M5MS03 - Applied Statistics; Coursework 4 - Autumn 2018

The dataset, `pima`, consists of 768 observations and 9 features. Each observation corresponds to one adult female Pima Indian living near Phoenix, Arizona. The data was collected by a U.S. government agency, the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) sometimes around 1988. The role of NIDDK is to research diabetes and related diseases, and to communicate scientific findings to the American public.

A 2004 paper connected with this dataset, authored by clinical researchers at the NIDDK's Phoenix research centre, states that the Pima Indians suffer from the highest reported rates of type 2 diabetes[1] in the world [1]. The paper's authors explain that the Pima Indian's limited environmental and genetic variability make them excellent candidates to understand the causes of type 2 diabetes, a disease that has both genetic and environmental components. The same paper highlights that research findings from previous studies of the Pima Indians with type 2 diabetes have been shown to be universally applicable. The symptoms of type 2 diabetes are unpleasant. The disease can also contribute to future health problems such as heart disease and nerve damage [4]. It is therefore desirable to understand the disease's etiology and to diagnose it as early as possible.

The dataset's features are as follows:

- `pregnant` - number of times pregnant.
- `glucose` - plasma glucose concentration at 2 hours in an oral glucose tolerance test (mg/dL)[2].
- `diastolic` - diastolic blood pressure (mm Hg).
- `triceps` - triceps skin fold thickness (mm).
- `insulin` - 2-hour fasting serum insulin level (mu U/ml, micro units per millilitre).
- `bmi` - body mass index (weight[kg]/height[m]^2).
- `diabetes` - diabetes pedigree function (see text for discussion).
- `age` - age (years).
- `test` - test whether the patient shows signs of diabetes (1 = shows signs).

We will be attempting to predict whether a patient has the symptoms of diabetes, (test=1), using the other features in the dataset. A priori of any analysis, it is clear that the datasets features are likely to predictive of whether a person has type 2 diabetes. Higher plasma glucose concentration, `glucose`, suggests higher insulin resistance and reduced insulin secretion, both of which occur in diabetics. Serum insulin level, `insulin`, has been shown to be predictive of type 2 diabetes, with a higher level corresponding to a higher prevalence [2]. The World Health Organisation classifies a person whose `bmi` is greater than 25 as overweight, while a score greater than 30 is obese[3]. Triceps skinfold measurement, feature `triceps`, involves using calipers to measure the thickness of skin on a person's triceps. This measurement can be used to derive an estimate of body fat percentage. As a result of variation in where fat is stored as a person ages, this estimate can be improved by including age as a variable in the calculation of body fat[4]. Obesity is a major risk factor in developing diabetes, so we would expect that `bmi` and `triceps` are predictive of `test` (although each may not be predictive after conditioning on the other). According to the NIDDK's informational materials, blood pressure is symptomatic of insulin resistance [3], so we might also expect it to be useful as a predictor.

The diabetes pedigree function, `diabetes`, is a feature developed in 1988 by researchers attempting to predict the onset of type 2 diabetes using neural networks [5]. In their words,

[1] Type 2 diabetes is characterised by high levels of glucose in blood plasma. It is caused by a combination of increased cellular resistance to insulin, a chemical that regulates glucose levels, and reduced insulin production.

[2] These units were inferred on the basis of the data's scale.

[3] `https://en.wikipedia.org/wiki/Body_mass_index`

[4] `https://en.wikipedia.org/wiki/Body_fat_percentage#Skinfold_methods`

> *We developed the Diabetes Pedigree Function (DPF) to provide a synthesis of the diabetes mellitus history in relatives and the genetic relationship of those relatives to the subject. The DPF uses information from parents, grandparents, full and half siblings, full and half aunts and uncles, and first cousins. It provides a measure of the expected genetic influence of affected and unaffected relatives on the subject's eventual diabetes risk.*

So DPF accounts for the heritability of diabetes, and therefore also seems likely to be predictive. With the relevance of these features in mind, we proceed with the analysis.

Question 1. Data pre-processing:

(a) Examine the data using simple graphs and numerical summaries.

(b) Can you find any obvious irregularities in the data?

Based on their nature, all feature values should be strictly positive and continuous, with exception of those for `test`, a binary factor, and `pregnant`, a nonnegative count. Summarizing the data reveals that feature vectors for `glucose`, `diastolic`, `triceps`, `insulin`, and `bmi` contained inexplicable 0 values. It seems plausible that `0` was used to code for `NA`, so we replace these values with `NA`. `pregnant` and `test` may have contained 0 values that were meant to represent `NA`, but identifying these would have required a time-consuming additional analysis. Instead, assume that no entries in these columns were missing. The missing values were broken down as follows:

```
  print(colSums(is.na(Pima))) # Number of missing values per column
```

```
##  pregnant   glucose diastolic   triceps   insulin       bmi  diabetes
##         0         5        35       227       374        11         0
##       age      test
##         0         0
```

```
  print(sum(rowSums(is.na(Pima)) > 0)) # Number of entries with > 0 missing values
```

```
## [1] 376
```

The analysis contained in this report omits all cases with missing values.

Plotting histograms for each feature reveals that many of the feature vectors contained extreme values, either in the form of isolated values or long tails. The features' marginal distributions tend to be skewed right, either due to a few women having (for example) more than a dozen children or a bmi above 50. Using boxplots such as those shown in Figure 1, it was possible to identify which features were plausibly predictive of `test`. Features that showed a large discrepancy in their median values when grouped by `test` value were `glucose`, `age`, `triceps`, `insulin`. The features `diabetes`, `pregnant`, and `diastolic` were borderline - there was a discrepancy, although it was small and possibly due to sample randomness.

The `NA`-omitted data set's correlation matrix was studied in anticipation of possible collinearity. Most features were only weakly correlated, with the exception of the pairs `(age, pregnant)`, `(glucose, insulin)`, and `(triceps, bmi)`. The correlation matrix is reproduced below.

```
  round(cor(Pima[complete.cases(Pima), -9]), 1)
```

```
##           pregnant glucose diastolic triceps insulin bmi diabetes age
## pregnant       1.0     0.2       0.2     0.1     0.1 0.0      0.0 0.7
## glucose        0.2     1.0       0.2     0.2     0.6 0.2      0.1 0.3
## diastolic      0.2     0.2       1.0     0.2     0.1 0.3      0.0 0.3
## triceps        0.1     0.2       0.2     1.0     0.2 0.7      0.2 0.2
## insulin        0.1     0.6       0.1     0.2     1.0 0.2      0.1 0.2
## bmi            0.0     0.2       0.3     0.7     0.2 1.0      0.2 0.1
## diabetes       0.0     0.1       0.0     0.2     0.1 0.2      1.0 0.1
## age            0.7     0.3       0.3     0.2     0.2 0.1      0.1 1.0
```

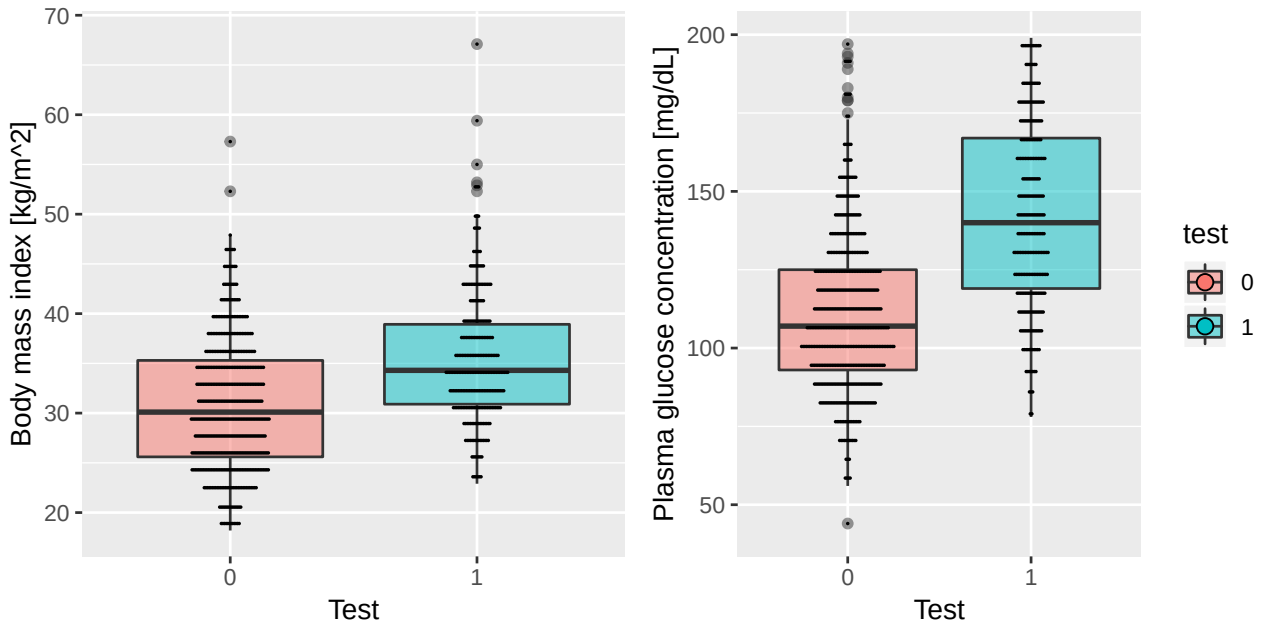

Figure 1: Combined box and dot plots for b.m.i. and diabetes pedigree function versus whether a person showed symptoms of diabetes (feature `test`).

Only complete cases were retained for further analysis, reducing the number of observations from 768 to 392. It was assumed that censorship was not correlated with the response. Future analyses could impute for the missing values by regressing the features onto one another - this may improve the resulting model.

Question 2. Data analysis for the full-order model.

(a) Fit to the data the full model (i.e. the model with all possible predictors included). Present results related to significance of predictors and provide appropriate comments.

A generalized linear model was fit to the data with each predictor included. Higher-order terms, interactions, and nonlinear transformations of the features were not included. Denoting response i's `test` probability of success[5] as π_i and its associated feature vector as x_i, this model has the form:

$$g(\pi_i) = x_i^T \beta \tag{1}$$

where g is the logit link function, $g(\pi) = \log \frac{\pi}{1-\pi}$. The formula for this model is:

`test ~ 1 + pregnant + diastolic + triceps + insulin + bmi + diabetes + age`

The AIC value for this model was 413, as compared to 500 for an intercept-only model. The p-values associated with each MLE coefficient are presented in the table below. We can see that the intercept, `insulin`, `bmi`, `diabetes`, and `age` coefficients are significantly non-zero. By contrast, the MLE `diastolic`, `pregnant`, and `triceps` coefficients could plausibly have been sampled when in fact their true values were zero[6].

```
formatC(summary(Pima.glm1)$coefficients[ , 4], format = "e", digits = 1)
```

```
## (Intercept)     insulin         bmi    diabetes         age    pregnant
##   "3.2e-12"   "4.0e-04"   "1.7e-02"   "5.9e-03"   "3.7e-03"   "1.6e-01"
##   diastolic     triceps
##   "5.5e-01"   "4.0e-01"
```

[5] 'success' in this context meaning having the symptoms of type 2 diabetes.

[6] This discussion makes no mention of the coefficients practical significance since that is not the focus of this exercise.

The p-values obtained for the MLE coefficients are supported by an analysis of the change in deviance of the model as each variable is included. Wilks' theorem states that, under the null hypothesis of the data being generated by the smaller model, the distribution of the change in deviance of nested models is asymptotically chi-squared distributed with degrees of freedom equal to the difference in parameter dimensionality. This means that a variable must reduce the deviance by an amount that is greater than what would occur if the feature were non-descriptive. As the table below shows, the the MLE values for `diastolic`, `triceps`, and `pregnant` are not statistically significant.

```
  anova(Pima.glm1, test='Chisq')
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: test
##
## Terms added sequentially (first to last)
##
##
##           Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                        391     498.10
## insulin    1   35.181       390     462.92 3.005e-09 ***
## bmi        1   18.661       389     444.26 1.562e-05 ***
## diabetes   1    9.529       388     434.73  0.002023 **
## age        1   34.621       387     400.11 4.005e-09 ***
## pregnant   1    2.061       386     398.05  0.151075
## diastolic  1    0.380       385     397.67  0.537853
## triceps    1    0.715       384     396.95  0.397786
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Crucially, the signs of the coefficients are consistent with the physical reasoning provided in the introduction (negative coefficients correspond to lower probability of `test=1`).

```
  round(summary(Pima.glm1)$coef[1:5, 1], 4)
```

```
## (Intercept)     insulin         bmi    diabetes         age
##     -6.6079      0.0040      0.0595      1.0219      0.0494
```

Question 3. Perform exploratory model selection, based on AIC, using forward selection and backward elimination. Comment on the results.

Backward elimination on the full model resulted in excluding `diastolic` then `triceps`. The AIC decreased from 413 for the full model to 410 for the reduced model. Forward selection from the intercept-only model decreased the AIC from 500 to 410, and resulted in a model that contained the same terms as were obtained via backward selection. This is a fortunate coincidence[7] that may suggest (but does not guarantee) the remaining features form the model with the globally lowest AIC. This being said, the AIC value decreased by only 0.06 when the feature `pregnant` was included during forward selection, as compared with decreases of 45, 25, 13, and 5 when variables `age`, `bmi`, `insulin`, and `diabetes` were added. This suggests that `pregnant` does not substantially affect the model's likelihood, which is used as a reason for its exclusion in the next question.

[7]Since forward selection and backward elimination do not, in general, converge to the same solution.

Question 4. Recommend one model for the test variable. In your answer:

(a) explain your reasoning

(b) comment on potential advantages and disadvantages for this model. Your answers may include arguments from both a statistical and physical perspective.

Three models were considered:

- Model 1: `test` $\sim$ `1 + insulin + bmi + diabetes + age + pregnant + diastolic + triceps`
- Model 2: `test` $\sim$ `1 + insulin + bmi + diabetes + age + pregnant`
- Model 3: `test` $\sim$ `1 + insulin + bmi + diabetes + age`

To supplement the AIC results of the previous section, along with the tests of significance discussed earlier, the models' predictive performances were evaluated by testing them against a disjoint training and testing subsets of the data set. ROC curves were generated based on the predicted test probabilities. These curves are shown in Figure 2.

Key metrics for these models are presented in Table 1. The ROC curves and AUC scores reveal that the models have similar predictive performance, since the size of the testing set (100 observations) makes it likely that the differences in AUC are not statistically significant. Stepwise methods obtained Model 2, although the inclusion of `pregnant` was borderline. In the absence of information suggesting that pregnancy and diabetes are linked, it seems reasonable to exclude this feature from the model. Model 3 is therefore chosen on the basis of parsimony, along with the fact that all of its features can be justified physically. Reduced insulin production, obesity, genetic risk, and age are all known risk factors for type 2 diabetes. Presumably `triceps` does not improve model fit since it is highly correlated with `bmi`. `diastolic` seems likely to provide little information after conditioning upon the other factors, particularly `insulin`, since high blood pressure is clinically interpreted as a sign of reduced insulin production.

Question 5. Using the full model from (2), predict the outcome for a woman with predictor values 1, 99, 64, -11, 76, 27, 0.25, 25. Give a confidence interval for your prediction, on the response scale. Do you have any reservations about this prediction?

The full-order model predicted that the probability of a `test` outcome of 1 was 0.054 with standard error 0.03. The specified query is nonsense however, since it contains `triceps=-11`. Otherwise its contents are reasonable. The predicted probability should not be taken literally, since the query is both physically ridiculous and lies outside of the data subspace on which the model was fit.

Question 6. It is useful to consider the extent to which a model learned from these data can be used as a diagnostic device. That is, if covariate measurements could be obtained for new subjects, it is possible to predict diabetes given the covariates?

(a) Randomly split this data set into two samples: a training set and a test set. For the full model, estimate model parameters on the training set.

The code on the following page splits the dataset as requested, then estimates the model parameters on the training dataset.

Quantity	Model 1	Model 2	Model 3
AIC	413	410	410
AUC	0.86	0.87	0.89
No. features	8	6	5

Table 1: Metrics for the models in Question 4. The AIC was computed on the full dataset after fitting the model on that dataset. The AUC is the test-set AUC.

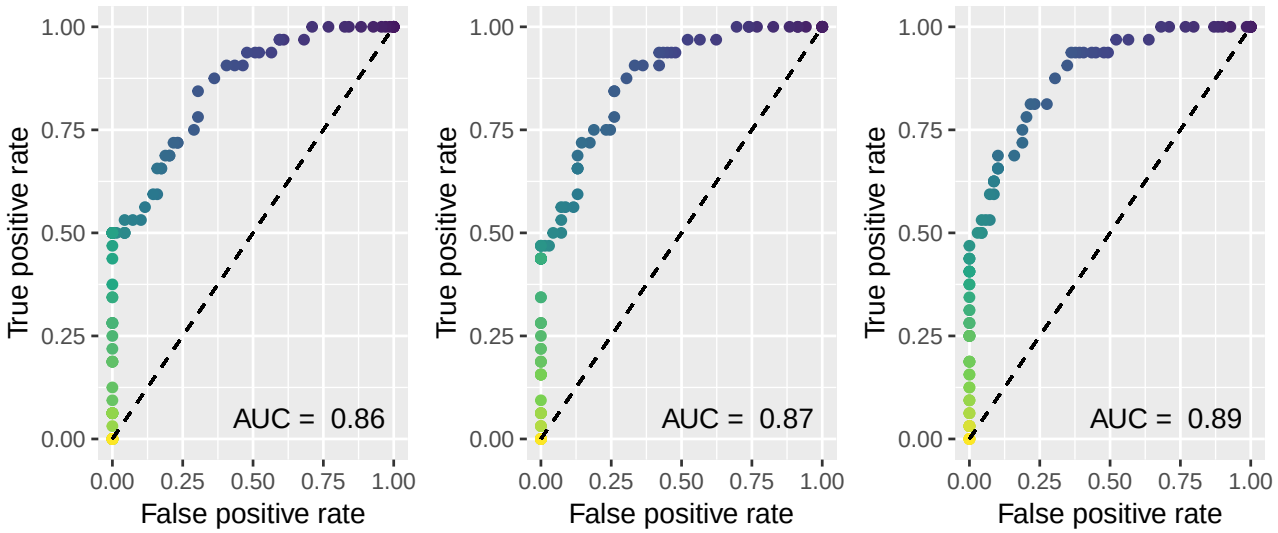

Figure 2: Test-set receiver operating characteristic (ROC) curves for the three models specified in the response to Question 4, with Model 1 on the left and Model 3 on the right. If the model is to be used to predict whether a person has symptoms of diabetes, then the ROC curve is relevant. It informs us of the proportion of people incorrectly referred for an in-person diagnosis, and the proportion of people with symptoms that are not referred for inspection. We can see that the choice of variables to include has a negligible effect on the ROC curve.

```
set.seed(23470)
shuffled_ix <- sample(1:nrow(Pima), size=nrow(Pima), replace=F)
training_ix <- shuffled_ix[1:292]
test_ix     <- shuffled_ix[293:392] # Test set contains 50 entries

Pima.full   <- glm(test ~ 1 + insulin + bmi + diabetes + age +
                     pregnant + diastolic + triceps, data=Pima[training_ix, ],
                     family=binomial(link='logit'))
```

(b) On the test set, compute (predict) the probability that test=1, given the covariate, for each instance. For the ith test set observation, denote this as P(test=1—xi), that is, the conditional probability that the ith observation is a 1.

This code computes the requested predictions.

```
pred_p <- predict(Pima.full, newdata=Pima[test_ix, ], type='response')
```

(c) Suppose we propose a decision rule that allocates a covariate vector to class 1 if $P(\texttt{test} = 1|x_i) > 0.5$ *(that is, the diabetic class), and class 0 otherwise. Apply this decision rule to the test set predictions. We now have a set of predicted classifications, and corresponding known labels. Compute a cross-classification of these two variables, and consider the sum of the off-diagonal components, expressed as a ratio of the the total number of test set observations. How do you interpret this number? Comment on the utility of this diagnostic mechanism for diabetes prediction.*

I assumed that this question asked for a confusion matrix. This matrix, C, is computed and printed below.

```
  pred_c  <- matrix(as.numeric(pred_p > 0.5))
  truth_c <- matrix(as.numeric(Pima$test[test_ix]) - 1)

  # Columns are actual class,
  # rows are predicted class,
  # cells are counts.
  #  | 0 | 1
  # 0|   |
  # 1|   |

  zero_zero <- sum(1 - pred_c[truth_c==0])
  zero_one  <- sum(1 - pred_c[truth_c==1])
  one_one   <- sum(pred_c[truth_c==1])
  one_zero  <- sum(pred_c[truth_c==0])
  C         <- matrix(c(zero_zero, zero_one,
                        one_zero, one_one), nrow=2, byrow=T)
  print(C) # Confusion matrix
```

```
##      [,1] [,2]
## [1,]   53   19
## [2,]   15   13
```

The sum of the confusion matrix's off-diagonal elements is the number of incorrect predictions that the classifier makes. Here, it is $(15+19)/100 = 0.34$. As a ratio of the total number of test set observations, this sum is the error rate of the classifier. By contrast, the ratio of the sum of the diagonal elements and the number of test set observations is the classifier's accuracy, $(53 + 13)/100 = 0.66$.

The confusion matrix is a valuable diagnostic tool in the context of most classification problems. It is particularly useful here, since it allows us to see that false-positive and false-negative rates simultaneously. Furthermore, it makes it clear that accuracy scores should be evaluated in the context of the dataset's class balance - the column sums reveal that there are only 32 `test=1` instances, compared to 68 `test=0` instances. If the imbalance were more extreme, a classifier could appear to achieve near-perfect accuracy simply by allocating all queries to the dominant class.

If this classifier were used commercially to decide whether a patient should be referred, the true-positive and false-negative rates would be important. A false-positive would mean that a person was referred to a doctor when they did not have symptoms of type 2 diabetes: while a nuisance, this outcome is not catastrophic. A false-negative, on the other hand, would result in a person failing to be referred to a doctor when they do in fact have type 2 diabetes, a circumstance that could have dire consequences. Reducing the decision boundary probability (0.5 in the question) would reduce the rate of dangerous false-negatives, while increasing the rate of irritating false-positives. Consultation with doctors and management staff would allow a suitable probability to be chosen.

In review, future analyses may consider attempting to exclude the outliers mentioned in the initial exploratory, or could impute for the missing values in the dataset in an attempt to fit a model with better predictive performance.

References

[1] Leslie J. Baier and Robert L. Hanson. "Genetic Studies of the Etiology of Type 2 Diabetes in Pima Indians". In: *Diabetes* 53.5 (2004), pp. 1181–1186. ISSN: 0012-1797. DOI: 10.2337/diabetes.53.5.1181. eprint: http://diabetes.diabetesjournals.org/content/53/5/1181.full.pdf. URL: http://diabetes.diabetesjournals.org/content/53/5/1181.

[2] Mercedes R. Carnethon et al. "Serum Insulin, Obesity, and the Incidence of Type 2 Diabetes in Black and White Adults". In: *Diabetes Care* 25.8 (2002), pp. 1358–1364. ISSN: 0149-5992. DOI: 10.2337/diacare.25.8.1358. eprint: http://care.diabetesjournals.org/content/25/8/1358.full.pdf. URL: http://care.diabetesjournals.org/content/25/8/1358.

[3] *National Institute of Diabetes and Digestive and Kidney Disease: Insulin Resistance Prediabetes*. https://www.niddk.nih.gov/health-information/diabetes/overview/what-is-diabetes/prediabetes-insulin-resistance. Accessed: 03-12-2018.

[4] *National Institute of Diabetes and Digestive and Kidney Disease: Type 2 Diabetes*. https://www.niddk.nih.gov/health-information/diabetes/overview/what-is-diabetes/type-2-diabetes. Accessed: 03-12-2018.

[5] Everhart J. E. Dickson W. C. Knowler W. C. Johannes R. S. Smith J. W. "Using the ADAP Learning Algorithm to Forecast the Onset of Diabetes Mellitus". In: *Proceedings of the Annual Symposium on Computer Application in Medical Care* (1988), pp. 261–265.