كلية الحوسبة والمعلوماتية |
College of Computing and Informatics
الجامعة السعودية الإلكترونية
SAUDI ELECTRONIC UNIVERSITY
2011-1432

Semester 1 – 2021/2022

Course Code	DS610
Course Name	Advanced Applied Statistics
Assignment type	Project
Module	ALL

Student ID	G200002614	G200007615
Student Name	Enad S. Alotaibi	Abdulaziz M. Alqumayzi

Task 1

Paper 1) A Multiple Linear Regression Approach for Estimating the Market Value of Football Players in Forward Position

The paper A Multiple Linear Regression Approach for Estimating the Market Value of Football Players in Forward Position is about estimating the market value of football players in forward position. Physical and performance factors during the 2017 - 2018 season are considered. Data is from 4 major European leagues. This data comprises forward players aged between 20 and 34 from Bundesliga, Series A, LaLiga, and Premier League.

This data contains the information about the league they are playing, the club they are playing, heights, dominant legs, their ages, outfitter, nationalities, direct contribution to goal, matches played in that season, card scores, and total time they have played in that season.

According to this model, 105 variables were used in this model based on metrics listed above, from this model it is found that the r-squared was 0.963(96.3%). This means that most of the variations were explained by the model. Hence, we can see that the models perform well.

Linear regression was appropriate here because the dependent variable (market value) was a quantitative variable. Therefore, giving regression model used to predict the market value of forward football players.

This method can be improved by trying other predictive models for instance decision tree. This model will help clubs to know where is the market value of their player of interest. The market value from previous data can be grouped into two classes then a model built. This will help the clubs to plan their budgets appropriately.

Paper 2) A study on multiple linear regression analysis.

This paper is about estimating the KPSS score based on program development, instructional techniques, counseling, educational psychology, and measurement & evaluation score. The data comes from Sakarya University. Regression was used because the dependent

variable is a quantitative variable. Therefore, regression was used to predict KPSS scores based on program development, instructional techniques, counseling, educational psychology, and measurement & evaluation scores. According to the result, the r squared was 0.87(87%). This means that 87% of variations were expressed by the model. 87% means the good performance of the model. On average we can see that KPSS scores are 9.811, the performance of other metrics are as follows:

- Measurement improves KPSS by 1.157
- Educational psychology improves KPSS by 0.090
- Teaching methods reduces KPSS by 0.339
- Guidance reduces KPSS by 0.195 and
- Curriculum development increases KPSS by 0.078

This method can be improved by trying other predictive models for instance decision tree. This model will help schools to know where is the KPSS value of their students of interest. The KPSS value from previous data can be grouped into two classes then a model built. This will help the schools to plan their education appropriately.

Task 2

```
# Cleaning environment
rm(list = ls())

# load packages
suppressPackageStartupMessages({
  library(tidyverse)
  library(psych)
})

# loading data
CarPrice = read_csv('CarPrice_Assignment.csv')
```

The type and structure of the data

```
sapply(X = CarPrice,FUN = class) %>%
  as.data.frame() %>%
  rownames_to_column(.data = .,var = 'Variables') %>%
```

```
  rename('Class' = '.') %>%
  knitr::kable()
```

Variables	Class
car_ID	numeric
symboling	numeric
CarName	character
fueltype	character
aspiration	character
doornumber	character
carbody	character
drivewheel	character
enginelocation	character
wheelbase	numeric
carlength	numeric
carwidth	numeric
carheight	numeric
curbweight	numeric
enginetype	character
cylindernumber	character
enginesize	numeric
fuelsystem	character
boreratio	numeric
stroke	numeric
compressionratio	numeric
horsepower	numeric
peakrpm	numeric
citympg	numeric
highwaympg	numeric
price	numeric

Interpretation

The data set car price has 205 observations and 26 variables. The data type of these variables is shown above.

Derive descriptive statistics:

```
num_df = CarPrice %>%
  select_if(is.numeric) %>%
  select(-c(car_ID))
```

```r
# descriptive statistics
num_df %>%
  describe() %>%
  as.data.frame() %>%
  select(mean,sd,median,min,max,range) %>%
  round(2) %>%
  rownames_to_column(.data = .,var = 'Variables') %>%
  knitr::kable()
```

Variables	mean	SD	median	min	max	range
symboling	0.83	1.25	1.00	-2.00	3.00	5.0
wheelbase	98.76	6.02	97.00	86.60	120.90	34.3
carlength	174.05	12.34	173.20	141.10	208.10	67.0
carwidth	65.91	2.15	65.50	60.30	72.30	12.0
carheight	53.72	2.44	54.10	47.80	59.80	12.0
curbweight	2555.57	520.68	2414.00	1488.00	4066.00	2578.0
enginesize	126.91	41.64	120.00	61.00	326.00	265.0
boreratio	3.33	0.27	3.31	2.54	3.94	1.4
stroke	3.26	0.31	3.29	2.07	4.17	2.1
compressionratio	10.14	3.97	9.00	7.00	23.00	16.0
horsepower	104.12	39.54	95.00	48.00	288.00	240.0
peakrpm	5125.12	476.99	5200.00	4150.00	6600.00	2450.0
citympg	25.22	6.54	24.00	13.00	49.00	36.0
highwaympg	30.75	6.89	30.00	16.00	54.00	38.0
price	13276.71	7988.85	10295.00	5118.00	45400.00	40282.0

Interpretation

The table above shows the descriptive statistics for each numeric variable. The statistics covered are: mean, standard deviation, median, minimum, maximum and range.

Fuel type:

```r
CarPrice %>%
  ggplot(aes(fueltype,fill = fueltype))+
  theme_bw()+
  geom_bar(show.legend = FALSE)+
  coord_flip()+
  labs(x = 'Fuel type', y = 'Frequency',title = 'Fuel type comparison')
```

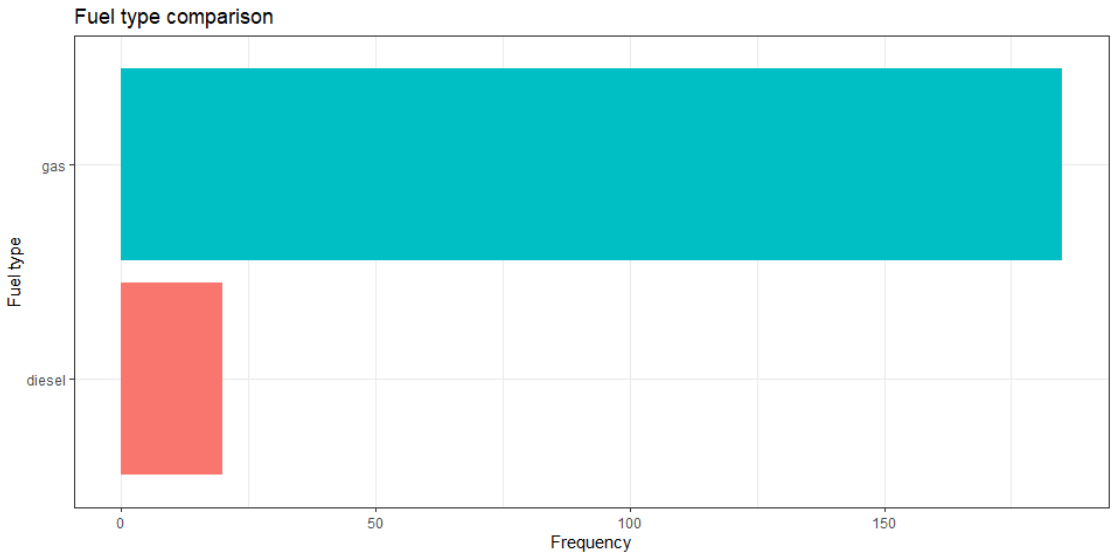

***Figure 1**: Fuel Types Comparison*

Number of doors:

```
CarPrice %>%
  ggplot(aes(doornumber,fill = doornumber))+
  theme_bw()+
  geom_bar(show.legend = FALSE)+
  coord_flip()+
  labs(x = 'Door number', y = 'Frequency',title = 'Door number comparison'
)
```

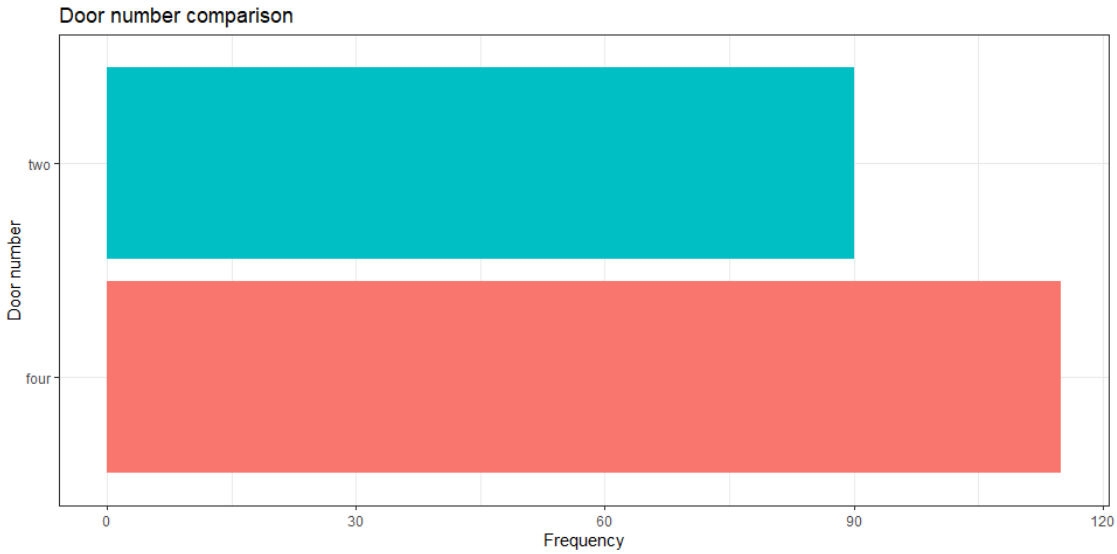

***Figure 2**: Number of doors*

Car body comparison:

```
CarPrice %>%
  ggplot(aes(carbody,fill = carbody))+
  theme_bw()+
  geom_bar(show.legend = FALSE)+
  coord_flip()+
  labs(x = 'Car body', y = 'Frequency',title = 'Car body comparison')
```

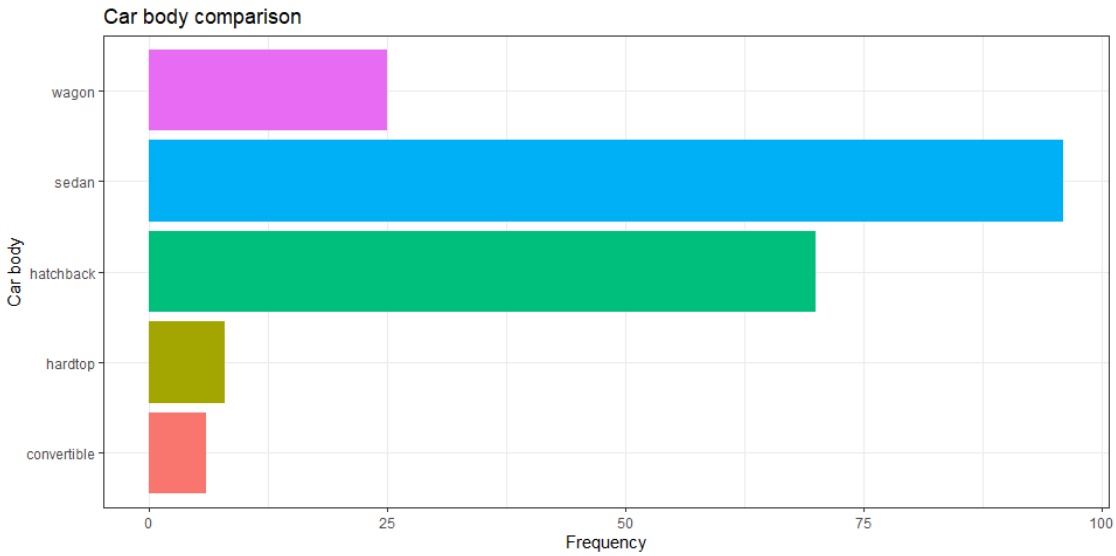

***Figure 3**:Car body comparison*

Interpretation

Generally, cars run on fuel. When it comes to fuel gas is preferred over diesel nearly 90% of cars run on gas and 10% are on diesel. Based on this data 44% of cars have 2 doors while 56% have 4 doors. There are five car body for instance: wagon, sedan, hatchback, hardtop, and convertible. Most of the cars are sedans (46%) and hatchback (34%). The least produced cars are convertibles (3.4%).

Task 3

```
set.seed(123)
index <- sample(1:nrow(CarPrice), 100)
sample_df = CarPrice[index,]
```

Independent t test:

```
# Visualization
width_height = sample_df %>%
  select(carwidth,carheight) %>%
  gather(key = Measure,value = Values)

boxplot(Values ~ Measure,
        data = width_height,
        main = "Car measurement by Values",
        xlab = "Car measurement",ylab = "Value",
        col = "red",border = "black")
```

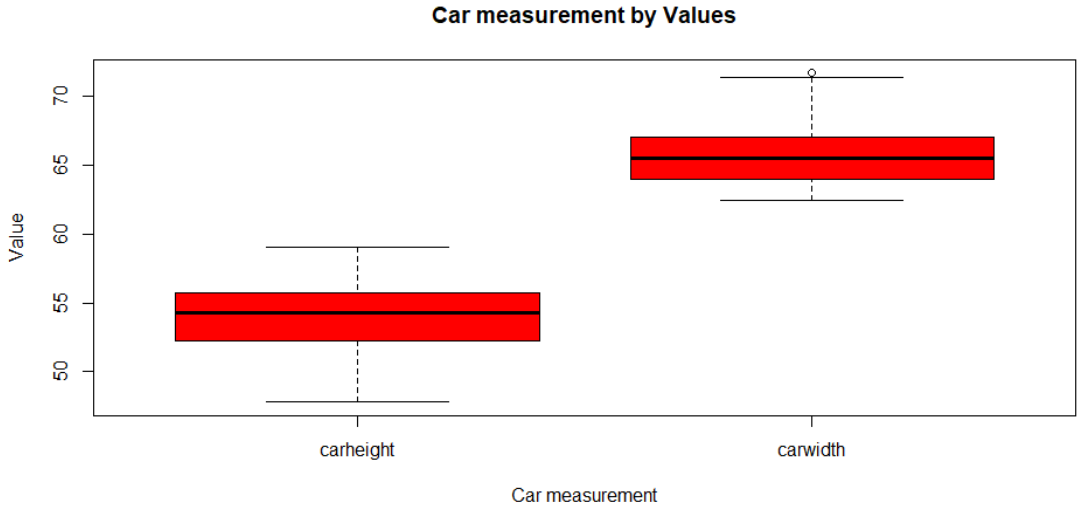

Figure 4: *Cars measurement by values*

```
# test
t.test(sample_df$carwidth, sample_df$carheight,var.equal = TRUE)


    Two Sample t-test

data:  sample_df$carwidth and sample_df$carheight
t = 37.56, df = 198, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 11.26005 12.50795
sample estimates:
mean of x mean of y
   65.886    54.002
```

In the result above:

t is the t-test statistic value (t = 37.56), DF is the degrees of freedom (DF= 198), p-value is the significance level of the t-test (p-value = < 2.2e-16). conf.int is the confidence interval of the mean at 95% (conf.int = [11.26005, 12.50795]); sample estimates mean value of the sample (mean = 65.886, 54.002).

The p-value of the test is < 2.2e-16, which is less than the significance level alpha = 0.05. We can conclude that the width of the car is significantly different from the height of the car with a p-value < 2.2e-16

Dependent t test:

```
t.test(sample_df$carwidth, sample_df$carheight,paired = TRUE)
```

```
    Paired t-test

data:  sample_df$carwidth and sample_df$carheight
t = 48.38, df = 99, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 11.3966 12.3714
sample estimates:
mean of the differences
                 11.884
```

In the result above:

- **t** is the t-test statistic value (t = 48.38),
- **DF** is the degrees of freedom (DF= 99),
- **p-value** is the significance level of the t-test (p-value < 2.2e-16).
- **conf.int** is the confidence interval (conf.int) of the mean differences at 95% is also shown (conf.int= [11.3966, 12.3714])
- **sample estimates** are the mean differences between pairs (mean = 11.884).

The p-value of the test is < 2.2e-16, which is less than the significance level alpha = 0.05. We can then reject the null hypothesis and conclude that the average width of the car is significantly different from the average height of the car with a p-value < 2.2e-16.

Task 4

```
model_df = sample_df[,22:26]

lm_model = lm(formula = (price ~ .),data = model_df)
summary(lm_model)
```

```
Call:
lm(formula = (price ~ .), data = model_df)

Residuals:
    Min      1Q  Median      3Q     Max
-8306.9 -1696.7  -103.4  1871.7 11693.5
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 19952.5591  6232.8113   3.201  0.00186 **
horsepower    142.0188    17.9331   7.919 4.45e-12 ***
peakrpm        -3.2091     0.8123  -3.951  0.00015 ***
citympg       253.6847   312.6194   0.811  0.41912
highwaympg   -365.4374   284.7448  -1.283  0.20248
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4064 on 95 degrees of freedom
Multiple R-squared:  0.7255,    Adjusted R-squared:  0.7139
F-statistic: 62.77 on 4 and 95 DF,  p-value: < 2.2e-16
```

Interpretation

Based on the results above, multiple linear regression has been used to predict car price using horsepower, peakrpm, citympg and highwaympg. From this result, horsepower and peakrpm are statistically significant in predicting price as their respective p-values are less than 0.05(5%). The p-values are 4.45e-12 and 0.00015 for horsepower and peakrpm respectively. citympg and highwaympg are not statistically significant in predicting price as the p-values are more than 0.05 (5%). These p-values are 0.41912 and 0.20248 for citympg and highwaympg respectively.

Based on this model, one unit of horsepower increases the price by 142.0188, peakrpm decreases the price by 3.2091, citympg increases price by 253.6847 and highwaympg decreases price by 365.4374.

The model has performed quite well, as the variation explained by this model is 0.7255 (72.55%).

Task 5

Analysis of variance (ANOVA)

```
# Visualization
boxplot(price ~ drivewheel,
data = sample_df,
main = "Price by Driver wheel",
xlab = "Driver Wheel",ylab = "Price",
col = "red",border = "black")
```

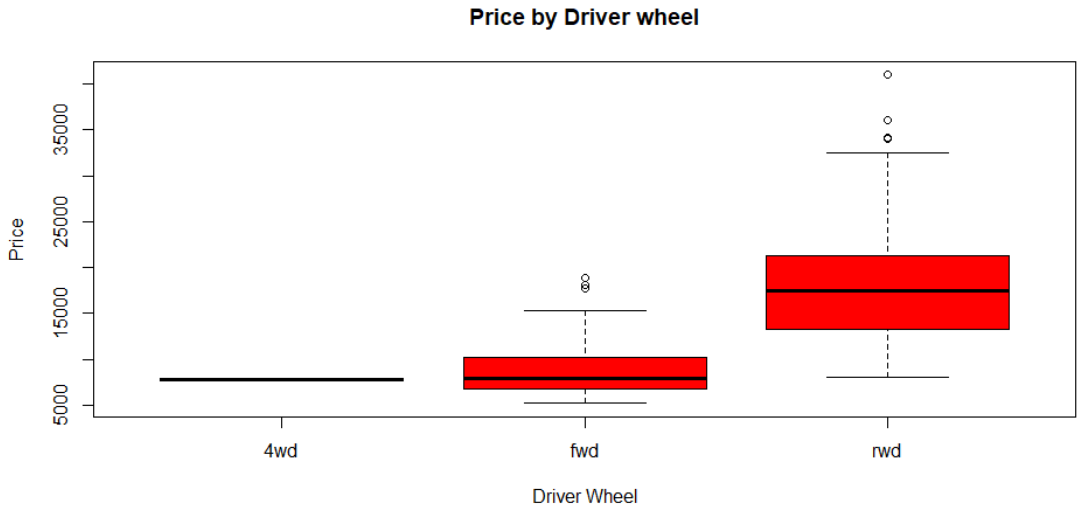

***Figure 5**: Price by driver wheel*

```
# Compute the analysis of variance
res.aov <- aov(price ~ drivewheel, data = sample_df)
# Summary of the analysis
summary(res.aov)
```

```
             Df    Sum Sq   Mean Sq F value   Pr(>F)
drivewheel    2 2.467e+09 1.233e+09   36.82 1.27e-12 ***
Residuals    97 3.250e+09 3.350e+07
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Interpretation

As the p-value = 1.27e-12 is less than the significance level 0.05, we can conclude that there are significant differences between the groups highlighted with "*" in the model summary. In other words, the mean prices among driver wheels are different.

Analysis of covariance (ANCOVA)

```
# Compute the analysis of variance
res.aov2 <- aov(price ~ drivewheel + horsepower, data = sample_df)
# Summary of the analysis
car::Anova(res.aov2, type="III")

Anova Table (Type III tests)

Response: price
               Sum Sq Df F value    Pr(>F)
(Intercept)    1190972  1  0.0681  0.794750
drivewheel   204598068  2  5.8456  0.004021 **
horsepower  1569636616  1 89.6921 2.042e-15 ***
Residuals   1680026006 96
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Interpretation

Based on the output above, we can see the p-values for drivewheel and horsepower are 0.004021 and 2.042e-15 are less than 0.05(5%) significance level. Therefore, drivewheel and horsepower contribute to the model significantly.

References

Peck, R., Olsen. C., & Devore, J. (2020) Introduction to Statistics and Data Analysis.

Field, A., Miles, J., & Field, Z. (2012). Discovering Statistics Using R. 978-1446200469

Koloğlu, Y., Birinci, H., Kanalmaz, I., & Özyılmaz, B. (2018). A Multiple Linear Regression Approach For Estimating the Market Value of Football Players in Forward Position. Retrieved from https://arxiv.org/ftp/arxiv/papers/1807/1807.01104.pdf

Uyanık, G. K., & Güler, N. (2013). A Study on Multiple Linear Regression Analysis. Procedia - Social and Behavioral Sciences, 106, 234–240. https://doi.org/10.1016/j.sbspro.2013.12.027

Brown, A. (2008). The strange origins of the Student's t-test. Retrieved from: https://doi.org/10.36866/pn.71.13

Mayfield, P. (2013). Using the paired t-test to compare Wegmans and Publix supermarkets. SigmaZone. Retrieved from: http://1989-6580.el-alt.com/Articles_PairedTTest.htm

Miller, G. A., & Chapman, J. P. (2001). Misunderstanding analysis of covariance. Journal of Abnormal Psychology, 110(1), 40–48. http://apsychoserver.psych.arizona.edu/JJBAReprints/PSYC501A/Readings/Miller_Chapman_JAP_2001.pdf