SDS 439 - Homework 02

Due Feb 10, 1:00 pm

Minnesota Water Usage

This homework concerns analysis of the minnesota water dataset, which is included in the R package for the textbook "Applied Linear Regression" by Sanford Weisberg. The dataset has municipal and agricultural water usage data for 24 years, along with agricultural and municipal precipitation data for the same time period. We are interested in studying how water usage is changing over time.

The dataset is located in the course repository in the file `datasets/mn_water.csv`.

A thorough description of the data can be found on page 32 of the `alr4` documentation: https://cran.r-project.org/web/packages/alr4/alr4.pdf

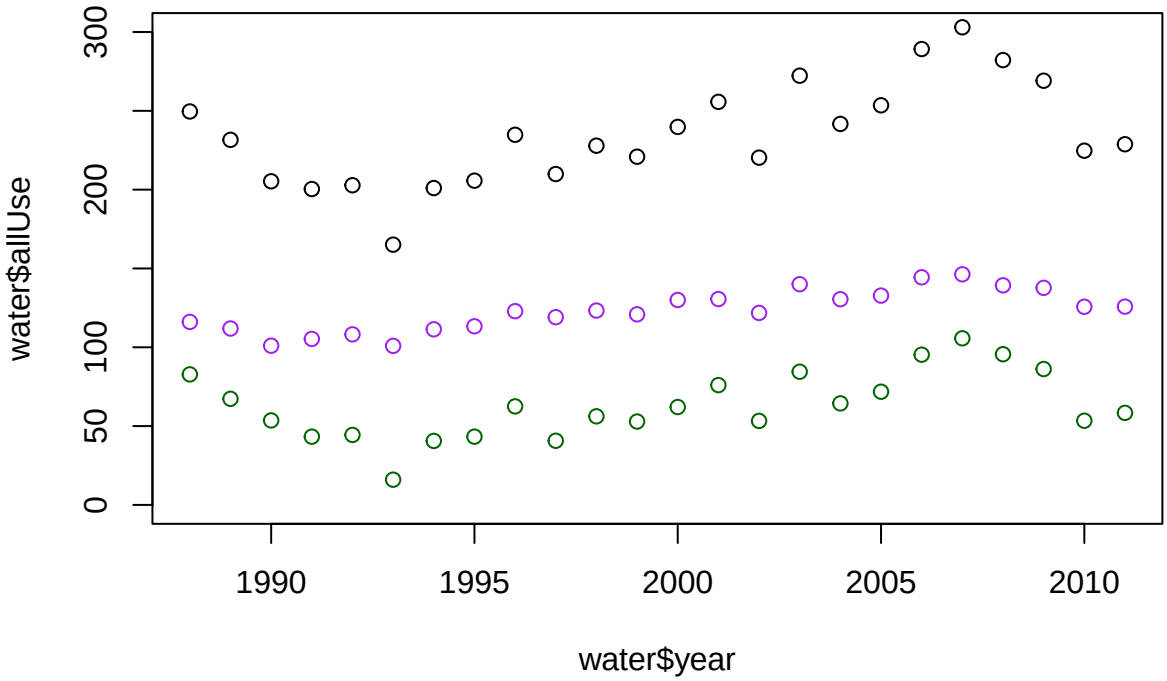
1. Read in the dataset and print out the first six rows.

```
water <- read.csv("../datasets/mn_water.csv")
head( water )
```

```
##   year allUse muniUse irrUse agPrecip muniPrecip statePop muniPop
## 1 2011  228.8   125.8   58.4     13.3       21.0  5332246 3354497
## 2 2010  224.7   125.7   53.4     15.7       25.6  5310584 3322498
## 3 2009  269.1   137.7   86.2     10.5       14.5  5281203 3288623
## 4 2008  282.2   139.3   95.6      9.4       15.1  5247018 3248005
## 5 2007  303.0   146.3  105.7      7.5       18.8  5207203 3209556
## 6 2006  289.2   144.4   95.3      7.7       17.4  5163555 3171804
```

2. You should try to start your analyses by looking at the data, so you have an idea of what to expect from your analysis. Make 4 informative plots of the data, to familiarize yourself with the major features of the data. Use good plotting practice, labeling your axes, using legends when appropriate, etc. A good plot should show several aspects of the data, without overwhelming you.

```
plot( water$year, water$allUse, ylim = c(0, 300) )
points( water$year, water$muniUse, col = "purple" )
points( water$year, water$irrUse, col = "darkgreen" )
```



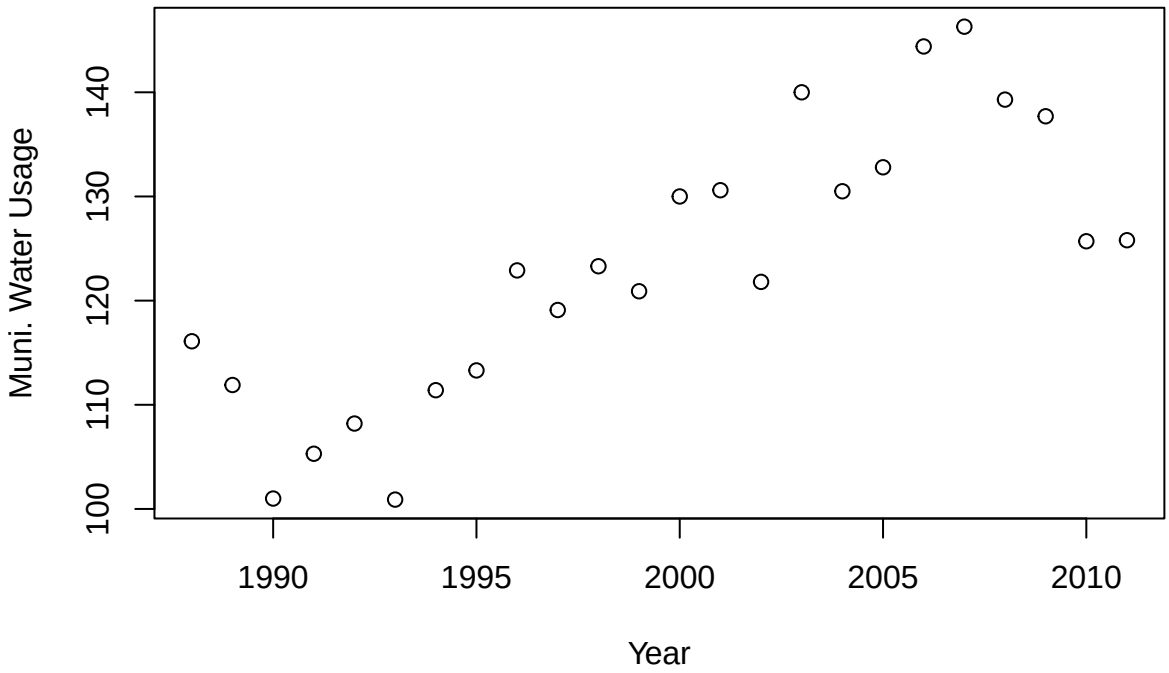
3. Use `lm` to regress the total water usage on the municipal usage and the irrigation usage, and print out the summary table. What do the results tell you about the relationship between these 3 variables?

```
summary( lm( allUse ~ muniUse + irrUse, data = water ) )
```

```
##
## Call:
## lm(formula = allUse ~ muniUse + irrUse, data = water)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.0721 -1.1282  0.4092  0.9033  2.8897
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 65.45520    4.54298   14.41 2.33e-12 ***
## muniUse      0.81381    0.04845   16.80 1.19e-13 ***
## irrUse       1.09657    0.02998   36.57  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.699 on 21 degrees of freedom
## Multiple R-squared:  0.9975, Adjusted R-squared:  0.9973
## F-statistic:  4263 on 2 and 21 DF,  p-value: < 2.2e-16
```

4. Let's first study the municipal water usage to understand how it is changing over time. Make a plot of municipal water usage over time.

```
plot( water$year, water$muniUse, xlab = "Year", ylab = "Muni. Water Usage" )
```



5. Regress municipal water usage on year. Print the results and answer the following questions: How do you interpret the slope? How do you interpret the intercept?

```
m1 <- lm( muniUse ~ year, data = water )
summary(m1)
```

```
##
## Call:
## lm(formula = muniUse ~ year, data = water)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.8980  -4.1003   0.2099   4.9290  11.6535
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2901.6827   459.6402  -6.313 2.36e-06 ***
## year            1.5129     0.2299   6.581 1.28e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.795 on 22 degrees of freedom
## Multiple R-squared:  0.6632, Adjusted R-squared:  0.6478
## F-statistic: 43.31 on 1 and 22 DF,  p-value: 1.282e-06
```

```
# Slope is the estimated expected in crease in water usage each year
# Intercept is the estimated water usage in year 0!
```

6. To facilitate interpretation of the intercept, create a new variable counting the years since 1988. Then regress municipal use on your new variable and print the results. What is the interpretation of this intercept in this model? Verify that the intercept for this model is equal to the fitted value for 1988 from the previous model.

```r
water$years_since_1988 <- water$year - 1988
summary( lm( muniUse ~ years_since_1988, data = water ) )
```

```
##
## Call:
## lm(formula = muniUse ~ years_since_1988, data = water)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.8980  -4.1003   0.2099   4.9290  11.6535
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      105.9020     3.0855  34.322  < 2e-16 ***
## years_since_1988   1.5129     0.2299   6.581 1.28e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.795 on 22 degrees of freedom
## Multiple R-squared:  0.6632, Adjusted R-squared:  0.6478
## F-statistic: 43.31 on 1 and 22 DF,  p-value: 1.282e-06
```

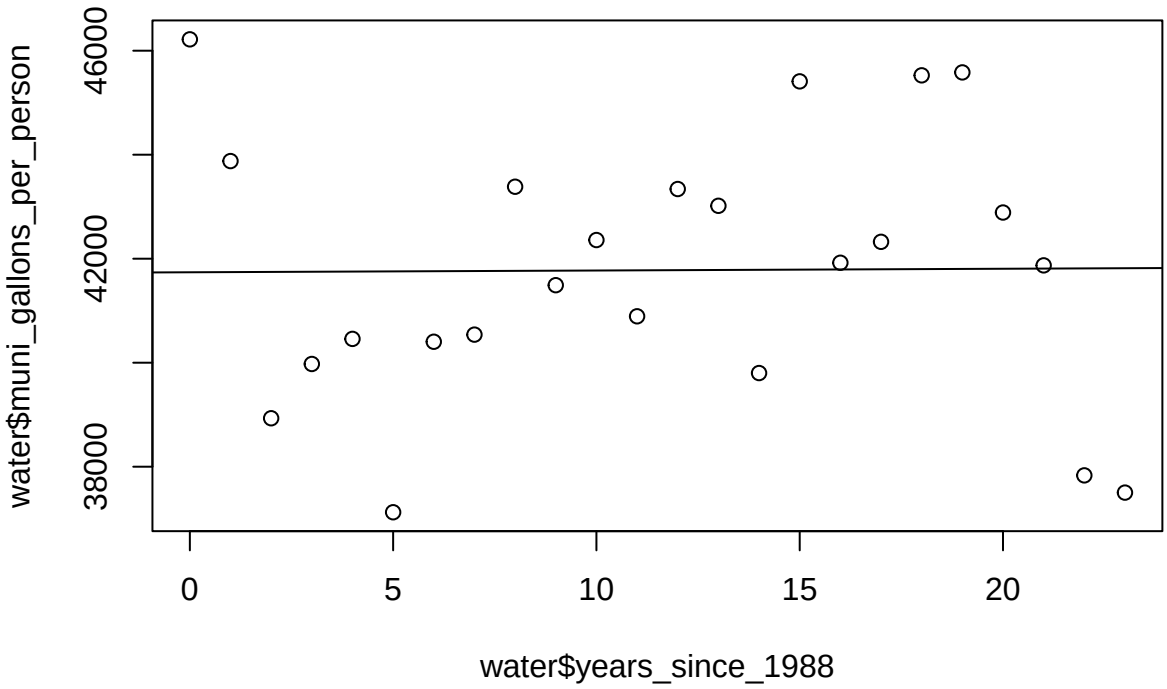
```r
m1$fitted.values[24]
```

```
##      24
## 105.902
```

```r
# Now the intercept is the estimated expect water usage in year 1988.
```

7. Perhaps the changes in municipal water usage can be explained by changes in municipal population. Regress municipal water use *per person* on years since
8. BUT make sure to use an appropriate unit for the response (billions of gallons per person is hard to interpret). Make a plot with the data and the regression line, and summarize your results in words.

```r
water$muni_gallons_per_person <- water$muniUse*1e9/water$muniPop
m1 <- lm( muni_gallons_per_person ~ years_since_1988, data = water )
plot( water$years_since_1988, water$muni_gallons_per_person )
abline(m1)
```



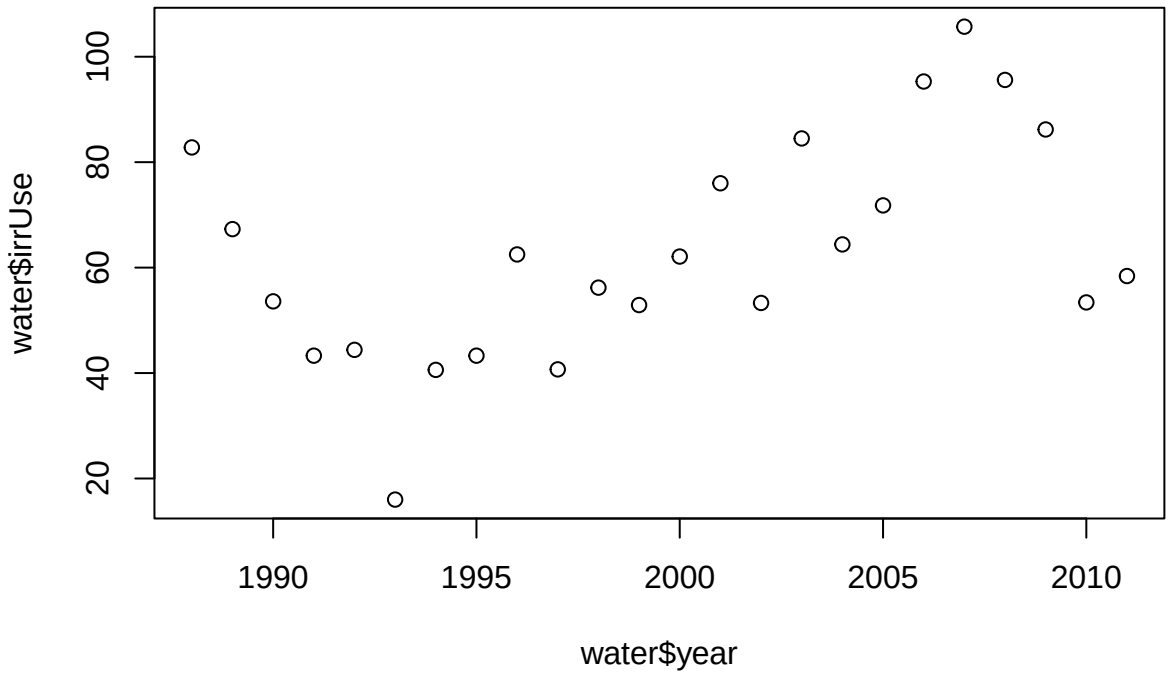

```
summary(m1)
```

```
##
## Call:
## lm(formula = muni_gallons_per_person ~ years_since_1988, data = water)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -4632  -1461     95   1575   4479
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      41739.017   1037.238  40.241   <2e-16 ***
## years_since_1988     3.386     77.275   0.044    0.965
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2621 on 22 degrees of freedom
## Multiple R-squared:  8.727e-05,  Adjusted R-squared:  -0.04536
## F-statistic: 0.00192 on 1 and 22 DF,  p-value: 0.9654
```

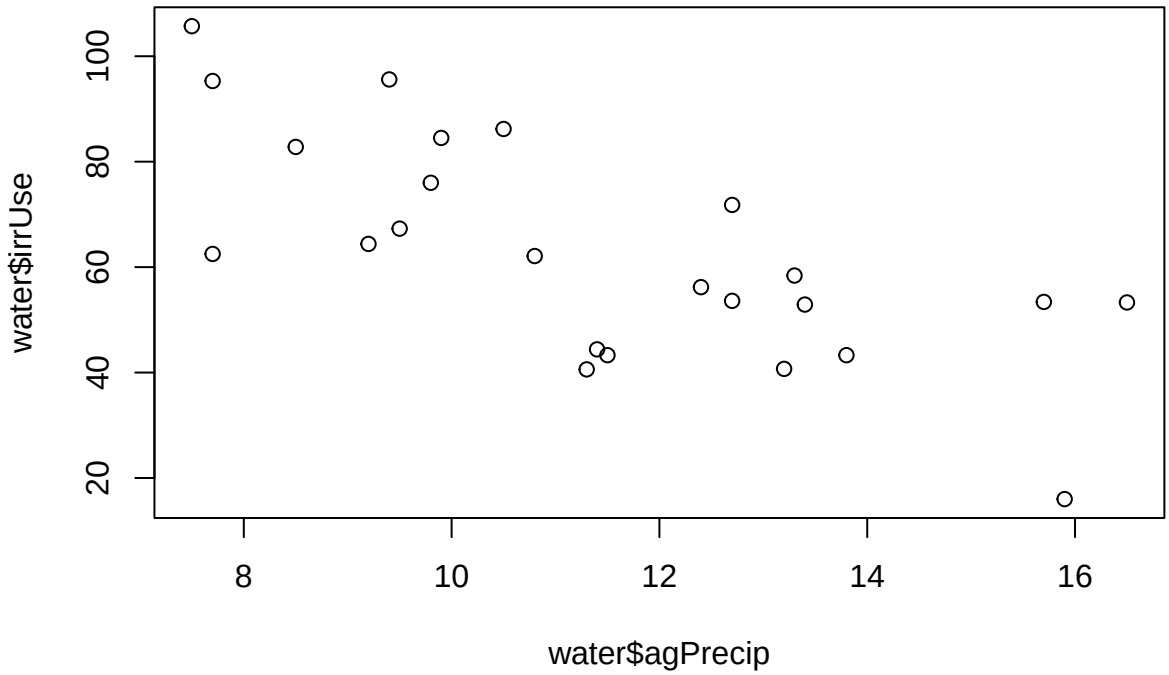
```
# this analysis converts the response to gallons per person. We can see from
# the plot and the fit that water usage in terms of gallons per person
# is not changing much overall, though the pattern is interesting. There were
# two early years with high usage, then two later years with low usage.
# Otherwise, usage was increasing in the middle.
```

8. Let's pivot to the irrigation usage. Make two plots of the irrigation usage, one against year, and one against agricultural precipitation

```
plot( water$year, water$irrUse )
```




```
plot( water$agPrecip, water$irrUse )
```



9. Perform a multiple regression of irrigation usage on years since 1988 and agricultural precipitation. Perform the following steps: (1) Write out the model in mathematical notation (like we do in class), (2) Print the summary table, (3) provide an interpretation of each of the 3 regression coefficients.

(1) $Y_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + \varepsilon_i$

```
# (2)
m2 <- lm( irrUse ~ agPrecip + years_since_1988, data = water )
summary(m2)
```

```
##
## Call:
## lm(formula = irrUse ~ agPrecip + years_since_1988, data = water)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -18.264  -7.777   1.064   8.229  19.576
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      113.2115    10.8155  10.468 8.65e-10 ***
## agPrecip          -5.8809     0.8668  -6.784 1.04e-06 ***
## years_since_1988   1.4723     0.3184   4.624 0.000146 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.8 on 21 degrees of freedom
## Multiple R-squared:  0.7638, Adjusted R-squared:  0.7413
## F-statistic: 33.95 on 2 and 21 DF,  p-value: 2.631e-07
```

(3) The intercept is the expected usage in 1988 when there is 0 precipitation. The agPrecip coefficient is the expected change in water usage when the precipitation increases by 1. The years_since_1988 coefficient is the expected change in water usage each year.

10. What do the regression results tell us about how irrigation usage is changing over time?

It seems to be increasing over time, after controlling for precipitation.

11. Provide an estimate of how much water would have been used in 2012 if the precipitation were 12 inches. Write down the estimate in terms of your model parameter estimates and provide the numerical value. Can you provide a standard error for your estimate?

```
vv <- vcov( m2 )
est <- m2$coefficients[1] + 12*m2$coefficients[2] + (2012-1988)*m2$coefficients[3]
est
```

```
## (Intercept)
##    77.97582
```

```
x <- c(1, 12, (2012-1988))
se <- sqrt( t(x) %*% vv %*% x )
se
```

```
##          [,1]
## [1,] 4.579781
```

12. Provide an estimate of how much less water would be used for irrigation in a rainy year (say 16 inches) versus in a dry year (say 8 inches). Write down the estimate in terms of your model parameters and provide the numerical value. Can you provide a standard error for your estimate?

```
est <- (16-8)*m2$coefficients[2]
x <- c(0, 16-8, 0)
se <- sqrt( t(x) %*% vv %*% x )
est
```

```
##  agPrecip
## -47.04713
```

```
se
```

```
##          [,1]
## [1,] 6.934632
```