B. Kedem, STAT 430

SAS Examples SAS8
==========

ssh xyz@glue.umd.edu, tap sas913, sas
https://www.statlab.umd.edu/sasdoc/sashtml/onldoc.htm

Multiple Regression
==========

0. Show How to Get AIC, BIC Using PROC MIXED.
a. Designed Regression: An example in which the covariates
   are uncorrelated.
b. Nonexperimental Regression (The common case). Use Stepwise Reg.
c. Dummy Variables.

a. Desingned Regression: An example in which the covariates
   are uncorrelated.
==========

```
OPTION PS=45 LS=70;

Example: Study the effect of a stimulant (dose) and exercise on
         weight loss. The stimulant and exercise are controlled!!!
         Not random!!! But the weight loss is random!!!

We have 24 students (subjects), 4 levels of stimulant and 3 levels of
exercise. Since there are 24 students and 3*4=12 stimulant-exercise
combinations, each combination is repeated twice on 2 different students.

DATA DOSERESP;
INPUT ID DOSAGE EXERCISE WLOSS;
EX2=EXERCISE**2;  <--- Need to have an additional model for AIC!!!
```

```
DATALINES;
1 100  0  -4
2 100  0   0
3 100  5  -7
4 100  5  -6
5 100 10  -2
6 100 10 -14
7 200  0  -5
8 200  0  -2
9 200  5  -5
10 200 5  -8
11 200 10 -9
12 200 10 -9
13 300  0  1
14 300  0  0
15 300  5 -3
16 300  5 -3
17 300 10 -8
18 300 10 -12
19 400  0  -5
20 400  0  -4
21 400  5  -4
22 400  5  -6
23 400 10  -9
24 400 10  -7
;



PROC REG DATA=DOSERESP;
/*P=PREDICTED VALUES, R=RESIDUALS*/;
MODEL WLOSS = DOSAGE EXERCISE/ P R;
RUN;
```

The REG Procedure
Model: MODEL1
Dependent Variable: WLOSS

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	162.97083	81.48542	11.19	0.0005
Error	21	152.98750	7.28512		
Corrected Total	23	315.95833			

Root MSE	2.69910	R-Square	0.5158
Dependent Mean	-5.45833	Adj R-Sq	0.4697
Coeff Var	-49.44909		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-2.56250	1.50884	-1.70	0.1042
DOSAGE	1	0.00117	0.00493	0.24	0.8151
EXERCISE	1	-0.63750	0.13495	-4.72	0.0001

Model: WLOSS = -2.56250 + 0.00117*DOSAGE - 0.63750*EXERCISE

We see 0.00117 is small with p-val=0.8151 ===> Can remove DOSAGE!!!

Note: t with df=21. thus: p-val=P(T_{21} < -1.70)*2 = 0.1039,
p-val=P(T_{21} < -0.24)*2 = 0.8126, etc.

But: P(T_{1} < -1.70)*2 = 0.3385061 FAR FROM 0.1039

Output Statistics

Obs	Dep Var WLOSS	Predicted Value	Std Error Mean Predict	Residual	Std Error Residual	Student Residual
1	-4.0000	-2.4458	1.1425	-1.5542	2.445	-0.636
2	0	-2.4458	1.1425	2.4458	2.445	1.000
3	-7.0000	-5.6333	0.9219	-1.3667	2.537	-0.539
4	-6.0000	-5.6333	0.9219	-0.3667	2.537	-0.145
5	-2.0000	-8.8208	1.1425	6.8208	2.445	2.789
6	-14.0000	-8.8208	1.1425	-5.1792	2.445	-2.118
7	-5.0000	-2.3292	0.9053	-2.6708	2.543	-1.050
8	-2.0000	-2.3292	0.9053	0.3292	2.543	0.129
9	-5.0000	-5.5167	0.6035	0.5167	2.631	0.196
10	-8.0000	-5.5167	0.6035	-2.4833	2.631	-0.944
11	-9.0000	-8.7042	0.9053	-0.2958	2.543	-0.116
12	-9.0000	-8.7042	0.9053	-0.2958	2.543	-0.116
13	1.0000	-2.2125	0.9053	3.2125	2.543	1.263
14	0	-2.2125	0.9053	2.2125	2.543	0.870
15	-3.0000	-5.4000	0.6035	2.4000	2.631	0.912
16	-3.0000	-5.4000	0.6035	2.4000	2.631	0.912
17	-8.0000	-8.5875	0.9053	0.5875	2.543	0.231
18	-12.0000	-8.5875	0.9053	-3.4125	2.543	-1.342
19	-5.0000	-2.0958	1.1425	-2.9042	2.445	-1.188
20	-4.0000	-2.0958	1.1425	-1.9042	2.445	-0.779
21	-4.0000	-5.2833	0.9219	1.2833	2.537	0.506
22	-6.0000	-5.2833	0.9219	-0.7167	2.537	-0.283
23	-9.0000	-8.4708	1.1425	-0.5292	2.445	-0.216
24	-7.0000	-8.4708	1.1425	1.4708	2.445	0.601

Obs	-2-1 0 1 2	Cook's D
1	*| |	0.029
2	|** |	0.073
3	*| |	0.013
4	| |	0.001
5	|***** |	0.566
6	****| |	0.326
7	**| |	0.047
8	| |	0.001
9	| |	0.001
10	*| |	0.016
11	| |	0.001
12	| |	0.001
13	|** |	0.067
14	|* |	0.032
15	|* |	0.015
16	|* |	0.015
17	| |	0.002
18	**| |	0.076
19	**| |	0.103
20	*| |	0.044
21	|* |	0.011
22	| |	0.004
23	| |	0.003
24	|* |	0.026

Sum of Residuals	0
Sum of Squared Residuals	152.98750
Predicted Residual SS (PRESS)	212.03588

Note: Cook's distance D is a metric for deciding whether a particular data point y_i alone affects regression estimates much. That is, if a certain y_i stands out. |D| > 2 is considered large. If so, this deserves some special scrutiny.

Estimate beta from y1,...,yN. Now delete yi and estimate beta from

the rest of the observations. We have b=Hat_beta and b(i)=Hat_beta(i).
Note b and beta(i) have the same length p!!!

D= {b-b(i)}'{X'X}{b-b(i)}/(p*s^2)

To get a residual plot:

```
PROC REG DATA=DOSERESP;
MODEL WLOSS = DOSAGE EXERCISE;
PLOT RESIDUAL. *WLOSS='O';      <---- TO GET RESIDUAL PLOT OF
RUN;                                  RESIDUAL VS WLOS. GET PLOT ON WINDOW.
```

Find sample correlation between all variables:

```
PROC CORR DATA=DOSERESP;
VAR DOSAGE EXERCISE WLOSS;
RUN;
```

Pearson Correlation Coefficients, N = 24
Prob > |r| under H0: Rho=0

	DOSAGE	EXERCISE	WLOSS
DOSAGE	1.00000	0.00000	0.03595
		1.0000	0.8676
EXERCISE	0.00000	1.00000	-0.71729
	1.0000		<.0001
WLOSS	0.03595	-0.71729	1.00000
	0.8676	<.0001	

WE SEE DOSAGE AND EXERCISE ARE NOT (sample) CORRELATED!!! by design!!!

We see the coeff. of DOSAGE can be taken as 0!!! as the
p-val=0.8151!!! So let's run a new regression only on EXERCISE.
Nothing will happen to the coeff of EXERCISE in the new regression
since DOSAGE AND EXERCISE ARE NOT CORRELATED by design!!!

So, regress on EXERCISE only:

```
PROC REG DATA=DOSERESP;
/*P=PREDICTED VALUES, R=RESIDUALS*/;
MODEL WLOSS = EXERCISE/ P R;
RUN;
```

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	162.56250	162.56250	23.31	<.0001
Error	22	153.39583	6.97254		
Corrected Total	23	315.95833			

Root MSE	2.64056	R-Square	0.5145
Dependent Mean	-5.45833	Adj R-Sq	0.4924
Coeff Var	-48.37661		

NOTE: R^2 almost the same!!!

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-2.27083	0.85224	-2.66	0.0142
EXERCISE	1	-0.63750	0.13203	-4.83	<.0001

Sum of Residuals	0
Sum of Squared Residuals	153.39583
Predicted Residual SS (PRESS)	187.39377

New model: WLOSS = -2.27083 - 0.63750*EXERCISE (b2 = -0.63750 same!)

The new residuals are almost the same!!

Obs	-2-1 0 1 2	Cook's D
1	*| |	0.028
2	|* |	0.048
3	*| |	0.008
4	| |	0.001
5	|***** |	0.411
6	****| |	0.267
7	**| |	0.069
8	| |	0.001
9	| |	0.001
10	*| |	0.021
11	| |	0.001
12	| |	0.001
13	|** |	0.100
14	|* |	0.048
15	|* |	0.020
16	|* |	0.020
17	| |	0.004
18	**| |	0.105
19	**| |	0.069
20	*| |	0.028
21	|* |	0.007
22	| |	0.001
23	| |	0.001
24	|* |	0.025

Now get AIC and BIC of the several models models:

AIC = - 2(maximum log likelihood) + 2(number of free parameters)
BIC = - 2(maximum log likelihood) + p log(N)

Note: Assume normal dist to get the likelihood!!!

Can get AIC, BIC, Cp by using "selection=Rsquare Aic bic cp":

```
PROC REG DATA=DOSERESP;
/*P=PREDICTED VALUES, R=RESIDUALS*/;
MODEL WLOSS = DOSAGE EXERCISE/ selection=Rsquare Aic bic cp;
RUN;
```

The REG Procedure
Model: MODEL1
Dependent Variable: WLOSS

R-Square Selection Method

Number in Model	R-Square	C(p)	AIC	BIC	
1	0.5145	1.0561	48.5192	51.0394	EXERCISE <--Best
1	0.0013	23.3143	65.8303	65.6490	DOSAGE
1	0.4906	2.1102	49.6740	52.0012	EX2
2	0.5158	3.0000	50.4553	53.2716	DOSAGE EXERCISE
2	0.5162	3.0000	50.4376	53.2539	EXERCISE EX2

Different than what we get from PROC MIXED!!!! It appears that using
PROC MIXED to get aic,bic,Cp, is probably better as we get
-2 Log Likelihood and from this check the AIC,BIC, computation.

Model I:

```
proc mixed DATA=DOSERESP method=ml;
MODEL WLOSS = DOSAGE EXERCISE;
RUN;
```

Note: p=4!!! the 3 betas and 1 sigma^2.

-2 Log Likelihood	112.6
AIC (smaller is better)	120.6 <-- 4 parameters!!!

```
                AICC (smaller is better)        122.7
                BIC (smaller is better)         125.3 <-- 4 parameters!!!

112.6 +2*4=120.6          OK!
112.6 +4*log(24)=125.3122 OK!

Model II:
proc mixed DATA=DOSERESP method=ml;
MODEL WLOSS = EXERCISE;
RUN;

                -2 Log Likelihood               112.6
                AIC (smaller is better)         118.6
                AICC (smaller is better)        119.8
                BIC (smaller is better)         122.2

Model III:
proc mixed DATA=DOSERESP method=ml;
MODEL WLOSS = EXERCISE EX2; <----- EX2=EXERCISE**2
RUN;

                -2 Log Likelihood               112.5
                AIC (smaller is better)         120.5
                AICC (smaller is better)        122.7
                BIC (smaller is better)         125.3
```

Model	R^2	p	AIC	BIC	
Dosage+Exercise	0.5158	4	120.6	125.3	
Exercise	0.5145	3	118.6	122.2	<-- Best model
Exercise+Ex2	0.5162	4	120.5	125.3	

We see model II gives the smallest AIC and BIC.

Can run several models simultaneously and store the results
in a new output set using the option OUTEST!!!

```
PROC REG DATA=DOSERESP OUTEST=NEW;
/*P=PREDICTED VALUES, R=RESIDUALS*/;
MODEL WLOSS = DOSAGE EXERCISE;
MODEL WLOSS = EXERCISE;
MODEL WLOSS = EXERCISE EX2;
RUN;


PROC PRINT DATA=NEW;
RUN;
```

Obs	_MODEL_	_TYPE_	_DEPVAR_	_RMSE_	Intercept	DOSAGE	EXERCISE	WLOSS	EX2
1	MODEL1	PARMS	WLOSS	2.69910	-2.56250	.001166667	-0.6375	-1	.
2	MODEL2	PARMS	WLOSS	2.64056	-2.27083	.	-0.6375	-1	.
3	MODEL3	PARMS	WLOSS	2.69810	-2.37500	.	-0.5125	-1	-0.0125

Note: Model II has the smallest RMSE!!! as we would have guessed.

b. Nonexperimental Regression (The common case). Use Stepwise Reg
==========

Before the cols of the design matrix were uncorrelated by design. Here
they may be correlated as is the usual case in applications when the
experiment is not controlled.

```
Do model selection by FORWARD, MAXR, AIC, BIC, C(p)

OPTION PS=45 LS=70;

DATA READING;
INPUT ID ACH6 ACH5 APT ATT INCOME;
DATALINES;
 1 7.5 6.6 104 60 67
 2 6.9 6.0 116 58 29
 3 7.2 6.0 130 63 36
 4 6.8 5.9 110 74 84
 5 6.7 6.1 114 55 33
 6 6.6 6.3 108 52 21
 7 7.1 5.2 103 48 19
 8 6.5 4.4  92 42 30
 9 7.2 4.9 136 57 32
10 6.2 5.1 105 49 23
11 6.5 4.6  98 54 57
12 5.8 4.3  91 56 29
13 6.7 4.8 100 49 30
14 5.5 4.2  98 43 36
15 5.3 4.3 101 52 31
16 4.7 4.4  84 41 33
17 4.9 3.9  96 50 20
18 4.8 4.1  99 52 34
19 4.7 3.8 106 47 30
20 4.6 3.6  89 58 27
;

PROC REG DATA=READING;
MODEL ACH6 = ACH5 APT ATT INCOME/ SELECTION = FORWARD;
RUN;
```

The REG Procedure

Model: MODEL1
Dependent Variable: ACH6

Forward Selection: Step 1

Variable ACH5 Entered: R-Square = 0.6691 and C(p) = 1.8755

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	12.17625	12.17625	36.40	<.0001
Error	18	6.02175	0.33454		
Corrected Total	19	18.19800			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	1.83725	0.71994	2.17866	6.51	0.0200
ACH5	0.86756	0.14380	12.17625	36.40	<.0001

Bounds on condition number: 1, 1

Forward Selection: Step 2

Variable APT Entered: R-Square = 0.7082 and C(p) = 1.7646

Model: MODEL1
Dependent Variable: ACH6

Forward Selection: Step 2

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	12.88735	6.44367	20.63	<.0001
Error	17	5.31065	0.31239		
Corrected Total	19	18.19800			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	0.64270	1.05398	0.11616	0.37	0.5501
ACH5	0.72475	0.16814	5.80435	18.58	0.0005
APT	0.01825	0.01210	0.71110	2.28	0.1497

Bounds on condition number: 1.464, 5.8559

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

Summary of Forward Selection

Step	Variable Entered	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	ACH5	1	0.6691	0.6691	1.8755	36.40	<.0001
2	APT	2	0.0391	0.7082	1.7646	2.28	0.1497

So, the model selected is:

ACH6 = 0.64270 + 0.72475*ACH5 + 0.01825*APT

Note: the FORWARD selection method compares the p-val of a NEW
entering covariate with 0.5: p-val < 0.5 YES! p-val > 0.5 NO!

Model	P-val From F test	Include?	
ACH6 = ACH5	0.0001 for ACH5	ACH5	Yes
ACH6 = ACH5+APT	0.1497 for APT	APT	Yes
ACH6 = ACH5+ATT	0.9504 for ATT	ATT	No
ACH6 = ACH5+INCOME	0.8686 for INCOME	INCOME	No

Let's look at the Corr. Matrix to see if we get support for our finding:

```
PROC CORR DATA=READING;
VAR ACH6 ACH5 APT ATT INCOME;
RUN;
```

Pearson Correlation Coefficients, N = 20
Prob > |r| under H0: Rho=0

	ACH6	ACH5	APT	ATT	INCOME
ACH6	1.00000	0.81798	0.62387	0.42559	0.31896
		<.0001	0.0033	0.0614	0.1705
ACH5	0.81798	1.00000	0.56297	0.51104	0.36326
	<.0001		0.0098	0.0213	0.1154
APT	0.62387	0.56297	1.00000	0.49741	0.09811
	0.0033	0.0098		0.0256	0.6807
ATT	0.42559	0.51104	0.49741	1.00000	0.62638
	0.0614	0.0213	0.0256		0.0031
INCOME	0.31896	0.36326	0.09811	0.62638	1.00000
	0.1705	0.1154	0.6807	0.0031	

We see ACH6 is highly correlated with ACH5 and only moderately corr.
with APT, but ACH6 is weakly correlated with ATT and Income. This
supports the FORWARD method which indeed selected only ACH5 and APT!!!

Now use MAXR selection using R^2 as a criterion:

```
PROC REG DATA=READING;
MODEL ACH6 = ACH5 APT ATT INCOME/ SELECTION = MAXR;
RUN;
```

The REG Procedure
Model: MODEL1
Dependent Variable: ACH6

Maximum R-Square Improvement: Step 1

Variable ACH5 Entered: R-Square = 0.6691 and C(p) = 1.8755

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	12.17625	12.17625	36.40	<.0001
Error	18	6.02175	0.33454		
Corrected Total	19	18.19800			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	1.83725	0.71994	2.17866	6.51	0.0200
ACH5	0.86756	0.14380	12.17625	36.40	<.0001

Bounds on condition number: 1, 1

The above model is the best 1-variable model found.

Maximum R-Square Improvement: Step 2

Variable APT Entered: R-Square = 0.7082 and C(p) = 1.7646

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	12.88735	6.44367	20.63	<.0001
Error	17	5.31065	0.31239		
Corrected Total	19	18.19800			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	0.64270	1.05398	0.11616	0.37	0.5501
ACH5	0.72475	0.16814	5.80435	18.58	0.0005
APT	0.01825	0.01210	0.71110	2.28	0.1497

Bounds on condition number: 1.464, 5.8559

The above model is the best 2-variable model found.

Maximum R-Square Improvement: Step 3

Variable ATT Entered: R-Square = 0.7109 and C(p) = 3.6194

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	12.93628	4.31209	13.11	0.0001
Error	16	5.26172	0.32886		
Corrected Total	19	18.19800			

	Parameter	Standard

Variable	Estimate	Error	Type II SS	F Value	Pr > F
Intercept	0.80014	1.15586	0.15759	0.48	0.4987
ACH5	0.74740	0.18223	5.53198	16.82	0.0008
APT	0.01973	0.01299	0.75862	2.31	0.1483
ATT	-0.00798	0.02068	0.04893	0.15	0.7048

Bounds on condition number: 1.6336, 14.16

The above model is the best 3-variable model found.

Maximum R-Square Improvement: Step 4

Variable INCOME Entered: R-Square = 0.7223 and C(p) = 5.0000

Maximum R-Square Improvement: Step 4

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	13.14492	3.28623	9.76	0.0004
Error	15	5.05308	0.33687		
Corrected Total	19	18.19800			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	0.91165	1.17841	0.20162	0.60	0.4512
ACH5	0.71374	0.18933	4.78747	14.21	0.0019
APT	0.02394	0.01419	0.95826	2.84	0.1124
ATT	-0.02116	0.02681	0.20983	0.62	0.4423
INCOME	0.00899	0.01142	0.20864	0.62	0.4435

Bounds on condition number: 2.4316, 31.793

The above model is the best 4-variable model found.
No further improvement in R-Square is possible.

Now SELECTION=STEPWISE regression : Can add but also remove covariates!!!

The stepwise selection process consists of a series of
alternating step-up and step-down phases. The former
adds variables to the model, while the latter removes
variables from the model.

```
PROC REG DATA=READING;
MODEL ACH6 = ACH5 APT ATT INCOME/ SELECTION = STEPWISE;
RUN;
```

The REG Procedure
Model: MODEL1
Dependent Variable: ACH6

Stepwise Selection: Step 1

Variable ACH5 Entered: R-Square = 0.6691 and C(p) = 1.8755

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	12.17625	12.17625	36.40	<.0001
Error	18	6.02175	0.33454		
Corrected Total	19	18.19800			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	1.83725	0.71994	2.17866	6.51	0.0200
ACH5	0.86756	0.14380	12.17625	36.40	<.0001

Bounds on condition number: 1, 1

Stepwise Selection: Step 2

Variable APT Entered: R-Square = 0.7082 and C(p) = 1.7646

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	12.88735	6.44367	20.63	<.0001
Error	17	5.31065	0.31239		
Corrected Total	19	18.19800			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	0.64270	1.05398	0.11616	0.37	0.5501
ACH5	0.72475	0.16814	5.80435	18.58	0.0005
APT	0.01825	0.01210	0.71110	2.28	0.1497

Bounds on condition number: 1.464, 5.8559

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.

Summary of Stepwise Selection

Step	Variable Entered	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value
1	ACH5		1	0.6691	0.6691	1.8755	36.40
2	APT		2	0.0391	0.7082	1.7646	2.28

Summary of Stepwise Selection

Step	Pr > F
1	<.0001
2	0.1497

Now get AIC, BIC!!! Must assume normal dist to get the likelihood.

Note: To get AIC, BIC I ran each model separately!!! using
PROC MIXED IC <--- IC=Information Criterion.

```
proc mixed IC DATA=READING method=ml;
MODEL ACH6 = ACH5;
MODEL ACH6 = ACH5 APT;
MODEL ACH6 = ACH5 APT ATT;
MODEL ACH6 = ACH5 APT ATT INCOME;
RUN;
```

Summary

Model	p	AIC	BIC	c(p)	R^2	
ACH5	3	38.8	41.7	1.87	0.669	
ACH5+APT	4	38.2	42.2	1.76	0.708	
ACH5+APT+ATT	5	40.1	45.0	3.62	0.711	
ACH5+APT+ATT+INCOME	6	41.2	47.2	5.00	0.722	
ACH5+ATT	4	40.7	44.7		0.669	<-- Did separately
ACH5+INCOME	4	40.7	44.7		0.670	<-- Did separately

```
AIC selects ACH5+APT as did STEPWISE and FORWARD!!!
BIC selects ACH5 only!!!

Note: C(p) keeps changing depending on all the covariates available.

       RSS(p)
C(p)= -------- - (n-2p)
        sigma^2

p = # of betas.
RSS(p)=Residual SS from a model with p parameters
s^2=Residual mean square from the largest regression equation.
RSS(p)/sigma^2=(n-p)sigma^2/sigma^2=n-p approximately!!!
Therefore E[C(p)]=p approximately. Look for p that minimizes C(p).


c. Dummy Variables
====================
Two   levels:  One Dummy variable   Z=0,1. Add aZ to the model
Three levels:  Two Dummy variables Z1=0,1, Z2=0,1.
                  Thus: (Z1,Z2)=(1,0)===> Level 1
                        (Z1,Z2)=(0,1)===> Level 2
                        (Z1,Z2)=(0,0)===> Level 3 (REFERENCE)

                  Add aZ1+bZ2 to the model.

EXAMPLE: Have turkey weights (Y) in pounds, turkey age (X) in weeks,
of 13 Thanksgiving turkeys. 4 turkeys were reared in GA (G), 4 in VA
(V), 5 in Wisconsin (W). We wish to fit a straight line and regress Y
on X, but the origin of the turkeys may be an important factor.
We consider 2 models:

Y=X
Y=X+Z1+Z2, Z1=1 if GA and 0 ow
           Z2=1 if VA and 0 ow
           If Z1=Z2=0, then Wisconsin


OPTION PS=45 LS=70;
```

```
DATA TURKEY;
INPUT X Y Z1 Z2;
DATALINES;
28 13.3 1 0
20  8.9 1 0
32 15.1 1 0   GA
22 10.4 1 0
----------------
29 13.1 0 1
27 12.4 0 1
28 13.2 0 1   VA
26 11.8 0 1
---------------
21 11.5 0 0
27 14.2 0 0
29 15.4 0 0   WI
23 13.1 0 0
25 13.8 0 0
;

Model I: Y=X
--------------------

PROC REG DATA=TURKEY;
/*P=PREDICTED VALUES, R=RESIDUALS*/;
MODEL Y=X/ P R;
RUN;
```

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	26.20192	26.20192	21.81	0.0007
Error	11	13.21500	1.20136		
Corrected Total	12	39.41692			

Root MSE	1.09607	R-Square	0.6647
Dependent Mean	12.78462	Adj R-Sq	0.6343
Coeff Var	8.57333		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	1.98333	2.33273	0.85	0.4133
X	1	0.41667	0.08922	4.67	0.0007

Model I: Y= 1.98333 + 0.41667*X

Obs	-2-1 0 1 2	Cook's D
1	| | |	0.007
2	***| |	0.542
3	| |	0.014
4	* | |	0.062
5	* | |	0.073
6	* | |	0.029
7	| |	0.011
8	* | |	0.039
9	| * |	0.100
10	| * |	0.039
11	| ** |	0.140
12	| *** |	0.174
13	| ** |	0.080

```
proc mixed IC DATA=TURKEY method=ml;
MODEL Y = X;
RUN;
```

Information Criteria

Neg2LogLike	Parms	AIC	AICC	HQIC	BIC	CAIC
37.1	3	43.1	45.8	42.8	44.8	47.8

Model II: Y=X+Z1+Z2

```
PROC REG DATA=TURKEY;
/*P=PREDICTED VALUES, R=RESIDUALS*/;
MODEL Y=X Z1 Z2/ P R;
RUN;
```

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	38.60575	12.86858	142.78	<.0001
Error	9	0.81118	0.09013		
Corrected Total	12	39.41692			

Root MSE	0.30022	R-Square	0.9794
Dependent Mean	12.78462	Adj R-Sq	0.9726
Coeff Var	2.34827		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	1.43088	0.65744	2.18	0.0575
X	1	0.48676	0.02574	18.91	<.0001
Z1	1	-1.91838	0.20180	-9.51	<.0001
Z2	1	-2.19191	0.21143	-10.37	<.0001

```
Model II: Y = 1.43088 + 0.48676*X - 1.91838*Z1 - 2.19191*Z2
Thus:
Y = 1.43088 + 0.48676*X - 1.91838*Z1 = -0.4875 + 0.48676*X for GA
Y = 1.43088 + 0.48676*X - 2.19191*Z2 = -0.7610 + 0.48676*X for VA
Y = 1.43088 + 0.48676*X              =  1.4308 + 0.48676*X for Wisconsin

For Model II get better behaving residuals!!!
                                                  Cook's
                      Obs   -2-1 0 1 2              D

                        1 |      |*     |       0.041
                        2 |   ***|      |       0.569
                        3 |      |      |       0.001
                        4 |      |*     |       0.069
                        5 |     *|      |       0.089
                        6 |      |      |       0.000
                        7 |      |**    |       0.137
                        8 |      |      |       0.012
                        9 |     *|      |       0.044
                       10 |    **|      |       0.150
                       11 |     *|      |       0.041
                       12 |      |***   |       0.240
                       13 |      |*     |       0.035

proc mixed IC DATA=TURKEY method=ml;
MODEL Y = X Z1 Z2;
RUN;

Neg2LogLike   Parms       AIC      AICC      HQIC       BIC      CAIC

        0.8       5      10.8      19.4      10.2      13.7      18.7

  Model          p   AIC    BIC    R^2
-----------------------------------------
 I   Y=X         3   43.1   44.8   0.6647
II   Y=X+Z1+Z2   5   10.8   13.7   0.9794

WE SEE MODEL II IS A MUCH BETTER MODEL!!!
```

```
Interesting to see that the FORWARD selection method let all the
covariates X,Z1,Z2 enter into the model!!! That is, Model II is selected.

PROC REG DATA=TURKEY;
MODEL Y = X Z1 Z2/ SELECTION = FORWARD;
RUN;


          All variables have been entered into the model.

                    Summary of Forward Selection

      Variable   Number   Partial    Model
Step  Entered    Vars In  R-Square  R-Square   C(p)    F Value  Pr > F

  1   X             1     0.6647    0.6647   137.620    21.81  0.0007
  2   Z2            2     0.1080    0.7728   92.3680     4.76  0.0542
  3   Z1            3     0.2066    0.9794    4.0000    90.37  <.0001


Now plot Model II
------------------

DATA TURKEY;
INPUT X Y Z1 Z2;
G = -0.4875 + 0.48676*X; /* For GA */;
V = -0.7610 + 0.48676*X; /* For VA */;
W =  1.4308 + 0.48676*X; /* For Wisconsin */;
DATALINES;
28 13.3 1 0
20  8.9 1 0
32 15.1 1 0
22 10.4 1 0
29 13.1 0 1
27 12.4 0 1
28 13.2 0 1
26 11.8 0 1
21 11.5 0 0
27 14.2 0 0
29 15.4 0 0
```

```
23 13.1 0 0
25 13.8 0 0
;

PROC PLOT DATA=TURKEY;
PLOT G*X='G' V*X='V' W*X='W'/OVERLAY;
RUN;

                Plot of G*X.  Symbol used is 'G'.
                Plot of V*X.  Symbol used is 'V'.
                Plot of W*X.  Symbol used is 'W'.

     G |
       |
17.249 +                                                                W
16.762 +
16.276 +
15.789 +                                                  W
15.302 +                                                        W       G
14.815 +                                                    W           V
14.329 +                                        W
13.842 +                                W                       G
13.355 +                                                        G   V
12.868 +                    W                               G   V
12.382 +          W                             G   V
11.895 +     W                          G   V
11.408 +W                               V
10.921 +               G
10.434 +          G    V
 9.948 +     G    V
 9.461 +G    V
 8.974 +V
       |
       -+----+----+----+----+----+----+----+----+----+----+----+----+-
       20   21   22   23   24   25   26   27   28   29   30   31   32

                                       X
NOTE: 9 obs hidden.

We see GA and VA are close, both far from Wisconsin.
```