Class Test 2 TEMPLATE

2416963E

Question 1

A farmer is interested in comparing the effect of different fertilizers on crop yield, and decides to undertake an experiment. He wants to compare three different fertilizers, labelled `A`, `B` and `C`, respectively, against a control group with no fertilizer, labelled `D` . He partitions his field of potatoes into 40 plots, and applies each of the four treatments A, B, C and D to 10 plots at random. At the end of the experiment he measures the total weight (in kilograms) of potatoes grown in each of the 40 plots. Use bootstrap confidence intervals to compare the effects of the fertilizers on the yield of potatoes harvested.

The results of this experiment are stored in `test1.csv`.

10 MARKS

```
q1data <- read_csv("test1.csv")

# Put the data in tidy format
data1 <- gather(data = q1data,
                key = treatment,   ## select column to collapse
                value = weight, ## column name for values
                1:4)
```

```
set.seed(123)
# Take 1000 bootstrap samples for each treatment
samples_A <- data1 %>%
  filter(treatment == "A") %>%
  specify(formula = weight ~ NULL) %>%
  generate(reps = 1000, type = "bootstrap") %>%
  calculate(stat = "mean")

samples_B <- data1 %>%
  filter(treatment == "B") %>%
  specify(formula = weight ~ NULL) %>%
  generate(reps = 1000, type = "bootstrap") %>%
  calculate(stat = "mean")

samples_C <- data1 %>%
  filter(treatment == "C") %>%
  specify(formula = weight ~ NULL) %>%
  generate(reps = 1000, type = "bootstrap") %>%
  calculate(stat = "mean")

samples_D <- data1 %>%
  filter(treatment == "D") %>%
  specify(formula = weight ~ NULL) %>%
  generate(reps = 1000, type = "bootstrap") %>%
  calculate(stat = "mean")

A_bar <- samples_A %>%
  summarize(mean_of_means = mean(stat))
B_bar <- samples_B %>%
```

```
  summarize(mean_of_means = mean(stat))
C_bar <- samples_C %>%
  summarize(mean_of_means = mean(stat))
D_bar <- samples_D %>%
  summarize(mean_of_means = mean(stat))

# Visualize the distribution of the bootstrap samples
pA <- samples_A %>%
  visualize(obs_stat = A_bar)
pB <- samples_B %>%
  visualize(obs_stat = B_bar)
pC <- samples_C %>%
  visualize(obs_stat = C_bar)
pD <- samples_D %>%
  visualize(obs_stat = D_bar)

# Caluclate bootstrap confidence intervals
# Plots of boot strap samples are roughly Normal (symmetric and bell-shaped) so we use "percentile meth
ci_A <- samples_A %>%
  get_ci(level = 0.95, type = "percentile")
ci_B <- samples_B %>%
  get_ci(level = 0.95, type = "percentile")
ci_C <- samples_C %>%
  get_ci(level = 0.95, type = "percentile")
ci_D <- samples_D %>%
  get_ci(level = 0.95, type = "percentile")
```

To compare the effect of three fertilizers (Plot A, Plot B, Plot C) to the control (Plot D), we use bootstrap samples to estimate the distribution of each treatment and create 95% confidence intervals. We are interested in seeing if there is an effect on the weight of potatoes grown using the three different fertilizers. The sample mean of Plot D is 12.39 is used a a point comparison. The condfidence intervals for Plots A is (8.82, 15.49), which contains the point estimate for the control (Plot D). We therefore conclude there is no evidence of a difference in the weight of potatoes grown using fertilizer A. The confidence interval for Plot B is (15.66, 16.83), which does not contain the point estimate for the control (12.39) so we conclude there is evidence that fertilizer B has an effect on the weight of potatos. The confidence interval for Plot C is (16.66, 18.03), which does not contain the point estimate for the control (12.39) so we conclude there is evidence that fertilizer C has an effect on the weight of potatos.

Question 2

A social scientist is interested in exploring what people consider the "ideal height" of a life partner. She sampled 100 male and 100 female adults and asked what they considered the ideal height (in centimetres) for a partner to be. The social scientist also recorded if the subject was a man or a woman and their height (in centimetres), as she was interested to see if there were differences between men and women and if ideal partner height can be predicted given an individual's height.

The data set is stored in the file `test2.csv`.

```
data2 <- read_csv("test2.csv")
```

(a) Generate numerical and graphical summaries that are appropriate for this data and the research questions. Comment briefly on the summaries with respect to the research questions.

5 MARKS

Table 1: Summary statistics on ideal height of a partner by gender of 100 adults.

Gender	n	Mean	St.Dev	Min	Q1	Median	Q3	Max
Man	100	177.5	6.9	161.5952	172.6241	177.5696	181.6523	194.1484
Woman	100	165.5	7.0	150.9097	160.2595	166.4957	169.8838	180.3053

Table 2: Summary statistics on height of an individual and ideal height of a partner of 100 adults.

Variable	Mean	SD	Minimum	1st quartile	Median	3rd quartile	Maximum
Height	171.51	9.17	150.91	166.03	171.14	177.72	194.15
Ideal.Height	171.84	9.73	148.8	164.26	172.17	178.73	194.76

```
#This is here in case ggpairs doesn't work

ggplot(data2[-3], aes(x = Height, y = Ideal.Height)) +
  geom_point() +
  labs(x = "Height", y = "Ideal Height")
```

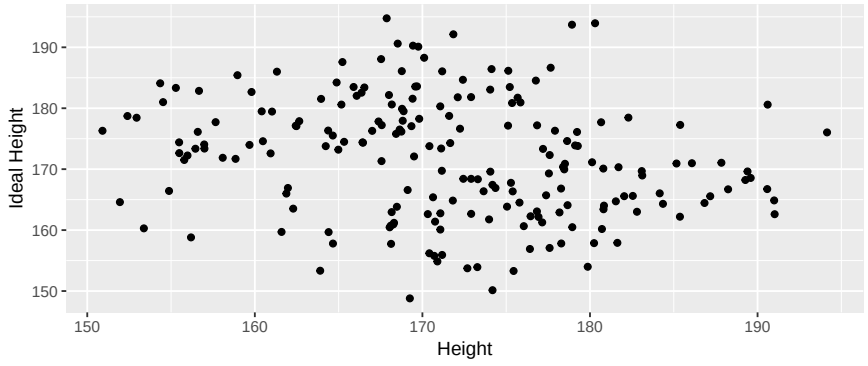

Figure 1: Pairs plot for question 2.

```
ggplot(data2, aes(x = Gender, y = Ideal.Height)) +
  geom_boxplot() +
```

```
  labs(x = "Gender", y = "Ideal Height")
```

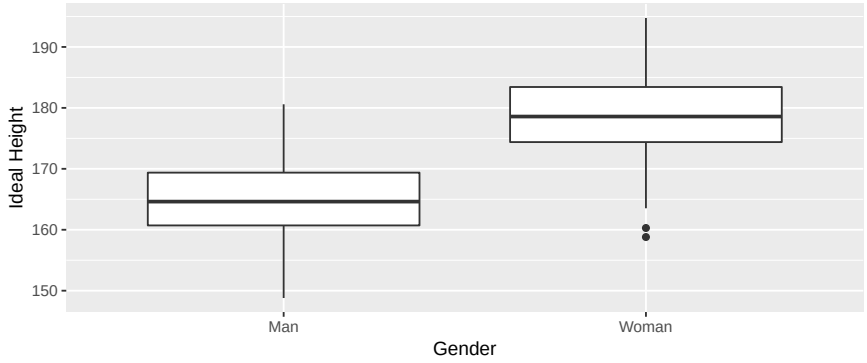

Figure 2: Pairs plot for question 2.

There doesn't appear to be any obvious patterns between `Ideal.Height` and `Height` though the correlation is -0.22, indicating a weak negative correlation between the height of an individual and their partner's ideal height, which seems counterintuitive. The boxplot of Gender against Ideal.Height indicates that women prefer taller partners than men as shown by the non-overlapping IQRs.

(b) Starting with the model with the most parameters, use a model selection method of your choosing to find the model which is most appropriate for this data (you do **not** need to check the model assumptions in this class test). Write down the equation(s) for your chosen fitted model and produce a plot which shows how your chosen model relates to the original data. Use your chosen fitted model to answer the research questions in clear, non-technical English.

6 MARKS

We begin by first fitting the full model with interaction terms.

```
model1 <- lm(Ideal.Height ~ Height*Gender, data = data2)

coeff.model1 <- model1 %>%
  coef() %>%
  as.numeric()

get_regression_table(model1) %>%
  kable(digits = 3, caption = '\\label{tab:model1}Parameter estimates obtained from the model Ideal Hei
  kable_styling(font_size = 10, latex_options = "hold_position")
```

Table 3: Parameter estimates obtained from the model Ideal Height Height * Gender

term	estimate	std_error	statistic	p_value	lower_ci	upper_ci
intercept	84.183	15.361	5.480	0.000	53.889	114.477
Height	0.454	0.086	5.254	0.000	0.284	0.625
GenderWoman	12.381	20.808	0.595	0.553	-28.656	53.417
Height:GenderWoman	0.043	0.121	0.354	0.724	-0.196	0.282

```
reg.tab1 <- get_regression_table(model1)
reg1 <- as.data.frame(reg.tab1)
```

From Table **??** below, the terms `GenderWoman` and `Height:GenderWoman` have a confidence interval containing

zero (-28.66, 53.42) and (-0.2, 0.28), respectively. The term `Height:GenderWoman` is a higher order interaction term so it dropped from the model.

Table 4: Parameter estimates obtained from the model Ideal Height Height + Gender

term	estimate	std_error	statistic	p_value	lower_ci	upper_ci
intercept	80.307	10.735	7.481	0	59.137	101.477
Height	0.476	0.060	7.885	0	0.357	0.595
GenderWoman	19.726	1.105	17.854	0	17.547	21.905

Without the interaction term between Height and Gender, `Height:GenderWoman`, all confidence interval doe not contain zero and this model is chosen as the best-fitting model for this data.

The equation for the best fitting model is:

$$\text{Ideal.Height}_i = \alpha + \beta_{\text{Height}} \cdot \text{Height}_i + \beta_{\text{Gender}} \cdot \mathbb{I}_{\text{Gender}}(i) + \epsilon_i$$

where

- ϵ_i is the error not captured in the model.
- α is the intercept of the regression line;
- β_{Height} is the slope of the regression line for both Females and Males;
- Height_i is the Height of the i^{th} observation
- β_{Gender} is the additional term added to α to get the intercept of the regression line for Gender; and
- $\mathbb{I}_{\text{Gender}}(i)$ is an indicator function such that

$$\mathbb{I}_{\text{Gender}}(i) = \begin{cases} 1 & \text{if the } i\text{th observation Woman,} \\ 0 & \text{Otherwise.} \end{cases}$$

```
ggplot(data2, aes(x = Height, y = Ideal.Height, color = Gender)) +
  geom_point() +
  labs(x = "Height", y = "Ideal Height", color = "Gender") +
  geom_smooth(method = "lm", se = FALSE)
```

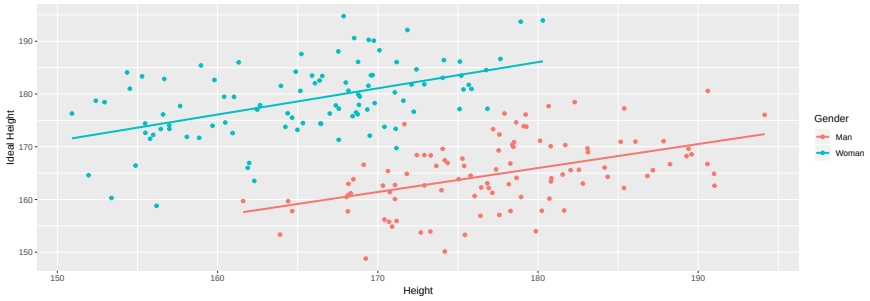

Figure 3: Scatterplot of ideal partner height by height and gender

For every 1 centimeter increase in the height of a women, her ideal height of a partner increases by0.48 centimeters. On average, women prefer partners 19.73 centimeters taller than their own height.

(c) Suppose someone else analysed this data but ignored the effect of gender. What model would they choose to best describe the data? Produce a graph showing their preferred model and compare its interpretation with your preferred model in part (b). What name is given to this apparent contradiction?

4 MARKS

If someone were to model this data without taking gender into account they might use the following model:

$$\text{Ideal.Height}_i = \alpha + \beta_{\text{Height}} \cdot \text{Height}_i + \epsilon_i$$

where

- α is the intercept of the regression line;
- β_{Height} is the slope of the regression line for both Females and Males;
- Height_i is the Height of the i^{th} observation, and
- ϵ_i is the error not captured in the model.

```
ggplot(data2, aes(x = Height, y = Ideal.Height)) +
  geom_point() +
  labs(x = "Height", y = "Ideal Height") +
  geom_smooth(method = "lm", se = FALSE)
```

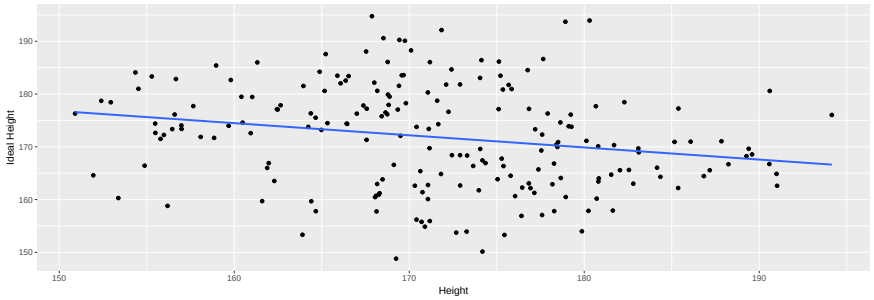

Figure 4: Scatterplot of ideal partner height by height.

This model shows the ideal height of a partner decreases with the height of the individual. When gender is taken into account, the ideal height of a partner increases. This is an example of Simpson's Paradox.