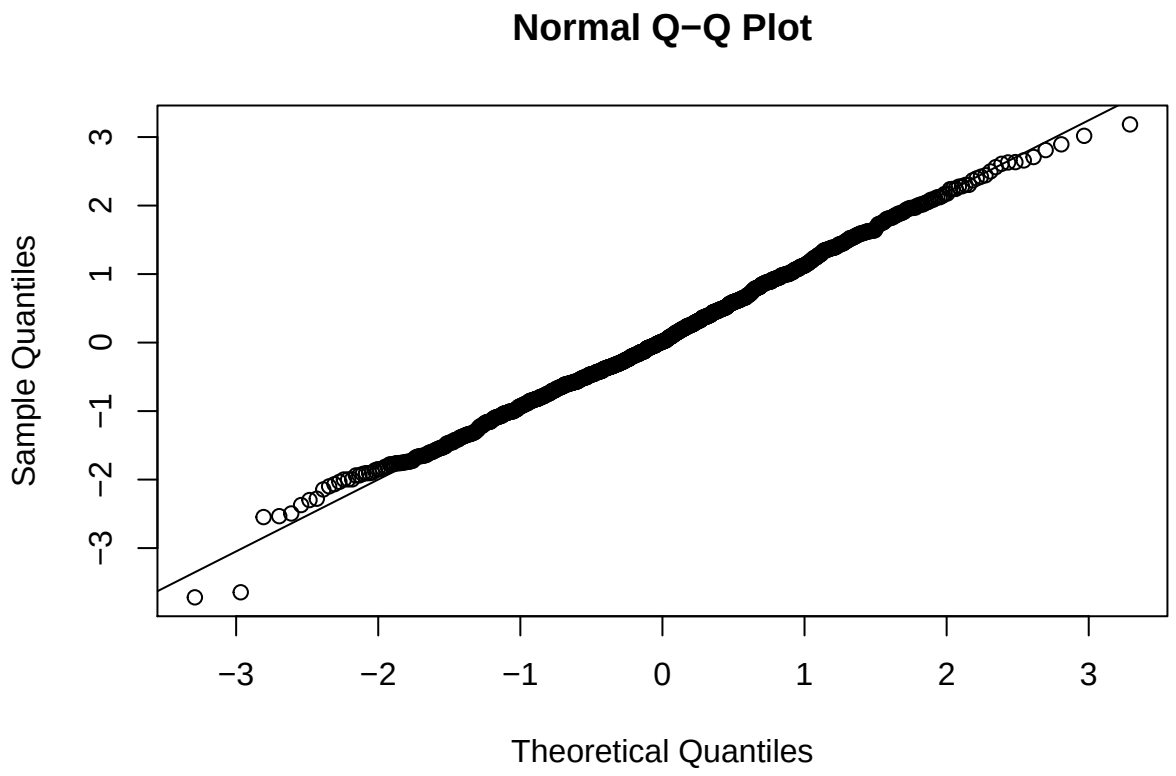
11/26 APLM Residual analysis

san teng

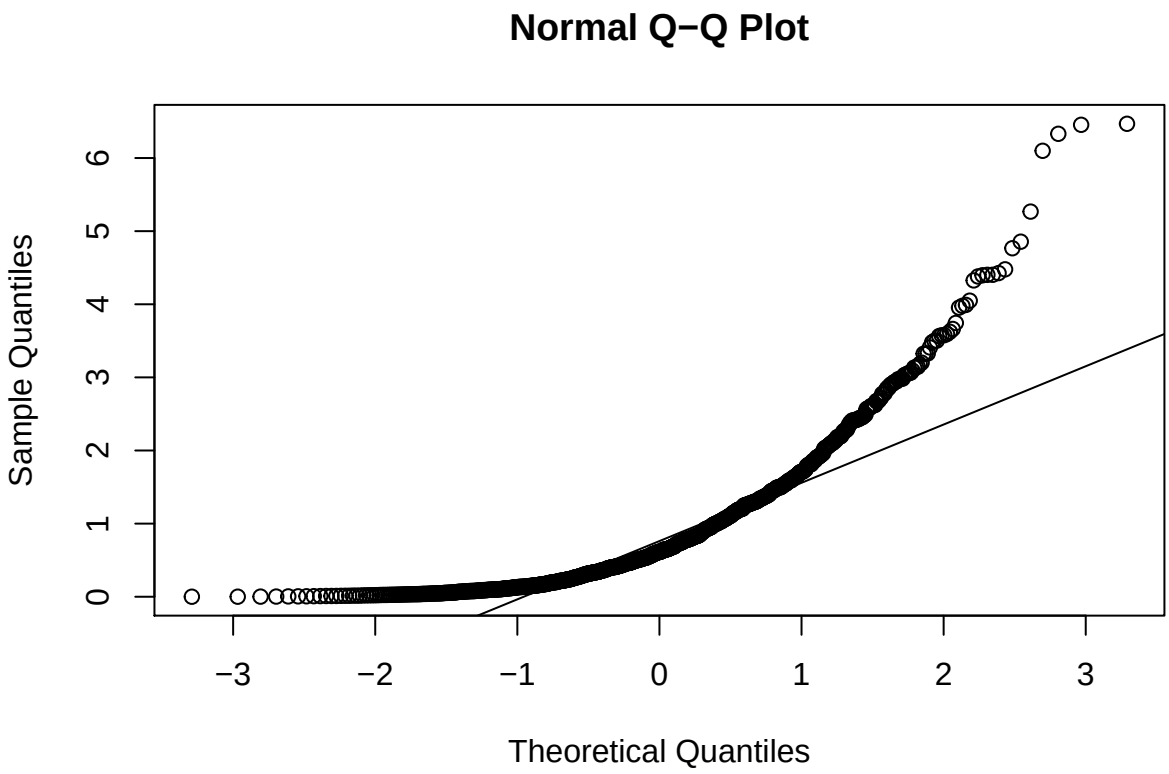
2018/11/26

(Problem 1)

```
##Genrate data from normal distribution
x <- rnorm(1000)
qqnorm(x) ## Normal probability plot
qqline(x)
```




```
##Genrate data from exponential distribution
x <- rexp(1000, rate = 1)
qqnorm(x) ## Normal probability plot
qqline(x)
```



(Q) Please interpret the above Normal Q-Q plots ?

Ans: if data comes from normal distrbution, they will lie on the straight line.

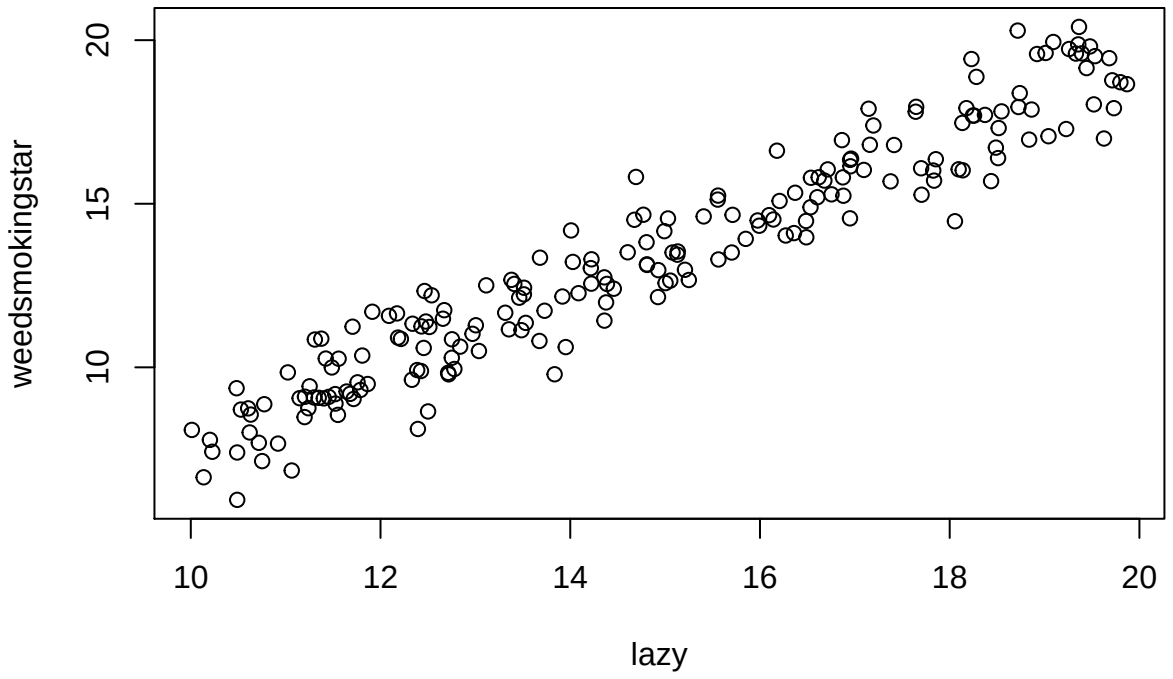
(Problem 2)

```
n<-200 ##The size of observed data
sport <- runif(n, 2, 6)
lazy <- runif(n, 10, 20)
lazy2 <- runif(n, 2, 4)
lazy3 <- runif(n, 2, 5)
lazy4 <- runif(n, 3, 7)

# true model:
error <- rnorm(n) ##random error
weedsmokingstar <- -3.5 - 0.25 * sport + 1.2 * lazy + error

##linear regression fit the data by least square
fit1 <- lm(weedsmokingstar ~ sport + lazy) ##fit regression model with
##predictor variables "sport"," lazy"
summary(fit1) ##show the summary information of least square estimator
```

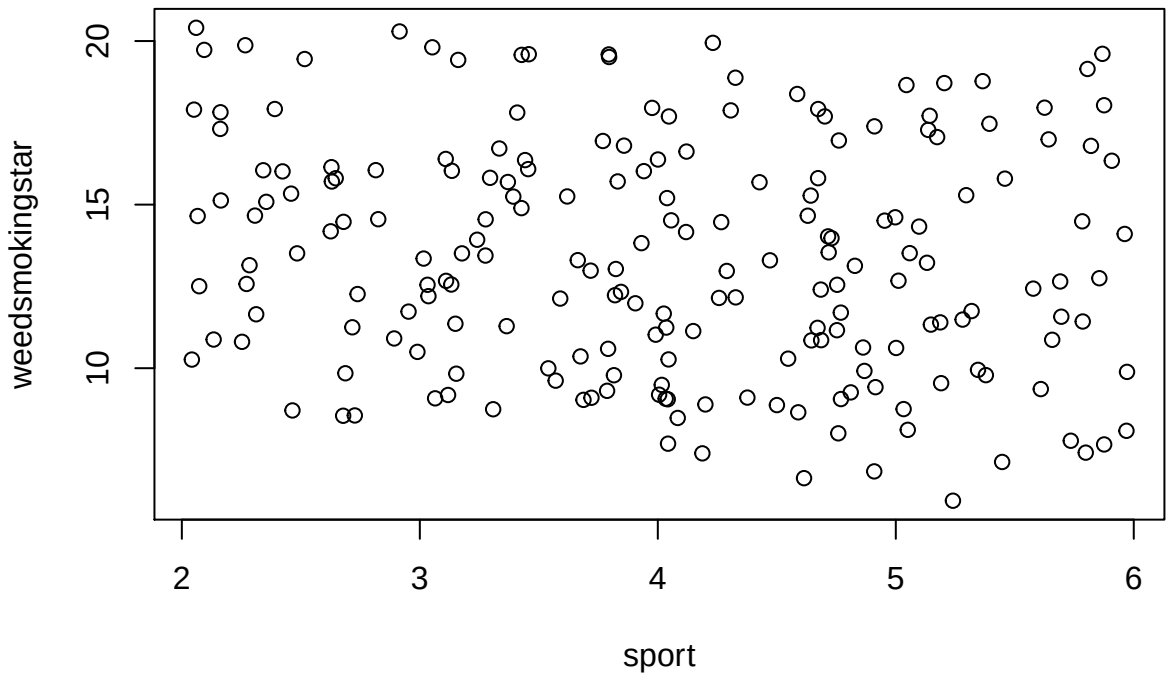
```
##
## Call:
## lm(formula = weedsmokingstar ~ sport + lazy)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
```

```
## -2.64599 -0.66038 -0.08455  0.71969  2.50324
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3.43678    0.46732  -7.354 5.04e-12 ***
## sport       -0.17688    0.06305  -2.805  0.00553 **
## lazy         1.17985    0.02423  48.686  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.972 on 197 degrees of freedom
## Multiple R-squared:  0.9252, Adjusted R-squared:  0.9244
## F-statistic:  1218 on 2 and 197 DF,  p-value: < 2.2e-16
```

```
plot(lazy, weedsmokingstar) #### scatter plot lazy v.s weedsmokingstar
```




```
plot(sport, weedsmokingstar) #### scatter plot sport v.s weedsmokingstar
```

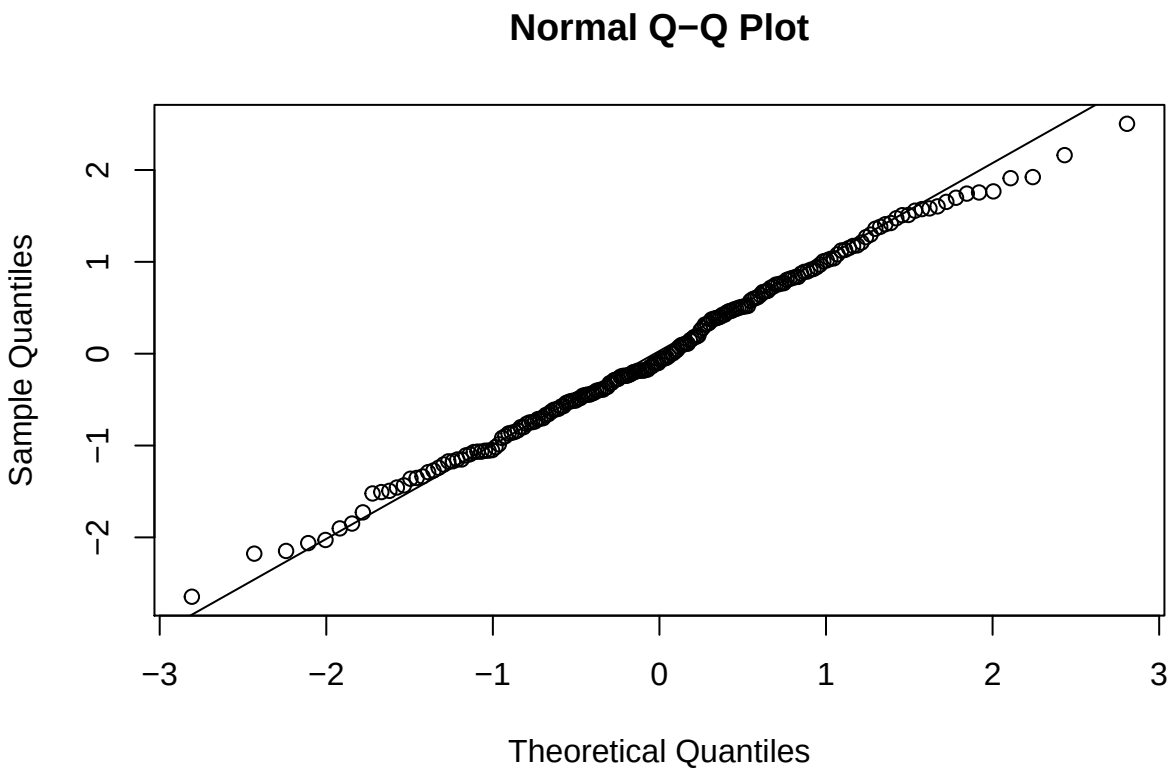


(Q) When you look the above scatter plot, do you believe the linear relationship between the responses variable and predictor variables ?

Ans: I think there is linear relationship between weedsmokingstar and Lazy. But, it's hard to identify linear relationship between weedsmokingstar and Sport. Actually,without any knowledge about the true model, I will initially conclude that there is no linear relationship between weedsmokingstar and Sport.

(Problem 3)

```
qqnorm(fit1$residuals)
qqline(fit1$residuals)
```



(Q) Based on the above Normal Q-Q plot, do you think the assumption of normal error term is appropriate ?

Ans: Yes. if data comes from normal distrbution, they will lie on the straight line.

(Problem 4)

```
## comparison of sse under the true model and overfitting model
fit2<- lm(weedsmokingstar ~ sport + lazy+lazy2+lazy3+lazy4)##fit data by predictor
##variable sport, lazy, lazy2, lazy3, lazy4
SSE1 <- sum(fit1$residuals^2)
SSE2 <- sum(fit2$residuals^2)
list(model = SSE1 ,mode2 = SSE2)
```

```
## $model
## [1] 186.1301
##
## $mode2
## [1] 183.6958
```

```
# summary(fit1)
# summary(fit2)
```

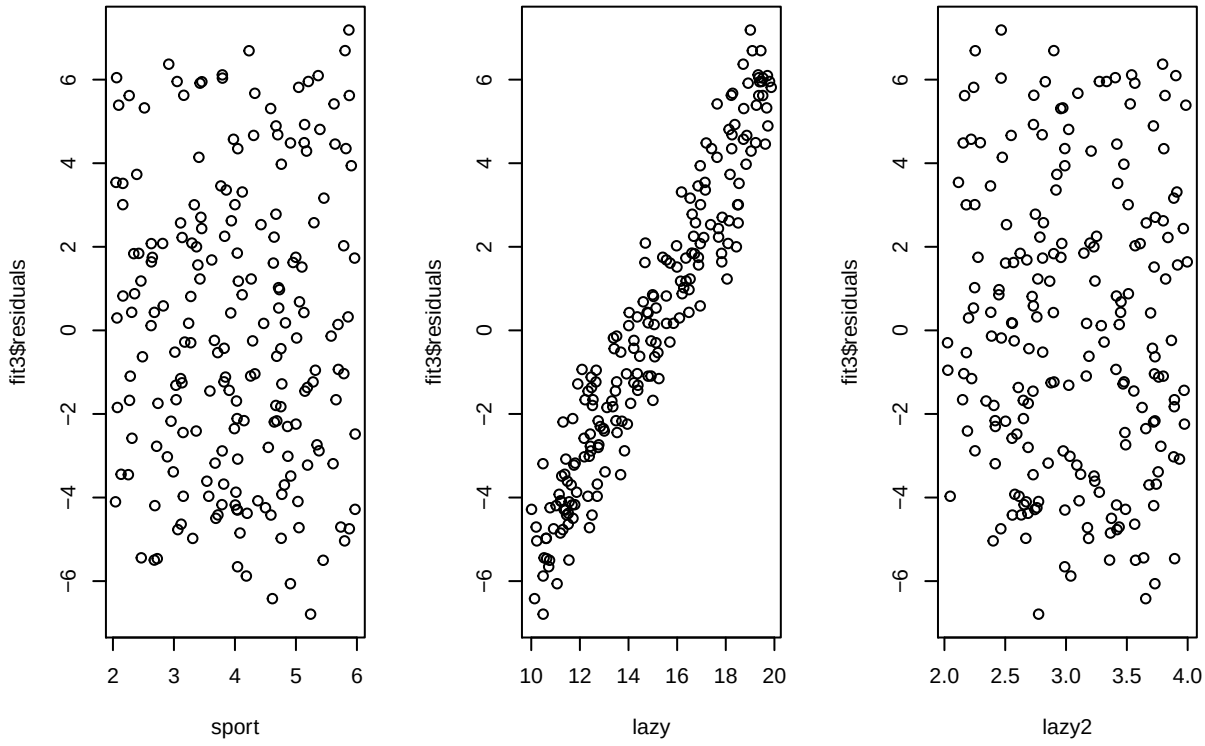
(Q) Is SSE1 greater than or equal to SSE2 ? How to explain it by the theoretical properties ?

Ans: SSE1 greater than SSE2.

(Problem 5)

```
## comparison of residual under the true model and underfitting model
fit3<- lm(weedsmokingstar ~ sport )###model underfitting
```

```
layout(matrix(1:3, ncol = 3))###present multiple plot together
plot(sport,fit3$residuals)###scatter plot residuals v.s sport (predictor variable)
plot(lazy,fit3$residuals)###scatter plot residuals v.s lazy (predictor variable)
plot(lazy2,fit3$residuals)###scatter plot residuals v.s lazy (predictor variable)
```

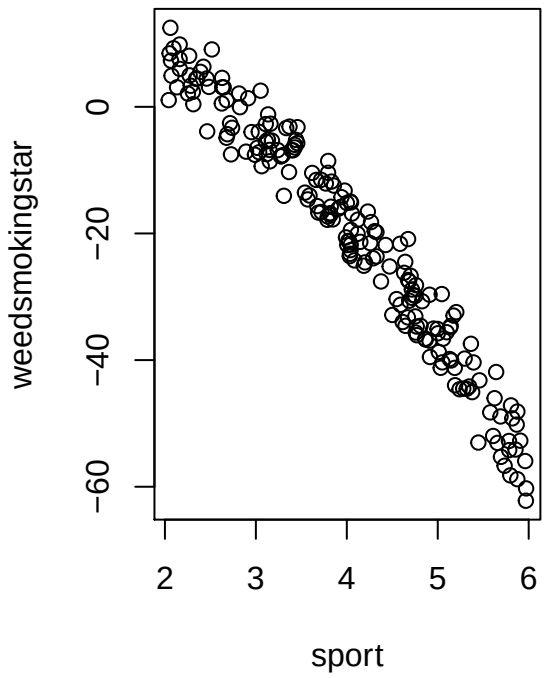


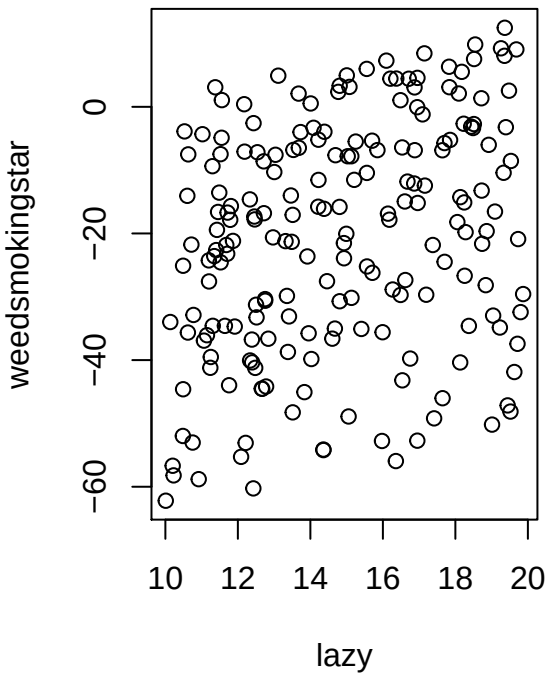
(Q) How do you explain the above scatter plot (residual v.s. predictor variable)

Ans: For Sport, there is no special pattern in the residual plot. It tell us the linear relationship between weedsmokingstar and Sport is suitable for the current model. For Lazy, there is linear relationship in the residual plot. We may consider the variable, Lazy, in our model. For Lazy2, there is no special pattern in the residual plot. We could conclude Lazy2 is not important variable for our model.

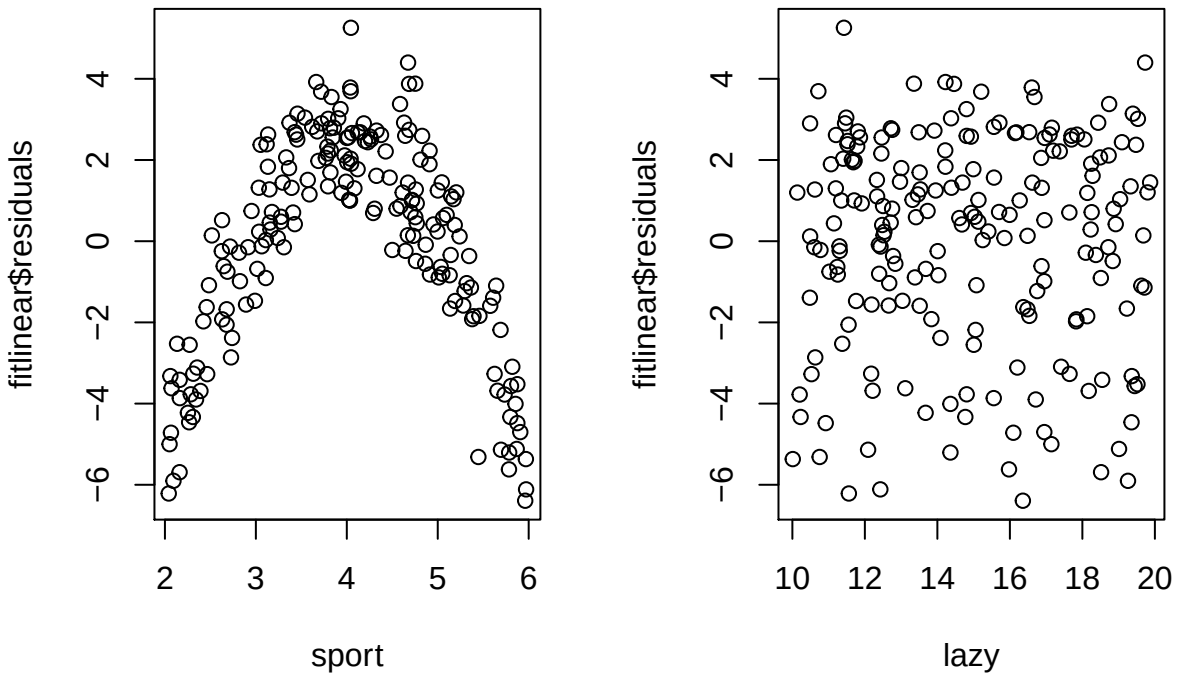
(Problem 6)

```
#(a) true model (nonlinear model):
error <- rnorm(n) ##random error
weedsmokingstar <- -3.5 - 2* sport^2 + 1.2* lazy + error ## true model is linear model

##scatter plot "X" v.s "weedsmokingstar "Y"
layout(matrix(1:2, ncol = 2))
plot(sport, weedsmokingstar)
plot(lazy, weedsmokingstar)
```






```
### Fit data by linear model
fitlinear<- lm(weedsmokingstar ~ sport + lazy)##linear model fitting
plot(sport,fitlinear$residuals)
plot(lazy,fitlinear$residuals)
```

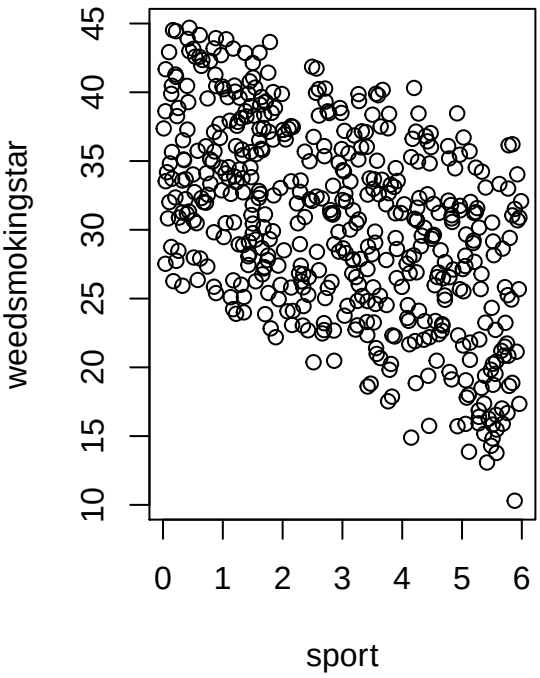


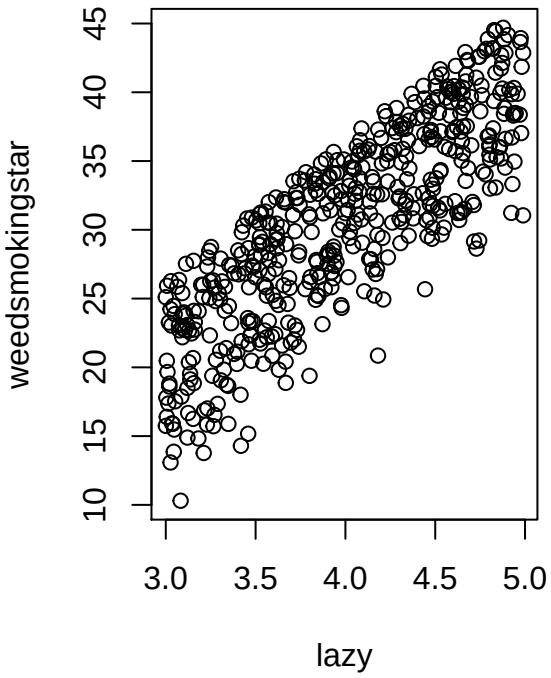
(Q) How to explain the above scatter (response variable v.s. predictor variable and residual v.s. predictor variable ?

Ans: It's a little hard to determine the relationship between Sport and weedsmokingstar. It seems to be linear but also familiar to quadratic.

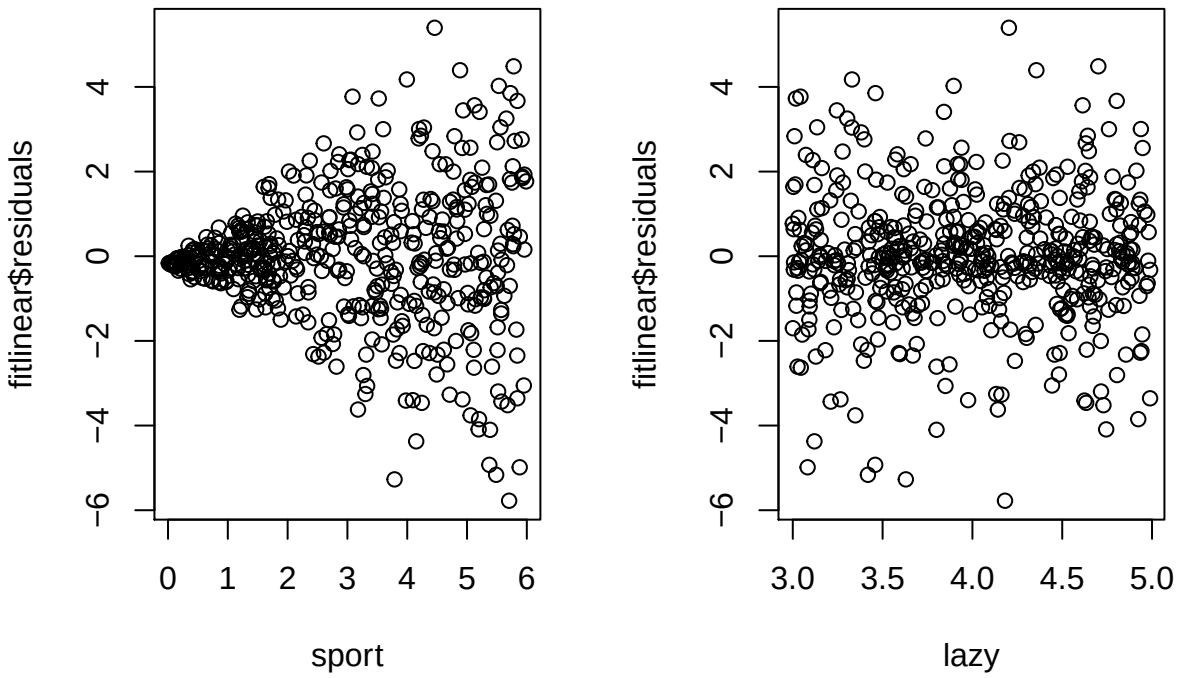
(Problem 7)

```
n <- 500## The size of observed data
sport <- runif(n, 0, 6)
lazy <- runif(n, 3, 5)
for (i in 1:n){
error[i] <- rnorm(1, mean = 0, sd=(0.5*sport[i])) ##random error
}
weedsmokingstar <- -3.5 - 2 * sport+ 10 * lazy + error ## true model is linear model

##scatter plot "X" v.s "weedsmokingstar "Y"
layout(matrix(1:2, ncol = 2))
plot(sport, weedsmokingstar)
plot(lazy, weedsmokingstar)
```






```
### Fit data by linear model
fitlinear<- lm(weedsmokingstar ~ sport + lazy)##linear model fitting
plot(sport,fitlinear$residuals)
plot(lazy,fitlinear$residuals)
```



(Q) How to explain the above scatter (response variable v.s. predictor variable and residual v.s. predictor variable ?

Ans: We find that the data points sread out with increasing of Sport value. It seems that the variance of error is not constant.