Coursework: EDA & Regression

Mark Muldoon <mark.muldoon@manchester.ac.uk> and
Diego Perez Ruiz <diego.perezruiz@manchester.ac.uk>

29 October – 12 November 2022

The coursework involves a dataset, `PimaDiabetes.csv`, derived from one originally collected by the USA's National Institute of Diabetes and Digestive and Kidney Diseases[1]. It lists various diagnostic measures recorded from 750 women along with a 0/1 variable, `Outcome`, that indicates whether the person eventually tested positive for diabetes. Table 1 shows the first few rows of the dataset while the diagnostic measures are explained below.

Pregnancies: number of times the woman has been pregnant

Glucose: plasma glucose concentration (mg/dl) at 2 hours in an oral glucose tolerance test (OGTT)

Blood Pressure: Diastolic blood pressure (mm Hg)

Skin Thickness: Triceps skin fold thickness (mm)

Serum Insulin: insulin concentration[2] (μ U/ml) at 2 hours in an OGTT

BMI: body mass index (weight in kg)/(height in m)^2

Diabete Pedigree: a numerical score designed to measure the genetic influence of both the woman's diabetic and her non-diabetic relatives on diabetes risk: higher scores mean higher risk. You can read more about this in Smith, Everhart, Dickson, Knowler, and Johannes (1988).

Age: in years

Outcome: 1 if the woman eventually tested positive for diabetes, zero otherwise

[1] You can read about the data in Smith *et al.* (1988), Using the ADAP Learning Algorithm to Forecast the Onset of Diabetes Mellitus, *Proceedings of the Annual Symposium on Computer Application in Medical Care*, 261–265.

[2] It's unclear from Smith *et al.* which units are being used here. Insulin concentration is sometimes reported in terms of *international units*, which measure biological activity rather than amount of molecules, but the concentrations reported here seem too high, as Wikipedia states (based on Iwase, Kobayashi, Nakajima, and Takatori (2001)) that "A typical blood level between meals is 8–11 μIU/mL".

Pregnancies	Glucose	Blood Pressure	Skin Thickness	Insulin	BMI	Diabetes Pedigree	Age	Outcome
6	148	72	35	0	33.6	0.627	50	1
1	85	66	29	0	26.6	0.351	31	0
8	183	64	0	0	23.3	0.672	32	1
1	89	66	23	94	28.1	0.167	21	0
0	137	40	35	168	43.1	2.288	33	1

Table 1: The first five rows of data in `PimaDiabetes.csv`

Prepare a 1000 word report that summarises your work on the following exercises.

1. Write a brief description of the data, including its origin and quality issues. You should imagine you are writing for a group who have no idea what this dataset is about. *[2 marks]*
2. Do an exploratory data analysis. *[4 marks]*
3. Add a column, `ThreeOrMoreKids`, to the dataset that answers the question "Does the woman have 3 or more children?", then fit an appropriate regression model to predict whether a woman will develop diabetes using `ThreeOrMoreKids` as a single predictor. With the help of the fitted model, answer the following questions (show your calculations, either by hand or with help of R or Python): *[5 marks]*
   - What is the probability that you get diabetes, given that you have two or fewer children?
   - What is the probability that you get diabetes, given that you have three or more children?
4. Using the data in `PimaDiabetes.csv`, fit appropriate regression models and use them to determine how likely the women whose data are listed in Table 2 are to develop diabetes. You are free to choose which explanatory variables to inclue in your model and may, if you like, compare several models, but make sure that you clearly state the final model chosen and the reasons behind this choice. With the help of your chosen model, interpret the results in terms of probability of developing diabetes (as you did for the model based on `ThreeOrMoreKids`). *[7 marks]*
5. Include R or Python code used to produce the analysis. *[2 marks]*

Illustrate your analysis with appropriate figures and tables. Figure and table captions, the contents of tables and your code do not count against the word limit.

Due Date: 17:00 on 12 November 2022, uploaded to BlackBoard as a PDF. Also note:

- We want to mark your work anonymously, so please don't include your name in your report. Instead, label it with your student ID number.
- Although there is no minimum or maximum number of references required, you should reference any sources (except for materials from this course) that you use when developing your code or preparing your report. The list of references should come at the end of the report and does not count against the word limit.

		Blood	Skin			Diabetes	
Pregnancies	Glucose	Pressure	Thickness	Insulin	BMI	Pedigree	Age
4	136	70	0	0	31.2	1.182	22
1	121	78	39	74	39	0.261	28
3	108	62	24	0	26	0.223	25
0	181	88	44	510	43.3	0.222	26
8	154	78	32	0	32.4	0.443	45

Table 2: Diagnostic measures for the women whose `Outcome` you should predict. These values are available in `ToPredict.csv`.

References

Iwase, H., Kobayashi, M., Nakajima, M., & Takatori, T. (2001). The ratio of insulin to C-peptide can be used to make a forensic diagnosis of exogenous insulin overdosage. *Forensic Science International*, *115*(1), 123-127. doi: 10.1016/S0379-0738(00)00298-X

Smith, J. W., Everhart, J., Dickson, W., Knowler, W., & Johannes, R. (1988). Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. In *Proceedings of the annual symposium on computer application in medical care* (pp. 261–265).