Problem Set 3: Multiple linear regression

Introduction to Statistical Thinking and Data Analysis

MSc in Epidemiology and MSc in Health Data Analytics, Imperial College London

24 October 2022

A) Consolidating concepts

A1. Which of the following regression models tests the null hypothesis that there is no association between BMI and age (in years), after controlling for gender:

a) $\text{BMI} = \beta_0 + \beta_1 \times \text{age} \times \text{gender}$
b) $\text{BMI} = \beta_0 + \beta_1 \times \text{age} + \text{gender}$
c) $\text{BMI} = \beta_0 + \beta_1 \times \text{age} + \beta_2 \times \text{gender}$
d) $\text{BMI} = \beta_0 + \beta_1 \times \text{age} + \beta_2 \times \text{gender} + \beta_3 \times \text{gender} \times \text{age}$

A2. Having fit the model, the estimated regression coefficients are:

Intercept = 21.43 (95% CI, 21.17-21.74)
Age = 0.14 (95% CI, 0.13 - 0.14)
Gender Male = 0.037 (95% CI -0.23 - 0.31)

Which of the following is correct?

a) After controlling for the association of age and BMI, there is a significant relationship between gender and BMI.
b) There is no relationship between BMI and age in our sample, having adjusted for gender.
c) For woman in our sample, an increase of age by 1 year is associated with an increase in BMI of 0.14.
d) We have evidence to suggest that the relationship between BMI and age is different for men and woman.

A3. In our sample, what BMI would be predicted for a 45 year old woman?

a) 23.13
b) 22.79
c) 27.77
d) 27.73

A4. If the correlation coefficient is 0.8, the percentage of variation in the response variable explained by the variation in the explanatory variable is

a) 2%
b) 80%
c) 64%
d) 20%

A5. When is it appropriate to include an interaction term in a regression model?

a) When an independent variable has a different effect on the outcome depending on the values of another independent variable.
b) When an independent variable influences both the dependent variable and independent variable.
c) Both of the above.
d) Neither of the above.

A6. Suppose there is a clinical trial investigating which dose of a new drug is optimal to lower blood pressure in men and women. In which scenario might we fit an interaction term between gender and dose?

a) If there is no difference in blood pressure across doses by gender.
b) If there a constant difference in blood pressure across doses by gender.
c) If the effect of the dose on blood pressure depends on gender.
d) Not enough information.

A7. Which regression equation corresponds to a model investigating blood pressure by drug dose and gender, assuming there is an interaction between drug dose and gender:

a) blood pressure $= \beta_0 + \beta_1 \times \text{dose} + \beta_2 \times \text{sex} + \beta_3 \times \text{dose} \times \text{sex}$
b) blood pressure $= \beta_0 + \beta_1 \times \text{dose} + \beta_2 \times \text{sex}$
c) blood pressure $= \beta_0 + \beta_1 \times \text{dose} \times \text{sex}$
d) blood pressure $= \beta_1 \times \text{dose} + \beta_2 \times \text{sex} + \beta_3 \times \text{dose} \times \text{sex}$

A8. Which violation of the assumptions underlying linear regression could be addressed by including a logarithmic transformation of one or more of the variables in a regression model?

a) Unequal standard deviations.
b) Positively skewed distribution.
c) Non-linear relationship between the outcome and exposure variable.
d) All of the above.

A9. How should exposure variables be chosen when building a regression model?

a) As many as possible to best explain the relationship.
b) Reviewing the literature and using external knowledge.
c) Based on statistical significance.
d) Stepwise regression.

B) Practicing skills

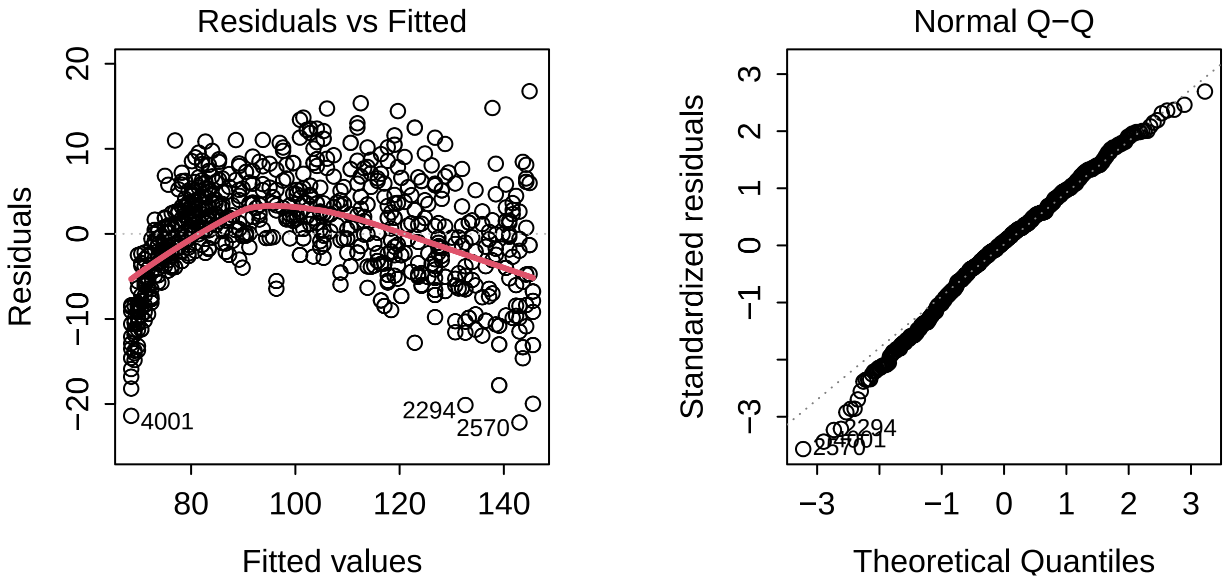
B1. Problem B2 of Problem Set 2 used the NHANES dataset to study child growth by estimating the relationship between height in centimeters age in months for children aged zero to 10 years (less than 120 months).

Residual analysis demonstrated that the linear regression assumptions were violated, particularly the assumptions of linearity and homoscedasticity (constant variance).

```
library(NHANES)
data(NHANES)
nhanes_child <- subset(NHANES, AgeMonths < 120)

nhanes_child$height_all <- ifelse(nhanes_child$AgeMonths < 24,
                                  nhanes_child$Length,
                                  nhanes_child$Height)
fit2 <- lm(height_all ~ AgeMonths, nhanes_child)
```

```
par(mfrow = c(1,2), tcl = -0.25, mgp = c(2, 0.5, 0))
plot(fit2, 1, lwd = 3); plot(fit2, 2)
```



a. Consider transformations for both your outcome variable height and predictor variable age in months in order to more satisfactorily meet the linear regression assumptions.

b. Interpret the parameter estimates of your final model from part (a).

B2. The data set `england_gcse.csv` contains data on average GCSE exam results for 3073 state run secondary schools in England for the year 2015-2016, and other information about the schools including whether the school has a selective admissions process, is single- or mixed-gender education, the percentage of students qualifies for free school meals, the percentage of the students who are girls, and the percentage of students for whom English is a second language.

Variable	Description
id	School ID number
admissions	Admissions policy: `comprehensive` or `selective`
entry_gender	Single- or mixed-gender school: `mixed`, `girls`, or `boys`
perc_free_meals	Percentage of students qualifying for free school meals
perc_girls	Percentage of female students
perc_sec_lang	Percentage of students for whom English is a second language
avg_score	Average GCSE exam score

a. What is the target population to which the sample is generalisable?

b. Describe the sample, including appropriate univariate summaries of each of the variables and bivariate associations among the variables.

c. Are average exam scores different for selective admissions schools than for comprehensive admissions schools? Do differences in socioeconomic backgrounds of students, measured by the percentage eligible for free school meals, explain differences in exam scores?

d. Using your model from part (c), what is the predicted difference in average exam scores for a selective school in which 3% of students are eligible for free meals and a comprehensive school in which 15% of students are eligible for free school meals?

e. Is the difference in exam scores for selective admissions schools different for single-gender schools than for mixed gender schools?

f. Fit a model to estimate how average test scores change with the percentage of female students and interpret the results. Then fit a model with percentage of female students, adjusted for whether the school is single- or mixed-gender, and a third model with an interaction between the percentage of female students and whether the school is single- or mixed-gender. How do the estimates change? How do you interpret each estimate?

C) Advanced learning: robust standard errors

Note: These questions are challenging and requires some R tools above a beginner's level and above the level expected to be successful in ISTDA. Make an effort to complete these questions if you would like to challenge your understanding of regression assumptions but do not be discouraged if you are not able to complete them.

C1. In Problem Set 2 Exercise B2, we found that regressing child height on age did not satisify the assumptions of linear regression. In exercise B1 above, we used transformations to develop a regression model that better satisfied the assumptions. Even though the linear regression between height and age in months did not satisfy the linear regression assumptions, we might nonetheless be interested in estimating the best-fitting linear trend between these variables, and estimate a standard error for this slope. This is called a *Generalized Estimating Equation* (GEE) approach.

Section 12.2 of Kirkwood and Sterne describes that one option to address deviations from the linear regression assumptions is to derive alternative standard errors that relax these assumptions. One such approach is to derive standard errors that rely solely on the large sample properties of the Central Limit Theorem, but do not require an uncorrelated normal distribution for the model residuals. These are typically referred to as *robust standard errors* because they are robust to the assumptions of least-squares linear regression.

The R package **`sandwich`** implements calculation of robust parameter covariance matrix. (The name *sandwich* comes from the form of the mathematical formula used to estimate the covariance matrix.). The function below wraps the key function `sandwich::vcovHC(...)` to produce 95% confidence intervals for the regression coefficients based on the robust standard errors.

```
## install.packages("sandwich")
library(sandwich)

confint_robust <- function(object, parm, level = 0.95, type = "HC3") {

  cf <- coef(object)
  vcov <- sandwich::vcovHC(object, type = type)

   pnames <- names(cf)
   if (missing(parm))
       parm <- pnames
   else if (is.numeric(parm))
       parm <- pnames[parm]
   a <- (1 - level)/2
   a <- c(a, 1 - a)
   pct <- stats:::format.perc(a, 3)
   fac <- qnorm(a)
   ci <- array(NA, dim = c(length(parm), 2L), dimnames = list(parm, pct))
   ses <- sqrt(diag(vcov))[parm]
   ci[] <- cf[parm] + ses %o% fac
   ci
}
```

Load the `confint_robust()` function into R and use it to calculate robust 95% CIs for the model you estimated in part (b).

How do the 95% CIs compare those estimated by the base `confint()` function? Intuitively, can you explain why you think this is the case?

C2. *Consequences of violating regression assumptions.* In exercise B2 from Problem Set 2, we demonstrated that regressing child height on age in months did not satisify several of the linear regression assumptions. In this exercise, we will use simulation to explore the consequences of this for our statistical inference.

Similarly to exercise C1 in Problem Set 2, this exercise will use the `nhanes_child` dataset constructed in exercise 2 as a 'true' population from which to simulate smaller samples and study the properties of statistical inference for linear regression coefficient estimates. Consider the regression parameter estimates using the full dataset in exercise 2(b) as true population values for β_0 and β_1.

Do the following steps *(Note: a., b., and c. are the same as C2. last week; d. and e. are additions building on the robust standard errors introduced in C1.)*:

a. Simulate smaller datasets by sampling rows from the `nhanes_child` dataset. Sample a large number of datasets with replacement of size 10, 25, 50, 100, and 500. Since both the outcome height and covariate AgeMonths are needed, entire rows must be resampled rather than simply resampling values from a vector. Instead of using the function `sample()`, use `sample.int()` to randomly sample rows to retain, and then subset the data frame to only these rows. Example code for generating one simulated dataset:

```
maxrow <- nrow(nhanes_child)
n <- <sample sizze>
df <- nhanes_child[sample.int(maxrow, n, replace = TRUE), ]
```

b. For each simulated dataset, fit a linear regression model for the combined height outcome and AgeMonths as the linear predictor (the same regression model estimated in 2(b)). Extract the $\hat{\beta}_1$ coefficient estimate for the slope associated with age. For each sample size, plot a histogram of the distribution of $\hat{\beta}_1$ estimates and compare them to the true best population value (the coefficient estimated in part (b) of Problem Set 2, exercise B2). What do you observe about the distribution of $\hat{\beta}_1$.

c. For the regression model fitted to each simulated dataset, extract the 95% confidence interval using the `confint()` function. For each simulated sample size, calculate the proportion of 95% confidence intervals that contain the true value of β_1. How does this change with sample size relative to the nominal 95% coverage target?

d. Repeat your confidence interval coverage calculation from part (c) using the `confint_robust()` function from question C1. For each sample size, compare the coverage of the robust confidence intervals with the coverage of the least-squares regression confidence intervals from part (c). Is this what you you expected?

e. For a bit of bonus fun, replicate this simulation study investigating a linear regression between the three outcome variables studied last week (Height, BMI, and AlcoholYear) and adult age as a covariate.