# STA138 Final Project

***Lam Vu***

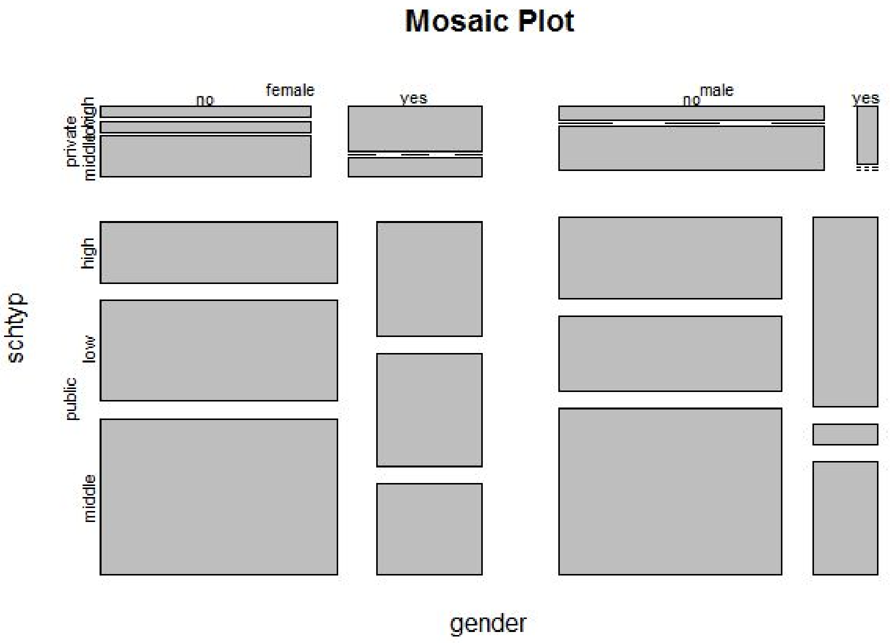
***March 20, 2017***

# Problem 1: Socioeconomic Class

# Introduction:

The goal in this analysis of the student dataset is to determine the relationship between the variables ses (Socioeconomic class), gender(male=1, female=0), schtype(school type with values public=1 and private=0), honors ( Whether or not the student is enrolled in an honors class, with values yes=1 and no=0), scores on the test with subjects read,write,math, and science. We perform data summary, model selection and identification of outliers and influential points, predictive power assessment and goodness of fit.

# Summary:



From the mosaic plot we can see the proportional distribution of the students.We notice there are more students who go to public schools than private schools. Most student are also not in the honors program. The dataset seems be approximate even proportion of female and male. Most student who go to a public school are middle class. Female students who go to a public school that are in honors program are evenly proportioned between economic classes. In comparison to male students in public school where most students in honors program are from high class. There is also less male students compared to female students in honors program overall. For further analysis I will fit multinomial logistic regression and perform diagnostics on the model to find an appropriate model to describe the data.

# Analysis:

### Best forward

```
## multinom(formula = ses ~ read + honors + schtyp + gender, data = student,
##     trace = FALSE)
```

AIC: 403.7529

---

### Best Backward

```
## multinom(formula = ses ~ gender + schtyp + honors + read, data = student,
##     trace = FALSE)
```

AIC: 403.7529

---

### Best Forward Backward

```
## multinom(formula = ses ~ read + honors + schtyp + gender, data = student,
##     trace = FALSE)
```

AIC: 403.7529

---

### Best Backward Forward

```
## multinom(formula = ses ~ gender + schtyp + honors + read, data = student,
##     trace = FALSE)
```

AIC: 403.7529

---

### Log Likelihood, number of parameter, degree of freedom, AIC and BIC

```
    LL        K     D.F.      AIC      BIC
-191.8765   10.0000  190.0000  403.7529  436.7361
```

---

### Split Student Dataset

### Best Subset Model for Low vs High based on AIC

```
## Morgan-Tatar search since family is non-gaussian.

## y ~ schtyp + read + math
```

---

### Best Subset Model for Medium vs High based on AIC

```
## Morgan-Tatar search since family is non-gaussian.

## y ~ read
```

---

### Goodness of Fit Diagnostic for Low vs High model

```
##
##  Hosmer and Lemeshow goodness of fit (GOF) test
##
## data:  LvsH.best.subset.AICmod$y, LvsH.best.subset.AICmod$fitted.values
## X-squared = 5.9254, df = 8, p-value = 0.6556
```

---

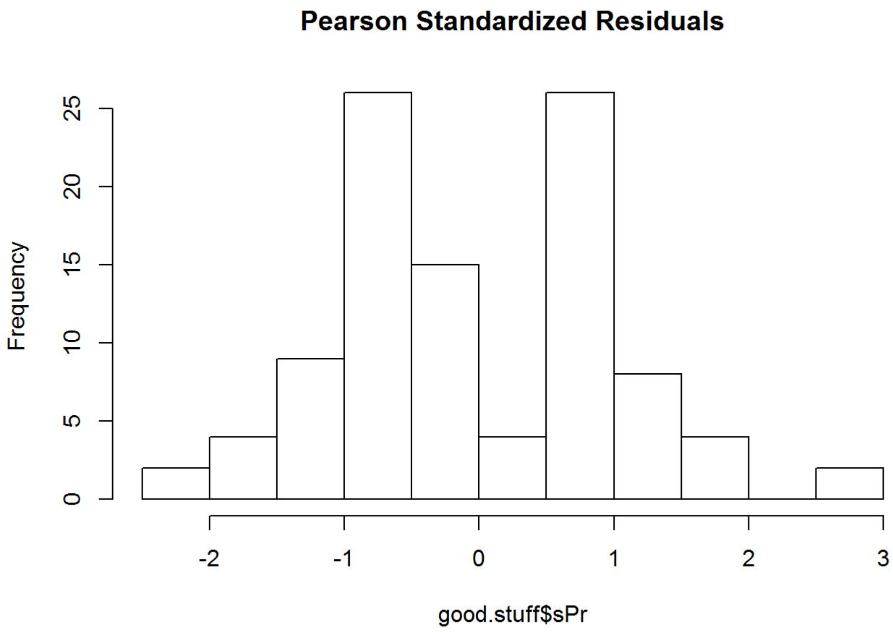
### Goodness of Fit Diagnostic for Medium vs High model

```
##
##  Hosmer and Lemeshow goodness of fit (GOF) test
##
## data:  MvsH.best.subset.AICmod$y, MvsH.best.subset.AICmod$fitted.values
## X-squared = 9.7497, df = 8, p-value = 0.283
```

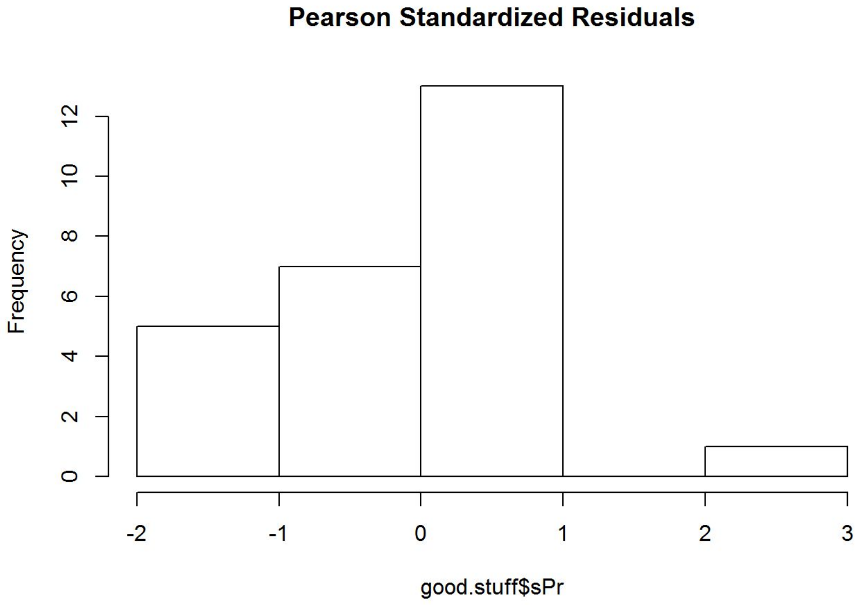
---

### Residuals and Influence Measures

Low vs High model



Medium vs High mode



**Measure of Predictive Power**

Low vs High Model Correlation

##[1] 0.4590329

proportional reduction in squared error

## [1] 0.2107065

Medium vs High Model

Correlation

##[1] 0.2402279

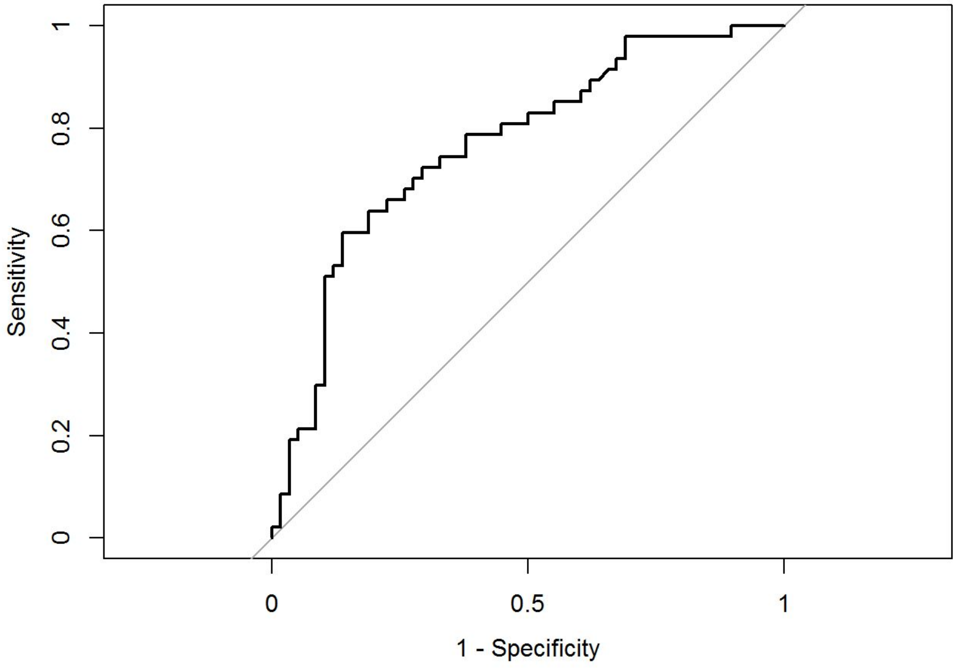
proportional reduction in squared error

## [1] 0.05767551

---

**AUC and ROC**

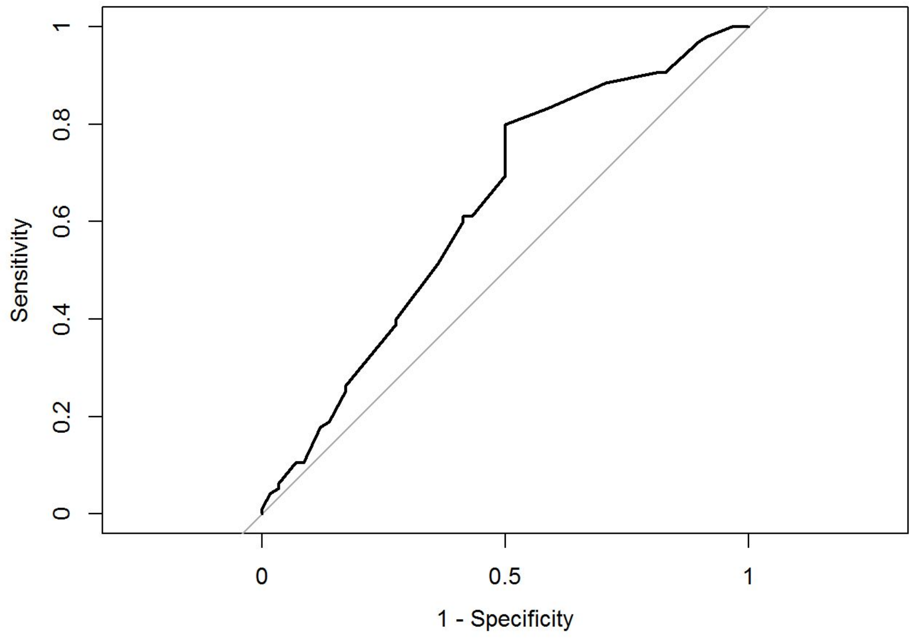
Low vs High model



## Area under the curve: 0.7654

## 95% CI: 0.6733-0.8575 (DeLong)

Medium vs High model



```
## Area under the curve: 0.6284
```

```
## 95% CI: 0.5343-0.7225 (DeLong)
```

### Error Matrix

Low vs High model

```
##      predict
## truth  0  1
##     0 44 14
##     1 16 31
```

Medium vs High model

```
##      predict
## truth  0  1
##     0 17 41
##     1 11 84
```

**Predict ses for female, who scored 60 on all her tests who went to private school and who was not in the honors program.**

Low vs High model

```
##          1 
## 0.07180827
```

Medium vs High model

```
##         1 
## 0.5500063
```

---

# Interpretation:

For model selection I used the different stepwise methods, where all the methods resulted in a model consisting of gender,schtyp,honors,and read:  y~gender+schtyp+honors+read with an AIC=403.7529 and BIC=436.7361. I chose this to be the best model for the data. I then performed diagnostics where I first split the data by ses into low and medium. Then chose the best model using an exhaustive subset selection based on AIC, the LvsH data is fitted with model  y ~ schtyp + read + math and the MvsH data is fitted with model  y ~ read. Then I performed a goodness of fit diagnostic LvsH model with a  X-squared = 5.9254 and p-value = 0.6556. The goodness of fit  for LvsH model with a  X-squared = 9.7497 and p-value = 0.283. Where both p-values do not reject the null hypothesis and conclude the current model fits well. From the histogram of the Pearson standardized residuals for both model which showed the residuals are both not normally distributed because it does  not follow a bell curve. The LvsH model appears bimodally distributed as well as having outliers. The MvsH model also has outliers. LvsH has a correlation of 0.4590329 and MvsH has a correlation of  0.2402279 which meant they are slightly positively correlated. MvsH has a low proportional reduction in squared error of 0.05767551 which meant no association and LvsH has 0.2107065 which mean low association between independent variable and prediction of dependent variable. LvsH has AUC=0.7654 and MvsH has AUC=0.6284 which are LvsH is relatively high suggesting the model does fit the data well and MvsH is considered low which suggest the model does not fit the data well.  From the error matrix, LvsH has a better prediction rate and MvsH has greater error in prediction rate. The percentage of total error for LvsH is 0.2857143 and for MvsH is 0.3398693. We then use the models to predict 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. LvsH has probability of 0.07180827 and MvsH has a probability of  0.5500063, this meant the probability of this student being from low is 7.18% and the probability of this student being from medium's 55%.

# Conclusion:

From this analysis I chose the best fitting model for the multinomial data to be y=gender+schtyp+honors+read based on the AIC=403.7529 from using all the step model selection methods. I then performed diagnostics where I first split the data by ses into low and medium. Then chose the best model using an exhaustive subset selection based on AIC, the LvsH data is fitted with model  y ~ schtyp + read + math and the MvsH data is fitted with model y ~ read. From the goodness of fit diagnostic I concluded both models fit well. From there Pearson standardized residual I conclude they are both not normal and had outliers. Both model had slightly positively correlation. MvsH had no association and LvsH had low association between independent variable and prediction of dependent variable based on the proportional reduction in squared error. From the AUC, LvsH is relatively high suggesting the model does fit the data well and MvsH is considered low which suggest the model does not fit the data well. From the error matrix LvsH has a better prediction rate and MvsH has greater error in prediction rate.

# STA138 Final Project

*Lam Vu*

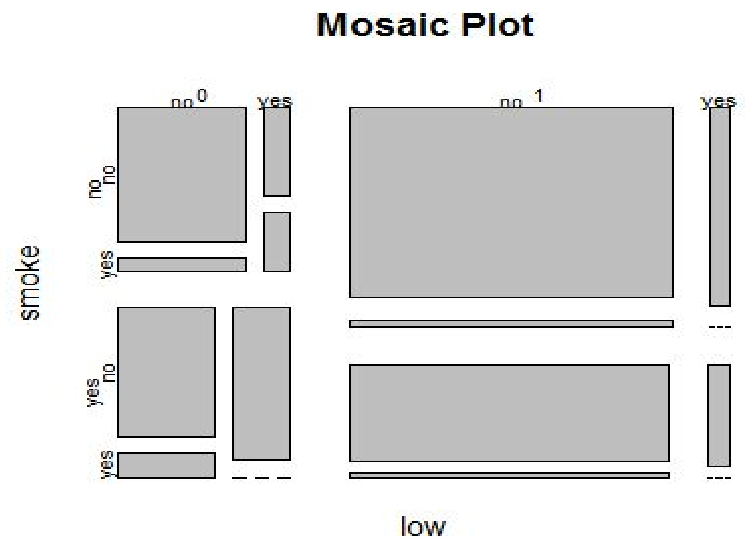
*March 21, 2017*

# Problem 2: Low birth rate or not

# Introduction:

The goal in this analysis of the baby dataset is to determine the relationship between the variables birth (low birth weight=1 and not low birth weight=0), age, weight(before pregnancy), smoke( smoker =1 and nonsmoker=0), pre(pre-mature labor=1, no pre-maturer labor=0),  hyp (hypertension=1, no hypertension=0), and visits. I will perform data summary, model selection and identification of outliers and influential points, predictive power assessment and goodness of fit test.

# Summary:



From the mosaic plot we can see the proportional distribution of the rate of low birth weight of babies. We notice more babies are not low birth weight. There does not seem to be a difference in proportion of babies from smokers to nonsmokers in low birth weight. But smokers seem to be more likely to have hypertension than non-smokers. More babies are low birth weight had mothers with premature labors. There does not seem to be a difference in hypertension levels to low birth weight. For further analysis I will fit a logistic regression and perform diagnostics on the model to find an appropriate model to describe the data.

# Analysis:

**Full Model**

```
##
## Call:  glm(formula = low ~ age + weight + smoke + pre + hyp + visits,
##     family = binomial(logit), data = baby)
##
## Coefficients:
## (Intercept)         age       weight     smokeyes       preyes
##    -2.02149     0.05909      0.01609     -0.51374     -1.79891
##      hypyes      visits
##    -1.77264     0.03211
```

```
##
## Degrees of Freedom: 188 Total (i.e. Null);  182 Residual
## Null Deviance:       234.7
## Residual Deviance: 202.2     AIC: 216.2
```

---

### Best Subset Model based on AIC

```
## Morgan-Tatar search since family is non-gaussian.
```

```
##
## Call:  glm(formula = low ~ age + weight + smoke + pre + hyp, family = binomial(link = logit),
##     data = Xy)
##
## Coefficients:
## (Intercept)          age       weight        smoke          pre
##    -2.03197      0.06032      0.01615     -0.51837     -1.79404
##         hyp
##    -1.78271
##
## Degrees of Freedom: 188 Total (i.e. Null);  183 Residual
## Null Deviance:       234.7
## Residual Deviance: 202.2     AIC: 214.2
```

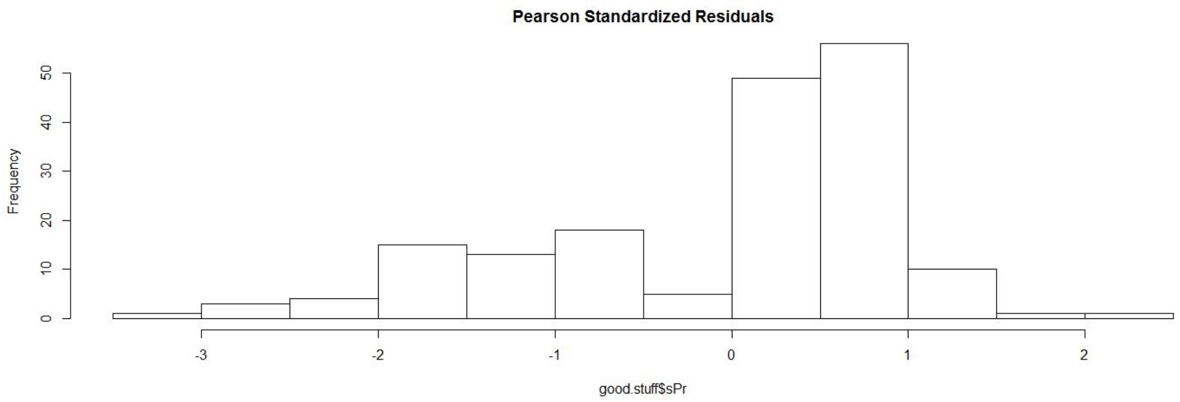
---

### Goodness of Fit

```
##
##  Hosmer and Lemeshow goodness of fit (GOF) test
##
## data:  best.subset.AICmod$y, best.subset.AICmod$fitted.values
## X-squared = 4.425, df = 8, p-value = 0.8169
```

---

### Residuals and Influence Measures



## Residuals and Influence Measures

## Correlation

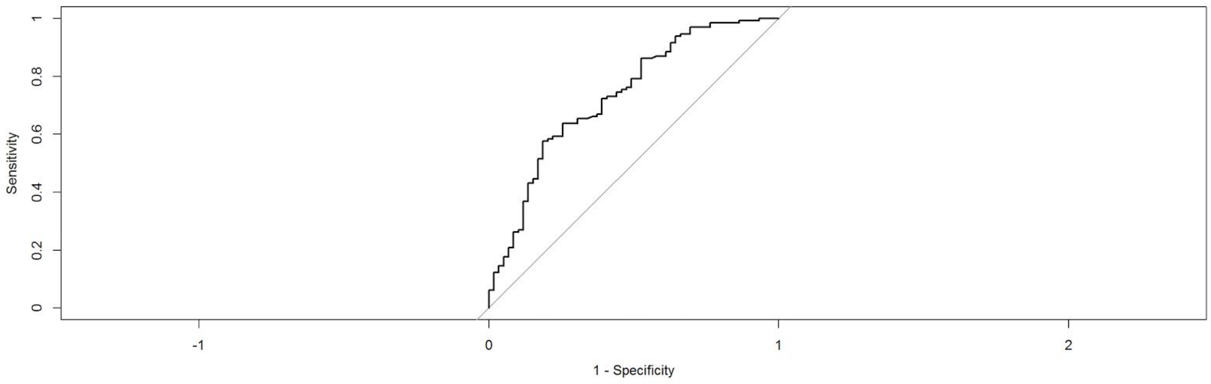
```
## [1] 0.4137879
```

## proportional reduction in squared error

```
## [1] 0.171218
```

---

## AUC and ROC Low vs High model



```
## Area under the curve: 0.734
```

```
## 95% CI: 0.6554-0.8127 (DeLong)
```

---

**Error Matrix Low vs High model**

```
##      predict
## truth   0   1
##     0  21  38
##     1  10 120
```

---

**Predict if a woman weighing 157 pounds, age 29, with no history of smoking, premature labor, or hypertension, and 10 visits will have a low birth weight child.**

```
##         1
## 0.9049528
```

# Interpretation:

For model selection I used the subset model based on AIC, which resulted in a model consisting of age, weight, smoke, pre, and hyp: low ~ age + weight + smoke + pre + hyp with an AIC= 214.2. I then performed diagnostics where I used a Goodness of Fit Test which resulted in a p-value= 0.8169 this meant to not reject the null hypothesis and conclude the current model fits well. I then measured the residuals and influential points, using correlation and proportional reduction in squared error. This resulted in a correlation coefficient of 0.4137879 and a proportional reduction in squared error of 0.171218. The correlation meant there is slight positive correlation between the explanatory variable and the response. The low proportional reduction in squared error meant there is no association between independent variable and prediction of dependent variable. I obtained an ROC which resulted in an AUC= 0.734 with a 95% confidence interval between 0.6554 to 0.8127, which is relatively high suggesting the model does fit the data well. I then obtained an error matrix with a percentage of total error = 0.2539 which is considered high.
I then use the model to predict the probability for a woman weighing 157 pounds, age 29, with no history of smoking, premature labor, or hypertension, and 10 visits will have a low birth weight child which resulted in 0.9049528 or 90.49%.

# Conclusion:

From this analysis I chose the best fitting model the the baby dataset to be low ~ age + weight + smoke + pre + hyp based on the AIC= 214.2 from using the subset model selection method. I then performed diagnostics which resulted in a slight positive correlation and no association between independent variable and prediction of dependent variable. The model fits the data well according to the AUC and the percentage of total error is considered high. I then use the model to predict the probability for a woman weighing 157 pounds, age 29, with no history of smoking, premature labor, or hypertension, and 10 visits will have a low birth weight child which resulted in 0.9049528 or 90.49%.