hw5_Sayan_Biswas

Sayan Biswas

21 March 2019

Part A

Problem 1

I chose the miniposter created by Christopher Tso for my Part A. The original dataset can be found here : https://data.cdc.gov/NCHS/NCHS-Leading-Causes-of-Death-United-States/bi63-dtpu

```
nchs <- read_csv("NCHS_-_Leading_Causes_of_Death__United_States.csv")
head(nchs,n=10)
```

```
## # A tibble: 10 x 6
##     Year `113 Cause Name`     `Cause Name`  State   Deaths `Age-adjusted D~
##    <dbl> <chr>                <chr>         <chr>    <dbl>            <dbl>
##  1  2016 Accidents (unintent~ Unintentiona~ Alabama   2755             55.5
##  2  2016 Accidents (unintent~ Unintentiona~ Alaska     439             63.1
##  3  2016 Accidents (unintent~ Unintentiona~ Arizona   4010             54.2
##  4  2016 Accidents (unintent~ Unintentiona~ Arkans~   1604             51.8
##  5  2016 Accidents (unintent~ Unintentiona~ Califo~  13213             32  
##  6  2016 Accidents (unintent~ Unintentiona~ Colora~   2880             51.2
##  7  2016 Accidents (unintent~ Unintentiona~ Connec~   1978             50.3
##  8  2016 Accidents (unintent~ Unintentiona~ Delawa~    516             52.4
##  9  2016 Accidents (unintent~ Unintentiona~ Distri~    401             58.3
## 10  2016 Accidents (unintent~ Unintentiona~ Florida  12561             54.9
```

Problem 2

Figure 1: Plot for Deaths from Heart Disease over time:

```r
nchs %>%
  filter(`Cause Name`%in% c("Heart disease","Stroke"))%>%
  group_by(Year) %>%
  summarise(t_d= sum(Deaths,na.rm = T))%>%
  ggplot(aes(x=Year,y=t_d))+
  geom_col()+
  labs(title = "Deaths from Heart Disease over time",
       x = "Year ",
       y = "Total Deaths from Heart Disease")
```

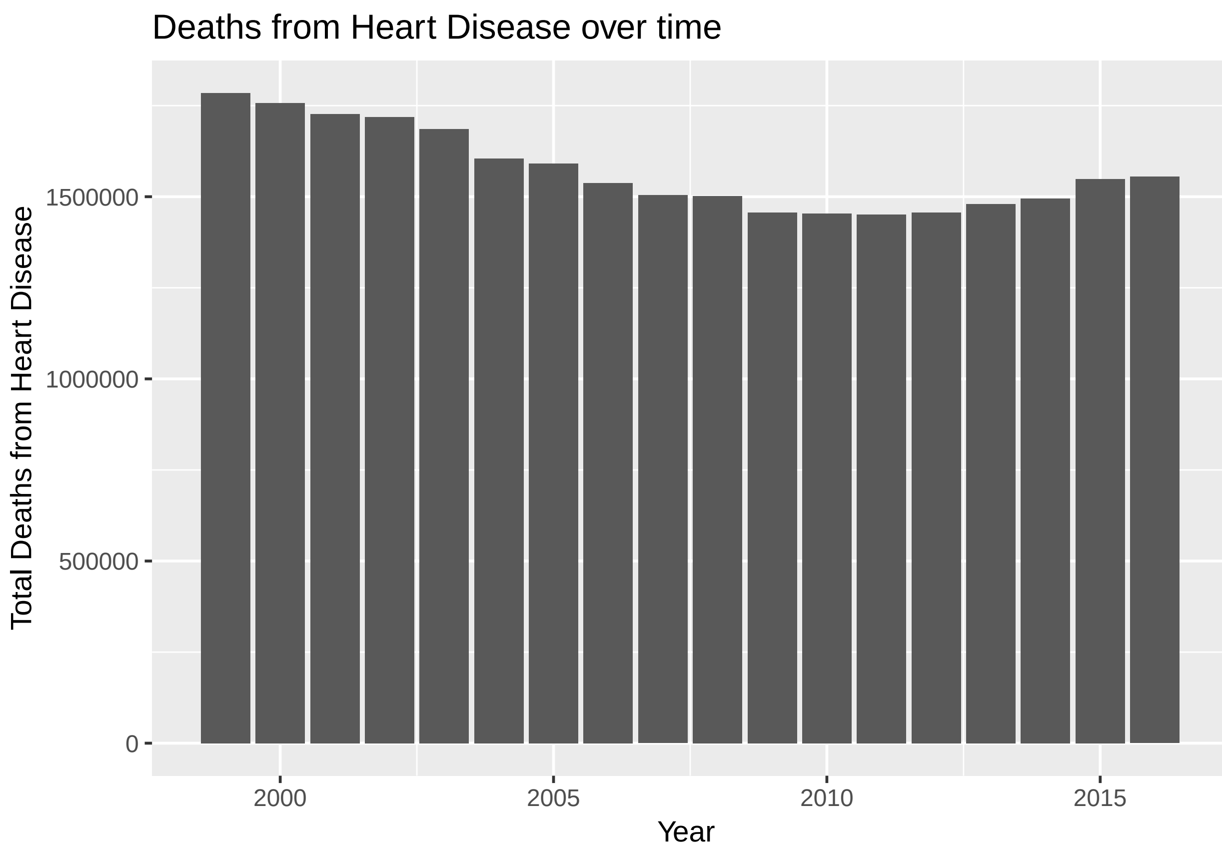


Figure 2: Plot for Deaths from Mental Illness over time:

```
nchs %>%
  filter(`Cause Name`%in% c("Alzheimer's disease","Suicide"))%>%
  group_by(Year) %>%
  summarise(t_d= sum(Deaths,na.rm = T))%>%
  ggplot(aes(x=Year,y=t_d))+
  geom_col()+
  labs(title = "Deaths from Mental Illness over time",
       x = "Year ",
       y = "Total Deaths from Mental Illness")
```

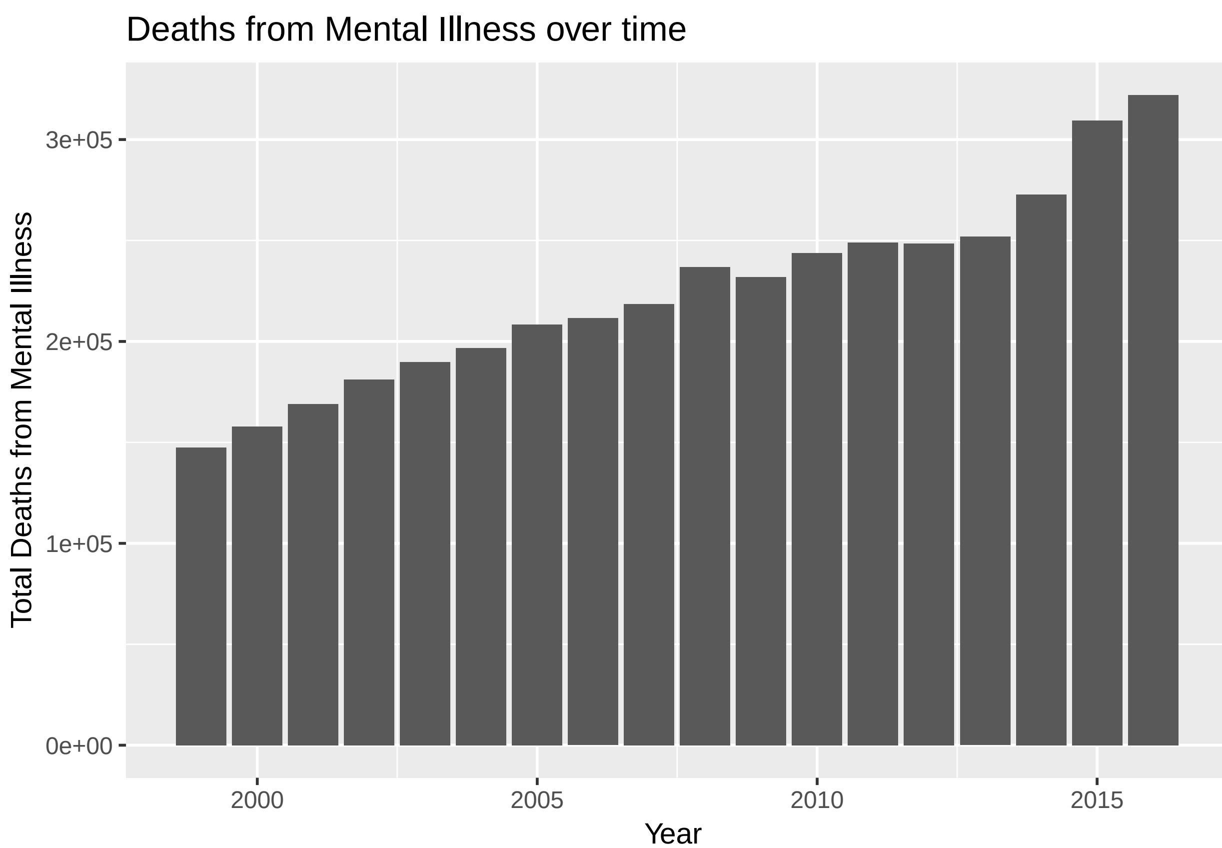


Figure 3: Plot for Deaths from Cancer over time:

```
nchs %>%
  filter(`Cause Name`=="Cancer")%>%
  group_by(Year) %>%
  summarise(t_d= sum(Deaths,na.rm = T))%>%
  ggplot(aes(x=Year,y=t_d))+
  geom_col()+
  labs(title = "Deaths from Cancer over time",
       x = "Year ",
       y = "Total Deaths from Cancer")
```

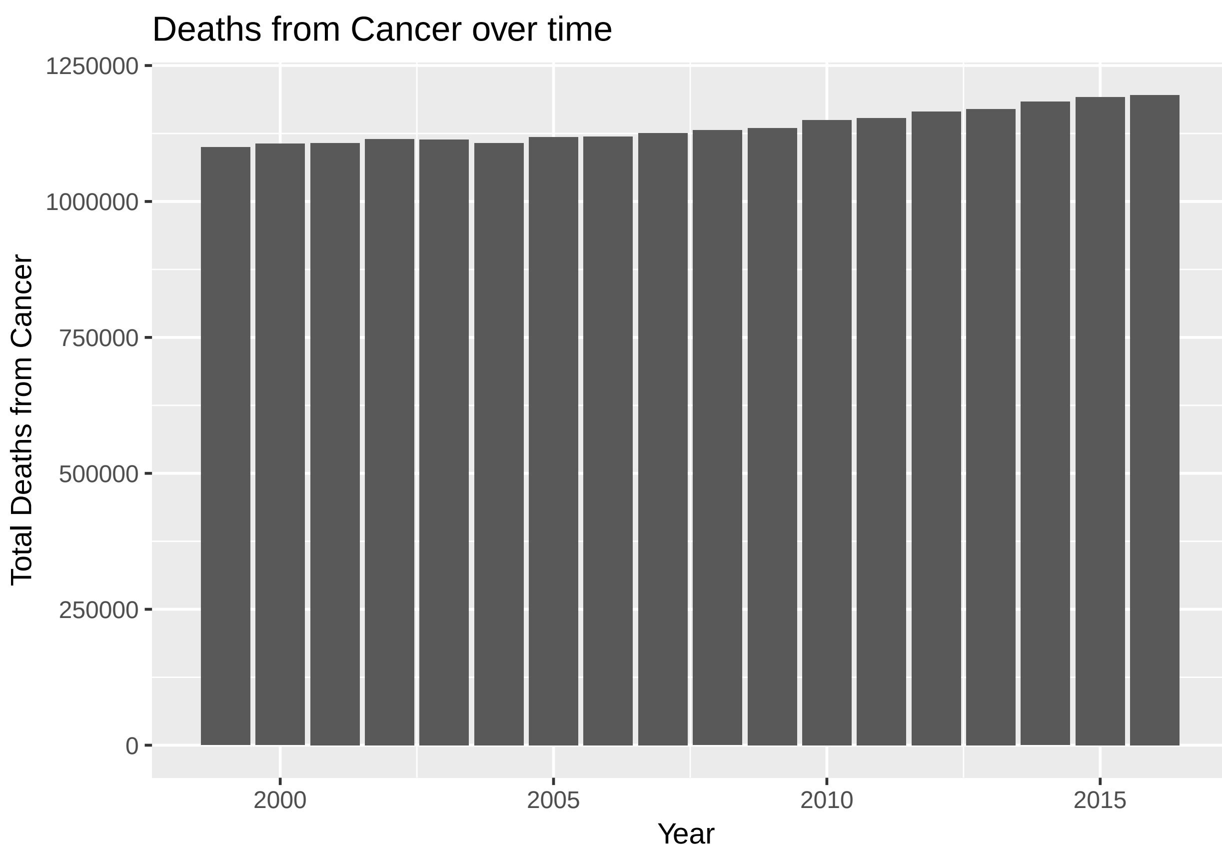
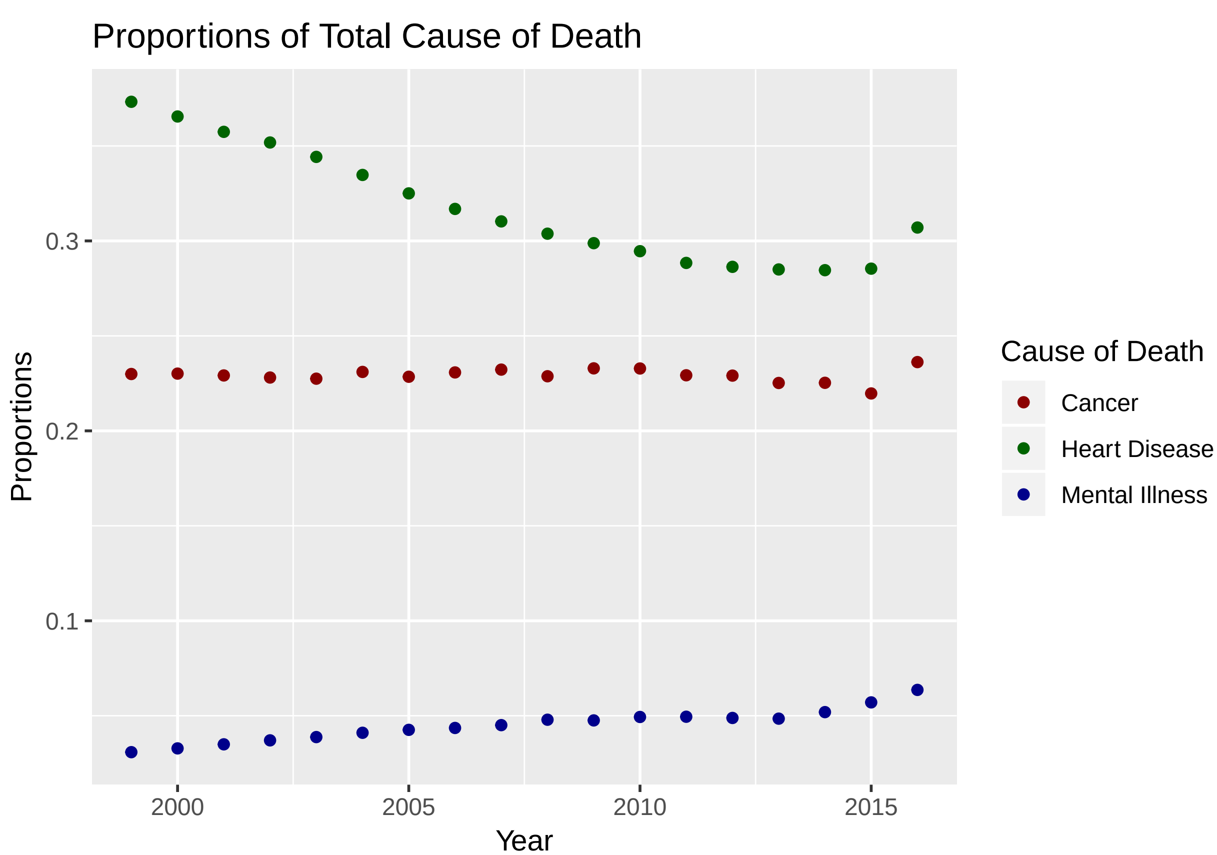


Figure 4: Plot for Proportions of Total Cause of Death

```r
nchs %>%
    group_by(Year,`Cause Name`)%>%
    summarise(sum=sum(Deaths,na.rm = T))%>%
    spread(key = `Cause Name`,value = sum)%>%
    ungroup()%>%
    transmute(
      Year= Year,
      `All causes`= `All causes`,
      d_can=Cancer,
      d_hd=`Heart disease`+ Stroke,
      d_mi=`Alzheimer's disease`+Suicide,
      Cancer = d_can/`All causes`,
      `Heart Disease`= d_hd/`All causes`,
      `Mental Illness`=d_mi/`All causes`)%>%
    gather(Cancer,`Heart Disease`,`Mental Illness`,
           key="Cause of Death",value=prop_v)%>%
    ggplot()+
    geom_point(aes(x=Year,y=prop_v,color=`Cause of Death`))+
    scale_color_manual(name="Cause of Death",
                       values=c("darkred","darkgreen", "darkblue"))+
    labs(title = "Proportions of Total Cause of Death",
      y="Proportions")
```



Part B

Problem 3

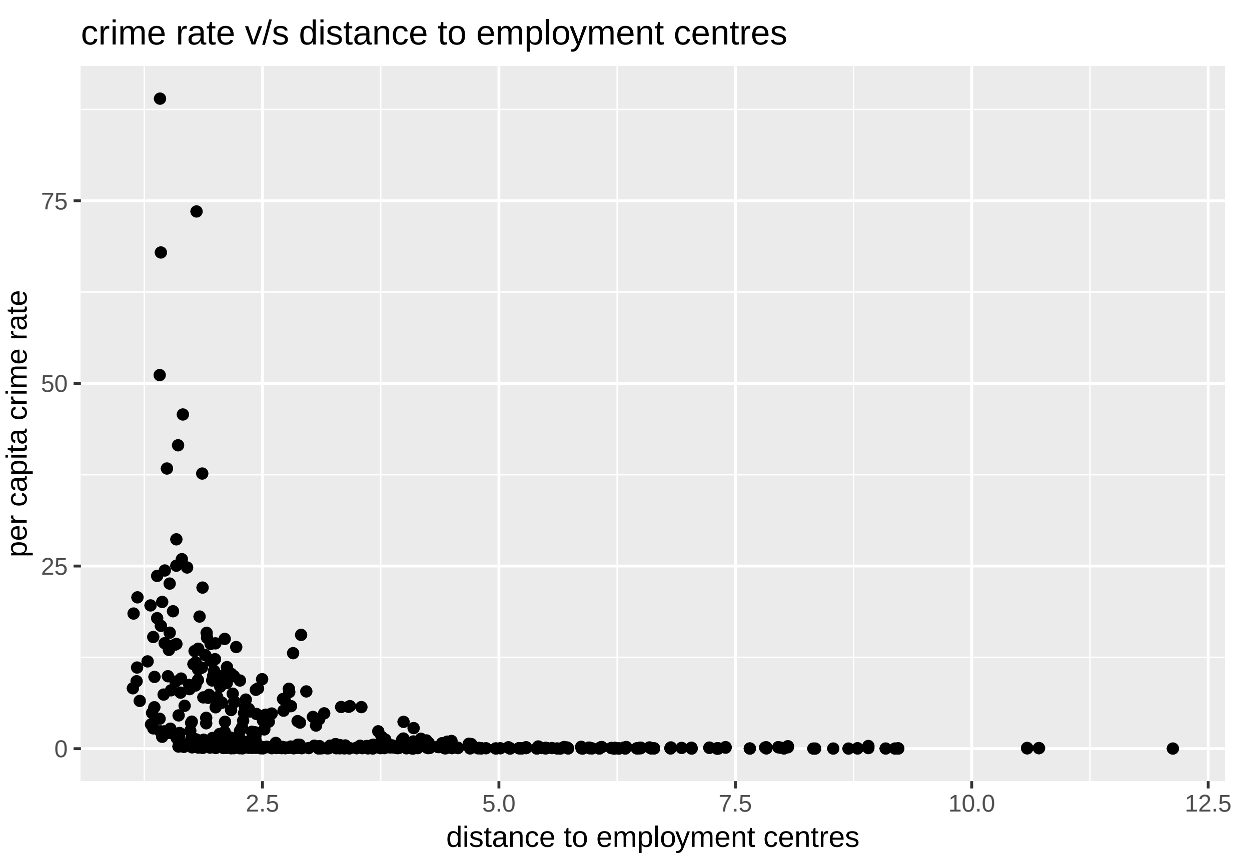
The variables dis,indus,nox,age,rad,tax,lstat,medv when plotted against crim does show some kind of linear relationship when the variables are log transformed.

But among all the variables in the dataset, the variable dis(weighted distance to five Boston employment centres) is the one which shows a strong linear relationship with the variable crim(per capita crime rate) when both the variables are log-transformed.

Although the plot of crim(per capita crime rate) versus dis(weighted distance to five Boston employment centres) does not show much of a linear relationship between them as shown below but it shows a trend that the crime rate is decreasing as the distance increasing.

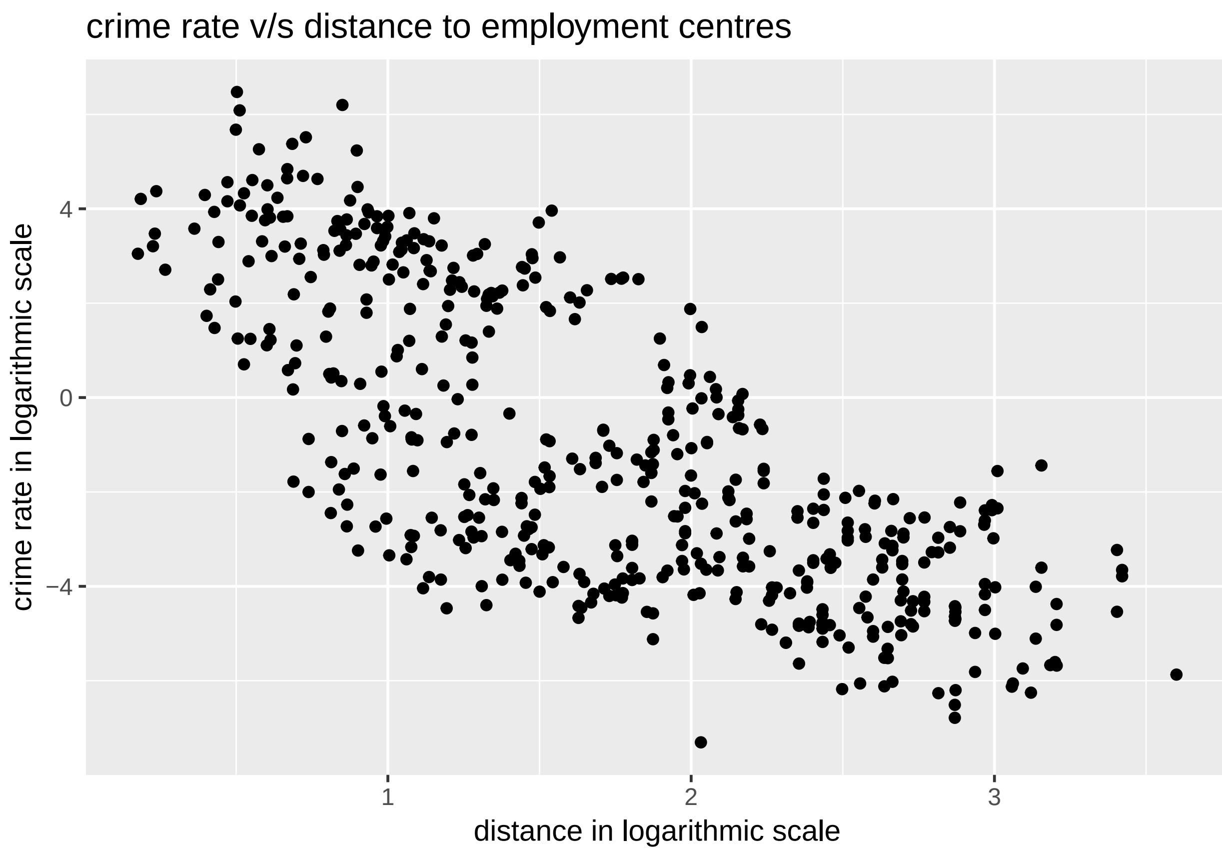
```r
data("BostonHousing")

BostonHousing %>%
  ggplot(aes(x=dis,y=crim))+
  geom_point()+
  labs(title = "crime rate v/s distance to employment centres",
       x = "distance to employment centres",
       y = "per capita crime rate")
```



Applying log tranformations to crim(crime rate) and dis(weighted distance to employment centres) variables makes the pattern linear and it establishes a strong relationship between the transformed variables as shown below:

```
BostonHousing %>%
  ggplot(aes(x=log2(dis),y=log2(crim)))+
  geom_point()+
  labs(title = "crime rate v/s distance to employment centres",
       x = "distance in logarithmic scale",
       y = "crime rate in logarithmic scale")
```



The pattern shows a strong linear relationship between the plotted variables and it shows that the crime rate decreases with the increase of the distance.

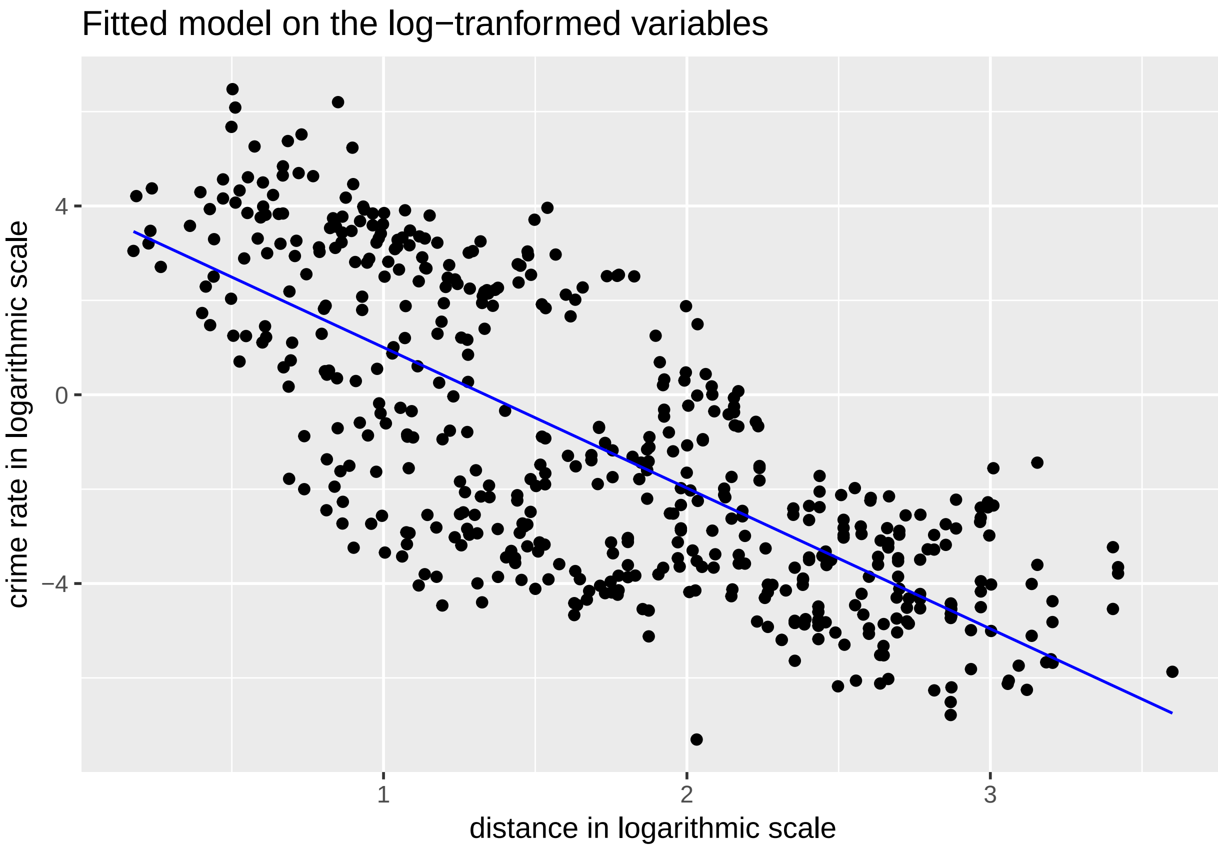
To make the pattern explicit, we fit a model.

```
fit_dis <- lm(log2(crim) ~ log2(dis), data = BostonHousing)
summary(fit_dis)
```

```
##
## Call:
## lm(formula = log2(crim) ~ log2(dis), data = BostonHousing)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -5.232 -1.608 -0.027  1.800  4.751
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.9835     0.2245   17.74   <2e-16 ***
## log2(dis)    -2.9810     0.1193  -24.99   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.086 on 504 degrees of freedom
## Multiple R-squared:  0.5534, Adjusted R-squared:  0.5525
## F-statistic: 624.6 on 1 and 504 DF,  p-value: < 2.2e-16
```

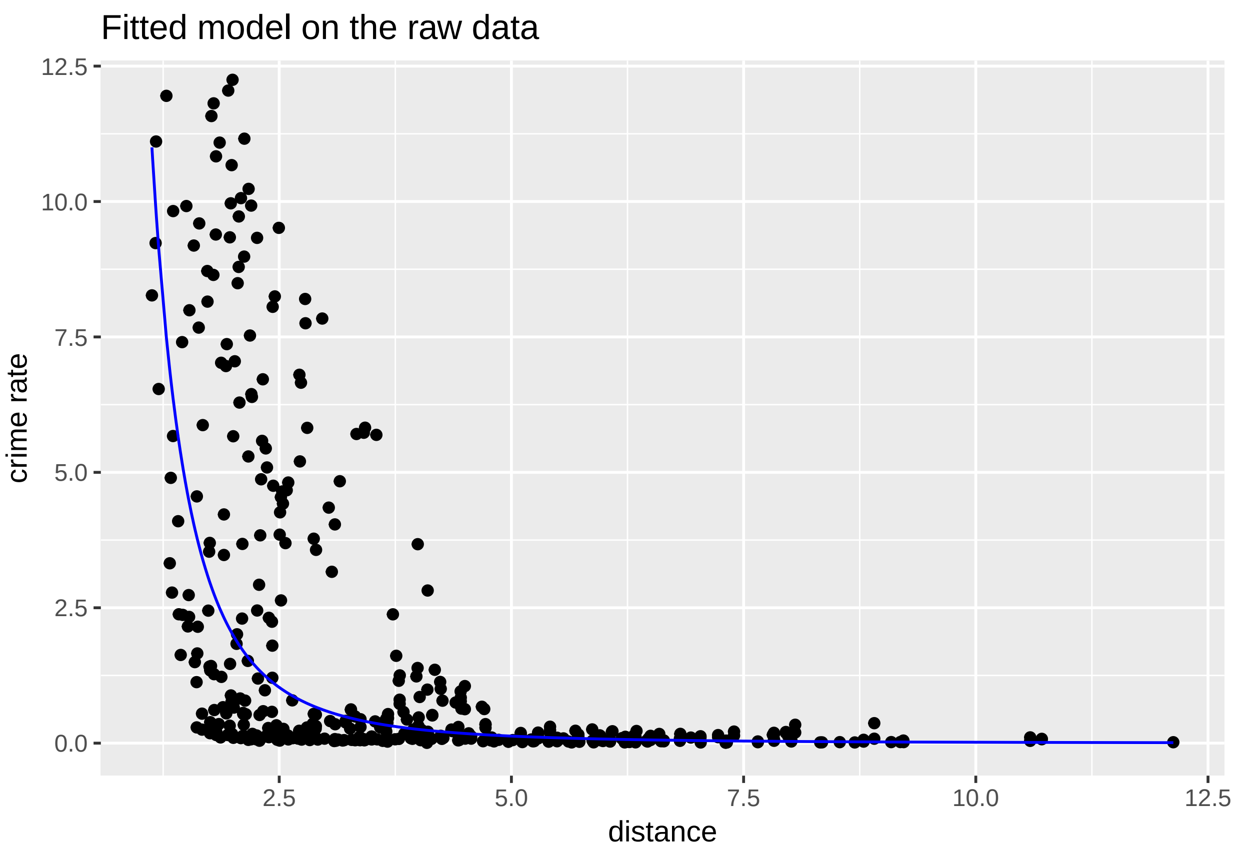
The plot below shows the fitted model on the log-transformed variables.

```
BostonHousing %>%
  add_predictions(fit_dis,"lpred")%>%
  ggplot(aes(x=log2(dis)))+
  geom_point(aes(y=log2(crim)))+
  geom_line(aes(y=lpred),color="blue")+
  labs(title = "Fitted model on the log-tranformed variables",
       x = "distance in logarithmic scale",
       y = "crime rate in logarithmic scale")
```


Fitted model on the log−tranformed variables
crime rate in logarithmic scale
4
0
−4
1
2
3
distance in logarithmic scale

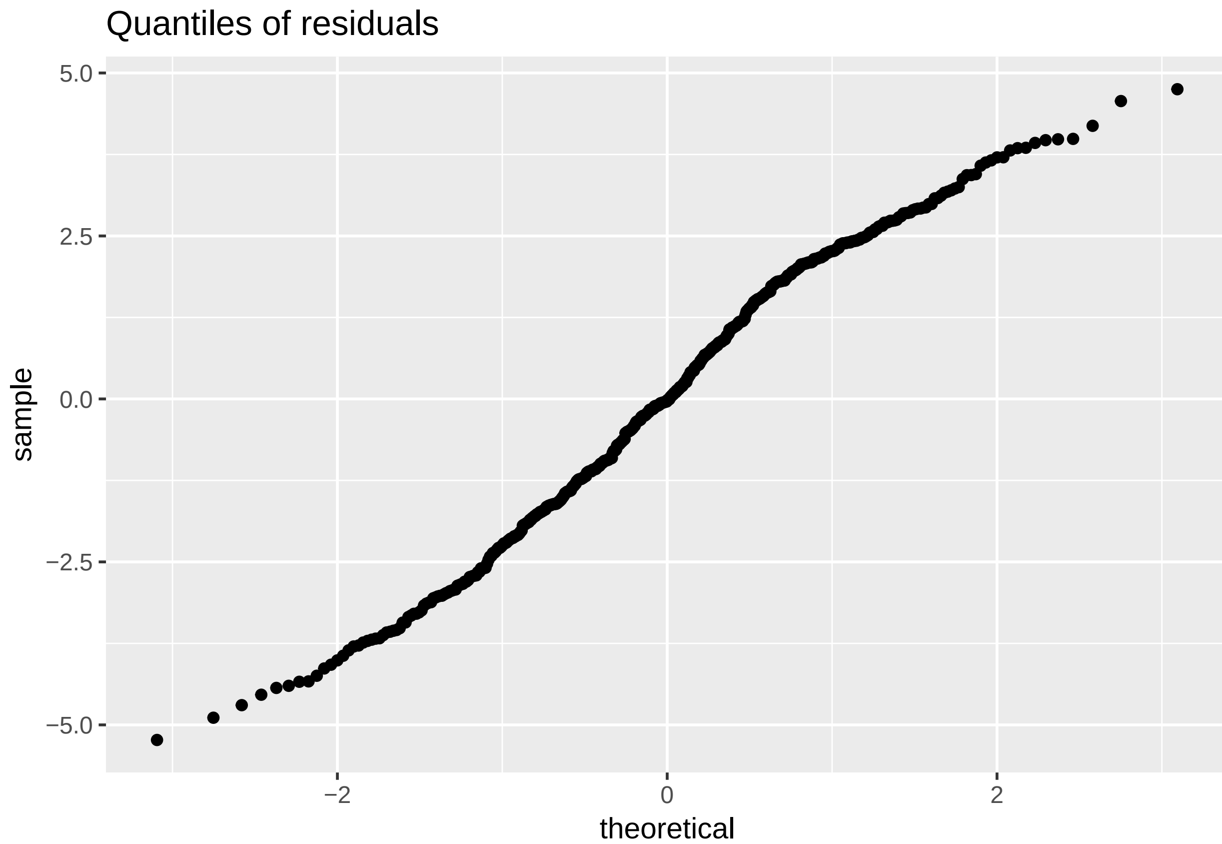
Fitting the model undoing the log transformation, so that predictions can be overlayed on the raw data.

```
BostonHousing %>%
  add_predictions(fit_dis,"lpred") %>%
  mutate(pred=2^lpred) %>%
  ggplot(aes(x=dis))+
  geom_point(aes(y=crim))+
  geom_line(aes(y=pred),color="blue")+
  labs(title = "Fitted model on the raw data",
       x = "distance",
       y = "crime rate")+
  coord_cartesian(ylim=c(0,12))
```



We can also plot the quantiles of residuals to verify that the distibution of residual errors is normal and verify the assumptions of linear modelling.

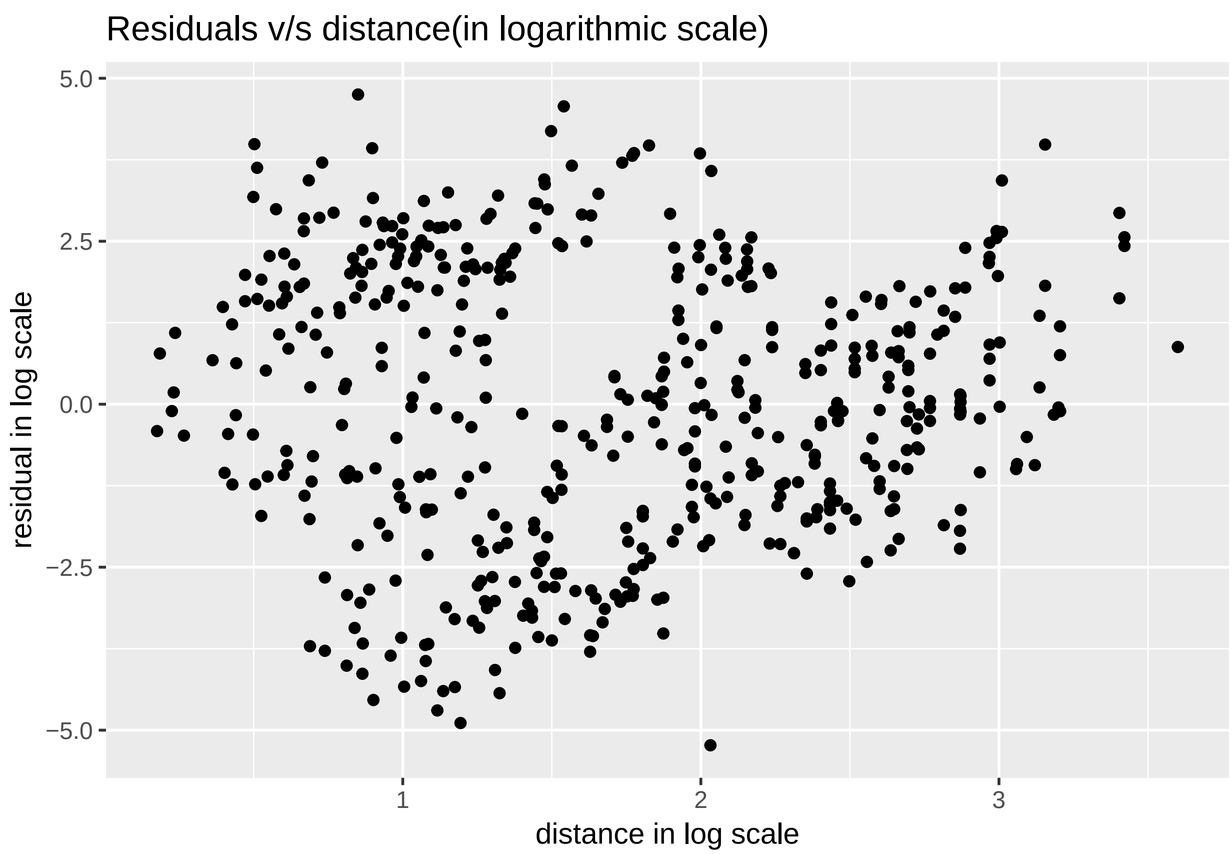
```
BostonHousing %>%
add_residuals(fit_dis) %>%
ggplot(aes(sample=resid)) +
geom_qq()+
  labs(title="Quantiles of residuals")
```


Quantiles of residuals
5.0
2.5
0.0
−2.5
−5.0
sample
−2
0
2
theoretical

Problem 4

In order to verify that the true error is randomly(normally in linear model) distributed, plotting the residuals against the predictor variable used for the model.

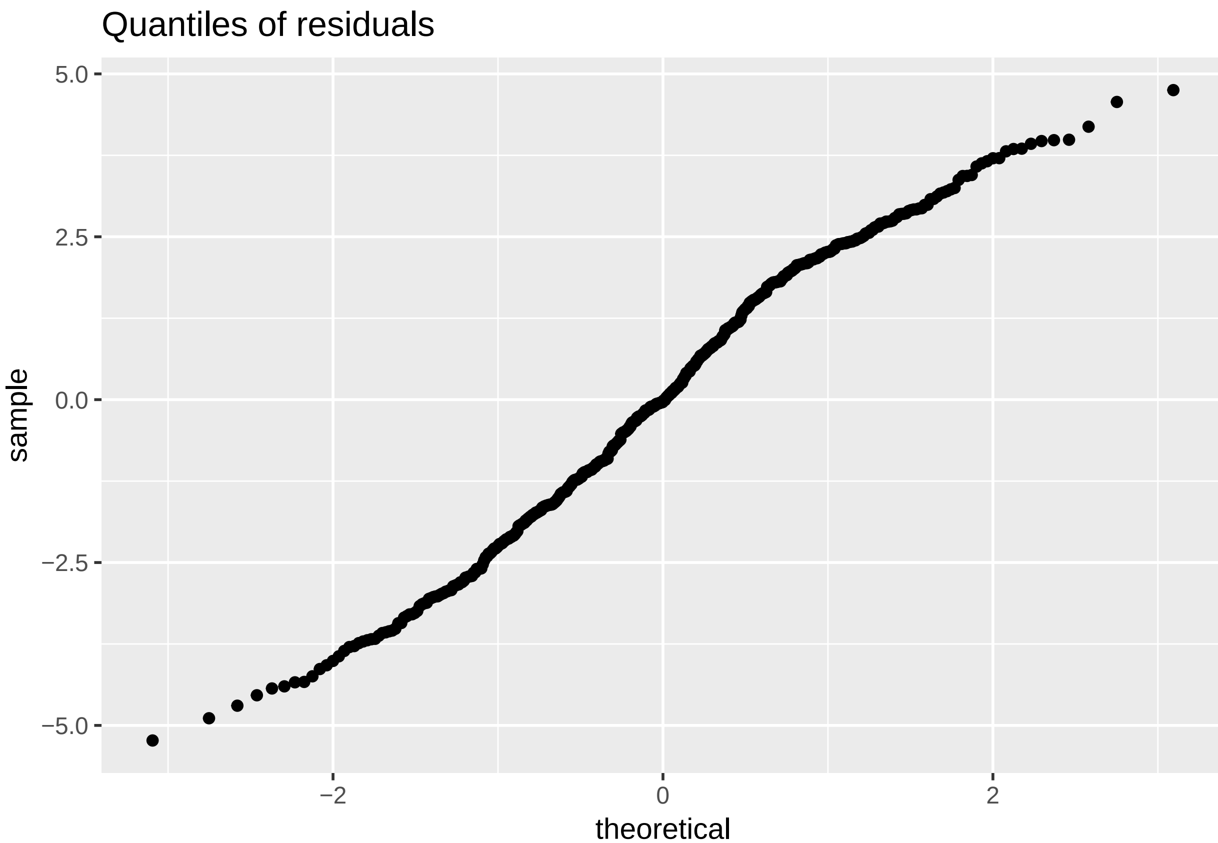
```
BostonHousing %>%
  add_residuals(fit_dis,"lres")%>%
  ggplot(aes(x=log2(dis),y=lres))+
  geom_point()+
  labs(title = "Residuals v/s distance(in logarithmic scale)",
       x = "distance in log scale",
       y = "residual in log scale")
```



The plot shows no systematic pattern when plotted against the predictor variable (dis) used in the model.

We can also plot the quantiles of residuals to verify that the distibution of errors is normal.

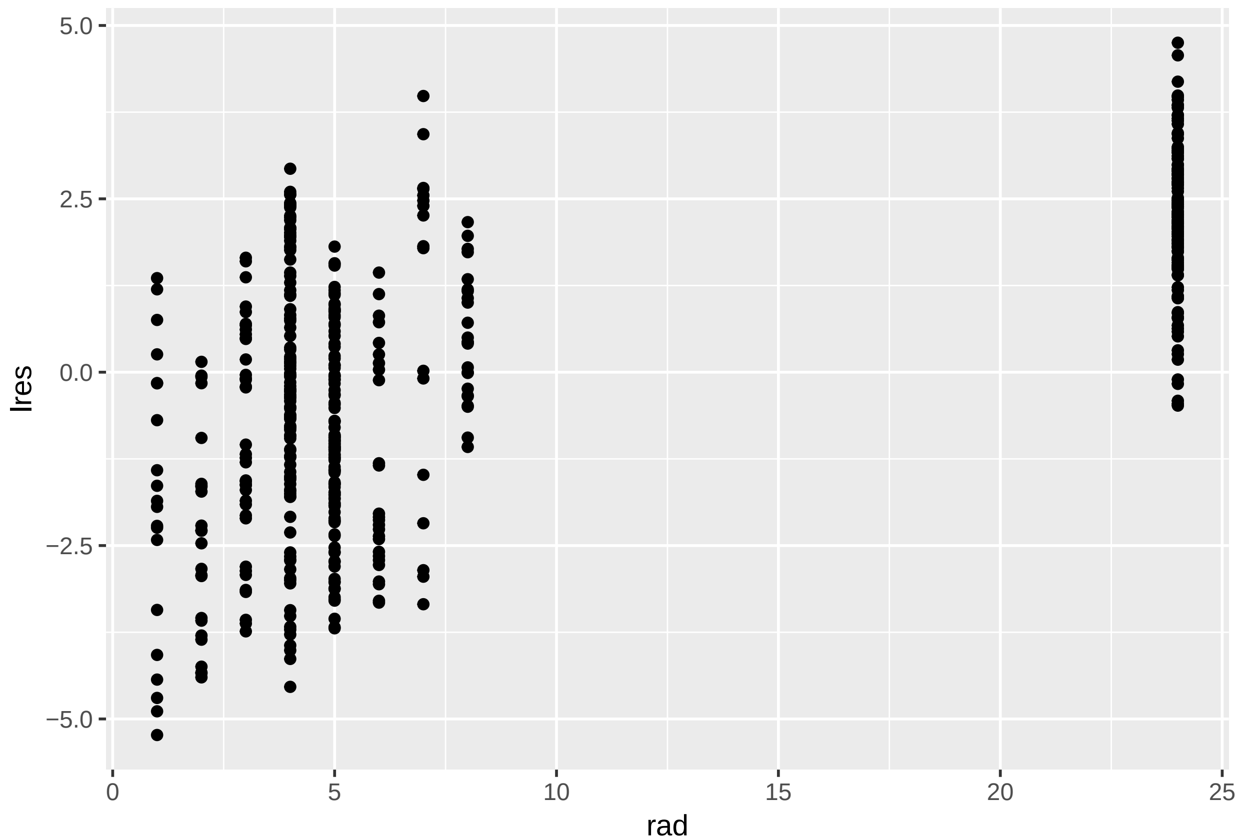
```
BostonHousing %>%
add_residuals(fit_dis) %>%
ggplot(aes(sample=resid)) +
geom_qq()+
  labs(title="Quantiles of residuals")
```



The plot being linear confirms that the distribution of error is normal.

While plotting the residuals of the fitted model with the other predictor variables, I found a pattern with residuals and rad.

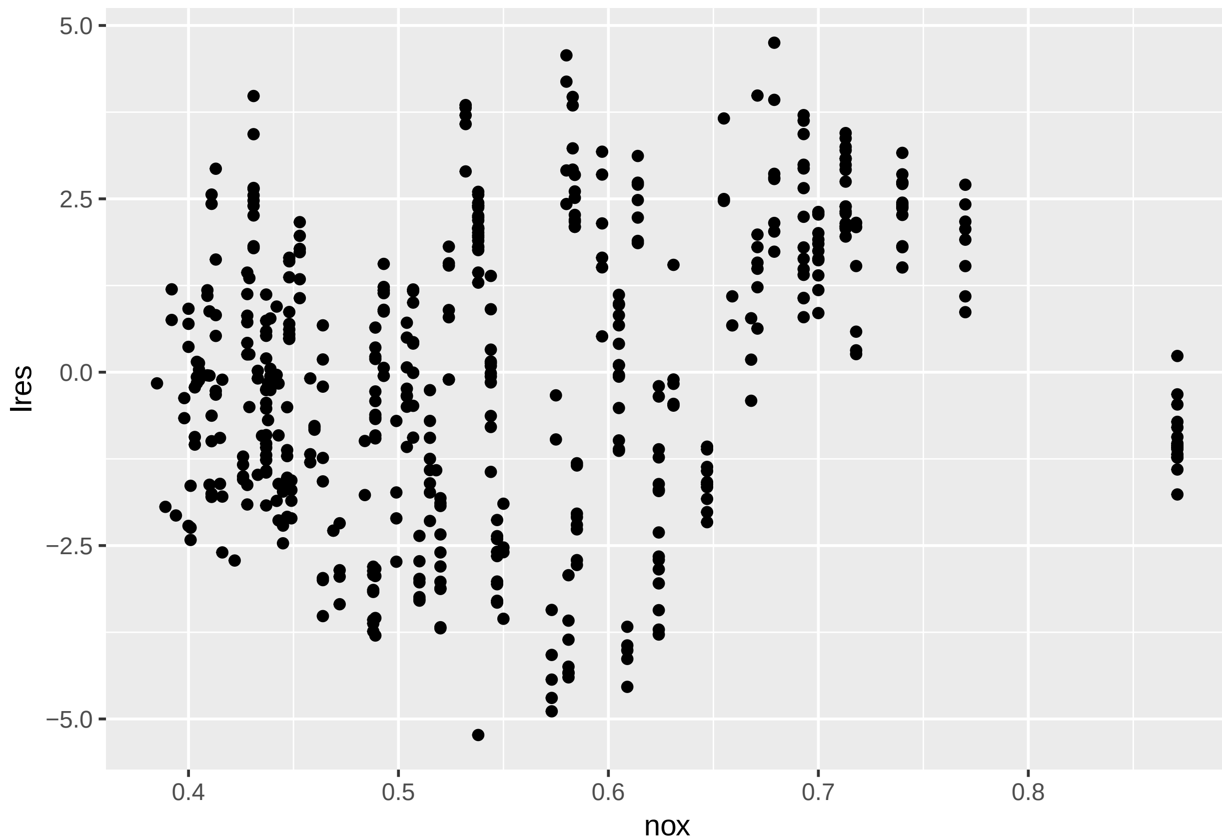
```
BostonHousing%>%
  add_residuals(fit_dis,"lres")%>%
  ggplot(aes(x=rad,y=lres))+
  geom_point()
```



From the plot I am able to see a systematic pattern which is not random taking zero lres=0 as the refrence line. The plot shows that for rad less than 10 my model is mostly overpredicting the crime rate and when rad is greater than 10 my model underpredicts it.

While plotting the residuals of the fitted model with the other predictor variables, I found a pattern with residuals and nox also.

```
BostonHousing%>%
  add_residuals(fit_dis,"lres")%>%
  ggplot(aes(x=nox,y=lres))+
  geom_point()
```



From the plot I am able to see a systematic pattern which is not random taking zero lres=0 as the refrence line. There are certain portions of the graph which underpredicts the model as nox is greater than 0.65, hence there is a scope of improvement by adding nox to the model.

For the other potential predictor variables as listed below, I did not see much of a systematic pattern when plotted with the residuals of my fitted model.

```
BostonHousing%>%
  add_residuals(fit_dis,"lres")%>%
  ggplot(aes(x=indus,y=lres))+
  geom_point()

BostonHousing%>%
  add_residuals(fit_dis,"lres")%>%
  ggplot(aes(x=log2(indus),y=lres))+
  geom_point()

BostonHousing%>%
  add_residuals(fit_dis,"lres")%>%
  ggplot(aes(x=age,y=lres))+
  geom_point()

BostonHousing%>%
  add_residuals(fit_dis,"lres")%>%
  ggplot(aes(x=log2(age),y=lres))+
  geom_point()

BostonHousing%>%
  add_residuals(fit_dis,"lres")%>%
  ggplot(aes(x=tax,y=lres))+
  geom_point()

BostonHousing%>%
  add_residuals(fit_dis,"lres")%>%
  ggplot(aes(x=log2(tax),y=lres))+
  geom_point()

BostonHousing%>%
  add_residuals(fit_dis,"lres")%>%
  ggplot(aes(x=lstat,y=lres))+
  geom_point()

BostonHousing%>%
  add_residuals(fit_dis,"lres")%>%
  ggplot(aes(x=log2(lstat),y=lres))+
  geom_point()

BostonHousing%>%
  add_residuals(fit_dis,"lres")%>%
  ggplot(aes(x=medv,y=lres))+
  geom_point()

BostonHousing%>%
  add_residuals(fit_dis,"lres")%>%
  ggplot(aes(x=log2(medv),y=lres))+
  geom_point()
```

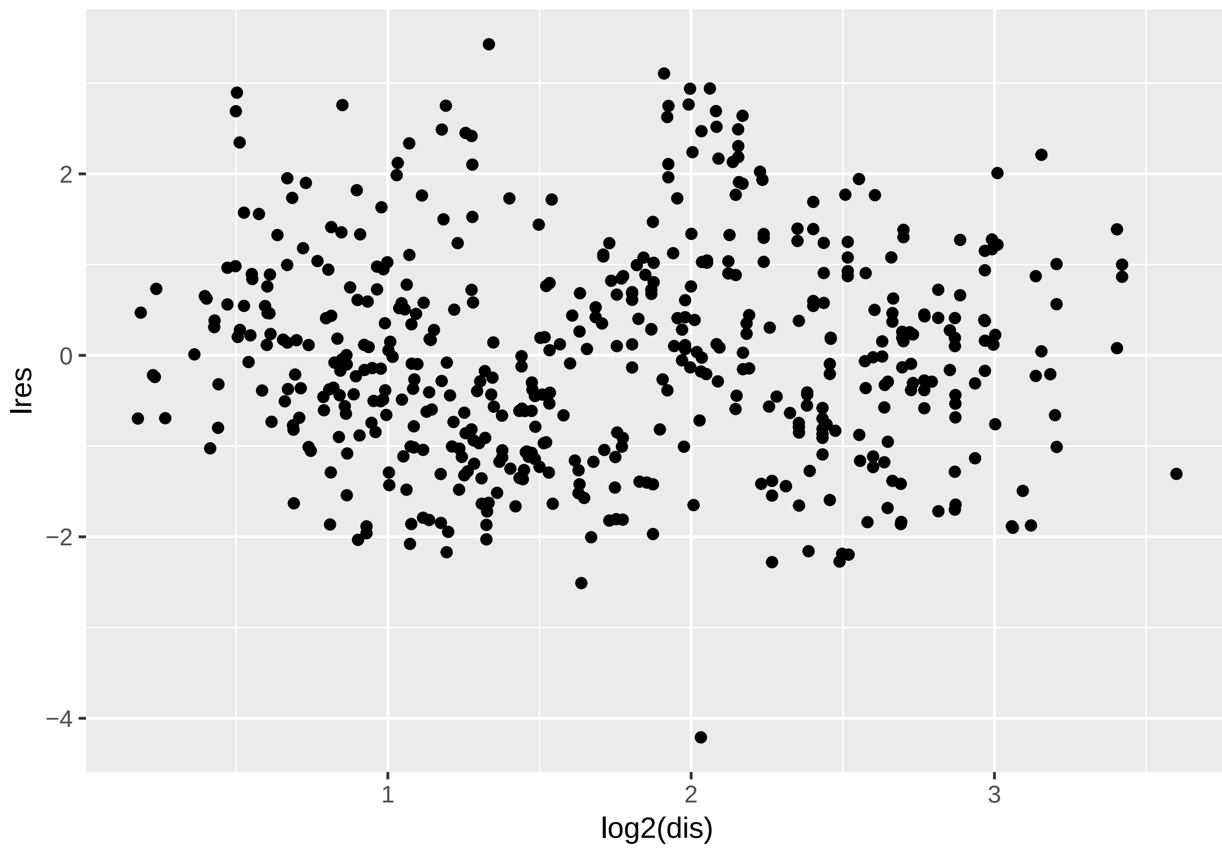
Problem 5

Based on the residuals plots plotted above, I am adding rad and nox as a predictor variable to my existing model.

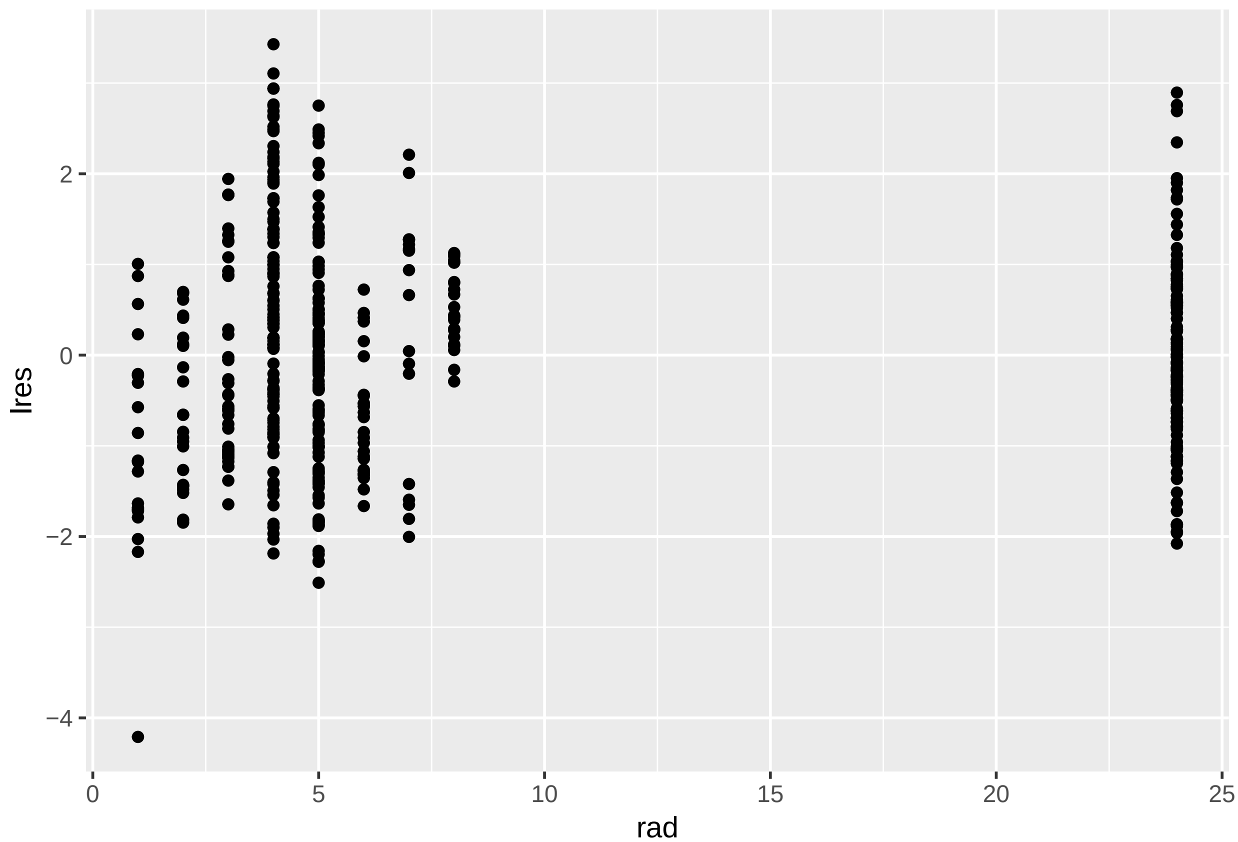
```
fit_dis_rad_nox <- lm(log2(crim)~log2(dis)+rad+log2(nox),data=BostonHousing)
summary(fit_dis_rad_nox)
```

```
##
## Call:
## lm(formula = log2(crim) ~ log2(dis) + rad + log2(nox), data = BostonHousing)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.2100 -0.8505 -0.0690  0.7566  3.4284
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.02429    0.26480   3.868 0.000124 ***
## log2(dis)   -0.57098    0.13310  -4.290 2.14e-05 ***
## rad          0.20328    0.00778  26.128  < 2e-16 ***
## log2(nox)    3.53691    0.37829   9.350  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.186 on 502 degrees of freedom
## Multiple R-squared:  0.8563, Adjusted R-squared:  0.8554
## F-statistic: 996.8 on 3 and 502 DF,  p-value: < 2.2e-16
```

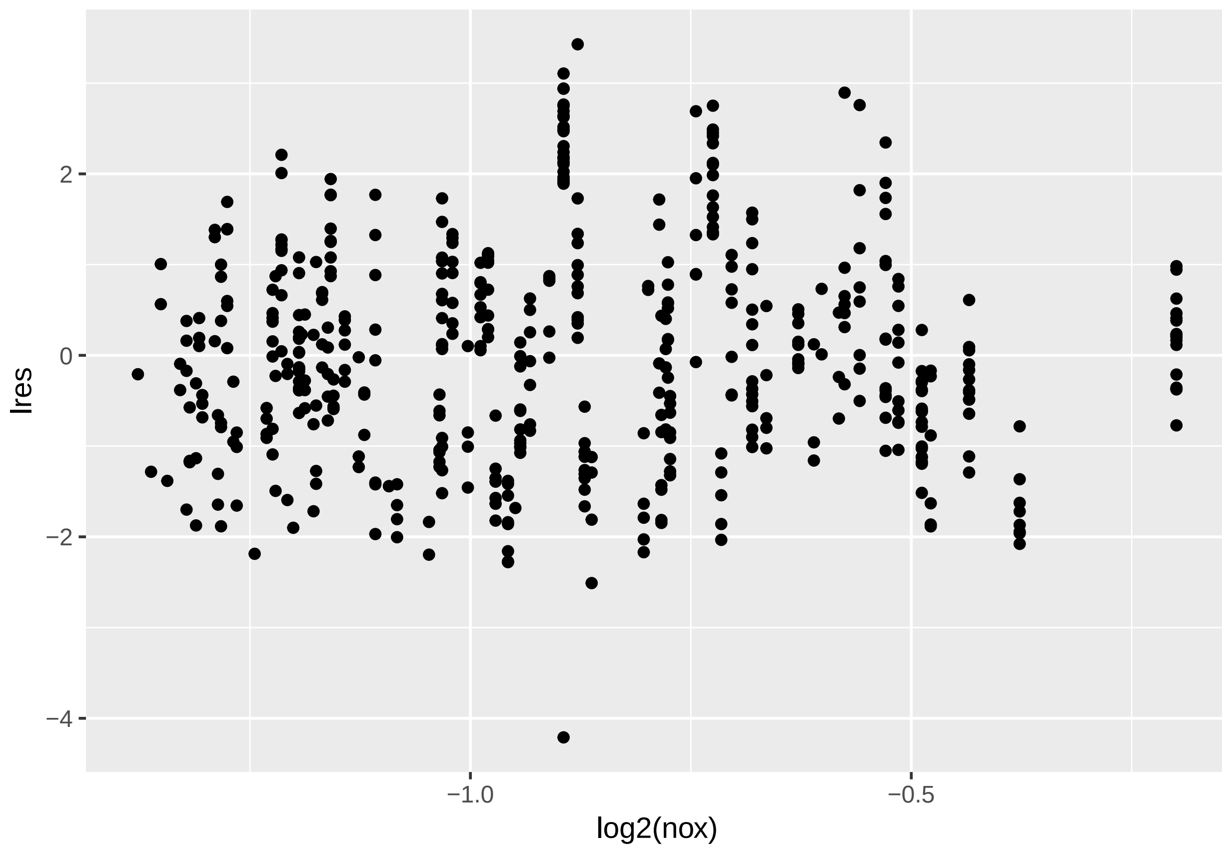
```r
BostonHousing%>%
  add_residuals(fit_dis_rad_nox,"lres")%>%
  ggplot(aes(x=log2(dis),y=lres))+
  geom_point()
```



```
BostonHousing%>%
  add_residuals(fit_dis_rad_nox,"lres")%>%
  ggplot(aes(x=rad,y=lres))+
  geom_point()
```



```r
BostonHousing%>%
  add_residuals(fit_dis_rad_nox,"lres")%>%
  ggplot(aes(x=log2(nox),y=lres))+
  geom_point()
```



The plot shows no systematic pattern when plotted against the predictor variables dis,rad and nox used in the model.

From the model we can interpret that the crime rate decreases as the distance to employment centre increases meaning the crime rate is more near to the employment centre or the industrial area,and the crime rate increases as the index of accessibility to radial highway increases. The crime rate also increases as the nitric oxide concentration increases probably the the nitric oxide concentration increases near to the industrial area hence the crime rate is more near to the industrial area.