p8130\_hw4

2022-11-13

library(tidyverse)

## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.2 ──  
## ✔ ggplot2 3.4.0 ✔ purrr 0.3.5   
## ✔ tibble 3.1.8 ✔ dplyr 1.0.10  
## ✔ tidyr 1.2.1 ✔ stringr 1.4.1   
## ✔ readr 2.1.3 ✔ forcats 0.5.2   
## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()

library(BSDA)

## Loading required package: lattice  
##   
## Attaching package: 'BSDA'  
##   
## The following object is masked from 'package:datasets':  
##   
## Orange

library(readxl)  
library(arsenal)  
library(knitr)

# Problem 1

blood\_data = c(125,123, 117, 123, 115, 112, 128 ,118, 124, 111, 116, 109 ,125,  
 120 ,113, 123, 112, 118, 121 ,118, 122, 115,  
 105, 118, 131)  
  
SIGN.test(blood\_data, md=120,alternative = "less")

##   
## One-sample Sign-Test  
##   
## data: blood\_data  
## s = 10, p-value = 0.2706  
## alternative hypothesis: true median is less than 120  
## 95 percent confidence interval:  
## -Inf 122.1203  
## sample estimates:  
## median of x   
## 118   
##   
## Achieved and Interpolated Confidence Intervals:   
##   
## Conf.Level L.E.pt U.E.pt  
## Lower Achieved CI 0.9461 -Inf 122.0000  
## Interpolated CI 0.9500 -Inf 122.1203  
## Upper Achieved CI 0.9784 -Inf 123.0000

wilcox.test(blood\_data, mu = 120, alternative = "less")

## Warning in wilcox.test.default(blood\_data, mu = 120, alternative = "less"):  
## cannot compute exact p-value with ties

## Warning in wilcox.test.default(blood\_data, mu = 120, alternative = "less"):  
## cannot compute exact p-value with zeroes

##   
## Wilcoxon signed rank test with continuity correction  
##   
## data: blood\_data  
## V = 112.5, p-value = 0.1447  
## alternative hypothesis: true location is less than 120

From the Sign test, the test statistic is 10, the p-value is 0.276, which is greater than 0.05. Therefore, we do not have significant evidence to reject the null hypothesis, there is no evidence that the blood sugar readings is less than 120.

From the Wilcoxon signed-rank test, the test statistic is 112.5, the p-value is 0.1447, which is greater than 0.05. Therefore, there is no significant evidence that the blood sugar level is less than 120.

# Problem 2

## a)

brain\_data= read\_excel("C:/Users/lenovo/Downloads/Brain.xlsx") %>%  
 janitor::clean\_names()   
  
  
non\_human\_reg= brain\_data %>%  
 filter(species!="Homo sapiens") %>%   
 lm(glia\_neuron\_ratio~ln\_brain\_mass,data= .)  
summary(non\_human\_reg)

##   
## Call:  
## lm(formula = glia\_neuron\_ratio ~ ln\_brain\_mass, data = .)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.24150 -0.12030 -0.01787 0.15940 0.25563   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.16370 0.15987 1.024 0.322093   
## ln\_brain\_mass 0.18113 0.03604 5.026 0.000151 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.1699 on 15 degrees of freedom  
## Multiple R-squared: 0.6274, Adjusted R-squared: 0.6025   
## F-statistic: 25.26 on 1 and 15 DF, p-value: 0.0001507

## b)

The relationship between glia-neuron ratio (denote as ) and brain mass (denote as ) is:

the glia-neuron ratio of Homo sapiens should be:

## c)

We find that the glia neuron ratio for human is 1.65, which is higher than other species. Therefore, the prediction interval interval for a single new observation is more appropriate since the value of glia neuron ratio for human can be considered as a new value. The predicted mean glia- neuron ratio at the given brain mass can only capture information of the given data.

## d)

new\_brain = brain\_data %>%   
 filter(species == "Homo sapiens")  
  
interval = as.tibble(  
 predict(non\_human\_reg, new\_brain, interval="predict"),  
) %>%   
 mutate(category = c("predict"))

## Warning: `as.tibble()` was deprecated in tibble 2.0.0.  
## ℹ Please use `as\_tibble()` instead.  
## ℹ The signature and semantics have changed, see `?as\_tibble`.

knitr::kable(interval)

| fit | lwr | upr | category |
| --- | --- | --- | --- |
| 1.471458 | 1.036047 | 1.906869 | predict |

brain\_data %>%   
 filter(species!="Homo sapiens") %>%   
 dplyr::select(glia\_neuron\_ratio) %>%   
 summary()

## glia\_neuron\_ratio  
## Min. :0.46   
## 1st Qu.:0.64   
## Median :1.02   
## Mean :0.94   
## 3rd Qu.:1.15   
## Max. :1.22

the true value of human brain after log transformation falls in the prediction interval of non-human distribution, thus human brain do not have excessive glia-neuron ratio for its mass

## e)

As seen from the plot, we can see that the glia neuron ration for human exceeds other specie’s ratio. So the prediction of human from this model may not be appropriate enough.

# Problem 3

heart\_data= read\_csv("C:/Users/lenovo/Downloads/HeartDisease.csv")%>%  
 janitor::clean\_names()

## Rows: 788 Columns: 10  
## ── Column specification ────────────────────────────────────────────────────────  
## Delimiter: ","  
## dbl (10): id, totalcost, age, gender, interventions, drugs, ERvisits, compli...  
##   
## ℹ Use `spec()` to retrieve the full column specification for this data.  
## ℹ Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

## a)

stat\_data =tableby( ~ totalcost + age + gender +  
 interventions + drugs+  
 e\_rvisits+complications+  
 comorbidities+duration,   
 data = heart\_data,  
 test = FALSE,   
 total = FALSE,  
 numeric.stats = c("meansd" ,"medianq1q3","range" ) )  
sum = summary(stat\_data,text = TRUE)  
sum

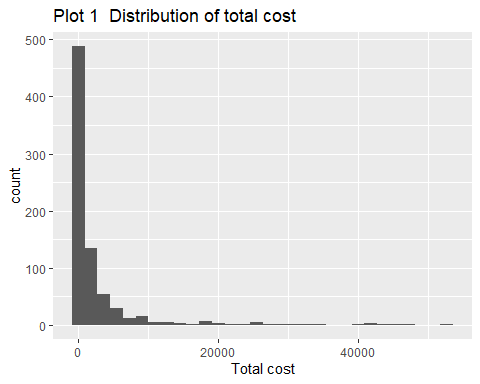
##   
##   
## | | Overall (N=788) |  
## |:------------------|:---------------------------:|  
## |totalcost | |  
## |- Mean (SD) | 2799.956 (6690.260) |  
## |- Median (Q1, Q3) | 507.200 (161.125, 1905.450) |  
## |- Range | 0.000 - 52664.900 |  
## |age | |  
## |- Mean (SD) | 58.718 (6.754) |  
## |- Median (Q1, Q3) | 60.000 (55.000, 64.000) |  
## |- Range | 24.000 - 70.000 |  
## |gender | |  
## |- Mean (SD) | 0.228 (0.420) |  
## |- Median (Q1, Q3) | 0.000 (0.000, 0.000) |  
## |- Range | 0.000 - 1.000 |  
## |interventions | |  
## |- Mean (SD) | 4.707 (5.595) |  
## |- Median (Q1, Q3) | 3.000 (1.000, 6.000) |  
## |- Range | 0.000 - 47.000 |  
## |drugs | |  
## |- Mean (SD) | 0.447 (1.064) |  
## |- Median (Q1, Q3) | 0.000 (0.000, 0.000) |  
## |- Range | 0.000 - 9.000 |  
## |e\_rvisits | |  
## |- Mean (SD) | 3.425 (2.637) |  
## |- Median (Q1, Q3) | 3.000 (2.000, 5.000) |  
## |- Range | 0.000 - 20.000 |  
## |complications | |  
## |- Mean (SD) | 0.057 (0.248) |  
## |- Median (Q1, Q3) | 0.000 (0.000, 0.000) |  
## |- Range | 0.000 - 3.000 |  
## |comorbidities | |  
## |- Mean (SD) | 3.766 (5.951) |  
## |- Median (Q1, Q3) | 1.000 (0.000, 5.000) |  
## |- Range | 0.000 - 60.000 |  
## |duration | |  
## |- Mean (SD) | 164.030 (120.916) |  
## |- Median (Q1, Q3) | 165.500 (41.750, 281.000) |  
## |- Range | 0.000 - 372.000 |

In this dataset, the main outcome is total cost. Other important covariate includes the age and gender, number of complications that happens during treatment, and duration of treatment condition.From the plot above, the possible important predictors are likely to be complications, drugs and ERvisits and interventions.

## b)

heart\_data %>%   
 ggplot(aes(totalcost))+  
 geom\_histogram()+  
 labs(x = "Total cost",  
 title = "Plot 1 Distribution of total cost")

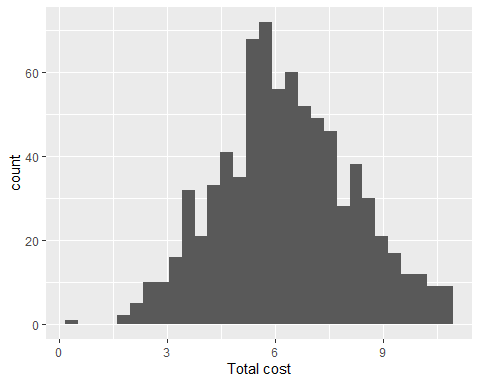
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



heart\_log =   
 heart\_data %>%   
 mutate(log\_totalcost = log(totalcost))  
  
plot =  
 heart\_log %>%   
 ggplot(aes(log\_totalcost))+  
 geom\_histogram()+  
 labs(x = "Total cost")  
plot

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

## Warning: Removed 3 rows containing non-finite values (`stat\_bin()`).

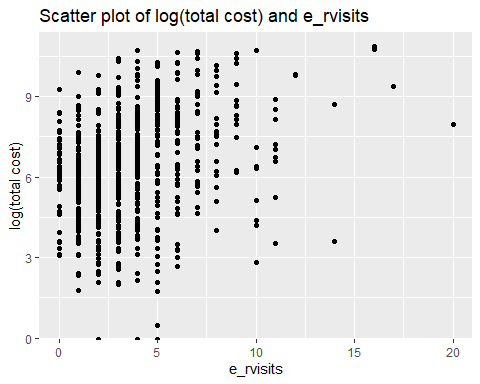


## c)

heart\_new =   
 heart\_log %>%   
 mutate(comp\_bin =   
 case\_when(  
 complications == 0 ~ "0",  
 complications != 0 ~ "1"))

## d)

heart\_new %>%   
 mutate(log\_totalcost = log(totalcost)) %>%   
 ggplot(aes(y = log\_totalcost, x = e\_rvisits))+  
 geom\_point()+  
 labs(x = "e\_rvisits",  
 y = "log(total cost)",  
 title = "Scatter plot of log(total cost) and e\_rvisits")



heart\_new\_data=   
 heart\_new %>%   
 filter(log\_totalcost >= 0 )  
  
fit = lm(log\_totalcost ~ e\_rvisits,heart\_new\_data)  
summary(fit)

##   
## Call:  
## lm(formula = log\_totalcost ~ e\_rvisits, data = heart\_new\_data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -6.2013 -1.1265 0.0191 1.2668 4.2797   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 5.53771 0.10362 53.44 <2e-16 \*\*\*  
## e\_rvisits 0.22672 0.02397 9.46 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.772 on 783 degrees of freedom  
## Multiple R-squared: 0.1026, Adjusted R-squared: 0.1014   
## F-statistic: 89.5 on 1 and 783 DF, p-value: < 2.2e-16

We can see that the p-value is extremely low, so we reject the null hypothesis that there isn’t a linear relationship between total cost and number of emergency visits. The intercept represents the expected value of (total cost + 1) after log transformation, in which case number of emergency visits equals to 0; The slope means that when one visit increases, the estimated value of (total cost + 1) after log transformation will increase 0.22529 on average. Based on the regression results, the of this model is only 0.098, which is quite small, illustrating poor performance on predicting.

## e)

heart\_new[is.na(heart\_new) | heart\_new=="-Inf"] = NA  
multi\_reg <- lm(log\_totalcost~e\_rvisits+comp\_bin,data = heart\_new)  
  
summary(multi\_reg)

##   
## Call:  
## lm(formula = log\_totalcost ~ e\_rvisits + comp\_bin, data = heart\_new)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -6.0741 -1.0737 -0.0181 1.1810 4.3848   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 5.5211 0.1013 54.495 < 2e-16 \*\*\*  
## e\_rvisits 0.2046 0.0237 8.633 < 2e-16 \*\*\*  
## comp\_bin1 1.6859 0.2749 6.132 1.38e-09 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.732 on 782 degrees of freedom  
## (3 observations deleted due to missingness)  
## Multiple R-squared: 0.1437, Adjusted R-squared: 0.1416   
## F-statistic: 65.64 on 2 and 782 DF, p-value: < 2.2e-16

## i)

lm(log\_totalcost~factor(comp\_bin)+e\_rvisits+factor(comp\_bin)\*e\_rvisits,data = heart\_new) %>%  
 summary()

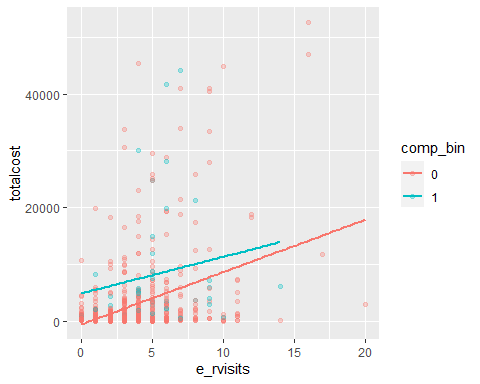
##   
## Call:  
## lm(formula = log\_totalcost ~ factor(comp\_bin) + e\_rvisits + factor(comp\_bin) \*   
## e\_rvisits, data = heart\_new)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -6.0852 -1.0802 -0.0078 1.1898 4.3803   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 5.49899 0.10349 53.138 < 2e-16 \*\*\*  
## factor(comp\_bin)1 2.17969 0.54604 3.992 7.17e-05 \*\*\*  
## e\_rvisits 0.21125 0.02453 8.610 < 2e-16 \*\*\*  
## factor(comp\_bin)1:e\_rvisits -0.09927 0.09483 -1.047 0.296   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.732 on 781 degrees of freedom  
## (3 observations deleted due to missingness)  
## Multiple R-squared: 0.1449, Adjusted R-squared: 0.1417   
## F-statistic: 44.13 on 3 and 781 DF, p-value: < 2.2e-16

lm(log\_totalcost~factor(comp\_bin)\*e\_rvisits,data = heart\_new) %>%   
 summary()

##   
## Call:  
## lm(formula = log\_totalcost ~ factor(comp\_bin) \* e\_rvisits, data = heart\_new)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -6.0852 -1.0802 -0.0078 1.1898 4.3803   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 5.49899 0.10349 53.138 < 2e-16 \*\*\*  
## factor(comp\_bin)1 2.17969 0.54604 3.992 7.17e-05 \*\*\*  
## e\_rvisits 0.21125 0.02453 8.610 < 2e-16 \*\*\*  
## factor(comp\_bin)1:e\_rvisits -0.09927 0.09483 -1.047 0.296   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.732 on 781 degrees of freedom  
## (3 observations deleted due to missingness)  
## Multiple R-squared: 0.1449, Adjusted R-squared: 0.1417   
## F-statistic: 44.13 on 3 and 781 DF, p-value: < 2.2e-16

heart\_new %>%   
 ggplot(aes(x = e\_rvisits, y = totalcost,color = comp\_bin))+  
 geom\_point(alpha = 0.3)+  
 geom\_smooth(method = "lm",se = F)

## `geom\_smooth()` using formula = 'y ~ x'



From the plot we can see that the slope of e\_rvisits change quite bit for different comp\_bin, there might be an interaction between e\_rvisits and comp\_bin. From the above summary, the model with the term “comp\_bin*e\_rvisits”, we fail to reject the null hypothesis that the coefficient of comp\_bin*e\_rvisits is 0, therefore, the interaction effect is not significant. So the comp\_bin is not a modifier.

## ii)

When adding comp\_bin into the model, the coefficient of e\_rvisits decrease from 0.22672 to 0.2046, it decreases about 10% , so binary complication variable is a counfounder of association between number of emergency visits and total cost.

## iii)

anova(multi\_reg)

## Analysis of Variance Table  
##   
## Response: log\_totalcost  
## Df Sum Sq Mean Sq F value Pr(>F)   
## e\_rvisits 1 281.16 281.160 93.680 < 2.2e-16 \*\*\*  
## comp\_bin 1 112.84 112.842 37.598 1.379e-09 \*\*\*  
## Residuals 782 2347.01 3.001   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Total cost of comp\_bin is significantly different. As a confounder, should be considered when finding the relationship between e\_rvisits and total cost.