Data Mining, Machine Learning, and Statistical Analysis

Lecture 6

Department of Mathematics and Statistics

PURDUE UNIVERSITY NORTHWEST

Correlation

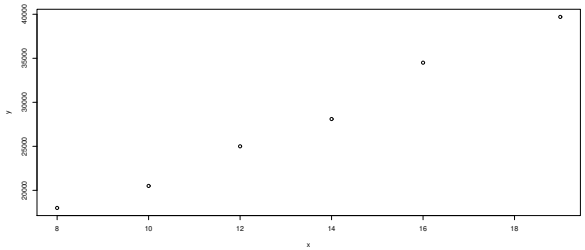
The tools used to explore the relationship between two continuous variable, such as height and weight, the concentration of an injected drug and heart rate, or the consumption level of some nutrient and weight gain etc., is the regression and correlation analysis. These tools can be used to find out if the outcome from one variable depends on the value of the other variable, which would mean a dependency from one variable on the other. Regression and correlation analysis can be used to describe the nature and strength of the relationship between two continuous variables. The first step in the investigation of the relationship between two continuous variables is a scatterplot. We create a scatterplot for the two variables and evaluate the quality of the relationship.

Example

Does the number of years invested in schooling pay of in the job market? Apparently so - the better educated you are, the more money you will earn. The data in the following table give the median annual income of full-time workers age 25 or older by the number of years of schooling completed. Let X: Years of schooling and Y: Salary(in Dollars)

X	8	10	12	14	16	19
Y	18,000	20,500	25,000	28,100	34,500	39,700

The scatteplot of the data is as below



Correlation Coefficient

The Pearson Product-Moment Correlation Coefficient (ρ), or correlation coefficient for short is a measure of the degree of linear relationship between two variables, usually labeled X and Y.
The correlation coefficient $\rho = \rho_{XY}$ is defined by

$$\rho = \frac{COV(X, Y)}{\sqrt{Var(X)Var(Y)}}.$$

where, $COV(X, Y) = E(X - \mu_X)(Y - \mu_Y) = E(XY) - \mu_X\mu_Y$ is called the covariance between X and Y.
Note that, $-1 \leq \rho \leq 1$.
In R use `cor(x,y)` to calculate the correlation coefficient and `cov(x,y)` to calculate the covariance between x and y.

Testing for Correlation Coefficient

- Pearson Correlation Coefficient
  cor.test(x,y)
- Spearman Correlation Coefficient
  `cor.test(x,y,method="spearman")`
- Kendall (tau-b) Correlation Coefficient.
  `cor(x,y,method="kendall")`

Pearson R Correlation Coefficient

Assumptions

- They must be approximately Gaussian distributed.
- There must be a significant linear relationship between them.
- They must be either interval or ratio measurements.
- There may not be any outliers.
- They must have similar variances.

The correlation coefficient $\rho = \rho_{XY}$ is defined by

$$\rho = \frac{COV(X, Y)}{\sqrt{Var(X)Var(Y)}}.$$

where, $COV(X, Y) = E(X - \mu_X)(Y - \mu_Y) = E(XY) - \mu_X\mu_Y$ is called the covariance between X and Y

Spearman Rank Correlation Coefficient

Assumptions

- They must be rank ordered.
- They are monotonically related.
- They need not be Gaussian distributed.
- They do not require the parameters of distribution.
- They do not require that the relationship between them being linear.
- They do not require to be measured on interval or ratio scale.

The Spearman rank correlation coefficient is Pearson's moment correlation formula applied to ordinals, X_i and $Y_i, i = 1, 2, ..., N$, with no ties in either X_i or Y_i.
We have

$$\rho = 1 - \frac{6 \sum_i d_i^2}{N(N^2 - 1)}$$

where, $d_i = R(X_i) - R(Y_i)$ is the difference in the ranks.

Kendall's Tau Coefficient

It is a non-parametric correlation coefficient like the Spearman correlation that can be used to find correlation between the variables X and Y. While Spearman rank correlation coefficient is Pearson correlation coefficient computed from ranked variables, the Kendall correlation rather represents the difference between the probabilities of the dependent variable Y increasing and decreasing with respect to X, and may not be necessary to rank order them. However, they are usually rank ordered to facilitate computation. If (U_i, V_i) and $(U_j, V_j), i, j \in N$ are two pairs of observations not necessarily rank ordered unlike the Spearman correlation coefficient, with no ties among them then if the pairs $(Ui - Uj)$ and $(Vi - Vj)$ are of the same sign for each i and j then these pairs are called `concordant` pairs c. If they have opposite signs then they are called `discordant` pairs d. The measure of correlation proposed by Kendall in case of no ties is

$$\tau = \frac{(N_c - N_d)}{N(N-1)/2}$$

where N_c and N_d are the number of concordant pairs and number of discordant pairs respectively.

Example

Twelve MBA graduates are studied to measure the strength of the relationship between their score on the GMAT which they took prior to entering graduate school, and their grade point average while they were in MBA program. Their GMAT scores(x) and GPA scores(y) are given below

GMAT	710	610	640	580	545	560	610	530	560	540	570	560
GPA	4.0	4.0	3.9	3.8	3.7	3.6	3.5	3.5	3.5	3.3	3.2	3.2

```
x<-c(710,610,640,580,545,560,610,530,560,540,570,560)
y<-c(4.0, 4.0, 3.9, 3.8, 3.7, 3.6, 3.5, 3.5, 3.5, 3.3, 3.2, 3.2)
cor(x,y)
cor.test(x,y)
cor.test(x, y, method="s")
cor.test(x,y, method="k")
```

The function `rcorr()` [in Hmisc package] can be used to compute the significance levels for pearson and spearman correlations. It returns both the correlation coefficients and the p-value of the correlation for all possible pairs of columns in the data table.

```
> library(Hmisc)
> data=mtcars[,c(1,2,3,4,5,6,7)]
> data=as.matrix(data)
> rcorr(data, type="pearson")
> pairs(data)

> library(corrplot)
> corrplot(cor(data))
> corrplot(cor(data), method="number")
> corrplot(cor(data), method="number", type="upper")
```

Regression Analysis

Regression analysis is a statistical technique for investigating and modeling the relationship between variables. Applications of regression are numerous and occur in almost every field, including engineering, the physical and chemical sciences, economics, management, life and biological sciences and the social science. It is the most widely used statistical technique.

We use regression analysis for explaining or modeling the relationship between a single variable y, called the response, output or dependent variable, and one or more predictor, input, independent or explanatory variables, $x_1, x_2, ..., x_p$. When $p = 1$, it is called simple regression but when $p > 1$ it is called multiple regression or sometimes multivariate regression. When there is more than one y, then it is called multivariate multiple regression. The response must be a continuous variable but the explanatory variables can be continuous, discrete or categorical.

Regression analysis have several possible objectives including

- A general description of data structure
- Variable screening
- Prediction of future observations
- Assessment of the effect of, or relationship between, explanatory variables on the response

Simple Linear Regression Model

Consider a random sample of n observations of the form $(x_1, y_1), (x_2, y_2), ...(x_n, y_n)$ where x is independent variable and y is the dependent variable, both being scalars. The model that is applicable in the simplest regression structure is the simple linear regression model. Here the term simple implies a single regressor variable x and the term linear implies linear in the coefficients $\beta's$. The model is given by

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

where β_0 and β_1 are the intercept and slope respectively, and ϵ is the model error.

Note that the variable x is often called the predictor or regressor variable and y as a response variable.

We assume that $\epsilon_i \sim N(0, \sigma^2)$ and we also assume that ϵ_i are uncorrelated from observation to observation. In addition, any error in the measurement of the x_i is assumed to be small compared to the range.

In a simple linear regression model we have

$$E(y_i) = \beta_0 + \beta_1 x_i$$

and the variance is σ^2.

Model Assumption

The standard analysis is based on the following assumptions about the regressor variable x and the random errors $\epsilon_i, i = 1, 2, ..., n$:

- The regressor variable is under the experimenter's control, who can set the values of $x_1, x_2, ..., x_n$. This means that x_i can be taken as constants, they are not random variables.
- $E(\epsilon_i) = 0, i = 1, 2, 3, ..., n$. This implies that $\mu_i = E(y_i) = \beta_0 + \beta_1 x_i, i = 1, 2, 3, \cdots, n$
- $Var(\epsilon_i) = \sigma^2$ is constant for all $i = 1, 2, \cdots, n$. This implies $Var(y_i) = \sigma^2$.
- Different errors ϵ_i and ϵ_j and hence the different responses y_i and y_j, are independent. This implies that $Cov(\epsilon_i, \epsilon_j) = 0$ for $i \neq j$.

In summary, the regression model $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ implies that the response y_i come from probability distributions whose means are $E(y_i) = \beta_0 + \beta_i x_i$ and whose variances are σ^2, the same for all levels of x. Furthermore, any two responses y_i and y_j are not correlated.

Objectives of the Analysis

Given a set of observations, we would like to answer the following questions:

- Can we establish a relationship between x and y?
- Can we predict y from x? To what extent can we predict y from x?
- Can we control y by using x?

In order to answer these questions within the simple regression framework we need to estimate the values of β_0, β_1 & σ^2 from the given data set. In particular, we are interested in β_1 as $\beta_1 = 0$ indicates the absence of linear association.

Parameter Estimation

Consider a simple linear regression model

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \qquad i = 1, 2, 3, \cdots, n$$

The method of least squares is used more extensively than any other estimation procedure for estimating the regression parameters β_0 and β_1.
The method is designed to provide estimators b_0 and b_1 of β_0 and β_1, that the residual sum of squares (RSS) or the error sum of squares (SSE)is minimized.
Note that

$$SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \widehat{y_i})^2 = \sum_{i=1}^{n} [y_i - b_0 - b_1 x_i]^2$$

Hence b_0 and b_1 must satisfies the

$$\frac{\partial}{\partial b_0} \left[\sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2 \right] = 0$$

$$\frac{\partial}{\partial b_1} \left[\sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2 \right] = 0$$

Parameter Estimation

$$\sum_{i=1}^{n} y_i = nb_0 + b_1 \sum_{i=1}^{n} x_i$$

$$\sum_{i=1}^{n} x_i y_i = b_0 \sum_{i=1}^{n} x_i + b_1 \sum_{i=1}^{n} x_i^2$$

Solving these normal equations simultaneously for b_0 and b_1

$$b_1 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})y_i}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

$$b_0 = \overline{y} - b_1\overline{x}$$

Our estimators for intercept and slope are

$$b_0 = \overline{y} - b_1\overline{x}$$

$$b_1 = \frac{S_{xy}}{S_{xx}}$$

where $S_{xy} = \sum_{i=1}^{n}(x_i - \overline{x})y_i$ and $S_{xx} = \sum_{i=1}^{n}(x_i - \overline{x})^2$.

Example

Table below gives the measurements of systolic blood pressure(SBP) and age for a sample of 15 individuals older than age 40 years.

SBP(y)	164	220	133	146	162	144	166	152	140	145	135	150	170	122	120
Age(x)	65	63	47	54	60	44	59	64	51	49	57	56	63	41	43

For this data we have the summary statistic

$$\begin{aligned} n &= 15, \quad \sum_{i=1}^{n} x_i = 816, \\ \sum_{i=1}^{n} y_i &= 2269, \quad \sum_{i=1}^{n} x_i^2 = 45318, \\ \sum_{i=1}^{n} y_i^2 &= 351475, \quad \sum_{i=1}^{n} x_i y_i = 125445 \end{aligned}$$

Now,

$$S_{xy} = \sum_{i=1}^{n}(x_i - \overline{x})y_i = \sum_{i=1}^{n} x_i y_i - \overline{x}\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} x_i y_i - \frac{1}{n}\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i = 2011.4$$

and

$$S_{xx} = \sum_{i=1}^{n}(x_i - \overline{x})^2 = \sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2 = 927.6$$

Hence,

$$b_1 = \frac{S_{xy}}{S_{xx}} = \frac{2011.4}{927.6} = 2.1684$$

and

$$b_0 = \overline{y} - b_1\overline{x} = 151.267 - 2.1684 \times 54.4 = 33.306$$

Therefore, resulting least square line is

$$\hat{y} = 33.306 + 2.1684x$$

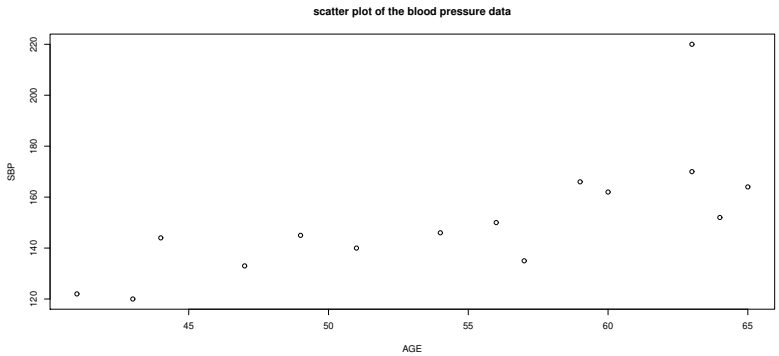
Example

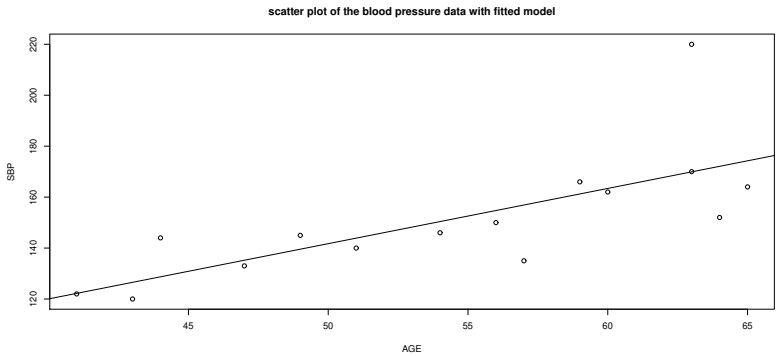
Using R to obtain the regression model

```
> x<-c( 65, 63, 47, 54, 60, 44, 59, 64, 51, 49, 57, 56, 63, 41, 43)
> y<-c(164,220,133,146,162,144,166,152,140,145,135,150,170,122,120)
> model=lm(y~x)
> model
Call:
lm(formula = y ~ x)
Coefficients:
(Intercept)            x
     33.306        2.168
```

Example

```
> x<-c( 65, 63, 47, 54, 60, 44, 59, 64, 51, 49, 57, 56, 63, 41, 43)
> y<-c(164,220,133,146,162,144,166,152,140,145,135,150,170,122,120)
> model=lm(y~x)
> plot(x,y,xlab="AGE",ylab="SBP",main="scatter plot of the blood pressure data"
> plot(x,y,xlab="AGE",ylab="SBP",main="scatter plot of the blood pressure data
> abline(model)
```



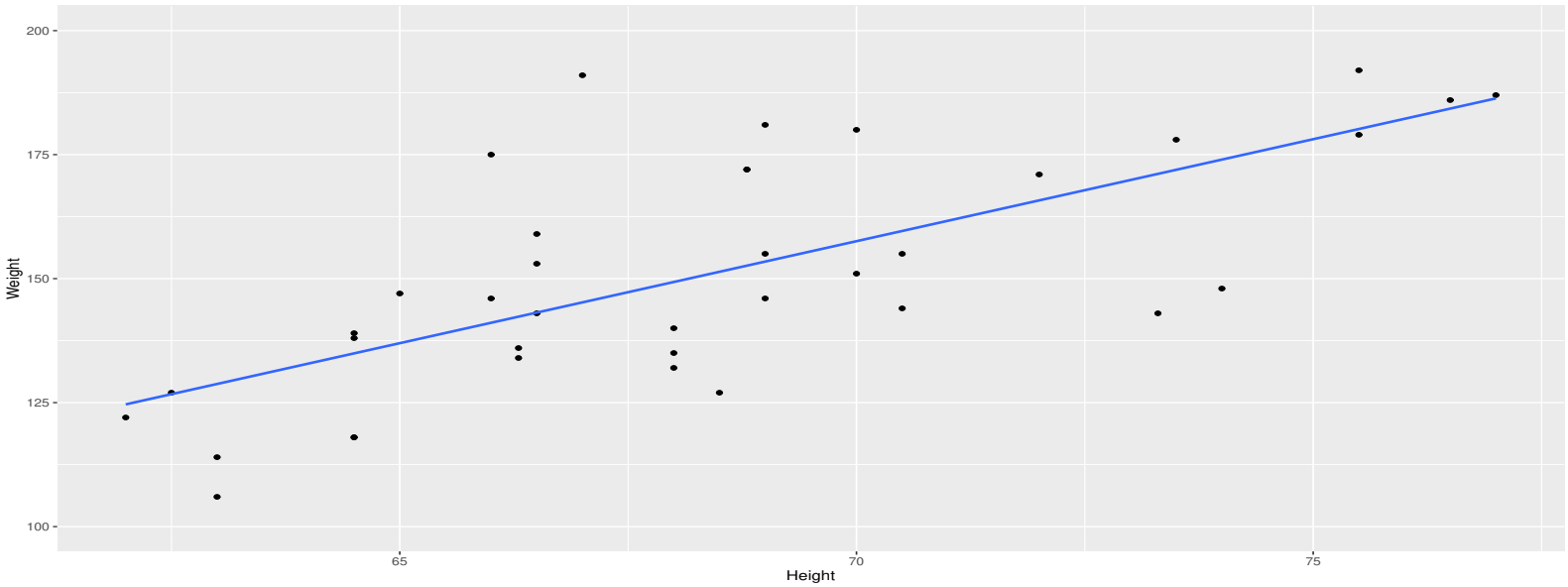


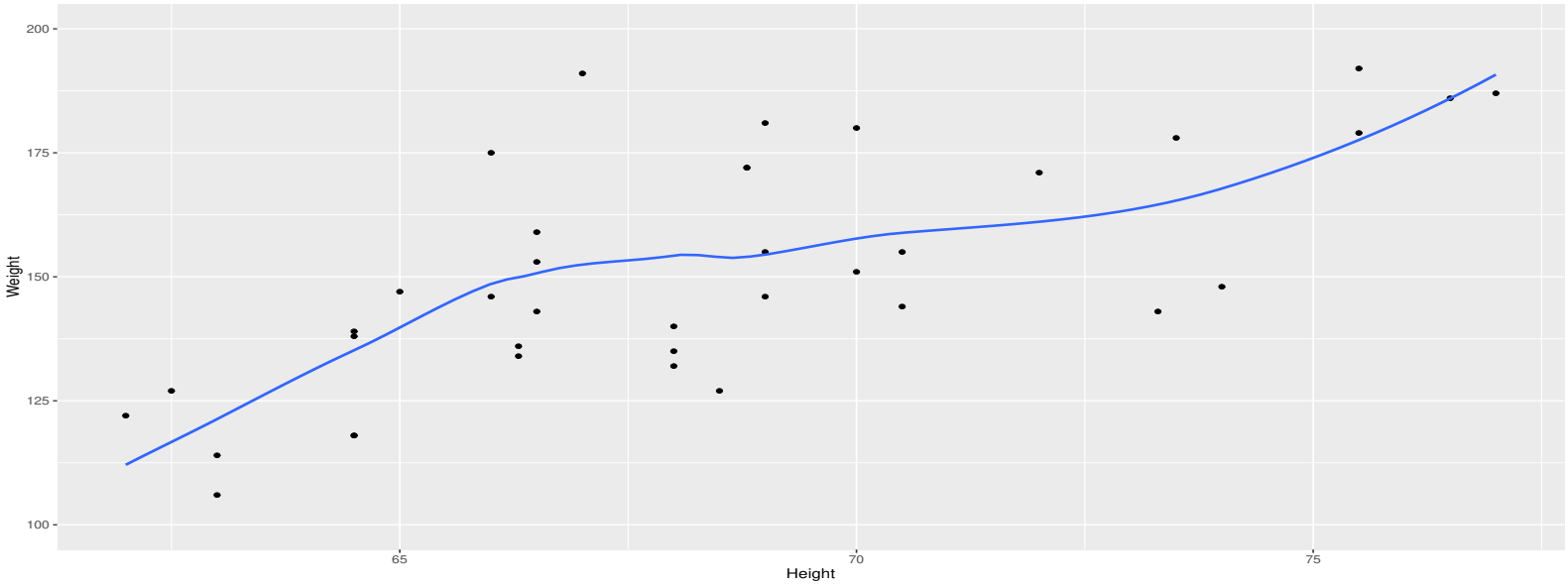
Regression with ggplot

Consider the data related to the height and weight of 38 people. The data are provided in the course website.

```
 > size=read.csv("C:\\aryal\\CS 59000\\Data Sets\\Size.csv")
> attach(size)
> head(size,2)
  Height Weight
1   64.5    118
2   73.3    143
> library(ggplot2)
ggplot(size,aes(x=Height,y=Weight))+geom_point()+ylim(100,200)
ggplot(size,aes(x=Height,y=Weight))+geom_point()+stat_smooth(method="lm")
ggplot(size,aes(x=Height,y=Weight))+geom_point()+ylim(100,200)+stat_smooth()
```

Regression models





Extractor Functions for the results of *lm*()

`summary()`	Returns summary information about the regression
`plot()`	makes diagnostic plots
`coef()`	returns the coefficients
`residuals()`	returns the residuals (can be abbreviated to resid())
`fitted()`	returns the fitted values
`confint()`	returns the confidence interval for the parameters
`deviance()`	returns RSS
`predict()`	performance predictions
`anova()`	finds various sums of squares
`AIC()`	is used for model selection

Properties of fitted regression line

The fitted regression line $\widehat{y} = b_0 + b_1 x$ has the following properties:

- The sum of the residuals is zero, i.e., $\sum_{i=1}^{n} e_i = 0$.
- The sum of the squared residuals, $\sum_{i=1}^{n} {e_i}^2$ is minimum.
- The sum of the observed values y_i equals the sum of the fitted values $\widehat{y}_i$,
  i.e., $\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \widehat{y}_i$
- The regression line always pases through the point $(\overline{x}, \overline{y})$.

Inference in Regression Analysis

Let $y = \beta_0 + \beta_1 x + \epsilon$ be a simple linear regression model with $\epsilon \sim N(0, \sigma^2)$ and ϵ_i are independent then

- b_0 and b_1 have normal distributions.
- b_0 and b_1 are unbiased estimators of β_0 and β_1 respectively
- $Var(b_0) = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right)$
- $Var(b_1) = \frac{\sigma^2}{S_{xx}}$

where σ^2 is the variance of ϵ_i
Note that $s^2 = \frac{1}{n-2} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$ is called the mean square error(MSE) or residual mean square. Therefore,

$$MSE = \frac{SSE}{n-2}$$

It can be shown for a simple linear regression model that MSE is an unbiased estimator of σ^2.

Conducting a Residual Analysis and Prediction

Conducting a Residual Analysis

The residuals are obtained using the `residuals` function in R. However these residuals don't have the same variance(heteroscedastic). We therefore use the studentized residuals, which have the same variance.

Predicting a new Value

Once a simple regression is developed we can use it to predict the corresponding y value for a given value of x. However the predicted value is of little interest without its corresponding confidence interval. We can use the R code `predict` to make a prediction.

Example-Heart Rate

15 people of varying ages are tested for their maximum heart rate.

Age	18	23	25	35	65	54	34	56	72	19	23	42	18	39	37
Max Rate	202	186	187	180	156	169	174	172	153	199	193	174	198	183	178

We can draw the confidence interval and prediction interval using the code below:

```
> x = c(18,23,25,35,65,54,34,56,72,19,23,42,18,39,37)
> y = c(202,186,187,180,156,169,174,172,153,199,193,174,198,183,178)
> plot(x,y)                          # make a plot
> abline(lm(y ~ x))                  # plot the regression line
> model= lm(y ~ x)
> confint(model)
>predict(model, data.frame(x=35), interval="pred")
```

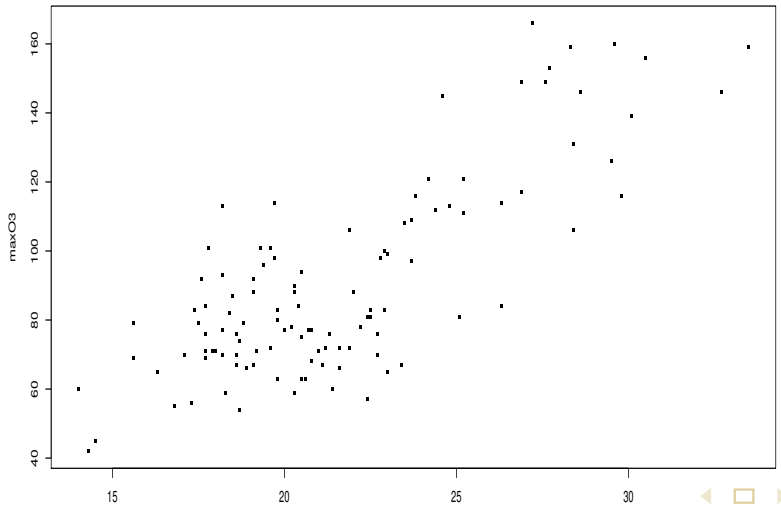
Interactive plot in ggplot

```
x = c(18,23,25,35,65,54,34,56,72,19,23,42,18,39,37)
y = c(202,186,187,180,156,169,174,172,153,199,193,174,198,183,178)
data=data.frame(x,y)
library(ggplot2)
library(ggiraph)
library(ggiraphExtra)
library(plyr)
fit=lm(y~x)
ggplot(data,aes(y=y,x=x))+geom_point()+geom_smooth(method="lm")
ggPredict(fit,se=TRUE,interactive=TRUE)
```

Example

Air Pollution is currently one of the most serious public health worries worldwide. Many epidemiological studies have proved that some chemical compounds such as sulphur dioxide (SO_2), nitrogen dioxide (NO_2), ozone (O_3) or other air-borne dust particles can have on our health. Link below contains 112 observations recorded during Summer 2001 in Rennes (France). Measurements for many variables including the ozone concentration(O3) and midday temperature (T12) are provided. We would like to study the relationship between the ozone level and the midday temperature.
http://math.agrocampus-ouest.fr/igagrocampus-ouest.fr/math/RforStat/ozone.txt

```
>ozone=read.table("http://math.agrocampus-ouest.fr/igagrocampus-ouest.fr/math/RforStat/ozone.txt", header=T)
>plot(maxO3~T12, data=ozone, pch=15, cex=0.5)
```



Example-Ozone

We will study the ozone data using R code below:

```
>ozone=read.table("http://math.agrocampus-ouest.fr/igagrocampus-ouest.fr/math/RforStat/ozone.txt", header=T)

>model=lm(maxO3~T12, data=ozone)
>summary(model)
>coef(model)
>res.simple<-residuals(model)
>plot(res.simple, pch=16, ylab="Residuals")
>abline(h=c(-2,0,2),lty=c(2,1,2))
>xnew<-20
>xnew=as.data.frame(xnew)
>colnames(xnew)<-"T12"
>predict(model, xnew, interval="pred")
```

Example-Heart Rate

15 people of varying ages are tested for their maximum heart rate.

Age	18	23	25	35	65	54	34	56	72	19	23	42	18	39	37
Max Rate	202	186	187	180	156	169	174	172	153	199	193	174	198	183	178

We can draw the confidence interval and prediction interval using the code below:

```
> library(UsingR)
> x = c(18,23,25,35,65,54,34,56,72,19,23,42,18,39,37)
> y = c(202,186,187,180,156,169,174,172,153,199,193,174,198,183,178)
> plot(x,y)                          # make a plot
> abline(lm(y ~ x))                  # plot the regression line
> lm(y ~ x)
> simple.lm(x,y,show.ci=TRUE,conf.level=0.90)
```

Example- Heart Rate

```
> x = c(18,23,25,35,65,54,34,56,72,19,23,42,18,39,37)
> y = c(202,186,187,180,156,169,174,172,153,199,193,174,198,183,178)
> model=lm(y~x)
> plot(x,y)                          # make a plot
> abline(model)                  # plot the regression line
> predict(model,data.frame(x=sort(x)), level=.9, interval="confidence")
> ci.lwr = predict(model,data.frame(x=sort(x)),
            + level=.9,interval="confidence")[,2]
> ci.upr = predict(model,data.frame(x=sort(x)), level=.9,
            + interval="confidence")[,3]
> points(sort(x), ci.lwr,type="l")#  or use lines()
> points(sort(x), ci.upr,type="l") # or use lines()
##### OR
>curve(predict(model,data.frame(x=x), interval="confidence")[,3],add=T)
>curve(predict(model,data.frame(x=x), interval="confidence")[,2],add=T)
```

The Analysis of Variance

Once we fit a model we want to check

- Does x, the regressor variable, truly influence y, the response?
- Is there an adequate fit of the data to the model?
- Will the model adequately predict the response?

In the first case, success can be quite often be achieved in answering the question through hypothesis testing on the slope β_1. We would like to test

$$\begin{aligned} H_0 &: \beta_1 = 0 \\ H_a &: \beta_1 \neq 0 \end{aligned}$$

Of course if H_0 is true, the implication is that the model reduces to $E(y) = \beta_0$, suggesting that the regressor variable doesn't influence the response(at least through the linear model). Rejection of H_0 in favor of H_a leads one to conclude that x significantly influence the response.

The Analysis of Variance

A simple F-test produced through the computation outlined in the ANOVA table can be used. Since we have

$$\frac{SS_{Reg}/1}{SS_{Res}/n-2} = \frac{MSR}{s^2} \sim \frac{\chi^2_1/1}{\chi^2_{(n-2)}/(n-2)}$$

under H_0, we have MS_{Reg}/s^2 follows the $F_{1,n-2}$ under H_0 and is thus a candidate for a test statistic for testing the hypothesis. Below is a standard ANOVA table

Source	*Sum of Squares*	*df*	*Mean Square*	*F*
Regression	$SS_{Reg} = \sum(\widehat{y}_i - \overline{y})^2$	1	$SS_{Reg}/1$	$F = \frac{MS_{Reg}}{MSE}$
Residual	$SS_{Res} = \sum(y_i - \widehat{y}_i)^2$	$n-2$	$MSE = s^2$	
Total	$SS_{Total} = \sum(y_i - \overline{y})^2$	$n-1$		

Remark: For a given α level, the F test of $H_0 : \beta_1 = 0$ Vs. $H_1 : \beta_1 \neq 0$ is equivalent to the two-tailed t-test.

Example

Table below provides data on the boiling point of water (in ^oF) and barometric pressure (inches of mercury)

Boiling Point	Barometric Pressure	Boiling Point	Barometric Pressure
199.5	20.79	201.9	24.02
199.3	20.79	201.3	24.01
197.9	22.40	203.6	25.14
198.4	22.67	204.6	26.57
199.4	23.15	209.5	28.49
199.9	23.35	208.6	27.76
200.9	23.89	210.7	29.64
201.1	23.99	211.9	29.88
		212.2	30.06

Example

```
>T<-c(199.5,201.9,199.3,201.3,197.9,203.6,198.4,204.6,199.4,209.5,199.9,208.6,200.9,210.7,201.1,211.9,212.2)
>P<-c(20.79,24.02,20.79,24.01,22.40,25.14,22.67,26.57,23.15,28.49,23.35,27.76,23.89,29.64,23.99,29.88,30.06)
> model=lm(P~T)
> model

Call: lm(formula = P ~ T) Coefficients: (Intercept)            T
   -95.7572       0.5937

> anova(model)
Analysis of Variance Table Response: P
          Df Sum Sq Mean Sq F value    Pr(>F)
T          1 141.65 141.654  226.04 1.879e-10 *** Residuals 15
9.40   0.627
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

For the given data we have the following ANOVA table

Source	*SS*	*df*	*MS*	*F*	*Pr* $>$ *F*
Regression	141.65393	1	141.65393	226.04	$<$ 0.0001
Error	9.40008	15	0.62667		
Total	151.05401	16			

Decision: Since $p < \alpha$ we reject the null hypothesis that $\beta_1 = 0$, which means there is a strong relationship between the temperature and the barometric pressure

Quality of Fitted Model

To answer

- Is there an adequate fit of the data to the model?
- Will the model adequately predict the response?

We compute the coefficient of Determination.

The coefficient of determination, often is denoted by R^2 and is defined by

$$R^2 = \frac{SS_{Reg}}{SS_{Total}} = \frac{\sum(\widehat{y}_i - \overline{y})^2}{\sum(y_i - \overline{y})^2}$$

which also can be written as

$$R^2 = 1 - \frac{SS_{Res}}{SS_{Total}}$$

It is clear that

$$0 \leq R^2 \leq 1$$

We may interpret R^2 as the proportion of variation in the response data that is explained by the model. Thus, the larger R^2 is, the more the total variation of y is reduced by introducing the predictor variable x. When all the observation fall on the fitted regression line then $SS_{Res} = 0$ and $R^2 = 1$ whereas when the fitted regression line is horizontal so that

Adjusted R^2

Adjustment is made for complexity of the model (i.e. penalty for higher number of variables). The Formula for adjusted R^2 is

$$R^2_{Adj.} = 1 - \frac{MS_{Res}}{MS_{Total}}$$

It should be noted that R^2 is a measure of the linear association between y and x. A small R^2 does not always imply a poor relationship between y and x.

Example

```
> x1<-c(10,8,13,9,11,14, 6, 4,12, 7, 5)
> y1<-c(8.04,6.95,7.58,8.81,8.33,9.96,7.24,4.26,10.84, 4.82, 5.68)
> x2<-c(10,8, 13, 9,11,14, 6, 4,12, 7, 5)
> y2<-c(9.14, 8.14,8.74,8.77,9.26,8.10,6.13,3.10,9.13,7.26,4.74)
> x3<-c(10,8,13,9,11,14,6,4,12,7, 5)
> y3<-c(7.46, 6.77,12.74,7.11,7.81,8.84,6.08,5.39, 8.15,6.42,5.73)
> x4<-c(8,  8,  8 , 8 , 8 , 8 , 8 ,19 , 8,  8,  8)
> y4<-c(6.58, 5.76, 7.71, 8.84,8.47,7.04,5.25,12.50,5.56,7.91,6.89)

>par(mfrow=c(2,2))
>par(mfrow=c(2,2),oma=c(0,0,2,0))
>plot(x1,y1,main=" Scatter plot of Data Set 1")
>plot(x2,y2,main=" Scatter plot of Data Set 2" )
>plot(x3,y3,main=" Scatter plot of Data Set 3")
>plot(x4,y4, main="Scatter plot of Data Set 4")
```

A Look at Residuals

We would like to see what type of information can be gained from the ordinary residuals which is given by $e_i = y_i - \widehat{y_i}$ called as the errors of fit. The residuals may be regarded as the observed error, in distinction to the unknown true error ϵ_i in the regression model:
$\epsilon_i = y_i - E(y_i)$
We know that from the normal theory assumption ϵ_i are assumed to be independent normal random variables with mean 0 and variance σ^2. If the model is appropriate for the data at hand, the observed residuals e_i should reflect the properties assumed for ϵ_i. This is the basic idea of residual analysis, a highly useful means of examining the aptness of a statistical model.
Properties of Residuals:
a. We have

$$\overline{e} = \frac{\sum e_i}{n} = 0$$

so it provides no information as to whether the true errors ϵ_i have expected value $E(\epsilon_i) = 0$.
b. We have

$$Var(e_i) = \frac{\sum_i (e_i - \overline{e})^2}{} = \frac{\sum e_i^2}{} = \frac{SSE}{} = MSE$$

Diagnostics of Residuals

Graphical analysis of residuals is very effective to investigate the adequacy of the fit of the regression model and to check the underlying assumptions. We look at the following plots of residuals in order to check the model assumptions

- Plots of the residuals against predictor variable.
- Plot of absolute or squared residuals against predictor variables
- Plots of residuals against fitted values.
- Box plots of residuals
- Normal probability plots of residuals.

If the simple linear regression model is not appropriate it may occur due to

- Nonlinearity of Regression function
- Nonconstancy of Error Variance
- Nonindependence of Error terms
- Nonnormality of Error terms
- Omission of important predictor variable
- Outlying observations

Example

```
 > x1<-c(10,8,13,9,11,14, 6, 4,12, 7, 5)
>  y1<-c(8.04,6.95,7.58,8.81,8.33,9.96,7.24,4.26,10.84, 4.82, 5.68)
>  x2<-c(10,8, 13, 9,11,14, 6, 4,12, 7, 5)
>  y2<-c(9.14, 8.14,8.74,8.77,9.26,8.10,6.13,3.10,9.13,7.26,4.74)
>  x3<-c(10,8,13,9,11,14,6,4,12,7, 5)
>  y3<-c(7.46, 6.77,12.74,7.11,7.81,8.84,6.08,5.39, 8.15,6.42,5.73)
>  x4<-c(8,  8,  8 , 8 , 8 , 8 , 8 ,19 , 8,  8,  8)
>  y4<-c(6.58, 5.76, 7.71, 8.84,8.47,7.04,5.25,12.50,5.56,7.91,6.89)
>  model1= lm(y1~x1)
>  model2=lm(y2~x2)
>  model3=lm(y3~x3)
>  model4= lm(y4~x4)
>  res1=resid(model1) # It computes the residules of the first model
>  res2=resid(model2)
>  res3=resid(model3)
>  res4=resid(model4)
>  fit1=fitted(model1) # It computes the Fitted values of the first model
>  fit2=fitted(model2)
>  fit3=fitted(model3)
>  fit4=fitted(model4)
>  par(mfrow=c(2,2))
>  plot(fit1,res1,main="Data Set 1: Fitted VS Residual plot")
>  plot(fit2,res2,main="Data Set 2: Fitted VS Residual plot")
>  plot(fit3,res3,main="Data Set 3: Fitted VS Residual plot")
>  plot(fit4,res4,main="Data Set 4: Fitted VS Residual plot")
```

Box-Cox Transformation

If the simple linear regression model is not appropriate for the data set there are two choices
a) Abandon the simple linear regression model and develop and use a more appropriate model,
b) Employ some transformation on the data so that the simple linear regression model is appropriate for the transformed data.

Box-Cox Transformation or power transformation:

It is often difficult to determine from the scatter plot which transformation is most appropriate for correcting the skewness of the distributions of error terms, unequal error variance and the nonlinearity of the regression function. The box cox transformation is given by

$$g(y_i) = \frac{y_i^{\lambda} - 1}{\lambda}$$

If $\lambda = 1$, no transformation is needed and we analyze the original data. If $\lambda = -1$ we analyze the reciprocal $1/y_i$. If $\lambda = 1/2$, we analyze the the $\sqrt{y_i}$. And we analyze $\ln(y_i)$ if $\lambda = 0$
The maximum likelihood estimator of λ minimizes $SSE(\lambda)$ where $SSE(\lambda)$ is the residuals sum of squares from fitting the regression model with

Example

```
x<-c(0, 0, 0, 0, 0, 1,1, 1, 1, 1, 2, 2, 2, 2, 2, 3, 3, 3, 3, 3, 4, 4, 4, 4, 4)
y<-c(13.44, 12.84, 11.91, 20.09, 15.60, 10.11, 11.38, 10.28,  8.96,  8.59, 9.83,  9.00,
8.65,  7.85,  8.88,  7.94,  6.01,  5.14, 6.90,  6.77,  4.86,  5.10,  5.67,  5.75, 6.23)

>model1<-lm(y~x)
> par(mfrow=c(2,2)) # We need to specify this  dimension
> plot(model1)

>library(MASS)
>b=boxcox(model1) #  It will search the value of the parameter [-2,2]
>b=boxcox(model1, lambda=seq(0,3, by=0.01)) # for any positive value in [0,3]
>y1<-y^(-0.5)
> model2=lm(y1~x)
>par(mfrow=c(2,2))
> plot(model2)


>library(moments)
>skewness(model1$resid)
>skewness(model2$resid)
```

Regression Through the Origin

In practice sometimes one might be interested in building a model with no intercept. For example in chemical experiment the yield of a chemical process is zero when the temperature is zero.
The no intercept model is

$$y = \beta_1 x_i + \epsilon$$

Let b_1 be the estimator of β_1. Given n observations $(x_i, y_i), i = 1, 2, \cdots, n$ the least square function is

$$SSE = \sum_{i=1}^{n} (y_i - b_1 x_i)^2$$

The only normal equation is

$$b_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i$$

Therefore, the least square estimator of the slope is

$$b_1 = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}$$

Example

Data below measures the temperature(x) vs. the chemical product yield (y).

Temp(x)	95	100	105	110	115	125	135	140	145	150	155
Yield(y)	8	10	9	10	11	13	10	11	12	13	11

```
> x=c(95, 100, 105, 110, 115, 125, 135, 140, 145, 150, 155)
> y=c(8, 10,  9, 10, 11, 13, 10, 11, 12, 13, 11)
> model1<-lm(y~x)
> model1
Call:
lm(formula = y ~ x)# Intercept model
Coefficients:
(Intercept)            x
    4.33838      0.05111
> model2<-lm(y~-1+x) # No intercept model
> model2
Call:
lm(formula = y ~ -1 + x)
Coefficients:
      x
```

Simple Linear Regression Model in Matrix Terms

Let us consider a simple linear regression model

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1, 2, \cdots, n$$

This implies

$$\begin{aligned} y_1 &= \beta_0 + \beta_1 x_1 + \epsilon_1 \\ y_2 &= \beta_0 + \beta_1 x_2 + \epsilon_2 \\ &\vdots \\ y_n &= \beta_0 + \beta_1 x_n + \epsilon_n \end{aligned}$$

Let $\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$, $\mathbf{x} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}$, $\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}$ and $\boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}$

Note that $\mathbf{x} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}$ is called the design matrix.

Then we write the model as $\mathbf{y} = \mathbf{x}\boldsymbol{\beta} + \boldsymbol{\epsilon}$