Linear Regression and Correlation in R Commander

1. Correlation Coefficient (r)

Once you have imported your dataset into R, use the following commands to calculate the correlation coefficient between two variables in a bivariate data set:

Statistics | Summaries | Correlation Matrix…

In the resulting dialog box, choose the two variables in your data set that you want to calculate the correlation coefficient for by clicking the variable names with your mouse (hold down the Ctrl key on your computer keyboard while clicking with the mouse to select the *second* variable). Leave the setting on the default correlation coefficient (Pearson product-moment), then click OK.

Each subject in a group took two word memory tests. For each subject, the score on memory test 1 and memory test 2 were recorded. Here is a portion of the data.

```
test1 test2
6     6
3     7
8     5
8     9
:     :
```

Running the above commands on this data set results in the following output to the output window:

```
> cor(Dataset[,c("test1","test2")], use="complete.obs")
       test1     test2
test1 1.0000000 0.3915253
test2 0.3915253 1.0000000
```

Hence, the correlation coefficient between a subject's scores on memory test 1 and memory test 2 is 0.3915253.

2. Fitting Least-squares Regression Line

To fit a regression line, select **Statistics | Fit models | Linear regression…** In the resulting dialog box, select the y variable as the "response variable" and the x variable as the "explanatory variable." Then click OK.

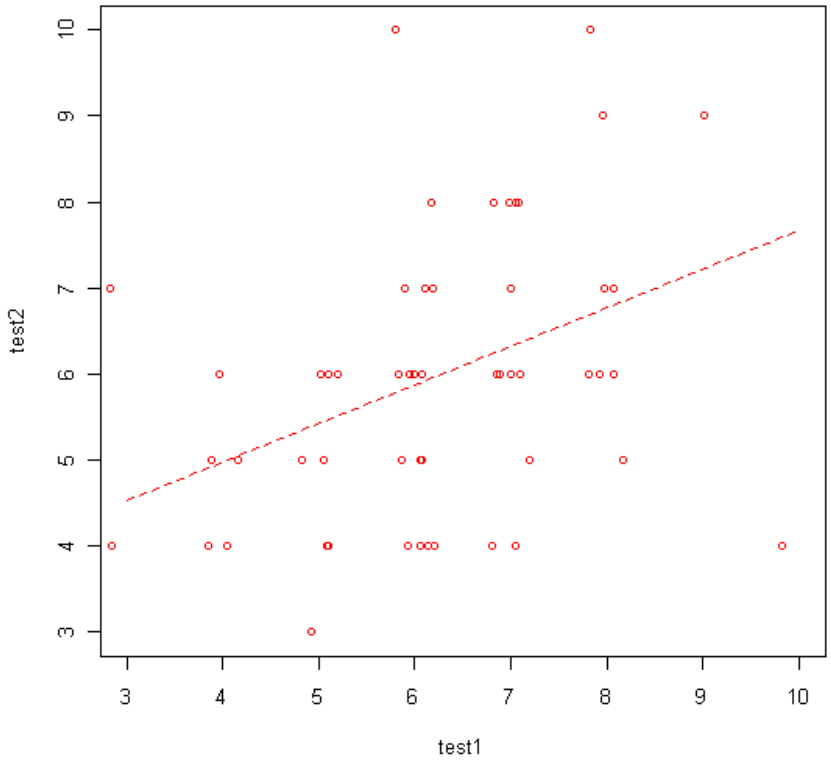
For the sample data set, the output window displays the following text:

```
> RegModel.2 <- lm(test2~test1, data=Dataset)
> summary(RegModel.2)

Call:
lm(formula = test2 ~ test1, data = Dataset)

Residuals:
     Min       1Q   Median       3Q      Max
-3.66271 -0.91880  0.02998  1.08120  4.13242

Coefficients:
```

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.1749     0.9303   3.413  0.00124 **
test1         0.4488     0.1449   3.098  0.00312 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.563 on 53 degrees of freedom
Multiple R-squared: 0.1533,   Adjusted R-squared: 0.1373
F-statistic: 9.595 on 1 and 53 DF,  p-value: 0.003117
```

The column titled "Estimate" gives the y-intercept and slope. The y-intercept is labeled "(Intercept)" and equals 3.1749. The slope is labeled "test1" since it is the coefficient of the x variable (test1) and equals 0.4488. Thus, the regression line is $y = 3.1749 + 0.4488x$. The correlation coefficient is the square root of "Multiple R-squared." So, $r = \sqrt{0.1533} = 0.3915$, which is what we arrived at using the correlation matrix command.

3. You can graph a scatterplot. If you check the option to "jitter" the x-variables, points that coincide will be moved slightly in the horizontal direction so they are all visible. Here is the scatterplot with the regression line:



4. Exercise: Use the regression line to predict the score on test 2 for a subject who go an 11 on test 1.

5. Exercise: Interpret the slope of the regression line in terms of the variables in the data set.

6. **Important caution:** Correlation does NOT imply cause and effect. Consider data x = number of TV's per household, y = life expectancy for 100 countries which has r = 0.80 (so the more TV's per hh, the longer the life expenctancy). Do TV's cause people to live longer???