Introduction

Introduction for Problem 1

The data includes observations which describe the information women and their babies. The purpose of the project is to find the best model which can best predict whether the infant will have a low birth weight or not. The model selected for this problem is Logistic Regression. It is because the response, whether the infant will have a low birth weight or not, is a binary.

Materials and Methods

Materials for Problem 1

There are 189 observations and 7 different columns. The response variable is if the birth weight of the infant was low or not. The predictor variables are the age of the mother, the weight (in pounds) of the mother (before pregnancy), smoking status during pregnancy (with levels yes and no), history of pre-mature labor (with levels yes and no), history of hypertension (with levels yes or no) and the number of visits during the first trimester in first three months. The scatter plot matrix among the variables is provided in Figure 1. We can see age and weight are skewed to right. The history of pre-mature labor (with levels yes and no) has high correlation with the response variable, if the birth weight of the infant was low or not.

Methods for Problem 1

The logistic regression is fitted to all first order variables. Also, the interaction variables age and weight, weight and hypertension, and weight and pre are included in the initial model. Then, we perform Hosmer and Lemeshow goodness of fit test to test the model specification. The Hosmer and Lemeshow goodness of fit test is performed. The null hypothesis for the test is that the model is correctly specified. The alternative hypothesis for the test is that the model is incorrectly specified. Next, the diagnostics are done for the initial model.

Then, the model selection is performed. It includes forward selection, backward selection, forward-backward selection and backward-forward selection. For each method, we use both AIC and BIC criteria to select the best model. The best model is selected based on the result in Hosmer and Lemeshow goodness of fit test. Since we would like to focus on having a good prediction result, we would use the result given by AIC since it has a larger model than the one selected by BIC.

Results

Results for Problem 1

The estimated initial model is given in Table 1. The diagnostic plot is given in Figure 2. We do not see any outliers or influential points from the plots.

The model selection is performed. It includes forward selection, backward selection, forward-backward selection and backward-forward selection. For each method, we use both AIC and BIC criteria to select the best model. The best model selected for each method is given in Table 2. Therefore, we have four model selected based on AIC criterion and another four model based on BIC criterion. For the four models based on AIC, the best one is the model with the lowest AIC. Using AIC, the best model selected is "birth ~ pre + I(age * weight) + hyp + smoke". For the four models based on BIC, the best one is the model with the lowest BIC. Using BIC, the best model selected is "birth~pre + I(age * weight) + hyp".

Based on the best model selected by AIC, the test statistic is 11.341 with p-value is 0.1831. Since the p-value is larger than 0.05, we fail to reject the null hypothesis at 5% significance level. Therefore, the model selected by AIC is correctly specified. Based on the best model selected by BIC, the test statistic is 7.5534 with p-value is 0.4783. Since the p-value is larger than 0.05, we fail to reject the null hypothesis at 5% significance level. Therefore, the model selected by BIC is correctly specified. Since the goal is to perform a good prediction, we decided to select the model based on AIC criterion since it provides a larger model than the model selected based on BIC.
The final model includes the predictors age, weigth, smoke, pre and hyp. The diagnostic plot for the final model is given in Figure 3. The distribution of the two types of residuals look similar. Therefore, there is no lack-of-fit.

The confusion matrix when we use the final model to predict the result is

		predict	
		0	1
Truth	0	21	38
	1	10	120

The accuracy rate is (21+120)/(21+38+10+120) = 74.60%.

Conclusion and Discussion

Conclusion and Discussion For Problem 1

The final model selected is

$$\hat{\pi} = \frac{\exp(-2.03 + 0.06Age + 0.0162Weight - 0.5184Smoke - 1.7940pre - 1.7827hyp)}{1 + \exp(-2.03 + 0.06Age + 0.0162Weight - 0.5184Smoke - 1.7940pre - 1.7827hyp)}$$

The misclassification rate of the final model is 25.4%. The interpretation for each coefficient is given below:

(1) As age is increased by 1 year, the odds ratio of that birth weight of the infant was low is increased by a factor exp(-1.794) = 1.0618.

(2) As the weight is increased by 1 pounds, the odds ratio of that birth weight of the infant was low is increased by a factor exp(0.06) = 1.0618.
(3) As we compare smoker vs non-smoker, the odds ratio of that birth weight of the infant was low is decreased by a factor exp(0.5184) = 1.679339.
(4) As we compare people with history of pre-mature labor to people who do not have the history of pre-mature labor, the odds ratio of that birth weight of the infant was low is decreased by a factor exp(1.7940) = 6.013458.
(5) As we compare people with history of hypertension to people who do not have the history of hypertension, the odds ratio of that birth weight of the infant was low is decreased by a factor exp(1.7827) = 5.945889, controlling for other variables.

Introduction for Problem 2

The data includes observations which describe health insurance consumers who had claims related to ischemic, also known as heart disease. The purpose of the project is to model the mean of the number emergency room visits as a function of 8 other variables provided in the data. The model selected for this problem is Poisson Regression. It is because the response, the number emergency room visits, is a count data.

Materials for Problem 2

There are 788 observations and 9 different columns. The response variable is number of emergency room visits. The predictor variables are total cost of claims made by subscriber (dollars), age of subscriber (years), gender of subscriber (1=male, 0=otherwise), total number of interventions or procedures carried out, number of tracked drugs prescribed, number of other complications that arose during the heart disease treatment, number of other diseases that the subscriber had during the period and number of days of duration of treatment condition.

Methods for Problem 2

There are two initial models. The first initial model is to fit a Poisson regression model with all first order predictors and its interaction with gender. The second initial model is to fit a Poisson regression model with all squared-root-transformed predictors (except gender) and its interaction with gender. Then, we perform a model diagnostic. The best model is selected by stepwise selection with AIC criterion.

Result for Problem 2

The first initial model with all first order predictors and its interaction with gender is given in Table 5. The second initial model with all squared-root-transformed predictors (except gender) and its interaction with gender is given in Table 6. Not all the predictors are significant in both initial models. Next, we perform the diagnostic plot. The Pearson residuals against the fitted value is given in Figure 3. We can see most of the values are under 5 counts. Figure 7 and 8 are the best model selected through stepwise selection with AIC criterion for the untransformed and transformed cases respectively. Since all the predictors in the best untransformed model are significant, we select it as the final model. It is given in Figure 9. The goodness of fit test is performed. For the best model, the test statistic is 1034.9 with degree of freedom 779. The p-value is about 0. We reject the null hypothesis and conclude that the model did not fit well.

Conclusion and Discussion For Problem 2

The best model selected is

$$\begin{aligned}\log(\hat{\mu}) = 0.484 + 1.46 \times 10^{-5} cost + 6.75 \times 10^{-3} age + 0.2166 Gender \\ + 1.074 \times 10^{-2}\ inter + 0.1927\ drugs + 0.239 \times complications \\ + 3.19 \times 10^{-4}\ duration - 0.3566 \times gender \times complications\end{aligned}$$

If the cost is increased by 1 unit, the expected count is increased by a factor of $exp(1.46 \times 10^{-5})$, with holding other predictors constant. If the age is increased by 1 unit, the expected count is increased by a factor of $exp(6.75 \times 10^{-3})$, with holding other predictors constant. If the inter is increased by 1 unit, the expected count is increased by a factor of $exp(1.074 \times 10^{-2})$, with holding other predictors constant. If the drugs is increased by 1 unit, the expected count is increased by a factor of $exp(0.1927)$, with holding other predictors constant. If the complications is increased by 1 unit, the expected count is increased by a factor of $exp(0.239)$ for female, with holding other predictors constant. If the complications is increased by 1 unit, the expected count is increased by a factor of $exp(0.239 - 0.3566)$ for male, with holding other predictors constant. If the duration is increased by 1 unit, the expected count is increased by a factor of $exp(3.19 \times 10^{-4})$, with holding other predictors constant.

Appendix A: Tables

Table 1: The initial model of the logistic regression

Estimate	Std. Error	z value	Pr(>|z|)
-3.2742	3.9296	-0.8332	0.4047
0.1165	0.1682	0.6927	0.4885
0.0259	0.0305	0.8495	0.3956
-0.5088	0.3545	-1.4351	0.1512
-2.6304	2.8681	-0.9171	0.3591
-1.6169	2.6786	-0.6036	0.5461
0.0292	0.1802	0.162	0.8713
-4.00E-04	0.0013	-0.3511	0.7255
-0.001	0.0172	-0.0565	0.9549
0.0065	0.0222	0.2921	0.7702

Table 2: Stepwise selection result with AIC as criterion

method	model	AIC
forward selection with AIC	birth~pre + I(age * weight) + hyp + smoke	212.9575326
backward selection with AIC	birth~age + weight + smoke + pre + hyp	214.1859568
forward then backward selection with AIC	birth~pre + I(age * weight) + hyp + smoke	212.9575326
backward then forward selection with AIC	birth~age + weight + smoke + pre + hyp	214.1859568

Table 3: Stepwise selection result with BIC as criterion

method	model	BIC
forward selection with BIC	birth~pre + I(age * weight) + hyp	213.1755461
backward selection with BIC	birth~weight + pre + hyp	215.4081003
forward then backward selection with BIC	birth~pre + I(age * weight) + hyp	213.1755461
backward then forward selection with BIC	birth~weight + pre + hyp	215.4081003

Table 3: Model selected with AIC

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.2595	0.5788	-0.4483	0.6539
pre	-1.8178	0.503	-3.6141	3.00E-04
I(age * weight)	6.00E-04	2.00E-04	2.9026	0.0037
hyp	-1.6557	0.6842	-2.4198	0.0155
smoke	-0.5188	0.3474	-1.4934	0.1353

Table 4: Model selected with BIC

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.5012	0.5541	-0.9045	0.3657
pre	-1.9215	0.4986	-3.8539	1.00E-04
I(age * weight)	6.00E-04	2.00E-04	2.981	0.0029
hyp	-1.6908	0.6957	-2.4303	0.0151

Table 5: Poisson Regression: Initial model with untransformed variables and interaction terms with gender

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.4768	0.2052	2.3234	0.0202
cost	0	0	5.1363	0
age	0.007	0.0035	2.0158	0.0438
gender1	0.2332	0.4025	0.5795	0.5622
inter	0.0102	0.0042	2.4504	0.0143
drugs	0.1894	0.015	12.653	0
complications	0.2436	0.0831	2.931	0.0034
comorbidities	0.0015	0.0041	0.3547	0.7228
duration	3.00E-04	2.00E-04	1.158	0.2469
cost:gender1	0	0	-0.8079	0.4191
age:gender1	-0.001	0.0068	-0.1466	0.8835
gender1:inter	0.0079	0.0114	0.6957	0.4866
gender1:drugs	0.0064	0.0287	0.2219	0.8244
gender1:complications	-0.3713	0.1307	-2.8396	0.0045
gender1:comorbidities	-0.0087	0.0095	-0.9108	0.3624
gender1:duration	3.00E-04	4.00E-04	0.6244	0.5324

Table 6: Poisson Regression: Initial model with transformed variables and interaction terms with gender.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.0049	0.3952	0.0124	0.9901
cost_sqrt	0.004	7.00E-04	5.7988	0
age_sqrt	0.0938	0.0515	1.8207	0.0686
gender1	0.0928	0.7776	0.1194	0.905
inter_sqrt	0.0209	0.0262	0.7967	0.4256
drugs_sqrt	0.4251	0.0319	13.3278	0
complications_sqrt	0.2117	0.0825	2.5662	0.0103
comorbidities_sqrt	-0.0188	0.0192	-0.9789	0.3276
duration_sqrt	0.0081	0.0048	1.6978	0.0895
cost_sqrt:gender1	-1.00E-04	0.0016	-0.0372	0.9704
age_sqrt:gender1	0.0214	0.1019	0.2098	0.8338
gender1:inter_sqrt	-0.0093	0.0547	-0.1694	0.8655
gender1:drugs_sqrt	-0.0292	0.0619	-0.4709	0.6377
gender1:complications_sqrt	-0.3184	0.1571	-2.0269	0.0427
gender1:comorbidities_sqrt	-0.0183	0.0406	-0.4499	0.6528
gender1:duration_sqrt	4.00E-04	0.0092	0.0464	0.963

Table 7: Poisson Regression: best model selected by AIC given the initial model with untransformed variables and interaction terms with gender

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.484	0.176	2.7502	0.006
cost	0	0	5.124	0
age	0.0068	0.003	2.2795	0.0226
gender1	0.2166	0.0452	4.7895	0
inter	0.0107	0.0038	2.8274	0.0047
drugs	0.1927	0.0123	15.68	0
complications	0.239	0.0826	2.892	0.0038
duration	3.00E-04	2.00E-04	1.8881	0.059
gender1:complications	-0.3566	0.1234	-2.8909	0.0038

Table 8: Poisson Regression: best model selected by AIC given the initial model with transformed variables and interaction terms with gender

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.0135	0.3387	-0.0397	0.9683
cost_sqrt	0.0043	4.00E-04	10.0521	0
age_sqrt	0.0998	0.0443	2.2521	0.0243
gender1	0.2052	0.0456	4.4957	0
drugs_sqrt	0.422	0.0269	15.7015	0
complications_sqrt	0.2171	0.0821	2.6431	0.0082
comorbidities_sqrt	-0.0239	0.0168	-1.429	0.153
duration_sqrt	0.0087	0.004	2.1666	0.0303
gender1:complications_sqrt	-0.3464	0.1486	-2.3309	0.0198

Table 6: Best Poisson Regression model

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.0135	0.3387	-0.0397	0.9683
cost_sqrt	0.0043	4.00E-04	10.0521	0
age_sqrt	0.0998	0.0443	2.2521	0.0243
gender1	0.2052	0.0456	4.4957	0
drugs_sqrt	0.422	0.0269	15.7015	0
complications_sqrt	0.2171	0.0821	2.6431	0.0082
comorbidities_sqrt	-0.0239	0.0168	-1.429	0.153
duration_sqrt	0.0087	0.004	2.1666	0.0303
gender1:complications_sqrt	-0.3464	0.1486	-2.3309	0.0198

Appendix B: Figures

Figure 1: the scatterplot matrix for the variables in problem 1

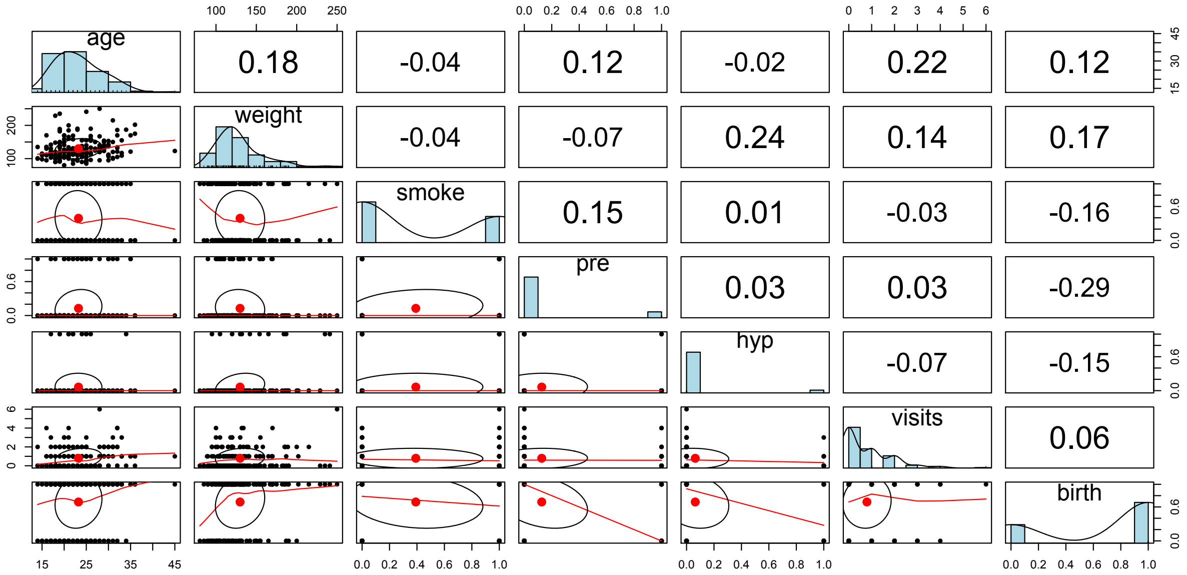


Figure 2: the diagnostic plots for the initial model for problem 1

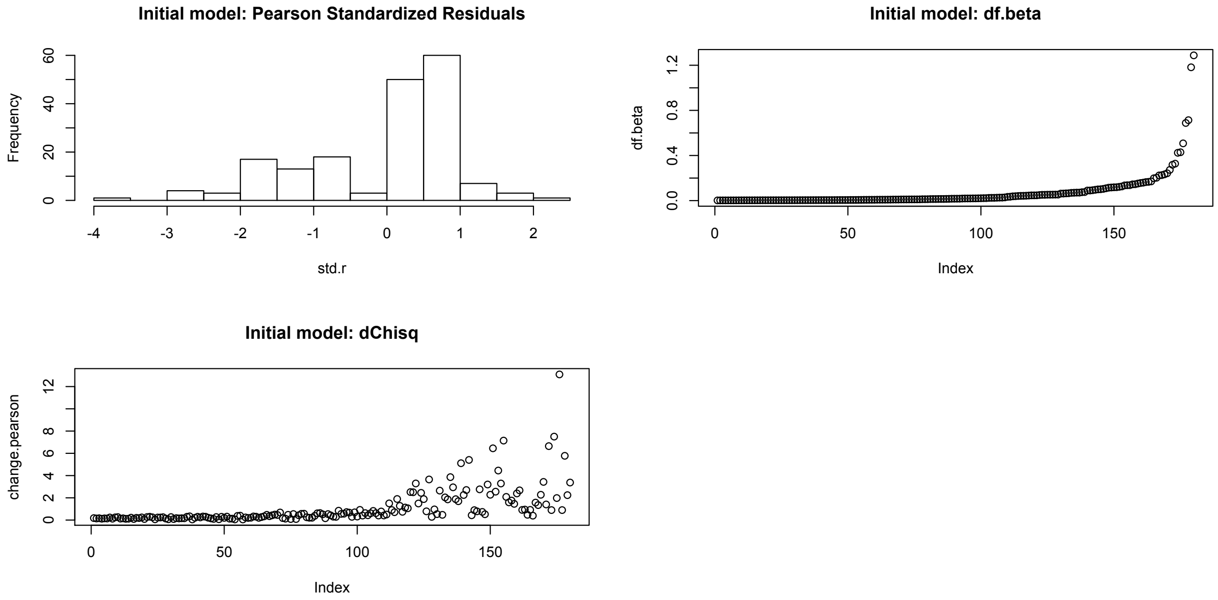


Figure 3: Distribution of Pearson and Deviance residuals for the final model

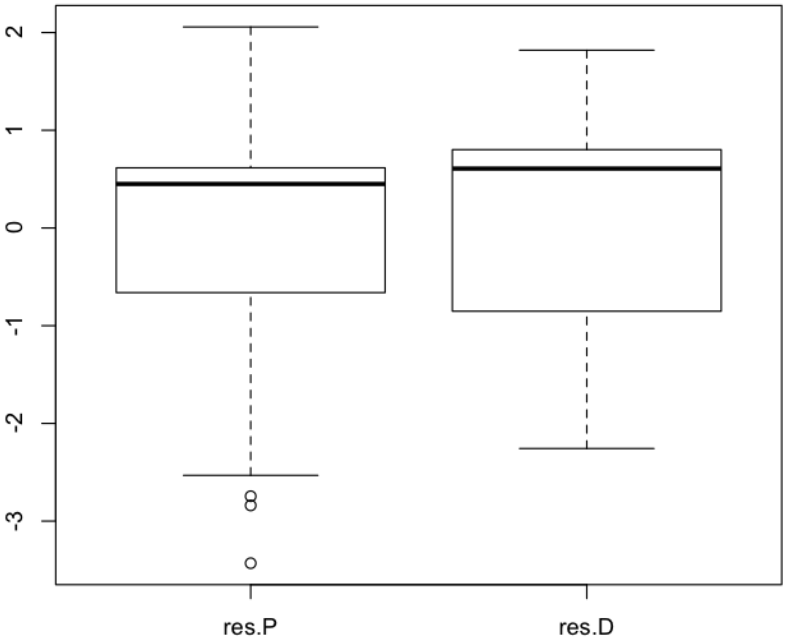
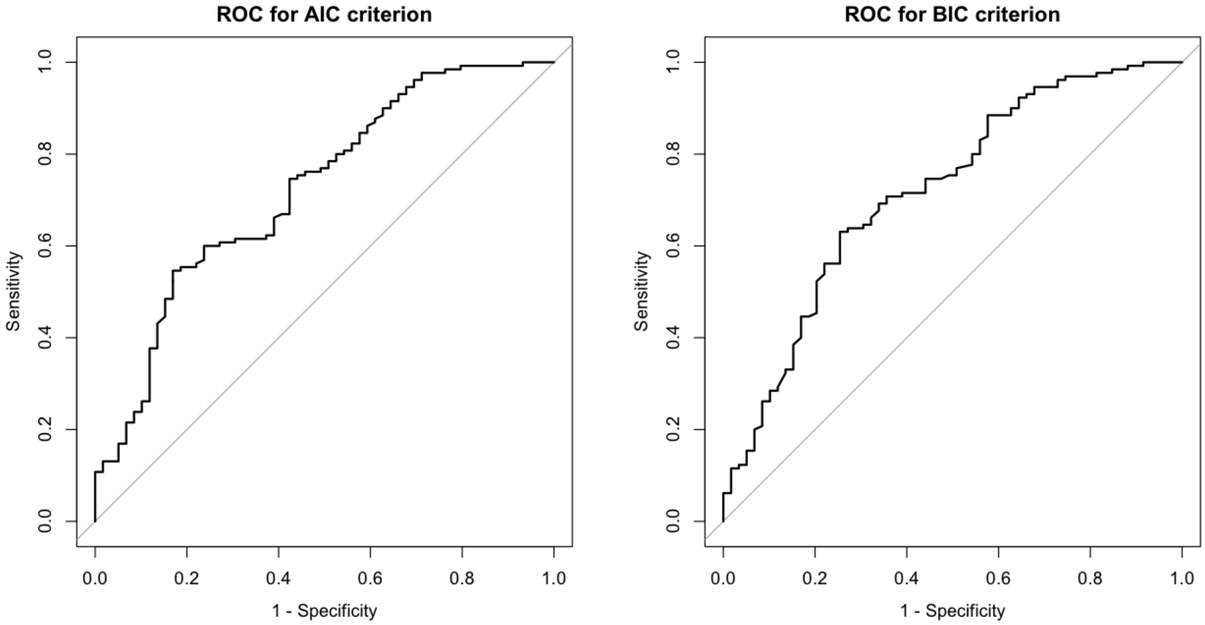
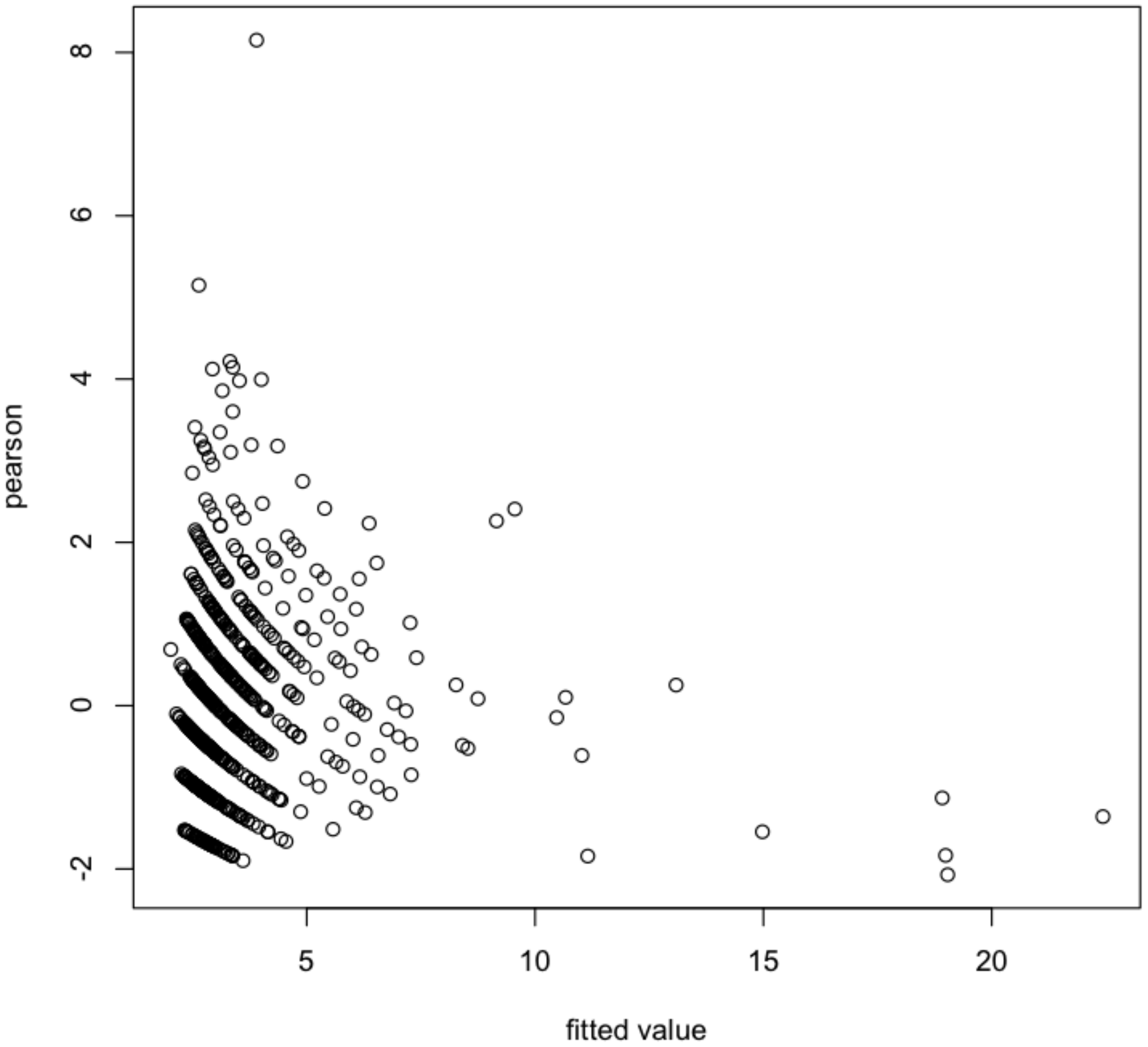


Figure 4: Prediction result based on model selected by AIC and BIC


ROC for AIC criterion

ROC for BIC criterion

Figure 5: Pearson residuals against fitted values.



Appendix C: Codes

```
library(readxl)
library(caret)
library(bestglm)
library(ResourceSelection)
library(LogisticDx)
library(pROC)
library(glmnet)

# Problem 1

baby = read_excel('baby.xls')

str(baby)
# Classes ‘tbl_df’, ‘tbl’ and 'data.frame':	189 obs. of  7 variables:
#  $ age   : num  19 33 20 21 18 21 22 17 29 26 ...
#  $ weight: num  182 155 105 108 107 124 118 103 123 113 ...
#  $ smoke : chr  "no" "no" "yes" "yes" ...
#  $ pre   : chr  "no" "no" "no" "no" ...
#  $ hyp   : chr  "no" "no" "no" "no" ...
#  $ visits: num  0 3 1 2 0 0 1 1 1 0 ...
#  $ birth : num  1 1 1 1 1 1 1 1 1 1 ...

baby$hyp = ifelse(baby$hyp == "yes", 1, 0)
baby$pre = ifelse(baby$pre == "yes", 1, 0)
baby$smoke = ifelse(baby$smoke == "yes", 1, 0)

# exploratory data analysis

pairs.panels(baby,
             method = "pearson", # correlation method
             hist.col = "lightblue",
             density = TRUE,  # show density plots
             ellipses = TRUE # show correlation ellipses
)

# fit model

initial_model = glm(birth ~ age + weight + smoke + pre + hyp + visits +
                    I(age*weight) + I(weight*hyp) + I(weight*pre), family=binomial(), data=baby)
summary(initial_model)$coef
write.csv(round(summary(initial_model)$coef, 4), "problem1_initial_model.csv")

empty_model = glm(birth ~ 1, data = baby, family = binomial(link = logit))
```

```
full_model = initial_model

# initial_model goodness of fit:
hoslem.test.initial_model = hoslem.test(initial_model$y, initial_model$fitted.values, g = 10);
hoslem.test.initial_model
#         Hosmer and Lemeshow goodness of fit (GOF) test
#
# data:  initial_model$y, initial_model$fitted.values
# X-squared = 5.3201, df = 8, p-value = 0.7229

# initial_model: leverage or outlier
par(mfrow=c(2,2))
good.stuff = dx(initial_model)
pear.r = good.stuff$Pr #Pearsons Residuals
deviance.r = good.stuff$dr #Deviance Residuals
std.r = good.stuff$sPr #Standardized residuals (Pearson)
df.beta = good.stuff$dBhat #DF Beta for removing each observation
change.pearson = good.stuff$dChisq #Change in pearson X^2 for each observation
change.LR = good.stuff$dDev #Change in LR-test G^2 for each observation
hist(std.r, main="Initial model: Pearson Standardized Residuals") #Histogram
plot(df.beta, main="Initial model: df.beta") # plot delta beta
good.stuff[df.beta > 0.1] # leverage points
plot(change.pearson, main="Initial model: dChisq") # plot chisq
good.stuff[change.pearson > 15] # outliers

# model selection

best.forward.AIC = step(empty_model, scope = list(lower = empty_model, upper = full_model),
                        direction = "forward", criterion = "AIC", trace = FALSE)
best.forward.AIC$formula #Forward selection
f1 = paste(best.forward.AIC$formula)
f1 = paste(f1[c(2,1,3)], collapse = '')

best.backward.AIC = step(full_model, scope = list(lower = empty_model, upper = full_model),
                         direction = "backward", criterion = "AIC", trace = FALSE)
best.backward.AIC$formula #Backward selection
f2 = paste(best.backward.AIC$formula)
f2 = paste(f2[c(2,1,3)], collapse = '')

best.FB.AIC = step(empty_model, scope = list(lower = empty_model, upper = full_model),
                   direction = "both", criterion = "AIC", trace = FALSE)
best.FB.AIC$formula #Forward/Backward Selectio
f3 = paste(best.FB.AIC$formula)
f3 = paste(f3[c(2,1,3)], collapse = '')

best.BF.AIC = step(full_model, scope = list(lower = empty_model, upper = full_model),
          direction = "both", criterion = "AIC", trace = FALSE)
best.BF.AIC$formula #Backward/Forward Selection
f4 = paste(best.BF.AIC$formula)
f4 = paste(f4[c(2,1,3)], collapse = '')

best.forward.BIC = step(empty_model, scope = list(lower = empty_model, upper = full_model),
             direction = "forward", k=log(n), trace = FALSE)
best.forward.BIC$formula #Forward selection
f5 = paste(best.forward.BIC$formula)
f5 = paste(f5[c(2,1,3)], collapse = '')

best.backward.BIC = step(full_model, scope = list(lower = empty_model, upper = full_model),
              direction = "backward", k=log(n), trace = FALSE)
best.backward.BIC$formula #Backward selection
f6 = paste(best.backward.BIC$formula)
f6 = paste(f6[c(2,1,3)], collapse = '')

best.FB.BIC = step(empty_model, scope = list(lower = empty_model, upper = full_model),
          direction = "both", k=log(n), trace = FALSE)
best.FB.BIC$formula #Forward/Backward Selection
f7 = paste(best.FB.BIC$formula)
f7 = paste(f7[c(2,1,3)], collapse = '')

best.BF.BIC = step(full_model, scope = list(lower = empty_model, upper = full_model),
          direction = "both", k=log(n), trace = FALSE)
best.BF.BIC$formula #Backward/Forward Selection
f8 = paste(best.BF.BIC$formula)
f8 = paste(f8[c(2,1,3)], collapse = '')

selection_result = data.frame(method = c("forward selection with AIC",
                     "backward selection with AIC",
                      "forward then backward selection with AIC",
                      "backward then forward selection with AIC",
                      "forward selection with BIC",
                      "backward selection with BIC",
                      "forward then backward selection with BIC",
                      "backward then forward selection with BIC"),
                model = c(f1,f2,f3,f4,f5,f6,f7,f8),
                `AIC/BIC` = c(AIC(best.forward.AIC),
                        AIC(best.backward.AIC),
                        AIC(best.FB.AIC),
                        AIC(best.BF.AIC),
                        AIC(best.forward.BIC),
                        AIC(best.backward.BIC),
                        AIC(best.FB.BIC),
                        AIC(best.BF.BIC)))
```

```
write.csv(selection_result, "problem1_model_selection.csv")

#The best AIC and BIC models:
best.model.AIC = best.forward.AIC
best.model.BIC = best.forward.BIC

#summary for AIC
write.csv(round(summary(best.model.AIC)$coef, 4), "problem1_AIC_model.csv")

#summary for BIC
write.csv(round(summary(best.model.BIC)$coef, 4), "problem1_BIC_model.csv")

## Hosmer and Lemeshow goodness of fit (GOF) test
# H0: the null hypothesis of the test is that the model is correctly specified

# AIC goodness of fit:
hoslem.test.AIC = hoslem.test(best.model.AIC$y, best.model.AIC$fitted.values, g = 10);
hoslem.test.AIC
#         Hosmer and Lemeshow goodness of fit (GOF) test
#
# data:  best.model.AIC$y, best.model.AIC$fitted.values
# X-squared = 11.341, df = 8, p-value = 0.1831

# BIC goodness of fit:
hoslem.test.BIC = hoslem.test(best.model.BIC$y, best.model.BIC$fitted.values, g = 10);
hoslem.test.BIC
#         Hosmer and Lemeshow goodness of fit (GOF) test
#
# data:  best.model.BIC$y, best.model.BIC$fitted.values
# X-squared = 7.5534, df = 8, p-value = 0.4783

###########

# leverage or outlier
good.stuff = dx(best.model.AIC)
pear.r = good.stuff$Pr #Pearsons Residuals
deviance.r = good.stuff$dr #Deviance Residuals
std.r = good.stuff$sPr #Standardized residuals (Pearson)
df.beta = good.stuff$dBhat #DF Beta for removing each observation
change.pearson = good.stuff$dChisq #Change in pearson X^2 for each observation
change.LR = good.stuff$dDev #Change in LR-test G^2 for each observation
hist(std.r, main="AIC model: Pearson Standardized Residuals") #Histogram
plot(df.beta, main="AIC model: df.beta") # plot delta beta
good.stuff[df.beta > 0.1] # leverage points
plot(change.pearson, main="AIC model: dChisq") # plot chisq
```

```
good.stuff[change.pearson > 15] # outliers

# prediction

par(mfrow=c(1,2))
the.roc1 = roc(best.model.AIC$y, best.model.AIC$fitted.values,auc = TRUE, ci =
TRUE,plot=TRUE, legacy.axes = TRUE, main="ROC for AIC criterion")
ci(the.roc1)
the.roc2 = roc(best.model.BIC$y, best.model.BIC$fitted.values,auc = TRUE, ci =
TRUE,plot=TRUE, legacy.axes = TRUE, main="ROC for BIC criterion")
ci(the.roc2)

# confusion matrix

pi0 = 0.50
my.table = table(truth = best.model.AIC$y,predict = ifelse(fitted(best.model.AIC)>pi0,1,0))
my.table

final_model = step(initial_model) # AIC
write.csv(round(summary(final_model)$coef, 4), "problem1_final_model.csv", row.names =
FALSE)

# model diagnostics

res.P = residuals(final_model, type="pearson")
res.D = residuals(final_model, type="deviance")
D_sq=sum(res.P)
G_sq=sum(res.D)
D_sq>qchisq(0.95,length(final_model)-5)
boxplot(cbind(res.P, res.D), labels = c("Pearson", "Deviance"))

# Problem 2

ischemic = read_excel("ischemic.xlsx")

# clean data
ischemic$gender=as.factor(ischemic$gender)

# fit poisson model
model1 = glm(visits ~ cost + age + gender + inter + drugs + complications + comorbidities +
duration, data=ischemic, family=poisson())
write.csv(round(summary(model1)$coef, 4), "problem2_initial_model.csv")
summary(ischemic.fit)

1-pchisq(1043.6,779)
```

```
# diagnostic

res.P = residuals(model1, type="pearson")
plot(model1$fitted.values,res.P,xlab="fitted value",ylab="pearson")

# stepwise selection
best.model1 = stepAIC(model1,direction="backward")

# using square term

model2 = glm(visits ~ .*.,data=ischemic)
best.model2 = stepAIC(model2,direction="backward")
write.csv(round(summary(best.model2)$coef, 4), "problem2_best.model2.csv")

# Deviance Residuals and Pearson Residuals plot.
res.P = residuals(model3, type="pearson")
res.D = residuals(model3, type="deviance") #or residuals(fit), by default
D_sq=sum(res.P)
G_sq=sum(res.D)
G_sq>qchisq(0.95,length(model3)-6)
boxplot(cbind(res.P, res.D), labels = c("Pearson", "Deviance"))
plot(model3$fitted.values,res.P,xlab="fitted value",ylab="pearson")

fit1 = glm(visits ~ cost + age + gender + inter + drugs + complications + comorbidities +
duration +

gender*cost+gender*age+gender*inter+gender*drugs+gender*complications+gender*comorbidi
ties+gender*duration, data=ischemic, family=poisson())

ischemic2 <- ischemic
ischemic2[,c("cost", "age", "inter", "drugs", "complications", "comorbidities", "duration")] =
sqrt(ischemic[,c("cost", "age", "inter", "drugs", "complications", "comorbidities", "duration")])
names(ischemic2) <- c("cost_sqrt", "age_sqrt", "gender", "inter_sqrt", "drugs_sqrt",
"complications_sqrt", "comorbidities_sqrt", "duration_sqrt", "visits")
fit2 = glm(visits ~ . + .*gender, data = ischemic2, family=poisson())

library(MASS)
fit1.best <- stepAIC(fit1)
summary(fit1.best)
fit2.best <- stepAIC(fit2)
summary(fit2.best)

# Deviance Residuals and Pearson Residuals plot.
res.P = residuals(fit1.best, type="pearson")
res.D = residuals(fit1.best, type="deviance") #or residuals(fit), by default
D_sq=sum(res.P)
```

```
G_sq=sum(res.D)
G_sq>qchisq(0.95,length(fit1.best)-6)
boxplot(cbind(res.P, res.D), labels = c("Pearson", "Deviance"))
plot(fit1.best$fitted.values,res.P,xlab="fitted value",ylab="pearson")
```