Week2 Lecture - Linear Regression

Unit Coordinator - Dr Liwan Liyanage

School of Computing, Engineering and Mathematics

WESTERN SYDNEY UNIVERSITY

Linear Regression

Linear Regression is to build a function of independent variables (also known as predictors) to predict a dependent variable (also called response or target).

- Example 1: Marketing Manager would like to assess the effect of Sales on the amount spent on advertising mediums, radio, TV, and Newspaper etc.
- Example 2: When testing the performance of an automobile one would like to assess how the horsepower relates to the fuel efficiency (miles per gallon).
- Example 3: Banks may wish to assess the risk of homeloan applicants based on their age, income, expenses, occupation, number of dependents, total credit limit, etc.

This lecture introduces basic concepts and presents examples of various regression techniques.

- Simple linear regression
  - Maximum likelihood estimates/ Least square estimates of the parameters
  - Important assumptions
  - Unreliability estimates for the parameters
  - ANOVA Table and critical value of F
  - Degree of scatter
  - 95% confidence intervals for the estimated parameters
  - Model checking
  - Prediction using the fitted model
- Logistic regression
- A collection of helpful R functions for regression analysis

WESTERN SYDNEY UNIVERSITY

Simple Linear Regression

Simple linear regression is a statistical method that allows us to summarize and study relationships between two continuous (quantitative) variables.

First we introduce the concept and basic procedures of simple linear regression.

We will also learn two measures that describe the strength of the linear association that we find in data.

Learning objectives & outcomes

Upon completion of this lecture, you should be able to do the following:

- Distinguish between a deterministic relationship and a statistical relationship.
- Understand the concept of the least squares criterion.
- Interpret the intercept α and slope β of an estimated regression equation.
- Know how to obtain the estimates α and β using R's fitted line plot and regression analysis output.
- Recognize the distinction between a population regression line and the estimated regression line.
- Summarize the four conditions that comprise the simple linear regression model.

Learning objectives & outcomes ctd...

- Know what the unknown population variance (σ^2) quantifies in the regression setting.
- Know how to obtain the estimate MSE of the unknown population variance (σ^2) from R's fitted line plot and regression analysis output.
- Know that the coefficient of determination (R^2) and the correlation coefficient (r) are measures of linear association. That is, they can be 0 even if there is perfect nonlinear association.
- Know how to interpret the (R^2) value.
- Understand the cautions necessary in using the (R^2) value as a way of assessing the strength of the linear association.
- Know how to calculate the correlation coefficient r from the (R^2) value.
- Know what various correlation coefficient values mean. There is no meaningful interpretation for the correlation coefficient as there is for the (R^2) value.

Regreesion Analysis

As mentioned before, Regression analysis is the statistical method you use when both the response variable and the explanatory variables are continuous variables.

i.e. real numbers with decimal places for example variables such as heights, weights, volumes, or temperatures

- Consider: Advertising Data set

Upload and View The Advertising Data set which contains Continuous Variables, Sales and Advertising Budget in three different types of Marketting Methods

Import the Data Set “Advertising”

```r
Advertising <- read.csv("Advertising.csv")
attach(Advertising)
names(Advertising)
```

```
## [1] "TV"        "Radio"     "Newspaper" "Sales"
```

```r
head(Advertising)
```

```
##      TV Radio Newspaper Sales
## 1 230.1  37.8      69.2  22.1
## 2  44.5  39.3      45.1  10.4
## 3  17.2  45.9      69.3   9.3
## 4 151.5  41.3      58.5  18.5
## 5 180.8  10.8      58.4  12.9
## 6   8.7  48.9      75.0   7.2
```

Seven important kinds of regression analysis:

- linear regression (the simplest, and much the most frequently used);
- polynomial regression (often used to test for non-linearity in a relationship);
- multiple regression (where there are numerous explanatory variables);
- non-linear regression (to fit a specified non-linear model to data);
- piecewise regression (two or more adjacent straight lines);
- robust regression (models that are less sensitive to outliers);
- non-parametric regression (used when there is no obvious functional form).

The first four cases are covered here

WESTERN SYDNEY UNIVERSITY

Simple Linear Regression

First, however, we need to select a model which describes the relationship between the response variable and the explanatory variable(s).

The simplest of all is the linear model:

In linear regression, the expected value of y_i given x_i is:

$E(y_i) = \alpha + \beta x_i$ for $i = 1, 2, ..., n$

y_i has a normal distribution with standard deviation σ. It is the random component of the model, which has a normal distribution.

The response variable is Y, and X is a single continuous explanatory variable. The parameters are α and β :

- The intercept is α: The value of Y when X = 0
- The slope is β: The change in Y divided by the change in X, which is the change in Y when X changes by one unit

Estimating the parameters

Linear regression finds the line that best fits the data points. There are actually a number of different definitions of "best fit," and therefore a number of different methods to find the parameters of linear regression.

By far the most common is "ordinary least-squares regression"; when someone just says "least-squares regression" or "linear regression" or "regression," they mean ordinary least-squares regression.

In ordinary least-squares regression, the "best" fit is defined as the line that minimizes the squared vertical distances between the data points and the line.

For a data point with an X value of X_1 and a Y value of Y_1, the difference between Y_1 and $\hat{Y}_1$ (the predicted value of Y at X_1) is calculated, then squared. This squared deviate is calculated for each data point, and the sum of these squared deviates measures how well a line fits the data. The regression line is the one for which this sum of squared deviates is smallest.

Let's start with the example which shows how the marketing dollar is influencing Sales

Consider the Uploaded data file “Advertising” you viewed earlier in Week 1

- How large is the sample?

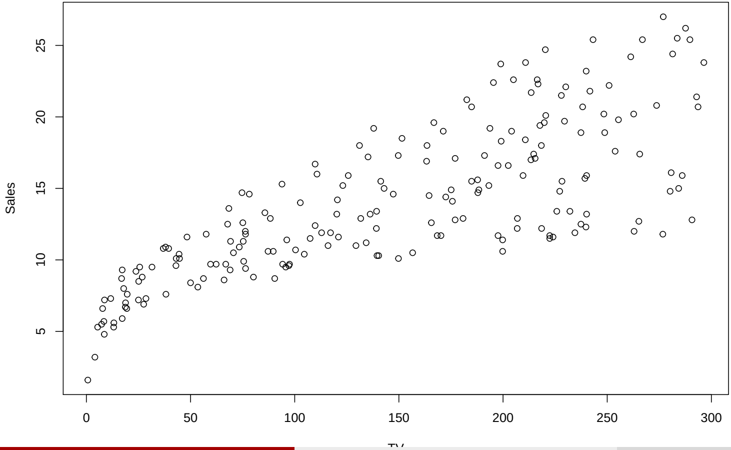
```
dim (Advertising)
```

```
## [1] 200   4
```

WESTERN SYDNEY UNIVERSITY

Construct a scatter plot of the data

```r
plot(Sales~TV)
```

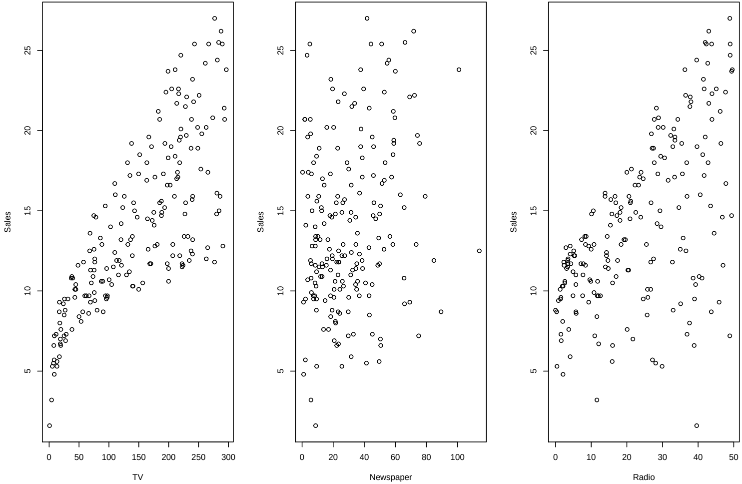


Compare the different media types: R Code

```
par(mfrow=c(1,3))
plot(Sales~TV, ylab='Sales', xlab='TV')
plot(Sales~Newspaper, ylab='Sales', xlab='Newspaper')
plot(Sales~Radio, ylab='Sales', xlab='Radio')
```

Compare the different media types: Plots

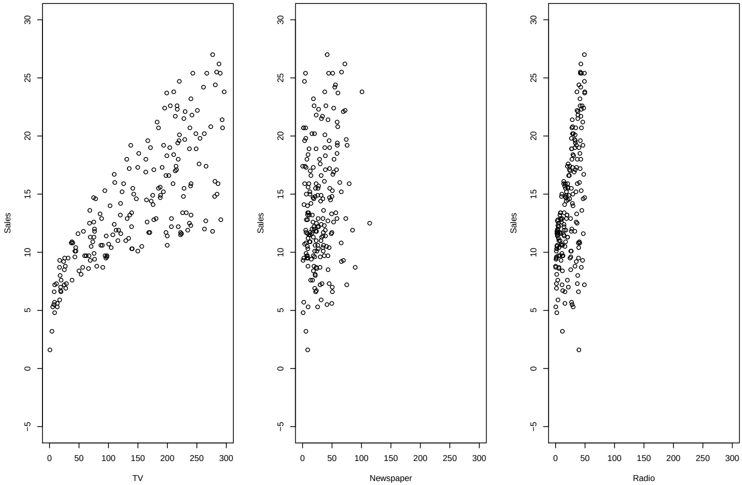
Note: Variable X has varying scales for different media types



Compare the different media types after Adjusting to reflect the same X scale

```
par(mfrow=c(1,3))
plot(Sales~TV, xlim=c(0,300), ylim=c(-5,30),
     ylab='Sales', xlab='TV')
plot(Sales~Newspaper, xlim=c(0,300), ylim=c(-5,30),
     ylab='Sales', xlab='Newspaper')
plot(Sales~Radio, xlim=c(0,300), ylim=c(-5,30),
     ylab='Sales', xlab='Radio')
```

Compare the different media types



Maximum likelihood estimates/ Least square estimates of the parameters

α and β

Given the data, and having selected a linear model, we want to find the values of the slope and intercept that make the data most likely.

The estimated linear model $\hat{y} = \alpha + \beta x$ for Sales Vs TV

```r
lm(Sales~TV)
```

```
##
## Call:
## lm(formula = Sales ~ TV)
##
## Coefficients:
## (Intercept)           TV
##     7.03259      0.04754
```

Maximum likelihood estimates/ Least square estimates ctd...

The estimated linear model $\hat{y} = \alpha + \beta x$ for Sales Vs Newspaper

```
lm(Sales~Newspaper)
```

```
##
## Call:
## lm(formula = Sales ~ Newspaper)
##
## Coefficients:
## (Intercept)    Newspaper
##    12.35141      0.05469
```

Maximum likelihood estimates/ Least square estimates ctd. . .

The estimated linear model $\hat{y} = \alpha + \beta x$ for Sales Vs Radio

```r
lm(Sales~Radio)
```

```
##
## Call:
## lm(formula = Sales ~ Radio)
##
## Coefficients:
## (Intercept)        Radio
##      9.3116       0.2025
```

WESTERN SYDNEY UNIVERSITY

We can now write the maximum likelihood equations as follows:

$$\text{Sales} = 7.03259 + 0.04754 \times \text{TV}$$

$$\text{Sales} = 12.35141 + 0.05469\text{xNewspaper}$$

$$\text{Sales} = 9.3116 + 0.2025\text{xRadio}$$

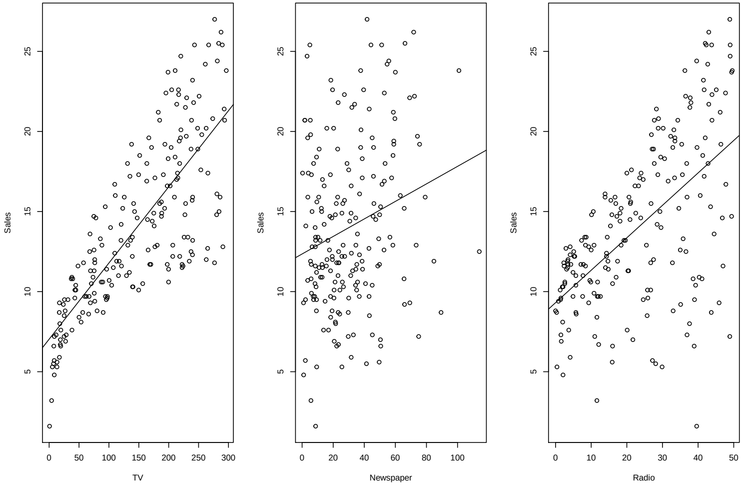
WESTERN SYDNEY UNIVERSITY

Compare the different media types: Using Scatter Plots with Regression line

R Code

```
par(mfrow=c(1,3))
plot(Sales~TV, ylab='Sales', xlab='TV')
abline(a=7.03259,b=0.04754)
plot(Sales~Newspaper, ylab='Sales', xlab='Newspaper')
abline(a=12.35141,b=0.05469)
plot(Sales~Radio, ylab='Sales', xlab='Radio')
abline(a=9.3116,b=0.2025)
```

Compare the different media types: Scatter Plots with Regression line



WESTERN SYDNEY UNIVERSITY

Important assumptions:

- The variance in y is constant (i.e. the variance does not change as y gets bigger).
- The explanatory variable, x, is measured without error.
- The difference between a measured value of y and the value predicted by the model for the same value of x is called a residual.
- Residuals are measured on the scale of y (i.e. parallel to the y axis).
- The residuals are normally distributed.

Under these assumptions, the maximum likelihood is given by the method of least squares

Unreliability estimates for the parameters

```r
model= (lm(Sales~TV))
summary (model)
```

```
##
## Call:
## lm(formula = Sales ~ TV)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.3860 -1.9545 -0.1913  2.0671  7.2124
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 7.032594   0.457843   15.36   <2e-16 ***
## TV          0.047537   0.002691   17.67   <2e-16 ***
## ---
```

ANOVA Table and critical value of F:

```r
anova(model)
```

```
## Analysis of Variance Table
##
## Response: Sales
##            Df Sum Sq Mean Sq F value    Pr(>F)
## TV          1 3314.6  3314.6  312.14 < 2.2e-16 ***
## Residuals 198 2102.5    10.6
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
```

```r
qf(0.95,1,198)
```

```
## [1] 3.888853
```

Degree of scatter

Residual standard error: 3.259 on 198 degrees of freedom Multiple R-squared: 0.6119, Adjusted R-squared: 0.6099 F-statistic: 312.1 on 1 and 198 DF, p-value: $< 2.2e-16$

- The summary.lm table shows everything you need to know about the parameters and their standard errors

Degree of scatter ctd ...

The residual standard error is the square root of the error variance from the ANOVA table = 3.259

Multiple R-squared is the fraction of the total variance explained by the model SSR/SSY = 0.6119.

The Adjusted R-squared is close to, but different from, the value of R2 we have just calculated. Instead of being based on the explained sum of squares SSR and the total sum of squares SSY, it is based on the overall variance (a quantity we do not typically calculate), = SSY/(n-1)

Large F Statistic or small p value indicates a significant linear relationship between Y and X.

95% confidence intervals for the estimated parameters

```r
confint(model)
```

```
##                  2.5 %     97.5 %
## (Intercept) 6.12971927 7.93546783
## TV          0.04223072 0.05284256
```

- These values are obtained by subtracting from, and adding to, each parameter estimate an interval which is the standard error times Student's t with given degrees of freedom
- If the interval do not include 0 it indicates that corresponding parameter value is significantly different from zero, which should confirm the established outcome by the earlier F tests.

Prediction using the fitted model

```r
predict(model,list(TV = 100.0))
```

```
##        1
## 11.78626
```

- Indicating a predicted Sales of 11.78626 million dollars where allocation of TV marketting budget is $100 thousand.
- To predict Sales at more than one level of marketting budget the list of values for the explanatory variable is specified as a vector

```r
predict(model,list(TV=c(50.0,100.0,150.0,200.0)))
```

```
##         1         2         3         4
##  9.409426 11.786258 14.163090 16.539922
```

Model checking

For instance, we should routinely plot the residuals against:

- the fitted values (to look for heteroscedasticity);
- the explanatory variables (to look for evidence of curvature);
- the sequence of data collection (to look for temporal correlation);
- standard normal deviates (to look for non-normality of errors).

Model checking ctd. . .

The assumptions we really want to be sure about are constancy of variance and normality of errors.

The simplest way to do this is with model-checking plots.

Six plots are currently available:

- a plot of residuals against fitted values;
- a scale-location plot of residuals against fitted values;
- a normal QQ plot;
- a plot of Cook's distances versus row labels;
- a plot of residuals against leverages;
- a plot of Cook's distances against leverage/(1-leverage).

By default four plots are provided (the first three plus the fifth):

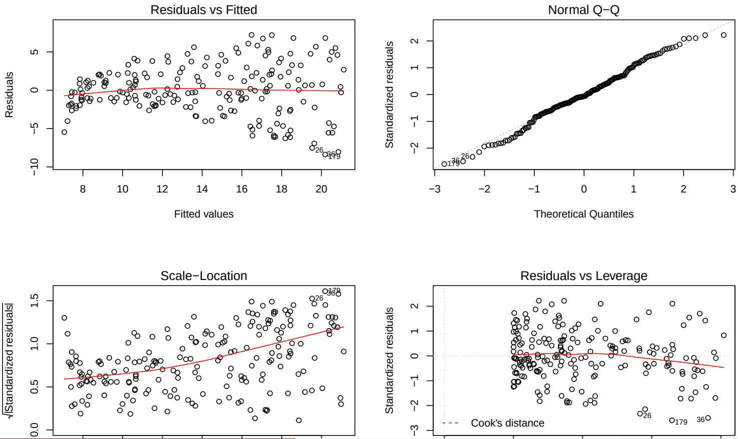
Model checking ctd. . .

```
influence.measures(lm(Sales ~ TV))
```

```
## Influence measures of
##   lm(formula = Sales ~ TV) :
##
##         dfb.1_    dfb.TV    dffit cov.r   cook.d     hat inf
## 1   -3.03e-02  0.087890  0.12624 1.003 7.94e-03 0.00970
## 2    4.22e-02 -0.032860  0.04281 1.021 9.20e-04 0.01217
## 3    5.79e-02 -0.048399  0.05798 1.025 1.69e-03 0.01649
## 4    4.27e-02  0.004851  0.09332 0.998 4.34e-03 0.00501
## 5   -9.67e-03 -0.023445 -0.06393 1.009 2.05e-03 0.00578
## 6   -1.03e-02  0.008765 -0.01031 1.029 5.34e-05 0.01805
## 7    6.27e-02 -0.046566  0.06443 1.017 2.08e-03 0.01047
## 8    7.64e-03 -0.003094  0.01034 1.016 5.38e-05 0.00549
## 9   -1.11e-01  0.094283 -0.11086 1.022 6.16e-03 0.01807
## 10   3.79e-03 -0.080304 -0.15310 0.983 1.16e-02 0.00690
```

Model checking ctd. . .

```r
par(mfrow=c(2,2))
plot(model)
```



WESTERN SYDNEY UNIVERSITY

Model checking ctd. . .

- If the model is linear plot 1 should look like the sky at night, with no pattern of any sort.
- In plot 2 line should be straight if the normality assumption of the residuals is valid
- The third graph is good for detecting non-constancy of variance (heteroscedasticity)
- The fourth graph shows any possible patterns in the standardized residuals as a function of the leverage. The graph also shows Cook's distance, highlighting the identity of particularly influential data points. Cook's distance is an attempt to combine leverage and residuals in a single measure.

Note: When data points are singled out as being influential, and When we were happier with other aspects of the model, we would repeat the modelling,leaving out each of these points in turn.

WESTERN SYDNEY UNIVERSITY

PRACTICE PROBLEMS: Cautions about R2

1. The coefficient of determination R^2 and the correlation coefficient r quantify the strength of a linear relationship. It is possible that $R^2 = 0\%$ and $r = 0$, suggesting there is no linear relation between x and y, and yet a perfect curved (or "curvilinear" relationship) exists.
2. A large R^2 value should not be interpreted as meaning that the estimated regression line fits the data well. Another function might better describe the trend in the data.
3. The coefficient of determination R^2 and the correlation coefficient r can both be greatly affected by just one data point (or a few data points).

PRACTICE PROBLEMS: Cautions about R2 ctd . . .

4. Correlation (or association) does not imply causation.
5. Ecological correlations - correlations that are based on rates or averages - tend to overstate the strength of an association.
6. A "statistically significant" R^2 value does not imply that the slope β_1 is meaningfully different from 0.
7. A large R^2 value does not necessarily mean that a useful prediction of the response y_{new}, or estimation of the mean response y_{new}, can be made. It is still possible to get prediction intervals or confidence intervals that are too wide to be useful.

WESTERN SYDNEY UNIVERSITY

Model Assumptions

- Residuals vs. Fits Plot
- Residuals vs. Predictor Plot
- Residuals vs. Order Plot
- A time trend
- Positive serial correlation
- Negative serial correlation
- Normal Probability Plot of Residuals
- Outliers & Influential Points
- Leverages
- Residuals
- Studentized residuals (or internally studentized residuals) [which Minitab calls standardized residuals] Difference in fits (DFITS)
- Cook's distance measure

WESTERN SYDNEY UNIVERSITY

Multicollinearity and other Regression Pitfalls

Extracting information from model objects by name

Examples:

- coef(model)
- summary(model)
- fitted(model)
- resid(model)
- effects(model)
- vcov(model)
- using $ to name the component, e.g. model$resid

Extracting information from model objects by name illustration

```r
vcov(model)
```

```
##               (Intercept)            TV
## (Intercept)  0.209620158 -1.064495e-03
## TV          -0.001064495  7.239367e-06
```

Logistic Regression

Recall the Simple Linear Regression Model $E(y_i) = \alpha + \beta x_i$ for $i = 1, 2, ..., n$

Where the response variable Y, and the predictor variable X are both continuous variables. Now lets assume that the response variable Y is a catogorical variable with 2 outcomes.

Examples: - eye colour ~ {brown, blue } - Hear Disease ~ {present, absent } - Insurance claim ~ {fraudulent, legitimate }.

Note that these Qualitative variables take values in an unordered set C

Can we use Linear Regression when Y is qualitative?

If we are to classify customers according to credit card Default then,

if we code

Y = 0 if Default is No

Y= 1 if Default is Yes

Question? Can we simply perform a linear regression of Y on X and classify as Yes if $\hat{Y}_i > 0.5$?

Limitations of using Linear Regression when Y is binary

Consider Y is Default and X is Balance,

Then simple linear regression model is $\hat{Y}_i = \hat{\alpha} + \hat{\beta} x_i$ for $i = 1, 2, ..., n$

In this case of a binary outcome, linear regression does a good job as a classifier

Why????

$E(\hat{Y}_i) = P(Y = 1)1 + P(Y = 0)0 = P(Y = 1) = P(Default)$

and $E(\hat{Y}_i) = \alpha + \beta x_i$

Therefore $P(Default) = \alpha + \beta x_i = E(\hat{Y}_i)$

However, linear regression might produce probabilities less than zero or bigger than one.

Thus we consider Logistic Regression

WESTERN SYDNEY UNIVERSITY

Logistic Regression NOTE:

I'm separating simple logistic regression, with only one independent variable, from multiple logistic regression, which has more than one independent variable.

Without lumping all logistic regression together, I think it's useful to treat simple logistic regression separately.

Contrast Between Logistic and Linear Regression

In linear regression, the expected value of y_i given x_i is

$E(Y_i) = \alpha + \beta x_i$ for $i = 1, 2, ..., n$

y_i has a normal distribution with standard deviation σ. It is the random component of the model, which has a normal distribution.

$\alpha + \beta x_i$ is the linear predictor.

In logistic regression, the expected value of d_i given x_i is

$E(d_i) = \pi_i = \pi[x_i]$ $logit(E(d_i)) = \alpha + \beta x_i$ for $i = 1, 2, ..., n$ d_i is dichotomous with probability of event $\pi_i = \pi[x_i]$

it is the random component of the model

logit is the link function that relates the expected value of the random component to the linear predictor.

Import the Data Set “Default”

To Install a package to R and call from an installed package use -

install.packages(“ISLR”)

installed.packages(“ISLR”)

Once the Package is installed, to upload the data set Use

```
library(ISLR)
attach(Default)
dim(Default)
```

```
## [1] 10000     4
```

```
View(Default)
summary(Default)
```

```
##  default    student       balance              income
##  No :9667   No :7056   Min.   :    0.0   Min.   :  772
```

Examples of Linear Regression and Logistic Regressin

- balance, income (Continuos quantitative variables)
- default, Student (qualitative binary varables)

Simple Linear Regression

- balance Vs income

Logistic Regression

- default Vs balance
- default Vs income

Import the Data Set "heart"

```
heart <- read.csv("heart.csv")
attach(heart)
dim(heart)
```

```
## [1] 303  15
```

```
names(heart)
```

```
##  [1] "X"          "Age"        "Sex"        "ChestPain"  "RestB
##  [6] "Chol"       "Fbs"        "RestECG"    "MaxHR"      "ExAng
## [11] "Oldpeak"    "Slope"      "Ca"         "Thal"       "AHD"
```

```
head(heart)
```

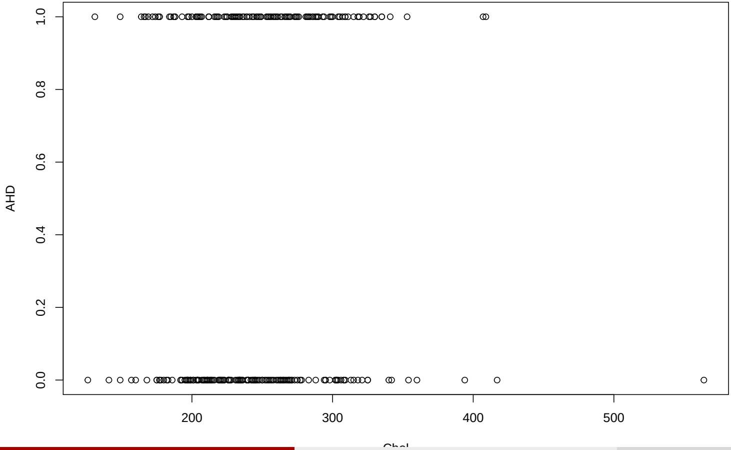
```
##   X Age Sex    ChestPain RestBP Chol Fbs RestECG MaxHR ExAn
## 1 1  63   1      typical    145  233   1       2   150
## 2 2  67   1 asymptomatic    160  286   0       2   108
```

Construct a scatter plot of the data

The following figure shows prevalence of heat disease in a sample of 303 patients as a function of their Cholesterol level. Patience are coded as 1 or 0 depending on whether they are with or without heart disease respectively.

Scatter plot of the hear disease data

```r
plot(AHD~Chol)
```



Simple Linear Regression Plot

We wish to predict prevelance of heart disease from Cholestorol in these patients.

Let $P(x)$ be the probability that a patient with Cholestorol level of x will have heat disease.

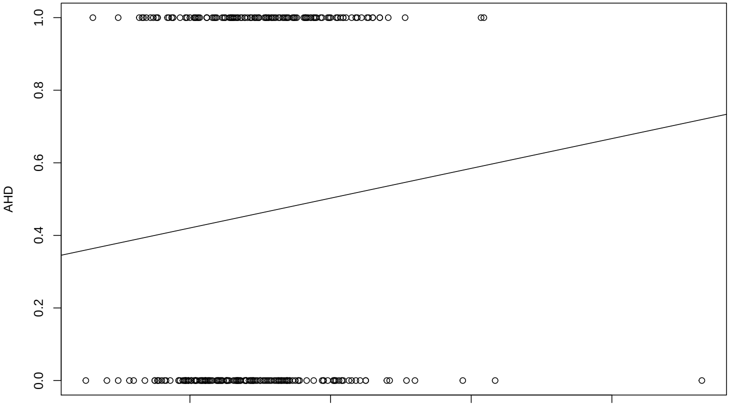
Note that linear regression would not work well here since it could produce probabilities less than zero or greater than one.

```
model2= (lm(AHD~Chol))
coef(model2)
```

```
##  (Intercept)         Chol
## 0.2562205371 0.0008209608
```

Simple Linear Regression Plot ctd. . .

```
plot(AHD~Chol)
abline(a=0.2562205371,0.0008209608)
```



Logistic Regression will be continued in Week 4 with Classification

TEXT BOOK

Lecture notes are based on the textbook.

For further reference refer;

Prescribed Textbook - Chapter 3

– James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An Introduction to Statistical Learning: with Applications in R Springer.

WESTERN SYDNEY UNIVERSITY