# Linear Regression

## Module 3 - Activity 4

Of the models with a convex representation of their parametricestimation, generalized linear models (GLM) are a crucial case.The two most frequent examples of GLM are linear regression andlogistic regression. Therefore, in this activity, several linear regressionand logistic regression exercises will be solved with R software andsome of its packages.

### Activities Problem 1: Warm Up

### 1. Section 3.7 Problem 8.

This question involves the use of simple linear regression on the Auto data set.

### a) Use the lm() function to perform a simple linear regression with mpg as the response and horsepower as the predictor. Use the summary() function to print the results. Comment on the output.

Adding the necessary libraries

```
library(ISLR2)
library(tidymodels)
library(dplyr)
library(modelsummary)
```

Loading dataframe

```
auto <- Auto
head(auto)
```

```
  mpg cylinders displacement horsepower weight acceleration year origin
1  18         8          307        130   3504         12.0   70      1
2  15         8          350        165   3693         11.5   70      1
3  18         8          318        150   3436         11.0   70      1
4  16         8          304        150   3433         12.0   70      1
5  17         8          302        140   3449         10.5   70      1
6  15         8          429        198   4341         10.0   70      1
                       name
1 chevrolet chevelle malibu
2         buick skylark 320
3        plymouth satellite
4             amc rebel sst
5               ford torino
6          ford galaxie 500
```

The Auto df counts with 9 columns that show information of different models of vehicles. For the present exercise is intended to evaluate if exists any relationship

```
linearmodelauto <- lm(mpg ~ horsepower, data = auto)
linearmodelauto
```

```

Call:
lm(formula = mpg ~ horsepower, data = auto)

Coefficients:
(Intercept)   horsepower
    39.9359      -0.1578
```

- Is there a relationship between the predictor and the response?

  Yes there is a relationship between the variables

```
summary(linearmodelauto)
```

```

Call:
lm(formula = mpg ~ horsepower, data = auto)

Residuals:
     Min       1Q   Median       3Q      Max
-13.5710  -3.2592  -0.3435   2.7630  16.9240
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 39.935861   0.717499   55.66   <2e-16 ***
horsepower  -0.157845   0.006446  -24.49   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.906 on 390 degrees of freedom
Multiple R-squared:  0.6059,    Adjusted R-squared:  0.6049
F-statistic: 599.7 on 1 and 390 DF,  p-value: < 2.2e-16
```

- How strong is the relationship between the predictor and the response

  With the p values of the model (<0.0001) we can assume that the model is significant to explain the relationship between our variables . Also the $R^2$ of 60% indicates that there is a highly correlation between this variables

- Is the relationship between the predictor and the response positive or negative?

  The relationship is negative

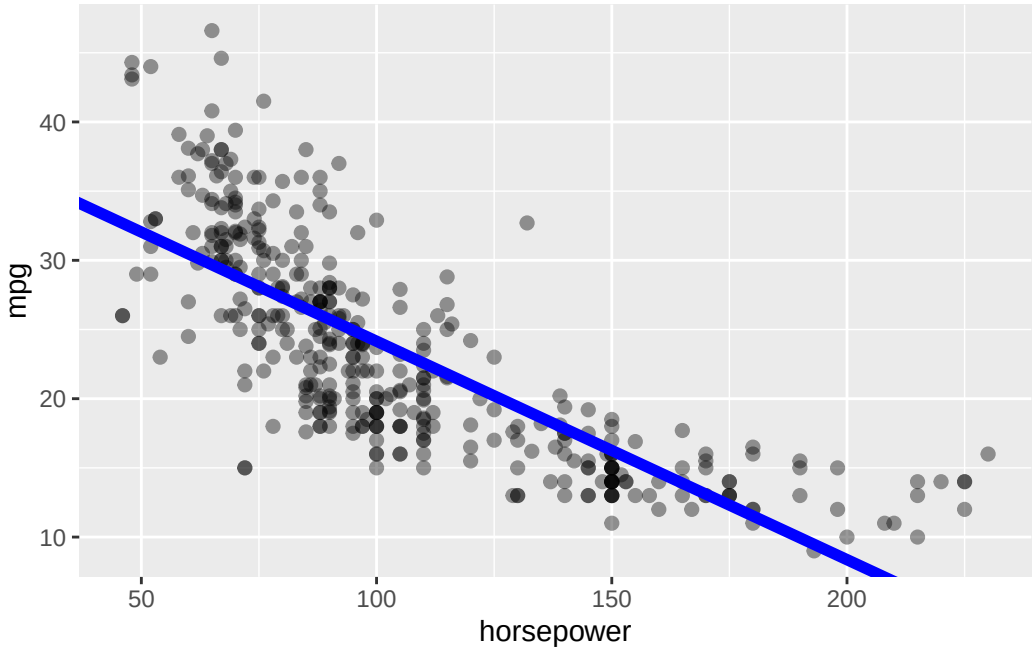
- What is the predicted mpg associated with a horsepower of 98? What are the associated 95 % confidence and prediction intervals?

```r
predict(linearmodelauto, tibble(horsepower=98), interval = "confidence")
```

```
       fit      lwr      upr
1 24.46708 23.97308 24.96108
```

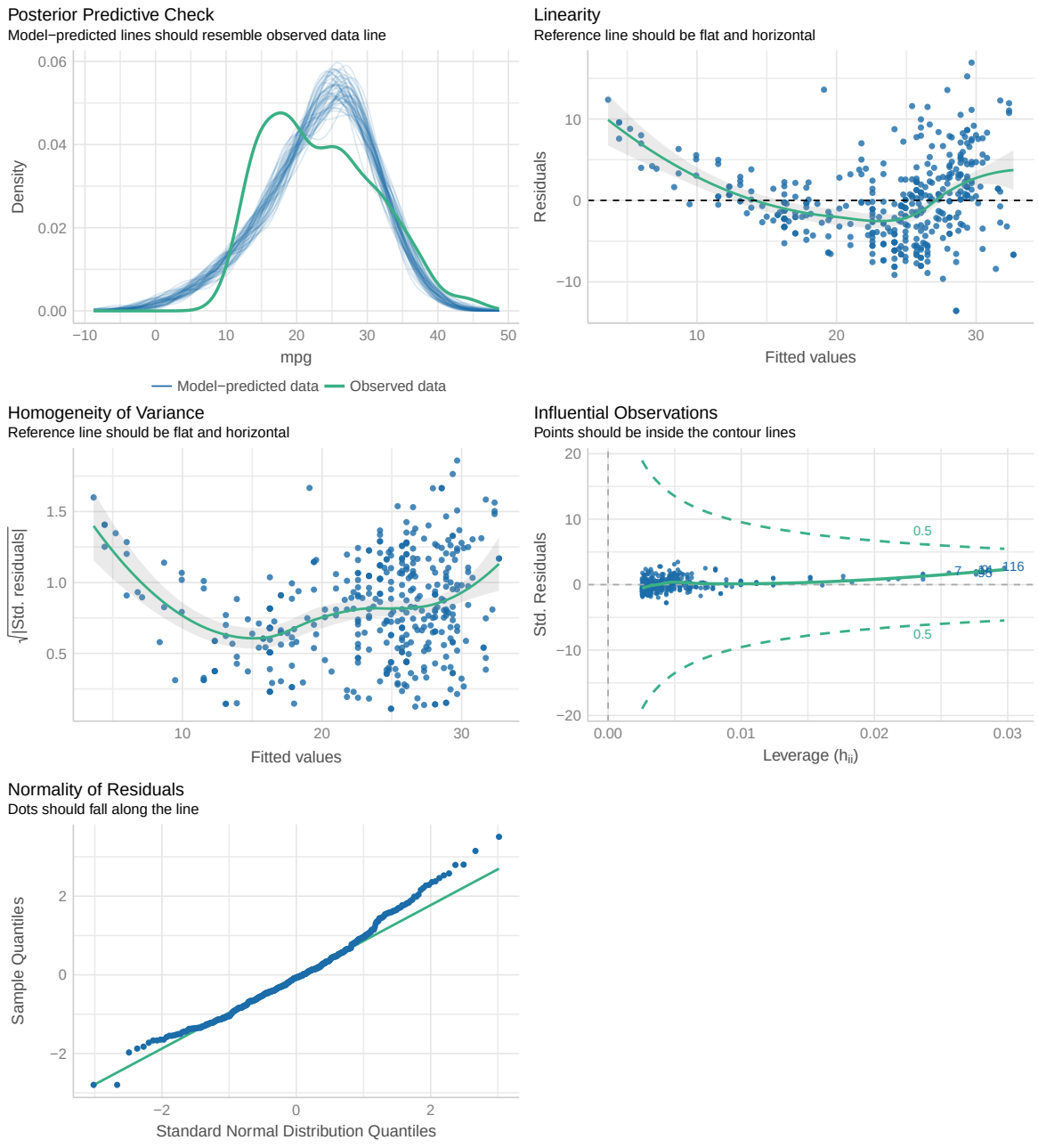
b) Plot the response and the predictor. Use the abline() functionto display the least squares regression line.

```r
auto %>%
  ggplot(aes(x = horsepower)) +
  geom_point(aes(y = mpg), size = 2, alpha = 0.4) +
  geom_abline(slope = coef(linearmodelauto)["horsepower"],
              intercept = coef(linearmodelauto)["(Intercept)"],
              size = 2, color = "blue")
```




**c) Use the plot() function to produce diagnostic plots of the least squares regression fit. Comment on any problems you see with the fit.**

```
linearmodelauto %>% performance::check_model()
```

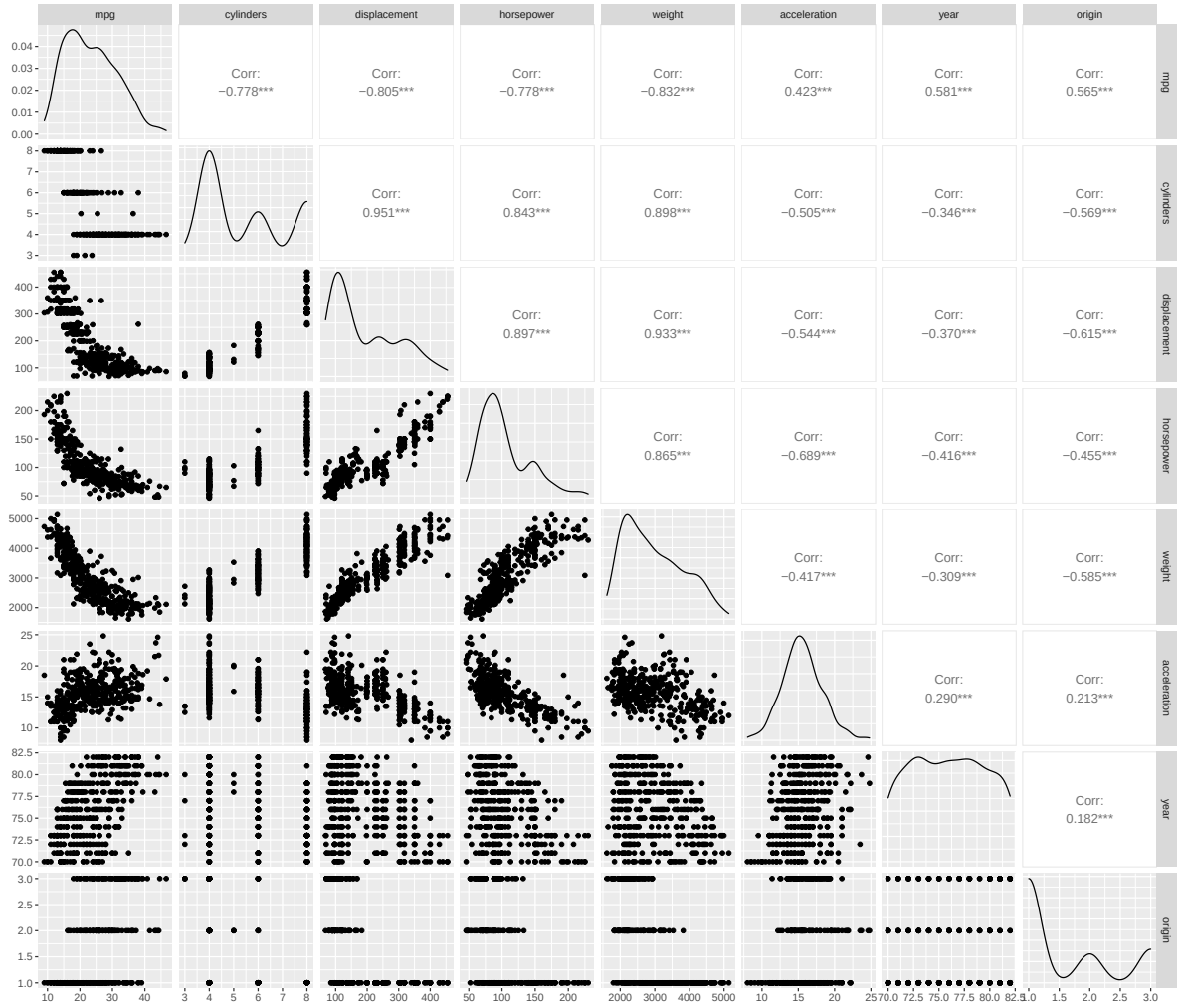



## 1. Section 3.7 Problem 9.

This question involves the use of simple linear regression on the Auto data set.

**a)Produce a scatter-plot matrix which includes all of the variables in the data set.**

```
p <- GGally::ggpairs(auto %>% select(-name))
print(p,progress=F)
```




**b)Compute the matrix of correlations between the variables using the function cor(). You will need to exclude the name variable, cor()which is qualitative.**

```
cor(auto %>% select(-name))
```

```
                    mpg  cylinders displacement horsepower     weight
mpg           1.0000000 -0.7776175   -0.8051269 -0.7784268 -0.8322442
cylinders    -0.7776175  1.0000000    0.9508233  0.8429834  0.8975273
```

```
displacement -0.8051269  0.9508233    1.0000000  0.8972570  0.9329944
horsepower   -0.7784268  0.8429834    0.8972570  1.0000000  0.8645377
weight       -0.8322442  0.8975273    0.9329944  0.8645377  1.0000000
acceleration  0.4233285 -0.5046834   -0.5438005 -0.6891955 -0.4168392
year          0.5805410 -0.3456474   -0.3698552 -0.4163615 -0.3091199
origin        0.5652088 -0.5689316   -0.6145351 -0.4551715 -0.5850054
             acceleration       year     origin
mpg             0.4233285  0.5805410  0.5652088
cylinders      -0.5046834 -0.3456474 -0.5689316
displacement   -0.5438005 -0.3698552 -0.6145351
horsepower     -0.6891955 -0.4163615 -0.4551715
weight         -0.4168392 -0.3091199 -0.5850054
acceleration    1.0000000  0.2903161  0.2127458
year            0.2903161  1.0000000  0.1815277
origin          0.2127458  0.1815277  1.0000000
```

**c)Use thelm() function to perform a multiple linear regression with mpgas the response and all other variables except name ast he predictors. Use thesummary() function to print the results.Comment on the output. For instance:**

```
multiplelinearmodel <- recipe(mpg ~ ., data = auto) %>% step_rm(name)

mlrworkflow <- workflow() %>% add_recipe(multiplelinearmodel) %>% add_model(linear_reg())

mlrworkflow
```

```
== Workflow ====================================================================
Preprocessor: Recipe
Model: linear_reg()

-- Preprocessor ----------------------------------------------------------------
1 Recipe Step

* step_rm()

-- Model -----------------------------------------------------------------------
Linear Regression Model Specification (regression)

Computational engine: lm
```

```r
mlrfit <- mlrworkflow %>% fit(data = auto)
mlrfitengine <- extract_fit_engine(mlrfit)
summary(mlrfitengine)
```

```
Call:
stats::lm(formula = ..y ~ ., data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-9.5903 -2.1565 -0.1169  1.8690 13.0604

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  -17.218435   4.644294  -3.707  0.00024 ***
cylinders     -0.493376   0.323282  -1.526  0.12780
displacement   0.019896   0.007515   2.647  0.00844 **
horsepower    -0.016951   0.013787  -1.230  0.21963
weight        -0.006474   0.000652  -9.929  < 2e-16 ***
acceleration   0.080576   0.098845   0.815  0.41548
year           0.750773   0.050973  14.729  < 2e-16 ***
origin         1.426141   0.278136   5.127 4.67e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.328 on 384 degrees of freedom
Multiple R-squared:  0.8215,    Adjusted R-squared:  0.8182
F-statistic: 252.4 on 7 and 384 DF,  p-value: < 2.2e-16
```

1. Is there a relationship between the predictors and the response?

   There is a relationship between the predictors and the response ($p < 0.001$) therefore the model can explain the relationship. Also the $R^2$ is $> 80\%$ explaining a lot of variance with this model

2. Which predictors appear to have a statistically significant relationship to the response?

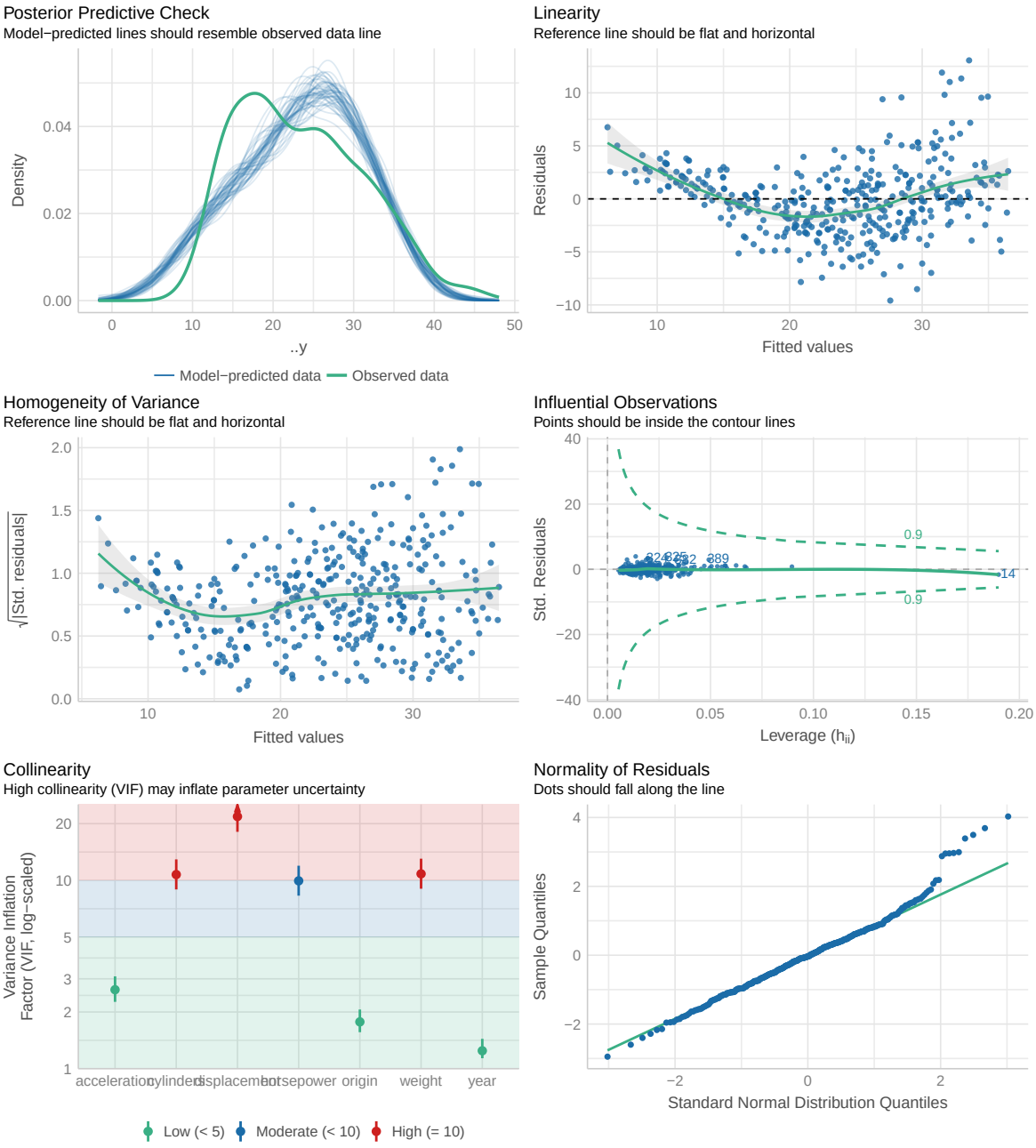
   Displacement, Weight, Year, Origin

3. What does the coefficient for the *year* variable suggest?

   The relationship between the mpg and the year is positive recent years make that the miles per gallon performance increase

**d)Use theplot()function to produce diagnostic plots of the linearregression fit. Comment on any problems you see with the fit.Do the residual plots suggest any unusually large outliers? Doesthe leverage plot identify any observations with unusually high leverage?**

```
mlrfit %>% extract_fit_engine() %>% performance::check_model()
```




**e)Use the $*$ and : symbols to fit linear regression models with interaction effects. Do any interactions appear to be statistically significant?**

```r
multiplelinearmodelint <- lm(mpg ~ cylinders * displacement , data = auto)

summary(multiplelinearmodelint)
```

```
Call:
lm(formula = mpg ~ cylinders * displacement, data = auto)

Residuals:
     Min       1Q   Median       3Q      Max
-16.0432  -2.4308  -0.2263   2.2048  20.9051

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)           48.22040    2.34712  20.545  < 2e-16 ***
cylinders             -2.41838    0.53456  -4.524 8.08e-06 ***
displacement          -0.13436    0.01615  -8.321 1.50e-15 ***
cylinders:displacement  0.01182    0.00207   5.711 2.24e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.454 on 388 degrees of freedom
Multiple R-squared:  0.6769,    Adjusted R-squared:  0.6744
F-statistic:   271 on 3 and 388 DF,  p-value: < 2.2e-16
```

**f)Try a few different transformations of the variables, such aslog(X),√X,X2. Comment on your findings.**