Chapter 4

Regression Analysis

In chapter 3 it was found that there were difference in the depth and salinity between two boats and 22 occasions.

In this chapter regression analysis is used to develop a predictive model for the salinity measurements. The determinants are depth, latitude, longitude and time. The main determinant of interest is depth. The other factors are covariates that may need to be taken into account in determining the relation between depth and salinity.

4.1 Distributions of Variables

Table 4.1 shows numerical summaries of the variables to be used in the regression model. These summaries include the sample size, mean, standard deviation, skewness and kurtosis coefficients, and the minimum and maximum values for the measurements. Note that for a normal distribution, the skewness and kurtosis coefficients are both 0.

There are 6 measurements on salinity for each sample, 2 samples from each boat, and 22 occasions. Thus we obtained 264 samples. The salinity measurements range from 19 to 29 ppt, and the mean is 23.96 ppt with standard deviation of 1.92 ppt. Depth ranges from 115.5 to 227.5 cm, the mean is 181.01 cm with standard deviation of 29.41 cm. The time ranges from 2 to 134 minutes after 9 am.

The histograms of salinity, depth, latitude, and longitude are shown in Figure 4.1. The distribution of salinity is unimodal with a very slight positive kurtosis. However the skewness is close to zero, which is the value for a normal distribution.

Numerical Summaries: Salinity(ppt)

Variable	Size	Mean	StDev	Skew	Kurt	Min	Max
Salinity(ppt)	264	23.98	1.92	0.16	0.50	19	29
Depth(cm)	264	181.01	29.41	-0.43	-0.53	115.5	227.5
Latitude	264	54.95	0.35	0.02	-0.40	54.24	55.62
Longitude	264	16.34	0.37	0.04	-0.79	15.74	17.01
Boat	264	1.50	0.50	0.00	-2.02	1	2
Occasion	264	11.50	6.36	0.00	-1.21	1	22
Time(minute)	264	60.14	39.81	0.15	-1.08	2	134

Table 4.1: Summaries of observed data

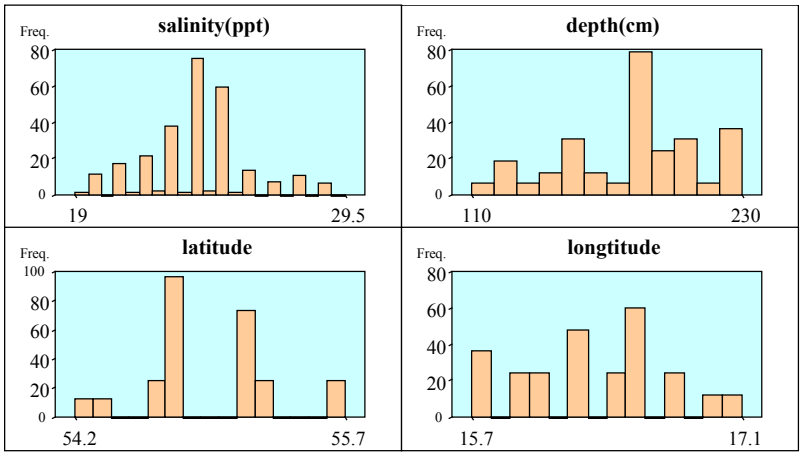

Figure 4.1: The histogram of observed data

4.2 The Correlation between the Variables

The relations between the salinity, time, latitude, longitude and depth are shown in Figure 4.2 using a scatterplot matrix. In Table 4.2, the correlation coefficients are shown for the associations between each variable. There was a relation between *salinity* and *depth* (r = 0.71) and between *salinity* and *latitude* (r = 0.59). In addition, there was a relation between *depth* and *latitude* (r = 0.58). However, there were no relations between *salinity* and *time* and between *salinity* and *longitude* with correlation coefficients approximately 0 (r = –0.06), and also slightly negative relations between *depth* and *time* and between *depth* and *longitude* with correlation coefficients of –0.16 and –0.18, respectively.

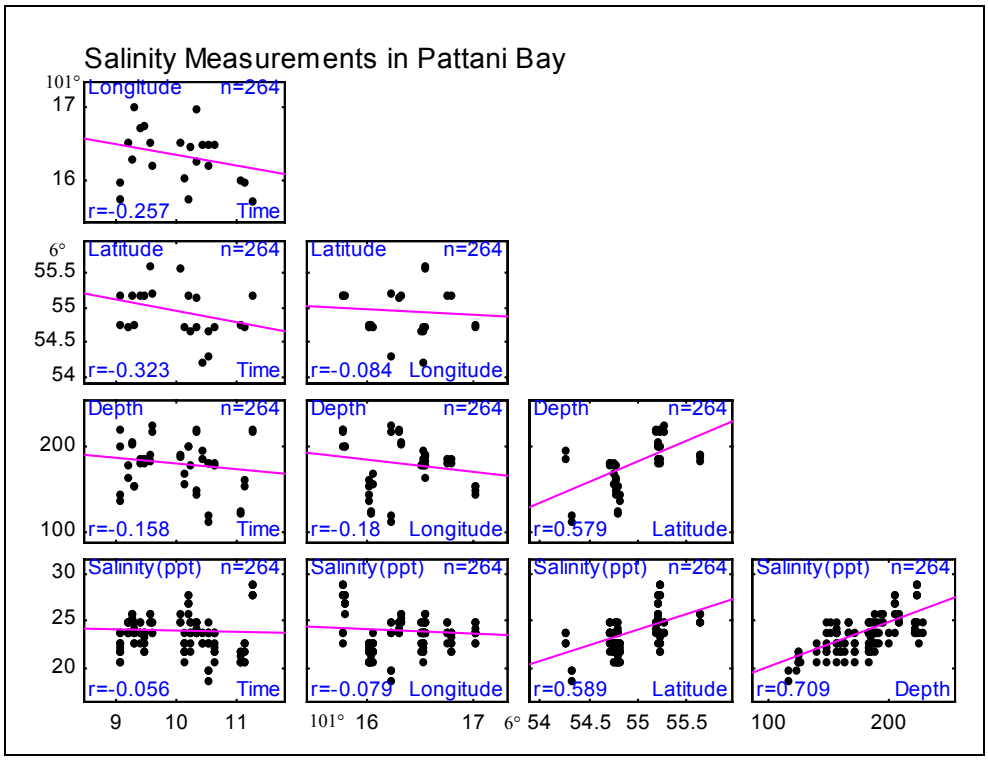

Figure 4.2: Scatterplot matrix between salinity and determinants

Variables	*salinity*	*Depth*	*Latitude*	*Longitude*	*Time*
Salinity	1.00				
Depth	0.71	1.00			
Latitude	0.59	0.58	1.00		
Longitude	-0.06	-0.18	-0.08	1.00	
Time	-0.06	-0.16	-0.32	-0.23	1.00

Table 4.2: The correlation between variables

Figure 4.3 shows the relationships between the variables in the conceptual model postulated in Chapter 1. Thus the variables that have the strongest relationships in this study are salinity and depth, salinity and latitude, and depth and latitude ($r > 0.5$). In addition, there is no evidence of relationships between time and any of the other variables.

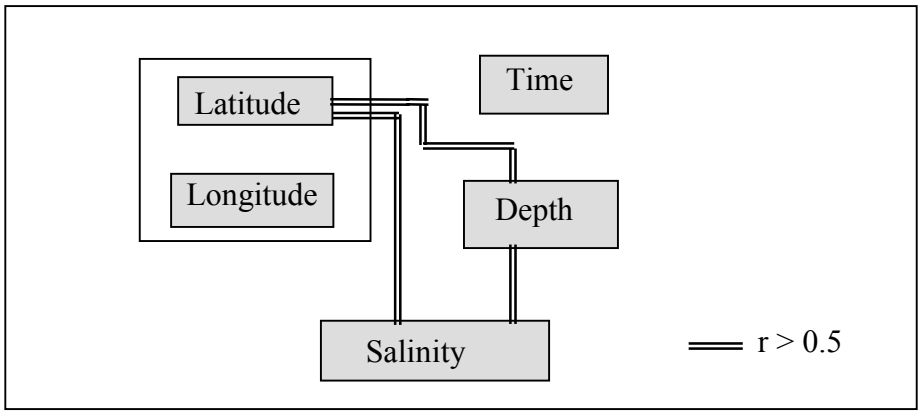

Figure 4.3: The relations between the variables

4.3 Linear Regression Analyses

Since the water depth is expected to be main determinant of salinity, we fitted a linear regression model with salinity taken as the response variable and the depth elevation, time and location as the determinants. For location in this study, we use both latitude and longitude, though from the above result there was only a relationship between salinity, latitude and depth.

Figure 4.4 shows the plot of regression model with salinity measurement and the depth elevation. We can see there was evidence of a linear relationship between salinity measurement and depth elevation. The normal scores plot of salinity measurement for water depth is a smooth linear trend, indicating that the normality assumption is reasonable for these data.

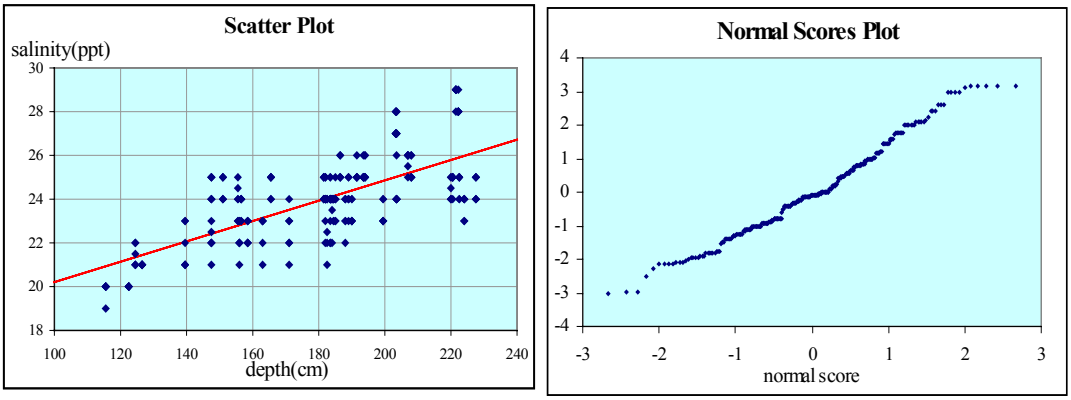

Figure 4.4: The relation between salinity and water depth

Table 4.3 presents the statistics of the regression analyses, in which the r-squared statistic is 50.7%, the residual standard deviation is 1.353 and the p-value is 0.000.

outcome: salinity(ppt) 264 cases

predictor	***coeff***	***SE***	***t***	***p-value***
constant	15.5499	0.5203	29.888	0.0000
depth(cm)	0.0465	0.0028	16.406	**0.0000**

r-sq: 0.507 Resid SS: 479.741 s: 1.3532 df: 262

Table 4.3: The linear regression analysis of salinity measurements

The plot of regression model with salinity measurement as the response and latitude as the determinant is shown in Figure 4.5. The slope of the linear relation between salinity and latitude is positive. It can be seen that the latitude is also associated with salinity. For this case the normal score plot is approximately linear, so that the normality assumption is reasonable for this data.

Table 4.4 shows the statistics of linear regression analyses , in which the r-squared statistic is 35.0%, the residual standard deviation is 1.553 and the p-value is 0.000.

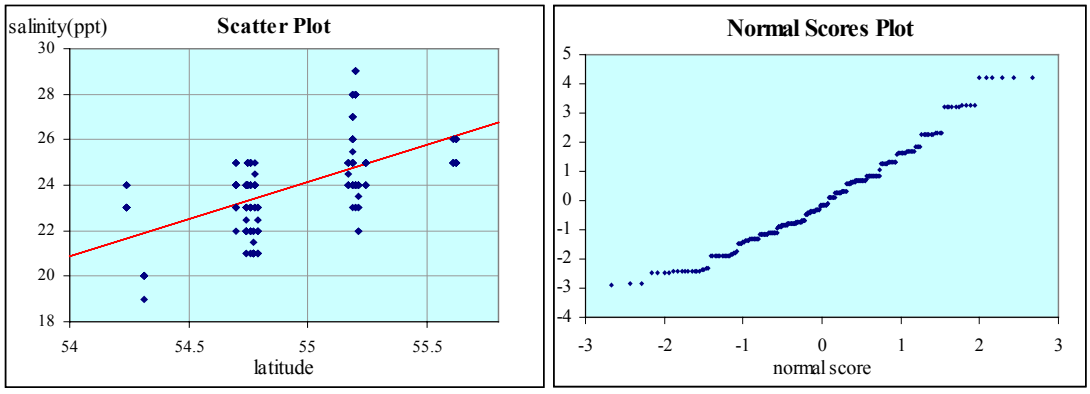

Figure 4.5: The relation between salinity and latitude

outcome: salinity(ppt) 264 cases

predictor	***coeff***	***SE***	***t***	***p-value***
constant	-153.8174	14.9742	-10.272	0.0000
latitude	3.2356	0.2725	11.873	**0.0000**

r-sq: 0.35 Resid SS: 632.336 s: 1.5535 df: 262

Table 4.4: The linear regression analysis of salinity measurements

Figure 4.6 shows the plot of regression model with salinity measurement and the time. There was no evidence of an association between salinity measurement and time. Also there is no evidence of an association between salinity and longitude, as shown in Figure 4.7. For these two cases the normal score plot is approximately linear, so that the normality assumption is reasonable for these data.

The statistics of linear regression analyses between salinity measurement and time are presented in Table 4.5. The r-squared statistic is 0.3%, and the residual standard deviation is 1.9237 with p-value 0.365.

Table 4.6 shows the statistics of linear regression analyses between salinity and longitude, in which the r-squared statistic is 0.6%, and the residual standard deviation is 1.9207 with p-value 0.2013.

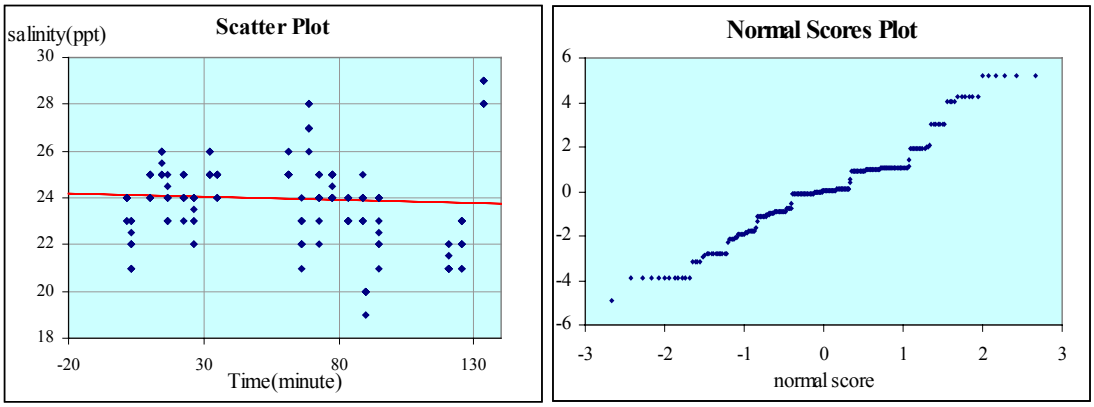

Figure 4.6: The relation between salinity and time

outcome:	salinity(ppt)			264 cases
predictor	***coeff***	***SE***	***t***	***p-value***
constant	24.1378	0.2148	112.395	0.0000
time	-0.0027	0.0030	-0.907	**0.3654**

r-sq: 0.003 Resid SS: 969.547 s: 1.9237 df: 262

Table 4.5: The linear regression analysis of salinity with time

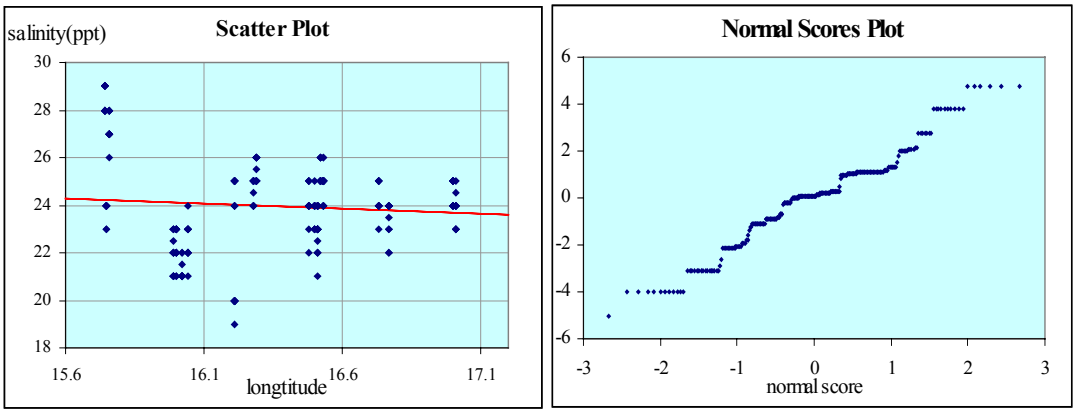

Figure 4.7: The relation between salinity and longitude

outcome:	salinity(ppt)			264 cases
predictor	***coeff***	***SE***	***t***	***p-value***
constant	30.7387	5.2805	5.821	0.0000
longtitude	-0.4140	0.3232	-1.281	**0.2013**

r-sq: 0.006 Resid SS: 966.535 s: 1.9207 df: 262

Table 4.6: The linear regression analysis of salinity with longitude

4.4 Multiple Regression Analysis

We fitted the multiple regression analysis with salinity taken as the response variable and the water depth, time, longitude and latitude as the determinants. First of all we performed the all possible regression by observing the r-squared and adjusted r-squared to determine the feasible determinants as shown in Table 4.7. We can see that for two determinants we got r-squared and adjusted r-squared highest in depth and latitude as determinants ($r^2 = 0.5545$ and $r^2_{adjust} = 0.5511$). For three determinants we got r-squared and adjusted r-squared highest in depth, time and latitude as determinants ($r^2 = 0.5704$ and $r^2_{adjust} = 0.5654$). In addition, for the full model with four determinants, namely depth, time, latitude and longitude, we got the highest r-squared and adjusted r-squared ($r^2 = 0.5779$ and $r^2_{adjust} = 0.5714$).

Model: determinants(s)	*coeff*	*SE*	*r-squared*	*adjust r-squared*
1: depth	0.0466	0.0028	**0.5067**	**0.5049**
2: latitude	3.2356	0.2725	0.3498	0.3474
3: longitude	-0.4140	0.3232	0.0062	0.0024
4: time	-0.1021	0.1788	0.0031	-0.0007
5: depth time	0.0471 0.1686	0.0029 0.1271	0.5100	0.5063
6: depth latitude	0.0363 1.4692	0.0033 0.2777	**0.5545**	**0.5511**
7: depth longitude	0.0472 0.2698	0.0029 0.2313	0.5093	0.5055
8: time latitude	0.4392 3.5041	0.1505 0.2840	0.3704	0.3657
9: latitude longitude	3.2221 -0.1542	0.2738 0.2627	0.3507	0.3457
10: time longitude	-0.2367 -0.5244	0.1845 0.3340	0.0125	0.0049
11: depth time latitude	0.0360 0.3860 1.7242	0.0033 0.1246 0.2853	**0.5704**	**0.5654**
12: depth time longitude	0.0483 0.2326 0.3944	0.0029 0.1327 0.2411	0.5150	0.5094
13: depth latitude longitude	0.0369 1.4613 0.2395	0.0034 0.2777 0.2204	0.5565	0.5514
14: time latitude longitude	0.4535 3.5200 0.0813	0.1581 0.2894 0.2718	0.3706	0.3634
15: depth time latitude longitude	0.0371 0.4695 1.7635 0.4847	0.0033 0.1297 0.2840 0.2259	**0.5779**	**0.5714**

Table 4.7:Regression models for determining salinity (all possible regression)

Table 4.8 shows the results from fitting various multiple regression models containing these four models as determinants. In this case, there is not much difference between

three and four determinants, in the r-squared and the adjusted r-squared ($r^2 = 0.5704$, $r^2_{adjust} = 0.5654$ and $r^2 = 0.5779$, $r^2_{adjust} = 0.5714$, respectively).

Model: determinants(s)	*coeff*	*SE*	*r-squared*	*adjusted r-squared*
1: depth	0.0466	0.0028	0.5067	0.5049
2: depth	0.0363	0.0033	0.5545	0.5511
latitude	1.4692	0.2777		
3: depth	0.0360	0.0033	0.5704	0.5654
time	0.3860	0.1246		
latitude	1.7242	0.2853		
4: depth	0.0371	0.0033	0.5779	0.5714
time	0.4695	0.1297		
latitude	1.7635	0.2840		
longitude	0.4847	0.2259		

Table 4.8: Regression models for determining salinity measurement with high r-squareds

The result of fitting a multiple regression model by using the backward elimination procedure for predicting salinity with all four determinants included is shown in Figure 4.8. This model containing four predictors accounts for 57.79% of the variability in the data, and the residual standard deviation is 1.259. Thus this model indicates that all the determinants including water depth, time, latitude, and longitude are statistically significant.

Figure 4.9 shows the regression analysis of the best model for predicting salinity with three determinants, namely, depth, time and latitude. The model gives a goodness-of-fit, measured by the r-squared statistics of 57.04%, and the residual standard deviation is 1.268. This model indicates that all the determinants including water depth, time and latitude are statistically significant.

In regression modeling, uncomplicated models are preferable. Consequently, we choose the model with three variables as the determinants; these are the depth, time and latitude.

linear regression analysis: response = Salinity(ppt)

predictor	coeff	St.Error	p-value
constant	7.2048	3.8756	0.064158
Depth	0.037069	0.0032871	0
Time	0.46946	0.12974	0.00035621
Latitude	1.7635	0.28398	2.0974e-009
Longitude	0.48472	0.22586	0.032797

r-sq: 0.57788(0.57136) rss: 410.5495 df: 259 sd: 1.259 p-value: 0

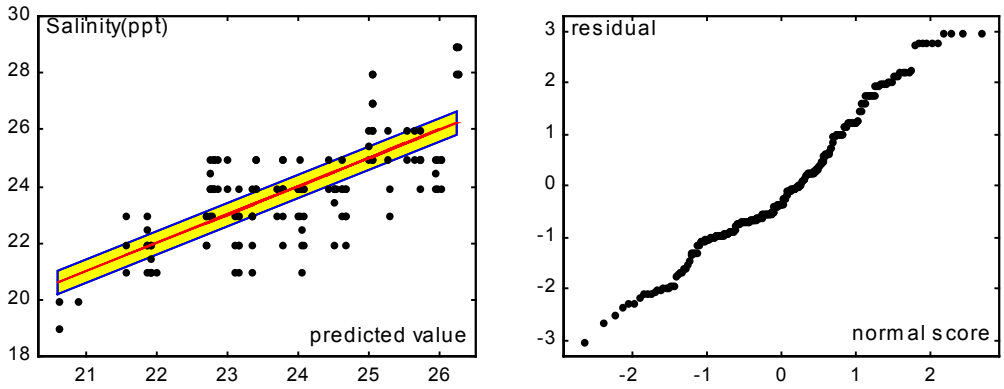

Figure 4.8: Multiple regression analysis of salinity in Pattani Bay (full model)

linear regression analysis: response = Salinity(ppt)

predictor	coeff	St.Error	p-value
constant	15.4461	0.52639	0
Depth	0.035951	0.003268	0
Time	0.38593	0.12462	0.0021695
Latitude	1.7242	0.28534	5.2292e-009

r-sq: 0.57037(0.56542) rss: 417.8499 df: 260 sd: 1.2677 p-value: 0

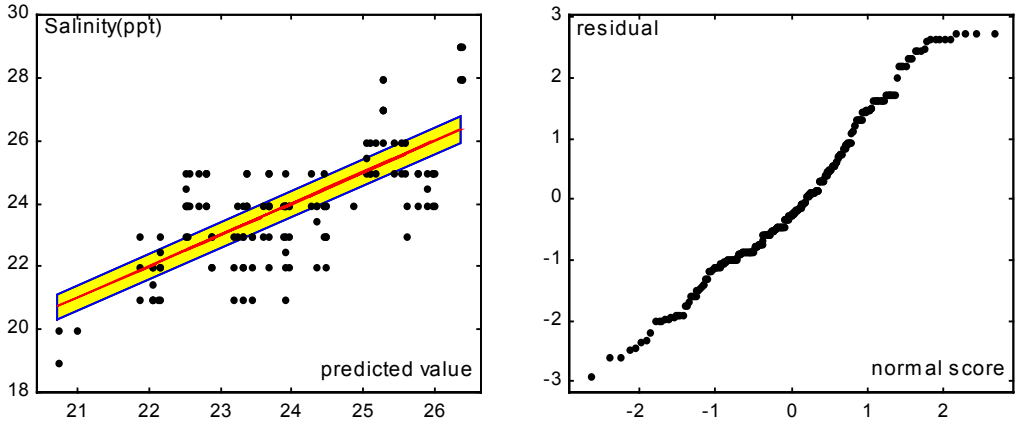

Figure 4.9: Multiple regression analysis of salinity in Pattani Bay (final model)

4.5 Model Assessment

As mentioned earlier, the fitted model needs to meet some criteria to be adequate. These criteria consist of the following.

1. The coefficient of determination (r-squared = r^2);

2. The relation between determinants and outcome in the population is linear (on the left bottom in Figure 4.9); and

3. Residual properties:

- r^2 is 0.5704 in the final model, for the model can explain approximately 57.04% of the variation in salinity. This indicates that model fit is quite high.

- From the plot of residual versus the fitted values (on the right bottom in Figure 4.9), we can see that all points are randomly distributed between –3 and +3. The constant variance assumption for the errors is not violated.

- The residuals versus normal scores plot on the right bottom of Figure 4.9 shows an approximately straight-line trend.

- The conclusion is that the assumptions of the model, such as normality, constant variance and linearity, are reasonable for this model.