final homework

Ruizhe Xu

2020/7/18

# 体检数据分析

## 数据的基本处理

首先将体检数据导入进R，体检数据依次包含：性别、年龄、体重、身高、是否吸烟、是否喝酒、葡萄糖、收缩压、舒张压、甘油三脂、高密度脂蛋白胆固醇与是否患代谢综合症。

phy\_test=read\_csv("data/phy\_test.csv")

## Parsed with column specification:  
## cols(  
## gender = col\_character(),  
## age = col\_double(),  
## weight = col\_double(),  
## height = col\_double(),  
## smoke = col\_character(),  
## drunk = col\_character(),  
## FPG = col\_double(),  
## sbp = col\_double(),  
## dbp = col\_double(),  
## TG = col\_double(),  
## HDL\_C = col\_double(),  
## whether\_sick = col\_character()  
## )

head(phy\_test)

## # A tibble: 6 x 12  
## gender age weight height smoke drunk FPG sbp dbp TG HDL\_C  
## <chr> <dbl> <dbl> <dbl> <chr> <chr> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 Male 56 66.3 171 Yes Yes 5.01 95 60 0.74 1.79  
## 2 Female 27 54.5 172 No No 4.44 112 58 0.75 2.17  
## 3 Male 64 59.9 159 No Yes 5.93 141 81 1.76 1.91  
## 4 Male 43 59.5 168. No No 5.68 100 69 5.03 1.2   
## 5 Female 59 47.8 160. No No 6.45 107 57 0.84 1.53  
## 6 Male 64 72.1 176. No No 5.34 137 80 0.83 1.81  
## # ... with 1 more variable: whether\_sick <chr>

将数据中的Yes与No的选项换成TRUE与FALSE的逻辑值，同时计算BMI值=体重(Kg)/(身高(m)\*身高(m))

phy\_test$smoke[which(phy\_test$smoke=="Yes")]=TRUE  
phy\_test$smoke[which(phy\_test$smoke=="No")]=FALSE  
phy\_test$drunk[which(phy\_test$drunk=="Yes")]=TRUE  
phy\_test$drunk[which(phy\_test$drunk=="No")]=FALSE  
phy\_test$whether\_sick[which(phy\_test$whether\_sick=="Yes")]=TRUE  
phy\_test$whether\_sick[which(phy\_test$whether\_sick=="No")]=FALSE  
phy\_test$whether\_sick=as.logical(phy\_test$whether\_sick)  
phy\_test$smoke=as.logical(phy\_test$smoke)  
phy\_test$drunk=as.logical(phy\_test$drunk)  
phy\_test$height=phy\_test$height/100  
phy\_test=phy\_test %>% mutate(BMI=weight/(height\*height))

## 代谢综合症患者的特折

根据数据的总结与可视化可知，患有代谢综合症的患者一般来说都具有超重、高血糖、高血压、高甘油三酯与低高密度脂蛋白胆固醇等特征

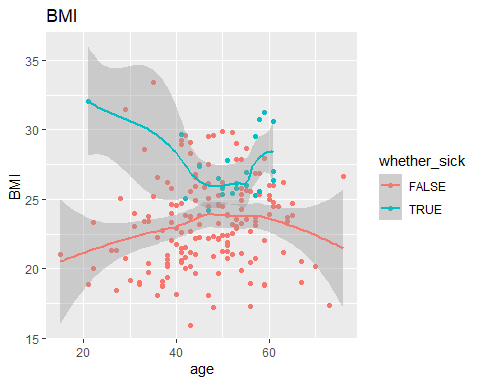
phy\_test %>% group\_by(whether\_sick) %>%  
 summarise(mean(BMI)) %>% ungroup()

## `summarise()` ungrouping output (override with `.groups` argument)

## # A tibble: 2 x 2  
## whether\_sick `mean(BMI)`  
## <lgl> <dbl>  
## 1 FALSE 23.3  
## 2 TRUE 27.2

ggplot(data = phy\_test)+  
 geom\_point(aes(x=age,y=BMI,color=whether\_sick))+  
 geom\_smooth(aes(x=age,y=BMI,color=whether\_sick))+  
 labs(title = "BMI")

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



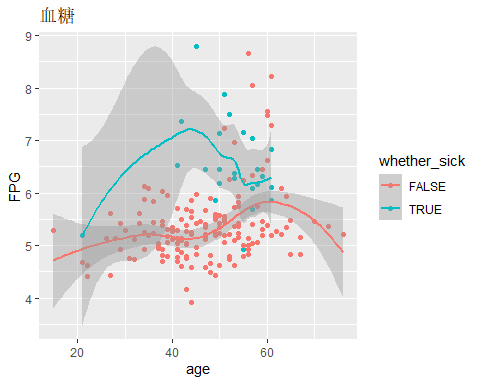
phy\_test %>% group\_by(whether\_sick) %>%  
 summarise(mean(FPG)) %>% ungroup()

## `summarise()` ungrouping output (override with `.groups` argument)

## # A tibble: 2 x 2  
## whether\_sick `mean(FPG)`  
## <lgl> <dbl>  
## 1 FALSE 5.36  
## 2 TRUE 6.52

ggplot(data = phy\_test)+  
 geom\_point(aes(x=age,y=FPG,color=whether\_sick))+  
 geom\_smooth(aes(x=age,y=FPG,color=whether\_sick))+  
 labs(title = "血糖")

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



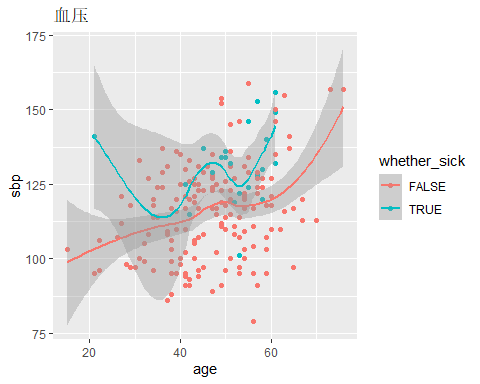
phy\_test %>% group\_by(whether\_sick) %>%  
 summarise(mean(sbp)) %>% ungroup()

## `summarise()` ungrouping output (override with `.groups` argument)

## # A tibble: 2 x 2  
## whether\_sick `mean(sbp)`  
## <lgl> <dbl>  
## 1 FALSE 116.  
## 2 TRUE 132.

ggplot(data = phy\_test)+  
 geom\_point(aes(x=age,y=sbp,color=whether\_sick))+  
 geom\_smooth(aes(x=age,y=sbp,color=whether\_sick))+  
 labs(title = "血压")

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



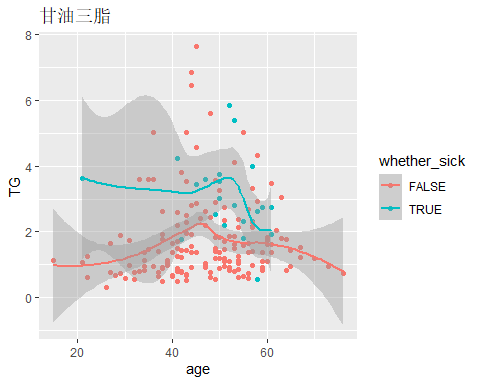
phy\_test %>% group\_by(whether\_sick) %>%  
 summarise(mean(TG)) %>% ungroup()

## `summarise()` ungrouping output (override with `.groups` argument)

## # A tibble: 2 x 2  
## whether\_sick `mean(TG)`  
## <lgl> <dbl>  
## 1 FALSE 1.72  
## 2 TRUE 2.93

ggplot(data = phy\_test)+  
 geom\_point(aes(x=age,y=TG,color=whether\_sick))+  
 geom\_smooth(aes(x=age,y=TG,color=whether\_sick))+  
 labs(title = "甘油三脂")

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



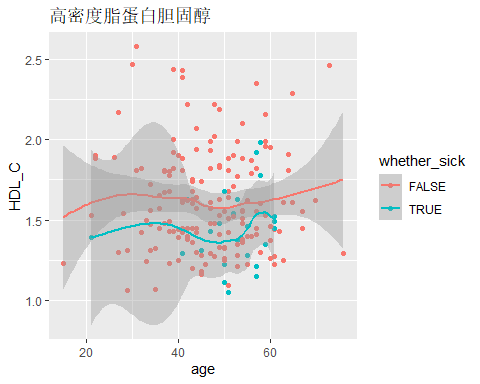
phy\_test %>% group\_by(whether\_sick) %>%  
 summarise(mean(HDL\_C)) %>% ungroup()

## `summarise()` ungrouping output (override with `.groups` argument)

## # A tibble: 2 x 2  
## whether\_sick `mean(HDL\_C)`  
## <lgl> <dbl>  
## 1 FALSE 1.62  
## 2 TRUE 1.45

ggplot(data = phy\_test)+  
 geom\_point(aes(x=age,y=HDL\_C,color=whether\_sick))+  
 geom\_smooth(aes(x=age,y=HDL\_C,color=whether\_sick))+  
 labs(title = "高密度脂蛋白胆固醇")

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



## 代谢综合症与不良生活习惯之间的关系

由下列的分析与数据可视化知，由抽烟习惯的人一般具有超重、高血糖、高血压、高甘油三酯与低高密度脂蛋白胆固醇这些特征，因此抽烟的人会有更大的概率换代谢综合症。

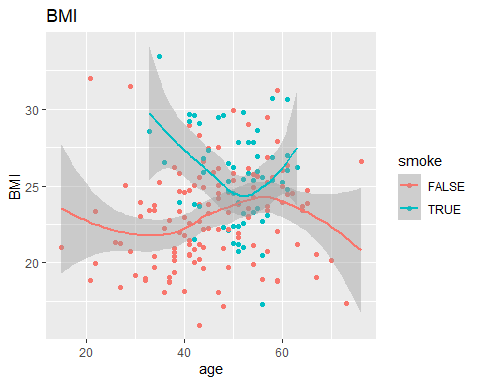
phy\_test %>% group\_by(smoke) %>%  
 summarise(mean(BMI)) %>% ungroup()

## `summarise()` ungrouping output (override with `.groups` argument)

## # A tibble: 2 x 2  
## smoke `mean(BMI)`  
## <lgl> <dbl>  
## 1 FALSE 23.0  
## 2 TRUE 25.5

ggplot(data = phy\_test)+  
 geom\_point(aes(x=age,y=BMI,color=smoke))+  
 geom\_smooth(aes(x=age,y=BMI,color=smoke))+  
 labs(title = "BMI")

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



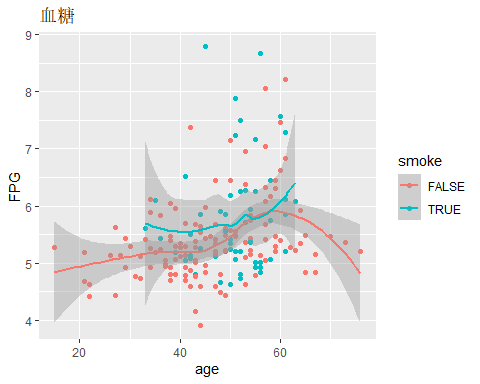
phy\_test %>% group\_by(smoke) %>%  
 summarise(mean(FPG)) %>% ungroup()

## `summarise()` ungrouping output (override with `.groups` argument)

## # A tibble: 2 x 2  
## smoke `mean(FPG)`  
## <lgl> <dbl>  
## 1 FALSE 5.40  
## 2 TRUE 5.75

ggplot(data = phy\_test)+  
 geom\_point(aes(x=age,y=FPG,color=smoke))+  
 geom\_smooth(aes(x=age,y=FPG,color=smoke))+  
 labs(title = "血糖")

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



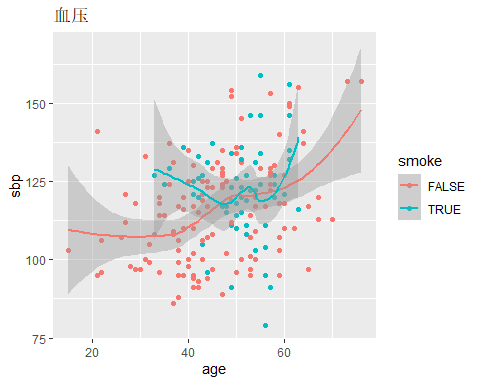
phy\_test %>% group\_by(smoke) %>%  
 summarise(mean(sbp)) %>% ungroup()

## `summarise()` ungrouping output (override with `.groups` argument)

## # A tibble: 2 x 2  
## smoke `mean(sbp)`  
## <lgl> <dbl>  
## 1 FALSE 116.  
## 2 TRUE 122.

ggplot(data = phy\_test)+  
 geom\_point(aes(x=age,y=sbp,color=smoke))+  
 geom\_smooth(aes(x=age,y=sbp,color=smoke))+  
 labs(title = "血压")

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



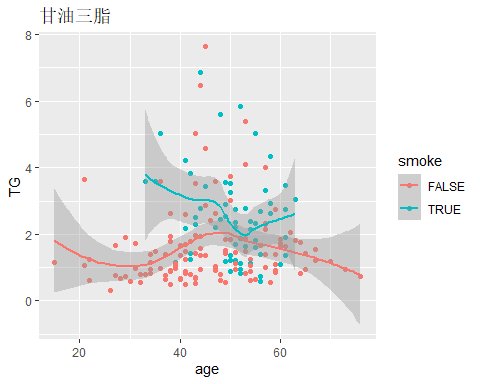
phy\_test %>% group\_by(smoke) %>%  
 summarise(mean(TG)) %>% ungroup()

## `summarise()` ungrouping output (override with `.groups` argument)

## # A tibble: 2 x 2  
## smoke `mean(TG)`  
## <lgl> <dbl>  
## 1 FALSE 1.58  
## 2 TRUE 2.52

ggplot(data = phy\_test)+  
 geom\_point(aes(x=age,y=TG,color=smoke))+  
 geom\_smooth(aes(x=age,y=TG,color=smoke))+  
 labs(title = "甘油三脂")

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



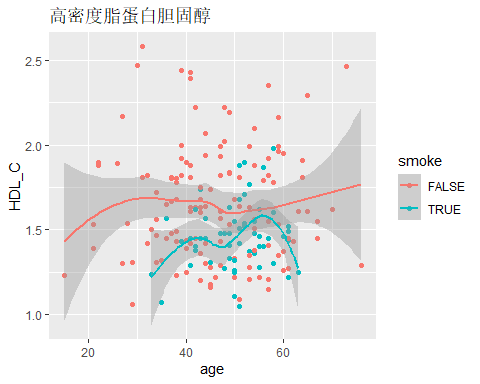
phy\_test %>% group\_by(smoke) %>%  
 summarise(mean(HDL\_C)) %>% ungroup()

## `summarise()` ungrouping output (override with `.groups` argument)

## # A tibble: 2 x 2  
## smoke `mean(HDL\_C)`  
## <lgl> <dbl>  
## 1 FALSE 1.65  
## 2 TRUE 1.48

ggplot(data = phy\_test)+  
 geom\_point(aes(x=age,y=HDL\_C,color=smoke))+  
 geom\_smooth(aes(x=age,y=HDL\_C,color=smoke))+  
 labs(title = "高密度脂蛋白胆固醇")

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



由下列的分析知，由饮酒习惯的人一般也都具有具有超重、高血糖、高血压、高甘油三酯与低高密度脂蛋白胆固醇这些特征，因此饮酒的人会得代谢综合症得概率更大一些

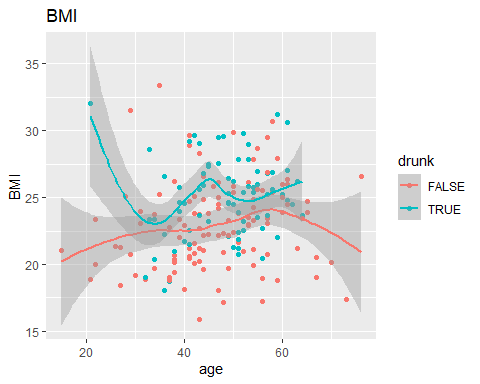
phy\_test %>% group\_by(drunk) %>%  
 summarise(mean(BMI)) %>% ungroup()

## `summarise()` ungrouping output (override with `.groups` argument)

## # A tibble: 2 x 2  
## drunk `mean(BMI)`  
## <lgl> <dbl>  
## 1 FALSE 23.0  
## 2 TRUE 25.0

ggplot(data = phy\_test)+  
 geom\_point(aes(x=age,y=BMI,color=drunk))+  
 geom\_smooth(aes(x=age,y=BMI,color=drunk))+  
 labs(title = "BMI")

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



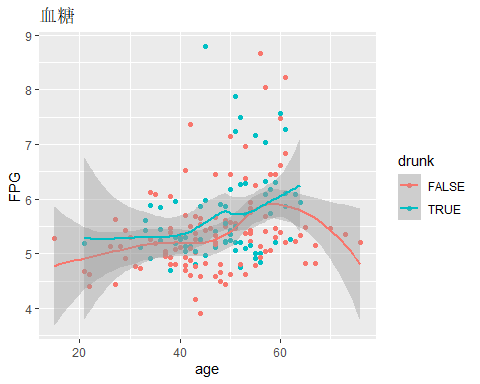
phy\_test %>% group\_by(drunk) %>%  
 summarise(mean(FPG)) %>% ungroup()

## `summarise()` ungrouping output (override with `.groups` argument)

## # A tibble: 2 x 2  
## drunk `mean(FPG)`  
## <lgl> <dbl>  
## 1 FALSE 5.41  
## 2 TRUE 5.67

ggplot(data = phy\_test)+  
 geom\_point(aes(x=age,y=FPG,color=drunk))+  
 geom\_smooth(aes(x=age,y=FPG,color=drunk))+  
 labs(title = "血糖")

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



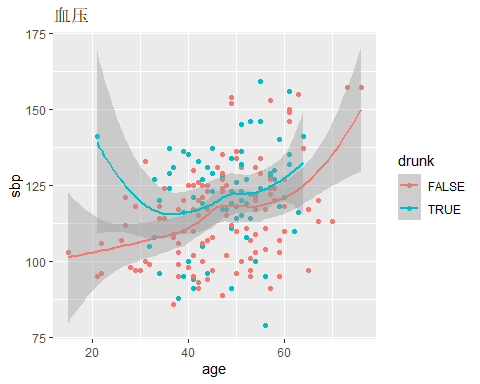
phy\_test %>% group\_by(drunk) %>%  
 summarise(mean(sbp)) %>% ungroup()

## `summarise()` ungrouping output (override with `.groups` argument)

## # A tibble: 2 x 2  
## drunk `mean(sbp)`  
## <lgl> <dbl>  
## 1 FALSE 116.  
## 2 TRUE 121.

ggplot(data = phy\_test)+  
 geom\_point(aes(x=age,y=sbp,color=drunk))+  
 geom\_smooth(aes(x=age,y=sbp,color=drunk))+  
 labs(title = "血压")

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



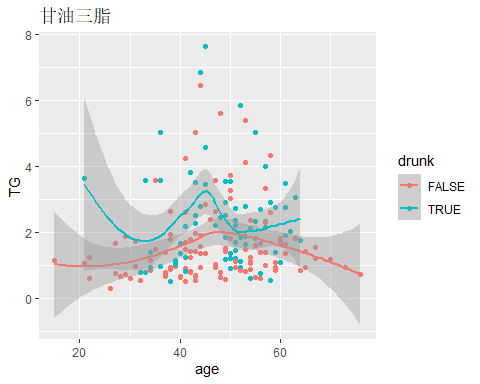
phy\_test %>% group\_by(drunk) %>%  
 summarise(mean(TG)) %>% ungroup()

## `summarise()` ungrouping output (override with `.groups` argument)

## # A tibble: 2 x 2  
## drunk `mean(TG)`  
## <lgl> <dbl>  
## 1 FALSE 1.60  
## 2 TRUE 2.29

ggplot(data = phy\_test)+  
 geom\_point(aes(x=age,y=TG,color=drunk))+  
 geom\_smooth(aes(x=age,y=TG,color=drunk))+  
 labs(title = "甘油三脂")

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



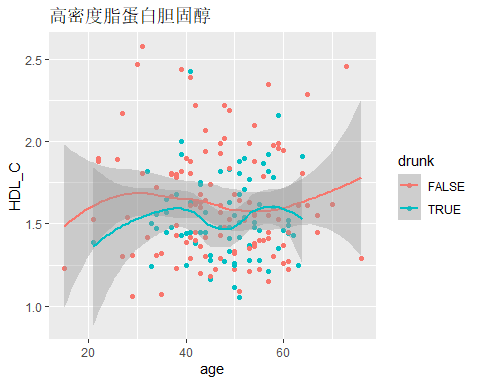
phy\_test %>% group\_by(drunk) %>%  
 summarise(mean(HDL\_C)) %>% ungroup()

## `summarise()` ungrouping output (override with `.groups` argument)

## # A tibble: 2 x 2  
## drunk `mean(HDL\_C)`  
## <lgl> <dbl>  
## 1 FALSE 1.63  
## 2 TRUE 1.55

ggplot(data = phy\_test)+  
 geom\_point(aes(x=age,y=HDL\_C,color=drunk))+  
 geom\_smooth(aes(x=age,y=HDL\_C,color=drunk))+  
 labs(title = "高密度脂蛋白胆固醇")

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



## 简单研究一下收缩压与舒张压之间的回归

先以收缩压为解释变量，舒张压为被解释变量进行线性回归

lm(data = phy\_test,sbp~dbp)

##   
## Call:  
## lm(formula = sbp ~ dbp, data = phy\_test)  
##   
## Coefficients:  
## (Intercept) dbp   
## 27.603 1.192

然后自己写一个求样本均方误差的函数mse，然后用nlm函数去拟合幂函数的回归形式

mse=function(x){  
 a=phy\_test$sbp  
 b=phy\_test$dbp  
 return(mean((a-x[1]-x[2]\*b^x[3])^2))  
}  
nlm(mse,c(27,1,1.1))

## $minimum  
## [1] 97.92024  
##   
## $estimate  
## [1] 42.5304550 0.4148847 1.2013052  
##   
## $gradient  
## [1] -0.1426453 -28.2106348 -52.8886300  
##   
## $code  
## [1] 4  
##   
## $iterations  
## [1] 100

将两种回归形式的图做出，蓝色的是线性回归，红色的是非线性回归，发现此处线性回归与非线性回归相差不大。

ggplot(data = phy\_test)+  
 geom\_point(aes(x=dbp,y=sbp))+  
 geom\_line(aes(x=dbp,y=27.6+1.192\*dbp),col='blue')+  
 geom\_line(aes(x=dbp,y=42.5+0.414\*dbp^1.2),col='red')+  
 labs(title = "收缩压与舒张压的关系")

