Homework 4

Xinyi Lin

11/9/2018

# Problem 1

## Question 1

Since

The expectation of ’s numerator is

So and are unbiased estimators of and .

## Question 2

As , and estimated regression model ,

when ,

so regression model always goes through the point .

## Question 3

The regression model is:

the likelihood of the linear model becomes:

the log-likelihood function:

to find , we let

so estimator of

since, is biased estimator of , while , is unbiased estimator of .

library(tidyverse)  
library(patchwork)

# Problem 2

First, we need to import data

HeartDisease\_df = read\_csv("./data/HeartDisease.csv")

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

head(HeartDisease\_df)

## # A tibble: 6 x 10  
## id totalcost age gender interventions drugs ERvisits complications  
## <int> <dbl> <int> <int> <int> <int> <int> <int>  
## 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  
## # ... with 2 more variables: comorbidities <int>, duration <int>

## Question 1

This dataset includes 788 observations and 10 variables. Among variables, main outcome is totalcost, main predictor is ERvisits and other important covariates including age, gender and duration.

Then, we show descriptive statistics for all variables of interest.

mean\_and\_sd = function(x) {  
   
 if (!is.numeric(x)) {  
 stop("Argument x should be numeric")  
 } else if (length(x) == 1) {  
 stop("Cannot be computed for length 1 vectors")  
 }  
   
 mean\_x = mean(x)  
 sd\_x = sd(x)  
  
 list(mean = mean\_x,   
 sd = sd\_x)  
}

totalcost

mean\_and\_sd(HeartDisease\_df$totalcost)

## $mean  
## [1] 2799.956  
##   
## $sd  
## [1] 6690.26

ERvisits

mean\_and\_sd(HeartDisease\_df$ERvisits)

## $mean  
## [1] 3.425127  
##   
## $sd  
## [1] 2.637474

age

mean\_and\_sd(HeartDisease\_df$age)

## $mean  
## [1] 58.71827  
##   
## $sd  
## [1] 6.754118

gender

summary(as.factor(HeartDisease\_df$gender))

## 0 1   
## 608 180

complications

summary(as.factor(HeartDisease\_df$complications))

## 0 1 3   
## 745 42 1

## Question 2

total\_plot =  
 HeartDisease\_df %>%   
 ggplot(aes(x = totalcost)) +  
 geom\_density() +  
 labs(title = "pdf of total cost")

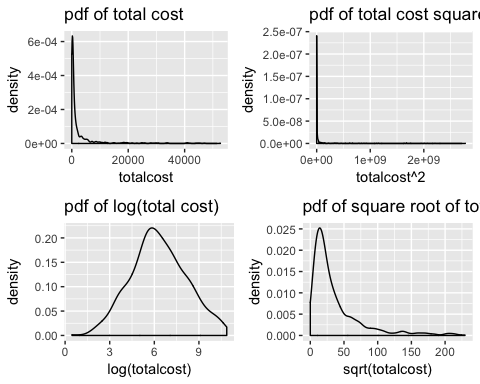
log\_plot =   
 HeartDisease\_df %>%   
 ggplot(aes(x = log(totalcost))) +  
 geom\_density() +  
 labs(title = "pdf of log(total cost)")

sqrt\_plot =   
 HeartDisease\_df %>%   
 ggplot(aes(x = sqrt(totalcost))) +  
 geom\_density() +  
 labs(title = "pdf of square root of total cost")

square\_plot =   
 HeartDisease\_df %>%   
 ggplot(aes(x = totalcost^2)) +  
 geom\_density() +  
 labs(title = "pdf of total cost square")

(total\_plot + square\_plot)/(log\_plot + sqrt\_plot)

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



Above are distribution of total cost, log(totalcost), suqre root of totalcost and totalcost square. We can find that apply log to total cost is the best transformations.

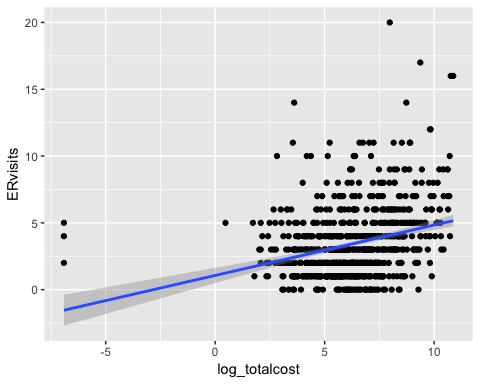
## Question 3

HeartDisease\_df =  
 HeartDisease\_df %>%   
 mutate(comp\_bin = ifelse(complications == 0, 0, 1)) %>%   
 mutate(totalcost = ifelse(totalcost == 0, 0.001, totalcost))  
  
head(HeartDisease\_df)

## # A tibble: 6 x 11  
## id totalcost age gender interventions drugs ERvisits complications  
## <int> <dbl> <int> <int> <int> <int> <int> <int>  
## 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  
## # ... with 3 more variables: comorbidities <int>, duration <int>,  
## # comp\_bin <dbl>

## Question 4

HeartDisease\_df %>%   
 mutate(log\_totalcost = log(totalcost)) %>%   
 ggplot(aes(x = log\_totalcost, y = ERvisits)) +  
 geom\_point() +  
 geom\_smooth(method = 'lm',formula = y~x)



reg\_Heart =   
 HeartDisease\_df %>%   
 mutate(log\_totalcost = log(totalcost)) %>%   
 #filter(is.finite(log\_totalcost)) %>%   
 lm(formula = log\_totalcost ~ ERvisits, data = .)   
  
reg\_Heart %>%   
 broom::tidy()

## # A tibble: 2 x 5  
## term estimate std.error statistic p.value  
## <chr> <dbl> <dbl> <dbl> <dbl>  
## 1 (Intercept) 5.49 0.114 48.2 3.56e-237  
## 2 ERvisits 0.225 0.0263 8.53 7.39e- 17

summary(reg\_Heart)

##   
## Call:  
## lm(formula = log\_totalcost ~ ERvisits, data = .)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -13.5255 -1.0922 0.0608 1.3147 4.3314   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 5.49384 0.11387 48.248 <2e-16 \*\*\*  
## ERvisits 0.22477 0.02635 8.531 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.949 on 786 degrees of freedom  
## Multiple R-squared: 0.08475, Adjusted R-squared: 0.08359   
## F-statistic: 72.78 on 1 and 786 DF, p-value: < 2.2e-16

According to the results, we can find that adjusted R-squared is 0.1014 which is very closed to 0 and means this simple linear model is not a proper model. However, p-value of slope is lower than 2.2e-16, which means the slope is significant and there are positive relationship between log of total cost and number of emergency room visits.

Interpretation: The slop of model is 0.227 which means if the number of emergency room vistis increases by 1 unit, the log of total cost will increase 0.452 units.

## Question 5

Test if comp\_bin is an effect modifier

reg\_modifier\_Heart =   
 HeartDisease\_df %>%   
 mutate(log\_totalcost = log(totalcost)) %>%   
 #filter(is.finite(log\_totalcost)) %>%   
 lm(formula = log\_totalcost ~ ERvisits + comp\_bin + ERvisits\*comp\_bin, data = .)   
  
reg\_modifier\_Heart %>%   
 broom::tidy()

## # A tibble: 4 x 5  
## term estimate std.error statistic p.value  
## <chr> <dbl> <dbl> <dbl> <dbl>  
## 1 (Intercept) 5.46 0.114 47.8 1.12e-234  
## 2 ERvisits 0.208 0.0271 7.70 4.01e- 14  
## 3 comp\_bin 2.22 0.602 3.69 2.39e- 4  
## 4 ERvisits:comp\_bin -0.0964 0.105 -0.921 3.57e- 1

summary(reg\_modifier\_Heart)

##   
## Call:  
## lm(formula = log\_totalcost ~ ERvisits + comp\_bin + ERvisits \*   
## comp\_bin, data = .)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -13.4051 -1.0559 0.0325 1.2269 4.4353   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 5.45548 0.11406 47.828 < 2e-16 \*\*\*  
## ERvisits 0.20837 0.02705 7.703 4.01e-14 \*\*\*  
## comp\_bin 2.22320 0.60233 3.691 0.000239 \*\*\*  
## ERvisits:comp\_bin -0.09639 0.10461 -0.921 0.357103   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.911 on 784 degrees of freedom  
## Multiple R-squared: 0.1227, Adjusted R-squared: 0.1193   
## F-statistic: 36.55 on 3 and 784 DF, p-value: < 2.2e-16

Since the corresponding p-value of ’ERvisits\*comp\_bin’ is 0.357 which is bigger than 0.05, we can conclude that there is no interaction between ERvisits and comp\_bin and comp\_bin is not a modifier.

Test if comp\_bin is a confunder.

reg\_confounder\_Heart =   
 HeartDisease\_df %>%   
 mutate(log\_totalcost = log(totalcost)) %>%   
 #filter(is.finite(log\_totalcost)) %>%   
 lm(formula = log\_totalcost ~ ERvisits + comp\_bin, data = .)   
  
reg\_confounder\_Heart %>%   
 broom::tidy()

## # A tibble: 3 x 5  
## term estimate std.error statistic p.value  
## <chr> <dbl> <dbl> <dbl> <dbl>  
## 1 (Intercept) 5.48 0.112 49.1 2.79e-241  
## 2 ERvisits 0.202 0.0261 7.73 3.33e- 14  
## 3 comp\_bin 1.74 0.303 5.75 1.27e- 8

summary(reg\_confounder\_Heart)

##   
## Call:  
## lm(formula = log\_totalcost ~ ERvisits + comp\_bin, data = .)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -13.3943 -1.0451 0.0252 1.2191 4.4397   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 5.47693 0.11165 49.054 < 2e-16 \*\*\*  
## ERvisits 0.20193 0.02613 7.728 3.33e-14 \*\*\*  
## comp\_bin 1.74365 0.30321 5.751 1.27e-08 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.911 on 785 degrees of freedom  
## Multiple R-squared: 0.1218, Adjusted R-squared: 0.1195   
## F-statistic: 54.41 on 2 and 785 DF, p-value: < 2.2e-16

When adding comp\_bin in model the association between log\_totalcost and ERvisits becomes smaller but still significant and the regression coefficient decreased by 10.2%, so comp\_bin is a confounder.

Since comp\_bin is a confounder but not a modifier, we use ‘Partial’ F-test to test whether we should include comp\_bin as a factor.

Model 1:

Model 2:

Among which, represents ER\_visits, represents comp\_bin.

Null hypothesis , alternative hypothesis

Decision rule:

where .

If , reject ;

If , fail to reject .

With , when , fail to reject , when , reject .

anova(reg\_confounder\_Heart, reg\_Heart) %>%   
 broom::tidy()

## Warning: Unknown or uninitialised column: 'term'.

## # A tibble: 2 x 6  
## res.df rss df sumsq statistic p.value  
## \* <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 785 2866. NA NA NA NA   
## 2 786 2987. -1 -121. 33.1 0.0000000127

According to results, p-value is smaller than 0.01 so we reject and conclude that Model 1 is ‘superior’.As a resuit, we should include comp\_bin should be added in the model and model 1 is ‘superior’.

## Question 6

reg\_added\_Heart =   
 HeartDisease\_df %>%   
 mutate(log\_totalcost = log(totalcost)) %>%   
 #filter(is.finite(log\_totalcost)) %>%   
 lm(formula = log\_totalcost ~ ERvisits + comp\_bin + age + gender + duration, data = .)   
  
reg\_added\_Heart %>%   
 broom::tidy()

## # A tibble: 6 x 5  
## term estimate std.error statistic p.value  
## <chr> <dbl> <dbl> <dbl> <dbl>  
## 1 (Intercept) 5.80 0.556 10.4 5.91e-24  
## 2 ERvisits 0.173 0.0246 7.05 4.07e-12  
## 3 comp\_bin 1.53 0.282 5.45 6.89e- 8  
## 4 age -0.0193 0.00945 -2.05 4.10e- 2  
## 5 gender -0.323 0.151 -2.14 3.26e- 2  
## 6 duration 0.00606 0.000533 11.4 6.76e-28

summary(reg\_added\_Heart)

##   
## Call:  
## lm(formula = log\_totalcost ~ ERvisits + comp\_bin + age + gender +   
## duration, data = .)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -12.1885 -0.9962 -0.0838 1.0099 4.3499   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 5.8016080 0.5559910 10.435 < 2e-16 \*\*\*  
## ERvisits 0.1732359 0.0245897 7.045 4.07e-12 \*\*\*  
## comp\_bin 1.5335773 0.2815738 5.446 6.89e-08 \*\*\*  
## age -0.0193389 0.0094493 -2.047 0.0410 \*   
## gender -0.3234418 0.1510875 -2.141 0.0326 \*   
## duration 0.0060629 0.0005325 11.386 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.769 on 782 degrees of freedom  
## Multiple R-squared: 0.2502, Adjusted R-squared: 0.2454   
## F-statistic: 52.18 on 5 and 782 DF, p-value: < 2.2e-16

According to results, we can find all p-value of covariates are smaller than 0.01, so all covariates have significant influence in total cost.

We use ‘Partial’ F-test to compare SLR and MLR models.

Model 1:

Model 2:

Among which, represents ERvisits, represents comp\_bin, represents age, represents gender, represents duration.

Null hypothesis , alternative hypothesis at least one of is not zero.

Decision rule:

where .

If , reject ;

If , fail to reject .

With , when , fail to reject , when , reject .

anova(reg\_Heart, reg\_added\_Heart) %>% broom::tidy()

## Warning: Unknown or uninitialised column: 'term'.

## # A tibble: 2 x 6  
## res.df rss df sumsq statistic p.value  
## \* <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 786 2987. NA NA NA NA   
## 2 782 2447. 4 540. 43.1 1.00e-32

According to the ANOVA results, p-value is smaller than 0.01 so we reject and conclude that Model 1 is ‘superior’.As a resuit, we should use MLR model.

# Problem 3

First, we import data

PatSatisfaction\_df = readxl::read\_xlsx("./data/PatSatisfaction.xlsx") %>%   
 janitor::clean\_names() %>%   
 reshape::rename(c(safisfaction = "satisfaction"))  
  
head(PatSatisfaction\_df)

## # A tibble: 6 x 4  
## satisfaction age severity anxiety  
## <dbl> <dbl> <dbl> <dbl>  
## 1 48 50 51 2.3  
## 2 57 36 46 2.3  
## 3 66 40 48 2.2  
## 4 70 41 44 1.8  
## 5 89 28 43 1.8  
## 6 36 49 54 2.9

## Question 1

PatSatisfaction\_df %>%   
 cor()

## satisfaction age severity anxiety  
## satisfaction 1.0000000 -0.7867555 -0.6029417 -0.6445910  
## age -0.7867555 1.0000000 0.5679505 0.5696775  
## severity -0.6029417 0.5679505 1.0000000 0.6705287  
## anxiety -0.6445910 0.5696775 0.6705287 1.0000000

According to the correlation matrix, we can find that all age, severity, anxirty have negative relationship with satisfaction and the relationship between age and satisfaction is stronger than severity and anxiety .

## Question 2

Assuming the model is

Among which, represents age, represents severity, represents anxiety.

Null hypothesis , alternative hypothesis at least one is not zero.

Decision rule:

If , reject ,

if , fail to reject .

with a significance level of 0.05,

reg\_all =   
 PatSatisfaction\_df %>%  
 lm(satisfaction ~ age + severity + anxiety, data = .)  
  
summary(reg\_all)

##   
## Call:  
## lm(formula = satisfaction ~ age + severity + anxiety, data = .)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -18.3524 -6.4230 0.5196 8.3715 17.1601   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 158.4913 18.1259 8.744 5.26e-11 \*\*\*  
## age -1.1416 0.2148 -5.315 3.81e-06 \*\*\*  
## severity -0.4420 0.4920 -0.898 0.3741   
## anxiety -13.4702 7.0997 -1.897 0.0647 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 10.06 on 42 degrees of freedom  
## Multiple R-squared: 0.6822, Adjusted R-squared: 0.6595   
## F-statistic: 30.05 on 3 and 42 DF, p-value: 1.542e-10

According to results, we can find = 30.05 > 2.8270487, so we reject and conclude that there is a regression relation.

## Question 3

reg\_all %>% broom::tidy()

## # A tibble: 4 x 5  
## term estimate std.error statistic p.value  
## <chr> <dbl> <dbl> <dbl> <dbl>  
## 1 (Intercept) 158. 18.1 8.74 5.26e-11  
## 2 age -1.14 0.215 -5.31 3.81e- 6  
## 3 severity -0.442 0.492 -0.898 3.74e- 1  
## 4 anxiety -13.5 7.10 -1.90 6.47e- 2

confint(reg\_all)

## 2.5 % 97.5 %  
## (Intercept) 121.911727 195.0707761  
## age -1.575093 -0.7081303  
## severity -1.434831 0.5508228  
## anxiety -27.797859 0.8575324

By using function confint, we get 95% CIs of all estimators. The 95% CIs of severity is (-1.4348, 0.5508) which means at significant level, we can conclude that the mean value of satisfaction changes somewhere between decreasing 1.4348 and increasing 0.5508 for each additional unit of the severity of the illness given all other values of predictors stay constant.

The estimated coefficient of severity is -0.442 which means if the value of severity increased by 1 units, the mean value of satisfaction will decrease 0.442 given all other values of predictors stay constant.

## Question 4

list(age = 35, severity = 42, anxiety = 2.1) %>%   
 predict(object = reg\_all, newdata = ., interval = "predict")

## fit lwr upr  
## 1 71.68332 50.06237 93.30426

By using predict function, we can get the prediction interval for the new patient’s satisfaction is (50.0624, 93.3042).

Interprest: We are 95% confident that the the new patient’s satisfaction fall within (50.0624, 93.3042) given age equals 35, severity equals 42 and anxiety equals 2.1

## Question 5

Model 1:

Model 2:

Among which, represents age, represents severity, represents anxiety.

We use ‘Partial’ F-test for nested models. Null hypothesis , alternative hypothesis

Decision rule:

where .

If , reject ;

If , fail to reject .

With , when , fail to reject , when , reject .

reg\_without\_anxiety =   
 PatSatisfaction\_df %>%  
 lm(satisfaction ~ age + severity, data = .)  
  
anova(reg\_all, reg\_without\_anxiety) %>%   
 broom::tidy()

## Warning: Unknown or uninitialised column: 'term'.

## # A tibble: 2 x 6  
## res.df rss df sumsq statistic p.value  
## \* <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 42 4249. NA NA NA NA   
## 2 43 4613. -1 -364. 3.60 0.0647

According to the ANOVA results, p-value is 0.0647 which is larger than 0.05, so we fail to reject and conclude that Model 1 is not ’superior` and we should use Model 2.