STAT 135, Lab Project, due 5/4/15
Type II Diabetes Detection

April 14, 2015

This project may be done in groups of up to 3 people. It requires you to download the dataset `diabetes.csv`.

1 Background

Diabetes is a major health issue in the US. In 2010, an estimated 285 million people, and 29.1 million Americans, or 9.3% of the population, had diabetes [1].

The file `diabetes.csv` contains cardiovascular data that was taken on a random sample of 375 African American patients in Central Virginia to study the prevalence of cardiovascular risk factors in this population. Type II diabetes is a common illness that one is not born with but rather develops during his or her lifetime. Trying to assess risk factors associated with type II diabetes is an important question as the prevalence of type II has gone up in recent decades. For instance, Type II Diabetes has been associated in past studies with obesity,

The level of glycosylated hemoglobin is increased in the red blood cells of people with poorly controlled diabetes. Since the glucose stays attached to hemoglobin for the life of the red blood cell (normally about 120 days), the level of glycosylated hemoglobin reflects the average blood glucose level over the past 3 months. A score of 7 or greater for glycosylated hemoglobin is considered type II diabetic. The outcome measure in the diabetes dataset is `glyhb`, which is Glycosolated hemoglobin. Possible predictors or indicators for diabetes are:

- `chol`: choesterol
- `stab.glu:` stabilized glucose levels
- `hdl:` high density lipoprotein
- `ratio:` chol/hdl
- `age`
- `height:` is in inches
- `weight:` is in pounds

[1] See more at: http://www.diabetes.org/diabetes-basics/statistics/#sthash.50n1N3pS.dpuf

- `bp.1s:` first systolic blood pressure
- `bp.1d:` first diastolic blood pressure
- `waist:` given in inches
- `hip:` circumference given in inches
- `frame:` complexity.
- `time.ppn:` minutes after eating that their glucose levels were measured (postprandial time)

Your goal is to perform a statistical analysis of the available data to gain insight on the effect of the different indicators on the presence of diabetes.

2 Accessing Data, Visualization and Summarization

We shall start by getting familiar with the data through some visualization and summarization.

1. Construct a histogram of `glyhb`, as well as a box-plot and *qq*-plot against a standard normal. Comment on the properties you observe.
2. Do the same for `chol`. Which feature is better approximated with a Gaussian distribution?
3. Construct a scatter-plot of `bp.1s` with `bp.1d`, and another of `age` versus `height`. Would you say the features are approximately independent?
4. As discussed earlier, diabetes is diagnosed according to $\texttt{glyhb} \geq 7$ or $\texttt{glyhb} < 7$. In order to identify potentially relevant features, compare for each available feature two box plots, conditional on whether $\texttt{glyhb}\geq 7$ or $\texttt{glyhb}< 7$.
5. Besides the available features, we can also include the `BMI` (Body Mass Index), defined as $703\frac{\texttt{weight}}{(\texttt{height})^2}$, and the waist-to-hip ratio, defined as $\texttt{WHR} = \frac{\texttt{waist}}{\texttt{hip}}$. Compute their empirical distributions and produce the conditional box-plots.
6. In light of these first experiments, which features seem more related (and unrelated) to the presence of type II diabetes?

3 Parametric Inference

We will now perform some parametric inference on selected features.

1. Use the method of moments to fit a Gamma distribution to the `BMI`. Use non-parametric bootstrap to construct 95% confidence intervals around the estimated parameters. Plot the fitted density over the observed histogram to examine visually goodness of fit.
2. The `WHR` has empirical histogram closer to a Gaussian distribution. Fit a Normal distribution using Maximum Likelihood and provide 95% confidence intervals for their parameters. Plot the fitted density over the observed histogram to examine visually goodness of fit.

3. The former ratios are known to behave differently according to gender, and our goal is to understand also their effect on the `glyhb` levels. For that purpose, repeat the parametric fits of `BMI` and `WHR`, but this time do a different fit for each population $\texttt{male} \cap (\texttt{glybh} > 7)$, $\texttt{female} \cap (\texttt{glybh} > 7)$, $\texttt{male} \cap (\texttt{glybh} < 7)$ and $\texttt{female} \cap (\texttt{glybh} < 7)$. Construct confidence intervals around each mean and variance parameter. What do you conclude?

4 Testing

We now advance towards our main objective, that is discover what are the factors influencing the type-II diabetes. For that purpose, you are asked to perform some comparison testing that might help us understand how to combine the features.

1. Are male and females equally exposed to type-II diabetes?

2. Pick your 5 favorite features (you may include the ratios if you want) and for each feature, test equality of means for those with type-II diabetes and those without. (use 5% significance level). Justify the test statistic (parametric or non-parametric) with arguments.

3. Estimate the probability $\pi_{\texttt{BMI}}$ that a diabetic male has larger `BMI` than a non-diabetic male. Do the same for $\pi_{\texttt{WHR}}$. Construct 95% confidence intervals for such estimated probability.

4. If you received a new male patient (assuming that it came from the same population) and only had access to his `WHR`, construct a test for type-II diabetes using the empirical distributions. Use significance $\alpha \leq 5\%$ and compute its power.

5. Using the categories of `BMI` given in Table 1, would you say that the male and female population sample has homogeneous distribution of such categories? What if you use the categories of `WHR` given by Table 2 instead ?

6. Use Tables 1 and 2 to test whether these features interact when it comes to detecting `glybh`. Use significance level 5%. Which of the two ratios is more sensitive to `glybh` in light of this analysis? (*Hint: compare the variability of means of `glybh` across groups of `BMI` with the variability of means of `glybh` across groups of `WHR`*). Is it consistent with the answer of Section 3, part 3 ?

5 Regression

Finally, let us apply the insights from previous Sections to predict type-II diabetes from the available features. Let $\rho(y, \lambda)$ denote the threshold function:

$$\rho(y, \lambda) = \begin{cases} 1 & \text{if } y > \lambda \ , \\ 0 & \text{otherwise.} \end{cases}$$

1. Perform linear regression on `glybh` using the features that you consider most relevant for predicting type-II diabetes (you can of course include the ratios). Denote by y_i the `glybh` level of patient i and by $\hat{y}_i$ the predicted value from the linear regression. As we did in earlier

sections, we consider that a patient has type-II diabetes when $d_i = \rho(y_i, 7) = 1$. If we predict that a subject has diabetes with $\hat{d}_i = \rho(\hat{y}_i, 7)$, compute your prediction error rate

$$\mathsf{err} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}(d_i \neq \hat{d}_i) \ .$$

2. Using the same threshold $\lambda = 7$, compute the False Positives Rate (incorrectly reporting a non-diabetic subject as being diabetic) and the False Negatives Rate (incorrectly reporting a diabetic subject as being non-diabetic). Now, suppose that a False Negatives Rate larger than 10% is unacceptable. How would you modify λ to meet with the specifications? What False Positive Rate do you achieve in that case?

3. What are the features with the largest influence? Test whether `hdl`, `bp 1s` and `bp 1d` have any predictive value at 5% significance. *Hint: You might want to standardize your features to better interpret the regression coefficients.*

4. If in Section 4, part 5 you concluded that both ratios interact, how would you use this information to improve your regression?

5. Plot the residuals as a function of `stab.glu` and the ratios `BMI`, `WHR`. Can you come up with a transformation of the variables that stabilizes the variance of the residual errors?

6. Perform Logistic Regression using the normalized input features and compare the performance with what you obtained in Section 5, part 2.

7. Finally, test your two predictors on the held-out data from `diabetes_test.csv`. Do you obtain similar performance as in the training set? Can you justify why? *Note: It is illegal to use the test set to re-train your models!*

6 Tables

Underweight	Healthy	Overweight	Level 1 Obese	Level 2 Obese	Level 3 Obese
< 18.5	$18.5 - 24.99$	$25 - 29.99$	$30 - 34.99$	$35 - 39.99$	> 40

Table 1: `BMI` Standards

gender	Age	Low	Moderate	High	Very High
men	$20 - 29$	< 0.83	$0.83 - 0.88$	$0.89 - 0.94$	> 0.94
	$30 - 39$	< 0.84	$0.84 - 0.91$	$0.92 - 0.96$	> 0.96
	$40 - 49$	< 0.88	$0.88 - 0.95$	$0.96 - 1.00$	> 1.00
	$50 - 59$	< 0.90	$0.90 - 0.96$	$0.97 - 1.02$	> 1.02
	> 60	< 0.91	$0.91 - 0.98$	$0.99 - 1.03$	> 1.03
women	$20 - 29$	< 0.71	$0.71 - 0.77$	$0.78 - 0.82$	> 0.82
	$30 - 39$	< 0.72	$0.72 - 0.78$	$0.79 - 0.84$	> 0.84
	$40 - 49$	< 0.73	$0.73 - 0.79$	$0.80 - 0.87$	> 0.87
	$50 - 59$	< 0.74	$0.74 - 0.81$	$0.82 - 0.88$	> 0.88
	> 60	< 0.76	$0.76 - 0.83$	$0.84 - 0.90$	> 0.90

Table 2: `WHR` Standards for Men and Women