The University of Hong Kong
School of Public Health

CMED6100/MMPH6002/CMED7100
Introduction to Biostatistics (Semester 1)
Practical 2 Suggested solution

Linear regression and correlation

Question 1

a) What is the Pearson correlation between height and weight at 7 years of age?

[Statistics → Summaries → Correlation test (Variables: hei7, wei7; Types of correlation: Pearson product moment; Alternative hypothesis: Two-sided)]
Answer: 0.739

b) Fit a linear regression model to predict the weight from the height at 7 years of age. What are the intercept and slope estimates?

[Statistics → Fit models → Linear regression (Response variable = wei7, Explanatory variables = hei7)]
Answer: Intercept estimate = -50.48, Slope estimate = 0.612

c) What is the value of R^2 in the regression model in (b)?
Multiple R-squared in the output in b)
Answer: 0.546

d) What is the relationship between the answer in (a) and (c)?

Answer: The square of Pearson correlation coefficient = R^2 or R = Pearson correlation coefficient

e) Calculate the variance of weight at 7 years of age.

[Statistics → Summaries → Numerical Summaries (Variables: wei7; check "standard deviation" in Options)]
Answer: 4.436749*4.436749 = 19.685

f) What is the relationship among the three quantities: (i) between-group sum of squares, (ii) within-group sum of squares, and (iii) total sum of squares?

Answer: Between-group sum of squares + Within-group sum of squares = Total sum of squares

g) Produce an ANOVA table of the regression model obtained in (b). What is the total sum of squares? Divide it by n – 1 where n is the sample size. How does this number relate to the answer in (e)?

[Models → Hypothesis tests → ANOVA table (Type of tests: Sequential (Type I); Sandwich Estimator: HC3)]

Answer: Total sum of squares = 9657.2+8039.4 = 17696.6; Divide it by (900 – 1) = 19.685 = Variance of weight

h) From the ANOVA table, what is the within-group sum of squares? What is the interpretation of this number?

Answer: Within-group sum of squares = 8039.4; it is the sum of the squared difference between the observed weight and the predicted weight by the regression model.

i) Divide the between-group sum of squares by the total sum of squares. How does this number relate to the answer in (c)? What is the interpretation of R^2?

Answer: 9657.2 / 17696.6 = 0.546 = R^2; R^2 is the proportion of variation or the proportion of sum of squares explained by the regression model.

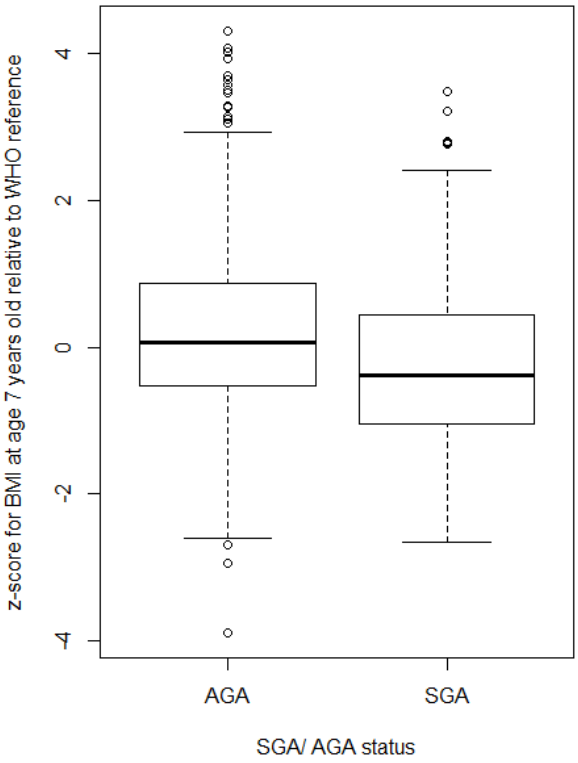
Comparing two means by t-test

Question 2

This question evaluates whether there is a difference in the BMI z-score between those children classified as AGA and those classified as SGA, by using an independent two sample t-test.

a) By using R Commander, draw the boxplot of the BMI z-score for children classified as SGA and AGA.

[Graphs → Boxplot → (Variable: bz7; Plot by groups: sga; Identify outliers: No; x-axis label: SGA/ AGA status; y-axis label: z-score for BMI at age 7 years old relative to WHO reference)]
Or you may refer to the code provided in Practical 1 solution to draw boxplots with colour change and axes drawn manually instead of using default ones.



b) What are the mean BMI z-scores and the corresponding standard deviations of the mean BMI z-scores for those children classified as AGA and those children classified as SGA? Construct the 95% confidence intervals for the mean BMIs for SGAs and AGAs.

[Statistics → Summaries → Numerical summaries (Variable: bz7; Summarize by groups: sga; Statistics: Mean, Standard deviation, Standard error of Mean)]

Confidence interval = Mean ± 1.96*SE (3 d.p.)

Characteristic	Mean	Standard deviation	95% Confidence interval
SGA	-0.199	1.201	(-0.334, -0.064)
AGA	0.221	1.166	(0.127, 0.315)

We want to determine whether there is a difference in the BMI z-scores between those children classified as AGA and those classified as SGA, using a two-sample t-test.

c) State the null hypothesis.

H_0: There is no difference between the mean BMI z-score for children in SGA versus AGA status

d) Evaluate whether the variances of BMIs of the AGAs and SGAs are equal or not. Explain your results based on the Levene's test of equality of variances.

[Statistics → Variances→ Levene's Test (Response Variable(s): bz7; Factors: sga; Center: Mean)]

H_0: The variances of BMI z-scores of the children in AGA and SGA status are equal.

Levene's Test Statistics	p-value
0.566	0.452

e) What are the t-value, the degree of freedom and the p-value?

[Statistics → Means→ Independent Samples T-test (Response Variable(s): bz7; Groups: sga; Options: Assume equal variances? Yes)]

t-value	Degree of freedom	p-value
5.05	898	<0.001

f) By using a 0.05 level of significance, state your conclusion based on your findings by using two sample t-test?

The mean BMI z-score of children between AGA and SGA status were statistically different at 5% level of significance.

Alternatively, we can use one-way ANOVA to determine whether there is a difference in the BMI between those children classified as AGA and those classified as SGA.

g) What are the F ratio and degree of freedom in the ANOVA? What is the p-value?

[Statistics → Means → One-Way ANOVA (Response variable: bz7; Groups: sga)]

F ratio	Degree of freedom	p-value
25.5	1; 898	<0.001

h) Present the p-values in (f) and (g) in 10 decimal places and compare them. What is your finding?

Both p-values are 0.000000534. The two tests are equivalent when there are two groups for comparison.