STATS 1000 / STATS 1004 / STATS 1504
Statistical Practice 1
Assignment 6
2020

DEADLINE:

- No need to submit

CHECKLIST

☐: Have you shown all of your working, including probability notation where necessary?

☐: Have you given all numbers to **3 decimal** places.

☐: Have you included all R output and plots to support your answers where necessary.

☐: Have you made sure that all plots and tables each have a caption.

☐: If before the deadline, have you submitted your assignment via the online submission on MyUni?

☐: Is your submission a single word document or pdf file - correctly orientated, easy to read? If not, penalties apply.

☐: Penalties for more than one document - 10% of final mark for each extra document. Note that you may resubmit and your final version is marked, but the final document should be a single file.

☐: If after the deadline, but within 24 hours, have you contacted us via the enquiry page on MyUni and then submitted your assignment online via the online submission on MyUni?

☐: Penalties for late submission - within 24 hours 40% of final mark. After 24 hours, assignment is not marked and you get zero.

☐: Assignments emailed instead of submitted by the online submission on MyUni will not be marked and will recieve zero.

☐: Have you checked that the assignment submitted if the correct one, as we cannot accept other submissions after the due date.

☐: Do not write directly on the question sheet.

1. **Chi-square test in R**

For full marks this answer must be typed in Word and plots and output included and captioned.

Myocardial infarction is where a blockage in the blood vessels to the heart cause damage to the heart muscle. It is often referred to as a "heart attack". A 1988 study looked at the effect of aspirin to prevent death after people have suffered a myocardial infection. The dataset is on MyUni and is called `aspirin.xlsx`. Load the data into R, then complete the following.

(a) Perform an Chi-square test to test for an association between treatment group and outcome. For full marks, include

i. your R output,

[1 mark]

```
        Pearson's Chi-squared test with Yates' continuity correction

data:  mytable
X-squared = 8.0691, df = 1, p-value = 0.004503
```

Table 1: R output for Chi-square test for the association between treatment group and outcome.

ii. the null and alternative hypotheses,

[2 marks]

H_0 : The variables treatment group and outcome are independent

H_a : The variables treatment group and outcome are dependent

iii. the value of the test statistic,

[1 mark]

The observed value of the test statistic is 8.069.

iv. the degrees of freedom,

[1 mark]

The degrees of freedom is 1.

v. the P-value, and

[1 mark]

The P-value is 0.005.

vi. whether you reject or retain the null hypothesis at the 5% significance level, and why?

[2 marks]

Reject the null hypothesis at the 5% significance level as the P-value is less than 0.05.

[Total: 8]

2. **Linear regression in R**

For full marks this answer must be typed in Word and plots and output included and captioned.

One of the original uses of linear regression was to examine the relationship between the height of fathers and their sons (both in inches). The dataset `pearson.xlsx` is a dataset obtained by Karl Pearson[1] to look at this relationship. In fact this relationship is why we call it regression[2]. Load the dataset into R and the complete the following.

(a) Produce a scatterplot of son's height (`son_height`) against father's height (`father_height`). Describe the relationship.

[3 marks]

The scatterplot is given in Figure 1.

We see a strong positive linear relationship.

[1] https://en.wikipedia.org/wiki/Karl_Pearson
[2] https://en.wikipedia.org/wiki/Regression_toward_the_mean

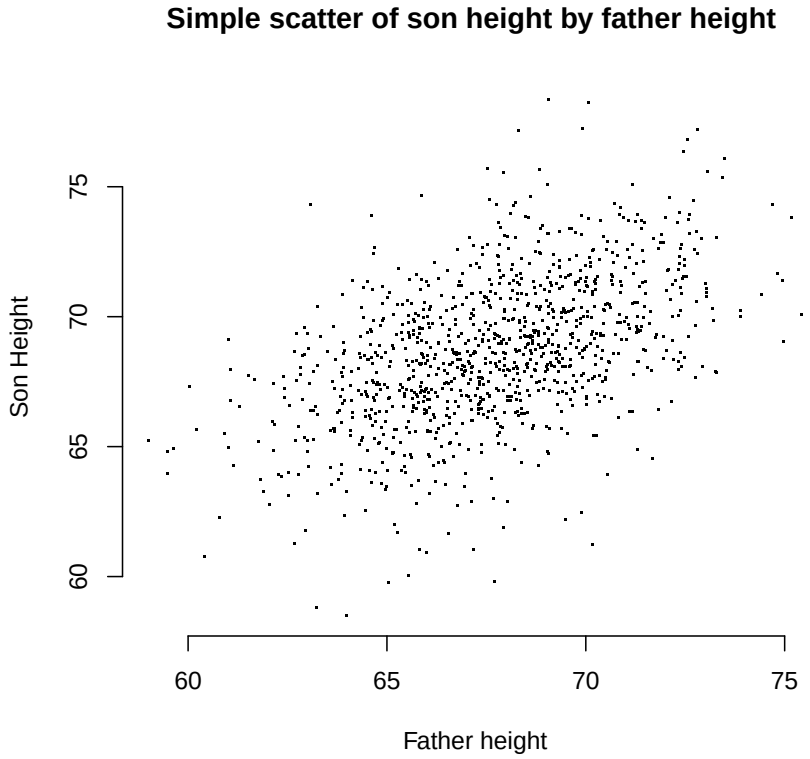

Figure 1:scatterplot of son's height (`son_height`) against father's height (`father_height`).

(b) Test for a statistically significant (5% level) linear relationship between `son_height` and `father_height`. Remember to include

i. your R output,

[1 mark]

The output is given in Table 2.

```
Call:
lm(formula = son.height ~ father.height, data = pearson)

Residuals:
    Min      1Q  Median      3Q     Max
-8.8772 -1.5144 -0.0079  1.6285  8.9685

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

```
(Intercept)   33.88660    1.83235   18.49   <2e-16 ***
father.height  0.51409    0.02705   19.01   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.437 on 1076 degrees of freedom
Multiple R-squared:  0.2513,        Adjusted R-squared:  0.2506
F-statistic: 361.2 on 1 and 1076 DF,  p-value: < 2.2e-16
```

Table 2: R output for simple linear regression between son's height (`son_height`) and father's height (`father_height` .

ii. the null and alternative hypotheses,

[2 marks]

$$H_0 : \beta_1 = 0,$$
$$H_0 : \beta_1 \neq 0,$$

where β_1 is the true slope.

iii. the observed value of test statistic,

[1 mark]

The observed value of the test statistic is 19.01.

iv. the P-value, and

[1 mark]

The P-value is <0.0001.

v. your conclusion. Do you reject or retain the null hypothesis? Why? Give conclusion in context.

[3 marks]

We reject the null hypothesis at the 5% significance level as the P-value is less than 0.05. Hence, there is a statistically significant linear relationship between `son_height` and `father_height`.

(c) Check the assumptions of the linear regression. Remember to include captioned plots where necessary.

[8 marks]

Linearity

What: Linearity.

Where: Residual versus fitted plot (Figure 2).

Expect: Random scatter around the zero line.

See: Random scatter about the zero line.

Conclude: The assumption of linearity seems reasonable.

Constant spread

What: Constant spread.

Where: Residual versus fitted plot (Figure 2).

Expect: Equal spread as fitted value increases.

See: Some decreased spread at the edges but that is fine as we have a large dataset.

Conclude: The assumption of constant spread seems reasonable.

Normality of residuals

What: Normality of residuals.

Where: Normal QQ-plot of residuals (Figure 3).

Expect: Roughly linear points.

See: Roughly linear points, but some deviation at the edges.

Conclude: The assumption of normality of the residuals seems reasonable.

Independence of residuals

What: Independence of residuals.

Where: Look at the experiment design.

Expect: That the subjects were randomly selected.

See: No information is given about how the subjects were obtained.

Conclude: Cannot assess as there is no enough information.

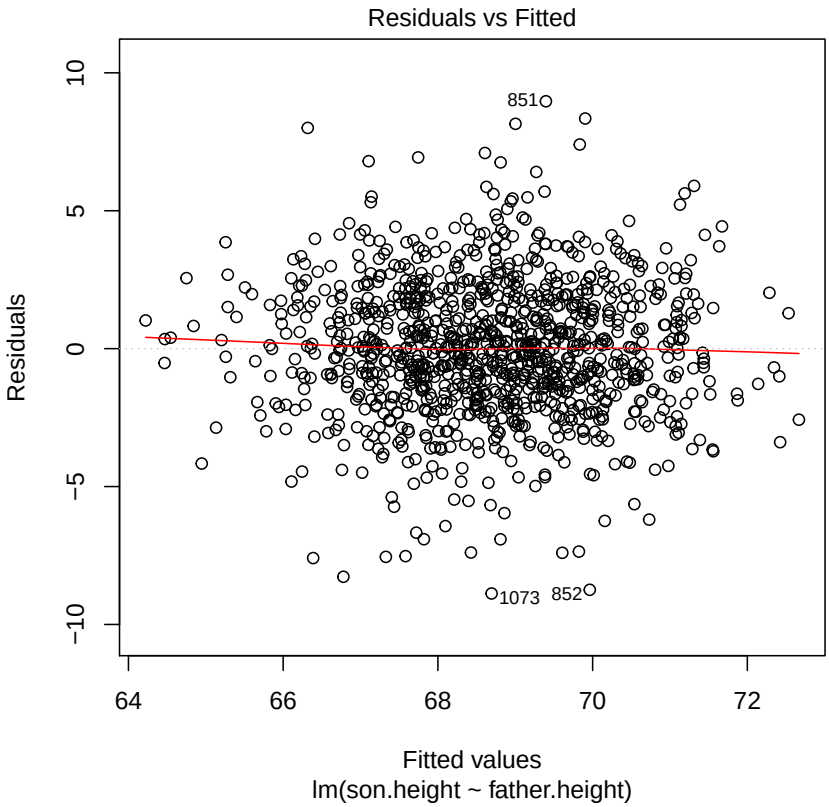

Figure 2:Resisual vs Fitted plot for the simple linear regerssion between son's height (`son_height`) and father's height (`father_height`).

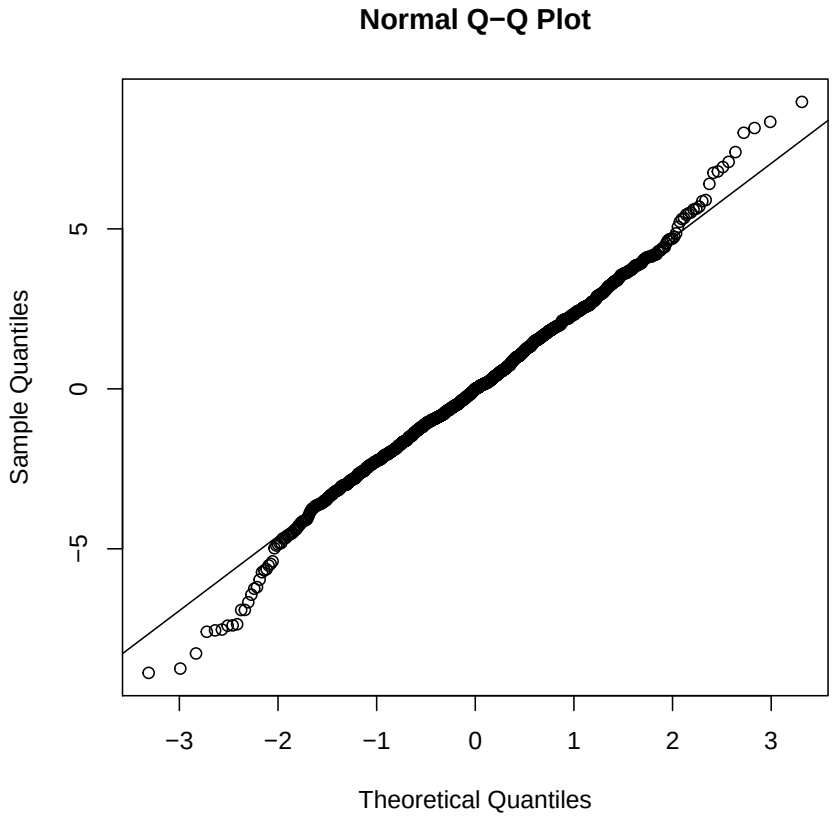

Figure 3:Q-Q plot for Resisuals of the simple linear regerssion between son's height (`son_height`) asnd father's height (`father_height`).

[Total: 19]

3. **One-way ANOVA in R**

   For full marks this answer must be typed in Word and plots and output included and captioned.

   Many studies have suggested that there is a link between exercise and healthy bones. It is suggested that exercise stresses the bones and this causes them to get stronger.

   One study examined the effect of jumping on the bone density of growing rats. The rats were randomly allocated to one of three treatments: a control with no jumping, a low-jump exercise, and a high-jump exercise. After 8 weeks of 10 jumps per day, for 5 days per week, the bone density of the rats (in mg/cm^3) was measured.

In this assignment question, we will use the `density.xlsx` dataset to look at how to perform one-way ANOVA in R. Download this dataset from MyUni and load it into R.

(a) A boxplot of the bone density for each exercise level is given in Figure 4. Compare the distribution for each group.

[4 marks]

- Shape: hard to definitely say from a boxplot, but Control and High Jump look symmetric, while Low Jump looks skewed. Mainly based on where the mediam lies in the box.
- Location: the median for High Jump is highest, then Low Jump and finally Control.
- Spread: Low Jump has the larger IQR.
- Outliers: Two outliers in the Control group. No outliers in the other groups.

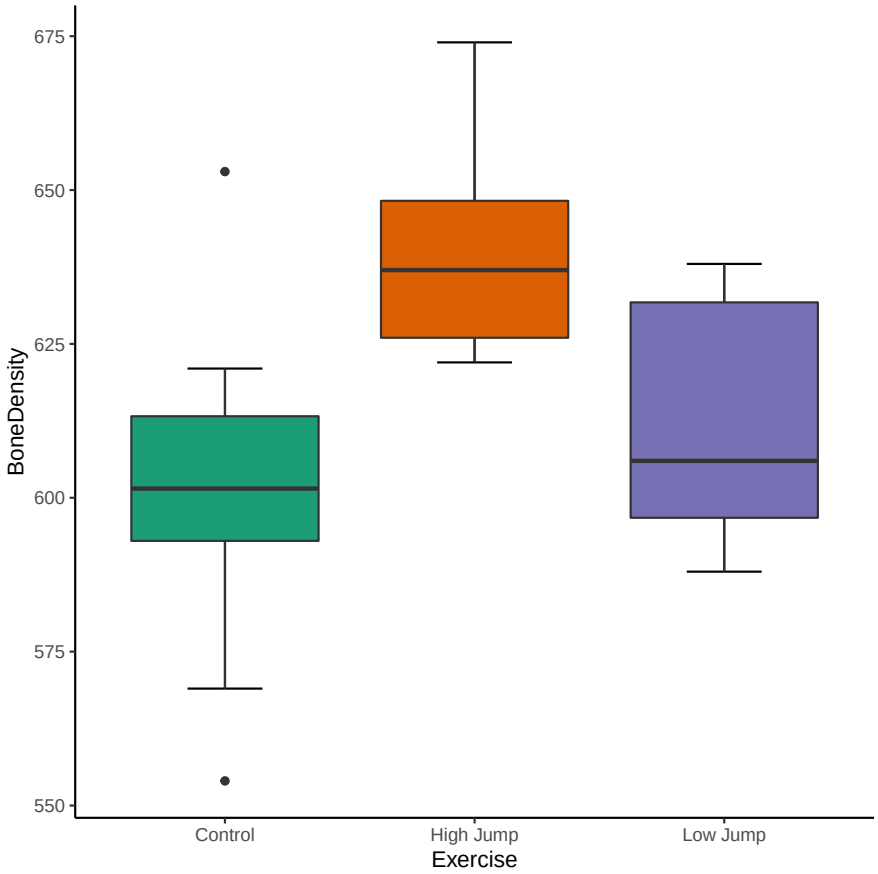

Figure 4:Boxplots of Bone Density for each Exercise for the density dataset.

(b) Use a one-way ANOVA to test for a significance difference between the mean bone density for each group with the following steps:

i. Write the appropriate null and alternative hypotheses. Remember to define all parameters used.

[2 marks]

$H_0 : \mu_1 = \mu_2 = \mu_3$,
H_a : at least one of the μ_is is different from the others,

where μ_1 is the mean bone density for rats in the control group, μ_2 is the mean bone density for rats in the high jump group, and μ_3 is the mean bone density for rats in the low jump group,

ii. Include the one-way ANOVA table in your assignment. Remember to caption it.

[1 mark]

The R ANOVA table is given in Table 3.

```
            Df Sum Sq Mean Sq F value Pr(>F)
Exercise     2   7434    3717   7.978 0.0019 **
Residuals   27  12579     466
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Table 3: R output for one-way ANOVA for Bone density according to Exercise.

iii. State the value of the observed test statistic.

[1 mark]

The value of the test statistic is 7.978.

iv. What is the distribution of the test statistic if the null hypothesis is true?

[2 marks]

The distribution of the test statistic if the null hypothesis is true is an F-distribution with 2 and 27 degrees of freedom.

v. State the P-value.

[1 mark]

The P-value is 0.002.

vi. Do you reject or retain the null hypothesis at the 5% significance level? Why?

[2 marks]

We reject the null hypothesis at the 5% significance level as the P-value is less than 0.05.

(c) Is the assumption of constant variance reasonable for this dataset? Remember to include any R output needed to support your conclusion.

[3 marks]

```
   Exercise BoneDensity
1   Control    27.36360
2 High Jump    16.59351
3  Low Jump    19.32902
```

Table 4: R output for sample standard deviations of Bone density for each Exercise group.

From Table 4, the largest standard deviation is 27.364, while the smallest standard deviation is 16.594 The ratio is 1.649 which is less than 2 and so the assumption is reasonable.

(d) In R, produce a multiple comparisons table using a Bonferroni adjustment. Using this table, which exercises are significantly different at the 5% significance level. Remember to include and caption your table.

[3 marks]

```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = BoneDensity ~ Exercise, data = density)

$Exercise
                     diff       lwr       upr     p adj
High Jump-Control    37.6  13.66604 61.533957 0.0016388
Low Jump-Control     11.4 -12.53396 35.333957 0.4744032
Low Jump-High Jump -26.2 -50.13396 -2.266043 0.0297843
```

Table 5: R output of the multiple comparison for the one-way ANOVA.

The multiple comparison table for the one-way ANOVA is given in Table 5. From this we can see that

- Control and high jump are significantly different,
- High jump and low jump are significantly different.

[Total: 19]

Presentation marks

Marks for use of word and captions for all figures and all tables.

[3 marks]

[Total: 3]

[[Assignment total: 49]]