Log Regression

**Descriptions:**

* Logistic regression is the nonparametric version of regression or discriminant analysis. Since log regression is nonparametric, you have more flexibility with variables because there are no normality assumptions.
* Log regression is used to predict group membership, so the DV/outcome variable is categorical.
* The IV/predictor variables can be a mix of categorical or continuous variables.

**Definitions:**

* IV(s) – these are your predictor variables. They can be any combination of types of variables: categorical, continuous, Likert…and they do not have to be normally distributed.
* DV – this variable is your group classification or group membership.
* Equation:
  + Regular regression: Y-hat = a + bx1 + bx2 + bx3 …
  + Log regression: Y-hat = eu / 1 + eu
  + U = a + bx1 + bx2 + bx3
  + Therefore, when you get predicted group membership, you are getting the *log odds* of a person being in one group over other.
    - Think about this like sports betting…when you bet on something, there are odds if you are going to win. Log odds is the probability that you will be in one group over another.

**Types of Log Regression:**

* Binary Logistic Regression – only two outcome categories.
* Multinomial Logistic Regression – three or more outcome categories.
* Subtypes of each of these:
  + Direct Logistic Regression – same as simultaneous regression. All predictors are entered into the equation at the same time. Usually used when there is no theory about order of variables or no “control” variables.
  + Sequential/Hierarchical Logistic Regression – you enter variables in steps or groups based on some pre-determined order. You will look at the addition or change in models when you add the new sets.
  + Statistical (stepwise) Logistic Regression – usually used in psychology as a screening procedure for correlated IVs. Enters variables into the equation based on their ability to predict – criticized for its lack of theory-based decisions.
  + Probit – often considered the half-way point between log regression with purely categorical DVs and regular regression with purely continuous DVs. Probit has the assumption that the DV is normally distributed, but is not totally continuous (like a likert scale!).

**Power:**

* Run it the same as a regular regression. Log regression will have more power than regular regression when the assumptions are not met.
* You can use the estimates from G\*Power and a regression model (depending on direct, hierarchical) and get a good number of subjects to run.

**Assumptions/Issues:**

* Categorical DVs – you have to have groups as your outcome variable. You can have as many as you want, but it does make interpretation more difficult the more groups you have.
* Ratio of Cases (small N) – you want to have non-small expected and observed frequencies in each outcome category. If you have a very small category, it will be hard to predict (and more probable to predict the big category). You can collapse categories that are small or collect more data.
  + Also you want to have more participants in your study than independent variables. You will get perfect predictors (bad) if you have a small number of people and lots of predictors.
* Additivity – you still do not want to use two predictor variables that overlap in variance a great deal (but you can use hierarchical log regression if you have that problem and don’t want to combine the predictors).
* Independence of errors – you can only be classified in **one** group. You cannot be part of two of the outcome groups. Therefore, no repeated measures variables.

**Theoretical Interpretations/Issues:**

1. Goodness of Fit Models:
   1. Null models – a model that includes no predictors.
   2. Residual/fitted models – models that include predictors.
   3. You want the fitted model to have a lower residual (error) than the null model.
2. Individual Predictors in the equation
   1. Wald test – Wald test is similar to testing if a b/beta is significantly different from zero, it’s a type of chi-square analysis.
   2. Remember, though, that we are predicting the odds/probability of group membership, so the positive predictors increase the odds of the group coded second, while negative predictors increase the odds of the group coded first.
3. Effect Sizes
   1. Cox and Snell – R2 based on likelihoods and sample size
      1. BUT never can reach 1, even if you achieve perfect fit.
   2. Nagelkerke R2 – adjusts Cox and Snell so that the upper limit is 1 – most people report this type of effect size.
   3. And a bunch more!
4. Coding (with multinomial log regression):
   1. You will have two log regressions.
      1. 0-1 equation.
      2. 0-2 equation.
      3. Etc.
   2. Hence why 0 needs to be your reference category you want to compare every other group to.
   3. If you want 0-1, 1-2, you have to recode or specify when you run the analysis.

# COMPLETE EXAMPLE

BINARY LOGISTIC

**Data set:** binary log.csv

**Research Question:** Can we correctly predict working and nonworking individuals by attitudes and demographics?

**Classification variable/DV:**

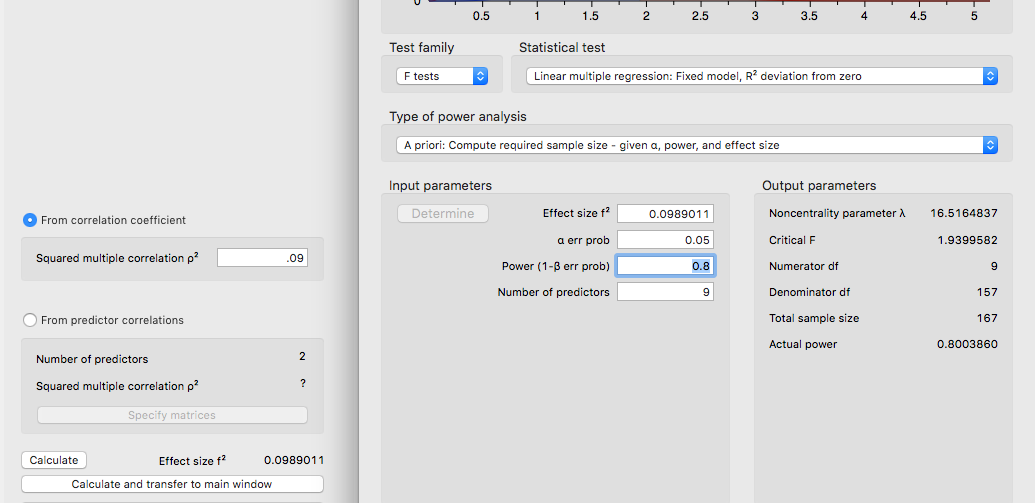
* Work status (working or not).

**IV(s):**

* Children – yes or no
* Race – white or other
* CONTROL – Locus of Control
* ATTMAR – attitude toward marriage
* ATTROLE – attitude toward role of women
* SEL – socioeconomic status
* ATTHOUSE – Attitude toward housework
* Age
* Education level

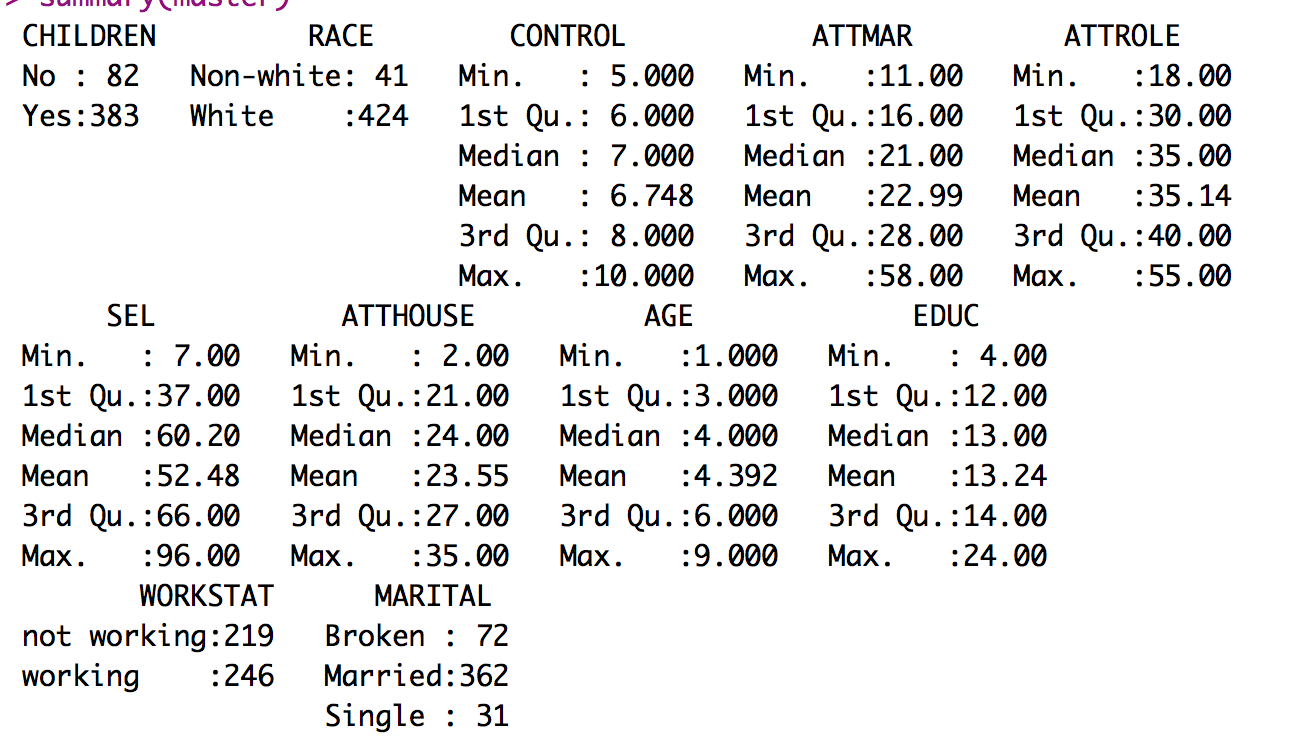
**Power:**

1. G\*Power Options:
   1. Test family: F Test
   2. Statistical test: Linear multiple regression: fixed model R2 deviation from zero
   3. Effect size: click determine 🡪 from correlation coefficient 🡪 estimate R2 🡪 calculate and transfer to main window.
   4. Alpha = .05
   5. Power (1-beta of .20) = .80
   6. Number of predictors = number of IVs.
2. Let’s test a medium effect size:
   1. R2 = .09
   2. Number of predictors: 9
   3. We need 167 individuals in our study with a medium effect size.



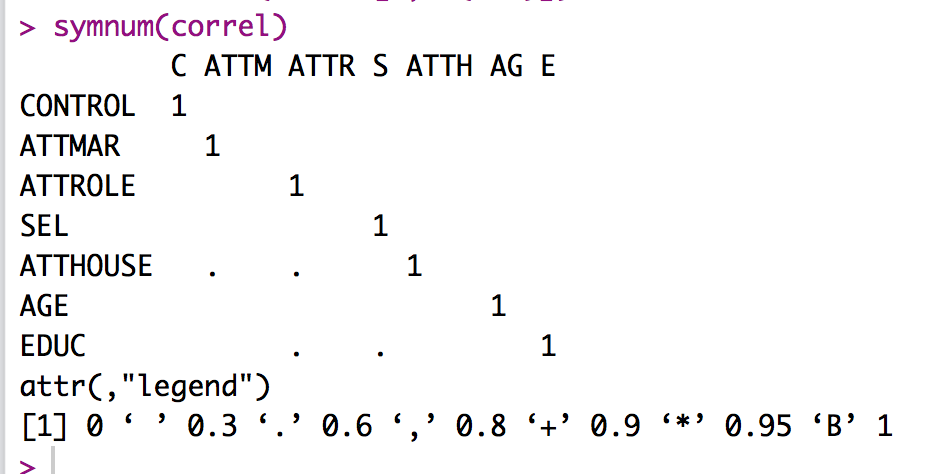
**Data screening and Assumptions:**

1. Accuracy and Ratio of Cases:
   1. Use the summary(*dataset name*) function to get the basic information for the data.



* 1. Here it is important to check a couple things:
     1. The continuous IVs do not exceed their min and max values.
     2. The categorical variables are set up as category for dummy coding and the DV.
     3. The categories for categorical variables do not have too small or empty categories.
        1. Our DV is ok but the ethnicity and children predictors are not good news.

1. Missing:
   1. I can see from my summary function that I do not have missing data. Remember that you will need at least twenty variables to estimate missing data for participants – so mostly you won’t be estimating for regression.
2. Additivity
   1. Get the correlations:
      1. correl = cor(*dataset*, use = “pairwise.complete.obs”)
   2. Get the symbols chart:
      1. symnum(correl)
   3. Appears we do not have serious issues (*r* = .90) or potential suppression issues (*r* = .70).

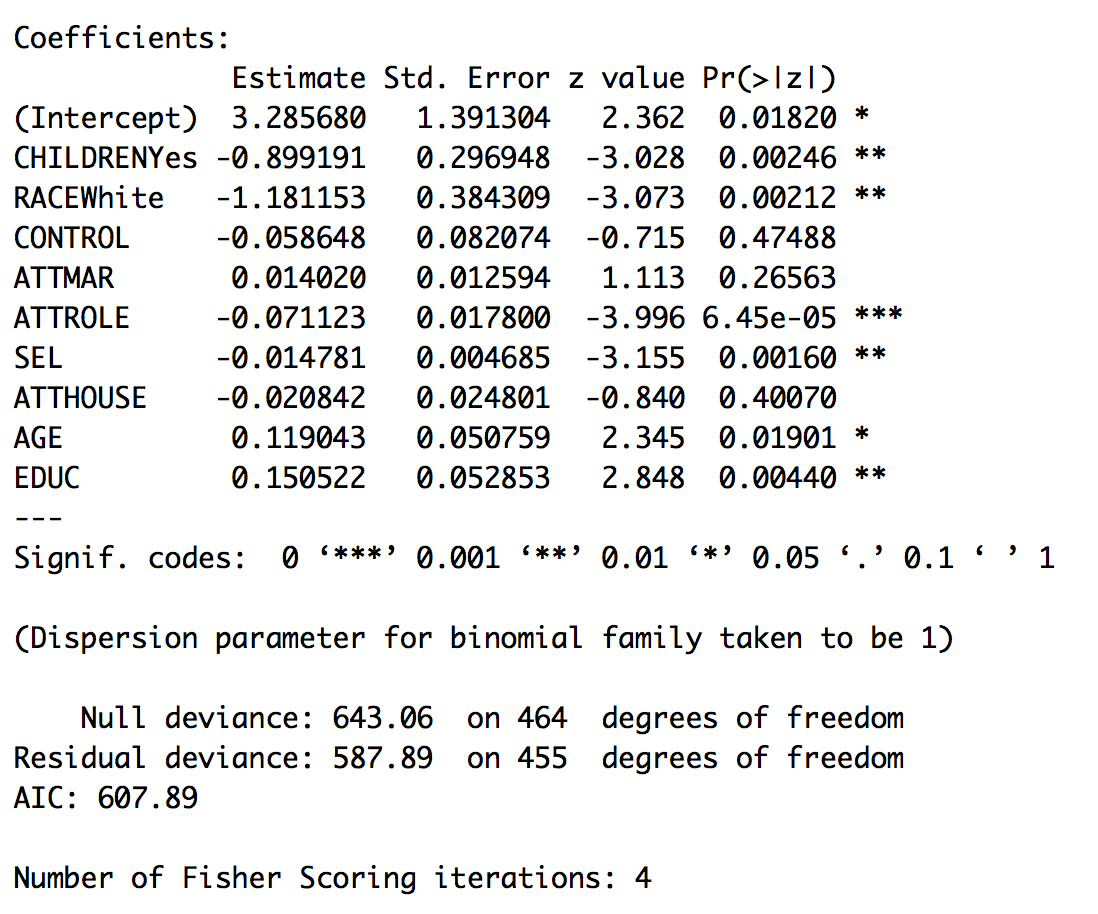


**Run the regression:**

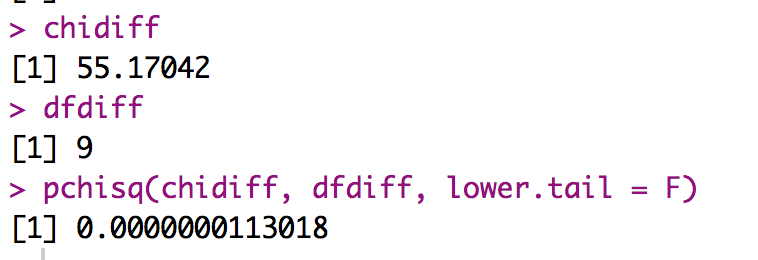
1. Run the analysis using the glm() command.
   1. glm(*DV* ~ *IV + IV + IV,*

family = binomial(link = ‘logit’),

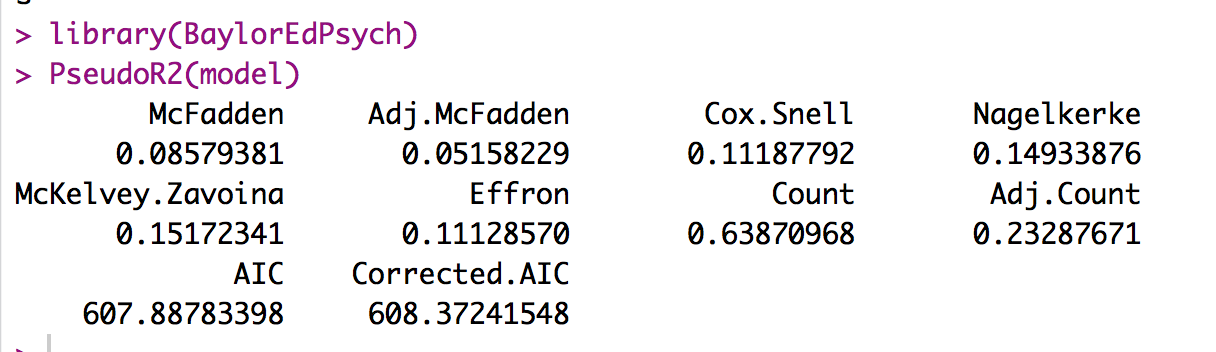
data = *dataset*)



1. Is the overall model significant?
   1. This output is not quite as obvious as regular regression output.
   2. Use the Null deviance and residual deviance to compare against each other.
      1. Null deviance is a null model that no predictors are included (kind of like a random guess).
      2. Residual deviation is the model with your predictors.
      3. Therefore, you want the error (deviance) to be lower with the predictors.
      4. You can use a chi-square change test to determine if it’s significant.
   3. Test the difference!
      1. Find the chi square difference between null and residual by subtracting:
         1. chidiff = model$null.deviance - model$deviance
      2. Find the df difference between null and residual by subtracting:
         1. dfdiff = model$df.null - model$df.residual
      3. Figure out if that difference score is significant:
         1. pchisq(chidiff, dfdiff, lower.tail = F)
      4. The logic of this test is the same as a hierarchical regression – we are seeing if the change in errors is significantly less than if we did not have any predictors in the model. So we figure out the change scores, then figure out the p value for those change scores (which we want to be less than .05).
   4. So, yes, the overall model is significant:
      1. *X*2(9) = 55.17, *p* < .001

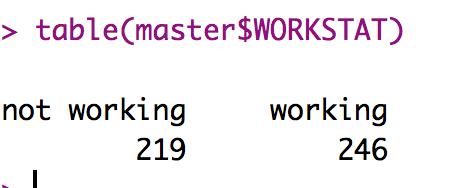


1. What about R2?
   1. Install and load the BaylorEdPsych library.
   2. Run PseudoR2(model).



* 1. Yes, the model is significant with a medium effect:
     1. *X*2(9) = 55.17, *p* < .001, Nagelkerke *R2* = .15

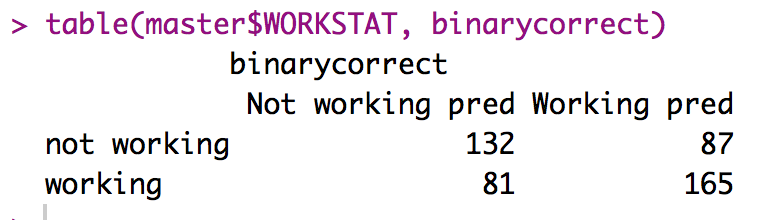
1. Interpret the coefficients:
   1. Use a table to figure out who is the base group versus other group.
   2. table(*dataset$DV)*



* 1. Make a table!

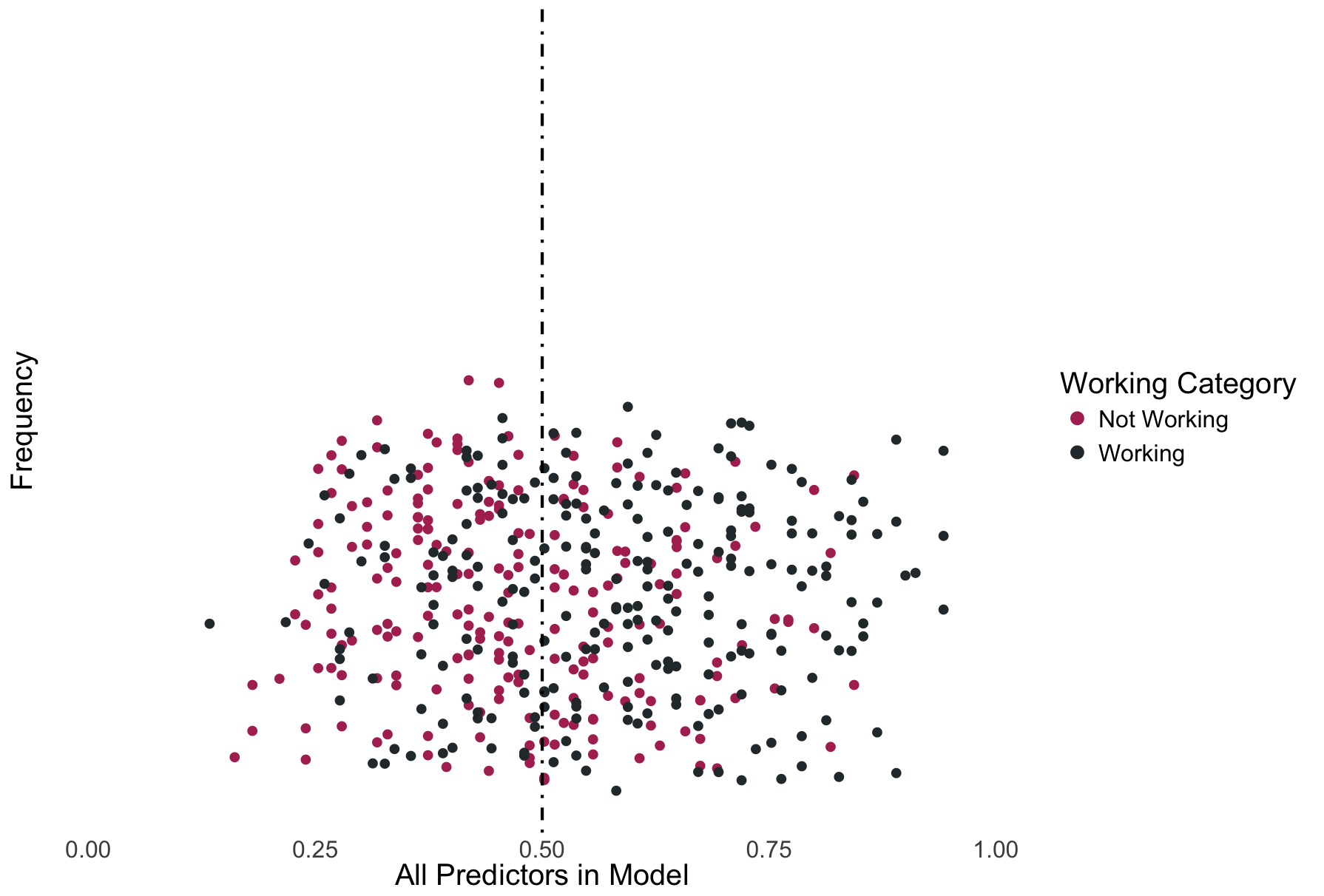
|  |  |  |
| --- | --- | --- |
|  | Not Working | Working |
| Children | -.90 (for Yes kids)  People who have kids are not working |  |
| Race | -1.18 (for Whites)  Whites are less likely to work |  |
| Control NS |  |  |
| Attitude Marriage NS |  |  |
| Attitude Role of Women | -.07  As attitudes go up (barefoot in the kitchen), less likely to work |  |
| SES | -.01  As income goes up, less likely to work |  |
| Attitude toward housework NS |  |  |
| Age |  | .12  As age increases more likely to be working |
| Education |  | .15  As education goes up more likely to be working |

1. Look at percent correct.
   1. How many people did we correctly classify with our results?
   2. We normally use the fitted values to determine if we’ve predicted better than chance – but in this case, fitted values are the probabilities of the second group. So we can use those values to see what group it would have predicted them be in.
   3. Run some code to get those values:
      1. correct = model$fitted.values
      2. binarycorrect = ifelse(correct > 0.5,1,0)
      3. I also ran a factor function to match this up with my categories.
   4. Then compare those values to the real values:
      1. table(*dataset*$*DV*, binarycorrect)



* 1. To get the percent correct for each group use the following formula:
     1. Percent match / total for that row
     2. 132 / (132 + 87) \* 100
     3. 165 / (81 + 165) \* 100
  2. To get the overall correct:
     1. Percent match of both / total N
     2. (132 + 165) / nrow(master) \* 100

1. You can also make a plot of the fitted values (in a continuous way) and see how well you did visually:



* 1. You wouldn’t necessarily publish this type of plot, but it helps see what’s happening.
     1. At the dashed line, to the left side is the lower coded group (not working) and to the right is the higher coded group (working).
     2. The x axis is probability of being in the higher coded group.
     3. So all the red dots SHOULD be on the left side, while the black dots should be on the right.
     4. Frequency is a bit of a mislabel, but each dot is a person.
  2. The code is included below.

1. Code explanation:
   1. Open ggplot2:
      1. library(ggplot2)
   2. Run some theme coding to turn off the ugly graph defaults.

theme = theme(panel.grid.major = element\_blank(),

panel.grid.minor = element\_blank(),

panel.background = element\_blank(),

axis.text.y=element\_blank(),

axis.ticks=element\_blank(),

axis.line.x = element\_line(color = "black"),

axis.line.y = element\_line(color = "black"),

text = element\_text(size=20),

legend.key = element\_blank())

* 1. Create a histogram:
     1. hist = ggplot(master, aes(*dataset*, color = *DV*, fill = *DV*))
  2. Add things to the histogram:
     1. Geom dotplot makes the dots rather than histogram bars. The jitter just makes sure that you can see the dots, otherwise they all stack together.
     2. Coord\_cartesian sets the x limit to 0 and 1 (because it’s probability).
     3. X lab adds the x axis label.
     4. Y lab adds the y axis label.
     5. Scale color and scale fill allow you to change the legend and dot colors – you have to include both.
     6. Geom v line creates the vertical dashed line at 50/50 guessing.

hist +

theme +

geom\_dotplot(binwidth = .01, position = "jitter") +

coord\_cartesian(xlim = c(0,1)) +

xlab("All Predictors in Model") +

ylab("Frequency") +

scale\_color\_manual(values = c("Maroon", "#2C3539"),

labels = c("Not Working", "Working"),

name = "Working Category")+

scale\_fill\_manual(values = c("Maroon", "#2C3539"),

labels = c("Not Working", "Working"),

name = "Working Category") +

geom\_vline(xintercept=c(.50), linetype="dotdash", size = 1)

**Write Up:**

**Results**

A direct binary logistic regression analysis was conducted to evaluate membership prediction for working and not working individuals using the presence of children (yes or not), race (Caucasian or other), locus of control, attitudes of the role of women, marriage, housework, socioeconomic status, age and education. 465 participants were included in this analysis: 246 for the working category and 219 for the not working category. Multicollinearity between variables was not present in the dataset.

The full model included the variables listed above and was significant, *X2*(9) = 55.17, *p*<.001, Nagelkerke *R*2=.15. Overall, 63.9% of the participants were correctly classified, with slightly better classification in the working group (67.1%) over the non-working group (60.3%).

Please see Table 1 for *b* values and their significance levels. Locus of control, marriage status attitude, and housework attitude were all non-significant predictors of working status. In general, participants who did not have children, were older, and had more education were likely to be classified in the working category. Caucasian participants, with high attitudes about the role of women, and high socioeconomic status were more likely to be classified in the non-working category.

Table 1. *Exponent B and Significance Values for Predictors.*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | *B* | *S.E.* | *Wald* | *p* |
| Children (yes) | -0.899 | 0.297 | 9.169 | 0.002 |
| Race (non-white) | 1.181 | 0.384 | 9.446 | 0.002 |
| Locus of Control | -0.059 | 0.082 | 0.511 | 0.475 |
| Marriage Attitude | 0.014 | 0.013 | 1.239 | 0.266 |
| Role of Women Attitude | -0.071 | 0.018 | 15.965 | 0.000 |
| Socioeconomic Status | -0.014 | 0.005 | 9.956 | 0.002 |
| Housework Attitude | -0.021 | 0.025 | 0.706 | 0.401 |
| Age | 0.119 | 0.051 | 5.500 | 0.019 |
| Education | 0.151 | 0.053 | 8.111 | 0.004 |
| Constant | 3.286 | 1.391 | 2.362 | 0.018 |
| *Note.* *df* = 1. |  |  |  |  |