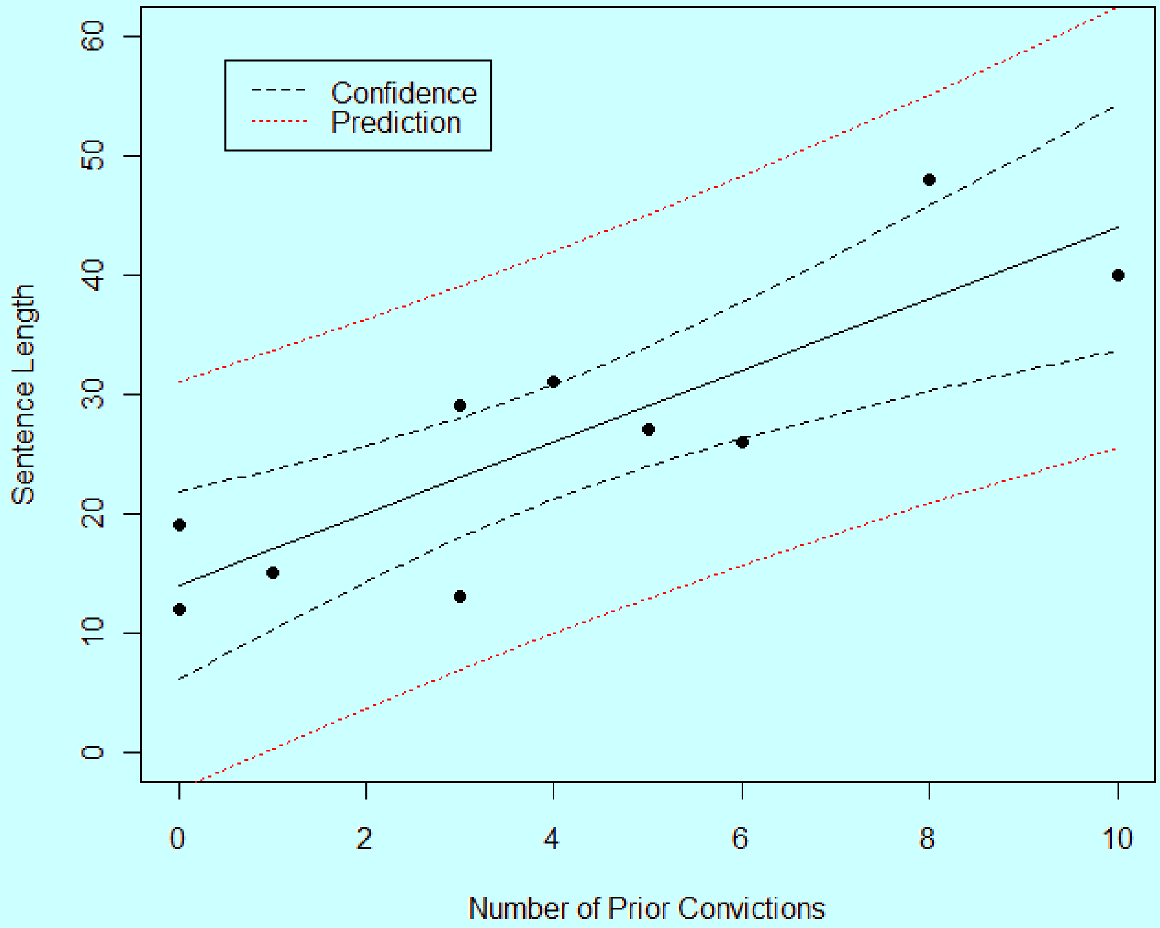
Regression Analysis



1. Univariate Linear Regression
2. Multivariate Linear Regression
3. Univariate Logistic Regression
4. Multivariate Logistic Regression

https://stats.idre.ucla.edu/other/dae/

1. Univariate Linear Regression

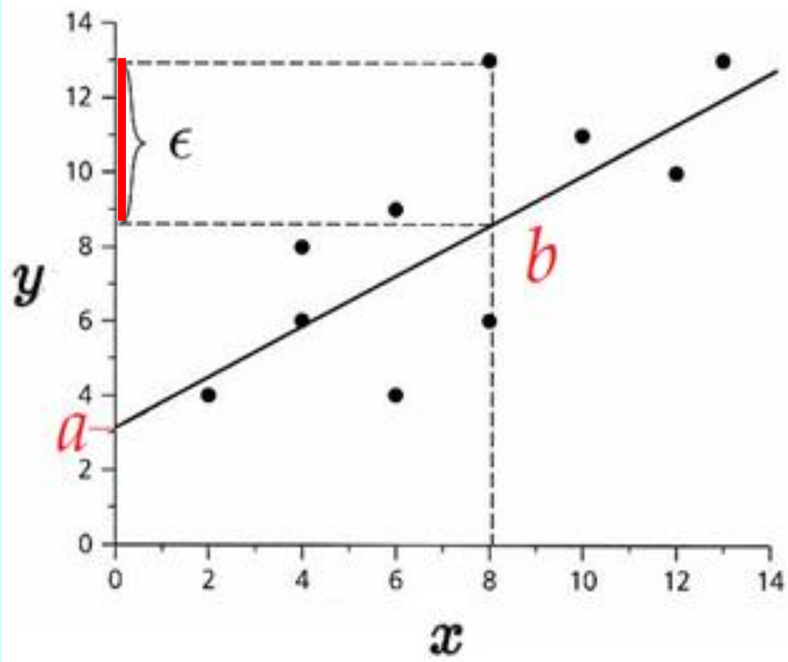
The **least-squares methods** in regression minimizes the sum of squared **residuals**, which is the difference between an observed *y* value and the fitted *y* value

Model: $y = a + bx + \varepsilon$

Least-squares fit:

Residuals $\varepsilon_i = y_i - a - bx_i$

$$a = \bar{y} - b\,\bar{x}, \quad b = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\mathrm{Cov}(x, y)}{\mathrm{Var}(x)}$$



Goodness of fit = how well a model describes the observed data

$$SS_{Total} = \sum (Y - Y_{Mean})^2$$

$$SS_{Residual} = \sum (Y - Y_{Predicted})^2$$

$$SS_{Model} = \sum (Y_{Predicted} - Y_{Mean})^2$$

$$SS_{Model} = SS_{Total} - SS_{Residual}$$

R-squared is the square of the correlation coefficient and ranges between 0 and 1.

$$R^2 = \frac{SS_M}{SS_T}$$

Coefficient of Determination

$$R^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2}$$

y_i : Observed data $\hat{y}_i$: Predicted data $\bar{y}$: Mean of the observed data

(bad) $0 \leq R^2 \leq 1.0$ (good)

#all of the *y* variables can be explained by *x*

- **Sample Data : longley**

longley {datasets} Longley's Economic Regression Data

A data frame with 7 economical variables, observed yearly from 1947 to 1962 (n=16).

```
> str(longley)
'data.frame':   16 obs. of  7 variables:
 $ GNP.deflator: num  83 88.5 88.2 89.5 96.2 ...
 $ GNP         : num  234 259 258 285 329 ...
 $ Unemployed  : num  236 232 368 335 210 ...
 $ Armed.Forces: num  159 146 162 165 310 ...
 $ Population  : num  108 109 110 111 112 ...
 $ Year        : int  1947 1948 1949 1950 1951 1952
 $ Employed    : num  60.3 61.1 60.2 61.2 63.2 ...
```

lm(formula, data, ...) {stats}

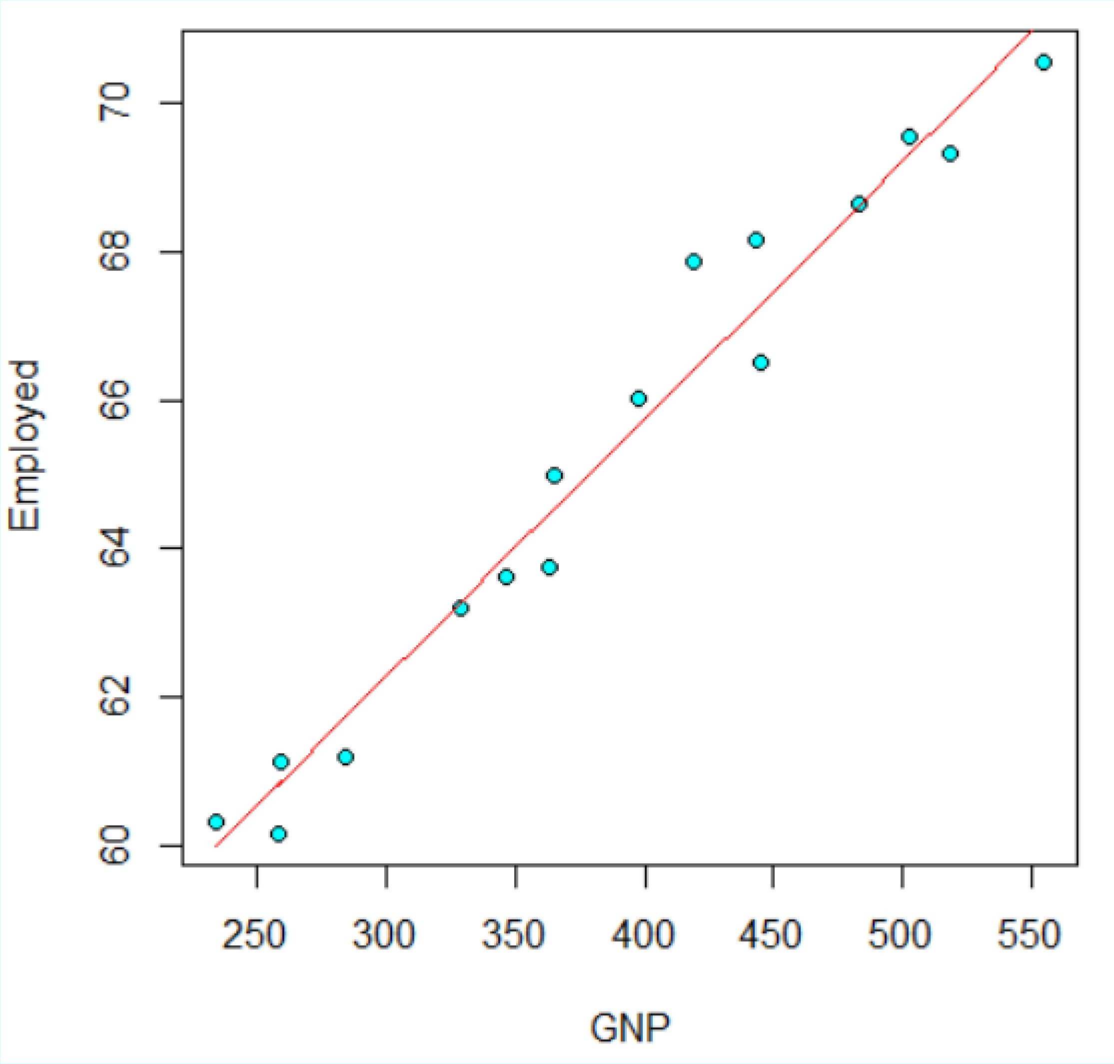
lm is used to fit linear models. It can be used to carry out regression.

```
> ur <- lm(Employed ~ GNP, data=longley)
> summary(ur)        #try anova(ur)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 51.843590   0.681372   76.09  < 2e-16 ***
GNP          0.034752   0.001706   20.37 8.36e-12 ***
---
Residual standard error: 0.6566 on 14 degrees of freedom
Multiple R-squared:  0.9674,    Adjusted R-squared:  0.965
F-statistic: 415.1 on 1 and 14 DF,  p-value: 8.363e-12
```

Regression equation:

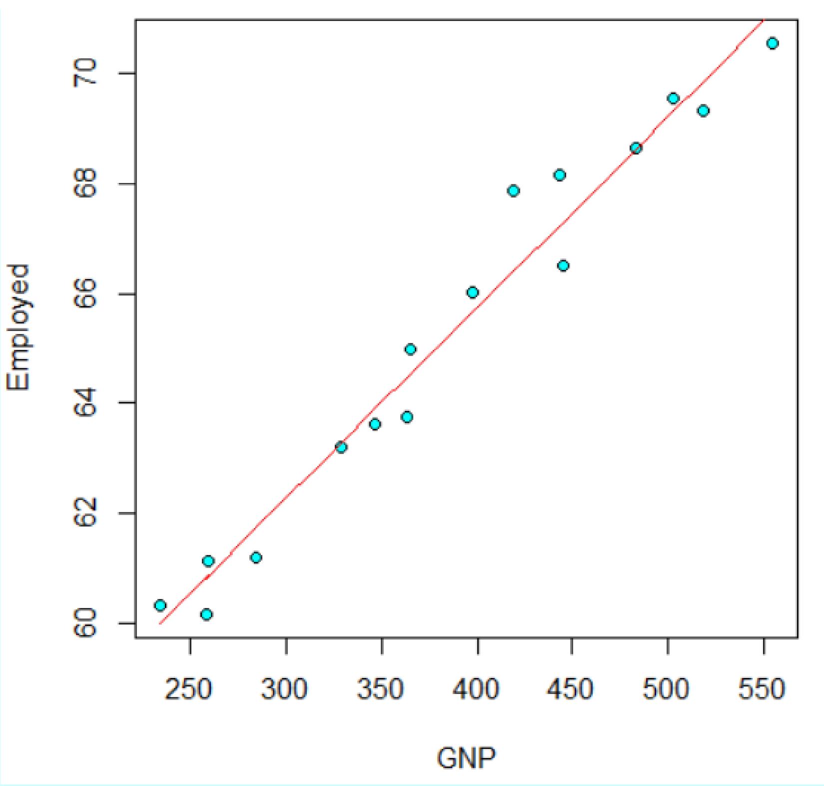
Employed = 51.844 + 0.035 GNP

```
# Plot with prediction
with(longley,plot(GNP,Employed,pch=21,bg='cyan'))
lines(longley$GNP,ur$fitted.values,col='red')
```



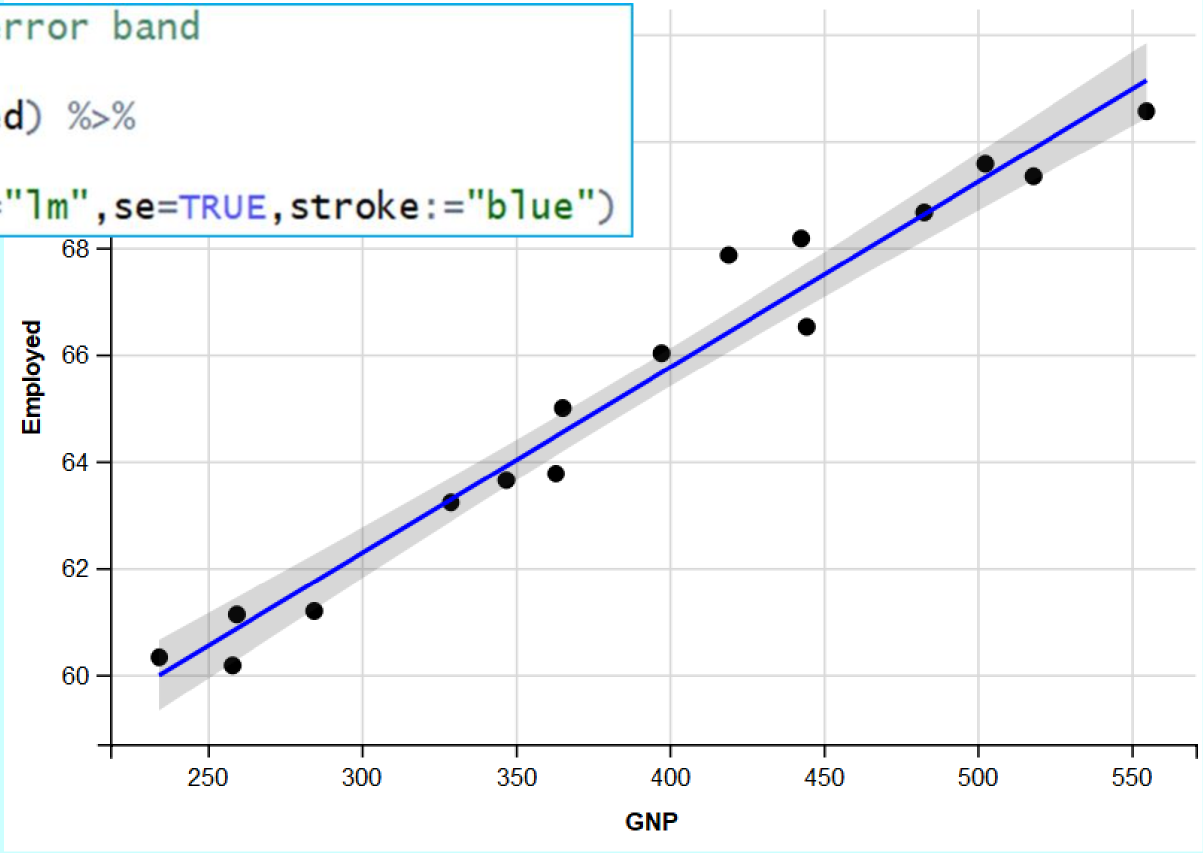
predict is a generic function for predictions from the results of various model fitting functions.

```
> predict(ur, list(GNP=300))
       1
62.26928
> predict(ur, list(GNP=c(300,500)))
       1        2
62.26928 69.21974
```



Quantity	Definition	R function
Mean	$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$	mean(x)
Standard Deviation	$s = \left[\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2\right]^{1/2}$	sd(x)
Standard Error	$\text{se} = \frac{s}{\sqrt{n}}$	sd(x)/sqrt(n)

```r
#Regression line with standard error band
library(ggvis)
longley %>% ggvis(~GNP, ~Employed) %>%
  layer_points() %>%
  layer_model_predictions(model="lm",se=TRUE,stroke:="blue")
```



2. Multivariate Linear Regression

Basic Model for Multiple Regression

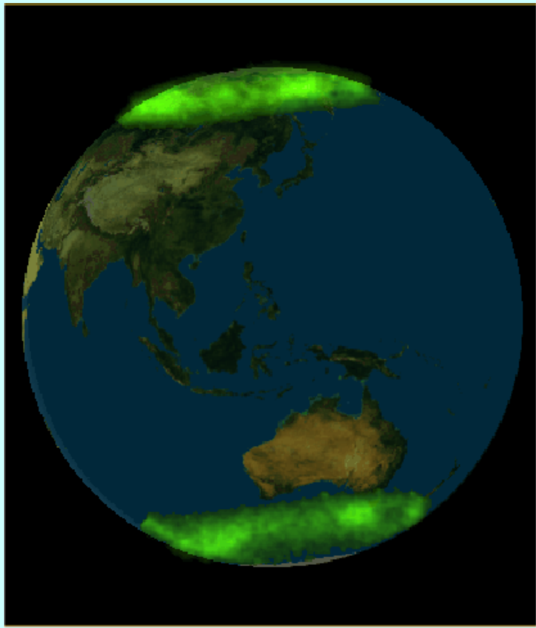
The basic model for multiple regression analysis is

$$\boxed{y = a_0 + a_1 x_1 + \cdots + a_k x_k + \epsilon}$$

where $x_1, \cdots, x_k$ are explanatory variables (predictors) and the parameters $a_0, \cdots, a_k$ can be estimated using the method of least squares.

- **Sample Data : flights**

```
> #Sample Data: flights
> library(nycflights13)
> flights_df <- as.data.frame(flights)
> head(flights_df,2)
  year month day dep_time sched_dep_time dep_delay arr_time
1 2013     1   1      517            515         2      830
2 2013     1   1      533            529         4      850
  sched_arr_time arr_delay carrier flight tailnum origin dest
1            819        11      UA   1545  N14228    EWR  IAH
2            830        20      UA   1714  N24211    LGA  IAH
  air_time distance hour minute           time_hour
1      227     1400    5     15 2013-01-01 05:00:00
2      227     1416    5     29 2013-01-01 05:00:00
```



[Task] What factors influence departure delay at JFK?

```
> ######Base R: multivariate regression
> library(dplyr)
> base_dat <- flights_df %>%
+   filter(origin=="JFK", dep_delay>0, arr_delay>0)
> form <- dep_delay ~ arr_delay + distance + air_time
> mfit1 = lm(form, data=base_dat)
> summary(mfit1)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 14.400617   0.260002   55.39   <2e-16 ***
arr_delay    0.926291   0.001883  491.90   <2e-16 ***
distance     0.072484   0.001016   71.35   <2e-16 ***
air_time    -0.579330   0.007913  -73.21   <2e-16 ***
---
Residual standard error: 19.22 on 29319 degrees of freedom
Multiple R-squared:  0.8936,    Adjusted R-squared:  0.8936
F-statistic: 8.207e+04 on 3 and 29319 DF,  p-value: < 2.2e-16
```

Regression equation:

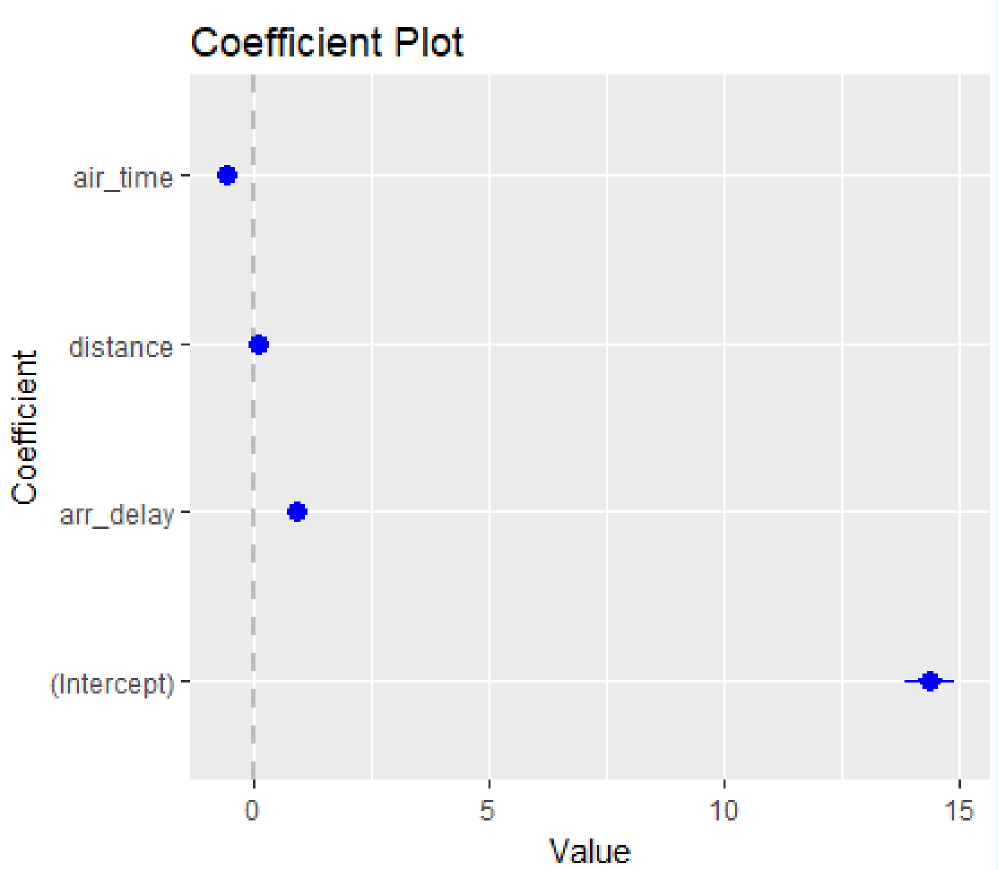
dep_delay = 14.40 + 0.926 arr_delay + 0.072 distance - 0.579 arr_time

Influence factors:
arr_delay, distance, air_time

coefplot(model, ...) {coefplot} Plotting Model Coefficients

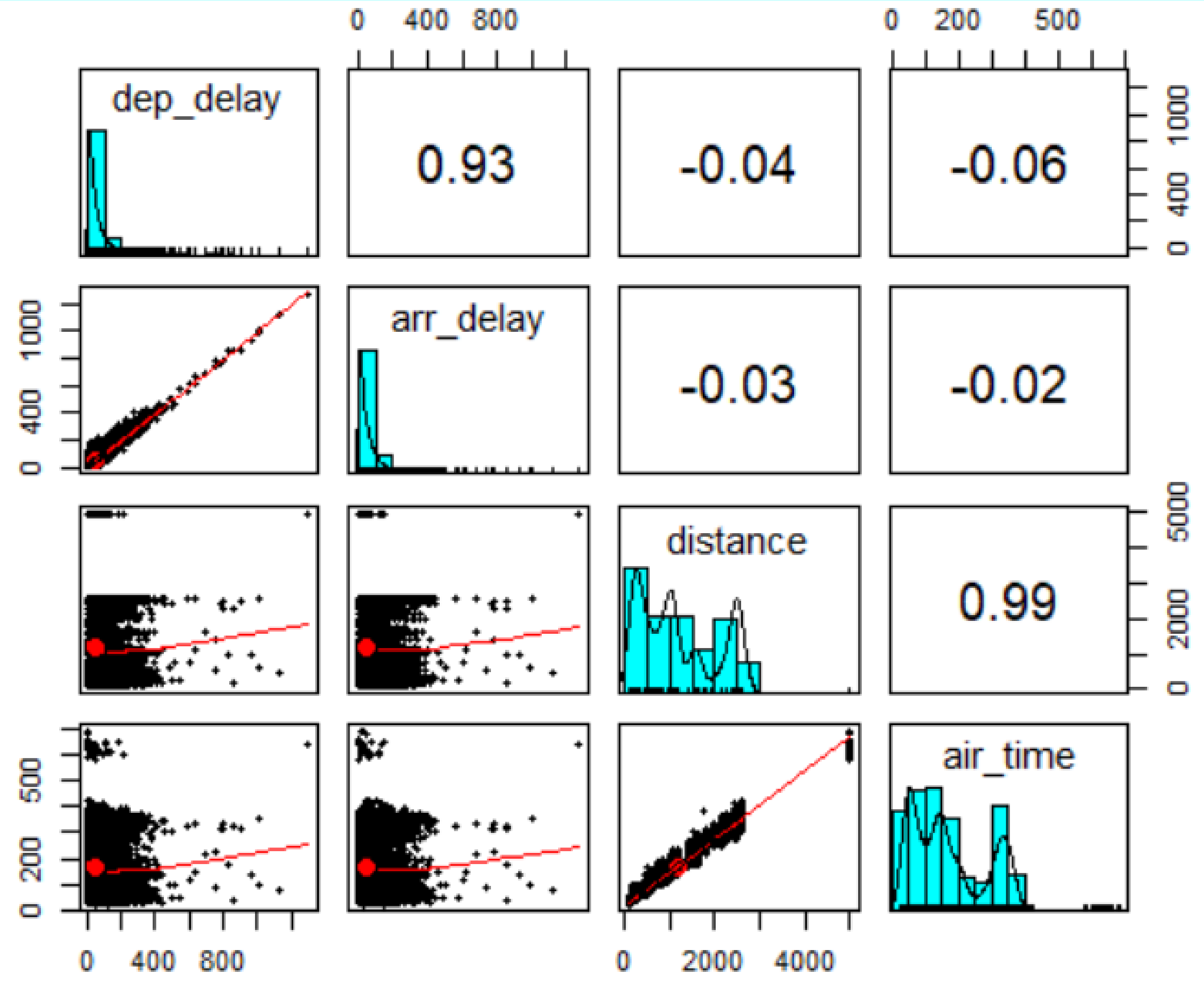
A graphical display of the coefficients and standard errors from a fitted model

```
#Coefficient plot
library(coefplot)
coefplot(mfit1)
```



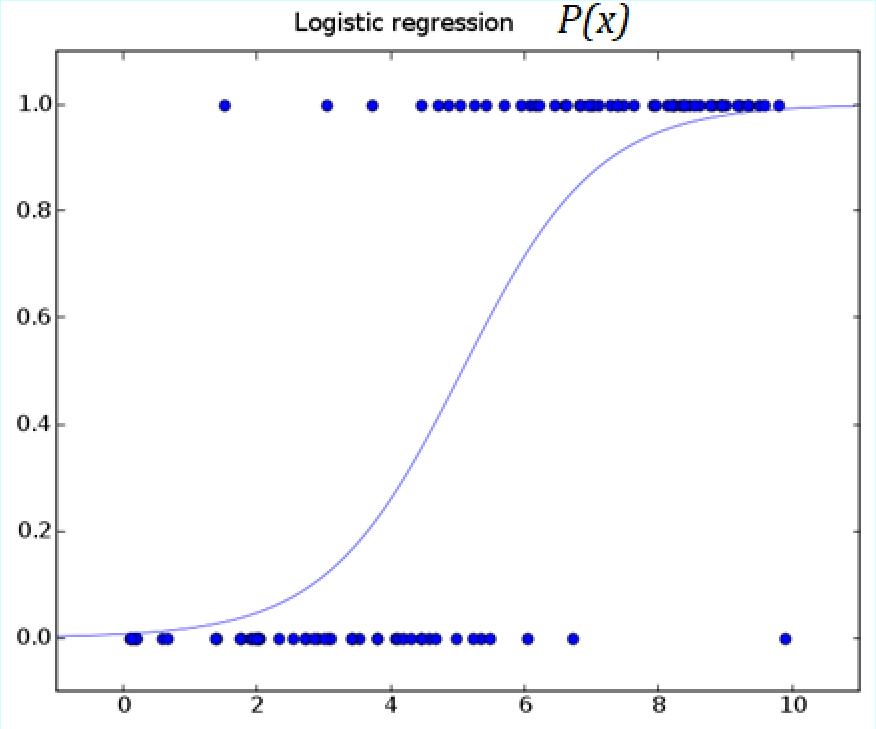
• Matrix Scatterplot


```
library(psych)
cordat <- base_dat %>%
  select(dep_delay,arr_delay,distance,air_time)
pairs.panels(cordat)
```



3. Univariate Logistic Regression

- Logistic regression is used to fit a regression curve, $\boldsymbol{y=P(x)}$.
- The dependent variable $\boldsymbol{y}$ is categorical, in general, binary.
- The predictors $\boldsymbol{x}$ can be continuous, ratio, interval or categorical.



The logistic function $P(x)$:

$$P(x) = \frac{1}{1+e^{-(\beta_0+\beta_1 x)}} = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}$$

$P(x)$ is interpreted as the probability of the dependent variable equaling a "success".

Odds

Odds is the ratio of success to ratio of failure. It ranges between 0 and positive infinity. The higher the odds, the better the chance for success.

$$\text{odds} = \frac{\text{probability(success)}}{\text{probability(failure)}} = \frac{\text{P}}{\text{1-P}} = e^{\beta_0+\beta_1 x}$$

Odds Ratio

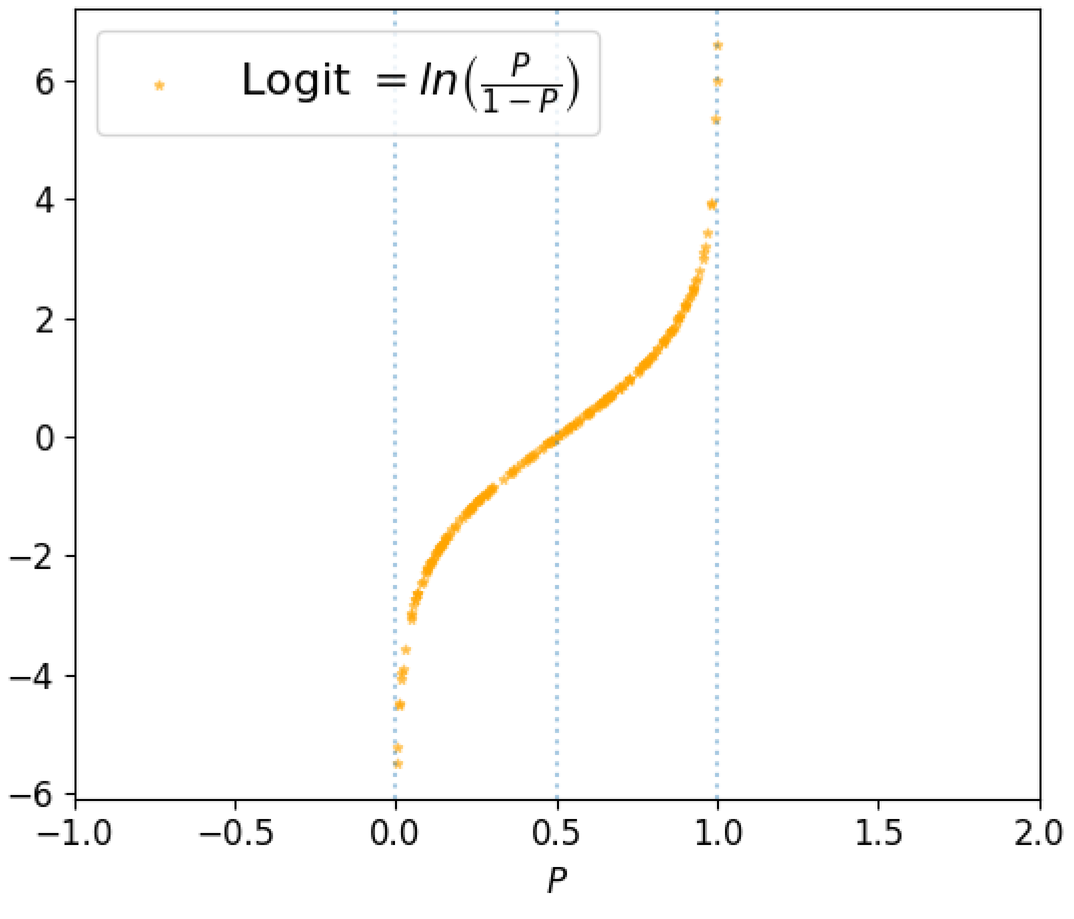
The ratio of odds, also between 0 and positive infinity. This represents the odds that an outcome will occur given a particular exposure, compared to the odds of the outcome occurring in the absence of that exposure.

$$OR = \text{odds}(x+1)/\,\text{odds}(x) = e^{\beta_1}$$

Logit (log of odds)

Transforms [0,1] to [negative infinity, positive infinity]

$$\text{logit}(P) = \log\frac{P}{1-P} = \beta_0 + \beta_1 x$$



$$\begin{aligned}\text{logit}(P) &= \log\frac{P}{1-P} \\ &= \ln\left[\frac{\frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}}{1-\frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}}\right] \\ &= \ln\left[\frac{\frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}}{\frac{1}{1+e^{\beta_0+\beta_1 X}}}\right] \\ &= \ln\left[e^{\beta_0+\beta_1 X}\right] \\ &= \beta_0 + \beta_1 X\end{aligned}$$

https://tinyurl.com/y5oyarkq

[Sample Data] Probability of passing an exam versus hours of study

[Reference] https://en.wikipedia.org/wiki/Logistic_regression

A group of 20 students spend between 0 and 6 hours studying for an exam. How does the number of hours spent studying affect the probability that the student will pass the exam?

Table: The number of hours each student spent studying, and whether they passed (1) or failed (0).

Hours	0.50	0.75	1.00	1.25	1.50	1.75	1.75	2.00	2.25	2.50	2.75	3.00	3.25	3.50	4.00	4.25	4.50	4.75	5.00	5.50
Pass	0	0	0	0	0	0	1	0	1	0	1	0	1	0	1	1	1	1	1	1

```
> Hours <- c(0.50,0.75,1.00,1.25,1.50,1.75,1.75,2.00,2.25,2.50,
+           2.75,3.00,3.25,3.50,4.00,4.25,4.50,4.75,5.00,5.50)
> Pass  <- c(0,0,0,0,0,0,1,0,1,0, 1,0,1,0,1,1,1,1,1,1)
> passhour <- data.frame(Hours, Pass)
> head(passhour,3); tail(passhour,3)
  Hours Pass
1  0.50    0
2  0.75    0
3  1.00    0
   Hours Pass
18  4.75    1
19  5.00    1
20  5.50    1
> table(passhour$Pass)

 0  1
10 10
```

glm(formula, family = gaussian, data, …)

glm is used to fit generalized linear models, specified by giving a symbolic description of the linear predictor and a description of the error distribution.

```
> ######Base R: Univariate logistic regression
> out1 <- glm(Pass~Hours,family=binomial(logit),data=passhour)
> summary(out1)
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -4.0777     1.7610  -2.316   0.0206 *
Hours         1.5046     0.6287   2.393   0.0167 *
---
    Null deviance: 27.726  on 19  degrees of freedom
Residual deviance: 16.060  on 18  degrees of freedom
AIC: 20.06
```

Interpretation of the output

(1) P-value=0.0167: Hours studying is significantly associated with the probability of passing the exam.

(2) $\text{logit}(P) = \beta_0 + \beta_1 x = -4.0777 + 1.5046 \times \text{Hours}$

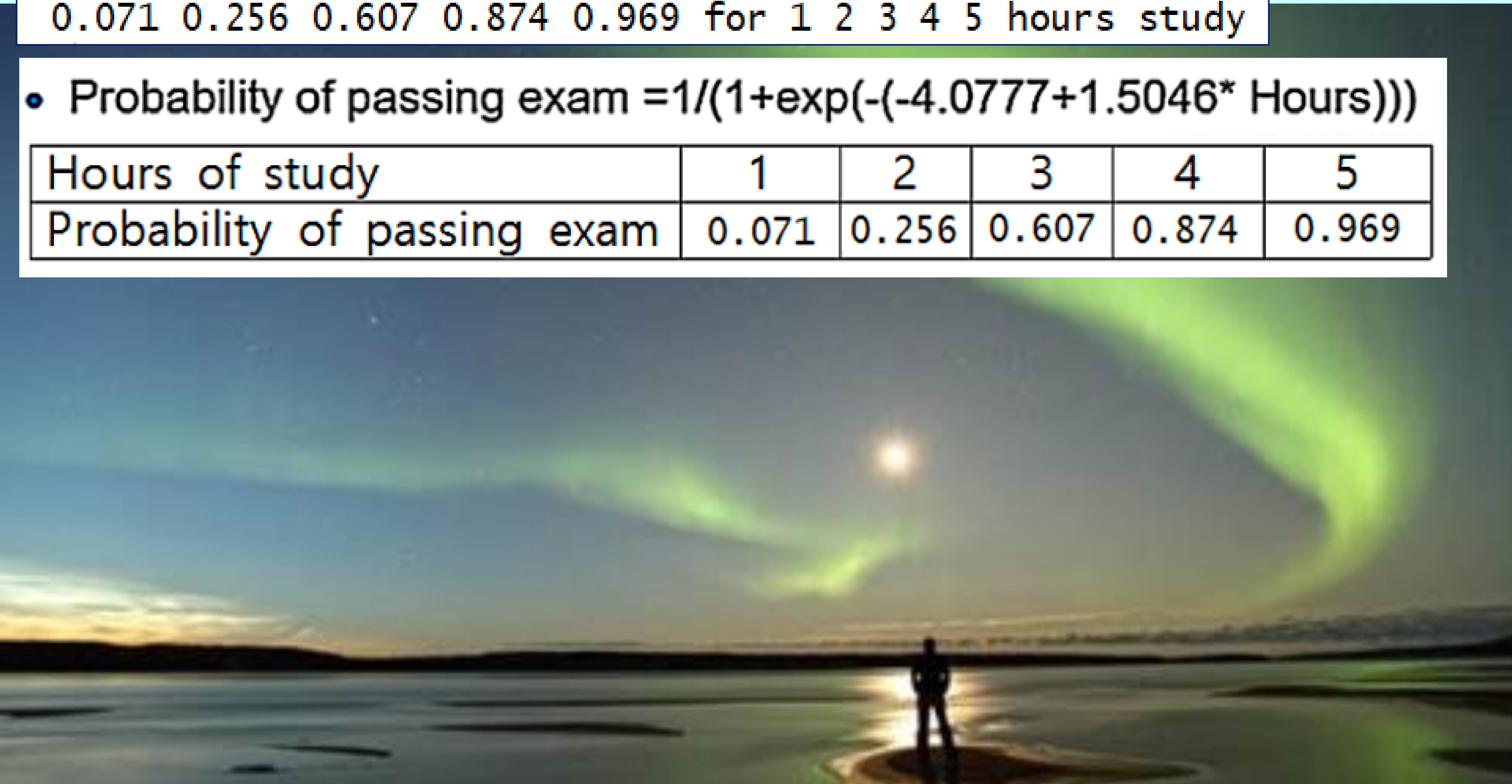
(3) OR=exp(1.5046)=4.502557: The odds ratio is 4.502557. The odds of passing an exam given an extra hour study, compared to the odds of passing an exam without that extra hour study, is 4.502557 times higher.

(3) Probability of passing the exam for 1~5 hours study

```
> b = coef(out1)
> x = 1:5; P = 1.0/(1+exp(-b[1]-b[2]*x))
> cat("Probabilities of passing exam:\n",
+      round(P,3),"for",x,"hours study")
Probabilities of passing exam:
 0.071 0.256 0.607 0.874 0.969 for 1 2 3 4 5 hours study
```

- Probability of passing exam =1/(1+exp(-(-4.0777+1.5046* Hours)))

Hours of study	1	2	3	4	5
Probability of passing exam	0.071	0.256	0.607	0.874	0.969

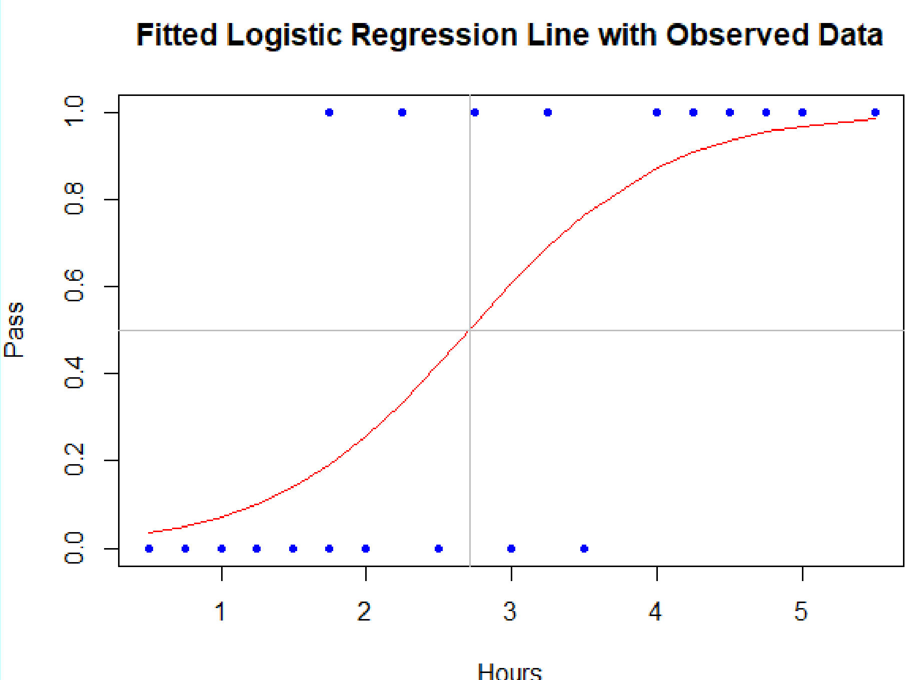


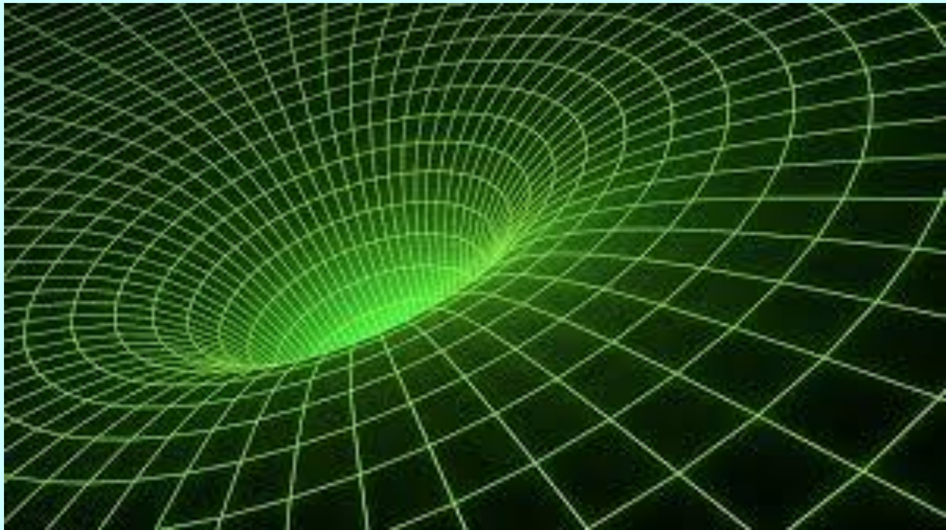
(4) Boundary hour to pass exam

```
> #logit(P)=0 at P=0.5 -> b[1]+b[2]*H=0, H=-b[1]/b[2]
> H = -b[1] / b[2]
> cat("Boundary hour to pass exam:",H,"\n")
Boundary hour to pass exam: 2.710083
```

(5) Visualization of Logistic Function

```
plot(Pass~Hours, pch=20,col="blue",
     main='Fitted Logistic Regression Line with Observed Data')
lines(Hours, out1$fitted, type="l", col="red")
abline(h=0.5,v=H,col="gray")
```





Multivariate Logistic Regression

The logit function now represents the probability of an event that depends on *k* covariates or independent variables

$$\text{logit}(P) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k$$

For simplicity, let us assume that the interaction between covariates does not exist in this class.

Correlations between Variables

[Sample Data] badhealth

data(**badhealth**) {COUNT}

The data may be evaluated as a logistic or other binary response model with the binary variable "badh" as the response.

numvisit : Number of visits to a physician during the year: 0 - 40

badh : 0=patient evaluates self as in good health, 1=patient in bad health

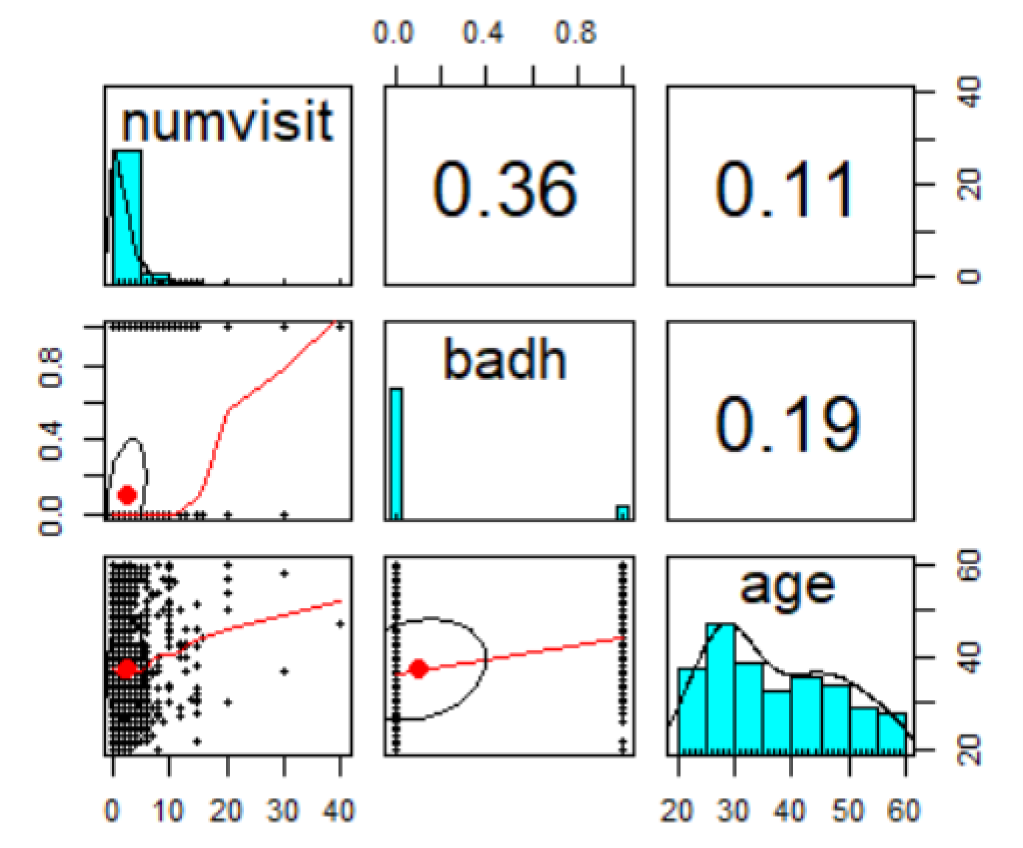
age : patient age 20~ 60

```
> data(badhealth, package="COUNT")
> head(badhealth,2)
  numvisit badh age
1       30    0  58
2       20    0  54
> #Look how many unique values
> sapply(badhealth, function(x) length(unique(x)))
numvisit     badh      age
      20        2       41
> table(badhealth$badh)

   0    1
1015  112
```

Correlations between Variables

```
library(psych)
pairs.panels(badhealth)
```



Conduct Multivariate Logistic Regression Analysis

```
> ### Base R
> mlr_fit <- glm(badh~numvisit+age,family=binomial(logit),data=badhealth)
> summary(mlr_fit)
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -5.04184    0.44998 -11.205  < 2e-16 ***
numvisit     0.22122    0.02628   8.419  < 2e-16 ***
age          0.05281    0.01007   5.244 1.57e-07 ***
---
    Null deviance: 729.66  on 1126  degrees of freedom
Residual deviance: 603.43  on 1124  degrees of freedom
AIC: 609.43
```

Interpretation of the output

(1) P-value<0.05 for both of numvisit and age

The numvisit and age are both statistically significant to the probability of badh.

(2) Probability of badh (bad health) = 1/[1+exp{-logit(P)}]

$$\mathrm{logit}(P) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

$$= -5.04184 + 0.22122\ \text{numvisit} + 0.05281\ \text{age}$$

(3) Odds Ratio


```
> #Odds Ratio
> exp(coef(mlr_fit)[2:3])
numvisit      age
1.247603 1.054233
```

For every one visit increase in **numvisit**, the odds of having reached bad health increases by **exp(0.22122) = 1.2476 times.**

For every one year increase in **age**, the odds of having reached bad health increases by **exp(0.05281) = 1.0542 times.**

(4) Hosmer-Lemeshow Test

HLtest(model, g = 10) {vcdExtra}

Hosmer-Lemeshow goodness of fit test for a binomial glm object in logistic regression

```
> library(vcdExtra)
> HLtest(model=mlr_fit)
Hosmer and Lemeshow Goodness-of-Fit Test
Call:
glm(formula = badh ~ numvisit + age, family = binomial(logit),
    data = badhealth)
 ChiSquare df   P_value
  8.210295  8 0.4132023
```

Small p-values in this model mean that the model is a poor fit. In our case, there is no evidence of a poor fit. => Ho is that the fitted model is correct.