Data were collected on 1970 to 1982 model cars. The response variable was: y = miles per gallon of gas in city driving, and the independent variables were as follows:

$$x_1 = \text{number of cylinders}, \quad x_2 = \text{displacement}, \quad x_3 = \text{horsepower},$$
$$x_4 = \text{weight}, \quad x_5 = \text{acceleration}, \quad x_6 = \text{model year}$$

The number of cars in the study was $n = 392$. The accompanying output (available at the back of this exam as a 3-page document) shows information and plots obtained from regression models fit to these data in an "R" session.

Use the information and the attachment above to answer the following 8 questions (i.e. the current one and the next 7 questions).

1. Consider the output for the model containing variables $x_1, \ldots, x_6$. Which of the following is the best conclusion to draw from the model utility test?

(a) All six of the independent variables are useful.

(b) The variable x_3 should definitely be included in the model.

(c) At least one of the 6 independent variables is useful.

(d) None of the independent variables is useful.

(e) The result of the model utility test is not provided.

2. A plot of the residuals versus x_4 from the model containing $x_1, \ldots, x_6$ is provided in the last page of the "R" attachment. Which of the following is the best conclusion to draw from this plot?

(a) The residuals are not Normally distributed.

(b) It appears that the residuals may be related to x_4 in a non-linear way, but y should still be linearly related to $x_1, \ldots, x_6$.

(c) There are numerous outliers.

(d) It appears that the residuals may be related to x_4 in a non-linear way, and y could be non-linearly related to $x_1, \ldots, x_6$.

(e) There are many influential data values.

3. Let $x_7 = x_3^2$ and $x_8 = x_4^2$. Call the model with variables $x_1, \ldots, x_6$ as model M_1, and the model with variables $x_1, \ldots, x_8$ as model M_2. COnsider testing the following hypotheses:

$$H_0 : \text{Correct model is } M_1 \quad \text{vs.} \quad H_a : \text{Correct model is } M_2$$

Then, the value of the F-statistic for testing the above hypotheses is:

(a) Cannot be determined from the information given.

(b) 110.2.

(c) 272.2.

(d) 296.9.

(e) 71.5.

4. Using the model containing $x_1, \ldots, x_8$, what would be a prediction and a rough measure of the standard error of the prediction for the miles per gallon of gas for a car with the following independent variable values: $x_1 = 8$, $x_2 = 400$, $x_3 = 200$, $x_4 = 4000$ $x_5 = 20$, $x_6 = 80$? (**Hint:** For the model with $x_1, \ldots, x_8$, $\hat{\beta}_0 + \hat{\beta}_1(8) + \hat{\beta}_2(400) + \hat{\beta}_3(200) + \hat{\beta}_4(4000) + \hat{\beta}_5(20) + \hat{\beta}_6(80) = -32.4346$.)

(a) 18.1 ± 2.94.

(b) -32.4 ± 3.44.

(c) 25.6 ± 2.94.

(d) -32.4 ± 2.94.

(e) 24.9 ± 8.6.

5. It turns out that the correlation coefficient between number of cylinders (x_1) and displacement (x_2) is 0.95. In this case, which of the following is the best conclusion?

(a) Both x_1 and x_2 are highly correlated with miles per gallon.

(b) We should not include either x_1 or x_2 in the model.

(c) Both x_1 and x_2 should be included in the model.

(d) A model with just one of x_1 and x_2 would probably have an R^2 value almost as large as that of a model with both x_1 and x_2.

(e) Eight-cylinder cars are sick!

6. The estimate of the error standard deviation in the model containing $x_1, \ldots, x_6$ is:

(a) 2.939.

(b) 3.435.

(c) 11.80.

(d) 8.64.

(e) 67.40.

7. The percentage of variance in miles per gallon explained by the model containing $x_1, \ldots, x_8$ is closest to: (**Note:** this question carries **4 points**.)

(a) 65%.

(b) 86%.

(c) 81%.

(d) 12%.

(e) 93%.

8. The following information was determined by fitting various models in "R".

Model name	Variables used	R^2	BIC
M_1	$x_1, \ldots, x_6$	0.8093	2120.7
M_2	$x_1, \ldots, x_8$	0.8611	2008.2
M_3	$x_2, \ldots, x_6$	0.8088	2115.7
M_4	$x_2, \ldots, x_8$	0.8609	2003.0
M_5	$x_1, x_3, \ldots, x_6$	0.8087	2115.8
M_6	$x_1, x_3, \ldots, x_8$	0.8607	2003.4

Based on this information, which of the following conclusions seems best?

(a) Model M_4 is much better than any of the other models because it has the smallest BIC.

(b) Model M_1 is best because it has the largest BIC.

(c) Model M_2 is best because it has the largest R^2.

(d) Either of models M_4 and M_6 would be good choices because they have the smallest BIC values and their R^2 values are very close to each other and close to the largest.

(e) It is impossible to distinguish between the models based on the given information.

9. Consider the following two regression models:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon \quad (1)$$
$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \epsilon \quad (2)$$

A motivation for model (2) over model (1) would be when:

(a) The error terms are known to have non-constant variances.

(b) It is known that the surface of averages is a plane.

(c) One suspects that the surface of averages has curvature due to interaction effects.

(d) The error terms are known to have a non-Normal distribution.

(e) The semester is almost over and my brain is just not working right.

10. For $0 \leq x_1 \leq 1$ and $1 \leq x_2 \leq 4$, the following regression model holds:

$$Y = 10 + x_1 - 3x_2 + \epsilon,$$

where ϵ has a Normal distribution with mean 0 and variance 4. If $x_1 = 0.5$ and $x_2 = 2$, then the probability that Y exceeds 7 is given by:

(a) 0.1056.

(b) 0.2660.

(c) 0.7340.

(d) 0.8944.

(e) Cannot be determined from the information given.

11. In a regression analysis, a multiple linear regression model was fitted for a response variable Y using a set of 9 independent variables $x_1, \ldots x_9$. The dataset used had $n = 1030$ observations and the residual sum of squares turned out to be 51682.

Suppose one now wants to predict the response Y given a specific choice of values for $x_1, \ldots x_9$. Let L denote the total width of a 95% prediction interval for Y given these choices of predictor values. Then, which of the following is a plausible value of L?

(a) 7.12.

(b) 13.95.

(c) 14.24.

(d) 22.76.

(e) 30.21.

12. A producer of orange juice wants to compare three different methods (1, 2 and 3) of processing juice. The amount of vitamin C per 8 oz. serving is the variable of interest. Five servings are chosen at random from each process, and the amount of vitamin C for each of the fifteen servings was measured.

The following information and a partial ANOVA table were obtained from the data (the blank entries in the table indicate information not provided to you):

$$\bar{X}_1 = 90, \qquad \bar{X}_2 = 120, \qquad \bar{X}_3 = 93$$

ANOVA Table

Source of variation	Degrees of freedom	Sum of squares	Mean square	F
Methods				
Error	12	130		
Total		2860		

Use the information and the table above to answer the following 4 questions (i.e. the current one and the next 3 questions).

The F-statistic for testing that the average amount of vitamin C is the same for all three processes is given by:

(a) 130/2860.

(b) (130/12)/(2860/14).

(c) [(2860 − 130)/2]/[130/12].

(d) [(2860 − 130)/3]/[130/12].

(e) Cannot be determined from the information given.

13. Which of the following is correct if we wish to test the null hypothesis that the process means are the same using $\alpha = 0.05$?

(a) If the F-statistic is larger than 3.89, then we reject the hypothesis of equal means.

(b) If the F-statistic is smaller than 3.89, then we reject the hypothesis of equal means.

(c) If the F-statistic is larger than 3.49, then we reject the hypothesis of equal means.

(d) If the F-statistic is smaller than 3.49, then we reject the hypothesis of equal means.

(e) If the F-statistic is smaller than 3.89, then we conclude that the three means are equal.

14. An estimate of the variance of vitamin C content per 8 oz. serving for process 1 will be:

(a) $2860/14$.

(b) $\sqrt{2860/14}$.

(c) $\sqrt{130/12}$.

(d) $130/12$.

(e) Cannot be determined from the information given.

15. If Tukey's procedure (with $\alpha = 0.05$) is used to compare the means, then two means are significantly different when their difference is at least:

(a) 2.75.

(b) 3.33.

(c) 3.92.

(d) 4.71.

(e) 5.55.

16. Which of the following is the <u>most</u> correct interpretation of the Cook's D statistic? (Remember: there is **no** partial credit!)

(a) It can help identify outliers.

(b) It shows how much the estimates of the regression coefficients change when a data value is excluded from the data set.

(c) It indicates whether or not the error terms are Normally distributed.

(d) Both options (a) and (b) are true.

(e) All of options (a), (b) and (c) are true.

17. Suppose we want to test two different null hypotheses H_{01} and H_{02} against their respective alternatives, and we wish to do so under a multiple testing framework controlling for the experimentwise error rate (EWER). A suitable multiple testing procedure is developed which is guaranteed to control the EWER at a pre-specified level.

(a) No type I error but one experimentwise error.

(b) One type I error and one experimentwise error.

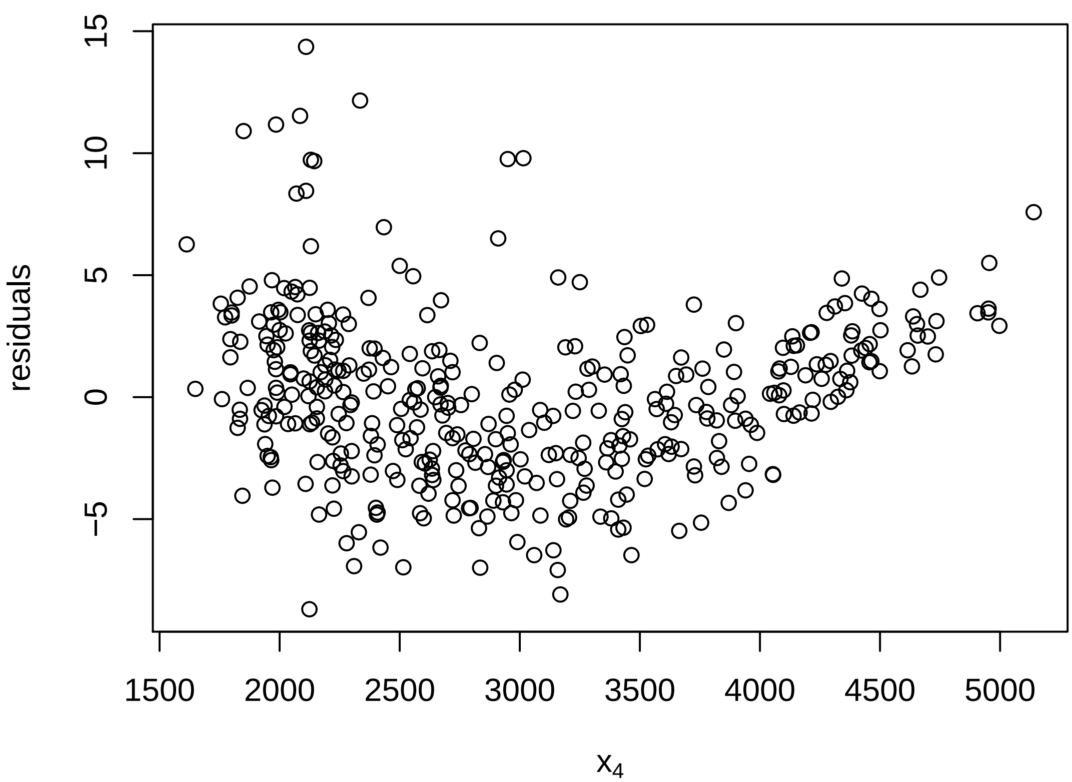
(c) One type II error and one experimentwise error.

(d) One type I error but no experimentwise error.

(e) One type I error and two experimentwise errors.

```
> fit=lm(y~x1+x2+x3+x4+x5+x6)
> summary(fit)

Call:
lm(formula = y ~ x1 + x2 + x3 + x4 + x5 + x6)

Residuals:
    Min      1Q  Median      3Q     Max 
-8.6927 -2.3864 -0.0801  2.0291 14.3607 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -1.454e+01  4.764e+00  -3.051  0.00244 ** 
x1          -3.299e-01  3.321e-01  -0.993  0.32122    
x2           7.678e-03  7.358e-03   1.044  0.29733    
x3          -3.914e-04  1.384e-02  -0.028  0.97745    
x4          -6.795e-03  6.700e-04 -10.141  < 2e-16 ***
x5           8.527e-02  1.020e-01   0.836  0.40383    
x6           7.534e-01  5.262e-02  14.318  < 2e-16 ***
---

Residual standard error: 3.435 on 385 degrees of freedom
Multiple R-squared:  0.8093,    Adjusted R-squared:  0.8063 
F-statistic: 272.2 on 6 and 385 DF,  p-value: < 2.2e-16

> anova(fit)
Analysis of Variance Table

Response: y
           Df  Sum Sq Mean Sq   F value    Pr(>F)    
x1          1 14403.1 14403.1 1220.5070 < 2.2e-16 ***
x2          1  1073.3  1073.3   90.9544 < 2.2e-16 ***
x3          1   403.4   403.4   34.1845  1.07e-08 ***
x4          1   975.7   975.7   82.6822 < 2.2e-16 ***
x5          1     1.0     1.0    0.0819    0.7749    
x6          1  2419.1  2419.1  204.9945 < 2.2e-16 ***
Residuals 385  4543.3    11.8                        
---


> fit1=lm(y~x1+x2+x3+x4+x5+x6+x7+x8)
> summary(fit1)

Call:
lm(formula = y ~ x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8)

Residuals:
    Min      1Q  Median      3Q     Max 
-8.4313 -1.6631 -0.0658  1.5147 12.6518 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  8.927e+00  4.527e+00   1.972   0.0493 *  
x1           2.562e-01  2.991e-01   0.857   0.3922    
x2          -7.373e-03  7.001e-03  -1.053   0.2930    
x3          -2.017e-01  4.031e-02  -5.003 8.60e-07 ***
x4          -1.467e-02  2.099e-03  -6.990 1.23e-11 ***
x5          -1.825e-01  1.016e-01  -1.796   0.0733 .  
x6           7.776e-01  4.562e-02  17.043  < 2e-16 ***
x7           6.231e-04  1.299e-04   4.797 2.31e-06 ***
x8           1.601e-06  2.793e-07   5.731 2.02e-08 ***
---

Residual standard error: 2.939 on 383 degrees of freedom
Multiple R-squared:  0.8611,    Adjusted R-squared:  0.8582 
F-statistic: 296.9 on 8 and 383 DF,  p-value: < 2.2e-16
```

```
> anova(fit1)
Analysis of Variance Table

Response: y
           Df  Sum Sq Mean Sq   F value    Pr(>F)    
x1          1 14403.1 14403.1 1667.7431 < 2.2e-16 ***
x2          1  1073.3  1073.3  124.2833 < 2.2e-16 ***
x3          1   403.4   403.4   46.7109 3.254e-11 ***
x4          1   975.7   975.7  112.9799 < 2.2e-16 ***
x5          1     1.0     1.0    0.1119    0.7382    
x6          1  2419.1  2419.1  280.1116 < 2.2e-16 ***
x7          1   952.0   952.0  110.2324 < 2.2e-16 ***
x8          1   283.7   283.7   32.8450 2.024e-08 ***
Residuals 383  3307.7     8.6                        
---
```


residuals
15
10
5
0
−5
1500
2000
2500
3000
3500
4000
4500
5000
x4