STOR 455 Homework #6

20 points - Due Tuesday 04/20 at 12:00pm

Theory Part

1. Suppose the log odds is 0, what is the probability of sucess?

Your answer: If the log odds is 0, then the odds is equal to 1 (since odds = e^(log odds)). The odds is defined as the ratio of the probability of success to the probability of failure. So, if the odds is 1, then the probability of success is equal to the probability of failure. Therefore, the probability of success in this case would be 0.5 (or 50%).

2. Suppose we included a categorical predictor with 5 categories in a logistic regression. What is the degrees of freedom of the drop-in-deviance test when we test the effectiveness of this categorical predictor?

Your answer: The degrees of freedom for the reference model is the total sample size minus 1 (df = n - 1), and the degrees of freedom for the full model is the total sample size minus the number of parameters estimated in the full model. In this case, the full model will have 5 parameters (one intercept and 4 dummy variables), so the degrees of freedom for the full model is n - 5. Therefore, the degrees of freedom for the drop-in-deviance test is: df = (n - 5) - (n - 1) = 4 So, the drop-in-deviance test has 4 degrees of freedom when testing the effectiveness of a categorical predictor with 5 categories in logistic regression.

Computing Part

Are Emily and Greg More Employable Than Lakisha and Jamal?

Bertrand, M., & Mullainathan, S. (2004). Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. *American Economic Review, 94*(4), pp. 991-1013.

Abstract

We perform a field experiment to measure racial discrimination in the labor market. We respond with fictitious resumes to help-wanted ads in Boston and Chicago newspapers. To manipulate perception of race, each resume is randomly assigned either a very African American sounding name or a very White sounding name. The results show significant discrimination against African-American names: White names receive 50 percent more callbacks for interviews. We also find that race affects the benefits of a better resume. For White names, a higher quality resume elicits 30 percent more callbacks whereas for African Americans, it elicits a far smaller increase. Applicants living in better neighborhoods receive more callbacks but, interestingly, this effect does not differ by race. The amount of

discrimination is uniform across occupations and industries. Federal contractors and employers who list �諄qual Opportunity Employer�� in their ad discriminate as much as other employers. We find little evidence that our results are driven by employers inferring something other than race, such as social class, from the names. These results suggest that racial discrimination is still a prominent feature of the labor market.

Variables	Descriptions
call	Was the applicant called back? (1 = yes; 0 = no)
ethnicity	indicating ethnicity ("Caucasian-sounding" vs. "African-American sounding" first name)
sex	indicating sex
quality	Indicating quality of resume.
experience	Number of years of work experience on the resume
equal	Is the employer EOE (equal opportunity employment)?

Use the *ResumeNames455* data on Sakai under "Resources/Data."

1) Construct a logistic model to predict if the job applicant was called back using *experience* as the predictor variable.

```
library(Stat2Data)
library(readr)
library(bestglm)

## Warning: 程辑包'bestglm'是用R版本4.2.3 来建造的

## 载入需要的程辑包：leaps

source("ShowSubsets.R")
data = read_csv("ResumeNames455.csv")

## Rows: 4870 Columns: 7

## ── Column specification
────────────────────────────────────────────────────────

## Delimiter: ","
## chr (5): name, sex, ethnicity, quality, equal
## dbl (2): call, experience
##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this
message.

callback_mod = glm(call ~ experience, data = data, family = binomial)
summary(callback_mod)

##
## Call:
## glm(formula = call ~ experience, family = binomial, data = data)
```

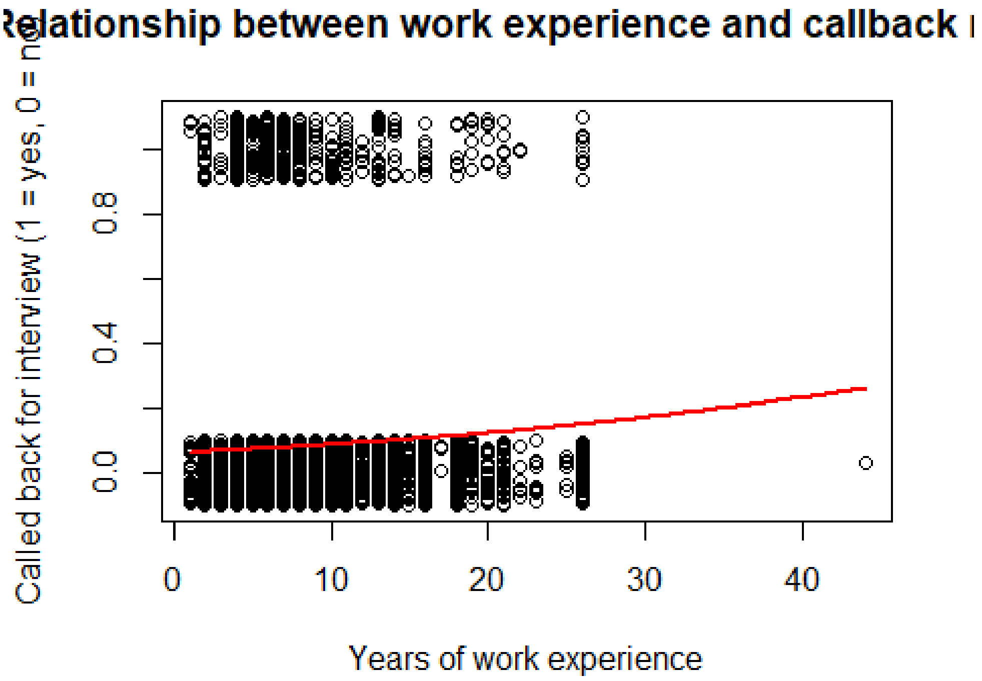
```
## 
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -0.7780  -0.4075  -0.3924  -0.3779   2.3598  
## 
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)    
## (Intercept) -2.75960    0.09620 -28.687  < 2e-16 ***
## experience   0.03908    0.00918   4.257 2.07e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for binomial family taken to be 1)
## 
##     Null deviance: 2726.9  on 4869  degrees of freedom
## Residual deviance: 2710.2  on 4868  degrees of freedom
## AIC: 2714.2
## 
## Number of Fisher Scoring iterations: 5
```

2) Plot the raw data and the sigmoid curve on the same axes.

```
sigmoid = function(B0, B1, x) {
  exp(B0 + B1 * x) / (1 + exp(B0 + B1 * x))
}

B0 = summary(callback_mod)$coef[1]
B1 = summary(callback_mod)$coef[2]

plot(jitter(call, amount = 0.1) ~ experience, data = data,
     xlab = "Years of work experience",
     ylab = "Called back for interview (1 = yes, 0 = no)",
     main = "Relationship between work experience and callback rates")
curve(sigmoid(B0, B1, x), add = TRUE, col = "red", lwd = 2)
```



3) For an applicant with 6 years of experience, what does your model predict is the probability of this applicant getting called back?

```
newdata = data.frame(experience = 6)
prob_callback = predict(callback_mod, newdata, type = "response")
prob_callback

##          1
## 0.07411543

"The predicted probability of an applicant with 6 years of experience getting
called back is approximately 0.074"

## [1] "The predicted probability of an applicant with 6 years of experience
getting called back is approximately 0.074"
```

4) Use the model from question #1 to perform a hypothesis test to determine if there is significant evidence of a relationship between *call* and *experience*. Cite your hypotheses, p-value, and conclusion in context.

```
summary(callback_mod)

##
## Call:
## glm(formula = call ~ experience, family = binomial, data = data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.7780  -0.4075  -0.3924  -0.3779   2.3598
```

```
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.75960    0.09620 -28.687  < 2e-16 ***
## experience   0.03908    0.00918   4.257 2.07e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2726.9  on 4869  degrees of freedom
## Residual deviance: 2710.2  on 4868  degrees of freedom
## AIC: 2714.2
##
## Number of Fisher Scoring iterations: 5
```

```
"H0:There is no significant relationship between experience and the
likelihood of being called back for a job.
Ha:There is a significant relationship between experience and the likelihood
of being called back for a job.
The p-value for the coefficient of experience is 2.07e-05, which is less than
the significance level of 0.05. Therefore, we reject the null hypothesis and
conclude that there is strong evidence of a significant relationship between
experience and the likelihood of being called back for a job. In other words,
the experience of a job applicant is a significant predictor of whether they
will be called back for a job interview."
```

```
## [1] "H0:There is no significant relationship between experience and the
likelihood of being called back for a job.\nHa:There is a significant
relationship between experience and the likelihood of being called back for a
job.\nThe p-value for the coefficient of experience is 2.07e-05, which is
less than the significance level of 0.05. Therefore, we reject the null
hypothesis and conclude that there is strong evidence of a significant
relationship between experience and the likelihood of being called back for a
job. In other words, the experience of a job applicant is a significant
predictor of whether they will be called back for a job interview."
```

5) Construct a confidence interval for the odds ratio for your model and include a sentence interpreting the interval in the context.

```
exp(confint.default(callback_mod))
```

```
##                  2.5 %     97.5 %
## (Intercept) 0.05243672 0.07645446
## experience  1.02131169 1.05873170
```

```
"The resulting confidence interval is (1.021, 1.058), which means that we are
95% confident that the odds ratio for a one-unit increase in experience lies
between 1.021 and 1.058. This suggests that for each additional year of
experience, the odds of being called back increase by a factor of 1.021 to
1.058 times, after controlling for the effect of other variables in the
model."
```

```
## [1] "The resulting confidence interval is (1.021, 1.058), which means that
we are 95% confident that the odds ratio for a one-unit increase in
experience lies between 1.021 and 1.058. This suggests that for each
additional year of experience, the odds of being called back increase by a
factor of 1.021 to 1.058 times, after controlling for the effect of other
variables in the model."
```

6) Does the number of years of work experience impact the relationship between *ethnicity*, *sex*, and an applicant getting called back? Construct a logistic model to predict if the job applicant was called back using *ethnicity*, *sex*, *experience*, and the interactions between *ethnicity* and *experience*, and *sex* and *experience* as the predictor variables.

```
callback_mod_interaction = glm(call ~ ethnicity*experience + sex*experience,
data = data, family = binomial)
summary(callback_mod_interaction)

##
## Call:
## glm(formula = call ~ ethnicity * experience + sex * experience,
##     family = binomial, data = data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.8867  -0.4320  -0.3941  -0.3458   2.4913
##
## Coefficients:
##                             Estimate Std. Error z value Pr(>|z|)
## (Intercept)                -3.111554   0.159022 -19.567  < 2e-16 ***
## ethnicitycauc               0.497435   0.196458   2.532  0.01134 *
## experience                  0.054109   0.014646   3.694  0.00022 ***
## sexmale                     0.351931   0.230841   1.525  0.12737
## ethnicitycauc:experience -0.006006   0.018719  -0.321  0.74831
## experience:sexmale       -0.057080   0.024796  -2.302  0.02133 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2726.9  on 4869  degrees of freedom
## Residual deviance: 2686.4  on 4864  degrees of freedom
## AIC: 2698.4
##
## Number of Fisher Scoring iterations: 5
```

```
"The coefficient for the interaction term between ethnicity and experience is
not significant (p-value = 0.74831), while the coefficient for the
interaction term between sex and experience is significant (p-value =
0.02133). This suggests that the number of years of work experience does not
have a significant impact on the relationship between ethnicity and getting
called back, but it does have a significant impact on the relationship
```

```
between sex and getting called back. Specifically, the odds of being called
back decrease by a factor of exp(-0.057080) = 0.944 for each additional year
of work experience for males, holding all other variables constant."
```

```
## [1] "The coefficient for the interaction term between ethnicity and
experience is not significant (p-value = 0.74831), while the coefficient for
the interaction term between sex and experience is significant (p-value =
0.02133). This suggests that the number of years of work experience does not
have a significant impact on the relationship between ethnicity and getting
called back, but it does have a significant impact on the relationship
between sex and getting called back. Specifically, the odds of being called
back decrease by a factor of exp(-0.057080) = 0.944 for each additional year
of work experience for males, holding all other variables constant."
```

7) Conduct a drop in deviance hypothesis test to determine the effectiveness of the *experience* terms in the model constructed in the previous question. Cite your hypotheses, p-value, and conclusion in context.

```
callback_mod_interaction = glm(call ~ ethnicity*experience + sex*experience,
data = data, family = binomial)
callback_mod_reduced = glm(call ~ ethnicity + sex, data = data, family =
binomial)
G = summary(callback_mod_reduced)$deviance -
summary(callback_mod_interaction)$deviance
p_value = 1 - pchisq(G, 2)
p_value
```

```
## [1] 1.25441e-05
```

```
"H0: The reduced model without experience terms is not significantly worse
than the full model with experience terms.
Ha: The full model with experience terms is significantly better than the
reduced model without experience terms.
With a p-value of  1.25441e-05, we fail to reject the null hypothesis and
conclude that the full model with interactions is significantly better than
the reduced model without interactions. Therefore, the experience terms are
effective in explaining the variation in the callback outcome."
```

```
## [1] "H0: The reduced model without experience terms is not significantly
worse than the full model with experience terms.\nHa: The full model with
experience terms is significantly better than the reduced model without
experience terms.\nWith a p-value of  1.25441e-05, we fail to reject the null
hypothesis and conclude that the full model with interactions is
significantly better than the reduced model without interactions. Therefore,
the experience terms are effective in explaining the variation in the
callback outcome."
```

8) Use an appropriate model selection method to construct a best model to predict if the job applicant was called back using any of the variables as predictors (except for *name*). You do not need to consider interaction terms. Why would you not want to use *name* as a predictor?

```
ResumeNames2 = within(data, {name = NULL})
ResumeNames3 = as.data.frame(ResumeNames2)
head(ResumeNames3)
```

```
##      sex ethnicity quality call experience equal
## 1 female      cauc     low    0          6   yes
## 2 female      cauc    high    0          6   yes
## 3 female      afam     low    0          6   yes
## 4 female      afam    high    0          6   yes
## 5 female      cauc    high    0         22   yes
## 6   male      cauc     low    0          6   yes
```

```
ResumeNames4 = ResumeNames3[,c(1:2,4:6,3)]
head(ResumeNames4)
```

```
##      sex ethnicity call experience equal quality
## 1 female      cauc    0          6   yes     low
## 2 female      cauc    0          6   yes    high
## 3 female      afam    0          6   yes     low
## 4 female      afam    0          6   yes    high
## 5 female      cauc    0         22   yes    high
## 6   male      cauc    0          6   yes     low
```

```
ResumeNames4$experience = as.numeric(ResumeNames4$experience)
ResumeNames4$ethnicity = as.factor(ResumeNames4$ethnicity)
ResumeNames4$quality = as.factor(ResumeNames4$quality)
ResumeNames4$sex = as.factor(ResumeNames4$sex)
ResumeNames4$call = as.factor(ResumeNames4$call)
ResumeNames4$equal = as.factor(ResumeNames4$equal)
ResumeNames4= as.data.frame(ResumeNames4)
bestglm(ResumeNames4, family=binomial)
```

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

```
## BIC
## BICq equivalent for q in (1.65719100997386e-07, 0.961933448817152)
## Best Model:
##               Estimate  Std. Error   z value     Pr(>|z|)
## (Intercept)  0.27461799 0.053679228  5.115908 3.122352e-07
## experience  -0.03628244 0.005830869 -6.222475 4.893720e-10
```

```
ResumeNames4.bestglm = bestglm(ResumeNames4, family = binomial)
```

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

```
ResumeNames4.bestglm$BestModels
```

```
##     sex ethnicity  call experience equal Criterion
## 1 FALSE     FALSE FALSE       TRUE FALSE  6719.928
## 2 FALSE     FALSE  TRUE       TRUE FALSE  6726.387
## 3 FALSE     FALSE FALSE       TRUE  TRUE  6727.850
## 4  TRUE     FALSE FALSE       TRUE FALSE  6728.268
## 5 FALSE      TRUE FALSE       TRUE FALSE  6728.419
```

```
"In our model, we exclude the name variable as it is a unique identifier and
does not provide any predictive value for whether an applicant is called
back. We also convert the relevant variables into factors and numerics to
prepare them for modeling.

Based on the output, our best model includes only the experience variable,
with a negative coefficient of -0.03628244. This suggests that as experience
increases, the probability of being called back decreases. No other variables
were found to be significant in the model.The best model selected by the BIC
criterion includes the experience predictor and the intercept term. The
estimated coefficient for experience is -0.0363 with a standard error of
0.0058. The z-value is -6.222 and the associated p-value is less than 0.0001,
indicating strong evidence that experience is a significant predictor of the
probability of a job applicant being called back."
```

```
## [1] "In our model, we exclude the name variable as it is a unique
identifier and does not provide any predictive value for whether an applicant
is called back. We also convert the relevant variables into factors and
numerics to prepare them for modeling.\n\nBased on the output, our best model
includes only the experience variable, with a negative coefficient of -
0.03628244. This suggests that as experience increases, the probability of
being called back decreases. No other variables were found to be significant
in the model.The best model selected by the BIC criterion includes the
experience predictor and the intercept term. The estimated coefficient for
experience is -0.0363 with a standard error of 0.0058. The z-value is -6.222
and the associated p-value is less than 0.0001, indicating strong evidence
that experience is a significant predictor of the probability of a job
applicant being called back."
```