Biostat 200B HW6

1. (a) $\hat{\beta}_{(-i)} = (X_{(-i)}'X_{(-i)})^{-1}X_{(-i)}'Y_{(-i)}$

$$= \left[(X'X)^{-1} + \frac{(X'X)^{-1}x_ix_i'(X'X)^{-1}}{1-h_i}\right](X'Y - x_iY_i)$$

$$= (X'X)^{-1}X'Y - (X'X)^{-1}x_iY_i + \frac{(X'X)^{-1}x_ix_i'(X'X)^{-1}}{1-h_i}X'Y - \frac{(X'X)^{-1}x_ix_i'(X'X)^{-1}}{1-h_i}x_iY_i$$

$$= \hat{\beta} - \frac{(X'X)^{-1}x_i}{1-h_i}[Y_i(1-h_i) - x_i'\underbrace{(X'X)^{-1}X'Y}_{\hat{\beta}} + \underbrace{x_i'(X'X)^{-1}x_i}_{h_i}Y_i]$$

$$= \hat{\beta} - \frac{(X'X)^{-1}x_i}{1-h_i}[Y_i - \cancel{Y_ih_i} - x_i'\beta + \cancel{h_iY_i}]$$

$$= \hat{\beta} - \frac{(X'X)^{-1}x_ie_i}{1-h_i} \quad \# \qquad (e_i = Y_i - \hat{Y} = Y_i - x_i'\hat{\beta})$$

(b) $y_i - x_i'\hat{\beta}_{(-1)} = y_i - x_i'\left(\hat{\beta} - \frac{(X'X)^{-1}x_ie_i}{1-h_i}\right)$

$$= y_i - x_i'\beta + \frac{\boxed{x_i'(X'X)^{-1}x_i}e_i}{1-h_i}$$

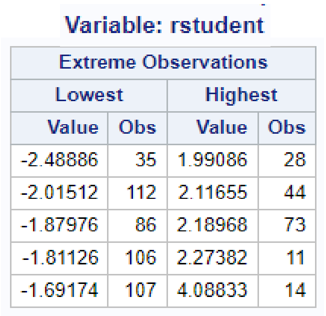
$$= e_i + \frac{h_ie_i}{1-h_i}$$

$$= \frac{e_i - e_ih_i + h_ie_i}{1-h_i} = \frac{e_i}{1-h_i} \quad \#$$

BIOSTAT 200B HW6

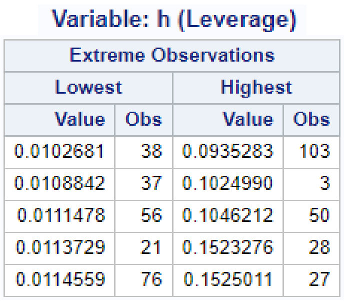
2.

```
proc reg data=senic;
        model loglength = xray census age    / influence r;
        output out=measures r=r   rstudent=rstudent h=h cookd=cookd ;
run; quit;
ods graphics on;
proc univariate data=measures plot;
    var rstudent h cookd ;
run;
ods graphics off;
```



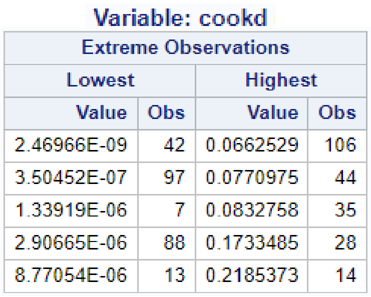
Variable: rstudent

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
-2.48886	35	1.99086	28
-2.01512	112	2.11655	44
-1.87976	86	2.18968	73
-1.81126	106	2.27382	11
-1.69174	107	4.08833	14



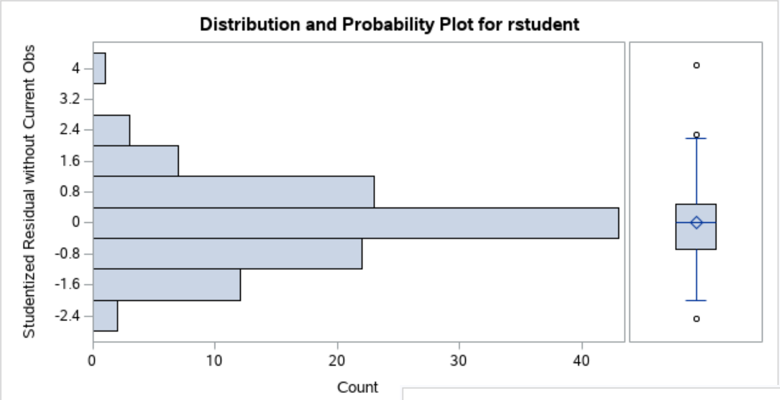
Variable: h (Leverage)

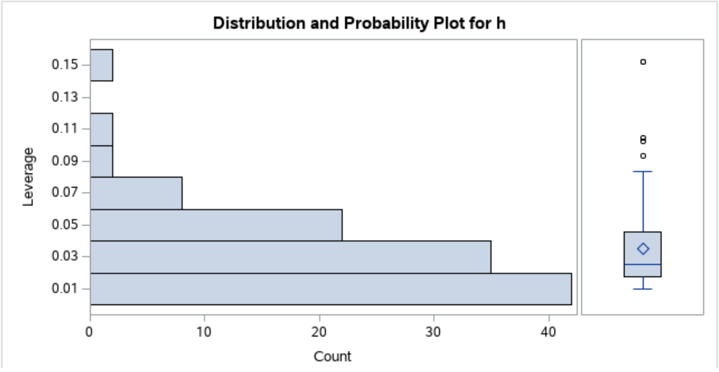
Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
0.0102681	38	0.0935283	103
0.0108842	37	0.1024990	3
0.0111478	56	0.1046212	50
0.0113729	21	0.1523276	28
0.0114559	76	0.1525011	27

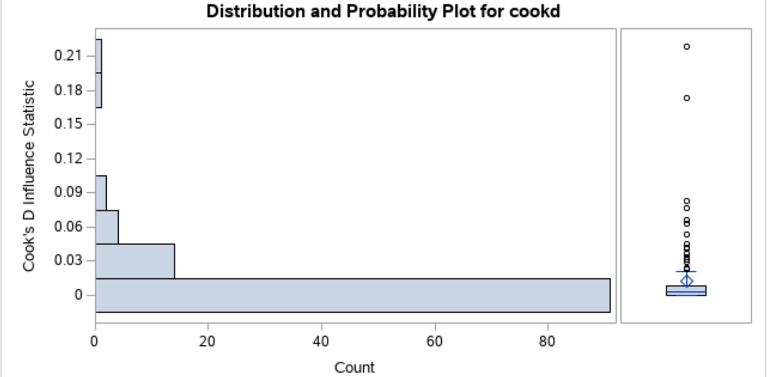


Variable: cookd

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
2.46966E-09	42	0.0662529	106
3.50452E-07	97	0.0770975	44
1.33919E-06	7	0.0832758	35
2.90665E-06	88	0.1733485	28
8.77054E-06	13	0.2185373	14







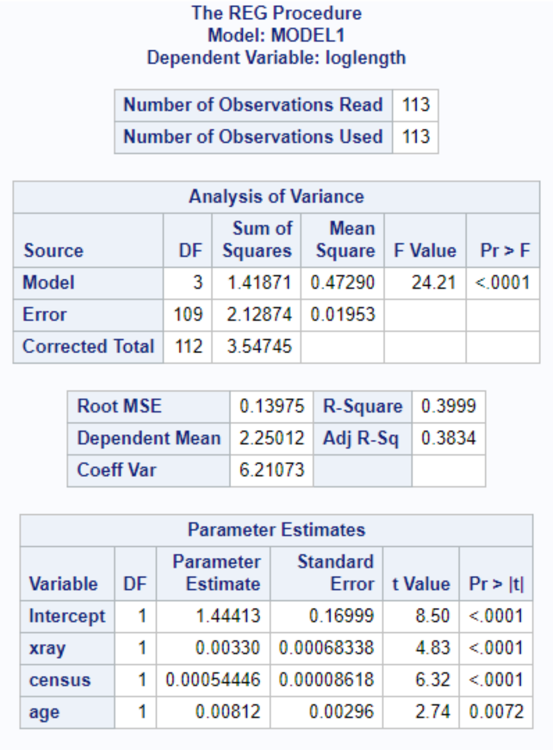
(a) From the table and the boxplots of studentized residual, leverage, and Cook's D, we can get the unusual observations for them. The unusual observations id number that have unusual studentized residual are 14, 11, 35. The unusual observations id number that have unusual leverage are 27, 28, 50, 3, 103. The unusual observations id number that have unusual Cooks' D are 14, 28, 35, 44, 106.

(b) The two hospitals with highest Cooks' D are 14 and 28. Hospital 14 has the highest studentized residual, and hospital 28 has a high leverage, which makes the, potentially influential.

(c)

```
proc reg data=senic;
		model loglength = xray census age;
run; quit;

data senic1;
set senic;
if id = 14 then delete; if id = 28 then delete;
run;
proc reg data=senic1;
		model loglength = xray census age;
run; quit;
```



The REG Procedure
Model: MODEL1
Dependent Variable: loglength

Number of Observations Read	113
Number of Observations Used	113

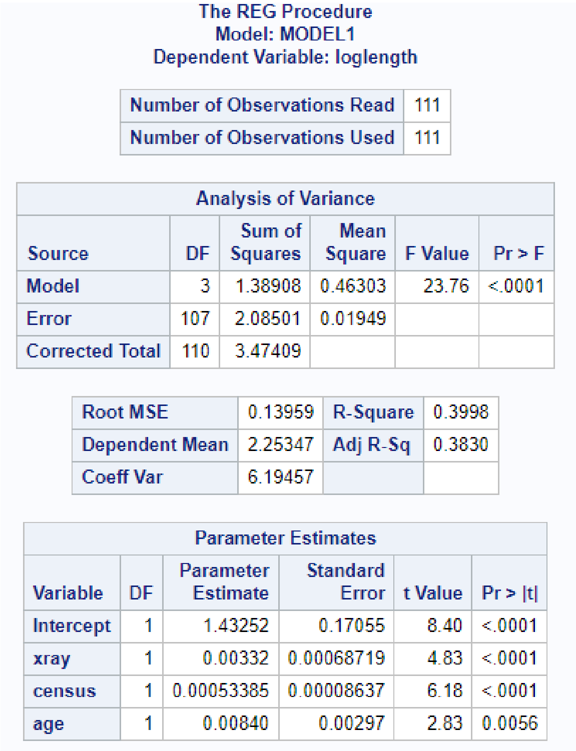
Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	1.41871	0.47290	24.21	<.0001
Error	109	2.12874	0.01953		
Corrected Total	112	3.54745			

Root MSE	0.13975	R-Square	0.3999
Dependent Mean	2.25012	Adj R-Sq	0.3834
Coeff Var	6.21073		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	1.44413	0.16999	8.50	<.0001
xray	1	0.00330	0.00068338	4.83	<.0001
census	1	0.00054446	0.00008618	6.32	<.0001
age	1	0.00812	0.00296	2.74	0.0072



The REG Procedure
Model: MODEL1
Dependent Variable: loglength

Number of Observations Read	111
Number of Observations Used	111

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	1.38908	0.46303	23.76	<.0001
Error	107	2.08501	0.01949		
Corrected Total	110	3.47409			

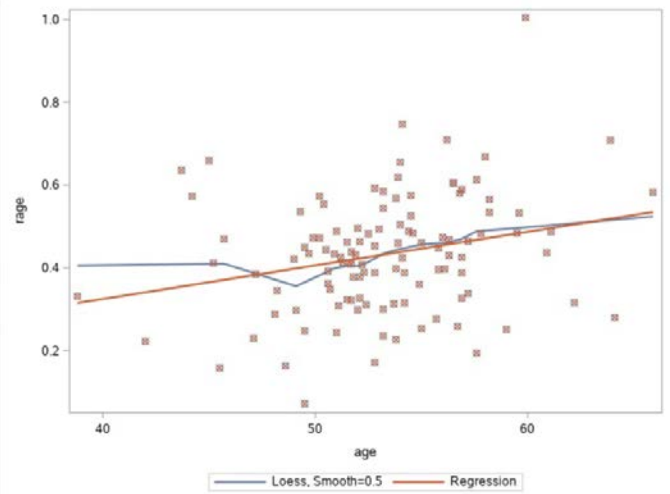
Root MSE	0.13959	R-Square	0.3998
Dependent Mean	2.25347	Adj R-Sq	0.3830
Coeff Var	6.19457		

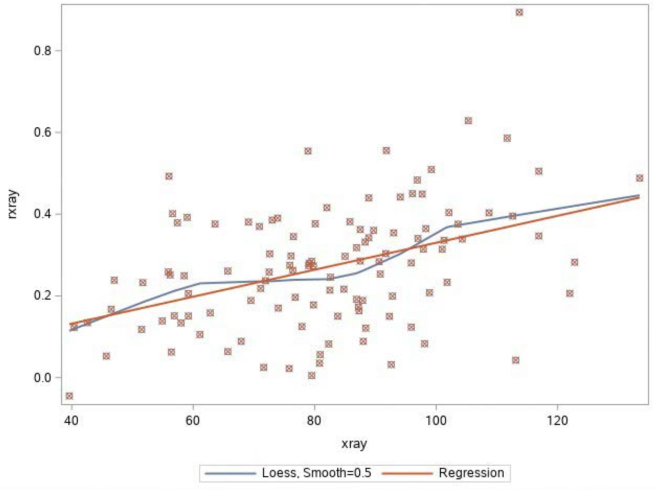
Parameter Estimates

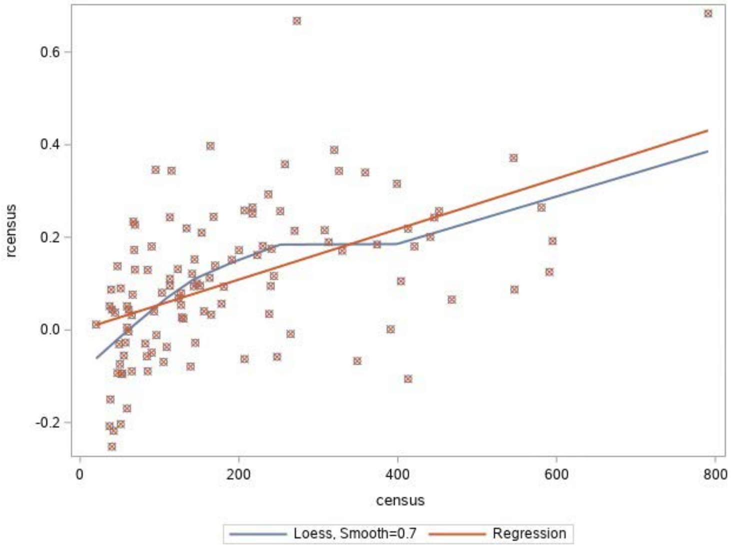
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	1.43252	0.17055	8.40	<.0001
xray	1	0.00332	0.00068719	4.83	<.0001
census	1	0.00053385	0.00008637	6.18	<.0001
age	1	0.00840	0.00297	2.83	0.0056

We conduct the sensitivity analyses, and the results are shown above. For two models, the regression coef estimates, p-values and root MSE are all pretty close.

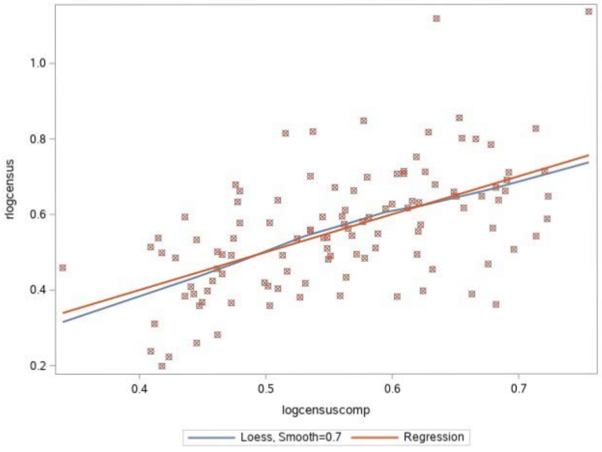
(d) The component-plus-residual plots are shown below.



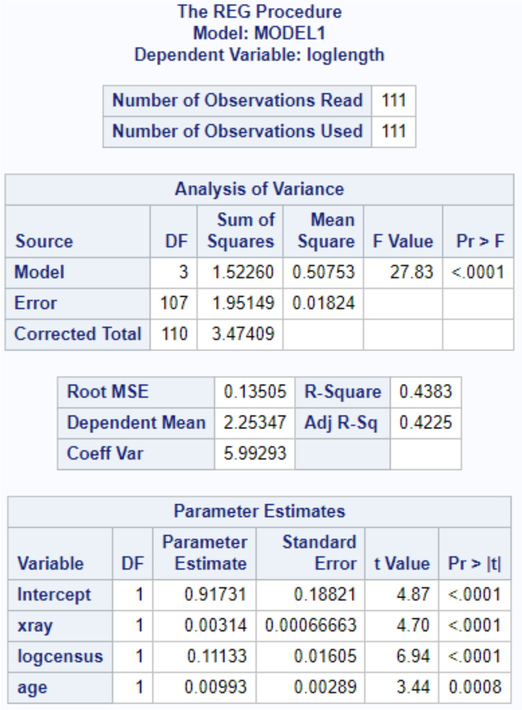




The predictor variable **census** needs to do log transformation to improve the model given the component-plus-residual plots shown above.



The final model will be ***loglength = xray logcensus age***. The ANOVA table and parameter estimates results are shown beside.



The REG Procedure
Model: MODEL1
Dependent Variable: loglength

Number of Observations Read	111
Number of Observations Used	111

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	1.52260	0.50753	27.83	<.0001
Error	107	1.95149	0.01824		
Corrected Total	110	3.47409			

Root MSE	0.13505	R-Square	0.4383
Dependent Mean	2.25347	Adj R-Sq	0.4225
Coeff Var	5.99293		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	0.91731	0.18821	4.87	<.0001
xray	1	0.00314	0.00066663	4.70	<.0001
logcensus	1	0.11133	0.01605	6.94	<.0001
age	1	0.00993	0.00289	3.44	0.0008

3.

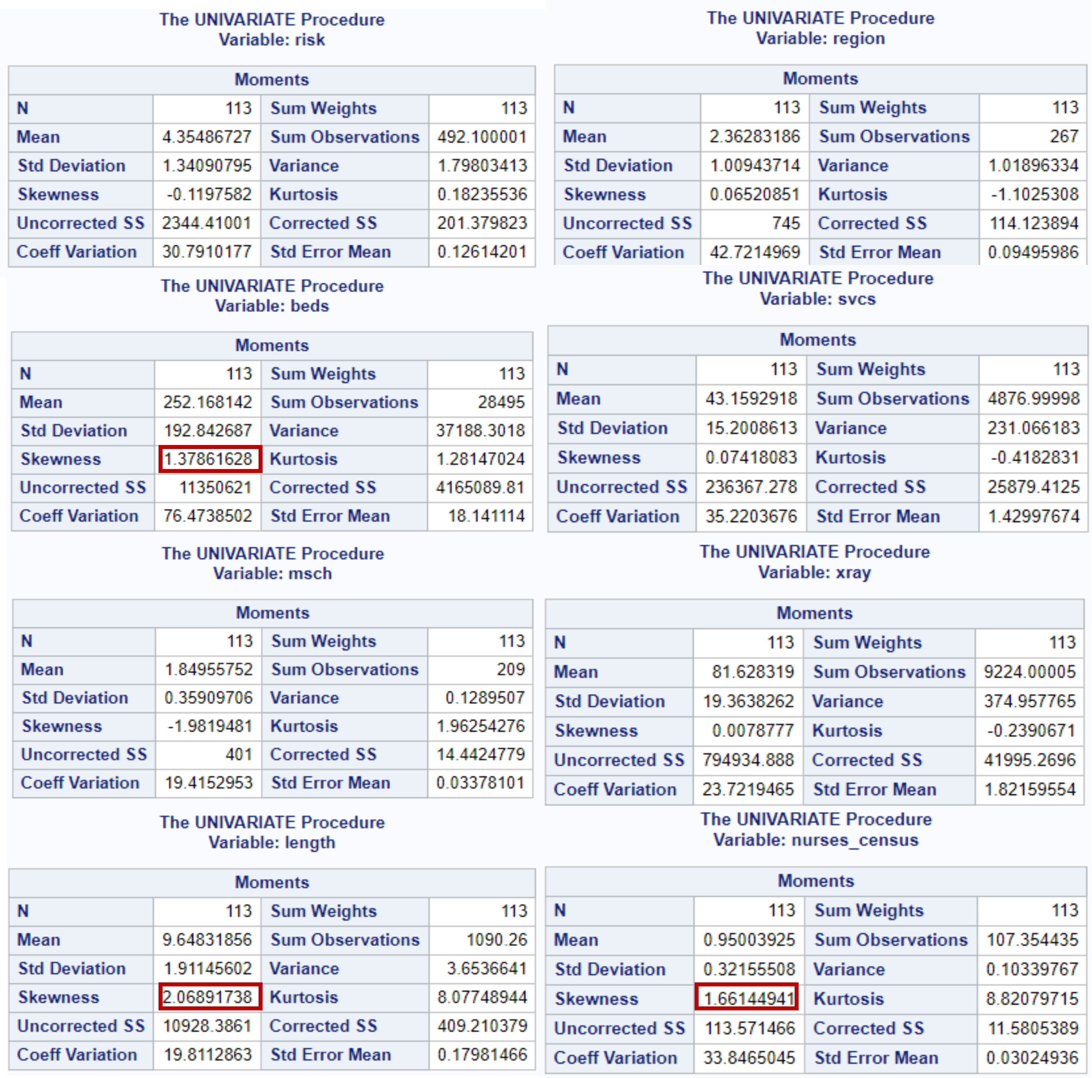
```
data d.senic;
 set d.senic;
 nurses_census = nurses/census;
run;

proc univariate data = d.senic plot;
 var risk region beds svcs msch xray length nurses_census;
run;

data d.senic;
 set d.senic;
 logbeds = log(beds);
 loglength = log(length);
 lognurses_census = log(nurses_census);
run;

proc reg data=d.senic;
 model risk = region logbeds svcs msch xray loglength lognurses_census/ selection=cp
aic bic;
run; quit;
```

We would like to log transform beds, length, and nurse/patient ratio, since the skewness of those variables are greater than 1, which can be considered as highly positive skewness.



The UNIVARIATE Procedure
Variable: risk

Moments			
N	113	Sum Weights	113
Mean	4.35486727	Sum Observations	492.100001
Std Deviation	1.34090795	Variance	1.79803413
Skewness	-0.1197582	Kurtosis	0.18235536
Uncorrected SS	2344.41001	Corrected SS	201.379823
Coeff Variation	30.7910177	Std Error Mean	0.12614201

The UNIVARIATE Procedure
Variable: region

Moments			
N	113	Sum Weights	113
Mean	2.36283186	Sum Observations	267
Std Deviation	1.00943714	Variance	1.01896334
Skewness	0.06520851	Kurtosis	-1.1025308
Uncorrected SS	745	Corrected SS	114.123894
Coeff Variation	42.7214969	Std Error Mean	0.09495986

The UNIVARIATE Procedure
Variable: beds

Moments			
N	113	Sum Weights	113
Mean	252.168142	Sum Observations	28495
Std Deviation	192.842687	Variance	37188.3018
Skewness	1.37861628	Kurtosis	1.28147024
Uncorrected SS	11350621	Corrected SS	4165089.81
Coeff Variation	76.4738502	Std Error Mean	18.141114

The UNIVARIATE Procedure
Variable: svcs

Moments			
N	113	Sum Weights	113
Mean	43.1592918	Sum Observations	4876.99998
Std Deviation	15.2008613	Variance	231.066183
Skewness	0.07418083	Kurtosis	-0.4182831
Uncorrected SS	236367.278	Corrected SS	25879.4125
Coeff Variation	35.2203676	Std Error Mean	1.42997674

The UNIVARIATE Procedure
Variable: msch

Moments			
N	113	Sum Weights	113
Mean	1.84955752	Sum Observations	209
Std Deviation	0.35909706	Variance	0.1289507
Skewness	-1.9819481	Kurtosis	1.96254276
Uncorrected SS	401	Corrected SS	14.4424779
Coeff Variation	19.4152953	Std Error Mean	0.03378101

The UNIVARIATE Procedure
Variable: xray

Moments			
N	113	Sum Weights	113
Mean	81.628319	Sum Observations	9224.00005
Std Deviation	19.3638262	Variance	374.957765
Skewness	0.0078777	Kurtosis	-0.2390671
Uncorrected SS	794934.888	Corrected SS	41995.2696
Coeff Variation	23.7219465	Std Error Mean	1.82159554

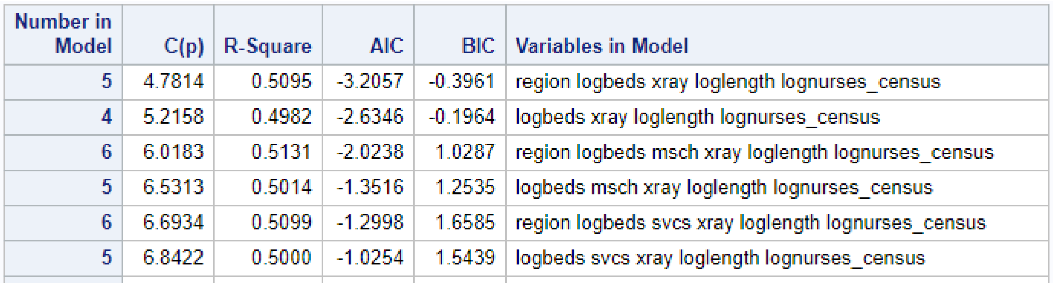
The UNIVARIATE Procedure
Variable: length

Moments			
N	113	Sum Weights	113
Mean	9.64831856	Sum Observations	1090.26
Std Deviation	1.91145602	Variance	3.6536641
Skewness	2.06891738	Kurtosis	8.07748944
Uncorrected SS	10928.3861	Corrected SS	409.210379
Coeff Variation	19.8112863	Std Error Mean	0.17981466

The UNIVARIATE Procedure
Variable: nurses_census

Moments			
N	113	Sum Weights	113
Mean	0.95003925	Sum Observations	107.354435
Std Deviation	0.32155508	Variance	0.10339767
Skewness	1.66144941	Kurtosis	8.82079715
Uncorrected SS	113.571466	Corrected SS	11.5805389
Coeff Variation	33.8465045	Std Error Mean	0.03024936

The model selection procedure suggests us to use the first model to predict **risk**, because it has the lowest Cp, AIC and BIC. We cannot really decide the best model by using R-square since it will only increase when the predictor variables add in, while we want the model as simple as possible (it is better to have fewer predictor variables).



Number in Model	C(p)	R-Square	AIC	BIC	Variables in Model
5	4.7814	0.5095	-3.2057	-0.3961	region logbeds xray loglength lognurses_census
4	5.2158	0.4982	-2.6346	-0.1964	logbeds xray loglength lognurses_census
6	6.0183	0.5131	-2.0238	1.0287	region logbeds msch xray loglength lognurses_census
5	6.5313	0.5014	-1.3516	1.2535	logbeds msch xray loglength lognurses_census
6	6.6934	0.5099	-1.2998	1.6585	region logbeds svcs xray loglength lognurses_census
5	6.8422	0.5000	-1.0254	1.5439	logbeds svcs xray loglength lognurses_census