HW0_Final.R

James

2019-04-13

```
setwd("~/GitHub/MMSS_311_2")

# Q1a) A vector with the numbers 1-5 in order
a <- c(1,2,3,4,5)

# Q1b) A scalar named Mindy that takes the value 12
Mindy <- 12

# Q1c) A 2×3 matrix with the numbers 1-6 in order by rows
b <- c(1,2,3,4,5,6)
matrix(b,2,3,TRUE)
```

```
##      [,1] [,2] [,3]
## [1,]    1    2    3
## [2,]    4    5    6
```

```
# Q1d) A 2×3 matrix with the numbers 1-6 in order by columns
matrix(b,2,3)
```

```
##      [,1] [,2] [,3]
## [1,]    1    3    5
## [2,]    2    4    6
```

```
# Q1e) A 10×10 matrix of 1's
matrix(1,10,10)
```

```
##       [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
##  [1,]    1    1    1    1    1    1    1    1    1     1
##  [2,]    1    1    1    1    1    1    1    1    1     1
##  [3,]    1    1    1    1    1    1    1    1    1     1
##  [4,]    1    1    1    1    1    1    1    1    1     1
##  [5,]    1    1    1    1    1    1    1    1    1     1
##  [6,]    1    1    1    1    1    1    1    1    1     1
##  [7,]    1    1    1    1    1    1    1    1    1     1
##  [8,]    1    1    1    1    1    1    1    1    1     1
##  [9,]    1    1    1    1    1    1    1    1    1     1
## [10,]    1    1    1    1    1    1    1    1    1     1
```

```r
# Q1f)  A vector consisting of the words THIS, IS, A, VECTOR (each word a separate element)
wordvec <- c("THIS", "IS", "A", "VECTOR")

# Q1g) A function that takes the sum of any three numbers
sum_of_three_numbers <- function(x,y,z) {
  x+y+z
}

# Q1h) A function that takes one number as input, returns "Yes" if the number is less than or eq
ual to 10 and "No" if the number is greater than 10
check <- function(x) {
  if (x<=10) {
    result <- "Yes"
  }
  else if (x>10) {
    result <- "No"
  }
  return(result)
}
check(9)
```

```
## [1] "Yes"
```

```r
# Q1i) Generate synthetic data by taking 1,000 draws from a normal distribution with a mean of 1
0 and a standard deviation of 1. Save these data to an object g.
g <- rnorm(1000,10,1)

# Q1j) Create a separate object called y with 1,000 draws from a normal distribution with a mean
of 5 and a standard deviation of 0.5.
y <- rnorm(1000,5,0.5)

# Q1k)Generate a variable x with 1,000 values, where each value is a mean of 10 samples from g,
 with replacement. (Hint: use a for loop)
x = NULL
for(i in 1:1000) {
  x [i] <- mean(sample(g, 10, TRUE))
}

# Q1)l Estimate a simple bivariate regression y on x and print your results. What do your result
s show?
# The results show that the OLS estimator of the coefficient of x is 0.02633, which is very smal
l. This shows that there is only a weak positive correlation between y and x.
reg <- lm(y ~ x)
print(reg)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Coefficients:
## (Intercept)            x
##     4.90030      0.01247
```

```
#Q2a Create an R script file that sets your working directory and loads the data.
setwd("~/GitHub/MMSS_311_2")

pums_chicago <- read.csv("pums_chicago.csv")

#2b How many variables are there in the dataset?
# There are 204 variables (from environment panel)

#2c What is the mean annual income, PINCP in this dataset?
PINCP_mean <- mean(pums_chicago$PINCP, na.rm = TRUE)

#2d Create a new variable in the PUMS dataframe called PINCP_LOG that is equal to the log of ann
ual income. Were NaN values produced? Why?
#NaN values were produced because we cannot take log of 0, which is the value of some annual inc
ome observations.
pums_chicago$PINCP_LOG <- log(pums_chicago$PINCP)
```

```
## Warning in log(pums_chicago$PINCP): NaNs produced
```

```
#2e Create a new variable GRAD.DUMMY that takes the value “grad” if the respondent has any post-
high school education, and “no grad” otherwise. Use the SCHL variable.
pums_chicago$GRAD.DUMMY <- ifelse(pums_chicago$SCHL > 17, "grad", "no grad")

#2f Drop the variable SERIALNO from the dataset.
pums_chicago$SERIALNO <- NULL

#2g Save your new dataset to a csv file in the working directory.
write.csv(pums_chicago,'editedPUMS_CHICAGO.csv')

#2h Use the variable ESR, create 5 new dataframes: under 16, employed, unemployed, in the armed
 forces, and not in the labor force.
under16 <- pums_chicago[pums_chicago$ESR == "NA", ]
employed <- pums_chicago[pums_chicago$ESR %in% c("1", "2"), ]
unemployed <- pums_chicago[pums_chicago$ESR %in% "3", ]
armedforces <- pums_chicago[pums_chicago$ESR %in% c("4", "5"), ]
notinlaborforce <- pums_chicago[pums_chicago$ESR %in% "6", ]

#2i Create a new dataframe that combines employed people and people in the armed forces.
employed_af <- pums_chicago[pums_chicago$ESR %in% c("1", "2", "4", "5"), ]

#2j In your new employed_af dataframe, keep only the variables AGEP, RAC1P, and PINCP_LOG
new_employed_af <- pums_chicago[c("AGEP", "RAC1P", "PINCP_LOG")]

#2ki Find the mean, median, and 80th percentile of travel time to work, JWMNP
summary(pums_chicago$JWMNP)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
##    1.00   20.00   30.00   34.84   45.00  149.00   27668
```

```
quantile(pums_chicago$JWMNP, probs=0.8, na.rm=TRUE)
```

```
## 80% 
##  45
```

```
#2kii Find the correlation between travel time to work JWMNP and annual wages WAGP
cor(pums_chicago$JWMNP, pums_chicago$WAGP, use="complete.obs")
```

```
## [1] -0.04205232
```

```
#2kiii Make a scatterplot of age and log income.
#2kiv Export this graph to your working directory in pdf format.
pdf("ageonLogincome.pdf")
plot(pums_chicago$AGEP, pums_chicago$PINCP_LOG, main="age on log income")
dev.off()
```

```
## png 
##   2
```

```
#2kv Create a crosstab of employment status ESR by race RAC1P
install.packages("gmodels", repos = "http://cran.us.r-project.org")
```

```
## Installing package into 'C:/Users/James/Documents/R/win-library/3.5'
## (as 'lib' is unspecified)
```

```
## package 'gmodels' successfully unpacked and MD5 sums checked
##
## The downloaded binary packages are in
##  C:\Users\James\AppData\Local\Temp\RtmpgvKp2C\downloaded_packages
```

```
library("gmodels")
CrossTable(pums_chicago$ESR, pums_chicago$RAC1P)
```

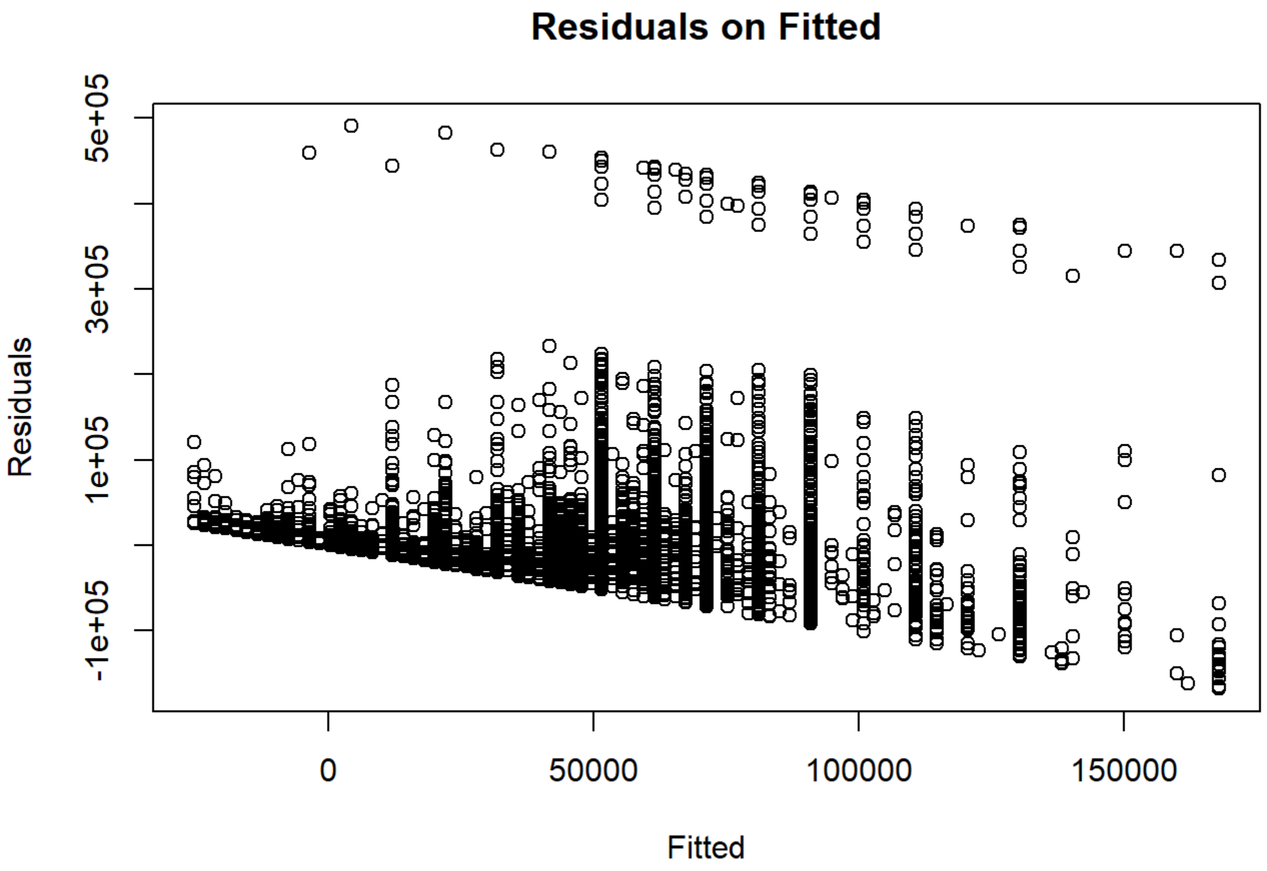
```
## 
## 
##    Cell Contents
## |-------------------------|
## |                       N |
## | Chi-square contribution |
## |           N / Row Total |
## |           N / Col Total |
## |         N / Table Total |
## |-------------------------|
## 
## 
## Total Observations in Table:  40348 
## 
## 
##                  | pums_chicago$RAC1P 
## pums_chicago$ESR |         1 |         2 |         3 |         4 |         5 |         6 | 
7 |         8 |         9 | Row Total | 
## -----------------|-----------|-----------|-----------|-----------|-----------|-----------|---
--------|-----------|-----------|-----------|
##                1 |     12870 |      5786 |        36 |         0 |        24 |      1746 | 
7 |      2502 |       521 |     23492 | 
##                  |   195.313 |   406.565 |     0.689 |     1.164 |     0.414 |     9.555 | 
0.591 |     4.491 |     3.239 |           | 
##                  |     0.548 |     0.246 |     0.002 |     0.000 |     0.001 |     0.074 | 
0.000 |     0.107 |     0.022 |     0.582 | 
##                  |     0.659 |     0.447 |     0.507 |     0.000 |     0.511 |     0.627 | 
0.778 |     0.607 |     0.630 |           | 
##                  |     0.319 |     0.143 |     0.001 |     0.000 |     0.001 |     0.043 | 
0.000 |     0.062 |     0.013 |           | 
## -----------------|-----------|-----------|-----------|-----------|-----------|-----------|---
--------|-----------|-----------|-----------|
##                2 |       258 |       147 |         0 |         0 |         0 |        31 | 
0 |        66 |         8 |       510 | 
##                  |     0.487 |     1.687 |     0.897 |     0.025 |     0.594 |     0.502 | 
0.114 |     3.730 |     0.576 |           | 
##                  |     0.506 |     0.288 |     0.000 |     0.000 |     0.000 |     0.061 | 
0.000 |     0.129 |     0.016 |     0.013 | 
##                  |     0.013 |     0.011 |     0.000 |     0.000 |     0.000 |     0.011 | 
0.000 |     0.016 |     0.010 |           | 
##                  |     0.006 |     0.004 |     0.000 |     0.000 |     0.000 |     0.001 | 
0.000 |     0.002 |     0.000 |           | 
## -----------------|-----------|-----------|-----------|-----------|-----------|-----------|---
--------|-----------|-----------|-----------|
##                3 |       794 |      1473 |         2 |         0 |         4 |       109 | 
0 |       268 |        57 |      2707 | 
##                  |   204.029 |   420.880 |     1.603 |     0.134 |     0.227 |    32.435 | 
0.604 |     0.252 |     0.041 |           | 
##                  |     0.293 |     0.544 |     0.001 |     0.000 |     0.001 |     0.040 | 
0.000 |     0.099 |     0.021 |     0.067 | 
##                  |     0.041 |     0.114 |     0.028 |     0.000 |     0.085 |     0.039 | 
0.000 |     0.065 |     0.069 |           | 
##                  |     0.020 |     0.037 |     0.000 |     0.000 |     0.000 |     0.003 | 
0.000 |     0.007 |     0.001 |           |
## -----------------|-----------|-----------|-----------|-----------|-----------|-----------|---
--------|-----------|-----------|-----------|
##                4 |         4 |         5 |         0 |         0 |         0 |         0 |
1 |         0 |         1 |        11 |
##                  |     0.331 |     0.613 |     0.019 |     0.001 |     0.013 |     0.759 |
405.558 |     1.123 |     2.661 |           |
##                  |     0.364 |     0.455 |     0.000 |     0.000 |     0.000 |     0.000 |
0.091 |     0.000 |     0.091 |     0.000 |
##                  |     0.000 |     0.000 |     0.000 |     0.000 |     0.000 |     0.000 |
0.111 |     0.000 |     0.001 |           |
##                  |     0.000 |     0.000 |     0.000 |     0.000 |     0.000 |     0.000 |
0.000 |     0.000 |     0.000 |           |
## -----------------|-----------|-----------|-----------|-----------|-----------|-----------|---
--------|-----------|-----------|-----------|
##                6 |      5618 |      5533 |        33 |         2 |        19 |       899 |
1 |      1283 |       240 |     13628 |
##                  |   146.443 |   308.317 |     3.392 |     2.597 |     0.615 |     1.846 |
1.369 |     8.421 |     5.537 |           |
##                  |     0.412 |     0.406 |     0.002 |     0.000 |     0.001 |     0.066 |
0.000 |     0.094 |     0.018 |     0.338 |
##                  |     0.287 |     0.427 |     0.465 |     1.000 |     0.404 |     0.323 |
0.111 |     0.311 |     0.290 |           |
##                  |     0.139 |     0.137 |     0.001 |     0.000 |     0.000 |     0.022 |
0.000 |     0.032 |     0.006 |           |
## -----------------|-----------|-----------|-----------|-----------|-----------|-----------|---
--------|-----------|-----------|-----------|
##     Column Total |     19544 |     12944 |        71 |         2 |        47 |      2785 |
9 |      4119 |       827 |     40348 |
##                  |     0.484 |     0.321 |     0.002 |     0.000 |     0.001 |     0.069 |
0.000 |     0.102 |     0.020 |           |
## -----------------|-----------|-----------|-----------|-----------|-----------|-----------|---
--------|-----------|-----------|-----------|
## 
## 
```

```
#2kvi Estimate a linear regression of annual wages WAGP on hours worked per week WKHP
wagp_on_wkhp <- lm(WAGP ~ WKHP, pums_chicago)

#2kvii Plot the residuals from this regression against the fitted values. What does this show?
#This shows that residuals tend to decrease as the fitted values increase. Furthermore, we can o
bserve that there are only a small number of large deviations for any given fitted value.
wagp_on_wkhp_res <- resid(wagp_on_wkhp)
wagp_on_wkhp_fitted <- fitted(wagp_on_wkhp)
plot(wagp_on_wkhp_fitted, wagp_on_wkhp_res,
     ylab = "Residuals", xlab="Fitted",
     main = "Residuals on Fitted")

#2li Estimate a linear regression of miles per gallon on weight
data(mtcars)
car_lm <- lm(mpg ~ wt, mtcars)

#2lii Estimate this regression separately for manual versus automatic transition
autoData <- mtcars[mtcars$am == "0",]
manualData <- mtcars[mtcars$am == "1",]
autocar_lm <- lm(mpg ~ wt, mtcars)
manualcar_lm <- lm(mpg ~ wt, mtcars)

#2liii Estimate a regression of miles per gallon on the log of horsepower.
mtcars$log.hp <- log(mtcars$hp)
mpg_on_lg.hp <- lm(mpg ~ log.hp, mtcars)

#2mi Make a scatterplot of weight against miles per gallon.
#2mii Color the points in your graph according to the transmission of the vehicle.
#2miii Change the shape of the points to correspond to the number of forward gears in the vehicl
e.
#2miv Change the x and y labels on the plot to make full words.
#2mv Change the background of the plot so that the panel background is not gray.
install.packages("ggplot2", repos = "http://cran.us.r-project.org")
```

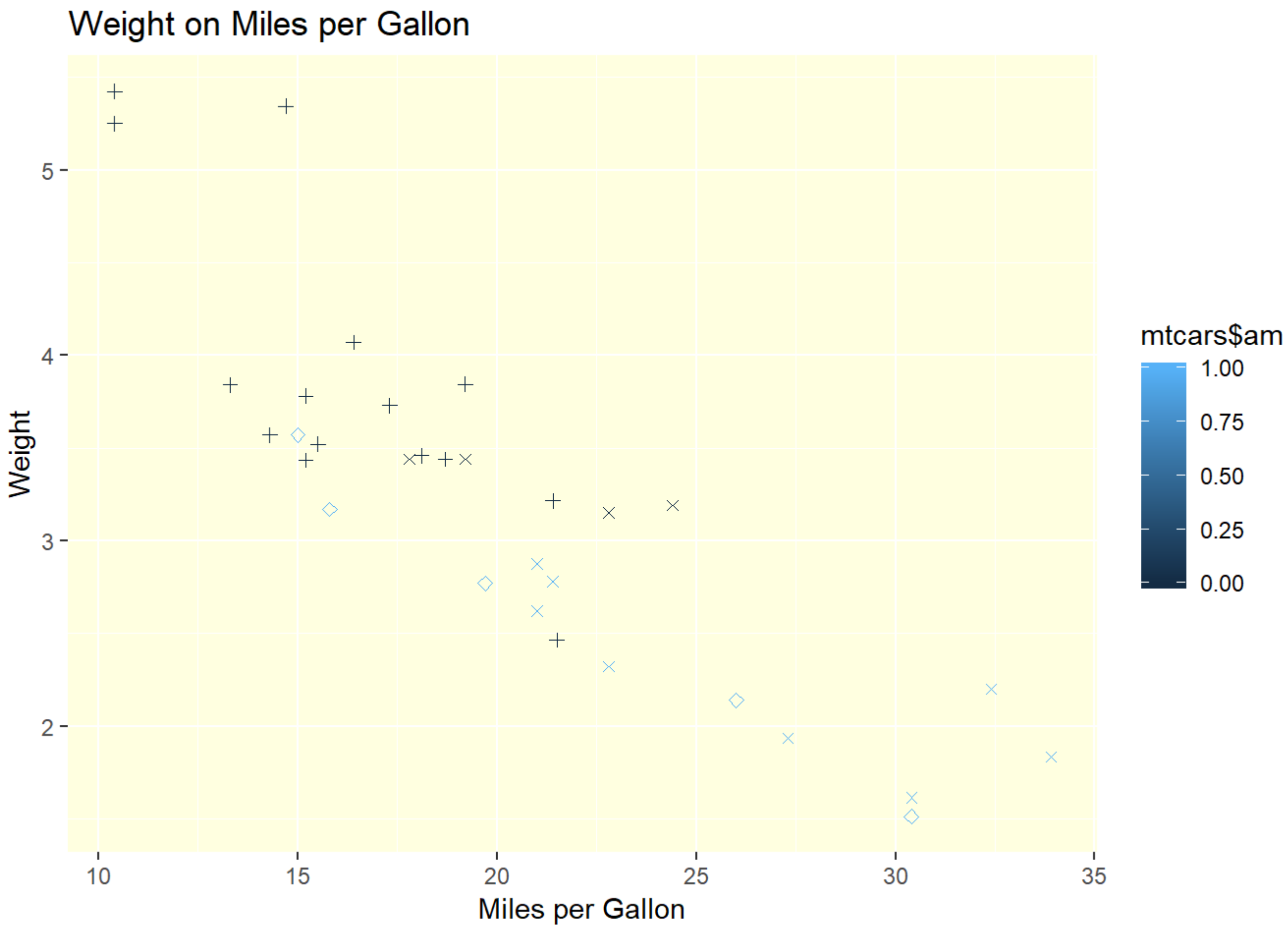
```
## Installing package into 'C:/Users/James/Documents/R/win-library/3.5'
## (as 'lib' is unspecified)
```

```
## package 'ggplot2' successfully unpacked and MD5 sums checked
##
## The downloaded binary packages are in
##  C:\Users\James\AppData\Local\Temp\RtmpgvKp2C\downloaded_packages
```

```
library(ggplot2)
```



```
ggplot(mtcars)+ geom_point(mapping = aes(x = mpg, y = wt,
                                        color = mtcars$am,
                                        shape = mtcars$gear)) + scale_shape_identity() +
  labs(title = "Weight on Miles per Gallon", x = "Miles per Gallon", y = "Weight") +
  theme(panel.background = element_rect(fill = "lightyellow"))
```


Weight on Miles per Gallon
Weight
5
4
3
2
Miles per Gallon
10
15
20
25
30
35
mtcars$am
1.00
0.75
0.50
0.25
0.00