Lab 4: Simple Linear Regression in R

STAT 310, Spring 2021

In this lab we will go over how to fit a simple linear regression (SLR) model in R. We will again use the NHANES data set, which was introduced in Lab 1.

```
# read in data set
nhanes <- readRDS(url("https://ericwfox.github.io/data/nhanes.rds"))
```

```
# check dimension (number of rows and columns)
dim(nhanes)
```

```
## [1] 1500   11
```

```
# get columns names
names(nhanes)
```

```
##  [1] "Gender"     "Age"        "Education"  "HHIncome"   "Weight"
##  [6] "Height"     "BPSysAve"   "BPDiaAve"   "HealthGen"  "PhysActive"
## [11] "Smoke100"
```

Type the following command to look at a scrollable, spreadsheet display of the data set:

```
View(nhanes)
```

Simple Linear Regression Model

We can use the `lm()` function in R to fit a simple linear regression model. Here we'll fit a model with systolic blood pressure (`BPSysAve`) as the response variable, and diastolic blood pressure (`BPDiaAve`) as the explanatory variable.[1]

```r
lm1 <- lm(BPSysAve ~ BPDiaAve, data = nhanes)
```

The function uses the formula notation `y ~ x`, where `y` is the response variable, and `x` is the explanatory variable.

Use the summary() function to print out important information about the linear regression model we just fit.

```r
summary(lm1)
```

```
##
## Call:
## lm(formula = BPSysAve ~ BPDiaAve, data = nhanes)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -33.666 -10.047  -2.328   7.451  99.100
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 89.65230    2.32651   38.53   <2e-16 ***
## BPDiaAve     0.44137    0.03267   13.51   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.28 on 1498 degrees of freedom
## Multiple R-squared:  0.1086, Adjusted R-squared:  0.108
## F-statistic: 182.5 on 1 and 1498 DF,  p-value: < 2.2e-16
```

The least squares estimates of the slope and intercept are given in the `Coefficients` table of the summary output. The equation of the least squares regression line can therefore be written as

$$\hat{y} = 89.6523 + 0.44137x$$

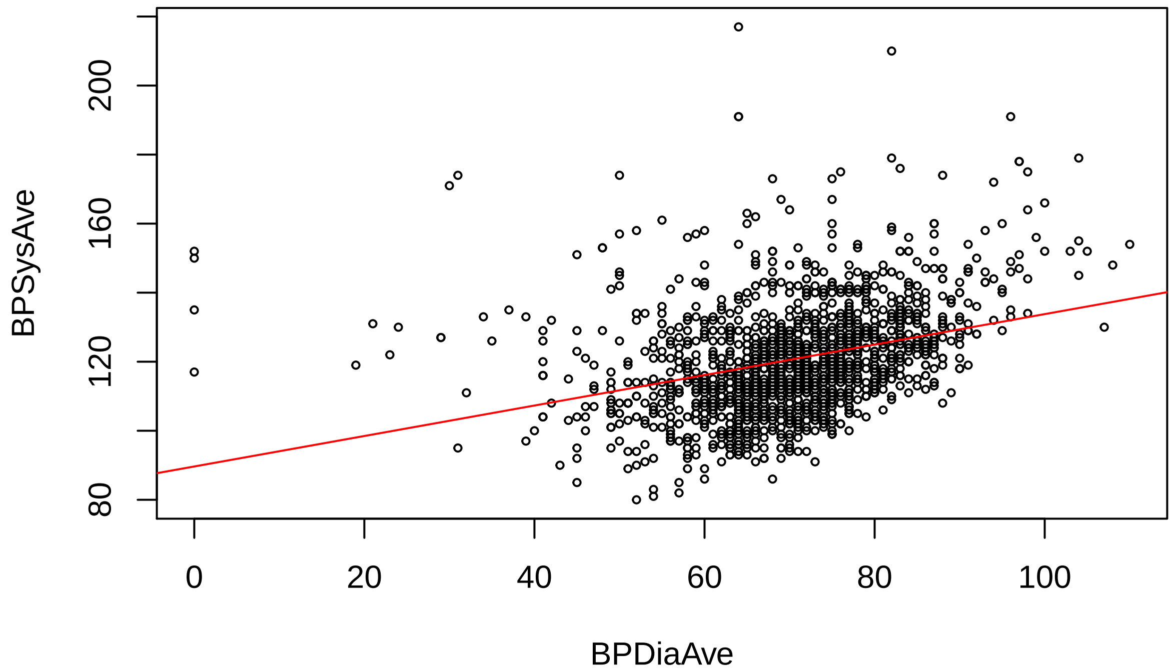
The summary output also gives an $R^2 = 0.1086$. This means that about 11% of the variability in systolic blood pressure (y) can be explained by diastolic blood pressure (x).

[1] Some background info about blood pressure: https://www.cdc.gov/bloodpressure/about.htm

Plot Least Squares Line

Next we make a scatter plot of the data, and add the least squares line:

```
plot(BPSysAve ~ BPDiaAve, data = nhanes, cex = 0.5)
abline(lm1, col = "red") # add least squares line
```



The scatterplot shows a positive linear association between diastolic and systolic blood pressure. However, there are some outliers – four individuals with a diastolic blood pressure reading of zero.

Note that `cex` controls the size of the points (magnification relative to 1); since there are 1500 points, I reduced the point size.

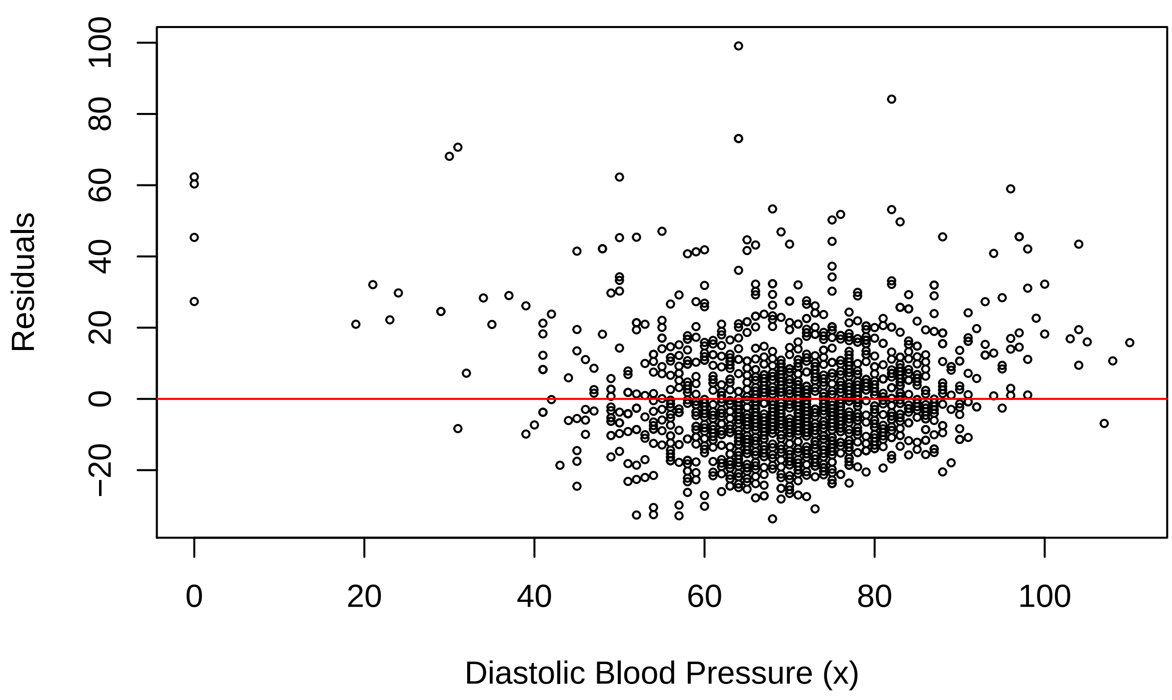
Check Conditions

Recall the conditions for SLR:

- Linearity: The data should follow a linear trend.
- Constant Variability: The variability of the points around the least squares line remains roughly constant.
- Normality: The residuals should have an approximate normal distribution with mean 0.
- Independence: Values of the response variable are independent of each other.

Based on the scatterplot the linearity condition seems satisfied. One useful plot for checking the constant variability condition is a plot of the residuals ($\hat{e}_i = y_i - \hat{y}_i$) versus the values of the explanatory variable (x_i):

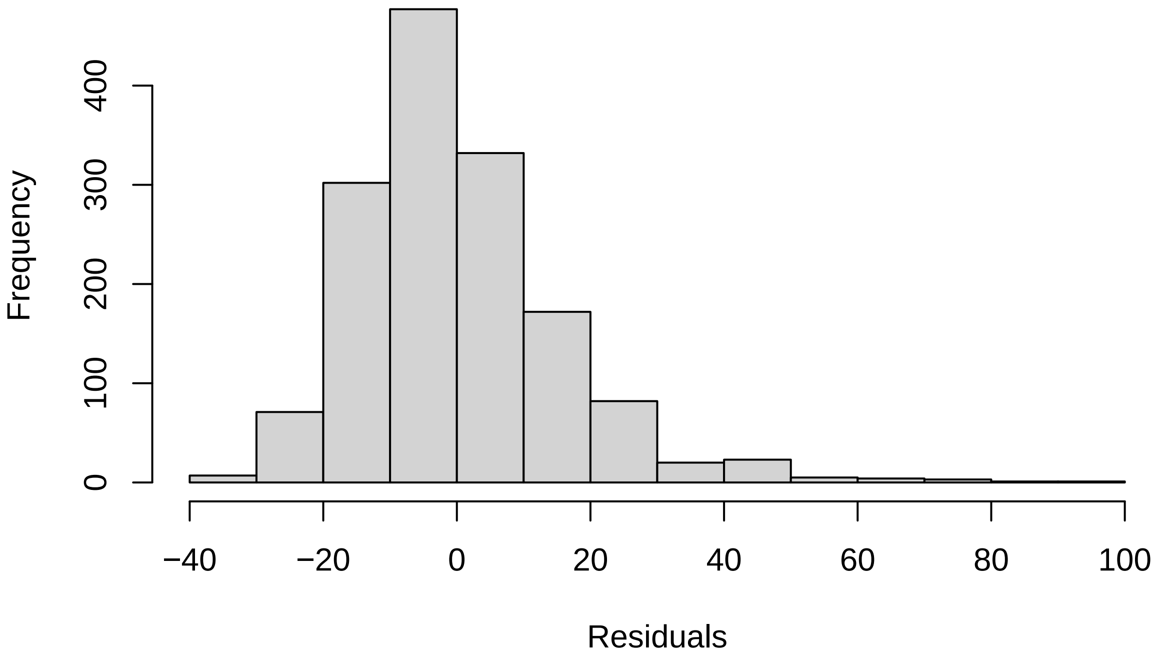
```
# residual plot
plot(nhanes$BPDiaAve, resid(lm1), cex = 0.5,
     xlab = "Diastolic Blood Pressure (x)", ylab = "Residuals")
abline(h=0, col = "red") # horizontal line at 0
```



The points look randomly scattered in the residual plot, so the constant variability condition is satisfied.

Next, to the check the normality condition, make a histogram of the residuals:

```r
hist(resid(lm1), xlab = "Residuals", main = "")
```



The distribution has a bell-curve shape. Although, the histogram looks a little right skewed, which indicates that there are some outliers.

Last, note that the data come from a random sample, so the independence condition is satisfied.

Overall, the conditions for SLR appear mostly satisfied with this data set. The main concern is that there are 4 points that are outliers (individuals with a diastolic blood pressure reading of zero). We should probably remove these outliers are refit the model.