University of California, Davis

Take Home Project 1

Question 2: Diet & Weight Loss

Ana Boeriu & Julia Tien

STA 106

Dr. Erin Melcon

February 9, 2018

I. Introduction

The goal of this project is to ultimately find out which diet is most effective in helping people lose weight using this sample of 76 people. We will perform different statistical measures such as exploratory data analysis, hypothesis tests, diagnostics, and form pairwise confidence intervals to analyze the data and reach conclusions. The approach we are going to take is using ANOVA which is typically the method used when testing differences between means of groups. More specifically we are using the Group Means Model. In this case, there is the categorical explanatory (X) variable with three diets (A, B, C) and the numerical response variable (Y) which is the difference in pounds. A positive number suggests a loss in weight while a negative amount means there was an increase in weight. Dieticians or anyone in the general population who is trying to lose weight may be interested in this outcome as it will help them be more effective in determining what lifestyle changes may need to be made.

II. Summary (This should include things like plots (histograms, box plots) including the interpretation of the plots, and summary values such as sample means and standard deviations. You should have an idea about the trend of the data from this section.)

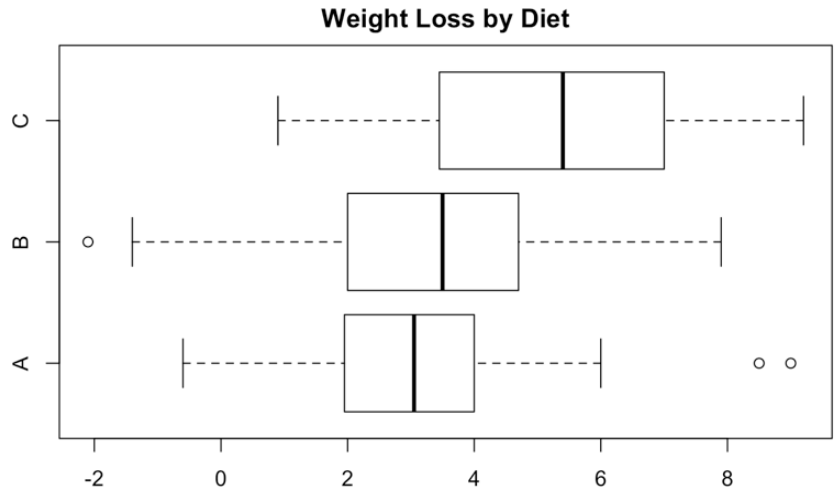
A. Summary of Data:

Diet	Number of people	Mean	Standard deviation
A	24	3.300	2.240148
B	25	3.2680	2.464535
C	27	5.2333	2.247734

Weight Loss Summary Statistics:

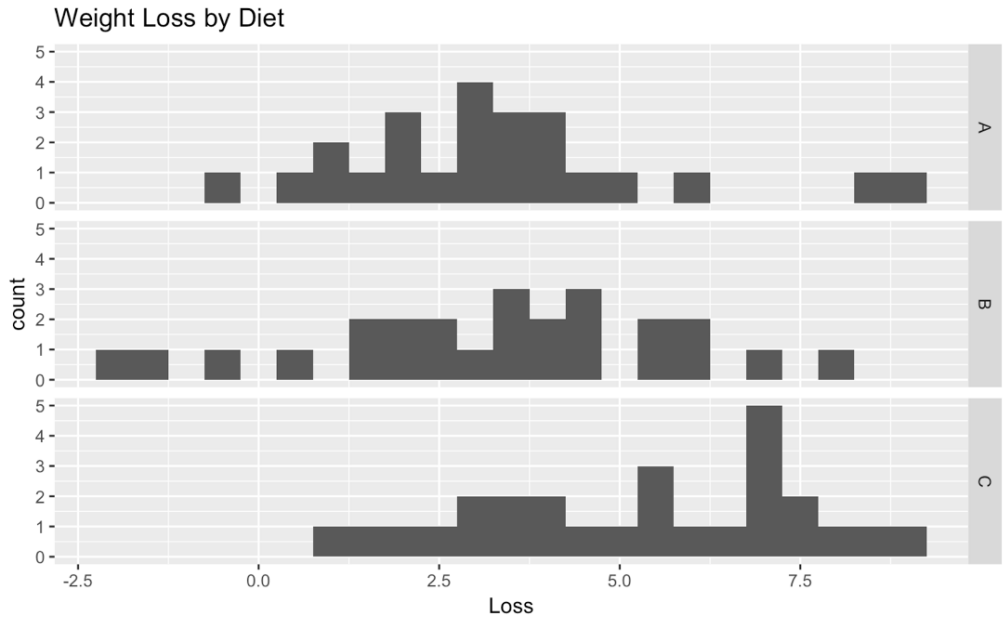
Min:	-2.100	Mean	3.976
1st Quartile	2.35	3rd Quartile	5.650
Median	3.700	Max	9.20
Standard Deviation	2.473156		

B. Box Plot



This boxplot shows that Diet C had greatest average weight loss as well as highest minimum weight loss. Diet B had the greatest spread of data while Diet A had the lowest mean and maximum.

C. Histogram



This histogram shows that none of the weight loss by diet groups are normally distributed although Diet A could be considered closest to it. Diet B seems to have the greatest spread of data.

The trends seen in this section are that Diet C seems to have the greatest weight loss when compared to Diets A and B.

III. Diagnostics

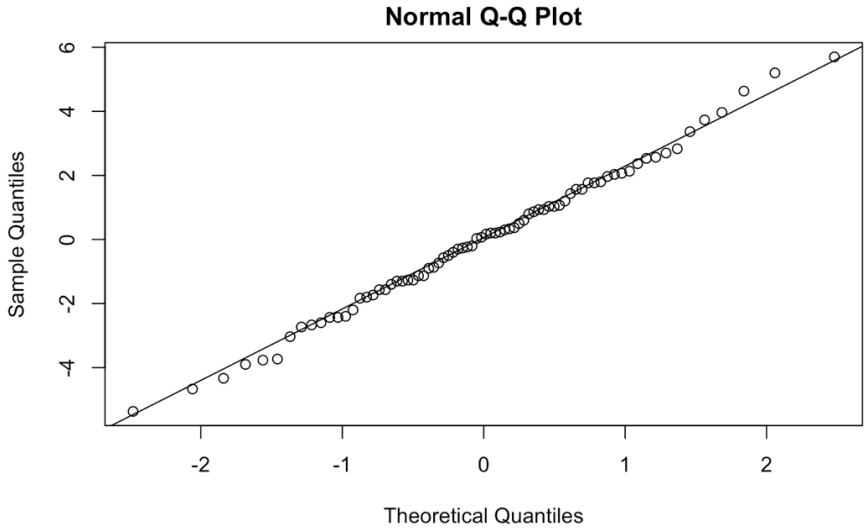
Diagnostics are performed to evaluate how well the data meets the assumptions for ANOVA. The assumptions of ANOVA are that the data is randomly sampled and independent, the different diet groups are independent, and the errors are normally distributed with a mean of 0 and a variance of σ^2. In this section we will first look for outliers and then use diagnostics to test these assumptions, specifically for normality and constant variance in the errors. We will be using a 5% significance level (alpha of .050) as that is conservative yet not too strict.

A. Outliers

While the boxplot from Section IIB may show that there are three outliers with Diet A, when completing the semi-studentized and standardized residual analyses we found that there are no outliers, so we have not removed any data points.

B. Testing for Normality

1. QQ Plot



This QQ plot shows most of the data falls on the X=Y which could signify normal errors, however this is a subjective judgement.

2. Shapiro-Wilks

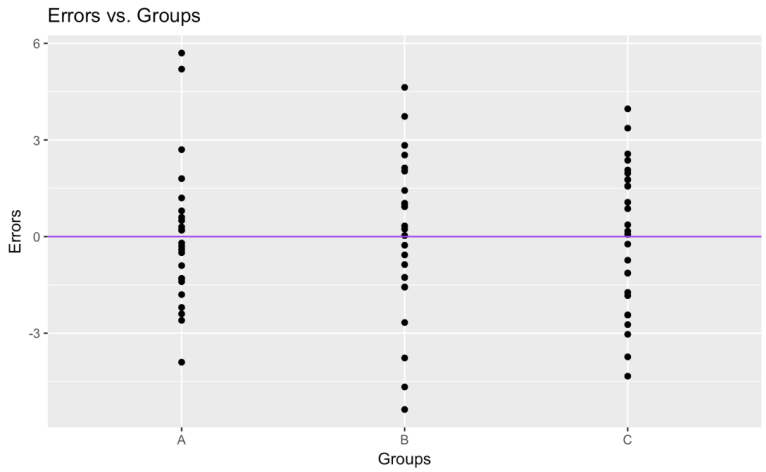
$Ho: Data\ is\ normal$ $Ha: Data\ is\ non-normal$

P-value = 0.9921

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 our data is approximately normally distributed.

C. Testing for Constant Variance

1. Errors vs. Groups



This plot shows that there does seem to be constant variance in the errors between groups.

2. **Brown-Forsythe Test**

Ho: *Residuals have equal variance* Ha: *Residuals do not have equal variance*

P-value: 0.694648

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

As seen in the plots and performing statistical tests, the data is approximately normal and has constant variance so this assumption for ANOVA is met and we can proceed to use ANOVA with the data and make inferences.

IV. Analysis & Interpretation

A. Model Fit

Group Means Model: $Y=\mu_i+\varepsilon_{ij}$

For μ_i the estimate for overall sample mean is 3.976. The individual sample means are estimated to be: $\underline{Y}_A$= 3.3 $\underline{Y}_B$= 3.268 $\underline{Y}_C$ = 5.233 as seen from the table in part IIA. The estimate of the variance of the errors would be MSE which is 5.377.

MSE	Overall Mean
5.377	3.976

B. Test-Statistics & P-Values

Ho: $There\ is\ no\ significant\ difference\ in\ true\ weight\ loss\ between\ diet\ groups$
Ha: $At\ least\ one\ of\ the\ true\ average\ weight\ loss\ per\ groups\ is\ unequal$

F-statistic	P-value
6.1537	0.00339

Because our p-value is less than alpha 0.05, we reject the null hypothesis and conclude that there is a significant difference in weight loss by diet.

C. Confidence Intervals with Tukey

We decided to use the Tukey multiplier because these are all pairwise intervals and when compared to Scheffe and Bonferroni, Tukey is the smallest.

Tukey	Scheffe	Bonferroni
2.392435	2.498841	2.450398

1. Diet A vs Diet B

Estimate	Lower bound	Upper bound
0.032000	-1.553445	1.617445

Because zero is contained in the interval we are overall 95% confident that there is no difference in true average weight loss between Diet A and Diet B.

2. Diet A vs Diet C

Estimate	Lower Bound	Upper bound
-1.9333333	-3.4897514	-0.3769153

We are overall 95% confident that the true average weight loss for people on Diet A is less than the true average weight loss for people on Diet C by between .037 and 3.48975. This means people on Diet A lost less weight than people on Diet C on average by between 0.377 and 3.49 pounds.

3. Diet B vs Diet C

Estimate	Lower Bound	Upper bound
-1.9653333	-3.5051835	-0.4254832

We are overall 95% confident that the true average weight loss for people on Diet B is less than the true average weight loss for people on Diet C by between .425 and 3.505. This means people on Diet B lost less weight than people on Diet C by between 0.42 and 3.51 pounds on average.

There is significant difference in Diets A and C and B and C but not between A and B. In comparing A and C and B and C, Diet C seems to be the most effective in helping people lose weight.

4. **Average of Diets A and Diet B vs Diet C**

Because there was no significant difference in means between Diets A and B, we calculated a contrast interval to further support that Diet C is the most effective in terms of weight loss.

Lower bound	Upper Bound
-3.0571150	-0.8415516

We are overall 95% confident that the true average of the means for Diets A and B is less than the mean of Diet C by between .084 and 3.05 pounds. Even after combining the averages of Diets A and B, we can still see that people lost more weight with Diet C.

D. Power Calculation

Power = 0.8777748

The power is 0.878. This means the probability of correctly concluding there is a difference in true weight loss between diets A, B and C when in reality there is a difference is .8778.

E. Sample size

Desired power	Sample size
0.90	n = 27.07278
0.95	n = 32.81085
0.98	n = 39.9276

These are some examples of how many people would need to be in each diet group in order to reach the respective desired power

---

V. Conclusion

In this project we started with a dataset of weight losses from diets A, B and C. We performed exploratory data analysis and hypothesized that the most effective results were from Diet C. Next we performed diagnostics to check our ANOVA assumptions and found that we did not need to remove outliers or transform the data and that the errors were normally distributed and had constant variance. Afterwards we created different confidence intervals and found no

significant difference for effects of Diets A and B but Diet C is significantly more effective in helping people lose weight. Dieticians or those who are trying to lose weight will hopefully take these results into account when deciding how to live a healthier lifestyle.

R Appendix

Introduction

Summary

A. Summary of the Data

```
summary(loseit)
aggregate(Loss ~ Diet, data = loseit, mean)
aggregate(Loss ~ Diet, data = loseit, sd)
sd(loseit$Loss)
sd
```

B. Box Plots

```
boxplot(Loss ~ Diet, data = loseit, main = "Weight Loss by Diet", horizontal = TRUE)
```

C. Histogram

```
library(ggplot)
ggplot(loseit, aes(x = Loss)) +
geom_histogram(binwidth = 0.5) + facet_grid(Diet ~.) +ggtitle("Weight Loss by Diet")
```

Diagnostics

A. Outliers

1. Semi-studentized

```
loseit.model = lm(Loss ~ Diet, data = loseit)

loseit$ei = loseit.model$residuals
nt = nrow(loseit)
a = length(unique(loseit$Diet))
SSE = sum(loseit$ei^2) #Sums and squares the errors (finds SSE)
MSE = SSE/(nt-a) #Finds MSE
eij.star = loseit.model$residuals/sqrt(MSE)
alpha = 0.6
t.cutoff= qt(1-alpha/(2*nt), nt-a)
CO.eij = which(abs(eij.star) > t.cutoff)
CO.eij
```

2. Studentized

```
rij = rstandard(loseit.model)
CO.rij = which(abs(rij) > t.cutoff)
CO.rij
```

B. Testing for Normality

1. QQ Plot

```
qqnorm(loseit.model$residuals)
qqline(loseit.model$residuals)
```

2. Shapiro-Wilks Test

```
ei = loseit.model$residuals
the.SWtest = shapiro.test(ei)
the.SWtest
```

C. Constant Variance

1. Errors vs. Groups

```
library(qqplot)
qplot(Diet, ei, data = loseit) +
ggtitle("Errors vs. Groups") +
xlab("Groups") + ylab("Errors") +
geom_hline(yintercept = 0,col = "purple")
```

2. **Brown Forsythe test**

```
library(car)
the.BFtest = leveneTest(ei~ Diet, data=loseit, center=median)
p.val = the.BFtest[[3]][1]
P.val
```

Analysis/Interpretation

A. Model fit

```
anova.table = anova(loseit.model)
Anova.table
SSE= anova.table[2,2]
SSE
SSA=anova.table[1,2]
SSA
SSTO = var(loseit$Loss)*(nrow(loseit) -1)
SSTO
```

B. Test statistic and p-value

```
anova.table = anova(loseit.model)
anova.table
```

C. Confidence Intervals

```
Tuk = qtukey(1-alpha,a,nt-a)/sqrt(2)
Tuk
S = sqrt((a-1)*qf(1-alpha, a-1, nt-a))
S
g=3
B = qt(1-alpha/(2*g),nt-a)
B
```

1. Diet A vs Diet B

```
group.means =  by(loseit$Loss, loseit$Diet,mean)
group.nis = by(loseit$Loss,loseit$Diet,length)
loseit.model = lm(Loss ~ Diet, data = loseit)
anova.table = anova(loseit.model)
MSE = anova.table[2,3]
nt = sum(group.nis)
a = length(group.means)
alpha = 0.05
Tuk = qtukey(1-alpha,a,nt-a)/sqrt(2)
Tuk
ci = c(1,-1,0)
give.me.CI(group.means,group.nis,ci,MSE,Tuk)
```

2. Diet A vs Diet C

```
ci = c(1,0,-1)
give.me.CI(group.means,group.nis,ci1, MSE,Tuk)
```

3. Diet Bvs Diet C

```
ci = c(0,1,-1)
give.me.CI(group.means,group.nis,ci,MSE,Tuk)
```

D. Power

```
the.power = give.me.power(group.means,group.nis,MSE,0.05)
the.power
```

E. Sample size

```
overall.mean = sum(group.means*group.nis)/sum(group.nis)
effect.size = sqrt( sum( group.nis/sum(group.nis) *(group.means -overall.mean)^2 )/MSE)
library(pwr)
pwr.anova.test(k = 3, f = effect.size, sig.level = 0.05, power = 0.95)
```

Functions used in R

CI

```
give.me.CI = function(ybar,ni,ci,MSE,multiplier){
```

```
  if(sum(ci) != 0 & sum(ci !=0 ) != 1){
    return("Error - you did not input a valid contrast")
  } else if(length(ci) != length(ni)){
    return("Error - not enough contrasts given")
  }
  else{
    estimate = sum(ybar*ci)
    SE = sqrt(MSE*sum(ci^2/ni))
    CI = estimate + c(-1,1)*multiplier*SE
    result = c(estimate,CI)
    names(result) = c("Estimate","Lower Bound","Upper
Bound")
    return(result)
  }
}
```

Power

```
give.me.power = function(ybar,ni,MSE,alpha){
  a = length(ybar) # Finds a
  nt = sum(ni) #Finds the overall sample size
  overall.mean = sum(ni*ybar)/nt # Finds the overall mean
  phi = (1/sqrt(MSE))*sqrt( sum(ni*(ybar - overall.mean)^2)/a)
#Finds the books value of phi
  phi.star = a *phi^2 #Finds the value of phi we will use for R
  Fc = qf(1-alpha,a-1,nt-a) #The critical value of F, use in R's
function
  power = 1 - pf(Fc, a-1, nt-a, phi.star)# The power,
calculated using a non-central F
  return(power)
}
```