Multivariate Statistical Methods for Engineering and Management

Master in Industrial Engineering and Management

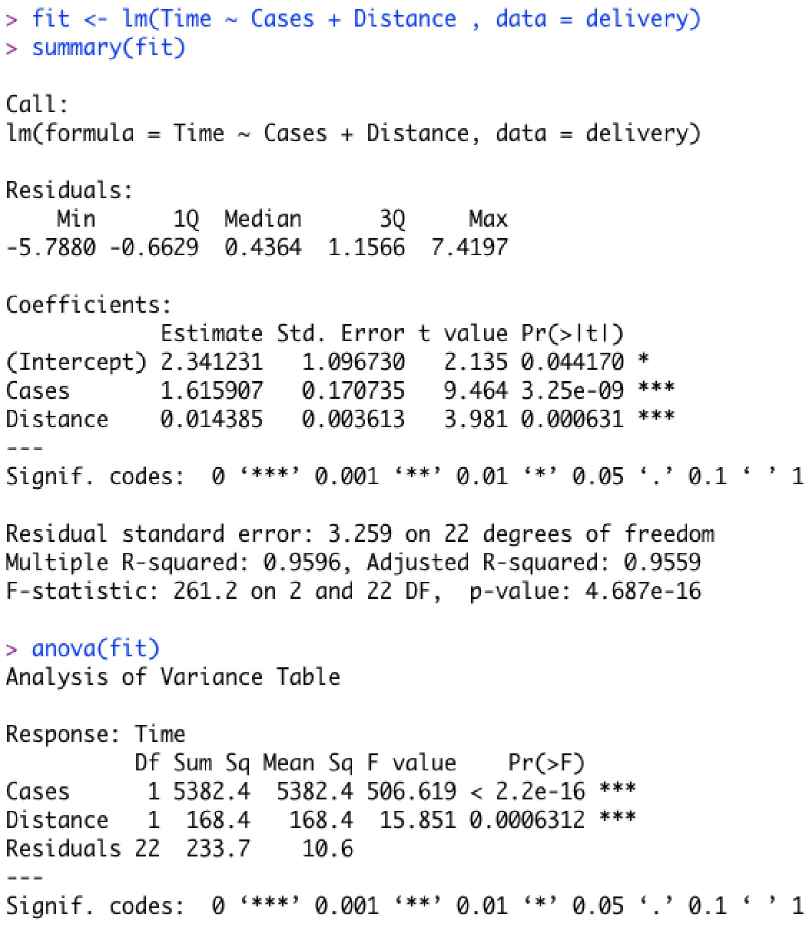
1st Semester – 2019/2020 **1st Exam**

07/01/2020 – 3:00 PM – Room: 1-2 **Duration: 3h**

Justify your answers

Group I 7 points

A soft drink bottler is analyzing the vending machine serving routes in his distribution system. The aim is predicting the time required by the distribution driver to service the vending machines in an outlet. This service activity includes stocking the machines with new beverage products and performing minor maintenance or housekeeping. It has been suggested that the two most important variables influencing delivery time (Y in min) are the number of cases of product stocked (x_1) and the distance walked by the driver (x_2 in feet). 25 observations on delivery times, cases stocked and walking times have been recorded and the output obtained for the adjust of a linear model to the **delivery** dataset in **R** are:



```
> fit <- lm(Time ~ Cases + Distance , data = delivery)
> summary(fit)

Call:
lm(formula = Time ~ Cases + Distance, data = delivery)

Residuals:
    Min      1Q  Median      3Q     Max 
-5.7880 -0.6629  0.4364  1.1566  7.4197 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 2.341231   1.096730   2.135 0.044170 *  
Cases       1.615907   0.170735   9.464 3.25e-09 ***
Distance    0.014385   0.003613   3.981 0.000631 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.259 on 22 degrees of freedom
Multiple R-squared: 0.9596, Adjusted R-squared: 0.9559 
F-statistic: 261.2 on 2 and 22 DF,  p-value: 4.687e-16

> anova(fit)
Analysis of Variance Table

Response: Time
          Df Sum Sq Mean Sq F value    Pr(>F)    
Cases      1 5382.4  5382.4 506.619 < 2.2e-16 ***
Distance   1  168.4   168.4  15.851 0.0006312 ***
Residuals 22  233.7    10.6                      
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

(a) State the estimated regression equation and the estimated variance of the response variable Y. (1.0)

(b) Using the Anova output:

i) Compute the coefficient of multiple determination and the adjust coefficient of multiple determination. Confirm the results with the **R** output. Comment the adequacy of the model to this dataset. (1.0)

ii) Test the significance of the regression, using $\alpha = 0.05$. State the hypotheses, test statistic, decision rule and the conclusion. Confirm your findings with the p-value returned by **R**. What extra assumption is necessary to perform this test? (2.0)

(c) Test the contribution of each variable to the model using the t-test. State the hypotheses, test statistic, decision rule and the conclusion based on the p-values for these tests obtained with **R**. Draw your conclusions. (1.5)

(d) Derive a 95% confidence interval for the intercept of the regression equation. (1.5)

Group II 6.0 points

A laboratory method gives the possibility of determining the weight percentage of clay (W_1), "fine" sand (W_2) and "coars" sand (W_3) in a gravel sample. (This does not determine all types of gravel i.e. $W_1+W_2+W_3$ is always smaller than 100). Such a determination is of great help in finding out what the gravel can be used for. Denoting by $X_i = (W_i - \bar{W}_i)$, for $i = 1, \ldots, 3$, below is the sample covariance matrix of $\mathbf{X}^T = (X_1, X_2, X_3)$ based on 31 determinations of the composition of gravel from different locations and its eigenvectors ($\hat{\gamma}_i$) and eigenvalues ($\hat{\lambda}_i$).

$$\mathbf{S} = \begin{pmatrix} 6.8491 & & \\ 5.1952 & 6.8433 & \\ 1.5193 & 1.5384 & 2.3077 \end{pmatrix};$$

$\hat{\gamma}_1$	$\hat{\gamma}_2$	$\hat{\gamma}_3$
-0.6917168	-0.19534679	0.69524640
-0.6917123	-0.09741748	-0.71557245
-0.2075139	**a**	0.06773847
$\hat{\lambda}_1$=12.500054	$\hat{\lambda}_2$= 1.850005	$\hat{\lambda}_3$= 1.650041

(a) Compute the missing value **a**, the total variance and the generalized variance of the sample. Justify your answer. (1.0)

(b) Compute the percentage of the total sample variability explained by each sample principal component. Comment. (1.0)

(c) Write the first sample principal component. Give a verbal description of gravel samples where the first principal component is "large" and of samples where it is "small". Would the result change if the sample correlation matrix is used to extract the principal components? (2.0)

(d) Denoting by $\hat{Y}_1$ the 1st sample principal component, show that $\widehat{\text{cov}}(X_i, \hat{Y}_1) = \hat{\lambda}_1 \hat{\gamma}_{1i}$, $i = 1, \ldots, 3$, and $\widehat{\text{var}}(\hat{Y}_1) = \hat{\lambda}_1$. Find the sample correlation between the first sample principal component and each variable. Compare the first sample principal component interpretation with your findings in part (c). (2.0)

Group III 3.0 points

Consider a random vector $\mathbf{X} \in I\!R^5$ with $E(\mathbf{X}) = \mathbf{0}$ and covariance matrix

$$\mathbf{\Sigma} = \begin{pmatrix} 4 & & & & \\ 1 & 3 & & & \\ 0 & 2 & 10 & & \\ 2 & 2 & 4 & 6 & \\ 4 & 3 & 4 & 6 & 12 \end{pmatrix},$$

and the two-factors orthogonal model with the loadings matrix $\mathbf{\Lambda} = \begin{pmatrix} 1 & 1 \\ 1 & 0 \\ 2 & -2 \\ 2 & 0 \\ 3 & 1 \end{pmatrix}$.

(a) Calculate the communalities and the unique variances associated with each manifest variable. Obtain the proportion of total variance that is explained by each common factor and by both common factors. Comment the results. (1.5)

(b) Using the rotation matrix $\mathbf{C} = \begin{pmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & -1/\sqrt{2} \end{pmatrix}$ rotate $\mathbf{\Lambda}$ to obtain a new loadings matrix $\mathbf{\Lambda}^*$. In terms of the proportion of total variance explained by each common factor, does the rotated solution seem better than the original one? (1.5)

Group IV 4.0 points

The classification of five US cities with two criminal indicators, x_1 and x_2, is:

	Atlanta	*Boston*	*Chicago*	*Dallas*	*Detroit*
x_1	16.5	4.2	11.6	18.1	13.0
x_2	24.8	13.3	24.7	34.2	35.7

with the following euclidean distance matrix:

$$\mathbf{D} = \begin{array}{l} \textit{Atlanta} \\ \textit{Boston} \\ \textit{Chicago} \\ \textit{Dallas} \\ \textit{Detroit} \end{array} \begin{bmatrix} 0 & & & & \\ 16.84 & 0 & & & \\ 4.90 & 13.59 & 0 & & \\ 9.54 & 25.10 & 11.51 & 0 & \\ 11.45 & 24.07 & 11.09 & 5.32 & 0 \end{bmatrix}.$$

(a) Use the euclidean distance matrix $\mathbf{D}$ to clustering the cities with the single linkage method. Draw the corresponding dendrogram. (2.0)

(b) Admit that based on the classification of x_1 and x_2 you want to group the five cities in 2 clusters. Assuming Boston and Dallas as the initial centroids of the 2 clusters partition, apply the K-means algorithm, with the euclidean distance, to obtain the final 2 clusters partition. (2.0)