

A Brief Introduction to R*

Also see Appendix B of *Linear Models with R*

```
> 1+1
[1] 2
> 2^3 # Two to the power 3
[1] 8

> 1:30
 [1]  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25
[26] 26 27 28 29 30

> gamma(.5)^2      # Gamma(1/2) = Sqrt(Pi)
[1] 3.141593

> x = 1           # Assigns the value 1 to x
> y = 2           # Assigns the value 2 to y
> x+y
[1] 3
> z = x+y
> z
[1] 3
> x = c(1,2,3,4,5,6) # Collect these numbers; x is now a vector

> z # No dynamic updating; it's not a spreadsheet
[1] 3
> x+y
[1] 3 4 5 6 7 8

> y = 1 + 2*x
> cbind(x,y)
      x y
[1,] 1 3
[2,] 2 5
[3,] 3 7
[4,] 4 9
[5,] 5 11
[6,] 6 13

> z = y[x>4]      # z gets y such that x > 4
> z
[1] 11 13
> z2 = subset(y,x>4); z2
[1] 11 13

> # If you put an array of integers inside the brackets, you get those
> # elements, in the order indicated.

> y[c(6,5,4,3,2,1)] # y in opposite order
[1] 13 11 9 7 5 3
> y[c(2,2,2,3,4)] # Repeats are okay
[1] 5 5 5 7 9
> y[7] # There is no seventh element. NA is the missing value code
[1] NA
```

* Copyright information is on the last page

```

> # Computing probabilities, etc.
>
> pnorm(0) # Area less than zero for a standard normal
[1] 0.5
>
> pnorm(160,mean=100,sd=15) # IQ of 160
[1] 0.9999683
>
> pcauchy(4)
[1] 0.9220209
>
> dnorm(0) # height of the curve
[1] 0.3989423
>
> dpois(0,lambda=3) # P(Y=0) for Y ~ Poisson(3)
[1] 0.04978707
>
> qnorm(0.975) # z value with P(Z<z) = 0.975
[1] 1.959964
>
> qf(0.975,df1=6,df2=122) # Critical value for F, not in any table
[1] 2.513606
>
> CriticalValue = qchisq(0.95,df=1:8)
> df=1:8; cbind(df,CriticalValue)
      df CriticalValue
[1,]  1      3.841459
[2,]  2      5.991465
[3,]  3      7.814728
[4,]  4      9.487729
[5,]  5     11.070498
[6,]  6     12.591587
[7,]  7     14.067140
[8,]  8     15.507313

```

Before the beginning of the Fall term, students in a first-year Calculus class took a diagnostic test with two parts: Pre-calculus and Calculus. Their High School Calculus marks and their marks in University Calculus were also available. In order, the variables in the data file are: Identification code, Mark in High School Calculus, Score on the Pre-calculus portion of the diagnostic test, Score on the Calculus portion of the diagnostic test, and mark in University Calculus. Thanks to Dr. Cleo Boyd for permission to use these data.

1	65	2	0	39
2	54	6	2	57
3	77	4	4	62
4	80	5	2	76
5	87	4	4	86
6	53	3	1	60
.
.
.
290	83	4	3	56
291	81	6	3	70
292	73	5	9	60
293	80	5	2	50
294	56	4	2	50
295	80	6	1	61

```

> math = read.table("http://www.utstat.toronto.edu/~brunner/data/legal/mathtest.txt")
> head(math)
  V1 V2 V3 V4 V5
1  1 65  2  0 39
2  2 54  6  2 57
3  3 77  4  4 62
4  4 80  5  2 76
5  5 87  4  4 86
6  6 53  3  1 60
> colnames(math) = c("ID", "HScalcMark", "PreCalcScore", "CalcScore", "UnivCalcMark")
> head(math)
  ID HScalcMark PreCalcScore CalcScore UnivCalcMark
1  1          65            2         0           39
2  2          54            6         2           57
3  3          77            4         4           62
4  4          80            5         2           76
5  5          87            4         4           86
6  6          53            3         1           60
> summary(math)
  ID HScalcMark PreCalcScore CalcScore UnivCalcMark
1  1          65            2         0           39
2  2          54            6         2           57
3  3          77            4         4           62
4  4          80            5         2           76
5  5          87            4         4           86
6  6          53            3         1           60
> mean(math$HScalcMark)
[1] 78.50847
> mean(HScalcMark)
Error in mean(HScalcMark) : object 'HScalcMark' not found
> attach(math)
> mean(HScalcMark)
[1] 78.50847

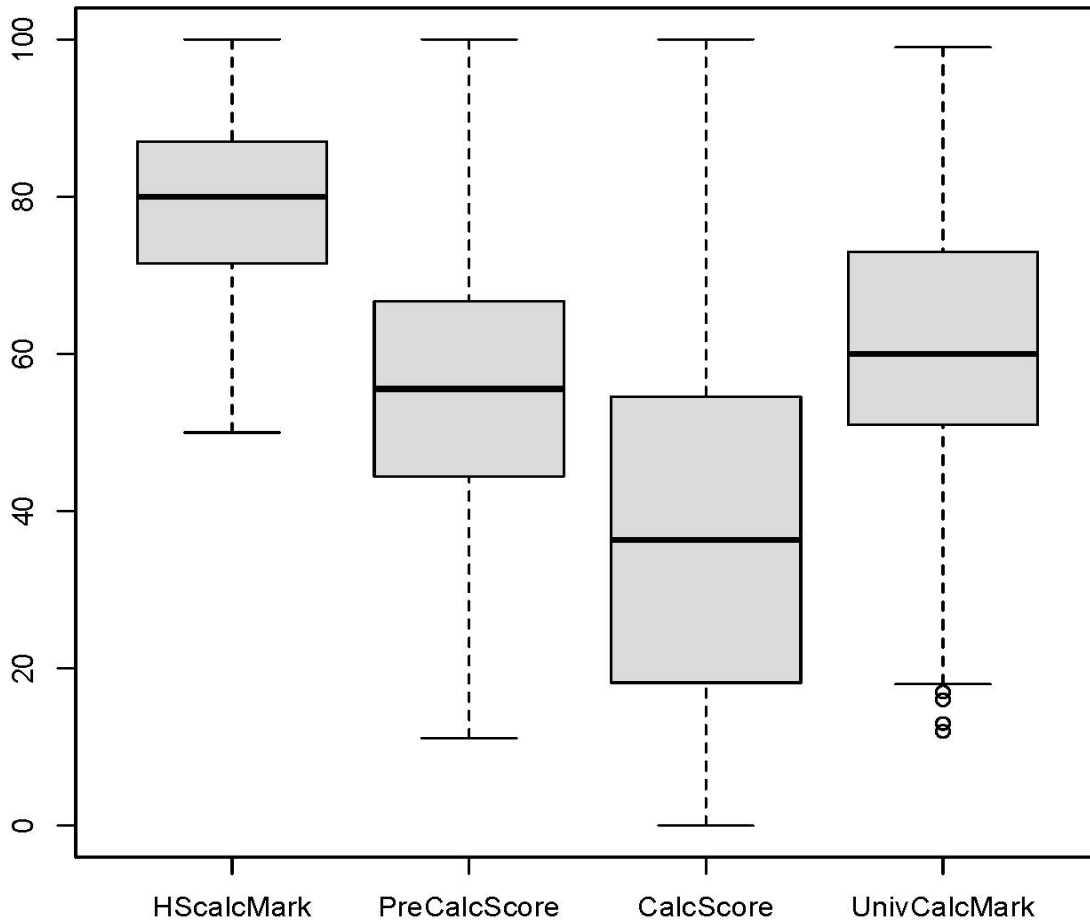
```

```

> # PreCalc score is out of 9 and Calc score is out of 11. Convert to percentages.
> PreCalcScore = 100 * PreCalcScore/9
> CalcScore = 100 * CalcScore/11
> ##### Descriptive stats #####
> datamat = cbind(HScalcMark,PreCalcScore,CalcScore,UnivCalcMark)
> boxplot(datamat); title("Box plots of the math data")

```

Box plots of the math data



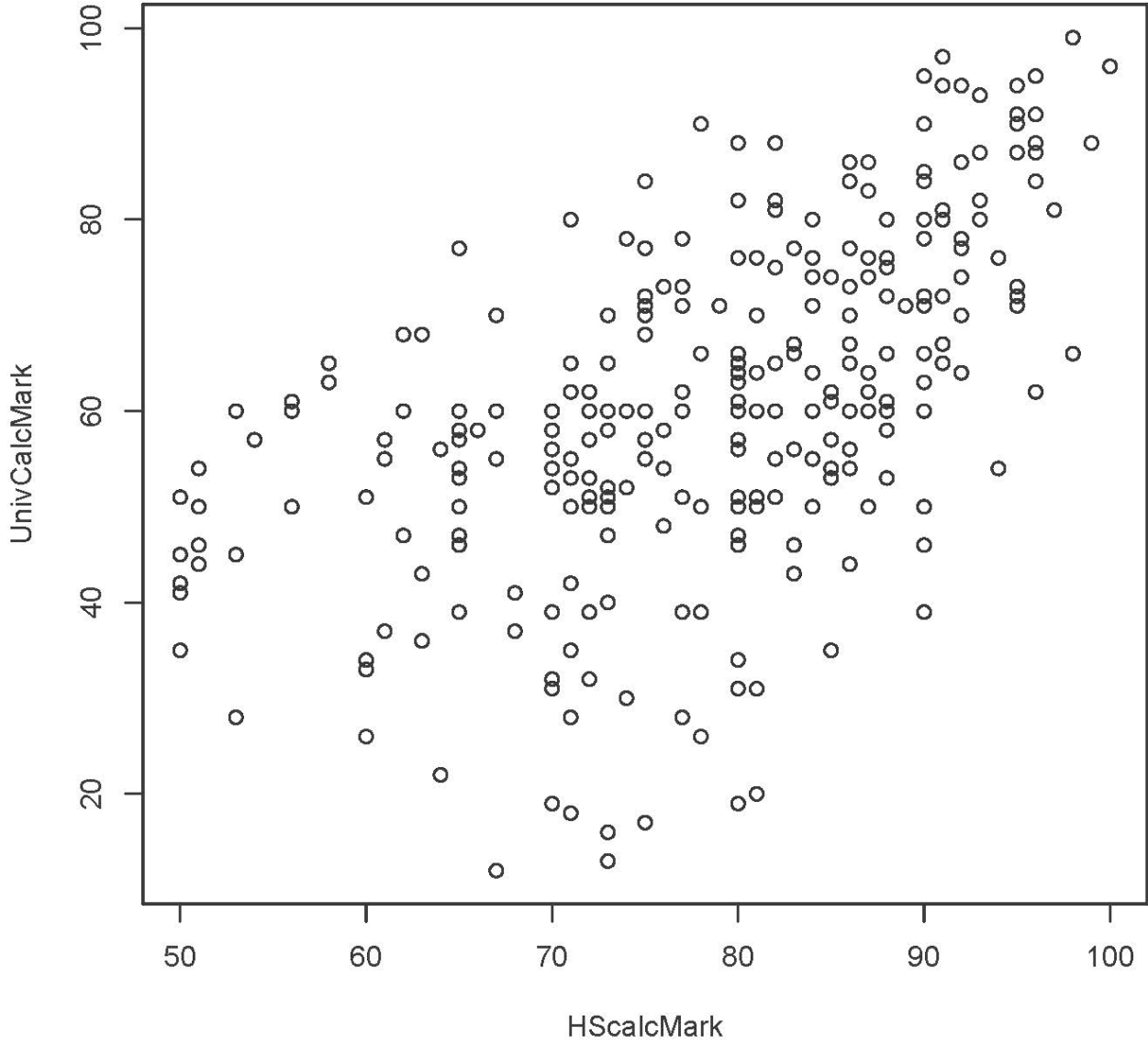
```

> # upper whisker = min(max(x), Q_3 + 1.5 * IQR)
> # lower whisker = max(min(x), Q_1 - 1.5 * IQR)
> summary(datamat)

```

HScalcMark	PreCalcScore	CalcScore	UnivCalcMark
Min. : 50.00	Min. : 11.11	Min. : 0.00	Min. : 12.00
1st Qu.: 71.50	1st Qu.: 44.44	1st Qu.: 18.18	1st Qu.: 51.00
Median : 80.00	Median : 55.56	Median : 36.36	Median : 60.00
Mean : 78.51	Mean : 53.75	Mean : 36.02	Mean : 60.91
3rd Qu.: 87.00	3rd Qu.: 66.67	3rd Qu.: 54.55	3rd Qu.: 73.00
Max. : 100.00	Max. : 100.00	Max. : 100.00	Max. : 99.00

```
>  
> plot(HScalcMark,UnivCalcMark)  
>
```

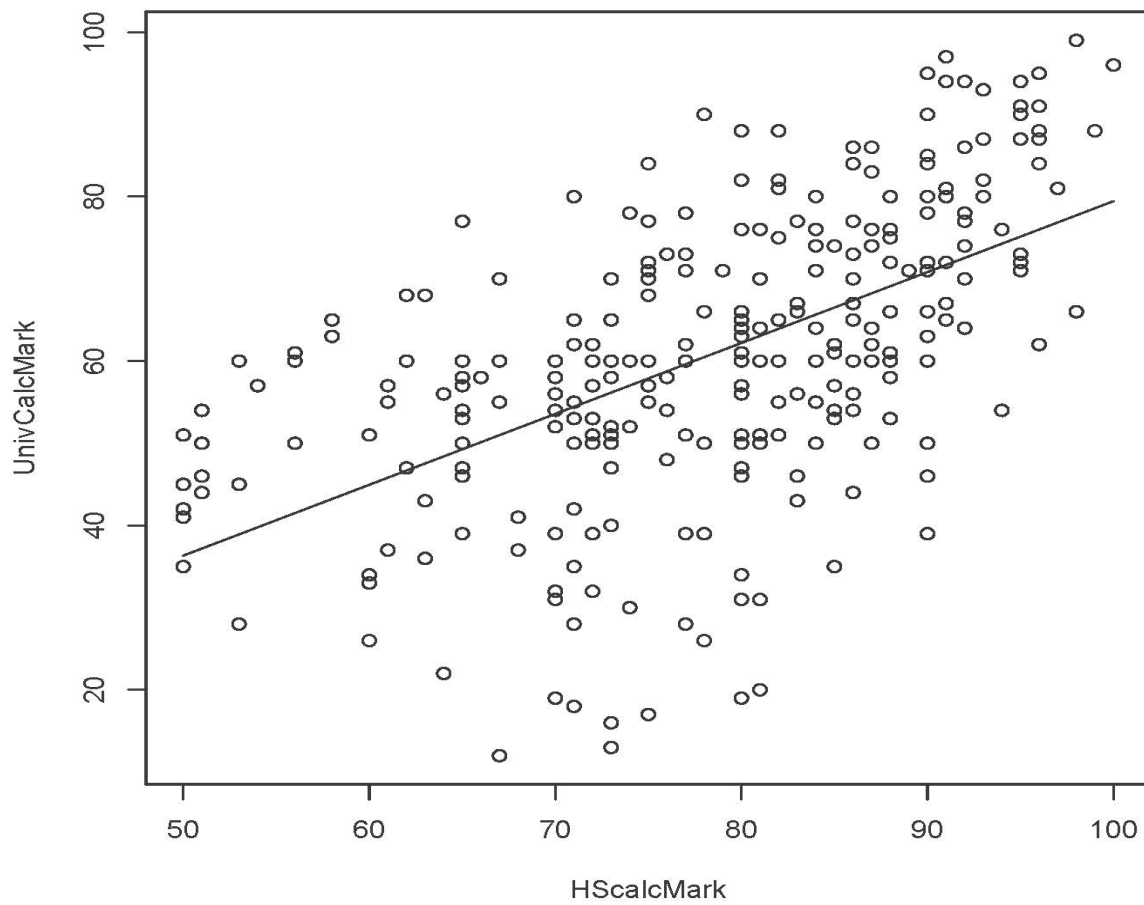


```

> # Add the least squares line to the plot.
> lm(UnivCalcMark ~ HScalcMark) # This means  $U_i = \beta_0 + \beta_1 H_i + \epsilon_i$ 
Call:
lm(formula = UnivCalcMark ~ HScalcMark)
Coefficients:
(Intercept)    HScalcMark
   -6.7891         0.8623

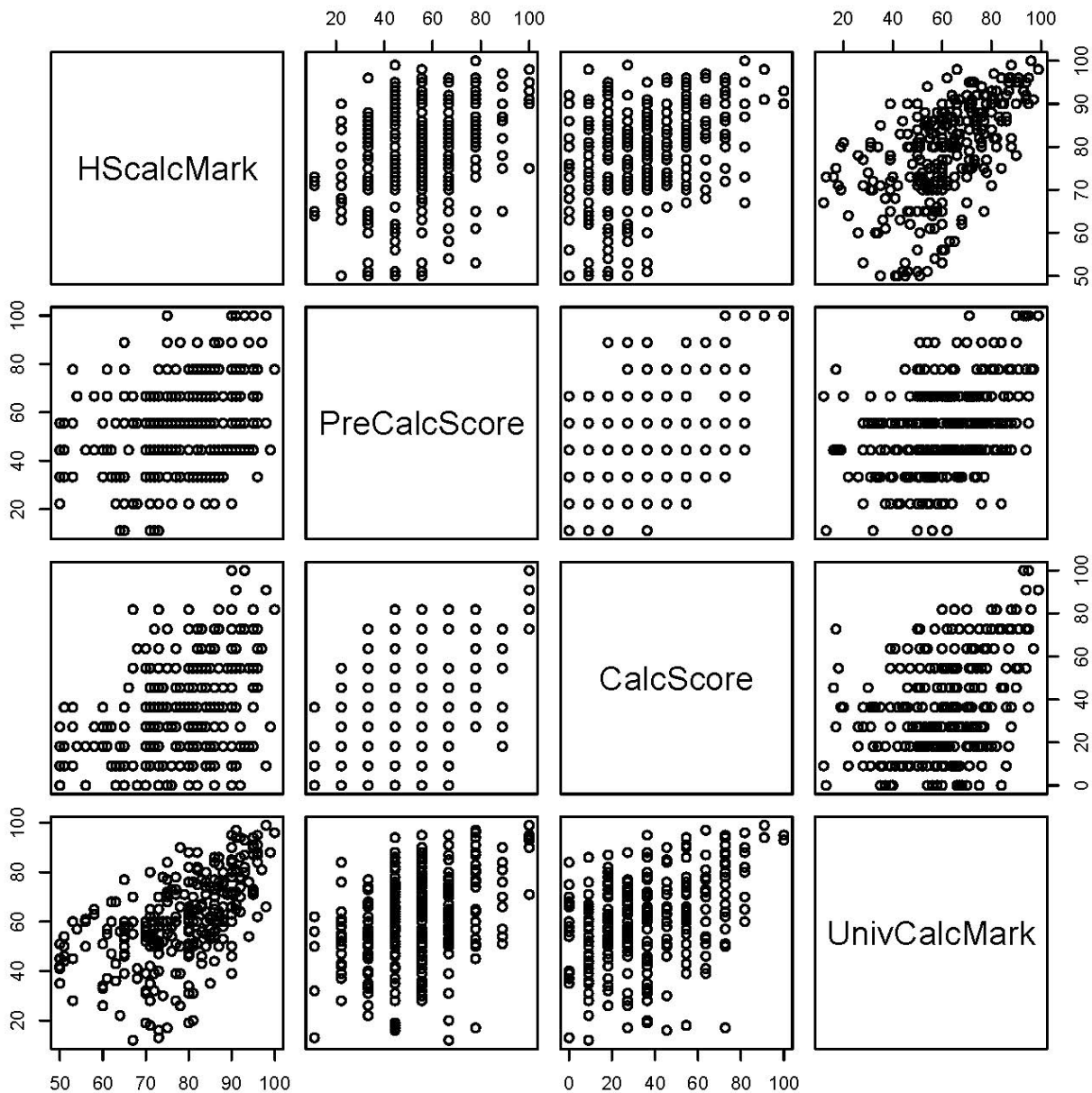
> # The intercept of the least squares line is -6.7891, and the slope is 0.8623.
> x = c(50,100); y = -6.7891 + 0.8623*x
> cbind(x,y)
      x      y
[1,] 50 36.3259
[2,] 100 79.4409
>
> plot(HScalcMark,UnivCalcMark)
> lines(x,y,lty=1) # Line Type 1 is a solid line.

```



```
> cor(datamat)
      HScalcMark PreCalcScore CalcScore UnivCalcMark
HScalcMark 1.0000000 0.3271197 0.4132186 0.5548571
PreCalcScore 0.3271197 1.0000000 0.4573164 0.3736003
CalcScore 0.4132186 0.4573164 1.0000000 0.3793398
UnivCalcMark 0.5548571 0.3736003 0.3793398 1.0000000
```

```
> pairs(datamat) # Matrix of scatterplots
```



This document was prepared by [Jerry Brunner](#), University of Toronto. It is licensed under a Creative Commons Attribution - ShareAlike 3.0 Unported License:

http://creativecommons.org/licenses/by-sa/3.0/deed.en_US. Use any part of it as you like and share the result freely. The Open Office document is available from the course website at <http://www.utstat.toronto.edu/~brunner/oldclass/302f20>