STA138 Final Project

My ex-girlfriend thinks I'm mean and my current one thinks I'm average...
Maybe you've set your bar too high?
$\overline{x}$

Ruriko Imai
Emily Ng

Problem 1 - student.csv

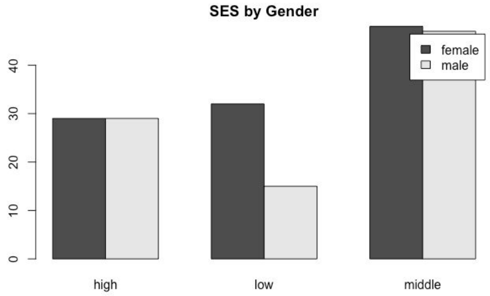
Introduction:

The goal of analyzing the student data is to able to predict the socioeconomic class (low, medium, high) of people. We will be using graphs, hypothesis tests, model selection, identification of outliers/influential points, predictive power assessment and goodness of fit by fitting the data into a multinomial logistic regression. The results will be of interest if we want to identify the predictors to one's socioeconomic status.

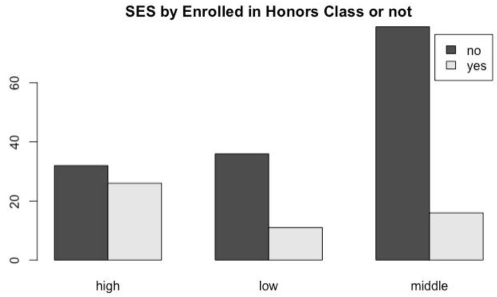
Summary:

The Student.csv is a data frame that consists of seven variables. The response variable is the socioeconomic class with 3 levels - low, medium, and high. There are 3 categorical variables which are gender (female, male), school type (public, private), honors (no, yes). The remaining 3 are continuous variables of standardized test scores of reading, writing, mathematics, and science.

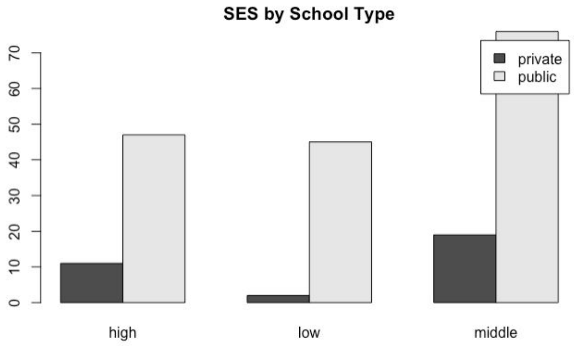
In order to attain an overview of the relationship between the response variable and its predictors, here are some visual representations.



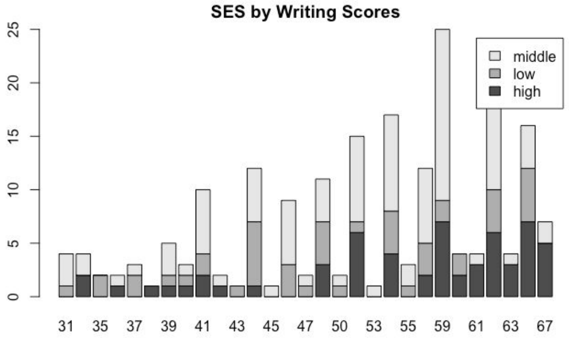
The SES compared to variable gender shows that middle and high ses groups are constructed of proportional number of both genders but low ses group is mainly of females.



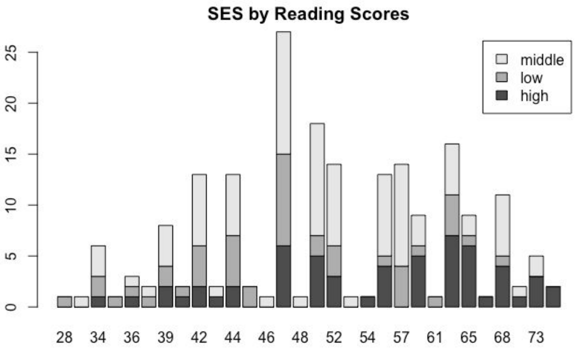
The SES compared to if a subject was enrolled in Honors class or not shows that all ses groups has a larger proportion of not enrolled in Honors class.



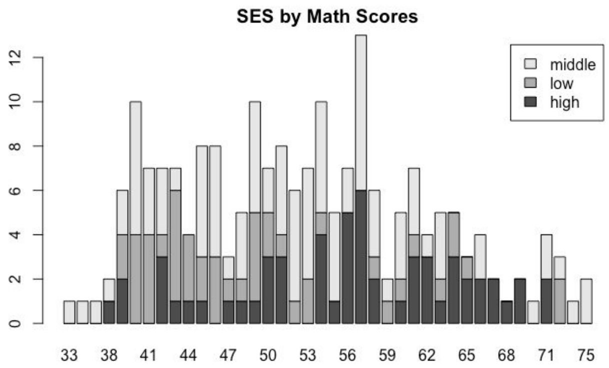
The SES compared to School type indicates that all ses groups have larger proportion of subjects from public schools.



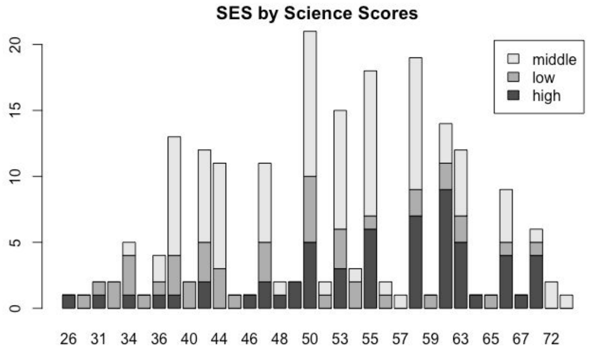
At a glance, middle ses group appears to be the larger proportion in all the writing scores.



Again, middle ses group appears to be the larger proportion in all the reading scores.



Most of the low ses group are observed in the lower range of the math scores, while middle ses group are in the lower through middle range and high ses group seems to almost have a normal distribution for the math standardized test scores.



All the ses groups are observed throughout the range of science scores, the middle group taking up a large proportion in the middle range.

From simple observations of these graphs, we are able to see some relationship between the response variable, ses, and its predictors. All the categorical variables (gender, school type, honors) seem to have some effect on ses. The graphs of continuous variables were harder to imply any valuable information, however, math scores is likely to have some effect on the ses since there is a pattern observed in all the ses groups. Now that we have some idea of what the relationship might be for the response variable, ses, let us conduct a quantitative analysis of the relationships.

Analysis:

Model Selection:

Out goal is to predict the response variable, ses. Larger models tend to have higher predictive power, so will use AIC and backward selection.

The full model using all the variables is, ses ~ factor(gender) + factor(schtyp) + factor(honors) + read + write + math + science. The empty model is ses ~ 1. Using backward stepwise selection and AIC gives us two models for the multinomial logistic regression with variables, (Intercept), read, factor(honors)yes, factor(schtyp)public, and factor(gender)male.

$$lm(\pi_{low}(x)/\pi_{high}(x)) \ = \ 2.644419 + (-0.07403496)X_1 + (-0.2937044)S_3 + (1.6367326)S_2 + (-0.8675304)S_1$$
$$ln(\pi_{middle}(x)/\pi_{high}(x)) \ = \ 2.304598 + (-0.02470590)X_1 + (-1.1706580)S_3 + (-0.0462631)S_2 + (-0.1910837)S_1$$

Residual Deviance: 383.7529
AIC: 403.7529

The coefficients are explained as follows:
S_1 = if gender is male, 0 otherwise
S_2 = if school type is public, 0 otherwise
S_3 = if student is enrolled in honors, 0 otherwise
X_1 = reading score
X_2 = writing score
X_3 = math score
X_4 = science score

CI:
The 95% profile-likelihood confidence intervals are as follows:

```
low
                           2.5 %       97.5 %
(Intercept)           -0.24282977  5.531667638
read                  -0.12360277 -0.024467154
factor(honors)yes     -1.35042458  0.763015878
factor(schtyp)public   0.03765038  3.235814731
factor(gender)male    -1.72806231 -0.006998487
```

```
middle
                          2.5 %      97.5 %
(Intercept)            0.1291376  4.48005772
read                  -0.0635126  0.01410079
factor(honors)yes     -2.0343238 -0.30699225
factor(schtyp)public  -0.9148289  0.82230270
factor(gender)male   -0.8949048  0.51273738
```

In order to conduct more diagnostics on the models, we will split them into simple logistic regression models by their ses, Low vs. High and Middle vs. High.

Then, we would use the logistic regression diagnostics on each. Importantly, the coefficients fit for two logistic regression models are not the same for the multinomial model. However, the "unusual observations" are the same for both.

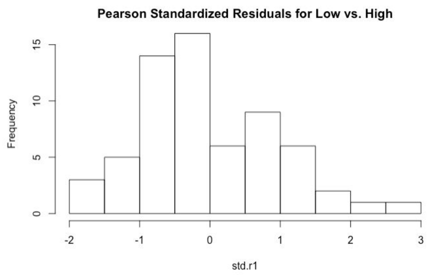
Split models:
The two split models are as follows:
model.LvsH = glm(ses ~ factor(gender) + factor(schtyp) + factor(honors) + read, data = LvsH, family = binomial(logit))
model.MvsH = glm(ses ~ factor(gender) + factor(schtyp) + factor(honors) + read, data = MvsH, family = binomial(logit))

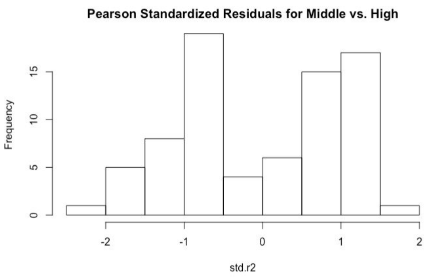
Diagnostics:

Hosmer Lemeshow goodness-of-fit:
H_0 : $The\ model\ fits\ well.$
H_A : $The\ model\ does\ not\ fit\ well.$

Low vs. High: X-squared = 3.0269, df = 8, p-value = 0.9327
Middle vs. High: X-squared = 8.8615, df = 8, p-value = 0.3541

Residuals and influence measures:



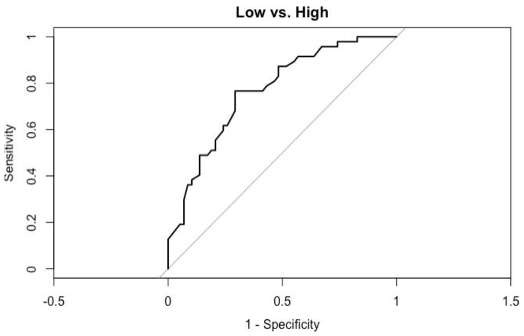


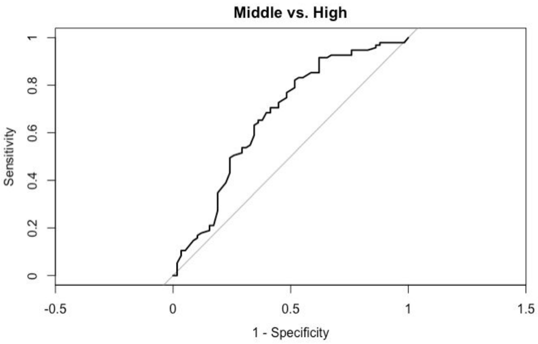
Influential Observations:
Both models do not have standard residuals or deviance greater than 3.

Predictive Power Assessments:

In general, higher values of simple correlation between Y_i and π_i are indicative of higher predictive power. However, these estimates can become highly biased upward, so they may not be reliable. Proportional Reduction in squared error: The higher this value is, the better.
Low vs. High: r1 = 0.4582591, prop.red1 = 0.2099446. Middle vs. High: r2 = 0.3042443, prop.red2 = 0.1049546.

The AUC are as follows:

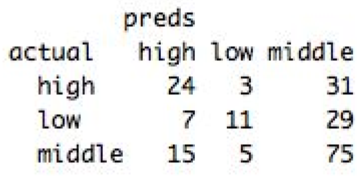




Area under the curve: 0.766
95% CI: 0.676-0.8559 (DeLong)

Area under the curve: 0.6712
95% CI: 0.5791-0.7634 (DeLong)

Error matrix and rate:



```
        preds
actual   high low middle
  high     24   3     31
  low       7  11     29
  middle   15   5     75
```

Error Rate = 0.45

Prediction:
The probability and the ses for a female, who scored 60 on all her tests, who went to private school, and who was not in the honors program.

Best.model: ses ~ read + honors + schtyp + gender
The ses is:

```
[1] middle
Levels: high low middle
```

The probabilities are:

```
     high        low     middle
0.29058933 0.04814572 0.66126495
```

Interpretation:

Interpretations of the coefficients for the selected model are as follows:

exp(1.6367326) indicates that the odds of having a low socioeconomic status compared to high socioeconomic status for male is multiplied by 5.138353 compared to females, holding all other variables constant.

exp(-0.0462631) indicates that the odds of having a middle socioeconomic status compared to high socioeconomic status for female is multiplied by 0.9547907 compared to females, holding all other variables constant.

exp(-0.07403496) indicates that the odds of having a low socioeconomic status compared to high socioeconomic status is multiplied by -0.07403496 for every 1 unit increase in reading score.

95% Profile-Likelihood Confidence Intervals:
Since the factor(honor)yes for the low socioeconomic class, and read, factor(schtyp)public, factor(gender)male for middle socioeconomic class contains 0 in their intervals, they do not hold significant effect towards the probability of identifying subjects in their correct ses. However, for all the interaction terms should still be included in the model since not all are insignificant.

We will split the data in order to conduct further diagnostic, since we will then be able to use the logistic regression diagnostic on each.

Hosmer and Lemeshow goodness-of-fit test:
Low vs. High
The null hypothesis is: H_0 : Our model is predicting the data well vs. H_A : Our model is not predicting the data well. The value of the test-statistic is: 3.0269; The value of the p-value is: 0.9327. The p-value is large, so we fail to reject H_0 and conclude that our model is fitting the data well.
Middle vs. High
The null hypothesis is: H_0 : Our model is predicting the data well vs. H_A : Our model is not predicting the data well. The value of the test-statistic is: 8.8615; The value of the p-value is: 0.3541. The p-value is large, so we fail to reject H_0 and conclude that our model is fitting the data well.

The Pearson Standardized Residuals for Low vs. High is normally distributed. The Pearson Standardized Residuals for Middle vs. High are slightly weighted on the positive side but the distribution is roughly normal. Therefore, we can assume normality and continue the diagnostic as a log linear models.

Since both models do not have standard residuals or deviance greater than 3, we can say that the model fits well.

AUC:

The AUC value for both LvsH and MvsH are relatively high and the confidence intervals are all above 50%. Therefore, it indicates that the predictive power is good (although since the value of AUC and confidence interval for MvsH are low, towards 50%, we cannot say that the predictive power is powerful).

Error Matrix:

The percentage of total error is 45%. Since the error rate is below 50% the model predicts relatively well.

Prediction:

The probability and the ses for a female, who scored 60 on all her tests, who went to private school, and who was not in the honors program is middle with 66.126495%.

Conclusion:

The conclusion reached from the above analysis and interpretation is that the best model selected by the AIC backward stepwise selection fits well. Although the prediction power is not too powerful and the error rate is higher than we would prefer, it is still better than a probability of a coin flip. Therefore, the model, ses ~ read + honors + schtyp + gender performs adequately in predicting the subject’s socioeconomic status.

PROBLEM 2 - Low Birth Rate or Not (Logistic Regression)

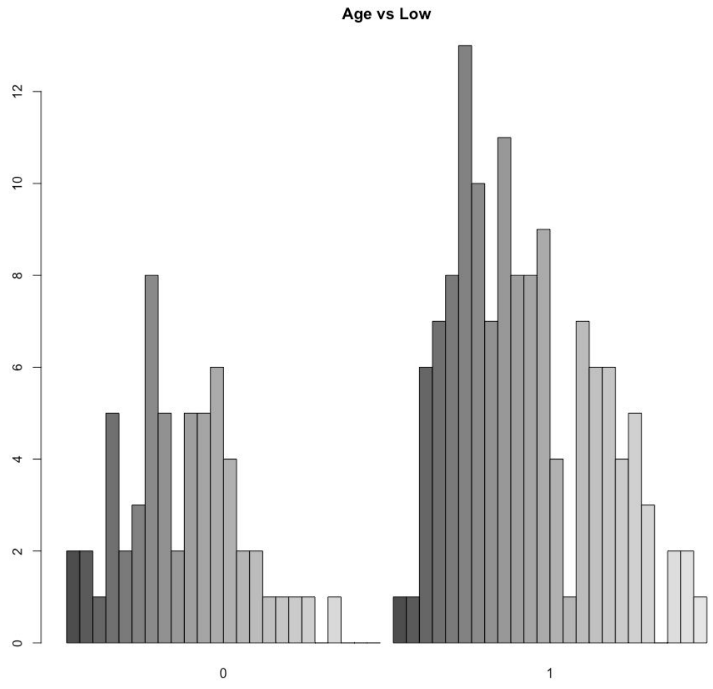
Introduction

We are given information on the age, weight, smoking status, history of premature labor, history of hypertension, and number of hospital visits for pregnant women. We would like to use the explanatory variables we are given to predict how likely it is for a pregnant woman to have a child of low birth weight. This would be interesting to know because we might want to see how the characteristics or habits of a mother can affect the likelihood of bearing a low birth weight child. Physicians might want to know which variables contribute to this occurrence so that they can provide recommendations or information to pregnant mothers to prevent having a low birth weight child.

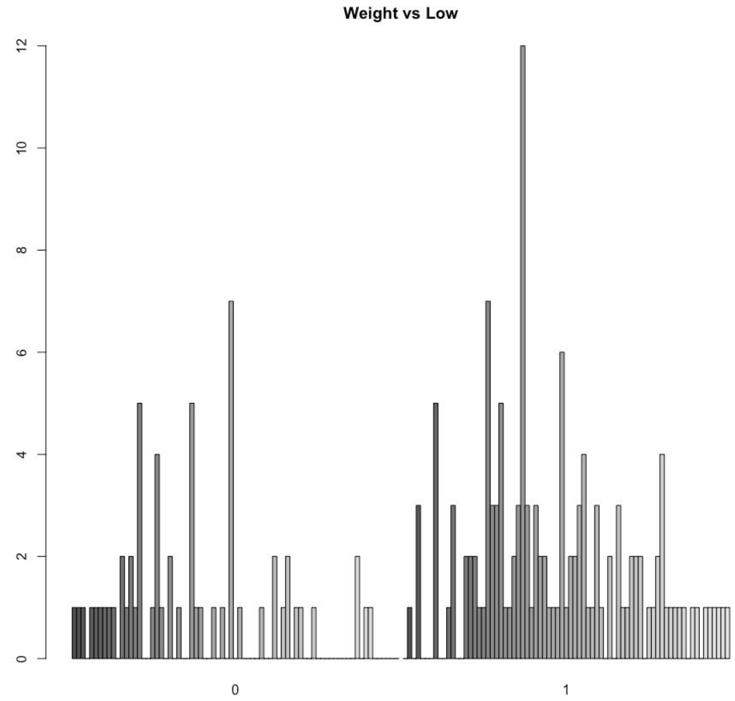
Summary

Since Age and Weight are continuous variables, we will plot them against our low-birth-weight variable to get an idea of how age and low birth rate are related, and how weight and low birth weight are related.

It seems that there are more occurrences of low birth rate when age is high.

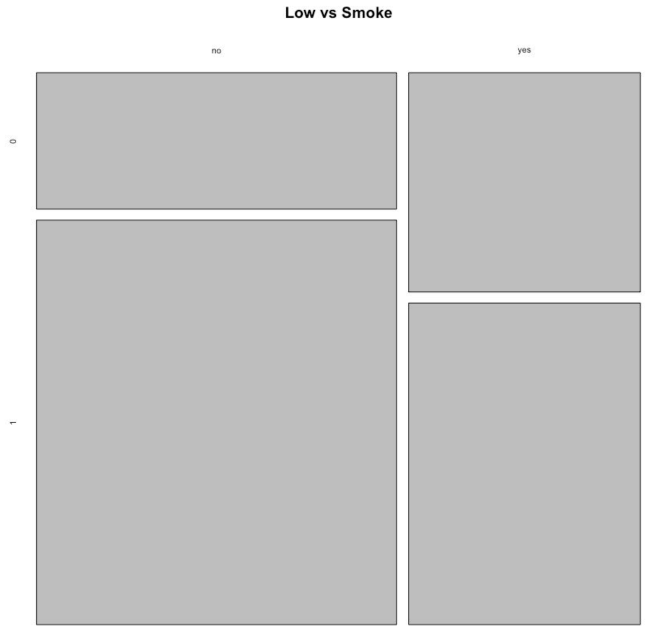


It seems that there are more occurrences of low birth weight when weight of the female is higher.



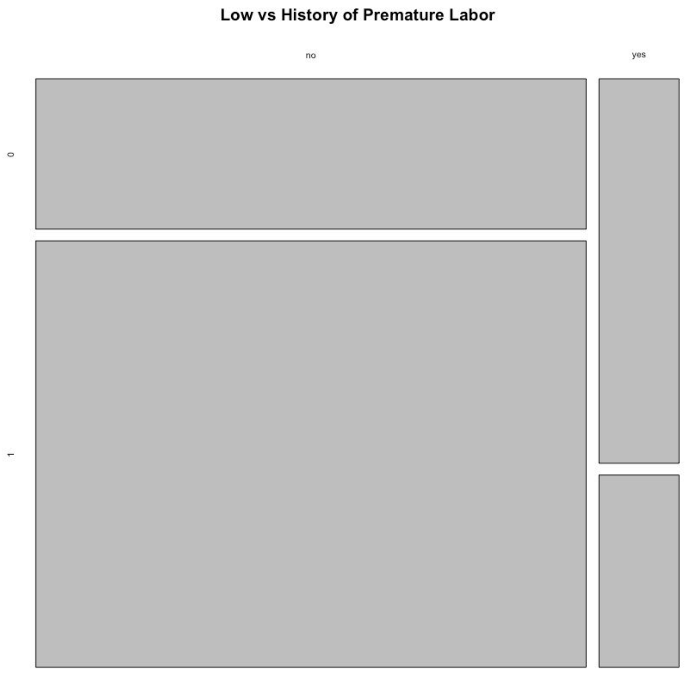
Since the other explanatory variables are categorical, we will use mosaic plots to see how each variable is related to the probability of low birth weight.

The mosaic plot of Low birth rate vs Smoking is below:



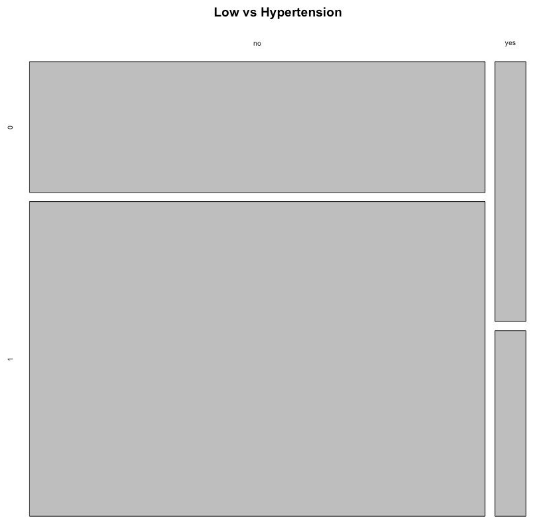
It seems that those who don’t smoke have higher instances of low birth weight.

Low vs History of Premature Labor:



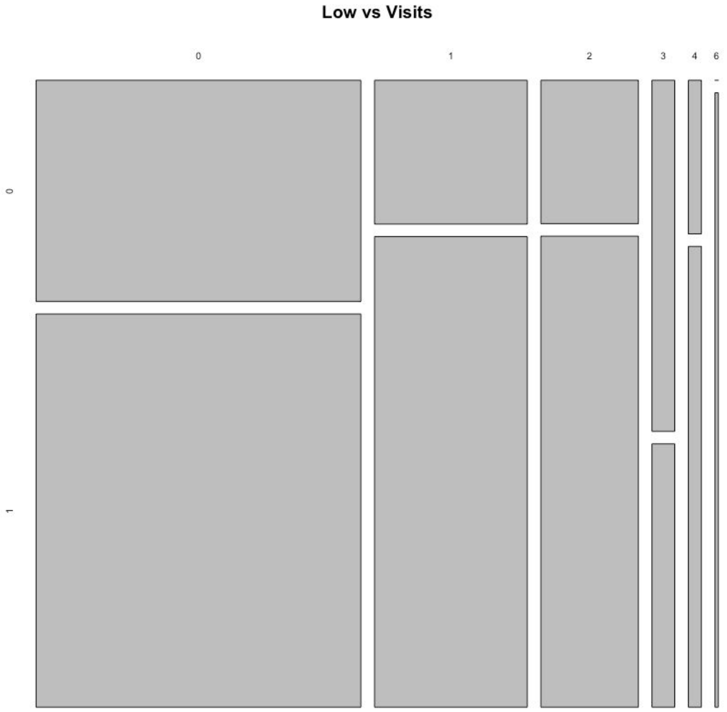
Those who don’t have a history of premature labor seem to have higher instances of low birth weight.

Low vs Hypertension:



Those with no hypertension seem to have higher instances of low birth weight.

Low vs Visits:



Those with the least number of visits have a higher probability of low birth rate.

Analysis and Interpretation

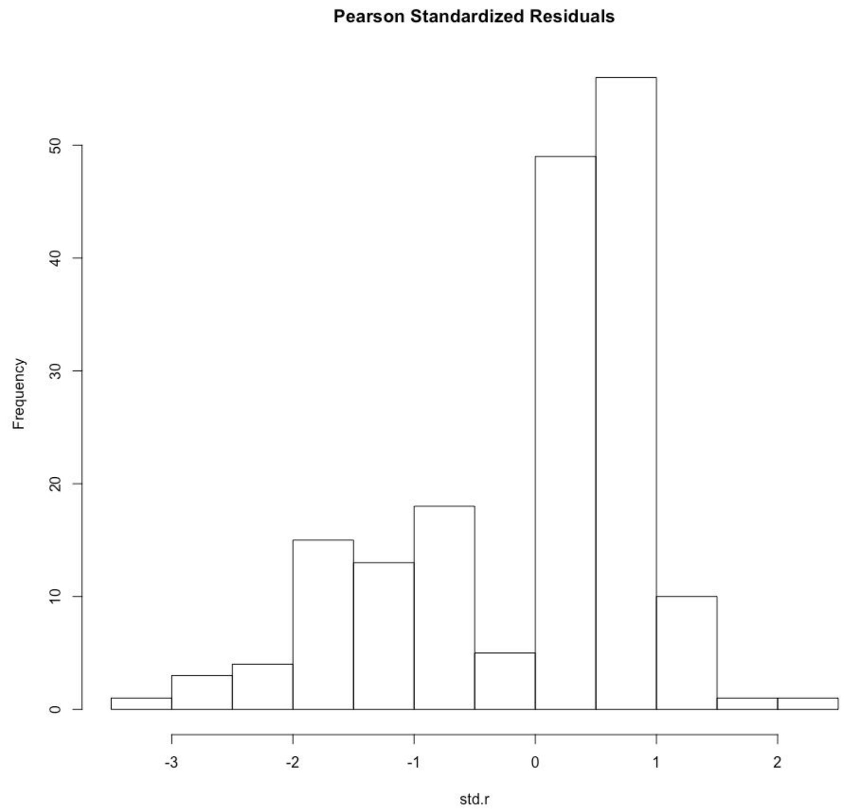
We want to know which explanatory variables contribute to having a low birth weight child, so we will use hypothesis testing and confidence intervals in R to test whether any of the betas = 0. The estimates that are fairly high relative to the others are SmokeYes, PreYes, and HypYes (-.5137, -1.799, -1.773). This suggests that these variables may have a significant impact on P(Y=1). For example, the odds of low birth weight is multiplied by e^(-.5137) for those who smoke compared to those who don't. We look at the Wald confidence interval to see which intervals contain zero, and which do not. We find that PreYes is the only interval that does not contain 0, so we conclude that this variable is the only significant variable in this equation. When we look at the Profile Log-Likelihood confidence intervals for the betas, we see that Weight, PreYes, and HypYes are the intervals which don't contain zero. We conclude that these betas have an impact on the odds of a low birth weight child.

We want to predict the odds of having a low birth weight child for a woman who is 157 lbs, age 29, doesn't smoke, does not have premature labor, does not have hypertension, and has 10 hospital visits. The prediction model in R gives us .9268. The odds of this occurrence is .9268, which is very high (> .5).

Now we want to do model selection, to determine which variables we want in our model to make this prediction. Since this is a medical study, it is preferred that more explanatory variables be left in the model in order to ensure accuracy. The more explanatory variables we have in our model, the more information we have to contribute to our prediction. Although some variables may not have a pronounced effect on the log-odds, the fact that they contribute a little bit to the log-odds makes it preferable to keep them. We will choose AIC as our criterion, because we may end up with such a large model with many predictors. AIC does not penalize large models as much as BIC does, so we will choose AIC. After doing backwards selection (we want to start with the largest model and remove predictor variables one by one until AIC is minimized) and using AIC criterion, we get that the "best model" is low ~ age + weight + smoke + pre + hyp. This means that we keep the variables age, weight, smoke, pre, and hyp as explanatory variables that have a significant contribution to the log-odds of having a low birth weight child.

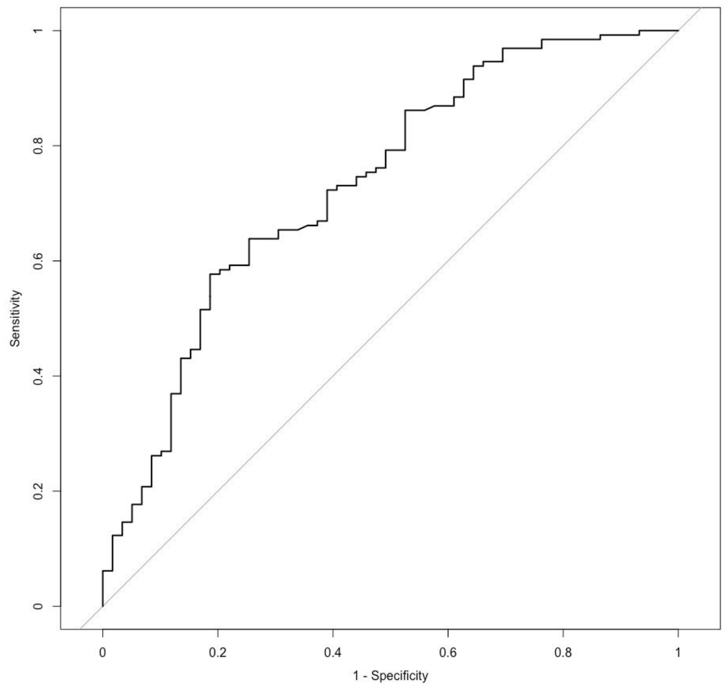
In order to test for goodness of fit for our model, we will use the Hosmer and Lemeshow good of fit test in R. Our null hypothesis is that the model fits well. Our alternative hypothesis is that the model does not fit well. We get a chi-squared value of 4.425, with a corresponding p-value of .8169. Since our p-value is greater than alpha (at all levels of alpha), we retain our null hypothesis that our "best model" fits well.

To conduct further diagnostics on our model, we will plot the standardized residuals for our model. The histogram is below:



Although the histogram is not perfectly bell shaped (which we would expect it to be because the residuals should be normally distributed with mean zero and standard deviation 1 when our null hypothesis holds), it is not so skewed so that there are many points lying far below -3 or far above 3. Therefore, we conclude that the model fits our data fairly well.

Now, we want to measure the predictive power of our model. We will calculate correlation in R to find the correlation between yi and π i. We get that our correlation is .4138. This is not exactly a high value because it not extremely close to 1 or -1. We conclude that our model does not have a very high predictive power. Another predictive power measure is Proportional Reduction in Squared Error, where we use ybar to estimate πi. In R, we get .1712. This is also not an extremely high value, since it is closer to zero than it is to 1, so we conclude again that this model does not have very high predictive power. Lastly, we will use AUC measure how sensitivity and specificity change with πo. From R, the AUC is .743, and the 95% confidence interval for the AUC is (.6554, .8127). This confidence interval is fairly close to 1, so we conclude that by using this measure, the predictive power of our model is high. As we can see from the plot below, the area under the curve is above the straight line, which means it is higher than .5.



Based on our error matrix, the sensitivity is 120/158 = .7595. The specificity is 21/31 = .6774. These are both over .5 which means that our model predicts accurately over half of the time. The percentage of total error is (10/31) + (38/158) = .3226 + .2405 = .5631, which is higher than .5. The model predicts wrongly over half of the time.

Conclusion

We used model selection to determine which explanatory variables to keep in our model in order to accurately predict the odds of having a low birth weight child. We used backwards selection and AIC in order to ensure our model was accurate for the purpose we were using it for (prediction for a medical study). Then, we conducted diagnostics to see how well our model performed in terms of predicting the odds accurately.

Code Appendix:

Part 1:

```
student = read.csv("/Users/rurikoimai/Downloads/student.csv")
head(student)
my.table1 = table(student$gender, student$ses)
my.table2 = table(student$schtyp, student$ses)
my.table3 = table(student$honors, student$ses)

barplot(my.table1, main = "SES by Gender", beside = TRUE, legend = rownames(my.table1))
barplot(my.table2, main = "SES by School Type", beside = TRUE, legend = rownames(my.table2))
barplot(my.table3, main = "SES by Enrolled in Honors Class or not", beside = TRUE, legend =
rownames(my.table3))
student_cat = table(student$gender,student$schtyp,student$honors,student$ses)
mosaicplot(student_cat, main = "Gender, School Type, Honors, Socieconomic Status")
my.table4 = table(student$ses, student$read)
my.table5 = table(student$ses, student$write)
my.table6 = table(student$ses, student$math)
my.table7 = table(student$ses, student$science)

barplot(my.table4, main = "SES by Reading Scores", beside = FALSE, legend = rownames(my.table4))
barplot(my.table5, main = "SES by Writing Scores", beside = FALSE, legend = rownames(my.table5))
barplot(my.table6, main = "SES by Math Scores", beside = FALSE, legend = rownames(my.table6))
barplot(my.table7, main = "SES by Science Scores", beside = FALSE, legend = rownames(my.table7))
#install.packages("nnet")
#install.packages("foreign")
library(foreign) #To read in the data
library(nnet)
the.model = multinom(ses ~ factor(gender) + factor(schtyp) + factor(honors) + read + write + math +
science, data = student, trace = FALSE)
null.model = multinom(ses ~ 1, data = student, trace = FALSE )
summary(the.model, digits = 4)
GOF.Multi = function(the.model){
  num.para = the.model$edf
  n = nrow(the.model$residuals)
  LL = logLik(the.model)
  df.model = n - num.para
  AIC = -2*logLik(the.model) +2*the.model$edf
  BIC = -2*logLik(the.model) +log(n)*the.model$edf
  the.results = c(LL,num.para,df.model,AIC,BIC)
  names(the.results) = c("LL","K","D.F.","AIC","BIC")
  return(the.results)
}
GOF.Multi(the.model)
the.model = multinom(ses ~ factor(gender) + factor(schtyp) + factor(honors) + read + write + math +
science, data = student, trace = FALSE)
null.model = multinom(ses ~ 1, data = student, trace = FALSE)
full.model = multinom(ses ~ factor(gender) + factor(schtyp) + factor(honors) + read + write + math +
science, data = student, trace = FALSE)
forward.model = step(null.model, scope = list(lower = null.model, upper = the.model), direction =
"forward", trace = FALSE)
forward.model
backward.model = step(null.model, scope = list(lower = null.model, upper = the.model), direction =
"backward", trace = FALSE)
FB.model = step(null.model, scope = list(lower = null.model, upper = the.model), direction = "both",
trace = FALSE)
BF.model = step(full.model, scope = list(lower = null.model, upper = the.model), direction = "both", trace
= FALSE)
best.model = forward.model
best.model
confint(best.model)
split.data = split(student, student$ses)
names(split.data)
LvsH = rbind(split.data[[2]], split.data[[1]]) #low vs. high
MvsH = rbind(split.data[[3]], split.data[[1]]) #middle vs. high
unique(LvsH$ses)
LvsH$ses = ifelse(LvsH$ses == "low",1,0)
MvsH$ses = ifelse(MvsH$ses == "middle",1,0)
model.LvsH = glm(ses ~ factor(gender) + factor(schtyp) + factor(honors) + read, data = LvsH, family =
binomial(logit))
model.LvsH
model.MvsH = glm(ses ~ factor(gender) + factor(schtyp) + factor(honors) + read, data = MvsH, family =
binomial(logit))
library(ResourceSelection)
HL.test1 = hoslem.test(model.LvsH$y, model.LvsH$fitted.values, g = 10)
HL.test1
HL.test2 = hoslem.test(model.MvsH$y, model.MvsH$fitted.values, g = 10)
HL.test2
library(LogisticDx)
good.stuff1 = dx(model.LvsH)
pear.r1 = good.stuff1$Pr #Pearsons Residuals
deviance.r1 = good.stuff1$dr #Deviance Residuals
std.r1 = good.stuff1$sPr #Standardized residuals (Pearson)
df.beta1 = good.stuff1$dBhat #DF Beta for removing each observation
change.pearson1 = good.stuff1$dChisq #Change in pearson X^2 for each observation
change.LR1 = good.stuff1$dDev #Change in LR-test G^2 for each observation
good.stuff2 = dx(model.MvsH)
pear.r2 = good.stuff2$Pr #Pearsons Residuals
deviance.r2 = good.stuff2$dr #Deviance Residuals
std.r2 = good.stuff2$sPr #Standardized residuals (Pearson)
df.beta2 = good.stuff2$dBhat #DF Beta for removing each observation
change.pearson2 = good.stuff2$dChisq #Change in pearson X^2 for each observation
change.LR2 = good.stuff2$dDev #Change in LR-test G^2 for each observation
hist(std.r1, main = "Pearson Standardized Residuals for Low vs. High")
hist(std.r2, main = "Pearson Standardized Residuals for Middle vs. High")
good.stuff1[deviance.r1 > 3,]
good.stuff1[std.r1 > 3,]
good.stuff2[deviance.r2 > 3,]
good.stuff2[std.r2 > 3,]
r1 = cor(model.LvsH$y,model.LvsH$fitted.values)
r1
prop.red1 = 1- sum((model.LvsH$y -model.LvsH$fitted.values)^2)/sum((model.LvsH$y -
mean(model.LvsH$y))^2)
prop.red1
r2 = cor(model.MvsH$y,model.MvsH$fitted.values)
r2
prop.red2 = 1- sum((model.MvsH$y -model.MvsH$fitted.values)^2)/sum((model.MvsH$y -
mean(model.MvsH$y))^2)
prop.red2
library(pROC)
the.roc1 = roc(model.LvsH$y, model.LvsH$fitted.values,auc = TRUE, ci = TRUE,plot=TRUE,
legacy.axes = TRUE, main = "Low vs. High")
auc(the.roc1)
ci(the.roc1)
the.roc2 = roc(model.MvsH$y, model.MvsH$fitted.values,auc = TRUE, ci = TRUE,plot=TRUE,
legacy.axes = TRUE, main = "Middle vs. High")
auc(the.roc2)
ci(the.roc2)
pi0 =0.50
my.table1 = table(truth = model.LvsH$y,predict = ifelse(fitted(model.LvsH)>pi0,1,0))
my.table1
pi0 =0.50
my.table2 = table(truth = model.MvsH$y,predict = ifelse(fitted(model.MvsH)>pi0,1,0))
my.table2
fitted(model.MvsH)
error =table(actual =student$ses,preds = predict(best.model)) #Error matrix
error
error.rate = 1 - sum(diag(error))/sum(error)
error.rate
the.model = multinom(ses ~ factor(gender) + factor(schtyp) + factor(honors) + read + write + math + science, data = student, trace = FALSE)
summary(the.model, digits = 4)
best.model = forward.model
best.model
predict(best.model, newdata = data.frame(gender = "female", schtyp = "private", honors="no", read = 60))
```

<u>Part 2 Appendix:</u>

```
Infant=read.csv("baby.csv",header=TRUE)
Infant

# Plots
head(Infant)
# Continuous
table1=table(Infant$age, Infant$low)
table2=table(Infant$weight, Infant$low)

table1
table2

barplot(table1, main = "Age vs Low", beside =TRUE)
barplot(table2,main="Weight vs Low",beside=TRUE)

table3=table(Infant$smoke,Infant$low)
table4=table(Infant$pre,Infant$low)
table5=table(Infant$hyp,Infant$low)
table6=table(Infant$visits,Infant$low)

mosaicplot(table3,main="Low vs Smoke")
mosaicplot(table4,main="Low vs History of Premature Labor")
mosaicplot(table5,main="Low vs Hypertension")
mosaicplot(table6,main="Low vs Visits")

# Categorical


table(Infant$low, Infant$smoke)
```

```
mosaicplot(table)
#Summary
big.logit = glm(low~age+weight+smoke+pre+hyp+visits, data = Infant, family = binomial)
big.logit

summary(big.logit)

estimates =  summary(big.logit)$coefficients[,1] # A vector of only the estimates
estimates
e=exp(estimates)
e
SE =  summary(big.logit)$coefficients[,2] #A vector of only the Wald SE's
alpha = 0.01
z.a.2 = qnorm(1-alpha/2)
upper.bounds = estimates +z.a.2*SE
lower.bounds = estimates -z.a.2*SE
Wald.CI = cbind(lower.bounds,upper.bounds)
Wald.CI
ew=exp(Wald.CI)
ew
#Profile Likelihood CI
confint(big.logit)

# Prediction

predict(big.logit, newdata =
data.frame(weight=157,age=29,smoke="no",pre="no",hyp="no",visits=10),type = "response")

# Model Selection

full.model = glm(low ~. , data = Infant,family = binomial(link=logit))
full.model

#AIC and BIC of full model.

full.AIC = AIC(full.model)
full.BIC = BIC(full.model)
c(full.AIC, full.BIC)
```

```
# Forward Stepwise Selection
empty.model = glm(low~ 1 ,data = Infant,family = binomial(link=logit))
#Forward selection
best.forward.AIC = step(empty.model,scope = list(lower = empty.model, upper =
full.model),direction = "forward", criterion = "AIC", trace = FALSE)
best.forward.AIC
best.forward.AIC$formula
#Backward Selection
step(full.model,scope = list(lower = empty.model, upper = full.model),direction = "backward")
best.backward.AIC = step(full.model,scope = list(lower = empty.model, upper =
full.model),direction = "backward", criterion = "AIC", trace = FALSE)
best.backward.AIC$formula

best.FB.AIC = step(empty.model,scope = list(lower = empty.model, upper =
full.model),direction = "both", criterion = "AIC", trace = FALSE)
best.BF.AIC = step(full.model,scope = list(lower = empty.model, upper = full.model),direction =
"both", criterion = "AIC", trace = FALSE)
best.FB.AIC$formula
best.BF.AIC$formula
# Using BIC
best.forward.BIC = step(empty.model,scope = list(lower = empty.model, upper =
full.model),direction = "forward", k = log(nrow(small.credit)), trace = FALSE)
best.backward.BIC = step(full.model,scope = list(lower = empty.model, upper =
full.model),direction = "backward", k = log(nrow(small.credit)), trace = FALSE)
best.FB.BIC = step(empty.model,scope = list(lower = empty.model, upper =
full.model),direction = "both", k = log(nrow(small.credit)), trace = FALSE)
best.BF.BIC = step(full.model,scope = list(lower = empty.model, upper = full.model),direction =
"both", k = log(nrow(small.credit)), trace = FALSE)
best.forward.BIC$formula
best.backward.BIC$formula
best.FB.BIC$formula
best.BF.BIC$formula


library(bestglm)
best.subset.AIC = bestglm(Xy = small.credit, family = binomial(link=logit),IC = "AIC",method
= "exhaustive")
best.subset.AIC
best.subset.BIC = bestglm(Xy = small.credit, family = binomial(link=logit),IC = "BIC",method
= "exhaustive")
best.subset.BIC

# Use AIC and backward for predictive power.

best.model=best.backward.AIC
best.model
best.model$formula

# Diagnostics. HL Goodness of Fit.
library(ResourceSelection)
HL.test = hoslem.test(best.model$y, best.model$fitted.values,g = 10)
HL.test

# Residuals and Influence Measures
library(LogisticDx)
good.stuff = dx(best.model)
pear.r = good.stuff$Pr #Pearsons Residuals
pear.r
deviance.r = good.stuff$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
change.LR

hist(std.r, main = "Pearson Standardized Residuals")
cutoff.beta = 0.20
df.beta[df.beta > cutoff.beta] #Shows the values
good.stuff[df.beta > cutoff.beta,] #what observations they were
cutoff.pearson = 6
change.pearson[change.pearson > cutoff.pearson] #Shows the values
good.stuff[change.pearson > cutoff.pearson,1:3] #what observations they were

# Measures of Predictive Power
r = cor(best.model$y,best.model$fitted.values)
r
```

```
prop.red = 1- sum((best.model$y -best.model$fitted.values)^2)/sum((best.model$y -
mean(best.model$y))^2)
prop.red

# AUC , ROC , classification tables.
library(pROC)
the.roc = roc(best.model$y, best.model$fitted.values,auc = TRUE, ci = TRUE,plot=TRUE,
legacy.axes = TRUE)
auc(the.roc)
ci(the.roc)

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