Problem Set 2: Linear regression

Introduction to Statistical Thinking and Data Analysis

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

17 October 2022

A) Consolidating concepts

A1. A regression analysis of the outcome blood pressure (measured in units of millimeters of mercury(mmHg)) as a function of weight (measured in kilograms(kg)) resulted in the following regression coefficients: $\beta_0 = 102.25$, $\beta_1 = 0.20$.

This implies that, in our sample:

y = mmHg, x = kg
beta1 = change in y per unit increase in x

a) for each 1kg increase in weight, blood pressure increased by 102.25mmHg.

b) for each 1mmHg increase in blood pressure, weight increased by 0.20kg.

c) for each 0.2kg increase in wight, blood pressure increased by 102.25mmHg.

d) for each 1kg in weight, blood pressure increased by 0.20mmHg.

A2. A residual is defined as:

a) The difference between the actual y values and the mean of y.

b) The difference between the actual y values and the predicted y values.

c) The predicted value of y for the average x value.

d) The square root of the slope.

A3. Regardless of the value of x, the standard deviation of the distribution of y values about the regression line is the same. This assumption of equal standard deviations of residuals about the regression line is called:

a) random error.

b) heteroscedasticity.

c) homoscedasticity.

d) residual analysis.

A4. In the regression equation $y = beta0 + beta1\ x$, beta0 is the:

a) slope of the line.

b) independent variable.

c) y intercept.

d) coefficient of determination.

A5. The equation of a regression line which describes the association of body weight (kg) and plasma volume (litres) of 8 healthy men is: *plasma volume = 0.0857 + 0.0436 * weight.* For a man weighing 85kg, what would you expect their plasma volume to be?

a) 3.79 litres.

b) 2.70 litres.

c) 3.71 litres.

d) 3.00 litres.

A6. Least square regression calculates the best fitting line for the observed data by minimising the sum of the squares of the _____ deviations.

a) Vertical.

b) Horizontal.

c) Both of these.

d) None of these.

A7. A simple linear regression is fit to a dataset of 10 samples. How many degrees of freedom are there?

a) 10

parameters estimated = beta0 and beta1
thus, df = n-p = 10-2 = 8

b) 9

c) 8

d) 7

A8. Which of the following is not an assumption underlying linear regression between outcome variable y and exposure variable x:

~~a) There is a linear relationship between x and y.~~

~~b) The residuals are independent of each other.~~

c) For any value of x, y is normally distributed.

d) The outcome variable is binary.

A9. The correlation coefficient allows us to measure the strength of association between an outcome and exposure variable. Which of the following is true of the correlation coefficient?

a) A correlation coefficient = 100 indicates perfect positive correlation.

b) The correlation coefficient is independent of the beta1 regression coefficient.

c) The correlation coefficient is the number of standard deviations that the outcome variable changes by for a standard deviation change in the exposure variable.

~~d) A correlation coefficient = 0 indicates negative correlation.~~

A10. Which of the following cannot be used to test for departures from normality?

a) Skewness and Kurtosis.

b) Shapiro-Wilk test.

c) Inverse normal plots.

d) Logarithmic transformation.

B) Practicing skills

B1. Revisit the `perulung_ems.csv` dataset from the week 1 tutorial sheet, introduced on page 27 of Kirkwood and Sterne.

Variable	Description
id	Participant ID number
fev1	Forced Expiratory Volume in 1 second
age	Age in years
height	Height in centimeters
sex	Sex (0 = female, 1 = male)
respsymptoms	Presence of respiratory symptoms (0 = no symptoms; 1 = symptoms)

a. Calculate parameter estimates for a linear regression of the outcome FEV1 as a function of height, and the associated standard errors, t-statistics, and p-values using the formulas from Chapter 10 of Kirkwood and Sterne. Plot the fit of your model compared to the data. Check that your calculations agree with the outputs of the `lm(...)` function in R.

b. Interpret the regression coefficients. State the null and alternative hypothesis that is being tested with the p-value reported for the height coefficient and your assessment of this hypothesis.

c. State and check the assumptions of your regression model—do each of them appear to be satsified?

d. Convert height into a categorical three groups: height below 120cm, between 120 to 130cm, and above 130cm. Refit your regression model for FEV1 this time using the height categories as your predictor. Interpret the regression coefficient estimates. *Hint: The function* `cut()` *will convert a numerical variable to a categorical variable.*

e. When analysing categorical predictors, it is often conventional to use the group with greatest frequency as the reference category (though there are other reasonable choices depending on the analysis). Change the reference category for your categorical height variable so that it is the group with the largest number of observations and refit your model. How do the parameter estimates change? What is the interpretation of the new parameters? *Hint: The functions* `levels()` *and* `relevel()` *will be helpful.*

f. Amongst the models estimated in parts (a), (d), and (e), which do you prefer and why?

g. Fit a linear regression model to test the null hypothesis that FEV1 is not associated with respiratory symptoms. What is your conclusions about this hypothesis? How do your effect estimates, statistical inference, and conclusions compare to what you estimated using the equal variance t-test last week?

B2. In this exercise we will use the NHANES dataset to study child growth by estimating the relationship between height in centimeters and age in months for children aged zero to 10 years (less than 120 months).

Age in months is only reported for children of all ages in the 2009 to 2010 data. For 2011 and 2012, age in months is available only for children aged 0 to 2 years. This is fine for the purposes of our analysis; we will retain the subset of data for which age in months is recorded and below 120 months:

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

Height is measured in two different ways depending on the child age. For children aged 2 and older, standing height is measured and recorded in the variable `Height`. For children aged 0 to 3 years, recumbent length is measured and recorded in the variable `Length`.

a. For children aged 24 to 47 months, both standing height (cm) and recumbent length (cm) were measured. Are standing height and recumbent length equivalent measures of height? Articulate, conduct, and report a hypothesis test to assess this question and report the estimated magnitude of any difference.

b. Proceed with your analysis assuming recumbent length is a reasonably good approximation for height for children under age 2. Create a single height variable that takes the variable `Length` for children under age 24 months and the variable `Height` for children age 24 to 119 months. Fit a linear regression model to estimate the relationship between age in months and height and give interpretations of both parameters. *Hint: Try the function* `ifelse()`.

c. Check the assumptions of your regression model fitted in part (b). Do any of the assumptions appear to be violated?

In week 3 we will look at how transforming variables and robust standard errors can help overcome violated regression assumptions.

C) Advanced learning

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 the central limit theorem, confidence intervals and regression assumptions but do not be discouraged if you are not able to complete them.

C1. *Distribution of the sample mean and confidence intervals.* This exercise will use the NHANES dataset as 'true' population from which to simulate samples and study the properties of the sample mean as an estimator of the population mean. As with last week, we will only consider the subset of the sample who are adults aged 20 years and older. First, lets revisit the the three variables used in question B3 from last weeks problem set:

- Height: Standing height in centimeters.
- BMI: Body Mass Index
- AlcoholYear: Number of days over the past year that participant drank alcoholic beverages

Next, create a vector of responses removing any `NA` values.

```
nhanes20pl <- NHANES[NHANES$Age >= 20, ]
   height <- nhanes20pl$Height[!is.na(nhanes20pl$Height)]
   bmi <- nhanes20pl$BMI[!is.na(nhanes20pl$BMI)]
   alcohol <- nhanes20pl$AlcoholYear[!is.na(nhanes20pl$AlcoholYear)]
```

Do the following steps for each variable:

a. Simulate a large number of samples with replacement of size 5, 10, 25, 50, 100, and 500. See the function `?sample`. (The function `?replicate` might also be helpful for simulating a large number of samples.)

b. For each sample, calculate the sample mean, the large-sample 95% CI using the normal distribution, and the small-sample 95% CI using the t-distribution. (The function `?apply` might be useful.)

c. Plot the distribution of the sample mean for each of the increasing sample size. What do you notice about the distribution of the sample mean with increasing sample size for each variable?

d. Calculate the coverage of the 95% confidence intervals by calculating for what proportion of samples the 95% confidence interval contains the true population means (calculated in exercise B2 (b) in problem set 1, last week:

	mean	sd
height	168.8	10.1
bmi	28.8	6.7
alcohol_year	75.7	103.6

Create a figure to visually summarise the analysis of 95% CI coverage at different sample sizes, for the large-sample and small-sample CI, and for different variables.

i. How do the coverage of the large-sample CI (using normal distribution) and small-sample CI (using t-distribution) compare at different sample sizes?
ii. For a given sample size, how does the coverage compare for each of the three variables? Can you relate this to your findings in exercise B3 (d) from week 1?

C2). *Consequences of violating regression assumptions.* In exercise B2, we demonstrated that regressing child height on age in months did not satisfy 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 above, this exercise will use the `nhanes_child` dataset 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 B2 (b). as true population values for β_0 and β_1.

Do the following steps:

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

b. For each simulated dataset, fit a linear regression model for the height outcome constructed in exercise B2(b) and AgeMonths as the linear predictor (the same regression model estimated in B2(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 exercise B2 (b). 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?