4. Simple Linear Regression

4.1 Introduction

Simple linear regression is the most commonly used technique for determining how one variable of interest (the response or dependent variable) is affected by the changes in another variable (the explanatory, predictor or independent variable).

Simple linear regression is used for two main purposes:

1. To describe the linear relationship between the response and the predictor
2. To predict values of response variable from given values of predictor

The scatter plot and correlation plays a main role in simple linear regression.

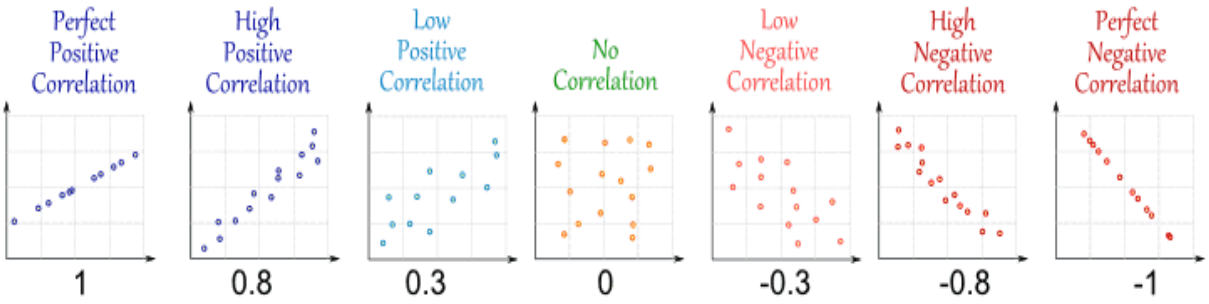
4.2 Scatter Plot and Correlation

Correlation coefficient measures the strength of linear relationship between response variable (y) and the predictor variable (x).
This can be expressed in graphical form using a scatter diagram.
Higher the correlation between the response and the predictor, better the regression fit would be.

Properties of correlation coefficient:

- The correlation values can range from -1 to +1
- If the correlation coefficient is +1, all the data fall on a line with positive slope
- If the correlation coefficient is -1, all the data fall on a line with negative slope
- If the correlation coefficient is 0, no linear relationship exists

Following are some scatter plots generated by using different correlation coefficients.



4.3 Simple Linear Regression Model

The equation that describes how y is related to x and an error term is called the regression model.

The simple linear regression model is:

$$y = \beta_0 + \beta_1 x + \varepsilon$$

Where,

- β_0 and β_1 are called parameters of the model
- ε is a random variable called the error term

Here β_0 is the intercept parameter and β_1 is the slope parameter.

For example, in the data set ***"faithful"*** in R, it contains sample data of two random variables named *waiting* and *eruptions*. The waiting variable denotes the waiting time (in minutes) between eruptions and eruptions denote the duration of the eruption (in minutes) for the Old Faithful geyser in Yellowstone National Park, Wyoming, USA.

Its linear regression model can be expressed as:

$$Eruptions = \beta_0 + \beta_1 * Waiting + \varepsilon$$

Estimated Regression Equation

If we choose the parameters β_0 and β_1 in the simple linear regression model so as to minimize the sum of squares of the error term ε, we will have the so called estimated simple regression equation. It allows us to compute fitted values of y based on the values of x.

We write it as,

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

Example:

Develop an estimated regression line to predict the duration of the eruption based on the waiting time.

```
> attach(faithful)
> fit<-lm(eruptions~waiting)
> fit
Call:
lm(formula = eruptions ~ waiting)
Coefficients:
(Intercept)      waiting
   -1.87402      0.07563
```

The estimated regression equation is,

$$\widehat{Eruptions} = -1.874 + 0.076 * Waiting$$

The slope parameter can be interpreted as, for each additional minute in waiting time, the next eruption occurs for 0.076 minutes longer.

Goodness of Fit Testing

The coefficient of determination or r^2 of a linear regression model is a measure of the variances of the fitted values and observed values of the dependent variable.

This measure can be used to comment how well the fitted equation describes the data.

- r^2 is the square of the correlation between x and y
- r^2 converted to a percentage is called the ***Coefficient of Determination***
- r^2 measures the percentage of variability in y explained by its association with x

If $r = 0.71$, then $r^2 \approx 0.50$, then 50% of the variability in y is explained by the regression fit.

***Example*:** Consider the same faithful data example.

```
> summary(fit)
Call:
lm(formula = eruptions ~ waiting)
Residuals:
     Min       1Q   Median       3Q      Max
-1.29917 -0.37689  0.03508  0.34909  1.19329
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.874016   0.160143  -11.70   <2e-16 ***
waiting      0.075628   0.002219   34.09   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’
0.1 ‘ ’ 1
```

```
Residual standard error: 0.4965 on 270 degrees of
freedom
Multiple R-squared: 0.8115,Adjusted R-squared: 0.8108
F-statistic: 1162 on 1 and 270 DF, p-value: < 2.2e-16
```

Here, r^2 is 0.8115, so 81% of the variability in eruptions is explained by the regression fit. It can be considered as a good fit.

Significance Testing

Assume that the error term ε in the linear regression model is independent of x, and is normally distributed, with zero mean and constant variance. These are called the assumptions of the error term. When these assumptions are met we can decide whether there is any significant relationship between x and y by testing the null hypothesis that $\beta_1 = 0$.

So the hypothesis of interest is,

$$H_0: \beta_1 = 0$$
$$H_a: \beta_1 \neq 0$$

This hypothesis can be tested using either as a t-test or as an F-test. Depending on which test is being used, the p-value corresponding to the calculated test statistic has to be taken using the t-distribution or F distribution.

Example:

All the information required for carrying out either t-test or F-test is available in the previous output. In the output,

- Calculated t – test statistic is given as 34.09 with the corresponding p-value < 2e-16

- Calculated F – test statistic is given as 1162 with corresponding p-value < 2.2e-16

In both tests, the p-value < 0.05, and hence the null hypothesis should be rejected at 5% level of significance. So we can conclude that, the slope parameter does not equal zero and hence, there is a significant relationship between the response and the predictor variable.

Residual Analysis

There are two objectives of carrying out residual analysis in regression.

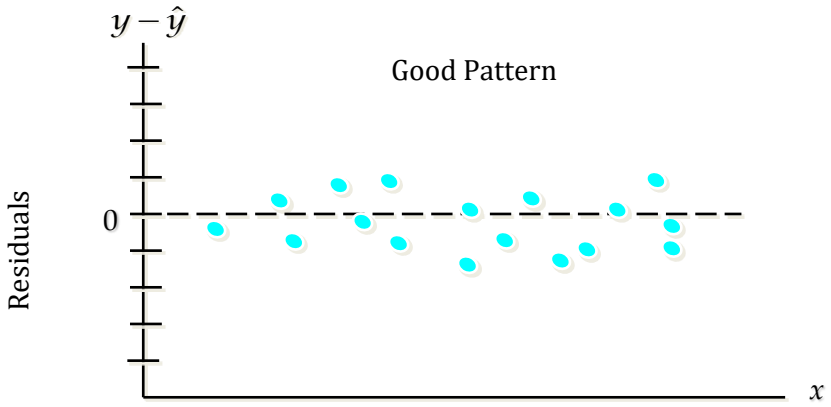
- The first and the main is to check the assumptions about the error term and
- The other is to identify unusual observations

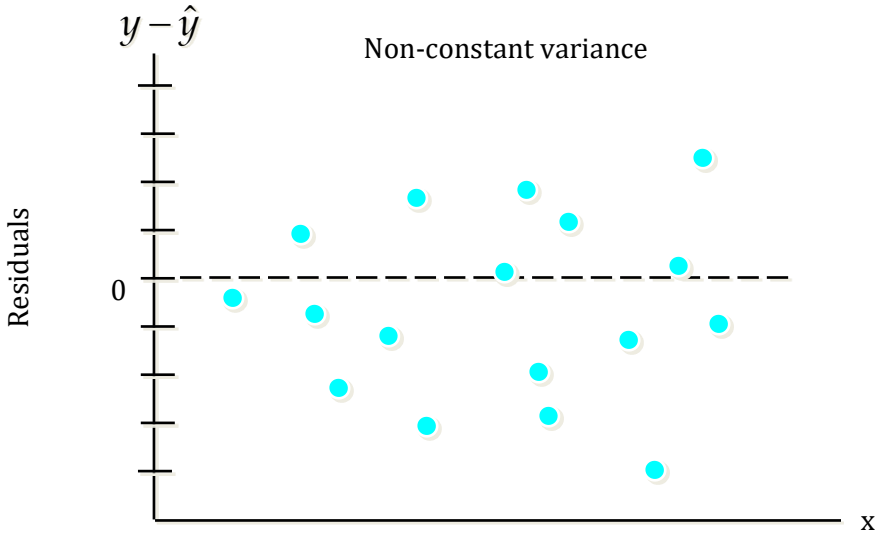
Checking the assumptions about error term:

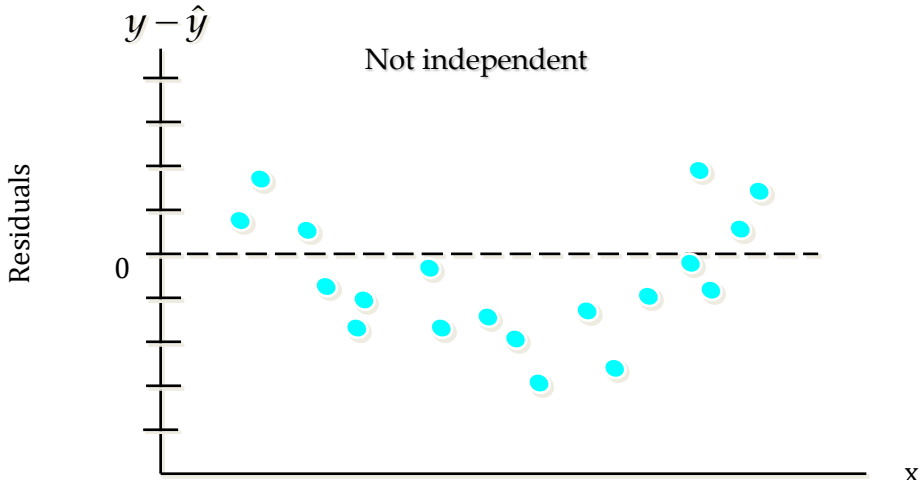
If the assumptions about the error term ε appear questionable, the hypothesis tests about the significance of the regression relationship may not be valid. So that the assumptions need to be checked before going for predictions. We do so by using two residual plots.

- Plot of Standardized residuals vs Predictor variable

If the assumption that the variance of ε is the same for all values of x is valid, and the assumed regression model is an adequate representation of the relationship between the variables, then this plot should give an overall pattern of a horizontal band of points.


$y - \hat{y}$
Good Pattern
Residuals
0
x

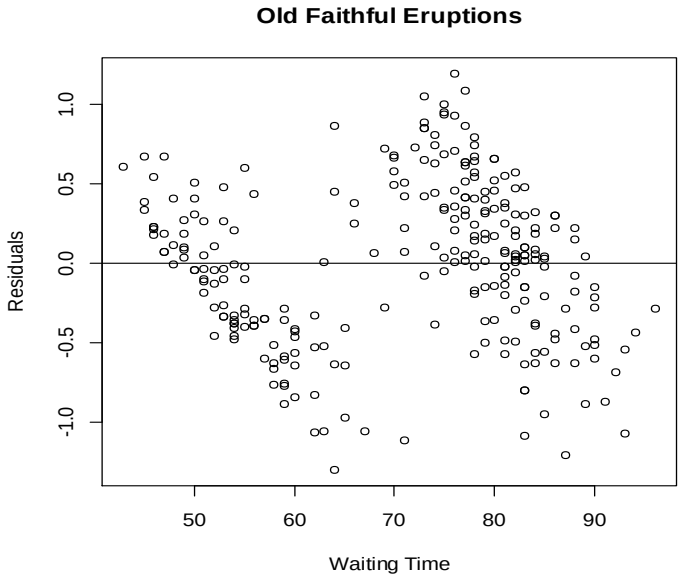

$y - \hat{y}$
Non-constant variance
Residuals
0
x


$y - \hat{y}$
Not independent
Residuals
0
x

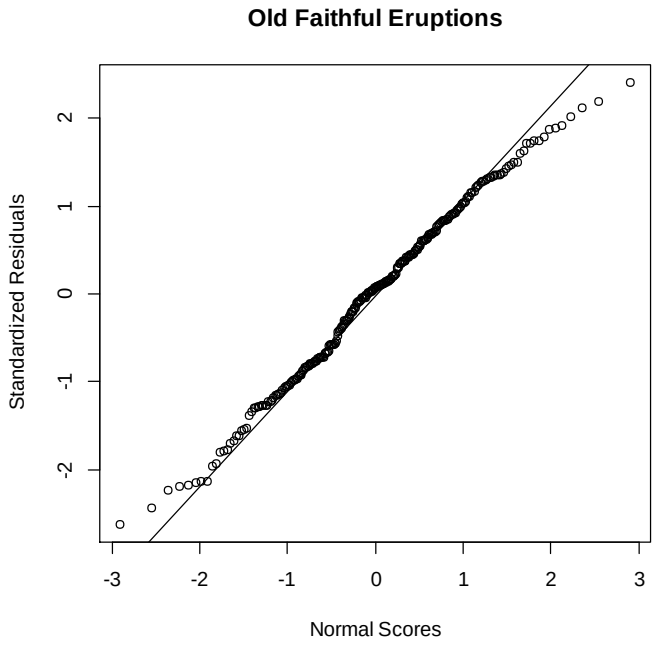
- Normal Probability Plot

With this display, each residual is plotted against its expected value when the distribution is Normal. A plot that is nearly linear suggests agreement with normality whereas a plot that departs substantially from linearity suggests that the error distribution is not Normal.

```
> fit.res<-resid(fit)
> plot(waiting, fit.res,ylab="Residuals",
xlab="Waiting Time",
+      main="Old Faithful Eruptions")
> abline(0, 0)                  # the horizon
```




```
> fit.stdres <- rstandard(fit)
> qqnorm(fit.stdres,ylab="Standardized
Residuals",xlab="Normal Scores",
+      main="Old Faithful Eruptions")
> qqline(fit.stdres)
```



Identifying unusual observations:

There are two types of unusual observations, ***'Outliers'*** and ***'Influential Observations'.***

Outliers

A regression outlier is a point that does not fit with the majority of the data. To examine outliers, we could use the standardized residuals. If the model is adequate and if the standardized residuals of a particular observation is beyond +2 or -2, then we suspect that the observation to be a regression outlier.

```
> fit.stdres[fit.stdres>2]
      51       70       76      151
2.021974 2.125149 2.408404 2.187389
> fit.stdres[fit.stdres < -2]
     17        46        58        95       158
197       211       249
-2.150451 -2.194719 -2.622667 -2.135205 -2.182591 -
2.438986 -2.244895 -2.139264
```

Influential observations

A point that pulls the regression line towards it, is known as an influential point. Cook's distance values can be used to identify such observations. Various cut-off values are given for the Cook's statistic. We use the cut-off value 1. For influential points the Cook's statistic would exceed this value.

```
> cook<-cooks.distance(fit)
> cook[cook>1]
named numeric(0)
```

Outlying and influential data should not be ignored, and they also should not be deleted without investigation. Bad data can often be corrected. Good observations that are unusual may provide insight into the structure of the data, and may motivate to consider a change in the structure of the model.

Prediction

Regression models predict a value of the Y variable, given known values of the X variables.

- When predictions are done within the range of values in the dataset used for model-fitting then it is known as ***interpolation***
- Predicting outside the range of the data is known as ***extrapolation***

Usually the interpolation is recommended using regression models.

Example:

Predict the duration of the eruption if the waiting time is 80 minutes.

```
> predict.80<-
fit$coefficients[1]+fit$coefficients[2]*80
> predict.80
(Intercept)
    4.17622
```

Alternative Way

```
> newdata<- data.frame(waiting=80)
> predict(fit, newdata)
      1
4.17622
```

If the waiting time is 80 minutes, then the duration of the next eruption will be 4.17 minutes.