1.a
x is a data set which contains 100 random numbers with mean =10 and sd = 1.
y is also a data set which contains 100 random numbers with mean =10 and sd = 1.
So they are similar 100 random numbers data set create by same condition (mean=10,sd=1).

b.

```
fit <- lm(y ~ x, data = df)
summary(fit)
```

The output is appropriate.(The intercept is 9.83, slope is 0.02)
With p-value very close to 1, we can conclude that there are no relationship exists between the response y and the predictor x. Whereas, the p-value corresponding to the F-statistic is very high, providing more evidence of no relationship between the predictor and response. Also, R-squared is quite low shows that even if there is a relationship, the relationship is not strong.

c.
The p-value change from 0.875 to 4.83e-09, F-statistic change from 0.02 to 41.17, Intercept change from 9.83 to 4.18 and slope of x change from 0.02 to 0.58.
With p-value very close to 0, we can conclude that a relationship exists between the response y and the predictor x. Whereas, the p-value corresponding to the F-statistic is very low, providing more evidence of no relationship between the predictor and response. And R-squared is around 30% shows that the relationship is not very strong.
The output changes because at first, the data are all close around (10,10) and all the data will form a shape like a circle. So there are no relationship between them. Then, we add a point (0,0) which makes the whole data set linearly. But from the output we can see that the model is not good enough and the relationship is not very strong. If we want to make the model more linearly and relationship more strong, we can add:

```
x<-rnorm(100,mean=0,sd=1)
y<-rnorm(100,mean=0,sd=1)
dff<-data.frame(x=x,y=y)
df<-rbind(df,dff)

x<-rnorm(100,mean=1,sd=1)
y<-rnorm(100,mean=1,sd=1)
dff<-data.frame(x=x,y=y)
df<-rbind(df,dff)

…

x<-rnorm(100,mean=9,sd=1)
y<-rnorm(100,mean=9,sd=1)
dff<-data.frame(x=x,y=y)
df<-rbind(df,dff)
```

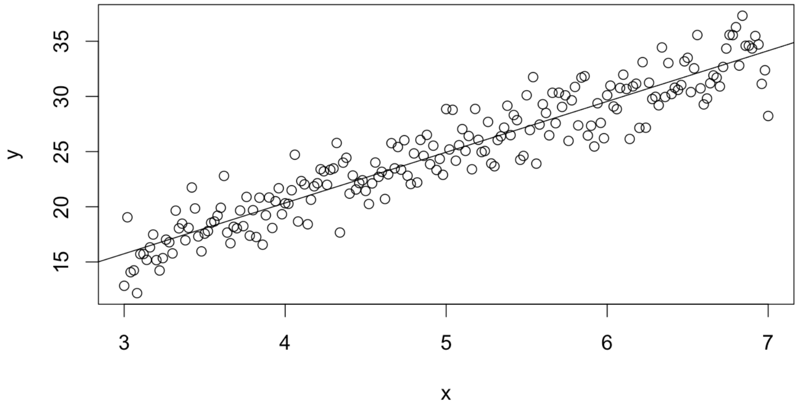
Then, the model will be strong linearly.

2.a
This code create a data set x from 3 to 7 with gap 0.02.
Then it create a data set r with 201 random numbers with mean = 0, sd =2.
Then it gives a function making data set y = 2+4.6*x +r.
Then it uses lm function to make a linearly regression to output called out.
Finally it summary the output result.

The intercept is 91.94, slope is 4.6. With p-value very close to 0, we can conclude that a relationship exists between the response y and the predictor x. Whereas, the p-value corresponding to the F-statistic is very low, providing more evidence of the relationship between the predictor and response. Also, R-squared is almost 90% shows that the relationship is strong.

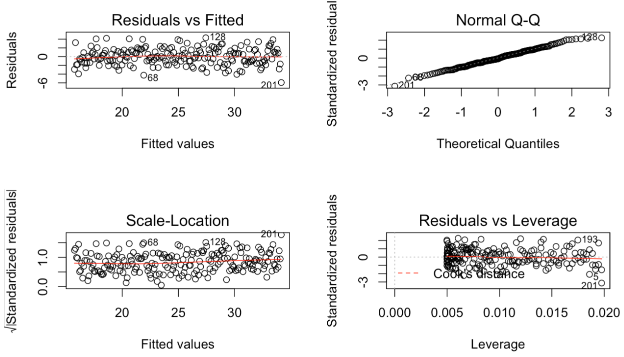
The true relationship between x and y is y = 4.6x + (r+2)

b.
plot(x,y)
abline(out)
This is the plot of data and the fitted line.



ß0 means intercept and ß1 means slope of x.
Population parameters, ß0 is (r+2)(around 2) and ß1 is 4.6.(r<-rnorm(201,mean=0,sd=2))
Their estimates in the model, ß0 is 1.94 and ß1 is 4.6.
We can see that the estimated coefficients are very close to the true coefficient value.

c.



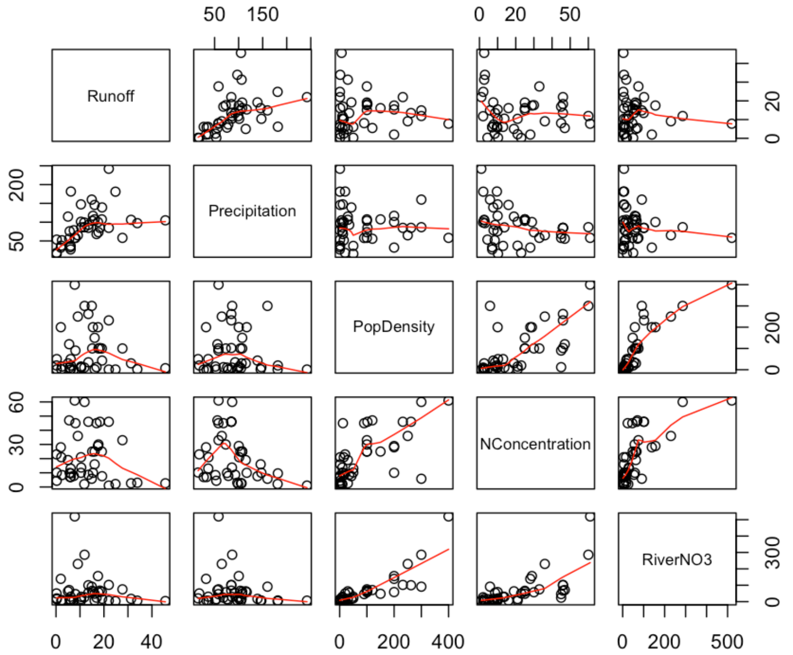
From the diagnostic plots we can see that:
The shape of line is quite flat in residuals vs fitted shows that our model is good.
Normal-q-q is largely shows that the residuals are normally distributed.
In the variance of residuals, the red line is relatively flat, showing homoscedasticity (i.e. constant variance along x).
Point 5, 193, 201 are high-leverage points.

d.

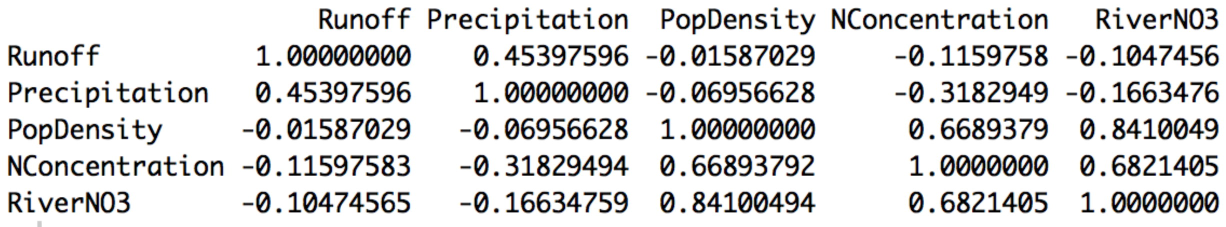
```
intercept_set<-rep(NA,1000)
slope_set<-rep(NA,1000)
for(i in 1:1000){
  set.seed(12)
  x<-seq(3,7,.02)
  r<-rnorm(201,mean=0,sd=2)
  y<-2+4.6*x+r
  out<-lm(y~x)
  intercept_set[i]<-coef(out)["(Intercept)"]
  slope_set[i]<-coef(out)["x"]
}
df<-data.frame(inter=intercept_set,slope=slope_set)
```

e.
If we use a seed(12) here, the results will stay the same as follows:
95% confidence interval of intercept is 1.94152~1.94152
95% confidence interval of slope is 4.601479~4.601479
But in the real test, we can get our 95% confidence interval as follow:
SEx = sqrt(((∑i)(xi - mean(x))^2/n-1)
And the 95% confidence interval will be:
(mean(x) - 1.96*SEx, mean(x) + 1.96*SEx)

3.a
Here is the scatterplot for each pair of numerical variables.



Here is the corresponding correlation matrix.



```
               Runoff Precipitation  PopDensity NConcentration   RiverNO3
Runoff          1.00000000    0.45397596 -0.01587029     -0.1159758 -0.1047456
Precipitation   0.45397596    1.00000000 -0.06956628     -0.3182949 -0.1663476
PopDensity     -0.01587029   -0.06956628  1.00000000      0.6689379  0.8410049
NConcentration -0.11597583   -0.31829494  0.66893792      1.0000000  0.6821405
RiverNO3       -0.10474565   -0.16634759  0.84100494      0.6821405  1.0000000
```

Comment:
With Precipitation increase, Runoff will also increase.
With PopDensity, NConcentration, RiverNO3 increase, Precipitation won’t change too much.
With RiverNO3 increase, PopDensity and NConcentration will increase and Runoff will decrease a little bit.
NConcentration increase with PopDensity increase.
With Runoff, Precipitation increase, NConcentration will first increase and then decrease.

b.

The best model I found to predict RiverNO3 is:

```
fit <- lm(RiverNO3 ~ I(PopDensity^2)*NConcentration, data = river)
```

The results shows that R-squared is 0.91 which is very high and shows the relationship is very strong. The p-value is quite close to 0, F-statistic is larger than other models and RSE is smaller than other models.

Assumptions:
PopDensity has strong relationship with RiverNO3.
NConcentration has relationship with RiverNO3.
PopDensity * NConcentration has strong relationship with RiverNO3.
PopDensity^2 has strong relation ship with RiverNO3.
Runoff and Precipitation have no relationship with RiverNO3.

The Null hypotheses is (PopDensity^2)*NConcentration have strong relationship with RiverNO3.

I have done tests as below:

```
fit <- lm(RiverNO3 ~. - RiverName - Country, data = river)
fit <- lm(RiverNO3 ~ Runoff + Precipitation, data = river)
fit <- lm(RiverNO3 ~ PopDensity, data = river)
fit <- lm(RiverNO3 ~ PopDensity + NConcentration, data = river)
fit <- lm(RiverNO3 ~ PopDensity * NConcentration, data = river)
fit <- lm(RiverNO3 ~ PopDensity * Precipitation, data = river)
fit <- lm(RiverNO3 ~ Precipitation * NConcentration, data = river)
fit <- lm(RiverNO3 ~ log(PopDensity), data = river)
fit <- lm(RiverNO3 ~ log(NConcentration), data = river)
fit <- lm(RiverNO3 ~ I(PopDensity^2), data = river)
fit <- lm(RiverNO3 ~ I(PopDensity^2)*NConcentration, data = river)
fit <- lm(RiverNO3 ~ I(PopDensity^2)*I(NConcentration^2), data = river)
fit <- lm(RiverNO3 ~ PopDensity*I(NConcentration^2), data = river)
fit <- lm(RiverNO3 ~ sqrt(PopDensity), data = river)
```

And finally I find out that

```
 fit <- lm(RiverNO3 ~ I(PopDensity^2)*NConcentration, data = river)
```

is the best model in these models.
The goodness of this model is R-squared is 0.91 which is very high and shows the relationship is very strong. The p-value is quite close to 0, F-statistic is larger than other models and RSE is smaller than other models.
And the potential problem I think is I am not sure that PopDensity^2 meaningful.

c.

```
fit <- lm(RiverNO3 ~ I(PopDensity^2)*NConcentration, data = river)
summary(fit)
```

With p-values close to 1, we can conclude that there is a relationship exists between the response RiverNO3 and the predictor PopDensity + NConcentration. Whereas, the p-value corresponding to the F-statistic is very low, providing more evidence of no relationship between the predictor and response. Also, R-squared is about 91% shows that the relationship is strong.

```
fit <- lm(RiverNO3 ~ PopDensity * NConcentration, data = river)
summary(fit)
```

With p-values close to 1, we can conclude that there is a relationship exists between the response RiverNO3 and the predictor PopDensity + NConcentration. Whereas, the p-value corresponding to the F-statistic is very low, providing more evidence of no relationship between the predictor and response. Also, R-squared is about 87% shows that the relationship is strong.

```
fit <- lm(RiverNO3 ~ PopDensity + NConcentration, data = river)
summary(fit)
```

With p-values close to 1, we can conclude that there is a relationship exists between the response RiverNO3 and the predictor PopDensity + NConcentration. Whereas, the p-value corresponding to the F-statistic is very low, providing more evidence of no relationship between the predictor and response. Also, R-squared is about 73% shows that the relationship is strong.

```
fit <- lm(RiverNO3 ~ Runoff + Precipitation, data = river)
summary(fit)
```

With p-values close to 0, we can conclude that there is no relationship exists between the response RiverNO3 and the predictor Runoff + Precipitation. Whereas, the p-value corresponding to the F-statistic is high, providing more evidence of no relationship between the predictor and response. Also, R-squared is low shows that even if there is a relationship, the relationship is not strong.

From the result we can see that the group’s claim is right. Higher nitrate concentrations in rivers are likely to be associated with natural phenomena and not due to human activity.