institute of health informatics

# Graduate Programme in Health Data Science

Assessed Coursework Submission

|  |  |
| --- | --- |
| **Student candidate number:** | LSRY9 |
| **Module:** | CHME0040: Data Science & Statistics |
| **Date due:** | 15 January 2025, 16:00 GMT |
| **Word count:** (excluding references, diagrams and appendices) | 2154 words total, Part A: 1271 words, Part B: 879 words |
| **Disability or other medical condition** for which UCL has granted special examination arrangements: |  |
| **Formative feedback:** | Please address in formative feedback: |
|  |
| Please ignore in formative feedback: |
|  |

**Part A: Introduction**

Diabetes is a chronic condition marked by elevated blood glucose levels (Diabetes 2025). In this study, a logistic regression model is applied to predict the presence of diabetes using a simulated numerical dataset. Logistic regression is well-suited for binary outcomes, making it an effective method for distinguishing between individuals with and without diabetes based on predictor variables. The model’s goal is to provide an accurate classification of diabetes status, leveraging various predictors from the dataset. This model can be used to demonstrate its effectiveness in binary prediction tasks within healthcare-related datasets.

**Variables**

The dataset consisted of ten variables: patient ID, pregnancies, glucose, blood pressure, skin thickness, insulin, BMI, diabetes genetic score, and age, as well as the presence of diabetes within the patient. Out of these ten, eight variables were incorporated into the model as independent predictors, while diabetes served as the dependent variable to be predicted. The patient ID was excluded from the analysis as it functions solely as a unique identifier and does not hold any predictive value or contribute meaningful information to the model’s performance. All values within the dataset were retained in their original form, and no data points were excluded, as there were no missing values in this data set. Given the vastness of the dataset, all recorded values, including extreme cases, were considered potentially accurate and logical under specific conditions. Outliers present in the data did not negatively impact the model’s performance in any way. To summarize the central tendency of the dataset, the mean was used for each variable. Although the mean can be influenced by outliers, the presence of these extreme values did not significantly skew the mean to a degree that justified their removal. The dispersion of the data was assessed through measures such as range and standard deviation, which provided insights into the consistency of the data and highlighted any potential outliers. The variables identified as having notable outliers were skin thickness, BMI, and insulin levels. While these values appeared unusually high, they were deemed reasonable and feasible within the context of extreme scenarios. Thus, these values were retained to preserve the dataset’s integrity and ensure that the model could account for a wide range of possible outcomes.

**Model**

For this analysis, a logistic regression model was employed to predict the presence of diabetes in the dataset. Logistic regression is a robust and widely used statistical technique that models the relationship between one or more independent variables and a binary dependent variable. In this case, the dependent variable is the presence or absence of diabetes, coded as a binary outcome. This model is particularly suitable for such binary outcomes because it generates probabilities, offering valuable insights into the likelihood of a patient developing diabetes based on the given predictors. Logistic regression assumes a linear relationship between the predictors and the log-odds of the outcome. It models the probability of the event (in this case, diabetes) occurring as a function of the independent variables, which include various factors such as glucose levels, age, BMI, insulin, and others. The model calculates a regression coefficient for each independent variable, and these coefficients represent the influence of each variable on the log-odds of having diabetes. A higher coefficient value indicates a stronger relationship between the predictor and the outcome, meaning that as the value of the predictor increases, so does the likelihood of the outcome to occur.

A graph of a curve

Description automatically generated

The model’s predictive performance was evaluated using receiver operating characteristic (ROC) curves. ROC curves are an essential tool for assessing the diagnostic ability of binary classification models. They graphically display the trade-off between the true positive rate (sensitivity) and the false positive rate (1-specificity). The area under the ROC curve (AUC) is a summary metric that quantifies the model’s ability to discriminate between the two groups—those with diabetes and those without. AUC values closer to 1.0 signify better model performance. In this analysis, the AUC values were 0.845 and 0.840, indicating that the model performs well in distinguishing between individuals with and without diabetes. These values suggest that the model is effective at predicting the presence or absence of diabetes, though there may be room for further refinement, particularly in reducing false positives and improving sensitivity. Additionally, the model’s performance was evaluated through sensitivity and specificity analysis, which were derived from the ROC curve. Sensitivity measures the proportion of true positives identified by the model, whereas specificity measures the proportion of true negatives. These metrics are particularly critical in healthcare, where the consequences of misclassifying patients can be significant. Maximizing sensitivity while minimizing false positives is essential to avoid incorrect diagnoses and unnecessary treatments. The ROC curve and AUC metrics provide a balanced view of the model’s performance, demonstrating its credibility and robustness for diabetes prediction.

**Sensitivity Analyses**

A graph of a number of probabilities

Description automatically generatedA graph of a number of probabilises

Description automatically generated

To enhance and validate the robustness of the model, the data was divided into ten deciles based on the presence or absence of diabetes. This approach allows for an assessment of whether the model is biased or underperforming in specific areas. In the plot above, the sum of M2PR represents the total predicted probabilities within each decile. By comparing the sum of actual diabetes cases to the sum of predicted probabilities (M2PR), it can be evaluated how well the model’s predictions align with the actual outcomes across each decile. This comparison helps to assess the model’s performance and identify any discrepancies between predicted and actual results.

**Results**

The logistic regression model was applied to predict the presence of diabetes for each individual, using a variety of predictors such as age, BMI, glucose levels and others. By examining the coefficients and their associated confidence intervals, one can assess the positive or negative relationships between these variables and the likelihood of developing diabetes. A higher coefficient value suggests that an increase in the corresponding predictor variable is associated with a higher probability of diabetes. Among the predictors, the diabetes genetic score and the number of pregnancies had the highest coefficient values, 2.58 and 1.21, respectively. These values indicate a strong correlation between an elevated genetic score and the likelihood of diabetes, as well as an increased number of pregnancies being associated with a higher probability of diabetes. Other variables, such as skin thickness, glucose levels, and BMI, showed moderate levels of correlation with the presence of diabetes, although their coefficient values were not as high as those for the genetic score and pregnancies. These findings suggest that while the genetic score and number of pregnancies may have the most significant impact on predicting diabetes, other factors like glucose levels, BMI, and skin thickness also play a meaningful role in determining the likelihood of developing the condition.

**Conclusion**

The logistic regression model demonstrates reasonable predictive power when considering the given variables and its overall ability to predict the presence of diabetes. The ROC curve and AUC values indicate that the model performs effectively at distinguishing between the two groups (those with and without diabetes). However, there is potential for further improvement. For instance, splitting the dataset into training and testing subsets, or employing cross-validation techniques, could help mitigate any potential bias and provide a more reliable assessment of model performance. Relying solely on the ROC curve and AUC as measures of sensitivity and specificity may not fully capture how well the model generalizes to new, unseen data. Additionally, incorporating feature selection could enhance the model’s accuracy by removing non-significant variables that do not contribute meaningfully to the prediction. This would allow for a better fit of the data, potentially improving the model’s interpretability and predictive capabilities. While the current model offers valuable insights into the presence of diabetes, refining it through these methods could lead to even more robust results and ensure that the model performs consistently across different datasets.

**Part B: Descriptive presentation of data**

This dataset consists of eleven variables, which provide a comprehensive view of the study participants. Among these, six variables were measured at baseline and were not considered time-varying covariates: sex, age, education, cardiovascular disease (CVD), diuretics, and dementia. Two additional variables, BMI and the number of cigarettes smoked, were treated as time-varying covariates since they were measured repeatedly throughout the study. The ID variable was excluded from the analysis because it did not correlate with BMI changes over time or relate to smoking cessation. The primary focus of this study was to assess the relationship between smoking cessation (the treatment variable) and the percentage change in BMI (the outcome variable).

A diagram of smoking cessation

Description automatically generated

The figures shown above relates the overall BMI change percent to smoking cessation at the end of the study. This box plot shows that those who underwent the treatment (smoking cessation) had an average lower BMI change percentage than those who did not undergo the treatment. With an average BMI change in those who stopped smoking of -5.85% and those who did stop smoking of -2.52%, the difference of these being 3.32%

**Results**

To examine the relationship between smoking cessation and BMI change after five years, a simple linear regression model was employed. This model was run in two versions: one without adjusting for other variables (i.e., an unadjusted model), and one adjusted for potential confounders (age, sex, education, CVD, diuretics, and dementia). In the unadjusted model, the results showed a significant relationship between smoking cessation and BMI change. Individuals who quit smoking generally experienced a lesser reduction in BMI compared to those who continued smoking. However, this model did not account for other variables that may influence BMI change, leading to a potential over- or under-estimation of the true effect. When adjusting for baseline covariates in the model, the estimated relationship between smoking cessation and BMI change was slightly attenuated, but the association remained significant. This suggests that while smoking cessation is strongly associated with BMI change, other baseline factors such as age, sex, and presence of CVD also play an important role in shaping the extent of BMI change over time

To further adjust for potential confounding factors, inverse probability weighting (IPW) was applied. This technique gives more weight to observations that are underrepresented in the treatment group (i.e., those who quit smoking), helping to balance the distribution of confounders between the treatment and control groups. By applying IPW, the analysis accounted for baseline differences in characteristics such as age, sex, and comorbidities. The results from the IPW-adjusted model showed that smoking cessation continued to have a significant effect on BMI change, reinforcing the findings from the outcome regression model. The use of IPW provides a more accurate estimate of the causal effect of smoking cessation on BMI change, as it adjusts for confounders that could otherwise bias the results.

**G-Formula BMI Change**

The G-formula was used to estimate the potential BMI changes under two hypothetical scenarios: (a) had nobody quit smoking, and (b) had all participants quit smoking. The G-formula allows us to estimate the counterfactual outcomes for both scenarios based on the observed data. In the scenario where nobody quit smoking, the average change in BMI after five years was estimated to be -5.85%. In contrast, if all participants had quit smoking, the average change in BMI was estimated to be -2.53%. The average causal effect of smoking cessation on BMI change, which is the difference between these two scenarios, was found to be 3.32%. This suggests that, on average, individuals who quit smoking experienced a 3.32% greater reduction in BMI compared to those who did not quit smoking.

**Comparison and Discussion**

The results of the study show that smoking cessation is associated with a significant change in BMI. Those who quit smoking had a lesser reduction in BMI compared to those who continued smoking. However, several important assumptions must be considered when interpreting these results. First, the model assumes that all variables have a linear relationship, which may not always be the case. If the relationships between the covariates and BMI change are non-linear, this assumption could lead to biased estimates. Additionally, although inverse probability weighting and the G-formula were used to adjust for confounding, there may still be unmeasured confounders that could influence the results. For example, factors such as diet, exercise, and overall lifestyle changes that often accompany smoking cessation were not accounted for in this study, and they may have further impacted BMI change. Another potential source of bias is selection bias. Individuals who choose to quit smoking may differ systematically from those who do not, and these differences might not be fully captured by the observed covariates. For instance, people who quit smoking might also engage in healthier behaviours that contribute to weight loss, independently of smoking cessation itself. Despite these potential sources of bias, the results of the study provide valuable insights into the relationship between smoking cessation and BMI change. While the findings suggest a significant causal effect of smoking cessation on BMI reduction, it is important to consider the limitations and assumptions of the models used to estimate these effects. Future research should aim to address unmeasured confounders and explore potential non-linear relationships to further refine an understanding of this relationship.

**Bibliography:**

Loke, Atul. “Diabetes.” *World Health Organization*, World Health Organization, www.who.int/health-topics/diabetes#tab=tab\_1. Accessed 15 Jan. 2025.

Part A Descriptive Statistics:

A graph of a box plot

Description automatically generatedA graph of a boxplot of bmi by diabetes status

Description automatically generatedA diagram of a boxplot of blood pressure

Description automatically generatedA graph of a boxplot of diabetes

Description automatically generatedA diagram of a box plot

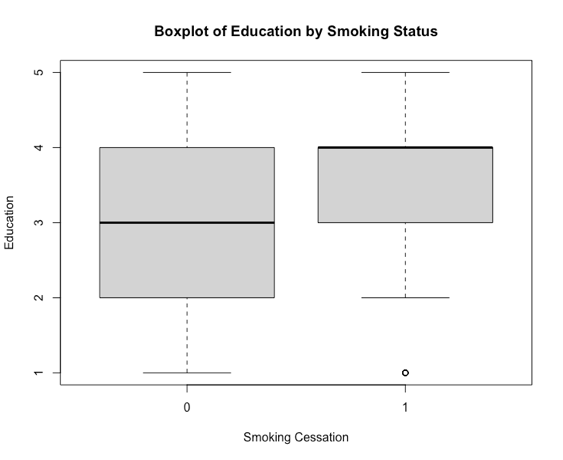
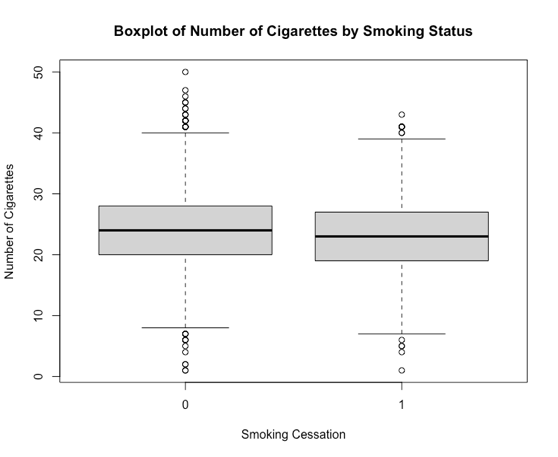
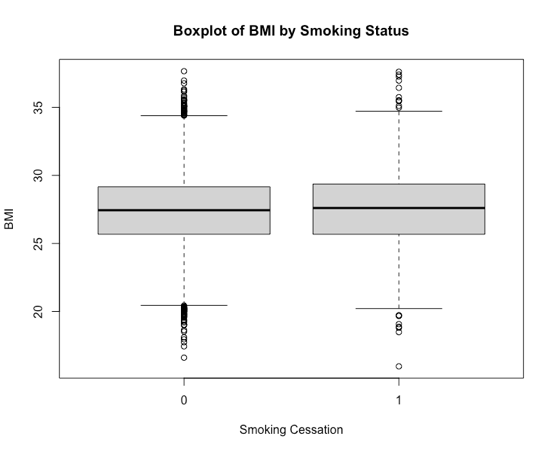
Description automatically generatedA graph of diabetes status

Description automatically generatedA graph of a boxplot

Description automatically generatedA graph of skin thickness by diabetes

Description automatically generated

Part B Descriptive Statistics:

A graph of smoking status

Description automatically generated