HW4-Rohit-Thakur

Rohit Thakur

2/24/2020

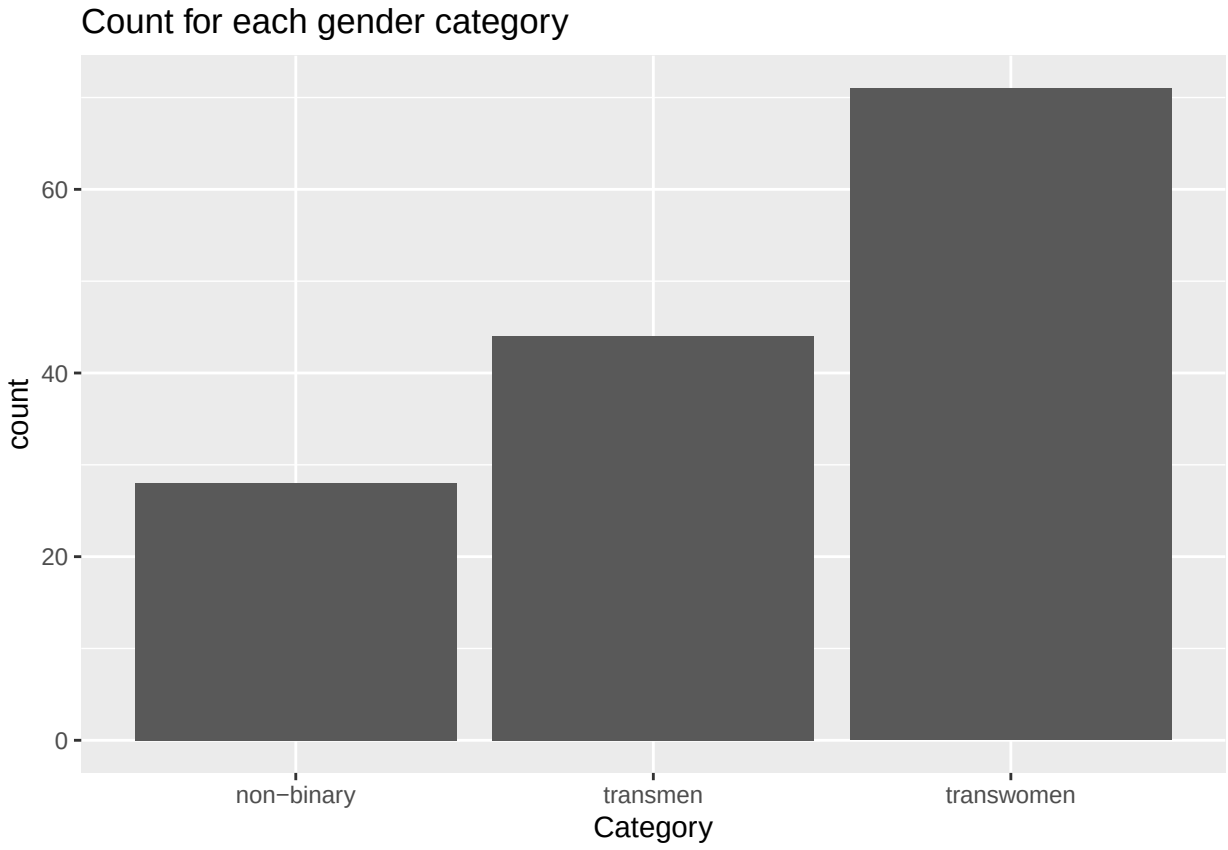
```
load("D:/Spring 20 Sem 2/DMP/ICPSR_31721/DS0001/31721-0001-Data.rda")
df<-da31721.0001
library(tidyr)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

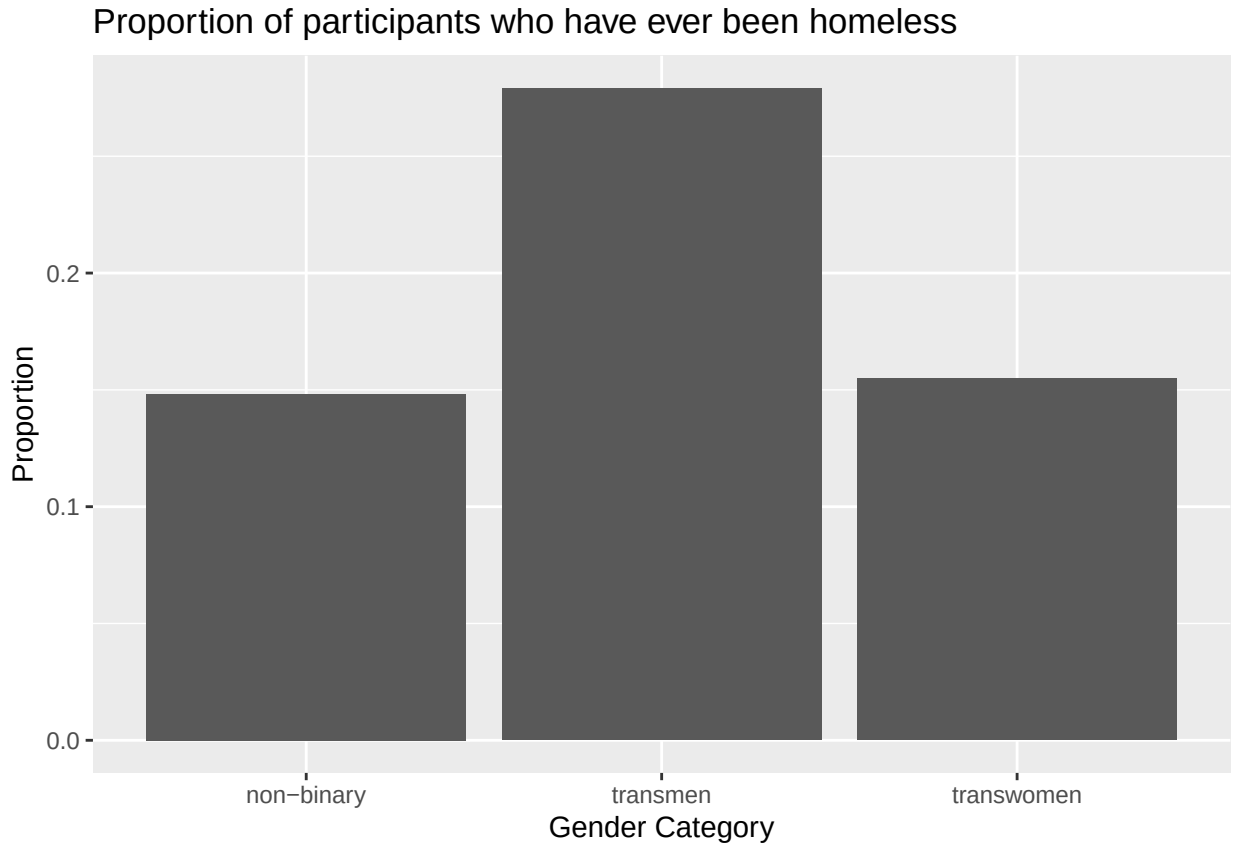
```
library(ggplot2)
```

PROBLEM 1

```
df1 <-
  mutate(df,Q6=ifelse(
      (df$Q6=="(1) Man" & df$Q5=="(2) Female"),"transmen",
      ifelse((df$Q6=="(2) Woman" & df$Q5=="(1) Male"),"transwomen",
      ifelse((df$Q6=="(4) Androgynous" | df$Q6=="(6) Gender Queer"),"non-binary",
      df$Q6))))

df1<- filter(df1,(Q6=="transmen" | Q6=="transwomen" | Q6=="non-binary"))
df1%>%ggplot()+
  geom_bar(aes(x=as.factor(Q6)))+
  xlab("Category")+
  ylab("count")+
  ggtitle("Count for each gender category")
```




```
df1<- filter(df1,Q88!="(10) Missing")
df1%>%dplyr::group_by(Q6)%>%
  summarise(y = sum(Q88=="(1) Yes"), n= sum(Q88 == "(2) No"))%>%
  mutate(prop=y/(y+n))%>%
ggplot() + geom_col(aes(x=Q6,y=prop))+
  xlab("Gender Category")+ylab("Proportion") +
  ggtitle("Proportion of participants who have ever been homeless")
```




```
#Proportion of homeless participants
df1%>%summarize(proportion_homeless=mean(Q88 == "(1) Yes"))
```

```
##   proportion_homeless
## 1           0.1914894
```

participants in our survey have experienced homelessness more than general US population. Problem 2

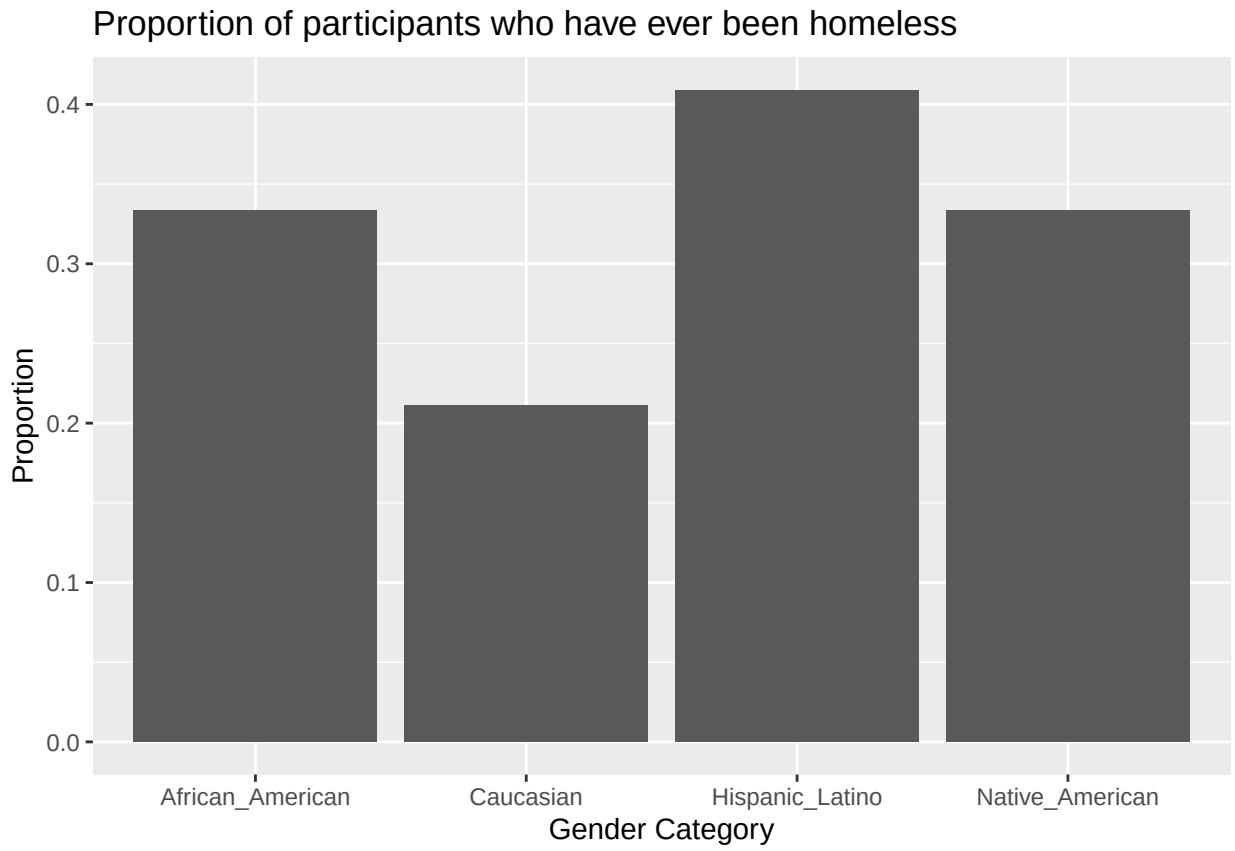
```
df2<-df
df2<-rename(df2,African_American=D9_1,Caucasian=D9_2,Hispanic_Latino=D9_3, Native_American=D9_4)
df2<-df2%>%
  gather(`African_American`,
         `Caucasian`,
         `Hispanic_Latino`,
         `Native_American`,
          key="Race",
          value="value")%>%
  filter(value=="(1) Selected")
summary(df2$Race)
```

```
##    Length     Class      Mode
##       368 character character
```

```r
df2<-df2%>% filter(Q88!="(10) Missing")
summary(df2$Race)
```

```
##    Length     Class      Mode
##       362 character character
```

```r
df2%>%
  group_by(Race)%>%
  summarise(y = sum(Q88=="(1) Yes"), n= sum(Q88 == "(2) No"))%>%
  mutate(prop=y/(y+n))%>%
  ggplot() + geom_col(aes(x=Race,y=prop))+
  xlab("Gender Category")+ylab("Proportion") +
  ggtitle("Proportion of participants who have ever been homeless")
```




```r
df2%>%summarize(proportion_homeless=mean(Q88 == "(1) Yes"))
```

```
##   proportion_homeless
## 1           0.2596685
```

From above proportion 25.96% participants from our study experience homelessness in their lifetime. This is greater than national US average.

```r
df3<-df
df3<-df3%>%mutate(Q133=ifelse(is.na(df3$Q133),"Missing",df3$Q133))
summary(df3$Q133)
```

```
##    Length     Class      Mode
##       350 character character
```

```r
#Mapped NAs to missing and yes response to 1 and no response to 2
df3%>%summarize(proportion_suicide_attempt=mean(df3$Q133=="1"))
```

```
##   proportion_suicide_attempt
## 1                  0.2542857
```

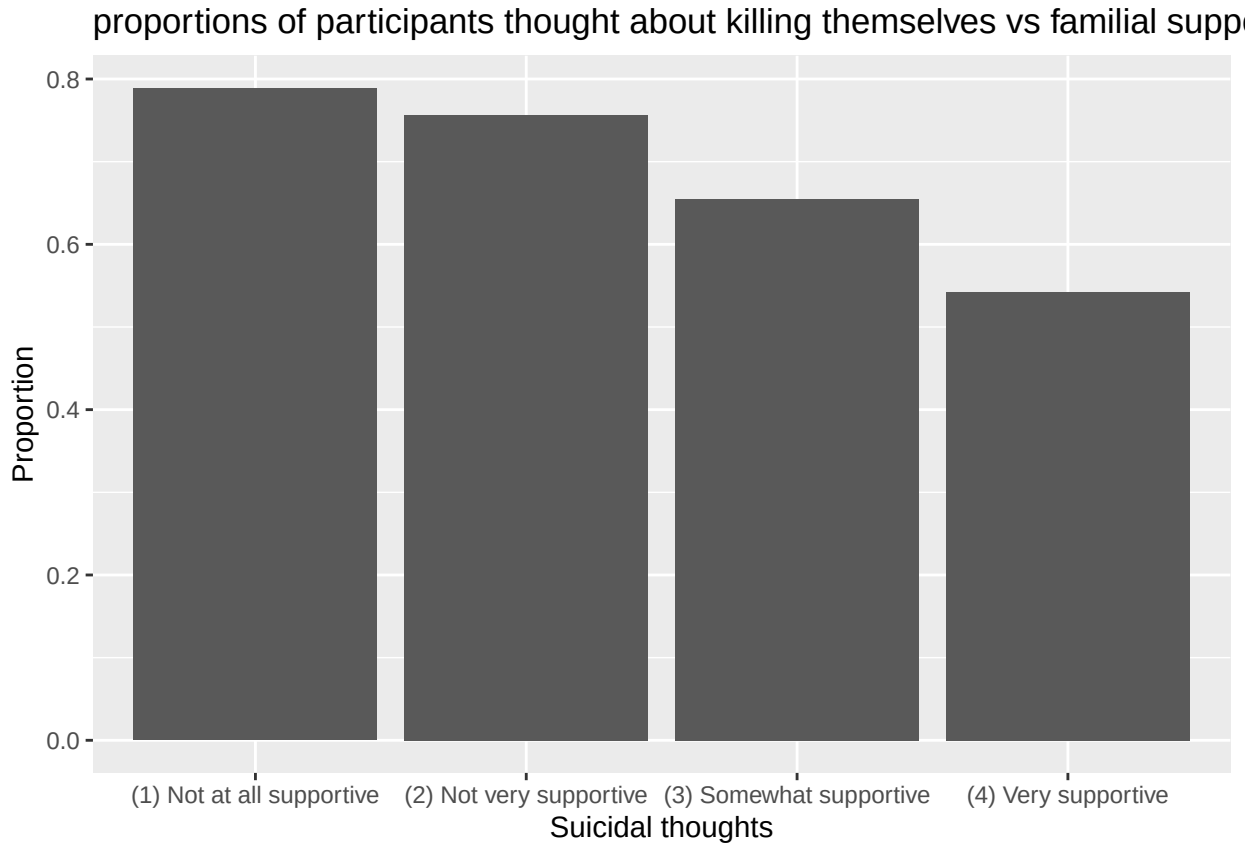
```r
summary(df3$Q119)
```

```
## (1) Not at all supportive   (2) Not very supportive   (3) Somewhat supportive
##                        71                         38                         88
##       (4) Very supportive  (5) Not applicable to me                       NA's
##                       120                         24                          9
```

```r
df4<-df3
df4<-df4[!is.na(df4$Q119),]
summary(df4$Q119)
```

```
## (1) Not at all supportive   (2) Not very supportive   (3) Somewhat supportive
##                        71                         38                         88
##       (4) Very supportive  (5) Not applicable to me
##                       120                         24
```

```r
df4<-filter(df4,Q119!="(5) Not applicable to me")
df4<-df4[!is.na(df4$Q131),]
df4%>%group_by(Q119)%>%
  summarise(y = sum(Q131=="(1) Yes"), n= sum(Q131 == "(2) No"))%>%
  mutate(prop=y/(y+n))%>%
  ggplot()+geom_col(aes(x=Q119,y=prop))+
  xlab("Suicidal thoughts")+ylab("Proportion")+
  ggtitle("proportions of participants thought about killing themselves vs familial support")
```



From above proportion of participants who have attempted suicide in Virginia in this survey is 25.14% which is below national average for trans people but much higher than average proportion of general population.

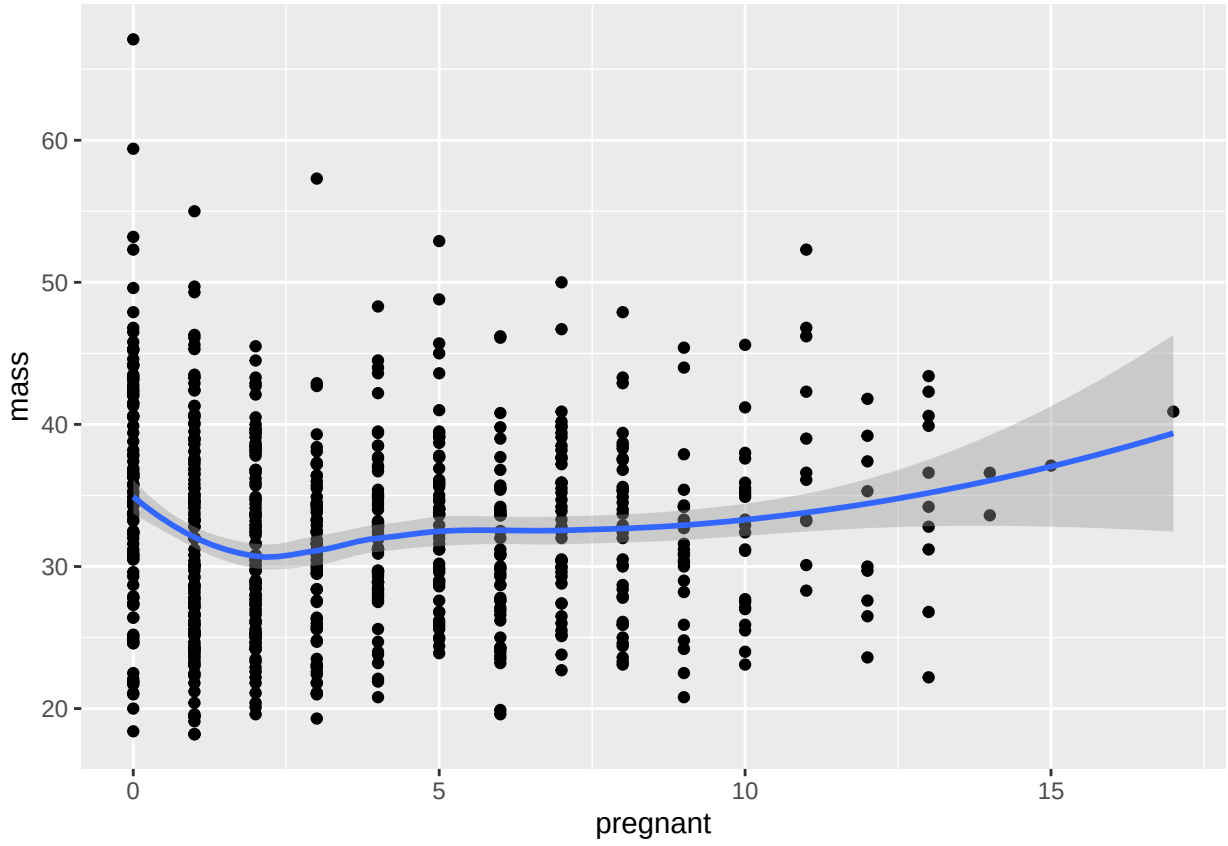
From above graph for participants who have thought of attempting suicide vs familial support levels we can conclude that there is higher percentage of participants who have positive suicidal thoughts overall in all familial support categories. But family support surely reduces percentage of population of participants that are having positive suicidal thoughts. Highest percentage of participants that have positive suicidal thoughts can be seen in participants with no supoortive family. This proportion get reduced as the family support level increases.

Problem 4

```
library(mlbench)
library(modelr)
```

```
data("PimaIndiansDiabetes2")
#First we will select pregnant column for mass.
PimaIndiansDiabetes2%>%ggplot(aes(x=pregnant,y=mass))+geom_point()+geom_smooth()
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'

## Warning: Removed 11 rows containing non-finite values (stat_smooth).

## Warning: Removed 11 rows containing missing values (geom_point).
```

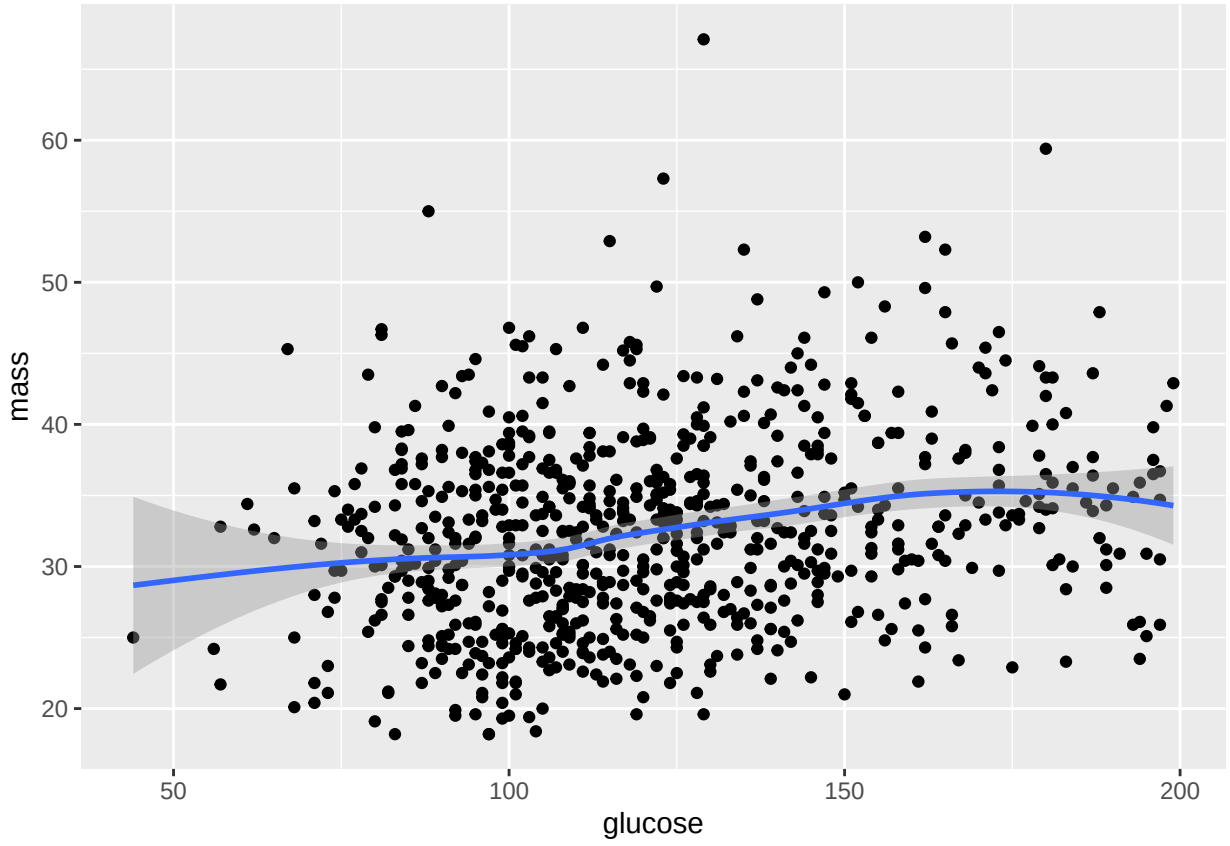



```
#Glucose vs mass
PimaIndiansDiabetes2%>%ggplot(aes(x=glucose,y=mass))+geom_point()+geom_smooth()
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```

```
## Warning: Removed 16 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 16 rows containing missing values (geom_point).
```

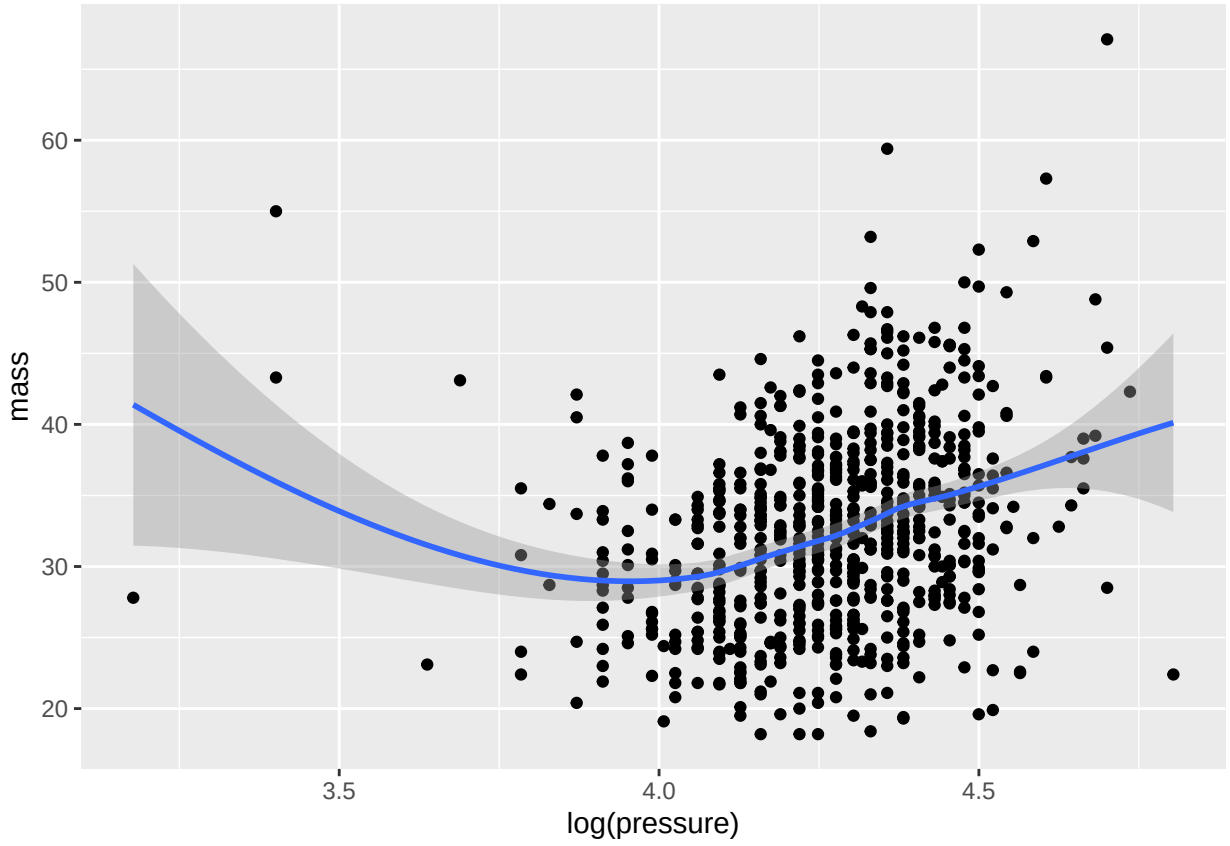



```
#Pressure vs mass
PimaIndiansDiabetes2%>%ggplot(aes(x=log(pressure),y=mass))+geom_point()+geom_smooth()
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```

```
## Warning: Removed 39 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 39 rows containing missing values (geom_point).
```

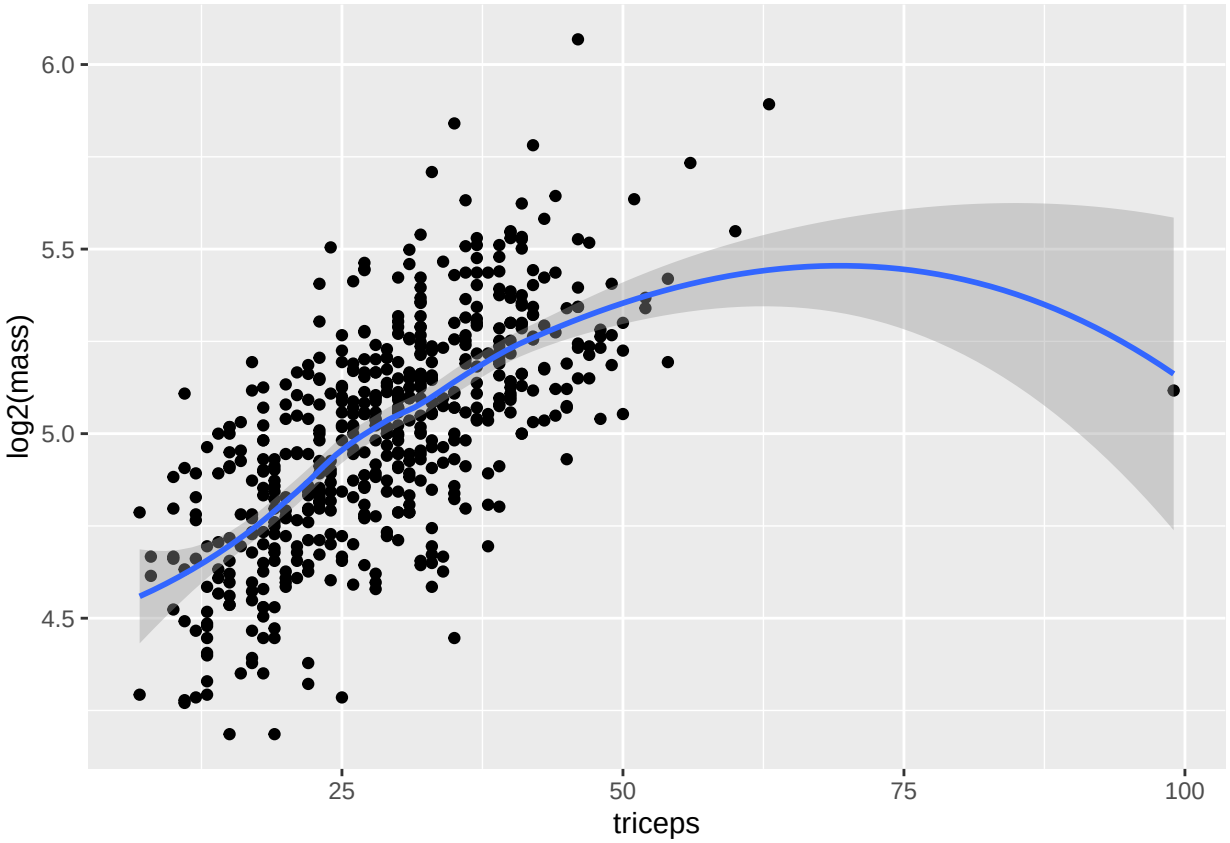



```
#from above graph there is a linear relationship with pressure and mass. Log og predictor variable
#better helps in visualizing this relationship.
#Triceps vs mass

PimaIndiansDiabetes2%>%ggplot(aes(x=triceps,y=log2(mass)))+geom_point()+geom_smooth()
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```

```
## Warning: Removed 229 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 229 rows containing missing values (geom_point).
```

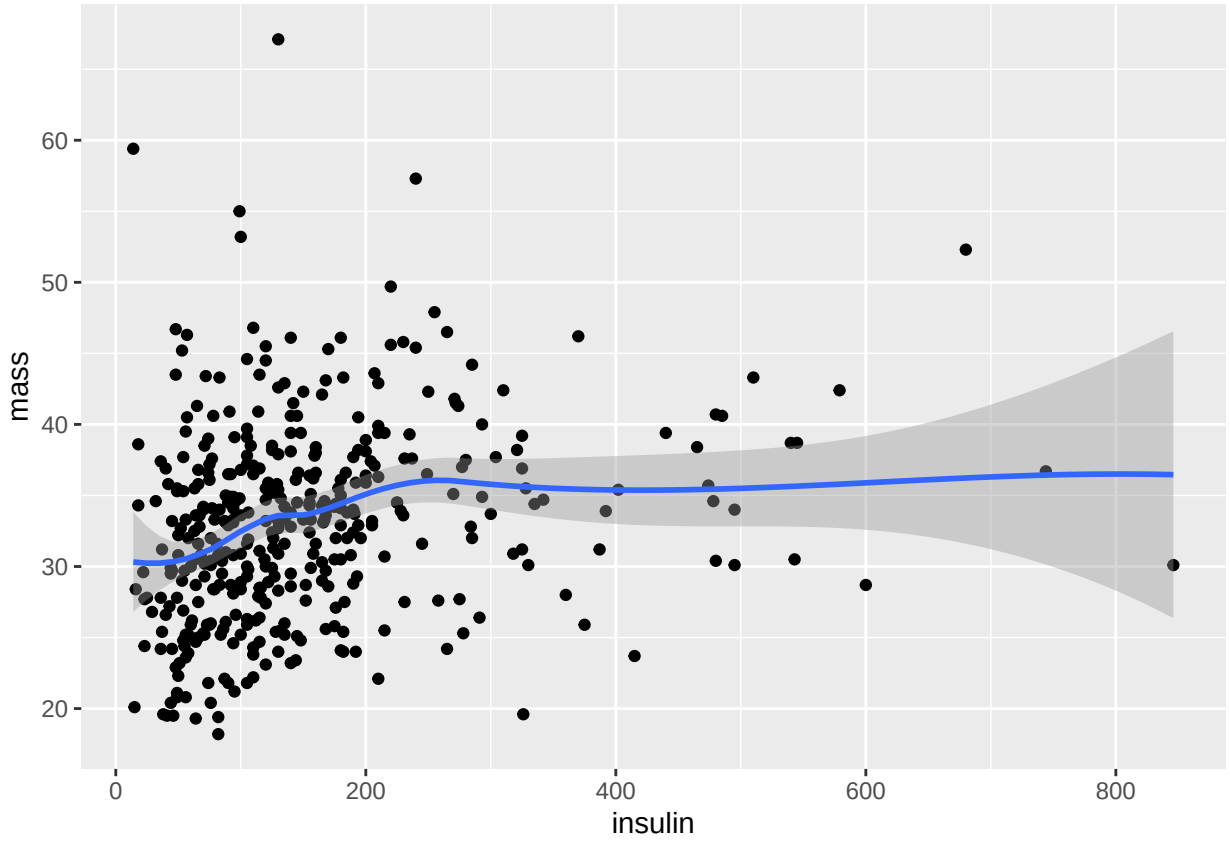



```
#From above we can see a linear relationship between log of response variable and triceps
#log of response variable better depicts above relationship.

#Insulin vs mass
PimaIndiansDiabetes2%>%ggplot(aes(x=insulin,y=mass))+geom_point()+geom_smooth()
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```

```
## Warning: Removed 375 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 375 rows containing missing values (geom_point).
```

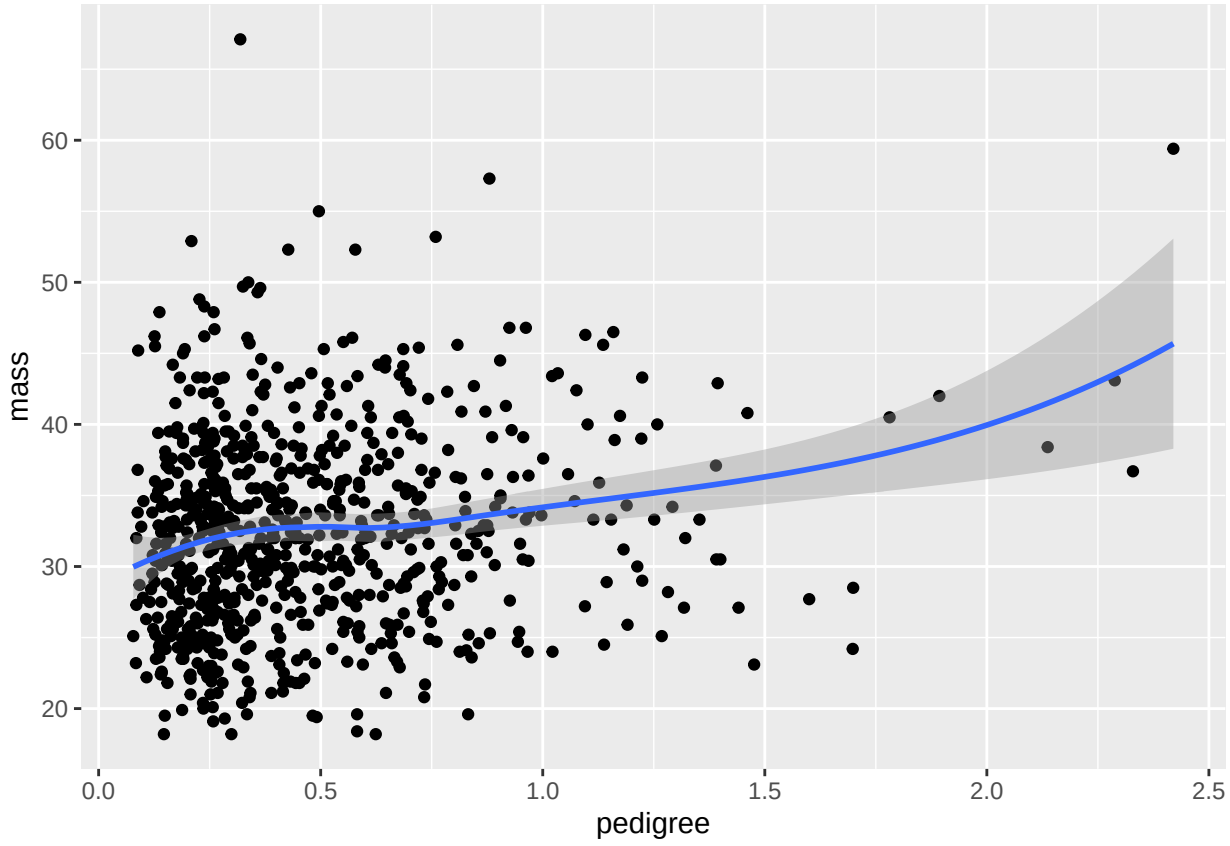



```
#pedigree vs mass
PimaIndiansDiabetes2%>%ggplot(aes(x=pedigree,y=mass))+geom_point()+geom_smooth()
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```

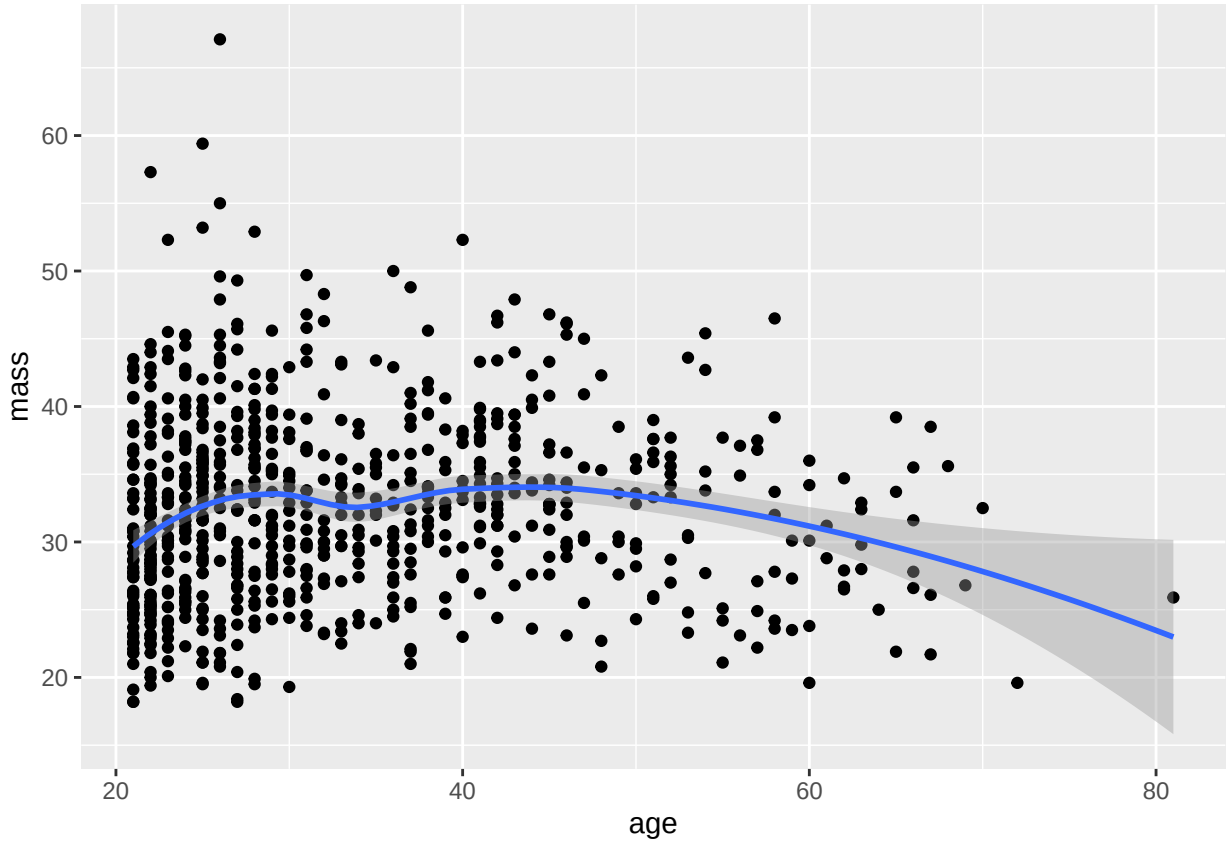
```
## Warning: Removed 11 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 11 rows containing missing values (geom_point).
```



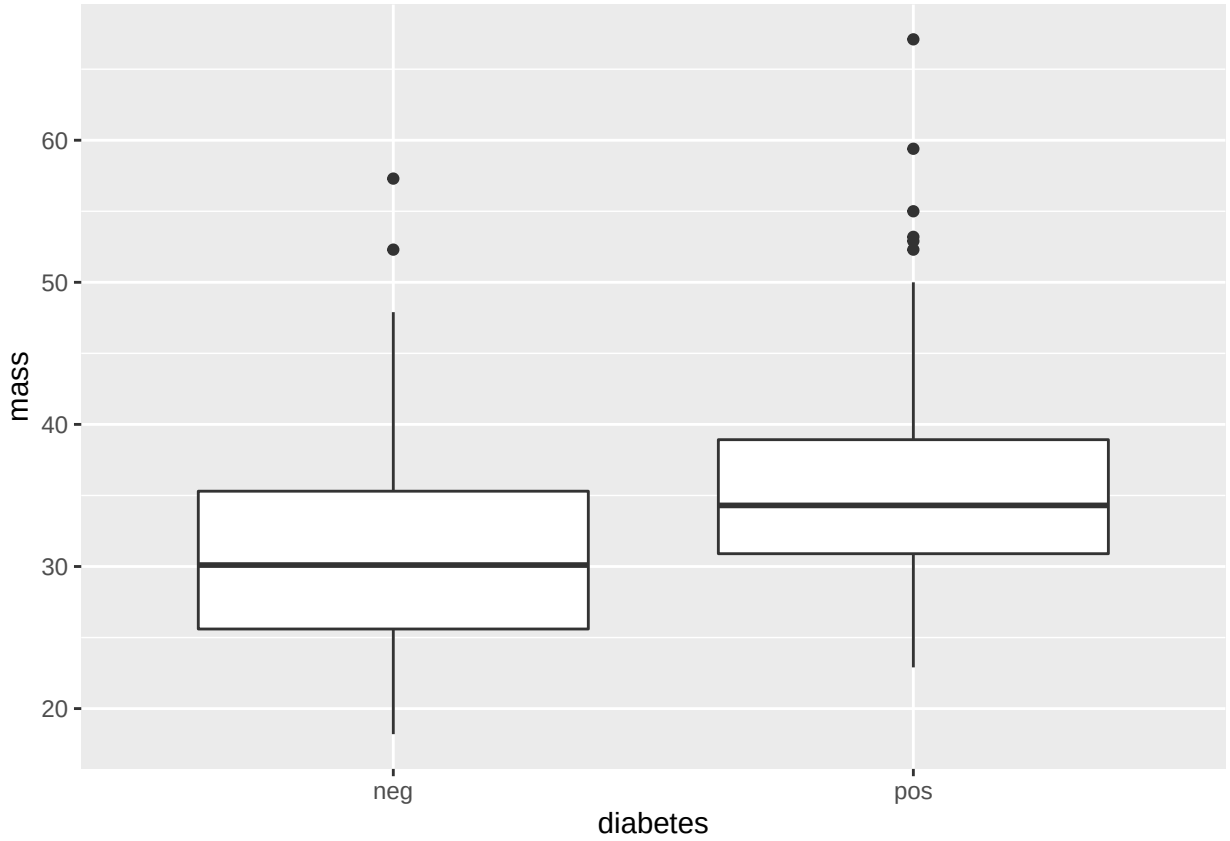

```
#age vs mass
PimaIndiansDiabetes2%>%ggplot(aes(x=age,y=mass))+geom_point()+geom_smooth()
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'

## Warning: Removed 11 rows containing non-finite values (stat_smooth).

## Warning: Removed 11 rows containing missing values (geom_point).
```




```
#diabetes vs mass
PimaIndiansDiabetes2%>%ggplot(aes(x=diabetes,y=mass))+geom_boxplot()
```

```
## Warning: Removed 11 rows containing non-finite values (stat_boxplot).
```



From above plots we can see that there is a linear relationship between (pressure,mass) and (triceps,mass). Thus we select these two columns as predictor variables. There are simply not enough categories to decide whether relationship between diabetes and mass is linear. Now fitting linear model for above two variables:

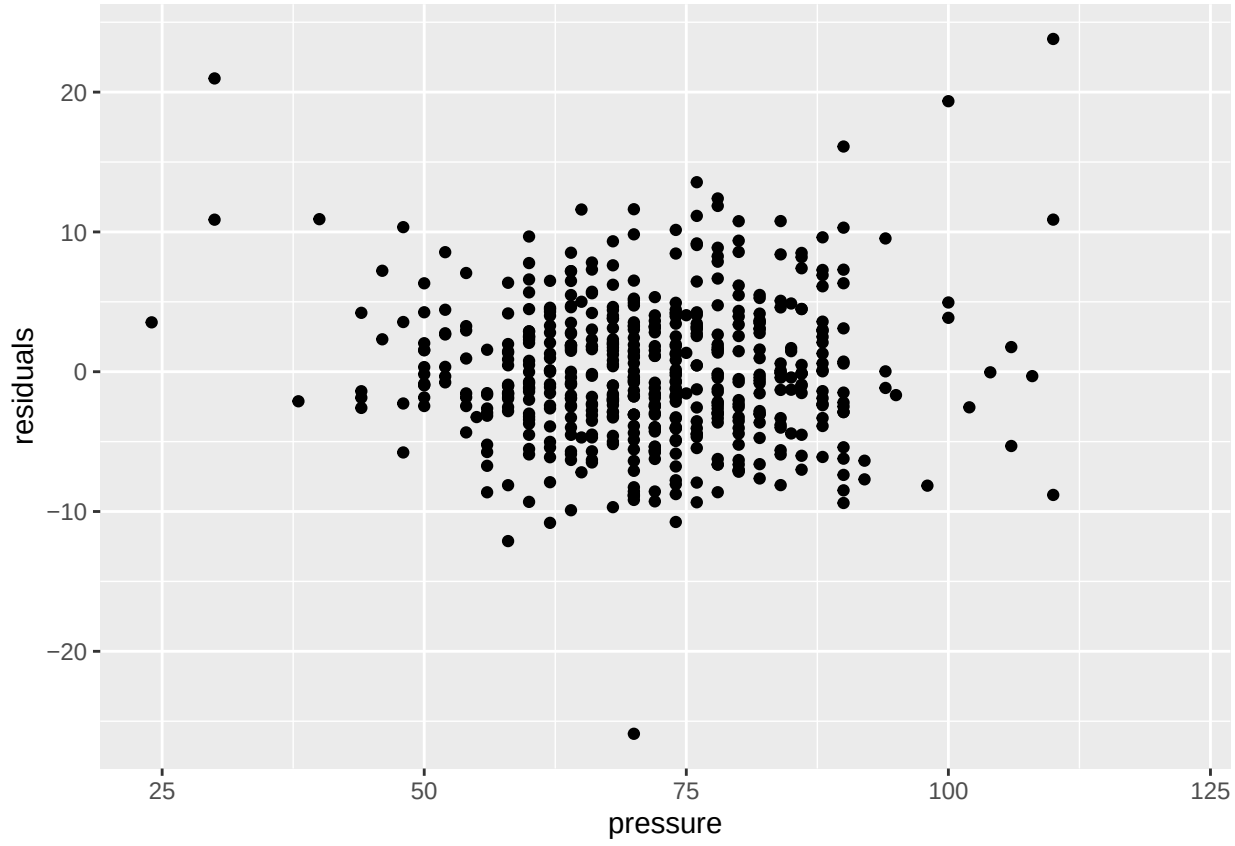
```
fit_1<-lm(mass~pressure+triceps,data=PimaIndiansDiabetes2)
summary(fit_1)
```

```
##
## Call:
## lm(formula = mass ~ pressure + triceps, data = PimaIndiansDiabetes2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -25.8992  -3.0776  -0.4222   3.0611  23.8029
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 14.38493    1.34402  10.703  < 2e-16 ***
## pressure     0.09602    0.01844   5.207 2.75e-07 ***
## triceps      0.39892    0.02159  18.473  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.116 on 534 degrees of freedom
##   (231 observations deleted due to missingness)
## Multiple R-squared:  0.4485, Adjusted R-squared:  0.4464
## F-statistic: 217.1 on 2 and 534 DF,  p-value: < 2.2e-16
```

Problem 5

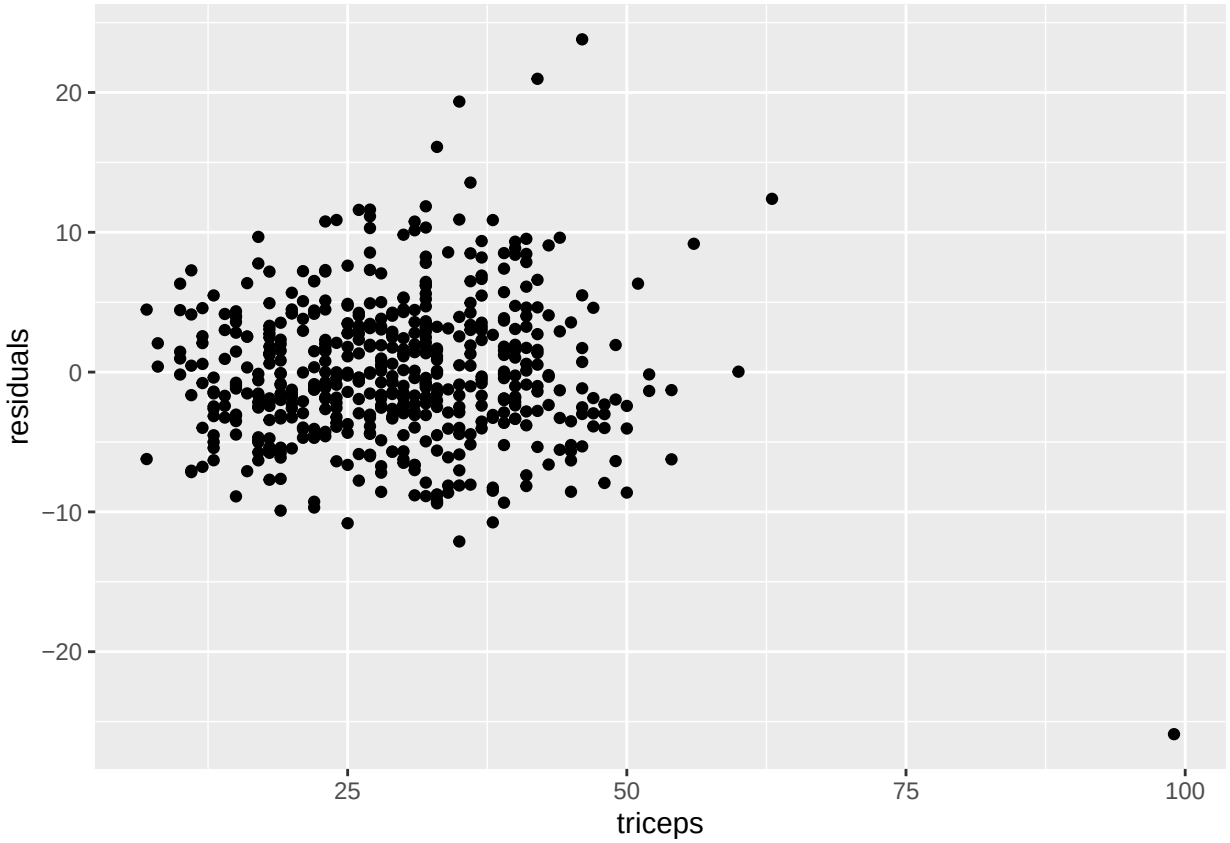
```
#residuals for predictor variable pressure
PimaIndiansDiabetes2%>%
add_residuals(fit_1, "residuals") %>%
ggplot(aes(x=pressure)) + geom_point(aes(y=residuals))
```

```
## Warning: Removed 231 rows containing missing values (geom_point).
```



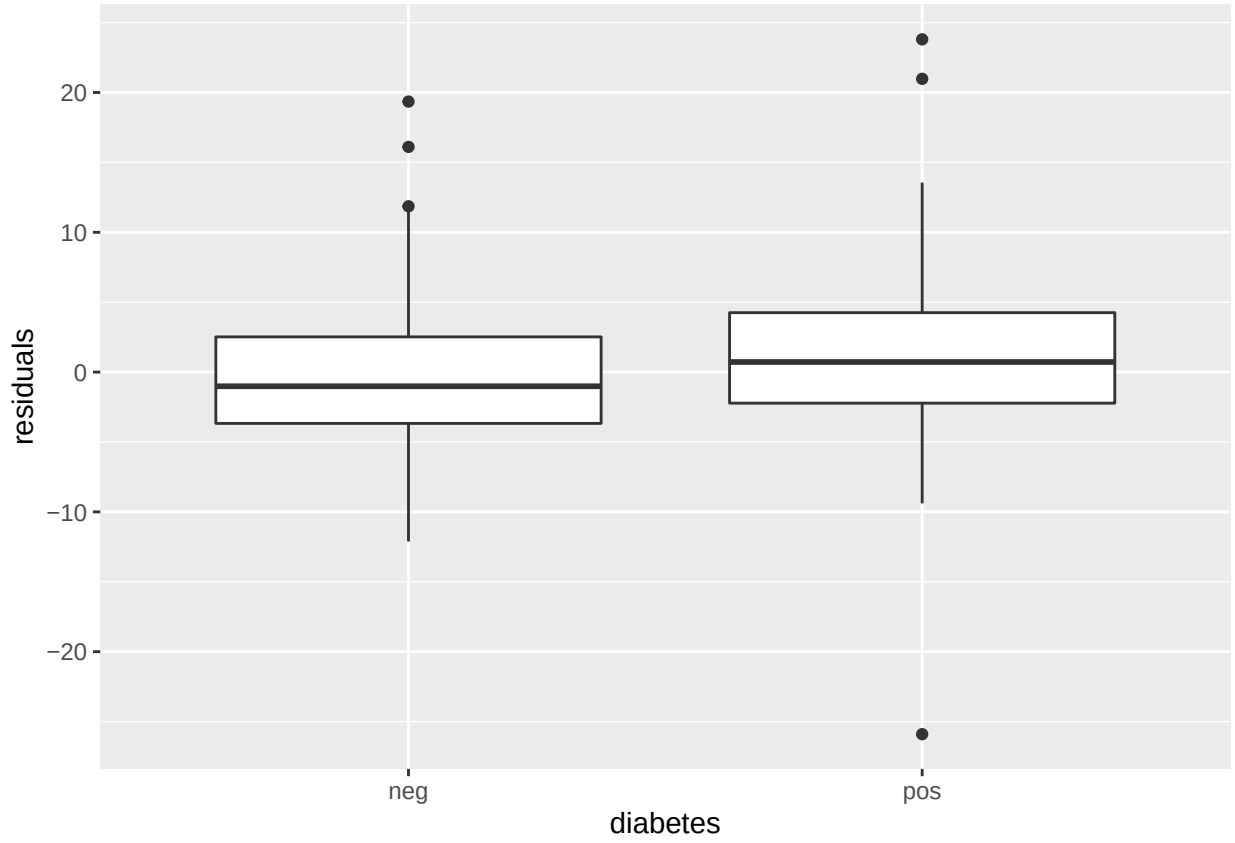

```
#residuals for predictor variable triceps

PimaIndiansDiabetes2%>%
add_residuals(fit_1, "residuals") %>%
ggplot(aes(x=triceps)) + geom_point(aes(y=residuals))
```

```
## Warning: Removed 231 rows containing missing values (geom_point).
```



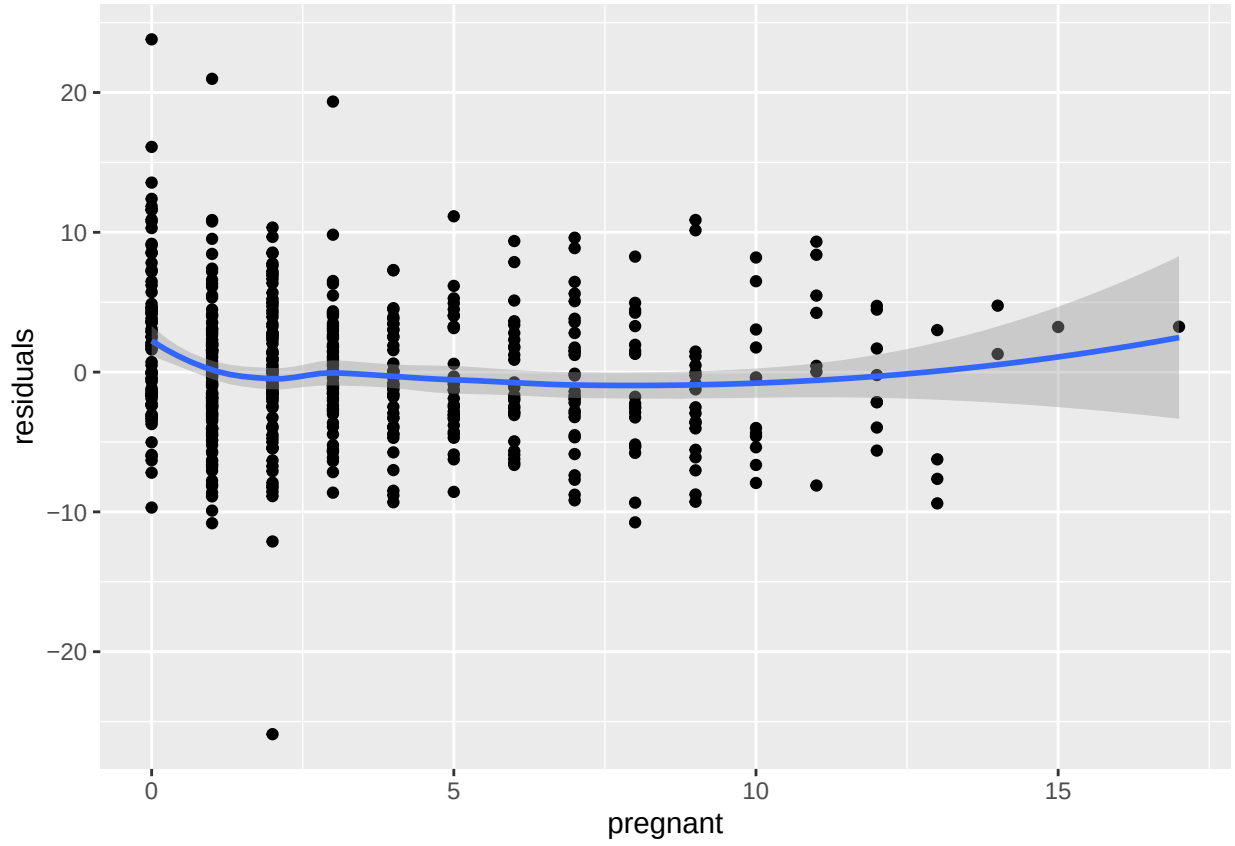

```
#From residual plots we can observe they are randomly dispersed around horizontal axis.
#This proves linear regression model is appropriate for this data.

#relationship between the residuals and the other potential predictor variables that are not
#currently in model

#for diabetes
PimaIndiansDiabetes2 %>%
add_residuals(fit_1, "residuals") %>%
ggplot(aes(diabetes, residuals)) + geom_boxplot()
```

```
## Warning: Removed 231 rows containing non-finite values (stat_boxplot).
```



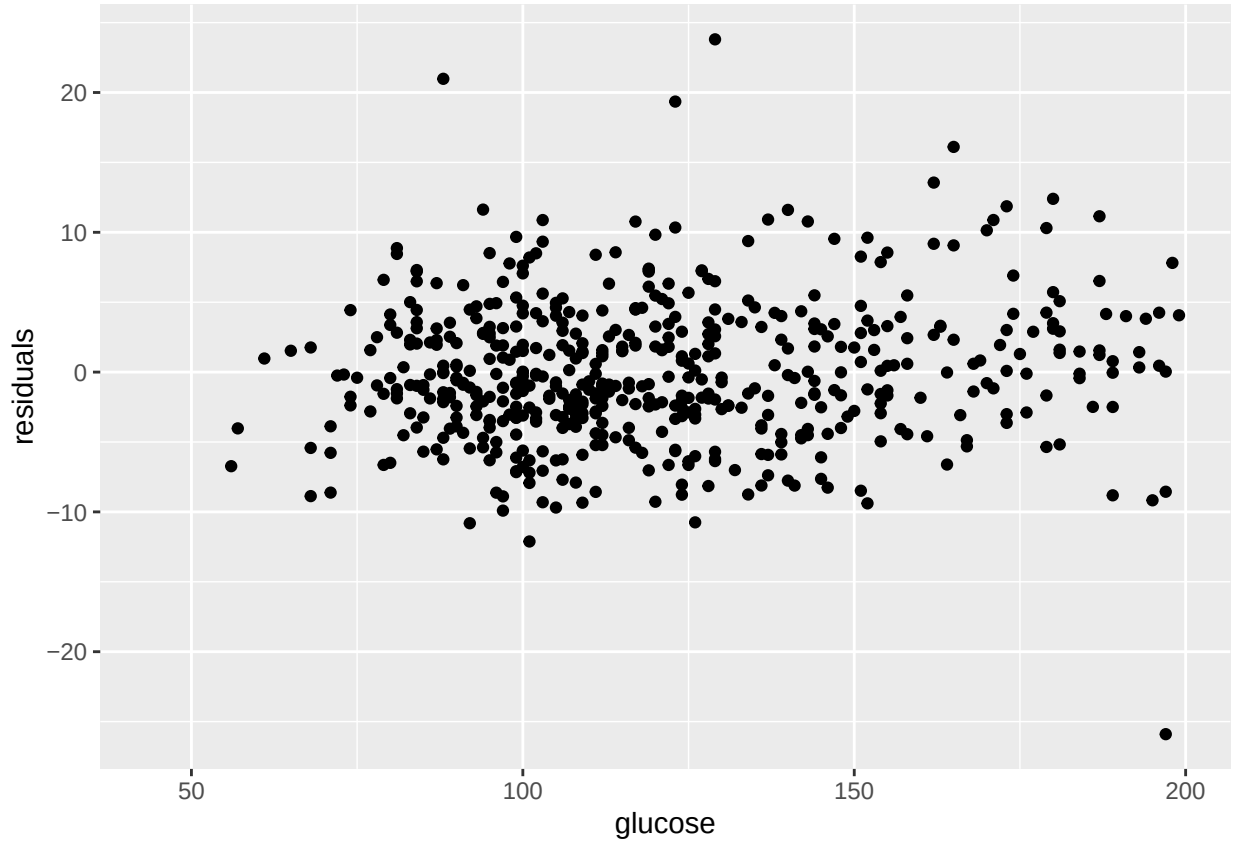

```
#for pregnant

PimaIndiansDiabetes2 %>%
add_residuals(fit_1, "residuals") %>%
ggplot(aes(pregnant, residuals)) + geom_point() + geom_smooth()
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'

## Warning: Removed 231 rows containing non-finite values (stat_smooth).

## Warning: Removed 231 rows containing missing values (geom_point).
```



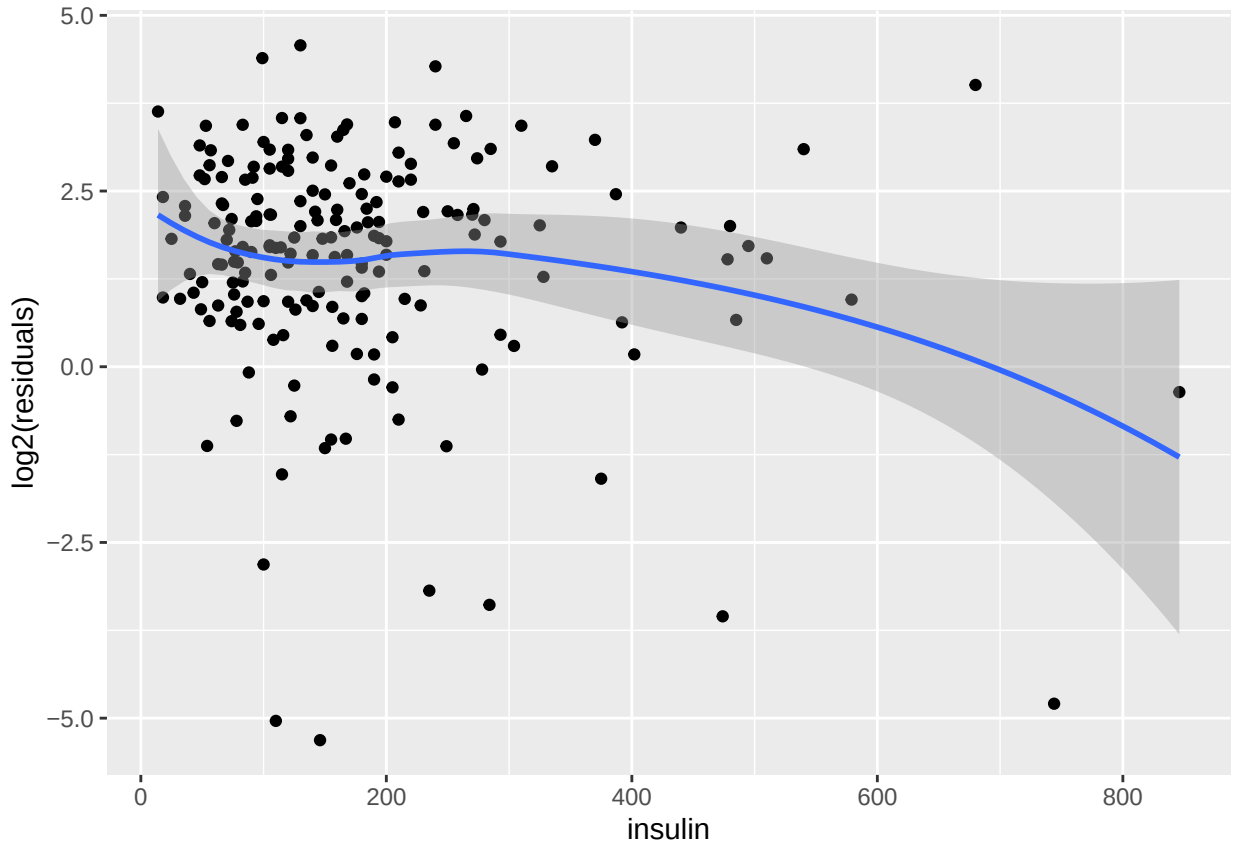

```
#for glucose

PimaIndiansDiabetes2 %>%
add_residuals(fit_1, "residuals") %>%
ggplot(aes(glucose, residuals)) + geom_point()
```

```
## Warning: Removed 236 rows containing missing values (geom_point).
```



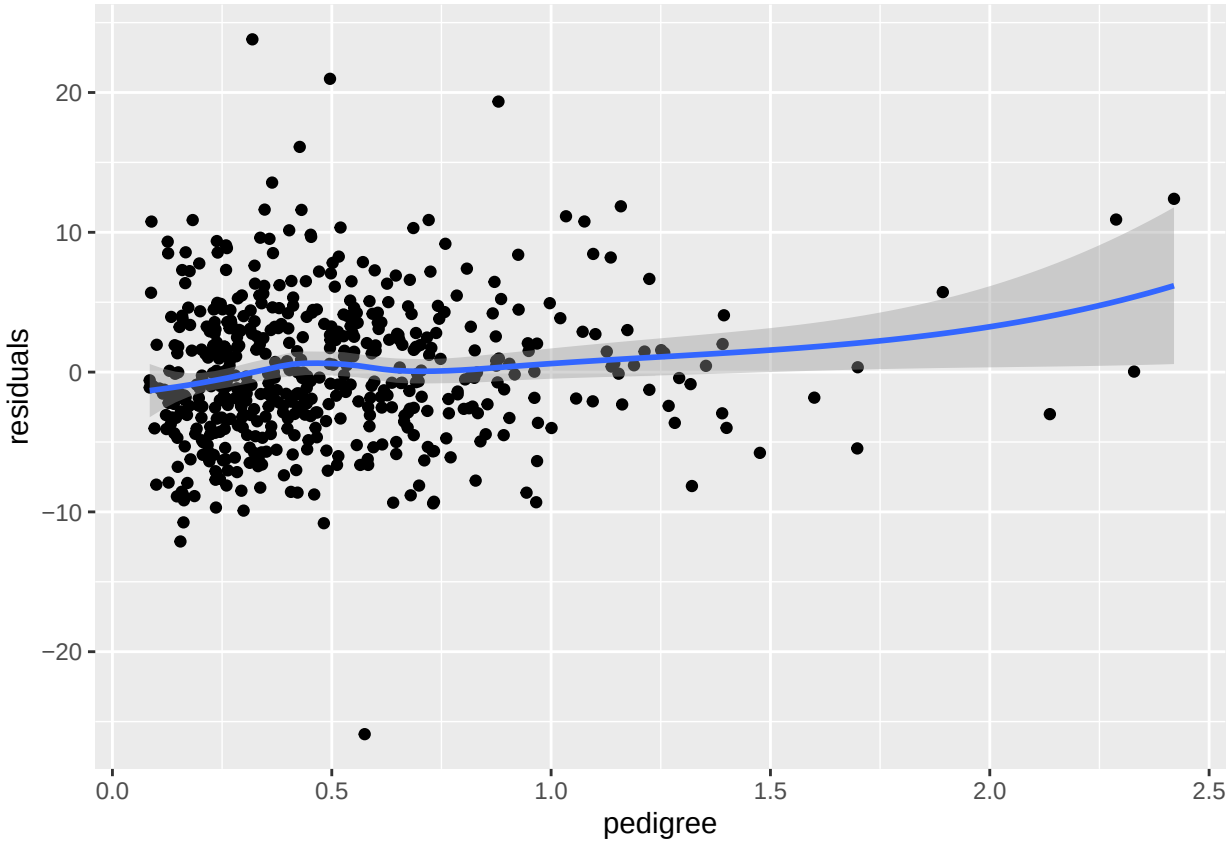

```
#for insulin

PimaIndiansDiabetes2 %>%
add_residuals(fit_1, "residuals") %>%
ggplot(aes(insulin, log2(residuals))) + geom_point() + geom_smooth()
```

```
## Warning in FUN(X[[i]], ...): NaNs produced

## Warning in FUN(X[[i]], ...): NaNs produced

## Warning in FUN(X[[i]], ...): NaNs produced

## `geom_smooth()` using method = 'loess' and formula 'y ~ x'

## Warning: Removed 578 rows containing non-finite values (stat_smooth).

## Warning: Removed 578 rows containing missing values (geom_point).
```



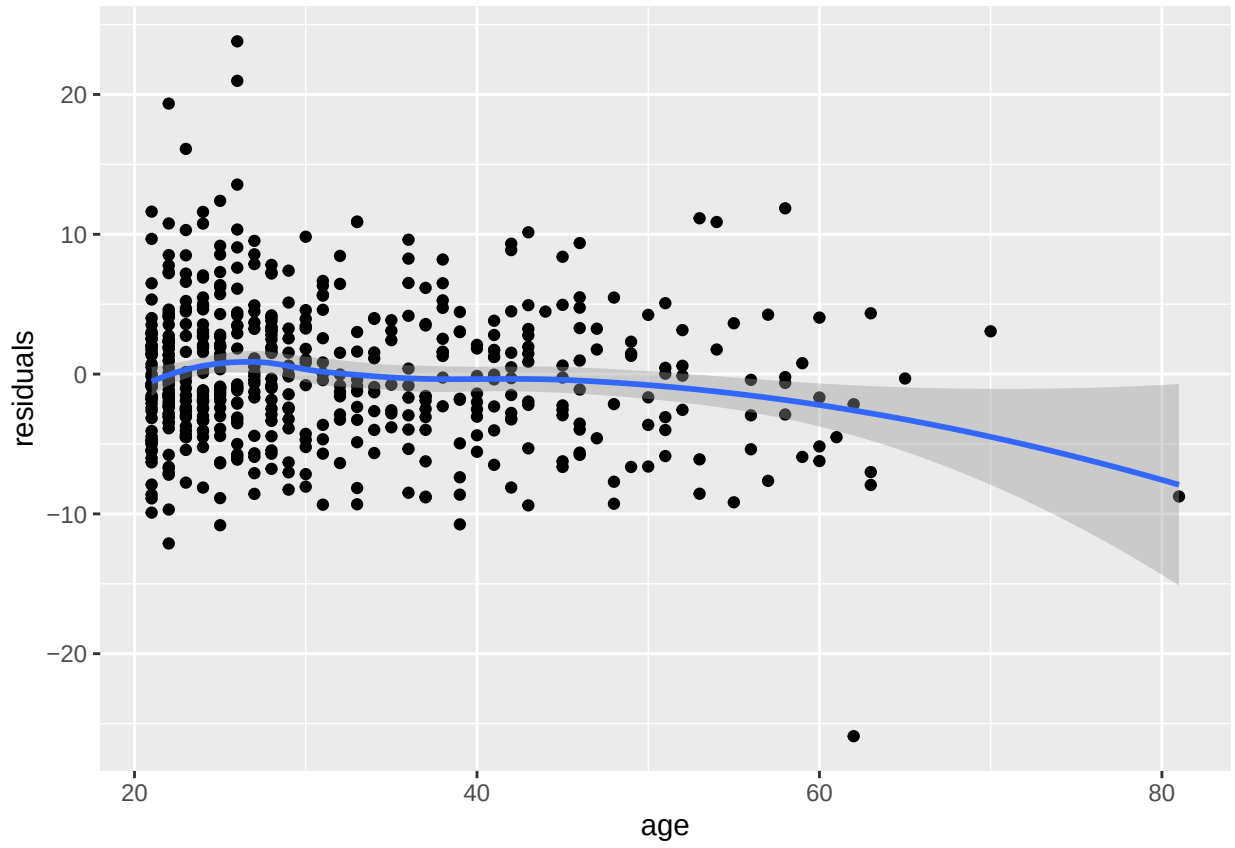

```
#for pedigree

PimaIndiansDiabetes2 %>%
add_residuals(fit_1, "residuals") %>%
ggplot(aes(pedigree, residuals)) + geom_point() + geom_smooth()
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'

## Warning: Removed 231 rows containing non-finite values (stat_smooth).

## Warning: Removed 231 rows containing missing values (geom_point).
```




```
#for age

PimaIndiansDiabetes2 %>%
add_residuals(fit_1, "residuals") %>%
ggplot(aes(age, residuals)) + geom_point() + geom_smooth()
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'

## Warning: Removed 231 rows containing non-finite values (stat_smooth).

## Warning: Removed 231 rows containing missing values (geom_point).
```




```
fit_3<-lm(mass~pressure+triceps+pregnant,data=PimaIndiansDiabetes2)
summary(fit_3)
```

```
##
## Call:
## lm(formula = mass ~ pressure + triceps + pregnant, data = PimaIndiansDiabetes2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -26.3649  -3.1533  -0.5769   3.0323  22.7713
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 14.25839    1.33764  10.659  < 2e-16 ***
## pressure     0.10528    0.01868   5.635 2.83e-08 ***
## triceps      0.40194    0.02151  18.687  < 2e-16 ***
## pregnant    -0.17771    0.06807  -2.611  0.00929 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.088 on 533 degrees of freedom
##   (231 observations deleted due to missingness)
## Multiple R-squared:  0.4555, Adjusted R-squared:  0.4524
## F-statistic: 148.6 on 3 and 533 DF,  p-value: < 2.2e-16
```

```r
# we will find rmse for two ,three, four variables

fit_4<-lm(mass~pressure+triceps+diabetes+pregnant,data=PimaIndiansDiabetes2)
summary(fit_4)
```

```
##
## Call:
## lm(formula = mass ~ pressure + triceps + diabetes + pregnant,
##     data = PimaIndiansDiabetes2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -26.4581  -3.3058  -0.5398   3.0874  21.6963
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 14.95739    1.32172  11.317  < 2e-16 ***
## pressure     0.09735    0.01842   5.285 1.83e-07 ***
## triceps      0.38006    0.02164  17.562  < 2e-16 ***
## diabetespos  2.25534    0.48971   4.605 5.16e-06 ***
## pregnant    -0.24752    0.06852  -3.613 0.000332 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.994 on 532 degrees of freedom
##   (231 observations deleted due to missingness)
## Multiple R-squared:  0.4764, Adjusted R-squared:  0.4724
## F-statistic:   121 on 4 and 532 DF,  p-value: < 2.2e-16
```

```r
rmse(fit_1,PimaIndiansDiabetes2)
```

```
## [1] 5.10139
```

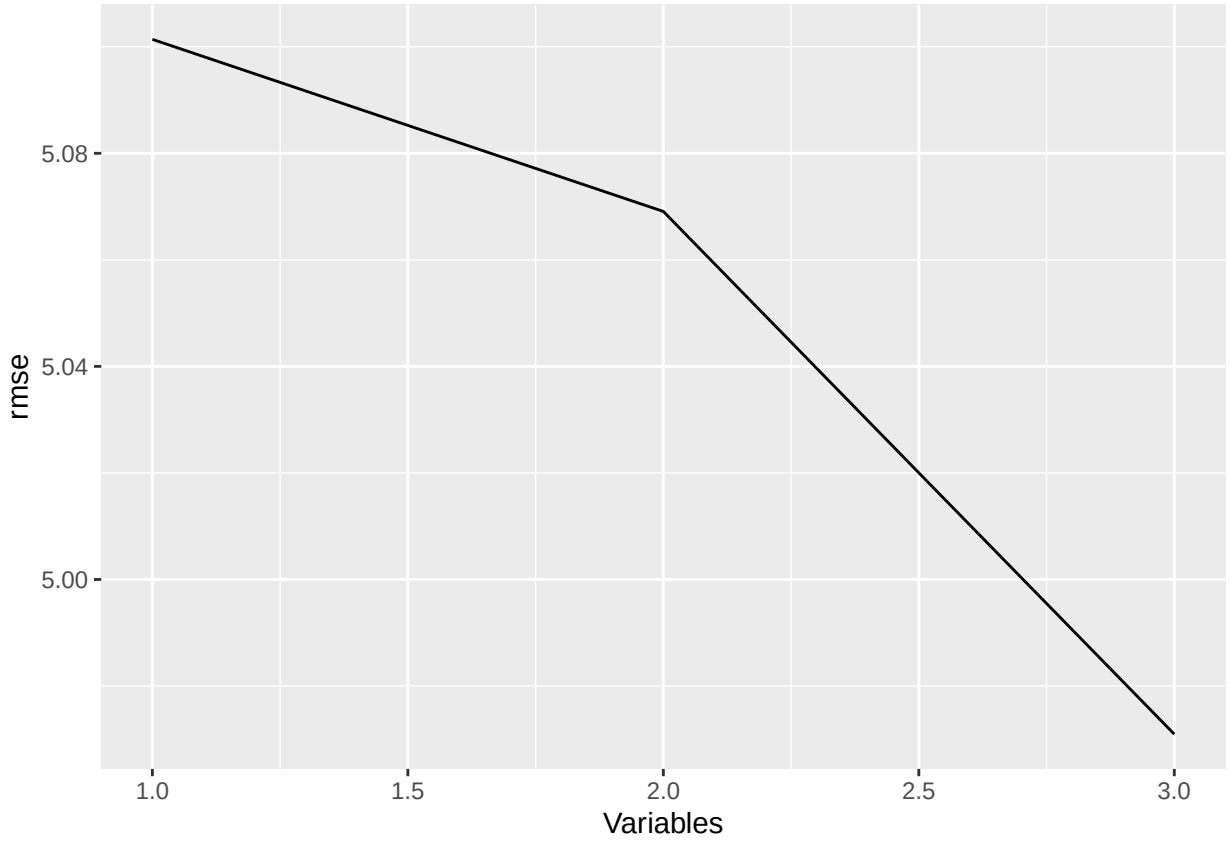
```r
rmse(fit_3,PimaIndiansDiabetes2)
```

```
## [1] 5.069083
```

```r
rmse(fit_4,PimaIndiansDiabetes2)
```

```
## [1] 4.970959
```

```r
# we can see rmse is decreasing as we add diabetes and pregnant predictor variable compared to
#two variable model.

#visualizing rmse
fits_rmse <- tibble(nvar = 1:3,
rmse = c(rmse(fit_1,PimaIndiansDiabetes2),
rmse(fit_3,PimaIndiansDiabetes2),
rmse(fit_4,PimaIndiansDiabetes2)))
ggplot(fits_rmse) + geom_line(aes(x=nvar, y=rmse)) +xlab("Variables")
```



From above residual plots we cannot see any definitive systematic pattern for variables. But for diabetes and pregnant variables there is some pattern. we'll add this predictor variables to our model. After calculating rmse for three and four variables model we compare it and find value to be lower.

Plotting RMSE we get linear decrease in rmse for these three fits.