# STAT460 – Solution 2
## Due: Feb. 4 at the start of class.

1. These data record the level of atmospheric ozone concentration from eight daily meteorological measurements made in the Los Angeles basin in 1976. We have the 330 complete cases[1]. We want to find climate/weather factors that impact ozone readings. Ozone is a hazardous byproduct of burning fossil fuels and can harm lung function. The data set for this problem is:

| Variable Name | Definition |
|---|---|
| ozone | Log Maximum Ozone |
| vh | Vandenberg 500 mb Height |
| wind | Wind Speed (mph) |
| humidity | Humidity (%) |
| temp | Sandburg AFB Temperature |
| ibh | Inversion Base Height |
| dpg | Daggot Pressure Gradient |
| ibt | Inversion Base Temperature |
| vis | Visibility (miles) |
| doy | Day of the Year |

```
#enter data and define variables
ozone = read.table('/Users/darrenho/Dropbox/teaching/STAT460/data/LAozone.txt',sep=",",hea

Y = ozone$ozone
X = ozone[,names(ozone)!=c('ozone')]
```

(a) Report the full linear regression of ozone on the other variables. Comment.

```
> summary(lm(Y˜.,data=X))
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 18.3792938 29.5045242   0.623  0.53377
vh          -0.0051340  0.0053950  -0.952  0.34200
wind        -0.0198304  0.1238829  -0.160  0.87292
humidity     0.0804923  0.0188345   4.274 2.54e-05 ***
temp         0.2743349  0.0497361   5.516 7.17e-08 ***
ibh         -0.0002497  0.0002950  -0.846  0.39798
dpg         -0.0036968  0.0112925  -0.327  0.74360
ibt          0.0292640  0.0136115   2.150  0.03231 *
vis         -0.0080742  0.0037565  -2.149  0.03235 *
doy         -0.0088490  0.0027199  -3.253  0.00126 **
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

[1] Note that this dataset violates some assumptions of linear regression. Do you know which one(s)? For this assignment, ignore this fact.

```
Residual standard error: 4.441 on 320 degrees of freedom
Multiple R-squared: 0.7011,Adjusted R-squared: 0.6927
F-statistic:  83.4 on 9 and 320 DF,  p-value: < 2.2e-16
```

(b) Report the selected variables using the following model selection techniques (use: either BIC, AIC, or Mallow's Cp)

i. All subsets (plot this using `regsubsets, plot`)

ii. Forward stepwise...

A. ... using the `step` approach

```
Step:  AIC=988.83
Y ~ temp + ibh + humidity + doy + ibt + vis

       Df Sum of Sq    RSS    AIC
<none>              6330.5 988.83
+ vh    1   16.4307 6314.1 989.98
+ dpg   1    1.4240 6329.1 990.76
+ wind  1    0.0003 6330.5 990.83

Call:
lm(formula = Y ~ temp + ibh + humidity + doy + ibt + vis, data = X)

Coefficients:
(Intercept)         temp          ibh     humidity          doy
 -9.4114549    0.2579707   -0.0003188    0.0798926   -0.0089918
        ibt          vis
  0.0250867   -0.0078422
```

B. ... using the `regsubsets` approach

```
> regfit.for = regsubsets ( x = X,y = Y, nvmax =19 ,method ="forward")
> regfit.for.sum = summary(regfit.for)
> regfit.for.sum$which[which.min(regfit.for.sum$cp),]
(Intercept)          vh        wind    humidity        temp
       TRUE       FALSE       FALSE        TRUE        TRUE
        ibh         dpg         ibt         vis         doy
       TRUE       FALSE        TRUE        TRUE        TRUE
```

iii. Backwards stepwise (choose any method)

```
> regfit.bac = regsubsets ( x = X,y = Y, nvmax =19 ,method ="backward")
> regfit.bac.sum = summary(regfit.bac)
> regfit.bac.sum$which[which.min(regfit.bac.sum$cp),]
(Intercept)          vh        wind    humidity        temp
       TRUE       FALSE       FALSE        TRUE        TRUE
        ibh         dpg         ibt         vis         doy
      FALSE       FALSE        TRUE        TRUE        TRUE
```

iv. Both stepwise (choose any method)

```
Step:  AIC=988.24
Y ~ temp + humidity + doy + ibt + vis

           Df Sum of Sq    RSS     AIC
<none>                  6357.4  988.24
```

```
+ vh        1     28.64 6328.8  988.74
+ ibh       1     26.93 6330.5  988.83
+ wind      1      0.70 6356.7  990.20
+ dpg       1      0.06 6357.4  990.23
- vis       1     96.89 6454.3  991.23
- doy       1    343.58 6701.0 1003.60
- ibt       1    532.30 6889.7 1012.77
- humidity  1    690.44 7047.9 1020.26
- temp      1    816.90 7174.3 1026.13

Call:
lm(formula = Y ~ temp + humidity + doy + ibt + vis, data = X)

Coefficients:
(Intercept)         temp     humidity          doy          ibt
 -10.318950     0.232690     0.085091    -0.010065     0.034929
        vis
  -0.008202
```

(c) Compare the outcome of these methods with the significant variables found in the full linear regression in part (a)

(d) Potentially, other transformations of covariates might be important. What happens if you attempt to do all subsets with the original covariates and their square? That is, for all covariates, put both

$$X \text{ and } X^2$$

as possible terms.

```
X.poly = cbind(X,X**2)
main.effects = names(X)
sq.effects   = paste(main.effects,'.Sq',sep='')
names(X.poly) = c(main.effects,sq.effects)

> regfit.exh = regsubsets ( x = X.poly,y = Y, nvmax =19 ,method ="exhaustive")
> regfit.exh.sum = summary(regfit.exh)
> regfit.exh.sum$which[which.min(regfit.exh.sum$bic),]
(Intercept)          vh        wind    humidity        temp
       TRUE       FALSE        TRUE       FALSE        TRUE
        ibh         dpg         ibt         vis         doy
      FALSE       FALSE        TRUE        TRUE        TRUE
      vh.Sq     wind.Sq humidity.Sq     temp.Sq      ibh.Sq
      FALSE       FALSE        TRUE        TRUE       FALSE
     dpg.Sq      ibt.Sq      vis.Sq      doy.Sq
       TRUE       FALSE        TRUE        TRUE
```