Exam Practice Questions

Calculate the estimators I:

Consider the following dataset, where Y is the dependent variable and X1 and X2 are the independent variables:

Observation	X1	X2	Y
1	1	2	5
2	2	3	10
3	3	4	15
4	4	5	20

Calculate the β as **X' * X** and **X' *Y.**

Result

We want to fit a linear regression model of the form Y = β0 + β1X1 + β2X2 + e

Compute the least squares estimates of β using the matrix formula:

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

Now, compute X^TX:

$$X^TX = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 2 & 3 & 4 \\ 2 & 3 & 4 & 5 \end{bmatrix} \begin{bmatrix} 1 & 1 & 2 \\ 1 & 2 & 3 \\ 1 & 3 & 4 \\ 1 & 4 & 5 \end{bmatrix} = \begin{bmatrix} 4 & 10 & 14 \\ 10 & 30 & 40 \\ 14 & 40 & 54 \end{bmatrix}$$

Next, compute X^TY:

$$X^TY = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 2 & 3 & 4 \\ 2 & 3 & 4 & 5 \end{bmatrix} \begin{bmatrix} 5 \\ 10 \\ 15 \\ 20 \end{bmatrix} = \begin{bmatrix} 50 \\ 150 \\ 200 \end{bmatrix}$$

We got the β0, β1 and β2 when we multiply (X^T * X)**-1 and X^T * Y.

Calculate the estimators II:

The inverse of XT * X is given:

The inverse of X^TX is:

$$(X^TX)^{-1} = \begin{bmatrix} 1.5 & -0.5 & 0 \\ -0.5 & 0.5 & 0 \\ 0 & 0 & 0.5 \end{bmatrix}$$

Find the estimators

Result

Final Answer:

$$\beta = \begin{bmatrix} 0 \\ 50 \\ 100 \end{bmatrix}$$

Final Model:

The linear regression model is:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

Substituting the values of β_0, β_1, and β_2:

$$Y = 0 + 50X_1 + 100X_2$$

Simplifying:

$$Y = 50X_1 + 100X_2$$

Given our estimated model, calculate the MS-reg and MS-res:

Result

- **Regression Sum of Squares (SSR):**

$$SSR = \sum(\hat{Y}_i - \bar{Y})^2$$
$$SSR = (250 - 12.5)^2 + (400 - 12.5)^2 + (550 - 12.5)^2 + (700 - 12.5)^2$$
$$SSR = (237.5)^2 + (387.5)^2 + (537.5)^2 + (687.5)^2$$
$$SSR = 56406.25 + 150156.25 + 288906.25 + 472656.25$$
$$SSR = 967125$$

- **Residual Sum of Squares (SSE):**

$$SSE = \sum(Y_i - \hat{Y}_i)^2$$
$$SSE = (5 - 250)^2 + (10 - 400)^2 + (15 - 550)^2 + (20 - 700)^2$$
$$SSE = (-245)^2 + (-390)^2 + (-535)^2 + (-680)^2$$
$$SSE = 60025 + 152100 + 286225 + 462400$$
$$SSE = 960750$$

Step 3: Degrees of Freedom

- **Degrees of Freedom for Regression (df Reg):**

$$df_{Reg} = k = 2$$

(number of predictors)

- **Degrees of Freedom for Residuals (df Res):**

$$df_{Res} = n - k - 1 = 4 - 2 - 1 = 1$$

Step 4: Calculate MS Reg and MS Res

- **Mean Square Regression (MS Reg):**

$$MS_{Reg} = \frac{SSR}{df_{Reg}} = \frac{967125}{2} = 483562.5$$

- **Mean Square Residual (MS Res):**

$$MS_{Res} = \frac{SSE}{df_{Res}} = \frac{960750}{1} = 960750$$

Now, form the hypothesis for our model, is our model significant (α = 0.05)?

Result

H0: β1 = β2 = 0
H1: At least one of our predictors are non-zero.

From the previous calculations:

- $MS_{\text{Reg}} = 483562.5$
- $MS_{\text{Res}} = 960750$

So:

$$F_{\text{observation}} = \frac{483562.5}{960750} \approx 0.503$$

- $df_1 = 2$ (numerator degrees of freedom, e.g., for regression)
- $df_2 = 1$ (denominator degrees of freedom, e.g., for residuals)

From the table:

- Go to the row for $df_2 = 1$.
- Go to the column for $df_1 = 2$.
- The critical F-value is **199.50**.

0.5 < 199: Fail to reject the H0, our model is **NOT SİGNİFİCANT.**

Since $F_{\text{observation}} < F_{\text{table}}$, we **fail to reject the null hypothesis**. This means that the regression model does not provide a statistically significant fit to the data at the $\alpha = 0.05$ level.

***Does adding X3, X4, X5, X6, and X7 to the model have a significant impact on improving the model's predictions? Form the hypothesis and test.

MLR EXAMPLE 11.1 - GAS MILEAGE 4

Model: MODEL2
Dependent Variable: MPG

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	2	1339.21617	669.60809	66.903	0.0001
Error	29	290.25101	10.00866		
C Total	31	1629.46719			

Root MSE	3.16365	R-square	0.8219
Dep Mean	20.29062	Adj R-sq	0.8096
C.V.	15.59166		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > |T|
INTERCEP	1	33.491421	1.47544969	22.699	0.0001
SIZE	1	-0.040148	0.00990172	-4.055	0.0003
HP	1	-0.016611	0.02250347	-0.738	0.4663

MLR EXAMPLE 11.1 - GAS MILEAGE 2

Model: MODEL1
Dependent Variable: MPG

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	7	1466.87390	209.55341	30.932	0.0001
Error	24	162.59329	6.77472		
C Total	31	1629.46719			

Root MSE	2.60283	R-square	0.9002
Dep Mean	20.29062	Adj R-sq	0.8711
C.V.	12.82774		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > |T|
INTERCEP	1	45.026054	12.37472908	3.639	0.0013
SIZE	1	0.077614	0.03039037	2.554	0.0174
HP	1	-0.032959	0.01916261	-1.720	0.0983
WEIGHT	1	-0.008264	0.00228236	-3.621	0.0014
SHAPE	1	1.283181	2.77643544	0.462	0.6481
CYLIN	1	-3.169670	1.44712974	-2.190	0.0385
TRANS	1	0.510739	3.58673059	0.142	0.8880
SPEEDS	1	1.869099	2.78918048	0.670	0.5092

Variable	DF	Type II SS
INTERCEP	1	89.690772
SIZE	1	44.187532
HP	1	20.040986
WEIGHT	1	88.815898
SHAPE	1	1.447078
CYLIN	1	32.501586
TRANS	1	0.137370
SPEEDS	1	3.042306

Result

H0: β3 = β4 = β5 = β6 = β7 = 0
H1: At least one of the predictors are non-zero.

R(β3, β4, β5, β6, β7 | β1, β2) **=** SSreg full - SSreg sub

Fobs = (SSreg full - SSreg sub / p) / MSres full model

To form this hypothesis we first look at the SSreg of the sub-model =1339,
Next, we look at the SSreg of the full model containing X3, X4, X5, X6 = 1466

1. SSreg full model - SSreg sub model = 1466 -1339 = 127
2. Divide it by the number of parameters in the model = 127 / 5 = 25.4
3. Divide it by MSres full model = 25.4 / 6.775 = 3.75 = Fobs
4. Look at the F table for (p, n-p-1) = F(5,24) = 2.62
5. Fobs (3.75) > Ftable (2.62), **Reject the null hypothesis**
6. Rejecting the H0 means **adding these predictors significantly improves the model's predictions.**

<u>F test:</u> $R(\beta_3, \beta_4, \beta_5, \beta_6, \beta_7 \mid \beta_1, \beta_2) = R(\beta_1, \beta_2, \ldots, \beta_7) - R(\beta_1, \beta$

$= 1466.874 - 1339.216$

$= 127.658$

$$F_{obs} = \frac{R(\beta_3, \beta_4, \beta_5, \beta_6, \beta_7 \mid \beta_1, \beta_2) / 5}{MSres_{full}} = \frac{127.658/5}{6.775}$$

$= 3.769 \sim F_{(5,24)}$

$F_{(5,24), .05} = 2.62 \Rightarrow$ reject H_0 $(.01 < p < .05)$

Does removing β4, β3, and β2 sequentially reduce the model's prediction quality significantly? (MSres full model = 0.0499, n = 20)

d) Type I SS = sequential SS

$R(\beta_1) = R(\beta_1) = 22.5590$

$R(\beta_2|\beta_1) = R(\beta_1, \beta_2) - R(\beta_1) = 0.2825$

$R(\beta_3|\beta_1, \beta_2) = R(\beta_1, \beta_2, \beta_3) - R(\beta_1, \beta_2) = 0.1353$

$R(\beta_4|\beta_1, \beta_2, \beta_3) = R(\beta_1, \beta_2, \beta_3, \beta_4) - R(\beta_1, \beta_2, \beta_3) = 0.0041$

notes:

(i) adding these sequential SS yields:

$R(\beta_1, \beta_2, \beta_3, \beta_4) = 22.9809 = SSreg_{full}$

Result

X4: Fobs = 0.0041 / 0.0499 = 0.082, F(1, 15) = 4.54, **Accept the H0** and **drop** X4,
X3: Fobs = 0.1353 / 0.0499 = 2.710, F(1,15) = 4.54, **Accept the H0** and **drop** X3,
X2: Fobs= 0.2825 / 0.0499 = 5.65, F(1,15) = 4.54, **Reject the H0** and **keep** X2.

e) example

Do we need X^4?

$$F_{obs} = \frac{R(\beta_4 | \beta_1, \beta_2, \beta_3) / 1}{MSres_{full}} = \frac{.0041}{.0499} = .082 \sim F_{(1,1}$$

$F_{(1,15), .95} = 4.54 \Rightarrow$ non-significant $\Rightarrow$ drop X

Do we need X^3?

$$F_{obs} = \frac{R(\beta_3 | \beta_1, \beta_2) / 1}{MSres_{full}} = \frac{.1353}{.0499} = 2.710 < 4.54$$

$\Rightarrow$ non-significant $\Rightarrow$ drop X^3

Do we need X^2?

$$F_{obs} = \frac{R(\beta_2 | \beta_1) / 1}{MSres_{full}} = \frac{.2825}{.0499} = 5.657 > 4.54$$

$\Rightarrow$ significant

$\Rightarrow$ stop and include X^2 and X $\quad \boxed{Y = \beta_0 + \beta_1 X + \beta_2 X^2}$

After deciding that 2^{nd} order model is adequate, we must run the 2^{nd} order model to obtain the final prediction equation.

<u>input</u>: PROC REG; MODEL Y = X X2; } to run 2^{nd} order model

<u>output</u> from this is on p.3

Final prediction equation is:

$$\hat{Y} = .1776 + 2.0469 X - .1756 X^2$$

Can we remove the β4 from the model?

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	7	1466.87390	209.55341	30.932	0.0001
Error	[illegible]	[illegible]	6.77472		
C Total	31	1629.46719			

Root MSE	2.60283	R-square	0.9002
Dep Mean	20.29062	Adj R-sq	0.8711
C.V.	12.82774		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > |T|
INTERCEP	1	45.026054	12.37472908	3.639	0.0013
SIZE	1	0.077614	0.03039037	2.554	0.0174
HP	1	-0.032959	0.01916261	-1.720	0.0983
WEIGHT	1	-0.008264	0.00228236	-3.621	0.0014
SHAPE	[illegible]	[illegible]	[illegible]	0.462	0.6481
CYLIN	1	-3.169670	1.44712974	-2.190	0.0385
TRANS	1	0.510739	3.58673059	0.142	0.8880
SPEEDS	1	1.869099	2.78918048	0.670	0.5092

Variable	DF	Type II SS
INTERCEP	1	89.690772
SIZE	1	44.187532
HP	1	20.040986
WEIGHT	1	88.815898
SHAPE	1	1.447078
CYLIN	1	32.501586
TRANS	1	0.137370
SPEEDS	1	3.042306

Result

To see if we can remove the β4 from the model we have to look at the predictors.

β1 = SIZE, B2 = HP, β3 = WEIGHT, β4 = SHAPE, ...

SHAPE predictor has 1.4471 Type I SS, which is equivalent to:

R(β4 | β1, β2, β3, β5, β6, β7) = SS full - SS sub

F-obs = (R(β4 | β1, β2, β3, β5, β6, β7) / p) / MS res full = ((SS full - SSsub) / p) / MSres full model

(1.4471 / 1) / 6.774 = 0.214

Now we want to compare the F-obs with F-table = F(1,24) = 4.26

We now can compare F-obs < F-table, **Fail to Reject H0**, We can drop **X4** because its contribution to the model is **not significant**.

Note: You can directly look at the **"T for H0: Parameter = 0"** value from the table to see if the parameter is significant according to Type II SS (partial SS)

!!! F-obs = T-obs^2 — 0.462^2 = 0.214

c) Type II SS = partial SS

Test of each vbl. as if it were the last to be added to the model. For each vbl. separately we ask whether it would be worth adding it all the other predictors were already in the model.

example 11.1 (p. 2 of output):

$R(\beta_1 | \beta_2, \beta_3, \beta_4, \beta_5, \beta_6, \beta_7) = 44.1875$

$R(\beta_2 | \beta_1, \beta_3, \beta_4, \beta_5, \beta_6, \beta_7) = 20.0410$

⋮

(ii) each of these can be tested as a partial F test.

example: $H_0: \beta_4 = 0$ in full model

$$F_{obs} = \frac{R(\beta_4 | \beta_1, \beta_2, \beta_3, \beta_5, \beta_6, \beta_7)/1}{MS\,res_{full}} = \frac{1.4471/1}{6.7747}$$

$= .214 \sim F_{(1,24)}$

$F_{(1,24),.95} = 4.26 \Rightarrow$ accept H_0 for $\alpha = .05$

$\Rightarrow$ we could delete X_4 from the full model without significant loss in predictability

(iii) SAS output contains

$t_{obs} = \sqrt{F_{obs}}$ under heading

("T FOR H_0: PARAMETER = 0"). Note that for

X_4 (SHAPE), $t_{obs} = .462 = \sqrt{.214} = \sqrt{F_{obs}}$,

p-value is .6481, consistent with our conclusion

Explain

Explain the following SAS inputs

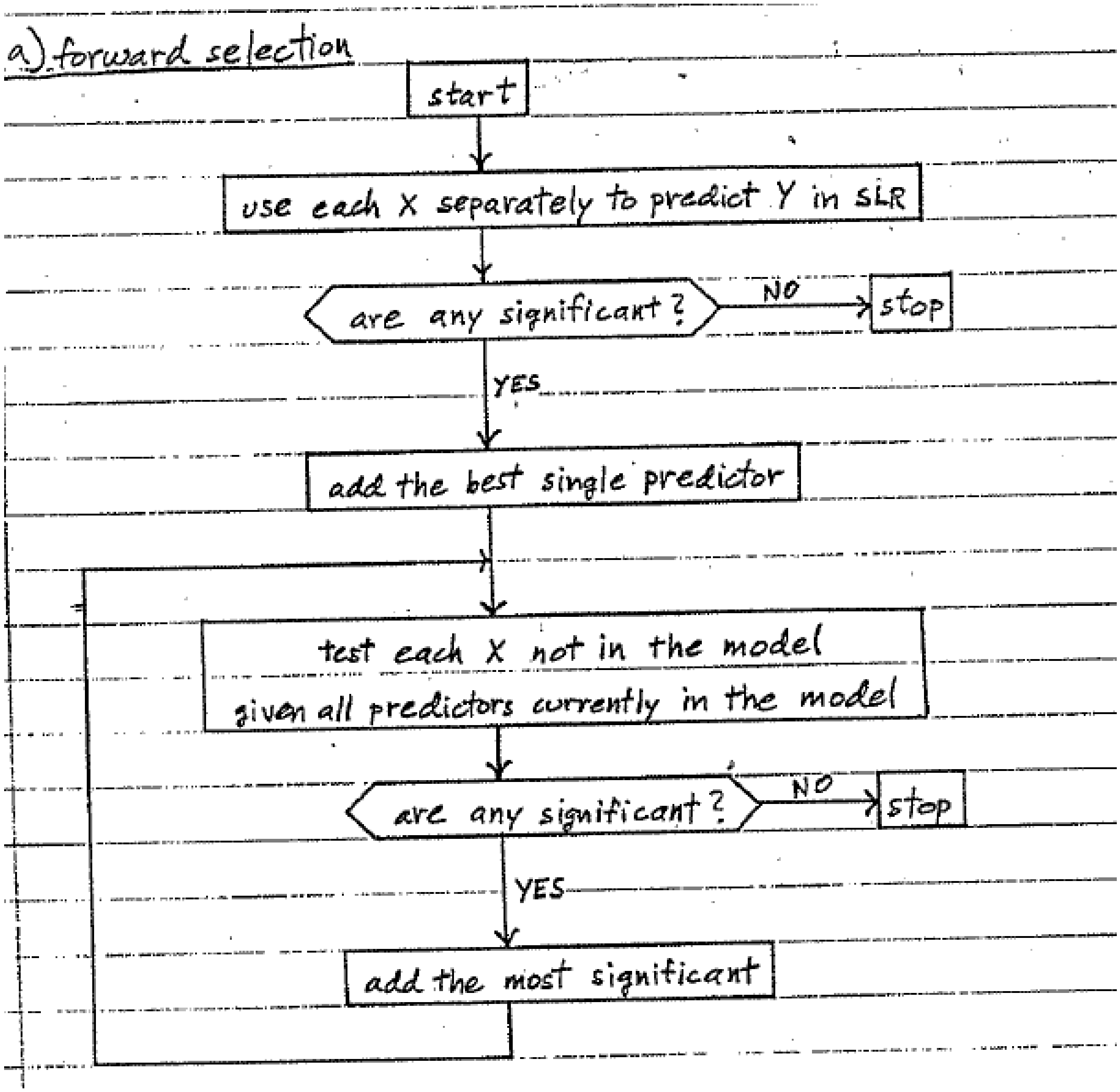
1. Proc PRINT;
2. Proc REG;
   - Model Y = X X2 X3 X4 / SS1;
3. Proc REG;
   - Model Y = X X2;
4. Proc REG;
   - MODEL MPG = SIZE WEIGHT SHAPE / P SS2;
     - RUN;

Result

1. Prints the data, the output contains the values of Y, X, X2, X3, X4 (ie: X2 = X^2)
2. To examine 4th order model (X, X^2, X^3, X^4), SS1 = Sequential SS
3. To examine 2th order model (X, X^2)
4. Builds a multiple linear regression model to predict Y (MPG) with X1 (Size), X2 (Weight) and X3 (Shape) features. Options: P = Requests predicted values ($\hat{Y}$) and residuals (e) for each observation. SS2: Provides the Type II SS = Partial SS for the model.

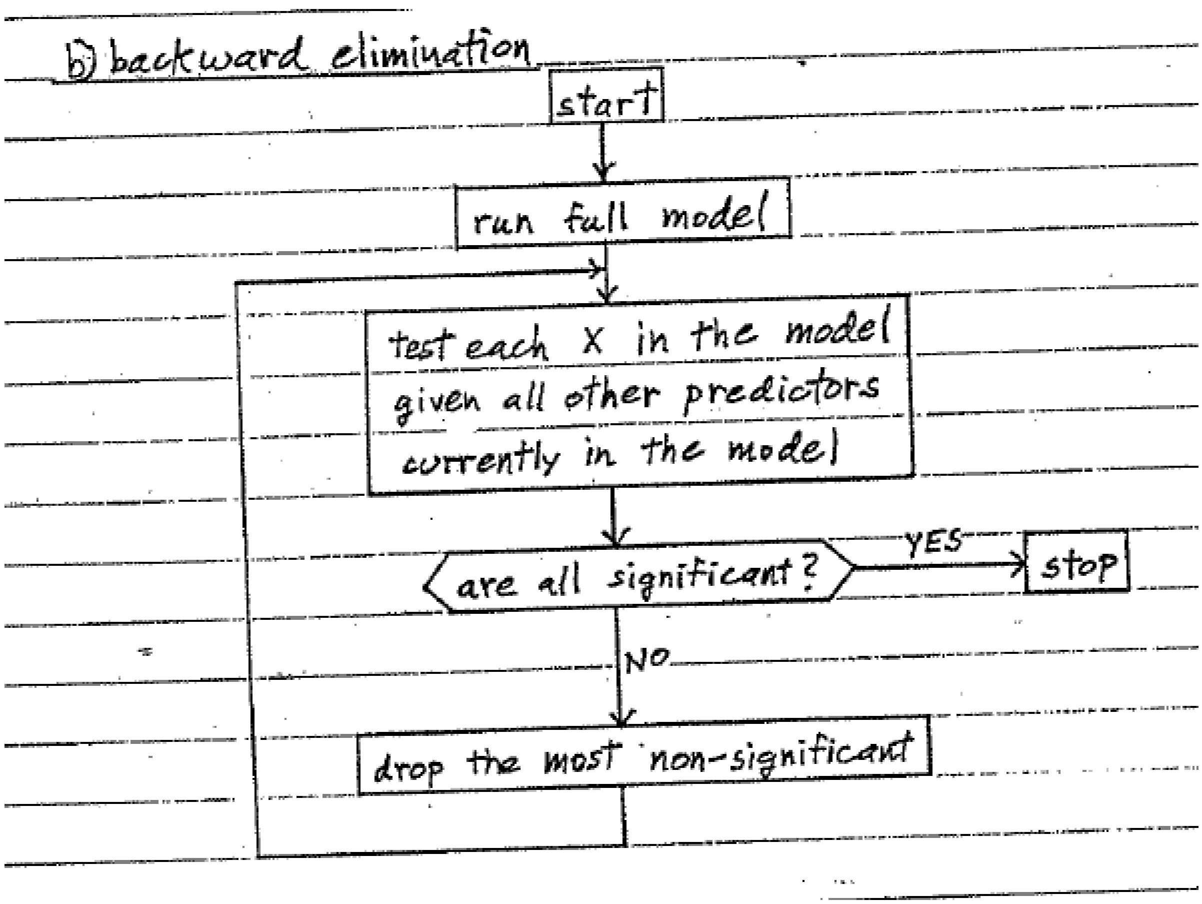
Draw the flow chart for the Forward Selection:

Result



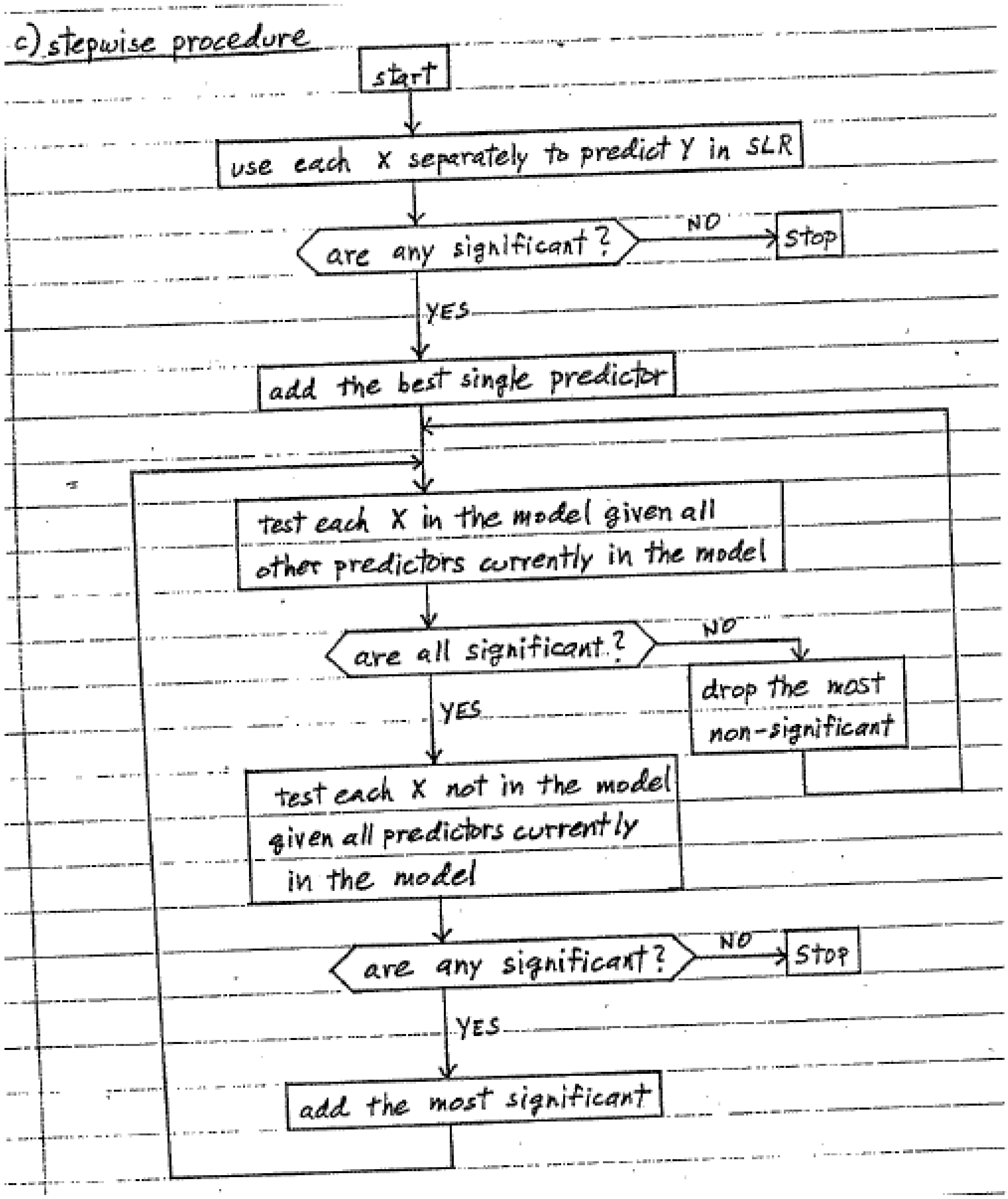
Draw the flow chart for the Backward Elimination:

Result



Draw the flow chart for the Stepwise Procedure:

Result



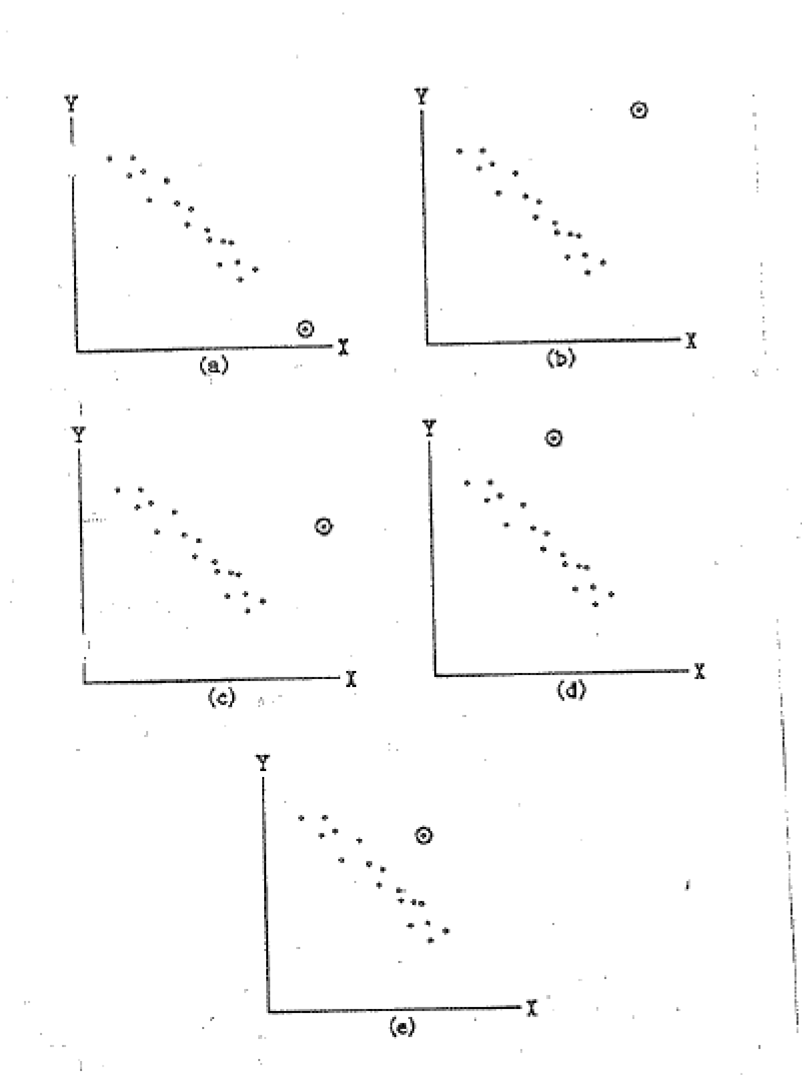
Explain the cross-validation procedure step by step.

Result

1. **Split Data:** Randomly divide data into training and validation sets.
2. **Fit Model:** Use the training set to select variables and fit the model.
3. **Predict:** Apply the model to the validation set to predict outcomes.

4. **Compare:** Check prediction accuracy by comparing predicted and actual values, often using sum of squared errors (SSE).
5. **Refit or Revise:** If the model performs well, refit using all data for stable estimates. If not, reduce predictors or revise the model to avoid **overfitting**.

Look at the Graps, and write the cause of outlier pattern or position (X or Y)



Result

a. Location X and Y

b. Location X and Y, and Pattern
c. Location X and Pattern
d. Location Y and Pattern
e. Only Pattern