Data Processing Assignment

Gaelan Gu, Sunil Prakash, Wang Ruoshi, Yu Yue

### Introduction

We will be preparing the salary dataset, extracted from the 1994 US Census, for a logistic regression.

We will determine whether a person makes over 50k a year; *class* will be the dependent variable.

### Data Import

train = read.csv('salary-train.csv')  
test = read.csv('salary-test.csv')  
str(train)

## 'data.frame': 32561 obs. of 14 variables:  
## $ age : int 39 50 38 53 28 37 49 52 31 42 ...  
## $ workclass : Factor w/ 9 levels " ?"," Federal-gov",..: 8 7 5 5 5 5 5 7 5 5 ...  
## $ fnlwgt : int 77516 83311 215646 234721 338409 284582 160187 209642 45781 159449 ...  
## $ education : Factor w/ 16 levels " 10th"," 11th",..: 10 10 12 2 10 13 7 12 13 10 ...  
## $ marital : Factor w/ 7 levels " Divorced"," Married-AF-spouse",..: 5 3 1 3 3 3 4 3 5 3 ...  
## $ occupation : Factor w/ 15 levels " ?"," Adm-clerical",..: 2 5 7 7 11 5 9 5 11 5 ...  
## $ relationship : Factor w/ 6 levels " Husband"," Not-in-family",..: 2 1 2 1 6 6 2 1 2 1 ...  
## $ race : Factor w/ 5 levels " Amer-Indian-Eskimo",..: 5 5 5 3 3 5 3 5 5 5 ...  
## $ sex : Factor w/ 2 levels " Female"," Male": 2 2 2 2 1 1 1 2 1 2 ...  
## $ capital.gain : int 2174 0 0 0 0 0 0 0 14084 5178 ...  
## $ capital.loss : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ hours.per.week: int 40 13 40 40 40 40 16 45 50 40 ...  
## $ native.country: Factor w/ 42 levels " ?"," Cambodia",..: 40 40 40 40 6 40 24 40 40 40 ...  
## $ class : Factor w/ 2 levels " <=50K"," >50K": 1 1 1 1 1 1 1 2 2 2 ...

str(test)

## 'data.frame': 16281 obs. of 14 variables:  
## $ age : int 25 38 28 44 18 34 29 63 24 55 ...  
## $ workclass : Factor w/ 9 levels " ?"," Federal-gov",..: 5 5 3 5 1 5 1 7 5 5 ...  
## $ fnlwgt : int 226802 89814 336951 160323 103497 198693 227026 104626 369667 104996 ...  
## $ education : Factor w/ 16 levels " 10th"," 11th",..: 2 12 8 16 16 1 12 15 16 6 ...  
## $ marital : Factor w/ 7 levels " Divorced"," Married-AF-spouse",..: 5 3 3 3 5 5 5 3 5 3 ...  
## $ occupation : Factor w/ 15 levels " ?"," Adm-clerical",..: 8 6 12 8 1 9 1 11 9 4 ...  
## $ relationship : Factor w/ 6 levels " Husband"," Not-in-family",..: 4 1 1 1 4 2 5 1 5 1 ...  
## $ race : Factor w/ 5 levels " Amer-Indian-Eskimo",..: 3 5 5 3 5 5 3 5 5 5 ...  
## $ sex : Factor w/ 2 levels " Female"," Male": 2 2 2 2 1 2 2 2 1 2 ...  
## $ capital.gain : int 0 0 0 7688 0 0 0 3103 0 0 ...  
## $ capital.loss : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ hours.per.week: int 40 50 40 40 30 30 40 32 40 10 ...  
## $ native.country: Factor w/ 41 levels " ?"," Cambodia",..: 39 39 39 39 39 39 39 39 39 39 ...  
## $ class : Factor w/ 2 levels " <=50K."," >50K.": 1 1 2 2 1 1 1 2 1 1 ...

We first import our datasets and determine which columns contain missing values.

From a glance, we can tell that the *workclass*, *occupation* and *native.country* columns contain missing values, indicated by question marks.

### Setting Entries with Question Marks as NA Values

# train set  
train$workclass = as.factor(gsub('?', NA, train$workclass, fixed = T))  
train$native.country = as.factor(gsub('?', NA, train$native.country, fixed = T))  
train$occupation = as.factor(gsub('?', NA, train$occupation, fixed = T))  
  
# test set  
test$workclass = as.factor(gsub('?', NA, test$workclass, fixed = T))  
test$native.country = as.factor(gsub('?', NA, test$native.country, fixed = T))  
test$occupation = as.factor(gsub('?', NA, test$occupation, fixed = T))

Since the missing values exist in both the training and testing datasets, therefore we have to indicate them as NA values before we may exclude them.

### Removing Incomplete Cases

train = train[complete.cases(train), ]  
test = test[complete.cases(test), ]  
str(train)

## 'data.frame': 30162 obs. of 14 variables:  
## $ age : int 39 50 38 53 28 37 49 52 31 42 ...  
## $ workclass : Factor w/ 8 levels " Federal-gov",..: 7 6 4 4 4 4 4 6 4 4 ...  
## $ fnlwgt : int 77516 83311 215646 234721 338409 284582 160187 209642 45781 159449 ...  
## $ education : Factor w/ 16 levels " 10th"," 11th",..: 10 10 12 2 10 13 7 12 13 10 ...  
## $ marital : Factor w/ 7 levels " Divorced"," Married-AF-spouse",..: 5 3 1 3 3 3 4 3 5 3 ...  
## $ occupation : Factor w/ 14 levels " Adm-clerical",..: 1 4 6 6 10 4 8 4 10 4 ...  
## $ relationship : Factor w/ 6 levels " Husband"," Not-in-family",..: 2 1 2 1 6 6 2 1 2 1 ...  
## $ race : Factor w/ 5 levels " Amer-Indian-Eskimo",..: 5 5 5 3 3 5 3 5 5 5 ...  
## $ sex : Factor w/ 2 levels " Female"," Male": 2 2 2 2 1 1 1 2 1 2 ...  
## $ capital.gain : int 2174 0 0 0 0 0 0 0 14084 5178 ...  
## $ capital.loss : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ hours.per.week: int 40 13 40 40 40 40 16 45 50 40 ...  
## $ native.country: Factor w/ 41 levels " Cambodia"," Canada",..: 39 39 39 39 5 39 23 39 39 39 ...  
## $ class : Factor w/ 2 levels " <=50K"," >50K": 1 1 1 1 1 1 1 2 2 2 ...

str(test)

## 'data.frame': 15060 obs. of 14 variables:  
## $ age : int 25 38 28 44 34 63 24 55 65 36 ...  
## $ workclass : Factor w/ 8 levels " Federal-gov",..: 4 4 2 4 4 6 4 4 4 1 ...  
## $ fnlwgt : int 226802 89814 336951 160323 198693 104626 369667 104996 184454 212465 ...  
## $ education : Factor w/ 16 levels " 10th"," 11th",..: 2 12 8 16 1 15 16 6 12 10 ...  
## $ marital : Factor w/ 7 levels " Divorced"," Married-AF-spouse",..: 5 3 3 3 5 3 5 3 3 3 ...  
## $ occupation : Factor w/ 14 levels " Adm-clerical",..: 7 5 11 7 8 10 8 3 7 1 ...  
## $ relationship : Factor w/ 6 levels " Husband"," Not-in-family",..: 4 1 1 1 2 1 5 1 1 1 ...  
## $ race : Factor w/ 5 levels " Amer-Indian-Eskimo",..: 3 5 5 3 5 5 5 5 5 5 ...  
## $ sex : Factor w/ 2 levels " Female"," Male": 2 2 2 2 2 2 1 2 2 2 ...  
## $ capital.gain : int 0 0 0 7688 0 3103 0 0 6418 0 ...  
## $ capital.loss : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ hours.per.week: int 40 50 40 40 30 32 40 10 40 40 ...  
## $ native.country: Factor w/ 40 levels " Cambodia"," Canada",..: 38 38 38 38 38 38 38 38 38 38 ...  
## $ class : Factor w/ 2 levels " <=50K."," >50K.": 1 1 2 2 1 2 1 1 2 1 ...

We run the *complete.cases* function to remove the NA values from both datasets. After that we use the *str* function again to ascertain that the variables are in the formats we need, without anymore missing entries.

## Full Model

fit = suppressWarnings(glm(formula = class ~ .,  
 family = binomial,  
 data = train))

We start to train our training set using a logistic classifier, with *class* as our target variable. We use the rest of the variables as input.

### Prediction the Test Set Results

prob\_pred = predict(fit, type = 'response', newdata = test[-14])  
y\_pred = ifelse(prob\_pred > 0.5, '>50K', '<=50K')  
  
# Confusion Matrix  
cm = table(test[, 14], y\_pred)  
cm

## y\_pred  
## <=50K >50K  
## <=50K. 10530 830  
## >50K. 1465 2235

### Computing the Accuracy and Error Rates

acc = sum(diag(cm)) / sum(cm)  
acc

## [1] 0.8476096

err = 1 - acc  
err

## [1] 0.1523904

Model has a **84.76%** accuracy rate / **15.24%** error rate.

Let us see if we can improve the error rate through feature selection.

summary(fit)

##   
## Call:  
## glm(formula = class ~ ., family = binomial, data = train)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -5.1182 -0.5148 -0.1885 0.0000 3.7839   
##   
## Coefficients:  
## Estimate Std. Error z value  
## (Intercept) -6.408e+00 7.636e-01 -8.392  
## age 2.550e-02 1.712e-03 14.890  
## workclass Local-gov -6.985e-01 1.130e-01 -6.184  
## workclass Private -5.055e-01 9.379e-02 -5.390  
## workclass Self-emp-inc -3.293e-01 1.239e-01 -2.658  
## workclass Self-emp-not-inc -9.972e-01 1.100e-01 -9.063  
## workclass State-gov -8.207e-01 1.254e-01 -6.544  
## workclass Without-pay -1.329e+01 1.972e+02 -0.067  
## fnlwgt 7.515e-07 1.762e-07 4.264  
## education 11th 9.462e-02 2.139e-01 0.442  
## education 12th 4.443e-01 2.784e-01 1.596  
## education 1st-4th -4.398e-01 4.960e-01 -0.887  
## education 5th-6th -3.956e-01 3.590e-01 -1.102  
## education 7th-8th -5.640e-01 2.433e-01 -2.318  
## education 9th -2.372e-01 2.702e-01 -0.878  
## education Assoc-acdm 1.269e+00 1.797e-01 7.063  
## education Assoc-voc 1.268e+00 1.729e-01 7.332  
## education Bachelors 1.899e+00 1.608e-01 11.807  
## education Doctorate 2.935e+00 2.231e-01 13.159  
## education HS-grad 7.735e-01 1.564e-01 4.945  
## education Masters 2.259e+00 1.719e-01 13.138  
## education Preschool -2.008e+01 1.987e+02 -0.101  
## education Prof-school 2.844e+00 2.071e-01 13.734  
## education Some-college 1.109e+00 1.587e-01 6.989  
## marital Married-AF-spouse 2.768e+00 5.766e-01 4.800  
## marital Married-civ-spouse 2.105e+00 2.747e-01 7.663  
## marital Married-spouse-absent 1.220e-02 2.404e-01 0.051  
## marital Never-married -4.861e-01 8.926e-02 -5.446  
## marital Separated -8.940e-02 1.656e-01 -0.540  
## marital Widowed 1.852e-01 1.582e-01 1.171  
## occupation Armed-Forces -1.165e+00 1.547e+00 -0.753  
## occupation Craft-repair 6.369e-02 8.076e-02 0.789  
## occupation Exec-managerial 8.054e-01 7.794e-02 10.334  
## occupation Farming-fishing -9.809e-01 1.408e-01 -6.968  
## occupation Handlers-cleaners -6.950e-01 1.447e-01 -4.803  
## occupation Machine-op-inspct -2.633e-01 1.027e-01 -2.564  
## occupation Other-service -8.245e-01 1.191e-01 -6.920  
## occupation Priv-house-serv -4.153e+00 1.723e+00 -2.411  
## occupation Prof-specialty 5.165e-01 8.253e-02 6.259  
## occupation Protective-serv 5.978e-01 1.263e-01 4.734  
## occupation Sales 2.943e-01 8.318e-02 3.538  
## occupation Tech-support 6.648e-01 1.117e-01 5.951  
## occupation Transport-moving -8.982e-02 1.001e-01 -0.898  
## relationship Not-in-family 4.522e-01 2.716e-01 1.665  
## relationship Other-relative -3.960e-01 2.477e-01 -1.599  
## relationship Own-child -7.322e-01 2.706e-01 -2.706  
## relationship Unmarried 3.358e-01 2.873e-01 1.169  
## relationship Wife 1.351e+00 1.057e-01 12.784  
## race Asian-Pac-Islander 8.280e-01 2.860e-01 2.896  
## race Black 4.359e-01 2.409e-01 1.810  
## race Other 1.255e-01 3.786e-01 0.332  
## race White 5.875e-01 2.291e-01 2.564  
## sex Male 8.648e-01 8.091e-02 10.689  
## capital.gain 3.225e-04 1.074e-05 30.022  
## capital.loss 6.420e-04 3.845e-05 16.696  
## hours.per.week 2.949e-02 1.702e-03 17.325  
## native.country Canada -8.113e-01 6.890e-01 -1.178  
## native.country China -1.916e+00 7.031e-01 -2.725  
## native.country Columbia -3.275e+00 1.031e+00 -3.177  
## native.country Cuba -7.738e-01 7.028e-01 -1.101  
## native.country Dominican-Republic -2.915e+00 1.220e+00 -2.390  
## native.country Ecuador -1.400e+00 9.587e-01 -1.461  
## native.country El-Salvador -1.745e+00 7.922e-01 -2.203  
## native.country England -8.348e-01 7.004e-01 -1.192  
## native.country France -5.604e-01 8.137e-01 -0.689  
## native.country Germany -6.860e-01 6.781e-01 -1.012  
## native.country Greece -2.126e+00 8.369e-01 -2.540  
## native.country Guatemala -1.396e+00 9.798e-01 -1.424  
## native.country Haiti -1.169e+00 9.273e-01 -1.261  
## native.country Holand-Netherlands -1.164e+01 8.827e+02 -0.013  
## native.country Honduras -2.306e+00 2.607e+00 -0.885  
## native.country Hong -1.355e+00 9.005e-01 -1.505  
## native.country Hungary -1.254e+00 9.905e-01 -1.266  
## native.country India -1.664e+00 6.682e-01 -2.491  
## native.country Iran -1.123e+00 7.578e-01 -1.482  
## native.country Ireland -6.158e-01 8.884e-01 -0.693  
## native.country Italy -3.295e-01 7.089e-01 -0.465  
## native.country Jamaica -1.125e+00 7.708e-01 -1.460  
## native.country Japan -9.413e-01 7.294e-01 -1.290  
## native.country Laos -1.883e+00 1.046e+00 -1.801  
## native.country Mexico -1.649e+00 6.648e-01 -2.481  
## native.country Nicaragua -1.880e+00 1.020e+00 -1.843  
## native.country Outlying-US(Guam-USVI-etc) -1.342e+01 2.095e+02 -0.064  
## native.country Peru -1.985e+00 1.053e+00 -1.884  
## native.country Philippines -8.782e-01 6.441e-01 -1.363  
## native.country Poland -1.146e+00 7.455e-01 -1.537  
## native.country Portugal -1.122e+00 8.849e-01 -1.268  
## native.country Puerto-Rico -1.440e+00 7.381e-01 -1.950  
## native.country Scotland -1.407e+00 1.085e+00 -1.297  
## native.country South -2.446e+00 7.356e-01 -3.325  
## native.country Taiwan -1.384e+00 7.540e-01 -1.835  
## native.country Thailand -1.831e+00 1.017e+00 -1.800  
## native.country Trinadad&Tobago -1.580e+00 1.060e+00 -1.490  
## native.country United-States -9.549e-01 6.302e-01 -1.515  
## native.country Vietnam -2.395e+00 8.452e-01 -2.834  
## native.country Yugoslavia -4.609e-01 9.193e-01 -0.501  
## Pr(>|z|)   
## (Intercept) < 2e-16 \*\*\*  
## age < 2e-16 \*\*\*  
## workclass Local-gov 6.26e-10 \*\*\*  
## workclass Private 7.06e-08 \*\*\*  
## workclass Self-emp-inc 0.007857 \*\*   
## workclass Self-emp-not-inc < 2e-16 \*\*\*  
## workclass State-gov 6.00e-11 \*\*\*  
## workclass Without-pay 0.946265   
## fnlwgt 2.01e-05 \*\*\*  
## education 11th 0.658185   
## education 12th 0.110525   
## education 1st-4th 0.375228   
## education 5th-6th 0.270456   
## education 7th-8th 0.020461 \*   
## education 9th 0.379942   
## education Assoc-acdm 1.63e-12 \*\*\*  
## education Assoc-voc 2.27e-13 \*\*\*  
## education Bachelors < 2e-16 \*\*\*  
## education Doctorate < 2e-16 \*\*\*  
## education HS-grad 7.61e-07 \*\*\*  
## education Masters < 2e-16 \*\*\*  
## education Preschool 0.919495   
## education Prof-school < 2e-16 \*\*\*  
## education Some-college 2.76e-12 \*\*\*  
## marital Married-AF-spouse 1.59e-06 \*\*\*  
## marital Married-civ-spouse 1.82e-14 \*\*\*  
## marital Married-spouse-absent 0.959518   
## marital Never-married 5.16e-08 \*\*\*  
## marital Separated 0.589277   
## marital Widowed 0.241607   
## occupation Armed-Forces 0.451591   
## occupation Craft-repair 0.430362   
## occupation Exec-managerial < 2e-16 \*\*\*  
## occupation Farming-fishing 3.21e-12 \*\*\*  
## occupation Handlers-cleaners 1.57e-06 \*\*\*  
## occupation Machine-op-inspct 0.010360 \*   
## occupation Other-service 4.50e-12 \*\*\*  
## occupation Priv-house-serv 0.015916 \*   
## occupation Prof-specialty 3.87e-10 \*\*\*  
## occupation Protective-serv 2.20e-06 \*\*\*  
## occupation Sales 0.000403 \*\*\*  
## occupation Tech-support 2.66e-09 \*\*\*  
## occupation Transport-moving 0.369448   
## relationship Not-in-family 0.095982 .   
## relationship Other-relative 0.109885   
## relationship Own-child 0.006812 \*\*   
## relationship Unmarried 0.242463   
## relationship Wife < 2e-16 \*\*\*  
## race Asian-Pac-Islander 0.003785 \*\*   
## race Black 0.070321 .   
## race Other 0.740237   
## race White 0.010343 \*   
## sex Male < 2e-16 \*\*\*  
## capital.gain < 2e-16 \*\*\*  
## capital.loss < 2e-16 \*\*\*  
## hours.per.week < 2e-16 \*\*\*  
## native.country Canada 0.238944   
## native.country China 0.006439 \*\*   
## native.country Columbia 0.001490 \*\*   
## native.country Cuba 0.270924   
## native.country Dominican-Republic 0.016847 \*   
## native.country Ecuador 0.144101   
## native.country El-Salvador 0.027612 \*   
## native.country England 0.233300   
## native.country France 0.491044   
## native.country Germany 0.311750   
## native.country Greece 0.011087 \*   
## native.country Guatemala 0.154334   
## native.country Haiti 0.207295   
## native.country Holand-Netherlands 0.989483   
## native.country Honduras 0.376267   
## native.country Hong 0.132285   
## native.country Hungary 0.205684   
## native.country India 0.012746 \*   
## native.country Iran 0.138389   
## native.country Ireland 0.488248   
## native.country Italy 0.642116   
## native.country Jamaica 0.144361   
## native.country Japan 0.196881   
## native.country Laos 0.071769 .   
## native.country Mexico 0.013106 \*   
## native.country Nicaragua 0.065264 .   
## native.country Outlying-US(Guam-USVI-etc) 0.948933   
## native.country Peru 0.059554 .   
## native.country Philippines 0.172770   
## native.country Poland 0.124250   
## native.country Portugal 0.204960   
## native.country Puerto-Rico 0.051129 .   
## native.country Scotland 0.194512   
## native.country South 0.000883 \*\*\*  
## native.country Taiwan 0.066463 .   
## native.country Thailand 0.071863 .   
## native.country Trinadad&Tobago 0.136166   
## native.country United-States 0.129711   
## native.country Vietnam 0.004601 \*\*   
## native.country Yugoslavia 0.616151   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 33851 on 30161 degrees of freedom  
## Residual deviance: 19486 on 30066 degrees of freedom  
## AIC: 19678  
##   
## Number of Fisher Scoring iterations: 13

## Model 1

We will try dropping the *race* variable as it does not appear to be significant from the p-values (mostly > 0.05).

fit\_1 = suppressWarnings(glm(formula = class ~ . - race,  
 family = binomial,  
 data = train))  
  
summary(fit\_1)

##   
## Call:  
## glm(formula = class ~ . - race, family = binomial, data = train)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -5.1125 -0.5152 -0.1898 0.0000 3.7969   
##   
## Coefficients:  
## Estimate Std. Error z value  
## (Intercept) -5.615e+00 7.100e-01 -7.907  
## age 2.567e-02 1.712e-03 14.995  
## workclass Local-gov -6.884e-01 1.126e-01 -6.114  
## workclass Private -4.879e-01 9.321e-02 -5.235  
## workclass Self-emp-inc -3.057e-01 1.234e-01 -2.477  
## workclass Self-emp-not-inc -9.791e-01 1.095e-01 -8.944  
## workclass State-gov -8.076e-01 1.251e-01 -6.455  
## workclass Without-pay -1.326e+01 1.973e+02 -0.067  
## fnlwgt 7.227e-07 1.742e-07 4.149  
## education 11th 9.811e-02 2.139e-01 0.459  
## education 12th 4.474e-01 2.783e-01 1.608  
## education 1st-4th -4.344e-01 4.956e-01 -0.877  
## education 5th-6th -3.992e-01 3.590e-01 -1.112  
## education 7th-8th -5.654e-01 2.434e-01 -2.323  
## education 9th -2.331e-01 2.699e-01 -0.864  
## education Assoc-acdm 1.283e+00 1.796e-01 7.143  
## education Assoc-voc 1.278e+00 1.728e-01 7.400  
## education Bachelors 1.912e+00 1.606e-01 11.904  
## education Doctorate 2.948e+00 2.228e-01 13.232  
## education HS-grad 7.806e-01 1.562e-01 4.996  
## education Masters 2.274e+00 1.718e-01 13.239  
## education Preschool -2.000e+01 1.973e+02 -0.101  
## education Prof-school 2.861e+00 2.070e-01 13.823  
## education Some-college 1.115e+00 1.586e-01 7.030  
## marital Married-AF-spouse 2.782e+00 5.768e-01 4.822  
## marital Married-civ-spouse 2.108e+00 2.746e-01 7.679  
## marital Married-spouse-absent 2.107e-03 2.399e-01 0.009  
## marital Never-married -4.861e-01 8.915e-02 -5.452  
## marital Separated -1.045e-01 1.651e-01 -0.633  
## marital Widowed 1.868e-01 1.581e-01 1.182  
## occupation Armed-Forces -1.227e+00 1.508e+00 -0.814  
## occupation Craft-repair 6.722e-02 8.071e-02 0.833  
## occupation Exec-managerial 8.083e-01 7.789e-02 10.377  
## occupation Farming-fishing -9.770e-01 1.406e-01 -6.948  
## occupation Handlers-cleaners -6.984e-01 1.446e-01 -4.829  
## occupation Machine-op-inspct -2.704e-01 1.026e-01 -2.635  
## occupation Other-service -8.330e-01 1.190e-01 -7.002  
## occupation Priv-house-serv -4.139e+00 1.710e+00 -2.421  
## occupation Prof-specialty 5.166e-01 8.240e-02 6.269  
## occupation Protective-serv 5.935e-01 1.262e-01 4.704  
## occupation Sales 2.975e-01 8.312e-02 3.579  
## occupation Tech-support 6.726e-01 1.116e-01 6.028  
## occupation Transport-moving -9.750e-02 9.998e-02 -0.975  
## relationship Not-in-family 4.532e-01 2.715e-01 1.669  
## relationship Other-relative -4.003e-01 2.479e-01 -1.615  
## relationship Own-child -7.286e-01 2.702e-01 -2.697  
## relationship Unmarried 3.291e-01 2.871e-01 1.146  
## relationship Wife 1.350e+00 1.057e-01 12.773  
## sex Male 8.702e-01 8.089e-02 10.758  
## capital.gain 3.218e-04 1.074e-05 29.968  
## capital.loss 6.429e-04 3.846e-05 16.717  
## hours.per.week 2.956e-02 1.702e-03 17.365  
## native.country Canada -1.056e+00 6.668e-01 -1.584  
## native.country China -1.926e+00 7.038e-01 -2.737  
## native.country Columbia -3.549e+00 1.013e+00 -3.502  
## native.country Cuba -1.028e+00 6.813e-01 -1.509  
## native.country Dominican-Republic -3.270e+00 1.206e+00 -2.711  
## native.country Ecuador -1.771e+00 9.373e-01 -1.889  
## native.country El-Salvador -1.983e+00 7.731e-01 -2.566  
## native.country England -1.090e+00 6.789e-01 -1.606  
## native.country France -8.087e-01 7.945e-01 -1.018  
## native.country Germany -9.294e-01 6.562e-01 -1.416  
## native.country Greece -2.372e+00 8.200e-01 -2.893  
## native.country Guatemala -1.657e+00 9.630e-01 -1.720  
## native.country Haiti -1.556e+00 9.098e-01 -1.711  
## native.country Holand-Netherlands -1.186e+01 8.827e+02 -0.013  
## native.country Honduras -2.551e+00 2.611e+00 -0.977  
## native.country Hong -1.375e+00 8.978e-01 -1.531  
## native.country Hungary -1.502e+00 9.750e-01 -1.541  
## native.country India -1.732e+00 6.674e-01 -2.594  
## native.country Iran -1.358e+00 7.434e-01 -1.827  
## native.country Ireland -8.421e-01 8.751e-01 -0.962  
## native.country Italy -5.712e-01 6.874e-01 -0.831  
## native.country Jamaica -1.502e+00 7.480e-01 -2.008  
## native.country Japan -1.041e+00 7.263e-01 -1.433  
## native.country Laos -1.876e+00 1.046e+00 -1.794  
## native.country Mexico -1.919e+00 6.419e-01 -2.990  
## native.country Nicaragua -2.153e+00 1.007e+00 -2.139  
## native.country Outlying-US(Guam-USVI-etc) -1.369e+01 2.102e+02 -0.065  
## native.country Peru -2.225e+00 1.040e+00 -2.140  
## native.country Philippines -9.029e-01 6.442e-01 -1.402  
## native.country Poland -1.386e+00 7.258e-01 -1.910  
## native.country Portugal -1.362e+00 8.680e-01 -1.569  
## native.country Puerto-Rico -1.789e+00 7.155e-01 -2.501  
## native.country Scotland -1.648e+00 1.072e+00 -1.538  
## native.country South -2.458e+00 7.361e-01 -3.339  
## native.country Taiwan -1.403e+00 7.539e-01 -1.861  
## native.country Thailand -1.875e+00 1.019e+00 -1.840  
## native.country Trinadad&Tobago -1.883e+00 1.056e+00 -1.784  
## native.country United-States -1.211e+00 6.062e-01 -1.998  
## native.country Vietnam -2.399e+00 8.456e-01 -2.837  
## native.country Yugoslavia -7.046e-01 9.035e-01 -0.780  
## Pr(>|z|)   
## (Intercept) 2.63e-15 \*\*\*  
## age < 2e-16 \*\*\*  
## workclass Local-gov 9.72e-10 \*\*\*  
## workclass Private 1.65e-07 \*\*\*  
## workclass Self-emp-inc 0.013236 \*   
## workclass Self-emp-not-inc < 2e-16 \*\*\*  
## workclass State-gov 1.08e-10 \*\*\*  
## workclass Without-pay 0.946429   
## fnlwgt 3.33e-05 \*\*\*  
## education 11th 0.646464   
## education 12th 0.107826   
## education 1st-4th 0.380651   
## education 5th-6th 0.266173   
## education 7th-8th 0.020155 \*   
## education 9th 0.387810   
## education Assoc-acdm 9.15e-13 \*\*\*  
## education Assoc-voc 1.36e-13 \*\*\*  
## education Bachelors < 2e-16 \*\*\*  
## education Doctorate < 2e-16 \*\*\*  
## education HS-grad 5.86e-07 \*\*\*  
## education Masters < 2e-16 \*\*\*  
## education Preschool 0.919269   
## education Prof-school < 2e-16 \*\*\*  
## education Some-college 2.06e-12 \*\*\*  
## marital Married-AF-spouse 1.42e-06 \*\*\*  
## marital Married-civ-spouse 1.60e-14 \*\*\*  
## marital Married-spouse-absent 0.992995   
## marital Never-married 4.97e-08 \*\*\*  
## marital Separated 0.526921   
## marital Widowed 0.237308   
## occupation Armed-Forces 0.415756   
## occupation Craft-repair 0.404877   
## occupation Exec-managerial < 2e-16 \*\*\*  
## occupation Farming-fishing 3.71e-12 \*\*\*  
## occupation Handlers-cleaners 1.37e-06 \*\*\*  
## occupation Machine-op-inspct 0.008409 \*\*   
## occupation Other-service 2.52e-12 \*\*\*  
## occupation Priv-house-serv 0.015477 \*   
## occupation Prof-specialty 3.62e-10 \*\*\*  
## occupation Protective-serv 2.55e-06 \*\*\*  
## occupation Sales 0.000345 \*\*\*  
## occupation Tech-support 1.67e-09 \*\*\*  
## occupation Transport-moving 0.329475   
## relationship Not-in-family 0.095025 .   
## relationship Other-relative 0.106368   
## relationship Own-child 0.007004 \*\*   
## relationship Unmarried 0.251693   
## relationship Wife < 2e-16 \*\*\*  
## sex Male < 2e-16 \*\*\*  
## capital.gain < 2e-16 \*\*\*  
## capital.loss < 2e-16 \*\*\*  
## hours.per.week < 2e-16 \*\*\*  
## native.country Canada 0.113109   
## native.country China 0.006206 \*\*   
## native.country Columbia 0.000462 \*\*\*  
## native.country Cuba 0.131195   
## native.country Dominican-Republic 0.006702 \*\*   
## native.country Ecuador 0.058833 .   
## native.country El-Salvador 0.010301 \*   
## native.country England 0.108287   
## native.country France 0.308779   
## native.country Germany 0.156655   
## native.country Greece 0.003816 \*\*   
## native.country Guatemala 0.085365 .   
## native.country Haiti 0.087122 .   
## native.country Holand-Netherlands 0.989278   
## native.country Honduras 0.328607   
## native.country Hong 0.125672   
## native.country Hungary 0.123393   
## native.country India 0.009475 \*\*   
## native.country Iran 0.067709 .   
## native.country Ireland 0.335890   
## native.country Italy 0.406006   
## native.country Jamaica 0.044620 \*   
## native.country Japan 0.151799   
## native.country Laos 0.072802 .   
## native.country Mexico 0.002792 \*\*   
## native.country Nicaragua 0.032462 \*   
## native.country Outlying-US(Guam-USVI-etc) 0.948048   
## native.country Peru 0.032333 \*   
## native.country Philippines 0.161005   
## native.country Poland 0.056103 .   
## native.country Portugal 0.116613   
## native.country Puerto-Rico 0.012389 \*   
## native.country Scotland 0.123997   
## native.country South 0.000841 \*\*\*  
## native.country Taiwan 0.062795 .   
## native.country Thailand 0.065710 .   
## native.country Trinadad&Tobago 0.074484 .   
## native.country United-States 0.045757 \*   
## native.country Vietnam 0.004549 \*\*   
## native.country Yugoslavia 0.435500   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 33851 on 30161 degrees of freedom  
## Residual deviance: 19501 on 30070 degrees of freedom  
## AIC: 19685  
##   
## Number of Fisher Scoring iterations: 13

### Prediction the Test Set Results

prob\_pred\_1 = predict(fit\_1, type = 'response', newdata = test[-14])  
y\_pred\_1 = ifelse(prob\_pred\_1 > 0.5, '>50K', '<=50K')  
  
# Confusion Matrix 1  
cm\_1 = table(test[, 14], y\_pred\_1)  
cm\_1

## y\_pred\_1  
## <=50K >50K  
## <=50K. 10537 823  
## >50K. 1470 2230

### Computing the Accuracy and Error Rates

acc\_1 = sum(diag(cm\_1)) / sum(cm\_1)  
acc\_1

## [1] 0.8477424

err\_1 = 1 - acc\_1  
err\_1

## [1] 0.1522576

This model has an accuracy rate of **84.77%**, which is only very slightly improved.

Model 1 is our best model so far.

## Model 2

We remove the *relationship* variable as well as it appears to be a less significant variable.

fit\_2 = suppressWarnings(glm(formula = class ~ . - race - relationship,  
 family = binomial,  
 data = train))  
  
summary(fit\_2)

##   
## Call:  
## glm(formula = class ~ . - race - relationship, family = binomial,   
## data = train)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -5.1465 -0.5077 -0.2119 0.0000 3.7692   
##   
## Coefficients:  
## Estimate Std. Error z value  
## (Intercept) -4.907e+00 6.515e-01 -7.532  
## age 2.596e-02 1.685e-03 15.407  
## workclass Local-gov -6.962e-01 1.117e-01 -6.234  
## workclass Private -4.847e-01 9.256e-02 -5.237  
## workclass Self-emp-inc -2.886e-01 1.234e-01 -2.339  
## workclass Self-emp-not-inc -9.597e-01 1.091e-01 -8.793  
## workclass State-gov -8.097e-01 1.244e-01 -6.510  
## workclass Without-pay -1.317e+01 1.987e+02 -0.066  
## fnlwgt 7.475e-07 1.736e-07 4.307  
## education 11th 8.503e-02 2.134e-01 0.398  
## education 12th 4.611e-01 2.778e-01 1.660  
## education 1st-4th -4.551e-01 4.958e-01 -0.918  
## education 5th-6th -4.098e-01 3.589e-01 -1.142  
## education 7th-8th -5.663e-01 2.432e-01 -2.329  
## education 9th -2.120e-01 2.694e-01 -0.787  
## education Assoc-acdm 1.319e+00 1.792e-01 7.361  
## education Assoc-voc 1.279e+00 1.724e-01 7.418  
## education Bachelors 1.932e+00 1.603e-01 12.051  
## education Doctorate 2.947e+00 2.225e-01 13.245  
## education HS-grad 7.914e-01 1.560e-01 5.073  
## education Masters 2.294e+00 1.713e-01 13.387  
## education Preschool -2.043e+01 1.890e+02 -0.108  
## education Prof-school 2.877e+00 2.068e-01 13.911  
## education Some-college 1.112e+00 1.583e-01 7.027  
## marital Married-AF-spouse 3.065e+00 5.032e-01 6.090  
## marital Married-civ-spouse 2.183e+00 6.809e-02 32.066  
## marital Married-spouse-absent 4.964e-02 2.360e-01 0.210  
## marital Never-married -5.141e-01 8.407e-02 -6.115  
## marital Separated -1.095e-01 1.616e-01 -0.678  
## marital Widowed 3.762e-02 1.547e-01 0.243  
## occupation Armed-Forces -1.331e+00 1.525e+00 -0.873  
## occupation Craft-repair 3.651e-03 7.939e-02 0.046  
## occupation Exec-managerial 7.757e-01 7.618e-02 10.183  
## occupation Farming-fishing -1.057e+00 1.404e-01 -7.531  
## occupation Handlers-cleaners -7.626e-01 1.441e-01 -5.294  
## occupation Machine-op-inspct -3.175e-01 1.017e-01 -3.123  
## occupation Other-service -8.348e-01 1.176e-01 -7.095  
## occupation Priv-house-serv -4.464e+00 1.726e+00 -2.586  
## occupation Prof-specialty 4.932e-01 8.077e-02 6.106  
## occupation Protective-serv 5.505e-01 1.258e-01 4.377  
## occupation Sales 2.461e-01 8.159e-02 3.016  
## occupation Tech-support 6.340e-01 1.098e-01 5.775  
## occupation Transport-moving -1.512e-01 9.918e-02 -1.525  
## sex Male 1.549e-01 5.398e-02 2.869  
## capital.gain 3.249e-04 1.068e-05 30.406  
## capital.loss 6.525e-04 3.838e-05 16.999  
## hours.per.week 2.950e-02 1.688e-03 17.473  
## native.country Canada -1.039e+00 6.706e-01 -1.549  
## native.country China -1.957e+00 7.083e-01 -2.763  
## native.country Columbia -3.598e+00 1.017e+00 -3.536  
## native.country Cuba -1.078e+00 6.821e-01 -1.580  
## native.country Dominican-Republic -3.142e+00 1.211e+00 -2.594  
## native.country Ecuador -1.836e+00 9.404e-01 -1.952  
## native.country El-Salvador -1.961e+00 7.769e-01 -2.524  
## native.country England -1.051e+00 6.824e-01 -1.540  
## native.country France -7.504e-01 7.999e-01 -0.938  
## native.country Germany -9.585e-01 6.589e-01 -1.455  
## native.country Greece -2.376e+00 8.231e-01 -2.886  
## native.country Guatemala -1.646e+00 9.695e-01 -1.698  
## native.country Haiti -1.600e+00 8.984e-01 -1.780  
## native.country Holand-Netherlands -1.293e+01 8.827e+02 -0.015  
## native.country Honduras -2.298e+00 2.325e+00 -0.988  
## native.country Hong -1.303e+00 9.066e-01 -1.438  
## native.country Hungary -1.476e+00 9.839e-01 -1.501  
## native.country India -1.755e+00 6.734e-01 -2.606  
## native.country Iran -1.373e+00 7.492e-01 -1.832  
## native.country Ireland -8.464e-01 8.845e-01 -0.957  
## native.country Italy -5.716e-01 6.902e-01 -0.828  
## native.country Jamaica -1.491e+00 7.505e-01 -1.986  
## native.country Japan -1.080e+00 7.309e-01 -1.478  
## native.country Laos -1.842e+00 1.058e+00 -1.741  
## native.country Mexico -1.932e+00 6.463e-01 -2.989  
## native.country Nicaragua -2.099e+00 1.009e+00 -2.081  
## native.country Outlying-US(Guam-USVI-etc) -1.354e+01 2.136e+02 -0.063  
## native.country Peru -2.231e+00 1.045e+00 -2.135  
## native.country Philippines -1.000e+00 6.472e-01 -1.545  
## native.country Poland -1.418e+00 7.297e-01 -1.943  
## native.country Portugal -1.238e+00 8.705e-01 -1.423  
## native.country Puerto-Rico -1.708e+00 7.186e-01 -2.377  
## native.country Scotland -1.478e+00 1.105e+00 -1.337  
## native.country South -2.401e+00 7.383e-01 -3.252  
## native.country Taiwan -1.410e+00 7.520e-01 -1.875  
## native.country Thailand -1.823e+00 9.969e-01 -1.828  
## native.country Trinadad&Tobago -1.776e+00 1.054e+00 -1.685  
## native.country United-States -1.211e+00 6.107e-01 -1.983  
## native.country Vietnam -2.459e+00 8.495e-01 -2.894  
## native.country Yugoslavia -7.617e-01 9.076e-01 -0.839  
## Pr(>|z|)   
## (Intercept) 5.00e-14 \*\*\*  
## age < 2e-16 \*\*\*  
## workclass Local-gov 4.55e-10 \*\*\*  
## workclass Private 1.63e-07 \*\*\*  
## workclass Self-emp-inc 0.019331 \*   
## workclass Self-emp-not-inc < 2e-16 \*\*\*  
## workclass State-gov 7.52e-11 \*\*\*  
## workclass Without-pay 0.947131   
## fnlwgt 1.66e-05 \*\*\*  
## education 11th 0.690349   
## education 12th 0.096984 .   
## education 1st-4th 0.358698   
## education 5th-6th 0.253590   
## education 7th-8th 0.019856 \*   
## education 9th 0.431193   
## education Assoc-acdm 1.82e-13 \*\*\*  
## education Assoc-voc 1.19e-13 \*\*\*  
## education Bachelors < 2e-16 \*\*\*  
## education Doctorate < 2e-16 \*\*\*  
## education HS-grad 3.92e-07 \*\*\*  
## education Masters < 2e-16 \*\*\*  
## education Preschool 0.913927   
## education Prof-school < 2e-16 \*\*\*  
## education Some-college 2.10e-12 \*\*\*  
## marital Married-AF-spouse 1.13e-09 \*\*\*  
## marital Married-civ-spouse < 2e-16 \*\*\*  
## marital Married-spouse-absent 0.833402   
## marital Never-married 9.63e-10 \*\*\*  
## marital Separated 0.497894   
## marital Widowed 0.807930   
## occupation Armed-Forces 0.382836   
## occupation Craft-repair 0.963320   
## occupation Exec-managerial < 2e-16 \*\*\*  
## occupation Farming-fishing 5.03e-14 \*\*\*  
## occupation Handlers-cleaners 1.20e-07 \*\*\*  
## occupation Machine-op-inspct 0.001790 \*\*   
## occupation Other-service 1.29e-12 \*\*\*  
## occupation Priv-house-serv 0.009697 \*\*   
## occupation Prof-specialty 1.02e-09 \*\*\*  
## occupation Protective-serv 1.20e-05 \*\*\*  
## occupation Sales 0.002558 \*\*   
## occupation Tech-support 7.71e-09 \*\*\*  
## occupation Transport-moving 0.127245   
## sex Male 0.004117 \*\*   
## capital.gain < 2e-16 \*\*\*  
## capital.loss < 2e-16 \*\*\*  
## hours.per.week < 2e-16 \*\*\*  
## native.country Canada 0.121407   
## native.country China 0.005732 \*\*   
## native.country Columbia 0.000406 \*\*\*  
## native.country Cuba 0.114067   
## native.country Dominican-Republic 0.009484 \*\*   
## native.country Ecuador 0.050954 .   
## native.country El-Salvador 0.011614 \*   
## native.country England 0.123658   
## native.country France 0.348227   
## native.country Germany 0.145750   
## native.country Greece 0.003897 \*\*   
## native.country Guatemala 0.089549 .   
## native.country Haiti 0.075013 .   
## native.country Holand-Netherlands 0.988318   
## native.country Honduras 0.323019   
## native.country Hong 0.150499   
## native.country Hungary 0.133444   
## native.country India 0.009156 \*\*   
## native.country Iran 0.066910 .   
## native.country Ireland 0.338616   
## native.country Italy 0.407571   
## native.country Jamaica 0.047015 \*   
## native.country Japan 0.139346   
## native.country Laos 0.081625 .   
## native.country Mexico 0.002799 \*\*   
## native.country Nicaragua 0.037463 \*   
## native.country Outlying-US(Guam-USVI-etc) 0.949445   
## native.country Peru 0.032747 \*   
## native.country Philippines 0.122232   
## native.country Poland 0.051960 .   
## native.country Portugal 0.154880   
## native.country Puerto-Rico 0.017455 \*   
## native.country Scotland 0.181291   
## native.country South 0.001147 \*\*   
## native.country Taiwan 0.060777 .   
## native.country Thailand 0.067504 .   
## native.country Trinadad&Tobago 0.091952 .   
## native.country United-States 0.047365 \*   
## native.country Vietnam 0.003803 \*\*   
## native.country Yugoslavia 0.401278   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 33851 on 30161 degrees of freedom  
## Residual deviance: 19767 on 30075 degrees of freedom  
## AIC: 19941  
##   
## Number of Fisher Scoring iterations: 13

### Prediction the Test Set Results

prob\_pred\_2 = predict(fit\_2, type = 'response', newdata = test[-14])  
y\_pred\_2 = ifelse(prob\_pred\_2 > 0.5, '>50K', '<=50K')  
  
cm\_2 = table(test[, 14], y\_pred\_2)  
cm\_2

## y\_pred\_2  
## <=50K >50K  
## <=50K. 10554 806  
## >50K. 1504 2196

### Computing the Accuracy and Error Rates

acc\_2 = sum(diag(cm\_2)) / sum(cm\_2)  
acc\_2

## [1] 0.8466135

err\_2 = 1 - acc\_2  
err\_2

## [1] 0.1533865

However, accuracy rate has decreased to **84.66**.

## Model 3

We will do more data cleaning, for it appears that there are many zero values present in the *capital.loss* and *capital.gain* columns. Let's remove these from our best model so far (Model 1) and see if the result improves.

fit\_3 = suppressWarnings(glm(formula = class ~ . - race - capital.gain - capital.loss,  
 family = binomial,  
 data = train))  
  
summary(fit\_3)

##   
## Call:  
## glm(formula = class ~ . - race - capital.gain - capital.loss,   
## family = binomial, data = train)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.6754 -0.5672 -0.2165 -0.0005 3.7071   
##   
## Coefficients:  
## Estimate Std. Error z value  
## (Intercept) -5.511e+00 6.843e-01 -8.053  
## age 2.901e-02 1.627e-03 17.834  
## workclass Local-gov -6.543e-01 1.078e-01 -6.071  
## workclass Private -4.388e-01 8.910e-02 -4.925  
## workclass Self-emp-inc -1.950e-01 1.171e-01 -1.665  
## workclass Self-emp-not-inc -8.986e-01 1.043e-01 -8.616  
## workclass State-gov -8.377e-01 1.201e-01 -6.977  
## workclass Without-pay -1.337e+01 1.953e+02 -0.068  
## fnlwgt 7.609e-07 1.658e-07 4.590  
## education 11th 1.601e-01 2.058e-01 0.778  
## education 12th 5.000e-01 2.615e-01 1.912  
## education 1st-4th -4.230e-01 4.704e-01 -0.899  
## education 5th-6th -3.329e-01 3.502e-01 -0.951  
## education 7th-8th -5.529e-01 2.352e-01 -2.351  
## education 9th -2.938e-01 2.623e-01 -1.120  
## education Assoc-acdm 1.338e+00 1.715e-01 7.799  
## education Assoc-voc 1.362e+00 1.647e-01 8.272  
## education Bachelors 2.005e+00 1.535e-01 13.064  
## education Doctorate 3.105e+00 2.112e-01 14.701  
## education HS-grad 8.293e-01 1.493e-01 5.554  
## education Masters 2.426e+00 1.638e-01 14.814  
## education Preschool -1.112e+01 1.085e+02 -0.103  
## education Prof-school 3.107e+00 1.954e-01 15.901  
## education Some-college 1.162e+00 1.515e-01 7.669  
## marital Married-AF-spouse 2.517e+00 5.661e-01 4.446  
## marital Married-civ-spouse 1.995e+00 2.662e-01 7.492  
## marital Married-spouse-absent -3.755e-02 2.212e-01 -0.170  
## marital Never-married -4.553e-01 8.161e-02 -5.579  
## marital Separated -8.774e-02 1.511e-01 -0.581  
## marital Widowed 1.759e-01 1.439e-01 1.222  
## occupation Armed-Forces -9.862e-01 1.282e+00 -0.769  
## occupation Craft-repair 6.640e-02 7.668e-02 0.866  
## occupation Exec-managerial 8.242e-01 7.359e-02 11.200  
## occupation Farming-fishing -9.415e-01 1.319e-01 -7.138  
## occupation Handlers-cleaners -7.195e-01 1.384e-01 -5.200  
## occupation Machine-op-inspct -3.031e-01 9.832e-02 -3.083  
## occupation Other-service -9.176e-01 1.139e-01 -8.058  
## occupation Priv-house-serv -2.541e+00 1.113e+00 -2.282  
## occupation Prof-specialty 5.255e-01 7.798e-02 6.739  
## occupation Protective-serv 5.474e-01 1.210e-01 4.523  
## occupation Sales 3.105e-01 7.850e-02 3.955  
## occupation Tech-support 6.265e-01 1.071e-01 5.852  
## occupation Transport-moving -1.221e-01 9.513e-02 -1.284  
## relationship Not-in-family 4.550e-01 2.636e-01 1.726  
## relationship Other-relative -4.417e-01 2.392e-01 -1.846  
## relationship Own-child -7.567e-01 2.643e-01 -2.863  
## relationship Unmarried 2.608e-01 2.771e-01 0.941  
## relationship Wife 1.366e+00 9.874e-02 13.836  
## sex Male 8.678e-01 7.387e-02 11.747  
## hours.per.week 3.013e-02 1.622e-03 18.572  
## native.country Canada -1.095e+00 6.434e-01 -1.702  
## native.country China -1.926e+00 6.824e-01 -2.823  
## native.country Columbia -3.778e+00 1.016e+00 -3.718  
## native.country Cuba -1.126e+00 6.584e-01 -1.710  
## native.country Dominican-Republic -2.779e+00 9.639e-01 -2.883  
## native.country Ecuador -1.828e+00 8.875e-01 -2.060  
## native.country El-Salvador -1.740e+00 7.417e-01 -2.346  
## native.country England -1.076e+00 6.579e-01 -1.636  
## native.country France -8.657e-01 7.876e-01 -1.099  
## native.country Germany -9.555e-01 6.322e-01 -1.511  
## native.country Greece -2.058e+00 7.774e-01 -2.647  
## native.country Guatemala -1.592e+00 8.907e-01 -1.788  
## native.country Haiti -1.669e+00 8.821e-01 -1.893  
## native.country Holand-Netherlands -1.073e+01 8.827e+02 -0.012  
## native.country Honduras -2.386e+00 2.100e+00 -1.136  
## native.country Hong -1.603e+00 8.770e-01 -1.827  
## native.country Hungary -1.481e+00 9.316e-01 -1.590  
## native.country India -1.776e+00 6.423e-01 -2.766  
## native.country Iran -1.322e+00 7.065e-01 -1.871  
## native.country Ireland -8.642e-01 8.525e-01 -1.014  
## native.country Italy -6.506e-01 6.646e-01 -0.979  
## native.country Jamaica -1.628e+00 7.172e-01 -2.270  
## native.country Japan -1.067e+00 6.967e-01 -1.531  
## native.country Laos -2.080e+00 1.052e+00 -1.978  
## native.country Mexico -1.987e+00 6.198e-01 -3.206  
## native.country Nicaragua -2.281e+00 1.001e+00 -2.278  
## native.country Outlying-US(Guam-USVI-etc) -1.400e+01 2.084e+02 -0.067  
## native.country Peru -2.286e+00 9.909e-01 -2.308  
## native.country Philippines -9.729e-01 6.218e-01 -1.565  
## native.country Poland -1.470e+00 7.090e-01 -2.074  
## native.country Portugal -1.480e+00 8.596e-01 -1.722  
## native.country Puerto-Rico -1.800e+00 6.904e-01 -2.607  
## native.country Scotland -1.957e+00 1.063e+00 -1.840  
## native.country South -2.417e+00 6.967e-01 -3.469  
## native.country Taiwan -1.590e+00 7.272e-01 -2.186  
## native.country Thailand -2.194e+00 1.006e+00 -2.181  
## native.country Trinadad&Tobago -1.953e+00 1.013e+00 -1.928  
## native.country United-States -1.216e+00 5.858e-01 -2.076  
## native.country Vietnam -2.235e+00 8.033e-01 -2.782  
## native.country Yugoslavia -8.269e-01 8.774e-01 -0.942  
## Pr(>|z|)   
## (Intercept) 8.08e-16 \*\*\*  
## age < 2e-16 \*\*\*  
## workclass Local-gov 1.27e-09 \*\*\*  
## workclass Private 8.42e-07 \*\*\*  
## workclass Self-emp-inc 0.095955 .   
## workclass Self-emp-not-inc < 2e-16 \*\*\*  
## workclass State-gov 3.01e-12 \*\*\*  
## workclass Without-pay 0.945409   
## fnlwgt 4.43e-06 \*\*\*  
## education 11th 0.436659   
## education 12th 0.055908 .   
## education 1st-4th 0.368534   
## education 5th-6th 0.341741   
## education 7th-8th 0.018728 \*   
## education 9th 0.262668   
## education Assoc-acdm 6.23e-15 \*\*\*  
## education Assoc-voc < 2e-16 \*\*\*  
## education Bachelors < 2e-16 \*\*\*  
## education Doctorate < 2e-16 \*\*\*  
## education HS-grad 2.79e-08 \*\*\*  
## education Masters < 2e-16 \*\*\*  
## education Preschool 0.918345   
## education Prof-school < 2e-16 \*\*\*  
## education Some-college 1.73e-14 \*\*\*  
## marital Married-AF-spouse 8.77e-06 \*\*\*  
## marital Married-civ-spouse 6.79e-14 \*\*\*  
## marital Married-spouse-absent 0.865167   
## marital Never-married 2.42e-08 \*\*\*  
## marital Separated 0.561557   
## marital Widowed 0.221520   
## occupation Armed-Forces 0.441618   
## occupation Craft-repair 0.386556   
## occupation Exec-managerial < 2e-16 \*\*\*  
## occupation Farming-fishing 9.44e-13 \*\*\*  
## occupation Handlers-cleaners 1.99e-07 \*\*\*  
## occupation Machine-op-inspct 0.002052 \*\*   
## occupation Other-service 7.78e-16 \*\*\*  
## occupation Priv-house-serv 0.022478 \*   
## occupation Prof-specialty 1.59e-11 \*\*\*  
## occupation Protective-serv 6.10e-06 \*\*\*  
## occupation Sales 7.66e-05 \*\*\*  
## occupation Tech-support 4.86e-09 \*\*\*  
## occupation Transport-moving 0.199116   
## relationship Not-in-family 0.084413 .   
## relationship Other-relative 0.064860 .   
## relationship Own-child 0.004195 \*\*   
## relationship Unmarried 0.346482   
## relationship Wife < 2e-16 \*\*\*  
## sex Male < 2e-16 \*\*\*  
## hours.per.week < 2e-16 \*\*\*  
## native.country Canada 0.088800 .   
## native.country China 0.004765 \*\*   
## native.country Columbia 0.000200 \*\*\*  
## native.country Cuba 0.087222 .   
## native.country Dominican-Republic 0.003940 \*\*   
## native.country Ecuador 0.039383 \*   
## native.country El-Salvador 0.018952 \*   
## native.country England 0.101862   
## native.country France 0.271686   
## native.country Germany 0.130686   
## native.country Greece 0.008129 \*\*   
## native.country Guatemala 0.073806 .   
## native.country Haiti 0.058415 .   
## native.country Holand-Netherlands 0.990299   
## native.country Honduras 0.255750   
## native.country Hong 0.067648 .   
## native.country Hungary 0.111904   
## native.country India 0.005681 \*\*   
## native.country Iran 0.061347 .   
## native.country Ireland 0.310691   
## native.country Italy 0.327626   
## native.country Jamaica 0.023180 \*   
## native.country Japan 0.125673   
## native.country Laos 0.047900 \*   
## native.country Mexico 0.001347 \*\*   
## native.country Nicaragua 0.022754 \*   
## native.country Outlying-US(Guam-USVI-etc) 0.946461   
## native.country Peru 0.021024 \*   
## native.country Philippines 0.117635   
## native.country Poland 0.038103 \*   
## native.country Portugal 0.085115 .   
## native.country Puerto-Rico 0.009128 \*\*   
## native.country Scotland 0.065725 .   
## native.country South 0.000522 \*\*\*  
## native.country Taiwan 0.028807 \*   
## native.country Thailand 0.029210 \*   
## native.country Trinadad&Tobago 0.053803 .   
## native.country United-States 0.037914 \*   
## native.country Vietnam 0.005397 \*\*   
## native.country Yugoslavia 0.345966   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 33851 on 30161 degrees of freedom  
## Residual deviance: 21424 on 30072 degrees of freedom  
## AIC: 21604  
##   
## Number of Fisher Scoring iterations: 13

AIC appears to have risen, which is a sign of a worse fit.

### Prediction the Test Set Results

prob\_pred\_3 = predict(fit\_3, type = 'response', newdata = test[-14])  
y\_pred\_3 = ifelse(prob\_pred\_3 > 0.5, '>50K', '<=50K')  
  
cm\_3 = table(test[, 14], y\_pred\_3)  
cm\_3

## y\_pred\_3  
## <=50K >50K  
## <=50K. 10415 945  
## >50K. 1616 2084

### Computing the Accuracy and Error Rates

acc\_3 = sum(diag(cm\_3)) / sum(cm\_3)  
acc\_3

## [1] 0.8299469

err\_3 = 1 - acc\_3  
err\_3

## [1] 0.1700531

Accuracy rate has decreased to **82.99%** in this case.

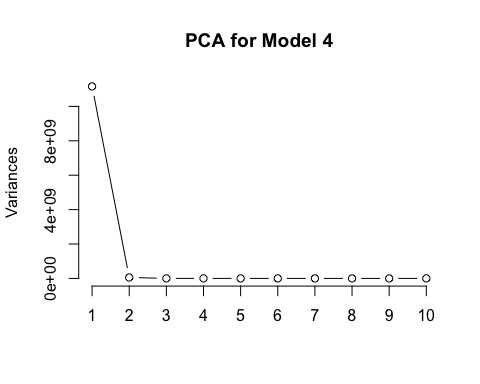
## Model 4

We will try feature transformation in this last model, by means of Principal Component Analysis (PCA). To prepare our datasets for this, we will need to create dummy variables.

# install.packages('dummies')  
library(dummies)

## dummies-1.5.6 provided by Decision Patterns

train\_4 = dummy.data.frame(train, names = c('workclass','education', 'marital','occupation',  
 'relationship','race','sex','native.country'))  
test\_4 = dummy.data.frame(test, names = c('workclass','education', 'marital','occupation',  
 'relationship','race','sex','native.country'))  
  
train\_4\_pca = prcomp(train\_4[-104])  
test\_4\_pca = prcomp(test\_4[-103])  
  
screeplot(train\_4\_pca, type = 'l', main = 'PCA for Model 4')



summary(train\_4\_pca)

## Importance of components:  
## PC1 PC2 PC3 PC4 PC5 PC6  
## Standard deviation 1.057e+05 7.406e+03 404.06991 13.26 11.67 0.8597  
## Proportion of Variance 9.951e-01 4.890e-03 0.00001 0.00 0.00 0.0000  
## Cumulative Proportion 9.951e-01 1.000e+00 1.00000 1.00 1.00 1.0000  
## PC7 PC8 PC9 PC10 PC11 PC12 PC13  
## Standard deviation 0.5582 0.5482 0.4821 0.4571 0.439 0.4211 0.3845  
## Proportion of Variance 0.0000 0.0000 0.0000 0.0000 0.000 0.0000 0.0000  
## Cumulative Proportion 1.0000 1.0000 1.0000 1.0000 1.000 1.0000 1.0000  
## PC14 PC15 PC16 PC17 PC18 PC19 PC20  
## Standard deviation 0.369 0.3481 0.3437 0.3354 0.3145 0.303 0.2918  
## Proportion of Variance 0.000 0.0000 0.0000 0.0000 0.0000 0.000 0.0000  
## Cumulative Proportion 1.000 1.0000 1.0000 1.0000 1.0000 1.000 1.0000  
## PC21 PC22 PC23 PC24 PC25 PC26 PC27  
## Standard deviation 0.2841 0.2703 0.2598 0.2397 0.2318 0.2254 0.2149  
## Proportion of Variance 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000  
## Cumulative Proportion 1.0000 1.0000 1.0000 1.0000 1.0000 1.0000 1.0000  
## PC28 PC29 PC30 PC31 PC32 PC33 PC34  
## Standard deviation 0.2107 0.2083 0.1985 0.1899 0.1886 0.1854 0.1814  
## Proportion of Variance 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000  
## Cumulative Proportion 1.0000 1.0000 1.0000 1.0000 1.0000 1.0000 1.0000  
## PC35 PC36 PC37 PC38 PC39 PC40 PC41  
## Standard deviation 0.1797 0.1776 0.1698 0.1642 0.1619 0.142 0.1404  
## Proportion of Variance 0.0000 0.0000 0.0000 0.0000 0.0000 0.000 0.0000  
## Cumulative Proportion 1.0000 1.0000 1.0000 1.0000 1.0000 1.000 1.0000  
## PC42 PC43 PC44 PC45 PC46 PC47 PC48  
## Standard deviation 0.1293 0.1237 0.1221 0.1178 0.115 0.1116 0.1041  
## Proportion of Variance 0.0000 0.0000 0.0000 0.0000 0.000 0.0000 0.0000  
## Cumulative Proportion 1.0000 1.0000 1.0000 1.0000 1.000 1.0000 1.0000  
## PC49 PC50 PC51 PC52 PC53 PC54  
## Standard deviation 0.09095 0.09028 0.07307 0.07207 0.06837 0.06381  
## Proportion of Variance 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000  
## Cumulative Proportion 1.00000 1.00000 1.00000 1.00000 1.00000 1.00000  
## PC55 PC56 PC57 PC58 PC59 PC60  
## Standard deviation 0.05941 0.05804 0.05754 0.05604 0.05528 0.0541  
## Proportion of Variance 0.00000 0.00000 0.00000 0.00000 0.00000 0.0000  
## Cumulative Proportion 1.00000 1.00000 1.00000 1.00000 1.00000 1.0000  
## PC61 PC62 PC63 PC64 PC65 PC66  
## Standard deviation 0.05193 0.04921 0.04794 0.04681 0.04656 0.04582  
## Proportion of Variance 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000  
## Cumulative Proportion 1.00000 1.00000 1.00000 1.00000 1.00000 1.00000  
## PC67 PC68 PC69 PC70 PC71 PC72  
## Standard deviation 0.04451 0.04386 0.04301 0.0404 0.03941 0.03868  
## Proportion of Variance 0.00000 0.00000 0.00000 0.0000 0.00000 0.00000  
## Cumulative Proportion 1.00000 1.00000 1.00000 1.0000 1.00000 1.00000  
## PC73 PC74 PC75 PC76 PC77 PC78  
## Standard deviation 0.03712 0.03559 0.03334 0.03298 0.03187 0.03117  
## Proportion of Variance 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000  
## Cumulative Proportion 1.00000 1.00000 1.00000 1.00000 1.00000 1.00000  
## PC79 PC80 PC81 PC82 PC83 PC84  
## Standard deviation 0.03024 0.02974 0.02841 0.0273 0.02519 0.02477  
## Proportion of Variance 0.00000 0.00000 0.00000 0.0000 0.00000 0.00000  
## Cumulative Proportion 1.00000 1.00000 1.00000 1.0000 1.00000 1.00000  
## PC85 PC86 PC87 PC88 PC89 PC90  
## Standard deviation 0.02417 0.02372 0.02364 0.02319 0.02266 0.02149  
## Proportion of Variance 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000  
## Cumulative Proportion 1.00000 1.00000 1.00000 1.00000 1.00000 1.00000  
## PC91 PC92 PC93 PC94 PC95 PC96  
## Standard deviation 0.02079 0.02 0.01917 0.01781 0.005818 1.05e-11  
## Proportion of Variance 0.00000 0.00 0.00000 0.00000 0.000000 0.00e+00  
## Cumulative Proportion 1.00000 1.00 1.00000 1.00000 1.000000 1.00e+00  
## PC97 PC98 PC99 PC100 PC101  
## Standard deviation 1.05e-11 1.05e-11 1.05e-11 1.05e-11 1.05e-11  
## Proportion of Variance 0.00e+00 0.00e+00 0.00e+00 0.00e+00 0.00e+00  
## Cumulative Proportion 1.00e+00 1.00e+00 1.00e+00 1.00e+00 1.00e+00  
## PC102 PC103  
## Standard deviation 1.05e-11 7.278e-14  
## Proportion of Variance 0.00e+00 0.000e+00  
## Cumulative Proportion 1.00e+00 1.000e+00

The first 2 principal components are sufficient to explain most, if not all of the variation in the variables. So we will use them for the fit.

# Joining PC1 and PC2 columns to the train\_4 dataset  
train\_4\_pca\_df = as.data.frame(train\_4\_pca$x)  
train\_4$PC1 = train\_4\_pca\_df$PC1  
train\_4$PC2 = train\_4\_pca\_df$PC2  
  
# Joining PC1 and PC2 columns to the test\_4 dataset  
test\_4\_pca\_df = as.data.frame(test\_4\_pca$x)  
test\_4$PC1 = test\_4\_pca\_df$PC1  
test\_4$PC2 = test\_4\_pca\_df$PC2  
  
fit\_4 = suppressWarnings(glm(class ~ PC1 + PC2,  
 family = 'binomial',  
 data = train\_4))  
  
summary(fit\_4)

##   
## Call:  
## glm(formula = class ~ PC1 + PC2, family = "binomial", data = train\_4)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -4.9964 -0.6879 -0.6832 -0.6521 1.8616   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -9.670e-01 1.543e-02 -62.689 <2e-16 \*\*\*  
## PC1 -1.888e-07 1.353e-07 -1.395 0.163   
## PC2 3.345e-04 8.916e-06 37.519 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 33851 on 30161 degrees of freedom  
## Residual deviance: 30698 on 30159 degrees of freedom  
## AIC: 30704  
##   
## Number of Fisher Scoring iterations: 6

### Prediction the Test Set Results

prob\_pred\_4 = predict(fit\_4, type = 'response', newdata = test\_4[c(104, 105)])  
y\_pred\_4 = ifelse(prob\_pred\_4 > 0.5, '>50K', '<=50K')  
  
cm\_4 = table(test\_4[, 103], y\_pred\_4)  
cm\_4

## y\_pred\_4  
## <=50K >50K  
## <=50K. 11200 160  
## >50K. 2970 730

### Computing the Accuracy and Error Rates

acc\_4 = sum(diag(cm\_4)) / sum(cm\_4)  
acc\_4

## [1] 0.7921647

err\_4 = 1 - acc\_4  
err\_4

## [1] 0.2078353

Accuracy rate has decreased to **79.22%** for this model, which shows that feature transformation does not improve our results.

## Summary & Conclusion

We have used the following models in our analysis to predict the *class* variable:

* Full Model (fit) - using **all** variables
* Model 1 (fit\_1) - using all **except** *race*
* Model 2 (fit\_2) - using all **except** *race* and *relationship*
* Model 3 (fit\_3) - using all **except** *race*, *capital.gain* and *capital.loss*
* Model 4 (fit\_4) - using 1st 2 components of PCA

We can conclude that **Model 1** produces the best result, with the *race* variable excluded. This model has the highest accuracy rate of **84.77%**.