Part A: Single Choice (3 marks each, Total 3*10 = 30 marks)

Please provide your choices in the table below:

1	2	3	4	5	6	7	8	9	10
D	A	C	E ×	C ×	B	B	A	C ×	C
			C	D				B	

1. Which of the following *WHERE* statement selects from a dataset only those observations for which the value of the variable *Color* is *RED*, or *BLUE*, or *GREEN*?

   A. WHERE Color = 'RED' OR 'BLUE' OR 'GREEN';

   B. WHERE Color IN 'RED' OR 'BLUE' OR 'GREEN';

   C. WHERE Color IN (RED, BLUE, GREEN);

   D. WHERE Color IN ('RED', 'BLUE', 'GREEN');

2. The following program is submitted to SAS:

```
DATA MA409.X;
    SET SASHELP.BASEBALL;
RUN;
```

   What must be submitted prior to this SAS program for the program to be executed successfully?

   A. A *LIBNAME* statement for the libref *MA409* only must be submitted.

   B. A *LIBNAME* statement for the libref *SASHELP* only must be submitted.

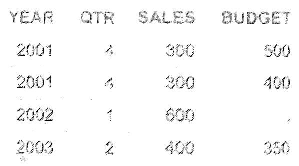
   C. *LIBNAME* statements for the librefs *MA409* and *SASHELP* must be submitted.

   D. No *LIBNAME* statement needs to be submitted.

3. Two SAS datasets named "ONE" and "TWO" are given below:

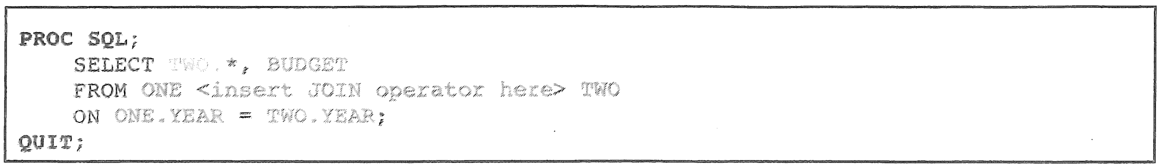
ONE			TWO		
YEAR	QTR	BUDGET	YEAR	QTR	SALES
2001	3	500	2001	4	300
2001	4	400	2002	1	600
2003	1	350	2003	2	400
2004	2	450			

   The following output is desired:



```
YEAR  QTR  SALES  BUDGET
2001    4    300     500
2001    4    300     400
2002    1    600       .
2003    2    400     350
```

Complete the following SAS program to generate the output above:



```
PROC SQL;
    SELECT TWO.*, BUDGET
    FROM ONE <insert JOIN operator here> TWO
    ON ONE.YEAR = TWO.YEAR;
QUIT;
```

A. INNER JOIN B. LEFT JOIN C. RIGHT JOIN D. FULL JOIN

4. Select the statement below that **incorrectly** interprets a 95% confidence interval (7.3, 9.2) for a population mean:

   A. A calculated 95% confidence interval may or may not contain the true population mean.

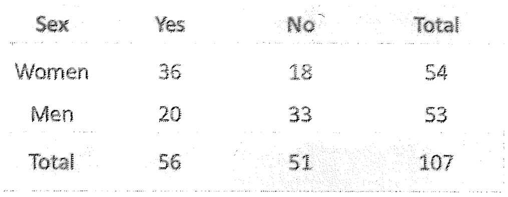
   B. The confidence interval would be narrower if a 90% confidence interval is calculated instead.

   C. 95% of the sample data lie within the interval (7.3, 9.2).

   D. Approximately 95% of the intervals calculated with this procedure will contain the true population mean.

   E. Both B and C.

5. A recent study compared the proportions of young women and men who use Instagram (a photo sharing APP). A total of 1069 young women and men were surveyed. The results are given in the table below:



Sex	Yes	No	Total
Women	36	18	54
Men	20	33	53
Total	56	51	107

   Let p_1 and p_2 denotes the proportions of young women and men who use Instagram, respectively. Which of the following statement is **correct**?

   A. The estimated standard error of $\hat{p}_1 - \hat{p}_2$ is 0.0966.

   B. Applying the two-sample z-test to test $H_0: p_1 = p_2$ vs. $H_1: p_1 > p_2$, the value of the test statistic is 2.9958.

   C. Applying the Pearson chi-square test to test whether using Instagram is independent of sex, the test statistic is 8.9746.

   D. Both B and C.

   E. All of the above.

6. Suppose that we performed 10 tests (e.g., test the association between 10 different risk factors and an outcome) and obtained the p-values 0.0141, 0.0029, 0.9535, 0.0031, 0.1053, 0.6415, 0.0071, 0.3017, 0.0053, 0.0452.

To control the family-wise error rate at 0.05, the Bonferroni and Holm adjustments are applied. How many hypotheses are rejected under the two adjustments?

A. 2 rejected under the Bonferroni adjustment and 3 rejected under the Holm adjustment.

B. 2 rejected under the Bonferroni adjustment and 4 rejected under the Holm adjustment.

C. 3 rejected under the Bonferroni adjustment and 4 rejected under the Holm adjustment.

D. 3 rejected under the Bonferroni adjustment and 5 rejected under the Holm adjustment.

7. Suppose that we fitted a linear model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$ to 20 observations and obtained $R^2 = 0.66$. To test $H_0: \beta_1 = \beta_2 = 0$ vs. $H_1: \beta_1 \neq 0$ or $\beta_2 \neq 0$, the F-test statistic is computed. What's the value of F-test statistic?

A. Less than 16

B. Greater or equal to 16, but less than 19.

C. Greater or equal to 19, but less than 22.

D. Greater or equal to 22.

8. When an explanatory variable is added to a linear regression model, which of the following values tend to increase or stay the same (i.e., can never decrease)?

A. R-square.

B. Adjusted R-square.

C. F-test statistic to test the overall significance of the model.

D. Both A and B.

E. Both A and C

9. Which of the following statement is **correct** about a fitted linear regression model $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon$?

A. The importance of explanatory variables can be evaluated by comparing their estimated regression coefficients, i.e., variables with larger $|\hat{\beta}_i|$ are more important.

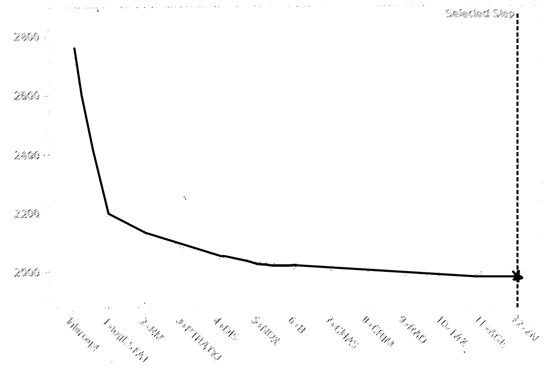
B. If X_i is standardized, i.e., $\tilde{X}_i = (X_i - \bar{X}_i)/s_{X_i}$ ($\bar{X}_i$ and s_{X_i} are the sample mean and standard deviation of X_i, respectively), the regression model is fitted again with $\tilde{X}_i$ substituting X_i, then only the estimates of β_0 and β_i will be different, the estimates of $\beta_1, \ldots, \beta_{i-1}, \beta_{i+1}, \ldots, \beta_p$ are the same.

C. If X_i is standardized, i.e., $\tilde{X}_i = (X_i - \bar{X}_i)/s_{X_i}$ ($\bar{X}_i$ and s_{X_i} are the sample mean and standard deviation of X_i, respectively), the regression model is fitted again with $\tilde{X}_i$ substituting X_i, then the estimates of all regression coefficients $\beta_0, \beta_1, \dots, \beta_p$ are different.

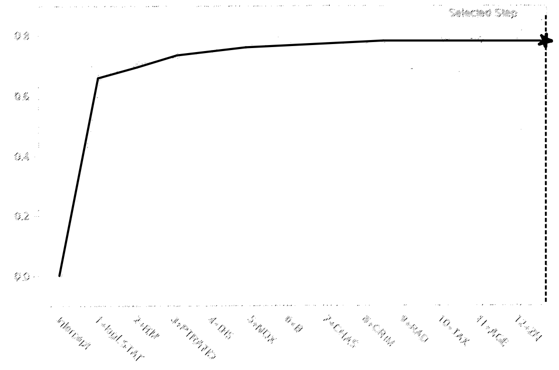
D. Both A and B.

E. Both A and C.

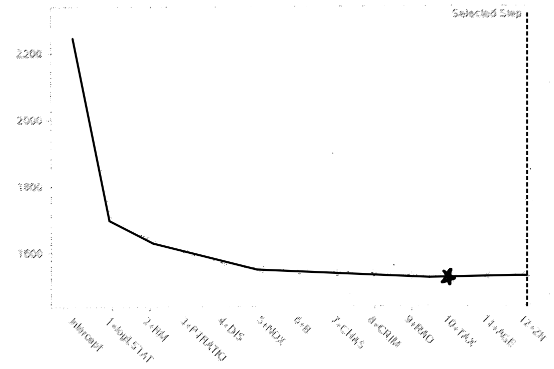
10. The following four plots are obtained using the stepwise selection method in a linear regression model to perform variable selection based on the adjusted R-square, Mallows's Cp, AIC, and BIC criteria. The step with a star indicates the best model under each criterion. Which of the plots corresponds to the BIC criteria?



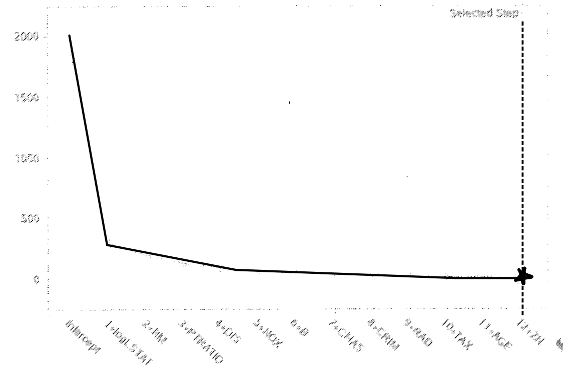
A.



B.



C.

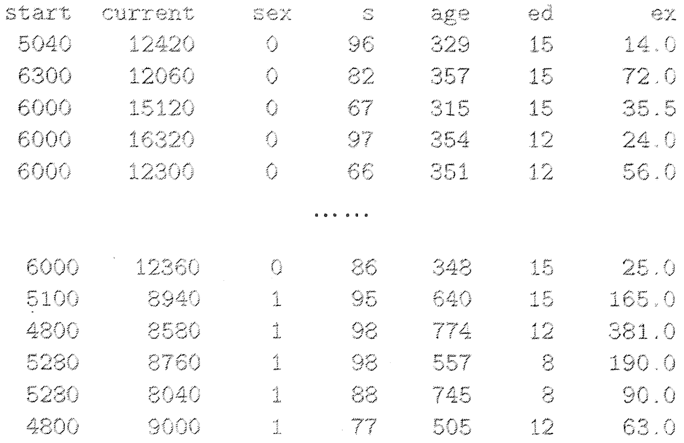


D.

Part B: Question Answering (Total 70 marks)

1. When we talked about Type I and Type II errors of hypothesis testing in classed, we said that these two types of error rates are traded off against each other. Can you think of any way to reduce the Type II error rate while not increasing the Type I error rate? Please provide a brief explanation. (5 marks)

2. Part of the data on employees from a specific job position in a bank is shown below. The data was used to study sex discrimination. There were 32 male employees and 61 female employees hired between 1998 and 2001. The variables recorded are: starting monthly salary (START), salary as of March 2001 (CURRENT), SEX (0 – male, 1 – female), seniority (资历) in months since first hired (S), age in months (AGE), education in years (ED), and experience prior to employment with the bank in months (EX).



```
start  current    sex      s    age    ed      ex
 5040    12420      0     96    329    15    14.0
 6300    12060      0     82    357    15    72.0
 6000    15120      0     67    315    15    35.5
 6000    16320      0     97    354    12    24.0
 6000    12300      0     66    351    12    56.0
                        ......

 6000    12360      0     86    348    15    25.0
 5100     8940      1     95    640    15   165.0
 4800     8580      1     98    774    12   381.0
 5280     8760      1     98    557     8   190.0
 5280     8040      1     88    745     8    90.0
 4800     9000      1     77    505    12    63.0
```

(1) Suppose that the dataset has been loaded into SAS and named "SexDis", write a SAS program to compute the mean and median starting salary (keep 3 decimal places) for male and female separately, ~~and test the normality assumption~~ of START for male and female separately. (5 marks)

(2) Let $\bar{X}_1$ and $\bar{X}_2$, S_1 and S_2 be the sample mean and standard deviation of the starting salary of male and female employees, respectively. Let μ_1 and μ_2, σ_1 and σ_2 be the population mean and standard deviation of the starting salary of male and female employees, respectively. Given $\bar{X}_1 = 5957$, $\bar{X}_2 = 5139$, $S_1 = 691$, $S_2 = 540$, test for equal variance: $H_0: \sigma_1^2 = \sigma_2^2 \ vs. H_1: \sigma_1^2 \neq \sigma_2^2$ at $\alpha = 0.05$. State the name of the test, the test statistic, the p-value and your conclusion clearly. (10 marks)

(3) Continued with (2), test whether there is a significant difference in mean starting salary between male and female. State the null and alternative hypotheses, state the name of the test, the test statistic, the p-value and your conclusion clearly. Use $\alpha = 0.05$. (10 marks)

(4) Another issue about sex discrimination is whether monthly salary increase tended to be higher for males than females. If a monthly raise of $100r\%$ is received in each of S successive months of employment (S is a variable in the dataset), then the current salary is

$$CURRENT = START \times (1 + r)^S.$$

Write a SAS program to compute $\log(1 + r)$ for each employee and test whether the mean of $\log(1 + r)$ are significantly different between male and female employees. (10 marks)

3. Suppose a linear regression model is fit, relating height (Y, in cm) to hand length (X_1, in cm) and foot length (X_2, in cm) for a sample of $n = 75$ adult females. The following results are obtained from a regression analysis of $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$:

ANOVA Table					
Source	DF	Sum of Squares	Mean Square	F value	P-value
Model		1105.52			
Error				-------	-------
Total		1793.85	-------	-------	-------

Coefficients Table				
Parameter	Estimate	Standard Error	t value	P-value
Intercept	74.41	7.97		
X_1	2.38	0.49		
X_2	1.73	0.37		

(1) Complete the tables. (10 marks)

(2) Compute the R-squared and adjusted R-squared of the model. (5 marks)

4. Suppose that $X_1, X_2, \cdots, X_n$ are i.i.d. samples from $N(\mu, \sigma^2)$ where the population mean μ is unknown while the population variance σ^2 is known. Now we would like to test $H_0: \mu = \mu_0$ vs. $H_1: \mu > \mu_0$ at significance level α.

(1) Write down the test statistic and determine its distribution under H_0. (5 marks)

(2) Suppose that the true underlying population mean is μ_1 ($\mu_1 > \mu_0$), compute the Type II error rate. (5 marks)

(3) Continued with (2), compute the minimum sample size needed to achieve power at least $1 - \beta$ of the test. (5 marks)