Homework #4\_FA2024\_BMEN7340

First\_Last\_7340\_HW4\_FA24

2024-11-06

# clean R environment

rm(list = ls(all=TRUE))  
graphics.off()  
shell("cls")

library(readxl)  
library(psych) # for pairs.panels.  
library(ggrepel) library(outliers) library(ggplot2) library(dplyr) library(epiR) library(MASS) library(nnet) library(glmnet) library(tidyverse) library(FSA) library(pwr) library(effsize) library(ggcorrplot) library(DAAG) library(caret) library(pROC) library(car) library(faraway) library(PerformanceAnalytics) library(pscl) library(ROCR) library(emmeans) library(ggpubr) library(SixSigma) library(rsm) library(AlgDesign) library(lattice) library(ez) library(nlme) library(lsmeans) library(Sleuth3) library(mosaic) library(mosaicData) library(multcompView) library(sjstats) library(lme4) library(lmerTest) library(MuMIn) library(carData) library(biotools) library(broom) library(gplots) library(ICSNP) library(heplots) library(rstatix) library(tidyr) library(magrittr) library(mvnormtest) library(mvtnorm) library(purrr) library(rio) library(MVN) library(rpart) library(rpart.plot) library(EnvStats) library(DescTools) library(RColorBrewer) library(survival) library(survminer)

#1.  
dnd <- matrix(c(1250,1387,991,1666),nrow = 2)  
colnames(dnd) <- c('Y10','Y14')  
rownames(dnd) <- c('Yes','No')  
  
# Part A:  
chisq.test(dnd)

##   
## Pearson's Chi-squared test with Yates' continuity correction  
##   
## data: dnd  
## X-squared = 54.942, df = 1, p-value = 1.241e-13

# H0: proportions of students who drove while drinking   
# are the same in the two calendar years   
# Ha: proportions of students who drove while drinking   
# are NOT the same in the two calendar years   
# p-value = 1.241e-13, therefore, reject the null hypothesis  
  
# Part B:  
# The behavior of college students before the legal drinking age  
# was raised is statistically different than the years afterwards  
rm(dnd)  
  
#2.  
smears <- matrix(c(1763,403,489,670),nrow = 2)  
colnames(smears) <- c('Present','Absent')  
rownames(smears) <- c('Present','Absent')  
  
smears

## Present Absent  
## Present 1763 489  
## Absent 403 670

sum(smears)

## [1] 3325

mcnemar.test(smears)

##   
## McNemar's Chi-squared test with continuity correction  
##   
## data: smears  
## McNemar's chi-squared = 8.0998, df = 1, p-value = 0.004427

smears2 <- matrix(c(2252,1073,2166,1159), nrow=2)  
smears2

## [,1] [,2]  
## [1,] 2252 2166  
## [2,] 1073 1159

sum(smears2) #3325 pairs

## [1] 6650

chisq.test(smears2)

##   
## Pearson's Chi-squared test with Yates' continuity correction  
##   
## data: smears2  
## X-squared = 4.8724, df = 1, p-value = 0.02729

#3  
alcohol <- read\_excel("alcohol.xlsx")  
View(alcohol)  
table(alcohol$genques)

##   
## 0 1   
## 86 151

table(alcohol$alcques)

##   
## 0 1   
## 121 116

test <- matrix(c(86, 151, 121, 116), nrow=2)  
colnames(test) <- c('genques','alcques')  
rownames(test) <- c('nondrinker','drinker')  
test

## genques alcques  
## nondrinker 86 121  
## drinker 151 116

chisq.test(test)

##   
## Pearson's Chi-squared test with Yates' continuity correction  
##   
## data: test  
## X-squared = 9.9141, df = 1, p-value = 0.00164

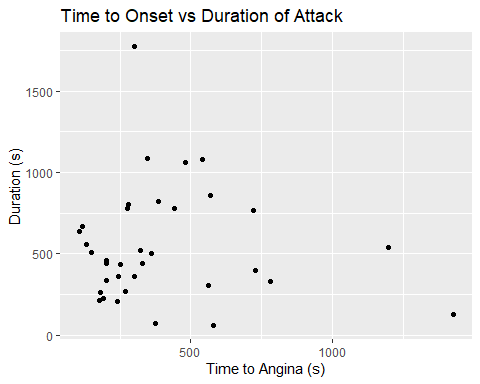
#or  
test1 <- matrix(c(86, 121, 151, 116), nrow=2)  
colnames(test1) <- c('nondrinker=0','drinker=1')  
rownames(test1) <- c('genques','alcques')  
test1

## nondrinker=0 drinker=1  
## genques 86 151  
## alcques 121 116

chisq.test(test1)

##   
## Pearson's Chi-squared test with Yates' continuity correction  
##   
## data: test1  
## X-squared = 9.9141, df = 1, p-value = 0.00164

#4  
ischemic <- read\_excel("ischemic.xlsx")  
# View(ischemic)  
# a:  
ggplot(ischemic) + geom\_point(aes(x=time,y=duration)) +   
 ggtitle('Time to Onset vs Duration of Attack') + xlab('Time to Angina (s)') +  
 ylab('Duration (s)')



# b:  
# Visually there does not appear to be a relationship; however, this can  
# be tested through a linear regression  
shapiro.test(ischemic$time)

##   
## Shapiro-Wilk normality test  
##   
## data: ischemic$time  
## W = 0.79258, p-value = 1.465e-05

# p-value = 1.465e-05, not normal  
shapiro.test(ischemic$duration)

##   
## Shapiro-Wilk normality test  
##   
## data: ischemic$duration  
## W = 0.90161, p-value = 0.00437

# p-value = 0.00437, not normal  
# therefore, the usual linear regression cannot be utilized  
  
# c:  
# I will use a spearman test since the data is not normally  
# distributed.  
  
# d:  
cor.test(ischemic$duration~ischemic$time,method='spearman')

## Warning in cor.test.default(x = mf[[1L]], y = mf[[2L]], ...): Cannot compute  
## exact p-value with ties

##   
## Spearman's rank correlation rho  
##   
## data: ischemic$duration and ischemic$time  
## S = 6598.9, p-value = 0.6653  
## alternative hypothesis: true rho is not equal to 0  
## sample estimates:  
## rho   
## 0.07578092

# or  
cor.test(ischemic$duration~ischemic$time,method='kendall')

## Warning in cor.test.default(x = mf[[1L]], y = mf[[2L]], ...): Cannot compute  
## exact p-value with ties

##   
## Kendall's rank correlation tau  
##   
## data: ischemic$duration and ischemic$time  
## z = 0.44033, p-value = 0.6597  
## alternative hypothesis: true tau is not equal to 0  
## sample estimates:  
## tau   
## 0.05218863

# H0: correlation between time to onset and the duration is equal to 0  
# Ha: correlation between time to onset and the duration is NOT equal to 0  
# p-value = 0.6653, we FTR the null hypothesis that there is not a correlation  
# between the time to onset and the duration of the attack  
  
# 5  
lowbwt <- read\_excel("lowbwt.xlsx")  
# View(lowbwt)  
# a:  
cor.test(lowbwt$sbp~lowbwt$apgar5,method='spearman')

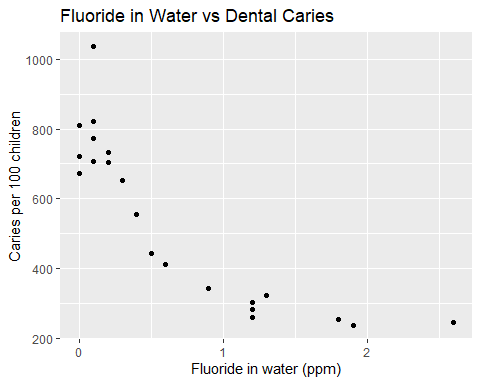
## Warning in cor.test.default(x = mf[[1L]], y = mf[[2L]], ...): Cannot compute  
## exact p-value with ties

##   
## Spearman's rank correlation rho  
##   
## data: lowbwt$sbp and lowbwt$apgar5  
## S = 148593, p-value = 0.2832  
## alternative hypothesis: true rho is not equal to 0  
## sample estimates:  
## rho   
## 0.1083551

# or  
cor.test(lowbwt$sbp~lowbwt$apgar5,method='kendall')

##   
## Kendall's rank correlation tau  
##   
## data: lowbwt$sbp and lowbwt$apgar5  
## z = 1.0698, p-value = 0.2847  
## alternative hypothesis: true tau is not equal to 0  
## sample estimates:  
## tau   
## 0.07841282

# H0: There is no correlation between the systolic blood pressure and APGAR score  
# Ha: There is a correlation between the systolic blood pressure and APGAR score  
# p-value = 0.2832, therefore, FTR the null hypothesis  
  
# b:  
# H0: There is no correlation between the systolic blood pressure and APGAR score  
# Due to the p-value being higher than 0.05, I am unable to reject the null hypothesis  
# and cannot statistically conclude a correlation  
  
#6.  
water <- read\_excel("waterFA23.xlsx")  
# View(water)  
# a:  
ggplot(water)+geom\_point(aes(x=fluoride,y=caries)) +   
 ggtitle('Fluoride in Water vs Dental Caries') +  
 ylab('Caries per 100 children') + xlab('Fluoride in water (ppm)')



# Part B:   
shapiro.test(water$fluoride)

##   
## Shapiro-Wilk normality test  
##   
## data: water$fluoride  
## W = 0.85133, p-value = 0.004478

# p-value = 0.004478, confirmed not normal  
shapiro.test(water$caries)

##   
## Shapiro-Wilk normality test  
##   
## data: water$caries  
## W = 0.90332, p-value = 0.0406

# p-value = 0.0406, confirmed not normal  
# non-parametric test  
cor.test(water$fluoride,water$caries,method='spearman')

## Warning in cor.test.default(x, y, ...): Cannot compute exact p-value with ties

##   
## Spearman's rank correlation rho  
##   
## data: x and y  
## S = 2947.7, p-value = 6.975e-09  
## alternative hypothesis: true rho is not equal to 0  
## sample estimates:  
## rho   
## -0.914098

log(water$caries)

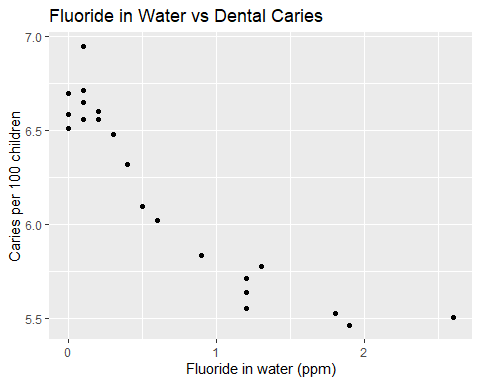
## [1] 6.697034 6.511745 6.582025 6.559615 6.712956 6.944087 6.648985 6.597146  
## [9] 6.555357 6.480045 6.320768 6.095825 6.021023 5.837730 5.552960 5.713733  
## [17] 5.638355 5.777652 5.529429 5.463832 5.505332

water$caries <- log(water$caries)  
cor.test(water$fluoride,water$caries,method='spearman')

## Warning in cor.test.default(x, y, ...): Cannot compute exact p-value with ties

##   
## Spearman's rank correlation rho  
##   
## data: x and y  
## S = 2947.7, p-value = 6.975e-09  
## alternative hypothesis: true rho is not equal to 0  
## sample estimates:  
## rho   
## -0.914098

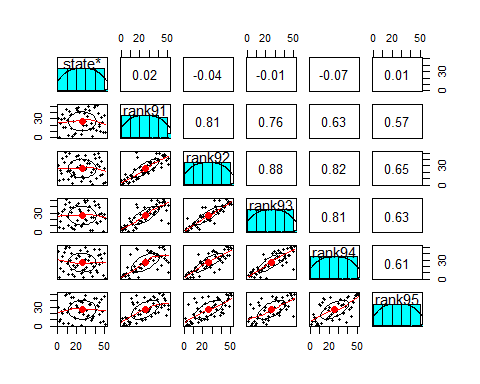
ggplot(water)+  
 geom\_point(aes(x=fluoride,y=caries)) +   
 ggtitle('Fluoride in Water vs Dental Caries') +  
 ylab('Caries per 100 children') + xlab('Fluoride in water (ppm)')



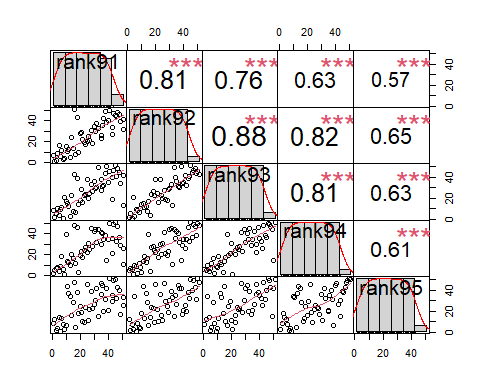
# H0: There is not a correlation between the number of dental caries per 100 children  
# and the fluoride concentration in the water in ppm  
# Ha: There is a correlation between the number of dental caries per 100 children  
# and the fluoride concentration in the water in ppm  
# p-value = 7.024e-07, therefore, reject the null hypothesis  
# tau = -0.8013917 (negative correlation)  
  
# Part C:  
# This correlation is significantly different from 0 because the p-value  
# of 7.024e-07 indicates we can reject the null hypothesis that the correlation is 0  
  
#7.  
actions <- read\_excel("actions.xls.xlsx")  
# View(actions)  
str(actions)

## tibble [51 × 6] (S3: tbl\_df/tbl/data.frame)  
## $ state : chr [1:51] "Alabama" "Alaska" "Arizona" "Arkansas" ...  
## $ rank91: num [1:51] 31 1 22 29 37 17 30 16 45 27 ...  
## $ rank92: num [1:51] 30 7 22 18 42 8 35 51 45 21 ...  
## $ rank93: num [1:51] 29 8 16 26 32 6 36 43 51 25 ...  
## $ rank94: num [1:51] 44 2 17 28 35 12 42 48 51 25 ...  
## $ rank95: num [1:51] 30 8 10 23 20 5 27 48 50 22 ...

# library(psych) # for pairs.panels()  
pairs.panels(actions)



# library(PerformanceAnalytics)  
chart.Correlation(actions[,2:6])



# Part A:  
# 1991 and 1992  
shapiro.test(actions$rank91)

##   
## Shapiro-Wilk normality test  
##   
## data: actions$rank91  
## W = 0.95531, p-value = 0.05269

# p-value = 0.04978, non-normal (since this is so close we can try both tests)  
shapiro.test(actions$rank92)

##   
## Shapiro-Wilk normality test  
##   
## data: actions$rank92  
## W = 0.95554, p-value = 0.05388

# p-value = 0.05388, normal  
  
# Parametric Test:  
summary(lm(actions$rank92~actions$rank91))

##   
## Call:  
## lm(formula = actions$rank92 ~ actions$rank91)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -14.058 -5.638 0.026 4.463 33.089   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 5.04322 2.51177 2.008 0.0502 .   
## actions$rank91 0.80421 0.08383 9.593 7.85e-13 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 8.852 on 49 degrees of freedom  
## Multiple R-squared: 0.6525, Adjusted R-squared: 0.6454   
## F-statistic: 92.02 on 1 and 49 DF, p-value: 7.854e-13

# H0: There is not a correlation between the severe actions  
# per 1000 doctors between 1991 and 1992  
# Ha: There is a correlation between the severe actions  
# per 1000 doctors between 1991 and 1992  
# p-value: 5.921e-13, reject the null hypothesis  
# Estimated slope: 0.81029   
  
# Non-Parametric Test:  
cor.test(actions$rank91,actions$rank92, method='spearman')

## Warning in cor.test.default(x, y, ...): Cannot compute exact p-value with ties

##   
## Spearman's rank correlation rho  
##   
## data: x and y  
## S = 4193.8, p-value = 5.921e-13  
## alternative hypothesis: true rho is not equal to 0  
## sample estimates:  
## rho   
## 0.810236

# p-value = 5.921e-13, reject the null hypothesis  
# rho = 0.810236  
  
# 1991 and 1993  
shapiro.test(actions$rank93)

##   
## Shapiro-Wilk normality test  
##   
## data: actions$rank93  
## W = 0.95513, p-value = 0.05179

# p-value = 0.0536, normal  
  
# Parametric Test:  
summary(lm(actions$rank93~actions$rank91))

##   
## Call:  
## lm(formula = actions$rank93 ~ actions$rank91)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -31.947 -4.679 -0.591 5.941 24.586   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 6.28412 2.76676 2.271 0.0276 \*   
## actions$rank91 0.75810 0.09235 8.209 9.16e-11 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 9.751 on 49 degrees of freedom  
## Multiple R-squared: 0.579, Adjusted R-squared: 0.5704   
## F-statistic: 67.39 on 1 and 49 DF, p-value: 9.158e-11

# H0: There is not a correlation between the severe actions  
# per 1000 doctors between 1991 and 1993  
# Ha: There is a correlation between the severe actions  
# per 1000 doctors between 1991 and 1993  
# p-value: 6.768e-11, reject the null hypothesis  
# Estimated slope: 0.76429  
  
# Non-Parametric Test:  
cor.test(actions$rank91,actions$rank93, method='spearman')

## Warning in cor.test.default(x, y, ...): Cannot compute exact p-value with ties

##   
## Spearman's rank correlation rho  
##   
## data: x and y  
## S = 5209.6, p-value = 6.768e-11  
## alternative hypothesis: true rho is not equal to 0  
## sample estimates:  
## rho   
## 0.764272

# p-value = 5.921e-13, reject the null hypothesis  
# rho = 0.764272   
  
# 1991 and 1994  
shapiro.test(actions$rank94)

##   
## Shapiro-Wilk normality test  
##   
## data: actions$rank94  
## W = 0.95462, p-value = 0.04929

# p-value = 0.05342, normal  
  
# Parametric Test:  
summary(lm(actions$rank94~actions$rank91))

##   
## Call:  
## lm(formula = actions$rank94 ~ actions$rank91)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -30.643 -6.244 -1.628 6.370 28.254   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 9.7364 3.3275 2.926 0.00519 \*\*   
## actions$rank91 0.6256 0.1111 5.633 8.56e-07 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 11.73 on 49 degrees of freedom  
## Multiple R-squared: 0.393, Adjusted R-squared: 0.3807   
## F-statistic: 31.73 on 1 and 49 DF, p-value: 8.558e-07

# H0: There is not a correlation between the severe actions  
# per 1000 doctors between 1991 and 1994  
# Ha: There is a correlation between the severe actions  
# per 1000 doctors between 1991 and 1994  
# p-value: 7.598e-07, reject the null hypothesis  
# Estimated slope: 0.6292  
  
# Non-Parametric Test:  
cor.test(actions$rank91,actions$rank94, method='spearman')

## Warning in cor.test.default(x, y, ...): Cannot compute exact p-value with ties

##   
## Spearman's rank correlation rho  
##   
## data: x and y  
## S = 8194.4, p-value = 7.598e-07  
## alternative hypothesis: true rho is not equal to 0  
## sample estimates:  
## rho   
## 0.6292114

# p-value = 7.598e-07, reject the null hypothesis  
# rho = 0.6292114   
  
# 1991 and 1995  
shapiro.test(actions$rank95)

##   
## Shapiro-Wilk normality test  
##   
## data: actions$rank95  
## W = 0.955, p-value = 0.05112

# p-value = 0.05342, normal  
  
# Parametric Test:  
summary(lm(actions$rank95~actions$rank91))

##   
## Call:  
## lm(formula = actions$rank95 ~ actions$rank91)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -22.131 -10.127 -1.844 10.227 27.729   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 11.1261 3.4963 3.182 0.00254 \*\*   
## actions$rank91 0.5715 0.1167 4.898 1.1e-05 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 12.32 on 49 degrees of freedom  
## Multiple R-squared: 0.3287, Adjusted R-squared: 0.315   
## F-statistic: 23.99 on 1 and 49 DF, p-value: 1.098e-05

# H0: There is not a correlation between the severe actions  
# per 1000 doctors between 1991 and 1995  
# Ha: There is a correlation between the severe actions  
# per 1000 doctors between 1991 and 1995  
# p-value: 1.084e-05, reject the null hypothesis  
# Estimated slope: 0.5736  
  
# Non-Parametric Test:  
cor.test(actions$rank91,actions$rank95, method='spearman')

## Warning in cor.test.default(x, y, ...): Cannot compute exact p-value with ties

##   
## Spearman's rank correlation rho  
##   
## data: x and y  
## S = 9423.9, p-value = 1.084e-05  
## alternative hypothesis: true rho is not equal to 0  
## sample estimates:  
## rho   
## 0.5735813

# p-value = 1.084e-05, reject the null hypothesis  
# rho = 0.5735813   
  
# As the years get further apart, the correlation coefficients   
# get smaller.  
  
# Part B:  
# All of the correlation calculations ran above proved significantly   
# different than the null hypothesis of 0.  
  
# Part C:  
#y=ranks  
#x=state   
# one way anova, or nonparametric  
kruskal.test(rank91+rank92+rank93+rank94+rank95~state,data=actions)

##   
## Kruskal-Wallis rank sum test  
##   
## data: rank91 + rank92 + rank93 + rank94 + rank95 by state  
## Kruskal-Wallis chi-squared = 50, df = 50, p-value = 0.4734

# H0: There is no difference between how the states discipline the doctors  
# Ha: There is a difference between how the states discipline the doctors  
# p-value = 0.4734, FTR the null hypothesis  
  
#8  
# Part A:  
shapiro.test(lowbwt$sbp)

##   
## Shapiro-Wilk normality test  
##   
## data: lowbwt$sbp  
## W = 0.97911, p-value = 0.1134

# p-value = 0.1134  
shapiro.test(lowbwt$gestage)

##   
## Shapiro-Wilk normality test  
##   
## data: lowbwt$gestage  
## W = 0.97995, p-value = 0.1316

# p-value = 0.1316  
lr <- lm(lowbwt$sbp~lowbwt$gestage)  
summary(lr)

##   
## Call:  
## lm(formula = lowbwt$sbp ~ lowbwt$gestage)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -23.162 -7.828 -1.483 5.568 39.781   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 10.5521 12.6506 0.834 0.40625   
## lowbwt$gestage 1.2644 0.4362 2.898 0.00463 \*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 11 on 98 degrees of freedom  
## Multiple R-squared: 0.07895, Adjusted R-squared: 0.06956   
## F-statistic: 8.401 on 1 and 98 DF, p-value: 0.004628

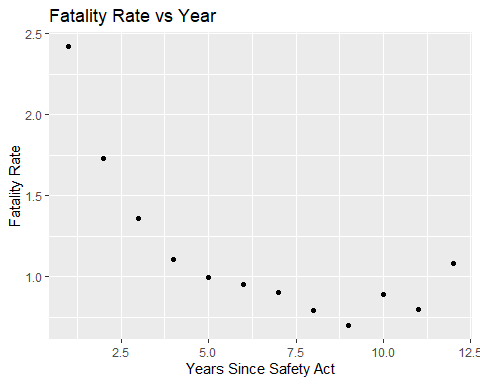
# H0: There is no relationship between sbp and gestation age  
# Ha: There is a relationship between sbp and gestation age  
# p-value: 0.004628, therefore, we reject the null hypothesis  
# Estimated slope = 1.264, meaning that the impact of the gestational  
# age on the sbp is approximately 1.264  
  
# Part B:  
# p < 0.05, Reject Ho, the b1 is not =0.   
  
# Part C:  
y <- 10.55 + 1.2644\*31  
#or  
new <- with(lowbwt, data.frame(gestage = 31))  
mean(predict(lr,new,type='response'))

## Warning: 'newdata' had 1 row but variables found have 100 rows

## Warning: 'newdata' had 1 row but variables found have 100 rows

## [1] 47.08

# 47.08 mmHg  
  
#9  
miner <- read\_excel("miner.xlsx")  
# View(miner)  
# Part A:  
ggplot(miner)+geom\_point(aes(x=Year,y=`Fatality Rate`)) +  
 ggtitle('Fatality Rate vs Year')+ylab('Fatality Rate')+  
 xlab('Years Since Safety Act')



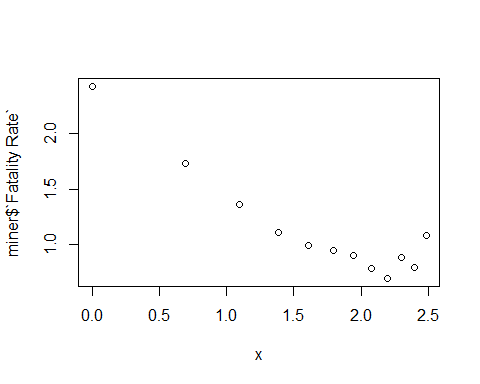
# This plot suggests a non-linear relationship.  
  
# Part B:  
lr1 <- lm(miner$`Fatality Rate`~miner$Year)  
summary(lr1)

##   
## Call:  
## lm(formula = miner$`Fatality Rate` ~ miner$Year)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.3013 -0.2112 -0.1647 0.1162 0.7150   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.80561 0.21012 8.593 6.26e-06 \*\*\*  
## miner$Year -0.10166 0.02855 -3.561 0.00518 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.3414 on 10 degrees of freedom  
## Multiple R-squared: 0.559, Adjusted R-squared: 0.515   
## F-statistic: 12.68 on 1 and 10 DF, p-value: 0.005176

# p-value: 0.005176  
# estimated slope: -0.10166  
# Adjusted R-squared: 0.515  
  
# Part C:  
x <- log(miner$Year)  
lr2 <- lm(miner$`Fatality Rate`~x)  
summary(lr2)

##   
## Call:  
## lm(formula = miner$`Fatality Rate` ~ x)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.20290 -0.12265 -0.09047 0.09815 0.42630   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2.13515 0.14879 14.350 5.35e-08 \*\*\*  
## x -0.59457 0.08193 -7.257 2.74e-05 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.2054 on 10 degrees of freedom  
## Multiple R-squared: 0.8404, Adjusted R-squared: 0.8244   
## F-statistic: 52.66 on 1 and 10 DF, p-value: 2.737e-05

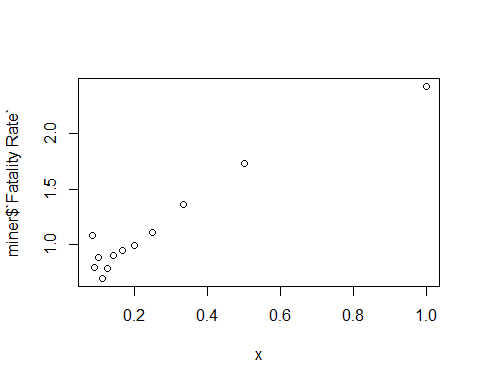
plot(miner$`Fatality Rate`~x)



# p-value: 2.737e-05  
# estimated slope: -0.59457   
# Adjusted R-squared: 0.8244   
  
# Part D:  
x <- 1/miner$Year  
lr3 <- lm(miner$`Fatality Rate`~x)  
summary(lr3)

##   
## Call:  
## lm(formula = miner$`Fatality Rate` ~ x)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.17711 -0.04879 -0.02905 0.04424 0.25612   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.67718 0.05063 13.38 1.05e-07 \*\*\*  
## x 1.80840 0.14019 12.90 1.48e-07 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.1224 on 10 degrees of freedom  
## Multiple R-squared: 0.9433, Adjusted R-squared: 0.9376   
## F-statistic: 166.4 on 1 and 10 DF, p-value: 1.476e-07

plot(miner$`Fatality Rate`~x)



# p-value: 1.476e-07  
# estimated slope: 1.80840  
# Adjusted R-squared: 0.9376   
  
# Part E:  
# Since the third model has the highest adjusted r squared value  
# it fits the linear regression model the best.  
AIC(lr1, lr2, lr3)

## df AIC  
## lr1 3 12.074630  
## lr2 3 -0.121211  
## lr3 3 -12.541961

# Part F:  
predict(lr3,Year=9) # 0.8781127

## 1 2 3 4 5 6 7 8   
## 2.4855799 1.5813796 1.2799795 1.1292795 1.0388595 0.9785794 0.9355223 0.9032294   
## 9 10 11 12   
## 0.8781127 0.8580194 0.8415794 0.8278794

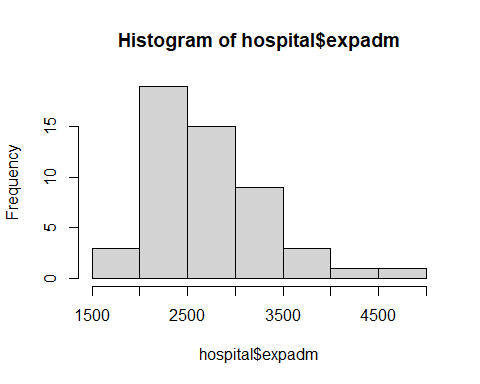
rm(x,lr1,lr2,lr3)  
  
#10  
hospital <- read\_excel("hospital.xlsx")  
# View(hospital)  
  
# Part A:  
summary(hospital$expadm)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 1772 2260 2600 2717 3054 4612

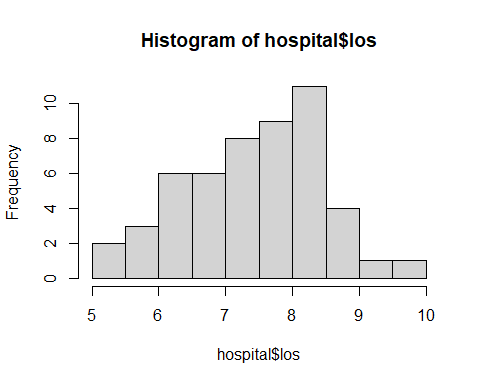
# median: 2600  
# mean: 2717  
summary(hospital$los)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 5.40 6.65 7.70 7.49 8.30 9.70

# median: 7.70  
# mean: 7.49  
  
# Part B:  
hist(hospital$expadm)



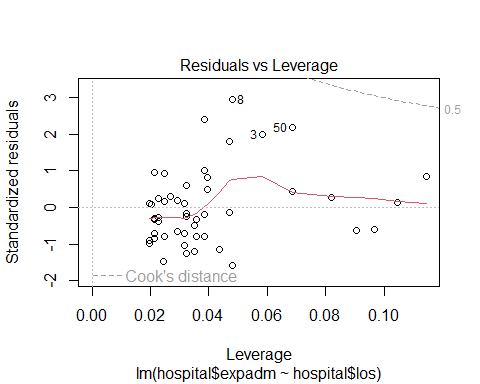
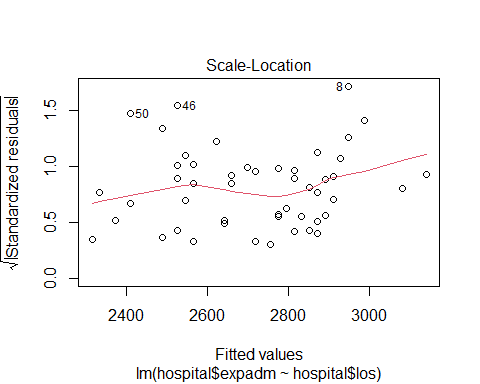
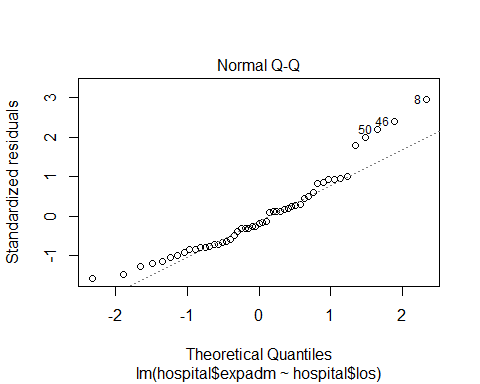
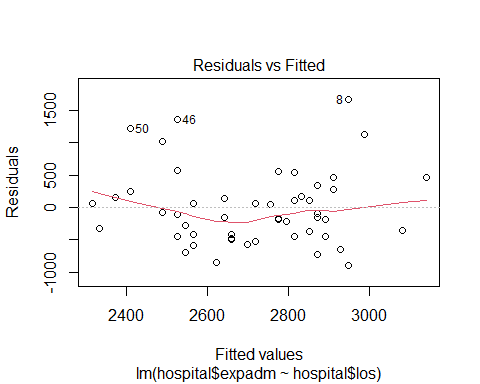
hist(hospital$los)



# Part C:  
summary(lm(hospital$expadm~hospital$los))

##   
## Call:  
## lm(formula = hospital$expadm ~ hospital$los)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -889.6 -428.1 -102.1 265.8 1663.4   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1281.96 608.10 2.108 0.0402 \*  
## hospital$los 191.56 80.47 2.381 0.0212 \*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 577.6 on 49 degrees of freedom  
## Multiple R-squared: 0.1037, Adjusted R-squared: 0.08538   
## F-statistic: 5.668 on 1 and 49 DF, p-value: 0.02121

# p-value: 0.02121  
  
# Part D: As the length of the stay increases, the expense of the stay increases by approximately $191.6  
  
# Part E:  
plot(lm(hospital$expadm~hospital$los))



# the residual plot gives us insight into how the outliers fit into the model as well.  
  
#11  
lowbwt <- read\_excel("lowbwt.xlsx")  
# View(lowbwt)  
# Part A:  
shapiro.test(lowbwt$sbp)

##   
## Shapiro-Wilk normality test  
##   
## data: lowbwt$sbp  
## W = 0.97911, p-value = 0.1134

# p-vue = 0.1134  
shapiro.test(lowbwt$gestage)

##   
## Shapiro-Wilk normality test  
##   
## data: lowbwt$gestage  
## W = 0.97995, p-value = 0.1316

# p-value = 0.1316  
lr <- lm(lowbwt$sbp~lowbwt$apgar5+lowbwt$gestage)  
summary(lr)

##   
## Call:  
## lm(formula = lowbwt$sbp ~ lowbwt$apgar5 + lowbwt$gestage)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -22.374 -8.180 -1.088 4.985 39.424   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 9.8034 12.6629 0.774 0.4407   
## lowbwt$apgar5 0.4875 0.4613 1.057 0.2932   
## lowbwt$gestage 1.1848 0.4424 2.678 0.0087 \*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 10.99 on 97 degrees of freedom  
## Multiple R-squared: 0.08944, Adjusted R-squared: 0.07066   
## F-statistic: 4.764 on 2 and 97 DF, p-value: 0.01063

# H0: The gestational age and APGAR scores are not correlated to systolic blood pressure measurement  
# Ha: The gestational age and APGAR scores are correlated to systolic blood pressure measurement  
  
# Part B:  
# b1 (APGAR5) = 0.4875, p-value = 0.2932  
# Since 0.2932 > 0.05, we FTR the null that the APGAR score significantly impacts the sbp measurement  
  
# Part C:  
# b1 is the increase in APGAR5 score (0.4875) per the increase in sbp  
# b2 is the increase in gestage (1.1848) per the increase in sbp  
  
# Part D:  
new <- data.frame(31,7)  
colnames(new) <- c('gestage','apgar5')  
new

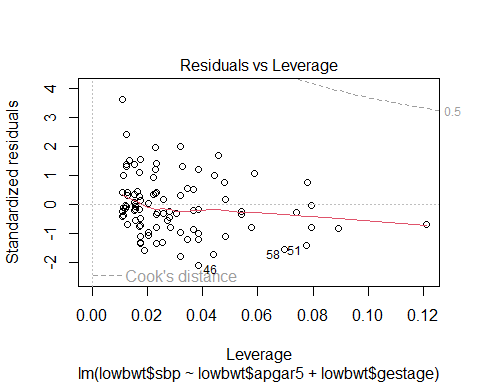
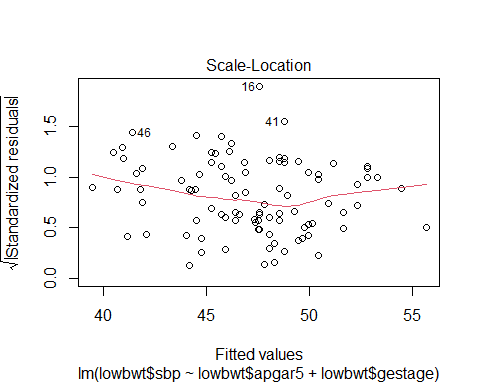
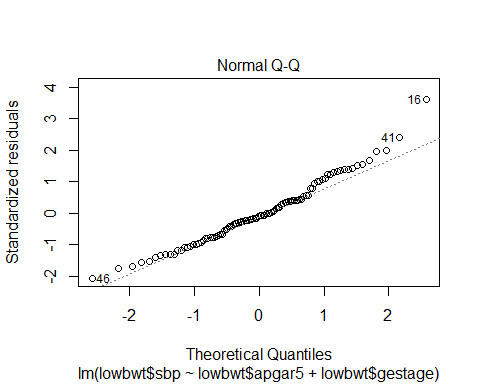
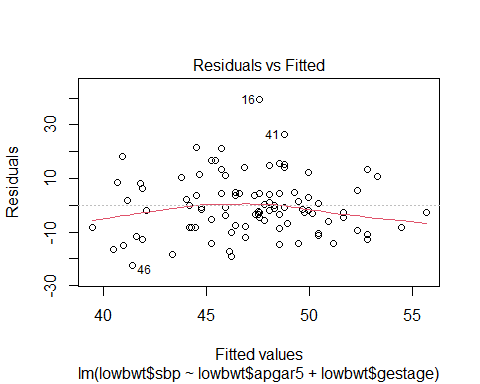
## gestage apgar5  
## 1 31 7

head(predict(lr,newdata=new),n=1)

## Warning: 'newdata' had 1 row but variables found have 100 rows

## 1   
## 47.57597

# 47.57 mmHg  
# or  
y <- 9.8 + 0.4875\*7+1.1848\*31   
# y=49.9 mmHg  
  
# Part E:  
# Adjusted R-squared: 0.07066   
# the adjusted R squared is the best indicator of proportion of the variance   
# for a response variable that’s explained by an explanatory variable  
# The addition of variables will increase R squared; however, the adjusted R  
# squared accounts for the addition of variables  
# In this instance, only 7% of the movement in the sbp is explained by the   
# gestage and the apgar5 score.  
  
# Part F:  
plot(lr) # passed



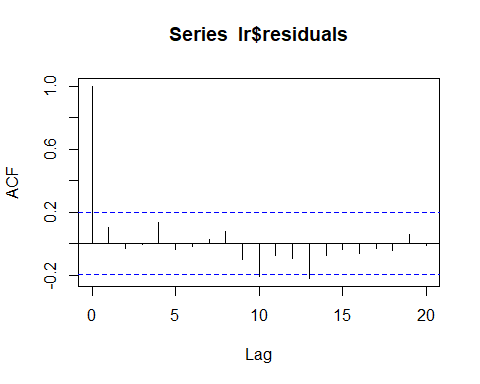
# Both of the explanatory variables are normally distributed  
vif(lr)

## lowbwt$apgar5 lowbwt$gestage   
## 1.029816 1.029816

# No Multicolinearity:  
# lowbwt$apgar5 lowbwt$gestage   
# 1.029816 1.029816   
  
# Homoscedasticity  
ncvTest(lr)

## Non-constant Variance Score Test   
## Variance formula: ~ fitted.values   
## Chisquare = 1.180479, Df = 1, p = 0.27726

# H0: The dataset has equal or similar variances  
# Ha: The dataset does NOT have equal or similar variances  
# p-value = 0.27726 > 0.05, therefore, FTR the null  
  
# No Autocorrelation  
print(acf(lr$residuals,pl=T))



##   
## Autocorrelations of series 'lr$residuals', by lag  
##   
## 0 1 2 3 4 5 6 7 8 9 10   
## 1.000 0.101 -0.032 -0.007 0.134 -0.034 -0.015 0.029 0.078 -0.097 -0.206   
## 11 12 13 14 15 16 17 18 19 20   
## -0.075 -0.096 -0.220 -0.073 -0.036 -0.063 -0.027 -0.045 0.056 -0.011

# No high autocorrelation  
  
#12  
# Part A:  
lr <- lm(sbp~gestage+apgar5+as.factor(sex),data=lowbwt)  
summary(lr)

##   
## Call:  
## lm(formula = sbp ~ gestage + apgar5 + as.factor(sex), data = lowbwt)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -21.854 -7.996 -1.046 5.092 38.737   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 9.3235 12.7371 0.732 0.46595   
## gestage 1.1853 0.4440 2.670 0.00892 \*\*  
## apgar5 0.4747 0.4635 1.024 0.30840   
## as.factor(sex)1 1.2434 2.2253 0.559 0.57764   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 11.03 on 96 degrees of freedom  
## Multiple R-squared: 0.09239, Adjusted R-squared: 0.06403   
## F-statistic: 3.257 on 3 and 96 DF, p-value: 0.02494

# interpretation of coefficient of sex: 1.24  
# compared with that of women (0=female), male has a higer sbp, on an average of 1.24 mmHg.  
  
# Part B:  
lr <- lm(sbp~gestage+apgar5+as.factor(sex)+gestage\*as.factor(sex),data=lowbwt)  
summary(lr)

##   
## Call:  
## lm(formula = sbp ~ gestage + apgar5 + as.factor(sex) + gestage \*   
## as.factor(sex), data = lowbwt)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -22.534 -7.660 -0.943 5.246 38.697   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 14.4355 15.2461 0.947 0.3461   
## gestage 1.0072 0.5313 1.896 0.0611 .  
## apgar5 0.4795 0.4651 1.031 0.3052   
## as.factor(sex)1 -15.7541 27.7402 -0.568 0.5714   
## gestage:as.factor(sex)1 0.5881 0.9567 0.615 0.5402   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 11.07 on 95 degrees of freedom  
## Multiple R-squared: 0.09599, Adjusted R-squared: 0.05792   
## F-statistic: 2.522 on 4 and 95 DF, p-value: 0.04606

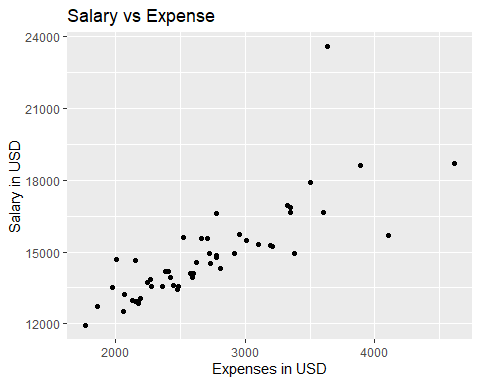
# since the interaction does not have a significance and it decreases the significance  
# of the gestage, my decision is to remove it from the model as it is not adding anything  
# p-value = 0.5402  
  
#13  
heart <- read\_excel("heart2.xlsx")  
# View(heart)  
# Parts A and B  
lr1 <- lm(pdi~as.factor(trtment),data=heart)  
summary(lr1)

##   
## Call:  
## lm(formula = pdi ~ as.factor(trtment), data = heart)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -47.771 -6.845 0.229 12.229 42.082   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 91.918 1.830 50.240 <2e-16 \*\*\*  
## as.factor(trtment)1 5.854 2.615 2.239 0.0268 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 15.63 on 141 degrees of freedom  
## (14 observations deleted due to missingness)  
## Multiple R-squared: 0.03432, Adjusted R-squared: 0.02747   
## F-statistic: 5.011 on 1 and 141 DF, p-value: 0.02676

# compared with that of trtment group 0, trtment group 1 has a higher pdi, on an average of 5.85 units.   
lr2 <- lm(mdi~as.factor(trtment),data=heart)  
summary(lr2)

##   
## Call:  
## lm(formula = mdi ~ as.factor(trtment), data = heart)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -56.400 -9.222 0.600 11.028 38.838   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 103.162 1.810 56.981 <2e-16 \*\*\*  
## as.factor(trtment)1 3.238 2.597 1.247 0.214   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 15.57 on 142 degrees of freedom  
## (13 observations deleted due to missingness)  
## Multiple R-squared: 0.01083, Adjusted R-squared: 0.003865   
## F-statistic: 1.555 on 1 and 142 DF, p-value: 0.2145

# compared with that of trtment group 0, trtment group 1 has a higher mdi, on an average of 0.214 units.  
  
# Part C:  
# PDI response: p-value = 0.0268  
# MDI response: p-value = 0.214  
# the pdi responded significantly to the trtment groups.   
  
  
# 14  
# Part A:  
ggplot(hospital) + geom\_point(aes(x=expadm,y=salary)) +   
 ggtitle('Salary vs Expense') + ylab('Salary in USD') +   
 xlab('Expenses in USD')



# this graph indicates a linear relationship between the two variables  
  
# Part B:  
lr1 <- lm(expadm~salary,data=hospital)  
summary(lr1)

##   
## Call:  
## lm(formula = expadm ~ salary, data = hospital)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1215.89 -176.14 -16.85 120.92 1182.44   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -905.7229 399.9285 -2.265 0.028 \*   
## salary 0.2439 0.0267 9.135 3.71e-12 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 371.1 on 49 degrees of freedom  
## Multiple R-squared: 0.6301, Adjusted R-squared: 0.6225   
## F-statistic: 83.45 on 1 and 49 DF, p-value: 3.709e-12

# as the salary increases, the average expense increases 0.2439 dollars  
  
lr2 <- lm(expadm~los,data=hospital)  
summary(lr2)

##   
## Call:  
## lm(formula = expadm ~ los, data = hospital)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -889.6 -428.1 -102.1 265.8 1663.4   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1281.96 608.10 2.108 0.0402 \*  
## los 191.56 80.47 2.381 0.0212 \*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 577.6 on 49 degrees of freedom  
## Multiple R-squared: 0.1037, Adjusted R-squared: 0.08538   
## F-statistic: 5.668 on 1 and 49 DF, p-value: 0.02121

# as the number of days increases, on average, the expense increases 191.56 dollars  
  
# Part C:  
lr3 <- lm(expadm~los+salary+los\*salary,data=hospital)  
summary(lr3)

##   
## Call:  
## lm(formula = expadm ~ los + salary + los \* salary, data = hospital)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -627.55 -130.35 20.39 97.50 794.48   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 6.073e+03 1.973e+03 3.079 0.003467 \*\*   
## los -1.044e+03 2.832e+02 -3.687 0.000587 \*\*\*  
## salary -3.097e-01 1.261e-01 -2.455 0.017827 \*   
## los:salary 8.154e-02 1.821e-02 4.478 4.8e-05 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 256.1 on 47 degrees of freedom  
## Multiple R-squared: 0.831, Adjusted R-squared: 0.8202   
## F-statistic: 77.04 on 3 and 47 DF, p-value: < 2.2e-16

anova(lr1, lr2, lr3)

## Analysis of Variance Table  
##   
## Model 1: expadm ~ salary  
## Model 2: expadm ~ los  
## Model 3: expadm ~ los + salary + los \* salary  
## Res.Df RSS Df Sum of Sq F Pr(>F)   
## 1 49 6746770   
## 2 49 16346819 0 -9600050   
## 3 47 3081890 2 13264929 101.15 < 2.2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

AIC(lr1, lr2, lr3)

## df AIC  
## lr1 3 752.1619  
## lr2 3 797.2954  
## lr3 5 716.2024

# the addition of both variables as well as an analysis of the interaction between the two (model lr3) does seem to improve the model's accuracy; however, this must be tested  
  
  
# 15  
detroit <- read\_excel("detroit.xlsx")  
# View(detroit)  
# Part A:  
lr <- lm(homicide~register+weekly+unemp+police,data=detroit)  
summary(lr)

##   
## Call:  
## lm(formula = homicide ~ register + weekly + unemp + police, data = detroit)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -6.0660 -0.7642 -0.0979 2.0019 4.0199   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -54.103370 11.119988 -4.865 0.00125 \*\*  
## register 0.018047 0.005398 3.344 0.01018 \*   
## weekly 0.095993 0.073464 1.307 0.22763   
## unemp 0.336170 0.538572 0.624 0.54989   
## police 0.167873 0.085514 1.963 0.08525 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3.275 on 8 degrees of freedom  
## Multiple R-squared: 0.9734, Adjusted R-squared: 0.9601   
## F-statistic: 73.1 on 4 and 8 DF, p-value: 2.461e-06

# p-value: 2.78e-06  
# Adjusted R-squared: 0.961  
  
# Part B:  
# The number of registered guns is the only variable that significantly  
# impacts the number of homicides p-value = 0.01018 \*   
  
# Part C:   
vif(lr)

## register weekly unemp police   
## 3.153817 10.913547 1.806433 17.930382

# register weekly unemp police   
# 3.010722 10.994870 1.734999 17.667117   
# remove police and try again  
lr1 <- lm(homicide~register+weekly+unemp,data=detroit)  
vif(lr1)

## register weekly unemp   
## 1.471500 1.495688 1.034609

# register weekly unemp   
# 1.461488 1.484048 1.024547  
summary(lr1)

##   
## Call:  
## lm(formula = homicide ~ register + weekly + unemp, data = detroit)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -8.3843 -0.9794 -0.3638 1.1534 4.6693   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -33.981258 4.948097 -6.868 7.33e-05 \*\*\*  
## register 0.025786 0.004231 6.094 0.000181 \*\*\*  
## weekly 0.229965 0.031212 7.368 4.25e-05 \*\*\*  
## unemp 1.027264 0.467766 2.196 0.055693 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3.758 on 9 degrees of freedom  
## Multiple R-squared: 0.9605, Adjusted R-squared: 0.9474   
## F-statistic: 73.03 on 3 and 9 DF, p-value: 1.225e-06

# p-value: 1.673e-06  
# Adjusted R-squared: 0.9436  
lr2 <- lm(homicide~register+police+unemp,data=detroit)  
vif(lr2)

## register police unemp   
## 2.249455 2.457336 1.246432

# register police unemp   
# 2.186117 2.384643 1.207249   
summary(lr2)

##   
## Call:  
## lm(formula = homicide ~ register + police + unemp, data = detroit)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -4.2641 -1.7954 -0.0929 2.2304 4.8048   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -65.065055 7.580651 -8.583 1.26e-05 \*\*\*  
## register 0.014271 0.004734 3.014 0.0146 \*   
## police 0.271673 0.032878 8.263 1.71e-05 \*\*\*  
## unemp -0.055656 0.464619 -0.120 0.9073   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3.401 on 9 degrees of freedom  
## Multiple R-squared: 0.9677, Adjusted R-squared: 0.9569   
## F-statistic: 89.84 on 3 and 9 DF, p-value: 5.002e-07

# p-value: 4.98e-07  
# Adjusted R-squared: 0.957   
lr3 <- lm(homicide~register,data=detroit)  
# p-value: 0.0007015  
anova(lr,lr1,lr2,lr3)

## Analysis of Variance Table  
##   
## Model 1: homicide ~ register + weekly + unemp + police  
## Model 2: homicide ~ register + weekly + unemp  
## Model 3: homicide ~ register + police + unemp  
## Model 4: homicide ~ register  
## Res.Df RSS Df Sum of Sq F Pr(>F)   
## 1 8 85.80   
## 2 9 127.12 -1 -41.33 3.8538 0.08525 .   
## 3 9 104.11 0 23.02   
## 4 11 1075.03 -2 -970.93 45.2672 4.345e-05 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

AIC(lr, lr1, lr2, lr3)

## df AIC  
## lr 6 73.42357  
## lr1 5 76.53526  
## lr2 5 73.93837  
## lr3 3 100.28942

# the model that has only the registered weapons is statistically  
# a better fit than the other models p < 2e-16  
  
# Part D:  
summary(lr3)

##   
## Call:  
## lm(formula = homicide ~ register, data = detroit)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -17.914 -2.901 -2.154 1.398 22.007   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.662063 5.708193 0.291 0.776337   
## register 0.043003 0.009175 4.687 0.000664 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 9.886 on 11 degrees of freedom  
## Multiple R-squared: 0.6663, Adjusted R-squared: 0.636   
## F-statistic: 21.97 on 1 and 11 DF, p-value: 0.0006642

# as the number of registered weapons increases, the number  
# of homicides increases by 0.042274   
  
# Part E:  
# Adjusted R-squared: 0.6325, meaning that the registered weapons can only  
# account for 63.25% of homicides  
rm(lr1,lr2,lr3,lr)  
  
#16  
# Part A:  
lr <- glm(grmhem~apgar5+tox,data=lowbwt,family='binomial')  
summary(lr)

##   
## Call:  
## glm(formula = grmhem ~ apgar5 + tox, family = "binomial", data = lowbwt)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.0865 -0.5400 -0.4840 -0.2898 2.4366   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -0.2177 0.6268 -0.347 0.7283   
## apgar5 -0.2334 0.1051 -2.222 0.0263 \*  
## tox -1.2974 1.0778 -1.204 0.2287   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 84.542 on 99 degrees of freedom  
## Residual deviance: 76.967 on 97 degrees of freedom  
## AIC: 82.967  
##   
## Number of Fisher Scoring iterations: 5

# APGAR5 p-value: 0.0263 \*  
  
# Part B:  
new <- data.frame(apgar5 = 3,tox = 1)  
stats::predict(lr,new,type="response")

## 1   
## 0.09837279

# 0.09837279, 9.83% likely to have experienced germinal matrix hem.  
  
# Part C:  
summary(lr)

##   
## Call:  
## glm(formula = grmhem ~ apgar5 + tox, family = "binomial", data = lowbwt)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.0865 -0.5400 -0.4840 -0.2898 2.4366   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -0.2177 0.6268 -0.347 0.7283   
## apgar5 -0.2334 0.1051 -2.222 0.0263 \*  
## tox -1.2974 1.0778 -1.204 0.2287   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 84.542 on 99 degrees of freedom  
## Residual deviance: 76.967 on 97 degrees of freedom  
## AIC: 82.967  
##   
## Number of Fisher Scoring iterations: 5

# b1 = -0.2334, p-value: 0.0263 \*  
# Therefore, we reject the null hypothesis  
  
# 17  
stenosis <- read\_excel("stenosis.xlsx")  
# View(stenosis)  
# Part A:  
lr <- glm(disease~smoke+sex,data=stenosis,family='binomial')  
summary(lr)

##   
## Call:  
## glm(formula = disease ~ smoke + sex, family = "binomial", data = stenosis)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.3630 -1.0555 -0.9783 1.0807 1.3905   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -0.4882 0.2159 -2.261 0.0238 \*  
## smoke 0.1946 0.2903 0.670 0.5026   
## sex 0.7199 0.2881 2.499 0.0125 \*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 297.94 on 214 degrees of freedom  
## Residual deviance: 289.64 on 212 degrees of freedom  
## AIC: 295.64  
##   
## Number of Fisher Scoring iterations: 4

# sex p-value: 0.0125 \*  
vif(lr)

## smoke sex   
## 4.457775 4.459868

new <- with(stenosis, data.frame(smoke = 1, sex=mean(sex, na.rm=TRUE)))  
stats::predict(lr,new,type='response')

## 1   
## 0.5153336

# 0.5153336, 51.5%  
new <- with(stenosis, data.frame(smoke = 0, sex=mean(sex, na.rm=TRUE)))  
stats::predict(lr,new,type='response')

## 1   
## 0.466738

# 0.466738, 46.7%  
  
# Part B:  
new <- with(stenosis, data.frame(sex = 1, smoke=mean(smoke, na.rm=TRUE)))  
stats::predict(lr,new,type='response')

## 1   
## 0.5785453

# 0.5785453, 57.8%  
new <- with(stenosis, data.frame(sex = 0, smoke=mean(smoke, na.rm=TRUE)))  
stats::predict(lr,new,type='response')

## 1   
## 0.4005706

# 0.4005706, 40%  
# it appears as though there is a difference, since our binomial model returned  
# a significant p value for sex, we reject the null hypothesis that it does not affect  
# the disease presence  
  
#18  
dialysis <- read\_excel("dialysis.xlsx")  
# View(dialysis)  
# Part A:  
lr1 <- glm(perito ~ age, data=dialysis,family='binomial')  
summary(lr1)

##   
## Call:  
## glm(formula = perito ~ age, family = "binomial", data = dialysis)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.4854 -1.4427 0.9172 0.9302 0.9438   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)  
## (Intercept) 0.770491 1.372803 0.561 0.575  
## age -0.002713 0.025542 -0.106 0.915  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 59.440 on 45 degrees of freedom  
## Residual deviance: 59.429 on 44 degrees of freedom  
## AIC: 63.429  
##   
## Number of Fisher Scoring iterations: 4

# p-value: 0.915  
lr2 <- glm(perito ~ sex, data=dialysis,family='binomial')  
summary(lr2)

##   
## Call:  
## glm(formula = perito ~ sex, family = "binomial", data = dialysis)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.6651 -1.3116 0.7585 1.0489 1.0489   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)  
## (Intercept) 0.3102 0.3970 0.781 0.435  
## sex 0.7885 0.6513 1.211 0.226  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 59.440 on 45 degrees of freedom  
## Residual deviance: 57.919 on 44 degrees of freedom  
## AIC: 61.919  
##   
## Number of Fisher Scoring iterations: 4

# p-value: 0.226  
lr3 <- glm(perito ~ race, data=dialysis,family='binomial')  
summary(lr3)

##   
## Call:  
## glm(formula = perito ~ race, family = "binomial", data = dialysis)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.8465 -1.3537 0.6335 1.0108 1.0108   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)  
## (Intercept) 0.4055 0.3450 1.175 0.240  
## race 1.0986 0.8545 1.286 0.199  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 59.440 on 45 degrees of freedom  
## Residual deviance: 57.542 on 44 degrees of freedom  
## AIC: 61.542  
##   
## Number of Fisher Scoring iterations: 4

# p-value = 0.199  
  
# Part B:  
lr <- glm(perito ~ age+sex+race, data=dialysis,family='binomial')  
summary(lr)

##   
## Call:  
## glm(formula = perito ~ age + sex + race, family = "binomial",   
## data = dialysis)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.9665 -1.2613 0.7478 0.8918 1.1579   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)  
## (Intercept) -0.105588 1.524711 -0.069 0.945  
## age 0.005228 0.027012 0.194 0.847  
## sex 0.658956 0.676989 0.973 0.330  
## race 0.962492 0.875976 1.099 0.272  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 59.440 on 45 degrees of freedom  
## Residual deviance: 56.569 on 42 degrees of freedom  
## AIC: 64.569  
##   
## Number of Fisher Scoring iterations: 4

new <- with(dialysis, data.frame(race = 0, sex=0,age=mean(age, na.rm=TRUE)))  
stats::predict(lr,new,type='response')

## 1   
## 0.5417879

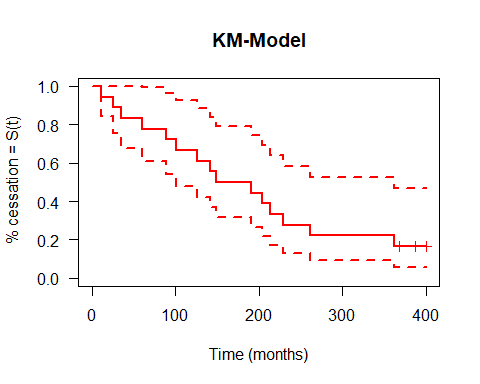
# 0.5417879, 54.2%  
  
# Part C:  
# The response variable is categorized as a dichotomous random variable. It should be listed as a numerical data, instead of a categorical data.   
  
  
#19  
cyto <- read\_excel("cyto.xlsx")  
# View(cyto)  
# Part A:  
km.cyto <- survfit(Surv(time[group==1], censor[group==1]) ~ 1, type = "kaplan-meier", data = cyto)  
km.cyto # median censor time is 170 months

## Call: survfit(formula = Surv(time[group == 1], censor[group == 1]) ~   
## 1, data = cyto, type = "kaplan-meier")  
##   
## n events median 0.95LCL 0.95UCL  
## [1,] 18 15 170 101 362

summary(km.cyto)

## Call: survfit(formula = Surv(time[group == 1], censor[group == 1]) ~   
## 1, data = cyto, type = "kaplan-meier")  
##   
## time n.risk n.event survival std.err lower 95% CI upper 95% CI  
## 11 18 1 0.944 0.0540 0.8443 1.000  
## 26 17 1 0.889 0.0741 0.7549 1.000  
## 35 16 1 0.833 0.0878 0.6778 1.000  
## 60 15 1 0.778 0.0980 0.6076 0.996  
## 89 14 1 0.722 0.1056 0.5423 0.962  
## 101 13 1 0.667 0.1111 0.4809 0.924  
## 126 12 1 0.611 0.1149 0.4227 0.883  
## 142 11 1 0.556 0.1171 0.3675 0.840  
## 149 10 1 0.500 0.1179 0.3150 0.794  
## 191 9 1 0.444 0.1171 0.2652 0.745  
## 204 8 1 0.389 0.1149 0.2179 0.694  
## 213 7 1 0.333 0.1111 0.1734 0.641  
## 229 6 1 0.278 0.1056 0.1319 0.585  
## 261 5 1 0.222 0.0980 0.0936 0.527  
## 362 4 1 0.167 0.0878 0.0593 0.468

plot(km.cyto,conf.int=T, xlab="Time (months)", ylab=" % cessation = S(t)", col=c("red"),  
 main="KM-Model", las=1, lwd=2, mark.time=T)



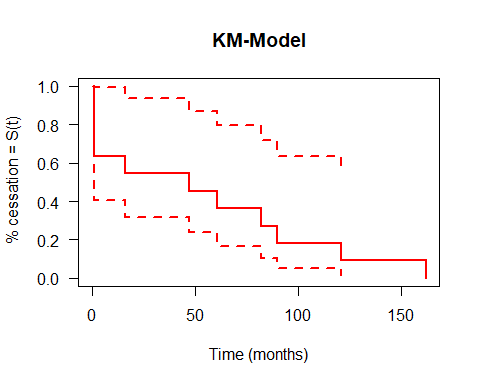
km.cyto <- survfit(Surv(time[group==2], censor[group==2]) ~ 1, type = "kaplan-meier", data = cyto)  
km.cyto # median censor time is 47 months

## Call: survfit(formula = Surv(time[group == 2], censor[group == 2]) ~   
## 1, data = cyto, type = "kaplan-meier")  
##   
## n events median 0.95LCL 0.95UCL  
## [1,] 11 11 47 1 NA

summary(km.cyto)

## Call: survfit(formula = Surv(time[group == 2], censor[group == 2]) ~   
## 1, data = cyto, type = "kaplan-meier")  
##   
## time n.risk n.event survival std.err lower 95% CI upper 95% CI  
## 1 11 4 0.6364 0.1450 0.4071 0.995  
## 16 7 1 0.5455 0.1501 0.3180 0.936  
## 47 6 1 0.4545 0.1501 0.2379 0.868  
## 61 5 1 0.3636 0.1450 0.1664 0.795  
## 82 4 1 0.2727 0.1343 0.1039 0.716  
## 90 3 1 0.1818 0.1163 0.0519 0.637  
## 121 2 1 0.0909 0.0867 0.0140 0.589  
## 162 1 1 0.0000 NaN NA NA

plot(km.cyto,conf.int=T, xlab="Time (months)", ylab=" % cessation = S(t)", col=c("red"),  
 main="KM-Model", las=1, lwd=2, mark.time=T)



# Part B:  
# it appears as though the individuals in group 1 (took drug) survive for longer  
# than the individuals in group 2 (did not), due to the median censor time  
# group 1: 170, group 2: 47  
  
# Part C:  
survdiff(Surv(time, censor)~ group, data=cyto)

## Call:  
## survdiff(formula = Surv(time, censor) ~ group, data = cyto)  
##   
## N Observed Expected (O-E)^2/E (O-E)^2/V  
## group=1 18 15 21.37 1.90 12.4  
## group=2 11 11 4.63 8.74 12.4  
##   
## Chisq= 12.4 on 1 degrees of freedom, p= 4e-04

# p= 4e-04, therefore, we reject the null hypothesis. There is difference in the survival times between the two groups  
  
#20  
bladder <- read\_excel("bladder.xlsx")  
# View(bladder)  
# Part A:  
km.bl <- survfit(Surv(time[group==1], censor[group==1]) ~ 1, type = "kaplan-meier", data = bladder)  
km.bl # median censor time is 16 months

## Call: survfit(formula = Surv(time[group == 1], censor[group == 1]) ~   
## 1, data = bladder, type = "kaplan-meier")  
##   
## n events median 0.95LCL 0.95UCL  
## [1,] 48 29 16 9 NA

summary(km.bl)

## Call: survfit(formula = Surv(time[group == 1], censor[group == 1]) ~   
## 1, data = bladder, type = "kaplan-meier")  
##   
## time n.risk n.event survival std.err lower 95% CI upper 95% CI  
## 1 47 1 0.979 0.0210 0.938 1.000  
## 2 45 4 0.892 0.0457 0.806 0.986  
## 3 41 7 0.739 0.0647 0.623 0.878  
## 5 33 2 0.695 0.0681 0.573 0.842  
## 6 31 2 0.650 0.0707 0.525 0.804  
## 7 29 1 0.627 0.0717 0.502 0.785  
## 9 27 2 0.581 0.0735 0.453 0.745  
## 10 25 1 0.558 0.0742 0.430 0.724  
## 12 23 2 0.509 0.0752 0.381 0.680  
## 16 20 1 0.484 0.0757 0.356 0.657  
## 17 19 1 0.458 0.0758 0.331 0.634  
## 18 18 1 0.433 0.0758 0.307 0.610  
## 25 15 1 0.404 0.0760 0.279 0.584  
## 28 13 1 0.373 0.0763 0.250 0.557  
## 29 12 1 0.342 0.0760 0.221 0.528  
## 35 6 1 0.285 0.0819 0.162 0.501

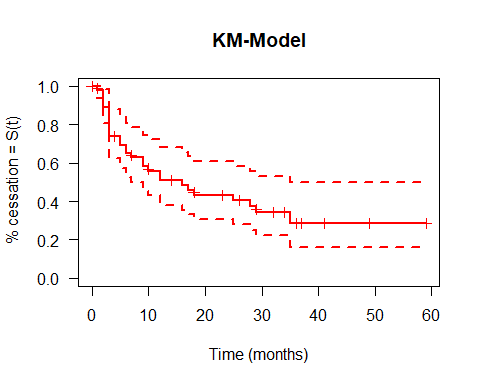
plot(km.bl,conf.int=T, xlab="Time (months)", ylab=" % cessation = S(t)", col=c("red"),  
 main="KM-Model", las=1, lwd=2, mark.time=T)  
  
km.b2 <- survfit(Surv(time[group==2], censor[group==2]) ~ 1, type = "kaplan-meier", data = bladder)  
km.b2 # median censor time is 26 months

## Call: survfit(formula = Surv(time[group == 2], censor[group == 2]) ~   
## 1, data = bladder, type = "kaplan-meier")  
##   
## n events median 0.95LCL 0.95UCL  
## [1,] 38 18 26 17 NA

summary(km.b2)

## Call: survfit(formula = Surv(time[group == 2], censor[group == 2]) ~   
## 1, data = bladder, type = "kaplan-meier")  
##   
## time n.risk n.event survival std.err lower 95% CI upper 95% CI  
## 1 38 2 0.947 0.0362 0.879 1.000  
## 2 34 4 0.836 0.0613 0.724 0.965  
## 3 30 1 0.808 0.0653 0.690 0.947  
## 4 29 2 0.752 0.0717 0.624 0.907  
## 5 27 1 0.724 0.0743 0.593 0.886  
## 6 26 2 0.669 0.0783 0.532 0.841  
## 17 21 2 0.605 0.0828 0.463 0.791  
## 22 18 1 0.571 0.0848 0.427 0.764  
## 24 16 1 0.536 0.0867 0.390 0.736  
## 26 12 1 0.491 0.0902 0.343 0.704  
## 38 11 1 0.446 0.0924 0.298 0.670

plot(km.bl,conf.int=T, xlab="Time (months)", ylab=" % cessation = S(t)", col=c("red"),  
 main="KM-Model", las=1, lwd=2, mark.time=T)



# Part B:  
# it appears as though the individuals in group 1 (placebo) have a shorter  
# time until the tumor reappears than individuals in group 2 (thiopeta)  
# group 1: 16, group 2: 26  
  
# Part C:  
survdiff(Surv(time, censor)~ group, data=bladder)

## Call:  
## survdiff(formula = Surv(time, censor) ~ group, data = bladder)  
##   
## N Observed Expected (O-E)^2/E (O-E)^2/V  
## group=1 48 29 24.9 0.671 1.52  
## group=2 38 18 22.1 0.757 1.52  
##   
## Chisq= 1.5 on 1 degrees of freedom, p= 0.2

# p= 0.2, therefore, we FTR the null hypothesis that the drug does not affect tumor regrowth