

6. Analysis of Cross-Sectional Studies

Cross-sectional study designs and their variations have been described in the foundation text - *Mother and Child Health: Research Methods* in Chapter 5. This chapter builds on the information contained therein.

The examples described in previous chapters are cross-sectional studies. In all such studies the investigator records observations and measurements on a number of variables at the same time... The objective is usually to make inferences about the effect of one or more explanatory variables on an outcome variable. In real life, however, the relationships between explanatory variables and an outcome are rarely straightforward. Several explanatory variables are usually involved in different ways with the outcome, whilst simultaneously influencing the effects of each other. In such cases inferences derived from bivariate analysis may not reveal the full story, and one stands the risk of drawing oversimplified conclusions. A form of analysis that models the real life situation and takes into account all the explanatory variables jointly together is more informative.

Multiple regression models the data. The regression equation acts as a compact summary of a complex state of affairs. In the clinical situation, where a variety of host factors like age, sex and lifestyle interact with extraneous influences in causing an outcome such a modelling of the data can help provide an insight into a complex situation. With such an insight the clinician is better able to predict outcomes. This is not to say that bivariate analysis has no value, but that by itself such analysis may not be enough. One complements the other. Another advantage of multiple regression analysis is that by taking into account simultaneously several explanatory variables it helps to correct for any possible confounding. In the case of bivariate analysis one would have to resort to stratified analysis to rule out confounders.

Several of these comments about the place of multiple regression analysis of cross-sectional studies in clinical medicine, and the contribution it can make, are now demonstrated by the following example:

Diabetes, Cholesterol and the Treatment of Hypertension

(Feher MD, Rains SGH, Richmond W, et al. *Postgrad.Med.J.*1988;64:926-930)

Diabetes is often associated with high blood pressure and raised cholesterol and triglyceride levels. Hypertension is usually treated with β -blockers. Even though these drugs are the treatment of choice for hypertension they are known to adversely affect the cholesterol profile in normal individuals. This means that treating hypertension with β -blockers in diabetics one runs the risk of making bad cholesterol profile worse, and thereby increase the risk of heart disease. To explore this possibility researchers studied the effects of β -blockers on high density lipoprotein fraction (HDL-2) in 71 hypertensive diabetics while taking into account other determinants of raised HDL-2. These were smoking, alcohol use, age, weight, and blood levels of other metabolically related compounds like triglycerides, C-peptide, and glucose. The subjects were all males. The data are given on the following pages:

HDL	Bta	Alc	Smk	Age	wt	Trigl.	C-pep	Gluc.
0.12	0	1	0	78	19.3	1.1	0.4	6.6
0.08	0	1	0	72	24.6	2.3	1.9	8.6
0.37	0	0	0	61	25.2	0.7	4.1	5.7
0.38	0	1	1	62	24.9	1	3.9	8.7
0.22	0	1	1	55	23.4	2.1	2.5	6.2
0.43	0	0	0	51	24.8	0.9	6.6	9.1
0.27	0	0	0	65	22.2	1.6	0.7	15.1
0.23	0	1	0	71	27.7	2.2	2.4	5.7
0.13	0	1	1	64	26.4	2.2	0.6	11.2
0.28	0	0	0	56	27.8	1.2	1.3	6.8
0.28	0	1	1	54	30.9	2.1	0.9	3.8
0.43	0	0	0	71	21.9	0.8	2.3	7.7
0.27	0	0	1	57	21.2	1.8	1.5	8.2
0.24	0	0	0	64	27.7	1.4	1.1	13.9
0.23	0	1	0	76	26.6	1.6	1.5	7.2
0.26	0	0	0	77	26.7	1.4	4.7	8.8
0.35	0	0	0	63	26.8	1.1	1.7	7.6
0.16	0	0	0	74	30.7	1.6	2.3	10.8
0.34	0	1	0	65	22.2	1.3	0.6	11.8
0.29	0	0	0	83	23.4	1.2	1.8	17.2
0.14	0	1	0	59	24.5	1.4	2.3	10.7
0.48	0	1	0	57	23.7	0.6	2.5	6.9
0.16	0	1	0	73	18.6	1.3	1.5	9.2
0.06	0	1	1	74	32.2	1.7	1.9	8.4
0.3	0	0	0	48	26.3	1.2	1.7	8.3
0.21	0	0	0	55	21.2	2.1	1.7	13.6
0.24	0	1	0	66	30.3	0.8	3.4	6
0.42	0	0	1	52	29.8	0.7	4.5	4
0.2	0	1	0	56	26.9	1.6	1	10.6
0.15	0	0	0	74	23.6	2	2.1	8.3
0.17	1	1	1	61	22.3	2.1	2.4	10.1
0.08	1	1	1	69	25.4	2.4	1.3	11.2
0.07	1	1	0	60	29.3	2.5	1.5	11.8
0.11	1	1	1	55	34.5	3.4	0.8	10.8
0.42	1	0	0	57	25.9	0.6	2.5	6.6
0.23	1	0	0	67	35	1.2	2.1	8.2
0.31	1	1	0	43	26.8	1.7	1.2	10.4
0.17	1	1	0	63	32.1	2.5	3.1	4.3
0.05	1	0	0	68	26.3	1.5	3.2	10.8
0.02	1	1	1	65	34.1	4.1	0.9	8.3
0.28	1	0	0	47	23.4	1.7	1.5	6.8
0.15	1	1	0	50	28.6	2.3	2.5	11.9
0.22	1	1	1	53	28.6	1.3	2.5	10.7
0.27	1	0	0	65	32.2	0.6	2.3	9.7
0.23	1	1	1	68	26.1	0.6	2.3	5.7
0.31	1	0	0	53	26.7	0.6	2.5	8.3
0.15	1	1	0	60	31.2	1.9	1.9	7.2
0.27	1	0	0	59	33.2	1.1	2.9	16.7
0.02	1	1	1	57	28.1	3.9	2.7	8.7
0.14	1	0	0	59	23.9	2.5	1.9	13.6
0.17	1	1	1	69	25.7	0.9	0.8	10.6
0.09	1	1	0	59	23.5	7.4	1.9	13.9
0.02	1	0	1	67	29.6	2.1	5.1	11.9
0.16	1	1	0	66	32.6	2.3	6.1	6
0.28	1	1	0	58	27.8	0.9	1.7	9.1
0.07	1	0	1	54	33.9	1.8	1.7	8.4
0.08	1	0	0	64	24.4	1.9	2.4	9.1
0.1	1	1	0	52	27.5	2.3	2.3	10.9
0.05	1	1	0	48	27.4	2.9	1.7	7.7
0.08	1	1	0	62	32.3	4.7	2.1	8.8
0.13	1	1	1	50	25.9	2.4	1.1	20.7
0.08	1	0	1	56	24	1.7	4.5	10.6
0.02	1	1	0	70	24.5	4.5	0.8	15.1
0.15	1	0	0	58	35.7	1.5	3.2	17
0.17	1	1	0	62	29.6	1.1	3.4	8.4
0.23	1	0	0	62	27.4	1	4.7	8.6
0.17	1	1	0	68	30.3	1.5	3.9	9.2
0.24	1	1	0	53	29.5	0.7	2.4	9.2
0.02	1	0	1	68	29.7	1.5	1.8	11.2
0.03	1	1	0	62	25.7	2.4	4.1	10.8
0.2	1	1	1	58	29.8	0.6	1.1	7.6

The abbreviations in the preceding data file mean as follows:

HDL = level of HDL in mmol/L
 Bta = β - blocker use (0 = No; 1 = Yes)
 Alc = Use of alcohol (0 = No; 1 = Yes)
 Smk = Smoking (0 = No; 1 = Yes).
 Age = Age in years
 Wt = Weight in Kg/m²
 Trigl. = Triglyceride in mmol/L
 C-pep = C-peptide in mU/L
 Gluc. = Glucose in mmol/L

We begin by exploring the data set by means of a correlation matrix

	HDL	Bta	Alc	Smk	Age	wt	Trigl.	C-pep	Gluc.
Bta	-0.460 0.000								
Alc	-0.292 0.014	0.134 0.265							
Smk	-0.222 0.062	0.117 0.331	0.180 0.134						
Age	-0.182 0.129	-0.289 0.014	-0.018 0.882	-0.101 0.401					
wt	-0.213 0.074	0.411 0.000	0.043 0.722	0.122 0.310	-0.117 0.331				
Trigl.	-0.598 0.000	0.279 0.018	0.331 0.005	0.074 0.541	-0.044 0.717	0.064 0.599			
C-pep	0.153 0.203	0.088 0.465	-0.244 0.040	-0.090 0.454	-0.022 0.853	0.157 0.191	-0.187 0.119		
Gluc.	-0.279 0.019	0.197 0.100	-0.124 0.303	-0.049 0.683	0.013 0.915	-0.084 0.489	0.246 0.038	-0.201 0.093	

Cell Contents: Correlation coefficient
 P-Value

There is no evidence of collinearity on the correlation matrix and we next proceed to regression analysis.

[In MINITAB → Stat → Regression → Regression In response box "HDL" In Predictor Box All explanatory variables]

The regression equation is

$$\text{HDL} = 0.711 - 0.0824 \text{ Bta} - 0.0173 \text{ Alc} - 0.0399 \text{ Smk} - 0.00455 \text{ Age} - 0.00214 \text{ wt} - 0.0444 \text{ Trig1.} + 0.00463 \text{ C-pep} - 0.00391 \text{ Gluc.}$$

Predictor	Coef	StDev	T	P	VIF
Constant	0.7110	0.1102	6.45	0.000	
Bta	-0.08244	0.02293	-3.59	0.001	1.5
Alc	-0.01726	0.02121	-0.81	0.419	1.3
Smk	-0.03995	0.02078	-1.92	0.059	1.1
Age	-0.004549	0.001179	-3.86	0.000	1.1
wt	-0.002140	0.002722	-0.79	0.435	1.3
Trigl.	-0.044372	0.009411	-4.71	0.000	1.3
C-pep	0.004633	0.007811	0.59	0.555	1.2
Gluc.	-0.003907	0.003239	-1.21	0.232	1.3

S = 0.07745 R-Sq = 59.5% R-Sq(adj) = 54.3%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	8	0.546958	0.068370	11.40	0.000
Residual Error	62	0.371915	0.005999		
Total	70	0.918873			

Source	DF	Seq SS
Bta	1	0.194174
Alc	1	0.049542
Smk	1	0.016268
Age	1	0.103049
wt	1	0.000221
Trigl.	1	0.170107
C-pep	1	0.004868
Gluc.	1	0.008728

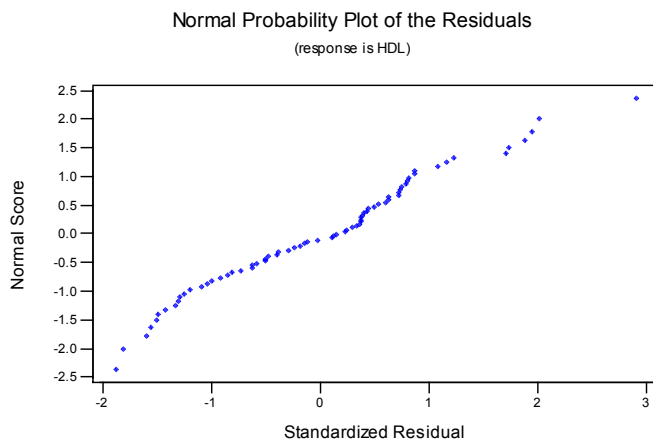
Unusual Observations

Obs	Bta	HDL	Fit	StDev Fit	Residual	St Resid
35	1.00	0.42000	0.27296	0.02541	0.14704	2.01R
52	1.00	0.09000	-0.08128	0.05017	0.17128	2.90RX

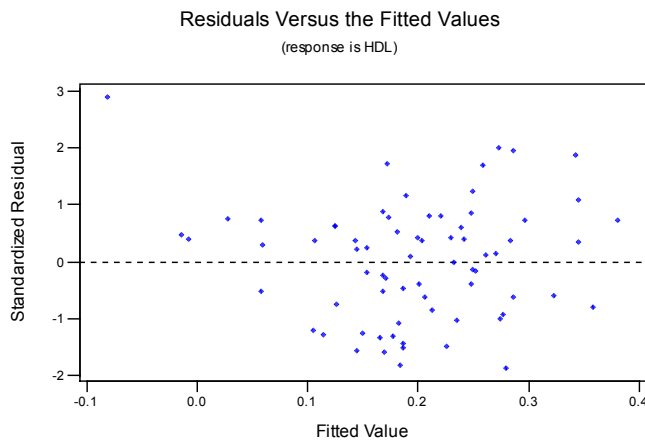
R denotes an observation with a large standardized residual
 X denotes an observation whose X value gives it large influence.

Durbin-Watson statistic = 1.90

VIF is no way near to 10, and the Durbin – Watson statistic is close to 2. Both confirm absence of collinearity. Next we look at some regression diagnostics

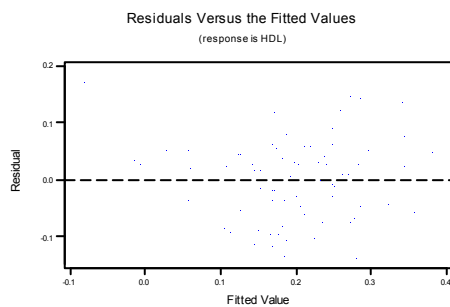


The normal probability plot is not a perfect straight line, especially in the top right hand corner, where some of the data points look odd.



One data point is nearing the value of 3 standardised residual, and this gives rise to some concern.

Note that on the vertical axis we have standardised residuals (and not ordinary residuals as in Chapter 5). Standardised residuals are ordinary residuals \div their estimated standard error. Any data point more than 2 standardised residuals is flagged as an outlier. A plot of ordinary residuals against fitted values is shown below for comparison.



We check for leverage

```

HI1
0.181733 0.070171 0.087033 0.141539 0.121798 0.260605
0.115232 0.087521 0.104175 0.105554 0.176969 0.085107
0.151667 0.104454 0.092070 0.132477 0.068508 0.113086
0.094609 0.212081 0.080127 0.092943 0.126087 0.155566
0.123053 0.118364 0.109092 0.193892 0.095440 0.094907
0.115601 0.099521 0.054418 0.136033 0.107613 0.138236
0.130008 0.109445 0.075766 0.168993 0.177965 0.079234
0.088015 0.097889 0.147971 0.091011 0.067327 0.164192
0.108174 0.090690 0.142762 0.419544 0.174203 0.204639
0.077796 0.142707 0.091118 0.062295 0.098294 0.154688
0.280699 0.161046 0.156353 0.210092 0.073576 0.092060
0.091527 0.092402 0.119365 0.094343 0.116527

```

In all 9 parameters including the constant have been calculated. The upper limit for leverage would be $(3 \times 9) \div 71 = 0.38$. None of the values of HI is above this limit.

We next check for leverage and Cook's D.

```

COOK1
0.055233 0.017102 0.001190 0.052844 0.002236 0.021229
0.002293 0.001456 0.003383 0.004502 0.008490 0.038961
0.000290 0.000204 0.002006 0.003174 0.004205 0.005655
0.017513 0.089326 0.034182 0.040034 0.017258 0.031832
0.010149 0.012641 0.005253 0.031223 0.011705 0.008347
0.005650 0.001018 0.003476 0.009152 0.054115 0.011097
0.012311 0.001919 0.023251 0.003437 0.000529 0.002199
0.002966 0.008008 0.014570 0.001509 0.000460 0.028952
0.003514 0.000382 0.007358 0.676665 0.033693 0.001531
0.006037 0.032843 0.018992 0.008796 0.039839 0.011309
0.005709 0.048034 0.004863 0.002525 0.001429 0.000006
0.000619 0.000258 0.025026 0.027943 0.000157

```

None of the values above is >1

Reading the output

In this example we have two types of variables. The independent variables like Age, Weight, Triglycerides, C-peptide, and Glucose are continuous variables. We also have categorical variables like use of β - blockers, smoking, and drinking alcohol. The latter have been coded as 1 = Yes and 0 = no.

A high level of HDL is protective against heart disease, so a +ve β coefficient for any of the independent variable is beneficial since it would mean a higher level of HDL. Except for C-peptide all the independent variables in the regression equation have negative coefficients, which means that they all bring about a lower level of HDL. In other words, by lowering the HDL levels they make the cholesterol profile worse. The research question of interest is about use of β - blockers (Is treatment of hypertension with these drugs in diabetics harmful?). The other variables are of lesser importance. They are included in order to take care of potential confounders that might obscure the effects of β - blockers. Also recall that the magnitude, the significance and the interpretation of a partial regression coefficient depends upon what other variates are included in the regression equation.

The test for goodness of fit (F ratio) is significant. This means that the regression equation describes the data well. The R-sq. value is 59.5%, which is rather low but significant. This means that the independent variables are not so striking as to make prediction for individual patients. Rather the results tell us about average changes in the sample as a whole.

Returning to the research question we find that the beta coefficient for use of β - blockers has a P value of 0.001, and highly significant. This means that even though β - blockers may lower the blood pressure in diabetics who are hypertensive they have the undesirable effect on HDL also. This means that some other treatment than β - blockers should be considered for the treatment of hypertension in diabetics.

Comment

The main purpose of regression analysis is to investigate the relationship between a response variable (in this case HDL-2) and several explanatory variables. From the T-ratios and their significance levels we see that among the explanatory variables entered in the regression analysis treatment with β -blockers, age, and serum triglycerides have the main effect on HDL-2.

Do the other explanatory variables serve any purpose? In an observational study of the kind under discussion all the subjects satisfying the inclusion criteria were included. Their ages differed, life-style and habits differed, and perhaps also other biological characteristics also differed. In such a situation the regression equation may be looked upon as depicting the relationship between HDL-2 and β -blockers while accounting for a number of confounding variables.

In routine practice it is advisable to check the validity of the model as has been done here, before drawing clinical conclusions based on the results. Residual plots, leverages, and Cook's statistics should be routinely looked at, particularly because in the clinical situation collinearity is likely. Commencing with a correlation matrix prior to the analysis and ending with plots of residuals against fitted values is good practice.

Observations identified as outliers or as influential should be carefully checked for errors or discussed with colleagues who originated the data. In the absence of adequate information the analysis should be performed with and without the suspect observation. Points that have values for the explanatory variables that are very different from those of other points (i.e. have high leverage) have the *potential to dominate* a regression analysis. High leverage points only have the potential to dominate the analysis. Observations that *actually do dominate* are called influential points, and Cook's distance is the commonly used method of measuring how influential an observation is.