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

# Brian Weinstein - bmw2148
# STAT W4201 001
# Homework 7
# 2016-03-28

# set working directory
setwd("~/Documents/advanced-data-analysis/homework_07")

# 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(GGally)
library(ggrepel)
library(MASS)
library(dplyr)
library(gridExtra)



# Problem 1: Ramsey 10.21
################################################################################

# no code needed



# Problem 2: Ramsey 10.22
################################################################################

# no code needed



# Problem 3: Ramsey 10.26
################################################################################

# load data
ozoneData <- Sleuth3::ex1026

# convert Surface to an indicator variable (lets us force the lm intercept to
0)
ozoneData <- ozoneData %>%
  mutate(Surface = as.integer(ifelse(Surface=="Surface", 1, 0)))

# plot of Inhibit vs UVB
ggplot(ozoneData, aes(x=UVB, y=Inhibit,
                      color=factor(Surface), shape=factor(Surface))) +
```

```
  geom_point(size=2)
ggsave(filename="writeup/3_eda.png", width=6.125, height=3.5, units="in")

# fit a linear model
lmFull <- lm(Inhibit ~ Surface*UVB, data=ozoneData)
summary(lmFull)$coefficients
confint(lmFull, level = 0.95)

# fit a linear model that forces the intercept to 0
lmZeroInt <- lm(Inhibit ~ 0 + Surface*UVB, data=ozoneData)
summary(lmZeroInt)$coefficients

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



# Problem 4: Ramsey 11.8
#########################################################################

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

# no code needed

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

# create a dataset with a high-leverage, low-influence obs
set.seed(1)
data_b <- data.frame(x=runif(n = 14, min = 0, max = 5)) %>%
  mutate(y= x + rnorm(n = 14, mean = 0, sd = 0.25))
data_b <- rbind(data_b, data.frame(x=7.5, y=7.55))

# plot
ggplot(data_b, aes(x=x, y=y)) + geom_point()
ggsave(filename="writeup/4b.png", width=6.125, height=3.5, units="in")

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

# create a dataset with a high-leverage, high-influence obs
set.seed(1)
data_c <- data.frame(x=runif(n = 14, min = 0, max = 5)) %>%
  mutate(y= x + rnorm(n = 14, mean = 0, sd = 0.25))
data_c <- rbind(data_c, data.frame(x=7.5, y=1.8))

# plot
ggplot(data_c, aes(x=x, y=y)) + geom_point()
ggsave(filename="writeup/4c.png", width=6.125, height=3.5, units="in")

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



# Problem 5: Ramsey 11.16
#########################################################################

# load data
```

```
fpmData <- Sleuth3::case1101

# fit a linear model of Metabol on Gastric*Sex
lmFpm <- lm(formula = Metabol ~ Gastric*Sex, data = fpmData)
summary(lmFpm)$coefficients

# plot of Metabol vs Gastric
ggplot(fpmData, aes(x=Gastric, y=Metabol, color=Sex, shape=Sex, label =
Subject)) +
  geom_point(size=2) +
  geom_text(data=filter(fpmData, Subject==32), nudge_x = 0.10, hjust = 0,
show.legend = FALSE) +
  theme(legend.position = c(0.85, 0.2))
ggsave(filename="writeup/5.png", width=6.125, height=3.5, units="in")

# calculate leverage for case 32
hatvalues(lmFpm)[32]

# calculate the studentized residual for case 32
studres(lmFpm)[32]

# calculate Cook's Distance for case 32
cooks.distance(lmFpm)[32]

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



# Problem 6: Ramsey 11.20
#########################################################################

# load data
dinoData <- Sleuth3::ex1120

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

# regression of Calcite on Carbonate with the full dataset
lmFull <- lm(formula = Calcite ~ Carbonate, data=dinoData)
summary(lmFull)

# regression of Calcite on Carbonate, excluding the smallest X
lmExcl1 <- lm(formula = Calcite ~ Carbonate,
              data=dinoData, subset = Carbonate > 22)
summary(lmExcl1)

# regression of Calcite on Carbonate, excluding the smallest two X's
lmExcl2 <- lm(formula = Calcite ~ Carbonate,
              data=dinoData, subset = Carbonate > 24)
summary(lmExcl2)

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

# compare total and residual sum of squares for lmFull
sum(anova(lmFull)$"Sum Sq")
(anova(lmFull)$"Sum Sq")[2]
```

```
# compare total and residual sum of squares for lmExcl1
sum(anova(lmExcl1)$"Sum Sq")
(anova(lmExcl1)$"Sum Sq")[2]

# compare total and residual sum of squares for lmExcl2
sum(anova(lmExcl2)$"Sum Sq")
(anova(lmExcl2)$"Sum Sq")[2]

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

# compute the case influence statistics on the full dataset
dinoData <- dinoData %>%
  mutate(leverageFull=hatvalues(lmFull),
         studResFull=studres(lmFull),
         cooksDistFull=cooks.distance(lmFull)) %>%
  mutate(sample=row_number())

# plot the case influence statistics on the full dataset
plot_leverageFull <- ggplot(dinoData, aes(x=sample, y=leverageFull)) +
  geom_point() +
  geom_hline(yintercept=(2*2/nrow(dinoData)), linetype="dotted") +
  labs(x="Sample Number", y="Leverage") +
  xlim(1, nrow(dinoData))
plot_studResFull <- ggplot(dinoData, aes(x=sample, y=studResFull)) +
  geom_point() +
  geom_hline(yintercept=c(-2, 2), linetype="dotted") +
  labs(x="Sample Number", y="Studentized Residual") +
  xlim(1, nrow(dinoData))
plot_cooksDistFull <- ggplot(dinoData, aes(x=sample, y=cooksDistFull)) +
  geom_point() +
  geom_hline(yintercept=1, linetype="dotted") +
  labs(x="Sample Number", y="Cook's Distance") +
  xlim(1, nrow(dinoData))
plot_Full <- grid.arrange(plot_leverageFull, plot_studResFull,
plot_cooksDistFull, nrow=3, ncol=1)
ggsave(filename="writeup/6c.png", plot=plot_Full, width=8, height=6,
units="in")

# scatterplot of Calcite on Carbonate
ggplot(dinoData, aes(x=Carbonate, y=Calcite)) +
  geom_point() +
  geom_text_repel(aes(label=sample), size=3.5)
ggsave(filename="writeup/6c_scatter.png", width=6.125, height=3.5, units="in")

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

# compute the case influence statistics, excluding the smallest X
dinoDataExcl1 <- dinoData %>%
  select(Carbonate, Calcite, sample) %>%
  filter(sample != 1) %>%
  mutate(leverageExcl1=hatvalues(lmExcl1),
         studResExcl1=studres(lmExcl1),
         cooksDistExcl1=cooks.distance(lmExcl1))
```

```
# plot the case influence statistics, excluding the smallest X
plot_leverageExcl1 <- ggplot(dinoDataExcl1, aes(x=sample, y=leverageExcl1)) +
  geom_point() +
  geom_hline(yintercept=(2*2/nrow(dinoDataExcl1)), linetype="dotted") +
  labs(x="Sample Number", y="Leverage") +
  xlim(1, 1+nrow(dinoDataExcl1))
plot_studResExcl1 <- ggplot(dinoDataExcl1, aes(x=sample, y=studResExcl1)) +
  geom_point() +
  geom_hline(yintercept=c(-2, 2), linetype="dotted") +
  labs(x="Sample Number", y="Studentized Residual") +
  xlim(1, 1+nrow(dinoDataExcl1))
plot_cooksDistExcl1 <- ggplot(dinoDataExcl1, aes(x=sample, y=cooksDistExcl1))
+
  geom_point() +
  geom_hline(yintercept=1, linetype="dotted") +
  labs(x="Sample Number", y="Cook's Distance") +
  xlim(1, 1+nrow(dinoDataExcl1))
plot_Excl1 <- grid.arrange(plot_leverageExcl1, plot_studResExcl1,
plot_cooksDistExcl1, nrow=3, ncol=1)
ggsave(filename="writeup/6d.png", plot=plot_Excl1, width=8, height=6,
units="in")

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