Practical Data Analysis Homework 1

Peter Shewmaker

September 19, 2019

Problem 1

The dataset mcalindon_Big.csv loaded on the Canvas site in the Datasets folder contains information on individuals who were involved in a clinical trial that measured their pain on 7 different days over the course of several weeks together with local weather information for each person. This is described in paper #8 on the syllabus.Construct a dataset which contains just the first observation for each individual (i.e., you should have as many rows as people)

Hint: use the rle function to determine the unique id numbers and the number of rows associated with each id. Then use functions like cumsum to construct the starting and ending row numbers for each individual. This will then allow you to pull off the first row for each person.

First, we read in the data set and observe how many unique ID numbers there are in the data by using the "rle" function. The "rle" function determines how many times equal values are repeated in a vector, and what those values are. This informs us not only of all the unique ID numbers, but also how many rows lie between the first usage of each unique ID.

From this we can obtain the row numbers of the first row associated with each ID: starting at row 1, we add the number of times that the first unique ID was repeated - this is the first row that the second unique ID was first used, and adding the number of times the second unique ID was repeated to this row number produces the first row that the third unique ID was first used. We can use the "cumsum" function to create a vector with all the row numbers, this function returns the cumulative sums of a vector, which is exactly what we need to perform the described process. Notice here that we do not need to know how many times the last ID is repeated, and thus that value is removed from the vector before the "cumsum" function is applied. Code for this process can be found at the end of the document.

a. Summarize the average pain score for each of the 7 days of the study.

The columns that contain a pain score can be selected out of the data frame using the "grepl" function since they all have the string "pain" in their column name. Once selected out, the average of each column can be calculated with the "colMeans" function. Since there are missing values, the argument "na.rm" must be set to TRUE. Then the "summary" function is called on the vector containing the average pain scores.

The average of the mean pain scores for each of the 7 days of the study is 7.451, and participants ranked their pain the highest on the first day, and the lowest on the last day.

```
##   pain.1   pain.2   pain.3   pain.4   pain.5   pain.6   pain.7
## 9.060458 8.004423 7.472159 7.277350 6.796253 7.028429 6.515129
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   6.515   6.912   7.277   7.451   7.738   9.060
```

b. Regress each pain score on age and use the summary function to create a summary table for each regression. Then find the 95% confidence interval for the regression slopes (e.g. use the confint function) and produce a table with the estimates, standard errors, p-values and confidence intervals of the 7 slopes and put these in a single table.

```
##           Estimate  Std. Error     Pr(>|t|)    CI_lower    CI_upper
## pain.1 -0.01762846 0.002918184 1.567778e-09 -0.02334844 -0.01190847
## pain.2 -0.03164871 0.003482514 1.165873e-19 -0.03847499 -0.02482243
## pain.3 -0.03814594 0.003825451 2.484881e-23 -0.04564442 -0.03064746
## pain.4 -0.02785784 0.003485622 1.432903e-15 -0.03469015 -0.02102554
## pain.5 -0.05828215 0.003560597 1.303944e-59 -0.06526145 -0.05130284
## pain.6 -0.03095032 0.003547161 2.979754e-18 -0.03790329 -0.02399735
## pain.7 -0.05485087 0.003920303 3.853399e-44 -0.06253529 -0.04716646
```

Each of the estimates for the slope of the regressions is negative, showing that as age increases, pain scores tend to decrease. The *p*-values are all quite close to zero, allowing us to reject the null hypothesis that the slope equals zero when $\alpha = 0.05$. This can also be seen in the 95% confidence intervals, since each of the intervals contain only negative values and thus do not contain zero.

c. For each individual fit a regression of pain on time. Summarize the slopes and intercepts produced.

The following table is the result of calling the "summary" function on a table containing the intercepts and slopes for each of the regressions.

```
##  (Intercept) Estimate (Intercept) Std. Error (Intercept) t value
##  Min.   :-36.505       Min.   : 0.000         Min.   :-3.000e+00
##  1st Qu.:  6.721       1st Qu.: 1.508         1st Qu.: 2.000e+00
##  Median : 10.577       Median : 2.435         Median : 5.000e+00
##  Mean   : 11.732       Mean   : 4.184         Mean   : 7.306e+13
##  3rd Qu.: 15.114       3rd Qu.: 4.942         3rd Qu.: 7.000e+00
##  Max.   : 48.475       Max.   :46.022         Max.   : 6.990e+15
##                        NA's   :12             NA's   :12
##  (Intercept) p value Slope Estimate      Slope Std. Error
##  Min.   :0.000000    Min.   :-0.144068   Min.   :0.00000
##  1st Qu.:0.001992    1st Qu.:-0.057677   1st Qu.:0.01474
##  Median :0.014797    Median :-0.024902   Median :0.02057
##  Mean   :0.126215    Mean   :-0.024204   Mean   :0.02540
##  3rd Qu.:0.135076    3rd Qu.: 0.002473   3rd Qu.:0.02940
##  Max.   :0.974124    Max.   : 0.214286   Max.   :0.12372
##  NA's   :12          NA's   :5           NA's   :12
##  Slope t value       Slope p value
##  Min.   :-50.22947   Min.   :0.000078
##  1st Qu.: -2.83726   1st Qu.:0.048253
##  Median : -1.44024   Median :0.156984
##  Mean   : -1.75528   Mean   :0.302892
##  3rd Qu.:  0.08248   3rd Qu.:0.560201
##  Max.   :  7.87109   Max.   :1.000000
##  NA's   :12          NA's   :12
```

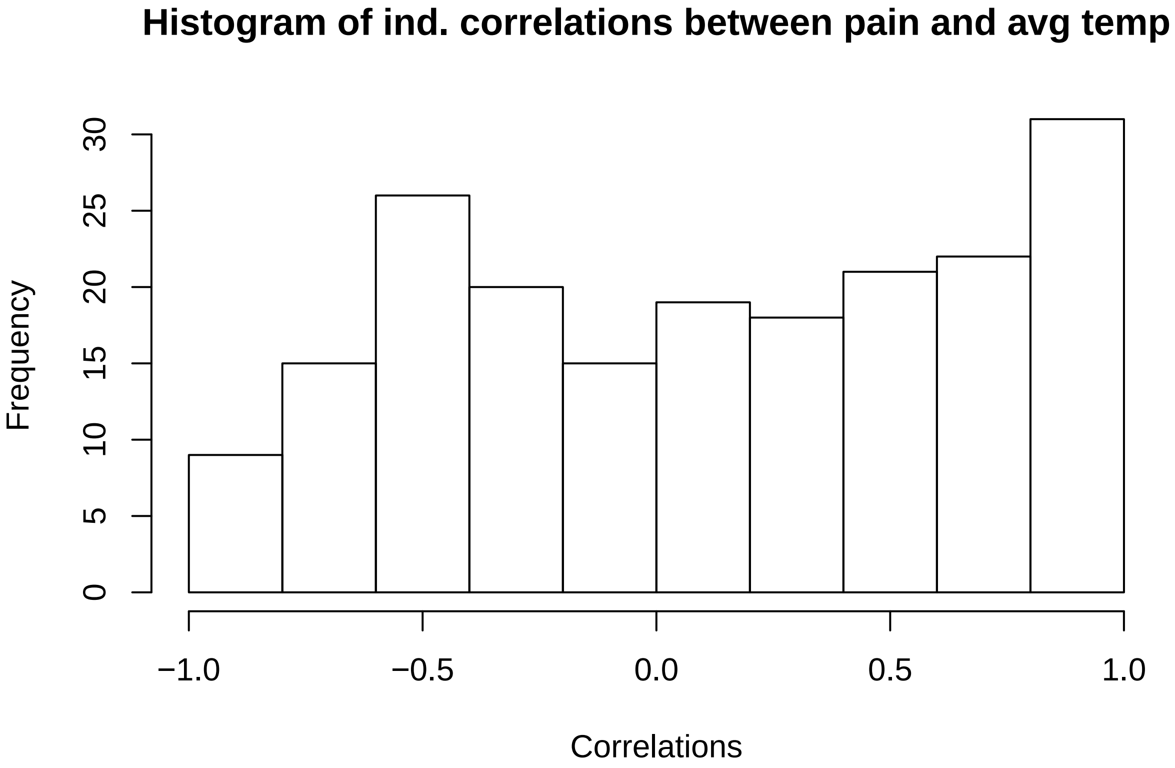
In the above regressions, the intercept represents what the pain score would be at time equal to zero. Notice here that the results show that the median intercept is 10.577 (the maximum score for pain is a 10) and that the median slope is -0.024902, implying that for the median patient pain decreases over time - decreasing by -0.024902 for each additional day.

Since the 3rd quartile for slope is positive, the upper quartile of patients had pain increase over time. However, note here that the median p value for the slope is 0.157, showing that for at least half of the regressions, the null hypothesis that time is unrelated to pain cannot be rejected even for relatively high levels of α.

d. Are the slopes or intercepts related to any of the patient characteristics (age, race, income, treatment, sex, occupation, working status, use of NSAIDs,)?

e. Use the whole database to compute the correlation for each individual between their pain scores and the average temperature on the dates the pain scores were taken. Construct a graph to display these correlations. Discuss whether pain is correlated with temperature.

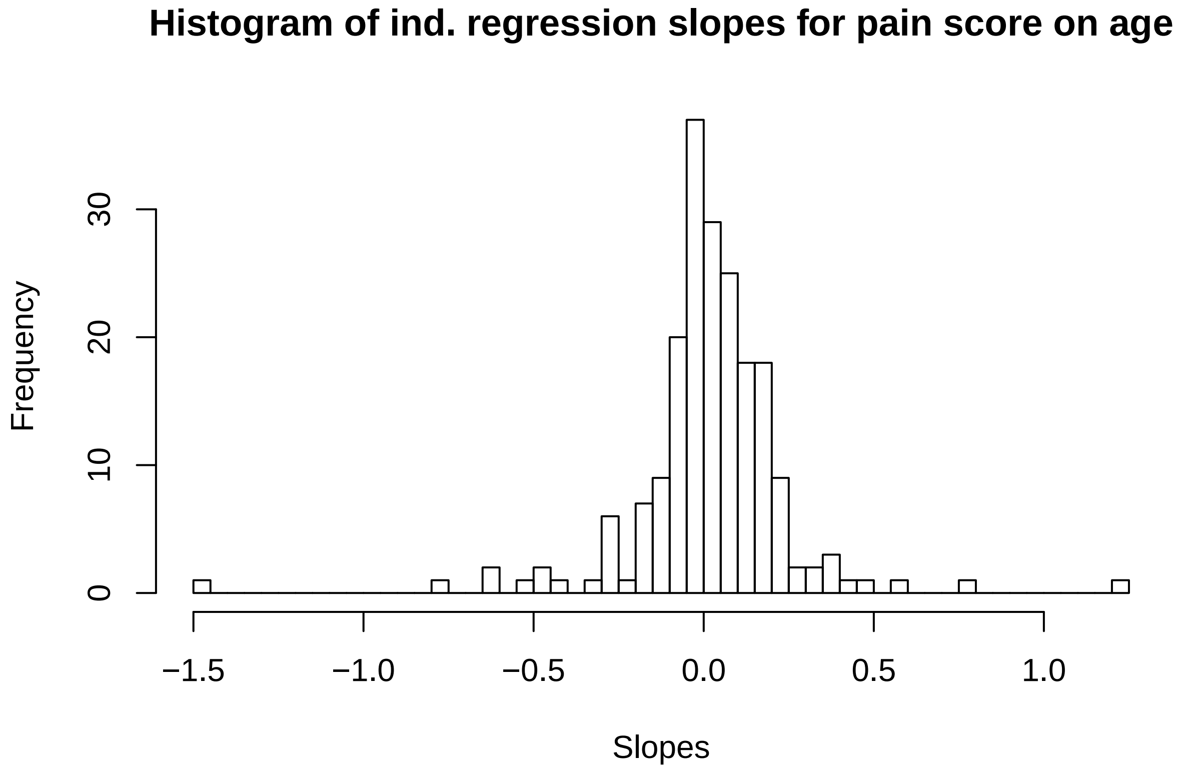
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
## -1.0000 -0.4016  0.1501  0.1181  0.6196  1.0000       9
```



Note that the correlations were calculated with "use = 'complete.obs' " to eliminate issues with missing values. From the above histogram, we can see that the correlation values are distributed close to uniformly over the interval [-1,1], since the expected value for each box would be close to 20 (196 participants in 10 boxes). Although the largest box is for values close to 1 and the smallest box is for values close to -1, many of the other boxes occur with a frequency near 20. From these results it would be difficult to point to an obvious pattern in the correlations between pain and average temperature.

f. Express these correlations as regressions and describe the slopes and intercepts as in problem (c). What do you notice about the distributions of the correlations in (e) and the slopes in (f)?

```
##    Intercepts            Slopes        
##  Min.   :-92.58333   Min.   :-1.50000  
##  1st Qu.: -0.04503   1st Qu.:-0.05254  
##  Median :  6.47043   Median : 0.02169  
##  Mean   :  6.40801   Mean   : 0.01649
##  3rd Qu.: 10.77769   3rd Qu.: 0.11226
##  Max.   :109.50000   Max.   : 1.25000
##                      NA's   :5
```



The median intercept of the regressions is 6.47043, and the median slope of the regressions is 0.02169, so that for the median patient each increase of 1 degree of average temperature implies an increase in pain score of 0.02169. However, the histogram of regression slopes appears to be normally distributed and centered close to zero. Compare this to the histogram of correlations between the two variables, which showed a near uniform distribution over [-1,1]. From this we might conclude that these two variables are not highly associated with each other, since if they were the average slope would lie further from zero, and the correlations would be grouped more around a non-zero value.

g. For each individual, do the multiple regression of pain on both time and temperature. What can you conclude about potential confounding between time and temperature?

```
##  (Intercept) Estimate (Intercept) Std. Error (Intercept) t value
##  Min.   :-342.896     Min.   : 0.000         Min.   :-6.000e+00
##  1st Qu.:   3.693     1st Qu.: 5.215         1st Qu.: 0.000e+00
##  Median :  11.179     Median : 8.325         Median : 1.000e+00
##  Mean   :  11.243     Mean   :12.172         Mean   : 1.295e+13
##  3rd Qu.:  20.996     3rd Qu.:15.574         3rd Qu.: 3.000e+00
##  Max.   : 151.071     Max.   :64.893         Max.   : 1.247e+15
##                       NA's   :24             NA's   :24
##  (Intercept) p value Temp Coeff Estimate Temp Coeff Std. Error
##  Min.   :0.00000     Min.   :-1.500000   Min.   :0.00000
##  1st Qu.:0.06659     1st Qu.:-0.097680   1st Qu.:0.05919
##  Median :0.23762     Median :-0.001584   Median :0.09944
##  Mean   :0.31792     Mean   : 0.007848   Mean   :0.15201
##  3rd Qu.:0.52005     3rd Qu.: 0.072839   3rd Qu.:0.17190
##  Max.   :0.97080     Max.   : 4.220779   Max.   :1.18433
##  NA's   :24          NA's   :5           NA's   :24
##  Temp Coeff t value  Temp Coeff p value Time Coeff Estimate
##  Min.   :-11.96478   Min.   :0.0029     Min.   :-0.678571
##  1st Qu.: -0.86863   1st Qu.:0.2698     1st Qu.:-0.061237
##  Median : -0.05645   Median :0.5300     Median :-0.026808
##  Mean   :  0.03632   Mean   :0.5182     Mean   :-0.015041
##  3rd Qu.:  0.62829   3rd Qu.:0.7715     3rd Qu.: 0.007682
##  Max.   : 24.25934   Max.   :0.9999     Max.   : 1.636364
##  NA's   :24          NA's   :24         NA's   :12
##  Time Coeff Std. Error Time Coeff t value Time Coeff p value
##  Min.   :0.00000       Min.   :-31.2637   Min.   :0.000395
##  1st Qu.:0.02096       1st Qu.: -1.9488   1st Qu.:0.121649
##  Median :0.03237       Median : -0.9058   Median :0.362699
##  Mean   :0.04283       Mean   : -1.1954   Mean   :0.380702
##  3rd Qu.:0.05130       3rd Qu.:  0.1776   3rd Qu.:0.607937
##  Max.   :0.26123       Max.   : 15.1785   Max.   :0.989779
##  NA's   :24            NA's   :24         NA's   :24
```

Notice that from the summary table, the median values for both the time and temperature coefficients are close to zero. Further more, the 1st quartile p-values for both the time and temperature coefficients are far higher than 0.05, showing that the results of the regression are not statistically significant in over 3/4 of the cases. In the case where time was the only term in the regression, the first quartile was 0.014, so that at least a quarter of the regressions had statistically significant results. Thus we can conclude that it is likely that the significant results from the regressions based on time alone did not account for confounding caused by the temperature on the dates the pain scores were recorded.

Problem 2

	Tai Chi (n = 20)	Attention Control (n = 20)	Total (n = 40)
Demographics			
Women, no. (%)	16 (80)	14 (70)	30 (75)
Age, years	63.2 ± 8.1	67.5 ± 7	65.4 ± 7.8
White, no. (%)	14 (70)	14 (70)	28 (70)
Greater than or equal to high school education, no. (%)	20 (100)	19 (95)	39 (98)
Body mass index, kg/m	30 ± 5.2	29.8 ± 4.3	29.9 ± 4.8
Disease condition			
Duration of knee pain (on study knee), years	9.7 ± 7	9.7 ± 8.3	9.7 ± 7.6
Radiograph score, no. (%)			
K/L grade 2	4 (20)	3 (15)	7 (18)
K/L grade 3	7 (35)	3 (15)	10 (25)
K/L grade 4	9 (45)	14 (70)	23 (57)
Knee surgery, no. (%)	6 (30)	8 (40)	14 (35)
Patient VAS (range 0–10 cm)	4.2 ± 2.1	4.8 ± 2	4.5 ± 2
Physician VAS (study knee; range 0–10 cm)	4.7 ± 1.7	5.8 ± 2.2	5.3 ± 2
WOMAC pain (range 0–500 mm)	209.3 ± 58.5	220.3 ± 101	214.8 ± 81.7
WOMAC physical function (range 0–1,700 mm)	707.6 ± 246.9	827 ± 258.8	767.3 ± 256.9
WOMAC stiffness (range 0–200 mm)	37.5 ± 8.5	32 ± 8.8	34.8 ± 9

	Tai Chi (n = 20)	Attention Control (n = 20)	Total (n = 40)
Receiving NSAIDs prior to study, no. (%)	9 (45)	13 (65)	22 (55)
Receiving analgesics prior to study, no. (%)	4 (20)	6 (30)	10 (25)
Self-reported comorbidities, no. (%)			
Congestive heart disease	1 (5)	4 (20)	5 (12)
Hypertension	7 (35)	12 (60)	19 (48)
Diabetes mellitus	0 (0)	4 (20)	4 (10)
Health-related quality of life and others			
SF-36 PCS (range 0–100)	37.5 ± 8.5	32 ± 8.8	34.8 ± 9
SF-36 MCS (range 0–100)	51.4 ± 12.2	50.8 ± 12.6	51.1 ± 12.3
CES-D (range 0–60)	13.6 ± 11.7	9.3 ± 9.2	11.4 ± 10.6
Self-efficacy score (range 1–5)	3.1 ± 1.1	3.3 ± 1	3.2 ± 1
Physical performance			
6-minute walk test, yards	416.7 ± 95.2	407.4 ± 91	412 ± 92
Balance score (range 0–5)	4 ± 0.7	3.8 ± 0.8	3.9 ± 0.7
Chair stand score, seconds	40.8 ± 13.4	35.6 ± 9.2	38.3 ± 11.7