hw6

AUTHOR
Hanbei Xiong

Question 1

Let $\boldsymbol{x}_i$ be the $p \times 1$ covariate vector for observation i (vector whose elements are the values of $p-1$ covariates for observation i and a leading 1). Let $\boldsymbol{X}_{(-i)}$ denote the design matrix $\boldsymbol{X}$ with the row corresponding to observation i removed, and $\boldsymbol{Y}_{(-i)}$ denote the vector of responses $\boldsymbol{Y}$ with y_i removed. We have the following identities:

$$(\boldsymbol{X}_{(-i)}{}'\boldsymbol{X}_{(-i)})^{-1} = (\boldsymbol{X}'\boldsymbol{X})^{-1} + \frac{(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_i\boldsymbol{x}_i'(\boldsymbol{X}'\boldsymbol{X})^{-1}}{1-h_i}$$
$$\boldsymbol{X}'_{(-i)}\boldsymbol{Y}_{(-i)} = \boldsymbol{X}'\boldsymbol{Y} - \boldsymbol{x}_i Y_i$$

a. Let $\widehat{\boldsymbol{\beta}}_{(-i)}$ denote the $p \times 1$ vector of least squares estimates of the regression coefficients obtained when the ith observation is omitted. Show that $\widehat{\boldsymbol{\beta}}_{(-i)} = \widehat{\boldsymbol{\beta}} - \frac{(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_i e_i}{1-h_i}$.

b. The quantity $Y_i - \boldsymbol{x}_i{}'\widehat{\boldsymbol{\beta}}_{(-i)}$ is the residual for the ith observation when $\boldsymbol{\beta}$ is estimated without the ith observation. This quantity is called the *predicted residual*, or the *PRESS* residual. Show that $y_i - \boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}_{(-i)} = \frac{e_i}{1-h_i}$.

(a) Answer:

$$
\begin{aligned}
\hat{\beta}_{(-i)} &= (X'_{(-i)}X_{(-i)})^{-1}X'_{(-i)}Y_{(-i)} \\
&= [(X'X)^{-1} + \frac{(X'X)^{-1}x_i x_i'(X'X)^{-1}}{1-h_i}](X'Y - x_iY_i) \\
&= (X'X)^{-1}X'Y - (X'X)^{-1}x_iY_i + (\frac{(X'X)^{-1}x_i x_i'(X'X)^{-1}}{1-h_i})X'Y - (\frac{(X'X)^{-1}x_i x_i'(X'X)^{-1}}{1-h_i})x_iY_i \\
&= \hat{\beta} - \frac{(X'X)^{-1}x_iY_i}{1-h_i} + (\frac{(X'X)^{-1}x_i x_i'(X'X)^{-1}}{1-h_i})x_iY_i \\
&+ (\frac{(X'X)^{-1}x_i x_i'(X'X)^{-1}}{1-h_i})X'Y - (\frac{(X'X)^{-1}x_i x_i'(X'X)^{-1}}{1-h_i})x_iY_i \\
&= \hat{\beta} - \frac{(X'X)^{-1}x_iY_i}{1-h_i} + (\frac{(X'X)^{-1}x_i x_i'(X'X)^{-1}}{1-h_i})X'Y \\
&= \hat{\beta} - \frac{(X'X)^{-1}x_iY_i}{1-h_i} + (\frac{(X'X)^{-1}x_i x_i'(X'X)^{-1}}{1-h_i})X'X\beta \\
&= \hat{\beta} - \frac{(X'X)^{-1}x_iY_i}{1-h_i} + \frac{(X'X)^{-1}x_i x_i'\beta}{1-h_i} \\
&= \hat{\beta} - \frac{(X'X)^{-1}x_i(Y_i - x_i'\beta)}{1-h_i} \\
&= \hat{\beta} - \frac{(X'X)^{-1}x_ie_i}{1-h_i}
\end{aligned}
$$

(b) Answer:

$$
\begin{aligned}
y_i - x_i'\hat{\beta}_{(-i)} &= y_i - x_i'(\hat{\beta} - \frac{(X'X)^{-1}x_ie_i}{1-h_i}) \\
&= y_i - x_i'\hat{\beta} + \frac{x_i'(X'X)^{-1}x_ie_i}{1-h_i} \\
&= \frac{(x_i'\hat{\beta} + e_i)(1-h_i)}{1-h_i} - \frac{x_i'\hat{\beta}(1-h_i)}{1-h_i} + \frac{h_ie_i}{1-h_i} \\
&= \frac{e_i(1-h_i)}{1-h_i} + \frac{h_ie_i}{1-h_i} \\
&= \frac{e_i(1-h_i+h_i)}{1-h_i} \\
&= \frac{e_i}{1-h_i}
\end{aligned}
$$

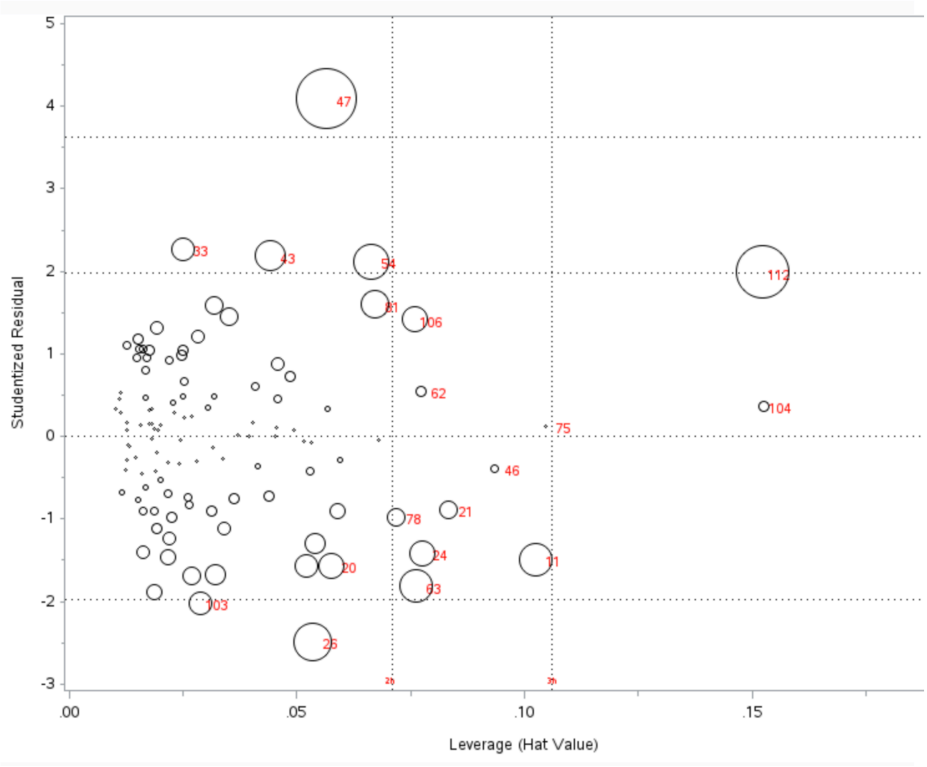
Question 2

For the SENIC data set, consider the model regressing loglength on xray, census and age.

a. Plot studentized residual, leverage and Cooks D by id and identify observations that have unusual leverage, residual or Cook's D. You can make three separate plots or produce a "bubble" plot.
b. Identify the two hospitals with the highest Cook's D. By examining influence statistics such as leverage and residual, and log(length) and predictor values, can you discover what are the characteristics of these hospitals that make them potentially influential (that is, give them a high Cook's D)?
c. Rather than rely completely on influence measures, it is better to **conduct sensitivity analyses** to see whether our inferences really do depend on one or a few unusual observations. Run the model after dropping these observations. (We are doing this to see what happens. I do not recommend that you automatically drop influential observations! Rather, you need to do an investigation.) How do the results compare to the results when fitting the model to the full data set? Compare regression coef estimates, p-values and root MSE.
d. Conduct model diagnostics for the partial relationships in the model using component-plus-residual plots. Decide whether any of the predictors should be transformed to improve the model. Provide the output for your final model.

(a) Answer:

Figure 1:



By observing Figure 1, we detect ids in red have high influence to the model. The size of the circle corresponds to the quantity of cook's D. The x axis is the leverage and y axis is the studentized residuals.

In general, here are the ids with unusual leverage, residual, cook's D:

leverage: 112, 104

residual: 47, 33, 43, 54

cook's D: 47, 112

(b) Answer:

Figure 2: id vs Cook's D

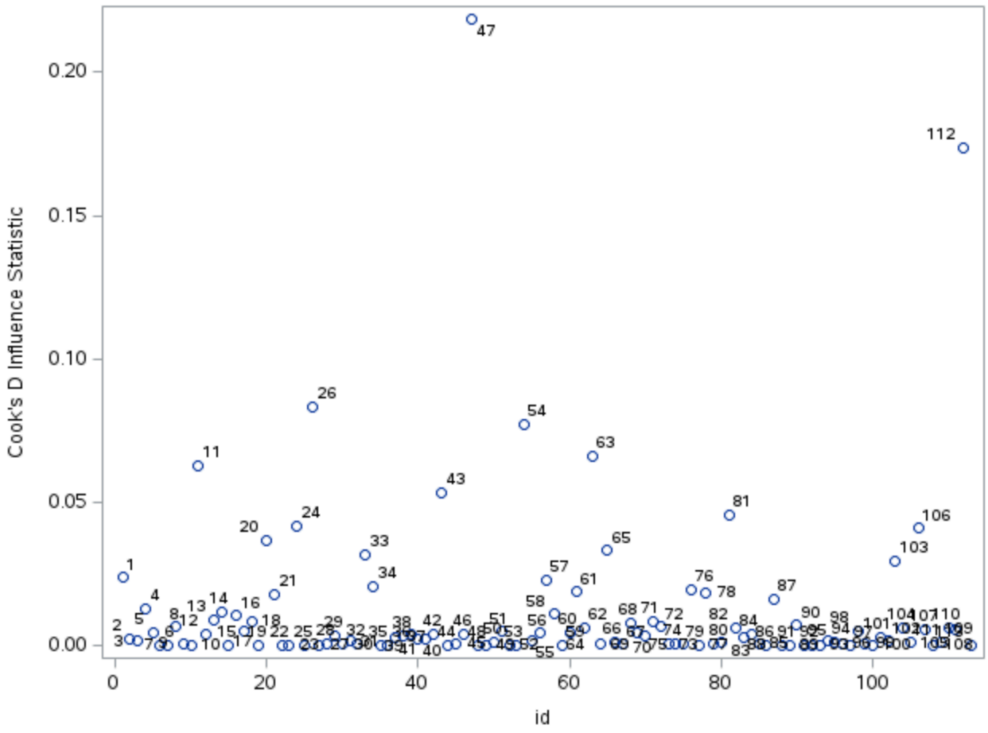


Figure 3: Studentized Residual without current observation vs Leverage

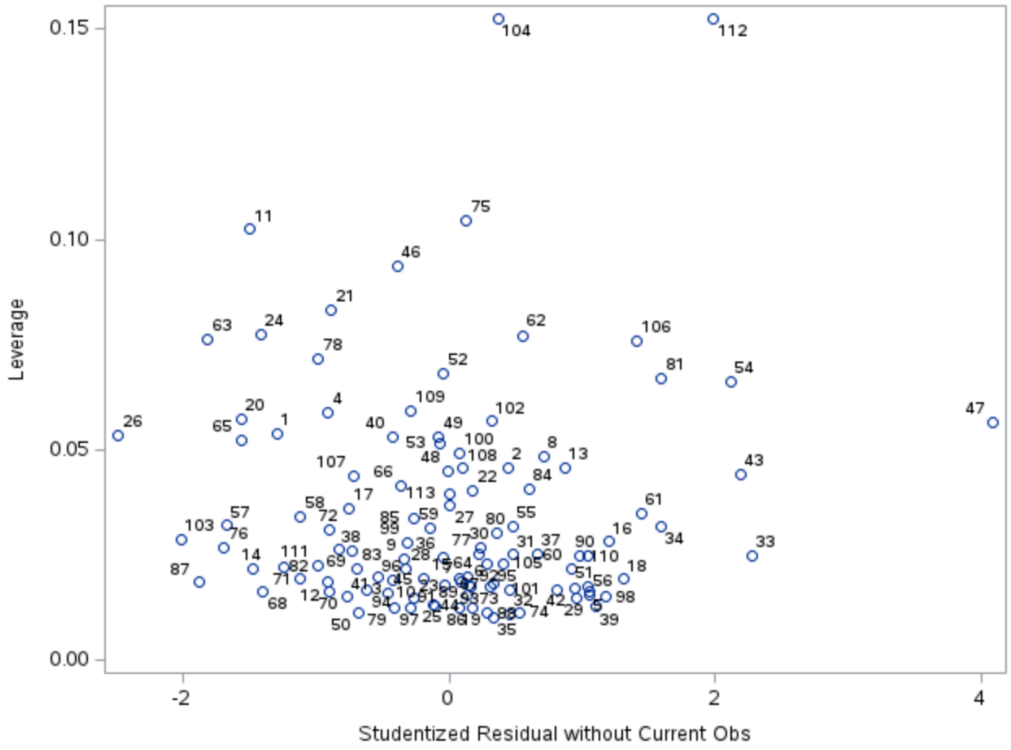


Figure 4: xray vs loglength

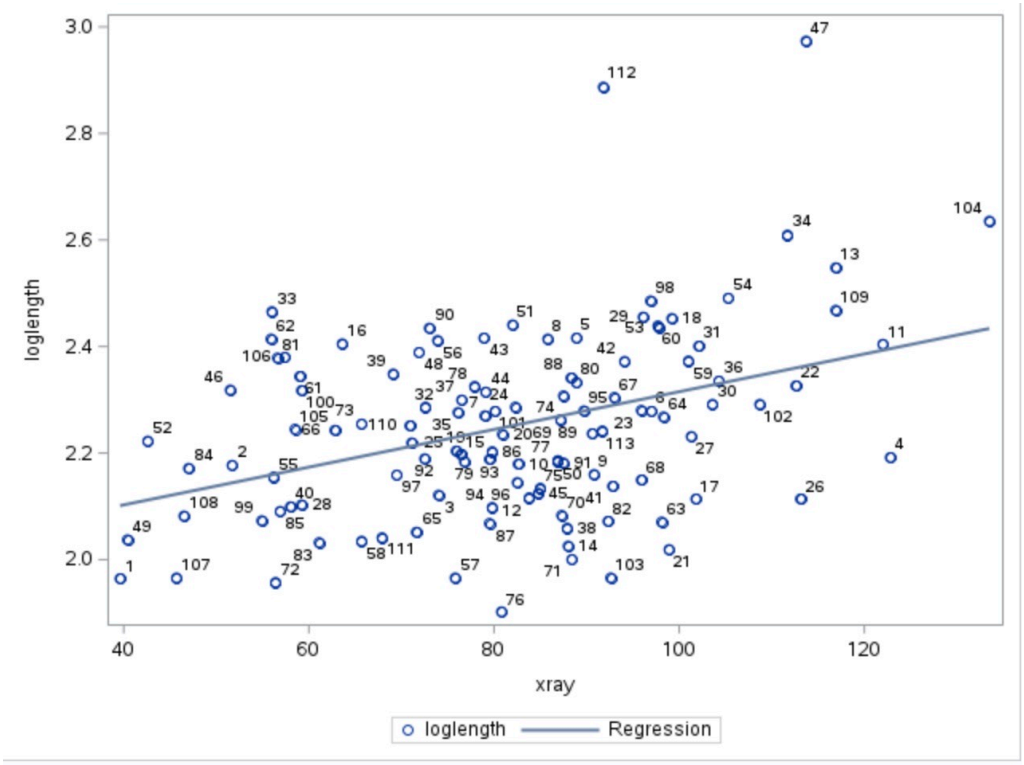


Figure 5: census vs loglength

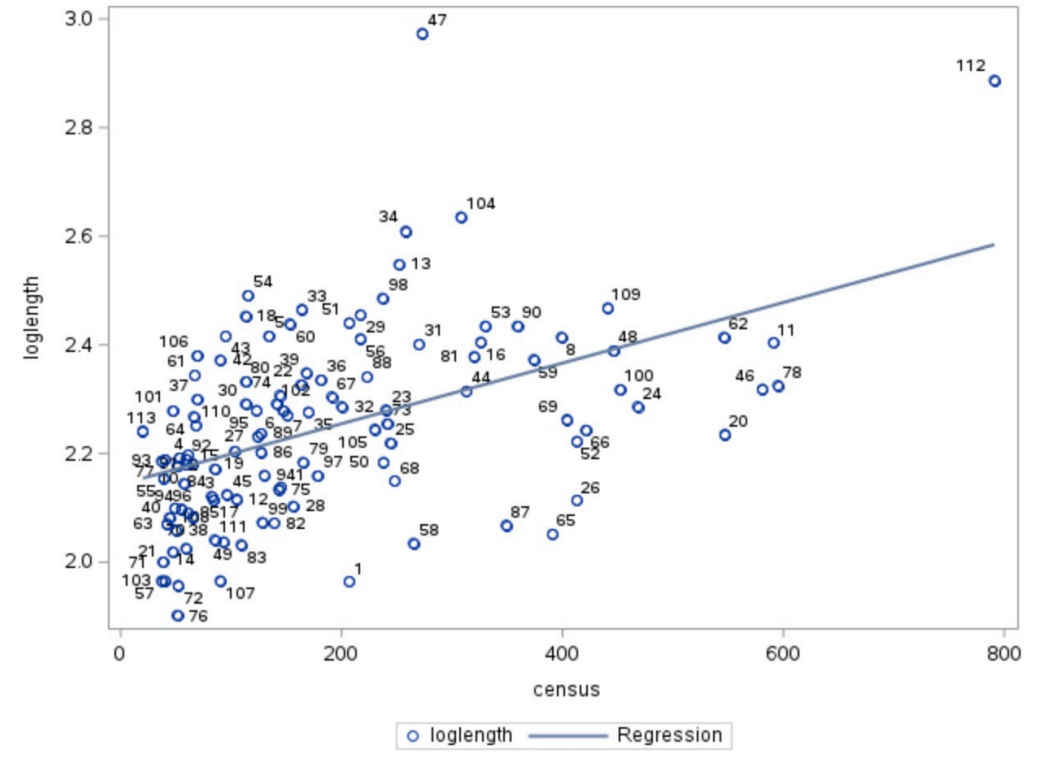
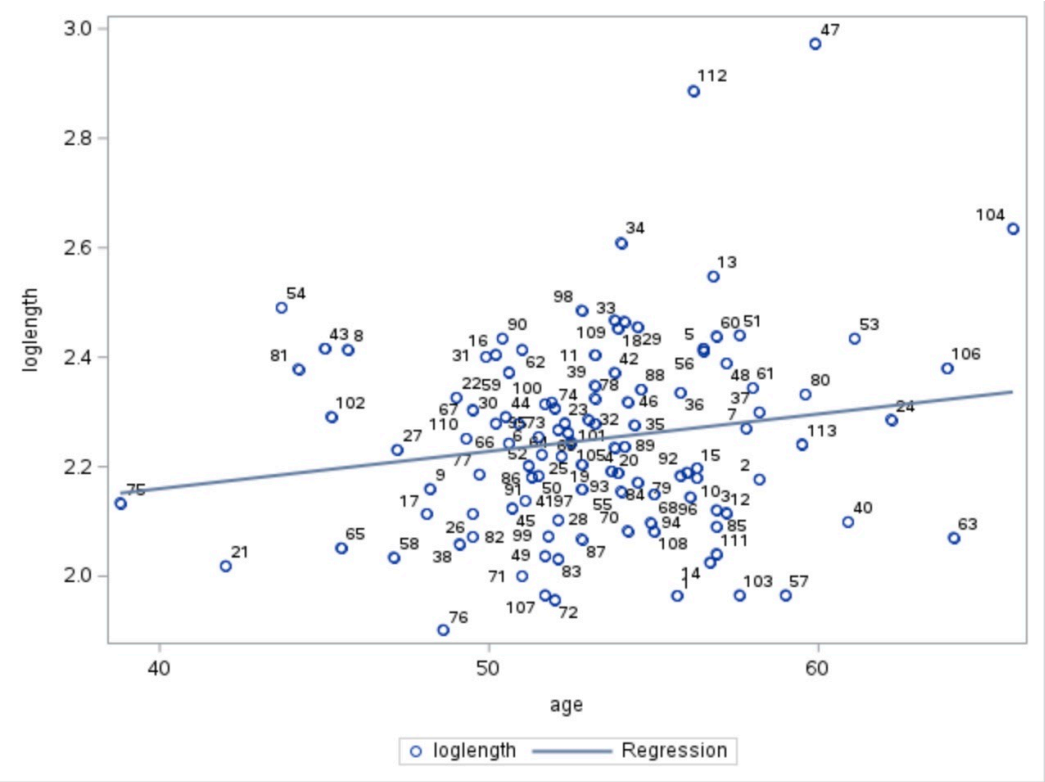


Figure 6: age vs loglength



By observing Figure 2, we can see hospital with id 47 and id 112 have the highest cook's id. By observing Figure 3, we can see hospital with id 47 has the highest studentized residual and id 112 have the highest leverage. By observing figure 4, 5 and 6, we can see hospital with id 47 and id 112 seem to be outliers. They have very high loglength compared to observation with same level of independent variables. The usual loglength of these two hospitals make them potentially influential.

(c) Answer:

Figure 7: Reduced Observation Model measures

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	0.84899	0.28300	17.27	<.0001
Error	107	1.75288	0.01638		
Corrected Total	110	2.60187			

Root MSE	0.12799	R-Square	0.3263
Dependent Mean	2.23787	Adj R-Sq	0.3074
Coeff Var	5.71939		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	1.61695	0.15986	10.11	<.0001
xray	1	0.00283	0.00063435	4.46	<.0001
census	1	0.00045298	0.00008524	5.31	<.0001
age	1	0.00578	0.00276	2.10	0.0385

Figure 8: Full Observation Model measures

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

Comparing Figure 7 and Figure 8, the regression parameter estimates in the reduced observation model are smaller than the regression parameter estimates in the full observation model. The p value for predict age has a big shift from 0.0072 (Figure 8) to 0.0385 (Figure 7). The statistical significance of predictor age in the model is waned after reducing the observations. The p values for overall F tests in both models remain the same. The root MSE in reduced observation model is smaller than that in full observation model which indicate the performance of the model improved after removing the outliers.

(d) Answer:

Figure 9: Final model Result

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	1.55646	0.51882	28.40	<.0001
Error	109	1.99100	0.01827		
Corrected Total	112	3.54745			

Root MSE	0.13515	R-Square	0.4388
Dependent Mean	2.25012	Adj R-Sq	0.4233
Coeff Var	6.00643		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	0.91575	0.18810	4.87	<.0001
xray	1	0.00314	0.00066240	4.74	<.0001
logcensus	1	0.11325	0.01598	7.09	<.0001
age	1	0.00973	0.00289	3.37	0.0010

Figure 9 is the result of the final model. I applied a log transformation on census after observing the component-plus-residual plots for each of the predictor. The numeric and graphic outputs generated in the procedure are displayed in Appendix.

Question 3

For the SENIC data set, use best subsets selection to select a best model. The outcome variable is risk.

a. The pool of candidate predictors is region, beds, services, medsch, xray, length and a new variable created as nurses/census (nurse/patient ratio).
b. Before model selection, log-transform positively skewed predictors.
c. Consider Cp, AIC and SBC/BIC to select among models.
d. To present your results, make a table listing the top models (about 5-8 models, using your discretion) and the values of their model selection criteria. Briefly describe the results and which model appears to be the “best”.

Answer:

Figure 10: Table listing top 5 models and selection criteria

Number in Model	C(p)	R-Square	AIC	BIC	Variables in Model
5	3.3371	0.5181	-5.2053	-2.2228	regw logbeds xray loglength lograti onp
6	4.2378	0.5232	-4.4022	-1.0963	regw logbeds msch xray loglength lograti onp
6	5.0489	0.5195	-3.5178	-0.3283	regw logbeds svcs xray loglength lograti onp
6	5.2821	0.5184	-3.2649	-0.1085	regs regw logbeds xray loglength lograti onp
6	5.3294	0.5182	-3.2136	-0.0639	regnc regw logbeds xray loglength lograti onp

I break the variable region to dummy variable (regw, regnc, regs). I applied log transformation on length, beds and ratio/census after observing positive right skewness in their distributions (Note that I name this variable as rationp to avoid containing dash in my variable name). The original distribution of predictors are displayed in Appendix.

From the result displayed in Figure 10, we can see the 5 predictors model containing regw, logbeds, xray, loglength, lograti onp are ranked as the top 1 among many subsets. These five predictors appear to be in each of the subset among the top 5 subsets. The top 1 model has the lowest AIC, BIC, and C(p) measures but R^2 in the second model outruns the top 1 model by 0.51.

Appendix

Appendix Table 1: Parameters used to build partial residuals in question 2d

The REG Procedure
Model: MODEL1
Dependent Variable: loglength

Number of Observations Read	113
Number of Observations Used	113

Analysis of Variance

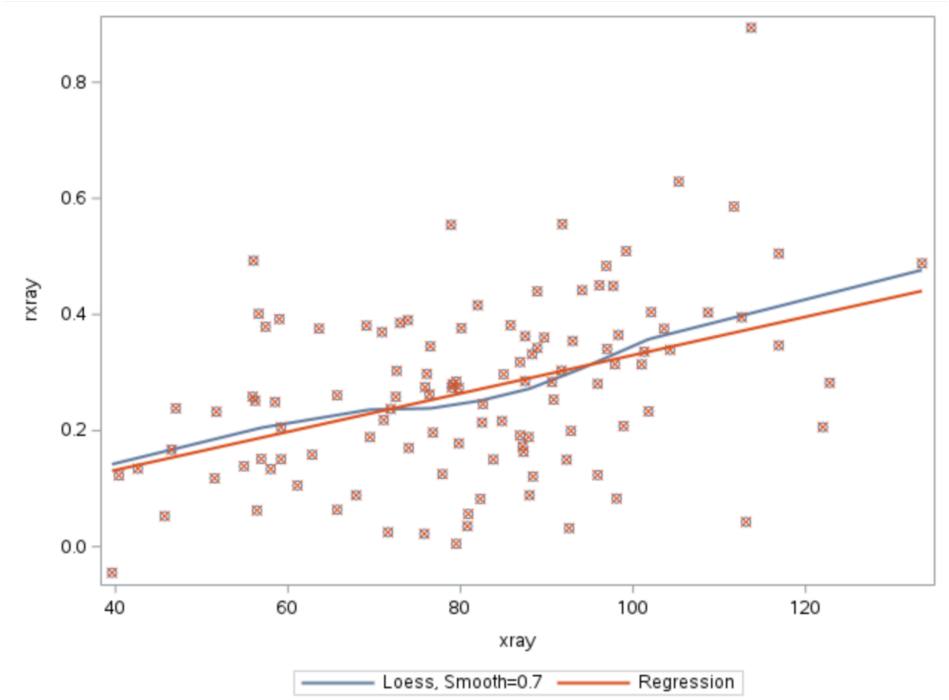
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	1.41871	0.47290	24.21	<.0001
Error	109	2.12874	0.01953		
Corrected Total	112	3.54745			

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

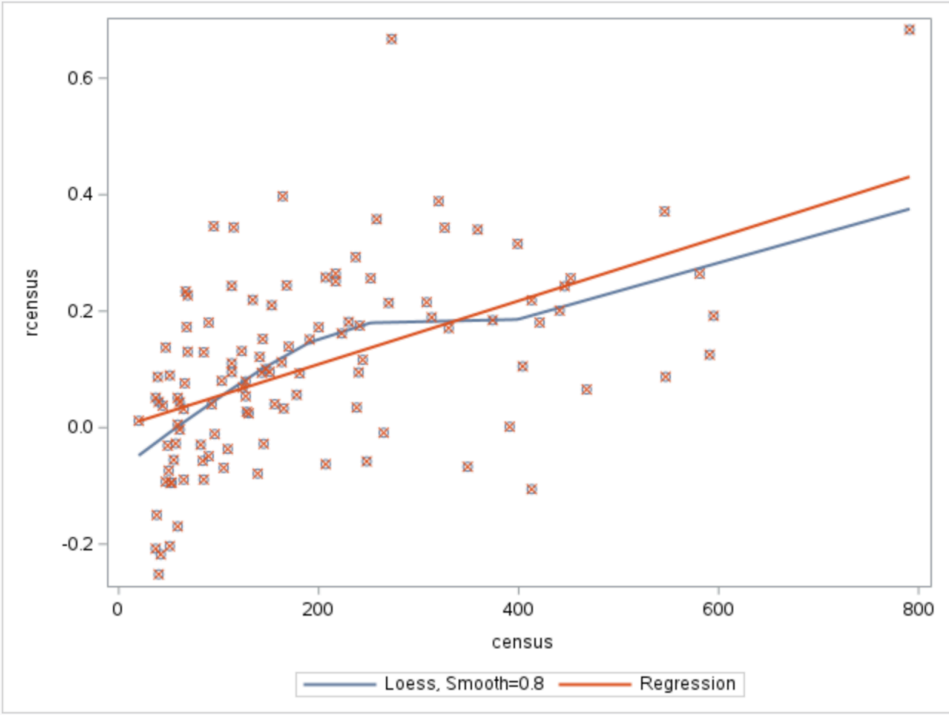
Parameter Estimates

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

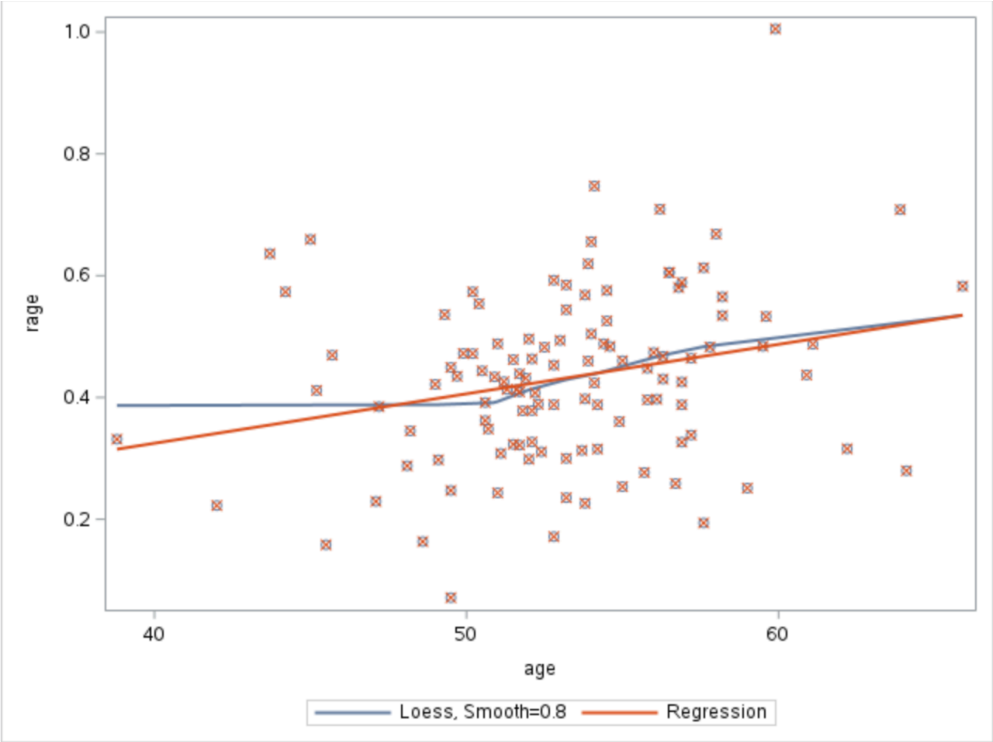
Appendix Graph 1: Component-plus-residual plot for xray in question 2d



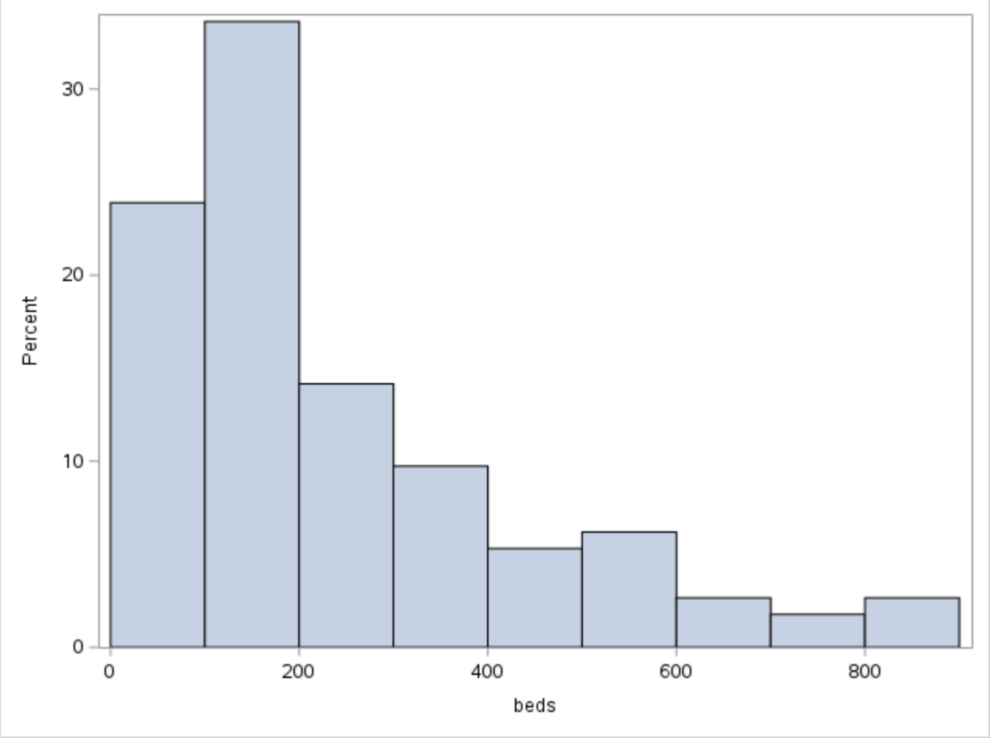
Appendix Graph 2: Component-plus-residual plot for census in question 2d



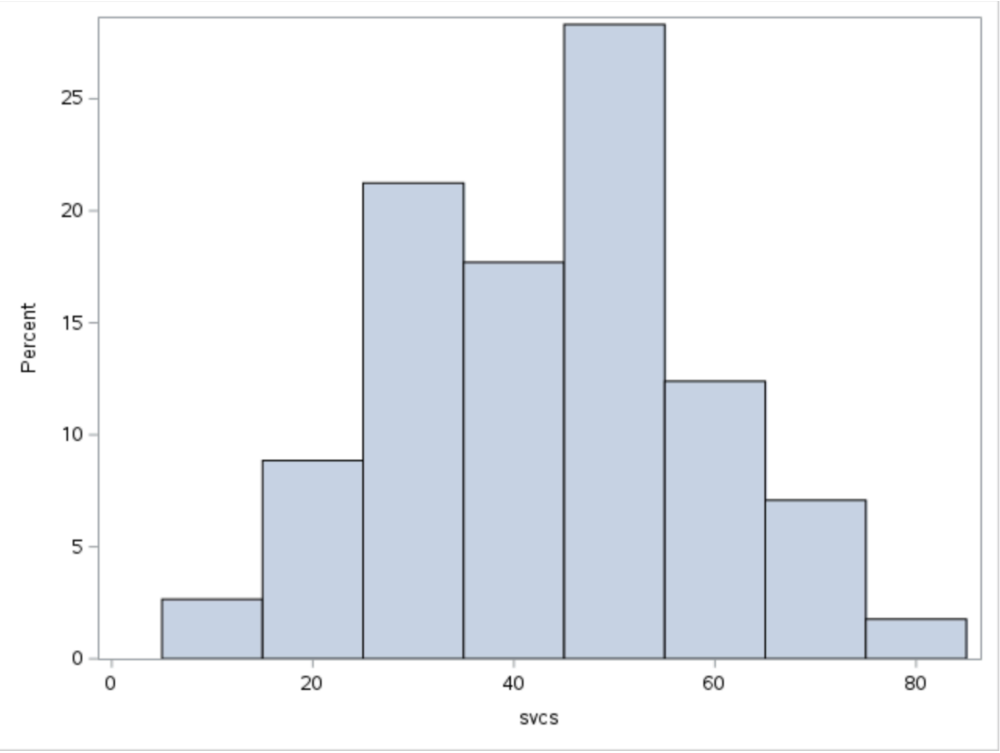
Appendix Graph 3: Component-plus-residual plot for age in question 2d



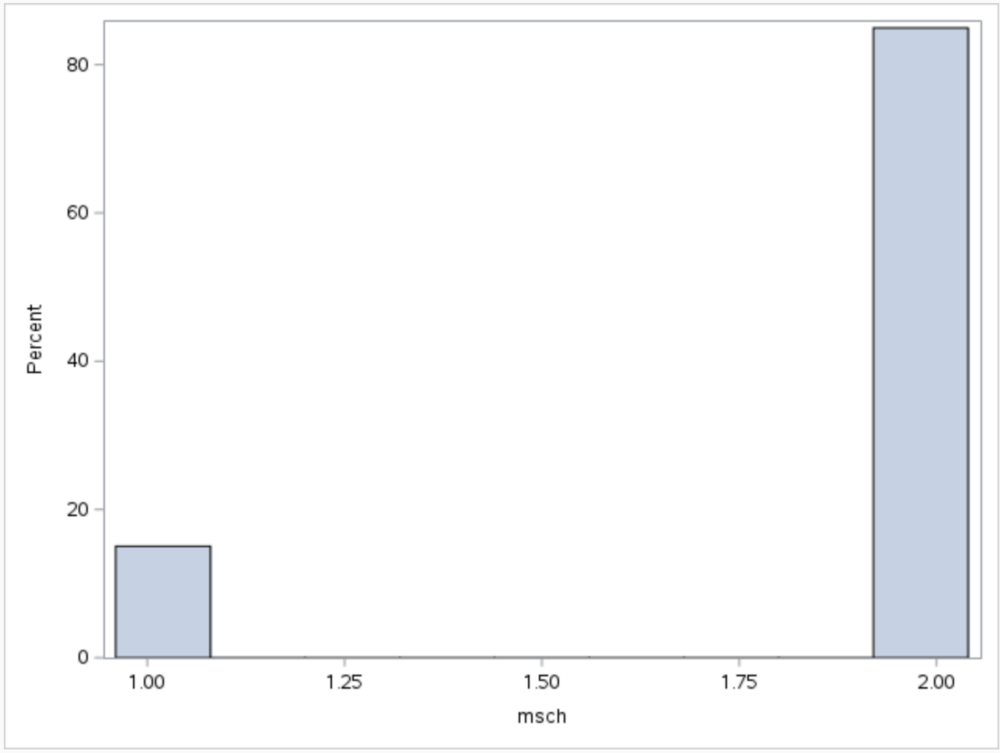
Appendix Graph 4: Histogram of beds in question 3b



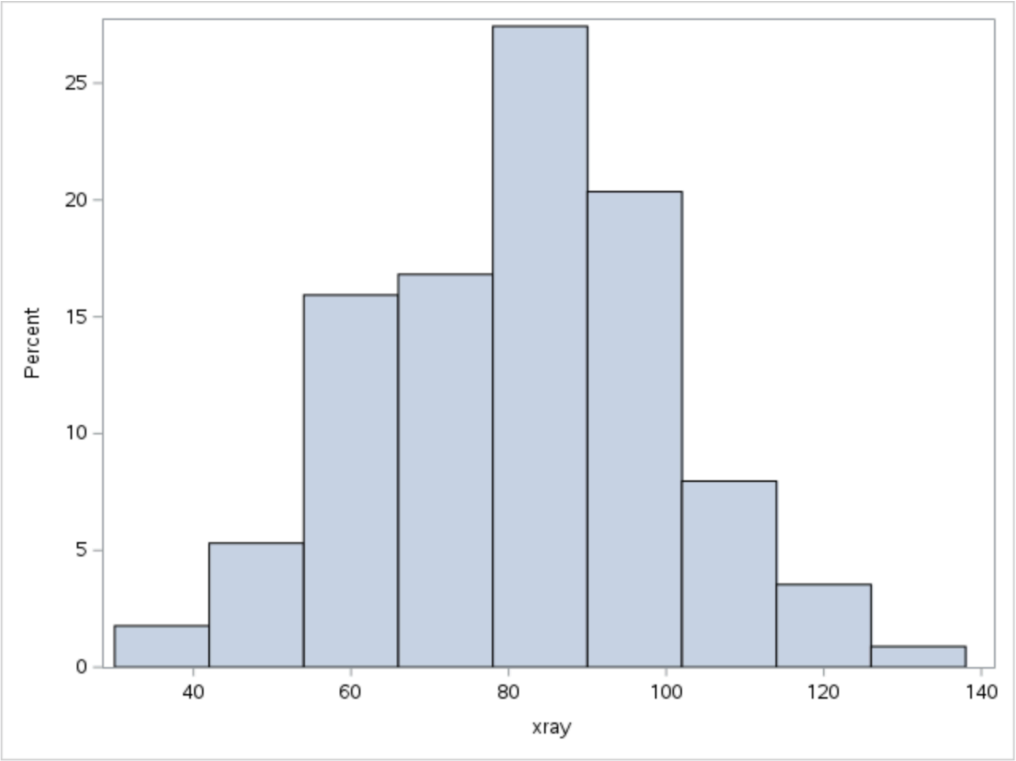
Appendix Graph 5: Histogram of svcs in question 3b



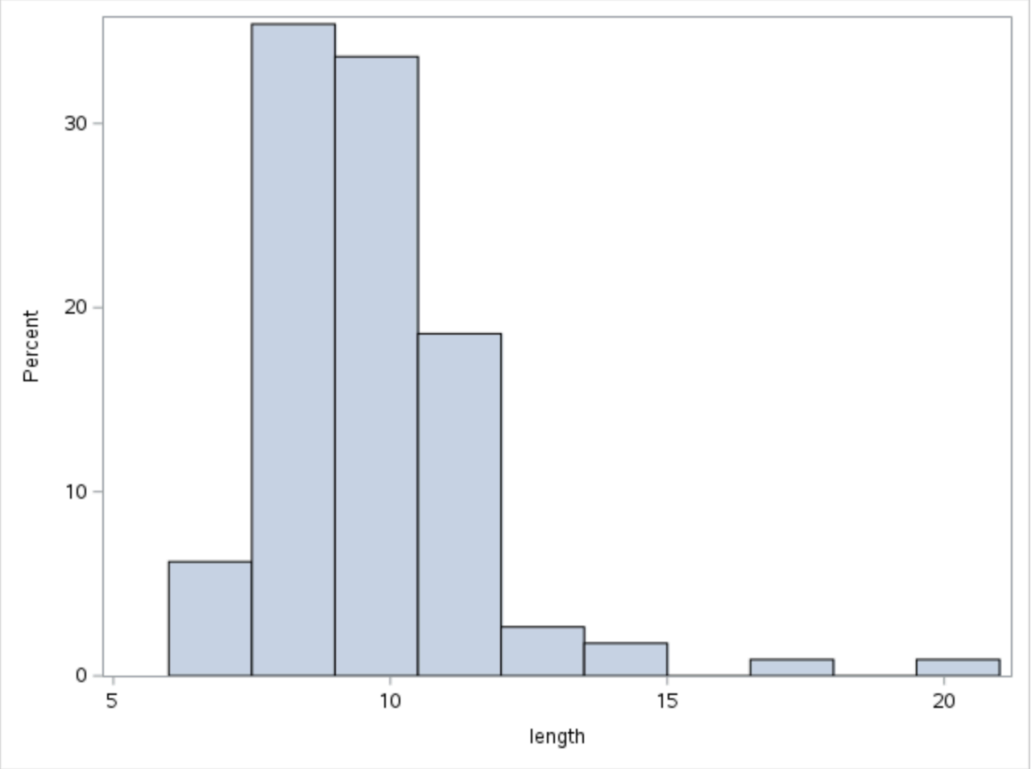
Appendix Graph 6: Histogram of msch in question 3b



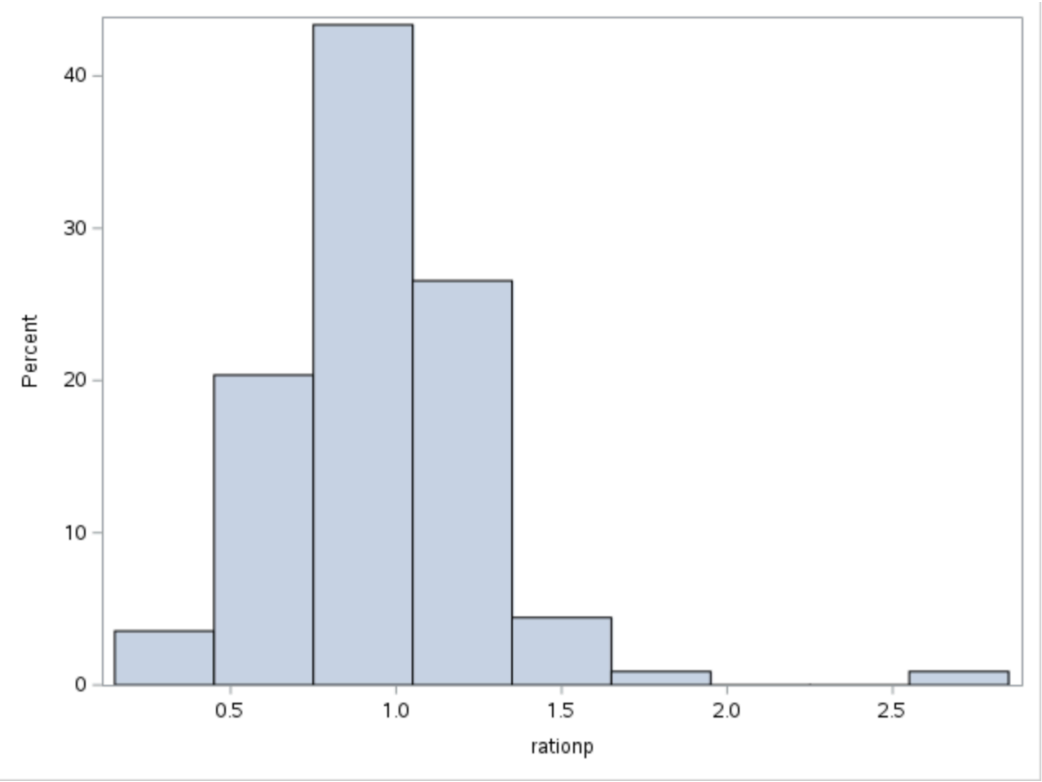
Appendix Graph 7: Histogram of xray in question 3b



Appendix Graph 8: Histogram of length in question 3b



Appendix Graph 9: Histogram of nurse/patient ratio


40
30
20
10
0
Percent
0.5
1.0
1.5
2.0
2.5
rationp