# 36-401 Class Notes 11 October 2001

**Case Study**

Prediction Models for Personal Ozone Exposure Assessment, Wypij and Liu, in *Case Studies in Biometry*, Wiley & Sons, 1994.

Exposure to elevated levels of ozone is known to cause increased incidence of cough, chest pain, and other respiratory symptoms and decreased lung function. Outdoor ozone concentrations have been well studied, but personal (individual) exposure is less well understood. It would be an aid to public health to study the relationships between personal exposure and various diseases. A personal ozone monitor is expensive and cumbersome. Other ozone monitoring equipment is much less expensive and easier to use. The current study is designed to find a predictive model for personal ozone exposure.

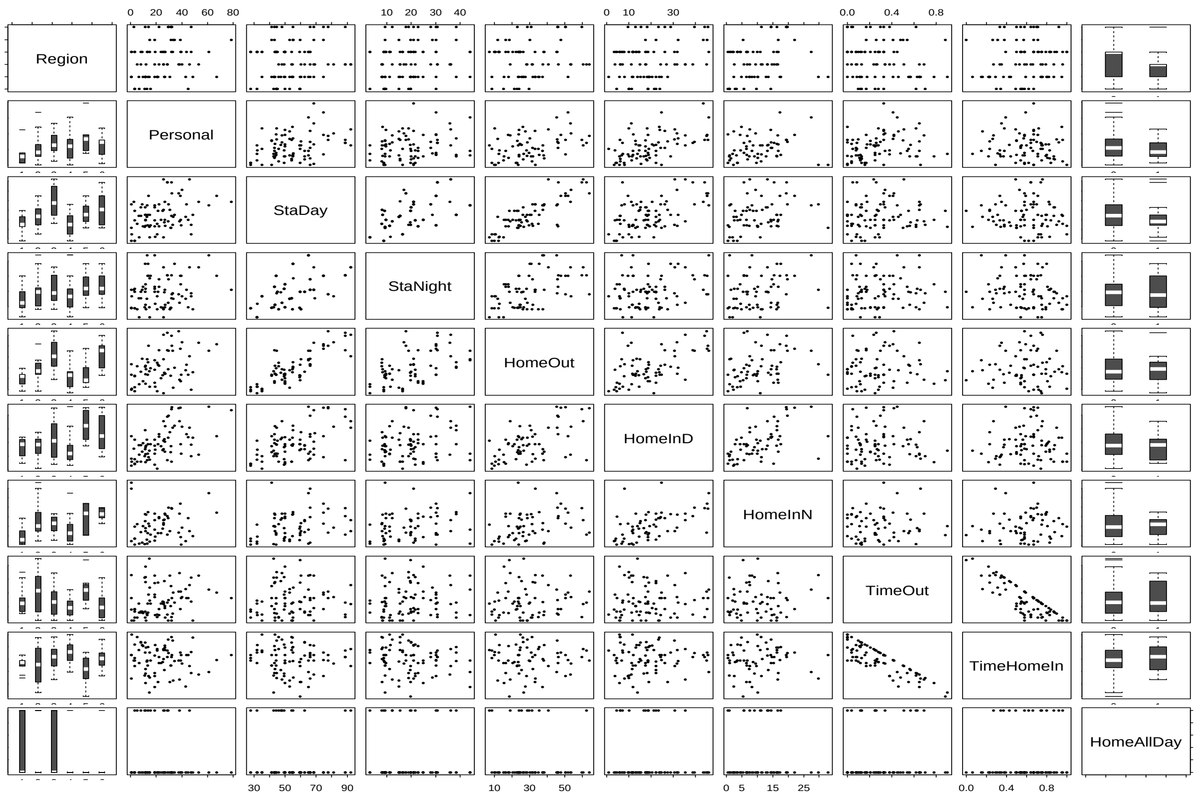
For each of 23 randomly chosen children who wore the monitors for 12 hours, we have

1. Subject identification number
2. Date [in MM/DD/YY format]
3. Home region, ranging from 1 to 6
4. 12-hour (8AM-8PM) average daytime personal ozone concentration (ppb)
5. 12-hour average daytime continuous ozone concentration at the stationary site (ppb)
6. 12-hour average nighttime continuous ozone concentration at the stationary site (ppb)
7. 24-hour average home outdoor passive ozone concentration (ppb)
8. 12-hour average home indoor daytime passive ozone concentration (ppb)
9. 12-hour average home indoor nighttime passive ozone concentration (ppb)
10. Fraction of time spent anywhere outdoors
11. Fraction of time spent at home indoors
12. Indicator variable for whether the child stayed near the home for the whole day

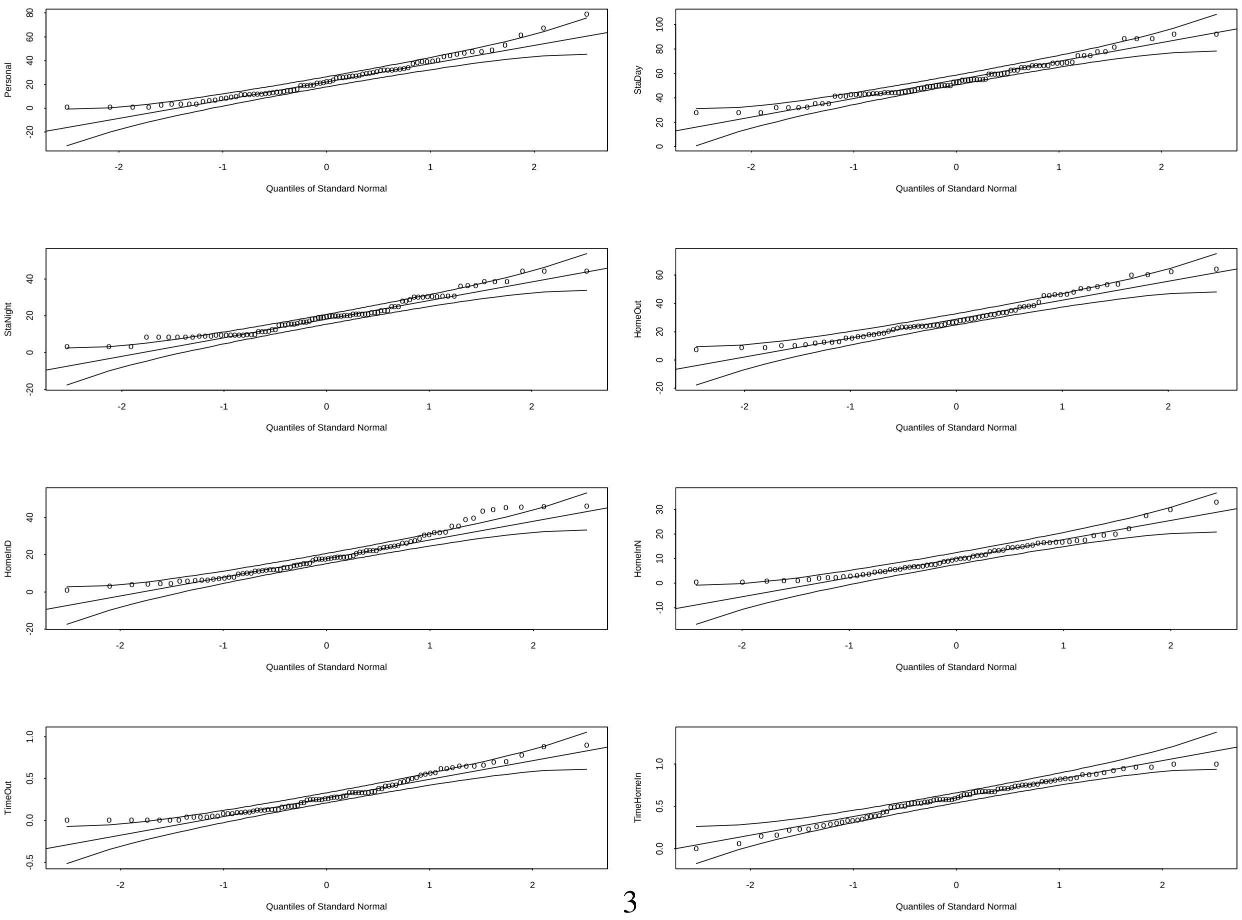
# 1 Initial Exploration

```
> oz _ read.table("ozone.dat", header=T, na.strings=".")
> oz[1:5,]
  Subject     Date Region Personal StaDay StaNight HomeOut HomeInD HomeInN
1       1 07/08/91      2    26.29  47.33    19.58   25.36   22.29   13.51
2       1 07/09/91      2    14.63  42.58     9.42   26.74   13.97   10.11
3       1 07/10/91      2    21.45  59.55    19.83   33.20   18.96    6.55
4       1 07/11/91      2    43.07  52.92    15.08   27.62      NA    2.37
5       2 07/09/91      1    11.18  42.58     9.42      NA   10.17    0.87

  TimeOut TimeHomeIn HomeAllDay
1    0.57       0.43          1
2    0.90       0.06          0
3    0.55       0.23          0
4    0.46       0.16          0
5    0.62       0.34          1
```

```
> oz$Subject
 [1]  1  1  1  1  2  2  2  3  3  3  4  4  4  4  5  5  5  5  5  6  6  6  6  7  7
[26]  7  7  7  8  8  8  8  9  9  9 10 10 10 10 10 11 11 11 11 11 12 12 12 12 12
[51] 13 13 13 13 13 14 14 14 14 14 15 15 15 15 15 16 17 18 19 19 19 19 19 20 20
[76] 20 20 20 20 21 21 21 21 21 21 22 22 23 23
> table(oz$Subject)
 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23
 4 3 3 4 5 4 5 4 3  5  5  5  5  5  5  1  1  1  5  6  6  2  2
> table(oz$Region)
  1  2  3  4 5  6
 10 20 16 25 8 10
> summary(dates(as.character(oz$Date)))
  Min. 1st Qu. Median  Mean 3rd Qu.  Max.
 11511 11524   11535  11536 11547   11560  # (No. of days since 1/1/1960)
> dates(.Last.value)
[1] 07/08/1991 07/21/1991 08/01/1991 08/02/1991 08/13/1991 08/26/1991

> dim(oz)          # 89 12
> dim(na.omit(oz)) # 55 12
> unlist(lapply(oz,function(x){sum(is.na(x))}))

 Subject Date Region Personal StaDay StaNight HomeOut HomeInD HomeInN TimeOut
       0    0      0        8      0        0      18       5      24       3

 TimeHomeIn HomeAllDay
          3          3

> source("mypairs.q") # Like pairs() but creates boxplots for discrete var’s
> mypairs(oz[,3:12],mincat=6)
```



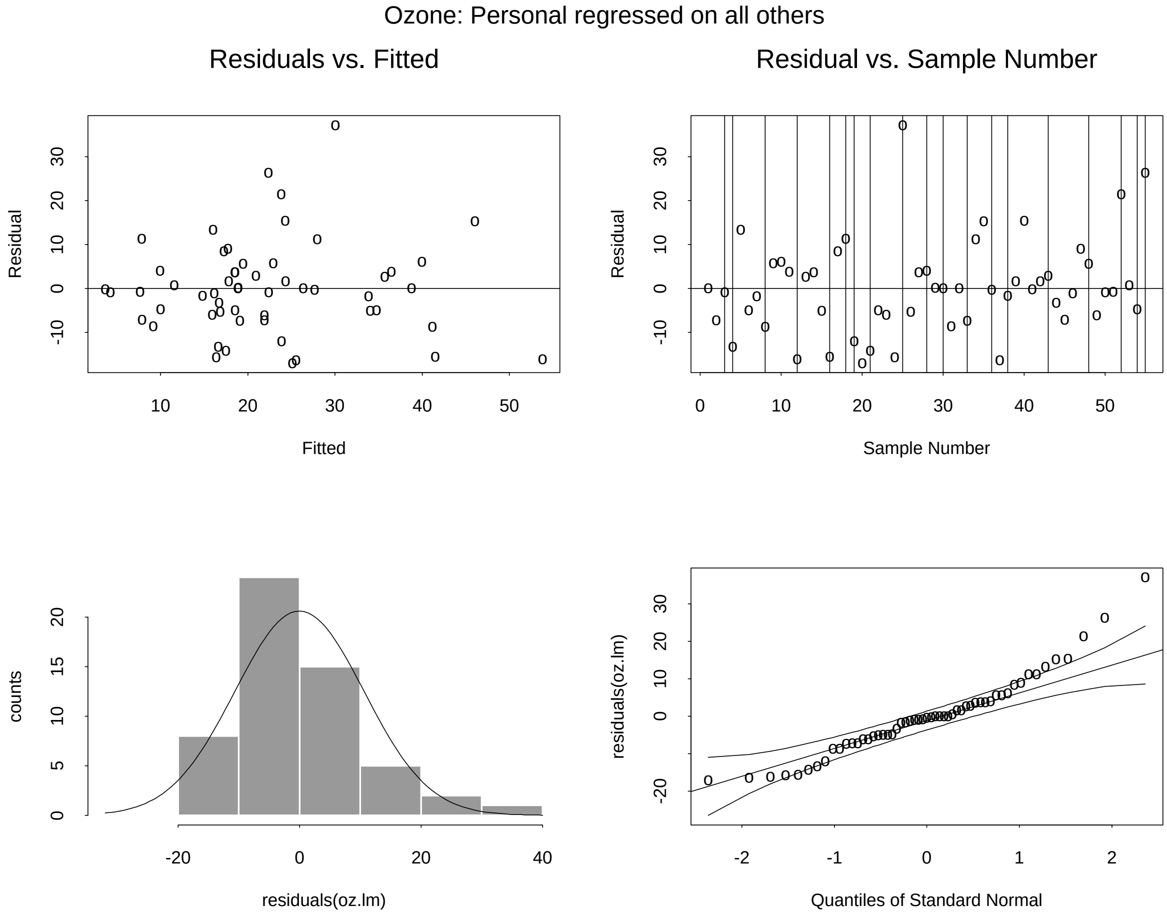
```
> round(cor(na.omit(oz[3:12])),2)
           Region Personal StaDay StaNight HomeOut HomeInD HomeInN TimeOut
    Region   1.00     0.12   0.16     0.17    0.23    0.28    0.17   -0.26
  Personal   0.12     1.00   0.45     0.25    0.43    0.61    0.21    0.18
    StaDay   0.16     0.45   1.00     0.58    0.88    0.54    0.29    0.08
  StaNight   0.17     0.25   0.58     1.00    0.55    0.29    0.17    0.09
   HomeOut   0.23     0.43   0.88     0.55    1.00    0.54    0.32    0.07
   HomeInD   0.28     0.61   0.54     0.29    0.54    1.00    0.57    0.02
   HomeInN   0.17     0.21   0.29     0.17    0.32    0.57    1.00    0.14
   TimeOut  -0.26     0.18   0.08     0.09    0.07    0.02    0.14    1.00
TimeHomeIn   0.23    -0.01  -0.04    -0.18   -0.04    0.05   -0.06   -0.82
HomeAllDay  -0.22    -0.01  -0.10    -0.02   -0.11   -0.08   -0.05    0.08

           TimeHomeIn HomeAllDay
    Region       0.23      -0.22
  Personal      -0.01      -0.01
    StaDay      -0.04      -0.10
  StaNight      -0.18      -0.02
   HomeOut      -0.04      -0.11
   HomeInD       0.05      -0.08
   HomeInN      -0.06      -0.05
   TimeOut      -0.82       0.08
TimeHomeIn       1.00       0.07
HomeAllDay       0.07       1.00

> par(mfrow=c(4,2))
> for (i in 4:11) qqn(na.omit(oz[,i]), ylab=names(oz)[i],pch="o"))
```



# 2 First Model: Try Everything!

```
> oz$r1 _ ifelse(oz$Region==1,1,0)
> oz$r2 _ ifelse(oz$Region==2,1,0)
> oz$r3 _ ifelse(oz$Region==3,1,0)
> oz$r4 _ ifelse(oz$Region==4,1,0)
> oz$r5 _ ifelse(oz$Region==5,1,0)
> oz$r6 _ ifelse(oz$Region==6,1,0)
> oz$Region_factor(oz$Region)
> ozOK_na.omit(oz)
> oz.lm _ lm(Personal ~ r1+r2+r3+r4+r5+r6 + StaDay + StaNight + HomeOut +
+ HomeInD + HomeInN + TimeOut + TimeHomeIn + HomeAllDay - 1, ozOK)
> summary(oz.lm)
 Coefficients: Value Std. Error  t value Pr(>|t|)
         r1 -17.9616  14.2326    -1.2620   0.2141
         r2 -18.8742  16.5427    -1.1409   0.2605
         r3 -18.2381  15.9873    -1.1408   0.2606
         r4 -13.1042  15.8024    -0.8293   0.4118
         r5 -19.8711  18.3739    -1.0815   0.2858
         r6 -22.9220  15.5583    -1.4733   0.1483
     StaDay   0.0199   0.2299     0.0867   0.9313
   StaNight   0.0235   0.2140     0.1098   0.9131
    HomeOut   0.2084   0.2775     0.7511   0.4569
    HomeInD   0.8239   0.2080     3.9612   0.0003
    HomeInN  -0.3695   0.3059    -1.2078   0.2341
    TimeOut  28.1214  17.7997     1.5799   0.1218
 TimeHomeIn  17.6135  17.8504     0.9867   0.3296
 HomeAllDay   1.5400   5.0228     0.3066   0.7607

 Residual standard error: 12.22 on 41 degrees of freedom
 Multiple R-Squared: 0.8464
 F-statistic: 16.13 on 14 and 41 degrees of freedom, the p-value is 2.126e-12

 Correlation of Coefficients:
                 r1      r2      r3      r4      r5      r6  StaDay StaNight
     StaDay -0.2230 -0.1683 -0.2792 -0.1949 -0.1463 -0.1353
   StaNight -0.2311 -0.2688 -0.1433 -0.3081 -0.2928 -0.2306 -0.3329
    HomeOut  0.1954  0.1263  0.0953  0.2136  0.1645  0.0129 -0.6662 -0.1853
    HomeInD  0.1129  0.2071  0.1551  0.1565  0.0615  0.1023 -0.1775  0.0141
    HomeInN -0.0225 -0.1547 -0.0746 -0.0399 -0.1461 -0.1467 -0.0030  0.0161
    TimeOut -0.7378 -0.8284 -0.7135 -0.8120 -0.7072 -0.6892 -0.1062  0.3122
 TimeHomeIn -0.8082 -0.8372 -0.7683 -0.8958 -0.7428 -0.7721 -0.1100  0.3794
 HomeAllDay  0.1668  0.2258  0.1258  0.3076  0.2533  0.1758  0.1099 -0.3043

           HomeOut HomeInD HomeInN TimeOut TimeHomeIn
   HomeInD -0.0814
   HomeInN  0.0509 -0.4722
   TimeOut -0.0929 -0.1382 -0.0273
TimeHomeIn -0.1055 -0.1459 -0.0289  0.8533
HomeAllDay  0.1754  0.0487  0.0374 -0.3883 -0.4205

> par(mfrow=c(2,2), oma=c(0,0,2.5,0))
> plot(fitted(oz.lm),residuals(oz.lm),xlab="Fitted",ylab="Residual",pch="o")
> title("Residuals vs. Fitted")
> abline(h=0)
```

```
> plot(1:nrow(ozOK),residuals(oz.lm), xlab="Sample Number",
+ ylab="Residual",pch="o")
> title("Residual vs. Sample Number")
> abline(h=0)
> for (i in cumsum(table(ozOK$Subject))) abline(v=i)
> ndhist(residuals(oz.lm))
> qqn(residuals(oz.lm),pch="o")
> mtext("Ozone: Personal regressed on all others",outer=T,cex=1.4)
```




```
> HighRes_oz.lm$res>quantile(oz.lm$res,0.9)    # oz.lm$res = residuals(oz.lm)
> data.frame(ozOK[HighRes,], Residual=oz.lm$res[HighRes])
   Subject     Date Region Personal  StaDay StaNight  HomeOut  HomeInD  HomeInN
11       4 07/15/91      6    29.28   48.08    12.33    28.87    23.40    14.47
39      10 08/01/91      2    67.18   74.67    30.17    51.94    24.76    17.37
59      14 08/08/91      4    61.19   64.92    44.08    45.60    46.04    27.47
70      19 08/19/91      4    39.59   44.33    30.25    23.90    12.01     0.46
83      21 08/22/91      1    45.13   49.83     9.75    23.97    24.00    14.35
88      23 08/25/91      2    48.65   54.75     8.42    26.10    23.74     4.83

   TimeOut TimeHomeIn HomeAllDay Residual
11    0.00       0.92          1 13.25925
39    0.41       0.59          0 37.12735
59    0.45       0.39          0 15.15337
70    0.33       0.67          0 15.31182
83    0.33       0.67          0 21.32053
88    0.00       0.95          0 26.28720
```

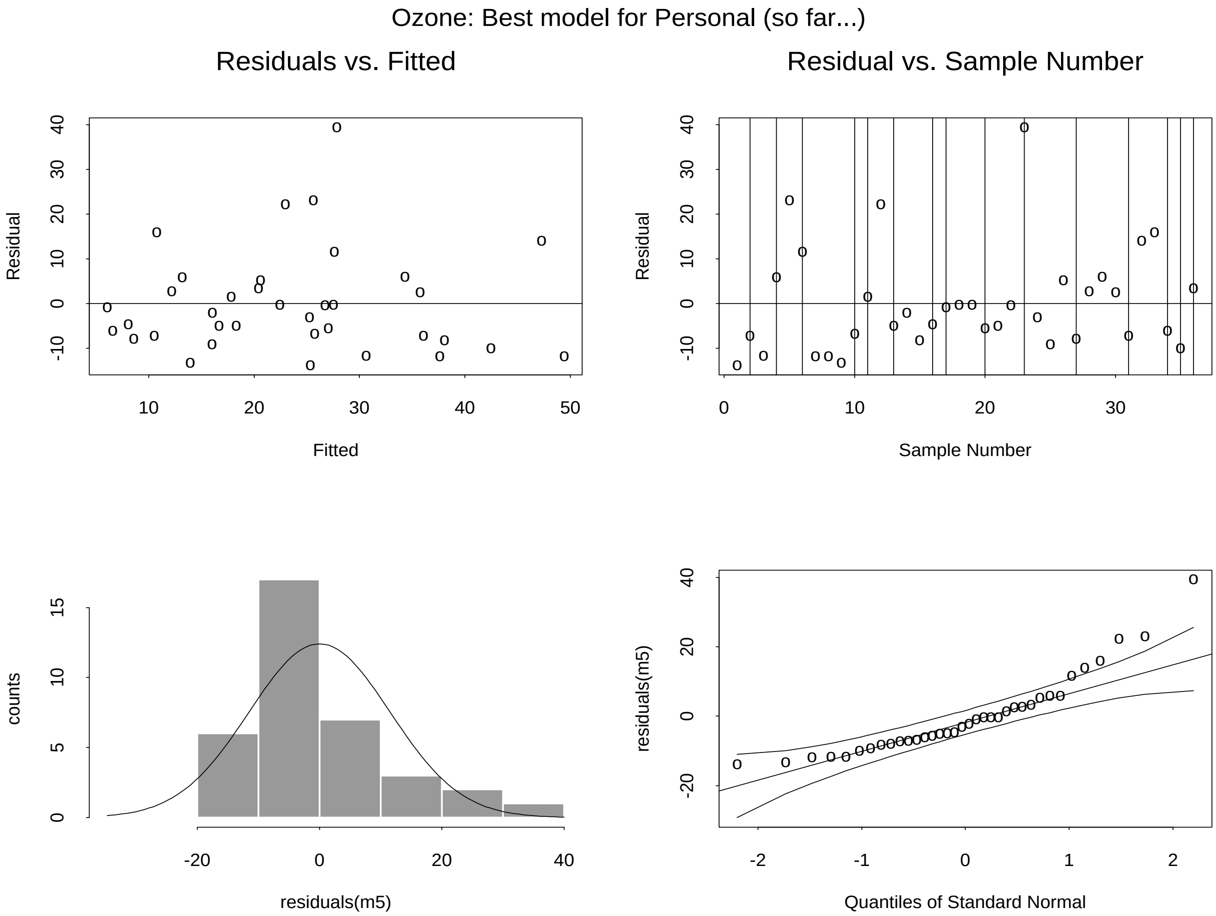
```
> names(oz)
 [1] "Subject"    "Date"       "Region"     "Personal"   "StaDay"
 [6] "StaNight"   "HomeOut"    "HomeInD"    "HomeInN"    "TimeOut"
[11] "TimeHomeIn" "HomeAllDay" "r1"         "r2"         "r3"
[16] "r4"         "r5"         "r6"
> collin(oz[,c(3,5:12)])
 StaDay 0.786
 HomeOut 0.775
 TimeOut 0.727
 TimeHomeIn 0.731
```

# 3 Try Automatic Variable Selection

Clearly we have some collinearity problems; maybe we can learn something from an automatic variable selection procedure. The idea here is to use exhaustive search to find the best model, as we have done in the past. In order to give us a good "check" on the model we choose, we will divide search for a good model on 2/3 of the data, and see how well the model predicts personal ozone readings on the remaining 1/3 of the data.

```
> samp_sample(1:nrow(ozOK), nrow(ozOK)*2/3)
> samp
 [1] 26 15 32 18 55 34 12 16 24 29 39 52 44 28  6 41 54 36 14  3 33  1 25 27 51
[26] 42 45 46 11 13 50 35 47 31  8 43
> ozFit_ozOK[samp,]
> ozCheck_ozOK[-samp,]

> # Try to find the best main effects models

> sw_stepwise(ozFit[,c(3,5:12)], ozFit$Personal, method="exh",nbest=5, plot=F)
> sum.step(sw, ozFit, "Personal", includeAIC=T, top=5, compact=T, sort=2)
         BIC    AIC   R2a StaNight HomeInD HomeInN TimeOut TimeHomeIn
1(#1) -143.6 -142.0 0.405        0       1       0       0          0
2(#1) -144.7 -142.3 0.413        0       1       1       0          0
3(#1) -145.6 -142.4 0.426        0       1       1       1          0
5(#1) -147.3 -142.5 0.453        1       1       1       1          1
4(#1) -146.5 -142.5 0.438        0       1       1       1          1

> # Now see if things get any better with some interactions

> ozFit$HTinD _ ozFit$HomeInD  * ozFit$TimeHomeIn
> ozFit$HTinN _ ozFit$HomeInN  * ozFit$TimeHomeIn
> ozFit$HTout _ ozFit$HomeOut  * ozFit$TimeOut
> ozFit$STout _ ozFit$StaNight * ozFit$TimeOut
> sw _ stepwise(ozFit[,5:16], ozFit$Personal, method="exh", nbest=5, plot=F)
> sum.step(sw, ozFit, "Personal", includeAIC=T, top=5, compact=T, sort=2)
#          BIC    AIC   R2a HomeInD HomeInN TimeHomeIn HTinN HTout STout
# 4(#1) -143.3 -139.3 0.530       1       0          1     1     0     1
# 4(#2) -143.4 -139.4 0.526       1       1          1     0     0     1
# 5(#1) -144.8 -140.0 0.524       1       0          1     1     1     1
# 5(#2) -144.9 -140.1 0.521       1       1          1     0     1     1
# 3(#1) -143.3 -140.1 0.494       1       1          0     0     0     1
```

```
> # The models with interactions seem slightly better...
> # Now we use them to predict y's in the 1/3 of the data we
> # set aside, to see which model has the best (smallest) prediction
> # error.

> m1 _ lm(Personal ~ HomeInD + TimeHomeIn + HTinN + STout, ozFit)
> m2 _ lm(Personal ~ HomeInD + HomeInN + TimeHomeIn + STout, ozFit)
> m3 _ lm(Personal ~ HomeInD + TimeHomeIn + HTinN+HTout + STout, ozFit)
> m4 _ lm(Personal ~ HomeInD + HomeInN + TimeHomeIn + HTout + STout, ozFit)
> m5 _ lm(Personal ~ HomeInD + HomeInN + STout, ozFit)

> ozCheck$HTinN _ ozCheck$HomeInN  * ozCheck$TimeHomeIn
> ozCheck$HTout _ ozCheck$HomeOut  * ozCheck$TimeOut
> ozCheck$STout _ ozCheck$StaNight * ozCheck$TimeOut
> n_nrow(ozCheck)  # 19
> round(sum((predict(m1,ozCheck)-ozCheck$Personal)^2)/(n-5),2) # 216.07
> round(sum((predict(m2,ozCheck)-ozCheck$Personal)^2)/(n-5),2) # 216.97
> round(sum((predict(m3,ozCheck)-ozCheck$Personal)^2)/(n-6),2) # 227.85
> round(sum((predict(m4,ozCheck)-ozCheck$Personal)^2)/(n-6),2) # 227.45
> round(sum((predict(m5,ozCheck)-ozCheck$Personal)^2)/(n-4),2) # 198.61

> # Model 5 does the best, so...

> summary(m5)
 Coefficients: Value Std. Error t value Pr(>|t|)
 (Intercept)  6.2882  3.9359     1.5977  0.1199
     HomeInD  0.9602  0.1981     4.8463  0.0000
     HomeInN -0.6466  0.3365    -1.9216  0.0636
       STout  0.9362  0.3737     2.5051  0.0175

 Residual standard error: 11.22 on 32 degrees of freedom
 Multiple R-Squared: 0.5376
 F-statistic: 12.4 on 3 and 32 degrees of freedom, the p-value is 1.513e-05

 Correlation of Coefficients:
         (Intercept) HomeInD HomeInN
 HomeInD -0.4644
 HomeInN -0.0901     -0.6378
   STout -0.3284     -0.0231 -0.2593
```




A variety of models were studied in an attempt to find ones that predict personal ozone exposure from cheaper and easier to obtain measurements. The predictors investigated were ozone concentration averages from standard station monitors separate from individual homes, indoor and outdoor home monitors, time spent indoors and outdoors as measured by diary records, and the interaction of time and concentration measures. The best model was chosen by calculating mean squared error of prediction for a reserved random subset of the data. The best prediction comes from using a model that includes the day and night average home indoor ozone concentrations as well as the product of fraction of time spent outdoors times the average nighttime concentration at the stationary site. This model explains only 54% of the variation in personal ozone exposure, so it may not be a satisfactory substitute for the use of the personal ozone monitors. A possible further problem with this model is that it has the counter-intuitive feature that increased home nighttime ozone is associated with decreased personal 8AM to 8PM average ozone concentration. Clearly further work needs to be done on this problem. Priorities should include collecting more data including from other cities, and considering models that take into account the possible effects of sample date and of within subject effects.