OCN 750: Homework 1 -- Maia Kapur

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see <http://rmarkdown.rstudio.com>.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

#Load Required Packages  
library(stats)  
library(dplyr)

## Warning: package 'dplyr' was built under R version 3.1.2

##   
## Attaching package: 'dplyr'  
##   
## The following object is masked from 'package:stats':  
##   
## filter  
##   
## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

library(tidyr)

## Warning: package 'tidyr' was built under R version 3.1.2

library(car)

## Warning: package 'car' was built under R version 3.1.3

library(reshape2)

## Warning: package 'reshape2' was built under R version 3.1.2

library(data.table)

## Warning: package 'data.table' was built under R version 3.1.3

##   
## Attaching package: 'data.table'  
##   
## The following objects are masked from 'package:dplyr':  
##   
## between, last

library(effects)

## Warning: package 'effects' was built under R version 3.1.2

##   
## Attaching package: 'effects'  
##   
## The following object is masked from 'package:car':  
##   
## Prestige

library(gdata)

## gdata: read.xls support for 'XLS' (Excel 97-2004) files ENABLED.  
##   
## gdata: read.xls support for 'XLSX' (Excel 2007+) files ENABLED.  
##   
## Attaching package: 'gdata'  
##   
## The following object is masked from 'package:dplyr':  
##   
## combine  
##   
## The following object is masked from 'package:stats':  
##   
## nobs  
##   
## The following object is masked from 'package:utils':  
##   
## object.size

library(ggplot2)

## Warning: package 'ggplot2' was built under R version 3.1.3

library(lmtest)

## Loading required package: zoo  
##   
## Attaching package: 'zoo'  
##   
## The following objects are masked from 'package:base':  
##   
## as.Date, as.Date.numeric

#Set Working Directory  
setwd("~/Dropbox/2015 Fall/OCN 750")

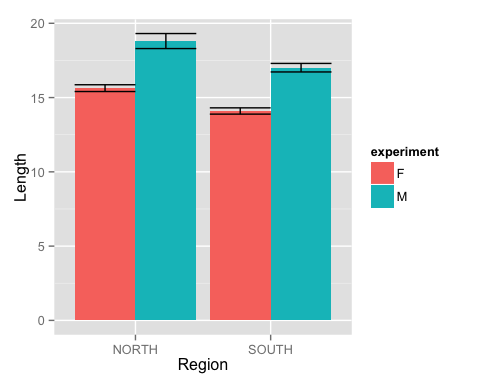
#### Question 1: Boxplot of Mean Length by Sex and Region

#Load appropriate CSV  
labcult = read.csv("lab\_cultures.csv")

#find means for sex and region  
means = aggregate(formula = Length..mm. ~ Sex + Region, data = labcult, FUN = mean)  
  
#calculate standard error for sex and region  
se = aggregate(formula = Length..mm. ~ Sex + Region, data = labcult, FUN = function(x) sqrt(var(x)/length(x)) )  
  
#create a list that contains the values from the above functions  
mean = c(means$Length..mm.)  
ses = c(se$Length..mm.)  
  
#designate the maximum and minimum values each test into an aes command - this makes  
#the ggplot code easier to write  
  
limits <- aes(ymax = mean + ses, ymin = mean - ses)  
  
#create a data frame that includes all the data you need to plot.  
#the melt function can quickly lay out your cross -interactions between the   
#Region and sex variables  
df = melt(data.frame(NORTH = c(15.63, 18.8), SOUTH = c(14.1, 17.0),  
 experiment = c("F", "M"),  
 variable\_name = "metric"))

## Using experiment, variable\_name as id variables

#plot it, telling ggplot to call from the dataframe (df) you just created  
#with Region ("variable") on the x - axis, Length ("value") on the Y, and filled by   
#Sex ("experiment")  
ggplot(df, aes(x = variable, value, fill = experiment)) +  
 geom\_bar(position = "dodge", stat = "identity") +  
 geom\_errorbar(limits, position = "dodge") +  
 xlab("Region") +  
 ylab("Length")



#### 2. Use lm(), and F-tests on the model, to test whether there is a significant difference between Sexes in mean length, whether there is a significant difference between Regions in mean length, and whether the effect of Region differs between Sexes. Report F-statistics and degrees of freedom. How do you interpret these results?

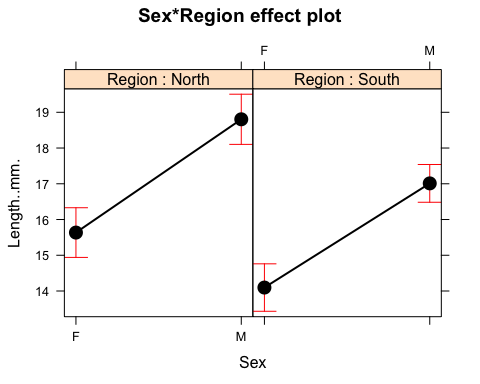
#Generate a linear model for variables of interest, testing for an interaction between the Sex and Region Factors  
lsr.noint = lm(labcult$Length..mm. ~ labcult$Region)  
lsr.interaction = lm(Length..mm. ~ Sex\*Region, data = labcult)  
culteff = effect("Sex\*Region", lsr.interaction)  
  
#Run ANOVA test using car package to examine significant differences  
Anova(lsr.interaction)

## Anova Table (Type II tests)  
##   
## Response: Length..mm.  
## Sum Sq Df F value Pr(>F)   
## Sex 554.12 1 85.5806 < 2.2e-16 \*\*\*  
## Region 168.00 1 25.9472 6.988e-07 \*\*\*  
## Sex:Region 0.96 1 0.1484 0.7004   
## Residuals 1592.81 246   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

#See F-statistics and Degrees of Freedom in above output.  
  
#It appears there is a significant difference in between sexes and regions in mean length, but a non-significant difference of Region's effect between sexes. Upon visual inspection, this appears to make sense; both Males and Females exhibit higher mean lengths in the North than the South.

#### 3 Plot the model-fitted group means and standard errors.

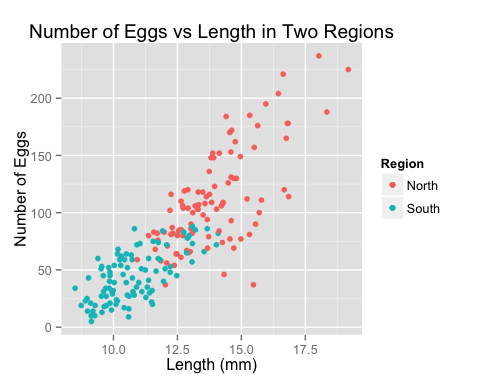
plot(culteff)



#### 4: Using the “Egg Data” worksheet, plot Number of Eggs vs. Length, and color code the plotted points by Region.

eggs = read.csv("eggs.csv")  
  
#Make new column and designate south/north regions for all populations  
eggs$Region = NA  
eggs$Region[1:50] = "NORTH"  
eggs$Region[51:200] = "NORTH"  
eggs$Region[51:150] = "SOUTH"  
  
#Subset individual populations  
nahant = subset(eggs, eggs$Population == "Nahant")  
Magnolia = subset(eggs, eggs$Population == "Magnolia")  
north = rbind(nahant, Magnolia)  
VIMS = subset(eggs, eggs$Population == "VIMS")  
CCVA = subset(eggs, eggs$Population == "CCVA")  
south = rbind(VIMS, CCVA)

#Plot Regionally-Colored data  
ggplot(data=eggs, aes(x=Length..mm., color = Region)) +  
 geom\_point(data = nahant, aes(y = Number.of.Eggs, color = "North")) +  
 geom\_point(data = CCVA, aes(y = Number.of.Eggs, color = "South")) +  
 geom\_point(data = VIMS, aes(y = Number.of.Eggs, color = "South")) +  
 geom\_point(data = Magnolia, aes(y = Number.of.Eggs, color = "North")) +  
 ylab("Number of Eggs") +  
 xlab("Length (mm)") +  
 ggtitle("Number of Eggs vs Length in Two Regions")



#### 5. Fit a linear model to test whether the relationship between Number of Eggs and Length differs between Regions. Perform and report the appropriate F-test to test this question. How might you interpret these results?

#Create a linear model that incorporates both the number of eggs and length for each region  
lmr = lm(Number.of.Eggs ~ Length..mm.\*Region, data = eggs) #All Regions  
lmn = lm(north$Number.of.Eggs ~ north$Length..mm.) #Just the North  
lms = lm(south$Number.of.Eggs ~ south$Length..mm.) #Just the South  
Anova(lmr)

## Anova Table (Type II tests)  
##   
## Response: Number.of.Eggs  
## Sum Sq Df F value Pr(>F)   
## Length..mm. 107208 1 179.5393 < 2.2e-16 \*\*\*  
## Region 6371 1 10.6689 0.001286 \*\*   
## Length..mm.:Region 5696 1 9.5382 0.002304 \*\*   
## Residuals 117037 196   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

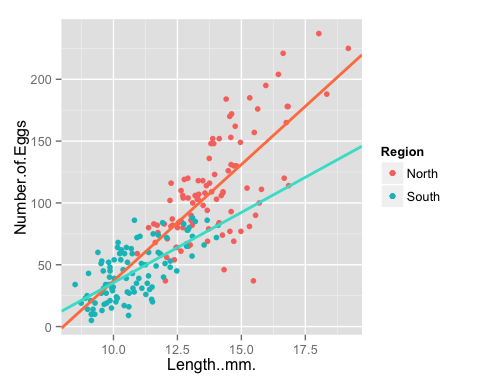
summary(lmr)

##   
## Call:  
## lm(formula = Number.of.Eggs ~ Length..mm. \* Region, data = eggs)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -102.922 -13.479 0.379 15.391 64.127   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -151.316 21.354 -7.086 2.43e-11 \*\*\*  
## Length..mm. 18.825 1.526 12.333 < 2e-16 \*\*\*  
## RegionSOUTH 73.091 29.456 2.481 0.0139 \*   
## Length..mm.:RegionSOUTH -7.454 2.414 -3.088 0.0023 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 24.44 on 196 degrees of freedom  
## Multiple R-squared: 0.7388, Adjusted R-squared: 0.7348   
## F-statistic: 184.8 on 3 and 196 DF, p-value: < 2.2e-16

#There is a significant difference for No. Eggs across lengths and among the two regions.

#### 6. For #5 you have essentially fit two linear regressions, one for each Region. Take your plot from #4 and add these two lines. abline() will help.

ggplot(data=eggs, aes(x=Length..mm., color = Region, group = 1)) +  
 geom\_point(data = nahant, aes(y = Number.of.Eggs, color = "North")) +  
 geom\_point(data = CCVA, aes(y = Number.of.Eggs, color = "South")) +  
 geom\_point(data = VIMS, aes(y = Number.of.Eggs, color = "South")) +  
 geom\_point(data = Magnolia, aes(y = Number.of.Eggs, color = "North")) +  
 geom\_abline(intercept = coef(lmn)[1], slope = coef(lmn)[2], col = "coral", size = 1) +  
 geom\_abline(intercept = coef(lms)[1], slope = coef(lms)[2], col = "turquoise", size = 1)

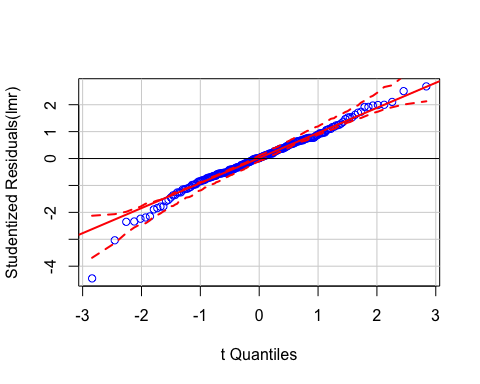


ggtitle("Number of Eggs vs Length in Two Regions")

## $title  
## [1] "Number of Eggs vs Length in Two Regions"  
##   
## attr(,"class")  
## [1] "labels"

#### 7. Model diagnostics: for the model you fit in #5, make some plots that explore whether the residuals of the model are normally distributed, whether the variance of the residuals increases as Length increases, and whether the variance of the residuals varies between regions.

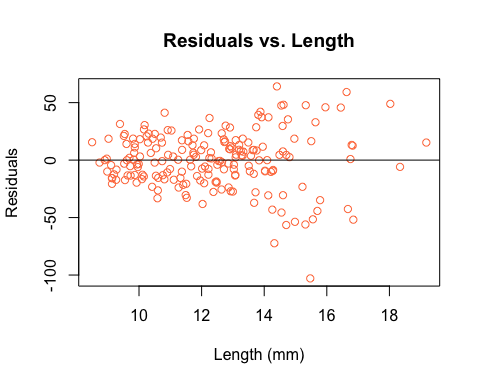
#Plot to inspect distribution of residuals  
qqPlot(lmr, col = "blue")  
abline(0,0)



#The residuals themselves appear normally distributed, with deviation from normality at the tails of the distribution.

# Variance of Residuals vs Length

#Calculate model residuals & add to data frame  
lmr.residual = resid(lmr)  
eggs = cbind(eggs, lmr.residual)  
  
#Plot the Length variable against the residuals  
plot(eggs$Length..mm., eggs$lmr.residual, col = "coral", xlab = "Length (mm)", ylab = "Residuals", main = "Residuals vs. Length")  
abline(0,0)



#The above plot suggests that variance in residuals increases with length.

# Variance of Residuals between Regions

boxplot(eggs$lmr.residual ~ eggs$Region,  
 main = "Variance of Residuals between Regions",  
 ylab = "residuals",  
 xlab = "region")

