Class Test 1 Marking Scheme

Successful upload of `.pdf` file.

2 MARKS

Report

Introduction

Introduction to the data being analysed and to the question of interest. No marks for copying the data description as given. 1 mark removed if the document title has not been changed.

2 MARKS

Exploratory data analysis

Summary statistics on driving accuracy and distance with appropriate comments. One mark removed if the output is simply 'copy-pasted' from R.

3 MARKS

Table 1: Summary statistics on driving accuracy and distance.

Variable	Missing	Complete	Mean	SD	Min.	1st Q.	Median	3rd Q.	Max.
Distance	0	1	287.61	8.55	261.4	282.0	287.0	293.3	315.1
Accuracy	0	1	63.36	5.46	49.0	59.5	63.1	66.9	80.4

Scatterplot of drive accuracy against distance. One mark removed if the plot is not appropriately labelled, and axis labels not adjusted accordingly.

2 MARKS

The regression line can be superimposed here or within the formal data analysis section.

1 MARK

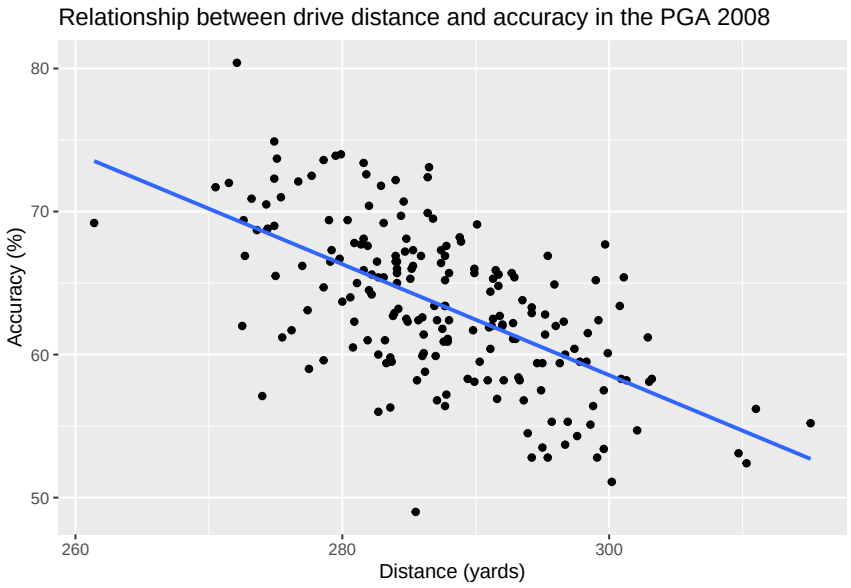

Figure 1: Relationship between driving accuracy and distance. The best-fitting line has been superimposed.

Comments on the scatterplot related to the question of interest.

2 MARKS

Formal data analysis

State the linear regression model being fitted. The linear regression model that will be fitted to the data is as follows:

$$y_i = \alpha + \beta x_i + \epsilon_i, \qquad \epsilon_i \sim N(0, \sigma^2),$$

where y_i and x_i denote the **accuracy** and **distance** of the i^{th} golfer, respectively. The intercept is denoted by α, while the slope is given by β. The i^{th} random error component is denoted by ϵ_i and are normally distributed with mean zero and variance σ^2.

1 MARK

Regression model output. One mark removed if the regression output is simply 'copy-pasted' from R.

2 MARKS

Table 2: Estimates of the intercept and slope from the fitted linear regression model.

term	estimate
intercept	174.925
Distance	-0.388

Appropriate comments on the regression coefficients and the relationship between drive accuracy and distance.

2 MARKS

Plots for checking model assumptions. One mark removed if not properly labelled.

3 MARKS

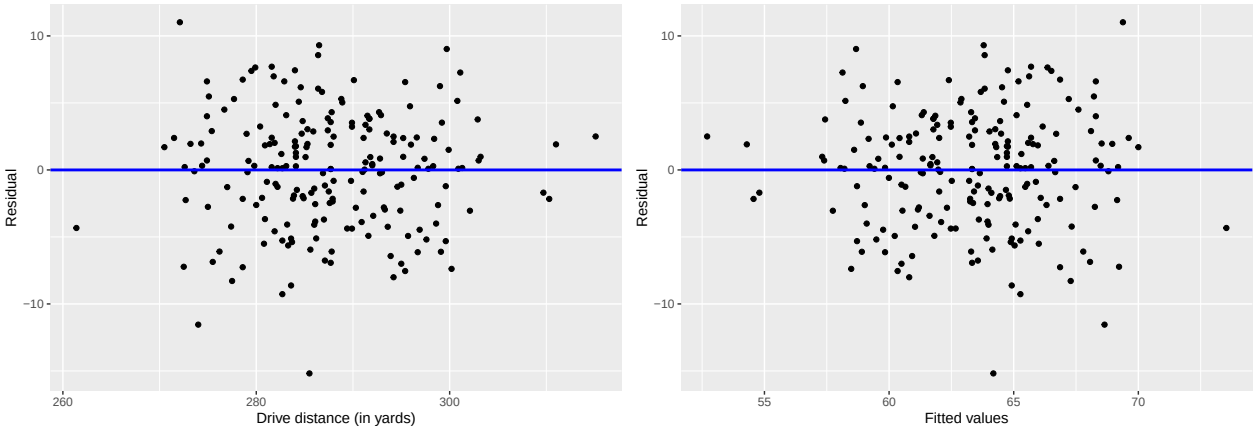

Figure 2: Scatterplots of the residuals against distance (left) and the fitted values (right).

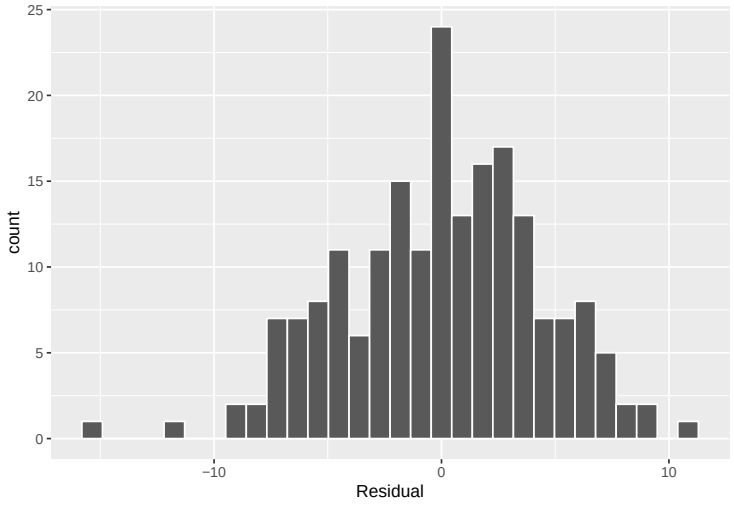

Figure 3: Histogram of the residuals.

Appropriate comments on the model assumptions.

3 MARKS

Conclusions

Overall conclusions with an answer to the question of interest.

2 MARKS

General report layout. This should include figure and table captions, with marks not awarded if these are not used. One mark removed if hyperlinks for sections and Figures not implemented (Tables are allowed no hyperlinks).

2 MARKS

Total: 25 MARKS

Further Question 1

```
q1data <- read_csv("EPL.csv")
```

Modify the data into `tidy` format and set the categorical variable to type `factor` using:

```
q1datatidy <- gather(data = q1data,
                     key = Year,
                     value = Points,
                     - Team)
q1datatidy$Team <- as.factor(q1datatidy$Team)
```

2 marks are awarded for converting to the `tidy` format, while 1 mark is awarded for converting `Team` to a factor.

3 MARKS

To graphically compare the point distributions of the six teams we produce boxplots.

```
ggplot(data = q1datatidy, aes(y = Points, x = Team)) +
         geom_boxplot() +
         scale_x_discrete(labels = c("Arsenal", "Chelsea", "Everton", "Liverpool",
                                     "Man United", "Spurs"))
```

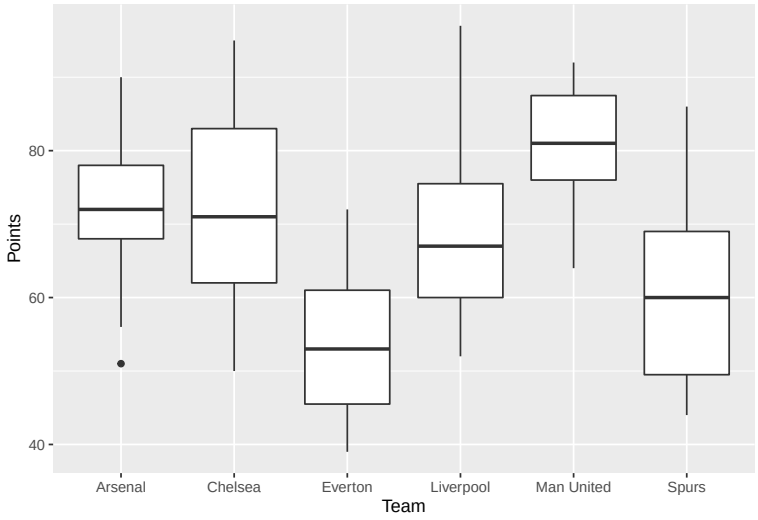

Figure 4: Distribtuion of the total points for each team.

2 marks are awarded for producing the appropriately labelled boxplot.

2 marks are awarded for appropriate comments relating to the data and boxplots.

4 MARKS

Total: 7 MARKS

Further Question 2

```r
set.seed(10)
n_sim <- 100
corr <- -0.65
mu <- c(12, 19)
VarX <- 2
VarY <- 1
sqrt.var <- sqrt(prod(c(VarX, VarY)))
sqrt.var.mat <- matrix(c(1, sqrt.var, sqrt.var, 1), 2, 2)
sigma <- matrix(c(VarX, corr, corr, VarY), 2, 2) * sqrt.var.mat
sim <- mvrnorm(n_sim, mu = mu, Sigma = sigma)
colnames(sim) <- c("X", "Y")
```

2 marks are awarded for correctly identifying the number of observations, the means of X and Y, and the covariance matrix. An additional 2 marks are awarded for the correct use of the `mvrnorm()` function.

4 MARKS

```r
sim <- as.data.frame(sim)
ggplot(data = sim, aes(x = X, y = Y)) +
  geom_point() +
  labs(x = "X", y = "Y")
```

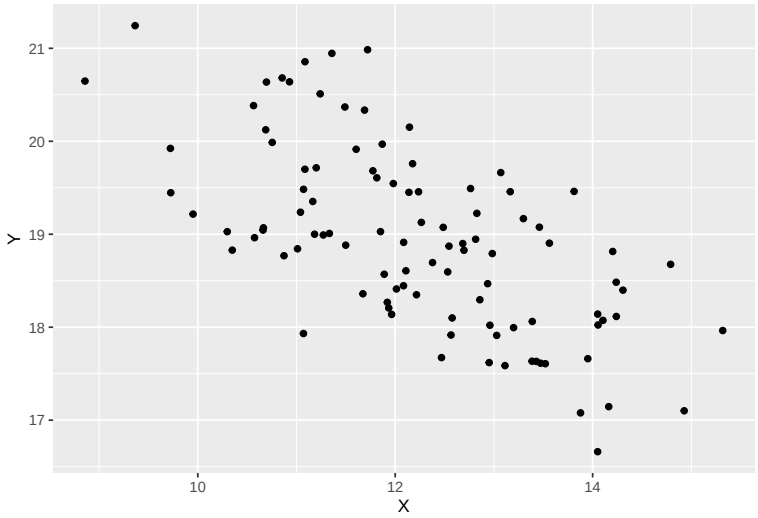

Figure 5: Scatterplot of Y against X.

```r
sim %>%
  summarize(cor(X, Y))
```

```
   cor(X, Y)
1 -0.6470412
```

1 mark for producing the scatterplot and comments, and 1 mark for obtaining the correlation coefficient.

2 MARKS

Total: 6 MARKS

Total: 40 MARKS