Lab Exercise 3 - Relationships Between Variables

The last two pages are an answer sheet as usual. Carefully remove it, put your name on it, and hand that in on Thursday (31st).

In this week's lab exercise we will continue our examination of relationships between variables. I will do the R for you. You will be interpreting the results. Although, if you want some advice, I think it would be a very good idea if you followed along in R on your own computer or in a computer lab somewhere. (Not required, but recommended, and you're going to have to do it eventually.)

We will be using a dataset called wages.txt, which is posted at the website. Go there, click on the link, and read all about it. Here is a summary.

```r
> ls()
[1] "MJ" "TR"
> WG = read.table(header=T, file="http://ww2.coastal.edu/kingw/psyc480/data/wages.txt")
> dim(WG)
[1] 534 11
> summary(WG)

<table>
<thead>
<tr>
<th></th>
<th>Min.</th>
<th>1st Qu.</th>
<th>Median</th>
<th>Mean</th>
<th>3rd Qu.</th>
<th>Max.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Educ</td>
<td>2.00</td>
<td>12.00</td>
<td>12.00</td>
<td>13.02</td>
<td>15.00</td>
<td>18.00</td>
</tr>
<tr>
<td>South</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.2921</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td>Sex</td>
<td>0.0000</td>
<td>1.0000</td>
<td>0.0000</td>
<td>0.4588</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td>Exper</td>
<td>0.0000</td>
<td>8.0000</td>
<td>15.00</td>
<td>17.82</td>
<td>26.00</td>
<td>55.00</td>
</tr>
<tr>
<td>Union</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.1798</td>
<td>0.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td>Wage</td>
<td>1.0000</td>
<td>5.2500</td>
<td>7.7800</td>
<td>9.0240</td>
<td>11.2500</td>
<td>44.5000</td>
</tr>
<tr>
<td>Age</td>
<td>18.00</td>
<td>28.0000</td>
<td>35.00</td>
<td>36.83</td>
<td>44.00</td>
<td>64.00</td>
</tr>
<tr>
<td>Race</td>
<td>1.0000</td>
<td>3.0000</td>
<td>3.0000</td>
<td>2.699</td>
<td>3.0000</td>
<td>3.0000</td>
</tr>
<tr>
<td>Occ</td>
<td>1.0000</td>
<td>3.0000</td>
<td>4.0000</td>
<td>4.148</td>
<td>6.0000</td>
<td>2.0000</td>
</tr>
<tr>
<td>Sect</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.2753</td>
<td>0.0000</td>
<td>2.0000</td>
</tr>
<tr>
<td>Marr</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.6554</td>
<td>0.0000</td>
<td>1.0000</td>
</tr>
</tbody>
</table>
```

**Question 1.** How large is this dataset?

**Question 2.** Several of the variables are coded using dummy coding. Which ones are they?

**Question 3.** The variable Marr, is the person married, is coded 0=no and 1=yes. What percentage of the people in this sample are married?

**Question 4.** Look at the above summary carefully. Look particularly at the numeric variables. Do you see anything that strikes you as odd? (Hint: look closely at the summary of Wage. The boxplot at right might help.)
It's always a good idea to look carefully at your data before you begin an analysis. That's not only true if it's someone else's data, it's also true if you created the dataset yourself. It's not hard to make a data transcription error or a typing error. Things you should look for include the following:

1) Missing values, which in R will be represented by NA at the bottom of the summary. For example, suppose we saw the following in the summary of Educ.

   Educ
   Min.   : 2.00
   1st Qu.:12.00
   Median :12.00
   Mean   :13.02
   3rd Qu.:15.00
   Max.   :18.00
   NAs    : 3

   NA, meaning not available, indicates that some data values are missing. Perhaps that bit of information was not obtained from some of the subjects (3 in this example). The presence of missing values in the data means we may have to take special care in the analysis. In wages.txt, there are no missing values.

2) Extreme values (min or max) that make no sense. For example, a minimum value of 0 for Age would be suspicious. A maximum value of 125 would also be suspicious.

3) Variables that appear to be coded numerically but are clearly not numeric variables. A good example here is Race. Another good example is Occ. This is exactly the way the data came from the authors, by the way, so we have to deal with that. We may have done the coding differently, but what we would have done doesn't matter. We work with what we have.

In this case, we have a very suspicious maximum value of Wage. The boxplot confirms that this value of 44.5 is an extreme outlier. Let's see who it is. We can ask R to find out for us by using a function called which().

   > which(WG$Wage > 40)   # which cases have values of Wage greater than 40?
   [1] 171
   > WG[171, ]   # examine case 171, all columns
       Educ South Sex Exper Union Wage Age Race Occ Sect Marr
   171  14     0   1     1     0 44.5  21    3   1    0    0

   First, I asked which values of Wage are greater than 40. R told me a value of Wage which fits that criterion occurs in case (row) 171. So I asked to see row 171 of the data frame (row 171, all columns). We see that this person was female, white, 21 years old with 1 year of experience at her job, worked in management, was unmarried, not a member of a union, not from the south, and probably went to a community college and got an associates degree (although we don't know that last one for sure).

   Question 5. Is it reasonable that this person is making $44.50 an hour (remember that this is 1985), or is this a suspicious value? Very briefly explain.

   I think it would be best if we deleted this case from the data frame, although with a data frame this large it probably wouldn't make much difference. (It would mess up some of our graphs.)

   > WG.bak = WG   # first I'm going to make a backup copy
   > WG = WG[-171, ]   # negating the case number deletes that case
A few rules: 1) Never modify a data set unless you have a backup copy; 2) If you have used the attach() function to tell R it’s okay to look inside the data frame, here’s a rule you would well advised to obey: NEVER MODIFY AN ATTACHED DATA FRAME! 3) You can't delete cases just because you don't like them. You have to have a good reason, preferably a very good reason.

Let's redo the summary now.

```r
> summary(WG)

                   Min. 1st Qu. Median   Mean 3rd Qu.   Max. 
Educ           2.00   12.00   12.00   13.02   15.00   18.00 
South          0.0000  0.0000  0.0000  0.2927  0.0000  0.2927 
Sex            0.0000  0.0000  0.0000  0.4578  0.0000  0.4578 
Exper          0.0000  8.0000 15.0000 17.8500 26.0000 55.0000 
Union          0.0000  0.0000  0.0000  0.1801  0.0000  0.1801 
Wage           1.0000  5.2500  7.7800  8.9580 11.2500 26.2900 
Age            18.0000 28.0000 35.0000 36.8600 44.0000 64.0000 
Race           1.0000  3.0000  3.0000  2.6980  3.0000  3.0000 
Occ            1.0000  8.0000 15.0000 4.1540  6.0000  6.0000 
Sect           0.0000  0.0000  1.0000  0.2758  2.0000  2.0000 
Marr           0.0000  0.0000  1.0000  0.6567  1.0000  1.0000
```

The only thing that should change is the maximum value of Wage (and the mean will change a smidge as well). Notice that the next highest value of Wage was $26.29, so $44.50 was way out of range. It's good we got rid of it. Now we are ready for some serious data analysis.

We are interested in finding variables that relate to Wage. Since the data frame is all numeric, one way we could begin is by looking at a correlation matrix. The basic R function is `cor()`.

```r
> round(cor(WG),3)  # round the correlations to 3 decimal places

                      Educ South Sex Exper Union Wage Age Race Occ Sect Marr 
Educ    1.000 0.001 -0.352 -0.244  0.395 0.149 0.096 -0.205 -0.189 -0.035
South  0.001 1.000 -0.020 -0.009 -0.078 -0.139 -0.040 -0.114  0.012  0.000
Sex     0.000 -0.020  1.000  0.078  0.156  0.230  0.082  0.026 -0.218 -0.170
Exper  -0.352 -0.009  0.078  1.000  0.117  0.110  0.978 -0.023 -0.027  0.110
Union  -0.024 -0.087 -0.156  1.000  0.117  0.176  0.119 -0.086  0.229  0.095
Wage    0.395 -0.139 -0.230  0.110  1.000  0.176  1.000  0.204  0.093 -0.026
Age     0.149 -0.040  0.082  0.978  0.119  0.204  1.000 -0.003 -0.074  0.075
Race    0.096 -0.114  0.026 -0.023 -0.086  0.093 -0.003  1.000  0.016  0.002
Occ    -0.205  0.012 -0.218 -0.027 -0.026 -0.074  0.016  1.000  0.364 -0.016
Sect   -0.189  0.000 -0.170  0.110  0.095  0.055  0.075  0.002  0.364  1.000
Marr   -0.035  0.005  0.140  0.268  0.092  0.124  0.276  0.045 -0.016  0.055
```

**Question 6.** Some of these correlations are meaningless. On the answer sheet, circle the correlations that are meaningless?

**Question 7.** When we think of correlation coefficients, we usually think of Pearson r. Not all of these correlations are Pearson r, however. Some are point-biserial correlations. On the answer sheet, circle the correlations that are point-biserial correlations?
Question 8. Which other variable has the strongest meaningful relationship to Wage?

Unlike some statistical software, R does not tell which correlations in a correlation matrix are significantly different from zero. So I looked it up. With 533 cases, we need the correlation to be more extreme, either positive or negative, than 0.085 to be significant at the alpha=.05 level.

Question 9. Which of the following correlations are significantly different from zero: Wage with Educ, Wage with South, Wage with Union, Wage with Race?

A positive correlation means higher values of one variable tend to go with higher values of the other variable and lower values with lower values. A negative correlation means that higher values of one variable tend to go with lower values of the other variable, and *vice versa*. Examine the correlation of Wage with South. South is coded 0=no (not from the south) and 1=yes (from the south).

Question 10. Which group has higher Wage, on the average, people from the South or people not from the South? Very briefly explain how you can tell from the correlation.

Question 11. You can see that the correlation of Wage with years of Educ is positive. Does that indicate that having more education means *a person* will make more money? (Firing up R and looking at a scatterplot might help you with this one if you're not sure or think I'm trying to trick you.)

Here is a t-test of Wage by Union membership.

```r
> t.test(Wage ~ Union, data=WG, var.eq=T)

Two Sample t-test

data:  Wage by Union
  t = -4.1184, df = 531, p-value = 4.424e-05
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  -3.315807  -1.174128
sample estimates:
mean in group 0  mean in group 1
  8.553158       10.798125
```

We can see that union members have a higher average wage than nonunion members and that this difference is statistically significant, p < .001.

Question 12. Is this difference large? Very briefly explain.

Occ is coded as follows.

```r
# Occ: Occupational category (1=Management, 2=Sales, 3=Clerical,
#      4=Service, 5=Professional, 6=Other).
```

To plot a graph of Wage by Occ, I would use boxplots. And I would pretty it up by putting the actual names of occupations on the horizontal axis rather than just labeling this axis 1 through 6.

```r
> occupations = c("Management","Sales","Clerical","Service","Professional","Other")
> boxplot(Wage ~ Occ, data=WG, names=occupations)
> title(ylab="Wage")  # add a y-axis label
```
Question 13. Which occupational group has the highest median Wage? (You'll have to look carefully, or use R to calculate the medians.)

Question 14. The person with the lowest Wage in this dataset is a member of which occupational group?

I am going to create a new occupations variable, and I'm going to put it in the workspace where I can get at it directly. The reason for this is that I don't like the way WG$Occ is coded. It wouldn't work as the IV in an analysis of variance, for example, because it is coded as numbers. I'm also going to pull Wage out of the data frame and put a copy in the workspace. I'm going to be careful to give them different names than the variables in the data frame so that we don't get confused.

> Occup = factor(WG$Occ, labels=occupations)  # declare a factor and name groups
> Wages = WG$Wage  # make a copy of the Wage variable in the workspace

The first command makes Occ into a factor, which is to say, a categorical variable that can be used as an IV. I'm going to code the levels of my new variable with the names we put into the occupations variable to do the boxplots. The second command just copies the Wage variable out of the data frame and into the workspace with a slightly different name. Now we can manipulate these variables directly without using that annoying $ notation. (And I prefer to do it this way rather than attaching the data frame, which is risky.) For example, here are the group means.

> tapply(Wages, Occup, mean)
  Management  Sales  Clerical  Service  Professional  Other
  12.115185 7.592632 7.422577 6.537470 11.947429 8.426474

Question 15. It appears that there are substantial differences in mean Wages by occupational group. What method would we use to determine if any of these differences are significant?

Question 16. Using this method (Question 15), it is actually possible to calculate a meaningful "correlation" (of sorts) between these two variables. That correlation is 0.424. How does this compare to how strongly other variables are related to Wage?
Let's pull Sex out of the data frame and recode it as a factor. Then we'll calculate the means.

```r
> Gender = factor(WG$Sex, labels=c("Male","Female"))
> tapply(Wages,Gender,mean)
  Male   Female
9.994913 7.728770
```

Finally, here is a table that shows how many people in each of the occupational groups are male or female.

```r
> table(Gender, Occup)

<table>
<thead>
<tr>
<th>Occup</th>
<th>Gender</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Male</td>
</tr>
<tr>
<td>Management</td>
<td>34</td>
</tr>
<tr>
<td>Sales</td>
<td>21</td>
</tr>
<tr>
<td>Clerical</td>
<td>21</td>
</tr>
<tr>
<td>Service</td>
<td>34</td>
</tr>
<tr>
<td>Professional</td>
<td>53</td>
</tr>
<tr>
<td>Other</td>
<td>126</td>
</tr>
</tbody>
</table>
```

For example, in the Management occupational group there are 34 males and 20 females. Are you beginning to see a problem?

**Question 17.** Notice that men make more money, on the average, than women. Notice also that there tends to be more men in the higher paying occupational groups such as Management, while there are more women in the lower paying occupational groups such as Clerical and Service. How do we know whether women make less money because they are women or because they are concentrated in lower paying occupations? (Don't give me a lot of politically correct drivel here. I don't want your opinion. Give me a statistical answer. In what way should we look at the data to answer this question?)

Here's another complicating factor.

```r
> tapply(WG$Exper, Gender, mean)
  Male   Female
16.96540 18.90574
```

Somewhat surprisingly, on the average, women have more years of experience at their jobs than men do, by about 2 years.

**Question 18.** If we could somehow adjust Wages for years of Exper, what effect would that have on the Gender disparity in Wages? I.e., would it make the difference larger or smaller?

**Question 19.** Examine the correlation between Age and Exper. What is it? Why is it so high?

**Question 20.** You will notice that the correlation between Educ and Exper is negative. That means the more Educ you have, the less years of Exper you have at your job. Why should that be true?

```r
> # create a married factor with group labels
> Married = factor(WG$Marr, labels=c("Unmarried","Married"))
> # means of Wages by Gender and Married
> tapply(Wages, list(Gender,Married), mean)

<table>
<thead>
<tr>
<th>Married</th>
<th>Gender</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Male</td>
<td>8.354752</td>
<td>10.876064</td>
</tr>
<tr>
<td></td>
<td>Female</td>
<td>7.817683</td>
<td>7.683765</td>
</tr>
</tbody>
</table>
> # draw a graph of those means (next page)
> interaction.plot(Married, Gender, Wages, type="b", pch=1:2)
```
Question 21. Describe the effects you see in the graph above?

```r
> # create a union factor with group labels
> Unio = factor(WG$Union, labels=c("Nonunion","Union"))
> # means of Wages by Gender and Married
> tapply(Wages, list(Gender, Unio), mean)
  Nonunion     Union
Male   9.596742 11.288971
Female 7.485417  9.606071
```

Question 22. Describe the effects you see in the graph on the next page?
Question 1. How large is this dataset? no. of cases ______ no. of variables ______

Question 2. Several of the variables are coded using dummy coding. Which ones are they? (Circle.)

> names(WG)
[1] "Educ" "South" "Sex" "Exper" "Union" "Wage" "Age" "Race" "Occ" "Sect" "Marr"

Question 3. The variable Marr, is the person married, is coded 0=no and 1=yes. What percentage of
the people in this sample are married?

percent married = _________ %

Question 4. Look at the summary carefully. Look particularly at the numeric variables. Do you see
anything that strikes you as odd?

Question 5. Is it reasonable that this person is making $44.50 an hour (remember that this is 1985), or
is this a suspicious value? Very briefly explain.

Question 6. Some of these correlations are meaningless. Circle the correlations that are meaningless?

> round(cor(WG), 3)

<table>
<thead>
<tr>
<th></th>
<th>Educ</th>
<th>South</th>
<th>Sex</th>
<th>Exper</th>
<th>Union</th>
<th>Wage</th>
<th>Age</th>
<th>Race</th>
<th>Occ</th>
<th>Sect</th>
<th>Marr</th>
</tr>
</thead>
<tbody>
<tr>
<td>Educ</td>
<td>1.000</td>
<td>-0.140</td>
<td>0.001</td>
<td>-0.352</td>
<td>-0.024</td>
<td>0.395</td>
<td>-0.149</td>
<td>0.096</td>
<td>-0.205</td>
<td>-0.189</td>
<td>-0.035</td>
</tr>
<tr>
<td>South</td>
<td>-0.140</td>
<td>1.000</td>
<td>-0.020</td>
<td>-0.009</td>
<td>-0.087</td>
<td>-0.139</td>
<td>-0.040</td>
<td>-0.114</td>
<td>0.012</td>
<td>0.000</td>
<td>0.005</td>
</tr>
<tr>
<td>Sex</td>
<td>0.001</td>
<td>-0.020</td>
<td>1.000</td>
<td>0.078</td>
<td>-0.156</td>
<td>-0.230</td>
<td>0.082</td>
<td>0.026</td>
<td>-0.218</td>
<td>-0.170</td>
<td>0.014</td>
</tr>
<tr>
<td>Exper</td>
<td>-0.352</td>
<td>-0.009</td>
<td>0.078</td>
<td>1.000</td>
<td>0.117</td>
<td>0.110</td>
<td>0.978</td>
<td>-0.023</td>
<td>-0.027</td>
<td>0.110</td>
<td>0.268</td>
</tr>
<tr>
<td>Union</td>
<td>-0.024</td>
<td>-0.087</td>
<td>-0.156</td>
<td>0.117</td>
<td>1.000</td>
<td>0.176</td>
<td>0.119</td>
<td>-0.086</td>
<td>0.229</td>
<td>0.095</td>
<td>0.092</td>
</tr>
<tr>
<td>Wage</td>
<td>0.395</td>
<td>-0.139</td>
<td>-0.230</td>
<td>0.110</td>
<td>0.176</td>
<td>1.000</td>
<td>0.204</td>
<td>0.093</td>
<td>-0.026</td>
<td>0.055</td>
<td>0.124</td>
</tr>
<tr>
<td>Age</td>
<td>-0.149</td>
<td>-0.040</td>
<td>0.082</td>
<td>0.978</td>
<td>0.119</td>
<td>0.204</td>
<td>1.000</td>
<td>-0.003</td>
<td>-0.074</td>
<td>0.075</td>
<td>0.276</td>
</tr>
<tr>
<td>Race</td>
<td>0.096</td>
<td>-0.114</td>
<td>0.026</td>
<td>-0.023</td>
<td>-0.086</td>
<td>0.093</td>
<td>-0.003</td>
<td>1.000</td>
<td>0.016</td>
<td>0.002</td>
<td>0.045</td>
</tr>
<tr>
<td>Occ</td>
<td>-0.205</td>
<td>0.012</td>
<td>-0.218</td>
<td>-0.027</td>
<td>0.229</td>
<td>-0.026</td>
<td>-0.074</td>
<td>0.016</td>
<td>1.000</td>
<td>0.364</td>
<td>-0.016</td>
</tr>
<tr>
<td>Sect</td>
<td>-0.189</td>
<td>0.000</td>
<td>-0.170</td>
<td>0.110</td>
<td>0.095</td>
<td>0.055</td>
<td>0.075</td>
<td>0.002</td>
<td>0.364</td>
<td>1.000</td>
<td>0.055</td>
</tr>
<tr>
<td>Marr</td>
<td>-0.035</td>
<td>0.005</td>
<td>0.014</td>
<td>0.268</td>
<td>0.092</td>
<td>0.124</td>
<td>0.276</td>
<td>0.045</td>
<td>-0.016</td>
<td>0.055</td>
<td>1.000</td>
</tr>
</tbody>
</table>

Question 7. When we think of correlation coefficients, we usually think of Pearson r. Not all of these
correlations are Pearson r, however. Some are point-biserial correlations. Circle the correlations that
are point-biserial correlations? (Correlation matrix for this is on next page.)
Question 8. Which other variable has the strongest meaningful relationship to Wage?

variable: _______________    correlation = __________

Question 9. Which of the following correlations are significantly different from zero?

Wage with Educ    yes / no / can't say    (circle)
Wage with South   yes / no / can't say    (circle)
Wage with Union   yes / no / can't say    (circle)
Wage with Race    yes / no / can't say    (circle)

Question 10. Which group has higher Wage, on the average, people from the South or people not from the South? Very briefly explain how you can tell from the correlation.

Question 11. You can see that the correlation of Wage with years of Educ is positive. Does that indicate that having more education means a person will make more money? Very briefly explain.

Question 12. Is the difference in wage between union and nonunion members large? Very briefly explain.

Question 13. Which occupational group has the highest median Wage? _______________

Question 14. The person with the lowest Wage is a member of which occupational group? ______________
**Question 15.** It appears that there are substantial differences in mean Wages by occupational group. What method would we use to determine if any of these differences are significant?

______________ (the answer is going to be something like t-test, ANOVA, chi-square, etc.)

**Question 16.** Using this method (Question 15), it is actually possible to calculate a meaningful "correlation" (of sorts) between these two variables. That correlation is 0.424. How does this compare to how strongly other variables are related to Wage?

**Question 17.** Notice that men make more money, on the average, than women. Notice also that there tends to be more men in the higher paying occupational groups such as Management, while there are more women in the lower paying occupational groups such as Clerical and Service. How do we know whether women make less money because they are women or because they are concentrated in lower paying occupations?

**Question 18.** If we could somehow adjust Wages for years of Exper, what effect would that have on the Gender disparity in Wages? I.e., would it make the difference larger or smaller?

larger / smaller (circle)

**Question 19.** Examine the correlation between Age and Exper. What is it? Why is it so high?

correlation = __________

**Question 20.** You will notice that the correlation between Educ and Exper is negative. That means the more Educ you have, the less years of Exper you have at your job. Why should that be true?
A hint for answering the last two questions: Don't get carried away describing every little variation you see on the graphs. These are sample data, remember, and sample data are noisy. Small differences in the means are probably not "real".

**Question 21.** Describe the effects you see in the graph of Wages ~ Gender + Married?

**Question 22.** Describe the effects you see in the graph of Wages ~ Gender + Union?