hw5-Deep-Bhalodia

Deep Bhalodia

3/16/2019

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

Loading required packages

```
library(ggplot2)
library(dplyr)
library(rlang)
library(readr)
library(corrplot)
library(mlbench)
library(modelr)
library(tidyverse)

eval = FALSE
```

Part A

Problem 1

Choose one of the "miniposters" created by your fellow classmates and posted on Piazza for Homework 3. Cite both the name of the student whose miniposter you chose and the original source of the dataset used in that miniposter. Download and import that dataset into R, put it into a tidy format (if necessary), and print the first ten observations of the dataset.

Miniposter Used

Name - Harsh Shah

Source - https://www.kaggle.com/uciml/breast-cancer-wisconsin-data

```
dataset <- read_csv("Cancer.csv")
```

```
## Warning: Missing column names filled in: 'X33' [33]
```

```
## Parsed with column specification:
## cols(
##   .default = col_double(),
##   diagnosis = col_character(),
##   X33 = col_character()
## )
```

```
## See spec(...) for full column specifications.
```

```
## Warning: 569 parsing failures.
## row col   expected     actual         file
##   1  -- 33 columns 32 columns 'Cancer.csv'
##   2  -- 33 columns 32 columns 'Cancer.csv'
##   3  -- 33 columns 32 columns 'Cancer.csv'
##   4  -- 33 columns 32 columns 'Cancer.csv'
##   5  -- 33 columns 32 columns 'Cancer.csv'
## ... ... ......... ......... ...........
## See problems(...) for more details.
```

```
colSums(is.na(dataset))
```

```
##                      id               diagnosis             radius_mean
##                       0                       0                       0
##            texture_mean         perimeter_mean               area_mean
##                       0                       0                       0
##         smoothness_mean        compactness_mean          concavity_mean
##                       0                       0                       0
##     concave points_mean          symmetry_mean  fractal_dimension_mean
##                       0                       0                       0
##               radius_se              texture_se            perimeter_se
##                       0                       0                       0
##                 area_se           smoothness_se          compactness_se
##                       0                       0                       0
##            concavity_se       concave points_se             symmetry_se
##                       0                       0                       0
##    fractal_dimension_se            radius_worst           texture_worst
##                       0                       0                       0
##         perimeter_worst              area_worst        smoothness_worst
##                       0                       0                       0
##       compactness_worst         concavity_worst    concave points_worst
##                       0                       0                       0
##          symmetry_worst fractal_dimension_worst                     X33
##                       0                       0                     569
```

```
dataset <- dataset[,-33]


dataset[1:10,]
```

```
## # A tibble: 10 x 32
##        id diagnosis radius_mean texture_mean perimeter_mean area_mean
##     <dbl> <chr>           <dbl>        <dbl>          <dbl>     <dbl>
##  1 8.42e5 M                18.0         10.4          123.      1001
##  2 8.43e5 M                20.6         17.8          133.      1326
##  3 8.43e7 M                19.7         21.2          130       1203
##  4 8.43e7 M                11.4         20.4           77.6      386.
##  5 8.44e7 M                20.3         14.3          135.      1297
##  6 8.44e5 M                12.4         15.7           82.6      477.
##  7 8.44e5 M                18.2         20.0          120.      1040
##  8 8.45e7 M                13.7         20.8           90.2      578.
##  9 8.45e5 M                13           21.8           87.5      520.
## 10 8.45e7 M                12.5         24.0           84.0      476.
## # ... with 26 more variables: smoothness_mean <dbl>,
```

```
## #   compactness_mean <dbl>, concavity_mean <dbl>, `concave
## #   points_mean` <dbl>, symmetry_mean <dbl>, fractal_dimension_mean <dbl>,
## #   radius_se <dbl>, texture_se <dbl>, perimeter_se <dbl>, area_se <dbl>,
## #   smoothness_se <dbl>, compactness_se <dbl>, concavity_se <dbl>,
## #   `concave points_se` <dbl>, symmetry_se <dbl>,
## #   fractal_dimension_se <dbl>, radius_worst <dbl>, texture_worst <dbl>,
## #   perimeter_worst <dbl>, area_worst <dbl>, smoothness_worst <dbl>,
## #   compactness_worst <dbl>, concavity_worst <dbl>, `concave
## #   points_worst` <dbl>, symmetry_worst <dbl>,
## #   fractal_dimension_worst <dbl>
```

```
summary(dataset)
```

```
##        id              diagnosis          radius_mean      texture_mean
##  Min.   :     8670   Length:569         Min.   : 6.981   Min.   : 9.71
##  1st Qu.:   869218   Class :character   1st Qu.:11.700   1st Qu.:16.17
##  Median :   906024   Mode  :character   Median :13.370   Median :18.84
##  Mean   : 30371831                      Mean   :14.127   Mean   :19.29
##  3rd Qu.:  8813129                      3rd Qu.:15.780   3rd Qu.:21.80
##  Max.   :911320502                      Max.   :28.110   Max.   :39.28
##  perimeter_mean     area_mean      smoothness_mean   compactness_mean
##  Min.   : 43.79   Min.   : 143.5   Min.   :0.05263   Min.   :0.01938
##  1st Qu.: 75.17   1st Qu.: 420.3   1st Qu.:0.08637   1st Qu.:0.06492
##  Median : 86.24   Median : 551.1   Median :0.09587   Median :0.09263
##  Mean   : 91.97   Mean   : 654.9   Mean   :0.09636   Mean   :0.10434
##  3rd Qu.:104.10   3rd Qu.: 782.7   3rd Qu.:0.10530   3rd Qu.:0.13040
##  Max.   :188.50   Max.   :2501.0   Max.   :0.16340   Max.   :0.34540
##  concavity_mean    concave points_mean symmetry_mean
##  Min.   :0.00000   Min.   :0.00000     Min.   :0.1060
##  1st Qu.:0.02956   1st Qu.:0.02031     1st Qu.:0.1619
##  Median :0.06154   Median :0.03350     Median :0.1792
##  Mean   :0.08880   Mean   :0.04892     Mean   :0.1812
##  3rd Qu.:0.13070   3rd Qu.:0.07400     3rd Qu.:0.1957
##  Max.   :0.42680   Max.   :0.20120     Max.   :0.3040
##  fractal_dimension_mean   radius_se        texture_se      perimeter_se
##  Min.   :0.04996        Min.   :0.1115   Min.   :0.3602   Min.   : 0.757
##  1st Qu.:0.05770        1st Qu.:0.2324   1st Qu.:0.8339   1st Qu.: 1.606
##  Median :0.06154        Median :0.3242   Median :1.1080   Median : 2.287
##  Mean   :0.06280        Mean   :0.4052   Mean   :1.2169   Mean   : 2.866
##  3rd Qu.:0.06612        3rd Qu.:0.4789   3rd Qu.:1.4740   3rd Qu.: 3.357
##  Max.   :0.09744        Max.   :2.8730   Max.   :4.8850   Max.   :21.980
##     area_se        smoothness_se      compactness_se      concavity_se
##  Min.   :  6.802   Min.   :0.001713   Min.   :0.002252   Min.   :0.00000
##  1st Qu.: 17.850   1st Qu.:0.005169   1st Qu.:0.013080   1st Qu.:0.01509
##  Median : 24.530   Median :0.006380   Median :0.020450   Median :0.02589
##  Mean   : 40.337   Mean   :0.007041   Mean   :0.025478   Mean   :0.03189
##  3rd Qu.: 45.190   3rd Qu.:0.008146   3rd Qu.:0.032450   3rd Qu.:0.04205
##  Max.   :542.200   Max.   :0.031130   Max.   :0.135400   Max.   :0.39600
##  concave points_se   symmetry_se       fractal_dimension_se
##  Min.   :0.000000   Min.   :0.007882   Min.   :0.0008948
##  1st Qu.:0.007638   1st Qu.:0.015160   1st Qu.:0.0022480
##  Median :0.010930   Median :0.018730   Median :0.0031870
##  Mean   :0.011796   Mean   :0.020542   Mean   :0.0037949
##  3rd Qu.:0.014710   3rd Qu.:0.023480   3rd Qu.:0.0045580
##  Max.   :0.052790   Max.   :0.078950   Max.   :0.0298400
```

```
##   radius_worst   texture_worst   perimeter_worst    area_worst
##  Min.   : 7.93   Min.   :12.02   Min.   : 50.41   Min.   : 185.2
##  1st Qu.:13.01   1st Qu.:21.08   1st Qu.: 84.11   1st Qu.: 515.3
##  Median :14.97   Median :25.41   Median : 97.66   Median : 686.5
##  Mean   :16.27   Mean   :25.68   Mean   :107.26   Mean   : 880.6
##  3rd Qu.:18.79   3rd Qu.:29.72   3rd Qu.:125.40   3rd Qu.:1084.0
##  Max.   :36.04   Max.   :49.54   Max.   :251.20   Max.   :4254.0
##  smoothness_worst  compactness_worst concavity_worst  concave points_worst
##  Min.   :0.07117   Min.   :0.02729   Min.   :0.0000   Min.   :0.00000
##  1st Qu.:0.11660   1st Qu.:0.14720   1st Qu.:0.1145   1st Qu.:0.06493
##  Median :0.13130   Median :0.21190   Median :0.2267   Median :0.09993
##  Mean   :0.13237   Mean   :0.25427   Mean   :0.2722   Mean   :0.11461
##  3rd Qu.:0.14600   3rd Qu.:0.33910   3rd Qu.:0.3829   3rd Qu.:0.16140
##  Max.   :0.22260   Max.   :1.05800   Max.   :1.2520   Max.   :0.29100
##  symmetry_worst   fractal_dimension_worst
##  Min.   :0.1565   Min.   :0.05504
##  1st Qu.:0.2504   1st Qu.:0.07146
##  Median :0.2822   Median :0.08004
##  Mean   :0.2901   Mean   :0.08395
##  3rd Qu.:0.3179   3rd Qu.:0.09208
##  Max.   :0.6638   Max.   :0.20750
```

```
table(dataset$diagnosis)
```

```
##
##   B   M
## 357 212
```

```
dataset$diagnosis <- factor(dataset$diagnosis, levels = c("B", "M"), labels = c("Benign", "Malignant"))


summary(dataset$radius_mean)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   6.981  11.700  13.370  14.127  15.780  28.110
```

```
summary(dataset$area_mean)
```

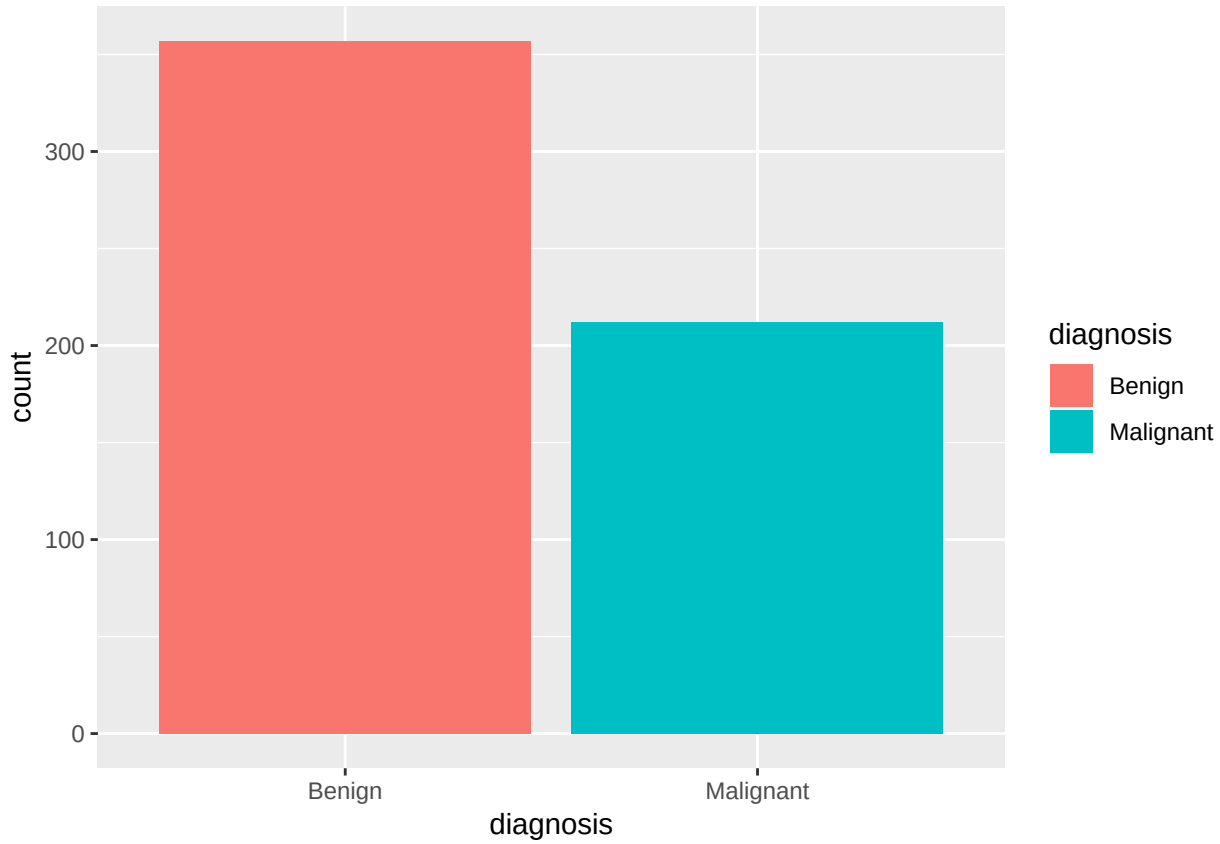
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   143.5   420.3   551.1   654.9   782.7  2501.0
```

```
# The data looks clean now therefore there is no need to further clean the data. We will use this data
```

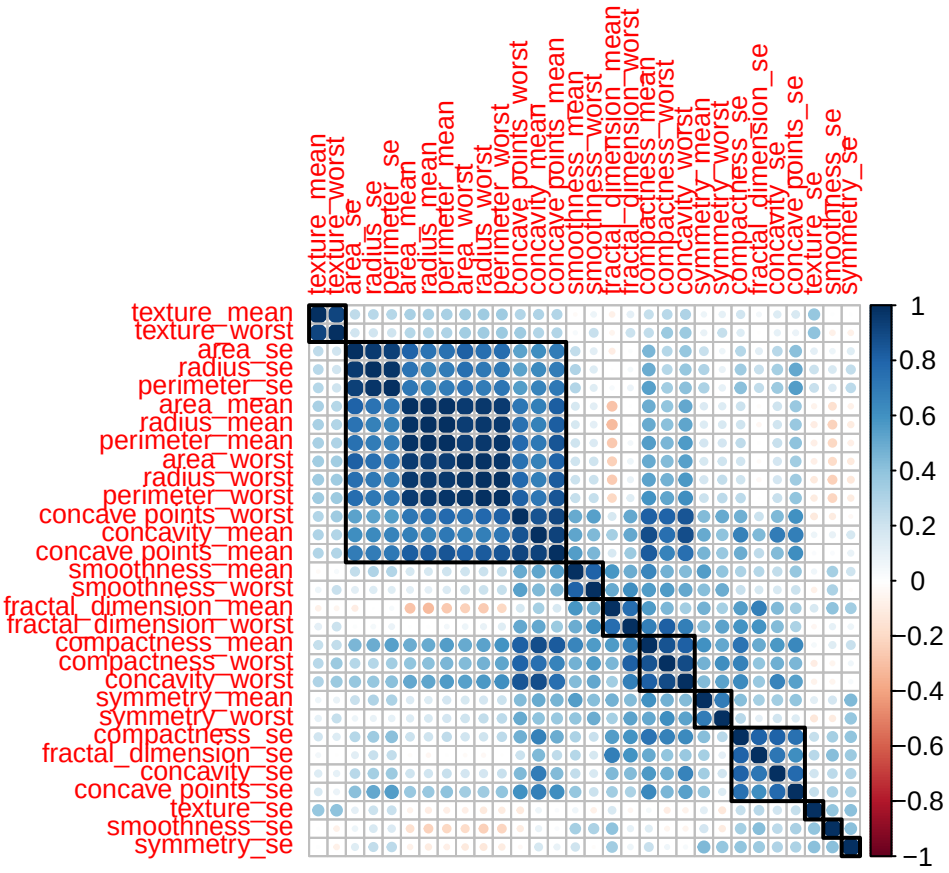
Problem 2

To the best of your ability, reproduce the figures from the miniposter you chose. You may contact the author of the original miniposter; if you do, cite and describe any information you receive from them. (If you are contacted for information on reproducing figures from your own miniposter, you may provide it, but you are not obligated respond.)

```
#The two figures from the miniposter are reproduced below:

ggplot(dataset) + geom_bar(aes(x=diagnosis,fill= diagnosis))
```




```
dataset_new <- dataset
dataset_new <- dataset_new[, -c(1:2)]

correlation_plot <- cor(dataset_new)

corrplot(correlation_plot, order = "hclust", tl.cex=0.8, addrect = 10)
```



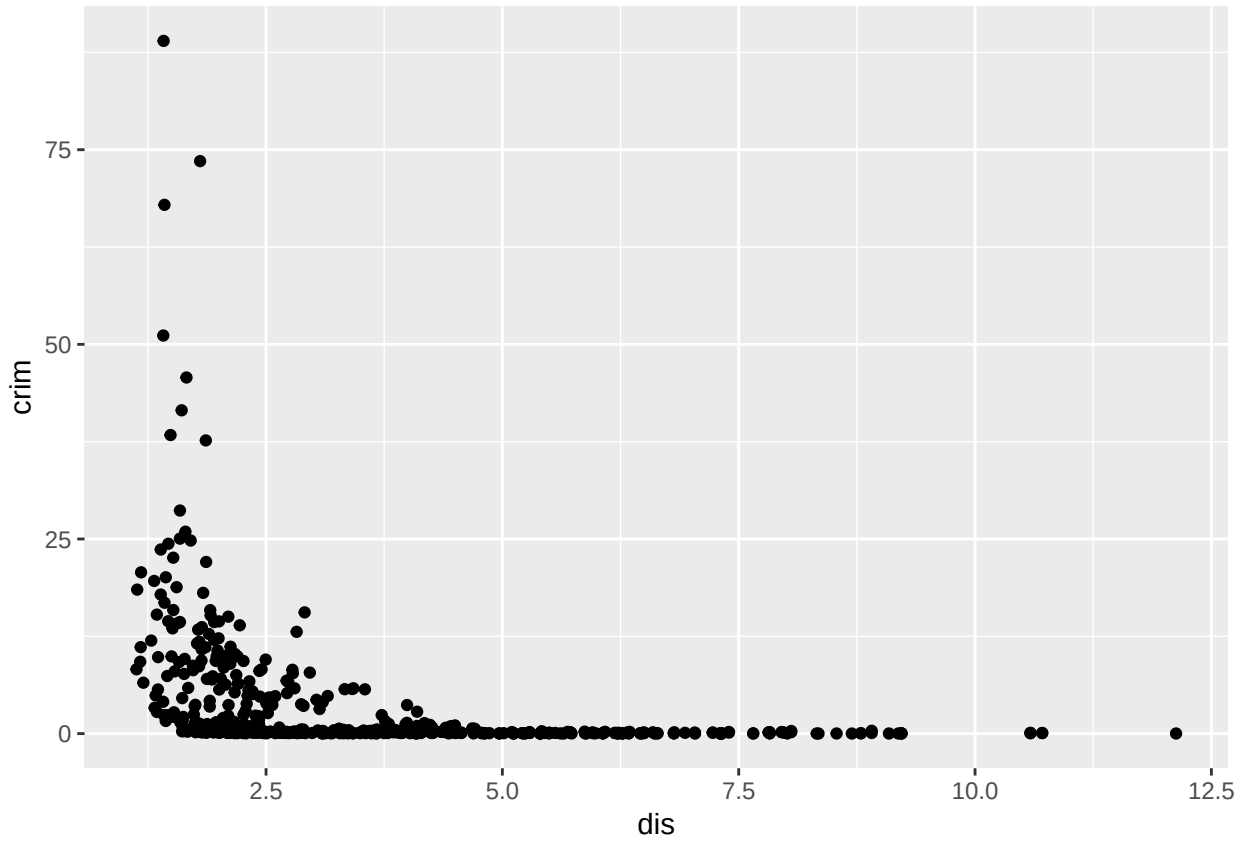

```
# We have successfully created both the plots harsh created in his miniposer
```

Part B

Problem 3

Fit a model that predicts per capita crime rate by town (crim) using only one predictor variable. Use plots to justify your choice of predictor variable and the appropriateness of any transformations you use. Print the values of the fitted model parameters.

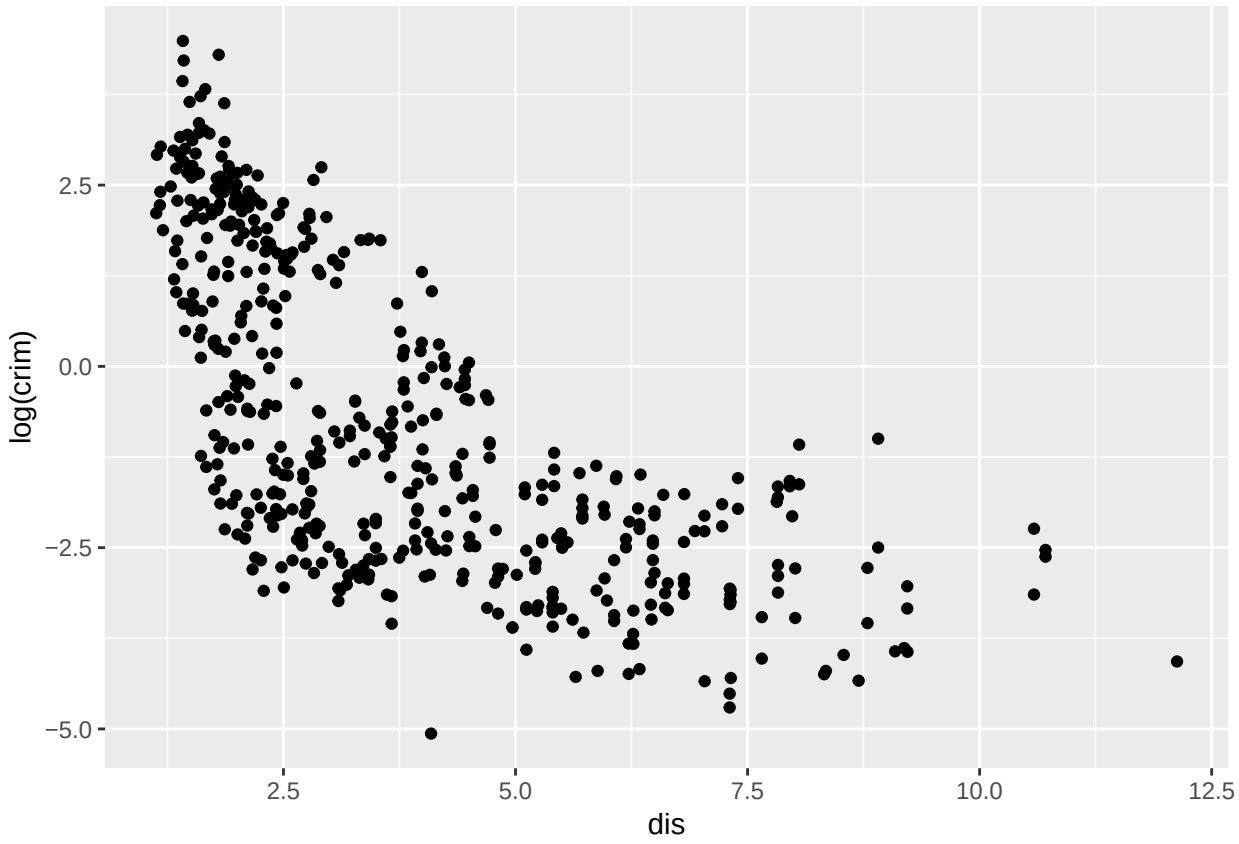
```
data(BostonHousing)

ggplot(BostonHousing, aes(x=dis, y=crim)) + geom_point()
```



Comment

From the scatterplot, there appears to be a negative association between dis and crim, but it's not linear

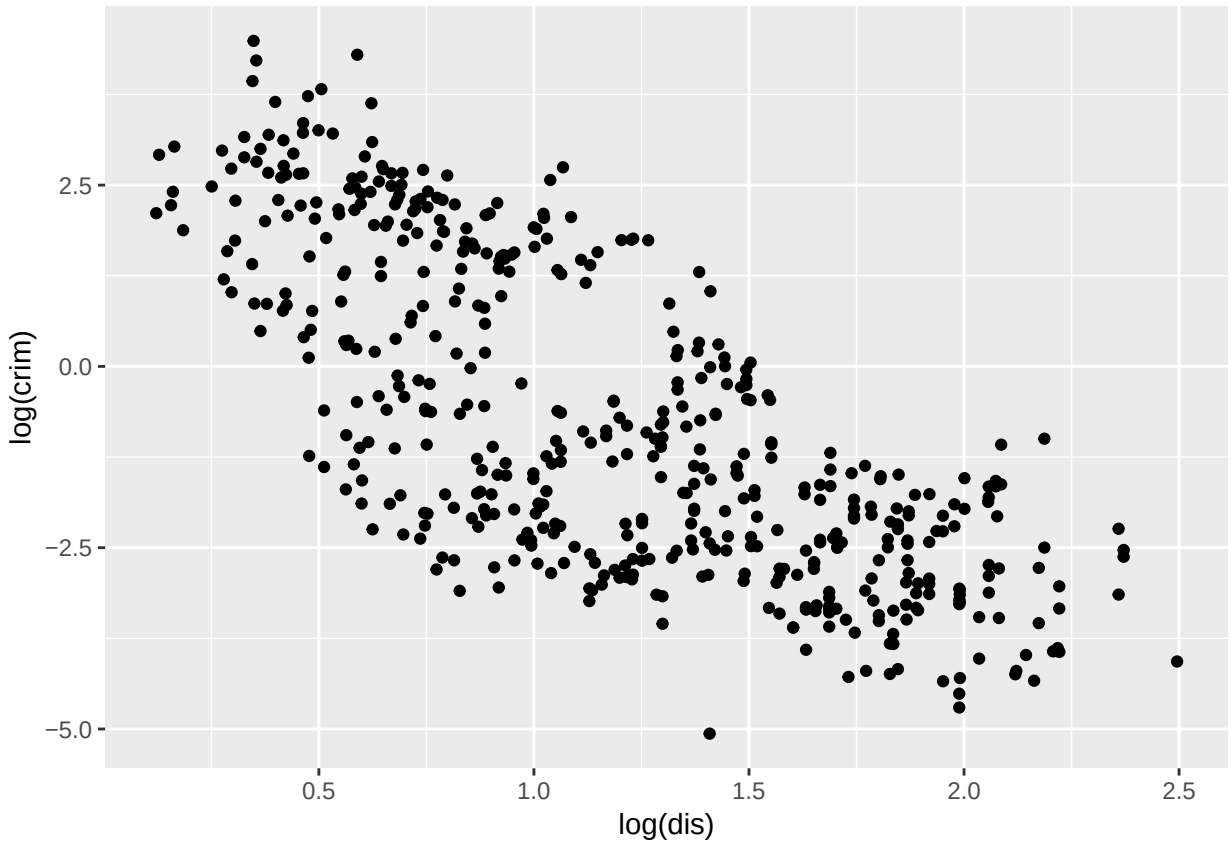
```
ggplot(BostonHousing, aes(x=dis, y=log(crim))) + geom_point()
```



Comment

Log transforming crim improves the relationship, but the relationship is still not quite linear

```
ggplot(BostonHousing, aes(x=log(dis), y=log(crim))) + geom_point()
```



Comment

Log transforming dis as well improves the relationship, making it much more linear. We will include dis as the predictor variable in our model.

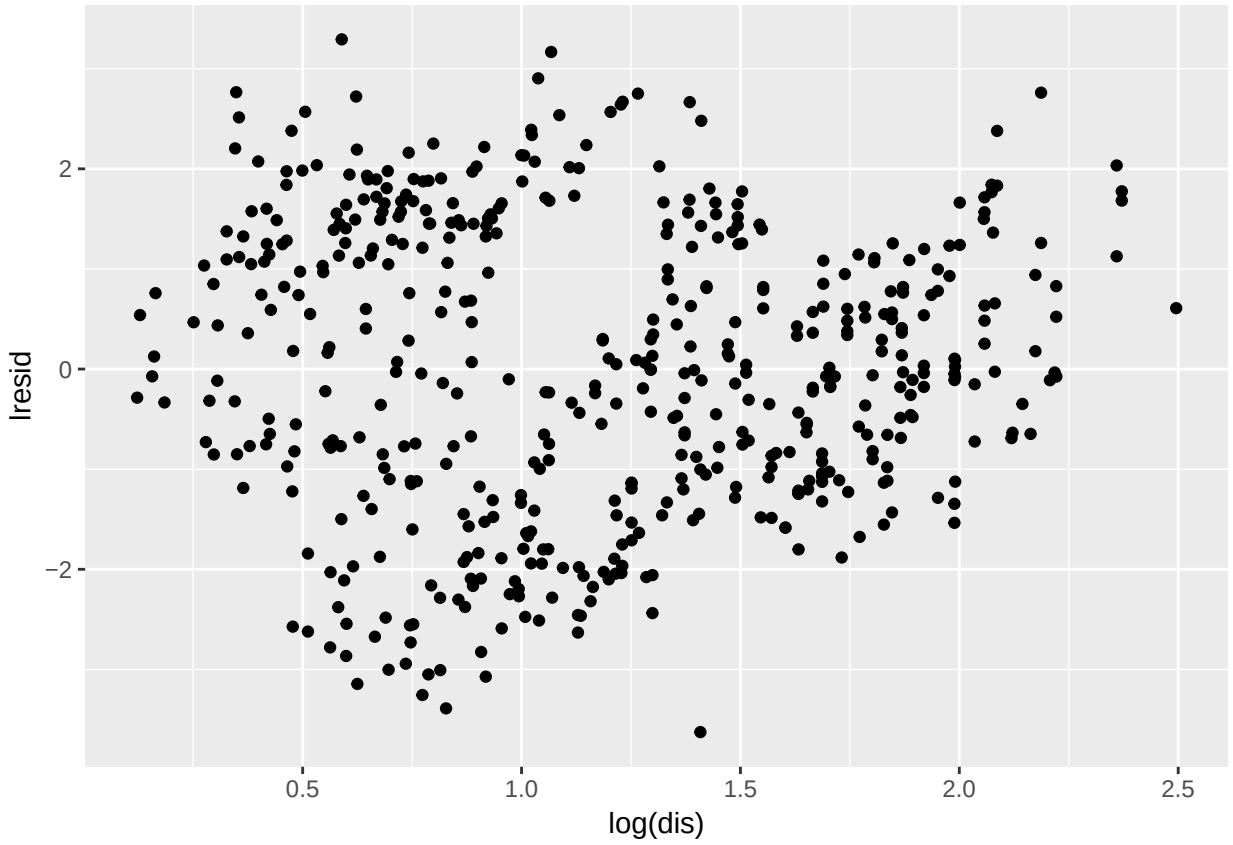
```
fit1 <- lm(log(crim) ~ log(dis), data=BostonHousing)
summary(fit1)
```

```
##
## Call:
## lm(formula = log(crim) ~ log(dis), data = BostonHousing)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.6262 -1.1145 -0.0187  1.2476  3.2931
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.7611     0.1556   17.74   <2e-16 ***
## log(dis)     -2.9810     0.1193  -24.99   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.446 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
```

Problem 4

Plot the residuals of the fitted model from Problem 3 against the predictor variable already in the model and against other potential predictor variables in the dataset. Comment on what you observe in each residual plot.

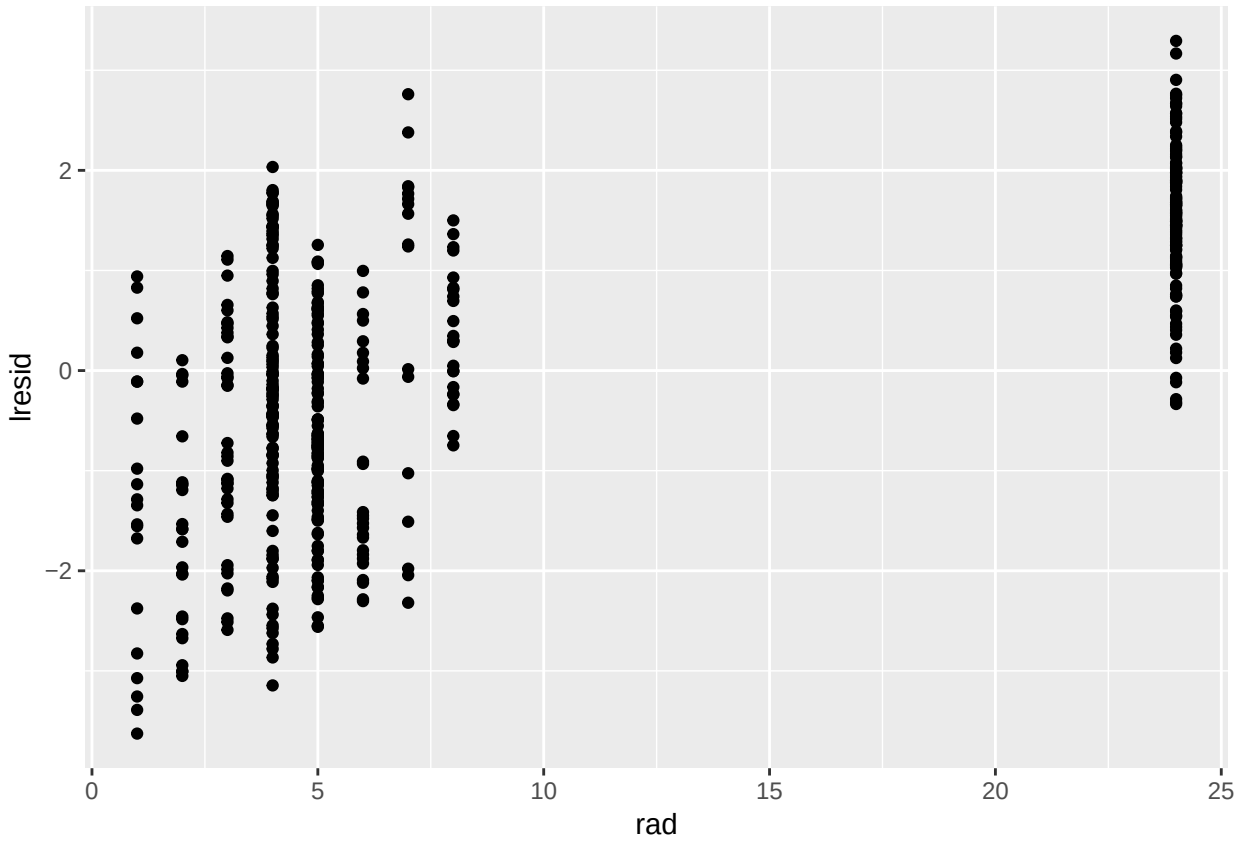
```
BostonHousing %>%
  add_residuals(fit1, "lresid") %>%
  ggplot(aes(x=log(dis), y=lresid)) +
  geom_point()
```



Comment

In the residual plot for log(dis), we mostly see simple random scatter and no systematic patterns, indicating no violation of model assumptions.

```
BostonHousing %>%
  add_residuals(fit1, "lresid") %>%
  ggplot(aes(x=rad, y=lresid)) +
  geom_point()
```



Comment

In the residual plot for rad, we see a positive linear relationship between the log residuals and rad, indicating that there is a relationship between rad and log(crim), so we should add rad as a predictor in the model.

Problem 5

Fit a new model for predicting per capita crime rate by town, adding or removing variables based on the residual plots from Problem 4. Interpret the model.

```
fit2 <- lm(log(crim) ~ log(dis) + rad, data=BostonHousing)
summary(fit2)
```

```
##
## Call:
## lm(formula = log(crim) ~ log(dis) + rad, data = BostonHousing)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.59042 -0.65587 -0.04422  0.57162  2.34690
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.445314   0.146923  -3.031  0.00256 **
## log(dis)    -1.552176   0.088618 -17.515  < 2e-16 ***
```

```
## rad          0.158011   0.005491  28.775  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.89 on 503 degrees of freedom
## Multiple R-squared:  0.8312, Adjusted R-squared:  0.8306
## F-statistic:  1239 on 2 and 503 DF,  p-value: < 2.2e-16
```

```
RMSE <- function(error) { sqrt(mean(error^2)) }

RMSE(fit2$residuals)
```

```
## [1] 0.8873146
```

Comments

It appears that crime has a negative relationship with distance from employment centers. Larger distances result in lower crime rates. Conversely, there is a positive relationship between crime and the index of accessibility to radial highways. Higher indices are associated with higher crime rates. “‘