Name: Nainan Mathew
Course: Regression Analysis
Session: Summer Session 2015

Correlation of Cardiac Arrest to Patient and Duration of Hospital Care

Introduction

This project analyzes the correlation of the age of the patient with cardiac arrest and duration for which patient spends in the hospital. It also tries to see how this affected by the sex of the patient.

Data

The original data was based on the survey of patients from the Worcester Heart Attack Community Surveillance Study. The patients included here are from the start of the study in December 1975 until 1988, although the study continued until 2006. he data were published in Hosmer D.W. and Lemeshow, S. (1998) Applied Survival Analysis: Regression Modeling of Time to Event Data, John Wiley and Sons Inc., New York, NY. Certain filters have been applied in order to control the variables in play and better demonstrate the understanding of the material learned. We will use 82 patients. The data used can be found in the appendix section.

Parameters

The model uses two quantitative explanatory variables (age and days spent in hospital). The dummy variable represents the sex of the patient (female or male represented as 1 or 0 respectively). In addition, the model expresses interactions between the quantitative and the qualitative explanatory variables. The variables are represented as follows:

$Y \equiv$ Peak Cardiac Enzyme

$X_1 \equiv$ Age (years)

$X_2 \equiv$ Days in hospital

$$D = \begin{cases} 1, & \text{female} \\ 0, & male \end{cases}$$

which results in the following full model equation:

$$Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \gamma D_i + \delta_1 X_{1i} D_i + \delta_2 X_{2i} D_i + \varepsilon_i$$

which in turn leads to the following regression equations for each group:

Female: $Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \gamma + \delta_1 X_{1i} + \delta_2 X_{2i} + \varepsilon_i$
Male: $Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i$

Results and Analysis

The data consists of 82 patients classified along three dimensions:

(i) Sex (male & female) (D)

(ii) Age in years (X_1)

(iii) Days in hospital (X_2)

The following 7 regression models were analyzed to determine the additional predictive power of each of the variables in the model.

Model	Terms	Description
1	$X_{1i}, X_{2i}, D_i, X_{1i} * D_i, X_{2i} * D_i$	Full model
2	$X_{1i}, X_{2i}, D_i, X_{1i} * D_i$	Removes interaction of days in hospital and sex
3	$X_{1i}, X_{2i}, D_i, X_{2i} * D_i$	Removes interaction of age and sex
4	X_{1i}, X_{2i}, D_i	Removes the interaction variables
5	X_{1i}, X_{2i}	Removes the sex variable and the higher order variables
6	$X_{1i}, D_i, X_{1i} * D_i$	Removes days in hospital and higher order variables
7	$X_{2i}, D_i, X_{2i} * D_i$	Removes age and higher order variables

Model 1

Model 1 is also known as the full model and is represented by the following equation:

$$Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \gamma D_i + \delta_1 X_{1i} D_i + \delta_2 X_{2i} D_i + \varepsilon_i$$

Using Regression tool in excel and further simplification we get the following equation for the full model:

$$Y_i = 5568.57 - 49.77 X_{1i} - 19.088 X_{2i} - 3909.67 D_i + 41.518 X_{1i} D_i + 23.48 X_{2i} D_i + \varepsilon_i$$

Regression Statistics	
Multiple R	0.316163496
R Square	0.099959356
Adjusted R Square	0.040746156
Standard Error	1596.139827
Observations	82

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	5	21503878.11	4300776	1.688126	0.147668658
Residual	76	193622338.3	2547662		
Total	81	215126216.5			

The implied regression equations for each group are as follows:

Female: $Y_i = 1658.897 - 8.25054X_{1i} + 4.392692X_{2i} + \varepsilon_i$

Male: $Y_i = 5568.57 - 49.77X_{1i} - 19.088X_{2i} + \varepsilon_i$

From these equations it can be observed that age and days in hospital seems to have a larger impact on the peak cardiac enzyme score for male patient compared to that of female patient.

The total degrees of freedom: 81, which is the number of data points (82 patients) less 1. The number of variables in the full model is the regression degrees of freedom which includes: age, days spent in hospital, sex, sex x age, and sex x days spent in hospital.

The correlation coefficient is 0.31616 which though not significant show a positive correlation.

MS (mean Sum of squares) is determined by taking the ratio of sum of squares and the corresponding degrees of freedom. The incremental F-statistic is in-turn determined by taking the ratio of Regression mean square and Residual Mean Sum of Squares.

Model 2

For Model 2, we remove the interaction of days spent in hospital and sex from the full model.

$Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \gamma D_i + \delta_1 X_{1i} D_i + \varepsilon_i$

The regression statistic obtained using the regression tool in excel is as shown below

Regression Statistics	
Multiple R	0.309915235
R Square	0.096047453
Adjusted R Square	0.049088879
Standard Error	1589.183778
Observations	82

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	4	20662325.16	5165581	2.045366	0.096272861
Residual	77	194463891.3	2525505		
Total	81	215126216.5			

Using Regression tool in excel and further simplification we get the following equation for Model 2:

$Y_i = 5429.954 - 49.6709X_{1i} - 1.33486X_{2i} - 3694.3D_i + 41.15705X_{1i}D_i + \varepsilon_i$

The regression equations for each group are as follows:

Female: $Y_i = \quad 1735.65 - 8.51389X_{1i} - 1.33486X_{2i} + \varepsilon_i$

Male: $Y_i = 5429.954 - 49.6709X_{1i} - 1.33486X_{2i} + \varepsilon_i$

From these equations it can be observed that age seems to have a larger impact on the peak cardiac enzyme score for male patient compared to that of female patient.

Model 3

For Model 3, we remove the interaction of age and sex from the full model. Hence we obtain the equation for Model 3 to be: $Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \gamma D_i + \delta_2 X_{2i} D_i + \varepsilon_i$
The regression statistic obtained using the regression tool in excel is as shown below

Regression Statistics	
Multiple R	0.286546423
R Square	0.082108853
Adjusted R Square	0.034426195
Standard Error	1601.389209
Observations	82

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	4	17663766.78	4415942	1.721986	0.153701503
Residual	77	197462449.7	2564447		
Total	81	215126216.5			

Using Regression tool in excel and further simplification we get the following equation for Model 3:

$Y_i = 3842.596 - 26.1423X_{1i} - 18.835X_{2i} - 810.45D_i + 422.55X_{2i}D_i + \varepsilon_i$

The regression equations for each group are as follows:

Female: $Y_i = 30332.149 - 26.1423X_{1i} + 3.713062X_{2i} + \varepsilon_i$
Male: $Y_i = 3842.596 - 26.1423X_{1i} - 18.8353X_{2i} + \varepsilon_i$

From these equations it can be observed that days spent in hospital seems to have a larger impact on the peak cardiac enzyme score for male patient compared to that of female patient.
F-test will be analyzed, however it is worth noting that the p value is high for this model and correlation coefficient is low which goes to show that the interaction of age and sex contributes significantly to the linear fit of the model

Model 4

For Model 4, we remove the interaction of age& sex and also the interaction of the days spent in hospital & sex of the patient from the full model. Hence we obtain the equation for Model 4 to be: $Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \gamma D_i + \varepsilon_i$
The regression statistic obtained using the regression tool in excel is as shown below

Regression Statistics	
Multiple R	0.280178858
R Square	0.078500192
Adjusted R Square	0.043057892
Standard Error	1594.215373
Observations	82

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	3	16887449.37	5629150	2.214873	0.092990152
Residual	78	198238767.1	2541523		
Total	81	215126216.5			

Using Regression tool in excel and by further simplification we get the following equation for Model 4:
$Y_i = 3842.596 - 26.1423X_{1i} - 18.835X_{2i} - 810.45D_i + 422.55X_{2i}D_i + \varepsilon_i$
The regression equations for each group are as follows:

Female: $Y_i = 3094.402\ -\ 26.2457X_{1i}\ -\ 1.78329X_{2i} + \varepsilon_i$
Male: $Y_i = 3723.866\ -\ 26.2457X_{1i}\ -\ 1.78329X_{2i} + \varepsilon_i$

From these equations it can be observed that days spent in hospital and age have similar effect irrespective of the gender. This model test if the interaction between age and sex and days spent in hospital and sex have any effect in predicting peak cardiac enzyme.

Model 5

For Model 5, we remove the dummy variable (sex) and all its higher order variables (keeping in line with principle of marginality) from the full model. Hence we obtain the equation for Model 5 to be: $Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i$

The regression statistic obtained using the regression tool in excel is as shown below

Regression Statistics	
Multiple R	0.206438897
R Square	0.042617018
Adjusted R Square	0.018379474
Standard Error	1614.640998
Observations	82

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	9168037.885	4584019	1.758306	0.179013293
Residual	79	205958178.6	2607066		
Total	81	215126216.5			

Using Regression tool in excel and further simplification we get the following equation for Model 5:

$Y_i = 3763.762 - 30.9466X_{1i} - 5.60561X_{2i} + \varepsilon_i$

This regression equations is the same for both male and female.

This model gets rid of the dummy variable from the full model, hence it tests if there is any impact sex on Peak Cardiac Enzyme. From these equations it can be observed that days spent in hospital seems to have a larger impact on the peak cardiac enzyme score for male patient compared to that of female patient.F-test will be analyzed, however it is worth noting that the p value is high for this model.

Model 6

For Model 6, we remove the days spent in hospital and all its higher order variables (keeping in line with principle of marginality) from the full model. Hence we obtain the equation for Model 6 to be:

$Y_i = \alpha + \beta_1 X_{1i} + \gamma D_i + \delta_1 X_{1i} D_i + \varepsilon_i$

The regression statistic obtained using the regression tool in excel is as shown below

Regression Statistics	
Multiple R	0.309804619
R Square	0.095978902
Adjusted R Square	0.06120886
Standard Error	1579.02371
Observations	82

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	3	20647578.02	6882526	2.760391	0.047690888
Residual	78	194478638.4	2493316		
Total	81	215126216.5			

Using Regression tool in excel and further simplification we get the following equation for Model 6:

$Y_i = 5419.531 - 49.66 X_{1i} - 3701.769 D_i - 41.211 X_{1i} D_i + \varepsilon_i$

The regression equations for each group are as follows:

Female: $Y_i = 1717.762 - 8.45252 X_{1i} + \varepsilon_i$

Male: $Y_i = 5419.531 - 49.6636 X_{1i} + \varepsilon_i$

This model gets rid of the variable representing the days spent in hospital from the full model. Higher orders of this variable are also removed in accordance with the principle of marginality. Hence it tests if there is any impact the days spent in hospital by the patient on the Peak Cardiac Enzyme.

Model 7

For Model 7, we remove the age and all its higher order variables (keeping in line with principle of marginality) from the full model. Hence we obtain the equation for Model 7 to be:

$Y_i = \alpha + \beta_2 X_{2i} + \gamma D_i + \delta_2 X_{2i} D_i + \varepsilon_i$

The regression statistic obtained using the regression tool in excel is as shown below

Regression Statistics	
Multiple R	0.230758893
R Square	0.053249667
Adjusted R Square	0.016836192
Standard Error	1615.909751
Observations	82

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	3	11455399.31	3818466	1.462362	0.231297948
Residual	78	203670817.1	2611164		
Total	81	215126216.5			

Using Regression tool in excel and further simplification we get the following equation for Model 6:

$Y_i = 1932.845 - 18.55X_{2i} - 907.20523.26D_i + 23.26X_{2i}D_i + \varepsilon_i$

The regression equations for each group are as follows:

Female: $Y_i = 1025.64 + 4.70611X_{2i} + \varepsilon_i$

Male: $Y_i = 1932.845 - 18.5557X_{2i} + \varepsilon_i$

This model gets rid of the variable representing the age from the full model. Higher orders of this variable are also removed in accordance with the principle of marginality. Hence it tests if age of the patient has any impact on the Peak Cardiac Enzyme. Note that the p-value is very high.

Summary

The full model equation is

$Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \gamma D_i + \delta_1 X_{1i} D_i + \delta_2 X_{2i} D_i + \varepsilon_i$

And the various models we have analysed are

Model	Terms	Description
1	$X_{1i}, X_{2i}, D_i, X_{1i} * D_i, X_{2i} * D_i$	Full model
2	$X_{1i}, X_{2i}, D_i, X_{1i} * D_i$	Removes interaction of days in hospital and sex
3	$X_{1i}, X_{2i}, D_i, X_{2i} * D_i$	Removes interaction of age and sex
4	X_{1i}, X_{2i}, D_i	Removes the interaction variables
5	X_{1i}, X_{2i}	Removes the sex variable and the higher order variables
6	$X_{1i}, D_i, X_{1i} * D_i$	Removes days in hospital and higher order variables
7	$X_{2i}, D_i, X_{2i} * D_i$	Removes age and higher order variables

The various scenarios with the corresponding information of degrees of freedom, regression sum of squares, mean squares are shown below

Model	df	Regression Sum of Squares	Mean Regression Sum of Squares
1	5	21,503,878.11	4,300,775.62
2	4	20,662,325.16	5,165,581.29
3	4	17,663,766.78	4,415,941.69
4	3	16,887,449.37	5,629,149.79
5	2	9,168,037.89	4,584,018.94
6	3	20,647,578.02	6,882,526.01
7	3	11,455,399.31	3,818,466.44

Model	Multiple R	Adjusted R2	Standard Error
1	0.316163496	0.04074616	1596.139827
2	0.309915235	0.04908888	1589.183778
3	0.286546423	0.0344262	1601.389209
4	0.280178858	0.04305789	1594.215373
5	0.206438897	0.01837947	1614.640998
6	0.309804619	0.06120886	1579.02371
7	0.230758893	0.01683619	1615.909751

The following table is Analysis of variance table and shows the incremental F-tests of the various variables' effects under study.

ANOVA Table Showing incremental F-test

Source	Model Contrasted	Sum of Squares	df	F-statistic	p-Value
Age	3- 7	6,208,367.47	1	2.40118	0.12514
Days Spent in hospital	2 - 6	14,747.14	1	0.00570	0.93999
Sex	4- 5	7,719,411.48	1	2.98559	0.08782
Ages * Sex	1 -3	3,840,111.33	1	1.48522	0.22650
Days Spent in hospital * Sex	1 -2	841,552.95	1	0.32548	0.56991
Residuals		196,502,026.09	76		
Total		215,126,216.45	81		

The hypotheses tested based on these model contrasts are illustrated below

Source	Models Contrasted	Description
Age	3 – 7	$\beta_1 = 0 \mid \delta_1 = 0$
Days Spent in hospital	2 – 6	$\beta_2 = 0 \mid \delta_2 = 0$
Sex	4 – 5	$\gamma = 0 \mid \delta_1 = \delta_2 = 0$
Ages *Sex	1 – 3	$\delta_1 = 0$
Days Spent in hospital * Sex	1 – 2	$\delta_2 = 0$

F-statistic and p-value (significance of F statistic) helps determine if the null hypothesis is to be rejected or accepted. Low p-value implies that the difference or fluctuations in dimension are random and hence the variable with lower p-value would have an effect on the Peak Cardiac enzyme level.

It is observed from the table above that the lowest p-value of the model contrasted is for sex. This variable has highest F-statistic and the lowest p-value compared to other variables and its corresponding effects. Model 5 removes the dummy variable (variable representing sex) and its higher variables from the full model. Upon comparing this model to the other models it is evident that Model 5 has the highest standard error and lowest correlation coefficient. Combining these results we can conclude that ignoring the sex of the patient leads to a worse linear fit used in predicting the Peak Cardiac Enzyme level for a patient.

We reject the null hypothesis $\gamma = 0 \mid \delta_1 = \delta_2 = 0$ since the p-value comparing the corresponding models 4 and 5 is 0.08579 which is less than 0.1. It can be concluded that sex of the patient contributes heavily in determining the peak cardiac enzyme level.

Upon further analysis we see that the comparison model of 2 and 6 which studies the impact of days spent in the hospital on the peak cardiac enzyme levels, has a high p-value. In fact it is very high and close to 1. Hence, we cannot reject the null hypothesis, $\beta_2 = 0 \mid \delta_2 = 0$. Therefore, it can be interpreted that the days spent in the hospital does not impact the response variable. Similarly other comparison models also have high value though not as high as the previously described comparison model of 2&6. Due to these high p-values, we conclude that their these variable combinations do not impact the peak cardiac enzyme levels.

Conclusion

By analyzing the comparison models and their F-statistic and relating correlation and standard error we find that sex of the patient is strongly correlated with the peak cardiac enzyme level. The days spent in hospital does not seem to much effect on the peak cardiac enzyme level. Though not significant, age of the patients seems to have more impact on the response variable than the days spent in hospital. However, this impact is not strong as is evident from the p-value of over 12%.

APPENDIX

The data used is as follows

	Age	Days in	
Peak cardiac enzyme	(years)	hospital	Sex: Female =1, Male =0
y	X1	x2	D
485	62	1	1
910	78	1	1
320	81	1	1
3290	79	1	1
2500	60	2	1
99	72	2	0
160	83	3	1
66	78	3	0
99	78	4	0
135	84	4	1
210	79	4	1
99	74	4	0
51	92	6	1
99	72	10	1
320	59	10	0
661	77	20	1
43	64	21	0
98	73	2	1
1606	46	3	1
413	90	3	0
3315	80	4	0
292	82	5	0
1610	84	5	0
301	82	6	1
699	64	9	0
613	76	13	1
254	83	14	0
3712	76	14	0
376	61	16	1
22	67	18	0
275	70	1	0
408	76	2	1
1084	69	2	0

1365	86	2	1
4200	68	3	0
1051	80	4	1
182	72	8	1
550	72	8	1
550	68	10	0
1435	62	12	1
533	70	13	0
94	82	15	1
326	81	15	1
649	84	19	1
2660	64	28	1
5330	43	2	0
914	71	2	0
2464	64	2	0
1030	77	3	1
1260	87	3	0
1415	77	7	0
1345	93	12	1
183	69	13	0
917	80	15	0
4160	71	38	0
1576	73	1	1
3040	79	2	1
724	81	3	0
3850	82	5	1
1354	81	7	0
607	50	10	1
2180	97	10	1
868	85	11	1
2350	67	14	1
200	68	20	1
986	84	1	0
5760	71	1	0
2228	78	2	1
9000	56	2	0
456	98	2	1
1765	80	2	1

523	68	2	0
826	75	3	1
6440	86	3	0
2290	50	4	0
514	82	8	1
1854	79	9	0
1209	65	10	1
3172	86	15	0
887	86	17	1
579	84	39	1
2203	74	68	1

Resources

https://www.statcrunch.com/5.0/shareddata.php?keywords=regression&startlimit=30