PROJECT 4

Kankanamge Harsha

The following data reflect information from 17 U.S. Naval hospitals at various sites around the world. The regressors are workload variables, factors that result in the need for personnel in a hospital. The data are saved in **PROJ4-HOSPITAL**.sav.
The variables are:
y = monthly labor hours
x1 = average daily patient load
x2 = monthly X-ray exposures
x3 = monthly occupied bed-days
x4 = eligible population in the area/1000
x5 = average length of the patient's stay, in days

1. Estimate Pearson's correlation coefficients between the variables. Summarize correlations between the dependent variable and the explanatory variables.

Correlations

		average daily patient load	monthly X-ray exposures	monthly occupied bed-days	eligible population in the area/1000	average length of patient's stay, in days	monthly labor-hours
average daily patient load	Pearson Correlation	1	.907**	1.000**	.936**	.671**	.986**
	Sig. (2-tailed)		.000	.000	.000	.003	.000
	N	17	17	17	17	17	17
monthly X-ray exposures	Pearson Correlation	.907**	1	.907**	.910**	.447	.945**
	Sig. (2-tailed)	.000		.000	.000	.072	.000
	N	17	17	17	17	17	17
monthly occupied bed-days	Pearson Correlation	1.000**	.907**	1	.933**	.671**	.986**
	Sig. (2-tailed)	.000	.000		.000	.003	.000
	N	17	17	17	17	17	17
eligible population in the area/1000	Pearson Correlation	.936**	.910**	.933**	1	.463	.940**
	Sig. (2-tailed)	.000	.000	.000		.061	.000
	N	17	17	17	17	17	17
average length of patient's stay, in days	Pearson Correlation	.671**	.447	.671**	.463	1	.579*
	Sig. (2-tailed)	.003	.072	.003	.061		.015
	N	17	17	17	17	17	17
monthly labor-hours	Pearson Correlation	.986**	.945**	.986**	.940**	.579*	1
	Sig. (2-tailed)	.000	.000	.000	.000	.015	
	N	17	17	17	17	17	17

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

Pearson Correlation is strongly significant between “average daily patient load” and “monthly labor hours”. There is a strong positive linear association between “average daily patient load” and “monthly labor hours”. Because Pearson correlation is 0.986.

Pearson Correlation is a strongly significant between “monthly X-ray exposures” and “monthly labor hours”. There is a strong positive linear association between “monthly X-ray exposures” and “monthly labor hours”. Because Pearson correlation is 0.945.

Pearson Correlation is strongly significant between “monthly occupied bed-days” and “monthly labor hours”. There is strong positive linear association between “monthly occupied bed-days” and “monthly labor hours”. Because Pearson correlation is 0.986.

Pearson Correlation is strongly significant between “eligible population in the area/1000” and “monthly labor hours”. There is strong positive linear association between “eligible population in the area/1000 ” and “monthly labor hours”. Because Pearson correlation is 0.940.

Pearson Correlation is moderate between “average length of the patient's stay, in days” and “monthly labor hours”. There is moderate positive linear association between “average length of the patient's stay, in days” and “monthly labor hours”. Because Pearson correlation is 0.579.

2. Estimate the multiple regression model by entering all 5 independent variables.

Descriptive Statistics

	Mean	Std. Deviation	N
monthly labor-hours	4978.4800	5560.53359	17
average daily patient load	148.2759	161.03858	17
monthly X-ray exposures	18163.24	21278.111	17
monthly occupied bed-days	4480.6182	4906.64206	17
eligible population in the area/1000	106.318	107.9542	17
average length of patient's stay, in days	5.8935	1.58407	17

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	.995[a]	.991	.987	642.08838	.991	237.790	5	11	.000

a. Predictors: (Constant), average length of patient's stay, in days, monthly X-ray exposures, eligible population in the area/1000, monthly occupied bed-days, average daily patient load

b. Dependent Variable: monthly labor-hours

R Square = 0.991 » 99% of variability in “monthly labor hours” is explained by “average daily patient load”,” monthly X-ray exposures”,” monthly occupied bed-days”,” eligible population in the area/1000” and “average length of the patient's stay, in days”.

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	490177488.1	5	98035497.62	237.790	.000[b]
	Residual	4535052.367	11	412277.488		
	Total	494712540.5	16			

a. Dependent Variable: monthly labor-hours

b. Predictors: (Constant), average length of patient's stay, in days, monthly X-ray exposures, eligible population in the area/1000, monthly occupied bed-days, average daily patient load

SSR= 490177488.1
SSE= 4535052.367
S2 =412277.488
F= 237.790 F is strongly significant.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B		Collinearity Statistics	
		B	Std. Error	Beta			Lower Bound	Upper Bound	Tolerance	VIF
1	(Constant)	1962.948	1071.362		1.832	.094	-395.103	4320.999		
	average daily patient load	-15.852	97.653	-.459	-.162	.874	-230.784	199.081	.000	9597.571
	monthly X-ray exposures	.056	.021	.214	2.631	.023	.009	.103	.126	7.941
	monthly occupied bed-days	1.590	3.092	1.403	.514	.617	-5.216	8.395	.000	8933.087
	eligible population in the area/1000	-4.219	7.177	-.082	-.588	.569	-20.014	11.577	.043	23.294
	average length of patient's stay, in days	-394.314	209.640	-.112	-1.881	.087	-855.728	67.099	.234	4.280

a. Dependent Variable: monthly labor-hours

Most of t values are not significant and most of the p-values are also high.

Multiple linear regression equation:

y=ß0+ß1x1+ß2x2+ß3x3+ß4x4+ß5x5+Ɛ

monthly labor hours ^ (estimated)= b0 + b1x1 + b2x2+b3x3+b4x4+b5x5

monthly labor hours ^ (estimated) = -15.852* average daily patient load + 0.056* monthly X-ray exposures+1.590* monthly occupied bed-days -4.219* eligible population in the area/1000 -394.314* average length of the patient's stay, in days +1962.948

3. Interpret the coefficient of multiple determination

   R2 = 0.991

   R Square = 0.99 » 99% Multiple coefficient of determination
   Using “average daily patient load”, “monthly X-ray exposures”,” monthly occupied bed-days”,” eligible population in the area/1000” and” average length of the patient's stay, in days” the model explains 99% of the total sample variation in “monthly labor hours”.

4. Test whether the regression explained by the model is significant at the 0.05 level of significance. Comment on the overall regression and quality of the fitted model.

   H0: ß1=ß2= ß3= ß4= ß5=0
   H1: At least one of the line model coefficient is non zero

   Test statistic F = 237.790

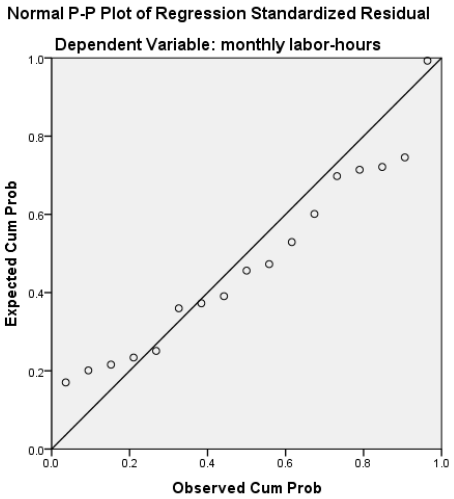
   p-value=0.000

   p-value<0.05 Reject H0

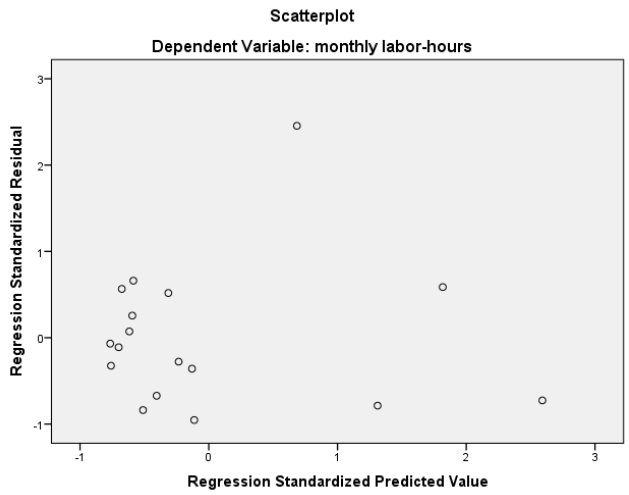
   The data provide evidence that at least one of the model coefficient is non zero. The overall model appears to be useful in predicting monthly labor hours.

   Test statistic also significant F= 237.790 and R^2 is also 0.991 high. Overall model is appeared to be good. But some of regression model coefficient are not significant (about three out of five).

5. Comment on the residual plots and normality tests of residuals.



Most of the data points are on or very closer to the line. We can assume residual are normal.



Residuals are not randomly distributed around the zero line. Therefore, variance of residuals is not homogeneous.

Check for normality

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
Standardized Residual	17	100.0%	0	0.0%	17	100.0%

Descriptives

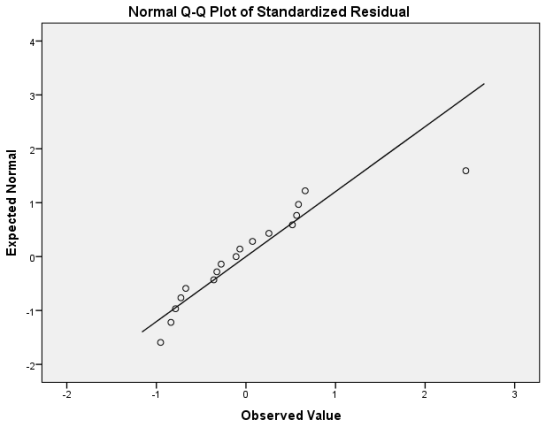
			Statistic	Std. Error
Standardized Residual	Mean		.0000000	.20109992
	95% Confidence Interval for Mean	Lower Bound	-.4263128	
		Upper Bound	.4263128	
	5% Trimmed Mean		-.0834195	
	Median		-.1102237	
	Variance		.688	
	Std. Deviation		.82915620	
	Minimum		-.95304	
	Maximum		2.45459	
	Range		3.40762	
	Interquartile Range		1.24078	
	Skewness		1.601	.550
	Kurtosis		3.858	1.063

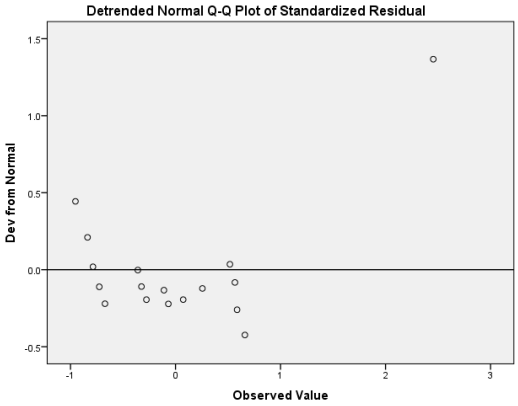
Tests of Normality

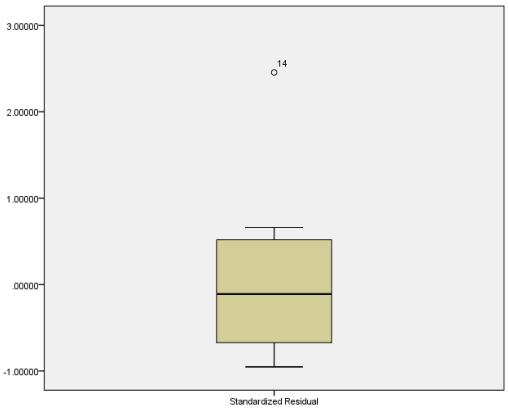
	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Standardized Residual	.154	17	.200*	.859	17	.015

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction







KS Test statistic is 0.154 P-value is 0.2>0.05. Residuals are normal. We can see there is an outlier showing on standardized residual plot. Data point number 14.

6. Test the hypothesis that ßi's = 0 at the 0.05 level of significance against the alternative ßi's ≠ 0 and interpret their significance.

ß1 is coefficient of average daily patient load

H0: ß1=0
H1: ß1≠0

Test statistic t = -0.162 p-value = 0.874
p-value>0.05 Do not reject H0

"average daily patient load" not lineally associated with "monthly labor hours" when others held constant.

ß2 is coefficient of monthly X-ray exposures

H0: ß2=0
H1: ß2≠0

Test statistic t = 2.631 p-value = 0.023
p-value<0.05 Reject H0

We can conclude that mean monthly labor-hours increases as monthly X-ray exposures increases when others variables are held constant. "monthly labor-hours" appears to be linearly associate with "monthly X-ray exposures" when others held constant.

ß3 is coefficient of monthly occupied bed-days
H0: ß3=0
H1: ß3≠0

Test statistic t = 0.514 p-value = 0.617
p-value>0.05 Do not reject H0

"monthly labor hours" not lineally associated with "monthly occupied bed-days" when others held constant.

ß4 is coefficient of eligible population in the area/1000

H0: ß4=0
H1: ß4≠0

Test statistic t = -0.588 p-value = 0.569
p-value>0.05 Do not reject H0

"monthly labor hours" not lineally associated with "eligible population in the area/1000" when others held constant.

ß5 is coefficient of average length of patient's stay, in days

H0: ß5=0
H1: ß5≠0

Test statistic t = -1.881 p-value = 0.087
p-value>0.05 Do not reject H0

"monthly labor hours" not lineally associated with "average length of patient's stay, in days" when others held constant

7. Do you suspect a problem with multicollinearity? Explain.

Yes. There are problems with multicollinearity.

Non-significant t's of coefficient of "average daily patient load", "monthly occupied bed-days", "eligible population in the area/1000" when F-statistic significant

Negative values of coefficient of "average daily patient load", "average length of the patient's stay", "eligible population in the area/1000" (not expected based on positive correlation between "monthly labor-hours" and those variables)

VIF>10 for coefficients of average daily patient load, monthly occupied bed-days, eligible population in the area/1000

8. Use the technique of **forward selection**, **backward elimination** and **stepwise regression** with a 0.05 level of significance to choose a linear regression model. Which variables predict 'monthly labor hours'? (Not necessary to show calculations)

Estimate the new multiple regression model for each method.

Forward selection

Model Summary[c]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.986[a]	.972	.970	957.85555
2	.993[b]	.987	.985	685.16852

a. Predictors: (Constant), monthly occupied bed-days

b. Predictors: (Constant), monthly occupied bed-days, monthly X-ray exposures

c. Dependent Variable: monthly labor-hours

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	480950231.626	1	480950231.626	524.204	.000[b]
	Residual	13762308.863	15	917487.258		
	Total	494712540.489	16			
2	Regression	488140157.951	2	244070078.975	519.900	.000[c]
	Residual	6572382.538	14	469455.896		
	Total	494712540.489	16			

a. Dependent Variable: monthly labor-hours

b. Predictors: (Constant), monthly occupied bed-days

c. Predictors: (Constant), monthly occupied bed-days, monthly X-ray exposures

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B		Collinearity Statistics	
		B	Std. Error	Beta			Lower Bound	Upper Bound	Tolerance	VIF
1	(Constant)	-28.129	319.041		-.088	.931	-708.149	651.892		
	monthly occupied bed-days	1.117	.049	.986	22.895	.000	1.013	1.221	1.000	1.000
2	(Constant)	-68.314	228.446		-.299	.769	-558.282	421.654		
	monthly occupied bed-days	.823	.083	.726	9.919	.000	.645	1.001	.177	5.647
	monthly X-ray exposures	.075	.019	.286	3.913	.002	.034	.116	.177	5.647

a. Dependent Variable: monthly labor-hours

Multiple linear regression equation:
Forward selection method

Variables are: monthly occupied bed-days and monthly X-ray exposures

y=ß0+ß1x1+ß2x2+Ɛ

monthly labor hours ^ (estimated)= b0 + b1x1 + b2x2

monthly labor hours ^ (estimated) = 0.823* monthly occupied bed-days + 0.075* monthly X-ray exposures -68.314

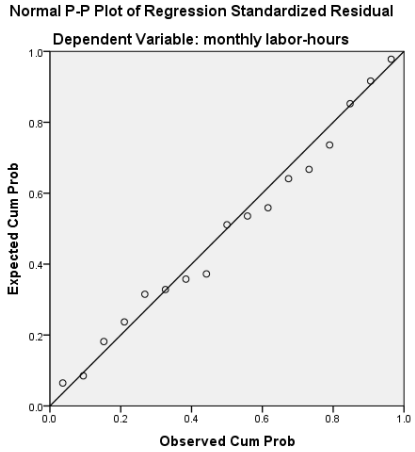
H0: ß1=ß2=0
H1: At least one of the line model coefficient is non zero

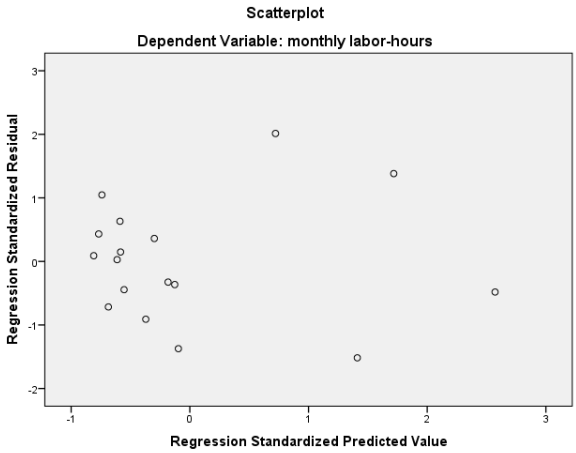
Test statistic F = 519.9

p-value=0.000

p-value<0.05 Reject H0

The data provide evidence that at least one of the model coefficient is non zero. The overall model appears to be useful in predicting monthly labor hours.





Most of the data points are on or very closer to the line. We can assume residual are normal.

Residuals are not randomly distributed around the zero line. Therefore, variance of residuals is not homogeneous.

Normality check for residuals

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
Standardized Residual	17	100.0%	0	0.0%	17	100.0%

Descriptives

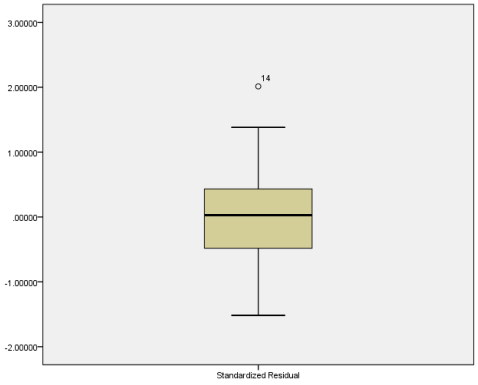
			Statistic	Std. Error
Standardized Residual	Mean		.0000000	.22687130
	95% Confidence Interval for Mean	Lower Bound	-.4809457	
		Upper Bound	.4809457	
	5% Trimmed Mean		-.0275699	
	Median		.0275690	
	Variance		.875	
	Std. Deviation		.93541435	
	Minimum		-1.51694	
	Maximum		2.01319	
	Range		3.53013	
	Interquartile Range		1.13052	
	Skewness		.420	.550
	Kurtosis		.078	1.063

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Standardized Residual	.107	17	.200*	.978	17	.940

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction



KS Test statistic is 0.107 P-value is 0.2. Residuals are normal. We can see there is an outlier showing on standardized residual plot. Data point number 14.

After removed the row 14

Model Summary[d]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.989[a]	.978	.976	856.70736
2	.997[b]	.993	.992	489.12639
3	.998[c]	.996	.995	387.15977

a. Predictors: (Constant), monthly occupied bed-days

b. Predictors: (Constant), monthly occupied bed-days, average length of patient's stay, in days

c. Predictors: (Constant), monthly occupied bed-days, average length of patient's stay, in days, monthly X-ray exposures

d. Dependent Variable: monthly labor-hours

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	453851336.7	1	453851336.7	618.370	.000[b]
	Residual	10275264.93	14	733947.495		
	Total	464126601.6	15			
2	Regression	461016421.4	2	230508210.7	963.483	.000[c]
	Residual	3110180.176	13	239244.629		
	Total	464126601.6	15			
3	Regression	462327889.4	3	154109296.5	1028.131	.000[d]
	Residual	1798712.218	12	149892.685		
	Total	464126601.6	15			

a. Dependent Variable: monthly labor-hours

b. Predictors: (Constant), monthly occupied bed-days

c. Predictors: (Constant), monthly occupied bed-days, average length of patient's stay, in days

d. Predictors: (Constant), monthly occupied bed-days, average length of patient's stay, in days, monthly X-ray exposures

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B		Collinearity Statistics	
		B	Std. Error	Beta			Lower Bound	Upper Bound	Tolerance	VIF
1	(Constant)	-70.230	286.004		-.246	.810	-683.648	543.187		
	monthly occupied bed-days	1.101	.044	.989	24.867	.000	1.006	1.196	1.000	1.000
2	(Constant)	2741.244	539.068		5.085	.000	1576.659	3905.829		
	monthly occupied bed-days	1.223	.034	1.098	36.304	.000	1.150	1.296	.563	1.775
	average length of patient's stay, in days	-572.249	104.567	-.166	-5.473	.000	-798.153	-346.345	.563	1.775
3	(Constant)	1946.802	504.182		3.861	.002	848.284	3045.320		
	monthly occupied bed-days	1.039	.068	.933	15.386	.000	.892	1.187	.088	11.396
	average length of patient's stay, in days	-413.758	98.598	-.120	-4.196	.001	-628.585	-198.931	.397	2.520
	monthly X-ray exposures	.039	.013	.149	2.958	.012	.010	.067	.128	7.828

a. Dependent Variable: monthly labor-hours

	After removed the row 14	**Before removed the row 14**
variables	monthly occupied bed-days, monthly X-ray exposures, average length of patient's stay	monthly occupied bed-days, monthly X-ray exposures
R^2	0.996	0.987
S^2	149892.685	469455.896
F/Sig	1028.131/0.000	519.9/0.000
Residual	1798712.218	6572382.538
t/Sig	15.386/0.00,-4.196/0.001,2.958/0.012	9.919/0.000,3.913/0.002
VIF	11.396, 2.520, 7.828	5.647, 5.647

After data row 14 removed model has improved a lot. S^2 and Residual have lower value and R^2, F have higher values after removed row 14.

So final forward selection model:

monthly labor hours ^ (estimated) = 1.039* monthly occupied bed-days + 0.039* monthly X-ray exposures -413.758*average length of patient's stay in days+1946.802

H0: ß1=ß2= ß3=0
H1: At least one of the line model coefficient is non zero

Test statistic F = 1028.131

p-value=0.000

p-value<0.05 Reject H0

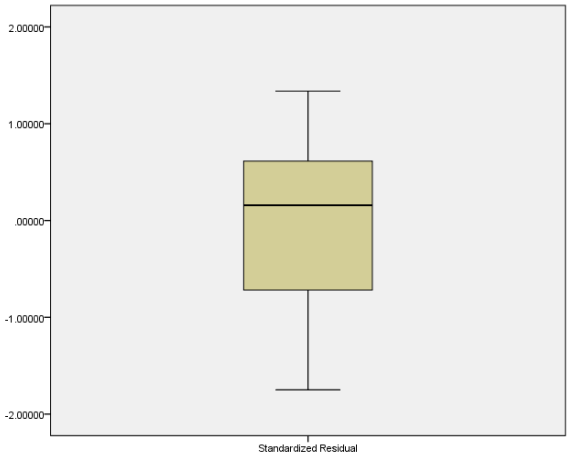
The data provide evidence that at least one of the model coefficient is non zero. The overall model appears to be useful in predicting monthly labor hours.

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Standardized Residual	.132	16	.200*	.968	16	.806

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction



Residuals are normal and no outliers on the plot.

backward elimination

Model Summary[d]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.995[a]	.991	.987	642.08838
2	.995[b]	.991	.988	615.48868
3	.995[c]	.990	.988	614.77942

a. Predictors: (Constant), average length of patient's stay, in days, monthly X-ray exposures, eligible population in the area/1000, monthly occupied bed-days, average daily patient load

b. Predictors: (Constant), average length of patient's stay, in days, monthly X-ray exposures, eligible population in the area/1000, monthly occupied bed-days

c. Predictors: (Constant), average length of patient's stay, in days, monthly X-ray exposures, monthly occupied bed-days

d. Dependent Variable: monthly labor-hours

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	490177488.1	5	98035497.62	237.790	.000[b]
	Residual	4535052.367	11	412277.488		
	Total	494712540.5	16			
2	Regression	490166624.6	4	122541656.2	323.477	.000[c]
	Residual	4545915.844	12	378826.320		
	Total	494712540.5	16			
3	Regression	489799142.0	3	163266380.7	431.975	.000[d]
	Residual	4913398.503	13	377953.731		
	Total	494712540.5	16			

a. Dependent Variable: monthly labor-hours

b. Predictors: (Constant), average length of patient's stay, in days, monthly X-ray exposures, eligible population in the area/1000, monthly occupied bed-days, average daily patient load

c. Predictors: (Constant), average length of patient's stay, in days, monthly X-ray exposures, eligible population in the area/1000, monthly occupied bed-days

d. Predictors: (Constant), average length of patient's stay, in days, monthly X-ray exposures, monthly occupied bed-days

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			95.0% Confidence Interval for B		Collinearity Statistics	
		B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound	Tolerance	VIF
1	(Constant)	1962.948	1071.362		1.832	.094	-395.103	4320.999		
	average daily patient load	-15.852	97.653	-.459	-.162	.874	-230.784	199.081	.000	9597.571
	monthly X-ray exposures	.056	.021	.214	2.631	.023	.009	.103	.126	7.941
	monthly occupied bed-days	1.590	3.092	1.403	.514	.617	-5.216	8.395	.000	8933.087
	eligible population in the area/1000	-4.219	7.177	-.082	-.588	.569	-20.014	11.577	.043	23.294
	average length of patient's stay, in days	-394.314	209.640	-.112	-1.881	.087	-855.728	67.099	.234	4.280
2	(Constant)	2032.188	942.075		2.157	.052	-20.417	4084.793		
	monthly X-ray exposures	.056	.020	.215	2.755	.017	.012	.100	.126	7.926
	monthly occupied bed-days	1.088	.153	.960	7.095	.000	.754	1.423	.042	23.927
	eligible population in the area/1000	-5.004	5.081	-.097	-.985	.344	-16.074	6.066	.079	12.706
	average length of patient's stay, in days	-410.083	178.078	-.117	-2.303	.040	-798.082	-22.084	.298	3.361
3	(Constant)	1523.389	786.898		1.936	.075	-176.600	3223.378		
	monthly X-ray exposures	.053	.020	.203	2.637	.021	.010	.096	.129	7.737
	monthly occupied bed-days	.978	.105	.863	9.305	.000	.751	1.206	.089	11.269
	average length of patient's stay, in days	-320.951	153.192	-.091	-2.095	.056	-651.902	10.001	.401	2.493

a. Dependent Variable: monthly labor-hours

Multiple linear regression equation:
backward elimination method

Variables are: monthly occupied bed-days, monthly X-ray exposures, average length of patient's stay, in days

y=ß0+ß1x1+ß2x2+ ß3x3+Ɛ

monthly labor hours ^ (estimated)= b0 + b1x1 + b2x2+ b3x3

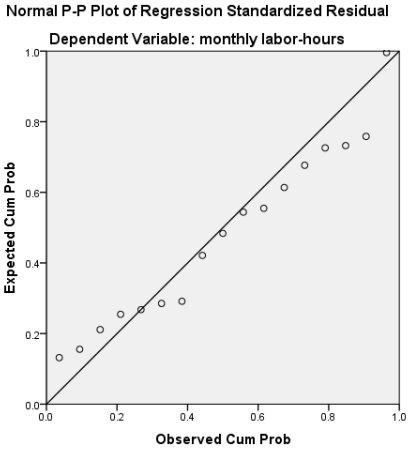
monthly labor hours ^ (estimated) = 0.978* monthly occupied bed-days + 0.053* monthly X-ray exposures -320.951*average length of patient's stay, in days+1523.389

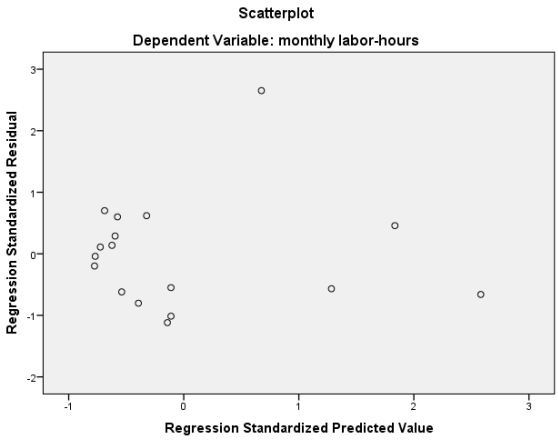
H0: ß1=ß2= ß3=0
H1: At least one of the line model coefficient is non zero

Test statistic F = 431.975

p-value=0.000
p-value<0.05 Reject H0

The data provide evidence that at least one of the model coefficient is non zero. The overall model appears to be useful in predicting monthly labor hours.





Most of the data points are on or very closer to the line. We can assume residual are normal.

Residuals are not randomly distributed around the zero line. Therefore, variance of residuals is not homogeneous.

Normality check for residuals

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
Standardized Residual	17	100.0%	0	0.0%	17	100.0%

Descriptives

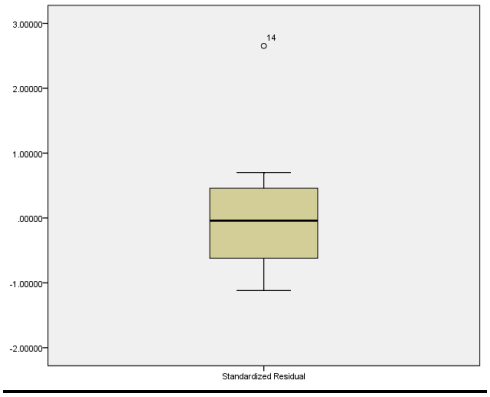
			Statistic	Std. Error
Standardized Residual	Mean		.0000000	.21861866
	95% Confidence Interval for Mean	Lower Bound	-.4634509	
		Upper Bound	.4634509	
	5% Trimmed Mean		-.0852247	
	Median		-.0407111	
	Variance		.813	
	Std. Deviation		.90138782	
	Minimum		-1.11813	
	Maximum		2.65217	
	Range		3.77031	
	Interquartile Range		1.16925	
	Skewness		1.523	.550
	Kurtosis		3.739	1.063

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Standardized Residual	.159	17	.200*	.872	17	.023

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction



KS Test statistic is 0.159 P-value is 0.2. Residuals are normal. We can see there is an outlier showing on standardized residual plot. Data point number 14.

After removed the row 14

Model Summary[d]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.998[a]	.997	.995	399.71176
2	.998[b]	.997	.995	381.55538
3	.998[c]	.996	.995	387.15977

a. Predictors: (Constant), average length of patient's stay, in days, monthly X-ray exposures, eligible population in the area/1000, monthly occupied bed-days, average daily patient load

b. Predictors: (Constant), average length of patient's stay, in days, monthly X-ray exposures, eligible population in the area/1000, monthly occupied bed-days

c. Predictors: (Constant), average length of patient's stay, in days, monthly X-ray exposures, monthly occupied bed-days

d. Dependent Variable: monthly labor-hours

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	462528906.7	5	92505781.34	578.995	.000[b]
	Residual	1597694.925	10	159769.492		
	Total	464126601.6	15			
2	Regression	462525172.1	4	115631293.0	794.255	.000[c]
	Residual	1601429.550	11	145584.505		
	Total	464126601.6	15			
3	Regression	462327889.4	3	154109296.5	1028.131	.000[d]
	Residual	1798712.218	12	149892.685		
	Total	464126601.6	15			

a. Dependent Variable: monthly labor-hours

b. Predictors: (Constant), average length of patient's stay, in days, monthly X-ray exposures, eligible population in the area/1000, monthly occupied bed-days, average daily patient load

c. Predictors: (Constant), average length of patient's stay, in days, monthly X-ray exposures, eligible population in the area/1000, monthly occupied bed-days

d. Predictors: (Constant), average length of patient's stay, in days, monthly X-ray exposures, monthly occupied bed-days

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			95.0% Confidence Interval for B		Collinearity Statistics	
		B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound	Tolerance	VIF
1	(Constant)	2270.415	670.786		3.385	.007	775.811	3765.020		
	average daily patient load	-9.297	60.810	-.274	-.153	.882	-144.790	126.196	.000	9334.456
	monthly X-ray exposures	.041	.014	.159	3.006	.013	.011	.072	.124	8.077
	monthly occupied bed-days	1.413	1.925	1.269	.734	.480	-2.877	5.703	.000	8684.208
	eligible population in the area/1000	-3.223	4.474	-.064	-.720	.488	-13.191	6.745	.043	23.005
	average length of patient's stay, in days	-467.861	131.627	-.135	-3.554	.005	-761.143	-174.578	.237	4.213
2	(Constant)	2311.275	587.301		3.935	.002	1018.634	3603.917		
	monthly X-ray exposures	.041	.013	.159	3.157	.009	.012	.070	.124	8.067
	monthly occupied bed-days	1.119	.095	1.005	11.736	.000	.909	1.329	.043	23.367
	eligible population in the area/1000	-3.682	3.163	-.073	-1.164	.269	-10.645	3.280	.079	12.624
	average length of patient's stay, in days	-477.169	111.398	-.138	-4.283	.001	-722.354	-231.984	.302	3.311
3	(Constant)	1946.802	504.182		3.861	.002	848.284	3045.320		
	monthly X-ray exposures	.039	.013	.149	2.958	.012	.010	.067	.128	7.828
	monthly occupied bed-days	1.039	.068	.933	15.386	.000	.892	1.187	.088	11.396
	average length of patient's stay, in days	-413.758	98.598	-.120	-4.196	.001	-628.585	-198.931	.397	2.520

a. Dependent Variable: monthly labor-hours

	After removed the row 14	**Before removed the row 14**
R^2	0.996	0.990
S^2	149892.685	377953.731
F/Sig	1028.131/0.000	431.975/0.000
Residual	1798712.218	4913398.503
t/Sig	15.386/0.00,-4.196/0.001,2.958/0.012	2.637/0.21,9.305/0.000,-2.095/0.056

After data row 14 removed model has improved. S^2 and Residual have lower value and R^2, F have higher values after removed row 14.

So final backward elimination model:

monthly labor hours ^ (estimated) = 1.039* monthly occupied bed-days + 0.039* monthly X-ray exposures -413.758*average length of patient's stay in days+1946.802

H0: ß1=ß2= ß3=0
H1: At least one of the line model coefficient is non zero

Test statistic F = 1028.131

p-value=0.000

p-value<0.05 Reject H0

The data provide evidence that at least one of the model coefficient is non zero. The overall model appears to be useful in predicting monthly labor hours.

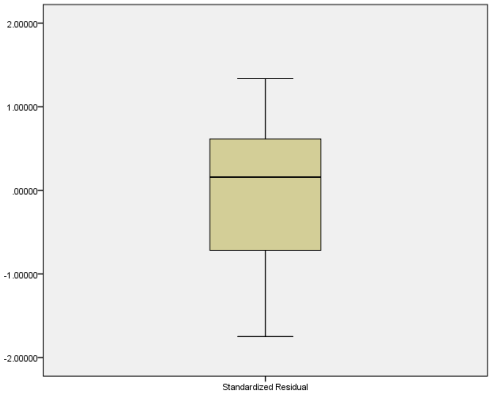
So final backward elimination model and final forward selection model are same.

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Standardized Residual	.132	16	.200*	.968	16	.806

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction



Residuals are normal and no outliers on the plot.

stepwise regression

Model Summary[c]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.986[a]	.972	.970	957.85555
2	.993[b]	.987	.985	685.16852

a. Predictors: (Constant), monthly occupied bed-days

b. Predictors: (Constant), monthly occupied bed-days, monthly X-ray exposures

c. Dependent Variable: monthly labor-hours

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	480950231.6	1	480950231.6	524.204	.000[b]
	Residual	13762308.86	15	917487.258		
	Total	494712540.5	16			
2	Regression	488140158.0	2	244070079.0	519.900	.000[c]
	Residual	6572382.538	14	469455.896		
	Total	494712540.5	16			

a. Dependent Variable: monthly labor-hours

b. Predictors: (Constant), monthly occupied bed-days

c. Predictors: (Constant), monthly occupied bed-days, monthly X-ray exposures

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B		Collinearity Statistics	
		B	Std. Error	Beta			Lower Bound	Upper Bound	Tolerance	VIF
1	(Constant)	-28.129	319.041		-.088	.931	-708.149	651.892		
	monthly occupied bed-days	1.117	.049	.986	22.895	.000	1.013	1.221	1.000	1.000
2	(Constant)	-68.314	228.446		-.299	.769	-558.282	421.654		
	monthly occupied bed-days	.823	.083	.726	9.919	.000	.645	1.001	.177	5.647
	monthly X-ray exposures	.075	.019	.286	3.913	.002	.034	.116	.177	5.647

a. Dependent Variable: monthly labor-hours

Multiple linear regression equation:
stepwise regression method

Variables are: monthly occupied bed-days and monthly X-ray exposures

y=ß0+ß1x1+ß2x2+Ɛ

monthly labor hours ^ (estimated)= b0 + b1x1 + b2x2

monthly labor hours ^ (estimated) = 0.823* monthly occupied bed-days + 0.075* monthly X-ray exposures -68.314

H0: ß1=ß2=0
H1: At least one of the line model coefficient is non zero

Test statistic F = 519.9

p-value=0.000

p-value<0.05 Reject H0

The data provide evidence that at least one of the model coefficient is non zero. The overall model appears to be useful in predicting monthly labor hours.

Forward selection regression model and stepwise regression model are exactly the same model.

So Normality check for residuals also same as forward selection method.

After removed the row 14

Model Summary[d]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.989[a]	.978	.976	856.70736
2	.997[b]	.993	.992	489.12639
3	.998[c]	.996	.995	387.15977

a. Predictors: (Constant), monthly occupied bed-days

b. Predictors: (Constant), monthly occupied bed-days, average length of patient's stay, in days

c. Predictors: (Constant), monthly occupied bed-days, average length of patient's stay, in days, monthly X-ray exposures

d. Dependent Variable: monthly labor-hours

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	453851336.7	1	453851336.7	618.370	.000[b]
	Residual	10275264.93	14	733947.495		
	Total	464126601.6	15			
2	Regression	461016421.4	2	230508210.7	963.483	.000[c]
	Residual	3110180.176	13	239244.629		
	Total	464126601.6	15			
3	Regression	462327889.4	3	154109296.5	1028.131	.000[d]
	Residual	1798712.218	12	149892.685		
	Total	464126601.6	15			

a. Dependent Variable: monthly labor-hours

b. Predictors: (Constant), monthly occupied bed-days

c. Predictors: (Constant), monthly occupied bed-days, average length of patient's stay, in days

d. Predictors: (Constant), monthly occupied bed-days, average length of patient's stay, in days, monthly X-ray exposures

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B		Collinearity Statistics	
		B	Std. Error	Beta			Lower Bound	Upper Bound	Tolerance	VIF
1	(Constant)	-70.230	286.004		-.246	.810	-683.648	543.187		
	monthly occupied bed-days	1.101	.044	.989	24.867	.000	1.006	1.196	1.000	1.000
2	(Constant)	2741.244	539.068		5.085	.000	1576.659	3905.829		
	monthly occupied bed-days	1.223	.034	1.098	36.304	.000	1.150	1.296	.563	1.775
	average length of patient's stay, in days	-572.249	104.567	-.166	-5.473	.000	-798.153	-346.345	.563	1.775
3	(Constant)	1946.802	504.182		3.861	.002	848.284	3045.320		
	monthly occupied bed-days	1.039	.068	.933	15.386	.000	.892	1.187	.088	11.396
	average length of patient's stay, in days	-413.758	98.598	-.120	-4.196	.001	-628.585	-198.931	.397	2.520
	monthly X-ray exposures	.039	.013	.149	2.958	.012	.010	.067	.128	7.828

a. Dependent Variable: monthly labor-hours

	After removed the row 14	**Before removed the row 14**
R^2	0.996	0.987
S^2	149892.685	469455.896
F/Sig	1028.131/0.000	519.9/0.000
Residual	1798712.218	6572382.538
t/Sig	15.386/0.00,-4.196/0.001,2.958/0.012	9.919/0.000,3.913/0.002

After data row 14 removed model has improved a lot. S^2 and Residual have lower value and R^2, F have higher values after removed row 14.

So final stepwise regression model:

monthly labor hours ^ (estimated) = 1.039* monthly occupied bed-days + 0.039* monthly X-ray exposures -413.758*average length of patient's stay in days+1946.802

H0: ß1=ß2= ß3=0
H1: At least one of the line model coefficient is non zero

Test statistic F = 1028.131

p-value=0.000

p-value<0.05 Reject H0

The data provide evidence that at least one of the model coefficient is non zero. The overall model appears to be useful in predicting monthly labor hours.

9. Compare the three variable screening methods. Which final model would you select? Explain.

	Forward, backward, stepwise (After removed the row 14)	**backward elimination (Before removed the row 14)**	**forward selection and stepwise (Before removed the row 14)**
variables	monthly occupied bed-days, monthly X-ray exposures, average length of patient's stay	monthly occupied bed-days, monthly X-ray exposures, average length of patient's stay	monthly occupied bed-days, monthly X-ray exposures
R^2	0.996	0.990	0.987
S^2	149892.685	377953.731	469455.896
F/Sig	1028.131/0.000	431.975/0.000	519.9/0.000
Residual	1798712.218	4913398.503	6572382.538
t/Sig	15.386/0.00,-4.196/0.001,2.958/0.012	2.637/0.21,9.305/0.000,-2.095/0.056	9.919/0.000,3.913/0.002
VIF	11.396, 2.520, 7.828	7.737, 11.269, 2.493	5.647, 5.647

Before remove the outlier forward and stepwise have same model. After data row 14 removed model has been improved a lot. S^2 and Residual have lower value and R^2, F have higher values after removed row 14. t values are significant.

Final model:

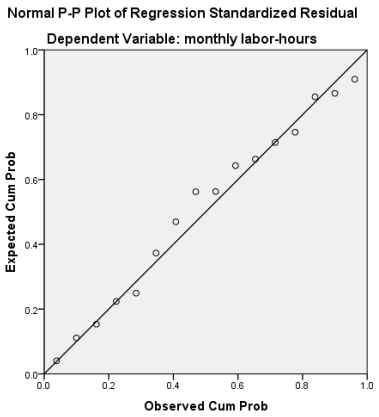
monthly labor hours ^ (estimated) = 1.039* monthly occupied bed-days + 0.039* monthly X-ray exposures -413.758*average length of patient's stay in days+1946.802

So finally we got the only one model from the all three method after removing the outlier. So our final model is:

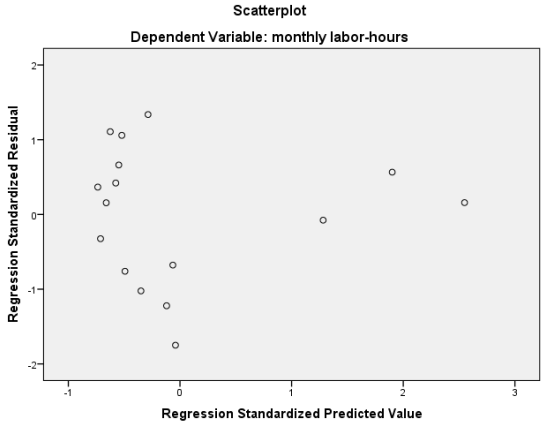
monthly labor hours ^ (estimated) = 1.039* monthly occupied bed-days + 0.039* monthly X-ray exposures -413.758*average length of patient's stay in days+1946.802

I did the comparison in above steps

10. Examine the residual plots of the final model.



Most of the data points are on or very closer to the line. We can assume residual are normal.



Residuals are not randomly distributed around the zero line. Therefore, variance of residuals is not homogeneous.

11)What improvements do you see in the final model? Make a table and compare with the original model.

	Final model	**Original model**
	monthly labor hours ^ (estimated) = 1.039* monthly occupied bed-days + 0.039* monthly X-ray exposures - 413.758*average length of patient's stay in days+1946.802	**monthly labor hours ^ (estimated) = -15.852* average daily patient load + 0.056* monthly X-ray exposures+1.590* monthly occupied bed-days -4.219* eligible population in the area/1000 - 394.314* average length of the patient's stay, in days +1962.948**
variables	monthly occupied bed-days monthly X-ray exposures average length of patient's stay in days	monthly occupied bed-days monthly X-ray exposures average length of patient's stay in days average daily patient load eligible population in the area/1000
R^2	0.996	0.991
S^2	149892.685	412277.488
F/Sig	1028.131/0.000	237.790/0.000
Residual	1798712.218	4535052.367
t/Sig	15.386/0.00,2.958/0.012,-4.196/0.001	0.514/0.617, 2.631/0.023, -1.881/0.087, -0.162/0.874, -0.588/0.569
VIF	11.396, 2.520, 7.828	8933.087, 7.941, 4.280, 9597.571, 23.294

Final model is better than the original model. Because final model has little higher R^2(0.996 verse 0.991), very higher F value (1028.131, 237.790), small mean square error compare with original model MSE and higher significant t values when comparing with original model t values. Final model has VIF less than 10 or very closer to 10 but original model has some VIF values very high.

Therefore, **monthly labor hours ^ (estimated) = 1.039* monthly occupied bed-days + 0.039* monthly X-ray exposures -413.758*average length of patient's stay in days+1946.802**

Model is better.