Investigating the Fluctuation of Blood Glucose Levels using Multiple Linear Regression

Dezhao Han

A Project for Regression Analysis Course

Summer 2012

offered by NEAS

1 Introduction and Data

The blood glucose level is the amount of sugar in the blood of a human. Glucose is the primary energy for the body, the mean normal blood glucose level in humans ranges from 4 mmol/L to 7 mmol/L. If someone's blood glucose level is outside the normal range, his or her medical condition is not as well as expected. A high level indicates hyperglycemia, while low levels are referred to as hypoglycemia.

It is believed that the blood glucose level is affected by other levels such as total cholesterol level, triglyceride level, insulin level, and glycated hemoglobin level, etc. In this project, we investigate how the blood glucose level is affected by these factors. In particular, we set up a multiple linear model to describe the interactions.

2 Data

Blood glucose levels of 27 diabetics as well as the level of total cholesterol, triglyceride, insulin, glycated hemoglobin and blood glucose are given in Table 1.

Table 1: Table 1

i	Total cholesterol (mmol/L) X_{1i}	Triglyceride (mmol/L) X_{2i}	Insulin (μ U/ml) X_{3i}	Glycated hemoglobin (%) X_{4i}	Blood glucose (mmol/L) Y_i
1	5.68	1.90	4.53	8.20	11.20
2	3.79	1.64	7.32	6.90	8.80
3	6.02	3.56	6.95	10.80	12.30
4	4.85	1.07	5.88	8.30	11.60
5	4.60	2.32	4.05	7.50	13.40
6	6.05	0.64	1.42	13.60	18.30
7	4.90	8.50	12.60	8.50	11.10
8	7.08	3.00	6.75	11.50	12.10
9	3.85	2.11	16.28	7.90	9.60
10	4.65	0.63	6.59	7.10	8.40
11	4.59	1.97	3.61	8.70	9.30
12	4.29	1.97	6.61	7.80	10.60
13	7.97	1.93	7.57	9.90	8.40
14	6.19	1.18	1.42	6.90	9.60
15	6.13	2.06	10.35	10.50	10.90
16	5.71	1.78	8.53	8.00	10.10
17	6.40	2.40	4.53	10.30	14.80
18	6.06	3.67	12.79	7.10	9.10
19	5.09	1.03	2.53	8.90	10.80
20	6.13	1.71	5.28	9.90	10.20
21	5.78	3.36	2.96	8.00	13.60
22	5.43	1.13	4.31	11.30	14.90
23	6.50	6.21	3.47	12.30	16.00
24	7.98	7.92	3.37	9.80	13.20
25	11.54	10.89	1.20	10.50	20.00
26	5.84	0.92	8.61	6.40	13.30
27	3.84	1.20	6.45	9.60	10.40

3 Model

We denote by Y_i the blood glucose level, X_{1i} the total cholesterol level, X_{2i} the triglyceride level, X_{3i} the insulin level, and X_{4i} the glycated hemoglobin level. Thus, Y_i is described as

$$Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \varepsilon_i,$$

where $\alpha, \beta_1, \beta_2, \beta_3, \beta_4$ are coefficients to be determined, ε_i follows a standard normal distribution.

Define $\hat{Y}_i$ as

$$\hat{Y}_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i}, \qquad i = 1, 2, \ldots, 27.$$

Thus,

$$\varepsilon_i = Y_i - \hat{Y}_i,$$

and

$$\sum E\left[\varepsilon_i\right] = 0.$$

We approximate the coefficients by the estimators that minimize

$$Q \triangleq \sum \varepsilon_i = \sum (Y_i - \hat{Y}_i)^2 = \sum \left[Y_i - (\alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i})\right]^2.$$

Taking derivatives of Q with respect to α, β_1, β_2, β_3, β_4 respectively leads to

$$\begin{aligned}
\frac{\partial Q}{\partial \alpha} &= \sum -2\left[Y_i - \alpha - \beta_1 X_{1i} - \beta_2 X_{2i} - \beta_3 X_{3i} - \beta_4 X_{4i}\right], \\
\frac{\partial Q}{\partial \beta_1} &= \sum -2X_{1i}\left[Y_i - \alpha - \beta_1 X_{1i} - \beta_2 X_{2i} - \beta_3 X_{3i} - \beta_4 X_{4i}\right], \\
\frac{\partial Q}{\partial \beta_1} &= \sum -2X_{2i}\left[Y_i - \alpha - \beta_1 X_{1i} - \beta_2 X_{2i} - \beta_3 X_{3i} - \beta_4 X_{4i}\right], \\
\frac{\partial Q}{\partial \beta_1} &= \sum -2X_{3i}\left[Y_i - \alpha - \beta_1 X_{1i} - \beta_2 X_{2i} - \beta_3 X_{3i} - \beta_4 X_{4i}\right], \\
\frac{\partial Q}{\partial \beta_1} &= \sum -2X_{4i}\left[Y_i - \alpha - \beta_1 X_{1i} - \beta_2 X_{2i} - \beta_3 X_{3i} - \beta_4 X_{4i}\right].
\end{aligned}$$

Let $\frac{\partial Q}{\partial \alpha}$, $\frac{\partial Q}{\partial \beta_1}$, $\frac{\partial Q}{\partial \beta_2}$, $\frac{\partial Q}{\partial \beta_3}$, $\frac{\partial Q}{\partial \beta_4}$ be zero respectively. The solutions to the coefficients are the roots of the following system of equations:

$$\sum Y_i - n\alpha - \beta_1 \sum X_{1i} - \beta_2 \sum X_{2i} - \beta_3 \sum X_{3i} - \beta_4 \sum X_{4i} = 0,$$

$$\begin{aligned}
&\sum X_{1i}Y_i - \alpha \sum X_{1i} - \beta_1 \sum X_{1i}^2 - \beta_2 \sum X_{1i}X_{2i} - \beta_3 \sum X_{1i}X_{3i} - \beta_4 \sum X1iX_{4i} = 0,\\
&\sum X_{2i}Y_i - \alpha \sum X_{2i} - \beta_1 \sum X_{1i}X_{2i} - \beta_2 \sum X_{2i}^2 - \beta_3 \sum X_{2i}X_{3i} - \beta_4 \sum X2iX_{4i} = 0,\\
&\sum X_{3i}Y_i - \alpha \sum X_{3i} - \beta_1 \sum X_{1i}X_{3i} - \beta_2 \sum X_{2i}X_{3i} - \beta_3 \sum X_{3i}^2 - \beta_4 \sum X3iX_{4i} = 0,\\
&\sum X_{4i}Y_i - \alpha \sum X_{4i} - \beta_1 \sum X_{1i}X_{4i} - \beta_2 \sum X_{2i}X_{4i} - \beta_3 \sum X_{3i}X_{4i} - \beta_4 \sum X_{4i}^2 = 0.
\end{aligned}$$

4 Analysis

4.1 The Classic Multiple Linear Regression

According to the first section, the estimators for $(\alpha, \beta_1, \beta_2, \beta_3, \beta_4)$ are

$$\begin{aligned}
\hat{\alpha} &= 5.9433,\\
\hat{\beta}_1 &= 0.1424,\\
\hat{\beta}_2 &= 0.3515,\\
\hat{\beta}_3 &= -0.2706,\\
\hat{\beta}_4 &= 0.6382.
\end{aligned}$$

Thus,

$$\hat{Y}_i = 5.9433 + 0.1424X_{1i} + 0.3515X_{2i} - 0.2706x_{3i} + 0.6382X_{4i}.$$

In the next, we show our multiple linear model is meaningful. That is to test the null hypothesis that all the regression slopes are 0:

$$H_0: \quad \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0.$$

Given the data in Table 1, test statistics are listed in Table 2. Given the P-value in Table 2, we reject the null hypothesis with a significance of 0.001. Hence, it is reasonable that all the factors affects the blood glucose level.

Note that

$$R^2 = \frac{\text{RegSS}}{\text{TSS}}$$

Table 2: Analysis-of-variance table

Source	Sum of squares	df	Mean Square	F	P
Regression	133.7107	4	33.4277	8.2778	0.000312129
Residuals	88.8412	22	4.038235		
Total	222.5519				

indicates how the model is fitted to the data. The closer is R^2 to 1, the better is the model fitted to the data. In our project

$$R^2 = \frac{133.7107}{222.5519} = 0.6008.$$

This indicates our model fits the data. That is the levels of total cholesterol, triglyceride, insulin, and Glycated hemoglobin do affects the blood glucose level.

4.2 The Importance of Each Factor

In the previous section, we carried out a multiple regression, assuming each factor has reasonable influences on the blood glucose level. However, under some circumstance, this is not the best model. In the next, we check the importance of each factor. If any factor has negligible influence on the blood glucose level, we kick out such factor and carry out another multiple linear regression with respect to the remaining variables. This procedure is also called the stepwise regression.

In doing that, we first evaluate t-values for each factor X_i and illustrate them in Table 3. The t_is are used to test the hypothesis $H_0 : \beta_i = 0$ for $i = 1, 2, 3, 4$. We know that t_i, $i = 1, 2, 3, 4$ follows a t distribution with a degree of freedom of $27 - 4 - 1 = 22$. Thus if

Table 3: t-values for each factor

X_i	X_1	X_2	X_3	X_4
t_i	0.3896	1.7211	-2.2290	2.6235

$|t_i| \leq t_{\alpha/2,22}$, $H_0 : \beta_i = 0$ is rejected. Since, given $\alpha = 0.05$, $t_{\alpha/2,22} = 2.074$, it is reasonable to assume that $\beta_1 = 0$, $\beta_2 = 0$. That is, levels of total cholesterol and triglyceride are less important in affecting the blood glucose level.

Another way to check the each factor's behavior is to compare their standardized coefficients. The smaller the standardized coefficient absolute value, the less important the corresponding factor is. To evaluate the standardized coefficient, we first standardize the data by

$$X_i' = \frac{X_i - \bar{X}_i}{S_i}, \qquad \text{for} i = 1, 2, 3, 4. \tag{4.1}$$

Then the standardized coefficient are obtained by carrying out multiple regression of Y with respect to X_i''s. Denote by β_i', $i = 1, 2, 3, 4$, as the standardized coefficient, they are given by

$$\beta_i' = \beta_i \sqrt{\sigma_i / \sigma_Y} \qquad \text{for} i = 1, 2, 3, 4.$$

where σ_i is the standard deviation of X_i, σ_Y is the standard deviation of Y. The values of β_i' are listed in Table 4.

Table 4: Standardized coefficient for each factor

i	1	2	3	4
β_i'	0.0776	0.3093	-0.3395	0.3977

Hence, the order of importance for these factors is

$$\text{total cholesterol} < \text{triglyceride} < \text{insulin} < \text{glycated hemoglobin}.$$

4.3 Improved Regression Model

In the previous section, we find that it is possible that some factors do not affect the blood glucose level, or the influence is negligible. By carrying out the stepwise regression, we find that the influence of total cholesterol level, that is X_1, is negligible.

Thus, we describe Y as

$$Y = \alpha + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \varepsilon = \hat{Y} + \varepsilon.$$

Here, we repeat what we did to the classic multiple regression model, and list the statistics in Table 5. Thus, the multiple linear regression model for the blood glucose level is

$$\hat{Y} = 6.4996 + 0.4023X_2 - 0.2871X_3 + 0.6632X_4.$$

Again, according to the standardized coefficient the glycated hemoglobin level is the most important factor and triglyceride level is the least influential factor.

Table 5: Statistics for the improved model

	coefficient estimator	σ	standardized coefficient	t-value	p-value
intercept	6.4996	2.3962	0	2.713	0.0124
X_2	0.4023	0.1540	0.3541	2.612	0.0156
X_3	-0.2870	0.1117	-0.3601	-2.570	0.0171
X_4	0.6632	0.2303	0.4133	2.880	0.0084

5 Conclusion

In this project, we investigate how the blood glucose level is affected by total cholesterol level, triglyceride level, insulin level, and glycated hemoglobin level. In particular, we describe the interactions by a multiple linear regression model.

At the beginning, we carried out the classic multiple linear regression, and find that

$$\hat{Y}_i = 5.9433 + 0.1424X_{1i} + 0.3515X_{2i} - 0.2706x_{3i} + 0.6382X_{4i}.$$

where Y_i is the blood glucose level, X_{1i} is the total cholesterol level, X_{2i} is the triglyceride level, X_{3i} is the insulin level, and X_{4i} is the glycated hemoglobin level. However, for this model $R^2 = 0.6008$. While a good model has a R^2 close to 1, the R^2 for the classic model indicates that we could improve this model.

In order to improve the result, we investigate each factor's influence by testing hypothesis $H_0 : \beta_i = 0$ for $i = 1, 2, 3, 4$. Moreover, we compare the standardized coefficients for each factor. The result is that insulin level and glycated hemoglobin level have impact on blood glucose bevels, while total cholesterol and triglyceride levels are not that important.

In the end, we carried out the stepwise regression. The result shows the influence of total cholesterol level can be omitted. After kicking out this factor, we repeat the multiple linear regression again. Finally, we suggest that the blood glucose level is affected by levels of triglyceride, insulin, and glycated hemoglobin. In particular, the model is described as

$$\hat{Y} = 6.4996 + 0.4023X_2 - 0.2871X_3 + 0.6632X_4.$$