p8130\_hw4\_rq2166

Ruoyuan Qian

2019/11/12

# Problem 1

brain = read\_excel("data\\Brain.xlsx") %>%   
 janitor::clean\_names()

## Warning in FUN(X[[i]], ...): strings not representable in native encoding  
## will be translated to UTF-8

## a)

brain\_non =   
 brain %>%   
 filter(!(species == "Homo sapiens"))  
  
fit = lm(glia\_neuron\_ratio ~ ln\_brain\_mass,brain\_non)   
summary(fit)

##   
## Call:  
## lm(formula = glia\_neuron\_ratio ~ ln\_brain\_mass, data = brain\_non)  
##   
## 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)

coef = summary(fit)$coefficients[2,1]  
err = summary(fit)$coefficients[1,1]  
  
ln\_brain\_mass\_human =   
 brain %>%   
 filter(species == "Homo sapiens") %>%   
 pull(ln\_brain\_mass)  
  
  
glia\_human = coef\*ln\_brain\_mass\_human + err

## c)

using PI

## d)

brain\_human =   
 brain %>%   
 filter(species == "Homo sapiens")  
  
pred\_fit <- predict.lm(fit, interval="prediction")

## Warning in predict.lm(fit, interval = "prediction"): predictions on current data refer to \_future\_ responses

PI\_all <- data.frame(cbind(brain$ln\_brain\_mass,  
 brain$glia\_neuron\_ratio,  
 pred\_fit))

## Warning in cbind(brain$ln\_brain\_mass, brain$glia\_neuron\_ratio, pred\_fit):  
## number of rows of result is not a multiple of vector length (arg 1)

human\_PI\_lwr = PI\_all[[1,4]]  
human\_PI\_upr = PI\_all[[1,5]]  
  
  
human\_PI\_upr > brain\_human$glia\_neuron\_ratio

## [1] FALSE

Yes, the upper bound of prediction for glia\_neuron\_ratio of human is smaller than the observed value, which means the human brain has excessive glia\_neuron\_ratio than other specise.

## e)

Compared to data of other species, human data is considered to be an outlier and would affect the regression if it was used to fit the model.

# Problem 4

heart = read\_csv("data\\HeartDisease.csv") %>%   
 janitor::clean\_names() %>%   
 mutate(gender = factor(gender,level=c(0:1),  
 labels = c("female","male")))

## Parsed with column specification:  
## cols(  
## id = col\_double(),  
## totalcost = col\_double(),  
## age = col\_double(),  
## gender = col\_double(),  
## interventions = col\_double(),  
## drugs = col\_double(),  
## ERvisits = col\_double(),  
## complications = col\_double(),  
## comorbidities = col\_double(),  
## duration = col\_double()  
## )

## a)

**Description**:

The dataset is collected by an insurance company about the total cost of heart disease and other relevant events caused by heart disease. The main outcome is total cost from the heart disease (totalcost), and the main predictor is the number of emergency room (ER) visits (e\_rvisits). There are other important variables like age and gender of subject, total number of interventions carried out(interventions), the number of tracked drugs (drugs), the number of other complications arose due to other disease instead of heart disease (complications), the number of other disease during the period (comorbidities), the number of days of treated duration (duration).

Descriptive statistics

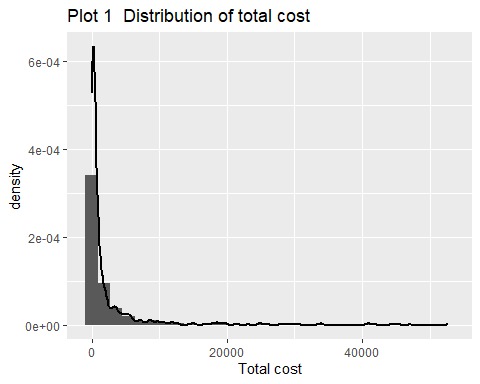
sum\_data <- arsenal::tableby( ~ totalcost + age + gender +  
 interventions + drugs+  
 e\_rvisits+complications+  
 comorbidities+duration,   
 data = heart,  
 test = FALSE,   
 total = FALSE,  
 numeric.stats = c("meansd" ,"medianq1q3","range" ) )  
 summary(sum\_data,text = TRUE)

##   
##   
## | | 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 | |  
## |- female | 608 (77.2%) |  
## |- male | 180 (22.8%) |  
## |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 |

## b)

heart %>%   
 ggplot(aes(totalcost,..density..))+  
 geom\_histogram()+  
 geom\_line(stat = 'Density',size = 1)+  
 labs(x = "Total cost",  
 title = "Plot 1 Distribution of total cost")

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



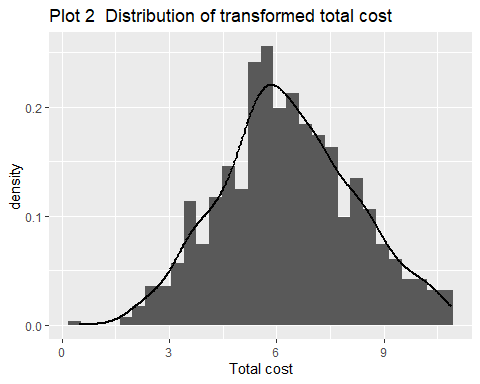
Different transformation

heart\_log =   
 heart %>%   
 mutate(log\_totalcost = log(totalcost))  
  
heart\_log %>%   
 ggplot(aes(log\_totalcost,..density..))+  
 geom\_histogram()+  
 geom\_line(stat = 'Density',size = 1)+  
 labs(x = "Total cost",  
 title = "Plot 2 Distribution of transformed total cost")

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

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

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

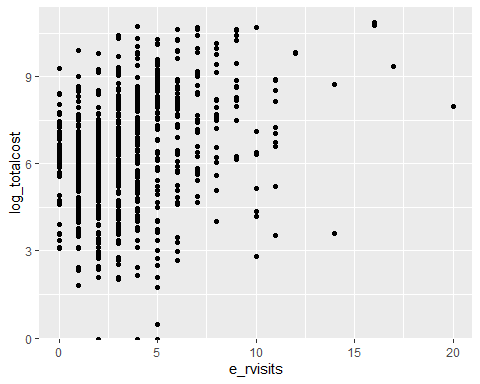


## c)

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

## d)

heart\_log\_bin %>%   
 mutate(log\_totalcost = log(totalcost)) %>%   
ggplot(aes(y = log\_totalcost, x = e\_rvisits))+  
 geom\_point()



heart\_fill =   
heart\_log\_bin %>%   
 filter(log\_totalcost >= 0 )  
  
fit = lm(log\_totalcost ~ e\_rvisits,heart\_fill)  
summary(fit)

##   
## Call:  
## lm(formula = log\_totalcost ~ e\_rvisits, data = heart\_fill)  
##   
## 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

## e)

### i)

reg2 = lm(log\_totalcost ~ e\_rvisits + factor(comp\_bin),heart\_fill)  
summary(reg2)

##   
## Call:  
## lm(formula = log\_totalcost ~ e\_rvisits + factor(comp\_bin), data = heart\_fill)  
##   
## 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 \*\*\*  
## factor(comp\_bin)1 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  
## Multiple R-squared: 0.1437, Adjusted R-squared: 0.1416   
## F-statistic: 65.64 on 2 and 782 DF, p-value: < 2.2e-16

coef\_after = summary(reg2)$coefficients[2,1]

comp\_bin is an effect modifier because the p-value of coefficient of comp\_bin in “having complication level” is significantly small at 0.05 significant level, which means we should reject the null and conclude that there is a significant difference between

### ii)

reg1 = lm(log\_totalcost ~ e\_rvisits ,heart\_fill)  
summary(reg1)

##   
## Call:  
## lm(formula = log\_totalcost ~ e\_rvisits, data = heart\_fill)  
##   
## 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

coef\_before = summary(reg1)$coefficients[2,1]

rate = (coef\_after - coef\_before)/coef\_before  
 rate > 0.1

## [1] FALSE

Since the changing rate of coefficient of ERvisits is smaller than 10% after putting the comp\_bin into model, the comp\_bin is not a confounder between total cost and ERvisits.

### iii)

anova(reg1,reg2)

## Analysis of Variance Table  
##   
## Model 1: log\_totalcost ~ e\_rvisits  
## Model 2: log\_totalcost ~ e\_rvisits + factor(comp\_bin)  
## Res.Df RSS Df Sum of Sq F Pr(>F)   
## 1 783 2459.8   
## 2 782 2347.0 1 112.84 37.598 1.379e-09 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

We reject the null and conclude that model with comp\_bin is superior, so it is DXXXable to add comp\_bin variable into the model.

## f)

### i)

reg\_m = lm(log\_totalcost ~ e\_rvisits + age + gender + duration, heart\_fill)  
summary(reg\_m)

##   
## Call:  
## lm(formula = log\_totalcost ~ e\_rvisits + age + gender + duration,   
## data = heart\_fill)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -5.3012 -1.0790 -0.1215 1.0567 4.2854   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 6.1967450 0.5162986 12.002 < 2e-16 \*\*\*  
## e\_rvisits 0.1946147 0.0225477 8.631 < 2e-16 \*\*\*  
## age -0.0249155 0.0087695 -2.841 0.00461 \*\*   
## gendermale -0.1189335 0.1408787 -0.844 0.39880   
## duration 0.0057155 0.0004941 11.569 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.639 on 780 degrees of freedom  
## Multiple R-squared: 0.2359, Adjusted R-squared: 0.232   
## F-statistic: 60.2 on 4 and 780 DF, p-value: < 2.2e-16

### ii)

I will use the MLR to address the interest problem of investigator. In the MLR, the coefficient of main predictor (ERvisits) is more precise due to the adjustment of other covariates, which may bias the reslut.