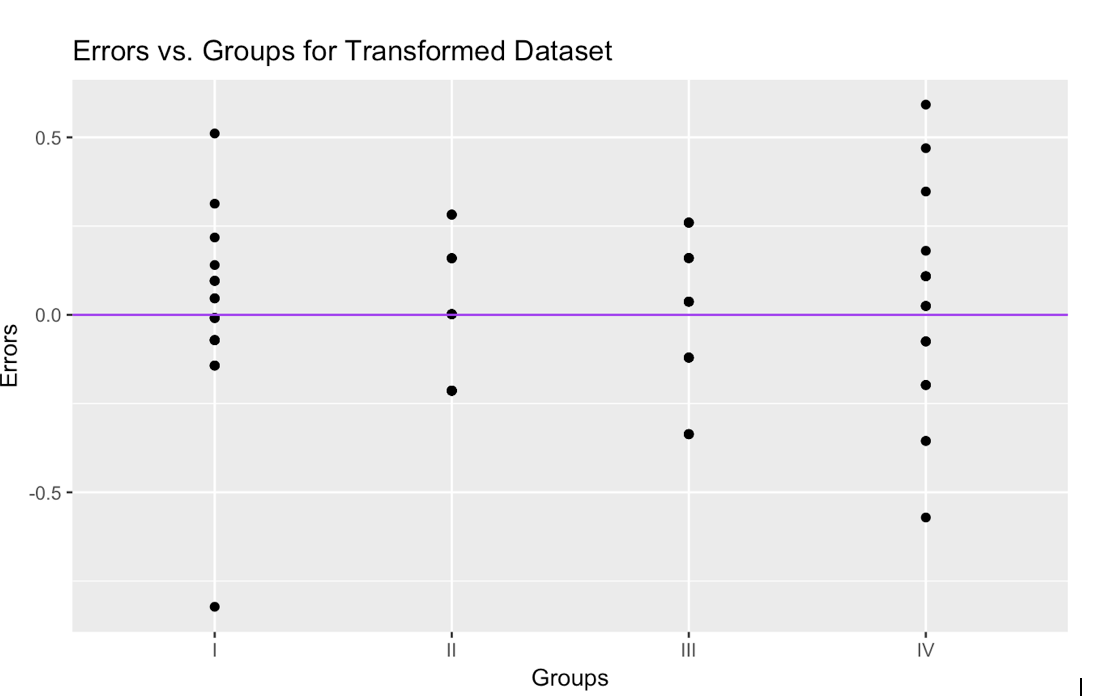
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| I | II | III | IV | MSE | Overall Mean |
| 11.20 | 4.15 | 5.00 | 7.60 | 10.297 | 6.987 |

**c. Outliers & Transformations**

1. **Box-Cox Transformations**
   1. **Shapiro-Wilks**

|  |
| --- |
| Lambda: -0.2322012 |

* + 1. **Plots**

Here we can see that the data points are more closely aligned to the perfectly normal distribution X=Y line and vertical spread is more equal.

* + 1. **Tests**

|  |
| --- |
| Shapiro-Wilks P-value: 0.109024 |

Fail to reject the null at alpha of 0.05, so conclude this transformed data has normality.

|  |
| --- |
| Brown-Forsythe P-value: 0.6467526 |

Fail to reject the null at an alpha of 0.05, so we conclude constant variance.

* 1. **Log-likelihood Transformation**

|  |
| --- |
| Lambda: -0.3964138 |

* + 1. **Tests**

|  |
| --- |
| Shapiro-Wilks P-value: 0.06738834 |

Fail to reject the null, so conclude this transformed data is normally distributed.

|  |
| --- |
| Brown-Forsythe P-value: 0.6596686 |

Fail to reject the null, so conclude constant variance.

Both transformations fulfill assumptions of normality and constant variance in the data. However, because the transformation using the Shapiro-Wilks transformation has a higher p-value for the Shapiro-Wilks normality test and therefore is more likely to be normal, we will proceed with the data that has been transformed using the Shapiro-Wilks lambda. Plots are shown for only this transformation above.

1. **Outliers**

We did not find any outliers using the semi-studentized method and an alpha of 0.05 with the best model of Shapiro-Wilks transformed data so no data points were removed.

**d. Discussion**

Transforming the data did help. Some downsides are that it is hard to interpret as the units change and we cannot use old units to interpret the values. We do believe the transformed data is a better fit because the p-values testing for normality and constant variance both increase and we are able to reject the nulls so that assumptions for normality and constant variance are met. For a client who wants to use this data set for ANOVA, we would suggest to use the Shapiro-Wilks transformation and no outliers would need to be removed. Those who may be looking into how many helicopters sheriff’s offices need on average may find this model useful to draw further conclusions from.

**PROJECT 2 - Salary**

**Two Factor ANOVA**

**I. Introduction**

The goal of this project is to build a model using data of annual salaries for Data Scientists, Software Engineers, and Bioinformatics Engineers in San Francisco and Seattle. We are using a Two Factor ANOVA approach and will test for various effects to build the best model, perform diagnostics, calculate confidence intervals and perform other statistical measures to analyze the data. People who are pursuing one of the professions we are looking at or students who are trying to get an idea of what to study may be interested in using this model to decide where they should work and which careers they can earn the most in.

**II. Summary of Data**

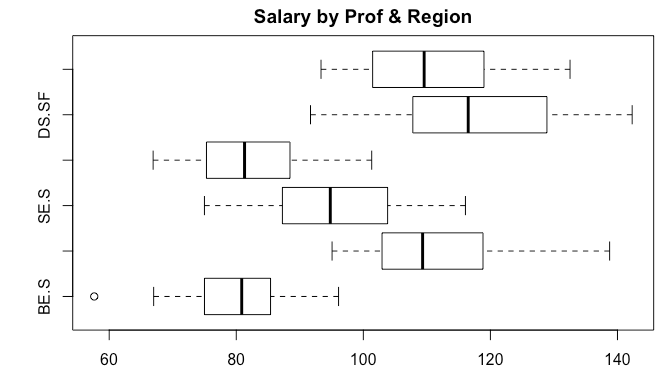
* 1. **Summary Statistics**

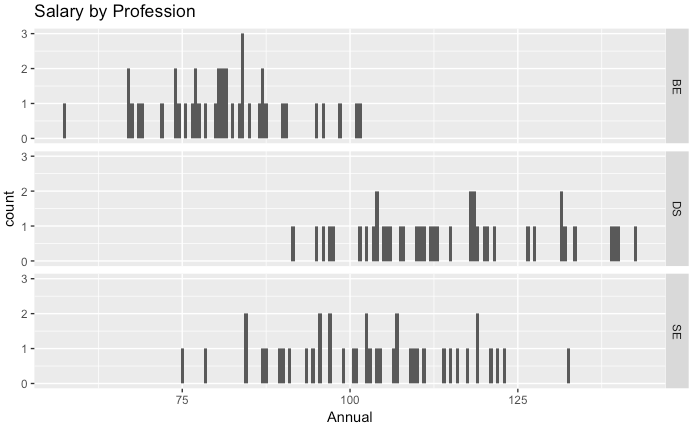
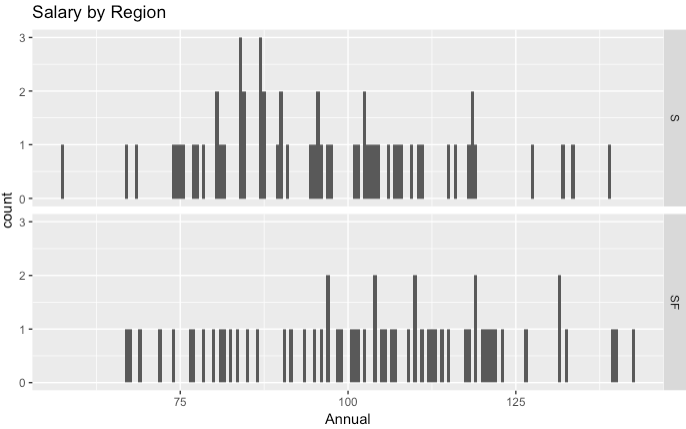
The first value in each box is the mean while the second value is the standard deviation. Units are in thousands of dollars.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  | j=1 | j=2 |  |
|  | (Units are thousands of dollars) | Seattle | San Francisco | Overall profession |
| i=1 | BE | 79.75485  (8.786628) | 82.41914  (10.521476) | 81.0870  (9.662515) |
| i=2 | DS | 112.52715  (12.838566) | 117.76883  (14.289227) | 115.1480  (13.668190) |
| i=3 | SE | 95.54875  (11.598722) | 110.26412  (10.551705) | 102.9064  (13.240313) |
|  | Overall region means | 95.94358  (17.41791) | 103.48403  (19.29842) | Overall mean:  99.71381  Overall SD:  18.69226 |

|  |  |  |
| --- | --- | --- |
| Total sample size | Factor A level (Profession) | Factor B levels (Region) |
| 120 | 3 | 2 |

* 1. **Boxplot & Histograms**

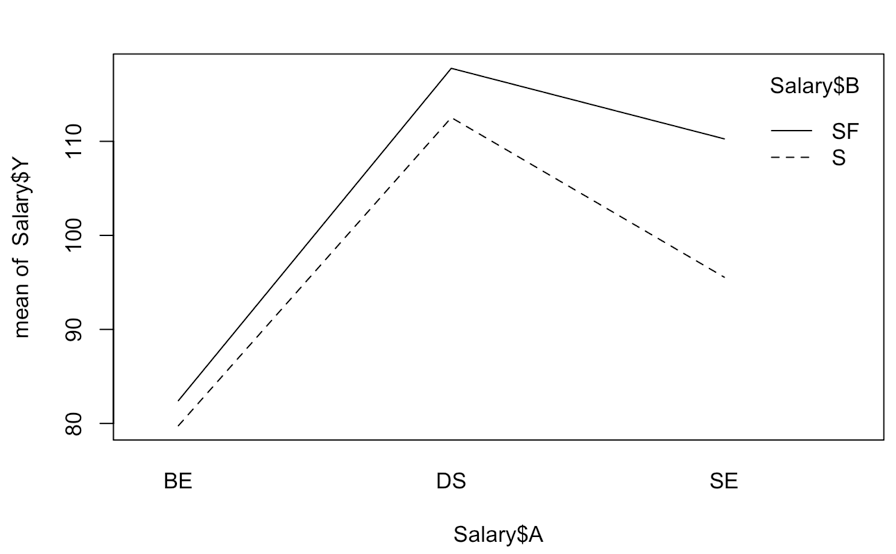


We can see from this section that salaries for Data Scientists have higher salaries on average than that of Bioinformatics Engineers and Software Engineers and Software Engineers have higher salaries on average than Bioinformatics Engineers. Another trend is that there is more variance in salaries in Seattle than in San Francisco.

**III. Analysis & Interpretation: Model Fit**

In this section we are choosing the best model to perform further analyses on. First we will test for interactions and if none are found, test for Profession effects and Region effects.

**A. Interaction Plot**

****

According to this interaction plot we can conclude that there are no interactions between profession and region. Furthermore to support this claim we can see in part IIA that the group means for each profession and region follow relatively the same trend.

**B. Test for Interaction Effects**

Full model**:** Yijk= μ.. + ɣi + 𝛿j+(ɣ𝛿)ij+ εijk

Reduced model: Yijk= μ.. + ɣi + 𝛿j+εijk

*:*

|  |  |
| --- | --- |
| F-statistic | P-value |
| 3.0098 | 0.05324 |

Because the p-value is greater than alpha 0.05, we fail to reject the null hypothesis and conclude that there is no interaction between profession and region.

If in reality the reduced model with no interaction term between Profession and Region was a better statistical fit, we would see our data or more extreme 5.32% of the time.

Because we have concluded a no interactions model, next we must test for Profession and Region effects.

**C. Test for Factor A Effects**

Full model : Yijk= 𝜇.. + ɣi+𝜹j +εijk

Reduced model: Yijk= 𝜇.. +𝜹j +εijk

|  |  |
| --- | --- |
| F-statistic | P-value |
| 86.014 | 2.2e-16 |

Because the p-value is less than alpha of 0.05 we reject the null hypothesis and conclude that Profession (Factor A) effects exist and so the full model with Profession effects fits significantly better than the reduced model.

1. **Partial R2**

R2{A+B|B} = 0.5972622

When we add Profession to a model that already contains Region we reduce error by 59.7%. This further shows that Profession is an important factor for this model since we significantly reduced the error.

**D. Test for Factor B effects**

Full model : Yijk= μ.. + ɣi + 𝛿j+εijk

Reduced model: Yijk= μ.. + ɣi + εijk

|  |  |
| --- | --- |
| F-stat | P-value |
| 6.2561 | 0.0005712 |

Because the p-value is less than alpha of 0.05 we reject the null and there is sufficient evidence to conclude that Region (Factor B) effects exist and so the full model with Region effects fits significantly better than the reduced model.

1. **Partial R2**

R2{A+B|A} = 0.09602243

When we add Region (Factor B) to a model that already has Profession (Factor A) effects we reduce our error by 9.6%. Although this value may seem small in comparison to the previous, it still shows a significant reduction in error.

**E. Final Model**

According to the above tests we can conclude that the best model is one with no interaction effect between profession and region but with both variables still in the model. Thus, the best model is:

|  |
| --- |
| Yijk= μ.. + ɣi + 𝛿j+ εijk |

Where the Gammas are:

|  |  |  |
| --- | --- | --- |
| BE (ɣ1) | DS (ɣ2) | SE (ɣ3) |
| -18.626812 | 15.434183 | 3.192629 |

And the Deltas are:

|  |  |
| --- | --- |
| Seattle (𝛿1) | San Francisco (𝛿2) |
| -3.770225 | 3.770225 |

A general interpretation of a gamma would be the change in the overall average salary when a Bioinformatic Engineer, Data Scientist, or Software Engineer ignoring interactions effects or effects of region.

For example the interpretation of BE would be: The average change in overall average salary when a subject is a Bioinformatic Engineer is a decrease of $18,626.81.

A general interpretation of a delta would be the change in overall average salary when in Seattle or San Francisco, ignoring effects of Profession.

For example the interpretation of Seattle would be: The average change in overall average salary when a subject works in Seattle is a decrease of $3,770.26.

**IV. Diagnostics**

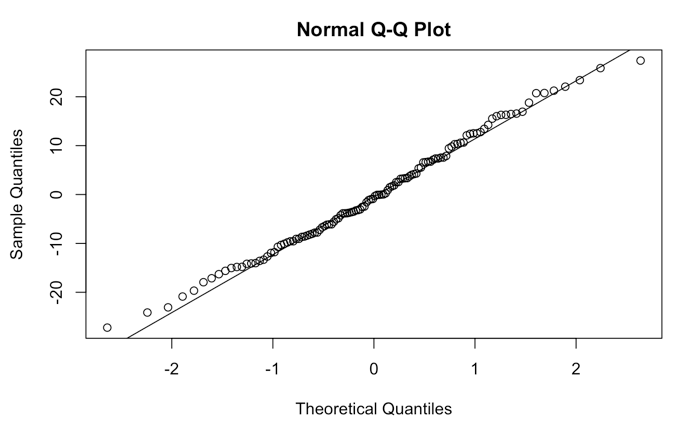
With our “best” model of no interaction effects, we will perform diagnostics to ensure that it meets the assumptions of ANOVA which are that all Yijk are randomly sampled, all group levels of both factors are independent, and errors are distributed normally with mean of 0 and variance of σ2ɛ. In this section we will be testing the assumptions of normality and constant variance.

1. **Outliers**

Using the Studentized and Semi-Studentized method, we found no outliers using alpha of 0.05 so we will not remove any data points.

**B. Testing for Normality**

1. **QQ Plot**



The residuals looks rather normal as the data points are lying very close to the perfect distribution Y=X line.

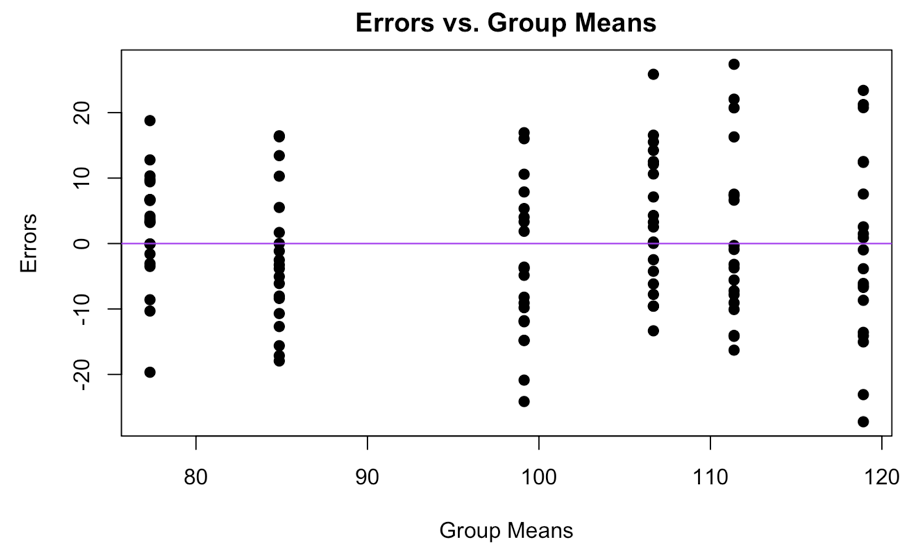
1. **Shapiro-Wilks**

|  |
| --- |
| P-value: 0.6698 |

Because our p-value is greater than the alpha of 0.05, there is evidence to fail to reject the null and we conclude that the errors are approximately normally distributed.

**C. Testing for Constant Variance**

1. **Errors vs. Groups**



This plot shows a rather equal spread of data across the plot.

1. **Brown-Forsythe Test**

|  |
| --- |
| P-value: 0.3048319 |

Because this p-value is greater than an alpha value of 0.05, we fail to reject the null hypothesis and there is evidence to conclude there is equal variance.

**D. Conclusion**

From this section we can conclude assumptions of normality and constant variance are met and no outliers need to be removed. We do not need to look at transformations of the data.

**V. Analysis & Interpretation: Confidence Intervals & Regression**

Now that we have concluded our model is reliable, we can draw conclusions from the data such as confidence intervals to further analyze the data of salaries.

Since there are equal sample sizes for each factor level we can use the equal weights formula.

1. **Pairwise Confidence Intervals** 
   1. **Multipliers**

|  |  |  |  |
| --- | --- | --- | --- |
| alpha | Bonferroni | Tukey | Scheffe |
| 0.05 | 2.429 | 2.374 | 2.480 |

For pairwise confidence intervals we can use either the Bonferroni, Tukey or Scheffe multiplier so in this case we will use Tukey because that is the smallest of the three for intervals i-iii.

**b. Intervals**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **i)** BE-SE | Lower Bound | Upper Bound |  | **ii)** BE-DS | Lower Bound | Upper Bound |
| -21.81944 | -28.10459 | -15.53429 |  | -34.06100 | -40.34614 | -27.77585 |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **iii)** DS-SE | Lower Bound | Upper Bound |  | **iv)** S-SF | Lower Bound | Upper Bound |
| 12.241555 | 5.956406 | 18.526703 |  | -7.540449 | -11.724029 | -3.356869 |

**i)** We are overall 95% confident that the true average annual salary for Software Engineers is greater than that of a Bioinformatics Engineer by between $28,104.59 and $15,534.29 thousands of dollars.

**ii)** We are overall 95% confident that the true average annual salary for Data Scientists is greater than that of Bioinformatics Engineers by between $27,775.85 and $40,346.14.

**iii)** We are overall 95% confident that the true average annual salary for Data Scientists is greater than that of Software Engineers by between $5,956.41 and $18,526.70.

**iv)** For this interval we assumed the number of intervals (g) is 1 and a multiplier of 1.981. We are overall 95% confident that the true average annual salary for those who reside in San Francisco is greater than that of those who reside in Seattle by between $3,356.87 and $11,724.03.

**2. Contrast Confidence Intervals**

**a. Multipliers**

|  |  |  |
| --- | --- | --- |
| Bonferroni | Tukey | Scheffe |
| 2.271 | 2.374 | 2.480 |

Because contrast intervals are of interest we will choose between Bonferroni or Scheffe. Typically Scheffe is more commonly used for contrasts, however, because Bonferroni is smaller we will be using that value for the following CI.

**b. Intervals**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **i)** μ31- | Lower bound | Upper bound |  | **ii)** μ22- | Lower bound | Upper bound |
| -0.592250 | -7.786276 | 6.601776 |  | 30.11703 | 22.92301 | 37.31106 |

**i)** We are overall 95% confident that there is no significant difference between the mean salary of Software Engineers in Seattle and the average of the mean salaries for Bioinformatic Engineers and DataScientists in Seattle.

**ii)**We are overall 95% confident that the true mean salary for Data Scientists in San Francisco is higher than the average of the mean salaries for Bioinformatic Engineers and Software Engineers in Seattle by between $22,923.01 and $37,311.06.

**C. Regression Formation**

Using regression formation, we can make further analyses which compare the different professions and regions to one another.

The line of best fit:

|  |
| --- |
|  |

**(i) (ii) (iii) (iv)**

**i)** The estimated average salary for Bioinformatic Engineers in Seattle is $77,316.77.

**ii)** Holding region in Seattle constant, the estimated difference in average salary for Data Scientists versus Bioinformatic Engineers is $34,061.

**iii)** Holding region in Seattle constant, the estimated difference in average salary for Software Engineers versus Bioinformatic Engineers is $21,819.44.

**iv)** Holding profession at Bioinformatic engineer constant, the estimated difference in average salary in San Francisco versus Seattle is $7,540.45.

**V**. **Conclusion**

Through this project we have found that an ANOVA Two Factor Model without interactions but with individual effects from Profession and Region have statistical significance in determining annual salary using this dataset. After performing diagnostics with that model in mind, we concluded assumptions of normality and constant variance were met and no outliers needed to be removed so we proceeded without performing transformations. Through calculating six confidence intervals, we formed various conclusions about salary in San Francisco and Seattle for the different professions. We then used regression formation to find average differences for various combinations between the variables. Those who are interested in these professions will be able to use this model to good use in determining where they should pursue their career.

**R Appendix**

**PROJECT 1**

**b. Diagnostics**

**Boxplot**

boxplot(Count ~ Shift, data = Helicopter, main = "Helicopter Count by Shift", horizontal = TRUE)

**Histogram**

library(ggplot2)

ggplot(Helicopter, aes(x = Count)) + geom\_histogram(binwidth = 0.5) + facet\_grid(Shift ~.) +ggtitle("Helicopter Count by Shift")

**Diagnostics**

**Testing for Normality**

**QQ Plot**

h.model = lm(Count ~ Shift, data = Helicopter)

qqnorm(h.model$residuals)

qqline(h.model$residuals)

**Shapiro-Wilkes Test**

h.ei = h.model$residuals

the.SWtest = shapiro.test(h.ei)

the.SWtest

**Testing for Constant Variance**

**Errors Vs. Groups**

qplot(Shift, h.ei, data = Helicopter) + ggtitle("Errors vs. Groups") + xlab("Groups") + ylab("Errors") + geom\_hline(yintercept = 0,col = "purple")

**Brown-Forsythe Test**

**Model Fit**

aggregate(Count ~ Shift, data = Helicopter, mean)

sd(Helicopter$Count)

anova.table = anova(h.model)

Anova.table

SSE= anova.table[2,2]

SSE

**Outliers and Transformations**

**Box-Cox**

library(EnvStats)

h.model = lm(Count ~ Shift, data = Helicopter)

boxcox(h.model ,objective.name = "PPCC")

boxcox(h.model ,objective.name = "Shapiro-Wilk")

boxcox(Helicopter$Count,objective.name = "Log-Likelihood")

L2 = boxcox(h.model ,objective.name = "Shapiro-Wilk",optimize = TRUE)$lambda

L3 = boxcox(Helicopter$Count,objective.name = "Log-Likelihood",optimize = TRUE)$lambda

L2

L3

**Transformed data(SW)**

par(mfrow = c(1,2))

YTS = (Helicopter$Count^(L2)-1)/L2

tS.data = data.frame(Count = YTS, Shift = Helicopter$Shift)

tS.model = lm(Count~Shift,data = tS.data)

plot(tS.data$Shift, tS.data$Count)

qqnorm(tS.model$residuals)

qqline(tS.model$residuals)

**Transformed data (LL)**

qqnorm(t.model$residuals)

qqline(t.model$residuals)

qplot(Shift, h.ei, data = t.data) + ggtitle("Errors vs. Groups") + xlab("Groups") + ylab("Errors") + geom\_hline(yintercept = 0,col = "purple")

h.ei.t = t.model$residuals

the.SWtest.t = shapiro.test(h.ei.t)

the.SWtest.t

library(car)

the.BFtest.t = leveneTest(h.ei.t~ Shift, data=t.data, center=median)

p.val.t = the.BFtest.t[[3]][1]

P.val.t

**Outliers**

t.data$ei = t.model$residuals

nt = nrow(t.data) #Calculates the total sample size

a = length(unique(t.data$Shift)) #Calculates the value of a

SSE.t = sum(t.data$ei^2) #Sums and squares the errors (finds SSE)

MSE.t = SSE.t/(nt-a) #Finds MSE

eij.star = t.model$residuals/sqrt(MSE.t)

alpha = 0.05

t.cutoff= qt(1-alpha/(2\*nt), nt-a)

CO.eij = which(abs(eij.star) > t.cutoff)

CO.eij

**PROJECT 2**

**Summary of the data**

aggregate(Y ~ A + B, data = Salary, sd)

sd(Salary$Y)

aggregate(Y ~ A, data = Salary, sd)

aggregate(Y ~ B, data = Salary, sd)

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)

}

the.means = find.means(the.data)

the.model = lm(Y ~ A\*B, data = the.data)

SSE = sum(the.model$residuals^2)

MSE = SSE/(nt-a\*b)

the.means

mean(the.data$Y)

the.data = Salary

nt = nrow(the.data)

nt

a = length(unique(the.data[,2]))

a

b = length(unique(the.data[,3]))

b

names(the.data) = c("Y","A","B")

Boxplot: boxplot(Annual ~ Prof + Region, data = Salary, main = "Salary by Prof & Region",

horizontal = TRUE)

1. **Model Fit**
2. **Interaction Plot**

names(Salary) = c("Y","A","B")

interaction.plot(Salary$A, Salary$B, Salary$Y)

**B. Testing for interaction effect**

names(Salary) = c("Y","A","B")

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"

SSE

anova(A.B,AB)

**C. Test for Factor A effects**

anova(B,A.B)

1. **Partial R2**

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)

}

Partial.R2(B,A.B)

**D.Test for Factor B effects**

anova(A,AB)

1. **Partial R2**

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)

}

Partial.R2(A,A.B)

**E. Final model**

A.B = lm(Y ~ A+B,Salary)

get.gamma.delta = function(the.model,the.data){

nt = nrow(the.data)

a = length(unique(the.data[,2]))

b = length(unique(the.data[,3]))

the.data$hat = the.model$fitted.values

the.ns = find.means(the.data,length)

a.vals = sort(unique(the.data[,2]))

b.vals= sort(unique(the.data[,3]))

muij = matrix(nrow = a, ncol = b)

rownames(muij) = a.vals

colnames(muij) = b.vals

for(i in 1:a){

for(j in 1:b){

muij[i,j] = the.data$hat[which(the.data[,2] == a.vals[i] & the.data[,3] == b.vals[j])[1]]

}

}

mi. = rowMeans(muij)

m.j = colMeans(muij)

mu.. = sum(muij)/(a\*b)

gammai = mi. - mu..

deltaj = m.j - mu..

gmat = matrix(rep(gammai,b),nrow = a, ncol = b, byrow= FALSE)

dmat = matrix(rep(deltaj,a),nrow = a, ncol = b,byrow=TRUE)

gamma.deltaij =round(muij -(mu.. + gmat + dmat),8)

results = list(Gam = gammai, Del = deltaj, GamDel = gamma.deltaij)

return(results)

}

Wow = get.gamma.delta(A.B, the.data)

Wow

**II. Diagnostics**

1. **Outliers**

theS.model = lm(Annual ~ Prof+Region, data = Salary)

Salary$ei = the.model$residuals

nt = nrow(Salary) #Calculates the total sample size

a = length(unique(Salary$Prof))

SSE = sum(Salary$ei^2) #Sums and squares the errors (finds SSE)

MSE = SSE/(nt-a) #Finds MSE

eij.star = the.model$residuals/sqrt(MSE)

alpha = 0.05

t.cutoff=qt(1-0.01,nt -a)

t.cutoff= qt(1-alpha/(2\*nt), nt-a)

CO.eij = which(abs(eij.star) > t.cutoff)

CO.eij

rij = rstandard(the.model)

CO.rij = which(abs(rij) > t.cutoff)

CO.rij

**B. Testing for Normality**

1. **QQ plot**

qqnorm(theS.model$residuals)

qqline(theS.model$residuals)

1. **SW test**

ei = theS.model$residuals

the.SWtest = shapiro.test(ei)

the.SWtest

**C. Testing for Constant Variance**

1. **Errors vs Groups**

plot(theS.model$fitted.values, theS.model$residuals, main = "Errors vs. Group Means",xlab = "Group Means",ylab = "Errors",pch = 19)

abline(h = 0,col = "purple")

**2. BF test**

the.BFtest = leveneTest(ei~ Prof\*Region, data=Salary, center=median)

p.val = the.BFtest[[3]][1]

p.val

**Confidence Intervals**

1. **Check for equal weight**

sum(the.data$A=="BE"& the.data$B=="SF")

sum(the.data$A=="SE"& the.data$B=="SF")

sum(the.data$A=="DS"& the.data$B=="SF")

sum(the.data$A=="BE"& the.data$B=="S")

sum(the.data$A=="DS"& the.data$B=="S")

sum(the.data$A=="SE"& the.data$B=="S")

**B. Multipliers**

all.mult = find.mult(alpha = 0.05, a = 3, b = 2, dfSSE = nt-a-b+1, g = 3, group = "A")

All.mult

**C.Pairwise**

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)

}

}

Tuk = find.mult(alpha = 0.05, a = 3, b = 2, df{SSE} = nt - a\*b, g = 3, group = "A")[1]

A.cs.1 = c(1,0,-1)

A.cs.2 = c(1,-1,0)

A.cs.3=c(0,1,-1)

scary.CI(the.data,MSE,equal.weights = TRUE,Tuk,"A",A.cs.1)

scary.CI(the.data,MSE,equal.weights = TRUE,Tuk,"A",A.cs.2)

scary.CI(the.data,MSE,equal.weights = TRUE,Tuk,"A",A.cs.3)

all.mult = find.mult(alpha = 0.05, a = 3, b = 2, dfSSE = nt-a-b+1, g = 1, group = "B")

all.mult

Bon = find.mult(alpha = 0.05, a = 3, b = 2, dfSSE = nt - a-b+1, g = 1, group = "B")[1]

A.cs.1 = c(1,-1)

scary.CI(the.data,MSE,equal.weights = TRUE,Bon,"B",A.cs.1)

**D. Contrasts**

all.mult = find.mult(alpha = 0.05, a = 3, b = 2, dfSSE =nt-a-b+1, g = 2, group = "AB")

All.mult

AB.cs = matrix(0,nrow = a, ncol = b)

AB.cs

the.means$AB

AB.cs[3,1] = 1

AB.cs[1,1] = -1/2

AB.cs[2,1]=-1/2

scary.CI(Salary,MSE,equal.weights = TRUE,Bon,"AB",AB.cs)

AB.cs = matrix(0,nrow = a, ncol = b)

AB.cs

the.means$AB

AB.cs[2,2] = 1

AB.cs[1,1] = -1/2

AB.cs[3,1]=-1/2

scary.CI(Salary,MSE,equal.weights = TRUE,Bon,"AB",AB.cs)

**D. Regression Formation**

names(Salary) = c("Y","A","B")

thebestS.model = lm(Y ~ A+B,data = Salary)

thebestS.model$coefficients