Homework 4

Yuki Joyama

2023-11-20

Problem 1

a) Let $d_1, d_2, ..., d_n$ be the differences between 25 pairs with and Δ be the median of d_i.
   $H_0 : \Delta \geq 0$
   $H_1 : \Delta < 0$

$n * p(1-p) \geq 5$ so I will apply normal-approximation to perform the one-sided sign test.
Let C be the number of negative differences, ignoring the zero differences; n* be the number of non-zero differences.
Now, C = 14 and n* = 24

The test statistics is:
$\frac{n^*}{2} + \frac{1}{2} + z_{1-\alpha}\sqrt{\frac{n^*}{4}} = 16.53 >$ C
p-value $= 1 - \Phi(\frac{C - \frac{n^*}{2} - \frac{1}{2}}{\sqrt{\frac{n^*}{4}}}) = 0.27$

Therefore, we fail to reject the null hypothesis. We do not have significant ($\alpha = 0.05$) evidence to support that the median sugar readings was less than 120.

b) H_0 : The median difference between blood sugar samples and 120 is equal to or greater than zero
   H_1 : The median difference between blood sugar samples and 120 is less than zero

In order to perform the Wilcoxon Signed-Rank Test (one-sided), I calculated the absolute differences between samples and 120 and their rank as follows.

```
bs = bs |>
  filter(sample != 120) |> # exclude difference = 0
  group_by(sample) |>
  mutate(
    d = sample - 120,
    abs_d = abs(d), # absolute differences
    positive_d = ifelse(d > 0, 1, 0),
    negative_d = ifelse(d < 0, 1, 0),
    same_n = n() # count numbers of same blood sugar samples
  ) |>
  ungroup() |>
  arrange(abs_d) |>
  mutate(
    rank = rank(abs_d) # assign average rank based on absolute differences
  ) |>
  print()
```

```
## # A tibble: 24 x 7
##    sample     d abs_d positive_d negative_d same_n  rank
##     <dbl> <dbl> <dbl>      <dbl>      <dbl>  <int> <dbl>
##  1    121     1     1          1          0      1   1
##  2    118    -2     2          0          1      4   4
##  3    118    -2     2          0          1      4   4
##  4    118    -2     2          0          1      4   4
##  5    122     2     2          1          0      1   4
##  6    118    -2     2          0          1      4   4
##  7    123     3     3          1          0      3   8.5
##  8    117    -3     3          0          1      1   8.5
##  9    123     3     3          1          0      3   8.5
## 10    123     3     3          1          0      3   8.5
## # i 14 more rows
```

Let R be the rank sum for negative differences.

R = 187.5
Since there are ties, the test statistics T is:
$T = \frac{|R - \frac{n^*(n^*+1)}{4}| - \frac{1}{2}}{\sqrt{(\frac{n^*(n^*+1)(2n^*+1)}{24} - \frac{\sum_{i=1}^{g}(t_i^3 - t_i)}{48})}} = 1.08 \sim$ N(0, 1) under H_0
p-value $= [1 - \Phi(T)] = 0.14$
Therefore, we failed to reject the null hypothesis and cannot conclude that there is a significant ($\alpha = 0.05$) evidence that median blood sugar reading was less than 120.

Problem 2

a)

```r
# exclude homo sapiens
df_brain_nonh = df_brain |>
  filter(species != "Homo sapiens")

# fit a regression model for the nonhuman data
reg_nonh = lm(glia_neuron_ratio ~ ln_brain_mass, df_brain_nonh)

reg_nonh |>
  broom::tidy() |>
  print()
```

```
## # A tibble: 2 x 5
##   term          estimate std.error statistic  p.value
##   <chr>            <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)      0.164    0.160       1.02 0.322
## 2 ln_brain_mass    0.181    0.0360      5.03 0.000151
```

b)

```r
# prediction intervals (95%)
predict(
  reg_nonh,
```

```r
  newdata = tibble(
    ln_brain_mass = df_brain |>
      filter(species == "Homo sapiens") |>
      pull(ln_brain_mass)
  ),
  interval = "prediction", level = 0.95
) |>
  round(3)
```

```
##     fit   lwr   upr
## 1 1.471 1.036 1.907
```

The predicted glia-neuron ratio for humans given the brain mass using the nonhuman primate relationship is 1.471.

c)

```r
# prediction intervals (95%)
predict(
  reg_nonh,
  newdata = tibble(
    ln_brain_mass = df_brain |>
      filter(species == "Homo sapiens") |>
      pull(ln_brain_mass)
  ),
  interval = "confidence", level = 0.95
) |>
  round(3)
```

```
##     fit  lwr   upr
## 1 1.471 1.23 1.713
```

The 95% prediction interval for the predicted human glia-neuron ratio given the brain mass is 1.036 - 1.907, and the 95% confidence interval is 1.230 - 1.713.
I would use prediction interval rather than confidence interval when it comes to prediction because the prediction interval is more conservative by accounting for both the uncertainty of estimating a value and the random variability of the sample.

d) Given the output in part (b), the 95% prediction interval is 1.036 - 1.907. The sample observation of human glia-neuron ratio is 1.65, which is within the range of the 95% prediction interval. Thus, using the regression model for nonhuman data, we can say that the human brain does not have an excessive glia-neuron ratio for its mass compared with other primates.

e) Because no other primates have brain mass as big as human, the regression model (based on primates' data) may not be able to accurately predict the `glia_neuron_ratio` with large `ln_brain_mass`.

Problem 3

a)

```
## # A tibble: 788 x 10
##       id totalcost   age gender interventions drugs e_rvisits complications
##    <dbl>     <dbl> <dbl>  <dbl>         <dbl> <dbl>     <dbl>         <dbl>
##  1     1      179.    63      0             2     1         4             0
##  2     2      319     59      0             2     0         6             0
##  3     3     9311.    62      0            17     0         2             0
##  4     4      281.    60      1             9     0         7             0
##  5     5    18727.    55      0             5     2         7             0
##  6     6      453.    66      0             1     0         3             0
##  7     7      323.    64      1             2     0         3             0
##  8     8     3874.    45      1             3     0         5             0
##  9     9     3244.    68      0             6     2         5             0
## 10    10      226.    64      1             3     0         2             0
## # i 778 more rows
## # i 2 more variables: comorbidities <dbl>, duration <dbl>
```

The data set consists of 10 variables and 788 observations.
The main outcome in this case is `totalcost` and the main predictor is `e_rvisits`. Other important covariates include `age`, `gender`, `complications`, and `duration`. (It is not specified but I will treat gender 0 as male, and 1 as female)

The descriptive statistics for all variables of interest is as follows.

Characteristic	**N = 788**[1]
Total cost (USD)	2,800.0 / 507.2 (6,690.3)
ER visits	3.4 / 3.0 (2.6)
Age	58.7 / 60.0 (6.8)
Female	180 (23%)
No. of complications	
0	745 (95%)
1	42 (5.3%)
3	1 (0.1%)
Duration of treatment condition (days)	164.0 / 165.5 (120.9)

[1]Mean / Median (SD); n (%)

b)

```r
# multiple linear regression model
reg_cost = lm(totalcost ~ e_rvisits + age + gender + complications + duration, data = df_hd)

reg_cost |>
  broom::tidy()
```

```
## # A tibble: 6 x 5
##   term          estimate std.error statistic  p.value
##   <chr>            <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)    4124.     1907.        2.16 3.09e- 2
```

```
## 2 e_rvisits       895.       84.4      10.6  1.30e-24
## 3 age             -93.8      32.4      -2.89 3.92e- 3
## 4 gender        -1053.      518.       -2.03 4.25e- 2
## 5 complications  3073.      886.        3.47 5.54e- 4
## 6 duration          7.21      1.83      3.95 8.47e- 5
```