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.

The model utility test tests the null hypothesis $H_0 : \beta_1 = \cdots = \beta_6 = 0$. The F-statistic and P-value for this test are 272.2 and $(2.2)10^{-16}$, and therefore we should reject H_0 and conclude that at least one of the 6 variables is useful.

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.

There is definitely a systematic pattern in the residual plot which in turn indicates that the relationship between y and $x_1, \ldots, x_6$ could be non-linear, i.e. a linear model for y vs. $x_1, \ldots, x_6$ may not be appropriate. Note further that the relationship between the residuals and x_4 also seems non-linear. However, this does *not* contribute to the previous claim of a possible non-linear relationship between y and $x_1, \ldots, x_6$. For that, all we needed was *some* systematic pattern (could be linear/non-linear etc.) as opposed to a random scatterplot.

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.

You should use the reduction method of testing here. The SSEs for the full and reduced models are 3307.7 and 4543.3, respectively. Therefore, the F-statistic is:

$$F = \frac{(SSE_r - SSE_f)/(8-6)}{SSE_f/(n-8-1)} = \frac{(4543.3 - 3307.7)/2}{3307.7/383} = 71.5.$$

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.

Using the output for the model with 8 independent variables, the prediction would be:

$$\begin{aligned} \hat{y} &= -32.4346 + \hat{\beta}_7\, 200^2 + \hat{\beta}_8\, 4000^2 \\ &= -32.4346 + (6.231 \cdot 10^{-4})200^2 + (1.601 \cdot 10^{-6})4000^2 \\ &= 18.1. \end{aligned}$$

A 'rough' measure of the standard error of the prediction is just the estimated standard error of the residuals from the fitted model, i.e. $\hat{\sigma}$, which from the output is given by $\hat{\sigma} = 2.94$. The reason why this is a 'roughly' acceptable estimate of the error of the prediction lies in the formula of prediction intervals in pg. 95 of Chapter 1C (for multiple linear regression), and in the formulae on pg. 37 of Chapter 1A (for the special case of simple linear regression). These were discussed in great detail in the solution to Question 3 of the Practice Midterm II exam. Please see that solution as the same argument applies here and won't be repeated.

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!

When two variables are highly correlated, one is almost a linear function of the other, and so it is redundant to have both variables in the model. This relates to the multi-collinearity issue and we discussed this in great detail in the context of the Mesquite Tree Data in class.

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.

We are looking for the estimate $\hat{\sigma}$ which is given by $\sqrt{MSE}$. This estimate can be found in the summary "R" output as "residual standard error" for the given model.

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%.

The value of R^2 is given in the summary "R" output as "multiple R-squared" for the model.

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 <u>much</u> 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.

Models M_4 and M_6 have nearly identical (and minimum) values of BIC. Moreover, both have the *same* size (6 variables each) as well. Lastly, even their R^2 values are similar and very close to the highest one (achieved, of course, by the full model with all 8 variables). So there is nothing really that distinctly separates these two models and they are both equally good (and best, since any smaller model has a distinctly lower R^2 and much higher BIC). Hence, based on all our (many!) discussions on model selection, (d) is the only reasonable answer.

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.

See pg. 83-84 of Chapter 1C notes. We discussed the implications of regression models with interaction effects in details while going through these slides in class.

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.

We need to calculate $P(Y > 7 \mid x_1 = 0.5, x_2 = 2)$. Using the linear model above, this equals:

$$\begin{aligned} P\left(10 + (0.5) - 3(2) + \epsilon > 7\right) &= P(\epsilon > 2.5) \\ &= P(Z > 2.5/2) \quad \text{where } Z \sim N(0,1), \\ &= P(Z > 1.25) \\ &= 1 - 0.8944 \\ &= 0.1056. \end{aligned}$$

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.

Noting that $n = 1030$ and $k = 9$ here, the length of a 95% prediction interval is given by:

$$2t_{1030-9-1;0.025}\hat{\sigma}\sqrt{1 + \boldsymbol{x}^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}} \;=\; 2t_{1020;0.025}\hat{\sigma}\sqrt{1 + \boldsymbol{x}^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}} \;\geq\; 2t_{1020;0.025}\hat{\sigma},$$

where the last step is true since the second term under $\sqrt{(\cdot)}$ satisfies $\boldsymbol{x}^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x} \geq 0$. (In case you are not sure, recall that this quantity appears by itself in the variance of the estimate of the mean of Y given $\boldsymbol{x}$ as well. So it must be positive for it to be a variance!)

Further, we have $SSE = 51682$. Hence, we have: $\hat{\sigma} = \sqrt{MSE} = \sqrt{SSE/(1030 - 9 - 1)} = \sqrt{51682/1020} = \sqrt{50.669} = 7.12$ and $t_{1020;0.025} \geq 1.96$ (it is actually almost equal to 1.96 since the degree of freedom 1020 is very large here). Therefore the length of the prediction interval is *at least* $2(1.96)(7.12) = 27.91$, and so the *only* plausible value is option (e)!

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.

The F-statistic is $MSTr/MSE = [SSTr/(k-1)]/[SSE/(N-k)]$, with $SSTr = SST - SSE$. Now use the fact that you are given $SST = 2860$, $SSE = 130$, $k = 3$ and $N = 15$.

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.

See pg. 125 of Chapter 2A notes and use the fact that $F_{2,12;0.05} = 3.89$. Note that option (e) is *not* correct since you *never* 'conclude' from a test that the null is true! In fact, you always begin by presuming that it *is* true and you are performing the test to see if data provides sufficient evidence against it (i.e. to reject it). Otherwise you simply 'fail to reject' it.

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.

The variance within each group is σ^2, and we estimate this by $MSE = 130/12$.

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.

The least significant difference (LSD) for Tukey's method in the case of 1-way ANOVA is:

$$Q_{\alpha,k,N-k}\sqrt{\frac{MSE}{2}\left(\frac{1}{n_i}+\frac{1}{n_j}\right)} = Q_{0.05,3,12}\sqrt{\frac{MSE}{2}\left(\frac{1}{5}+\frac{1}{5}\right)} = 3.77\sqrt{\frac{130/12}{5}} = 5.55.$$

As dictated by the method of Tukey's HSD, you consider a pair of treatment means significantly different only when the absolute difference between the sample means in the respective treatment groups is at least as large as the LSD value above. Hence the correct choice is (e).

16. Which of the following is the most 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.

See pg. 110-111 of Chapter 1C notes. We discussed the role of Cook's D statistic in detail in class. (Note also that it does not help in any way in diagnosing the Normality assumption for the errors). Since this was straight out of the notes and a fair warning was also given, you will **not** get any partial credit for picking out only one of these answers, i.e. options (a) and (b). You get credit only if you mark the *most* correct and inclusive option which is (d).

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.

The procedure was then implemented on an observed dataset to test both the null hypotheses and the test results turned out to be: (i) reject H_{01} and (ii) fail to reject H_{02}.

Assume now that both the null hypotheses H_{01} and H_{02} were actually true. Then, what errors were committed while doing the tests from the data?

(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.

Given that both the null hypotheses H_{01} and H_{02} are actually true, by definition an experimentwise error occurs whenever *at least* one of the two hypotheses are rejected. Since in this case H_{01} was rejected, this implies an experimentwise error has certainly occurred.

(**Note:** *even if* both nulls were rejected, this would have continued to count as **one** experimentwise error! This is a consequence of how it is defined and the essence behind why it is a more strict measure of error. Also, note that the test might be designed to control for the *probability* of such errors (i.e. the EWER). But that does *not* mean it won't ever happen!)

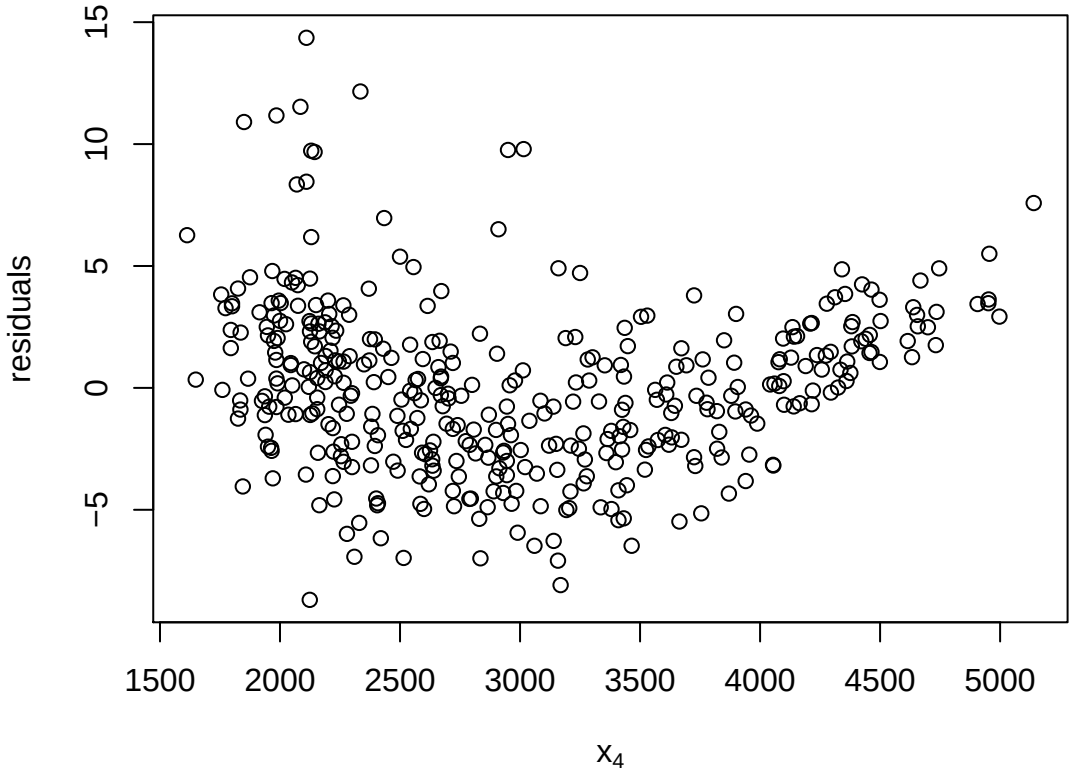
Moreover, since H_{01} is being rejected while it is actually true, by definition there has been a Type I error as well. But H_{02} is true and was not rejected, thus leading to no further errors. Hence, overall there has been one type I error and one experimentwise error committed.

18. Bonus question (6 points). To claim your bonus, **make sure to answer it!**

I hope you got this right :-)

```
> 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