Instructions: Attempt all questions. Short and specific answers are preferred. Given explanation when required, but keep it as short and simple as possible. Give only one answer to each question – if you give alternative answers, the worst answer will be graded.

Question A Question A refers to the following data: average butterfat content (percentages) of milk for random samples of twenty cows (ten two-year old and ten mature, i.e., greater than four years old) from each of five breeds (four cows in each breed). The average butterfat content was *repeatedly measured* an equal number of times on each cow. The data are from Canadian records of pure-bred dairy cattle. There are 3 variables in the dataset:

`Butterfat`: butter fat content by percentage

`Breed`: a factor with 5 levels, "Ayrshire," "Canadian," "Guernsey," "Holstein-Fresian," and "Jersey,"

`Age`: a factor with 2 levels, "2year" and "Mature"

A model of the form

```
Butterfat ~ Breed + Age + Breed:Age
```

was fit to the data and a *sequential ANOVA* table created:

Response: Butterfat

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Breed	**4**	**0.0034321**	**0.00085803**	**49.5651**	**<2e-16**
Age	**1**	**0.0000274**	**0.00002735**	**1.5801**	**0.2120**
Breed:Age	**4**	**0.0000514**	**0.00001285**	**0.7421**	**0.5658**
Residuals	**90**	**0.0015580**	**0.00001731**		

(1). (1 pt) How many replicates were collected for each *level combinations* of `Breed` and `Age` in this data?

(2). (2 pts) If the order of effects entering ANOVA model is changed to

```
Butterfat ~ Age + Breed + Breed:Age,
```

would the p-value of the test of `Breed` become larger, smaller, or staying the same? Explain.

The residuals against fitted-value plot is given below:

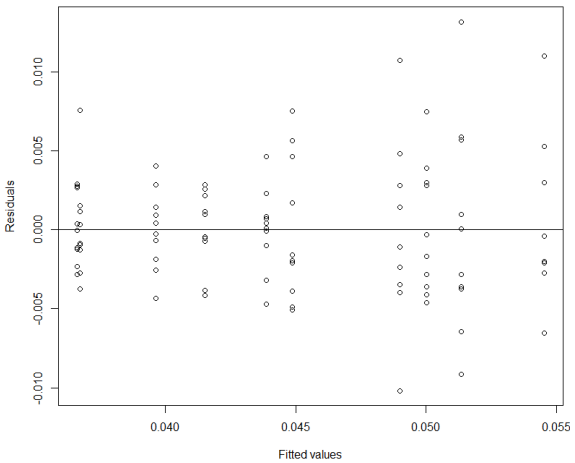

Figure 1

(3). (2 pts) What is the purpose of this plot? What does this particular plot indicate?

(4). (1 pt) In Figure 1, there are 10 groups of points, each corresponding to a distinct $\hat{y}$-value. What are the averages of the residuals in each groups? Explain how you get the answer.

(5). (2 pts) The ANOVA table suggests a simpler model `Butterfat ~ Breed`. Perform a lack-of-fit test for this simpler model and report the value of its test statistic. Explain how you get the answer.

The Box-Cox transformation method is applied on the response after removing the insignificant effects in the above ANOVA output. A plot of the log-likelihood for the transformation parameter is shown below:

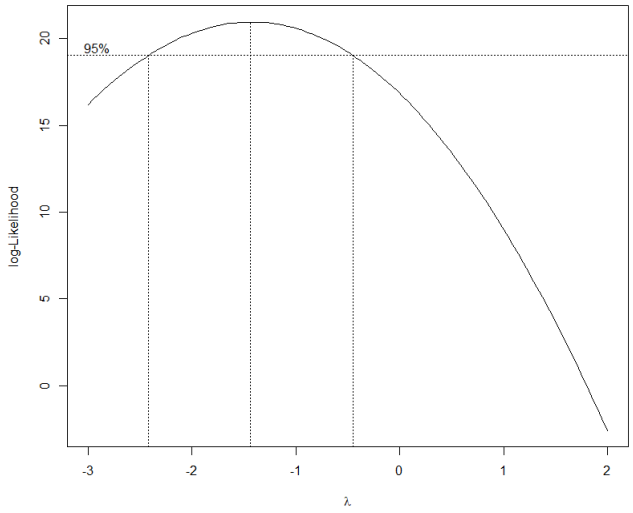

Figure 2

(6). (2 pts) What transformation of the response is suggested by the Box-Cox method and yet retain some interpretability? Is it really necessary to transform the response? Explain your answers.

A model is refitted with the transformation determined using the Box-Cox method shown above. The *regression output* of this fit is given below:

```
Coefficients:
                    Estimate  Std. Error   t value   Pr(>|t|)
(Intercept)         24.7298   0.4217       58.648    < 2e-16
BreedCanadian       -2.0592   0.5963       -3.453    0.000829
BreedGuernsey       -4.3417   0.5963       -7.281    9.56e-11
BreedHolstein-Fresian 2.6425  0.5963       4.431     2.51e-05
BreedJersey         -5.6092   0.5963       -9.406    3.07e-15

Residual standard error: 1.886 on 95 degrees of freedom
Multiple R-squared:  ??,  Adjusted R-squared:  ??
F-statistic: 61.85 on 4 and 95 DF,  p-value: < 2.2e-16
```

(7). (1 pt) What is the value of the *adjusted* R-squared of this fit? Explain.

(8). (1 pt) Interpret the meaning of the estimated coefficient `-2.0592`.

(9). (1 pt) Use the regression output to predict the *mean butterfat content of milk* for Jersey cows.

(10). (2 pts) Does this data show a statistically significant difference between Canadian and Guernsey cows in the means of the transformed response? Explain.

(11). (2 pts) A similar follow-up experiment is planned for a different population of cows. Would a new sample of 4 replicates for each level combinations be adequate to detect a difference *of similar size* between Ayrshire and Guernsey cows within this new population? Explain.

(12). (1 pt) No correlation is a common assumption on the error distribution of a linear model. What feature of this data would lead us to question this assumption? Explain.

The output from a Huber-M estimate of the regression model is given below:

```
Coefficients:
                    Value     Std. Error   t value
(Intercept)         24.6787   0.4144       59.5501
BreedCanadian       -1.9501   0.5861       -3.3273
BreedGuernsey       -4.3716   0.5861       -7.4592
BreedHolstein-Fresian 2.7974  0.5861       4.7731
BreedJersey         -5.5742   0.5861       -9.5111

Residual standard error: 1.954 on 95 degrees of freedom
```

(13). (1.5 pts) Why are overall F-statistic and R^2 not computed in the output of Huber-M robust regression?

Question B

Question B refers to the following data: natural gas usage in a house. The weekly gas consumption (in 1000 cubic feet) and the average outside temperature (in degrees Celsius) was recorded for 26 weeks before and 18 weeks after cavity-wall insulation had been installed. The house thermostat was set at 20C throughout. There are 3 variables in this dataset:

`Insulate`: a factor with 2 levels, After and Before, coded by 1 and 0 respectively

`Temp`: outside temperature

`Gas`: weekly consumption in 1000 cubic feet

A scatter plot of `Gas` against `Temp` is given below, in which Before points are marked by △ and After points by ○:

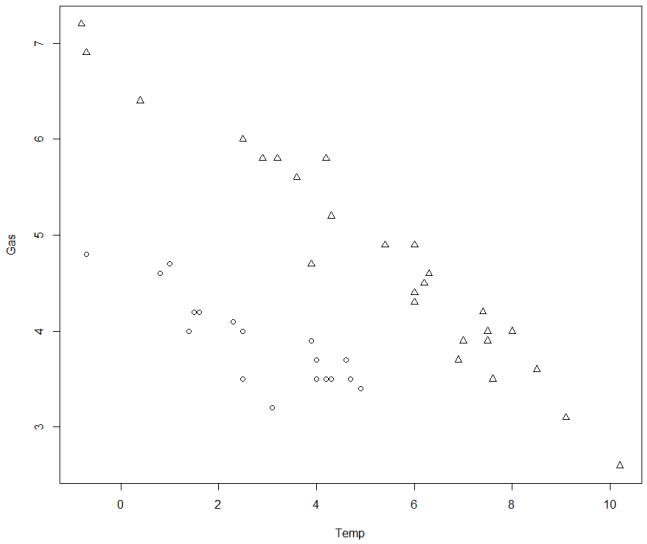

Figure 3

A regression model `Gas ~ Insulate + Temp + Insulate:Temp` is fitted to this data and output obtained is given below:

```
Coefficients:
             Estimate  Std. Error   t value   Pr(>|t|)
(Intercept)  6.85383   0.11362      60.320    < 2e-16
Insulate     -2.26321  0.17278      -13.099   4.71e-16
Temp         -0.39324  0.01879      -20.925   < 2e-16
Insulate:Temp 0.14361  0.04455      3.224     0.00252

Residual standard error: 0.2699 on 40 degrees of freedom
Multiple R-squared:  ??,  Adjusted R-squared:  ??
F-statistic: 194.8 on 3 and 40 DF,  p-value: < 2.2e-16
```

(14). (1 pt) In addition to the usual diagnostics, what specific assumption about the errors needs to be checked for this particular kind of data? Explain.

(15). (1 pt) What does the significance of the interaction term `Insulate:Temp` mean?

(16). (1 pt) What does the value `-2.26321` say about the difference in mean weekly gas consumption before and after the installation of insulation?

(17). (1 pt) It can be clearly observed from Figure 3 that the installation of insulation can reduce gas consumption. Suppose that the predicted average outside temperature for next week is 4°C. How much gas consumption do you expect *to save* in next week due to the installation of insulation?

(18). (2 pts) A half-normal plot for the Cook's distances are shown below in Figure 4. What is the purpose of this plot? Since the points follow a curve (i.e., not like a straight line), the Cook's distances do not follow a half-normal distribution. What is the practical consequence of this fact?

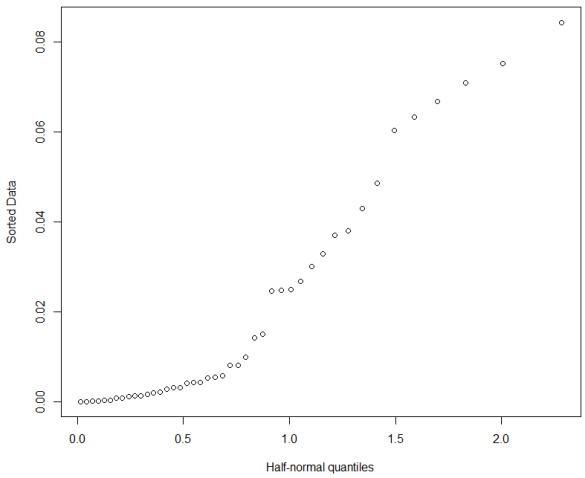

Figure 4

(19). (2 pts) If we fit the model `Gas ~ Insulate` (i.e., removing `Temp` and `Insulate:Temp` from the above model) to this data to estimate mean gas usage difference between the two groups, would you expect the new model to over-estimate, under-estimate, or correctly estimate the amount of gas saved by the installation of insulation? Explain.

(20). (1 pt) The side-by-side boxplot of residuals against `Insulate` is given below in Figure 5. It shows a sign of non-constant variance. Suggest a formal test (i.e., a numerical method, not a graphical one) to examine whether the error variances before and after the installation of insulation can be regarded as equal.

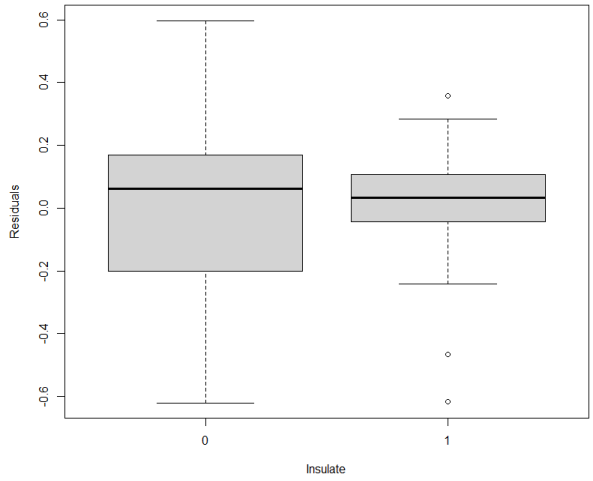

Figure 5

(21). (1 pt) Suppose that some data were missing and it was observed that the missing proportion in the Before group is much higher than that in the After group. Based on this information, what missingness mechanism would you infer it is, missing completely at random (MCAR), missing at random (MAR), or missing not at random (MNAR)? What assumption on the missing probability is required to support your answer?

Question C

Question C refers to the following data. Investigators studied physical characteristics and ability in 13 (American) football punters. Each volunteer punted a football ten times. The investigators recorded the average distance for the ten punts, in feet. There are 6 variables in the data:

`Distance`: average distance over 10 punts

`RStr`: right leg strength in pounds

`LStr`: left leg strength in pounds

`RFlex`: right hamstring muscle flexibility in degrees

`LFlex`: left hamstring muscle flexibility in degrees

`OStr`: overall leg strength in foot pounds

A linear model with `Distance` as the response and the other variables as the predictors is fitted. The regression output is given below:

```
Coefficients:
             Estimate  Std. Error   t value   Pr(>|t|)
(Intercept) -29.58047 65.70042     -0.450    0.666
RStr        0.27877   0.45638      0.611     0.561
LStr        0.06971   0.48388      0.144     0.890
RFlex       1.24146   1.44927      0.857     0.420
LFlex       -0.39535  0.74472      -0.531    0.612
OStr        0.22369   0.13053      1.714     0.130
```

```
Residual standard error: 14.65 on 7 degrees of freedom
Multiple R-squared:  0.8144
```

(22). (1 pt) In this regression output, the R^2 value is high (81.47%), but none of the predictors are significant. What could be the possible problem in these predictors that causes it?

(23). (1 pt) Because the responses are averages of several records, would weighted least square be a better choice for this data than ordinary least square in estimating the coefficients? Explain.

A *sequential* ANOVA output is given below:

```
Response: Distance
            Df  Sum Sq    Mean Sq   F value   Pr(>F)
RStr         1  5069.8    5069.8    23.6226   0.001835
OStr         1  1279.6    1279.6    5.9620    0.044646
RFlex        1  172.5     172.5     0.8037    0.399765
LFlex        1  64.7      64.7      0.3014    0.600097
LStr         1  4.5       4.5       0.0208    0.889513
Residuals    7  1502.3    214.6
```

(24). (2 pts) Can we use this ANOVA output to determine what would be the final fitted model after performing a testing-based backward model selection on this data? Explain. [Hint. Pay attention to the p-values in the ANOVA table.]

The result of eigen-decomposing the *correlation matrix* of the 5 predictors is given below:

```
Eigenvalues:
[1]  ??????  0.7599  0.5217  0.1183  0.0797

Eigenvectors:
      PC1       PC2       PC3       PC4       PC5
RStr  -0.4736   0.4268    -0.1599   0.5916    -0.4669
LStr  -0.4763   0.3342    -0.3769   -0.1716   0.7000
RFlex -0.5077   -0.1394   -0.0551   -0.6996   -0.4800
LFlex -0.3521   -0.8287   -0.2215   0.3534    0.1234
OStr  -0.4089   0.0033    0.8833    0.0789    0.2152
```

(25). (1 pt) Can you give the first principal component (PC1) an interpretation? Explain.

(26). (1.5 pts) What is the condition number in the collinearity diagnostic of the standardized predictors?

(27). (1.5 pts) If we perform a principal component regression using the model:

```
Distance ~ PC1 + PC2 + PC3 + PC4 + PC5,
```

and construct a confidence region for the coefficients of PC1 and PC2. What is the geometric shape of this confidence region? Explain.

A ridge regression is applied on the model with Distance as the response and the other variables in the data as the predictors, and the ridge trace for various λ is shown in the plot given below:

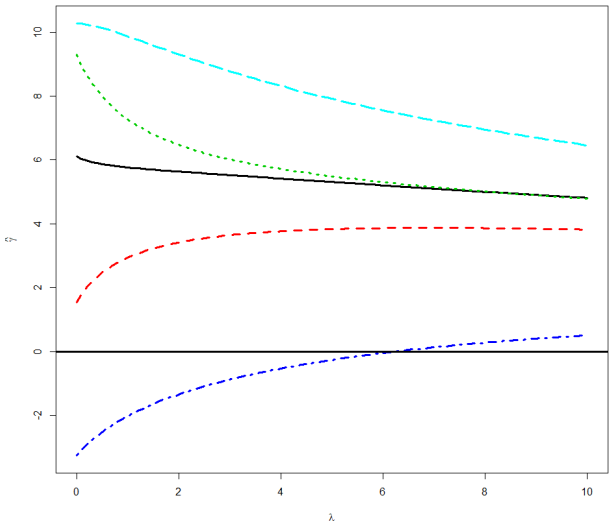

Figure 6

(28). (1.5 pt) What is the predictor whose ridge estimates converge to zero the fastest in Figure 6? Explain.

(29). (1 pt) Give at least one advantage and one disadvantage of ridge estimates over the ordinary least square estimates.