**RSCH 6120/8120: HW 5**

Kristen Zhang | Nov 28

**This HW uses the eclsk.csv data. Context: We are going to investigate students that attend public school in the eclsk data with several student and school variables. Review the eclsk codebook to understand the different variables available.**

**1. Use str(eclsk) and the eclsk codebook to identify dichotomous variables that would be an appropriate outcome for logistic regression.**

> str(eclsk)

'data.frame': 2435 obs. of 20 variables:

$ X : int 4 5 6 7 8 9 10 11 12 13 ...

$ schid : int 15 15 15 15 15 15 15 15 15 15 ...

$ parent.ed : int 3 5 8 5 4 3 5 1 3 5 ...

$ female : int 0 1 0 1 1 1 0 1 1 1 ...

$ income : num 15000 40000 35000 75000 14000 50000 37000 10000 25000 10000 ...

$ math : num 32.7 34.1 37.7 57.9 46.8 ...

$ read : num 36.2 39 76.2 74.5 32.5 ...

$ gen : num 24.1 27.7 31.4 37 23.2 ...

$ minority : int 0 1 0 0 0 0 0 0 0 0 ...

$ age : num 73 63.6 79 77.9 70.1 ...

$ public : int 1 1 1 1 1 1 1 1 1 1 ...

$ attend : num 90.5 90.5 90.5 90.5 90.5 ...

$ gifted : num 0 0 0 0 0 0 0 0 0 0 ...

$ free.lunch : num 52.4 52.4 52.4 52.4 52.4 ...

$ suburban : int 0 0 0 0 0 0 0 0 0 0 ...

$ urban : int 0 0 0 0 0 0 0 0 0 0 ...

$ enrollment : int 1 1 1 1 1 1 1 1 1 1 ...

$ min.percent : int 1 1 1 1 1 1 1 1 1 1 ...

$ parent.ed.cat: chr "A - HS or less" "B - College Experience" "C - Graduate Schooling" "B - College Experience" ...

$ income.cat : chr "A - Low" "B - Middle" "B - Middle" "C - High" ...

From the results, dichotomous variables that would be an appropriate outcome for logistic regression include: $female, $minority, $public, $suburban, $urban. Since $suburban and $urban are interchangeable, we will only consider the first four.

**2. Fit a logistic regression model to predict if a student attends public school (public) with at least three predictors.**

**a) Provide code and results**

**b) Identify if each predictor is significant or not and interpret the coefficients**

> model.log <- glm(public~attend+parent.ed+gen,data=eclsk,family=binomial(logit))

> summary(model.log)

Call:

glm(formula = public ~ attend + parent.ed + gen, family = binomial(logit),

data = eclsk)

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) 23.31092 2.59820 8.97 < 0.0000000000000002 \*\*\*

attend -0.20374 0.02712 -7.51 0.000000000000057 \*\*\*

parent.ed -0.20275 0.03097 -6.55 0.000000000058996 \*\*\*

gen -0.04778 0.00791 -6.04 0.000000001559502 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 2461.7 on 2434 degrees of freedom

Residual deviance: 2239.0 on 2431 degrees of freedom

AIC: 2247

Number of Fisher Scoring iterations: 5

> exp(summary(model.log)$coefficients[,"Estimate"])

(Intercept) attend parent.ed gen

13298599160.6537 0.8157 0.8165 0.9533

Interpretation: all the three variables (attend, parent.ed, and gen) are significant predictors of the odds of students being in public schools.

* For every one unit (%) increase in attendance rate, the predicted log odds of being in a public school go down 0.20. For every one 1%increase in attendance rate, the odds of it being a public school increase by exp(-0.20374) = 0.8157 times… 18% reduction in odds.
* For every one unit (%) increase in parent education, the predicted log odds of being in a public school go down 0.20. For every one 1%increase in attendance rate, the odds of it being a public school increase by exp(-0.20275) = 0.8165 times… 18% reduction in odds.
* For every one unit (%) increase in general IRT scores, the predicted log odds of being in a public school go down 0.05. For every one 1%increase in attendance rate, the odds of it being a public school increase by exp(-0.04778) = 0.9533 times… 4.7% reduction in odds.

**3. Model comparison: now add “female” to the model and run.**

**a) Is female a significant predictor?**

**b) Compare the models (i.e., use deviance and pseudo-R2). Are they significantly different? Which do you think is better and why?**

> model.log2 <- glm(public~attend+parent.ed+gen+female,data=eclsk,family=binomial(logit))

> summary(model.log2)

Call:

glm(formula = public ~ attend + parent.ed + gen + female, family = binomial(logit),

data = eclsk)

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) 23.26180 2.59741 8.96 < 0.0000000000000002 \*\*\*

attend -0.20384 0.02710 -7.52 0.000000000000055 \*\*\*

parent.ed -0.20467 0.03103 -6.60 0.000000000042181 \*\*\*

gen -0.04808 0.00793 -6.06 0.000000001362204 \*\*\*

female 0.16217 0.10601 1.53 0.13

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 2461.7 on 2434 degrees of freedom

Residual deviance: 2236.6 on 2430 degrees of freedom

AIC: 2247

Number of Fisher Scoring iterations: 5

> exp(summary(model.log2)$coefficients[,"Estimate"])

(Intercept) attend parent.ed gen female

12661151583.6741 0.8156 0.8149 0.9531 1.1761

1. Female is not a significant predictor.
2. The residual deviance for the first model is 2239.0, for the second is 2236.6. Lower is better, so the second model is better.

> pR2(model.log)

fitting null model for pseudo-r2

llh llhNull G2 McFadden r2ML r2CU

-1119.47794 -1230.85307 222.75026 0.09049 0.08742 0.13742

> pR2(model.log2)

fitting null model for pseudo-r2

llh llhNull G2 McFadden r2ML r2CU

-1118.30509 -1230.85307 225.09597 0.09144 0.08830 0.13880

Based on the **pseudo-R2**, McFadden's would be higher for the model with the greater likelihood. So the McFadden for the first model is lower than that for the second (0.09049 < 0.09144), and the second one is better.

But the two models are not significantly different. By Chi-square, the p > .05.

**4. Identify one more variable and add it to the model from question 3.**

**a) What variable did you select?** I selected income.

**b) Is it a significant predictor of public-school attendance?** Yes. p-value = 0.000049921706724 < 0.001.

**c) Of the three models you have fit, which is the best? Why?** The final one is the best. As the residual deviance is the lowest (2218.7), which means that it explained more variance.

> model.log3 <- glm(public~attend+parent.ed+gen+female+income, data=eclsk, family=binomial(logit))

> summary(model.log3)

Call:

glm(formula = public ~ attend + parent.ed + gen + female + income,

family = binomial(logit), data = eclsk)

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) 23.00666003 2.60319607 8.84 < 0.0000000000000002 \*\*\*

attend -0.20297513 0.02714546 -7.48 0.000000000000076 \*\*\*

parent.ed -0.14466851 0.03441146 -4.20 0.000026214637928 \*\*\*

gen -0.04306197 0.00803039 -5.36 0.000000082133107 \*\*\*

female 0.16803406 0.10660225 1.58 0.11

income -0.00000494 0.00000122 -4.06 0.000049921706724 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 2461.7 on 2434 degrees of freedom

Residual deviance: 2218.7 on 2429 degrees of freedom

AIC: 2231

Number of Fisher Scoring iterations: 5

> exp(summary(model.log3)$coefficients[,"Estimate"])

(Intercept) attend parent.ed gen female income

9809920727.0957 0.8163 0.8653 0.9579 1.1830 1.0000

**5. Broadly, regression approaches help us understand the relationship between variables. Identify an outcome/dependent variable from your substantive area and possible independent variables you would include in a regression analysis.**

**Example: I’m interested in student achievement in mathematics (outcome) and would consider gender, race, socio-economic status, and previous mathematics achievement scores as important dependent variables in a regression analysis involving student mathematic achievement.**

I’m interested in people’s adoption intentions of a new technology, so I consider their self-efficacy in technology (how much they believe they can operate them), scientific knowledge, educational level, and income as important dependent variables in a regression analysis involving their intentions to adopt a new media technology.