Project 3

Manav Mandhani, mm58926

Introduction

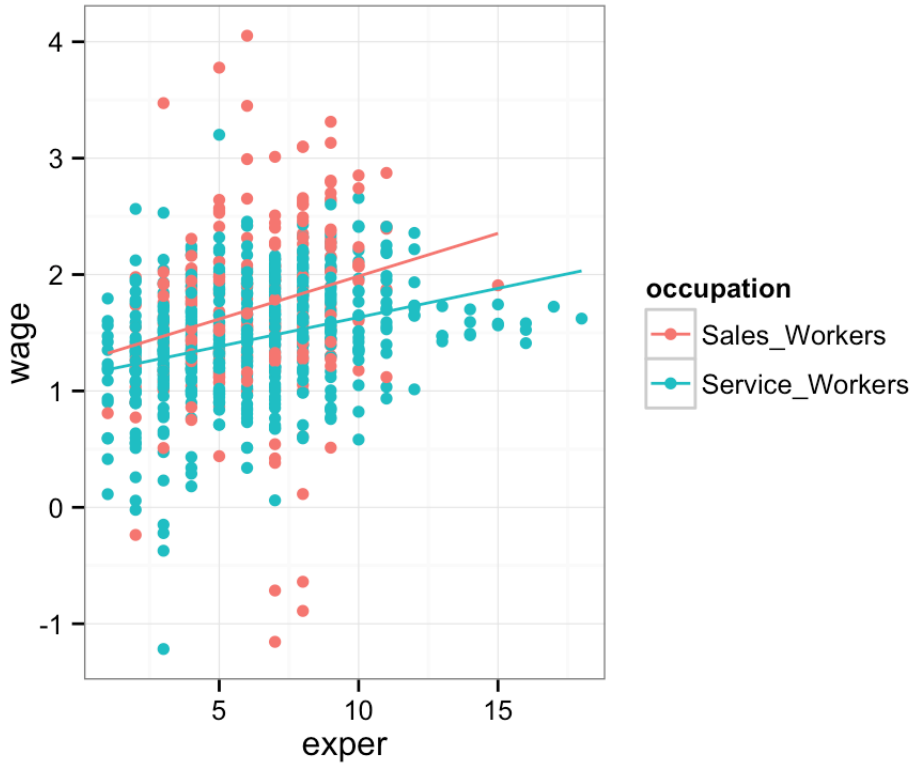
The dataset I will be looking at is the Males dataset which is a part of the ‘plm’ pakcage. The Males dataset contains data on wages and education of young males between 1980 and 1987. The dataset contains 4360 observations and 12 different attributes.

Wage contains the log of hourly wages, occupation, industry, ethn and residence are the categorical variables while all others are continuous.

Question One

Which occupation has the highest room for growth in terms of wages, between Sales and Service workers?

```
# Pick the relevant columns and occupations
Males %>% select(wage, exper, occupation) %>% group_by(occupation) %>% filter(occupation %in% c("Sales_Workers", "Service_Worke
rs"))-> new_males

# Plot experiance against wages and color by occupation. Smooth the points of the scatter plot by forming a line of best fit
ggplot(new_males, aes(x=exper, y=wage, color = occupation)) + geom_point() + geom_smooth(aes(group=occupation), method=lm, s
e=F)
```




```
# Create new dataframes using sales and service workers respectively
Males %>% filter(occupation == "Sales_Workers") -> sales
Males %>% filter(occupation == "Service_Workers") -> service

# Perform linear regressions on those frames
fit_sales<-lm(wage~exper, data=sales)
fit_service<-lm(wage~exper, data=service)

# Outputting summary of the results
summary(fit_sales)
```

```
##
## Call:
## lm(formula = wage ~ exper, data = sales)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.92015 -0.38426 -0.01143  0.36997  2.36048
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.24989    0.13814   9.048  < 2e-16 ***
## exper        0.07358    0.02032   3.622 0.000359 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7056 on 231 degrees of freedom
## Multiple R-squared:  0.05374,    Adjusted R-squared:  0.04964
## F-statistic: 13.12 on 1 and 231 DF,  p-value: 0.0003594
```

```
summary(fit_service)
```

```
##
## Call:
## lm(formula = wage ~ exper, data = service)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.49925 -0.28061 -0.00371  0.31868  1.81833
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.132729   0.047545  23.824  < 2e-16 ***
## exper       0.049927   0.006887   7.249 1.58e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4832 on 507 degrees of freedom
## Multiple R-squared:  0.09391,    Adjusted R-squared:  0.09212
## F-statistic: 52.55 on 1 and 507 DF,  p-value: 1.577e-12
```

To solve this question, I plotted the wages of sales and service workers over time on a scatter plot. The blue and red dots had various different wages for different years of experience which made any trend very hard to discern using the points. So, the points were smoothed to provide a linear visualization of all the data points. This showed a slightly higher slope for sales workers over service workers.

Then, I ran two linear regressions that model the wage against experience for the sales workers and service workers separately. The regressions were performed to determine the slope of the smoothing lines found in the plots above. The p-values for both these regressions were low enough to be considered signifianct. Since the slope of the sales workers (0.07358) was higher than that of service workers (0.049927), it can be concluded that sales workers have a slightly higher room for growth than service workers over time.

Question Two

Can the attributes provided in the Males dataset be used to predict whether a person is married or not?

```r
# Logistical regression on all attribtues
glm.out <- glm(married ~ nr + year + school + exper + union + ethn + health + occupation + wage + industry + residence,
               data = Males,
               family = binomial)
summary(glm.out)
```

```
##
## Call:
## glm(formula = married ~ nr + year + school + exper + union +
##     ethn + health + occupation + wage + industry + residence,
##     family = binomial, data = Males)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0983  -0.9375  -0.5545   1.0577   2.3141
##
## Coefficients:
##                                                Estimate Std. Error
## (Intercept)                                  -2.023e+02  7.241e+01
## nr                                            2.772e-05  1.307e-05
## year                                          9.980e-02  3.666e-02
## school                                        1.524e-01  3.214e-02
## exper                                         1.754e-01  3.289e-02
## unionyes                                      2.053e-01  1.010e-01
## ethnblack                                    -9.336e-01  1.384e-01
## ethnhisp                                      8.560e-02  1.413e-01
## healthyes                                     2.843e-01  3.230e-01
## occupationManagers, Officials_and_Proprietors 8.772e-02  1.826e-01
## occupationSales_Workers                      -2.225e-01  2.219e-01
## occupationClerical_and_kindred               -4.829e-02  1.756e-01
## occupationCraftsmen, Foremen_and_kindred      6.112e-01  1.626e-01
## occupationOperatives_and_kindred              1.812e-01  1.703e-01
## occupationLaborers_and_farmers                5.456e-02  1.951e-01
## occupationFarm_Laborers_and_Foreman           6.314e-01  4.432e-01
```

```
## occupationService_Workers                       -3.057e-01  1.859e-01
## wage                                             4.711e-01  9.443e-02
## industryMining                                  -4.863e-01  4.631e-01
## industryConstruction                            -3.765e-01  3.289e-01
## industryTrade                                   -2.876e-01  2.992e-01
## industryTransportation                          -3.691e-01  3.228e-01
## industryFinance                                 -1.077e-01  3.632e-01
## industryBusiness_and_Repair_Service             -5.746e-01  3.243e-01
## industryPersonal_Service                        -5.225e-01  4.327e-01
## industryEntertainment                           -1.823e+00  5.740e-01
## industryManufacturing                           -2.720e-01  2.985e-01
## industryProfessional_and_Related Service        -2.015e-01  3.268e-01
## industryPublic_Administration                    1.898e-01  3.499e-01
## residencenorth_east                              4.127e-02  2.642e-01
## residencenothern_central                         4.007e-01  2.604e-01
## residencesouth                                   4.918e-01  2.528e-01
##                                                 z value Pr(>|z|)
## (Intercept)                                      -2.794 0.005210 **
## nr                                                2.120 0.033984 *
## year                                              2.722 0.006485 **
## school                                            4.741 2.13e-06 ***
## exper                                             5.334 9.60e-08 ***
## unionyes                                          2.033 0.042042 *
## ethnblack                                        -6.748 1.50e-11 ***
## ethnhisp                                          0.606 0.544735
## healthyes                                         0.880 0.378774
## occupationManagers, Officials_and_Proprietors     0.480 0.631009
## occupationSales_Workers                          -1.003 0.316000
## occupationClerical_and_kindred                   -0.275 0.783285
## occupationCraftsmen, Foremen_and_kindred          3.759 0.000171 ***
## occupationOperatives_and_kindred                  1.064 0.287266
## occupationLaborers_and_farmers                    0.280 0.779685
## occupationFarm_Laborers_and_Foreman               1.425 0.154211
## occupationService_Workers                        -1.644 0.100122
## wage                                              4.989 6.06e-07 ***
```

```
## industryMining                                  -1.050 0.293728
## industryConstruction                            -1.145 0.252275
## industryTrade                                   -0.961 0.336470
## industryTransportation                          -1.143 0.252938
## industryFinance                                 -0.297 0.766847
## industryBusiness_and_Repair_Service             -1.772 0.076403 .
## industryPersonal_Service                        -1.207 0.227265
## industryEntertainment                           -3.176 0.001495 **
## industryManufacturing                           -0.911 0.362163
## industryProfessional_and_Related Service        -0.617 0.537398
## industryPublic_Administration                    0.542 0.587550
## residencenorth_east                              0.156 0.875838
## residencenothern_central                         1.539 0.123874
## residencesouth                                   1.945 0.051739 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4237.4  on 3114  degrees of freedom
## Residual deviance: 3676.1  on 3083  degrees of freedom
##   (1245 observations deleted due to missingness)
## AIC: 3740.1
##
## Number of Fisher Scoring iterations: 4
```

```
# Remove residence due to it's high p-value
glm.out <- glm(married ~ nr + year + school + exper + union + ethn + health + occupation + wage + industry,
               data = Males,
               family = binomial)
summary(glm.out)
```

```
##
## Call:
```

```
## glm(formula = married ~ nr + year + school + exper + union +
##     ethn + health + occupation + wage + industry, family = binomial,
##     data = Males)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0309  -0.9765  -0.5568   1.0551   2.6243
##
## Coefficients:
##                                                Estimate Std. Error
## (Intercept)                                   -5.641e+01  5.657e+01
## nr                                             2.813e-05  1.033e-05
## year                                           2.658e-02  2.866e-02
## school                                         1.328e-01  2.593e-02
## exper                                          2.199e-01  2.549e-02
## unionyes                                       1.789e-01  8.349e-02
## ethnblack                                     -1.088e+00  1.203e-01
## ethnhisp                                      -1.760e-01  1.015e-01
## healthyes                                     -1.998e-01  2.667e-01
## occupationManagers, Officials_and_Proprietors  8.832e-02  1.560e-01
## occupationSales_Workers                       -2.735e-01  1.864e-01
## occupationClerical_and_kindred                -1.126e-01  1.518e-01
## occupationCraftsmen, Foremen_and_kindred       5.341e-01  1.375e-01
## occupationOperatives_and_kindred               1.483e-01  1.427e-01
## occupationLaborers_and_farmers                -4.956e-02  1.649e-01
## occupationFarm_Laborers_and_Foreman            4.345e-01  3.651e-01
## occupationService_Workers                     -3.168e-01  1.593e-01
## wage                                           4.345e-01  7.693e-02
## industryMining                                -1.390e-01  3.587e-01
## industryConstruction                          -3.027e-01  2.665e-01
## industryTrade                                 -5.177e-01  2.474e-01
## industryTransportation                        -3.116e-01  2.706e-01
## industryFinance                               -2.890e-01  2.999e-01
## industryBusiness_and_Repair_Service           -7.127e-01  2.681e-01
## industryPersonal_Service                      -5.331e-01  3.659e-01
```

```
## industryEntertainment                          -2.006e+00  4.620e-01
## industryManufacturing                          -4.676e-01  2.461e-01
## industryProfessional_and_Related Service       -5.617e-01  2.748e-01
## industryPublic_Administration                  -2.692e-02  2.930e-01
##                                                z value Pr(>|z|)
## (Intercept)                                     -0.997 0.318674
## nr                                               2.722 0.006494 **
## year                                             0.928 0.353556
## school                                           5.122 3.03e-07 ***
## exper                                            8.625  < 2e-16 ***
## unionyes                                         2.142 0.032181 *
## ethnblack                                       -9.042  < 2e-16 ***
## ethnhisp                                        -1.734 0.082932 .
## healthyes                                       -0.749 0.453650
## occupationManagers, Officials_and_Proprietors    0.566 0.571267
## occupationSales_Workers                         -1.467 0.142425
## occupationClerical_and_kindred                  -0.742 0.458305
## occupationCraftsmen, Foremen_and_kindred         3.884 0.000103 ***
## occupationOperatives_and_kindred                 1.039 0.298774
## occupationLaborers_and_farmers                  -0.301 0.763713
## occupationFarm_Laborers_and_Foreman              1.190 0.233972
## occupationService_Workers                       -1.989 0.046753 *
## wage                                             5.647 1.63e-08 ***
## industryMining                                  -0.388 0.698270
## industryConstruction                            -1.136 0.256082
## industryTrade                                   -2.093 0.036341 *
## industryTransportation                          -1.152 0.249526
## industryFinance                                 -0.964 0.335171
## industryBusiness_and_Repair_Service             -2.658 0.007854 **
## industryPersonal_Service                        -1.457 0.145168
## industryEntertainment                           -4.342 1.41e-05 ***
## industryManufacturing                           -1.900 0.057440 .
## industryProfessional_and_Related Service        -2.044 0.040944 *
## industryPublic_Administration                   -0.092 0.926795
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 5979.2  on 4359  degrees of freedom
## Residual deviance: 5234.6  on 4331  degrees of freedom
## AIC: 5292.6
##
## Number of Fisher Scoring iterations: 4
```

```
# Remove health due to it's high p-value
glm.out <- glm(married ~ nr + year + school + exper + union + ethn + occupation + industry + wage,
               data = Males,
               family = binomial)
summary(glm.out)
```

```
##
## Call:
## glm(formula = married ~ nr + year + school + exper + union +
##     ethn + occupation + industry + wage, family = binomial, data = Males)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0514  -0.9756  -0.5562   1.0554   2.6253
##
## Coefficients:
##                                          Estimate Std. Error
## (Intercept)                            -5.728e+01  5.655e+01
## nr                                      2.817e-05  1.033e-05
## year                                    2.702e-02  2.865e-02
## school                                  1.329e-01  2.593e-02
## exper                                   2.196e-01  2.549e-02
## unionyes                                1.817e-01  8.341e-02
## ethnblack                              -1.086e+00  1.203e-01
```

```
## ethnhisp                                       -1.745e-01  1.015e-01
## occupationManagers, Officials_and_Proprietors   8.718e-02  1.560e-01
## occupationSales_Workers                        -2.766e-01  1.864e-01
## occupationClerical_and_kindred                 -1.179e-01  1.517e-01
## occupationCraftsmen, Foremen_and_kindred        5.308e-01  1.374e-01
## occupationOperatives_and_kindred                1.432e-01  1.425e-01
## occupationLaborers_and_farmers                 -5.309e-02  1.648e-01
## occupationFarm_Laborers_and_Foreman             4.360e-01  3.651e-01
## occupationService_Workers                      -3.198e-01  1.593e-01
## industryMining                                 -1.361e-01  3.587e-01
## industryConstruction                           -2.958e-01  2.663e-01
## industryTrade                                  -5.124e-01  2.472e-01
## industryTransportation                         -3.079e-01  2.705e-01
## industryFinance                                -2.843e-01  2.997e-01
## industryBusiness_and_Repair_Service            -7.132e-01  2.680e-01
## industryPersonal_Service                       -5.303e-01  3.657e-01
## industryEntertainment                          -2.002e+00  4.619e-01
## industryManufacturing                          -4.635e-01  2.460e-01
## industryProfessional_and_Related Service       -5.581e-01  2.747e-01
## industryPublic_Administration                  -2.305e-02  2.928e-01
## wage                                            4.353e-01  7.694e-02
##                                                z value Pr(>|z|)
## (Intercept)                                     -1.013 0.311152
## nr                                               2.726 0.006406 **
## year                                             0.943 0.345614
## school                                           5.124 2.99e-07 ***
## exper                                            8.617  < 2e-16 ***
## unionyes                                         2.178 0.029393 *
## ethnblack                                       -9.030  < 2e-16 ***
## ethnhisp                                        -1.720 0.085383 .
## occupationManagers, Officials_and_Proprietors    0.559 0.576255
## occupationSales_Workers                         -1.484 0.137772
## occupationClerical_and_kindred                  -0.777 0.437067
## occupationCraftsmen, Foremen_and_kindred         3.862 0.000112 ***
## occupationOperatives_and_kindred                 1.004 0.315156
```

```
## occupationLaborers_and_farmers                 -0.322 0.747328
## occupationFarm_Laborers_and_Foreman             1.194 0.232343
## occupationService_Workers                      -2.008 0.044683 *
## industryMining                                 -0.379 0.704375
## industryConstruction                           -1.111 0.266551
## industryTrade                                  -2.073 0.038160 *
## industryTransportation                         -1.139 0.254887
## industryFinance                                -0.949 0.342828
## industryBusiness_and_Repair_Service            -2.661 0.007792 **
## industryPersonal_Service                       -1.450 0.147006
## industryEntertainment                          -4.334 1.46e-05 ***
## industryManufacturing                          -1.884 0.059500 .
## industryProfessional_and_Related Service       -2.032 0.042200 *
## industryPublic_Administration                  -0.079 0.937259
## wage                                            5.657 1.54e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 5979.2  on 4359  degrees of freedom
## Residual deviance: 5235.2  on 4332  degrees of freedom
## AIC: 5291.2
##
## Number of Fisher Scoring iterations: 4
```

```
# Remove occupation as most of it's categories have a high p-value
glm.out <- glm(married ~ nr + year + school + exper + union + ethn + industry + wage,
               data = Males,
               family = binomial)
summary(glm.out)
```

```
##
## Call:
```

```
## glm(formula = married ~ nr + year + school + exper + union +
##     ethn + industry + wage, family = binomial, data = Males)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9345  -0.9814  -0.5749   1.0642   2.5579
##
## Coefficients:
##                                        Estimate Std. Error z value
## (Intercept)                           -4.813e+01  5.603e+01  -0.859
## nr                                     2.748e-05  1.023e-05   2.686
## year                                   2.260e-02  2.839e-02   0.796
## school                                 1.162e-01  2.496e-02   4.654
## exper                                  2.235e-01  2.532e-02   8.827
## unionyes                               1.609e-01  8.055e-02   1.997
## ethnblack                             -1.134e+00  1.191e-01  -9.521
## ethnhisp                              -1.951e-01  1.000e-01  -1.951
## industryMining                        -2.092e-01  3.239e-01  -0.646
## industryConstruction                  -2.694e-01  2.187e-01  -1.231
## industryTrade                         -7.553e-01  1.933e-01  -3.907
## industryTransportation                -4.365e-01  2.259e-01  -1.932
## industryFinance                       -5.941e-01  2.529e-01  -2.349
## industryBusiness_and_Repair_Service   -7.726e-01  2.194e-01  -3.521
## industryPersonal_Service              -8.815e-01  3.254e-01  -2.709
## industryEntertainment                 -2.318e+00  4.334e-01  -5.348
## industryManufacturing                 -5.205e-01  1.947e-01  -2.673
## industryProfessional_and_Related Service -8.377e-01  2.223e-01  -3.768
## industryPublic_Administration         -2.790e-01  2.470e-01  -1.130
## wage                                   4.576e-01  7.555e-02   6.056
##                                       Pr(>|z|)
## (Intercept)                           0.390381
## nr                                    0.007236 **
## year                                  0.425945
## school                                3.25e-06 ***
## exper                                  < 2e-16 ***
```

```
## unionyes                                  0.045785 *
## ethnblack                                  < 2e-16 ***
## ethnhisp                                  0.051111 .
## industryMining                            0.518403
## industryConstruction                      0.218162
## industryTrade                             9.35e-05 ***
## industryTransportation                    0.053369 .
## industryFinance                           0.018820 *
## industryBusiness_and_Repair_Service       0.000430 ***
## industryPersonal_Service                  0.006745 **
## industryEntertainment                     8.91e-08 ***
## industryManufacturing                     0.007510 **
## industryProfessional_and_Related Service 0.000165 ***
## industryPublic_Administration             0.258592
## wage                                      1.39e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 5979.2  on 4359  degrees of freedom
## Residual deviance: 5292.0  on 4340  degrees of freedom
## AIC: 5332
##
## Number of Fisher Scoring iterations: 4
```

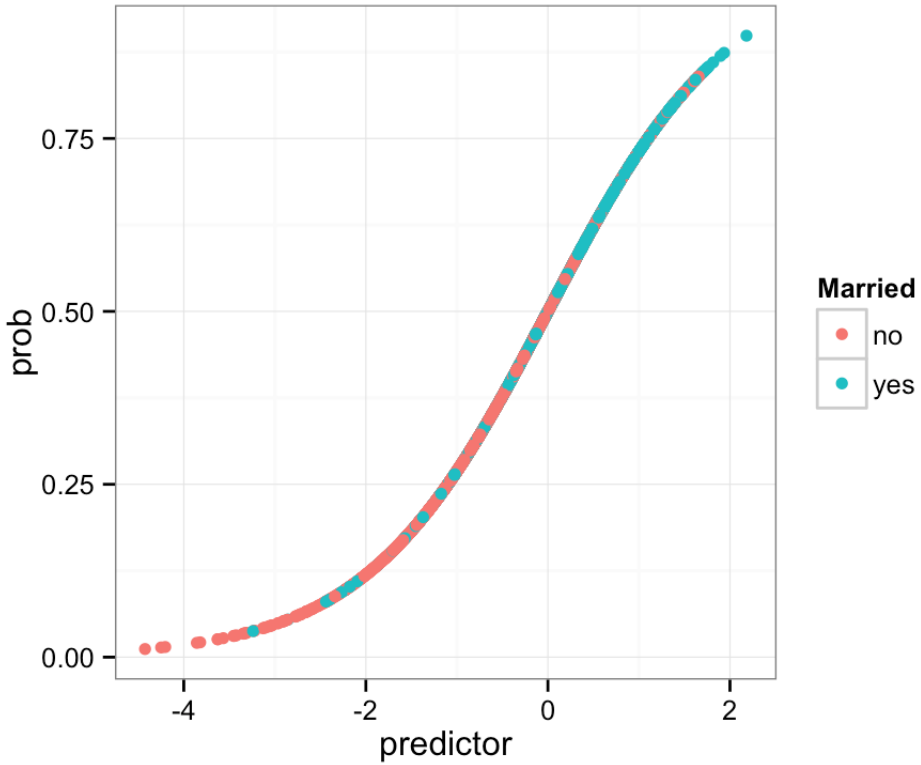
```
#Remove year due to it's high p-value
glm.out <- glm(married ~ nr + school + exper + union + ethn + industry + wage,
               data = Males,
               family = binomial)
summary(glm.out)
```

```
##
## Call:
```

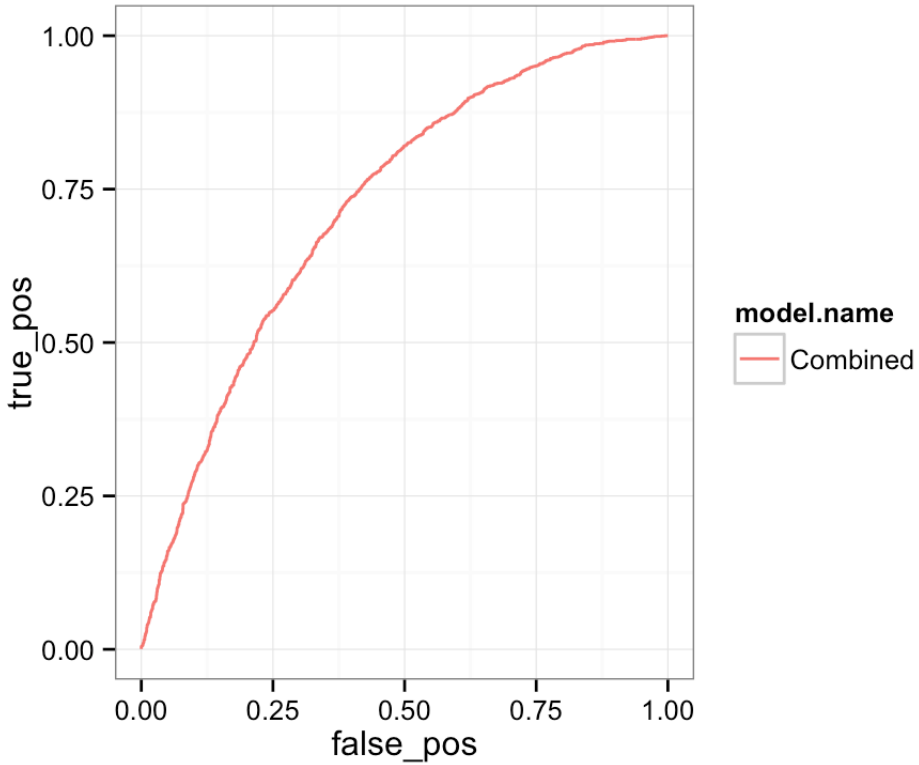
```
## glm(formula = married ~ nr + school + exper + union + ethn +
##     industry + wage, family = binomial, data = Males)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9393  -0.9828  -0.5758   1.0666   2.5602
##
## Coefficients:
##                                          Estimate Std. Error z value
## (Intercept)                             -3.519e+00  3.426e-01 -10.271
## nr                                       2.686e-05  1.020e-05   2.634
## school                                   1.247e-01  2.254e-02   5.534
## exper                                    2.405e-01  1.381e-02  17.413
## unionyes                                 1.573e-01  8.040e-02   1.957
## ethnblack                               -1.137e+00  1.190e-01  -9.555
## ethnhisp                                -1.944e-01  1.001e-01  -1.943
## industryMining                          -2.077e-01  3.239e-01  -0.641
## industryConstruction                    -2.650e-01  2.188e-01  -1.211
## industryTrade                           -7.467e-01  1.932e-01  -3.866
## industryTransportation                  -4.262e-01  2.256e-01  -1.889
## industryFinance                         -5.773e-01  2.521e-01  -2.290
## industryBusiness_and_Repair_Service     -7.602e-01  2.190e-01  -3.472
## industryPersonal_Service                -8.636e-01  3.247e-01  -2.660
## industryEntertainment                   -2.305e+00  4.334e-01  -5.318
## industryManufacturing                   -5.111e-01  1.945e-01  -2.628
## industryProfessional_and_Related Service -8.233e-01  2.217e-01  -3.714
## industryPublic_Administration           -2.641e-01  2.463e-01  -1.072
## wage                                     4.615e-01  7.542e-02   6.119
##                                         Pr(>|z|)
## (Intercept)                              < 2e-16 ***
## nr                                      0.008449 **
## school                                  3.14e-08 ***
## exper                                    < 2e-16 ***
## unionyes                                0.050366 .
## ethnblack                                < 2e-16 ***
```

```
## ethnhisp                                  0.052004 .
## industryMining                            0.521424
## industryConstruction                      0.225811
## industryTrade                             0.000111 ***
## industryTransportation                    0.058908 .
## industryFinance                           0.022003 *
## industryBusiness_and_Repair_Service       0.000517 ***
## industryPersonal_Service                  0.007810 **
## industryEntertainment                     1.05e-07 ***
## industryManufacturing                     0.008583 **
## industryProfessional_and_Related Service 0.000204 ***
## industryPublic_Administration             0.283734
## wage                                      9.42e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 5979.2  on 4359  degrees of freedom
## Residual deviance: 5292.6  on 4341  degrees of freedom
## AIC: 5330.6
##
## Number of Fisher Scoring iterations: 4
```

```
# Convert into data frame using predictors and probabilities
lr_data <- data.frame(predictor=glm.out$linear.predictors, prob=glm.out$fitted.values, Married=Males$married)

# Plot the data using a scatter plot
ggplot(lr_data, aes(x=predictor, y=prob, color=Married)) + geom_point()
```


0.75
0.50
0.25
0.00
prob
-4
-2
0
2
predictor
Married
no
yes

```r
# Function to calculate ROC values
calc_ROC <- function(probabilities, known_truth, model.name=NULL)
  {
  outcome <- as.numeric(factor(known_truth))-1
  pos <- sum(outcome) # total known positives
  neg <- sum(1-outcome) # total known negatives
  pos_probs <- outcome*probabilities # probabilities for known positives
  neg_probs <- (1-outcome)*probabilities # probabilities for known negatives
  true_pos <- sapply(probabilities,
                     function(x) sum(pos_probs>=x)/pos) # true pos. rate
  false_pos <- sapply(probabilities,
                     function(x) sum(neg_probs>=x)/neg)
  if (is.null(model.name))
    result <- data.frame(true_pos, false_pos)
  else
    result <- data.frame(true_pos, false_pos, model.name)
  result %>% arrange(false_pos, true_pos)
  }

ROC <- calc_ROC(probabilities=glm.out$fitted.values, # predicted probabilities
                known_truth=Males$married,      # the known truth, i.e., true species assignment
                model.name="Combined")

# Plot the ROC curve to interpret how good the model is
ggplot(data=NULL, aes(x=false_pos, y=true_pos)) +
  geom_line(data=ROC, aes(color=model.name))
```




```
# Calculate the value of AUC
ROC1 %>% mutate(delta=false_pos-lag(false_pos)) %>%
  summarize(AUC=sum(delta*true_pos, na.rm=T))
```

```
##         AUC
## 1 0.8463542
```

First, I performed a logsitical regression on all the attributes provided in the dataset against marriage. Logistical regression can be used to predict binary outputs using categorical and continuous variables which is useful for our task. Using the process of backwards selection, I was able to remove certain attributes until I was left with the following attribtues: nr, school, exper, union, ethn, industry, wage. Although some of the categories in the industry column were not statistically significant, a majority of them were which is why I decided to hold on to the variable.

After doing the regression, I plotted the data using a logistical curve. Although the data was not perfectly sergregated, the curve mostly had blue dots after a predictor unit of 0 and red dots before. The points seemed to separate at a probability value of ~ 0.6. This shows a decent predictive model of whether a person is married or not.

To understand how good the model is, I plotted a ROC curve and calculated the Area under the curve, which yielded a value of 0.7228822. This isn't an ideal number of 1 but also does not indicate a random chance. These results lead me to believe that this model is able to predict whether a person is married or not with a decent accuracy but is far from being an ideal model for prediction.