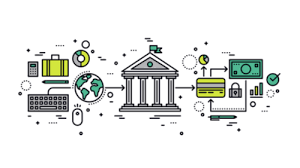
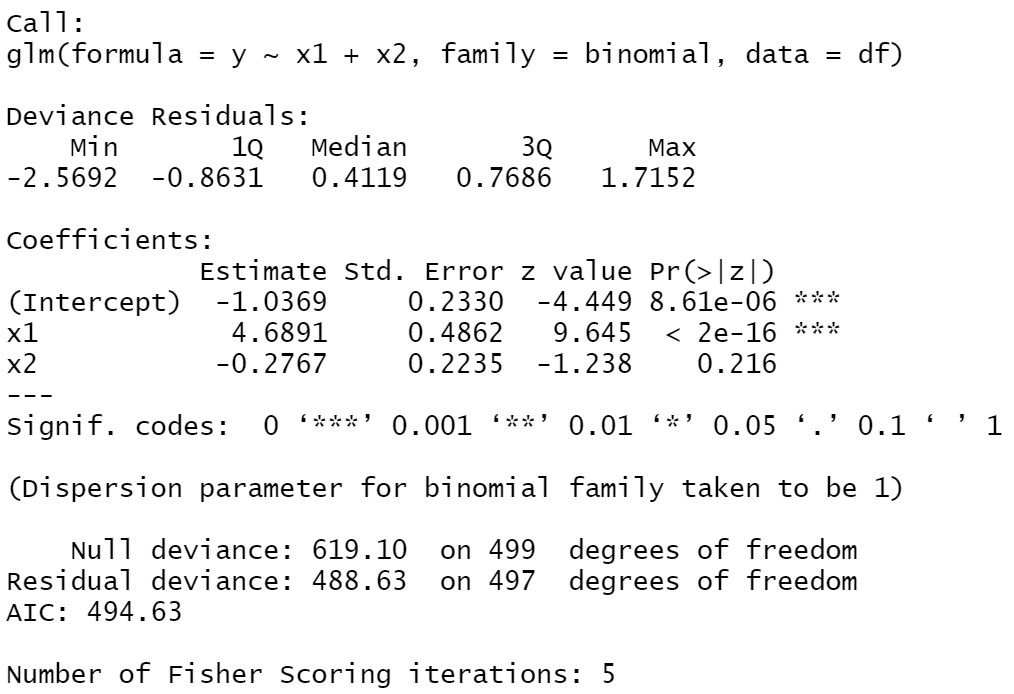
**Binomial Regression and Validation – Assignment 3**

Chris Steege and Shuki Saito

**A Simulation Study:**

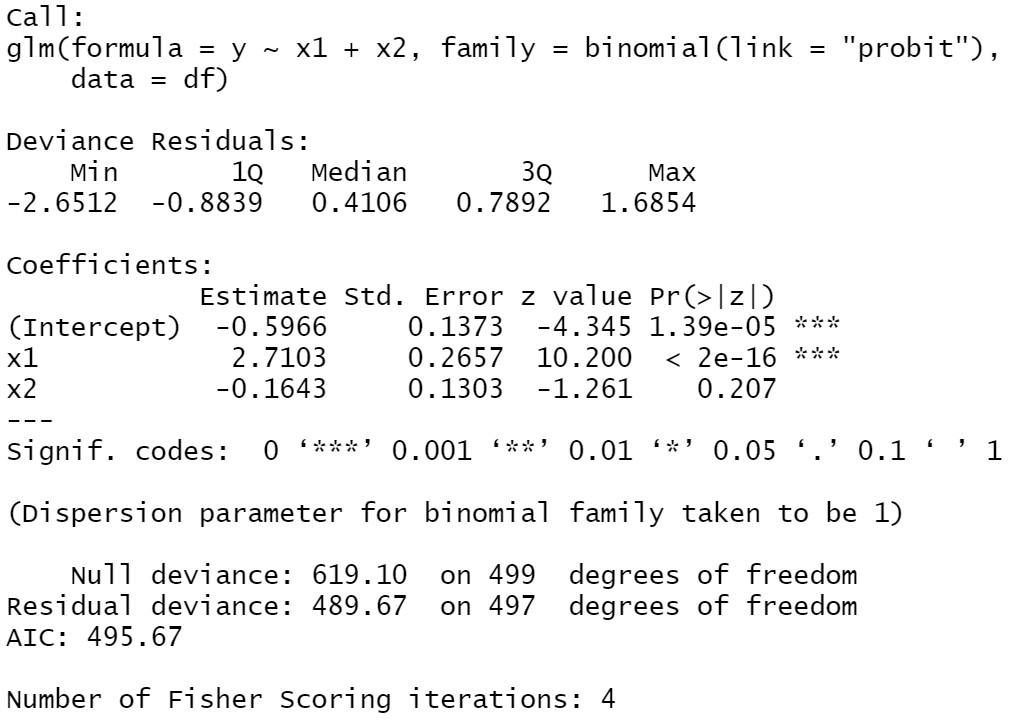
A simulation study will be conducted below using logistic and probit regression. The data was simulated according to the assumption that y|x is binary and p = E(y|x). We will use logit(pi) = -1.1+5x1i-0.4\*x2i and convert this result into a binary outcome variable. This is done by converting logit(pi) into a probability and sampling from a Bernoulli distribution to be used as response variable in building our regression models. X1 is generated using a uniform distribution between 0 and 1, while X2 is generated is 1 when odd and 0 when even by index.

1. **Logistic Regression:**

**Figure 1: Logistic Regression Summary**

1. **Probit Regression:**

**Figure 2: Probit Regression Summary**



**Summary:**

The results of the logit and probit regression models are very similar. In both cases, the intercept and X1 variable were considered significant in their ability to predict the response variable. This makes sense that X2 was not significant, because its relevance is essentially determined like a coin flip which gives it little predictive ability for the response variable.

Notably, our estimated coefficients and deviations of these coefficients were cut by about 40% in each case. This is because the logit link uses the cumulative distribution function of the logistic distribution for defining the Y estimate of our predictor variables, while the probit link uses the cumulative distribution function of the standard normal distribution.

**German Credit Scoring Data:**

We created a binary logistic regression model for German Credit data in order to find what factor/s have a significant impact on credit status. The response variable is binary, 0 or 1, where 0 indicates “good credit” and 1 indicates as “bad credit”. The variables with keys are given below:

* Status of existing checking account -> V1
* Duration in month -> V2
* Credit history -> V3
* Purpose -> V4
* Credit amount -> V5
* Savings account/bonds -> V6
* Present employment since -> V7
* Installment rate in percentage of disposable income -> V8
* Personal status and sex -> V9
* Other debtors / guarantors -> V10
* Present residence since -> V11
* Property -> V12
* Age in years -> V13
* Other installment plans -> V14
* Housing -> V15
* Number of existing credits at this bank -> V16
* Job -> V17
* Number of people being liable to provide maintenance for -> V18
* Telephone -> V19
* foreign worker -> V20

1. **Variable Selection**

In order to find best model, we conducted Variable Selection with Stepwise Approach. We will be using both directions in this variable selection.

Selection with AIC

Best Model: Credit Status = V1 + V2 + V3 + V4 + V5 + V6 + V7 + V8 + V10 + V13 + V14 + V20   
**AIC = 719.81**

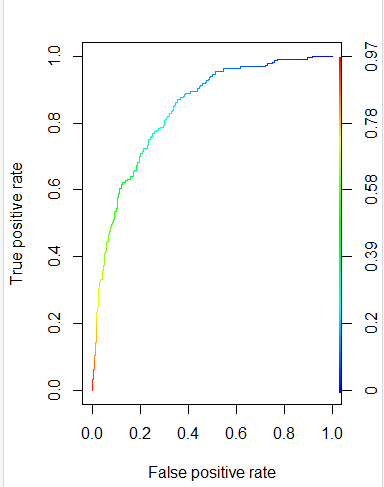
Selection with BIC

Best Model: Credit Status = Credit Status = V1 + V3 + V5 + V8   
**BIC = 789.39**

In this analysis, we will use the model: **Credit Status = V1 + V2 + V3 + V4 + V5 + V6 + V7 + V8 + V10 + V13 + V14 + V20.**

1. **In-sample analysis**

**Figure 3: ROC Curve**



The shape of this curve indicated that our predictions are much better than random because there is a large area underneath the curve. A straight line usually indicates a model is no better than a guess, therefore our model appears to be efficient.

**AUC = 0.8399**

Because AUC > 0.7, we will accept our model.

For German credit data, recommended cost ratio is 5:1, so that cut-off probability is 1/6 for the misclassification rate table. The misclassification rate table is below.

**Figure 4: Confusion Matrix**



**(16 + 213) / (281 + 190 + 16 + 213) ~=~ 0.33**

Thus, the misclassification rate is 0.33; meaning 33% of variables are possibly misclassified.

1. **Out of Sample Analysis**

Using the remaining 30% of the data, we can conduct an out-of-sample analysis.

**AUC = 0.7639**

Even though the AUC is lower compare to AUC for In-sample analysis, the value is still more than 0.7. Thus, this model is acceptable.

**AMR = 0.62**

This indicates that 62% of out of sample predictions are possibly misclassified.

**Bankruptcy Data:**

We created a binary logistic regression model using the bankruptcy data, in order to find what factors have a significant impact on credit status. The response variable is binary, 0 or 1, where 0 indicates “good credit” and 1 indicates as “bad credit”. The covariances are below.

R1=Working Capital/Total Asset;

R2=Retained Earning/Total Asset;

R3=Earning Before Interest & Tax/Total Asset;

R4=Market Capital / Total Liability;

R5=SALE/Total Asset;

R6=Total Liability/Total Asset

R7=Current Asset/Current Liability;

R8=Net Income/Total Asset;

R9=LOG(SALE);

R10=LOG(Market Cap)

1. **Variable Selection**

In order to find best model for the response value, we conducted Variable Selection with Stepwise Approach. Its direction is “both” in this variable selection.

Selection with AIC

Best Model: Credit Status = R1+R2+R3+R4+R6+R7+R8+R9+R10  
**AIC = 2158.4**

Selection with BIC

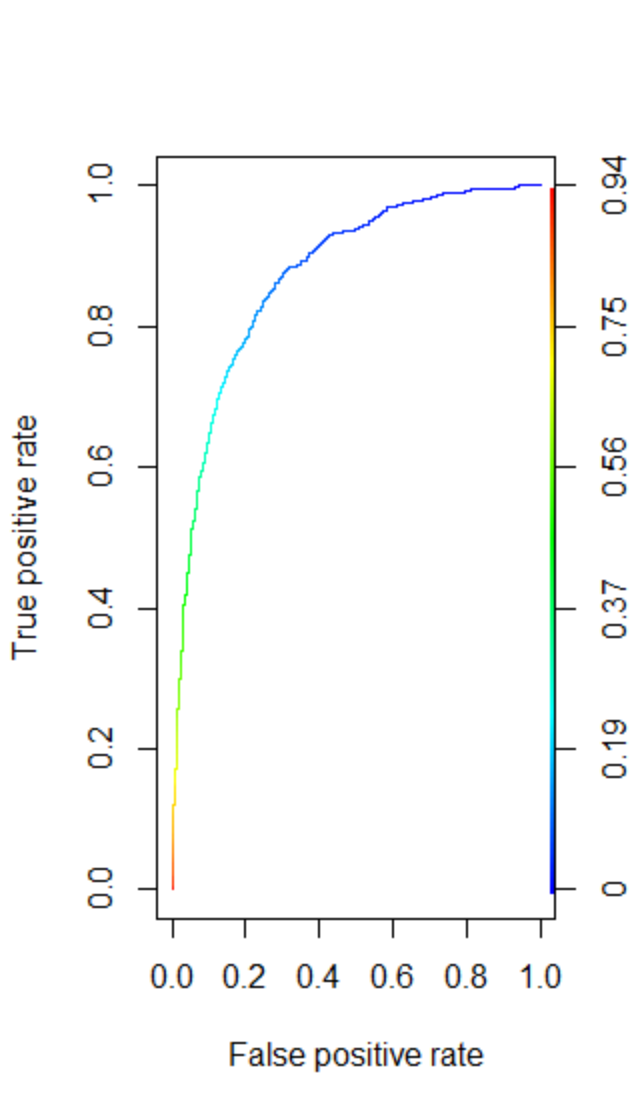
Best Model: Credit Status = R1+R2+R3+R6+R7+R8+R9+R10  
**BIC = 2216**

In this analysis, we will use the model: **Credit Status = R1+R2+R3+R4+R6+R7+R8+R9+R10.**

1. **In Sample Analysis**

For the model above, the ROC curve is shown below.

**Figure 5: ROC curve**



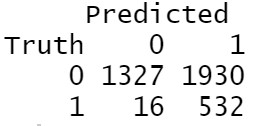
The shape of this curve indicated that our predictions are much better than random because there is a large area underneath the curve. A straight line usually indicates a model is no better than a guess, therefore our model appears to be efficient.

**AUC = 0.87**

Because AUC > 0.7, we will accept this model.

For German credit data, recommended cost ratio is 5:1, so that cut-off probability is 1/6 for the misclassification rate table. The misclassification rate table is below.

**Figure 6: Confusion Matrix**



**(16 + 1930) / (1327 + 532 + 16 + 1930) ~=~ 0.51**

Thus, misclassification rate is 0.51, meaning 51% of variables are possibly misclassified.

1. **Out of Sample Analysis**

**AUC = 0.89**

Even though the AUC is lower compare to AUC for In-sample analysis, the value is still more than 0.7. Thus, this model is acceptable.

**AMR = 0.67**

This indicates that 67% of out of sample predictions are possibly misclassified.