

OM7080 Statistical Research Techniques
Final Course Project – Analysis of Data

Final Course Project – Analysis of Data

Arman Kanooni

Capella University

OM7080 – Statistical Research Techniques

Professor: Dr. Mary Robinson

December 8, 2006

Table of Contents

1) Independent Samples t-test.....	3
2) One-Way.....	6
3) Mann-Whitney U test.....	10
4) Kruskal-Wallis H test.....	12
5) Chi Square test.....	14
6) Regression analysis.....	16
Summary.....	20

This is the abstract for the Corwin article in which the RENAL.sav file was collected and analyzed. The citation is in the back of the Norusis text. The RENAL.sav file will be used for both the mid-term AND final projects.

We performed a case-control study to identify risk factors for the development of acute *renal* failure after cardiac operations. Forty-two cases of acute *renal* failure were identified in a total of 572 patients who underwent cardiac operations. They were matched with a control population of patients having cardiac operations without acute *renal* failure. Discriminant analysis performed with preoperative variables revealed preoperative serum creatinine values, concurrent valve and bypass surgery, and age to be significant variables for identifying patients at risk for acute *renal* failure. The use of these three variables in a discriminant model correctly classified 77% of patients. The addition of intraoperative variables did not significantly improve the ability of the model to correctly classify patients. Acute *renal* failure was associated with a significant increase in the number of postoperative complications, mortality, and length of hospitalization and intensive care unit stay. For the FINAL PROJECT, you must do the following:

1) Independent Samples t-test

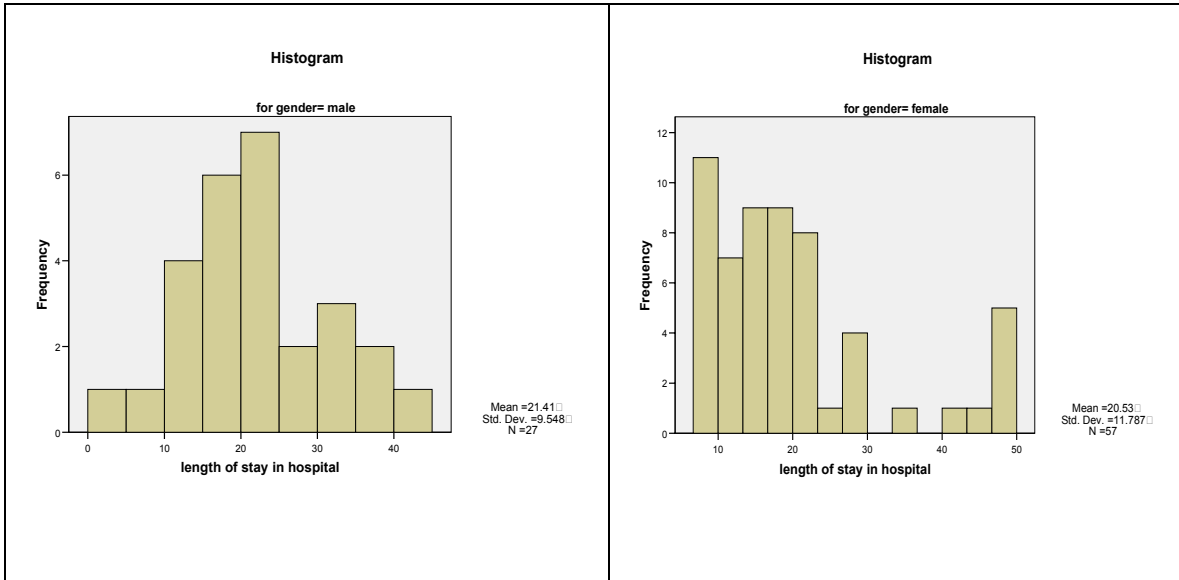
- **Choose any scale variable and any nominal variable with two levels from the RENAL.sav file (e.g., GENDER, DIABETES, ALIVE, BYPASS, etc.).**
- **Write out the assumptions for the Independent Samples t-test and evaluate that the variables are appropriate for the test (*if not, you must select another set of variables*).**
- **State the hypothesis to be tested.**
- **Test for the equality of variances before proceeding.**
- **Conduct an Independent Samples t-test on the scale variable with the nominal variable as a grouping variable.**
- **State the *reject / not reject* decision and conclusion.**
- **State any insights (in English) you can draw from the results; if you reject the null, your insights should follow from the conclusion, and if you fail to reject the null, then your insight should be non-conclusive but possibly offering thoughts about why the results were as they were and whether it makes intuitive sense.**

For this independent samples t-test, the scale variable LOS (Length of Stay in Hospital) and the nominal variable Gender are chosen. For this test, there must be two unrelated samples from normal distribution, or the sample size must be large enough to compensate for non-normality. We are using histograms, stem-and-leaf plots, and boxplots of LOS for Gender to evaluate normality.

OM7080 Statistical Research Techniques
Final Course Project – Analysis of Data

Case Processing Summary

		Cases					
		Valid		Missing		Total	
		N	Percent	N	Percent	N	Percent
length of stay in hospital	male	27	100.0%	0	.0%	27	100.0%
	female	57	100.0%	0	.0%	57	100.0%



Stem-and-Leaf Plots

length of stay in hospital Stem-and-Leaf Plot for gender= male

Frequency	Stem & Leaf
1.00	0 . 3
1.00	0 . 9
4.00	1 . 0013
6.00	1 . 556678
7.00	2 . 2234444
2.00	2 . 78
3.00	3 . 003
2.00	3 . 55
1.00	4 . 4

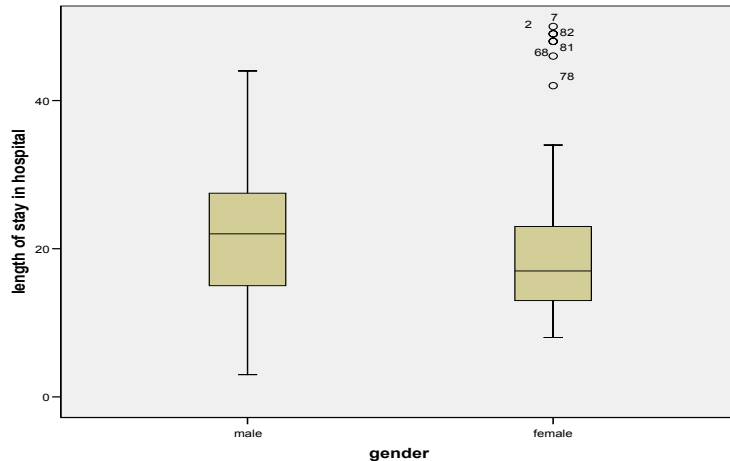
Stem width: 10
Each leaf: 1 case(s)

length of stay in hospital Stem-and-Leaf Plot for gender= female

Frequency	Stem & Leaf
4.00	0 . 8899
20.00	1 . 00000001233333444444
9.00	1 . 555777899
11.00	2 . 00011223333
5.00	2 . 57899
1.00	3 . 4
7.00	Extremes (>=42)

Stem width: 10
Each leaf: 1 case(s)

OM7080 Statistical Research Techniques
Final Course Project – Analysis of Data



The data distribution for the male population seems more normal compared to the female population. It is due to some outliers values in female data. But, these exceptions aside, the assumption for normal distribution is maintained. The assumption about the independence of observations is satisfied since there are two distinct male and female populations. The null hypothesis is that there is no difference in the length of stay in hospital between males and females. The average length of stay in hospital for male is 21.41 days compared to 20.53 for female. The difference is 0.88 day. The observed significance level, while the equal variances assumed, for this difference is 0.736 which is greater than 0.05. Therefore, we can't reject the null hypothesis.

Group Statistics

	gender	N	Mean	Std. Deviation	Std. Error Mean
length of stay in hospital	male	27	21.41	9.548	1.838
	female	57	20.53	11.787	1.561

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
									Lower	Upper
length of stay in hospital	Equal variances assumed	.448	.505	.339	82	.736	.881	2.599	-4.290	6.052
	Equal variances not assumed			.365	62.067	.716	.881	2.411	-3.939	5.701

The review of the independent samples tests showed that there is no difference between men or women regarding the duration of stay in hospital. So, the medical treatment of medical conditions which requires the patients to stay in hospital has nothing to do with the gender of individual. So, it seems the medical treatment affect almost the same both group in term of length of stay in hospital.

2) One-Way

ANOVA

- **Create a new variable AGE3 in which ages from 0-59 are recoded as a 1, ages 60-69 are recoded as a 2, and ages 70 and older are recoded as a 3.**
- **Choose any scale variable from the RENAL.sav file.**
- **Write out the assumptions for the ANOVA and evaluate that the variables are appropriate for the test (*if not, you must select another set of variables*).**
- **State the hypothesis to be tested.**
- **Conduct a One-Way ANOVA on the scale variable with AGE3 as a factor. Use the Bonferroni Post Hoc test *as appropriate*.**
- **State the *reject / not reject* decision and conclusion.**
- **State any insights (in English) you can draw from the results; if you reject the null, your insights should follow from the conclusion, and if you fail to reject the null, then your insight should be non-conclusive but possibly offering thoughts about why the results were as they were and whether it makes intuitive sense.**

The scale variable selected is hours on bypass pump (pumphrs). The assumptions for one-way ANOVA are independence, normality and equality of variance. Since the data is grouped into three separate groups, therefore the assumption of independence is verified. The histogram and stem-and-leaf diagram show a normal distribution. For the equality of variance, we use the boxplot. The boxplot of pumphrs variables for each age group are lineup well considering that the size of sample data.

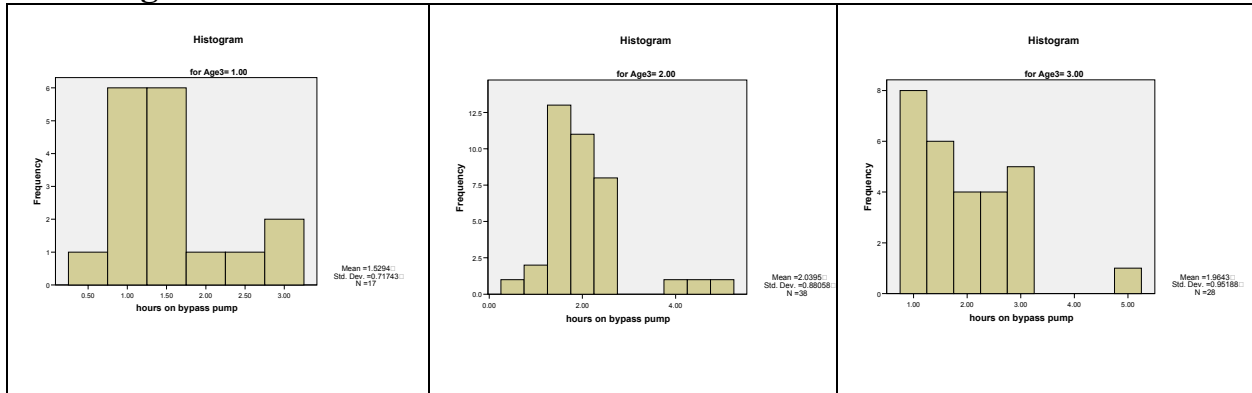
The hypothesis to be tested in a one-way ANOVA is that the average hours on a bypass pump, HOURS ON BYPAS PUMP, is the same for all three age groups, AGE3.

Case Processing Summary

		Cases					
		Valid		Missing		Total	
		N	Percent	N	Percent	N	Percent
hours on bypass pump	Age3 1.00	17	100.0%	0	.0%	17	100.0%
	2.00	38	97.4%	1	2.6%	39	100.0%
	3.00	28	100.0%	0	.0%	28	100.0%

OM7080 Statistical Research Techniques
Final Course Project – Analysis of Data

Histograms



Stem-and-Leaf Plots

hours on bypass pump Stem-and-Leaf Plot for Age3 = 1.00

Frequency	Stem &	Leaf
	.00	0 .
1	1.00	0 . 5
6	6.00	1 . 000000
6	6.00	1 . 555555
1	1.00	2 . 0
3	3.00	Extremes (>=2.5)

Stem width: 1.00
Each leaf: 1 case(s)

hours on bypass pump Stem-and-Leaf Plot for Age3 = 2.00

Frequency	Stem &	Leaf
	.00	0 .
1	1.00	0 . 5
2	2.00	1 . 00
13	13.00	1 . 55555555555555
11	11.00	2 . 000000000000
8	8.00	2 . 55555555
	.00	3 .
	.00	3 .
1	1.00	4 . 0
2	2.00	Extremes (>=4.5)

Stem width: 1.00
Each leaf: 1 case(s)

hours on bypass pump Stem-and-Leaf Plot for Age3= 3.00

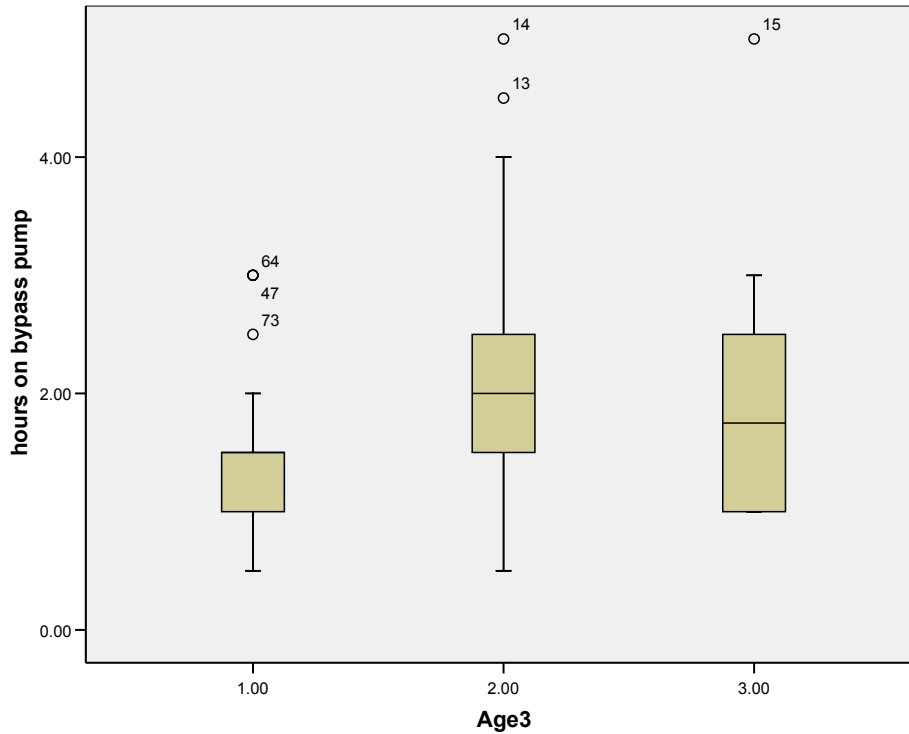
Frequency	Stem &	Leaf
-----------	--------	------

OM7080 Statistical Research Techniques
 Final Course Project – Analysis of Data

```

8.00      1 . 00000000
6.00      1 . 555555
4.00      2 . 0000
4.00      2 . 5555
5.00      3 . 00000
1.00 Extremes (>=5.0)
    
```

Stem width: 1.00
 Each leaf: 1 case(s)



ANOVA

hours on bypass pump

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	3.182	2	1.591	2.073	.132
Within Groups	61.390	80	.767		
Total	64.572	82			

OM7080 Statistical Research Techniques
Final Course Project – Analysis of Data

Multiple Comparisons

Dependent Variable: hours on bypass pump

Bonferroni

(I) Age 3	(J) Age 3	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
1.00	2.00	-.51006	.25561	.148	-1.1351	.1150
	3.00	-.43487	.26934	.331	-1.0935	.2238
2.00	1.00	.51006	.25561	.148	-.1150	1.1351
	3.00	.07519	.21818	1.000	-.4583	.6087
3.00	1.00	.43487	.26934	.331	-.2238	1.0935
	2.00	-.07519	.21818	1.000	-.6087	.4583

The p-value (i.e., SIG) of 0.132 exceeds 0,05, thus the null hypothesis that average hours on bypass pump is the same for all age groups cannot be rejected.

It seems that there are no significance differences between hours on bypass pump with the three age groups. But, the significance between age group 1 and 2 is smallest. It means the younger individuals expend less hours on bypass pump than the older individuals in group 3. This is due to the fact that the younger individual might be in better shape overall than the older individuals.

3) Mann-Whitney U test

- Choose any scale variable and any nominal variable with two levels from the RENAL.sav file (e.g., GENDER, DIABETES, ALIVE, BYPASS, etc.).
- Write out the assumptions for the Mann-Whitney U test.
- State the hypothesis to be tested.
- Conduct a Mann Whitney U test on the scale variable with the nominal variable as a grouping variable.
- State the *reject / not reject* decision and conclusion.
- State any insights (in English) you can draw from the results; if you reject the null, your insights should follow from the conclusion, and if you fail to reject the null, then your insight should be non-conclusive but possibly offering thoughts about why the results were as they were and whether it makes intuitive sense.

We selected the scale variable “number of complications” and the nominal variable “alive discharge”. A non-parametric test (distribution-free) used to compare two independent groups of sampled data. Unlike the parametric t-test, this non-parametric makes no assumptions about the distribution of the data (e.g., normality). This test is an alternative to the independent group t-test, when the assumption of normality or equality of variance is not met. This, like many non-parametric tests, uses the ranks of the data rather than their raw values to calculate the statistic. Since this test does not make a distribution assumption, it is not as powerful as the t-test.

The null hypothesis is that the number of complications has no effect on the number of patient being discharged from the hospital alive.

Mann-Whitney Test

Ranks

	alive discharge	N	Mean Rank	Sum of Ranks
number of complications	no	12	71.46	857.50
	yes	72	37.67	2712.50
	Total	84		

Test Statistics^a

	number of complications
Mann-Whitney U	84.500
Wilcoxon W	2712.500
Z	-5.023
Asymp. Sig. (2-tailed)	.000

a. Grouping Variable: alive discharge

The test statistic for the Mann-Whitney test is U. If U exceeds the critical value for U at some significance level (usually 0.05) it means that there is evidence to reject the null

OM7080 Statistical Research Techniques
Final Course Project – Analysis of Data

hypothesis in favor of the alternative hypothesis. Since the significance level is 0.000 and this is lower than 0.05 (the normal social science cutoff), we reject the null hypothesis.

This seems to be logical! Since when a patient suffer more complications during his or her medical treatment in hospital, the chances of survival become less probable.

4) Kruskal-Wallis H test

- Use the AGE3 variable from the ANOVA you performed.
- Choose any scale variable from the RENAL.sav file.
- Write out the assumptions for the Kruskal-Wallis test.
- State the hypothesis to be tested.
- Conduct a Kruskal-Wallis H test on the scale variable with AGE3 as a factor.
- State the *reject / not reject* decision and conclusion.
- State any insights (in English) you can draw from the results; if you reject the null, your insights should follow from the conclusion, and if you fail to reject the null, then your insight should be non-conclusive but possibly offering thoughts about why the results were as they were and whether it makes intuitive sense.

We use AGE3 variable from the ANOVA performed in problem 2 and NUMBER OF CARDIAC RISK FACTORS for the Kruskal-Wallis H test. This test is a nonparametric alternative to ANOVA. It is computed like Mann-Whitney test, except that there are more groups. The assumptions are that all population means are equal, the data must be independent samples from populations with the same shape. The assumption of equal variances is important.

The null hypothesis to be tested here is that the NUMBER OF CARDIAC RISK FACTORS is the same for all three AGE3 groups. A Kruskal-Wallis H Test was performed on the NUMBER OF CARDIAC RISK FACTORS with AGE3 as a factor.

Kruskal-Wallis Test

Ranks

	Age 3	N	Mean Rank
number of cardiac risk factors	1.00	17	38.29
	2.00	39	40.46
	3.00	28	47.89
	Total	84	

Test Statistics^{a,b}

	number of cardiac risk factors
Chi-Square	2.611
df	2
Asymp. Sig.	.271

a. Kruskal Wallis Test

b. Grouping Variable: Age 3

Since the observed significance level of 0.271 exceeded 0.05 the null hypothesis cannot be rejected. Consequently, it can be concluded that there are no significant differences in

OM7080 Statistical Research Techniques
Final Course Project – Analysis of Data

the number of cardiac risk factors among the three age groups. Looking at the mean ranks, however, it appears there may be some tendency towards a higher number of complications for the older age group.

It seems that the average number of cardiac risk factors associated with the three age groups is generally similar. But for the mean rank increases by age group from 38.29 for group 1 to 40.46 for group 2 and to 47.89 for group 3. This seems to be consistent with the aging process since the older an individual get, the number of cardiac risk factors is increasing.

5) Chi Square test

- Choose any two nominal variables from the RENAL.sav file.
- Write out the assumptions for the Chi Square test.
- State the hypothesis to be tested.
- Create a Cross-tabulation of the two nominal variables.
- Compute the Chi-Square, Lambda and Gamma.
- Discuss the computed values of Lambda and Gamma.
- State the *reject / not reject* decision and conclusion.
- State any insights (in English) you can draw from the results; if you reject the null, your insights should follow from the conclusion, and if you fail to reject the null, then your insight should be non-conclusive but possibly offering thoughts about why the results were as they were and whether it makes intuitive sense.

We selected two nominal variables Gender and Preexisting congestive heart failure for the Chi Square test. The assumptions needed are that the observations must be independent. This means that an individual can appear only once in a table. It also means that the categories of variable can't overlap. The most expected counts must be greater than 5 and none less than 1.

The null hypothesis is that gender and preexisting congestive heart failure are independent. A cross tabulation of the two nominal variables Gender and Preexisting congestive was created as was a Chi square, Lambda and Gamma statistics.

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
preexisting congestive heart failure * gender	84	100.0%	0	.0%	84	100.0%

preexisting congestive heart failure * gender Crosstabulation

Count		gender		Total
		male	female	
preexisting congestive heart failure	no	12	42	54
	yes	15	15	30
Total		27	57	84

OM7080 Statistical Research Techniques
Final Course Project – Analysis of Data

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	6.823 ^b	1	.009		
Continuity Correction ^a	5.608	1	.018		
Likelihood Ratio	6.697	1	.010		
Fisher's Exact Test				.014	.009
Linear-by-Linear Association	6.741	1	.009		
N of Valid Cases	84				

a. Computed only for a 2x2 table

b. 0 cells (.0%) have expected count less than 5. The minimum expected count is 9.64.

Directional Measures

			Value	Asymp. Std. Error ^a	Approx. T ^b	Approx. Sig.
Nominal by Nominal	Lambda	Symmetric	.053	.161	.322	.748
		preexisting congestive heart failure Dependent	.100	.164	.578	.563
		gender Dependent	.000	.203	.000	1.000
	Goodman and Kruskal tau	preexisting congestive heart failure Dependent	.081	.062		.009 ^c
		gender Dependent	.081	.062		.009 ^c

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

c. Based on chi-square approximation

Symmetric Measures

		Value	Approx. Sig.
Nominal by Nominal	Phi	-.285	.009
	Cramer's V	.285	.009
	Contingency Coefficient	.274	.009
N of Valid Cases		84	

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

The observed significance level for the Pearson Chi-Square value is 6.823 with 1 degree of freedom 0.009. This means that if the null hypothesis is true, one expects to see a chi-square value with 1 degree of freedom at least as large as 6.823 about 9 times out of 1000. Since the observed significance level is small, we can reject the null hypothesis. Therefore, it appears that men have more preexisting congestive heart failure than women.

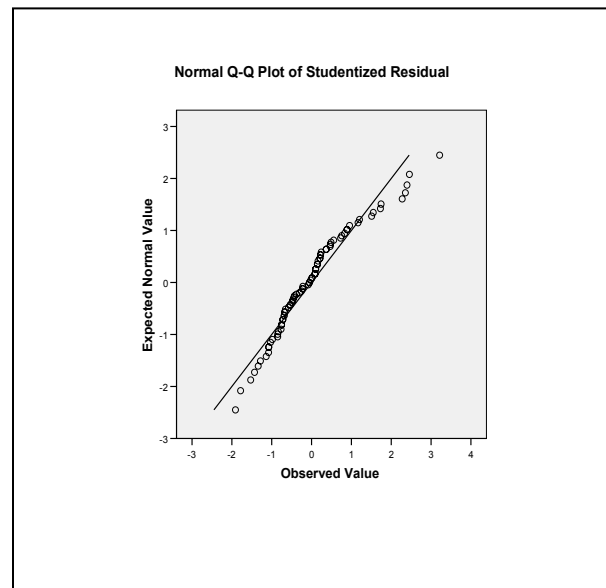
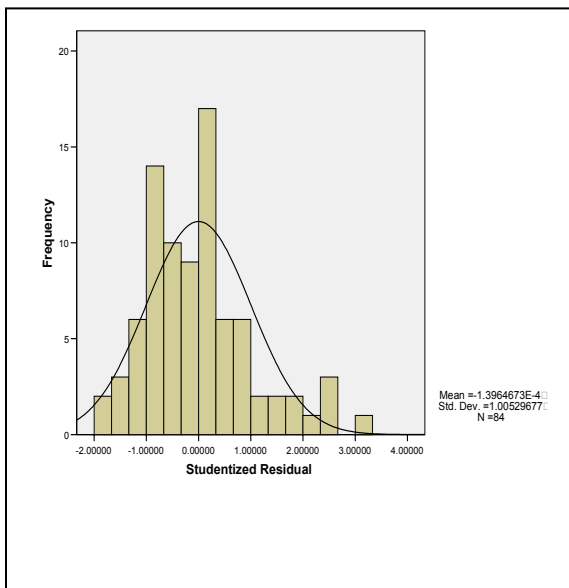
6) Regression analysis

- Choose any two scale variables from the RENAL.sav file.
- Write out the assumptions for the Regression Analysis.
- Conduct tests for the assumptions; evaluate the assumptions by analyzing the residuals.
- Create a scatterplot with regression line.
- State the hypothesis to be tested.
- Conduct a linear regression analysis and test for correlation.
- Discuss the computed R, R-squared and the regression model.
- State the *reject / not reject* decision and conclusion.
- State any insights (in English) you can draw from the results; if you reject the null, your insights should follow from the conclusion, and if you fail to reject the null, then your insight should be non-conclusive but possibly offering thoughts about why the results were as they were and whether it makes intuitive sense.

The two scale variables are Age and Hours in Operation Room. The assumptions for the regression analysis are as follow:

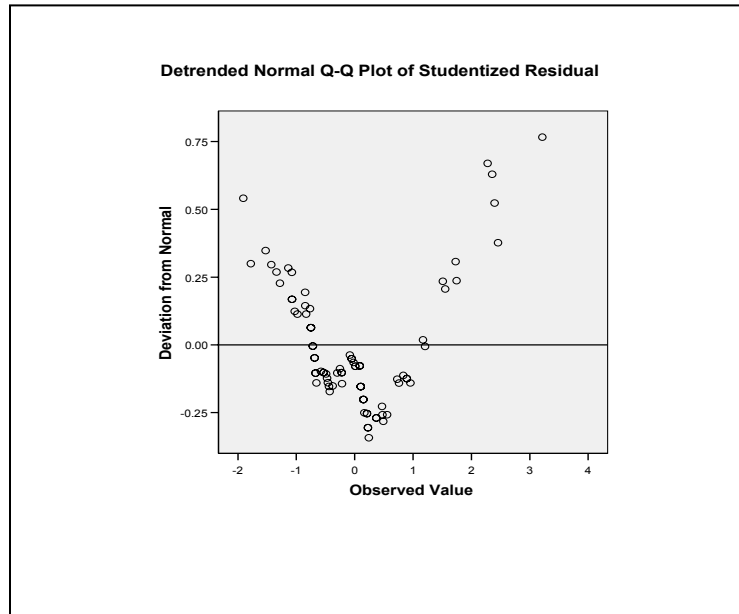
- 1) For each value of the independent variable, the distribution of the values of the dependent variable must be normal.
- 2) The variance of the distribution of the dependent variable must be the same for all values of the independent variable.
- 3) The relationship between the dependent and the independent variable must be linear in the population.
- 4) All of the observations must be independent. Inclusion of one case in the sample must not influence the inclusion of another case.

1) The histogram plot of Studentized Residual and the Q-Q plot of the Studentized residuals show that the distribution of the values of the dependent variable is generally normal.

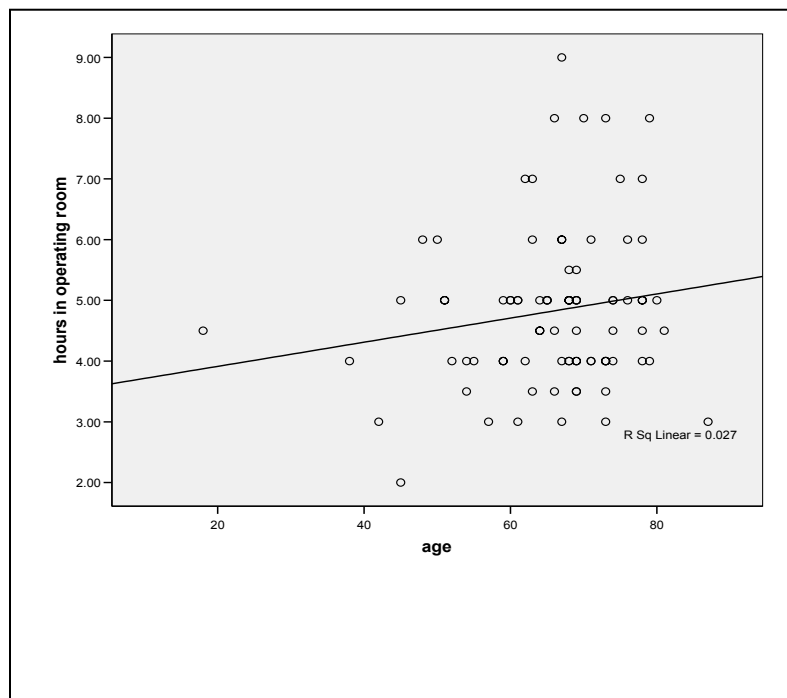


OM7080 Statistical Research Techniques
Final Course Project – Analysis of Data

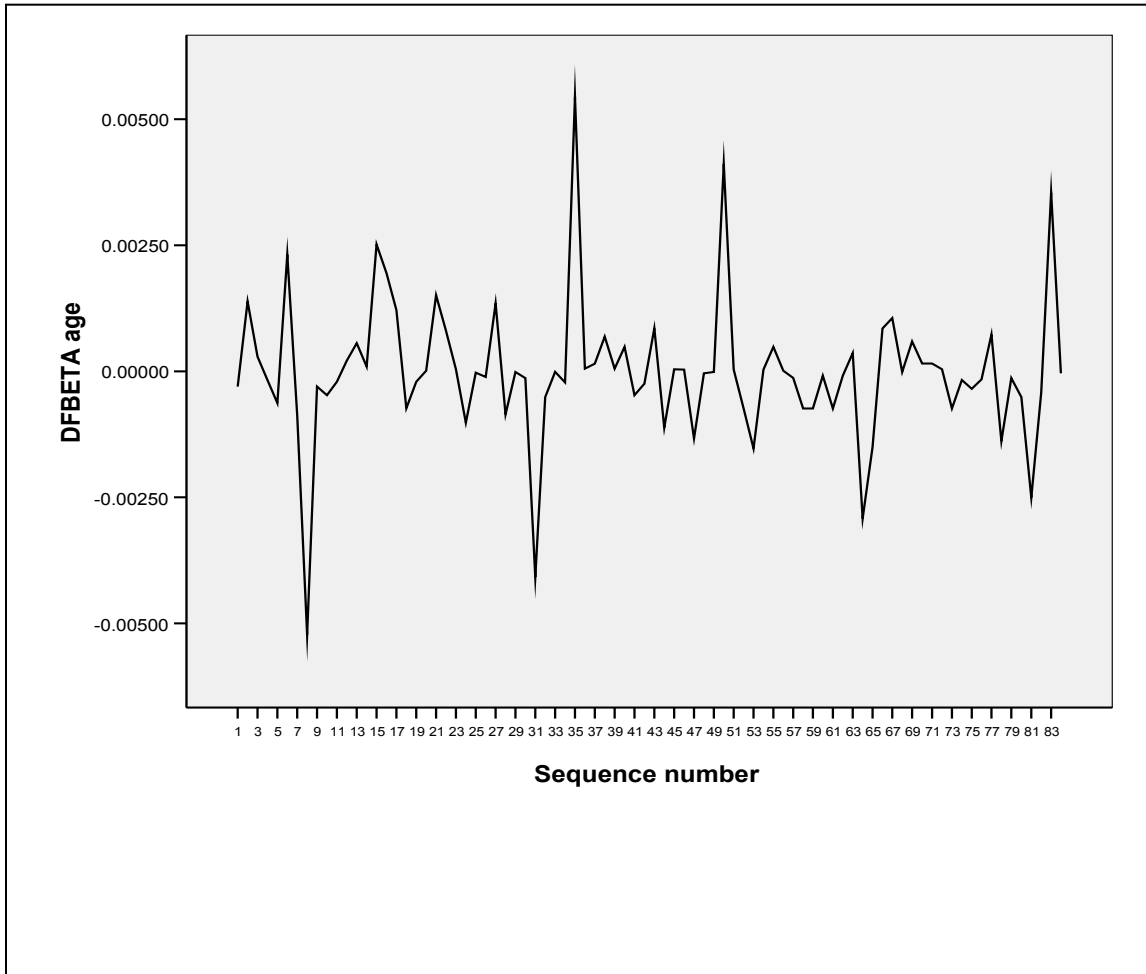
2) The detrended normal Q-Q plot of studentized residual shows the observed value and deviation from normal. The assumption of constant variance is verified because we don't see any pattern around the line of 0.



3) The assumption of linearity between the dependent variable HOURS IN OPERATING ROOM and the independent variable AGE can be verified by looking at the following plot. All the observations are not stacked against the regression line, but we can observe that the data distribution is not parabolic. So, the relationship between Age and HOURS IN OPERATING ROOM is assumed linear.



4) The assumption of independence can be verified by plotting the sequence plot of Studentized residuals against the sequence variable. There is no relationship, hence the independence is assumed.



The null hypothesis is that the population slope is 0. To test, a linear regression is performed for HOURS IN OPERATING ROOM and AGE. We conduct a linear regression analysis to test the null hypothesis for HOURS IN OPERATING ROOM against AGE as shown before in scatter plot during the linearity test.

Regression:

Model Summary^b

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.164 ^a	.027	.015	1.29953

a. Predictors: (Constant), age

b. Dependent Variable: hours in operating room

OM7080 Statistical Research Techniques
Final Course Project – Analysis of Data

ANOVA^b

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	3.841	1	3.841	2.274	.135 ^a
	Residual	138.480	82	1.689		
	Total	142.321	83			

a. Predictors: (Constant), age

b. Dependent Variable: hours in operating room

Coefficients^a

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	3.515	.878		4.005	.000
	age	.020	.013	.164	1.508	.135

a. Dependent Variable: hours in operating room

Residuals Statistics^a

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	3.8729	5.2446	4.8214	.21512	84
Std. Predicted Value	-4.409	1.967	.000	1.000	84
Standard Error of Predicted Value	.142	.645	.187	.072	84
Adjusted Predicted Value	3.6681	5.3841	4.8217	.22773	84
Residual	-2.40963	4.15301	.00000	1.29168	84
Std. Residual	-1.854	3.196	.000	.994	84
Stud. Residual	-1.908	3.215	.000	1.005	84
Deleted Residual	-2.55272	4.20377	-.00025	1.32163	84
Stud. Deleted Residual	-1.940	3.418	.006	1.024	84
Mahal. Distance	.001	19.442	.988	2.345	84
Cook's Distance	.000	.108	.012	.021	84
Centered Leverage Value	.000	.234	.012	.028	84

a. Dependent Variable: hours in operating room

The value of slope is 0.02 with the observed significance of 0.135 which is greater than 0.05. Therefore, the null hypothesis is accepted. So, there is no linear relationship between the hours in operation room and the age of patient. The correlation coefficient of 0.164 yields an R-Square indicating only 3% of the variability in HOURS IN OPERATING ROOM is explained by AGE. This is in accordance with the fact that the null hypothesis can not be rejected.

The regression analysis equation is:

$$\text{HOURS IN OPERATING ROOM} = 3.515 + (0.02 + \text{AGE})$$

OM7080 Statistical Research Techniques
Final Course Project – Analysis of Data

These results showed that the length of time and duration of a major medical procedure in operation room is not really depend on age of the patient and rather it is maybe determined by the nature of medical procedure itself. This seems to make sense in general. But, one can reasonably investigate the duration of time when a patient recuperates from a major operation. In that case, the age might play a role explaining the duration of stay in hospital. Since the

Summary

The analysis of SPSS output for selected variables used during this exercise reveals that following the independent sample t-test, there is no apparent affect of gender differences between men and women in relation with the duration of stay in the hospital knowing that the sample size used in this study showed a proportion of 2 females per 1 male.

Next, we used the one-way ANOVA analysis to see if there is a difference between hours on bypass pump with the three age groups of both men and women together. We concluded that there are no significant differences. There are some minor differences between the youngest and the oldest group. Generally, this makes sense. But, what is important to notice is that the medical procedure bypass pump seems to be not really affected by the age of the participant.

We pay our attention to Mann-Whitney U test of patient suffering number of complications and its affects on their chance of survival. Again, the results are conclusive and indicate that the presence of complications in medical treatment decrease the chance of survival.

Using the Kruskal-Wallis H test of cardiac risk factors associated with the three age groups showed that there are similarities with the fact that the mean value increases by age group which is logical and follow general wisdom that the aging process affects the cardiac risk factors. Investigating further the effect of congestive heart failure among men and women using the Chi Square test, we discovered that men had more preexisting congestive heart failure than women.

Finally, the regression analysis confirms that the amount of time a patient expends in the operation room is not correlated to the age of patient.