1

20 points

We have following variables:

- Name: University name
- SAT: average SAT score of new freshmen
- Top10: percentage of new freshmen in top 10
- Accept: percentage of applicants accepted
- Sfratio: Student-faculty ratio
- Expenses: Estimated annual expenses (divided by 1000)
- Grad: Graduation rate

Read **univ.txt**. Exclude Name from the analysis.

(a) Interpret a scatter diagram matrix.

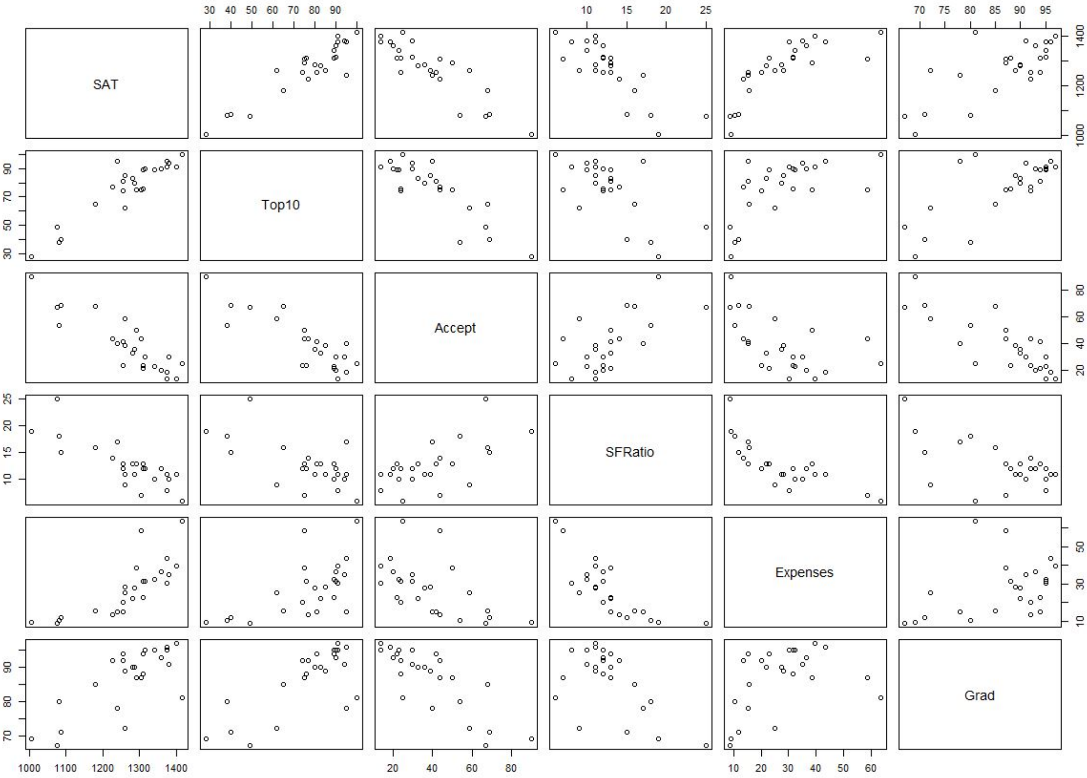
(b) Run multiple regression (response: SAT, predictors: Top10, Accept, Sfratio, Expenses, Grad), Report R^2, MSE, and interpret the overall F test and individual t tests. **(+20 points)**

(c) Conduct residual analysis (Standardized residuals vs. Fitted, Histogram and Q-Q plot of residuals) with the model and assess the assumptions.

Solution. (a) Using **R**,

```
> univ_data = univ[,c(2,3,4,5,6,7)]
> pairs(univ_data)
```

while the result is

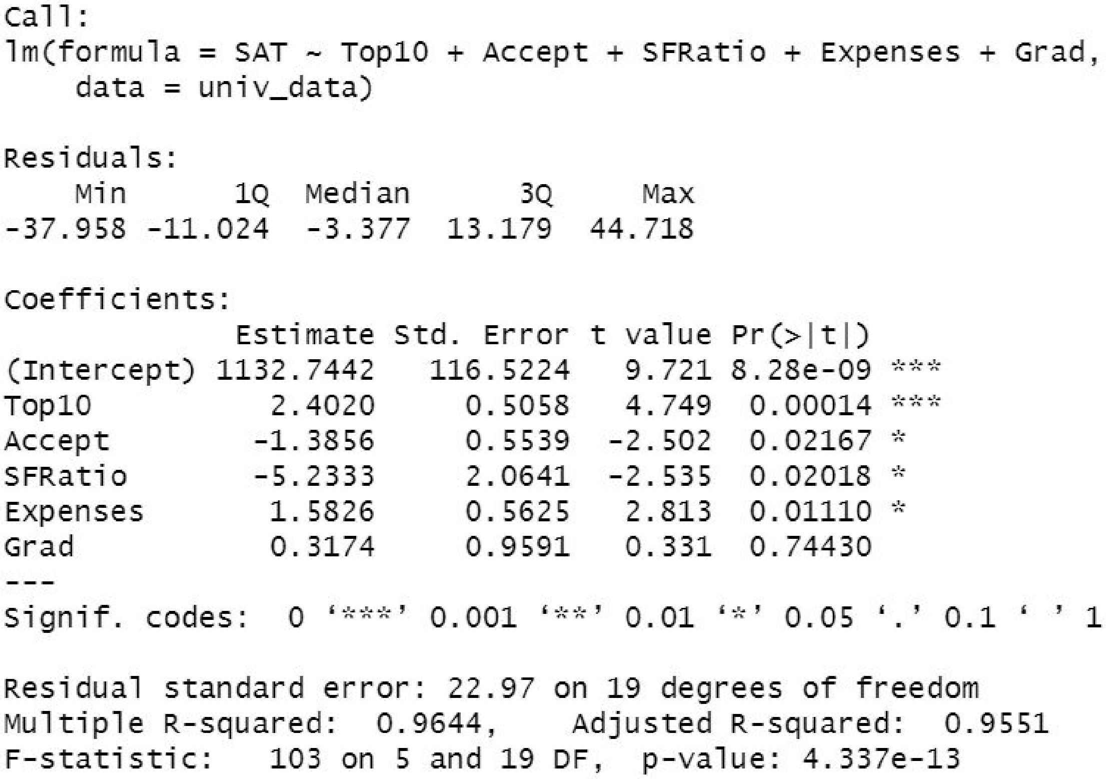


which means that the pair (SAT,Top10) has positive correlation and the pairs (SAT, Accept), (Top10, Accept) have negative correlation.

(b) Using **R**,

```
> univ_regression = lm(SAT~Top10+Accept+SFRatio+Expenses+Grad,data=univ_data)
> summary(univ_regression)
```

while the result is



```
Call:
lm(formula = SAT ~ Top10 + Accept + SFRatio + Expenses + Grad,
    data = univ_data)

Residuals:
    Min      1Q  Median      3Q     Max 
-37.958 -11.024  -3.377  13.179  44.718 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 1132.7442   116.5224   9.721 8.28e-09 ***
Top10          2.4020     0.5058   4.749  0.00014 ***
Accept        -1.3856     0.5539  -2.502  0.02167 *  
SFRatio       -5.2333     2.0641  -2.535  0.02018 *  
Expenses       1.5826     0.5625   2.813  0.01110 *  
Grad           0.3174     0.9591   0.331  0.74430    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 22.97 on 19 degrees of freedom
Multiple R-squared:  0.9644,	Adjusted R-squared:  0.9551 
F-statistic:   103 on 5 and 19 DF,  p-value: 4.337e-13
```

which implies that R^2 is 0.9644 **(+5 points)** and MSE is 527.5726 **(+5 points)** by

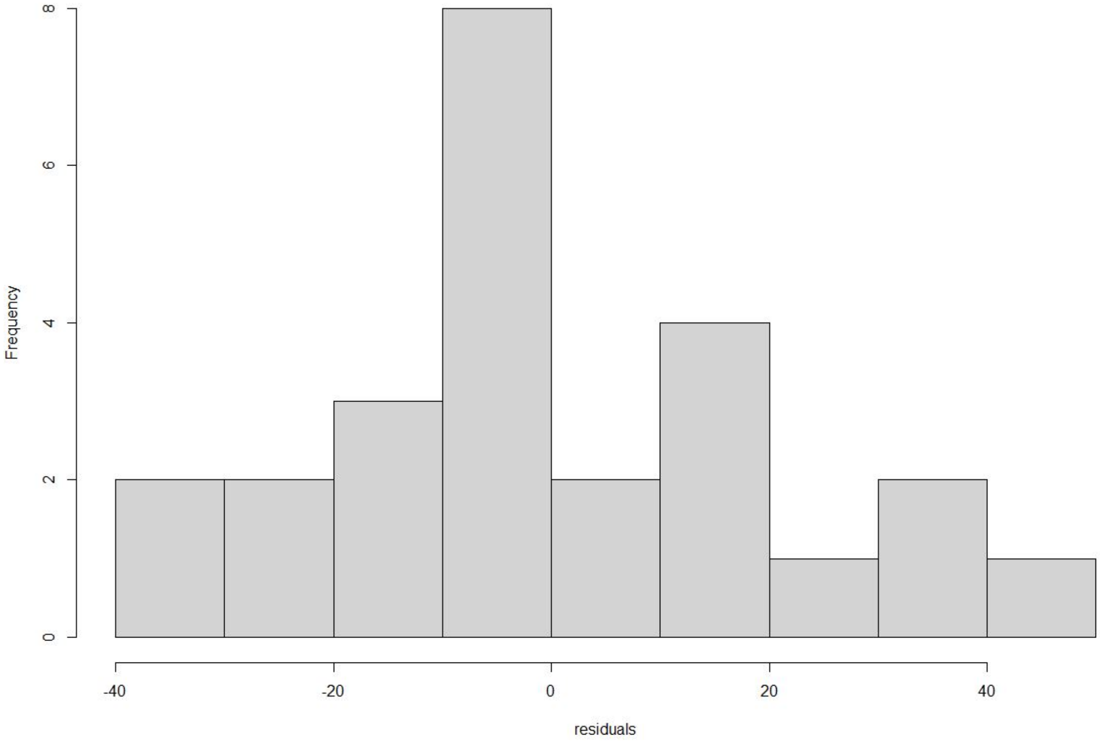
```
> sum(univ_regression$residuals^2)/univ_regression$df.residual
```

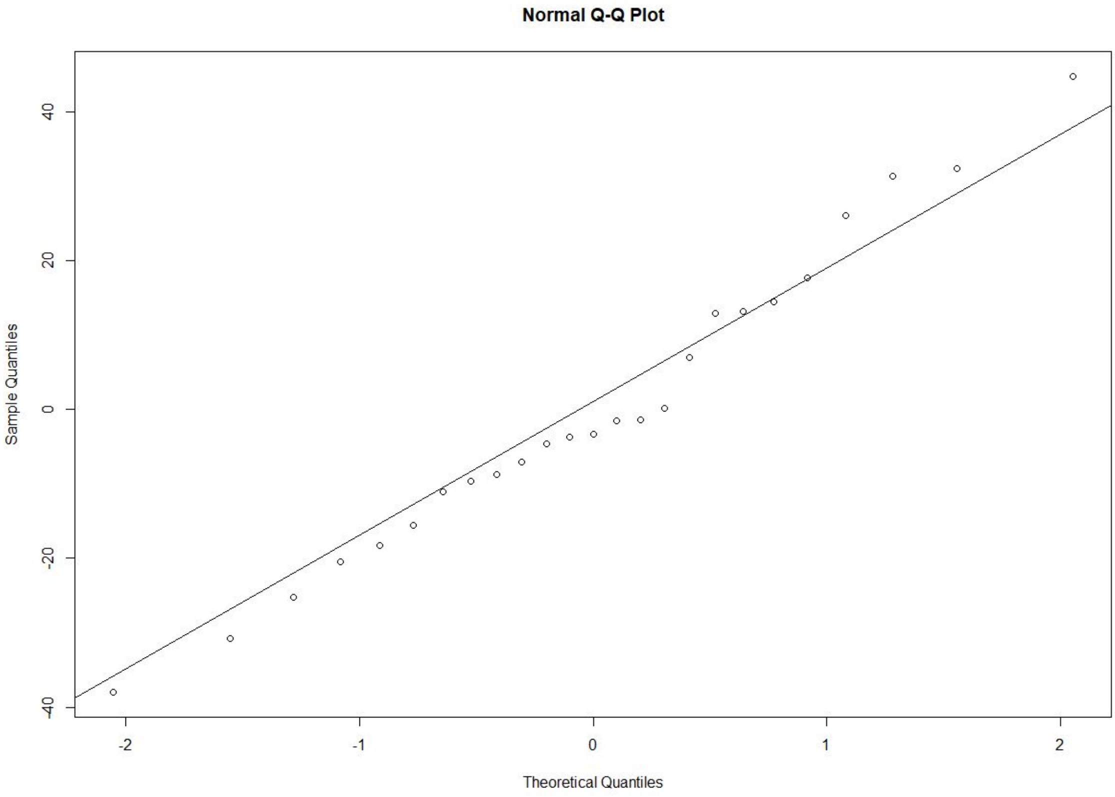
The p-value of overall F test is $4.337 \cdot 10^{-13}$ which is smaller than $\alpha = 0.05$. Therefore, we have to reject our hypothesis so that there is strong evidence to say that at least one of predictors has correlation with SAT. Among the p-values of individual t tests, the p-values except for Grad are smaller than $\alpha = 0.05$. Therefore, there is strong evidence to say Top10, Accept, SFRatio, and Expenses are correlated to SAT **(+10 points)**.

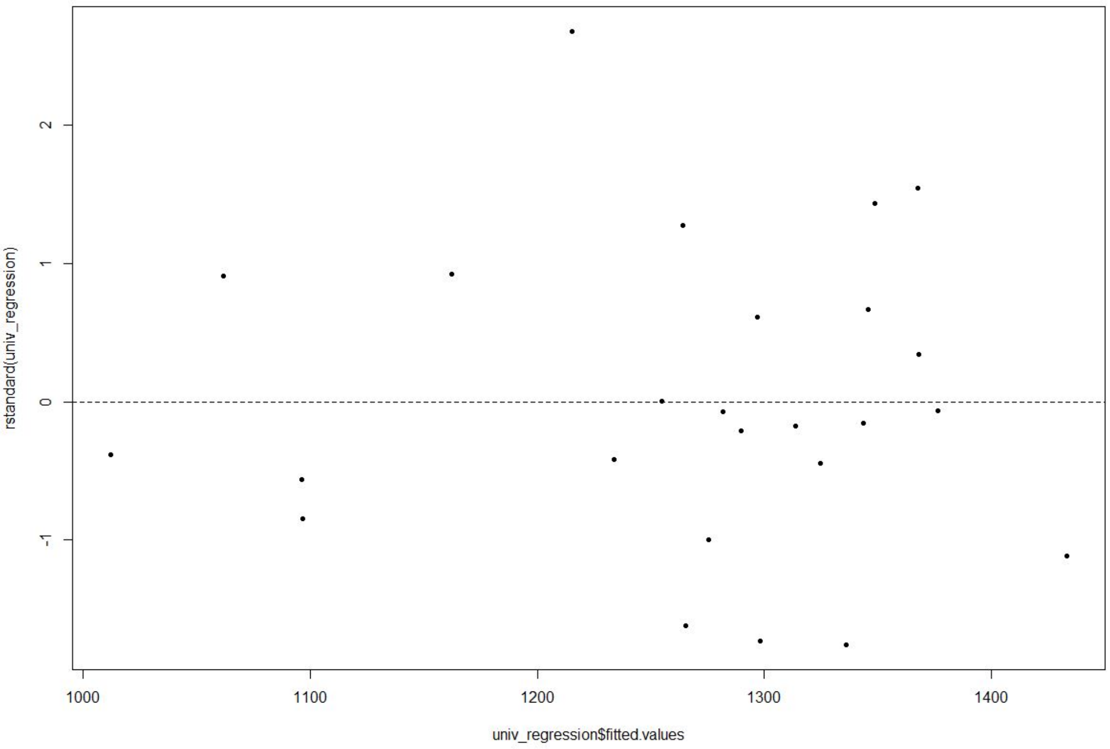
(c) For graph, use following codes

```
> hist(univ_regression$residuals,xlab=’residuals’)
> qqnorm(univ_regression$residuals)
> qqline(univ_regression$residuals)
> plot(univ_regression$fitted.values, rstandard(univ_regression), type=’p’, pch=20)
> abline(h=0,lty=2)
```

which lead to




Normal Q-Q Plot
Sample Quantiles
Theoretical Quantiles
40
20
0
-20
-40
-2
-1
0
1
2


rstandard(univ_regression)
univ_regression$fitted.values
2
1
0
-1
1000
1100
1200
1300
1400

2

30 points

Four chemical plants, producing the same products and owned by the same company, discharge effluents into streams in the vicinity of their locations. To monitor the extent of pollution created by the effluent and to determine whether this differs from plant to plant, the company collected random samples of liquid waste, five specimens from each plant.

Group	Sample Size	Sample Mean	Sample standard deviation
A	5	1.568	0.1366
B	5	1.772	0.2160
C	5	1.546	0.1592
D	5	1.916	0.1689

(a) Write the statistical model and identify factor and response. State the null hypothesis. **(+10 points)**

(b) Complete the following ANOVA table. Specify the null hypothesis of the test and conduct an analysis of variance at $\alpha = 0.05$. **(+10 points)**

Source	D.F.	Sum of Squares	Mean of Squares	F	p-value
Model					0.0107
Error					
Total		0.94			

(c) Interpret the multiple testing results (at the overall significance level $\alpha = 0.05$) by Tukey. **(+10 points)**

```
Tukey multiple comparisons of means
95% family-wise confidence level

Fit: aov(formula = effluent ~ plant, data = anovaprob)

$plant
diff          lwr         upr      p adj
B-A  0.204 -0.10836258 0.51636258 0.2796961
C-A -0.022 -0.33436258 0.29036258 0.9969809
D-A  0.348  0.03563742 0.66036258 0.0264399
C-B -0.226 -0.53836258 0.08636258 0.2047920
D-B  0.144 -0.16836258 0.45636258 0.5647642
D-C  0.370  0.05763742 0.68236258 0.0176858
```

Solution. (a) The statistical model of the problem is

$$X_{ij} = \mu_i + e_{ij}, \ \ i = 1, 2, 3, 4, \ \ j = 1, 2, 3, 4, 5 \quad \textbf{(+5 points)}$$

where X_{ij} is the j-th observation from the i-th group, μ_i is the sample mean of i-th group, and e_{ij} is the residual of the j-th observation from the i-th group. Also the null hypothesis H_0 is

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 \quad \textbf{(+5 points)}$$

(b) SS_B and SS_W can be calculated as follow:

$$SS_B = 5\{(1.568 - 1.7)^2 + (1.772 - 1.7)^2 + (1.546 - 1.7)^2 + (1.916 - 1.7)^2\} = 0.46$$
$$SS_W = 4(0.1366^2 + 0.216^2 + 0.1592^2 + 0.1689^2) = 0.48$$

Therefore, the ANOVA table will be

Source	D.F.	Sum of Squares	Mean of Squares	F	p-value
Model	3	0.46	0.155	5.2	0.0107
Error	16	0.48	0.03		
Total	19	0.94			

(+5 points)

Since the p-value is smaller than $\alpha = 0.05$, we should reject H_0 at $\alpha = 0.05$ which means that there is strong evidence to say that there exists differences between sample means. **(+5 points)**

(c) Among p-values in the results, the p-values of D-A difference and D-C difference are less than $\alpha = 0.05$. Therefore, there is strong evidence to say that there exist differences between group D&A and group D&C. **(+10 points)**

- For the ANOVA table in (b), **(−1 point)** will be deducted for each wrong or blank entries.

3

50 points

(**by R**) Suppose you want to determine whether the brand of laundry detergent used and the temperature affects the amount of dirt removed from your laundry. To this end, you buy two different brands of detergent (”Super” and ”Best”) and choose three different temperature levels (”cold”, ”warm”, and ”hot”). With each combination 4 loads were washed and the dirt removed were recorded.

(a) Write a model with an interaction. **(+10 points)**

(b) Draw a side-by-side boxplot and make comments. **(+10 points)**

(c) Is there a difference in the mean of two detergents? Use $\alpha = 0.05$. **(+10 points)**

(d) Is there a difference in the mean of three temperature levels? Use $\alpha = 0.05$. **(+10 points)**

(e) Does there exist an interaction between detergent and temperature? Use $\alpha = 0.05$. **(+10 points)**

Solution. (a) The model of the problem is

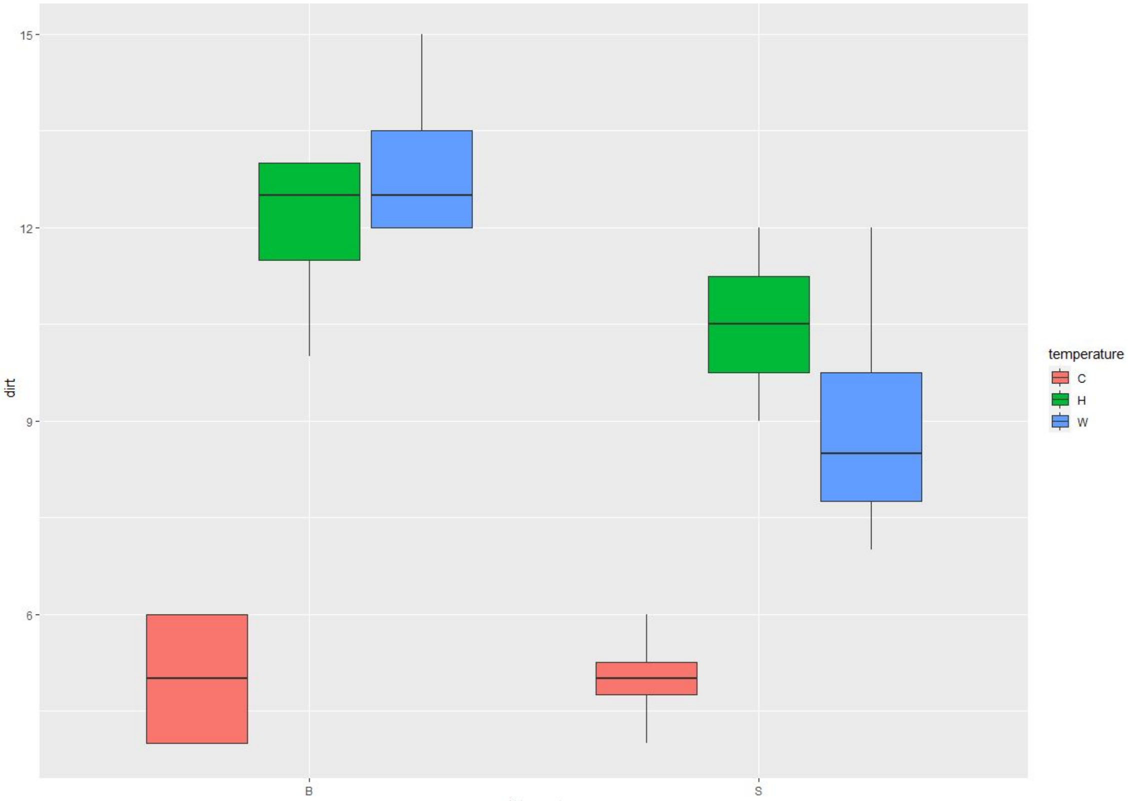
$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ijk} \quad \textbf{(+10 points)}$$

where Y_{ijk} is the amount of dirt removed from k-th observation of i brand with j temperature, α_i is the factor of i brand, β_j is the factor of j temperature, γ_{ij} is the interaction factor of i brand and j temperature, and ϵ_{ijk} is the residual from k-th observation of i brand with j temperature.

(b) Using **R**,

```
ggplot(aes(y=dirt,x=detergent,fill=temperature), data=detergent)+geom_boxplot()
```

while the result is

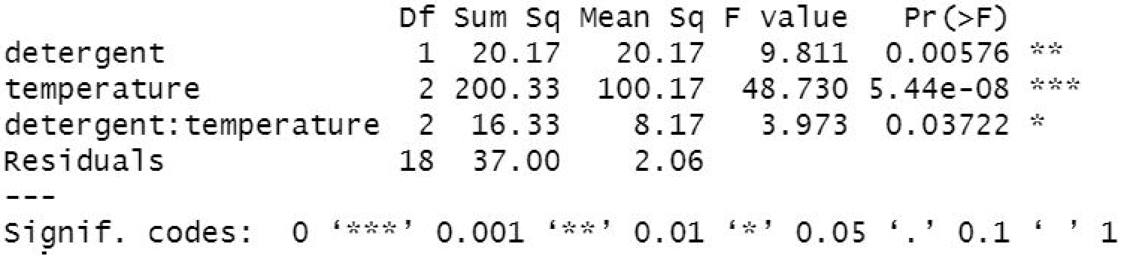


(+10 points) At the cold level, both Super and Best have almost the same mean. At warm and hot, Best outperformed Super, most notably at warm. For the temperature levels, cold did the worst, and the mean of warm higher than hot when used with Best but lower with Super. This is an indicator that an interaction term may be significant.

(c) Using **R**,

```
summary(aov(dirt~detergent*temperature, data=detergent))
```

while the result is



```
                      Df Sum Sq Mean Sq F value   Pr(>F)    
detergent              1  20.17   20.17   9.811  0.00576 ** 
temperature            2 200.33  100.17  48.730 5.44e-08 ***
detergent:temperature  2  16.33    8.17   3.973  0.03722 *  
Residuals             18  37.00    2.06                     
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Detergent has a p-value of .006 **(+5 points)**, which is less than $\alpha = 0.05$, so we would reject that the means of both detergents are the same. **(+5 points)**

(d) The p-value of temperature is approximately 0 **(+5 points)**, which is also less than α, so we have evidence that there is a difference among the means for cold, warm and hot. **(+5 points)**

(e) The interaction term has a p-value of .037 **(+5 points)** which is less than α, which indicates that what we observed in the individual value plot is correct, an interaction term is needed. **(+5 points)**

4(10.5) (**by R**) Four standard chemical procedures are used to determine the magnesium content in a certain chemical compound. Each procedure is used four times on a given compound with the following data resulting.

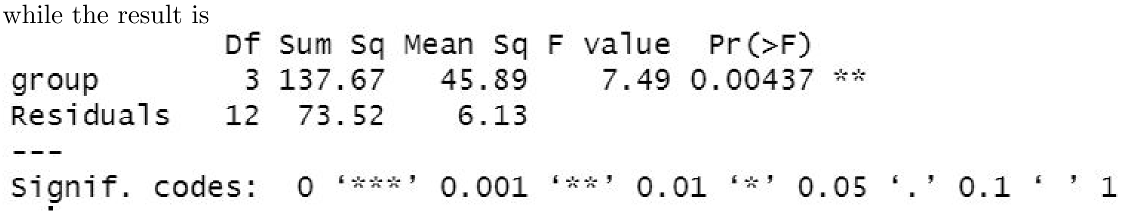
Method			
1	2	3	4
76.42	80.41	74.20	86.20
78.62	82.26	72.68	86.04
80.40	81.15	78.84	84.36
78.20	79.20	80.32	80.68

Do the data indicate that the procedures yield equivalent results?

Solution. Our hypothesis is $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$. Using **R**,

```
> X = data.frame(group=c("1","1","1","1","2","2","2","2","3","3","3","3","4","4","4","4"),
    content = c(76.42,78.62,80.40,78.20,80.41,82.26,81.15,79.20,74.20,72.68,78.84,80.32,86.42,
    86.04,84.36,80.68))
> result = aov(content~group, data=X)
> summary(result)
```

while the result is



```
            Df Sum Sq Mean Sq F value  Pr(>F)
group        3 137.67   45.89    7.49 0.00437 **
Residuals   12  73.52    6.13
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Since the p-value 0.00437 is less than $\alpha = 0.05$, we should reject our hypothesis which means that there is strong evidence to say that the procedures yield different results.

4(10.9) The following data relate to the ages at death of a certain species of rats that were fed 1 of 3 types of diets. Thirty rats of a type having a short life span were randomly divided into 3 groups of size 10 each. The sample means and sample variances of the ages at death (measured in months) of the 3 groups are as follows:

	Very Low Calorie	Moderate Calorie	High Calorie
Sample mean	22.4	16.8	13.7
Sample variance	24.0	23.2	17.1

Test the hypothesis, at the 5 percent level of significance, that the mean lifetime of a rat is not affected by its diet. What about at the 1 percent level?

Solution. Our hypothesis is $H_0 : \mu_L = \mu_M = \mu_H$. Then by calculation,

$$SS_W = (10-1)(24+23.2+17.1) = 578.7$$

$$SS_B = 10\{(22.4-52.9/3)^2 + (16.8-52.9/3)^2 + (13.7-52.9/3)^2\} = 388.8667$$

$$T = \frac{SS_B/(3-1)}{SS_W/(3(10-1))} = 9.072$$

Since $P(F_{2,27} > 9.072) = 0.00097 < 0.01 < 0.05$, we should reject our hypothesis for both $\alpha = 0.05$ and $\alpha = 0.01$.

4(10.13) (**by R**) Five servings each of three different brands of processed meat were tested for fat content. The following data (in fat percentage per gram) resulted.

Brand	1	2	3
Fat	32	41	36
Content	34	32	37
	31	33	30
	35	29	28
	33	35	33

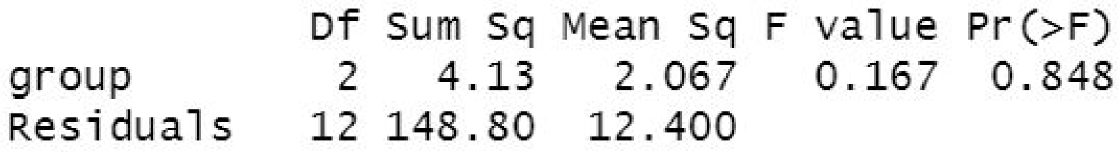
(a) Does the fat content differ depending on the brand?

(b) Find confidence intervals for all quantities $\mu_i - \mu_j$ that, with 95 percent confidence, are valid.

Solution. (a) Our hypothesis is $H_0 : \mu_1 = \mu_2 = \mu_3$. Using **R**,

```
> X = data.frame(group=c("1","1","1","1","1","2","2","2","2","2","3","3","3","3","3"),
    content = c(32,34,31,35,33,41,32,33,29,35,36,37,30,28,33))
> result = aov(content~group, data=X)
> summary(result)
```

while the result is



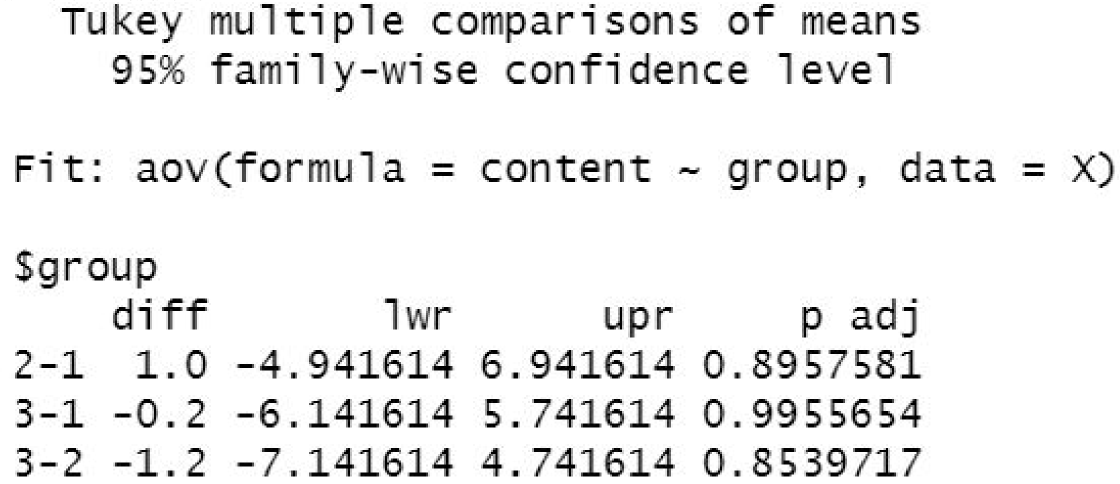
```
            Df Sum Sq Mean Sq F value Pr(>F)
group        2   4.13   2.067   0.167  0.848
Residuals   12 148.80  12.400
```

Since the p-value 0.848 is bigger than $\alpha = 0.05$, we should not reject our hypothesis which means that there is no strong evidence to say that the fat content differ depending on the brand.

(b) Using **R**,

```
> TukeyHSD(aov(content~group, data=X))
```

while the result is



```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = content ~ group, data = X)

$group
    diff       lwr      upr     p adj
2-1  1.0 -4.941614 6.941614 0.8957581
3-1 -0.2 -6.141614 5.741614 0.9955654
3-2 -1.2 -7.141614 4.741614 0.8539717
```

which means that 95 percent confidence intervals of $\mu_2 - \mu_1, \mu_3 - \mu_1$ and $\mu_3 - \mu_2$ are $(-4.941614, 6.941614), (-6.141614, 5.741614)$ and $(-7.141614, 4.741614)$ respectively.