"Building" a Regression Model I: Selection of X's

Preamble and 8.1

Types of Application and Role of X's

• Controlled Expt's ... Factors(s) manipulated by investigator

(authors' use of term "control variables" here is confusing)

without /

with supplemental X variables

- may be 'imbalanced' w.r.t. the (primary) factor of interest, and thus need to be 'adjusted for';

- (even if balanced) are a source of additional variation (noise) that one wishes to remove, in order to see the 'signal' more clearly

• Non-exp'tl studies (all data are "observational")

Primary variable(s), other explanatory variables

or

Search for explanatory variables

Might also categorize applications into predictive (focus on $\hat{Y}$ ' s) vs. descriptive (focus on individual $\hat{\beta}$ ' s)

Reduction of No. of Explanatory Variables

• Controlled Expt's with supplemental X variables

if n small, imbalances with respect to other important X's are possible (making for a biased comparison); this bias can be synthetically "removed" by including these important X variables (the objective is to make the comparison over the level(s) of the primary variable "**fairer**").

no matter whether n is large or small, and even if other important X's are well balanced over the levels of the primary X, (i.e. even if other X's are *"orthogonal"* to the primary X), supplementary variables that explain a lot of the extraneous 'noise' in Y should be included in order to remove this extraneous variation (a matter of making comparison "**sharper**").

(These objectives are the subject of the later Chapter on Analysis of Covariance)

Usually, the number of such variables is small, and their influence is known *a priori*.

See "Making comparisons sharper and fairer" in article "Appropriate Uses of Multivariate Analysis" by JH under "articles" at top of www material for this Course (697)

See also article "Modeling and Variable Selection in Epidemiology" by S. Greenland under "Chapter 16" in www material for Course 678

Reduction of No. of Explanatory Variables

• <u>Non-Expt'l Data</u>

Focus on a "Primary" X

other X's retained for comparability with other investigations even if they don't substantially reduce the variability in Y, or even if they are not highly correlated with the Primary X. (provided they aren't so many, and that the number of observations so few, that they create rather than eliminate, noise in the $\hat{\beta}$ of interest. Their use for bias-reduction and noise-reduction is no different than their use in the "Controlled Exp't" situation above.

Search for Explanatory X's

- helpful to have 'hierarchy' of X's
- ask if the objective is focus is on good $\hat{Y}$ ' s or good $\hat{\beta}$ ' s [conditional on other X's]
- no one "best" model.

<u>no matter whether n is large or small,</u> and <u>even if other important X's are well balanced over</u>

8.2   Surgical Unit Example ... general comments

[I would have preferred the title "prediction of survival following a liver operation"]

- Essentially a prognostic application

[notice that there is no examination of specific $\hat{\beta}$ ' s , their magnitudes, or even their signs!!]

? Wonder if X_2 (prognostic score that includes age") also includes parts of X_1, X_3, and X_4 as well.

?   Wouldn't a histogram of the 54 Y's (survival times) be helpful before starting out?

Distribution of survival data often has a long right tail, corresponding to "cures".

Also, survival times often "censored" -- if analysis performed before everyone has reached endpoint of interest, and so not amenable to parameter estimation by the LS criterion.

A bit surprising that post-surgery course was "uniformly fatal" and that the survival times were all within such a narrow interval. The text does not even give the (1 mo.- 27 mo.) range of Y's!

?   Why did authors choose log[Y] transform, rather than say $Y^{1/2}$ or other Y transform ?

8.2 Surgical Unit Example ... "Log-Normal" Distributions

Note: an important distinction and a connection to generalized linear models.

To keep example simple, X = the X_4 variable dichotomized at approximately its median. Also, take logs to base e to correspond with generalized linear model.

	Variable	N	Mean	Std Dev	Minimum	Maximum
X=0	Y	28	132.75	55.13	34	217 (days)
X=1	Y	26	266.54	178.39	72	830 (days)
	ratio		***2.01***			
X=0	LOG_eY	28	4.79	0.49	3.53	5.38 (log days)
X=1	LOG_eY	26	5.39	0.69	4.28	6.72 (log days)
	difference		***0.60***			

Notice that, on average, survival is twice as long when X = 1 i.e. when X_4 > median.

•1• When **fit model to Y using a <u>log link</u> with (constant) Gaussian errors**

i.e. $\log_e [\, \mu_Y \mid X \,] = \beta_0 + \beta_1 X$;   SD[Y | X] same for both X=0 and X=1

in generalized linear model, we obtain

$\log_e$ average(survival | X) = 4.89 + 0.70 X ; SD(survival | X) = 127 days

with the model: Y | X ~ Gaussian(exp[4.89 + 0.70 X] ; SD = 127 days)

•2• When fit to **log Y using the identity link and Gaussian errors**

i.e. $\mu_{\log Y \mid X} = \gamma_0 + \gamma_1 X$

in ("regular") linear model] , we obtain

average(log survival | X) = 4.79 + 0.60 X ; SD(log survival | X) = .56

This latter is modeling **survival** in the 2 X-groups as **log-normal,**

i.e. log(Y | X) ~ Gaussian(4.79 + 0.60 X , SD = .56)

exp[0.60] = 1.82 ≠ the 2.01 ratio in average survival when we model survival directly -- we lose something in trying to go back from log[survival] to survival.

Remember from mathematical-statistics courses: If log Y ~ Gaussian[mu, sigma], then

E[Y] = exp[mu + sigma^2/2]   ≠ exp[mu] if sigma > 0

Var[Y] = exp[2 mu + sigma^2] (Exp[sigma^2] - 1)

It is difficult to get <u>all</u> aspects to be correct & useful: Within-X variation in actual Y's is neither constant or Gaussian so model •1• has trouble reflecting the actual distribution of the Y's around each μ.

One should ask what <u>the important parameters</u> are in each specific application. If the objective is communication with the patient, then compare this prognosis example with a chest physician's use of say a patient's age, gender, height and weight to estimate the average FEV (<u>F</u>orced <u>E</u>xpiratory <u>V</u>olume) in patients with these "X" values, and to express this patient's FEV in relation to that average. Does it matter if the chest physician models the log of FEV?

8.3 ALL POSSIBLE REGRESSIONS

"Pool" of **P - 1** possible **terms** + the "1" term implicit in the β_0 = **P** possible terms in all.
Reduce to p - 1 **terms** + the "1" term implicit in the β_0 = p terms in all.

Try to find a few (3-6) "good" subsets for closer examination
- can specify target p in many software packages

Criteria for "good" [all indexed by p]	Principle / Comments
• R^2 (or SSE)	$R^2 = 1 - \frac{SSE}{SS_{total}}$ - increases with p - over-optimistic if p -> n
• MSE (same as $R^2_{adjusted}$)	$R^2_{adjusted} = 1 - \frac{SSE / (n-p)}{SS_{total} / (n-1)}$ does not continue to increase with p - at some point, SSR(extra term) < MSE_p
• C_p (" Mallow's C_p")	idea of "bias" or error in fitted model $\frac{\text{Total Mean Squared Error}}{\text{True Error Variance}}$ Uses the assumption that error variance estimated from model with all P terms is an unbiased estimate (i.e. that the P terms contain all the important ones and that no important ones were overlooked)
• PREdiction Sum Of Squares (PRESS)	how well does prediction based on n-1 of the observations do in predicting the omitted observation? $PRESS_p = \Sigma (Y - \hat{Y}$ from **other** n-1 X's $)^2$ cf. $SSE_p = \Sigma (Y - \hat{Y}$ from **same** n X's $)^2$

More details on C_p

$\underline{X}$ is a P-dimensional vector $(1, X_1, X_2, \ldots, X_{P-1})$ at which one of the n Y's is observed.

$\underline{x}$ is a p-dimensional vector (1, ...) , a subset of $\underline{X}$.

If a particular model, based on a subset $\underline{x}$ with p < P components, is not "correct" (some important components of $\underline{X}$ omitted), then, over all possible datasets with same set of n $\underline{X}$'s,

the average $\hat{Y} \mid \underline{x} \neq$ the "correct" $E[\, Y \mid \underline{X}\,] = \mu[\, Y \mid \underline{X}\,]$

i.e. $\hat{Y} \mid \underline{x}$ is a biased estimator of $\mu[Y \mid \underline{X}\,]$; denote the difference as " bias[P,p] ".

In any one application (dataset) ...

$\hat{Y} \mid \underline{x} = \mu[Y \mid \underline{X}\,]$ + bias[P,p] + random sampling error in $\hat{Y} \mid \underline{x}$

(random sampling error is around $\mu[Y \mid \underline{X}\,]$ + bias[P,p])

average $[\, \{ \hat{Y} \mid \underline{x} - \mu[Y \mid \underline{X}\,] \}^2 \,] = \{ \text{bias}[P,p] \}^2 + \text{variance} \{ \hat{Y} \mid \underline{x} \}$

Sum these biases over all n $\underline{X}$'s

Σ average squared error = $\Sigma \{ \text{bias}[P,p] \}^2 + \Sigma \{\text{variance} \{ \hat{Y} \mid \underline{x} \}$

Scale this sum by dividing by "true" error variance $\sigma[\varepsilon]^2$.

$$\Gamma_p = \frac{\Sigma \{ \text{bias[using p rather than P]} \}^2 + \Sigma \{\text{variance} \{ \hat{Y} \mid \underline{x} \}}{\sigma[\varepsilon]^2}$$

Estimate this by substituting $MSE_{\mathbf{P}}$ for $\sigma[\varepsilon]^2$ and MSE_p for the variance term in var $\{ \hat{Y} \mid \underline{x} \}$ †

$$C_p = \frac{SSE_p}{MSE_{\mathbf{P}}} - (n - 2p).$$

If model with p variables is unbiased, then $E[C_p]$ p . Otherwise, C_p will tend to be > p

So, look for "elbow" in "C_p vs p" plot.

---

† **Note 3** (p344):

(a) why $\Sigma\, \text{var}[\, \hat{Y} \mid \underline{x}\,] = p\, \sigma^2$: The n × 1 vector $\hat{\mathbf{Y}} = H_p \mathbf{Y}$, where H_p is the n × n hat matrix $X(X^TX)^{-1}X^T$, so $\text{var}[\,\hat{\mathbf{Y}}\,] = H_p(\sigma^2 I)H_p = \sigma^2 H_p$; thus $\Sigma\, \text{var}(\hat{\mathbf{Y}}) = \sigma^2 (\text{trace } H_p) = \sigma^2 p$.

(b) why $E[\, SSE_p] = \Sigma\, \text{bias}^2 + (n - p)\, \sigma^2$: $\mathbf{e}_p = (I-H_p)\mathbf{Y}$; $SSE_p = \mathbf{e}_p^T \mathbf{e}_p$. The ith element of $\mathbf{e}_p$ has expectation = bias_i, and variance = $(1-h_{ii})\sigma^2$ where h_{ii} is the ith diagonal entry of H. The expected squared residual = bias^2 + var, so their Σ over n = $\Sigma\, \text{bias}^2 + \sigma^2 (\text{trace } [I-H_p]) = \Sigma\, \text{bias}^2 + (n - p)\sigma^2$.

8.3 "Best" Subsets REGRESSION (Text p 346)

```
proc reg; model l_100km = Cylinder EngSize rpm domestic van
                          sporty front_dr allwh_dr /selection = cp best = 5 ;
```

C(p)	R-sq	# Vars in Model	
2.51	0.684	4	CYLINDER ENGSIZE VAN ALLWH_DR
2.93	0.675	3	CYLINDER ENGSIZE VAN
3.33	0.681	4	CYLINDER ENGSIZE VAN FRONT_DR
3.49	0.688	5	CYLINDER ENGSIZE DOMESTIC VAN ALLWH_DR
4.00	0.678	4	CYLINDER ENGSIZE DOMESTIC VAN

`/selection = cp stop = 3 ;`

C(p)	R-sq	# Vars in Model	
2.935	0.675	3	CYLINDER ENGSIZE VAN
4.319	0.670	3	CYLINDER VAN FRONT_DR
5.189	0.667	3	CYLINDER VAN ALLWH_DR
5.200	0.659	2	CYLINDER VAN
5.853	0.664	3	CYLINDER RPM VAN
6.253	0.663	3	ENGSIZE VAN ALLWH_DR
6.698	0.653	2	ENGSIZE VAN
6.781	0.661	3	CYLINDER VAN SPORTY
7.102	0.659	3	CYLINDER DOMESTIC VAN
7.116	0.659	3	ENGSIZE DOMESTIC VAN
7.455	0.658	3	ENGSIZE VAN FRONT_DR
7.899	0.656	3	ENGSIZE VAN SPORTY
8.681	0.654	3	ENGSIZE RPM VAN
27.976	0.582	3	RPM VAN FRONT_DR
31.649	0.560	2	RPM VAN
32.308	0.565	3	RPM VAN ALLWH_DR
32.345	0.565	3	RPM DOMESTIC VAN
33.432	0.561	3	RPM VAN SPORTY
37.181	0.547	3	CYLINDER ENGSIZE ALLWH_DR
41.111	0.533	3	CYLINDER RPM ALLWH_DR
41.175	0.525	2	CYLINDER ALLWH_DR
42.050	0.529	3	ENGSIZE DOMESTIC ALLWH_DR
42.689	0.527	3	CYLINDER SPORTY ALLWH_DR
42.698	0.519	2	ENGSIZE ALLWH_DR
42.914	0.526	3	CYLINDER DOMESTIC ALLWH_DR
43.160	0.525	3	CYLINDER FRONT_DR ALLWH_DR
43.235	0.525	3	ENGSIZE FRONT_DR ALLWH_DR
44.502	0.520	3	ENGSIZE SPORTY ALLWH_DR
44.648	0.519	3	ENGSIZE RPM ALLWH_DR
57.175	0.473	3	CYLINDER ENGSIZE FRONT_DR
57.654	0.463	2	CYLINDER FRONT_DR
58.346	0.468	3	CYLINDER RPM FRONT_DR
58.707	0.467	3	CYLINDER SPORTY FRONT_DR
59.485	0.464	3	CYLINDER DOMESTIC FRONT_DR
61.597	0.449	2	CYLINDER ENGSIZE
61.991	0.455	3	CYLINDER ENGSIZE DOMESTIC
63.522	0.449	3	CYLINDER ENGSIZE SPORTY
63.559	0.449	3	CYLINDER ENGSIZE RPM
64.440	0.431	1	CYLINDER
64.456	0.438	2	ENGSIZE FRONT_DR
64.584	0.437	2	CYLINDER RPM
64.619	0.445	3	ENGSIZE DOMESTIC FRONT_DR
65.306	0.442	3	CYLINDER RPM DOMESTIC
66.048	0.439	3	ENGSIZE SPORTY FRONT_DR
66.124	0.432	2	CYLINDER DOMESTIC
66.254	0.439	3	ENGSIZE RPM FRONT_DR
66.279	0.431	2	CYLINDER SPORTY
66.378	0.438	3	CYLINDER RPM SPORTY
67.310	0.435	3	VAN FRONT_DR ALLWH_DR
67.494	0.427	2	ENGSIZE DOMESTIC
67.997	0.432	3	CYLINDER DOMESTIC SPORTY
68.148	0.417	1	ENGSIZE

8.3 "Best" Subsets REGRESSION (Text p 346)

```
proc reg; model l_100km = Cylinder EngSize rpm domestic van
                        sporty front_dr allwh_dr

                                    /selection = adjrsq best = 5;
```

```
Adj.        # Vars in Model
R-sq  R-sq

0.670 0.688 5  CYLINDER ENGSIZE DOMESTIC VAN ALLWH_DR
0.670 0.684 4  CYLINDER ENGSIZE VAN ALLWH_DR
0.668 0.689 6  CYLINDER ENGSIZE DOMESTIC VAN SPORTY ALLWH_DR
0.667 0.685 5  CYLINDER ENGSIZE VAN SPORTY ALLWH_DR
0.666 0.681 4  CYLINDER ENGSIZE VAN FRONT_DR
```

```
                                    /selection = maxr stop = 4;
```

Step 1 CYLINDER Entered R-sq = 0.431 C(p) = 64.44

	DF	SS	MS	F	Prob>F
Regression	1	76.11	76.11	68.17	0.0001
Error	90	100.47	1.11		
Total	91	176.58			

Step 2 VAN Entered R-sq = 0.659 C(p) = 5.20

	DF	SS	MS	F	Prob>F
Regression	2	116.47	58.23	86.23	0.0001
Error	89	60.11	0.67		
Total	91	176.58			

Step 3 ENGSIZE Entered R-sq = 0.675 C(p) = 2.93

	DF	Sum of Sq	MS	F	Prob>F
Regression	3	119.28	39.76	61.06	0.0001
Error	88	57.30	0.65		
Total	91	176.58			

Step 4 ALLWH_DR Entered R-sq = 0.684 C(p) = 2.51

	DF	Sum of Sq	MS	F	Prob>F
Regression	4	120.88	30.22	47.20	0.0001
Error	87	55.70	0.64		
Total	91	176.58			

Variable	Parameter Estimate	Standard Error	Type II SS	F	Prob>F
INTERCEP	5.30	0.365	135.02	210.88	0.0001
CYLINDER	0.34	0.141	3.78	5.91	0.0171
ENGSIZE	0.39	0.179	3.08	4.81	0.0309
VAN	2.00	0.326	24.16	37.75	0.0001
ALLWH_DR	0.48	0.306	1.59	2.49	0.1181

8.4 Forward, Forward Stepwise, and Backward Elimination Search Procedures

[if large (> a dozen, say) terms ... each produces a single "best" model]

Forward Stepwise Procedure

Terms In Model {in} | Terms Not In Model {out}.

cycle (i)

From {out}, identify X_{max} with max F*

Add X_{max} if F* > "F-to-Enter"

From {in}, identify X_{min} with min F* →

Take out X_{min} if F* < "F-to-Stay"

...

N.B. Set F-to-Enter (or P-to-Enter) more extreme than F-to-Stay (or P-to-Stay). (Also, meaning of P-values not exact with with repeat testing)

Forward Selection Procedure

From {out}, identify X_{max}

Add X_{max} if > "F-to-Enter"

...

Backward Elimination Procedure

From {in}, identify X_{min} →

Take out X_{min} if F* < "F-to-Stay"

...

Backward Stepwise Procedure

From {in}, identify X_{min} →

Take out X_{min} if F* < "F-to-Stay"

From {out}, identify X_{max} (has max F*)

Add X_{max} if F* > "F-to-Enter"

...

```
/ selection = stepwise  slentry=0.05 slstay=0.10  details

         Stepwise Procedure for Dependent Variable L_100KM

                    Statistics for Entry: Step 1
                              DF = 1,90

                                 Model
     Variable    Tolerance        R**2            F       Prob>F

     CYLINDER     1.000000     0.4310       68.1748       0.0001
     ENGSIZE      1.000000     0.4172       64.4188       0.0001
     RPM          1.000000     0.3001       38.5971       0.0001
     DOMESTIC     1.000000     0.0258        2.3814       0.1263
     VAN          1.000000     0.3446       47.3186       0.0001
     SPORTY       1.000000     0.0011        0.1017       0.7505
     FRONT_DR     1.000000     0.1441       15.1564       0.0002
     ALLWH_DR     1.000000     0.0871        8.5871       0.0043

Step 1   CYLINDER Entered  R-square = 0.431   C(p) = 64.44

                    Statistics for Entry: Step 2
                              DF = 1,89

                                 Model
     Variable    Tolerance        R**2            F       Prob>F

     ENGSIZE      0.206596     0.4491        2.9201       0.0910
     RPM          0.436376     0.4379        1.0966       0.2978
     DOMESTIC     0.913235     0.4322        0.1849       0.6682
     VAN          0.968509     0.6596       59.7628       0.0001
     SPORTY       0.999808     0.4316        0.0941       0.7598
     FRONT_DR     0.899742     0.4638        5.4432       0.0219
     ALLWH_DR     0.999671     0.5253       17.6810       0.0001

Step 2   VAN Entered       R-square = 0.659   C(p) =  5.20

                    Statistics for Removal: Step 3
                              DF = 1,89

                              Partial     Model
                  Variable       R**2      R**2

                  CYLINDER     0.3150    0.3446
                  VAN          0.2286    0.4310

                    Statistics for Entry: Step 3
                              DF = 1,88

                                 Model
     Variable    Tolerance        R**2            F       Prob>F

     ENGSIZE      0.206534     0.6755        4.3171       0.0406
     RPM          0.436085     0.6646        1.3192       0.2538
     DOMESTIC     0.912307     0.6600        0.0946       0.7591
     SPORTY       0.982056     0.6612        0.4062       0.5255
     FRONT_DR     0.875287     0.6703        2.8699       0.0938
     ALLWH_DR     0.765756     0.6671        1.9836       0.1625

Step 3   ENGSIZE Entered   R-square = 0.675 C(p) =  2.93
```

/ selection = stepwise slentry=0.05 slstay=0.10 details CONTINUED

Statistics for Removal: Step 4
DF = 1,88

Variable	Partial R**2	Model R**2
CYLINDER	0.0215	0.6540
ENGSIZE	0.0159	0.6596
VAN	0.2264	0.4491

Statistics for Entry: Step 4
DF = 1,87

Variable	Tolerance	Model R**2	F	Prob>F
RPM	0.306882	0.6755	0.0019	0.9654
DOMESTIC	0.832604	0.6790	0.9450	0.3337
SPORTY	0.978957	0.6777	0.5890	0.4449
FRONT_DR	0.830951	0.6815	1.6310	0.2050
ALLWH_DR	0.762499	0.6845	2.4920	0.1181

All variables left in the model are significant at the 0.10 level.
No other variable met the 0.05 significance level for entry into the model.

Summary of Stepwise Procedure for Dependent Variable L_100KM

Step	Variable Entered	Removed	Number In	Partial R**2	Model R**2	C(p)	F	Prob>F
1	CYLINDER		1	0.431	0.4310	64.4402	68.1748	0.0001
2	VAN		2	0.228	0.6596	5.2001	59.7628	0.0001
3	ENGSIZE		3	0.015	0.6755	2.9352	4.3171	0.0406

selection = backward slentry=0.05 slstay=0.10 details

```
        Backward Elimination Procedure for Dependent Variable L_100KM

Step 0    All Variables Entered     R-square = 0.69019826   C(p) =  9.00000000

                DF         Sum of Squares      Mean Square          F    Prob>F

 Regression      8           121.87968268      15.23496034      23.11    0.0001
 Error          83            54.70679842       0.65911805
 Total          91           176.58648111

                Parameter        Standard          Type II
 Variable        Estimate           Error   Sum of Squares          F    Prob>F

 INTERCEP      5.62729863      1.23885070      13.59957045      20.63    0.0001
 CYLINDER      0.31820912      0.14663650       3.10387561       4.71    0.0329
 ENGSIZE       0.42816411      0.23612525       2.16719942       3.29    0.0734
 RPM          -0.00008224      0.00031917       0.04375972       0.07    0.7973
 DOMESTIC     -0.21009530      0.19595870       0.75764680       1.15    0.2868
 VAN           2.02835261      0.34330399      23.00871758      34.91    0.0001
 SPORTY        0.15755552      0.25574944       0.25015032       0.38    0.5395
 FRONT_DR     -0.01048277      0.28485068       0.00089265       0.00    0.9707
 ALLWH_DR      0.45557051      0.41213654       0.80536392       1.22    0.2722

                        Statistics for Removal: Step 1
                                  DF = 1,83

                                     Partial     Model
                         Variable       R**2      R**2

                         CYLINDER     0.0176   0.6726
                         ENGSIZE      0.0123   0.6779
                         RPM          0.0002   0.6900
                         DOMESTIC     0.0043   0.6859
                         VAN          0.1303   0.5599
                         SPORTY       0.0014   0.6888
                         FRONT_DR     0.0000   0.6902
                         ALLWH_DR     0.0046   0.6856

Step 1   Variable FRONT_DR Removed  R-square = 0.69019321   C(p) =  7.00135431

                DF         Sum of Squares      Mean Square          F    Prob>F

 Regression      7           121.87879003      17.41125572      26.73    0.0001
 Error          84            54.70769107       0.65128204
 Total          91           176.58648111

                Parameter        Standard          Type II
 Variable        Estimate           Error   Sum of Squares          F    Prob>F

 INTERCEP      5.61220184      1.16198895      15.19258476      23.33    0.0001
 CYLINDER      0.31756644      0.14472481       3.13582849       4.81    0.0310
 ENGSIZE       0.43138882      0.21795704       2.55132087       3.92    0.0511
 RPM          -0.00008156      0.00031673       0.04318430       0.07    0.7974
 DOMESTIC     -0.21143635      0.19139269       0.79483632       1.22    0.2724
 VAN           2.02598220      0.33519684      23.79252403      36.53    0.0001
 SPORTY        0.15958308      0.24825546       0.26911950       0.41    0.5221
 ALLWH_DR      0.46537255      0.31262849       1.44315684       2.22    0.1403
```

selection = backward slentry=0.05 slstay=0.10 details CONTINUED

Statistics for Removal: Step 2

DF = 1,84

Variable	Partial R**2	Model R**2
CYLINDER	0.0178	0.6724
ENGSIZE	0.0144	0.6757
RPM	0.0002	0.6899
DOMESTIC	0.0045	0.6857
VAN	0.1347	0.5555
SPORTY	0.0015	0.6887
ALLWH_DR	0.0082	0.6820

Step 2 Variable RPM Removed R-square = 0.68994866 C(p) = 5.06687261

	DF	Sum of Squares	Mean Square	F	Prob>F
Regression	6	121.83560574	20.30593429	31.52	0.0001
Error	85	54.75087537	0.64412795		
Total	91	176.58648111			

Variable	Parameter Estimate	Standard Error	Type II Sum of Squares	F	Prob>F
INTERCEP	5.32860210	0.36841192	134.75080686	209.20	0.0001
CYLINDER	0.31998873	0.14362339	3.19735729	4.96	0.0285
ENGSIZE	0.45905989	0.18858048	3.81695798	5.93	0.0170
DOMESTIC	-0.19936982	0.18454546	0.75176871	1.17	0.2831
VAN	2.02929857	0.33310460	23.90577213	37.11	0.0001
SPORTY	0.16375942	0.24636075	0.28460444	0.44	0.5080
ALLWH_DR	0.46453503	0.31088987	1.43812271	2.23	0.1388

Statistics for Removal: Step 3

DF = 1,85

Variable	Partial R**2	Model R**2
CYLINDER	0.0181	0.6718
ENGSIZE	0.0216	0.6683
DOMESTIC	0.0043	0.6857
VAN	0.1354	0.5546
SPORTY	0.0016	0.6883
ALLWH_DR	0.0081	0.6818

Step 3 Variable SPORTY Removed R-square = 0.68833696 C(p) = 3.49866847

	DF	Sum of Squares	Mean Square	F	Prob>F
Regression	5	121.55100129	24.31020026	37.99	0.0001
Error	86	55.03547981	0.63994744		
Total	91	176.58648111			

Variable	Parameter Estimate	Standard Error	Type II Sum of Squares	F	Prob>F
INTERCEP	5.33975349	0.36683352	135.59656477	211.89	0.0001
CYLINDER	0.32636222	0.14283718	3.34088490	5.22	0.0248
ENGSIZE	0.44966581	0.18743894	3.68302269	5.76	0.0186
DOMESTIC	-0.18722506	0.18304191	0.66953176	1.05	0.3092
VAN	1.98963874	0.32665239	23.74220664	37.10	0.0001
ALLWH_DR	0.49280572	0.30696596	1.64935958	2.58	0.1121

selection = backward slentry=0.05 slstay=0.10 details CONTINUED

Statistics for Removal: Step 4
DF = 1,86

Variable	Partial R**2	Model R**2
CYLINDER	0.0189	0.6694
ENGSIZE	0.0209	0.6675
DOMESTIC	0.0038	0.6845
VAN	0.1345	0.5539
ALLWH_DR	0.0093	0.6790

Step 4 Variable DOMESTIC Removed R-square = 0.68454544 C(p) = 2.51446793

	DF	Sum of Squares	Mean Square	F	Prob>F
Regression	4	120.88146953	30.22036738	47.20	0.0001
Error	87	55.70501157	0.64028749		
Total	91	176.58648111			

Variable	Parameter Estimate	Standard Error	Type II Sum of Squares	F	Prob>F
INTERCEP	5.30355882	0.36521981	135.02094830	210.88	0.0001
CYLINDER	0.34456671	0.14176168	3.78271792	5.91	0.0171
ENGSIZE	0.39280256	0.17905259	3.08150191	4.81	0.0309
VAN	2.00521810	0.32638377	24.16802325	37.75	0.0001
ALLWH_DR	0.48453473	0.30694096	1.59556738	2.49	0.1181

Statistics for Removal: Step 5
DF = 1,87

Variable	Partial R**2	Model R**2
CYLINDER	0.0214	0.6631
ENGSIZE	0.0175	0.6671
VAN	0.1369	0.5477
ALLWH_DR	0.0090	0.6755

Step 5 Variable ALLWH_DR Removed R-square = 0.67550982 C(p) = 2.93522907

	DF	Sum of Squares	Mean Square	F	Prob>F
Regression	3	119.28590216	39.76196739	61.06	0.0001
Error	88	57.30057895	0.65114294		
Total	91	176.58648111			

Variable	**Parameter Estimate**	**Standard Error**	**Type II Sum of Sq**	**F**	**Prob>F**
INTERCEP	**5.37**	**0.365**	**141.14122366**	**216.76**	**0.0001**
CYLINDER	**0.34**	**0.142**	**3.79891080**	**5.83**	**0.0178**
ENGSIZE	**0.37**	**0.180**	**2.81103746**	**4.32**	**0.0406**
VAN	**2.25**	**0.287**	**39.98361647**	**61.41**	**0.0001**

selection = backward slentry=0.05 slstay=0.10 details CONTINUED

Statistics for Removal: Step 6

DF = 1,88

Variable	Partial R**2	Model R**2
CYLINDER	0.0215	0.6540
ENGSIZE	0.0159	0.6596
VAN	0.2264	0.4491

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

Summary of Backward Elimination Procedure for Dependent Variable L_100KM

Step	Variable Removed	Number In	Partial R**2	Model R**2	C(p)	F	Prob>F
1	FRONT_DR	7	0.0000	0.6902	7.0014	0.0014	0.9707
2	RPM	6	0.0002	0.6899	5.0669	0.0663	0.7974
3	SPORTY	5	0.0016	0.6883	3.4987	0.4418	0.5080
4	DOMESTIC	4	0.0038	0.6845	2.5145	1.0462	0.3092
5	ALLWH_DR	3	0.0090	0.6755	2.9352	2.4920	0.1181