**Workshop 10 – Rough Solutions**

**1.(a)**  
Writing about your data analysis in a scientific report in an **Introduction** section is rare. If there is a need to, it is typically to describe the motivation for the method used to analyse your data.

In the **Materials and Methods** section, you:

* Describe the data you are working with.
* The data analysis method you will apply to analyse your data.
* The plan on how to “extract” the results from the analysis you will present and discuss in the report.

In the **Results** section, you present the results of your data analysis. In particular, you want to summarise key “patterns” from your results, which can be done visually (or as a table). It is important that you *do not* discuss the results of your analysis.

In the **Discussion** section, you want to contextualise the results of your data analysis in terms of the “science” you have conducted, e.g. an experiment or a study. In particular, clearly state how your analysis’s (statistical) inferences help you answer the research questions (aka why you collected data in the first place). This is also the section where you should outline the limitations of your analysis (if applicable).

In the **Conclusions** section, you would only present the key result(s) and inference(s), especially if the bulk of the scientific report is about the data analysis.

**1.(b)**  
In an **Executive Summary**, you would only present key result(s) and inference(s), similar to a scientific report’s **Conclusions** section. However, you may include caveats if it is for a crucial decision-making “forum”.

In contrast, a **Technical Report** describes all steps, from data wrangling to data analysis, that helped shape the **Executive Summary**. In particular, in a fashion where another technically minded person can easily reproduce and verify the results presented in the **Executive Summary**.

**2.(a)**  
One fact we can leverage is that: BMI=Weight/Height2. So Height=Weight/BMI.

library(dplyr)

NHANES.df <- read.csv("NHANES.csv")

NHANES.df.2 <- NHANES.df %>%

mutate(Height = sqrt(Weight / BMI))

summary(NHANES.df.2$Height)

Min. 1st Qu. Median Mean 3rd Qu. Max. NA's

1.448 1.690 1.738 1.738 1.785 2.018 5868

**2.(b)**  
Potentially? Let’s begin with visualising how missing values occur in this dataset.

library(naniar)

# Visualise it with the original data

gg\_miss\_upset(NHANES.df)

It seems that BMI is typically missing value when one of the iron variables *or* Haemoglobin is.

So, the answer is “maybe” and is dependent on how solid your understanding of human biology is. Mathematically speaking, we only need to see if Weight is a good proxy for BMI, as Weight was measured for all respondents.

library(visdat)

# Check missingness

vis\_miss(NHANES.df)

library(ggplot2)

# Visualise BMI being explained by Weight

ggplot(NHANES.df, aes(x = Weight, y = BMI)) +

geom\_hex(binwidth = c(5, 2.5)) +

scale\_fill\_continuous(type = "viridis") +

labs(x = "Weight (kg)", y = "BMI", fill = "Frequency")

Warning: Removed 5868 rows containing non-finite values (`stat\_binhex()`).

# Visualise Height being explained by sqrt( Weight / BMI )

NHANES.df.2 %>%

mutate(x = sqrt(Weight / BMI)) %>%

ggplot(aes(x = x, y = Height)) +

geom\_hex(bins = 50) +

scale\_fill\_continuous(type = "viridis") +

labs(x = "sqrt( Weight / BMI )", fill = "Frequency")

Warning: Removed 5868 rows containing non-finite values (`stat\_binhex()`).

That is… surprisingly linear. So if we “skip” the human biology aspect, it seems that we can recover Height given that BMI=Weight/Height2.

**2.(b) – Bonus Question**

# Data-driven imputation

library(missRanger)

# Let's use all the variables to account for anything a naive data analyst might not know about human biology here...

NHANES.df.3 <- missRanger(Height ~ ., data = NHANES.df.2)

Missing value imputation by random forests

Variables to impute: Height

Variables used to impute: Cancer.Incidence, Cancer.Death, Age, Smoke, Weight, Sex, Height

iter 1

|

| | 0%

|

|======================================================================| 100%

iter 2

|

| | 0%

|

|======================================================================| 100%

iter 3

|

| | 0%

|

|======================================================================| 100%

**2.(c)**  
Biologically speaking, yes it can. In particular, the Serum.Iron measurement should explain a person’s Haemoglobin.

ggplot(NHANES.df, aes(x = Serum.Iron, y = Haemoglobin)) +

geom\_hex(bins = 50) +

scale\_fill\_continuous(type = "viridis") +

labs(x = "Serum Iron (micrograms p/L)", y = "Haemoglobin", fill = "Frequency")

Warning: Removed 1728 rows containing non-finite values (`stat\_binhex()`).

**2.(d)**

NHANES.df %>%

mutate(Age\_Band = cut\_interval(Age, n = 5)) %>%

ggplot(aes(x = Serum.Iron, y = Haemoglobin)) +

geom\_hex(bins = 25) +

scale\_fill\_continuous(type = "viridis") +

labs(x = "Serum Iron (micrograms p/L)", y = "Haemoglobin", fill = "Frequency") +

facet\_wrap( ~ Age\_Band)

Warning: Removed 1728 rows containing non-finite values (`stat\_binhex()`).

It looks like the relationship does not change notably when we “condition” on an age band. The scatter of the points does not drastically change as we scan across the age bands.

**3.(a)**  
A two-sample *t*-test, where the hypothesised value of the difference is 0.

**3.(b)**

thyroid.df <- read.csv("thyroid.csv")

library(ggplot2)

ggplot(thyroid.df, aes(x = thyroid)) +

geom\_dotplot() +

facet\_wrap(~ group)

Bin width defaults to 1/30 of the range of the data. Pick better value with

`binwidth`.

t.test(thyroid ~ group, data = thyroid.df)

Welch Two Sample t-test

data: thyroid by group

t = -2.6153, df = 12.854, p-value = 0.02154

alternative hypothesis: true difference in means between group control and group drug is not equal to 0

95 percent confidence interval:

-4.1564459 -0.3935541

sample estimates:

mean in group control mean in group drug

14.075 16.350

The new drug, on average, does affect the weight of the thyroid gland. With 95% confidence, we estimate that the true mean thyroid weight of the drug group is somewhere 0.4 and 4.2 milligrams lighter than that of the control group.

**3.(c)**

ggplot(thyroid.df, aes(x = body, fill = group)) +

geom\_dotplot()

Bin width defaults to 1/30 of the range of the data. Pick better value with

`binwidth`.

**3.(d)**  
It does not seem that the laboratory animals were effectively identical based on their body weights before the experiment began. This might be problematic because to conclude that the new drug, on average, affects the weight of the thyroid gland, we need to be certain it is only a function of the drug and no other factors.

**3.(d) – Bonus Question**

# Fit a non-parallel lines model

thyroid.fit <- lm(thyroid ~ body \* group, data = thyroid.df)

anova(thyroid.fit)

Analysis of Variance Table

Response: thyroid

Df Sum Sq Mean Sq F value Pr(>F)

body 1 51.569 51.569 71.9607 2.053e-06 \*\*\*

group 1 0.706 0.706 0.9852 0.3405

body:group 1 2.202 2.202 3.0734 0.1051

Residuals 12 8.600 0.717

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

# Fit a parallel lines model as we can drop the interaction

thyroid.fit.2 <- lm(thyroid ~ body + group, data = thyroid.df)

anova(thyroid.fit.2)

Analysis of Variance Table

Response: thyroid

Df Sum Sq Mean Sq F value Pr(>F)

body 1 51.569 51.569 62.0624 2.645e-06 \*\*\*

group 1 0.706 0.706 0.8496 0.3734

Residuals 13 10.802 0.831

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

# Fit a simple linear regression model as we can drop group

thyroid.fit.3 <- lm(thyroid ~ body, data = thyroid.df)

anova(thyroid.fit.3)

Analysis of Variance Table

Response: thyroid

Df Sum Sq Mean Sq F value Pr(>F)

body 1 51.569 51.569 62.736 1.538e-06 \*\*\*

Residuals 14 11.508 0.822

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

The “star-gazing” approach (sequential dropping based on *p*-values) does suggest we should drop the group altogether. However… is this wise? We note that the amount of variability of the non-parallel lines model is higher than the ideal “star-gazing” model.

The new drug, on average, does not affect the weight of the thyroid gland after we controlled for the animals’ body weight (*p*-value = 0.3405). We likely have to re-run the experiment and ensure that the laboratory animals used in the experiment are truly “as similar” as possible to confirm if this is indeed the case.

**4.(a)**  
120 mm Hg is considered “normal” systolic blood pressure in New Zealand.

**4.(b)**  
A one-sample *t*-test, where the hypothesised value of the true mean is 120 mm Hg.

peru.df <- read.csv("peru.csv")

t.test(BP ~ 1, data = peru.df, mu = 120)

One Sample t-test

data: BP

t = 3.5298, df = 38, p-value = 0.001107

alternative hypothesis: true mean is not equal to 120

95 percent confidence interval:

123.1604 131.6601

sample estimates:

mean of x

127.4103

**4.(c)**

library(ggplot2)

ggplot(peru.df, aes(x = age, y = BP)) +

geom\_point() +

labs(x = "Age (years)", y = "Blood pressure (mm Hg)")

ggplot(peru.df, aes(x = years, y = BP)) +

geom\_point() +

labs(x = "Years since migration (years)", y = "Blood pressure (mm Hg)")

ggplot(peru.df, aes(x = weight, y = BP)) +

geom\_point() +

labs(x = "Weight (kg)", y = "Blood pressure (mm Hg)")

ggplot(peru.df, aes(x = height, y = BP)) +

geom\_point() +

labs(x = "Height (mm)", y = "Blood pressure (mm Hg)")

**4.(d)**  
Across all four plots, blood pressure does not seem to depend strongly on any of the other numeric variables.

**4.(d) – Bonus Question**

# If you have done DATAX221: Let's fit a multiple linear regression model here

lm(BP ~ weight + height + years + age, data = peru.df) %>%

summary()

Call:

lm(formula = BP ~ weight + height + years + age, data = peru.df)

Residuals:

Min 1Q Median 3Q Max

-17.263 -6.561 0.875 5.644 22.320

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 90.17743 52.47507 1.718 0.0948 .

weight 1.49865 0.31726 4.724 3.91e-05 \*\*\*

height -0.02761 0.03674 -0.752 0.4575

years -0.53803 0.21951 -2.451 0.0195 \*

age -0.16133 0.27948 -0.577 0.5676

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10.42 on 34 degrees of freedom

Multiple R-squared: 0.4344, Adjusted R-squared: 0.3679

F-statistic: 6.529 on 4 and 34 DF, p-value: 0.0005195

# ...

lm(BP ~ weight + years, data = peru.df) %>%

summary()

Call:

lm(formula = BP ~ weight + years, data = peru.df)

Residuals:

Min 1Q Median 3Q Max

-17.469 -7.878 1.076 6.292 24.113

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 50.3191 15.8184 3.181 0.00302 \*\*

weight 1.3541 0.2672 5.067 1.22e-05 \*\*\*

years -0.5718 0.1879 -3.043 0.00436 \*\*

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10.25 on 36 degrees of freedom

Multiple R-squared: 0.4208, Adjusted R-squared: 0.3886

F-statistic: 13.08 on 2 and 36 DF, p-value: 5.385e-05

# If you have not done DATAX221: Let's fit a simple linear regression model here

lm(BP ~ weight, data = peru.df) %>%

summary()

Call:

lm(formula = BP ~ weight, data = peru.df)

Residuals:

Min 1Q Median 3Q Max

-20.294 -8.491 0.446 6.662 35.040

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 66.5969 16.4639 4.045 0.000255 \*\*\*

weight 0.9629 0.2591 3.716 0.000665 \*\*\*

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 11.34 on 37 degrees of freedom

Multiple R-squared: 0.2718, Adjusted R-squared: 0.2521

F-statistic: 13.81 on 1 and 37 DF, p-value: 0.0006654

**4.(e)**

*Executive Summary*  
The average blood pressure for those weighing more than 60 kilograms was greater than 120 mm Hg. **[DATAX221 specific-knowledge]** We also found that the time since they migrated also explains the average blood pressure once we hold weight constant (*p*-value = 0.0044).

ggplot(peru.df, aes(x = weight, y = BP)) +

geom\_point() +

geom\_smooth(method = "lm") +

geom\_hline(yintercept = 120, colour = "red", linetype = 2) +

labs(x = "Weight (kg)", y = "Blood pressure (mm Hg)")

`geom\_smooth()` using formula = 'y ~ x'

*Technical Report*  
**[DATAX221 specific-knowledge]** We fitted a multiple linear regression model to the data, with blood pressure as the response variable and all other variables as explanatory variables. Using a backwards selection approach to simplify the model, we identified that the height and age of a person were not necessary to explain the average blood pressure.

Then, write some comments on verifying why the final model you used to generate your inferences is appropriate. Usually, this will involve “diagnosing” your (statistical) model to ensure that the assumptions for inference have been met, as this implies that we can trust the confidence intervals, *p*-values, etc., at face value.