Running head: PREDICTORS OF JOB RETENTION

Overtime Hours and Tips as Predictors of Job Retention in a Restaurant Setting

Daniel Pinedo

Psych 308d: Assignment 3

PREDICTORS OF JOB RETENTION

Overtime Hours and Tips as Predictors of Job Retention in a Restaurant Setting

Results

Data analysis is in Appendix A. Observations did not contain any missing parameters in the dataset. Initial dataset ($N = 100$) contained an additional categorical level for retention (plans on staying an additional year) *border* was requested to not be examined by client and was subsequently removed ($N = 69$), with categories for *Yes* and *No* remaining and all predictor variables centered. Analysis continued with tests of assumptions and binary logistic regression models. The first assumption of independence of observations for tested observed variables passed. The second assumption of normal distribution test for overtime hours passed (*skew* = 0.02, *kurtosis* = -1.51) although the distribution appears bimodal and therefore likely non-normal, and for tips passed (*skew* = 0.09, *kurtosis* = -0.37). The test for multicollinearity of the saturated model passed, *tolerance* = .97.

Model 1 tested if overtime hours predicted retention which was significant, χ^2 (68) = 67.40 , $p < .001$, *Classification* = .86, indicating 86% projected overall accuracy in predictions for this model. Overtime hours had an odds ratio of 12.70 (*logit* = 2.97) indicating that a person who worked more hours is 12.70 times more likely to intend to remain on the job versus leaving within an additional year.

Model 2 tested if tips predicted retention which was significant, χ^2 (68) = 58.30 , $p < .001$, *Classification* = .77, indicating 77% projected overall accuracy in predictions for this model. Tips had an odds ratio of 1.03 (*logit* = 0.03) indicating that a person who received more tips is 1.03 times more likely to intend to remain on the job versus leaving within an additional year.

Model 3 [χ^2 (67) = 45.30 , *p* < .001, *Classification* = .87] tested the hypothesis that there was a significant difference between the most accurate single predictor model (Model 2) and the model with both predictors added, which was significant, $\Delta\chi^2$ (1) = 22.10, *p* < .001. Tips had an odds ratio of 1.03 (*logit* = 0.03), and overtime hours had an odds ratio of 12.63 (*logit* = 2.54).

Discussion

In response to the HR manager's query about the star employee Trudy based on her average weekly overtime of seven hours and $100 in tips, we were able to predict her likelihood to remain for another year applying Model 3 with a predicted probability of 100% (*logit* = 20.95 , *odds ratio* = 1,254,496,332). Although Model 3 was identified as the most significant model, the odds ratio for tips did not add much useful impact to predicting whether an employee intends stay or quit within the next year.

The first major limitation of this study is that the most impactful predictor variable (hours) was also bimodal reducing the validity of this statistical analysis. Hours can be transformed into a binary categorical variable (e.g. over 2.5 hours overtime or under 2.5 hours overtime) and dummy coded. Alternately, a larger sample size may result in a more normal distribution with only one mode. The second major limitation was removing the borderline responses from the dataset. A multinomial logistic regression can be completed with that data to test if people that are undecided can be predicted to remain or leave within a year as well.

Recommendations for the HR director at this restaurant for future research are to test if intention to stay/leave accurately predicts actual turnover, indicating if it is a good indicator of turnover or not for the restaurant. In addition, attributes of star performers such as personality or psychological capital can be measured and used for selection of new employees. Finally, completing a mixed-methods study including interviews of employees indicating why they

would decide to stay versus leave may point towards other data points or survey items related to

person-organization fit or job satisfaction that may be included in the future to a predictive

model for why employees leave or stay in this organization.

Appendix A

Statistical Analysis in R

Daniel Pinedo

April 16, 2019

```
## Warning: package 'knitr' was built under R version 3.5.3
```

You have been hired as an Organizational Psychologist for a local restaurant. The Head of HR is concerned about high turnover amongst their servers. Specifically, she is interested in figuring out what predicts whether a server will stay at the restaurant for another year or not. Although a survey of her staff included responses of uncertainty of staying or not, HR *only* cares about those who are planning to stay or leave.

Analyses: After speaking with the managers, you think that the two best predictors will be number of overtime hours worked per week and amount earned in tips each week. You decide to survey the wait staff to see whether (a) tips, (b) overtime hours, or (c) both tips AND overtime hours should be used by the HR manager in predicting someone's retention status.

Additional Discussion Question: Additionally, the HR manager is particularly worried that she is going to lose her star waitress Trudy. Given that, on average, Trudy works 7 hours of overtime a week and makes $100 in tips, what would you tell the HR manager about the probability of Trudy staying for another year? *Please address this concern in your discussion section.*

Variables: 1. Hours - continuous, average overtime hours worked per week (in hours) 2. Tips - continuous, average amount of tips earned each week (in dollars) 3. Re (Retention) a. "Yes" (plans on staying at the restaurant for another year) b. "No" (does not plan on staying at the restaurant for another year) c. "border" (is unsure whether or not they will stay for another year)

TIP: Please center your predictor variables for your main analyses and when using it to calculate the likelihood of Trudy staying!

```
library(psych)
library(jmv)

## Warning: package 'jmv' was built under R version 3.5.3

##
## Attaching package: 'jmv'
```

```
## The following object is masked from 'package:psych':

##

##     pca

## The following object is masked from 'package:stats':

##

##     anova
```

library(aod)

```
## Warning: package 'aod' was built under R version 3.5.3
```

library(QuantPsyc)

```
## Warning: package 'QuantPsyc' was built under R version 3.5.3

## Loading required package: boot

##
## Attaching package: 'boot'

## The following object is masked from 'package:psych':

##

##     logit

## Loading required package: MASS

##
## Attaching package: 'QuantPsyc'

## The following object is masked from 'package:base':

##

##     norm
```

library(popbio)

dat <- **read.csv**("https://www.dropbox.com/s/jej8t73qnelvijp/PSY.308d.DA3-4.csv?dl=1")
head(dat)

PREDICTORS OF JOB RETENTION

```
##   Hours Tips    Re
## 1  2.10  467 border
## 2  2.22  591 border
## 3  2.35  541 border
## 4  2.41  444 border
## 5  2.57  572 border
## 6  2.63  483 border

dim(dat)

## [1] 100   3
```

Subset dataset and check for missing parameters

```
#remove observations that are not "yes" or "no" for Retention variable
dat.subset <- dat[which(dat$Re!='border'), ] # N=100 changes to N=69
dat.subset <- droplevels(dat.subset) # change levels for Retention variable by dropping
"border"

#see what is missing
#run descriptives
desc <- descriptives(data = dat.subset,
             vars = c('Re', 'Hours', 'Tips'),
             sd = TRUE,
             skew = TRUE,
             kurt = TRUE,
             freq = TRUE,
             hist = TRUE)
desc

##
##  DESCRIPTIVES
##
## Descriptives
## -----------------------------------------------
##                Re   Hours   Tips
```

PREDICTORS OF JOB RETENTION

```
## ------------------------------------------------
## N                      69      69      69
## Missing                0       0       0
## Mean                          2.97    509
## Median                        3.03    521
## Standard deviation            0.518   84.2
## Minimum                       2.13    321
## Maximum                       3.80    693
## Skewness                      0.0184  0.0913
## Std. error skewness           0.289   0.289
## Kurtosis                      -1.51   -0.374
## Std. error kurtosis           0.570   0.570
## ---------------------------------------------
##
##
## FREQUENCIES
##
## Frequencies of Re
## ------------------------------------------------
##   Levels   Counts   % of Total   Cumulative %
## ------------------------------------------------
##   No       33       47.8         47.8
##   Yes      36       52.2         100.0
## ------------------------------------------------
```

PREDICTORS OF JOB RETENTION
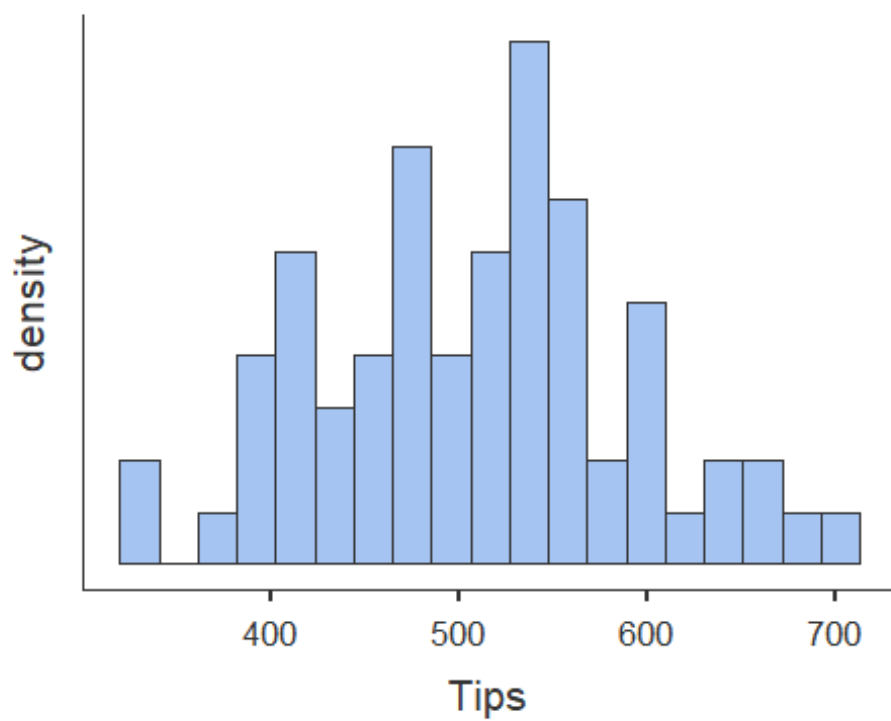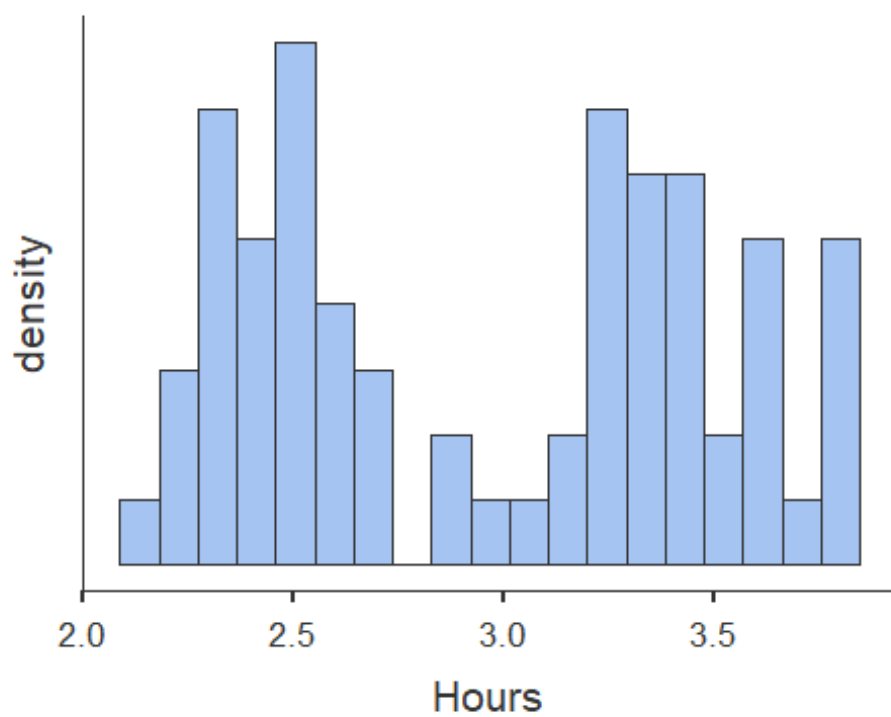
 # baseline classification success is equal to the reference frequency for Retention (No = 48%)

 # it also looks like Hours is bimodal - likely non-normal distribution

PREDICTORS OF JOB RETENTION

Assumptions 1. Independence of Observations 2. Predictor Variables Normally Distributed *(Hours is bimodal)* 3. Multicollinearity

Correlations

```
# Correlations of continuous variables
cortable <- corrMatrix(data = dat.subset,
              vars = c('Hours', 'Tips'),
              flag = TRUE)
cortable

##
##  CORRELATION MATRIX
##
##  Correlation Matrix
##  -----------------------------------------
##                   Hours    Tips
##  -----------------------------------------
##   Hours   Pearson's r      —      0.436
##        p-value          —     < .001
##
##   Tips    Pearson's r          —
##        p-value             —
##  -----------------------------------------
##   Note. * p < .05, ** p < .01, *** p <
##   .001
```

Logistic Plots

```
#Transform binary outcome to integer for the plot to work
dat.subset$Re.int[dat.subset$Re == "No"] <- 0
dat.subset$Re.int[dat.subset$Re == "Yes"] <- 1


#Center Predictors
dat.subset$HoursC <- dat.subset$Hours - round(mean(dat.subset$Hours), digits = 2)
dat.subset$TipsC <- dat.subset$Tips - round(mean(dat.subset$Tips), digits = 2)
```

PREDICTORS OF JOB RETENTION

#Show Plots for centered predictors

logi.hist.plot(dat.subset**$**TipsC, dat.subset**$**Re.int, boxp=FALSE, type="hist", col="gray", xlabel = "Tips Centered")
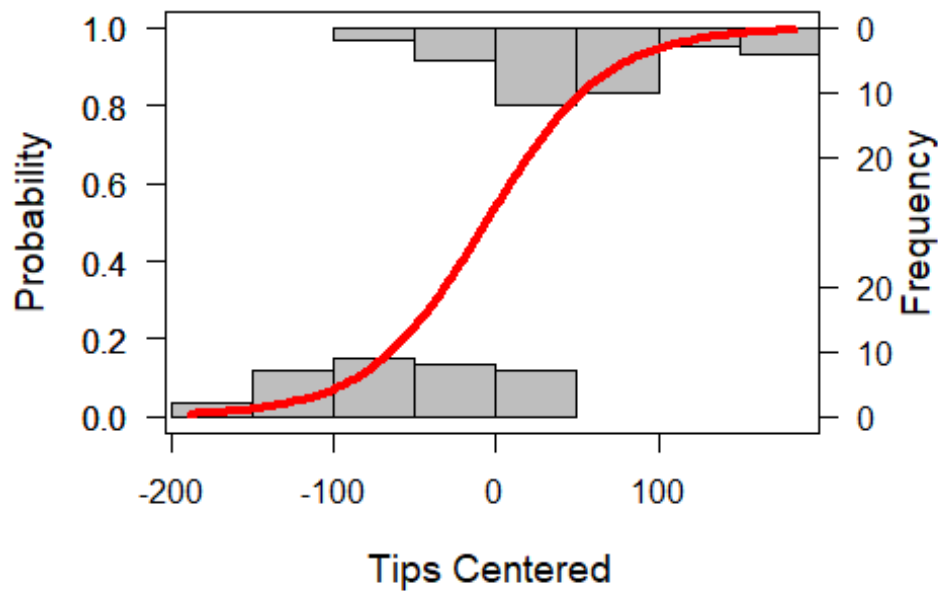
 # when tips are above average, people tend to stay vs. below average they go

logi.hist.plot(dat.subset**$**HoursC, dat.subset**$**Re.int, boxp=FALSE, type="hist", col="gray", xlabel = "Hours Centered")
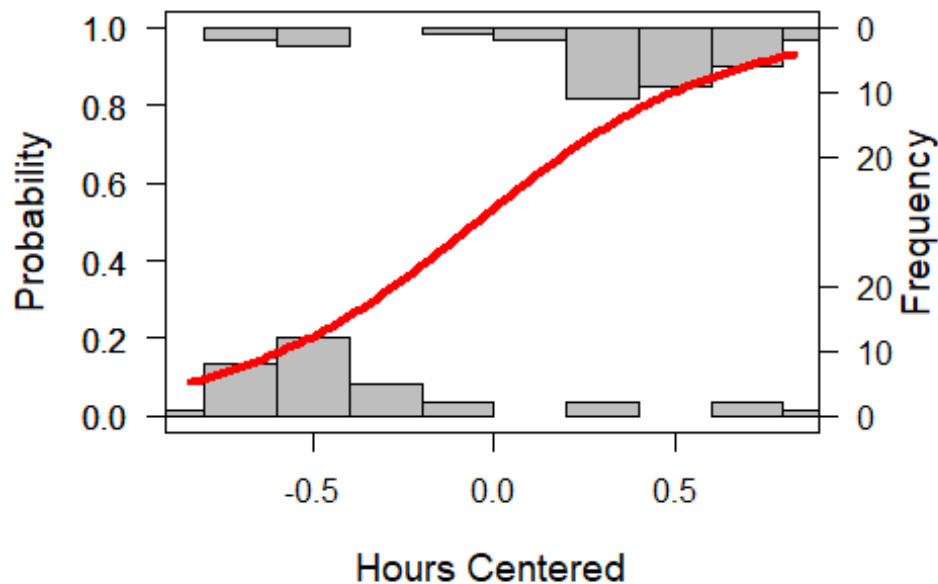
when OT hours are above average, people tend to stay vs. below they go (caveat: bimodal dsitribution so it is inconclusive)

BiLoRe Models Null model

```
# Null deviance = Chi squared for the model
# df = N - (# of parameters) - 1
model0 <- glm(dat.subset$Re ~ 1, family = binomial)
summary(model0)

##
## Call:
## glm(formula = dat.subset$Re ~ 1, family = binomial)
##
## Deviance Residuals:
##    Min      1Q   Median      3Q      Max
## -1.215  -1.215   1.141   1.141   1.141
##
## Coefficients:
```

```
##           Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.08701   0.24100   0.361    0.718
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 95.524  on 68  degrees of freedom
## Residual deviance: 95.524  on 68  degrees of freedom
## AIC: 97.524
##
## Number of Fisher Scoring iterations: 3
```

```r
print("Logit")
```

```
## [1] "Logit"
```

```r
coef(model0)
```

```
## (Intercept)
##  0.08701138
```

```r
model0.odds <- exp(coef(model0)) #converts coefficient to odds [P(outcome)/(1-P(outcome))]
print("Odds")
```

```
## [1] "Odds"
```

```r
model0.odds
```

```
## (Intercept)
##    1.090909
```

```r
model0.probs <- model0.odds / (1 + model0.odds) #
print("Probabilities")
```

```
## [1] "Probabilities"
```

```r
model0.probs
```

```
## (Intercept)
##   0.5217391
```

```
print("Columns = Observed, Rows = Predicted")

## [1] "Columns = Observed, Rows = Predicted"

print("Null model")

## [1] "Null model"

ClassLog(model0, dat.subset$Re) # classification success under the null model (baseline)

## $rawtab
##      resp
##       No Yes
##   TRUE 33  36
##
## $classtab
##      resp
##       No Yes
##   TRUE  1   1
##
## $overall
## [1] 0.4782609
##
## $mcFadden
## [1] 0
```

Model 1 - Hours predicting Retention

```
#Multicollinearity
 #Tolerance = 1 - R squared --> for our purpose < .4 is bad
 #VIF = 1/Tolerance
 #Small VIF values indicates low correlation among variables under ideal conditions
 #Multicollinearity occurs when two or more predictors in the model are correlated and provide
redundant information   about the response. Multicollinearity was measured by variance inflation
factors (VIF) and tolerance. If VIF value   exceeding 4.0, or by tol- erance less than 0.2 then
there is a problem with multicollinearity (Hair et al., 2010).
```

```
# when odds ratio < 1 just flip (invert) the result(in relation to "no" instead of in relation to "yes")

# Deviance score is the chi-squared for this model
# AIC is used to compare non-nested models for fit (lower means better fit)
# top chi-squared indicates the change of chi-squared vs the null model (Deviance + chi
squared)
# top df score indicates the change of df vs the null model
# df = N - (# of predictors) - 1

model1.jmv <- jmv::logRegBin( # Multicollinearity is not relevant for this answer
  data = dat.subset,
  dep = Re,
  covs = vars(HoursC),
  blocks = list(
    list(
      'HoursC')),
  refLevels = list(
    list(
      var = 'Re',
      ref = 'No')),
  modelTest = TRUE,
  OR = TRUE,
  class = TRUE,
  acc = TRUE,
  collin = TRUE)

model1.jmv

##
##  BINOMIAL LOGISTIC REGRESSION
##
##  Model Fit Measures
##  ---------------------------------------------------------------
##    Model   Deviance   AIC    R²-McF   <U+03C7>²    df   p
```

PREDICTORS OF JOB RETENTION

```
## ----------------------------------------------------------
##      1       67.4   71.4    0.295   28.1    1   < .001
## ----------------------------------------------------------
##
##
## MODEL SPECIFIC RESULTS
##
## MODEL 1
##
## Model Coefficients
## ----------------------------------------------------------------
##   Predictor   Estimate   SE      Z       p        Odds ratio
## ----------------------------------------------------------------
##   Intercept      0.139   0.304   0.457   0.648         1.15
##   HoursC         2.973   0.667   4.458   < .001       19.54
## ----------------------------------------------------------------
##   Note. Estimates represent the log odds of "Re = Yes" vs. "Re
##   = No"
##
##
## ASSUMPTION CHECKS
##
## Collinearity Statistics
## ------------------------------
##          VIF     Tolerance
## ------------------------------
##   HoursC   1.00       1.00
## ------------------------------
##
##
## PREDICTION
##
## Classification Table – Re
```

PREDICTORS OF JOB RETENTION

```
## -------------------------------------
##   Observed   No   Yes   % Correct
## -------------------------------------
##       No   28    5       84.8
##       Yes   5   31       86.1
## -------------------------------------
##   Note. The cut-off value is set
##   to 0.5
##
##
## Predictive Measures
## ------------------
##   Accuracy
## ------------------
##      0.855
## ------------------
##   Note. The
##   cut-off value
##   is set to 0.5
```

Model 2 - Tips predicting Retention

#Multicollinearity

#Tolerance = 1 - R squared --> for our purpose < .4 is bad

#VIF = 1/Tolerance

#Small VIF values indicates low correlation among variables under ideal conditions

#Multicollinearity occurs when two or more predictors in the model are correlated and provide redundant information   about the response. Multicollinearity was measured by variance inflation factors (VIF) and tolerance. If VIF value   exceeding 4.0, or by tol- erance less than 0.2 then there is a problem with multicollinearity (Hair et al., 2010).

 # when odds ratio < 1 just flip (invert) the result(in relation to "no" instead of in relation to "yes")

 # Deviance score is the chi-squared for this model

 # AIC is used to compare non-nested models for fit (lower means better fit)

PREDICTORS OF JOB RETENTION

```
# top chi-squared indicates the change of chi-squared vs the null model (Deviance + chi
squared)
# top df score indicates the change of df vs the null model
# df = N - (# of predictors) - 1

model2.jmv <- jmv::logRegBin( # Multicollinearity is not relevant for this answer
  data = dat.subset,
  dep = Re,
  covs = vars(TipsC),
  blocks = list(
    list(
      'TipsC')),
  refLevels = list(
    list(
      var = 'Re',
      ref = 'No')),
  modelTest = TRUE,
  OR = TRUE,
  class = TRUE,
  acc = TRUE,
  collin = TRUE)

model2.jmv

##
##  BINOMIAL LOGISTIC REGRESSION
##
##  Model Fit Measures
##  -------------------------------------------------------------
##   Model   Deviance   AIC    R²-McF   <U+03C7>²   df   p
##  -------------------------------------------------------------
##     1       58.3     62.3   0.389    37.2        1    < .001
##  -------------------------------------------------------------
##
```

PREDICTORS OF JOB RETENTION

```
##
##  MODEL SPECIFIC RESULTS
##
##  MODEL 1
##
##  Model Coefficients
##  ------------------------------------------------------------------
##    Predictor    Estimate    SE        Z       p        Odds ratio
##  ------------------------------------------------------------------
##    Intercept     0.1660    0.32620   0.509   0.611        1.18
##    TipsC         0.0271    0.00646   4.199   < .001       1.03
##  ------------------------------------------------------------------
##    Note. Estimates represent the log odds of "Re = Yes" vs. "Re =
##    No"
##
##
##  ASSUMPTION CHECKS
##
##  Collinearity Statistics
##  -----------------------------
##         VIF      Tolerance
##  -----------------------------
##    TipsC   1.00       1.00
##  -----------------------------
##
##
##  PREDICTION
##
##  Classification Table – Re
##  --------------------------------------
##    Observed   No   Yes   % Correct
##  --------------------------------------
##        No    23    10       69.7
```

PREDICTORS OF JOB RETENTION

```
##      Yes    6   30        83.3
##  --------------------------------------
##    Note. The cut-off value is set
##    to 0.5
##
##
##  Predictive Measures
##  -------------------
##    Accuracy
##  -------------------
##      0.768
##  -------------------
##    Note. The
##    cut-off value
##    is set to 0.5
```

Model 3 - Comparing Hours Model to Full Model

```
#Multicollinearity
 #Tolerance = 1 - R squared --> for our purpose < .4 is bad
 #VIF = 1/Tolerance
  #Small VIF values indicates low correlation among variables under ideal conditions
  #Multicollinearity occurs when two or more predictors in the model are correlated and provide
redundant information   about the response. Multicollinearity was measured by variance inflation
factors (VIF) and tolerance. If VIF value   exceeding 4.0, or by tol- erance less than 0.2 then
there is a problem with multicollinearity (Hair et al., 2010).


# when odds ratio < 1 just flip (invert) the result(in relation to "no" instead of in relation to "yes")


# Deviance score is the chi-squared for this model
# AIC is used to compare non-nested models for fit (lower means better fit)
# top chi-squared indicates the change of chi-squared vs the null model (Deviance + chi
squared)
# top df score indicates the change of df vs the null model
# df = N - (# of predictors) - 1
```

PREDICTORS OF JOB RETENTION

```
model2.jmv <- jmv::logRegBin( # Multicollinearity is relevant for this answer
  data = dat.subset,
  dep = Re,
  covs = vars(HoursC, TipsC),
  blocks = list(
    list(
      'HoursC'),
    list(
      'TipsC')),
  refLevels = list(
    list(
      var = 'Re',
      ref = 'No')),
  modelTest = TRUE,
  OR = TRUE,
  class = TRUE,
  acc = TRUE,
  collin = TRUE)

model2.jmv

##
##  BINOMIAL LOGISTIC REGRESSION
##
##  Model Fit Measures
##  ------------------------------------------------------------
##   Model   Deviance   AIC    R²-McF   <U+03C7>²   df   p
##  ------------------------------------------------------------
##     1       67.4     71.4   0.295     28.1      1   < .001
##     2       45.3     51.3   0.526     50.3      2   < .001
##  ------------------------------------------------------------
##
##
```

PREDICTORS OF JOB RETENTION

```
## Model Comparisons
##  ---------------------------------------------
##   Model      Model   <U+03C7>²   df   p
##  ---------------------------------------------
##     1    -     2    22.1    1   < .001
##  ---------------------------------------------
##
##
## MODEL SPECIFIC RESULTS
##
## MODEL 1
##
## Model Coefficients
##  ----------------------------------------------------------------
##   Predictor   Estimate   SE      Z       p        Odds ratio
##  ----------------------------------------------------------------
##   Intercept    0.139    0.304   0.457   0.648       1.15
##   HoursC       2.973    0.667   4.458   < .001      19.54
##  ----------------------------------------------------------------
##   Note. Estimates represent the log odds of "Re = Yes" vs. "Re
##   = No"
##
##
## ASSUMPTION CHECKS
##
## Collinearity Statistics
##  ------------------------------
##          VIF     Tolerance
##  ------------------------------
##   HoursC   1.00      1.00
##  ------------------------------
##
##
```

PREDICTORS OF JOB RETENTION

```
## PREDICTION
##
## Classification Table – Re
## ------------------------------------
##   Observed    No   Yes   % Correct
## ------------------------------------
##       No   28    5       84.8
##       Yes   5   31       86.1
## ------------------------------------
##   Note. The cut-off value is set
##   to 0.5
##
##
## Predictive Measures
## -------------------
##   Accuracy
## -------------------
##      0.855
## -------------------
##   Note. The
##   cut-off value
##   is set to 0.5
##
##
## MODEL 2
##
## Model Coefficients
## -------------------------------------------------------------------
##   Predictor   Estimate   SE      Z      p      Odds ratio
## -------------------------------------------------------------------
##   Intercept    0.1737   0.37858   0.459   0.646      1.19
##   HoursC       2.5360   0.78704   3.222   0.001      12.63
##   TipsC        0.0256   0.00735   3.490   < .001     1.03
```

PREDICTORS OF JOB RETENTION

```
## -----------------------------------------------------------------------
##    Note. Estimates represent the log odds of "Re = Yes" vs. "Re =
##    No"
##
##
## ASSUMPTION CHECKS
##
## Collinearity Statistics
## -----------------------------
##          VIF    Tolerance
## -----------------------------
##   HoursC   1.03     0.968
##   TipsC    1.03     0.968
## -----------------------------
##
##
## PREDICTION
##
## Classification Table – Re
## ------------------------------------
##   Observed   No   Yes   % Correct
## ------------------------------------
##       No    29    4       87.9
##       Yes    5   31       86.1
## ------------------------------------
##    Note. The cut-off value is set
##    to 0.5
##
##
## Predictive Measures
## ------------------
##    Accuracy
## ------------------
```

PREDICTORS OF JOB RETENTION

```
##      0.870
## -------------------
##   Note. The
##   cut-off value
##   is set to 0.5
```

Use regression equation to calculate predicted logit, odds, and probability

```
#Discussion: star performer Trudy
print("Given that Trudy works 7 hours of ovetime and makes $100 in tips, the odds she will
remain for another year:")
```

```
## [1] "Given that Trudy works 7 hours of ovetime and makes $100 in tips, the odds she will
remain for another year:"
```

```
# Let OT = Overtime Hours, T = tips
OT = 7
T = 100

print("Model - Full model")
```

```
## [1] "Model - Full model"
```

```
predlogit <- .17 + (2.54*OT) + (.03*T)
predodds <- exp(predlogit)
predprob <- predodds / (1 + predodds)

print("Predicted Logit")
```

```
## [1] "Predicted Logit"
```

```
predlogit
```

```
## [1] 20.95
```

```
print("Predicted Odds")
```

```
## [1] "Predicted Odds"
```

PREDICTORS OF JOB RETENTION

predodds

[1] 1254496332

```
print("Predicted Probability")
```

[1] "Predicted Probability"

predprob

[1] 1