TASK

Using the data on more than 8,000 patients, determine if a particular type of treatment helps in reducing the hemoglobin A1C level in human body. High level of hemoglobin A1C is associated with diabetes.

SOLUTION

The analysis below is an abridged version of the solution provided to the client. In particular, client's version included the SPSS output of all regression models and tests that we ran. Here we display only the most relevant SPSS output.

TREATING DIABETES

The objective

The purpose of this study is to see if a particular type of treatment helps in reducing the hemoglobin A1C level. We compare two groups of patients, those that were treated and those that were not. The group of treated patients (group A) is composed of more than 6,000 people, while the group of untreated patients (group B) is composed of more than 2,000 people.

When comparing the two groups we want to match the patients according to characteristics F1, F2, F3 and F4. Each of the characteristic is recorded as a numerical variable in our data set. The characteristics represent demographic information, employment history and medical history of the patients. As such, they may be useful in predicting hemoglobin A1C level. They may have influence of their own, so we want to account for that influence. We want to make sure that the difference in hemoglobin A1C levels that we see between the groups can be attributed to the treatment only. We do that by choosing the method of multiple linear regression. In the regression, the hemoglobin A1C level will be the response variable and certain functions of factors F1 – F4 will be the predictors.

In general, the model of multiple linear regression states that the response variable is a linear function of several predictors plus random noise, which has zero mean. Let us rephrase this statement mathematically:

$$Y = B0 + B1 * X1 + ... Bp * Xp + epsilon,$$

where Y is the response variable, X1,..., Xp are the predictors, epsilon is the random noise and B0,...,Bp are the regression coefficients. Typically the regression coefficients are estimated using the method of Ordinary Least Squares or Generalized Least Squares.

When estimating several candidate regression models, one should pay particular attention to the following information:

1) Which regression coefficient estimates are statistically significant? If a regression coefficient estimate is statistically significant, this means that true regression coefficient is unlikely to be 0, i.e. the corresponding predictor variable has influence on the response.

2) What are the values of the regression coefficients of the significant predictors? The values indicate the magnitude of influence of the predictors and its direction.

The significance of regression coefficients is determined by examining the p-values of the associated t-tests, which should be less than 0.05... The process of identifying the best multiple linear regression model can be done according to the following algorithm.

1) Estimate the model containing all the predictors of interest. It is called the full model.

2) Identify the insignificant predictors and drop some or all of them from the next candidate regression model.

3) Estimate the new model. If some of the predictors are still insignificant, drop some or all of them from the next candidate regression model.

4) Keep repeating step 3) until the resulting model is fully significant.

5) Try adding several predictors dropped at earlier stages in groups of one or two. See if they become significant in the new model.

6) Keep repeating step 5) until you get the largest model which is fully significant.

We use the algorithm above in our study. In our case, the response variable is A1C_Level, which codes the hemoglobin A1C level of the patient. For the candidate predictors we choose:

1) variables F1, F1^2, F2, F3 and F4;

2) the interaction of variables F1 and F3, which is defined as F1*F3;

3) variable Group, indicating which group the patient belongs to - if the patient is part of group A then the variable equals 1, otherwise it is 0.

Thus, the equation for the full regression model is:

$$A1C\_Level = B0 + B1 * F1 + B11 * F1^2 + B2 * F2 + B3 * F3 + B4 * F4 + B13 * F1 * F3 + B5 * Group + \\ + epsilon. \quad \mathbf{(*)}$$

Two notes have to be given here. First, we include variable F1^2 as a candidate predictor because we recognize the possibility of A1C_Level having non-linear dependence of F1, for the reasons pertaining to the meaning of variable F1 (not disclosed in this case study). By allowing both F1 and F^2 into the model we allow for a parabolic term in equation (*). This parabolic term is displayed in green. It is a simple non-linear function and has a chance of serving as a good approximation of the true non-linear relationship that we are trying to identify.

Second, we include interaction of F1 and F3 (displayed in blue) because we think that the influence of F1 on A1C_Level may change with different levels of F3. Again, our conjecture is based on the meaning of variables F1 and F3. The regression coefficient of the interaction indicates by how much the increase of A1C_Level with a unit increase in F1 increases with a unit increase in F3 (read this sentence twice and think carefully about what it means).

If we compare different patients with different F1 - F4 and some of them were treated while some were not, the regression is telling us how much of the hemoglobin A1C level of the patient is determined by F1 - F4 and how much is determined by the treatment, or the lack of it. If the regression coefficient of a specific predictor is significantly different from 0, we know that that the predictor has influence on the hemoglobin level. By examining the absolute value of the regression coefficient (or its beta version) we see the magnitude of influence of the predictor on the hemoglobin level. In particular, the regression coefficient of variable Group indicates by how much on average the hemoglobin level of the treated patient is higher than that of the untreated patient. If the regression coefficient is statistically significant and negative, we know that the treatment helps in reducing the hemoglobin level.

Results

Consistent with the estimation algorithm described above, we run several versions of the regression model. We notice that there are very many missing values of variable F4, which reduces the effective size of the data set if F4 is considered. Therefore, in the first part of the analysis we consider models without F4 and in the second part of the analysis we estimate the same models with F4 added. Below is the SPSS output for estimating the full model without F4.

Coefficients[a]

		Unstandardized Coefficients		Standardized Coefficients			95.0% Confidence Interval for B	
Model		B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound
1	(Constant)	9.44	.51		18.62	.000	8.44	10.43
	Group	-4.29	.06	-.62	-73.53	.000	-4.41	-4.18
	F1	.03	.02	.09	1.74	.08	-.004	.07
	F1^2	.000	.000	-.09	-1.62	.11	.000	.000
	F2	.03	.068	.004	.45	.65	-.10	.17
	F3	.81	.34	.12	2.42	.016	.15	1.47
	F1 * F3	-.014	.006	-.12	-2.29	.022	-.026	-.002

a. Dependent Variable: A1C_Level

Obviously, we are unhappy that several predictors are insignificant, and we drop some of them in the subsequent model selection process. After several steps, we arrive at the optimal model without F4, which is displayed below.

Coefficients[a]

		Unstandardized Coefficients		Standardized Coefficients			95.0% Confidence Interval for B	
Model		B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound
1	(Constant)	10.35	.055		189.69	.000	10.24	10.45
	Group	-4.30	.058	-.62	-73.87	.000	-4.41	-4.19
	F3	.71	.26	.12	2.70	.007	.19	1.22
	F1 * F3	-.012	.005	-.11	-2.62	.009	-.021	-.003

a. Dependent Variable: A1C_Level

We see that our original conjecture about non-linear dependence of A1C_Level on F1 was wrong. Or, to be more correct, the dependence structure is clearly not parabolic, because term F1^2 does not belong to the final model. Nonetheless, F1 has influence on A1C_Level, through its interaction with F3. The regression equation can be written as

$$\mathit{A1C\_Level} = 10.35 - 4.30 * \mathit{Group} + 0.71 * \mathit{F3} – 0.012 * \mathit{F1} * \mathit{F3} + \mathit{epsilon}.$$

Variable Group is highly significant. The regression coefficient indicates that, on average, the hemoglobin level of a treated patient is 4.30 units lower than that of an untreated patient. Seems like a promising result.

Next, we move to models with variable F4. The model selection process is similar. The ultimate (optimal) model is displayed below:

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	10.82	.14		77.66	.000	10.55	11.10
	Group	-4.61	.10	-.57	-44.30	.000	-4.82	-4.41
	F3	1.32	.48	.17	2.73	.006	.37	2.26
	F1 * F3	-.022	.008	-.16	-2.68	.007	-.039	-.006
	F4	-.051	.024	-.027	-2.10	.036	-.099	-.003

a. Dependent Variable: A1C_Level

Again, we see dependence of A1C_Level on F1 only through the interaction term F1 * F3. No dependence on F2 whatsoever. The regression equation is the following:

$$\mathit{A1C\_Level} = 10.82 - 4.61 * \mathit{Group} + 1.32 * \mathit{F3} – 0.022 * \mathit{F1} * \mathit{F3} – 0.051 * \mathit{F4} + \mathit{epsilon}.$$

Again, variable Group is highly significant. The regression coefficient indicates that, on average, the hemoglobin level of a treated patient is 4.61 units lower than that of an untreated patient.

The performance of Group is consistent over the two models above. It is interesting to see if the performance is similar when we do not match the groups on age, gender and duration. We run a simple linear regression model with Group as the only predictor:

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	10.36	.049		211.36	.000	10.27	10.46
	Group	-4.31	.058	-.62	-74.37	.000	-4.42	-4.19

a. Dependent Variable: A1C_Level

The regression equation is now

$$\mathit{A1C\_Level} = 10.36 - 4.31 * \mathit{Group} + \mathit{epsilon}.$$

Variable Group is highly significant. The regression coefficient indicates that, on average, the hemoglobin level of a treated patient is 4.31 units lower than that of an untreated patient.

Conclusions

We make the following conclusions.

- F1 does not have non-linear influence on the hemoglobin level. Predictor F1^2 is insignificant in all regressions and, therefore, it should not be in the model... Nonetheless, F1 has influence on the hemoglobin level through its interaction with F3.

- Variable F2 is highly insignificant in all models.

- The models with F4 and without F4 are not much different. The factors which are important in the first case remain important in the second case.

- The best model without F4 is

  *A1C_Level = 10.35 - 4.30 * Group + 0.71 * F3 – 0.012 * F1 * F3 + epsilon.*

  All the predictors are highly significant and the regression coefficients are estimated accurately.

- The best model with F4 is

  *A1C_Level = 10.82 - 4.61 * Group + 1.32 * F3 – 0.022 * F1 * F3 – 0.051 * F4 + epsilon.*

  All the predictors are significant and the regression coefficients are estimated accurately.

- F4 is an important predictor for the hemoglobin level. A unit increase in F4 results in a 0.051 decrease in the hemoglobin level, on average.

- The performance of Group is consistent over different models:

  1) on average, the hemoglobin level of a treated patient is 4.30 units lower than that of an untreated patient, according to the best model without F4;

  2) on average, the hemoglobin level of a treated patient is 4.61 units lower than that of an untreated patient, according to the best model with F4;

  3) on average, the hemoglobin level of a treated patient is 4.31 units lower than that of an untreated patient, according to the linear regression model with Group as the only predictor.

  In all three models, the treated patients have much better life than the untreated patients. The hemoglobin A1C level is significantly lower in the treated group than in the untreated group. The three models predict the similar magnitudes of the effect.

The treatment works. And matching the groups on characteristics F1 – F4 improves the accuracy of the results but does not change the conclusions materially.

Statistical & Financial Consulting by Stanford PhD

consulting@stanfordphd.com