Simple Linear Regression Model Building

In this lecture

- Simple linear regression
  - Loading the data from .txt file
  - Plot the data
  - Build linear model
  - Look at summary of the model

Loading data

- Dataset 'bonds' is given in ".txt" format
- To load data from the file the function used is `read.delim( )`

read.delim()

Reads a file in table format and creates a data frame from it

SYNTAX

```
read.delim(file,row.names=1)
```

file	the name of the file which the data are to be read from. Each row of the table appears as one line of the file.
row.names	a vector of row names. This can be a vector giving the actual row names, or a single number giving the column of the table which contains the row names, or character string giving the name of the table column containing the row names.

Loading data

- Assuming that bonds.txt is in your current working directory

```
bonds <- read.delim("bonds.txt", row.names=1)
```

- The data is saved into a data frame 'bonds'

Viewing data

- `View(bonds)` will display the dataframe in a tabular format

Filter

	CouponRate	BidPrice
1	7.000	92.94
2	9.000	101.44
3	7.000	92.66
4	4.125	94.50

- `head(bonds)` and `tail(bonds)` will display the first and last six rows from the dataframe

Description of dataset

- The data has two variables CouponRate and BidPrice.
- CouponRate refers to the fixed interest rate that the issuer pays to the lender.
- BidPrice is the price someone is willing to pay for the bond.

Structure of the data

- Each variable and its data type
- `str()`- input is dataframe
- See whether each of the variable datatypes are same as you expect them to be
- If not coerce

```
> str(bonds)
'data.frame':   35 obs. of  2 variables:
 $ CouponRate: num  7 9 7 4.12 13.12 ...
 $ BidPrice  : num  92.9 101.4 92.7 94.5
```

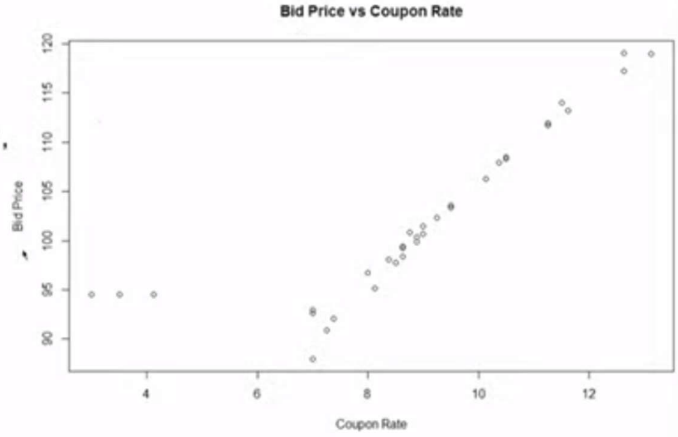
Summary of the data

- Gives mean and five number summary

```
> summary(bonds)
   CouponRate        BidPrice
 Min.   : 3.000   Min.   : 88.00
 1st Qu.: 8.062   1st Qu.: 95.95
 Median : 8.875   Median :100.38
 Mean   : 8.921   Mean   :102.14
 3rd Qu.:10.438   3rd Qu.:108.11
 Max.   :13.125   Max.   :119.06
```

Plotting the data

```
plot(bonds$CouponRate,bonds$BidPrice,
     main = "Bid Price vs Coupon Rate",
     xlab = "Coupon Rate",
     ylab = "Bid Price")
```



Building linear regression model

- Building linear model using the function `lm()`
- Syntax: `lm(formula,data)`

`lm(dependent var~independent var)`

```
bondsmod <- lm(bonds$BidPrice~bonds$CouponRate)
```

or

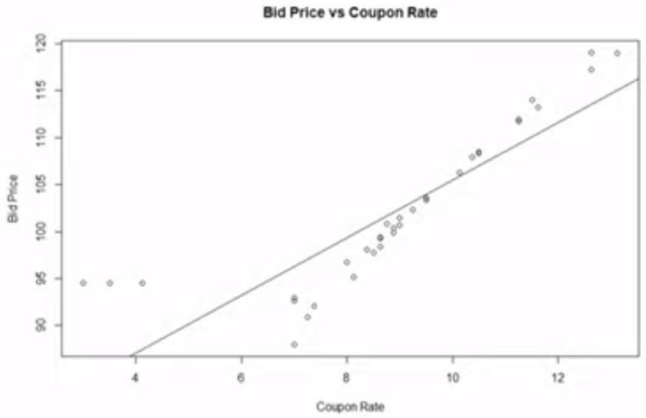
```
bondsmod <- lm(BidPrice~CouponRate,data = bonds)
```

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \epsilon_i$$

Intercept Slope

Fitting the regression line over the plot

```
plot(bonds$CouponRate,bonds$BidPrice,
     main = "Bid Price vs Coupon Rate",
     xlab = "Coupon Rate",
     ylab = "Bid Price")
abline(bondsmod)
```



Model summary

```
bondsmod <- lm(BidPrice~CouponRate,data = bonds)
summary(bondsmod)

Call:
lm(formula = BidPrice ~ CouponRate, data = bonds)

Residuals:
   Min     1Q Median     3Q    Max
-8.249 -2.470 -0.838  2.550 10.515

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  74.7866     2.8267  26.458  < 2e-16 ***
CouponRate    3.0661     0.3068   9.994 1.64e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.175 on 33 degrees of freedom
Multiple R-squared:  0.7516,    Adjusted R-squared:  0.7441
F-statistic: 99.87 on 1 and 33 DF,  p-value: 1.645e-11
```

Data science for Engineers

Model summary


```
Call:
lm(formula = BidPrice ~ CouponRate, data = bonds)

Residuals:
   Min     1Q Median     3Q    Max
-8.249 -2.470 -0.838  2.550 10.515

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  74.7866     2.8267  26.458  < 2e-16 ***
CouponRate    3.0661     0.3068   9.994 1.64e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.175 on 33 degrees of freedom
Multiple R-squared:  0.7516,    Adjusted R-squared:  0.7441
F-statistic: 99.87 on 1 and 33 DF,  p-value: 1.645e-11
```

Model summary

```
Call:
lm(formula = BidPrice ~ CouponRate, data = bonds)

Residuals:
    Min      1Q  Median      3Q     Max
-8.249 -2.470 -0.838   2.550 10.515

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  74.7866     2.8267  26.458  < 2e-16 ***
CouponRate    3.0661     0.3068   9.994 1.64e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.175 on 33 degrees of freedom
Multiple R-squared:  0.7516,    Adjusted R-squared:  0.7441
F-statistic: 99.87 on 1 and 33 DF,  p-value: 1.645e-11
```

Difference between observed and predicted/fitted values

Five number summary of residuals

Estimates of slope and intercept parameter

$\hat{\beta}_0$ $\hat{\beta}_1$

$s_{\hat{\beta}_0}$ $s_{\hat{\beta}_1}$

Estimated standard deviation for the slope and intercept

Model summary

```
Call:
lm(formula = BidPrice ~ CouponRate, data = bonds)

Residuals:
    Min      1Q  Median      3Q     Max
-8.249 -2.470 -0.838   2.550 10.515

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  74.7866     2.8267  26.458  < 2e-16 ***
CouponRate    3.0661     0.3068   9.994 1.64e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.175 on 33 degrees of freedom
Multiple R-squared:  0.7516,    Adjusted R-squared:  0.7441
F-statistic: 99.87 on 1 and 33 DF,  p-value: 1.645e-11
```

Probability of seeing a 't' random variable will be greater than the observed t-statistic

Significance level above which the null hypothesis will be rejected

Ratio of estimate by the standard error. Used as a criterion for hypothesis testing

F statistic to test whether slope=0