QAC 201: Applied Data Analysis

Logistic Regression: More with interpretations of models

1. We would like to explore the relationship between number of kids less than 6 and employment status for women in the workforce in 1975. This can be accomplished with a logistic regression since the response variable (employment status (inlf) is binary categorical coded as 0=no, 1=yes).

```
model1<-glm(inlf~kidslt6, family="binomial",data=mroz)
summary(model1)
```

```
##
## Call:
## glm(formula = inlf ~ kidslt6, family = "binomial", data = mroz)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.3869  -1.3869   0.9815   0.9815   1.7392
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.48006    0.08265   5.808 6.32e-09 ***
## kidslt6     -0.87179    0.15705  -5.551 2.84e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1029.75  on 752  degrees of freedom
## Residual deviance:  994.75  on 751  degrees of freedom
## AIC: 998.75
##
## Number of Fisher Scoring iterations: 4
```

```
exp(model1$coefficients)
```

```
## (Intercept)     kidslt6
##   1.6161718   0.4182015
```

The conclusion here would be that there is a significant relationship between number of kids less than 6 and hourly wage (OR=0.42, p-value<0.001). In particular, number of kids less than 6 is significantly and negatively associated with likelihood of employment. In particular, for each additional kid less than 6, the odds of employment is expected to change by a factor of 0.48. (Notice that the odds changing by a multiplicative less than 1, denotes a decrease in likelihood).

2. Now suppose we are looking to explore the relationship between whether they live in the city (0=no, yes=1) and employment status. While we could explore this relationship with an appropriate bivariate test (here, that would be chi-square), suppose instead we wish to construct an appropriate regression model.

```
model2<-glm(inlf~factor(city), family="binomial",data=mroz)
summary(model2)
```

```
##
## Call:
## glm(formula = inlf ~ factor(city), family = "binomial", data = mroz)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.304  -1.292   1.056   1.067   1.067
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)     0.2920     0.1232   2.369   0.0178 *
## factor(city)1  -0.0260     0.1536  -0.169   0.8656
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1029.7  on 752  degrees of freedom
## Residual deviance: 1029.7  on 751  degrees of freedom
## AIC: 1033.7
##
## Number of Fisher Scoring iterations: 4
```

```
exp(model2$coefficients)
```

```
##   (Intercept) factor(city)1
##     1.3391304     0.9743352
```

The model suggests that there is not a significant relationship between city status and employment (OR=0.97, p-value=0.8656).

3. We can continue to build the model to include additional control variables. Suppose now we wish to examine the relationship between city status and wage after controlling for experience and spouse's wage.

```
model3<-glm(inlf~factor(city)+exper+huswage, family="binomial",data=mroz)
summary(model3)
```

```
##
## Call:
## glm(formula = inlf ~ factor(city) + exper + huswage, family = "binomial",
##     data = mroz)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.8056  -1.0501   0.5749   1.0245   1.5550
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -0.615352   0.206438  -2.981  0.00287 **
## factor(city)1 -0.002297   0.173811  -0.013  0.98946
## exper          0.104551   0.011981   8.726  < 2e-16 ***
## huswage       -0.019152   0.019329  -0.991  0.32177
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1029.75  on 752  degrees of freedom
## Residual deviance:  930.09  on 749  degrees of freedom
## AIC: 938.09
##
## Number of Fisher Scoring iterations: 4
```

```
exp(model3$coefficients)
```

```
##   (Intercept) factor(city)1         exper       huswage
##     0.5404507     0.9977058     1.1102119     0.9810302
```

City status is still not significantly related to employment status (OR=0.9977, p-value=0.9895) after controlling for work experience and spouse wage. You can continue making other interpretations in your model as well. For instance, it appears that experience (labeled exper in data) is significantly and positively associated with likelihood of employment (OR=1.11, p-value<0.001) after controlling for city and spouse wage.

4. Now we are looking to explore the relationship between maximum educational attainment level (defined as High School or Less, College BA/BS, or Grad School) and employment status.

```
model<-glm(inlf~education, family="binomial",data=mroz)
summary(model)
```

```
##
## Call:
## glm(formula = inlf ~ education, family = "binomial", data = mroz)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6828  -1.2201   0.7457   1.1353   1.1353
##
## Coefficients:
##                               Estimate Std. Error z value Pr(>|z|)
## (Intercept)                     0.4285     0.1960   2.186   0.0288 *
## educationGrad School            0.7094     0.3020   2.349   0.0188 *
## educationHigh School or Less   -0.3286     0.2141  -1.535   0.1248
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1029.7  on 752  degrees of freedom
## Residual deviance: 1009.0  on 750  degrees of freedom
## AIC: 1015
##
## Number of Fisher Scoring iterations: 4
```

```
exp(model$coefficients)
```

```
##                   (Intercept)         educationGrad School
##                     1.5348837                    2.0327273
## educationHigh School or Less
```

```
##                     0.7199623
```

Notice that even though we have only a single explanatory variable (education) there are two slope terms/odds ratios to consider for education. Every categorical explanatory variable that you use in a model will have a reference level and each level that appears in your model will be compared against that reference level. In this case, the reference level is 'College'.

In this model we can conclude that those who have a graduate school education are significantly more likely to be employed ***compared to people with a maximum educational attainment level of College*** *(OR=2.03, p-value=0.0188). This model estimates that those who have a graduate school education have an odds of employment that is 2.03 times higher than those with a maximum educational attaintment level of College.*

In this model we we do not have enough evidence to show that those who have have a maximum educational attainment of High School education or less have a significantly different likelihood of employment ***compared to those with a maximum attaingment of college*** *(OR=0.7200, p-value=0.1248). While the model estimates that those who have a high school education or less are less likely to be employed compared to those with a College education, the difference is not deemed significant.*

You may now be wondering, is there a difference in employment likelihood between Grad School educations compared to those with High School or Less? Our current model does not allow you to test this, so if it is of interest, you would need to change the reference level. We only recommend you change the reference level, if it is interesting to your research question.

```r
# Code below changes the reference level to Grad School
mroz$education<-as.factor(mroz$education)
mroz$education<-relevel(mroz$education, ref="Grad School")

# Re-run your model code and your output will give you
# two different slope terms
model<-glm(inlf~education, family="binomial",data=mroz)
summary(model)
```

```
##
## Call:
## glm(formula = inlf ~ education, family = "binomial", data = mroz)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6828  -1.2201   0.7457   1.1353   1.1353
##
## Coefficients:
##                              Estimate Std. Error z value Pr(>|z|)
## (Intercept)                    1.1378     0.2298   4.951 7.39e-07 ***
## educationCollege              -0.7094     0.3020  -2.349   0.0188 *
## educationHigh School or Less  -1.0379     0.2454  -4.229 2.35e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1029.7  on 752  degrees of freedom
## Residual deviance: 1009.0  on 750  degrees of freedom
## AIC: 1015
##
## Number of Fisher Scoring iterations: 4
```

```r
exp(model$coefficients)
```

```
##                   (Intercept)             educationCollege
##                     3.1200000                   0.4919499
## educationHigh School or Less
##                     0.3541854
```

There is enough evidence to suggest that those who have a maximum educational attainment level of High School or Less are significantly less likely to be be employed compared to those with Grad School educations (OR=-2.877, p-value<0.001). The model estimates that those with High School or Less have an odds of employment that is expected to be 0.35 times the odds of employment for someone with Grad School education.

This next part should sound very familiar (look at our previous model). In this model we can conclude that those who have a maximum educational attainment of college are significantly less likely to be employed ***compared to people with a maximum educational attainment level of Graduate School*** *(OR=0.49, p-value=0.0188). This model estimates that those who have a high school or less education have an odds of employment that is expected to be 0.49 times that of someone who has a grad school education.*

Running a third model will not be particularly useful since there are no more comparisons to make.

The models that you obtain by changing the reference are algebraically equivalent. Why?