HW1: Regression Modeling

110065508 Cheng-En Lee

due on 10/11 (Tue)

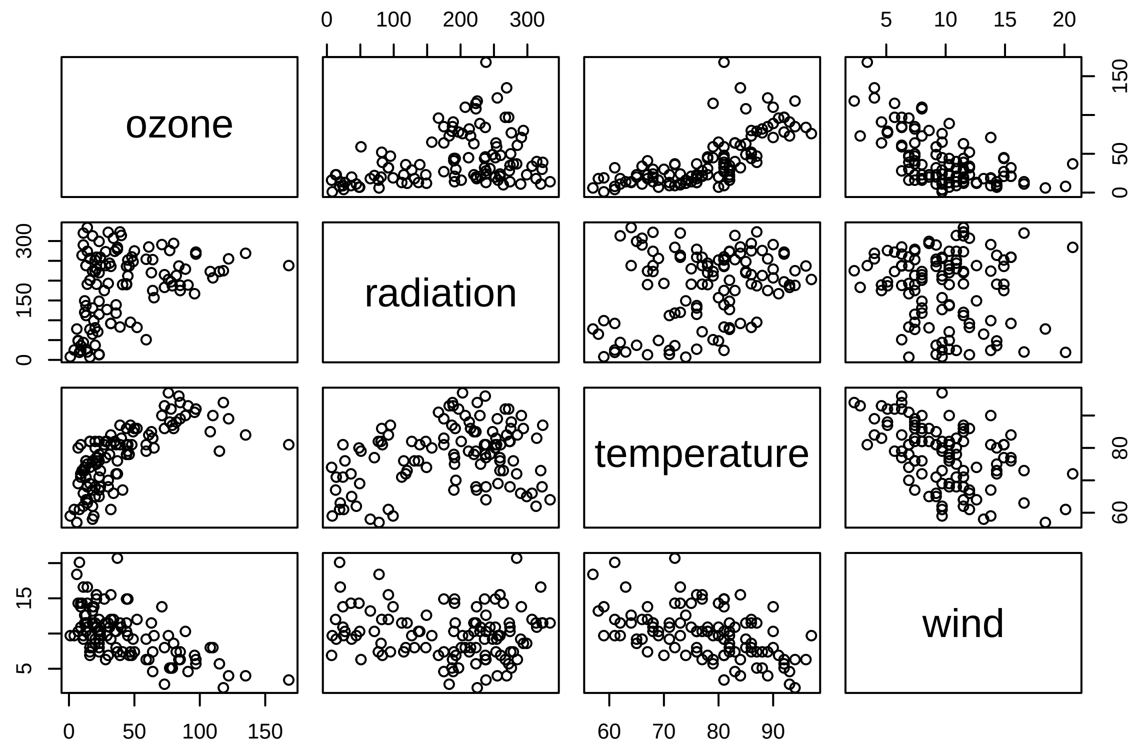
Problem1

(a) Exploratory data analysis (EDA) among 4 variables

```
oz <- read.csv("ozone.csv")
head(oz)
```

```
##   ozone radiation temperature wind
## 1    41       190          67  7.4
## 2    36       118          72  8.0
## 3    12       149          74 12.6
## 4    18       313          62 11.5
## 5    23       299          65  8.6
## 6    19        99          59 13.8
```

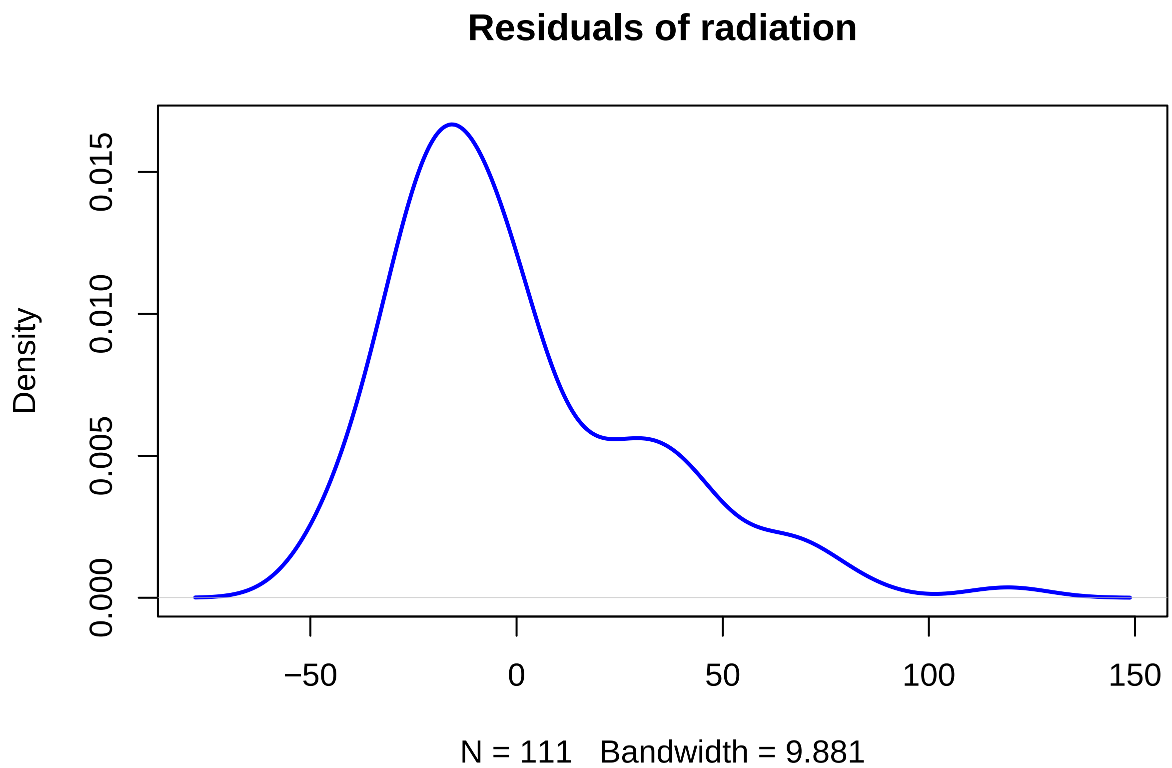
```
pairs(oz)
```



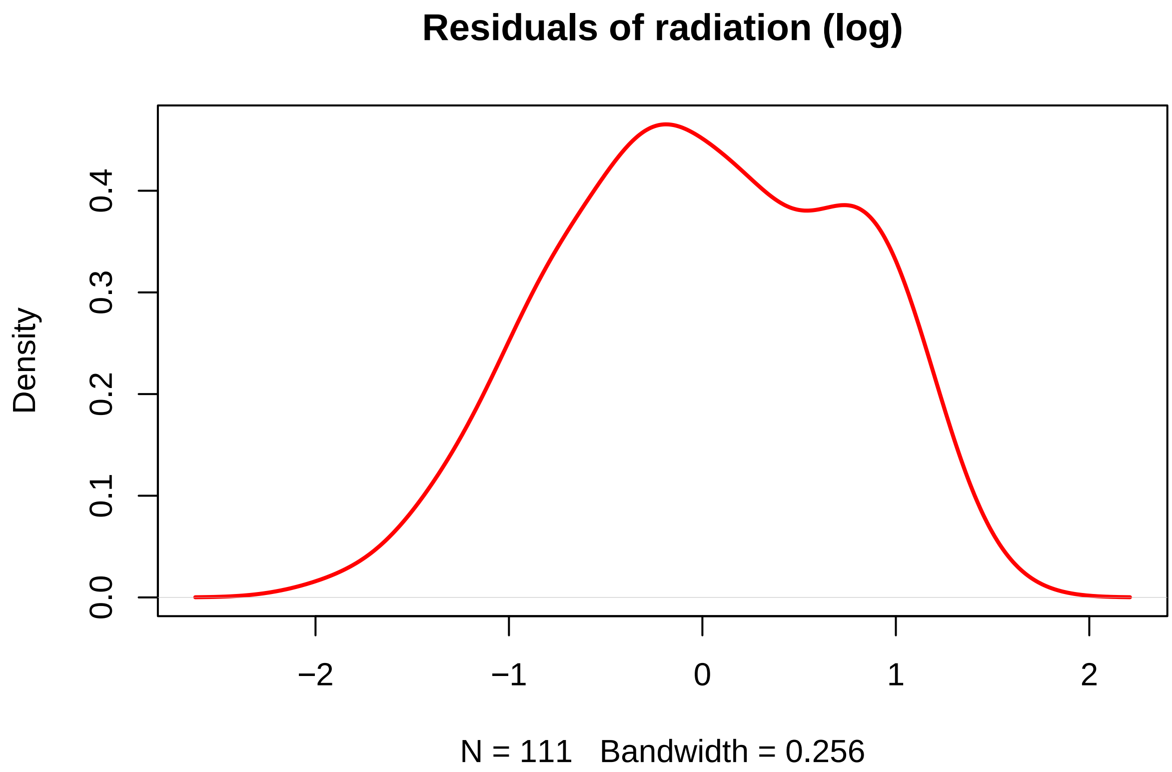
By the scatter plots, it looks like all the variables have a positive or negative relation with `ozone`. However, the variable `radiation` and `temperature` seems to have non-linearity relations with `ozone`.

To investigate the relation, let's plot the density plots of the residuals. My intuition is: If the log-transformed dataset produce better-distributed residuals, then let's do the log transformed multiple regression.

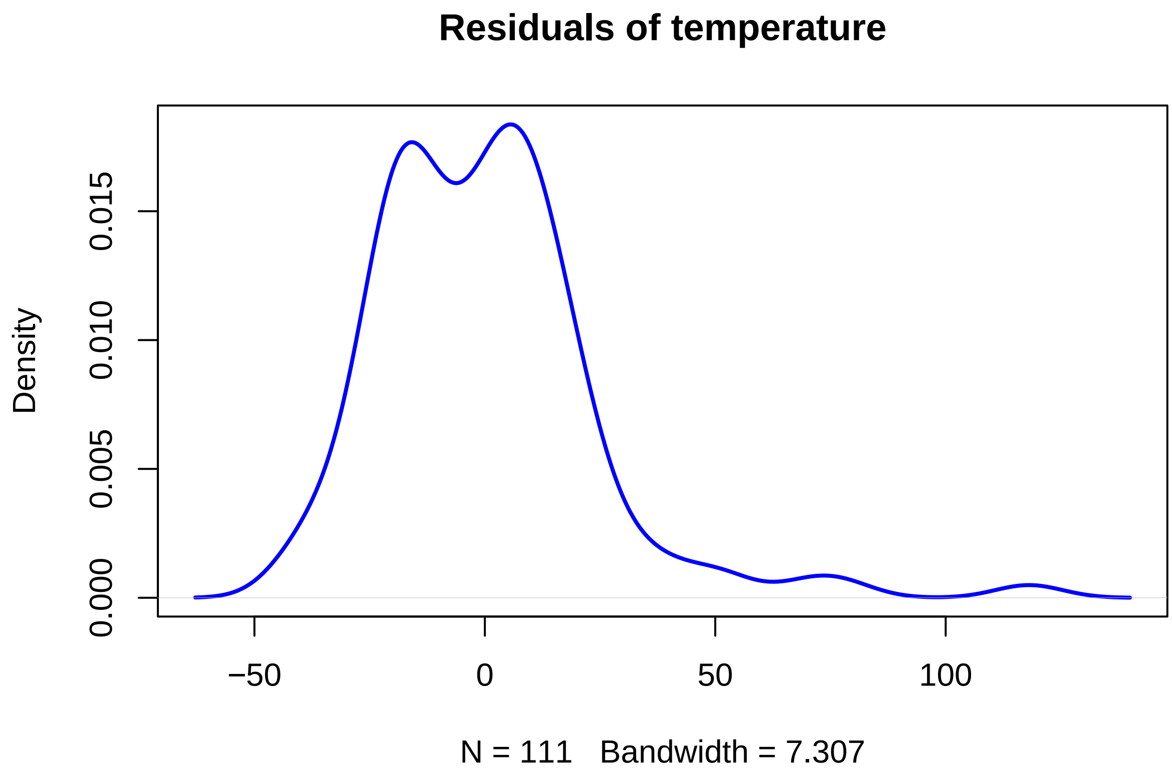
```
oz_log <- with(oz, data.frame(log(ozone), log(radiation), log(temperature),
                              log(wind)))
regr_1 = lm(ozone ~ radiation, data=oz)
regr_1_log = lm(log.ozone. ~ log.radiation., data=oz_log)
regr_2 = lm(ozone ~ temperature, data=oz)
regr_2_log = lm(log.ozone. ~ log.temperature., data=oz_log)
regr_3 = lm(ozone ~ wind, data=oz)
regr_3_log = lm(log.ozone. ~ log.wind., data=oz_log)
plot(density(regr_1$residuals), main="Residuals of radiation", col="blue", lwd=2)
```



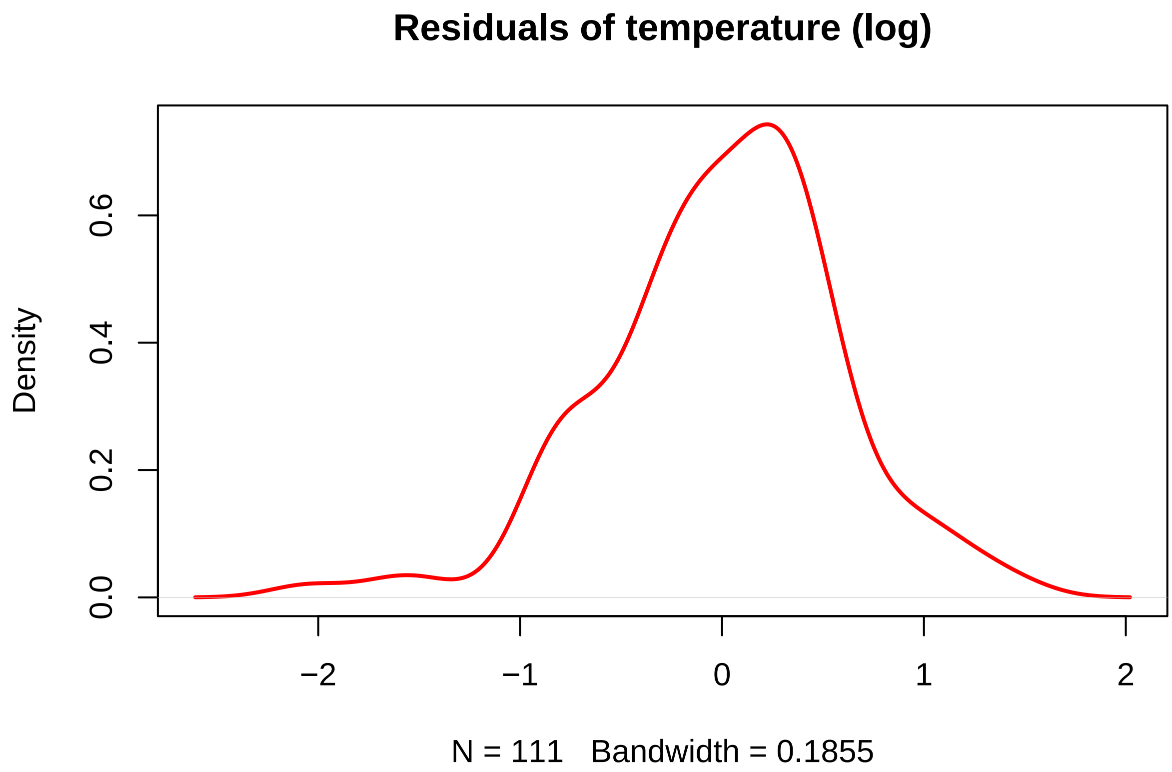

```
plot(density(regr_1_log$residuals), main="Residuals of radiation (log)",
     col="red", lwd=2)
```



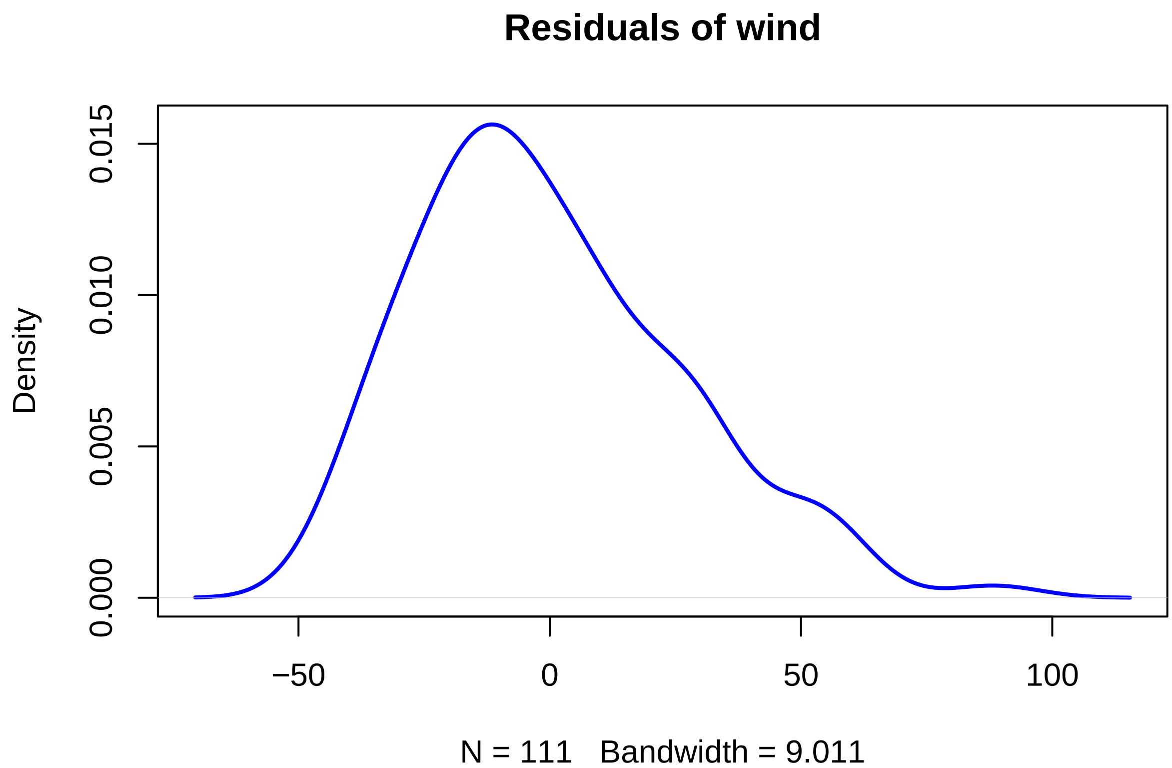

```
plot(density(regr_2$residuals), main="Residuals of temperature", col="blue", lwd=2)
```



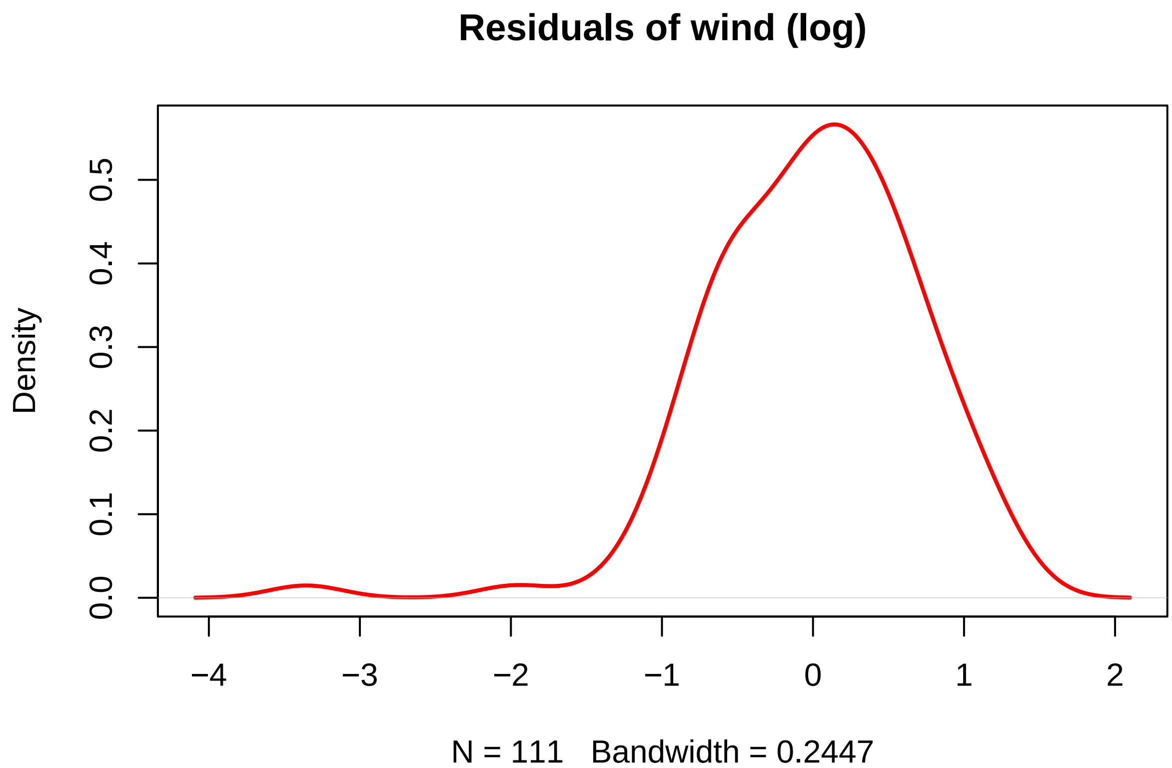

```
plot(density(regr_2_log$residuals), main="Residuals of temperature (log)",
     col="red", lwd=2)
```




```
plot(density(regr_3$residuals), main="Residuals of wind", col="blue", lwd=2)
```




```
plot(density(regr_3_log$residuals), main="Residuals of wind (log)",
     col="red", lwd=2)
```



The residuals of the log-transformed data are like having normal distribution.

(b) Regression model fitting and model summaries.

First, we compare the convention linear regression (as a baseline) and log transformed multiple regression.

```
regr <- lm(ozone ~ radiation + temperature + wind, data=oz)
summary(regr)
```

```
##
## Call:
## lm(formula = ozone ~ radiation + temperature + wind, data = oz)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -40.485 -14.210  -3.556  10.124  95.600
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -64.23208   23.04204  -2.788  0.00628 **
## radiation     0.05980    0.02318   2.580  0.01124 *
## temperature   1.65121    0.25341   6.516 2.43e-09 ***
## wind         -3.33760    0.65384  -5.105 1.45e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 21.17 on 107 degrees of freedom
## Multiple R-squared:  0.6062, Adjusted R-squared:  0.5952
## F-statistic: 54.91 on 3 and 107 DF,  p-value: < 2.2e-16
```

```
regr_log <- lm(log.ozone. ~ log.radiation. + log.temperature. + log.wind.,
               data=oz_log)
summary(regr_log)
```

```
##
## Call:
## lm(formula = log.ozone. ~ log.radiation. + log.temperature. +
##     log.wind., data = oz_log)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.63961 -0.30073 -0.00097  0.34414  1.11545
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -10.55570    2.08818  -5.055 1.79e-06 ***
## log.radiation.     0.30500    0.05868   5.198 9.73e-07 ***
## log.temperature.   3.20478    0.46019   6.964 2.79e-10 ***
## log.wind.         -0.66305    0.13751  -4.822 4.74e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4907 on 107 degrees of freedom
## Multiple R-squared:  0.6876, Adjusted R-squared:  0.6788
## F-statistic: 78.49 on 3 and 107 DF,  p-value: < 2.2e-16
```

The R^2 improved from 0.6062 to 0.6876, which is huge. However, there are about 1/3 variance that are unexplained. Perhaps adding interaction terms improves.

(c) Model selection and diagonostics

My model selection strategy is backward selection. Since the model is not large, I start with all the dependent variables `temperature`, `radiation`, and `wind` and the interaction terms `log.radiation.*log.temperature.`, `log.temperature.*log.wind.`, and `log.radiation.*log.wind.`. The stopping criteria is all the p-value are significance (<0.001).

```
regr_log <- lm(log.ozone. ~ log.radiation. + log.temperature. + log.wind. +
               log.radiation.*log.temperature. +
               log.temperature.*log.wind. +
               log.radiation.*log.wind.,
               data=oz_log)
summary(regr_log)
```

```
##
## Call:
## lm(formula = log.ozone. ~ log.radiation. + log.temperature. +
##     log.wind. + log.radiation. * log.temperature. + log.temperature. *
##     log.wind. + log.radiation. * log.wind., data = oz_log)
```

```
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.60475 -0.29291  0.00275  0.33937  1.09598
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     -16.1815    17.3019  -0.935    0.352
## log.radiation.                    0.3867     2.3635   0.164    0.870
## log.temperature.                  3.9701     4.0410   0.982    0.328
## log.wind.                         2.4115     4.7165   0.511    0.610
## log.radiation.:log.temperature.   0.0794     0.5200   0.153    0.879
## log.temperature.:log.wind.       -0.4872     1.1513  -0.423    0.673
## log.radiation.:log.wind.         -0.1788     0.2018  -0.886    0.377
##
## Residual standard error: 0.493 on 104 degrees of freedom
## Multiple R-squared:  0.6935, Adjusted R-squared:  0.6759
## F-statistic: 39.23 on 6 and 104 DF,  p-value: < 2.2e-16
```

```
regr_log <- lm(log.ozone. ~ log.radiation. + log.temperature. + log.wind. +
               log.temperature.*log.wind. +
               log.radiation.*log.wind.,
               data=oz_log)
summary(regr_log)
```

```
##
## Call:
## lm(formula = log.ozone. ~ log.radiation. + log.temperature. +
##     log.wind. + log.temperature. * log.wind. + log.radiation. *
##     log.wind., data = oz_log)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.56485 -0.29505  0.00379  0.34347  1.09681
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                -18.2241    10.9227  -1.668   0.0982 .
## log.radiation.               0.7406     0.4619   1.603   0.1119
## log.temperature.             4.4301     2.6807   1.653   0.1014
## log.wind.                    2.5714     4.5774   0.562   0.5755
## log.temperature.:log.wind.  -0.5159     1.1305  -0.456   0.6491
## log.radiation.:log.wind.    -0.1857     0.1958  -0.948   0.3451
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4907 on 105 degrees of freedom
## Multiple R-squared:  0.6935, Adjusted R-squared:  0.6789
## F-statistic: 47.51 on 5 and 105 DF,  p-value: < 2.2e-16
```

```
regr_log <- lm(log.ozone. ~ log.radiation. + log.temperature. + log.wind. +
               log.radiation.*log.wind.,
               data=oz_log)
summary(regr_log)
```

```
##
## Call:
## lm(formula = log.ozone. ~ log.radiation. + log.temperature. +
##     log.wind. + log.radiation. * log.wind., data = oz_log)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.57529 -0.30129 -0.00387  0.33570  1.10823
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)              -13.4287     2.9720  -4.518 1.63e-05 ***
## log.radiation.             0.8432     0.4019   2.098   0.0383 *
## log.temperature.           3.2249     0.4587   7.031 2.08e-10 ***
## log.wind.                  0.5222     0.8864   0.589   0.5570
## log.radiation.:log.wind.  -0.2297     0.1697  -1.353   0.1788
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4888 on 106 degrees of freedom
## Multiple R-squared:  0.6929, Adjusted R-squared:  0.6813
## F-statistic: 59.78 on 4 and 106 DF,  p-value: < 2.2e-16
```

```
regr_log <- lm(log.ozone. ~ log.radiation. + log.temperature. + log.wind.,
               data=oz_log)
summary(regr_log)
```

```
##
## Call:
## lm(formula = log.ozone. ~ log.radiation. + log.temperature. +
##     log.wind., data = oz_log)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.63961 -0.30073 -0.00097  0.34414  1.11545
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -10.55570    2.08818  -5.055 1.79e-06 ***
## log.radiation.     0.30500    0.05868   5.198 9.73e-07 ***
## log.temperature.   3.20478    0.46019   6.964 2.79e-10 ***
## log.wind.         -0.66305    0.13751  -4.822 4.74e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4907 on 107 degrees of freedom
## Multiple R-squared:  0.6876, Adjusted R-squared:  0.6788
## F-statistic: 78.49 on 3 and 107 DF,  p-value: < 2.2e-16
```

It looks like the interaction terms do not helps. Hence, simply applying the log-transformed technique is enough. Note that the best results I found on the Internet is $R^2 = 0.705$ using Poisson regression [1]. My result is near.

(d) Comments on your prediction results and scientific findings. (state at least 3 viewpoints with data evidence)

The model is

$$\log(\texttt{ozone}) = -10.55570 + 0.30500 \times \log(\texttt{radiation}) + 3.20478 \times \log(\texttt{temperature}) - 0.66305 \times \log(\texttt{wind})$$

My comments:

1. One more `radiation` leads to 30.5% increase in `ozone`, and one more `temperature` leads to 320.4% increase in `ozone`. Perhaps the more radiation and higher temperature makes oxygen easier to transform into ozone.
2. One more `wind` leads to 66.3% decrease in `ozone`. The oxygen molecules may be difficult to react with each other since the wind is strong.
3. The interactions between `radiation`, `temperature`, and `wind` are not significant. They may be not chemically related.

Problem2

(a) EDA

```
pro <- read.csv("Prostate.csv")
dim(pro)
```

```
## [1] 97 10
```

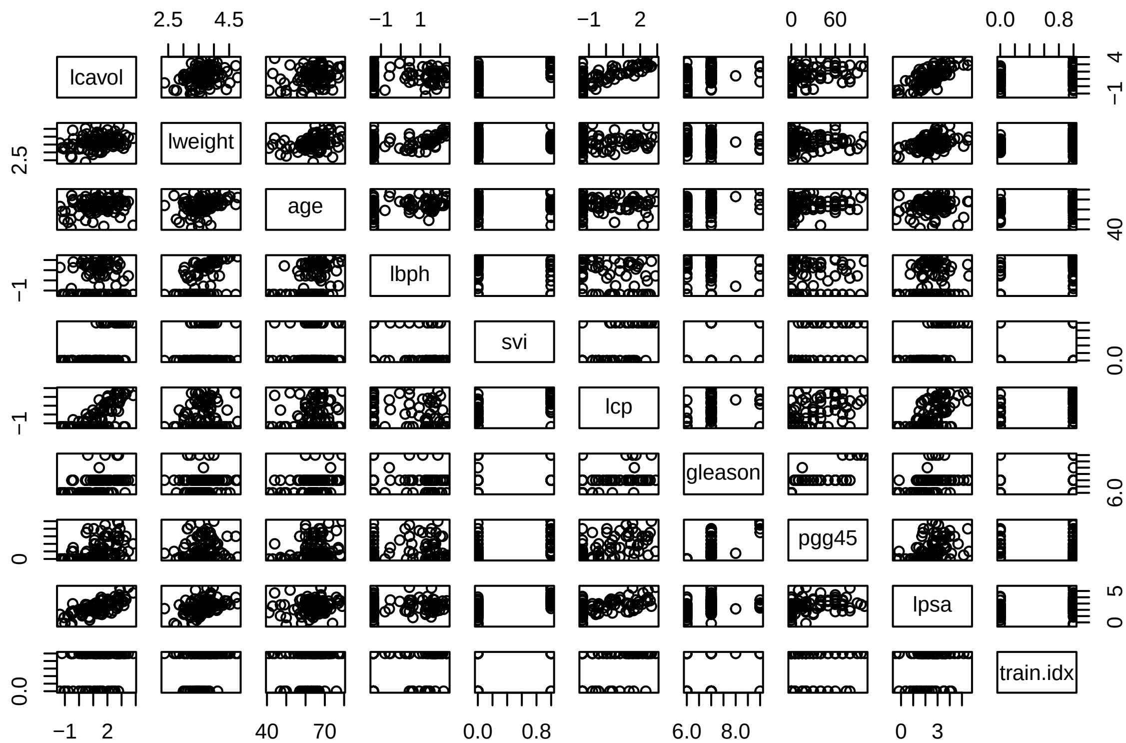
```
train <-subset(pro, train.idx == 1)
test <-subset(pro, train.idx == 0)
head(round(train,3))
```

```
##    lcavol lweight age   lbph svi    lcp gleason pgg45   lpsa train.idx
## 1  -0.580   2.769  50 -1.386   0 -1.386       6     0 -0.431         1
## 3  -0.511   2.691  74 -1.386   0 -1.386       7    20 -0.163         1
## 4  -1.204   3.283  58 -1.386   0 -1.386       6     0 -0.163         1
## 8   0.693   3.540  58  1.537   0 -1.386       6     0  0.854         1
## 10  0.223   3.245  63 -1.386   0 -1.386       6     0  1.047         1
## 15  1.206   3.442  57 -1.386   0 -0.431       7     5  1.399         1
```

```
head(round(test,3))
```

```
##    lcavol lweight age   lbph svi    lcp gleason pgg45   lpsa train.idx
## 2  -0.994   3.320  58 -1.386   0 -1.386       6     0 -0.163         0
## 5   0.751   3.432  62 -1.386   0 -1.386       6     0  0.372         0
## 6  -1.050   3.229  50 -1.386   0 -1.386       6     0  0.765         0
## 7   0.737   3.474  64  0.615   0 -1.386       6     0  0.765         0
## 9  -0.777   3.540  47 -1.386   0 -1.386       6     0  1.047         0
## 11  0.255   3.604  65 -1.386   0 -1.386       6     0  1.267         0
```

```
pairs(pro)
```



This dataset is about cancer. The dependent `gleason` is not a category but a score. The target `lpsa` looks having linear relation with `lcp` and `lcavol`.

Since there are many variables now, including all the potential interaction terms is not reasonable. My model selection strategy is:

(1) Using backward selection, until all the p-values <0.01.

(2) Investigate the interaction for the remaining variables.

(3) Compared the model with log-transformed model if necessary. Note that some variables has already log-transformed.

(b) Determine a good regression model for predicting `lpsa`

Using backward selection

```
regr_baseline <- lm(lpsa ~ lcavol + lweight + age + lbph + factor(svi) + lcp + gleason
           + pgg45, data=train)
summary(regr_baseline)
```

```
##
## Call:
```

```
## lm(formula = lpsa ~ lcavol + lweight + age + lbph + factor(svi) +
##     lcp + gleason + pgg45, data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.7239 -0.3500 -0.0441  0.3290  1.5922
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.1871196  1.4546024   0.129   0.8981
## lcavol        0.5992435  0.1052536   5.693  3.8e-07 ***
## lweight       0.6500283  0.2166712   3.000   0.0039 **
## age          -0.0176933  0.0129691  -1.364   0.1775
## lbph          0.0423207  0.0673668   0.628   0.5322
## factor(svi)1  0.5851166  0.3076684   1.902   0.0619 .
## lcp          -0.0558661  0.1216291  -0.459   0.6476
## gleason       0.0239192  0.1824911   0.131   0.8962
## pgg45        -0.0001328  0.0046882  -0.028   0.9775
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6912 on 61 degrees of freedom
## Multiple R-squared:  0.6691, Adjusted R-squared:  0.6257
## F-statistic: 15.42 on 8 and 61 DF,  p-value: 4.03e-12
```

```r
regr <- lm(lpsa ~ lcavol + lweight + age + lbph + factor(svi) + lcp + gleason
           , data=train)
summary(regr)
```

```
##
## Call:
## lm(formula = lpsa ~ lcavol + lweight + age + lbph + factor(svi) +
##     lcp + gleason, data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.72232 -0.35124 -0.04014  0.33002  1.58754
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.20410    1.31469   0.155   0.8771
## lcavol        0.59955    0.10384   5.774 2.67e-07 ***
## lweight       0.65019    0.21484   3.026   0.0036 **
## age          -0.01775    0.01274  -1.393   0.1685
## lbph          0.04224    0.06676   0.633   0.5293
## factor(svi)1  0.58442    0.30419   1.921   0.0593 .
## lcp          -0.05685    0.11565  -0.492   0.6248
## gleason       0.02127    0.15544   0.137   0.8916
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6856 on 62 degrees of freedom
## Multiple R-squared:  0.6691, Adjusted R-squared:  0.6317
## F-statistic: 17.91 on 7 and 62 DF,  p-value: 9.078e-13
```

```r
regr <- lm(lpsa ~ lcavol + lweight + age + lbph + factor(svi) + lcp, data=train)
summary(regr)
```

```
##
## Call:
## lm(formula = lpsa ~ lcavol + lweight + age + lbph + factor(svi) +
##     lcp, data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.72632 -0.36092 -0.03876  0.32535  1.58443
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.32903    0.93857   0.351  0.72709
## lcavol        0.60051    0.10279   5.842 1.97e-07 ***
## lweight       0.64591    0.21089   3.063  0.00322 **
## age          -0.01717    0.01193  -1.440  0.15496
## lbph          0.04162    0.06608   0.630  0.53113
## factor(svi)1  0.57357    0.29138   1.968  0.05342 .
## lcp          -0.04839    0.09697  -0.499  0.61952
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6802 on 63 degrees of freedom
## Multiple R-squared:  0.669,  Adjusted R-squared:  0.6375
## F-statistic: 21.22 on 6 and 63 DF,  p-value: 1.884e-13
```

```r
regr <- lm(lpsa ~ lcavol + lweight + age + lbph + factor(svi), data=train)
summary(regr)
```

```
##
## Call:
## lm(formula = lpsa ~ lcavol + lweight + age + lbph + factor(svi),
##     data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.72191 -0.36831 -0.03299  0.34030  1.61006
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.33779    0.93289   0.362  0.71848
## lcavol        0.57702    0.09085   6.351 2.52e-08 ***
## lweight       0.65190    0.20931   3.114  0.00276 **
## age          -0.01672    0.01182  -1.414  0.16222
## lbph          0.03685    0.06500   0.567  0.57276
## factor(svi)1  0.48553    0.23054   2.106  0.03912 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6762 on 64 degrees of freedom
```

```
## Multiple R-squared:  0.6677, Adjusted R-squared:  0.6417
## F-statistic: 25.72 on 5 and 64 DF,  p-value: 3.941e-14
```

```
regr <- lm(lpsa ~ lcavol + lweight + age + factor(svi), data=train)
summary(regr)
```

```
##
## Call:
## lm(formula = lpsa ~ lcavol + lweight + age + factor(svi), data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.6803 -0.3325 -0.0359  0.3513  1.6723
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.07341    0.80373   0.091 0.927502
## lcavol        0.57443    0.09026   6.364 2.27e-08 ***
## lweight       0.69433    0.19445   3.571 0.000676 ***
## age          -0.01480    0.01127  -1.313 0.193751
## factor(svi)1  0.46175    0.22550   2.048 0.044638 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6727 on 65 degrees of freedom
## Multiple R-squared:  0.666,  Adjusted R-squared:  0.6455
## F-statistic:  32.4 on 4 and 65 DF,  p-value: 7.524e-15
```

```
regr <- lm(lpsa ~ lcavol + lweight + factor(svi), data=train)
summary(regr)
```

```
##
## Call:
## lm(formula = lpsa ~ lcavol + lweight + factor(svi), data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.69204 -0.37619 -0.04593  0.36557  1.63435
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.54339    0.65575  -0.829  0.41029
## lcavol        0.56620    0.09053   6.254 3.35e-08 ***
## lweight       0.60441    0.18299   3.303  0.00155 **
## factor(svi)1  0.47345    0.22656   2.090  0.04050 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6764 on 66 degrees of freedom
## Multiple R-squared:  0.6571, Adjusted R-squared:  0.6416
## F-statistic: 42.17 on 3 and 66 DF,  p-value: 2.439e-15
```

Now the remaining dependent variables, say `lcavol`, `lweight`, and `svi` are significant. Let's take the interaction terms into consideration. Using backward selection again to drop the insignificant interaction terms.

```
regr <- lm(lpsa ~ lcavol + lweight + factor(svi)
           + lcavol*lweight + lweight*factor(svi) + lcavol*factor(svi), data=train)
summary(regr)
```

```
##
## Call:
## lm(formula = lpsa ~ lcavol + lweight + factor(svi) + lcavol *
##     lweight + lweight * factor(svi) + lcavol * factor(svi), data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.44007 -0.37345 -0.03987  0.44306  1.57354
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)            -1.6977     0.8907  -1.906   0.0612 .
## lcavol                  1.3714     0.7494   1.830   0.0720 .
## lweight                 0.9415     0.2518   3.740   0.0004 ***
## factor(svi)1            1.9373     2.3594   0.821   0.4147
## lcavol:lweight         -0.2317     0.2027  -1.143   0.2575
## lweight:factor(svi)1  -0.5700     0.6305  -0.904   0.3694
## lcavol:factor(svi)1    0.2877     0.2580   1.115   0.2690
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6625 on 63 degrees of freedom
## Multiple R-squared:  0.686,  Adjusted R-squared:  0.6561
## F-statistic: 22.94 on 6 and 63 DF,  p-value: 3.762e-14
```

```
regr <- lm(lpsa ~ lcavol + lweight + factor(svi)
           + lcavol*lweight + lcavol*factor(svi), data=train)
summary(regr)
```

```
##
## Call:
## lm(formula = lpsa ~ lcavol + lweight + factor(svi) + lcavol *
##     lweight + lcavol * factor(svi), data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.54219 -0.39618 -0.03854  0.37846  1.57600
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)           -1.8126     0.8803  -2.059 0.043569 *
## lcavol                 1.7859     0.5920   3.017 0.003661 **
## lweight                0.9747     0.2487   3.919 0.000219 ***
## factor(svi)1          -0.1132     0.6489  -0.174 0.862032
## lcavol:lweight        -0.3426     0.1611  -2.127 0.037317 *
```

```
## lcavol:factor(svi)1   0.2526     0.2547   0.992 0.325063
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6616 on 64 degrees of freedom
## Multiple R-squared:  0.6819, Adjusted R-squared:  0.657
## F-statistic: 27.44 on 5 and 64 DF,  p-value: 1.002e-14
```

```
regr_trained <- lm(lpsa ~ lcavol + lweight + factor(svi) + lcavol*lweight, data=train)
summary(regr_trained)
```

```
##
## Call:
## lm(formula = lpsa ~ lcavol + lweight + factor(svi) + lcavol *
##     lweight, data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.54002 -0.42961 -0.04901  0.38347  1.63654
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -1.7396     0.8771  -1.983 0.051565 .
## lcavol            1.7308     0.5893   2.937 0.004576 **
## lweight           0.9439     0.2467   3.826 0.000295 ***
## factor(svi)1      0.4915     0.2218   2.216 0.030169 *
## lcavol:lweight   -0.3183     0.1592  -1.999 0.049779 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6615 on 65 degrees of freedom
## Multiple R-squared:  0.677,  Adjusted R-squared:  0.6571
## F-statistic: 34.06 on 4 and 65 DF,  p-value: 2.574e-15
```

Hence our model looks like

$$\texttt{lpsa} = -1.7396 + 1.7308 \times \texttt{lcavol} + 0.9439 \times \texttt{lweight} + 0.4915 \times \texttt{svi} - 0.3183 \times \texttt{lcavol} * \texttt{lweight}$$

Compare to the first model, which $R^2 = 0.6691$. This model not only improved R^2 to 0.677 but also more easily to interpret. Also, since the variables `lpsa`, `lcavol` and `lweight` has already log-transformed, for interpretation, it is not necessary to to log-transformed again.

(c) Describe the important main effects and interaction effects.

1. The main effects are `lcavol`, `lweight` and `svi`
2. The interaction term is `lcavol*lweight`. This suggest that an increase in cancer volume of 1 unit is associated with increased prostate weight of 173.08%; An increase in prostate weight of 1 unit is associated with increased cancer volume of 94.39%.

(d) Predict lpsa for the validation data set based on the fitted model, with their prediction intervals. And compared the prediction results to the true observations. Comment on your model performance.

```r
test_actual <- test$lpsa
test_pred <- predict(regr_trained, test)
err <- test_actual - test_pred # error on testing set
NRMSE <- sqrt(mean((test_pred - test_actual)^2) / mean(test_actual^2)) # RMS
```

```r
test_actual <- test$lpsa
test_pred <- predict(regr_baseline, test)
NRMSE <- sqrt(mean((test_pred - test_actual)^2) / mean(test_actual^2)) # RMS
```

My model gives NRMSE=0.3503, while the baseline model gives NRMSE=0.3363. This shows that our model still fits well.

Reference

[1] Finding a Suitable Linear Model for Ozone Prediction, https://www.datascienceblog.net/post/machine-learning/improving_ozone_prediction/