<Regression Analysis Final Exam 2015> Hakbun: _____

(total points : 100) Name in Korean :_____

(괄호 안에 답만 적으시오. 과정생략. 괄호 밖의 내용은 인정하지 않음)

[1] The following output is the summary of p-values from the regressions of FuelCon on Pop, Area, or Gastax variables.

	pop	area	gastax
model fuelcon = pop area gastax ;	0.0296	0.0543	0.1353
model fuelcon = pop area ;	0.0380	0.8240	
model fuelcon = pop gastax;	0.0986		0.0560
model fuelcon = area gastax ;		0.7478	0.0727
model fuelcon = pop ;	0.0359		
model fuelcon = area ;		0.8087	
model fuelcon = gastax ;			0.0718

1-1 Using “forward selection with SLE=0.06”, write your selected variables : (5pts)

(Pop Gastax Area)

1-2 Using “backward elimination with SLS= 0.08”, write your selected variables : (5pts)

(Pop)

1-3 Using “Stepwise Regression with SLE= 0.06 and SLS = 0.08”, write your selected variables (5pts) :

(Gastax)

1-4 When you try the following SAS code, write your selected variables (5pts) :

```
proc reg ;
    model fuelcon = gastax pop area / Selection = F sle = 0.10 include =1 ;
run;
```

(Pop Gastax Area)

[2] The statistic, C(p), is called as ___(a)____ ’s Cp, and it is recommended that regressions with __(b)__ Cp values and those with values near __(c)___ where K is the number of independent variables in the regression. (10pts)

(a) = (Mallow)

(b) = (large , small)

(c) = (K+1)

[3] The following results are from “all possible regression” technique. Answer the questions. (20 pts)

	R^2	$Adj.\ R^2$	Cp	$\sqrt{MSE}$
Model 1: x1, x2	85.5%	84.2%	1.61	90.75
Model 2: x1, x2, x3	85.7%	83.6%	3.33	92.26
Model 3: x1, x2, x4	85.8%	83.8%	3.11	91.75

3-1 Based on R^2 , we could choose Model 3 as the best model. (True , False)

3-2 Based on $Adj.\ R^2$, Model 1 is the best. (True , False)

3-3 The chosen model based on $Adj.\ R^2$ is always same to that of $\sqrt{MSE}$

(True , False)

3-4 Considering the given information, the best model is

(Model1 , Model 2 , Model 3)

[4] The Furniture monthly sales (Y) are regressed on the time (Jan. 1992 - Dec. 2002: starting 1 through 130). “LagY" is the variable representing the previous month sales amount (Lagged dependent variable). JAN is the indicator variable having 1 for January and 0 for other months. We used 11 indicator variables for 12 months.

$$\hat{y} = 1552 + 8\,Time + 0.5\,LagY - 645\,JAN - 404\,FEB - 80\,MAR - 450\,APL - 180\,MAY - 300\,JUN - 280\,JUL - 180\,AUG - 350\,SEP - 240\,OCT + 10\,NOV$$

4-1 To test if there is seasonal variation (seasonal effect), we need to calculate F-ratio which has ___(a)___ and ___(b)___ degrees of freedom. (5pts)

(a) = (11)

(b) = (115)

4-2 According to the above regression line, the highest season for the furniture sales is November. (5pts) (True , False)

4-3 Suppose the sale amount on Dec. 2002 is 4500. Then, predict the sales amount on Jan. 2003. (5pts) (4205)

4-4 The reason to use Lagged dependent variable (LagY) in the regression equation is to eliminate the autocorrelation between dependent variable. (5pts) (True , False)

[5] Consider the regression model : $y_t = \beta_0 + \beta_1\, t + \beta_2\, t^2 + \epsilon_t$

y_t	10	20	25	30
t	1	2	3	4

Construct the matrix and vectors to make $\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$

5-1 where (5pts)

$$X = \left(\quad\right), \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} \qquad X = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 2 & 4 \\ 1 & 3 & 9 \\ 1 & 4 & 16 \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix}$$

5-2 With X and y , complete the following least square estimator formula : (5pts)

$$\hat{\boldsymbol{\beta}} = (\qquad\qquad)\boldsymbol{y} \qquad\qquad \hat{\boldsymbol{\beta}} = ((X^T X)^{-1} X^T)\boldsymbol{y}$$

[6] Fill the blanks:

6-1 One caution should be observed in using high-order polynomial regression models. Correlations between powers of a variable can result in m___(a)_____ problems. To reduce the possibility of computational difficulties, the use of explanatory variables that has been c____(b)____ is recommended. (10 pts)

(a) multicollinearity

(b) centered

6-2 Although R^2 has a nice interpretation, there is a drawback to its use in multiple regression. As more explanatory variables are added to the regression model, the value of R^2 will never d___(a)____ , even if the additional variables are explaining ___(b)___ proportion of the variation in y. (10 pts)

(a) = decrease

(b) = (an insignificant , a significant)