STAT448 - Advanced Data Analysis

Homework 4

Name: Xinyan Yang

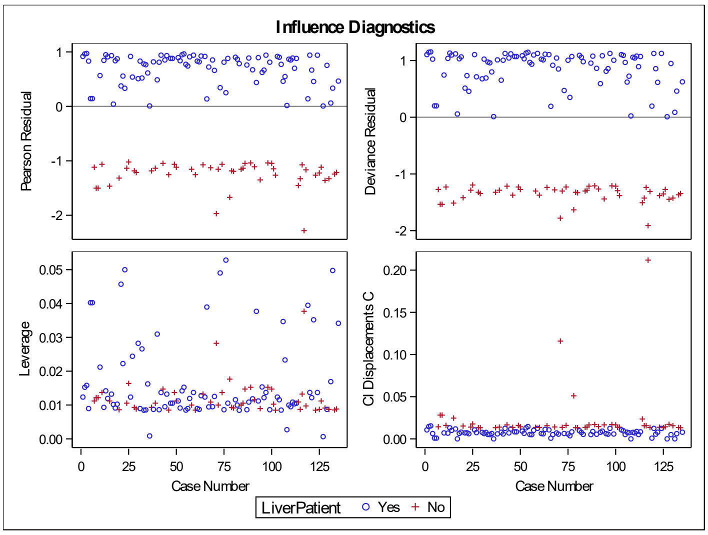
Exercise 1 Solution:

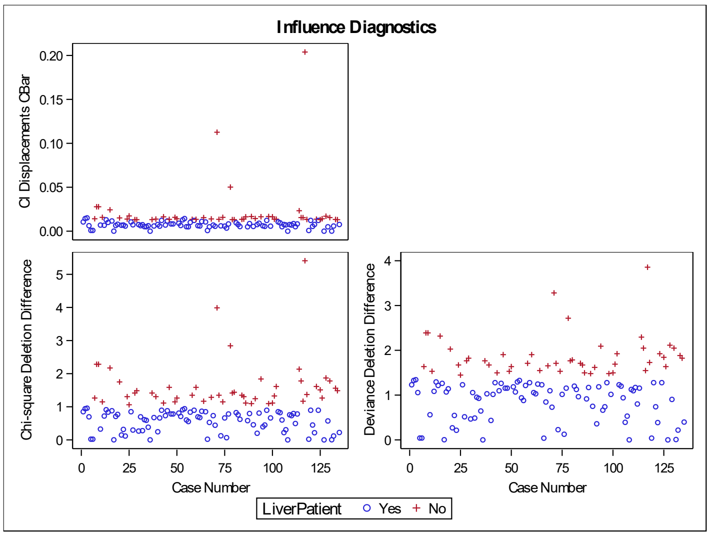
(a).

Testing Global Null Hypothesis: BETA=0			
Test	**Chi-Square**	**DF**	**Pr > ChiSq**
Likelihood Ratio	13.3381	1	0.0003
Score	7.1865	1	0.0073
Wald	5.8270	1	0.0158

Analysis of Maximum Likelihood Estimates					
Parameter	**DF**	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Pr > ChiSq**
Intercept	1	-0.1213	0.3061	0.1571	0.6918
Aspartate	1	0.0164	0.00679	5.8270	0.0158

Odds Ratio Estimates			
Effect	**Point Estimate**	**95% Wald Confidence Limits**	
Aspartate	1.017	1.003	1.030





Obs	Age	Gender	TB	DB	Alkphos	Alamine	Aspartate	TP	ALB	AGRatio	LiverPatient	resid_fem1	cd_fem1
117	28	Female	1	0.3	90	18	108	6.8	3.1	0.8	No	-2.28188	0.20391

First, I use the stepwise method to fit a logistic regression model with all predictors and get the above results. From the influence plots we can see that, there are no observations with cook's distance (which refers to CBar in the plots) greater than 0.5 in this model. However, I do find an observation with its absolue value of Pearson residual greater than 2, and I decide to remove it and refit the model.

(b).

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	17.2614	1	<.0001
Score	7.9344	1	0.0049
Wald	7.0431	1	0.0080

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-0.3120	0.3440	0.8224	0.3645
Aspartate	1	0.0238	0.00897	7.0431	0.0080

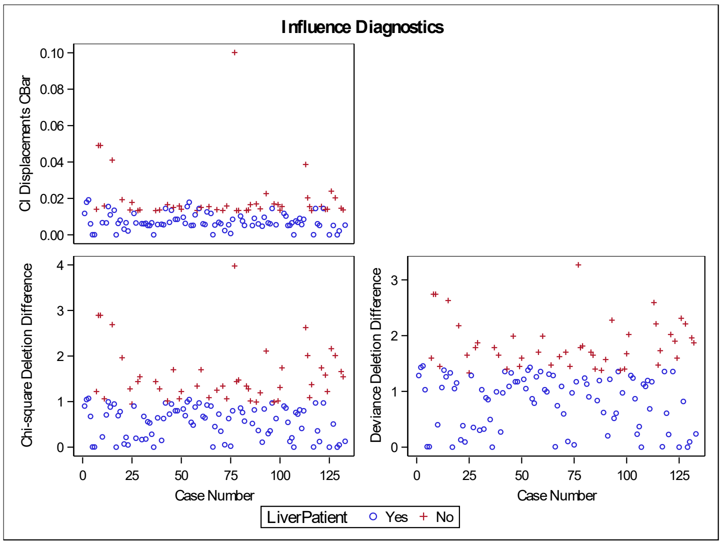
Odds Ratio Estimates

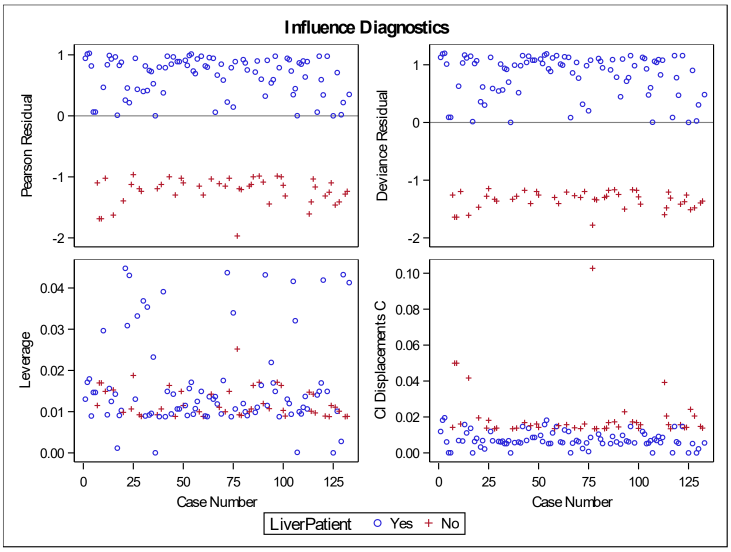
Effect	Point Estimate	95% Wald Confidence Limits	
Aspartate	1.024	1.006	1.042

Partition for the Hosmer and Lemeshow Test

		LiverPatient = Yes		LiverPatient = No	
Group	Total	Observed	Expected	Observed	Expected
1	11	5	5.45	6	5.55
2	13	8	6.68	5	6.32
3	14	10	7.54	4	6.46
4	12	7	6.70	5	5.30
5	13	6	7.52	7	5.48
6	14	10	8.47	4	5.53
7	15	7	9.70	8	5.30
8	13	7	9.32	6	3.68
9	13	12	10.94	1	2.06
10	15	15	14.67	0	0.33

Hosmer and Lemeshow Goodness-of-Fit Test		
Chi-Square	**DF**	**Pr > ChiSq**
8.9558	8	0.3460





I have removed two influential points with pearson residual value <-2 and I refit the logistic model and get only one significant predictor which is aspartate.

The p-value of aspartate is 0.008, which means that aspartate is significant under the significance level of 5%.

The P-value of Hosmer-Lemeshows test result is 0.346, which means that we should not reject the null hypothesis and we conclude that there is no lackness of fit in this model.

From the influence plots we can find that, most points have a cook’s distance which less than 0.05, and there is one observation has a really large Cbar value.

(c).

The confidence interval of odds ratio doesn’t include 0, which means that the predictor aspartate is statistically significant. The point estimate of aspartate is 1.024, which means that the odds of female having a liver would increase by exp^1.024 with one unit increase in aspartate.

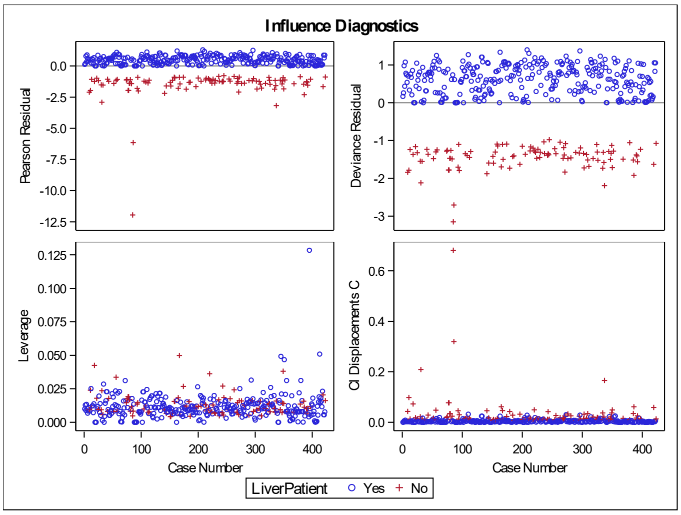
2. Solution:

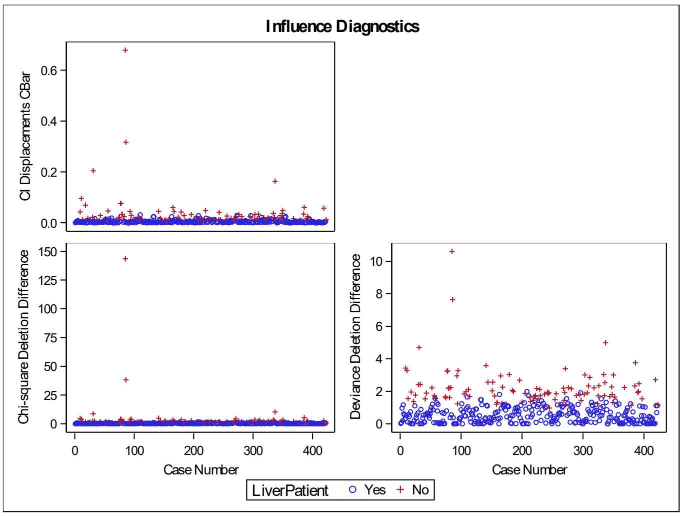
(a)

Testing Global Null Hypothesis: BETA=0			
Test	**Chi-Square**	**DF**	**Pr > ChiSq**
Likelihood Ratio	80.8655	4	<.0001
Score	41.2153	4	<.0001
Wald	34.9743	4	<.0001

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-0.2194	0.6509	0.1136	0.7360
Age	1	0.0190	0.00819	5.3685	0.0205
DB	1	0.5046	0.1723	8.5752	0.0034
Alamine	1	0.0182	0.00526	11.9229	0.0006
AGRatio	1	-0.8511	0.4140	4.2261	0.0398

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
Age	1.019	1.003	1.036
DB	1.656	1.182	2.322
Alamine	1.018	1.008	1.029
AGRatio	0.427	0.190	0.961





Obs	Age	Gender	TB	DB	Alkphos	Alamine	Aspartate	TP	ALB	AGRatio	LiverPatient	resid_ma1	cd_ma1
85	50	Male	5.8	3	661	181	285	5.7	2.3	0.67	No	-11.9466	0.67834

Obs	Age	Gender	TB	DB	Alkphos	Alamine	Aspartate	TP	ALB	AGRatio	LiverPatient	resid_ma1	cd_ma1
85	50	Male	5.8	3.0	661	181	285	5.7	2.3	0.67	No	-11.9466	0.67834
86	50	Male	7.3	3.6	1580	88	64	5.6	2.3	0.60	No	-6.1538	0.31694

After perform the stepwise selection, we have 4 predictors left. From the influence plots we can find that, there is one observation with cook's distance greater than 0.5 and there are two observations with residual less than -5. I decide to refit the model after removing these three observations.

(b)

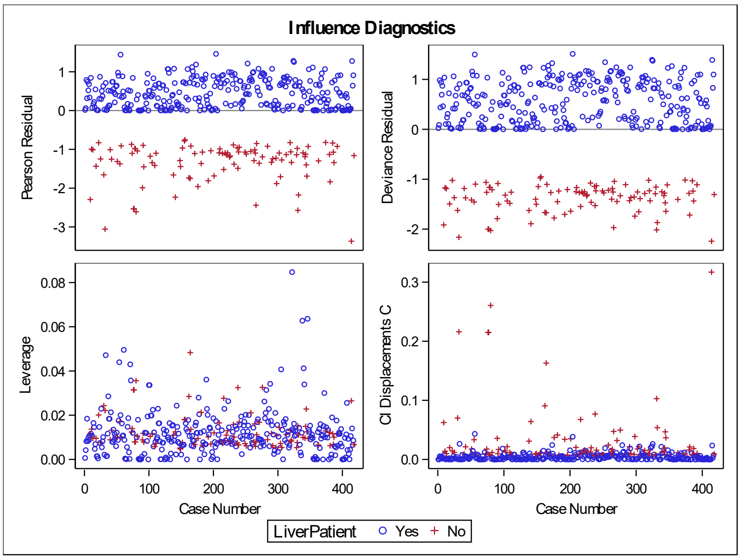
Model Fit Statistics		
Criterion	**Intercept Only**	**Intercept and Covariates**
AIC	466.539	371.029
SC	470.574	391.207
-2 Log L	464.539	361.029

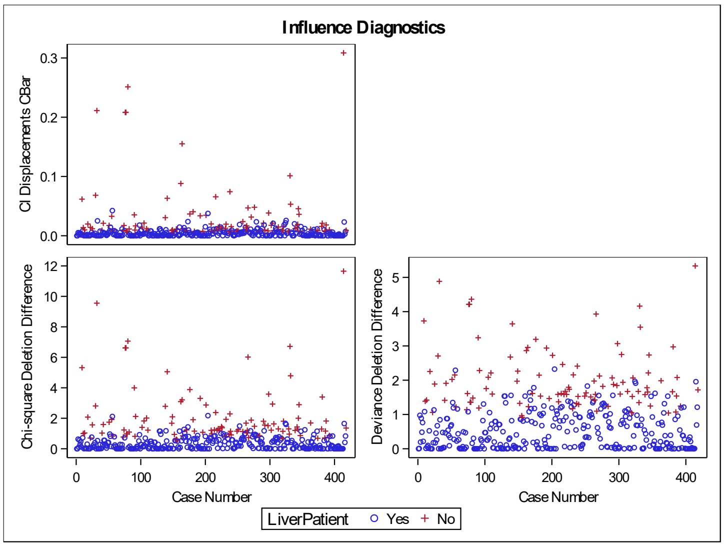
Testing Global Null Hypothesis: BETA=0			
Test	**Chi-Square**	**DF**	**Pr > ChiSq**
Likelihood Ratio	103.5096	4	<.0001
Score	46.5621	4	<.0001
Wald	37.4811	4	<.0001

Analysis of Maximum Likelihood Estimates					
Parameter	**DF**	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Pr > ChiSq**
Intercept	1	-2.1636	0.5644	14.6970	0.0001
Age	1	0.0176	0.00838	4.4040	0.0359
DB	1	0.7113	0.2590	7.5405	0.0060
Alkphos	1	0.00631	0.00185	11.6205	0.0007
Aspartate	1	0.00910	0.00364	6.2581	0.0124

Odds Ratio Estimates			
Effect	**Point Estimate**	**95% Wald Confidence Limits**	
Age	1.018	1.001	1.035
DB	2.037	1.226	3.384
Alkphos	1.006	1.003	1.010
Aspartate	1.009	1.002	1.016

Hosmer and Lemeshow Goodness-of-Fit Test		
Chi-Square	**DF**	**Pr > ChiSq**
5.2159	8	0.7343





After removing all influential points and fit the logistic model, I get the above results.

The model choose 4 predictors which are age, DB, alkphos, aspartate.

The p-value of all 4 predictors are all less than 0.05, which mean that these 4 predictors all significant under the significance level of 5%.

The P-value of Hosmer-Lemeshows test result is 0.7343, which means that we should not reject the null hypothesis and we conclude that there is no lackness of fit in this model.

From the influence plots we can find that, all points have a cook’s distance which less than 0.5, but there still exists some points which have a larger value than the others, and the residuals are between -3 and 2.

(c).

The confidence interval of all 4 odds ratios don’t include 0, which means that all 4 predictors are statistically significant.

The point estimate of age is 1.018, which means that the odds of male having a liver would increase by exp^1.018 with one unit increase in age.

The point estimate of DB is 2.037, which means that the odds of male having a liver would increase by exp^2.037 with one unit increase in DB.

The point estimate of alkphos is 1.006, which means that the odds of male having a liver would increase by exp^1.006 with one unit increase in alkphos.

The point estimate of aspartate is 1.009, which means that the odds of having a liver would increase by exp^1.009 with one unit increase in aspartate.

The difference between the model for male and female is that the model for female has only 1 predictor and the model for male has 4 predictors.

3. Solution:

(a)

Model Information	
Data Set	WORK.HYPER
Distribution	Gamma
Link Function	Log
Dependent Variable	bp

Analysis Of Maximum Likelihood Parameter Estimates

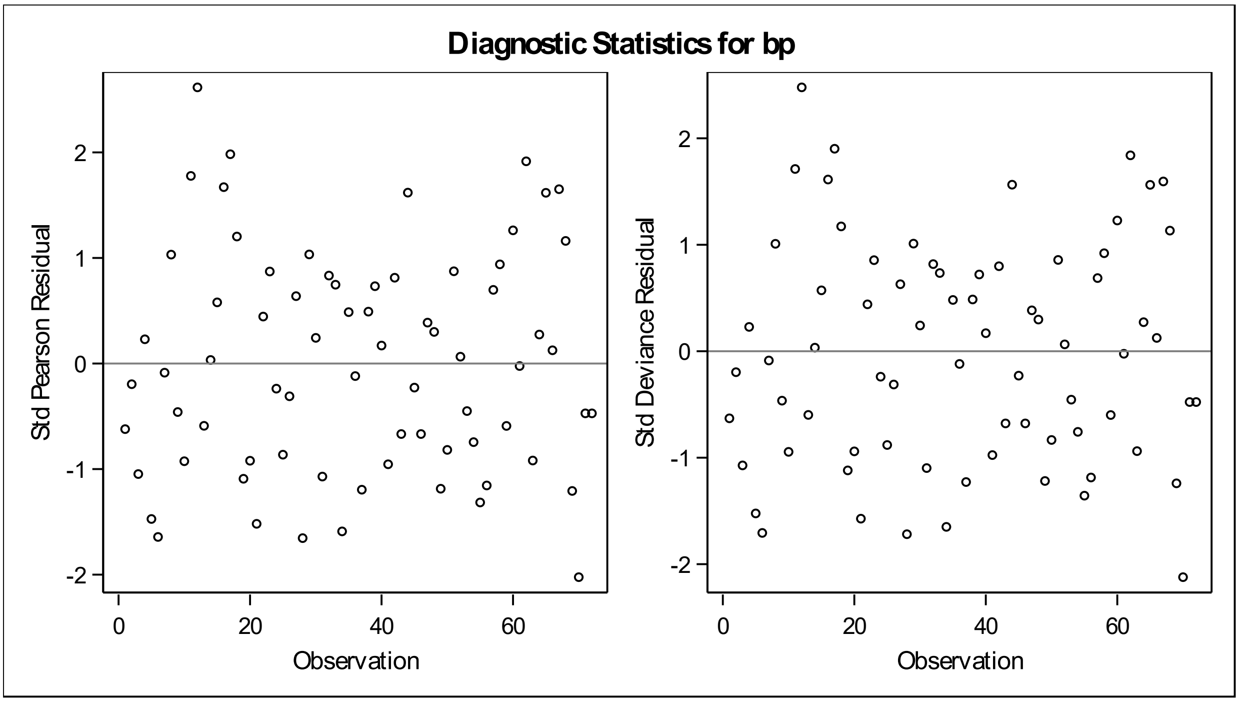
Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept		1	5.1612	0.0182	5.1256	5.1968	80724.3	<.0001
drug	X	1	-0.0740	0.0198	-0.1129	-0.0352	13.94	0.0002
drug	Y	1	0.0138	0.0198	-0.0251	0.0526	0.48	0.4879
drug	Z	0	0.0000	0.0000	0.0000	0.0000	.	.
diet	N	1	0.0907	0.0162	0.0589	0.1224	31.36	<.0001
diet	Y	0	0.0000	0.0000	0.0000	0.0000	.	.
biofeed	A	1	0.0580	0.0162	0.0262	0.0897	12.82	0.0003
biofeed	P	0	0.0000	0.0000	0.0000	0.0000	.	.
Scale		1	211.9407	35.2957	152.9181	293.7446		

LR Statistics For Type 1 Analysis

Source	2*LogLikelihood	DF	Chi-Square	Pr > ChiSq
Intercept	-617.1508			
drug	-603.9951	2	13.16	0.0014
diet	-581.2634	1	22.73	<.0001
biofeed	-569.4689	1	11.79	0.0006

LR Statistics For Type 3 Analysis

Source	DF	Chi-Square	Pr > ChiSq
drug	2	19.57	<.0001
diet	1	26.03	<.0001
biofeed	1	11.79	0.0006



From the above results we can see that, all 3 predictors(drug, diet, biofeed) are significant under the significance level of 5%. And the MLE results tell us that drug Y, diet N and biofeed A would increase the blood pressure.

From the residual plot we can see that, there's no obvious trend in residuals and most fall in -2 to 2, therefore the assumptions of the model seem reasonable.

(b).

Model Information	
Data Set	WORK.HYPER
Distribution	Poisson
Link Function	Log
Dependent Variable	bp

Criteria For Assessing Goodness Of Fit			
Criterion	**DF**	**Value**	**Value/DF**
Deviance	67	62.3950	0.9313
Scaled Deviance	67	62.3950	0.9313
Pearson Chi-Square	67	62.6547	0.9351
Scaled Pearson X2	67	62.6547	0.9351
Log Likelihood		56056.9547	
Full Log Likelihood		-285.0646	
AIC (smaller is better)		580.1291	

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
AICC (smaller is better)		581.0382	
BIC (smaller is better)		591.5125	

Analysis Of Maximum Likelihood Parameter Estimates

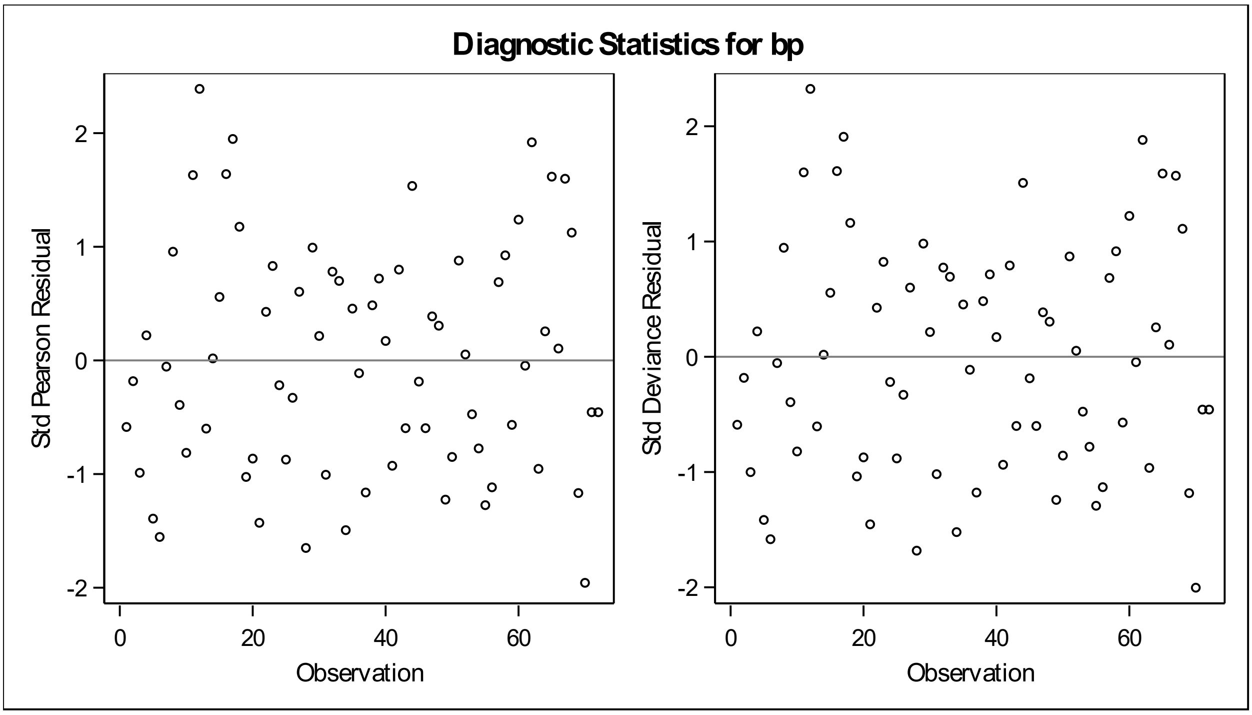
Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept		1	5.1613	0.0196	5.1229	5.1997	69435.9	<.0001
drug	X	1	-0.0758	0.0215	-0.1179	-0.0338	12.50	0.0004
drug	Y	1	0.0132	0.0210	-0.0279	0.0543	0.40	0.5293
drug	Z	0	0.0000	0.0000	0.0000	0.0000	.	.
diet	N	1	0.0922	0.0174	0.0582	0.1263	28.17	<.0001
diet	Y	0	0.0000	0.0000	0.0000	0.0000	.	.
biofeed	A	1	0.0578	0.0174	0.0238	0.0919	11.10	0.0009
biofeed	P	0	0.0000	0.0000	0.0000	0.0000	.	.
Scale		0	1.0000	0.0000	1.0000	1.0000		

LR Statistics For Type 1 Analysis

Source	Deviance	Num DF	Den DF	F Value	Pr > F	Chi-Square	Pr > ChiSq
Intercept	121.7935						
drug	101.7019	2	67	10.79	<.0001	21.57	<.0001
diet	73.4968	1	67	30.29	<.0001	30.29	<.0001
biofeed	62.3950	1	67	11.92	0.0010	11.92	0.0006

LR Statistics For Type 3 Analysis

Source	Num DF	Den DF	F Value	Pr > F	Chi-Square	Pr > ChiSq
drug	2	67	10.79	<.0001	21.57	<.0001
diet	1	67	30.29	<.0001	30.29	<.0001
biofeed	1	67	11.92	0.0010	11.92	0.0006



From the above results we can see that, all 3 predictors(drug, diet, biofeed) are significant under the significance level of 5%. And the MLE results tell us that drug Y, diet N and biofeed A would increase the blood pressure.

The scaled deviance is less than 1, so the model is under diepersed.

From the residual plot we can see that, there's no obvious trend in residuals, therefore the assumptions of the model seem reasonable.

(c).

Results from the ANOVA model:

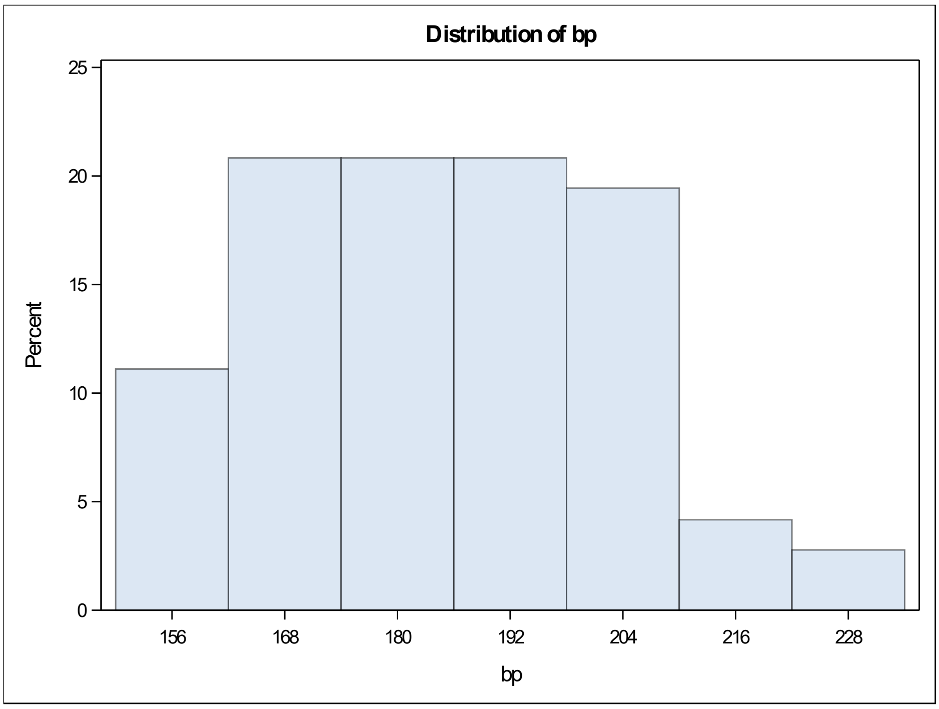
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	10925.00000	2731.25000	15.68	<.0001
Error	67	11669.00000	174.16418		
Corrected Total	71	22594.00000			

R-Square	Coeff Var	Root MSE	bp Mean
0.483535	7.152915	13.19713	184.5000

Source	DF	Anova SS	Mean Square	F Value	Pr > F
drug	2	3675.000000	1837.500000	10.55	0.0001
diet	1	5202.000000	5202.000000	29.87	<.0001
biofeed	1	2048.000000	2048.000000	11.76	0.0010

The similarities between these three models are that both 3 predictors are significant.

I think both the gamma model and poisson model would be better than the ANOVA model, because anova assumes normality of the data but from the histogram we can see than , the response variable actually has a left-skewed distribution. So I would prefer the gamma and poisson model.


Distribution of bp
Percent
bp