

1. Variable selection using the SMSA data. I had you eliminate eliminate noxpot (perfectly correlated with nox) and hcpot since highly correlated with nox (but see SAS results if we leave hcpot in, surprising results).

i) Model selection using adjusted R^2 , C_p and AIC. There are many, many models that are indistinguishable from one another by any of the criteria and will produce similar fitting values. *For example, THERE ARE 183 COMBINATIONS OF VARIABLES GIVEN WITH ADJUSTED R-SQUARE RUNNING BETWEEN .7036 and .69003!* This clearly doesn't help too much with choosing among models

sorted by adjR2

	(Intercept)	jantemp	julytemp	relhum	rain	educ	popden	pnwhite	pw	pop			
9	1	1	1	0	1	1	1	1	1	1	0		
8	1	1	1	0	1	1	1	1	1	1	0		
7	1	1	1	0	1	0	1	1	1	1	0		
	perhouse	income	so2pot	nox	rsquared	sse	adjR2	Cp	AIC				
9	1	0	1	0	0.7449062	57649.29	0.6980522	6.946659	426.1912				
8	0	0	1	0	0.7424629	58201.47	0.7012569	5.386748	424.7536				
7	0	0	1	0	0.7373596	59354.77	0.7013109	4.305935	423.9113				

BEST MODELS VIA adjR2

1. jantemp,julytemp,rain educ,popden,pnwhite,pw,perhouse,so2pot (9 variables, p = 10)
2. jantemp,julytemp,rain,educ,popden,pnwhite,pw, so2pot (8 variables, p = 9)
3. jantemp,julytemp,rain, popden,pnwhite,pw, so2pot (7 variables, p = 8)

sorted by Cp

```
> sortcp<-subsetinfo[order(Cp),]
> sortcp
```

	(Intercept)	jantemp	julytemp	relhum	rain	educ	popden	pnwhite	pw	pop			
6	1	1	1	0	1	1	0	1	0	0			
7	1	1	1	0	1	0	1	1	1	0			
6	1	1	0	0	1	0	1	1	1	0			
	perhouse	income	so2pot	nox	rsquared	sse	adjR2	Cp	AIC				
6	0	0	1	0	0.7266009	61786.16	0.6950548	4.243757	424.2799				
7	0	0	1	0	0.7373596	59354.77	0.7013109	4.305935	423.9113				
6	0	0	1	0	0.7243113	62303.57	0.6925011	4.656138	424.7720				

BEST MODELS VIA Mallows Cp

1. jantemp,julytemp,rain educ, pnwhite, so2pot (6 variables, p = 7)
2. jantemp,julytemp,rain, popden,pnwhite,pw,so2pot (7 variables, p = 8)
3. jantemp, rain, popden,pnwhite,pw,so2pot (6 variables, p = 7)

sorted by AIC

	(Intercept)	jantemp	julytemp	relhum	rain	educ	popden	pnwhite	pw	pop			
7	1	1	1	0	1	0	1	1	1	0			
6	1	1	1	0	1	1	0	1	0	0			
8	1	1	1	0	1	1	1	1	1	0			
	perhouse	income	so2pot	nox	rsquared	sse	adjR2	Cp	AIC				
7	0	0	1	0	0.7373596	59354.77	0.7013109	4.305935	423.9113				
6	0	0	1	0	0.7266009	61786.16	0.6950548	4.243757	424.2799				
8	0	0	1	0	0.7424629	58201.47	0.7012569	5.386748	424.7536				

BEST MODELS VIA AIC

1. jantemp,julytemp,rain, popden,pnwhite,pwc,so2pot (7 variables, p = 8)
2. jantemp,julytemp,rain,educ, pnwhite, so2pot (6 variables, p = 7)
3. jantemp,julytemp,rain,educ,popden,pnwhite,pwc,so2pot (8 variables, p = 9)

Of course you get the same results in SAS.

Now consider forward, backward and stepwise. We'll do this in R first.

Using Forward selection in R.

FINAL STEP.

Step: AIC=424.28

```
mortal ~ pnwhite + educ + jantemp + so2pot + rain + julytemp
```

	Df	Sum of Sq	RSS	AIC
<none>			61786	424.28
+ popden	1	1012.02	60774	425.31
+ perhouse	1	980.34	60806	425.34
+ pwc	1	973.56	60813	425.34
+ pop	1	707.67	61078	425.60
+ relhum	1	162.74	61623	426.12
+ income	1	115.27	61671	426.17
+ nox	1	110.19	61676	426.17

Above shows what happens to AIC if add a variable (none means do nothing)

```
lm(formula = mortal ~ pnwhite + educ + jantemp + so2pot + rain +  
julytemp, data = data)
```

Coefficients:

```
(Intercept)      pnwhite      educ      jantemp      so2pot      rain  
1214.0281      4.9416     -14.7342     -1.5586      0.2517      1.3872  
julytemp  
-2.4792
```

```
> stepAIC(full,direction="backward") #backward selection
```

BACKWARD. FINAL STEP

	Df	Sum of Sq	RSS	AIC
<none>			59355	423.91
- julytemp	1	2949	62304	424.77
- popden	1	4277	63631	426.02
- pwc	1	6970	66325	428.46
- so2pot	1	8075	67430	429.44
- jantemp	1	8691	68046	429.97
- rain	1	14321	73676	434.66
- pnwhite	1	55938	115293	461.08

Above shows what happens to AIC if remove a variable.

```
lm(formula = mortal ~ jantemp + julytemp + rain + popden + pnwhite +  
pwc + so2pot, data = data)
```

Coefficients:

```
(Intercept)      jantemp      julytemp      rain      popden      pnwhite  
1.099e+03     -1.446e+00     -2.134e+00     1.616e+00     7.152e-03     4.939e+00  
pwc      so2pot
```

-2.399e+00 2.253e-01

STEPWISE

```
> stepAIC(null,scope=list(lower=null,upper=full) ,direction="both") #stepwise
```

LAST STEP

Step: AIC=424.28

```
mortal ~ pnwhite + educ + jantemp + so2pot + rain + julytemp
```

	Df	Sum of Sq	RSS	AIC
<none>			61786	424.28
+ popden	1	1012	60774	425.31
+ perhouse	1	980	60806	425.34
+ pwc	1	974	60813	425.34
+ pop	1	708	61078	425.60
- julytemp	1	3958	65745	425.94
+ relhum	1	163	61623	426.12
+ income	1	115	61671	426.17
+ nox	1	110	61676	426.17
- educ	1	6082	67868	427.82
- rain	1	9031	70817	430.33
- jantemp	1	10474	72260	431.52
- so2pot	1	11877	73663	432.65
- pnwhite	1	57481	119267	461.08

Shows what happen to AIC if add (+) or subtract (-) a variable.

```
lm(formula = mortal ~ pnwhite + educ + jantemp + so2pot + rain +  
julytemp, data = data)
```

Coefficients:

(Intercept)	pnwhite	educ	jantemp	so2pot	rain
1214.0281	4.9416	-14.7342	-1.5586	0.2517	1.3872
julytemp					
-2.4792					

Using SAS we get a different answer. This is because SAS and many other variable selection method determine whether to enter or remove a variable based on the p-value (if small enough to enter, or big enough to remove). This rather than whether the AIC decreases determines the stopping rule.

First using the same pool of variables as above:

Summary of Forward Selection							
Step	Variable Entered	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	pnwhite	1	0.4180	0.4180	49.8217	40.94	<.0001
2	educ	2	0.1455	0.5635	25.6138	18.67	<.0001
3	jantemp	3	0.0730	0.6365	14.4700	11.04	0.0016
4	so2pot	4	0.0440	0.6805	8.5531	7.43	0.0086
5	rain	5	0.0286	0.7091	5.3986	5.21	0.0264
6	julytemp	6	0.0175	0.7266	4.2438	3.33	0.0737
7	popden	7	0.0045	0.7311	5.4372	0.85	0.3611
8	pwc	8	0.0114	0.7425	5.3867	2.21	0.1434
9	perhouse	9	0.0024	0.7449	6.9467	0.47	0.4965
10	nox	10	0.0028	0.7477	8.4476	0.53	0.4713

add four more beyond what stepAIC does

Backward Elimination: Step 7

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	960.50822	48.02034	479360	400.08	<.0001
jantemp	-1.50274	0.53581	9424.29396	7.87	0.0071
rain	1.35740	0.43738	11540	9.63	0.0031
popden	0.00707	0.00379	4176.05048	3.49	0.0676
pnwhite	4.39089	0.63275	57697	48.15	<.0001
pwc	-2.45423	0.99406	7303.25177	6.10	0.0169
so2pot	0.24225	0.08611	9483.09302	7.91	0.0069

All variables left in the model are significant at the 0.1000 level.

Doesn't have julytemp with stepAIC keeps

STEPWISE

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	1214.02811	127.39854	107899	90.81	<.0001
jantemp	-1.55861	0.52495	10474	8.82	0.0045
julytemp	-2.47923	1.35832	3958.35690	3.33	0.0737
rain	1.38724	0.50319	9030.81166	7.60	0.0080
educ	-14.73416	6.51272	6081.53709	5.12	0.0279
pnwhite	4.94164	0.71048	57481	48.38	<.0001
so2pot	0.25172	0.07962	11877	10.00	0.0026

All variables left in the model are significant at the 0.1500 level.

No other variable met the 0.1500 significance level for entry into the model.

Same as step AIC

I had an older set of analyses where I had eliminated nox (perfectly correlated with noxpote) but left both hcpote and noxpote as potential variables. It is interesting/surprising that both forward and backward have models with both noxpote and hcpote in them given their high pairwise correlation. But note that the collection of other variables is not the same as above. Stepwise (preferred by many among the automatic procedures) is the same here as above.

 ** SUMMARY OF FOWARD SELECTION USING DEFAULT OF P-VALUE < .5 TO ENTER. ***

No other variable met the 0.5000 significance level for entry into the model.

Step	Variable Entered	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	pnwhite	1	0.4180	0.4180	52.5492	40.94	<.0001
2	educ	2	0.1455	0.5635	27.6594	18.67	<.0001
3	jantemp	3	0.0730	0.6365	16.1736	11.04	0.0016
4	so2pot	4	0.0440	0.6805	10.0507	7.43	0.0086
5	rain	5	0.0286	0.7091	6.7620	5.21	0.0264
6	julytemp	6	0.0175	0.7266	5.5251	3.33	0.0737
7	popden	7	0.0045	0.7311	6.6976	0.85	0.3611
8	pwc	8	0.0114	0.7425	6.5938	2.21	0.1434
9	hcpote	9	0.0026	0.7451	8.1155	0.50	0.4840
10	noxpote	10	0.0073	0.7524	8.7574	1.42	0.2385
11	perhouse	11	0.0055	0.7579	9.7410	1.07	0.3068
12	pop	12	0.0034	0.7613	11.1193	0.65	0.4249

 ** SUMMARY OF BACKWARD SELECTION USING DEFAULT OF P-VALUE > .10 TO REMOVE

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
----------	--------------------	----------------	------------	---------	--------

Intercept	1131.34191	104.79824	136306	116.54	<.0001
jantemp	-1.21286	0.62543	4398.42573	3.76	0.0581
julytemp	-2.54711	1.44280	3645.16421	3.12	0.0836
rain	1.19093	0.50118	6604.12151	5.65	0.0214
popden	0.00923	0.00353	7999.98574	6.84	0.0118
pnwhite	4.96479	0.71293	56721	48.50	<.0001
pwc	-2.28620	0.98958	6242.58784	5.34	0.0250
hcpot	-0.87778	0.32279	8648.84416	7.39	0.0090
noxpot	1.70967	0.61839	8939.89408	7.64	0.0080

** SUMMARY OF STEPWISE SELECTION USING DEFAULT OF P-VALUE > .15 TO REMOVE AND p-VALUE < .15 TO ENTER.

Step	Variable Entered	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	pnwhite		1	0.4180	0.4180	52.5492	40.94	<.0001
2	educ		2	0.1455	0.5635	27.6594	18.67	<.0001
3	jantemp		3	0.0730	0.6365	16.1736	11.04	0.0016
4	so2pot		4	0.0440	0.6805	10.0507	7.43	0.0086
5	rain		5	0.0286	0.7091	6.7620	5.21	0.0264
6	julytemp		6	0.0175	0.7266	5.5251	3.33	0.0737

2. Continue to use the SMSA data and consider fitting mortality as a function of all of the variables (without noxpot and hcpot as before).

i) See partial results

ii) See plot. For many cases the studentized and regular residuals are the same but there are a few where they are a couple that are quite different; these correspond to the two highest leverage values.

iii) Using the leverage values there are six values (observations 47,31,37,18,58,28) that are potential outliers in the x space and warrant further investigation. Want to check to be sure there are no erroneous values and also watch for how these behave with respect to influence (e.g., via Cook's distance). These values are also the ones that will have the greatest impact when we studentize since the divisor involves $1 - h_{ii}$.

iv) The biggest Cook's distance is .5936, which is less than even $F(.2, p, n-p)$ so no apparent influential observations.

v) Applying the rough rule in the notes, there is one positive values (observation 36 with a value of almost 4) above the cutoff of $3.581952 = t(1 - (.05/2n), n - p)$

```
> regout<-lm(mortal~jantemp + julytemp +relhum+ rain+
+ educ+ popden+ pnwhite+ pwc +pop +perhouse+ income+ so2pot+ nox,data=data, na.action=na.exclude)
> summary(regout)
```

Call:

```
lm(formula = mortal ~ jantemp + julytemp + relhum + rain + educ +
    popden + pnwhite + pwc + pop + perhouse + income + so2pot +
    nox, data = data, na.action = na.exclude)
```

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	1.349e+03	2.831e+02	4.764	2.01e-05 ***
jantemp	-1.547e+00	7.702e-01	-2.009	0.0506 .
julytemp	-2.722e+00	1.954e+00	-1.393	0.1704
relhum	1.574e-01	1.166e+00	0.135	0.8932
rain	1.230e+00	5.650e-01	2.177	0.0348 *
educ	-1.009e+01	9.091e+00	-1.110	0.2728
popden	4.660e-03	4.393e-03	1.061	0.2944

```

pnwhite      5.332e+00  9.179e-01  5.808 6.01e-07 ***
pwc          -1.730e+00  1.238e+00 -1.397 0.1691
pop          2.641e-06  4.136e-06  0.639 0.5263
perhouse     -3.140e+01  4.054e+01 -0.775 0.4427
income       -3.097e-04  1.305e-03 -0.237 0.8135
so2pot       2.195e-01  1.014e-01  2.165 0.0357 *
nox          -1.405e-01  1.888e-01 -0.744 0.4607
Residual standard error: 35.42 on 45 degrees of freedom
Multiple R-squared: 0.7502, Adjusted R-squared: 0.678
F-statistic: 10.39 on 13 and 45 DF, p-value: 1.233e-09

```

NOTE: YOU CAN'T JUST USE THE P-VALUES FROM THE FULL FIT ABOUT TO DO MODEL BUILDING!

```

> xmat<-model.matrix(regout)
> p=ncol(xmat)
> n=nrow(xmat)
> resid<-residuals(regout)
> stresid<-rstudent(regout) #gets studentized residual
> cooks<-cooks.distance(regout) #gets cook's distances
> leverage<-hat(xmat) #gets leverage values
> obs<-seq(1,n)
> sumall<-cbind(mortal,stresid,cooks,leverage)

```

```

strsort<-sumall[order(stresid),]
> strsort
      mortal      stresid      cooks      leverage
31  861.44 -2.175550914 3.640349e-01 0.53834511
27  844.05 -2.096147799 5.740149e-02 0.16436254
58  911.82 -1.941168328 5.936231e-01 0.70070619

2   997.87  2.791282608 4.786139e-02 0.09006575
36 1113.16  3.663475194 2.144260e-01 0.22204147

```

```

> cooksort<-sumall[order(cooks),]
> cooksort
      mortal      stresid      cooks      leverage
17  936.23  0.005632865 1.951853e-07 0.07766833
29  989.26  0.013411935 3.722886e-06 0.22076699

36 1113.16  3.663475194 2.144260e-01 0.22204147
31  861.44 -2.175550914 3.640349e-01 0.53834511
58  911.82 -1.941168328 5.936231e-01 0.70070619

```

```

> levsort<-sumall[order(leverage),]
> levsort
      mortal      stresid      cooks      leverage
44  874.28 -1.096703394 6.022786e-03 0.06578776
25  968.67  0.158032337 1.478561e-04 0.07500680

8   899.53 -0.492630404 1.471523e-02 0.45492393
47  911.70  0.085665715 5.764738e-04 0.51818589
31  861.44 -2.175550914 3.640349e-01 0.53834511
37  994.65  0.318375436 9.510959e-03 0.56282199
18  871.77  0.635937315 5.291185e-02 0.64380475

```

```

58  911.82 -1.941168328 5.936231e-01 0.70070619
28  861.26 -0.501672364 1.184529e-01 0.86630364
> # Values to compare things to
> levcut = 2*p/n
> cat("compare leverage to ", levcut, "\n")

compare leverage to  0.4745763

> cookcut=qf(.5,p,n-p)
> cookcut2=qf(.2,p,n-p)
cat("compare Cook's distance to ", cookcut," or " ,cookcut2, "\n")

compare Cook's distance to  0.9671291  or  0.6592054

> residcut<-qt(1-(.05/(2*n)),n-p-1)
> cat("compare stud. or stud-del residual to (with alpha = .05) ", residcut, "\n")

compare stud. or stud-del residual to (with alpha = .05)  3.581952

```

3. Models with sex discrimination.

i)

- Regress wage on age, educ, sex, educ*sex and age*sex

$$E(Y) = \beta_0 + \beta_1 age + \beta_2 educ + \beta_3 sex + \beta_4 educ * sex + \beta_5 age * sex.$$

Sex	Model
0	$\beta_0 + \beta_1 age + \beta_2 educ$
1	$\beta_0 + \beta_1 age + \beta_2 educ + \beta_3 + \beta_4 educ + \beta_5 age$ $= \beta_0 + \beta_3 + (\beta_1 + \beta_5) age + (\beta_2 + \beta_4) educ$

This model allows a separate first order/additive model with separate intercept and coefficients for age and educ for each sex. β_3 , β_4 and β_5 represent differences in these coefficient for sex =1 compared to sex = 0.

- Regress wage on age, educ, sex, educ*sex, age*sex and age*educ*sex.

$$E(Y) = \beta_0 + \beta_1 age + \beta_2 educ + \beta_3 sex + \beta_4 educ * sex + \beta_5 age * sex + \beta_6 age * educ * sex$$

Sex	Model
0	$\beta_0 + \beta_1 age + \beta_2 educ$
1	$\beta_0 + \beta_1 age + \beta_2 educ + \beta_3 + \beta_4 educ + \beta_5 age + \beta_6 educ * age$ $= \beta_0 + \beta_3 + (\beta_1 + \beta_5) age + (\beta_2 + \beta_4) educ + \beta_6 educ * age$

This model is just “additive” in educ and age for sex = 0 but for sex = 1 there is an education by sex interaction. This would be a strange choice.

ii) We want a model like the one for sex = 1 in the previous item (with four parameters) to hold also for sex = 0. So $4 \times 2 = 8$. If we add $+\beta_7 age * educ$ to the previous model that will do it. The model for sex = 1 is as above and the one for sex = 0 would add the $\beta_7 educ * age$ term.

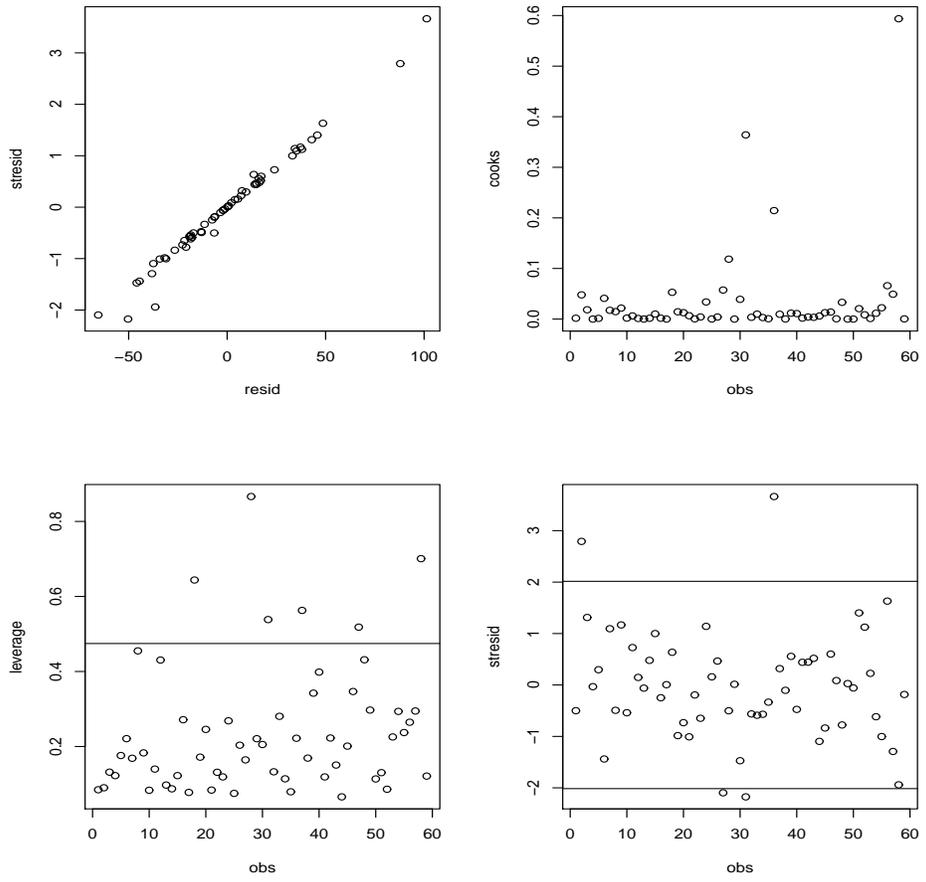


Figure 1: Plots for diagnostics with SMSA data.