Class Test 2 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

Need to begin by filtering out the missing values (NA's). 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) ideal partner height and height by gender.

Gender	Variable	Mean	SD	Minimum	1st quartile	Median	3rd quartile	Maximum
Man	Height	177.97	7.11	159.41	173.39	178.34	182.54	198.21
Man	Ideal.Height	165.36	7.15	146.85	160.32	165.47	170.71	182.79
Woman	Height	166.54	7.54	143.02	161.6	166.25	171.92	186.36
Woman	Ideal.Height	177.95	7.57	155.22	172.42	178.12	182.83	200.18

Table 2: Correlation between partner height and height by gender.

Gender	Correlation
Man	0.495
Woman	0.596

Scatterplot of ideal partner height against height by gender. 1 mark removed if the plot is not appropriately labelled, and axis labels not adjusted accordingly.

2 MARKS

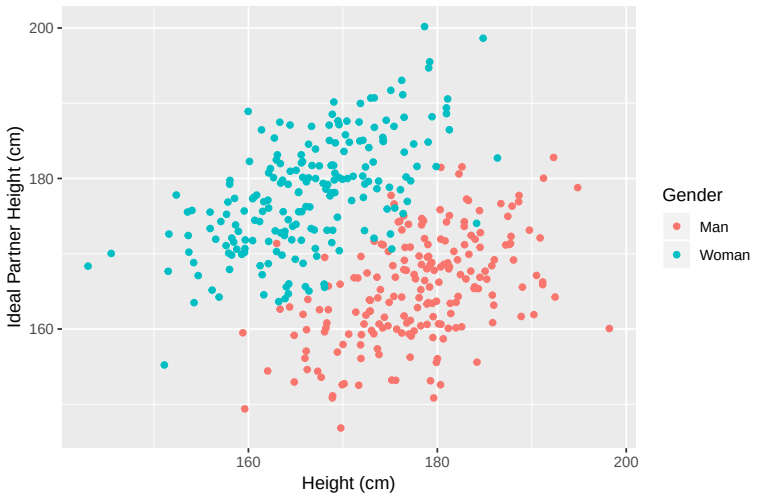

Figure 1: Relationship between ideal partner height and height by gender.

Comments on the scatterplot related to the question of interest.

2 MARKS

Formal data analysis

State the multiple linear regression model being fitted. The 'full' model should contain an interaction term between height and gender.

1 MARK

Stepwise regression (forward/backward selection based on AIC), or some other valid method (examining p-values), for reducing the 'full' model should be implemented to obtain the 'final' model.

1 MARK

Regression model output. 1 mark removed if the regression output is simply 'copy-pasted' from R.

1 MARK

Table 3: Estimates of the regression coefficients from the final model.

term	estimate	std_error	statistic	p_value	lower_ci	upper_ci
intercept	67.256	7.531	8.930	0	52.449	82.062
Height	0.551	0.042	13.048	0	0.468	0.634
GenderWoman	18.896	0.784	24.099	0	17.354	20.437

Hence, from Table 3 we obtain the following regression lines:

$$\widehat{\text{Ideal Height}}_{\text{man}} = 67.256 + 0.551 \cdot \text{Height} \quad (1)$$

$$\widehat{\text{Ideal Height}}_{\text{woman}} = 86.152 + 0.551 \cdot \text{Height} \quad (2)$$

Appropriate comments on the regression coefficients and the relationship between the response and explanatory variables.

2 MARKS

The regression line(s) can be superimposed in the exploratory analysis section or here in the formal data analysis section. However, if different regression lines are subsequently fitted after the modelling process then the corresponding regression lines will need to superimposed again.

1 MARK

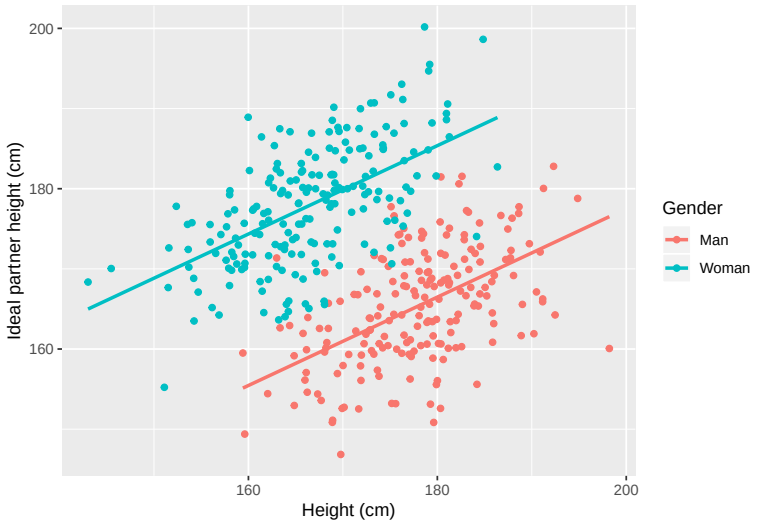

Figure 2: Relationship between ideal partner height and height by gender. The parallel regression lines have been superimposed.

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

3 MARKS

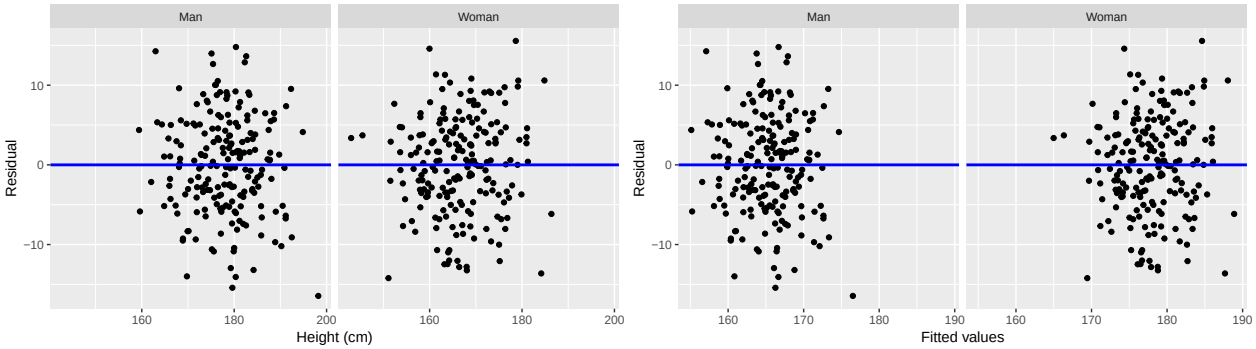

Figure 3: Scatterplots of the residuals against height (left) and the fitted values (right) by gender.

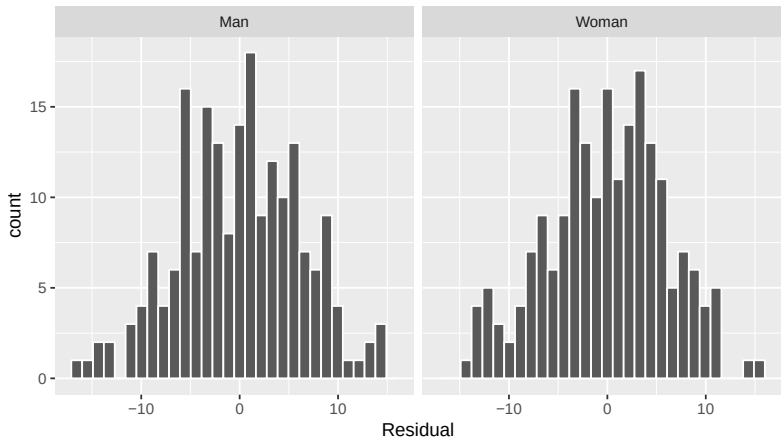

Figure 4: Histogram of the residuals by gender.

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).

2 MARKS

Total: 25 MARKS

Further Question 1

```
log.model <- glm(Gender ~ Height, data = Ideal, family = binomial(link = "logit"))
log.model %>%
  summary()
```

```
Call:
glm(formula = Gender ~ Height, family = binomial(link = "logit"),
    data = Ideal)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.35980  -0.70792   0.07466   0.70179   2.44037

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
```

```
(Intercept) 36.11525    3.57736   10.10   <2e-16 ***
Height      -0.20949    0.02073  -10.11   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 551.74  on 397  degrees of freedom
Residual deviance: 365.32  on 396  degrees of freedom
AIC: 369.32

Number of Fisher Scoring iterations: 5
```

2 MARKS

```
mod.coefs <- log.model %>%
              summary() %>%
              coef()

odds.lower <- exp(mod.coefs["Height", "Estimate"]
                  - 1.96 * mod.coefs["Height", "Std. Error"])
odds.lower
```

```
[1] 0.7787138
```

```
odds.upper <- exp(mod.coefs["Height", "Estimate"]
                  + 1.96 * mod.coefs["Height", "Std. Error"])
odds.upper
```

```
[1] 0.844628
```

The odds of being female decrease by between 15.5% and 22.1% for every 1 cm increase in height.

2 marks for obtaining the confidence interval and 1 mark for the interpretation.

3 MARKS

```
exp(mod.coefs["(Intercept)", "Estimate"] + mod.coefs["Height", "Estimate"] * 171)
```

```
[1] 1.340782
```

The odds of being female given a height of 171 cm are 34% greater than being male.

1 mark for obtaining the estimate, and 1 mark for the interpretation.

2 MARKS

Total: 7 MARKS

Further Question 2

For count data we want to fit a Poisson regression model, with the logarithm as the link function.

2 MARKS

```
poisson.model <- glm(bikes ~ temp, data = bikes, family = poisson(link = "log"))
poisson.model %>%
  summary()
```

```
Call:
glm(formula = bikes ~ temp, family = poisson(link = "log"), data = bikes)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-20.571   -9.868   -2.022    5.333   27.480

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) 3.948244   0.006885   573.5   <2e-16 ***
temp        2.431680   0.014389   169.0   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 251586  on 2133  degrees of freedom
Residual deviance: 222771  on 2132  degrees of freedom
AIC: 236105

Number of Fisher Scoring iterations: 5
```

1 mark for using the `glm()` function. 1 mark for correctly identifying `bikes` and `temp` as the response/explanatory variable. 2 marks for correctly identifying the `family` to use in the `glm()` function and for outputting the results of the fitted model.

4 MARKS

Total: 6 MARKS

Total: 40 MARKS