36-401 Class Notes 10/5/00
Howard Seltman

Case Study

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

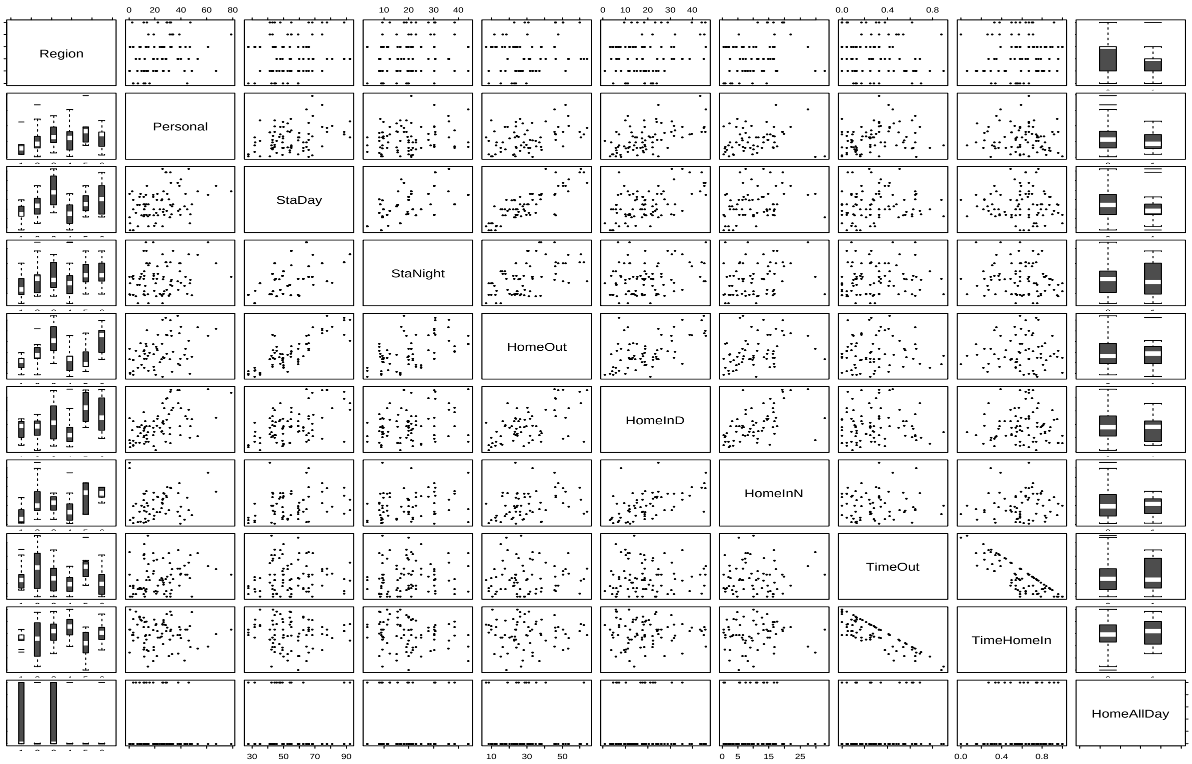
Exposure to elevated levels of ozone is known to cause increased incidence of cough, chest pain, and other respiratory symptoms and decreased lung function. Outdoor ozone concentrations have been well studied, but personal (individual) exposure is less well understood. It would be an aid to public health 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

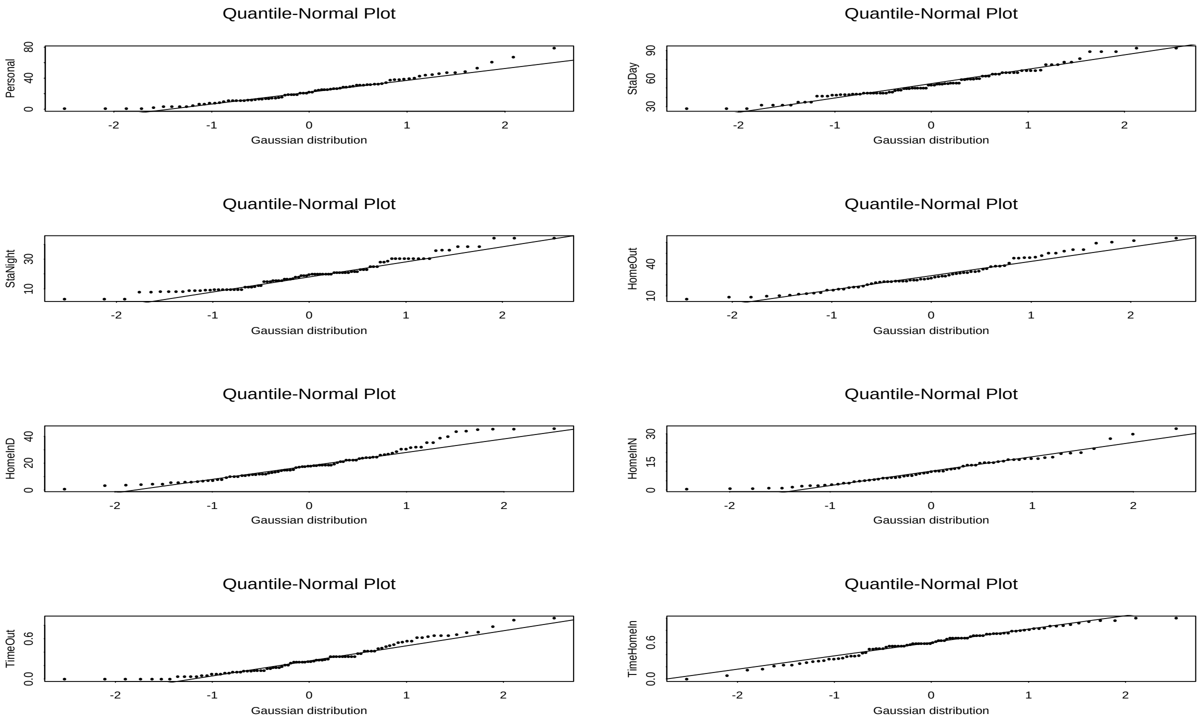
1. Subject identification number
2. Date [in MM/DD/YY format]
3. Home region, ranging from 1 to 6
4. 12-hour (8AM-8PM) average daytime personal ozone concentration (ppb)
5. 12-hour average daytime continuous ozone concentration at the stationary site (ppb)
6. 12-hour average nighttime continuous ozone 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

```
oz_read.table("ozone.dat", T, na.strings=".")
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  TimeHomeIn HomeAllDay
#        0    0      0        8      0        0      18       5      24       3           3          3
names(oz)
#  [1] "Subject"    "Date"       "Region"     "Personal"   "StaDay"    "StaNight"
#  [7]   "HomeOut"    "HomeInD"    "HomeInN"    "TimeOut"  "TimeHomeIn" "HomeAllDay"
oz[1:5,]
#   Subject     Date Region Personal StaDay StaNight HomeOut HomeInD HomeInN TimeOut TimeHomeIn HomeAllDay
# 1       1 07/08/91      2    26.29  47.33    19.58   25.36   22.29   13.51    0.57       0.43          1
# 2       1 07/09/91      2    14.63  42.58     9.42   26.74   13.97   10.11    0.90       0.06          0
# 3       1 07/10/91      2    21.45  59.55    19.83   33.20   18.96    6.55    0.55       0.23          0
# 4       1 07/11/91      2    43.07  52.92    15.08   27.62      NA    2.37    0.46       0.16          0
# 5       2 07/09/91      1    11.18  42.58     9.42      NA   10.17    0.87    0.62       0.34          1
summary(dates(as.character(oz$Date)))             table(oz$Region)
#   Min. 1st Qu. Median  Mean 3rd Qu.  Max.         #   1  2  3  4 5  6
#  11510   11520  11540 11540   11550 11560         #  10 20 16 25 8 10
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
```

```
mypairs(oz[,3:12],mincat=6)
collin(oz[c(3,5:12)])
# StaDay 0.786
# HomeOut 0.775
# TimeOut 0.727
# TimeHomeIn 0.731
stepwise(oz[,c(3,5,6,8:12)],HomeOut,method="forward", plot=F)
#       Region StaDay StaNight HomeInD HomeInN TimeOut TimeHomeIn HomeAllDay
# 1(+2)      F      T        F       F       F       F          F          F
round(cor(na.omit(oz[3:12])),2)
           Region Personal StaDay StaNight HomeOut HomeInD HomeInN TimeOut TimeHomeIn HomeAllDay
    Region   1.00     0.12   0.16     0.17    0.23    0.28    0.17   -0.26       0.23      -0.22
  Personal   0.12     1.00   0.45     0.25    0.43    0.61    0.21    0.18      -0.01      -0.01
    StaDay   0.16     0.45   1.00     0.58    0.88    0.54    0.29    0.08      -0.04      -0.10
  StaNight   0.17     0.25   0.58     1.00    0.55    0.29    0.17    0.09      -0.18      -0.02
   HomeOut   0.23     0.43   0.88     0.55    1.00    0.54    0.32    0.07      -0.04      -0.11
   HomeInD   0.28     0.61   0.54     0.29    0.54    1.00    0.57    0.02       0.05      -0.08
   HomeInN   0.17     0.21   0.29     0.17    0.32    0.57    1.00    0.14      -0.06      -0.05
   TimeOut  -0.26     0.18   0.08     0.09    0.07    0.02    0.14    1.00      -0.82       0.08
TimeHomeIn   0.23    -0.01  -0.04    -0.18   -0.04    0.05   -0.06   -0.82       1.00       0.07
HomeAllDay  -0.22    -0.01  -0.10    -0.02   -0.11   -0.08   -0.05    0.08       0.07       1.00
```

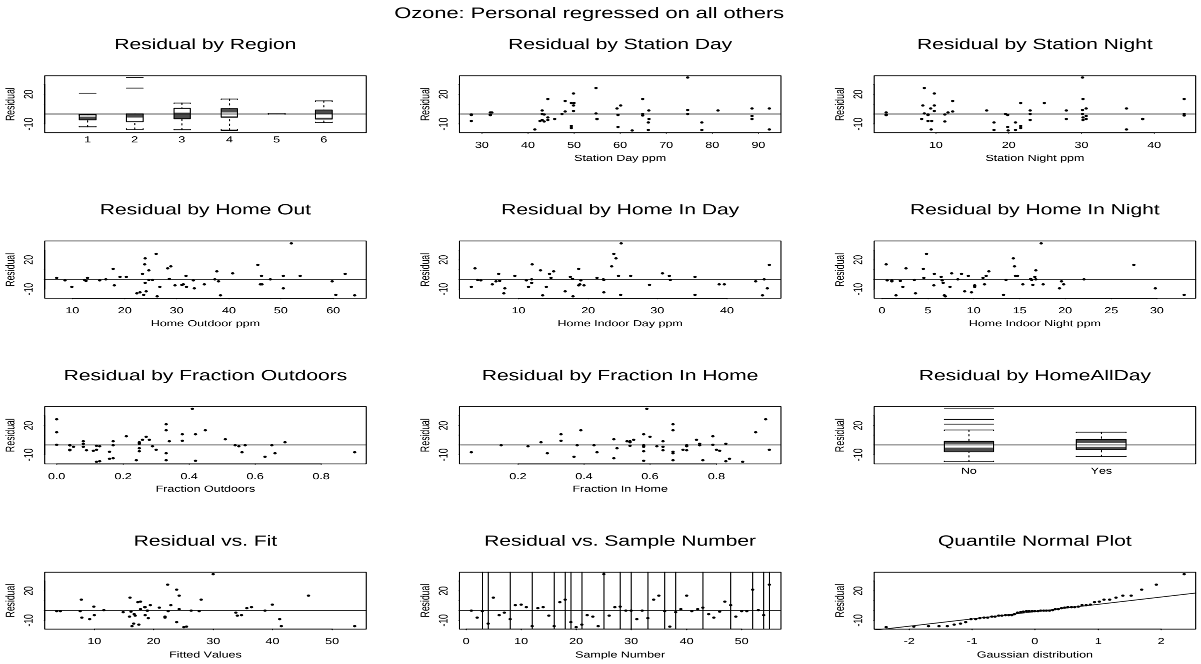


```
par(mfrow=c(4,2))
for (i in 4:11) qqn(na.omit(oz[,i]), ylab=names(oz)[i])
```




```
detach("oz");
oz$Region_factor(oz$Region)
ozOK_na.omit(oz)
oz.lm_lm(Personal~Region+StaDay+StaNight+HomeOut+HomeInD+HomeInN+TimeOut+TimeHomeIn+HomeAllDay-1, ozOK)
summary(oz.lm)
# Coefficients: Value Std. Error  t value Pr(>|t|)
#    Region1 -17.9616  14.2326    -1.2620   0.2141
#    Region2 -18.8742  16.5427    -1.1409   0.2605
#    Region3 -18.2381  15.9873    -1.1408   0.2606
#    Region4 -13.1042  15.8024    -0.8293   0.4118
#    Region5 -19.8711  18.3739    -1.0815   0.2858
#    Region6 -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:
#            Region1 Region2 Region3 Region4 Region5 Region6  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
par(mfrow=c(4,3), oma=c(0,0,2.5,0))
par(mai=c(0.5,0.5,0.5,0.1))
boxplot(split(oz.lm$res,ozOK$Region), ylab="Residual", names=paste(1:6))
abline(h=0)
title("Residual by Region")
mtext("Ozone: Personal regressed on all others", outer=T, cex=1.4)
plot(ozOK$StaDay, oz.lm$res, xlab="Station Day ppm", ylab="Residual")
title("Residual by Station Day")
abline(h=0)
...
plot(1:nrow(ozOK),oz.lm$res, xlab="Sample Number",ylab="Residual")
title("Residual vs. Sample Number")
abline(h=0)
for (i in cumsum(table(ozOK$Subject))) abline(v=i)
qqn(oz.lm$res, ylab="Residual")
```

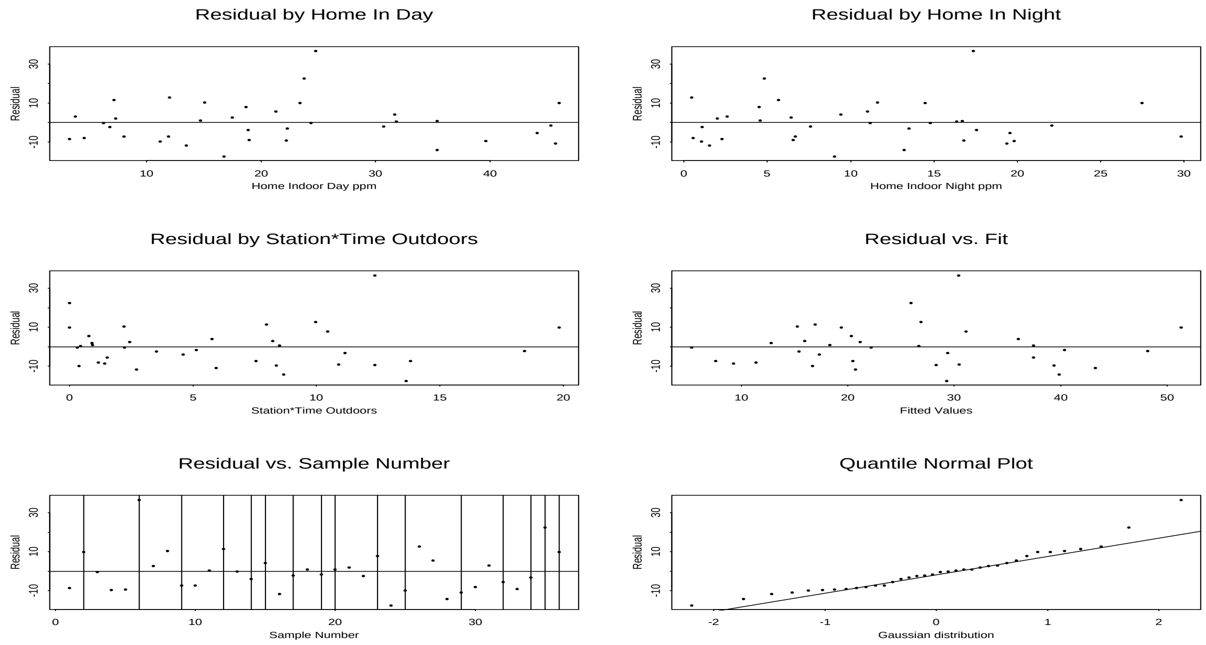


```
HighRes_oz.lm$res>quantile(oz.lm$res,0.9)
data.frame(ozOK[HighRes,], Residual=oz.lm$res[HighRes])
#    Subject     Date Region Personal StaDay StaNight HomeOut HomeInD HomeInN TimeOut TimeHomeIn HomeAllDay FRegion Residual
# 11       4 07/15/91      6    29.28  48.08    12.33   28.87   23.40   14.47    0.00       0.92          1       6 13.25925
# 39      10 08/01/91      2    67.18  74.67    30.17   51.94   24.76   17.37    0.41       0.59          0       2 37.12735
# 59      14 08/08/91      4    61.19  64.92    44.08   45.60   46.04   27.47    0.45       0.39          0       4 15.15337
# 70      19 08/19/91      4    39.59  44.33    30.25   23.90   12.01    0.46    0.33       0.67          0       4 15.31182
# 83      21 08/22/91      1    45.13  49.83     9.75   23.97   24.00   14.35    0.33       0.67          0       1 21.32053
# 88      23 08/25/91      2    48.65  54.75     8.42   26.10   23.74    4.83    0.00       0.95          0       2 26.28720
```

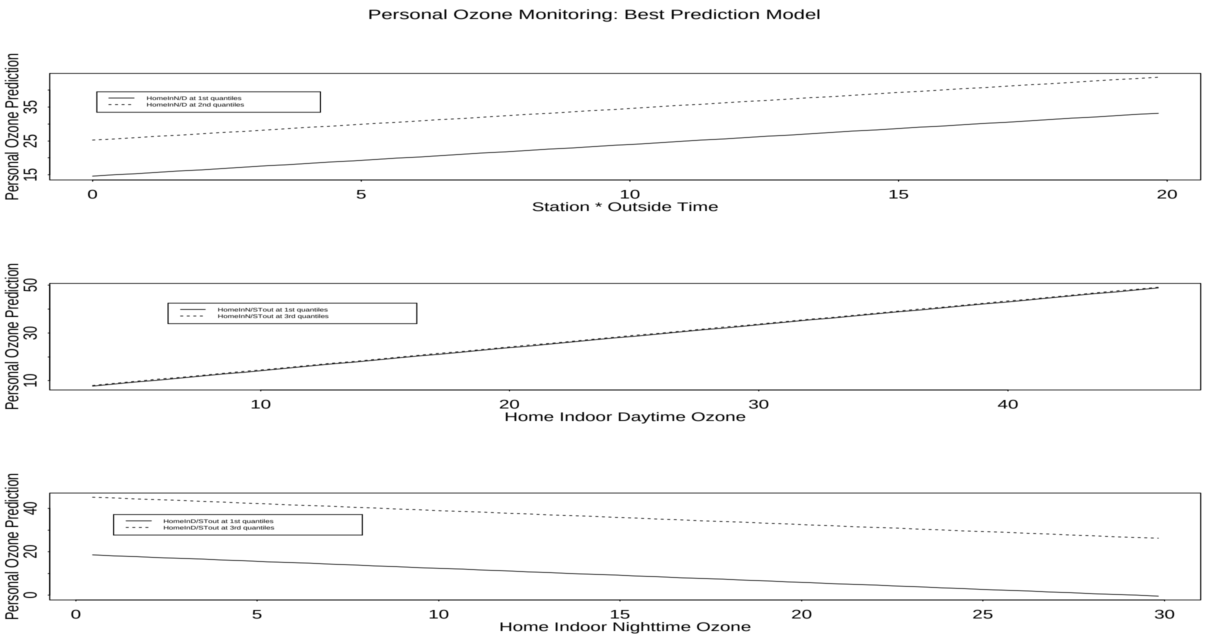
```
samp_sample(1:nrow(ozOK), nrow(ozOK)*2/3)
samp
#  [1] 45  5 54  6 29 25 43 17 31 44 36  9 14 22 11 37 10 39 30 13 46 38 34 19 51
# [26] 40 42 16  8 50 18  7  3  1 55 35
ozFit_ozOK[samp,]
ozCheck_ozOK[-samp,]
```

```
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
```

```
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

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(m1,ozCheck)-ozCheck$Personal)^2)/(n-5),2) # 216.97
round(sum((predict(m1,ozCheck)-ozCheck$Personal)^2)/(n-6),2) # 227.85
round(sum((predict(m1,ozCheck)-ozCheck$Personal)^2)/(n-6),2) # 227.45
round(sum((predict(m1,ozCheck)-ozCheck$Personal)^2)/(n-4),2) # 198.61

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
```




```
par(mfrow=c(3,1))
s1_summary(ozFit$HomeInN)
s2_summary(ozFit$HomeInD)
s3_summary(ozFit$STout)
SToutSeq_seq(min(ozFit$STout), max(ozFit$STout), len=50)
new1_data.frame(HomeInD=rep(s1[2],50), HomeInN=rep(s2[2],50), STout=SToutSeq)
new2_data.frame(HomeInD=rep(s1[5],50), HomeInN=rep(s2[5],50), STout=SToutSeq)
y1_predict(m5, new1)
y2_predict(m5, new2)
plot(range(SToutSeq), range(y1,y2), xlab="Station * Outside Time",
     ylab="Personal Ozone Prediction", type="n", cex=1.2)
lines(SToutSeq, y1, lty=1)
lines(SToutSeq, y2, lty=2)
legend(locator(1), c("HomeInN/D at 1st quantiles","HomeInN/D at 2nd quantiles"), lty=1:2)
```



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.