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 of the data with appropriate comments. 1 mark removed if the output is simply 'copy-pasted' from R.

3 MARKS

Table 1: Mean, median and standard deviation (sd) of accuracy by gender.

Gender	Accuracy (Mean)	Accuracy (Median)	Accuracy (sd)
Female	67.59	68.3	5.77
Male	63.36	63.1	5.46

Boxplot of accuracy by gender. 1 mark removed if the plot is not appropriately labelled, and axis labels not adjusted accordingly.

2 MARKS

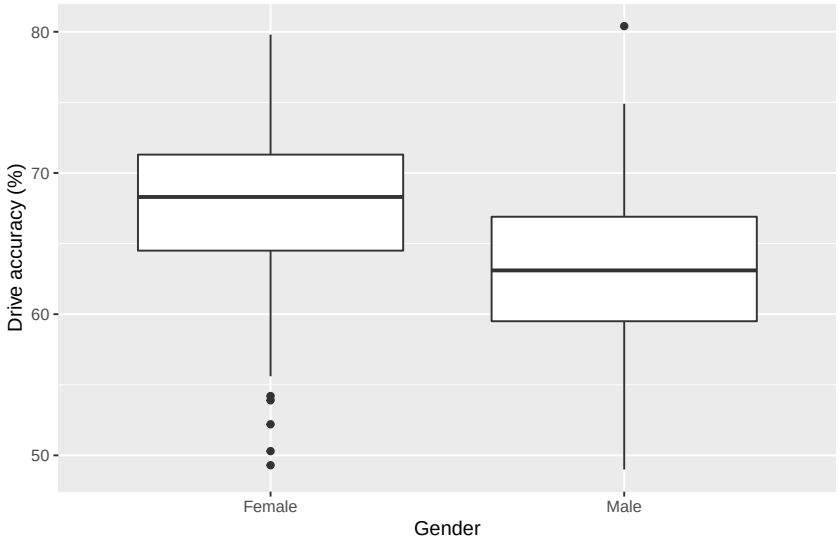

Figure 1: Driving accuracy between PGA and LPGA Tour golfers in 2008.

Comments on the boxplot related to the question of interest.

2 MARKS

Formal data analysis

State the linear regression model being fitted, i.e.

$$\widehat{\text{accuracy}} = \hat{\alpha} + \hat{\beta}_{\text{Male}} \cdot \mathbb{I}_{\text{Male}}(x)$$

where

- the intercept $\hat{\alpha}$ is the mean accuracy for the baseline category (females);
- $\hat{\beta}_{\text{Male}}$ is the difference in the mean accuracy of males relative to the baseline category (females); and
- $\mathbb{I}_{\text{Male}}(x)$ is an indicator function such that

$$\mathbb{I}_{\text{Male}}(x) = \begin{cases} 1 & \text{if the gender of } x\text{th observation is Male,} \\ 0 & \text{Otherwise.} \end{cases}$$

2 MARKS

Regression model output. 1 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	67.591
Gender: Male	-4.226

Appropriate comments on the regression coefficients and differences between males and females.

2 MARKS

Plots for checking model assumptions. 1 mark removed if not properly labelled.

3 MARKS

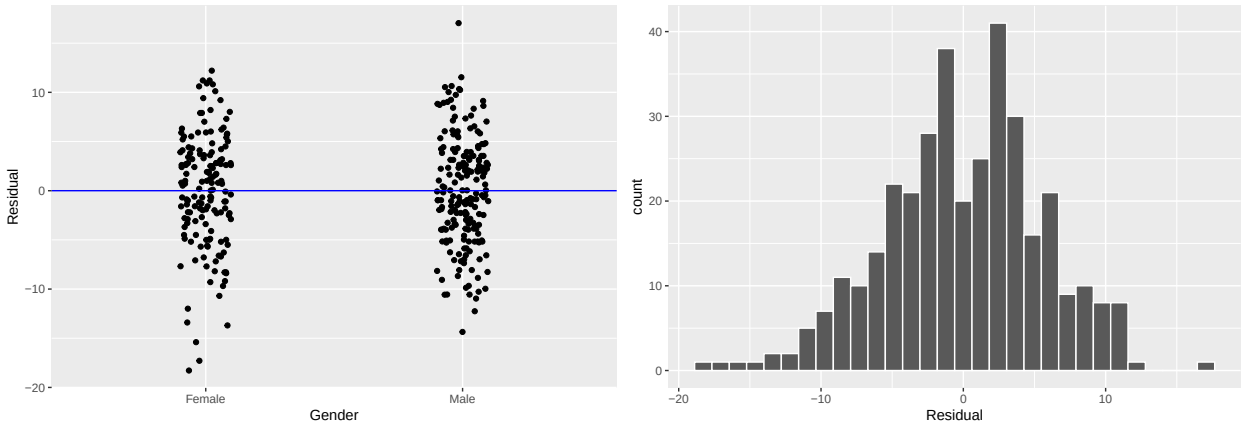

Figure 2: Scatterplots of the residuals by gender (left) and a histogram of the residuals (right).

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. 1 mark removed if hyperlinks for sections and figures not implemented (tables are allowed no hyperlinks). 1 mark removed for consistently poor spelling mistakes/errors.

2 MARKS

Total: 25 MARKS

Further Question 1

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

```
q1datatidy <- gather(data = q1data,
                     key = Fertilizer,
                     value = Yield)
q1datatidy$Fertilizer <- as.factor(q1datatidy$Fertilizer)
```

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

3 MARKS

To graphically compare the distribution of the yields from the four fertilizers we produce boxplots.

```
ggplot(data = q1datatidy, aes(y = Yield, x = Fertilizer)) +
         geom_boxplot()
```

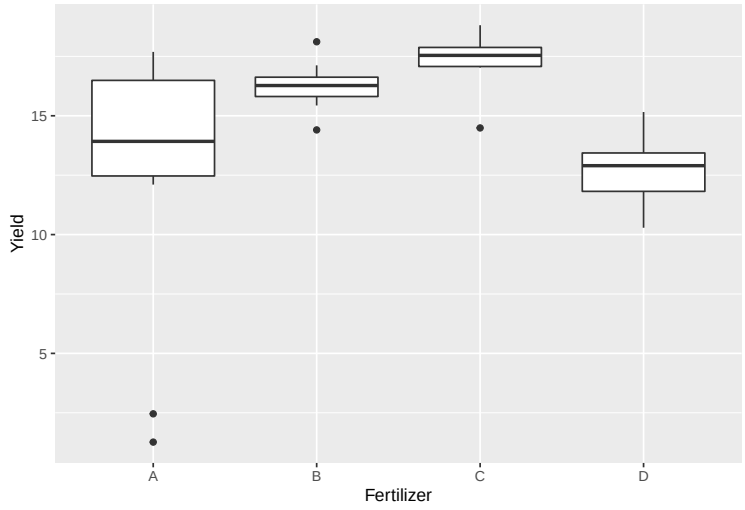

Figure 3: Distribtuion of the crop yields for each fertilizer (A-C) and control group (D).

2 marks are awarded for producing the appropriately labelled boxplot.

The boxplots suggest that there are **two outliers** (i.e. extremely small yields compared to the other 8 yields) for fertilizer A. These should be at the very least investigated further and potentially removed from the data. The boxplots also suggest fertilizer C produces the **biggest crop** yields, then B, then A (with **large variation**) with the control group (D) having the smallest yields.

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

4 MARKS

Total: 7 MARKS

Further Question 2

```
set.seed(10)
n_sim <- 100
corr <- 0.6
mu <- c(10, 18)
sigma <- matrix(c(1, corr, corr, 1), 2, 2)
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 correlation matrix. An additional 2 marks are awarded for the correct use of the `mvrnorm()` function.

4 MARKS

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

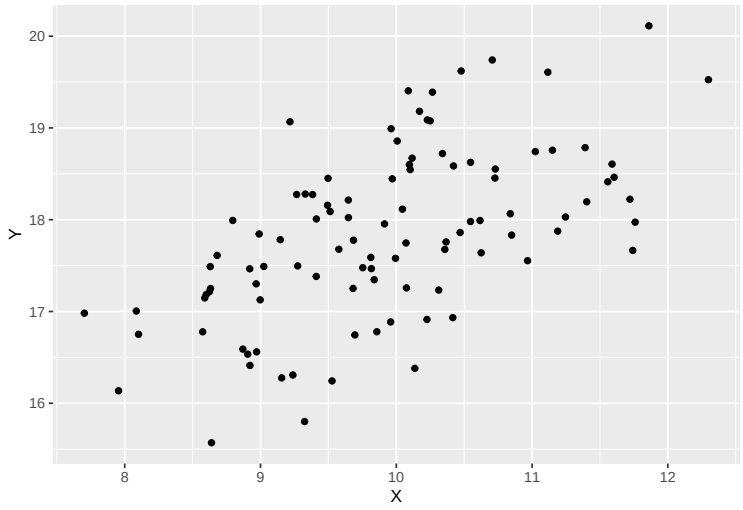

Figure 4: Scatterplot of Y against X.

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

```
  cor(X, Y)
1 0.5813982
```

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

2 MARKS

Total: 6 MARKS

Total: 40 MARKS