Workshop 3.2 Solutions

Veerasak Kritsanapraphan

We’ll begin by loading some packages.

library(MASS)  
library(plyr)  
library(ggplot2)  
library(reshape)

Let’s form our favourite birthwt data set.

# Rename the columns to have more descriptive names  
colnames(birthwt) <- c("birthwt.below.2500", "mother.age", "mother.weight",   
 "race", "mother.smokes", "previous.prem.labor", "hypertension", "uterine.irr",   
 "physician.visits", "birthwt.grams")  
  
# Transform variables to factors with descriptive levels  
birthwt <- transform(birthwt,   
 race = as.factor(mapvalues(race, c(1, 2, 3),   
 c("white","black", "other"))),  
 mother.smokes = as.factor(mapvalues(mother.smokes,   
 c(0,1), c("no", "yes"))),  
 hypertension = as.factor(mapvalues(hypertension,   
 c(0,1), c("no", "yes"))),  
 uterine.irr = as.factor(mapvalues(uterine.irr,   
 c(0,1), c("no", "yes")))  
 )

### ANOVA with birthwt data

**(a)** Create a new factor that categorizes the number of physician visits into four levels: 0, 1, 2, 3 or more.

phys.visit.binned <- birthwt$physician.visits  
class(phys.visit.binned)

## [1] "integer"

phys.visit.binned[phys.visit.binned >= 3] <- "3.or.more"  
class(phys.visit.binned)

## [1] "character"

birthwt <- transform(birthwt, phys.visit.binned = as.factor(phys.visit.binned))  
  
birthwt$phys.visit.binned

## [1] 0 3.or.more 1 2 0 0 1   
## [8] 1 1 0 0 1 0 2   
## [15] 0 0 0 3.or.more 0 1 2   
## [22] 3.or.more 1 0 2 0 0 2   
## [29] 0 1 1 1 1 1 0   
## [36] 2 2 0 2 1 2 2   
## [43] 1 0 0 0 3.or.more 0 2   
## [50] 0 1 0 0 2 0 0   
## [57] 0 0 0 0 0 2 0   
## [64] 0 0 1 2 3.or.more 1 2   
## [71] 0 2 1 0 0 0 1   
## [78] 3.or.more 0 0 1 0 0 0   
## [85] 0 0 0 0 0 1 0   
## [92] 2 0 0 0 1 1 0   
## [99] 0 1 1 0 0 1 0   
## [106] 0 1 0 2 3.or.more 2 1   
## [113] 2 1 0 1 0 0 2   
## [120] 1 1 0 1 0 2 2   
## [127] 1 0 1 1 0 2 0   
## [134] 0 0 0 1 1 0 1   
## [141] 0 0 0 1 0 2 2   
## [148] 0 0 0 1 2 0 0   
## [155] 0 0 3.or.more 1 0 0 0   
## [162] 1 0 0 0 0 3.or.more 0   
## [169] 1 0 1 0 0 0 0   
## [176] 0 1 3.or.more 0 2 1 3.or.more  
## [183] 0 0 2 2 0 0 3.or.more  
## Levels: 0 1 2 3.or.more

**Hint**: One way of doing this is with mapvalues, by mapping all instances of 3, 4,… etc, to “3 or more”.

**(b)** Run an ANOVA to determine whether the average birth weight varies across number of physician visits.

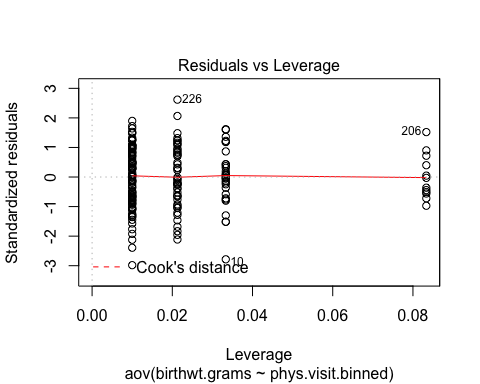
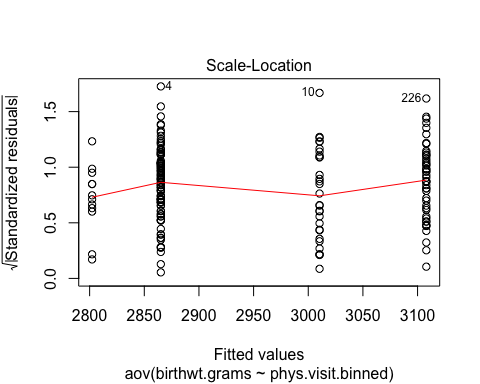
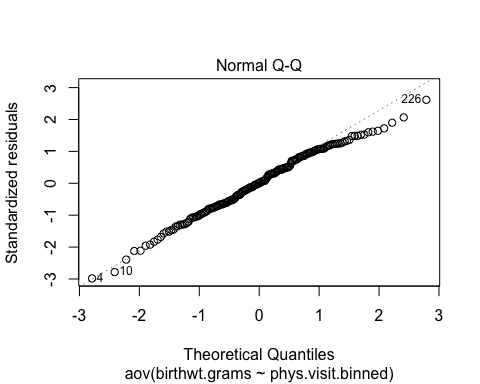
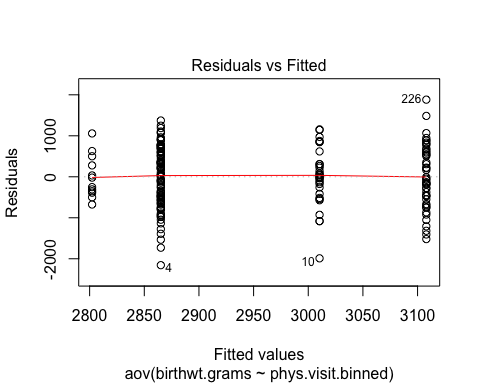
aov.birthwt <- aov(birthwt.grams ~ phys.visit.binned , data = birthwt)  
summary(aov.birthwt)

## Df Sum Sq Mean Sq F value Pr(>F)  
## phys.visit.binned 3 2259057 753019 1.426 0.237  
## Residuals 185 97710599 528165

The p-value is greater than 0.05, so the variation in birthweight across number of physician visits is not statistically significant.

## Check Normality

plot(aov.birthwt)



## Shapiro-Wilk Test

# Extract the residuals  
aov\_residuals <- residuals(object = aov.birthwt )  
# Run Shapiro-Wilk test  
shapiro.test(x = aov\_residuals )

##   
## Shapiro-Wilk normality test  
##   
## data: aov\_residuals  
## W = 0.99084, p-value = 0.2731

### Tukey multiple pairwise-comparisons

posthoc <- TukeyHSD(x=aov.birthwt, 'phys.visit.binned', conf.level=0.95)  
posthoc

## Tukey multiple comparisons of means  
## 95% family-wise confidence level  
##   
## Fit: aov(formula = birthwt.grams ~ phys.visit.binned, data = birthwt)  
##   
## $phys.visit.binned  
## diff lwr upr p adj  
## 1-0 242.86000 -90.35099 576.0710 0.2360659  
## 2-0 145.19333 -247.01844 537.4051 0.7724282  
## 3.or.more-0 -62.89000 -638.49952 512.7195 0.9920501  
## 2-1 -97.66667 -537.96332 342.6300 0.9394260  
## 3.or.more-1 -305.75000 -915.14100 303.6410 0.5636139  
## 3.or.more-2 -208.08333 -851.63434 435.4677 0.8361567

plot(posthoc)

