BUEC-333, Fall 2012
Hand-in assignment #2
Solutions + Point distribution

Deadline: July 24, 14:30

Rules

1. Use the data that is provided on the course website
2. Use R to answer the questions. Other software is not accepted.
3. Hand in a hardcopy of typed answers, font Times New Roman, font size 12.
4. For **each** question:
   (a) If applicable: copy-paste the code that you used to get that outcome: **without** code, you get **0 points**
   (b) If applicable: copy-paste the R output supporting the answer (this might involve tables and pictures)
   (c) Type **2 or 3 lines** of explanation: **fewer than 2 lines, or more than 3 lines: 0 points**

Part I: Unions

Last Fall, the labor unions CUPE and TSSU were on strike. One of the reasons to form a union, and one of the reasons to organize a strike, is to increase your bargaining power in labor negotiations with the employer. It is sometimes believed that unions and strikes improve the wages of the members of that union. In this assignment, you are going to investigate the effects of unions and strikes on wages empirically using linear regression in R. Load the data for our analysis of unions. A description of this data can be found in Table 1 of this paper: [link]

```
> ## Load the "foreign" software
> require(foreign)
> require(ggplot2)
> ## Download the data from my website and store it in a data.frame.
> unionData <- read.dta("http://www.sfu.ca/~cmuris/2013-Summer-333/wagepan.dta")
> ## Delete missing data
> unionData <- na.omit(unionData)
```

To get an idea about the data set, use "str", "summary", "head", etc. Once you are familiar with the data, answer the following questions:

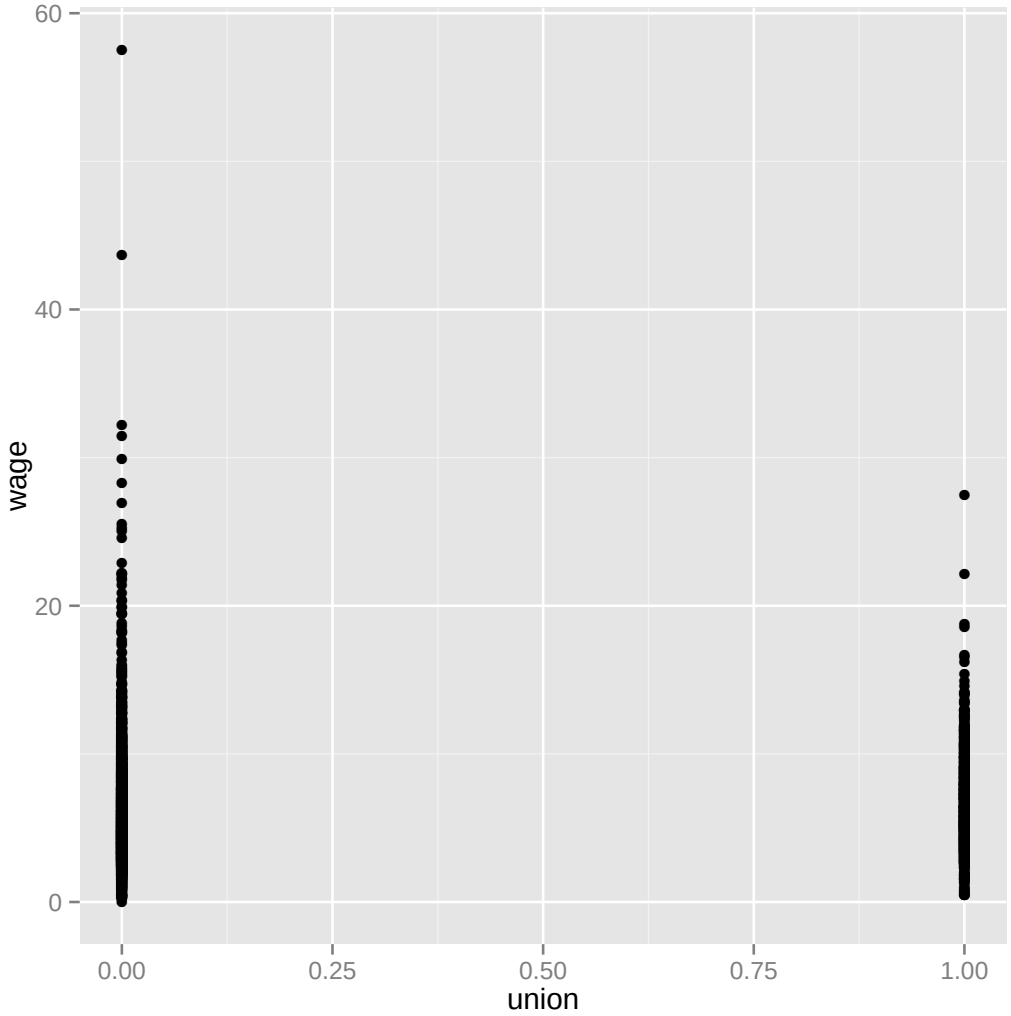
1. There is no variable that contains the wage, as only the log of wage is provided. Generate a new variable, "unionData$wage", that contains the hourly wage in US dollars. **1 point for the formula that follows or an alternative that gives the same result.**

```
> ## Construct a new variable containing wage
> unionData$wage <- exp(unionData$lwage)
> summary(unionData$wage)

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.0279  3.8600  5.3180  5.9190  7.3230 57.5000
```

2. Make a scatterplot (you can use the package "ggplot", for example) of "wage" and the variable "union", which measures whether somebody is a part of a union or not. What do you conclude from this scatterplot? Does it make a difference when you do it for the log og wage? [Include code and the plot you made. **1 point for the obvious plot that follows, and then 1 point if they conclude that (i) it is difficult to say anything based on that graph (ii) find an alternative way to say something about the relationship. -1 if they use "lwage" without an explanation.**

```
> require(ggplot2)
> qplot(union,wage,data=unionData)
```



3. Run a regression of "wage" on "union". Interpret the regression coefficient estimate for "union". **3 points. 1 point for the code+output, 1 point for the interpretation of coefficient estimate, 1 point for the CI. -1 if they forget "ceteris paribus".**

```
> reg <- lm(wage~union,data=unionData)
> summary(reg)

Call:
lm(formula = wage ~ union, data = unionData)

Residuals:
   Min     1Q Median     3Q    Max
-6.132 -2.013 -0.598  1.345 51.797

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.70771    0.05541 103.017  < 2e-16 ***
union        0.86652    0.11216   7.726 1.37e-14 ***
```

```
---
Signif. codes:  0
```

Report the 90% confidence interval for the regression coefficient estimate of union: what do you conclude? **1 point, by hand or using commands like "conf.int". Conclusion is about statistical significance, so that there is a true effect of unions on wage.**

```
> confint(reg,2,level=0.90)

            5 %     95 %
union 0.6820009 1.051043

> 0.86652+1.64*c(-.11216,.11216)

[1] 0.6825776 1.0504624
```

4. Do the same for "lwage". **2 points. -1 if the coefficient of union is not interpreted as a semi-elasticity.**

```
> reg.log <- lm(lwage~union,data=unionData)
> summary(reg.log)

Call:
lm(formula = lwage ~ union, data = unionData)

Residuals:
    Min      1Q  Median      3Q     Max
-5.1845 -0.2903  0.0197  0.3321  2.4465

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.605400   0.009181 174.866   <2e-16 ***
union       0.179264   0.018584   9.646   <2e-16 ***
---
Signif. codes:  0

> confint(reg.log,2,level=0.90)

           5 %      95 %
union 0.148689 0.2098394

> 0.179264+1.64*c(-.018584,.018584)

[1] 0.1487862 0.2097418
```

5. I prefer the model with "lwage" over the model with "wage". However, there is a problem because we did not include any variables other than "union". Explain what the problems are with this, from a theoretical/statistical point of view. **1 point. Omitted variable bias. BONUS point for an example of such an omitted variable. For example, "age" could be one: older people generally earn more, and are more likely to be member of a union (?).**

6. Run a regression of "lwage" on union, hours, year, occ1, occ2, occ3, occ4, occ5, occ6, occ7, occ8, occ9. R refuses to give an estimate for the regression coefficient on occ9. Why? **1 point. Multicollinearity. occ1+...+oc99=1.**

```
> reg3 <- lm(lwage~union+hours+year+occ1+occ2+occ3+occ4+occ5+occ6+occ7+occ8+occ9,dat
> summary(reg3)

Call:
lm(formula = lwage ~ union + hours + year + occ1 + occ2 + occ3 +
    occ4 + occ5 + occ6 + occ7 + occ8 + occ9, data = unionData)

Residuals:
    Min      1Q  Median      3Q     Max
-5.4186 -0.2533  0.0293  0.3059  2.1956

Coefficients: (1 not defined because of singularities)
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.166e+02  6.674e+00 -17.470  < 2e-16 ***
union        2.295e-01  1.780e-02  12.890  < 2e-16 ***
hours       -5.909e-05  1.379e-05  -4.286 1.86e-05 ***
year         5.954e-02  3.368e-03  17.677  < 2e-16 ***
occ1         4.226e-01  3.199e-02  13.212  < 2e-16 ***
occ2         3.740e-01  3.363e-02  11.121  < 2e-16 ***
occ3         3.183e-01  3.926e-02   8.107 6.66e-16 ***
occ4         1.887e-01  3.116e-02   6.057 1.50e-09 ***
occ5         2.699e-01  2.727e-02   9.897  < 2e-16 ***
occ6         1.828e-01  2.752e-02   6.643 3.45e-11 ***
occ7         8.913e-02  3.284e-02   2.714  0.00666 **
occ8        -4.992e-02  6.588e-02  -0.758  0.44865
occ9                NA         NA      NA       NA
---
Signif. codes:  0
```

7. Interpret the estimate of the regression coefficient of "year". **1 point. Everything else constant, wages have increased by 5.95% per year.**

8. I wonder whether I should include a person's work experience, as measured by the variable "exper". Try it, and argue why it should or should not be included. You can be informal.

```
> reg4 <- lm(lwage~union+hours+year+exper+occ1+occ2+occ3+occ4+occ5+occ6+occ7+occ8+oc
> summary(reg4)

Call:
lm(formula = lwage ~ union + hours + year + exper + occ1 + occ2 +
    occ3 + occ4 + occ5 + occ6 + occ7 + occ8 + occ9, data = unionData)

Residuals:
    Min      1Q  Median      3Q     Max
-5.4205 -0.2571  0.0359  0.3041  2.1743

Coefficients: (1 not defined because of singularities)
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.452e+02  1.147e+01 -12.654  < 2e-16 ***
union        2.294e-01  1.779e-02  12.897  < 2e-16 ***
hours       -5.777e-05  1.378e-05  -4.192 2.82e-05 ***
year         7.399e-02  5.798e-03  12.762  < 2e-16 ***
exper       -1.430e-02  4.672e-03  -3.061  0.00222 **
occ1         4.106e-01  3.219e-02  12.753  < 2e-16 ***
occ2         3.670e-01  3.368e-02  10.898  < 2e-16 ***
occ3         3.126e-01  3.927e-02   7.960 2.18e-15 ***
occ4         1.848e-01  3.116e-02   5.933 3.21e-09 ***
occ5         2.762e-01  2.733e-02  10.108  < 2e-16 ***
occ6         1.881e-01  2.755e-02   6.827 9.84e-12 ***
occ7         9.454e-02  3.285e-02   2.878  0.00403 **
occ8        -3.831e-02  6.593e-02  -0.581  0.56125
occ9                NA         NA      NA       NA
---
Signif. codes:  0
```

1 point. I would accept several arguments here. One would be: include, because coefficient estimate is statistically significantly different from zero, and including it also affects the estimate for "union". Well-explained reference to change in the R^2 are also acceptable. Alternatively, it would be ok if they said not to include it, for example because the sign does not make sense (you'd expect it to be positive).

9. The relationship between wages and experience seems suspicious. There may still be an omitted variable problem. Include "educ". What do you conclude about education and experience, and their relationship with wages.

```
> reg5 <- lm(lwage~union+hours+year+exper+educ+occ1+occ2+occ3+occ4+occ5+occ6+occ7+oc
> summary(reg5)

Call:
lm(formula = lwage ~ union + hours + year + exper + educ + occ1 +
    occ2 + occ3 + occ4 + occ5 + occ6 + occ7 + occ8 + occ9, data = unionData)
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-5.4042 -0.2364  0.0384  0.2948  2.3645

Coefficients: (1 not defined because of singularities)
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -5.697e+01  1.238e+01  -4.601 4.32e-06 ***
union        2.064e-01  1.733e-02  11.911  < 2e-16 ***
hours       -6.772e-05  1.339e-05  -5.057 4.44e-07 ***
year         2.888e-02  6.272e-03   4.604 4.26e-06 ***
exper        3.317e-02  5.390e-03   6.154 8.21e-10 ***
educ         8.691e-02  5.331e-03  16.302  < 2e-16 ***
occ1         3.052e-01  3.192e-02   9.563  < 2e-16 ***
occ2         2.938e-01  3.301e-02   8.903  < 2e-16 ***
occ3         2.419e-01  3.837e-02   6.305 3.16e-10 ***
occ4         1.561e-01  3.030e-02   5.152 2.69e-07 ***
occ5         2.775e-01  2.653e-02  10.459  < 2e-16 ***
occ6         1.975e-01  2.675e-02   7.384 1.84e-13 ***
occ7         1.070e-01  3.190e-02   3.352 0.000808 ***
occ8        -2.834e-02  6.401e-02  -0.443 0.657980
occ9                NA         NA      NA       NA
---
Signif. codes:  0
```

2 points. EDUC was an omitted variable. It is negatively correlated with work experience (the years you spent in school detract from your years of work experience) and we expect it to matter for wages, also after conditioning on work experience. Conclude that both the years of schooling and work experience have positive effects on a person's wages, in expectation, ceteris paribus.

10. Are unions good for employees' wages? **2 points.**