Homework # 12 (correlation and regression)

Do *NOT* use R for problems 1 - 5:

1) You compare the height (cm) and weight (kg) of 5 adult women. You get the following results:

height	153.6	165.9	169.7	162.7	159.0
weight	54.1	62.6	67.7	60.2	59.6

(a) Construct a scatter plot of height (x) vs. weight (y). (Don't just "sketch", be a little careful (see also problem 3(c))).

(b) Calculate the following (*show* your work for SS_{cp}):

SS_x SS_y SS_{cp} $\bar{x}$ $\bar{y}$

(c) Calculate the correlation coefficient (r)

2) Use the information from problem 1. Perform a complete test of the hypothesis that the population correlation coefficient (ρ) is 0. Show all steps (note - obviously this should be a one sided test! *Make sure you know why!*).

3) Now let's assume you wanted to predict weights from heights. In other words, now let's use the same data from problem (1) and do a regression instead.

(a) Calculate b_0 and b_1.

(b) Give the equation for the least squares regression line.

(c) Carefully draw your least square regression line on the plot you made in 1(a). (Don't just "sketch", be a little careful).

4) Let's continue working with these data:

(a) Now do a significance test of H_0: $\beta_1 = 0$. Show all your calculations (including your residual calculations). Again, note that this should be a one sided test (*why?*).

(b) Compare your t^* from 4(a) with your t^* from problem 2. Are they the same? This is not a coincidence, although once you do more complicated analyses, you can't rely on this "equivalence".

5) Continuing with this data set:

(a) Create a residual plot (by hand) for the regression in problem (4) and interpret. Are there any serious problems?

(b) Calculate R^2 and interpret.

You MUST use R for problems 6 - 8:

6) Now let's do some R. First we'll explore our Irises a bit more. Let's extract the data we want. This time we'll use all 50 values for sepal length and petal length for *Iris versicolor* (the middle of the three species in the data set - sepals are the "leaves" that surround the petals before a flower opens).

Without much explanation*, let's put the data into two variables called `slength` and `plength`:

```
slength <- iris$Sepal.Length[51:100]
plength <- iris$Petal.Length[51:100]
```

* in brief, these commands pull out the middle 50 values for sepal length and petal length from the built in iris data set. Type "`iris`" at the command prompt to see the data set, and you should be able to figure out how these commands work.

(a) Now perform a correlation test of sepal length vs. petal length. Incidentally, should this test be one sided? *Why or why not?*

(b) Now perform the correlation test again, this time do petal length vs. sepal length. The results should be identical. ***Why??***

(c) Create a scatterplot of the data. Why do you think the graph has such an odd appearance?

7) You investigate the relationship between dbh (diameter (cm) at breast height) and height (m) of oak trees. You get the following results:

dbh (cm):	40	57	39	13	34	46	26	14	20	29	38	31	60	11	18	48	43	44	51	49
height(m):	13.4	13.4	8.6	6.7	11.1	9.3	9.5	9.6	6.9	11.2	12.3	9.4	13.3	6.8	7.6	12.3	10.8	13.4	14.1	13.0

Read the data into R.

(a) Calculate the following using R: $\bar{x}$, $\bar{y}$, SS_x, SS_y.

(Note that to get SS_y or SS_x, you can just ask R for the variance (`var`) or standard deviation, and then do the appropriate calculation).

(b) Perform a complete test of the hypothesis that there is no difference in height as dbh increases. *Write out all the appropriate steps of a regular hypothesis test* (give H_0, H_1, α, your decision, etc.)

(c) Give the equation of the least squares line *(Write it out, don't just hand in a printout!!)*

8) Finally do the following:

(a) create a scatter plot and residual plot for the analysis you did in (**6**) and comment on the residual plot (is it okay or do you see any problems?).

(b) create a *q-q* plot of the residuals and comment on it.

(c) Write down the value of R^2, and ***interpret*** it (**do *not* use "adjusted" R^2**).

Be prepared to discuss these problems in recitation the week of April 28th.

Computer notes (R instructions):

1) For regression:

Make sure you have your data in two columns (in other words, two variables).

Although you don't need to name your regression, it will be a lot easier if you do. So give your regression a name as follows (In this example, I've named it “prob6” (so maybe it's the regression you're doing for problem 6)).

```
prob6 <- lm(y ~ x)
```

Note the “~” symbol. It is on your keyboard, but you may have to look a bit (try the upper left or near the space bar)

Now type:

```
summary(name-of-your-regression)
```

Of course, you'll use the right name for your regression (e.g., “prob6”). You will get a printout that looks a bit like this:

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  10.7778     1.4265   7.555 6.57e-05 ***
height       -0.9537     0.2842  -3.356  0.00999 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.321 on 8 degrees of freedom
Multiple R-squared: 0.5847, Adjusted R-squared: 0.5328
F-statistic: 11.26 on 1 and 8 DF,  p-value: 0.009989
```

Now you need to interpret this result:

*The important bits are highlighted in **bold** above.*

The “Estimate” column for the row labeled (Intercept) is the value of b_0 (the intercept)

The “Estimate” column for the row labeled with your variable name is the value of b_1 (the slope)

The probability (last column) in the row labeled with your variable name is the p-value that tells you if the regression was significant.

IMPORTANT! R will automatically give you a two sided p-value. How do you get a one sided p-value? Divide this p-value by 2. See below for an example.

Note that R will also print the R^2 value (it may be labeled as “multiple” (don't use “adjusted” R^2)).

So we see that the intercept is 10.778, and the slope is -0.9537

You should arrange this into a regression equation:

$$\hat{Y} = 10.778 - 0.9537\, X$$

Note that the given *p*-value is 0.00999. ***If the test is one sided, then you need to do:***

0.00999/2 = .004995 (in either case, the test would be significant at $\alpha = 0.01$)

Finally, notice that the R^2 is 0.5845. If everything else is okay (e.g. residual plots), we can say that X explains 58.45% of the variation in Y.

2) To get your scatterplot (which should be part of any regression or correlation), do:

```
plot(x,y)
```

Make sure you don't have x and y backwards, or your axes will be wrong.

Now add your regression line (if you're doing regression) by doing:

```
abline(name-of-your-regression)
```

Again, make sure that "name-of-your-regression" is the actual name of your regression (e.g., "prob6" in the example above).

(The "`abline`" command essentially draws a line with the given intercept and slope. Type "`name-of-your-regression`" (without the "`summary`") to see how it works.

3) To get your residual plot do:

```
plot(x,name-of-your-regression$residuals)
abline(0,0)
```

This will give you a residual plot as well as a line as a reference.

4) For correlation:

This is pretty simple. Make sure you have your two data variables, then do:

```
cor.test(x,y)
```

or for a one sided correlation:

```
cor.test(x,y,alternative = "greater")
```

(or, of course, you can use `"less"`)

The results should be pretty straight forward. You're given a *p*-value, the actual correlation estimate, as well as a number of other statistics. You should be able to figure it out.