Stat 5100 Handout 3.2.1 – SAS: Variable Selection

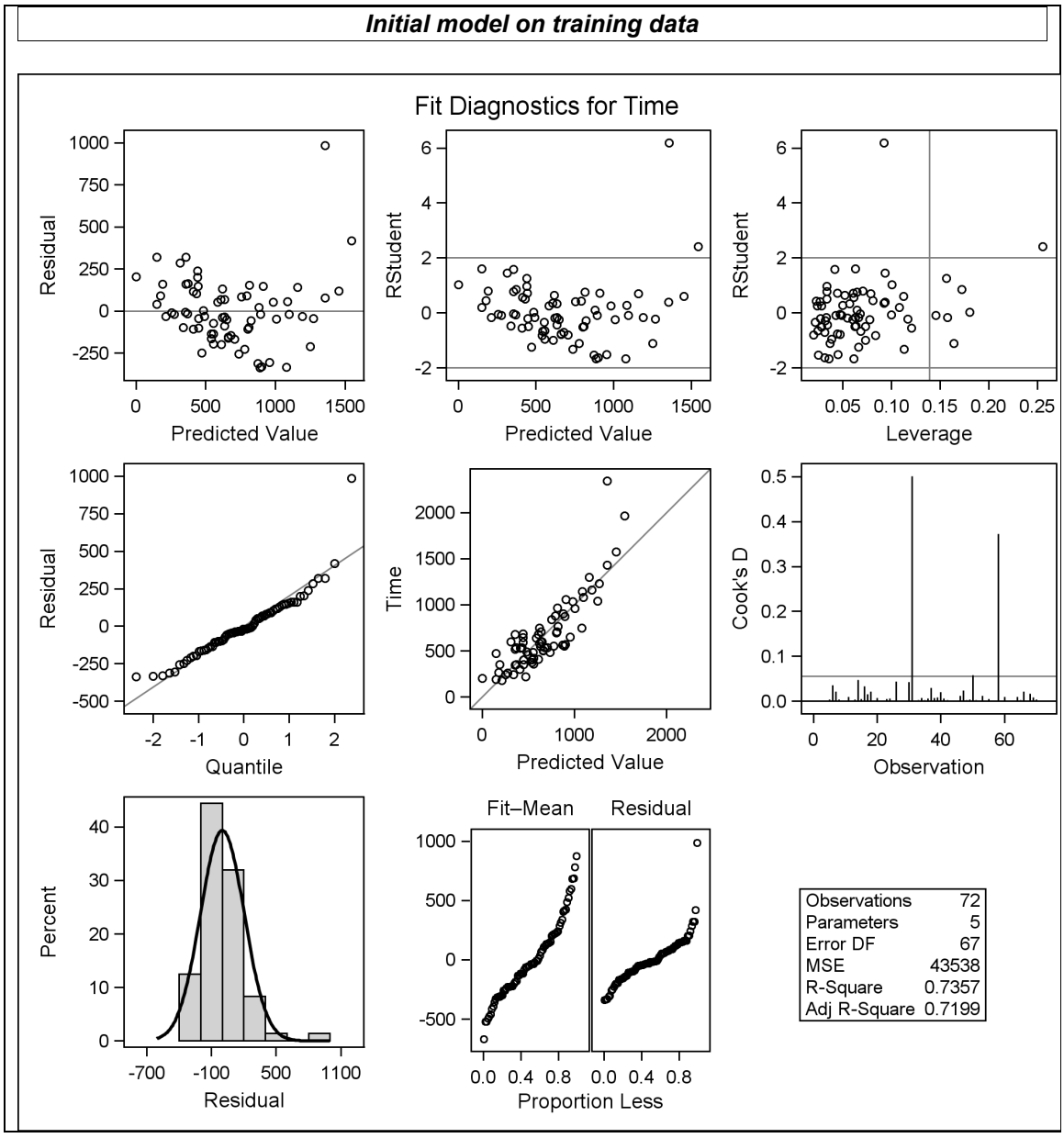
Example: (Textbook tables 9.1 & 9.5) A hospital surgical unit was interested in predicting survival time for patients who undergo a particular liver operation. Data are reported for 108 patients on the following variables: blood-clotting score, prognostic index, enzyme function test score, liver function test score, age (in years), gender (0=male, 1=female), indicators of alcohol use (none, moderate, heavy), and survival time (in days). Which (if any) of these predictors should be used in a linear model?

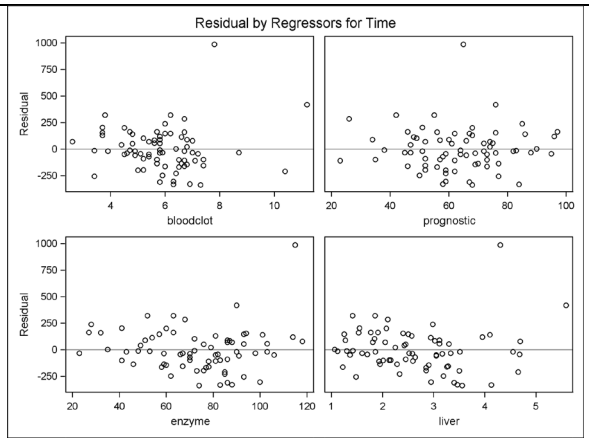
```
/* Input data -- see Table 9.1 in text */
data surgical;
  infile '<filename>' delimiter = '09'x;
       /* '09'x indicates tab-delimited .txt file */
  input bloodclot prognostic enzyme liver age gender
        modAlcohol heavyAlcohol Time;
run;

/* Randomly select training and test sets */
data surgical; set surgical;
  U = uniform(1234);
  ID = _n_;
proc sort data=surgical;
  by U;
proc print data=surgical;
  var U ID Time;
  title1 'Sorted Surgical Data (by U)';
run;
```

Sorted Surgical Data (by U)

Obs	U	ID	Time
1	0.00276	27	545
2	0.00722	101	1158
…			
107	0.97760	38	362
108	0.98587	84	881

```
data train; set surgical;
  if _n_ <= 72;
data test; set surgical;
  if _n_ > 72;
run;

/*************************************************************/

/* Check initial residual assumptions */
proc reg data=train;
  model Time = bloodclot prognostic enzyme liver;
  output out=out1 r=resid p=pred;
  title1 'Initial model on training data';
run;
```






```
/* Define shortcut macro, using line copied from
Canvas page
*/
%macro resid_num_diag(dataset,...
/* Call shortcut macro */
%resid_num_diag(dataset=out1, datavar=resid,
   label='Residual', predvar=pred, predlabel='Predicted');
run;
```

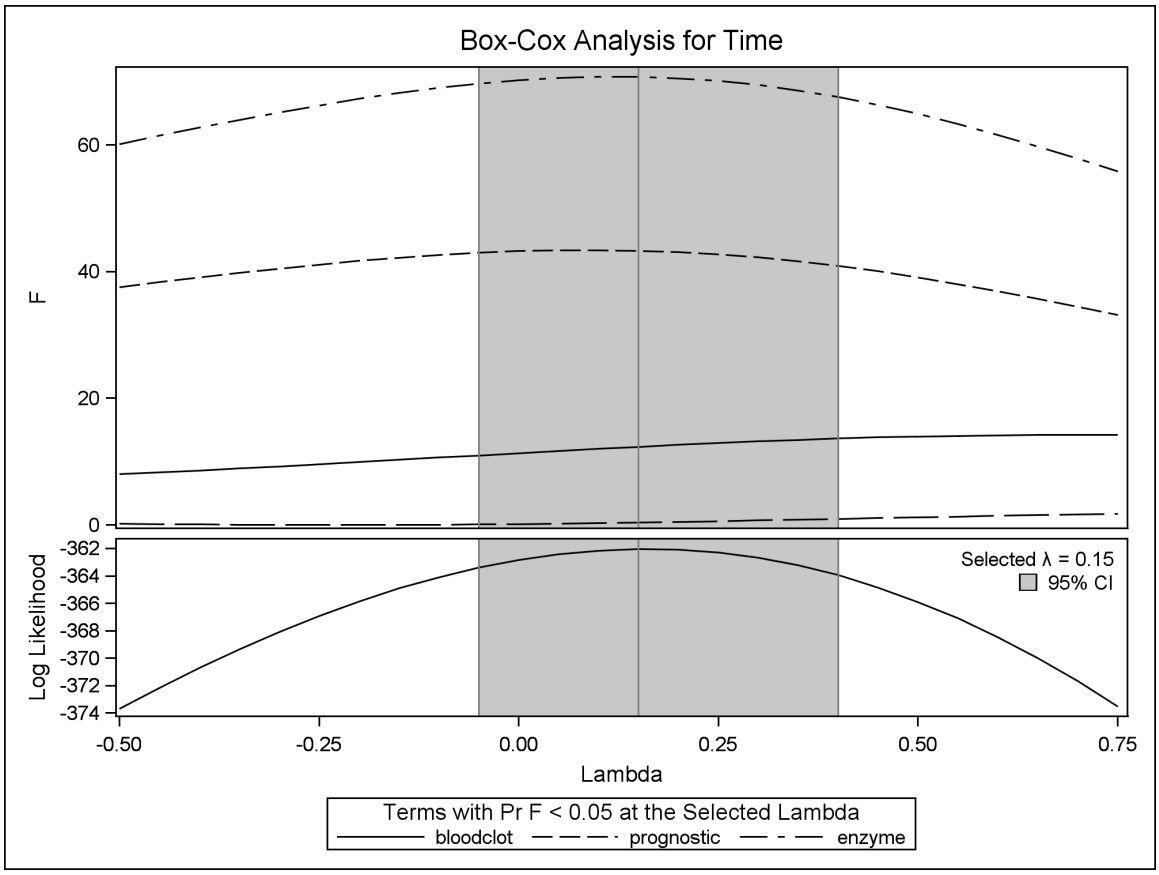
P-value for Brown-Forsythe test of constant variance in Residual vs. Predicted

Obs	t_BF	BF_pvalue
1	1.10680	0.27217

Output for correlation test of normality of Residual (Check text Table B.6 for threshold)

Pearson Correlation Coefficients, N = 72 Prob > |r| under H0: Rho=0		
	resid	**expectNorm**
resid Residual	1.00000	0.94169 <.0001
expectNorm	0.94169 <.0001	1.00000

```
/* Check possible transformation */
proc transreg data=train;
 model boxcox(Time / lambda = -.5 to .75 by .05)
       = identity(bloodclot prognostic enzyme liver);
 title1 'Box-Cox transformation on training data';
run;
```




```
/* Make transformation */
data train; set train;
  logTime = log(Time);
run;
```

```
/*************************************************************/
```

```
/* Look at some 'all possible regressions' approaches: */

proc reg data=train;
  model logTime = bloodclot prognostic enzyme liver
                  / selection=Rsquare;
  title1 'R-square Selection';
run;
```

R-square Selection

Number in Model	R-Square	Variables in Model
1	0.5474	enzyme
1	0.4175	liver
1	0.2690	prognostic
1	0.0307	bloodclot
2	0.7040	prognostic enzyme
2	0.6166	enzyme liver
2	0.5808	bloodclot enzyme
2	0.5265	prognostic liver
2	0.4249	bloodclot liver
2	0.3407	bloodclot prognostic
3	0.7688	bloodclot prognostic enzyme
3	0.7303	prognostic enzyme liver
3	0.6203	bloodclot enzyme liver
3	0.5273	bloodclot prognostic liver
4	0.7692	bloodclot prognostic enzyme liver

```
proc reg data=train;
  model logTime = bloodclot prognostic enzyme liver
                  / selection=AdjRSq;
  title1 'Adjusted R-square Selection';
run;
```

Adjusted R-square Selection

Number in Model	Adjusted R-Square	R-Square	Variables in Model
3	**0.7586**	0.7688	bloodclot prognostic enzyme
4	**0.7554**	0.7692	bloodclot prognostic enzyme liver
3	**0.7184**	0.7303	prognostic enzyme liver
2	**0.6954**	0.7040	prognostic enzyme
2	**0.6055**	0.6166	enzyme liver
3	**0.6036**	0.6203	bloodclot enzyme liver
2	**0.5686**	0.5808	bloodclot enzyme
1	**0.5409**	0.5474	enzyme
2	**0.5128**	0.5265	prognostic liver
3	**0.5064**	0.5273	bloodclot prognostic liver
1	**0.4092**	0.4175	liver
2	**0.4082**	0.4249	bloodclot liver
2	**0.3216**	0.3407	bloodclot prognostic
1	**0.2586**	0.2690	prognostic
1	**0.0168**	0.0307	bloodclot

```
proc reg data=train;
  model logTime = bloodclot prognostic enzyme liver
                  / selection=Cp;
  title1 'Mallows Cp Selection';
run;
```

Mallows Cp Selection

Number in Model	C(p)	R-Square	Variables in Model
3	**3.1274**	0.7688	bloodclot prognostic enzyme
4	**5.0000**	0.7692	bloodclot prognostic enzyme liver
3	**14.3147**	0.7303	prognostic enzyme liver
2	**19.9321**	0.7040	prognostic enzyme
2	**45.3107**	0.6166	enzyme liver
3	**46.2329**	0.6203	bloodclot enzyme liver
2	**55.7184**	0.5808	bloodclot enzyme
1	**63.4064**	0.5474	enzyme
2	**71.4633**	0.5265	prognostic liver
3	**73.2405**	0.5273	bloodclot prognostic liver
2	**100.9613**	0.4249	bloodclot liver
1	**101.1208**	0.4175	liver
2	**125.4071**	0.3407	bloodclot prognostic
1	**144.2297**	0.2690	prognostic
1	**213.4195**	0.0307	bloodclot

```
proc reg data=train;
  model logTime = bloodclot prognostic enzyme liver
                  / selection=AdjRSq Cp AIC SBC;
  title1 'Compare Selection Criteria';
run;
```

Compare Selection Criteria

Number in Model	Adjusted R-Square	R-Square	C(p)	AIC	SBC	Variables in Model
3	**0.7586**	0.7688	3.1274	-187.9550	-178.84833	bloodclot prognostic enzyme
4	**0.7554**	0.7692	5.0000	-186.0918	-174.70842	bloodclot prognostic enzyme liver
3	**0.7184**	0.7303	14.3147	-176.8567	-167.75005	prognostic enzyme liver
2	**0.6954**	0.7040	19.9321	-172.1735	-165.34349	prognostic enzyme
2	**0.6055**	0.6166	45.3107	-153.5422	-146.71221	enzyme liver
3	**0.6036**	0.6203	46.2329	-152.2428	-143.13611	bloodclot enzyme liver
2	**0.5686**	0.5808	55.7184	-147.1065	-140.27651	bloodclot enzyme
1	**0.5409**	0.5474	63.4064	-143.5924	-139.03909	enzyme
2	**0.5128**	0.5265	71.4633	-138.3479	-131.51793	prognostic liver
3	**0.5064**	0.5273	73.2405	-136.4647	-127.35805	bloodclot prognostic liver
1	**0.4092**	0.4175	101.1208	-125.4255	-120.87213	liver
2	**0.4082**	0.4249	100.9613	-124.3507	-117.52075	bloodclot liver
2	**0.3216**	0.3407	125.4071	-114.5126	-107.68262	bloodclot prognostic
1	**0.2586**	0.2690	144.2297	-109.0774	-104.52411	prognostic
1	**0.0168**	0.0307	213.4195	-88.7607	-84.20735	bloodclot

```
/*************************************************************/
```

```
/* Now look at three stepwise approaches: */


proc reg data=train;
  model logTime = bloodclot prognostic enzyme liver
                  / selection=backward slstay=.10;
  title1 'Backward Elimination';
run;
```

Backward Elimination

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

Summary of Backward Elimination

Step	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	liver	3	0.0004	0.7688	3.1274	0.13	0.7223

```
proc reg data=train;
  model logTime = bloodclot prognostic enzyme liver
                  / selection=forward slentry=.10;
  title1 'Forward Selection';
run;
```

Forward Selection

No other variable met the 0.1000 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	enzyme	1	0.5474	0.5474	63.4064	84.66	<.0001
2	prognostic	2	0.1566	0.7040	19.9321	36.51	<.0001
3	bloodclot	3	0.0648	0.7688	3.1274	19.05	<.0001

```
proc reg data=train;
  model logTime = bloodclot prognostic enzyme liver
              / selection=stepwise slentry=.10 slstay=.10;
  title1 'Stepwise Selection';
run;
```

Stepwise Selection

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

No other variable met the 0.1000 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	Pr > F
1	enzyme		1	0.5474	0.5474	63.4064	84.66	<.0001
2	prognostic		2	0.1566	0.7040	19.9321	36.51	<.0001
3	bloodclot		3	0.0648	0.7688	3.1274	19.05	<.0001

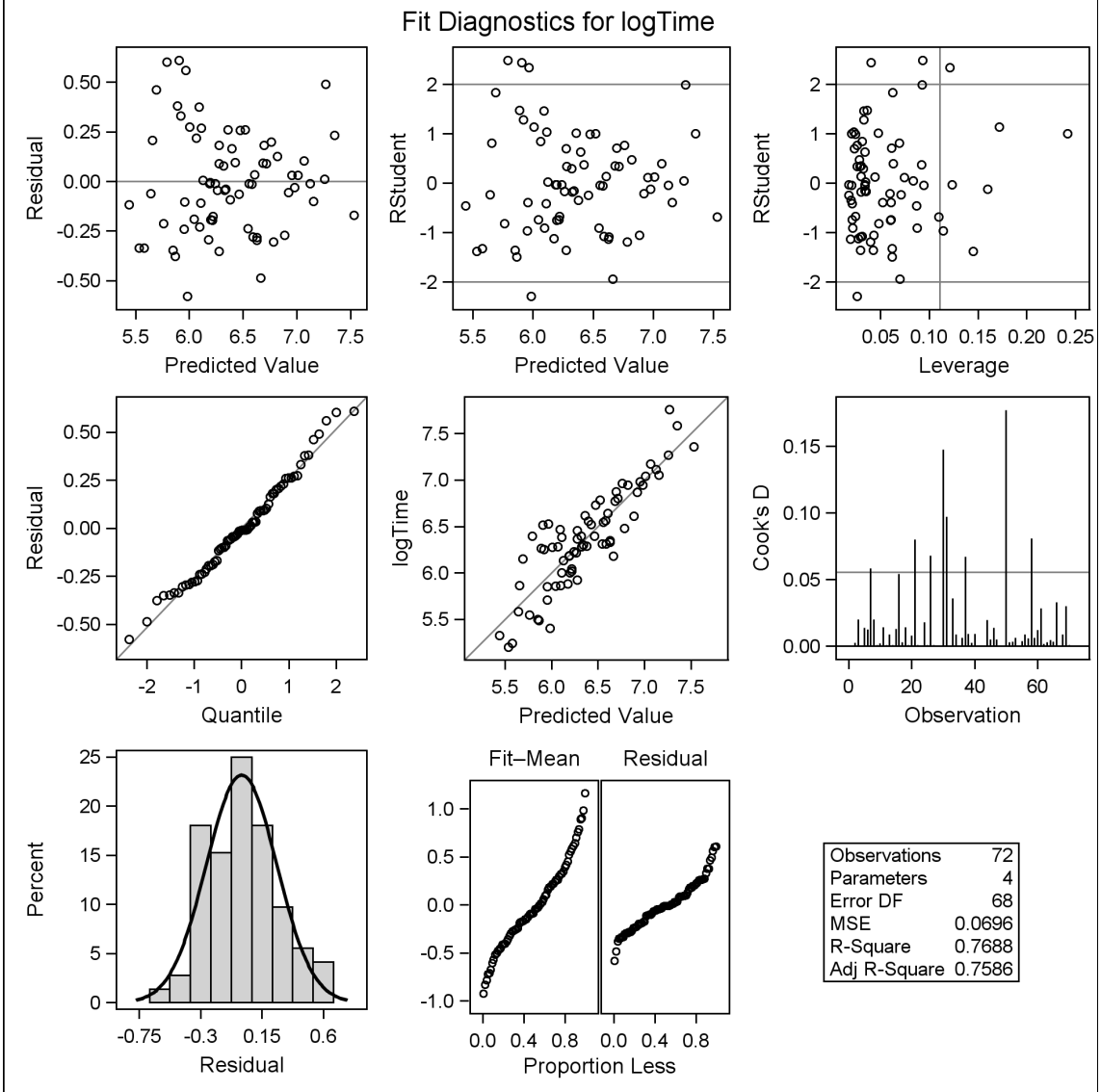
```
/*************************************************************/

/* Validity check of tentative model */

proc reg data=train;
  model logTime = bloodclot prognostic enzyme;
  output out=out2 r=resid p=pred;
  title1 'Tentative Model';
run;
```

Tentative Model

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	15.74523	5.24841	75.37	<.0001
Error	68	4.73541	0.06964		
Corrected Total	71	20.48065			

Parameter Estimates					
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	**1**	3.62880	0.21572	16.82	<.0001
bloodclot	**1**	0.09656	0.02212	4.36	<.0001
prognostic	**1**	0.01523	0.00205	7.44	<.0001
enzyme	**1**	0.01649	0.00147	11.22	<.0001



```
%resid_num_diag(dataset=out2, datavar=resid,
   label='Residual', predvar=pred, predlabel='Predicted');
run;
```

P-value for Brown-Forsythe test of constant variance in Residual vs. Predicted

Obs	t_BF	BF_pvalue
1	2.39814	0.019148

Output for correlation test of normality of Residual (Check text Table B.6 for threshold)

Pearson Correlation Coefficients, N = 72 Prob > |r| under H0: Rho=0		
	resid	expectNorm
resid Residual	1.00000	0.99273 <.0001
expectNorm	0.99273 <.0001	1.00000

```
data test; set test;
  logTime = log(Time);
  logTimehat = 3.62880 + 0.09656*bloodclot
               + 0.01523*prognostic + 0.01649*enzyme;
  SqPredError = (logTime - LogTimehat)**2;
proc means data=test mean;
  var SqPredError;
  title1 'MSPR for test set';
run;
```

MSPR for test set

Mean
0.0763624