```
################################################################################

# Brian Weinstein - bmw2148
# STAT W4201 001
# Homework 5
# 2016-03-02

# set working directory
setwd("~/Documents/advanced-data-analysis/homework_05")

# prevent R from printing large numbers in scientific notation
options(scipen=5)

# load packages
library(Sleuth3) # Data sets from Ramsey and Schafer's "Statistical Sleuth
(3rd ed)"
library(ggplot2); theme_set(theme_bw())
library(gridExtra)
library(GGally)
library(dplyr)



# Problem 1: Ramsey 7.18
################################################################################

# define coefficients and constants from Display 7.12
beta0 <- 6.9836
beta1 <- -0.7257
sigmaHat <- 0.08226
n <- 10
xBar <- 1.190
sampleVarX <- 0.6344
x0 <- log(5)

# Part a ### ### ### ### ### ### ### ### ### ### ### ### ### ### ### ### ### ###
###

# calculate the standard error of the prediction
sep <- sigmaHat * sqrt(1 + (1/n) + ((x0 - xBar)^2 /  ((n-1) * sampleVarX))) ;
sep

# Part b ### ### ### ### ### ### ### ### ### ### ### ### ### ### ### ### ### ###
###

# predicted value at x0=log(5)
pred <- beta0 + (beta1 * x0) ; pred

# 95% prediction confidence interval
t <- qt(p=0.975, df=(n-2)) ; t
lowerBound <- pred - (t * sep)
upperBound <- pred + (t * sep)
c(lowerBound, upperBound)
```

```
rm(list = ls()) # clear working environment



# Problem 2: Ramsey 7.24
######################################################################

# load data
birthData <- Sleuth3::ex0724

# Part a ### ### ### ### ### ### ### ### ### ### ### ### ### ### ### ### ###
###

# fit a linear model to Denmark ~ Year
lmDenmark <- lm(formula=Denmark~Year, data=birthData)
summary(lmDenmark)$coefficients

# fit a linear model to Netherlands ~ Year
lmNetherlands <- lm(formula=Netherlands~Year, data=birthData)
summary(lmNetherlands)$coefficients

# fit a linear model to Canada ~ Year
lmCanada <- lm(formula=Canada~Year, data=birthData)
summary(lmCanada)$coefficients

# fit a linear model to USA ~ Year
lmUsa <- lm(formula=USA~Year, data=birthData)
summary(lmUsa)$coefficients

# Part b ### ### ### ### ### ### ### ### ### ### ### ### ### ### ### ### ###
###

# Denmark
summary(lmDenmark)$coefficients
c(-2.072598, 4.423828e-02/2)

# Netherlands
summary(lmNetherlands)$coefficients
c(-5.710196, 9.636921e-07/2)

# Canada
summary(lmCanada)$coefficients
c(-4.016653, 7.375947e-04/2)

# USA
summary(lmUsa)$coefficients
c(-5.779212, 1.439109e-05/2)

# Part c ### ### ### ### ### ### ### ### ### ### ### ### ### ### ### ### ###
###

# no code needed

# Part d ### ### ### ### ### ### ### ### ### ### ### ### ### ### ### ### ###
```

```
###

# calculate the RSS for the USA model
sum((lmUsa$residuals)^2)

# calculate the RSS for the Canada model
sum((lmCanada$residuals)^2)

# Part e ### ### ### ### ### ### ### ### ### ### ### ### ### ### ### ### ### ###
###

# no code needed

rm(list = ls()) # clear working environment



# Problem 3: Ramsey 7.28
#########################################################################

# load data
brainData <- Sleuth3::ex0728

# Part a ### ### ### ### ### ### ### ### ### ### ### ### ### ### ### ### ### ###
###

# create a variable groupin the subjects into "control" and "player" groups
brainData <- brainData %>%
  mutate(Group=as.factor(ifelse(Years==0, "control", "player")))

# boxplots of brain activity for each group
ggplot(brainData, aes(x=Group, y=Activity)) +
  # geom_violin(alpha=0.15) +
  geom_boxplot() +
  labs(y="Neuronal activity index", x="Group")
ggsave(filename="writeup/3_orig.png", width=6.125, height=3.5, units="in")

# create a log(Activity) variable
brainData <- brainData %>%
  mutate(LogActivity=log(Activity))

# boxplots of brain activity for each group
ggplot(brainData, aes(x=Group, y=LogActivity)) +
  # geom_violin(alpha=0.15) +
  geom_boxplot() +
  labs(y="Neuronal activity index (log scale)", x="Group")
ggsave(filename="writeup/3_log.png", width=6.125, height=3.5, units="in")

# compare group standard deviations on the original and log scales
brainData %>%
  group_by(Group) %>%
  summarize(sd(Activity), sd(LogActivity)) %>%
  as.data.frame()

# Perform a two-sample t-test
```

```
tt <- t.test(formula=LogActivity~Group, data=brainData,
             var.equal=TRUE, conf.level=0.95, alternative="two.sided")
tt

# take antilog of the estimate
exp(diff(tt$estimate)[[1]])

# Perform a two sided t-test for the confidence interval and take antilog
exp(-tt$conf.int)

# Part b ### ### ### ### ### ### ### ### ### ### ### ### ### ### ### ### ### ###
###

# plot Activity vs Years
ggplot(brainData, aes(x=Years, y=Activity)) +
  geom_point() +
  geom_smooth(method=lm)
ggsave(filename="writeup/3b_fit.png", width=6.125, height=3.5, units="in")

# create a linear regression model of Activity on Years
lmBrain <- lm(formula=Activity~Years, data=brainData)
summary(lmBrain)$coefficients

# check the residuals of the fitted model
ggplot(lmBrain, aes(x=.fitted, y=.resid)) +
  geom_point() +
  geom_hline(yintercept=0, linetype="dashed") +
  labs(x="Fitted values", y="Residuals")
ggsave(filename="writeup/3b_resid.png", width=6.125, height=3.5, units="in")

rm(list = ls()) # clear working environment



# Problem 4: Ramsey 8.17
#########################################################################

# load data
pestData <- Sleuth3::ex0817

# Part a ### ### ### ### ### ### ### ### ### ### ### ### ### ### ### ### ### ###
###

# Mass vs Load
p1 <- ggplot(pestData, aes(x=Load, y=Mass)) +
  geom_point() +
  geom_smooth(method=lm) +
  labs(title="(1) Mass vs Load")

# log(Mass) vs Load
p2 <- ggplot(pestData, aes(x=Load, y=log(Mass))) +
  geom_point() +
  geom_smooth(method=lm) +
  labs(title="(2) log(Mass) vs Load")
```

```
# Mass vs log(Load)
p3 <- ggplot(pestData, aes(x=log(Load), y=Mass)) +
  geom_point() +
  geom_smooth(method=lm) +
  labs(title="(3) Mass vs log(Load)")

# log(Mass) vs log(Load)
p4 <- ggplot(pestData, aes(x=log(Load), y=log(Mass))) +
  geom_point() +
  geom_smooth(method=lm) +
  labs(title="(4) log(Mass) vs log(Load)")

# combine all 4 plots
pGrid <- grid.arrange(p1, p2, p3, p4, nrow=2, ncol=2)
ggsave(filename="writeup/4a.png", plot=pGrid, width=11, height=5, units="in")

# Part b ### ### ### ### ### ### ### ### ### ### ### ### ### ### ### ### ###
###

# create log-transformed variables
pestData <- pestData %>%
  mutate(LogLoad=log(Load),
         LogMass=log(Mass))

# create a linear regression model of Activity on Years
lmPest <- lm(formula=LogMass~LogLoad, data=pestData)
summary(lmPest)

# residuals
residuals(lmPest)

# fitted values
fitted(lmPest)

# Part c ### ### ### ### ### ### ### ### ### ### ### ### ### ### ### ### ###
###

# check the residuals of the fitted model
ggplot(lmPest, aes(x=.fitted, y=.resid)) +
  geom_point() +
  geom_hline(yintercept=0, linetype="dashed") +
  labs(x="Fitted values", y="Residuals")
ggsave(filename="writeup/4c.png", width=6.125, height=3.5, units="in")

rm(list = ls()) # clear working environment



# Problem 5: Ramsey 8.20
#########################################################################

# load data
voteData <- Sleuth3::ex0820

# Part a ### ### ### ### ### ### ### ### ### ### ### ### ### ### ### ### ###
```

```
###

# plot DemPctOfAbsenteeVotes vs DemPctOfMachineVotes
plot.all <- ggplot(voteData, aes(x=DemPctOfMachineVotes,
y=DemPctOfAbsenteeVotes)) +
  geom_point(aes(shape=Disputed, color=Disputed), size=2.5)
plot.all
ggsave(filename="writeup/5a.png", width=6.125, height=3.5, units="in")

# Part b ### ### ### ### ### ### ### ### ### ### ### ### ### ### ### ### ###
###

# create a linear regression model of DemPctOfAbsenteeVotes on
# DemPctOfMachineVotes, excluding the disputed election
lmVoteExclDisputed <- lm(formula=DemPctOfAbsenteeVotes~DemPctOfMachineVotes,
                         data=voteData, subset=(Disputed=="no"))
summary(lmVoteExclDisputed)

# compute the prediction interval band
predVoteExclDisputed <- cbind(filter(voteData, Disputed=="no"),
                              predict(lmVoteExclDisputed,
interval="prediction"))

# plot the scatterplot and prediciton interval
ggplot(predVoteExclDisputed, aes(x=DemPctOfMachineVotes,
y=DemPctOfAbsenteeVotes)) +
  geom_point(data=voteData, aes(shape=Disputed, color=Disputed), size=2.5) +
  geom_ribbon(aes(ymin=lwr, ymax=upr), fill="darkgray", alpha=0.4) +
  geom_smooth(method=lm) +
  labs(title='Linear Regression of
       DemPctOfAbsenteeVotes on DemPctOfMachineVotes
       (model excludes the Disputed=="yes" observation)', size=10)
ggsave(filename="writeup/5b.png", width=6.125, height=3.5, units="in")

# Part c ### ### ### ### ### ### ### ### ### ### ### ### ### ### ### ### ###
###

# find the predicted value of DemPctOfAbsenteeVotes at
DemPctOfMachineVotes=49.3
predDisputed <-  predict(lmVoteExclDisputed, newdata=filter(voteData,
Disputed=="yes"),
                         interval="prediction", se.fit=TRUE)
predDisputed

# calculate the standard error of the predicted value
predDisputedSe <- sqrt((predDisputed$se.fit)^2 +
(summary(lmVoteExclDisputed)$sigma)^2)
predDisputedSe

# calculate how many SEs away the observed pct is from the predicted value
obs <- voteData[voteData$Disputed=="yes", "DemPctOfAbsenteeVotes"] ; obs
pred <- predDisputed$fit[[1]] ; pred
tstat <- abs(obs-pred)/predDisputedSe ; tstat

# calculate the 2-sided p-value
```

```
pval <- 2 * pt(q=(-1 * abs(tstat)), df=summary(lmVoteExclDisputed)$df[[2]]) ;
pval

# Part d ### ### ### ### ### ### ### ### ### ### ### ### ### ### ### ### ###
###

# adjusting the pvalue with the Bonferroni correction
p.adjust(p=pval, method="bonferroni", n=22)

rm(list = ls()) # clear working environment



# Problem 6: Ramsey 9.12
######################################################################

# load data
mammalData <- Sleuth3::case0902

# Part a ### ### ### ### ### ### ### ### ### ### ### ### ### ### ### ### ###
###

# define logged variables
mammalData <- mammalData %>%
  mutate(LogBrain=log(Brain),
         LogBody=log(Body),
         LogGestation=log(Gestation),
         LogLitter=log(Litter))

# plot a matrix of pairwise scatterplots (log scale)
plot.pairsLog <- ggpairs(data=select(mammalData,
                                     LogBody, LogGestation, LogLitter,
LogBrain),
                         lower=list(continuous=wrap("points", size=0.7)))
png(filename="writeup/6a.png", width=11, height=9, units="in", res=300)
print(plot.pairsLog)
dev.off()

# Part b ### ### ### ### ### ### ### ### ### ### ### ### ### ### ### ### ###
###

# fit a multiple linear regression of log brain weight
# on log body weight, log gestation, and log litter size
lmMammal <- lm(formula=LogBrain ~ LogBody + LogGestation + LogLitter,
data=mammalData)
summary(lmMammal)$coefficients

# Part c ### ### ### ### ### ### ### ### ### ### ### ### ### ### ### ### ###
###

# plot a matrix of pairwise scatterplots (log scale, except Litter)
plot.pairsLogExceptLitter <- ggpairs(data=select(mammalData,
                                                 LogBody, LogGestation,
Litter, LogBrain),
                                     lower=list(continuous=wrap("points",
```

```
size=0.7)))
png(filename="writeup/6c.png", width=11, height=9, units="in", res=300)
print(plot.pairsLogExceptLitter)
dev.off()

rm(list = ls()) # clear working environment
```