Homework # 12 (correlation and regression)

As usual, 3rd edition references in **[bold]**, 4th edition reference in *{italics}*.

{Also, due to serious silliness in the 4th edition, please read the problems carefully}.

1) Exercise 12.25, p. 538 **[12.32, p. 564]**, *{Once again, not in the 4th edition, so reproduced below}*

2nd and **[3rd]** edition:

a & b only.

{4th edition}:

A veterinary anatomist measured the density of nerve cells at specified sites in the intestine of nine horses. Each density value is the average of counts of nerve cells in five equal sections of tissue. The results are given in the accompanying table for site I (midregion of jejunum) and site II (mesenteric region of jejunum). *{Note: the data are given in table 8.5.1, on page 322, so I did not reproduce them here. The difference column is obviously irrelevant for this problem (it's used for something else)}*

$SS_x = 1419.82$ $SS_y = 853.396$ $SS_{cp} = 893.689$

(a) Calculate the correlation coefficient between the densities at the two sites.
(b) Construct a scatterplot of the data.

2) Exercise 12.26, p. 538 **[12.33, p. 565]** *{Again, not in the 4th edition... so reproduced below}* (you'll need the (modified) answers from # 1)

2nd and **[3rd]** edition:

Do the problem as stated.

{4th edition}:

Refer to Exercise (1). Test the hypothesis that the true (population) correlation coefficient is zero against the alternative that it is positive. Let $\alpha = 0.05$.

3) Exercise 12.5, p. 512 **[12.5, p. 537 (caution - in 3rd ed. problem continues on to the next page)]** *{4th edition see below}*

2nd and **[3rd]** edition:

Do the problem as stated, but don't bother with (c) or (e).

{4th edition: see problem 12.2.5, p. 491, then note:}

$SS_x = 20{,}209.0$ $SS_y = 11{,}831.8$ $SS_{cp} = -14{,}563.1$ $SS_{resid} = 1{,}337.3$

(a) Calculate the linear regression of Y on X
(b) Plot the data and draw the regression line on your graph
(d) Calculate s_y and $s_{y|x}$ and specify the units of each

4) Exercise 12.22, p. 538 **[12.28, p. 564]** *{12.5.3}*. All editions: do not do the problem as stated. Instead, do the following (basically you can just use the data from (3) and then follow my instructions):

Read the data into R (see problem 3 for the data).

(a) Calculate the regression of Y on X using R (verify what you got in problem (3)).

(b) Perform a complete hypothesis test of the regression of of Y on X using R.

(note that (a) and (b) are basically done in the same step in R)

(c) Create a residual plot using R *and comment on it.*

Note: don't just hand in a printout. Write out they hypotheses, give the value for α you want to use, write out your conclusions, etc.

5) Exercise 12.2, p. 511 **[12.2, p. 536]** *{Again, not in 4th edition, so see below...}*

2nd and **[3rd]** editions:

Also do:

d) perform a complete hypothesis test to see if the regression line is significant.

{4th edition:}

Here are some data:

						mean
X:	3	7	6	7	2	5
Y:	10	2	9	4	15	8

(a) Compute the linear regression of Y on X and compute $\hat{y}$ for each data point.
(b) Plot the data and also the values of $\hat{y}$.
(c) compute the residual SS.

Then do (d) as above for the 2nd and 3rd edition

6) Exercise 12.47, p. 564 **[12.58, p. 592]** *{12.S.15}*. Same for all editions. It shouldn't take you too long to type this into R. Yes, use R for this problem, or you'll be at it way too long ***(seriously!)***). *Show all plots and calculations.*

In addition, do residual plots *(they'll really let you know what's happening).*

Make sure you actually answer the questions!

BIOL 214: Be prepared to discuss these problems in recitation Monday, July 22nd.

BIOL 312: Problems are due on Tuesday, July 23rd, as specified by your lab instructor.

(Computer notes start on next page)

Computer notes (R instructions):

Please note that there are no regression instructions posted under R-notes yet. But this should be enough for you to finish the problems.

Using the command line:

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 "prob9" (so maybe it's the regression you're doing for problem 9)).

prob9 <- 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., "prob9"). 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:

Look for the last bit of the results. 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.

- 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, the slope is -0.9537, the p-value is 0.00999 (so the regression is significant of $\alpha < .01$), and finally, the R^2 is 0.5845

You should arrange this into a regression equation:

$$Y = 10.778 - 0.9537 X$$

2) To get your scatterplot (which is part of your regression plot), 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 by doing:

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

Again, make sure that "name-of-your-regression" is the actual name of your regression (e.g., "prob9" in the example above).

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)
```

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.

Using R-commander:

1) To do just a regression:

a) "Statistics" --> "Fit models" --> "Linear regression"

b) Pick your y variable as the "response" and your x variable as "explanatory".

(incidentally, note the model name at the top - you may need it again)

c) Click OK

d) Now you need to interpret the results (see lower down):

Look for the last bit of the results.

- 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.

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

Sample printout from a R regression problem:

```
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
```

So we see that the intercept is 10.778, the slope is -0.9537, the p-value is 0.00999 (so the regression is significant of $\alpha < .01$), and finally, the R^2 is 0.5845

You should arrange this into a regression equation:

$$Y = 10.778 - 0.9537 X$$

2) To get a fitted regression plot:

a) "Graphs" --> "Scatterplot"

b) pick the appropriate x and y variables, and make sure that only the "Least-squares line" is checked under the "Options" column.

c) Click OK

3) To get your residual plots:

a) verify that the correct model name is indicated next to "Model: *some model name*" (it's on the far left, the line just below the menu bar). If "*some model name*" doesn't match what you had in 1) b), then it won't work or give you the wrong result. If it doesn't match, click on "*some model name*" and select the right one.

b) “Models” --> “Graphs” --> “Basic diagnostic plots”

The plots you need are the first two (residuals vs. fitted and normal q-q). *Ignore the other two plots.*

c) If it doesn't work, make sure you did (a) correctly.

4) To get a correlation (and do a correlation test):

a) Statistics --> Summaries --> Correlation test

b) select both your variables (the variables you want to get the correlation for).

c) make sure the “Pearson product-moment” box is checked, then click OK

d) the t-value and probability should be pretty obvious (near the top of the printout). The estimate for the correlation coefficient (r) is near the bottom of the printout and should also be pretty obvious. R will even print a CI if you want it.