# Part 2

Multiple binary logistic regression

# Example
# What factors predict happiness?

Does hamster ownership, marital status, and number of hours free time an individual has predict response to the following survey question:

Are you happy?

- Yes
- No

- Predictors: Hamster ownership (yes/no), marital status (single, cohabiting, married, divorced), and hours free time (continuous)

- Outcome: Happiness (Yes/No)

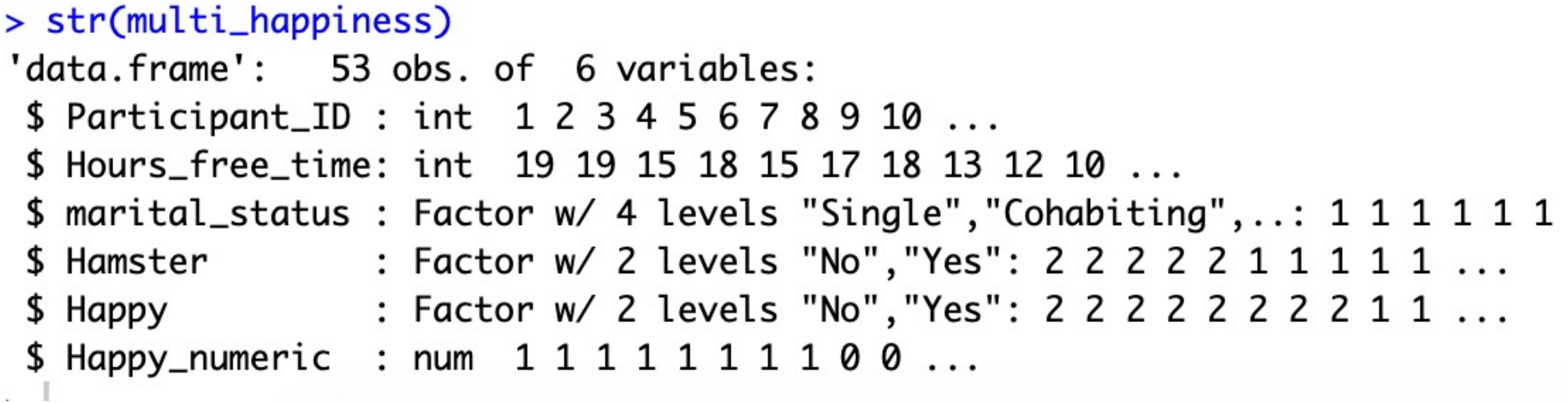
# 1. Prepare dataset:

- Outcome: binary:
  - Set as a numeric variable, where 1 is the outcome we are interested in (e.g. happiness = yes) and 0 is the other level (e.g. happiness = no)

- Predictors:
  - Categorical: factors with first level as the reference category:
    - Hamster ownership = no
    - Marital status = single
  - Continuous: numeric/integer variable
    - Hours free time

Check using the “str” function and adjust variables as required

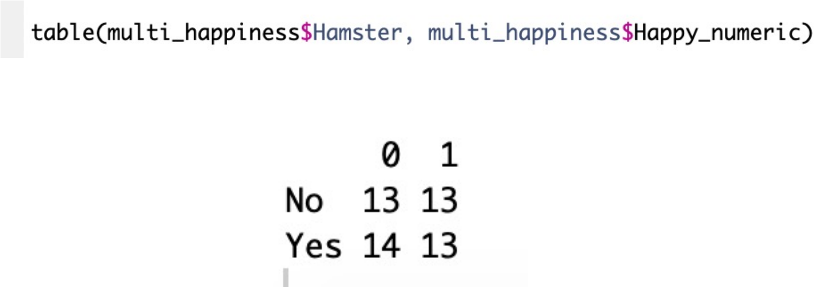
# 1. Prepare dataset



```
> str(multi_happiness)
'data.frame':	53 obs. of  6 variables:
 $ Participant_ID : int  1 2 3 4 5 6 7 8 9 10 ...
 $ Hours_free_time: int  19 19 15 18 15 17 18 13 12 10 ...
 $ marital_status : Factor w/ 4 levels "Single","Cohabiting",..: 1 1 1 1 1 1
 $ Hamster        : Factor w/ 2 levels "No","Yes": 2 2 2 2 2 1 1 1 1 1 ...
 $ Happy          : Factor w/ 2 levels "No","Yes": 2 2 2 2 2 2 2 2 1 1 ...
 $ Happy_numeric  : num  1 1 1 1 1 1 1 1 0 0 ...
```

# 2. Explore the data and check for separation
# Categorical variables: use ‘table’

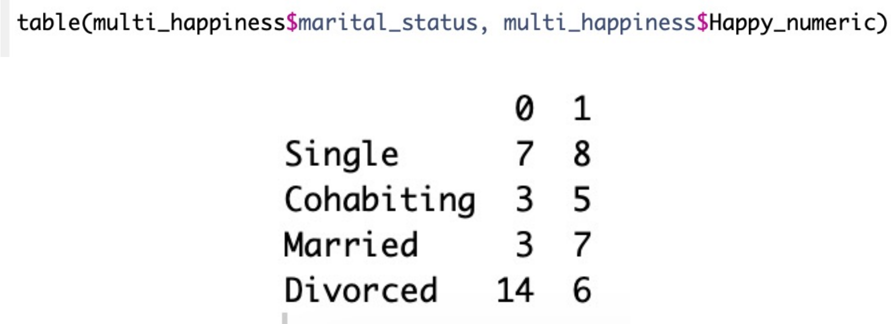
Hamster ownership



```
table(multi_happiness$Hamster, multi_happiness$Happy_numeric)
```

```
      0  1
No   13 13
Yes  14 13
```

Marital status

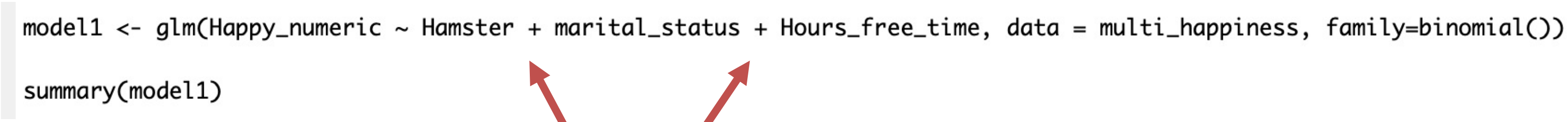


```
table(multi_happiness$marital_status, multi_happiness$Happy_numeric)
```

```
            0  1
Single      7  8
Cohabiting  3  5
Married     3  7
Divorced   14  6
```

No evidence of complete separation or quasi-complete separation for either variable

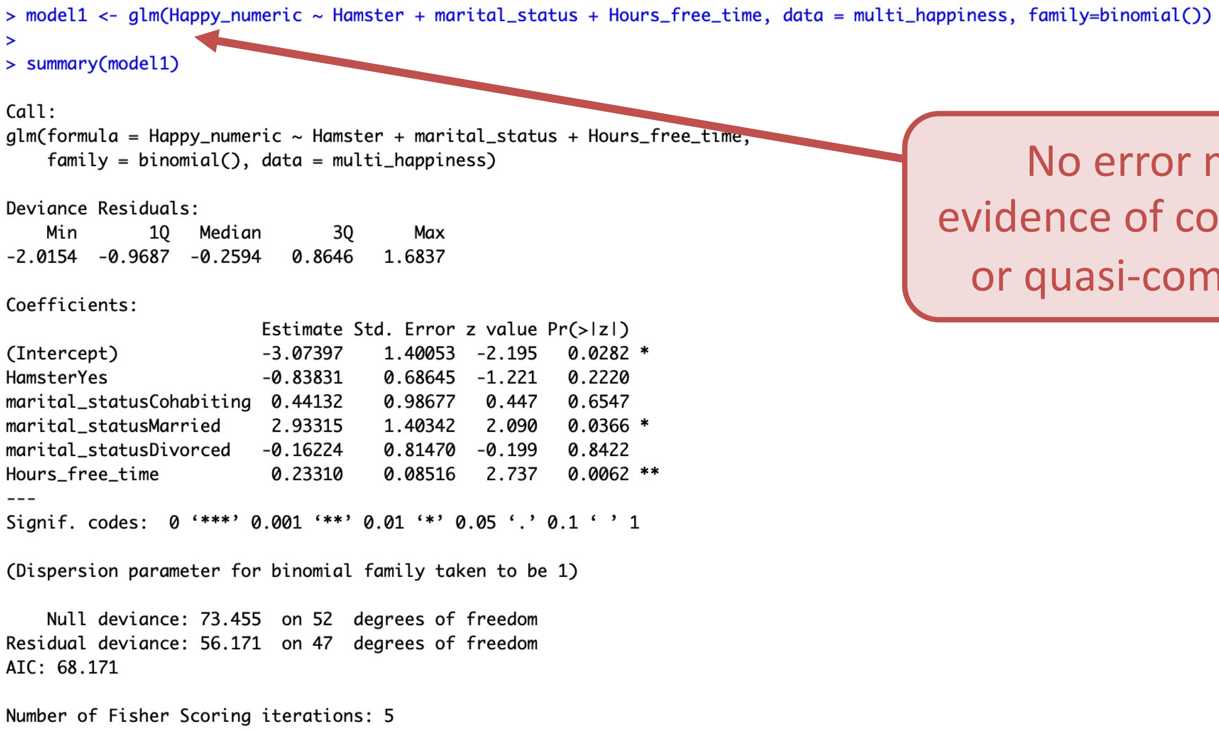
# 3. Run the model



```
model1 <- glm(Happy_numeric ~ Hamster + marital_status + Hours_free_time, data = multi_happiness, family=binomial())

summary(model1)
```

Separate variables with a “ + “

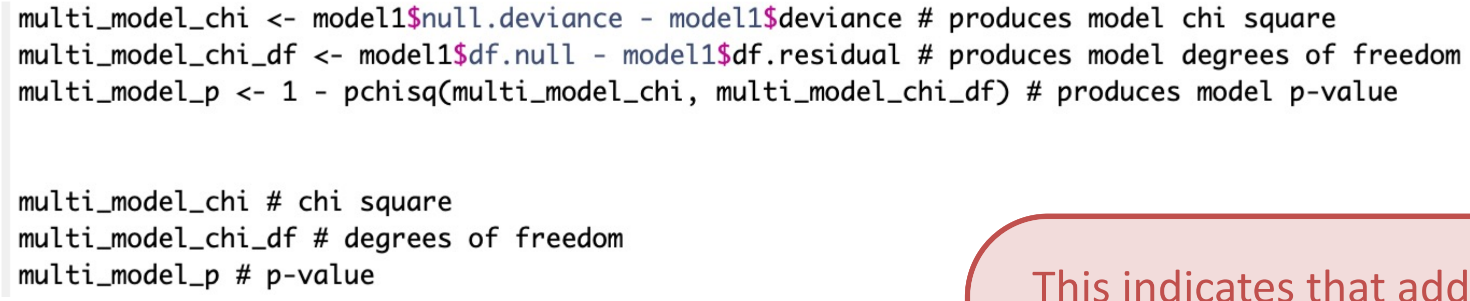
# 3. Run the model



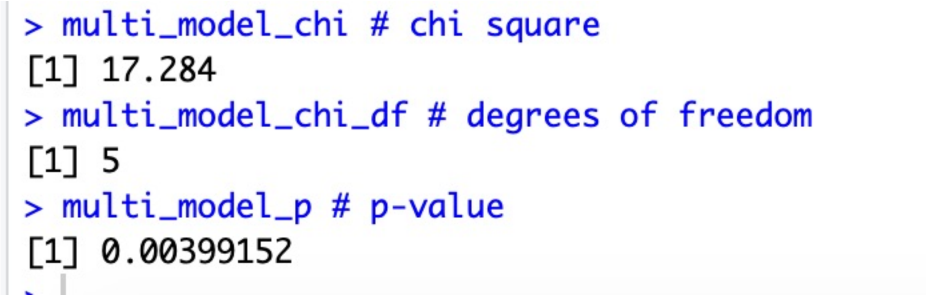
```
> model1 <- glm(Happy_numeric ~ Hamster + marital_status + Hours_free_time, data = multi_happiness, family=binomial())
>
> summary(model1)

Call:
glm(formula = Happy_numeric ~ Hamster + marital_status + Hours_free_time,
    family = binomial(), data = multi_happiness)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0154  -0.9687  -0.2594   0.8646   1.6837

Coefficients:
                         Estimate Std. Error z value Pr(>|z|)
(Intercept)              -3.07397    1.40053  -2.195   0.0282 *
HamsterYes               -0.83831    0.68645  -1.221   0.2220
marital_statusCohabiting  0.44132    0.98677   0.447   0.6547
marital_statusMarried     2.93315    1.40342   2.090   0.0366 *
marital_statusDivorced   -0.16224    0.81470  -0.199   0.8422
Hours_free_time           0.23310    0.08516   2.737   0.0062 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 73.455  on 52  degrees of freedom
Residual deviance: 56.171  on 47  degrees of freedom
AIC: 68.171

Number of Fisher Scoring iterations: 5
```

No error messages – no evidence of complete separation or quasi-complete separation

# 4. Evaluate the model
# Comparing to the intercept-only model



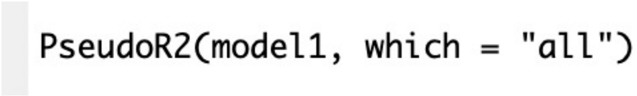
```
multi_model_chi <- model1$null.deviance - model1$deviance # produces model chi square
multi_model_chi_df <- model1$df.null - model1$df.residual # produces model degrees of freedom
multi_model_p <- 1 - pchisq(multi_model_chi, multi_model_chi_df) # produces model p-value


multi_model_chi # chi square
multi_model_chi_df # degrees of freedom
multi_model_p # p-value
```



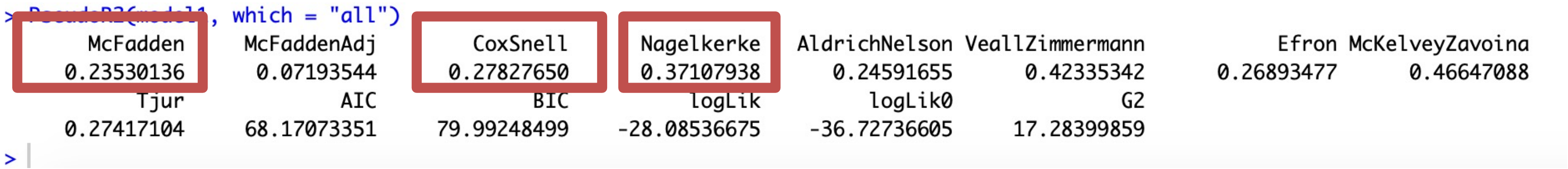
```
> multi_model_chi # chi square
[1] 17.284
> multi_model_chi_df # degrees of freedom
[1] 5
> multi_model_p # p-value
[1] 0.00399152
```

This indicates that adding the hamster ownership, marital status and hours free time variable to our model significantly improved the fit, compared to the null model containing intercept only

# 4. Pseudo R$^2$s



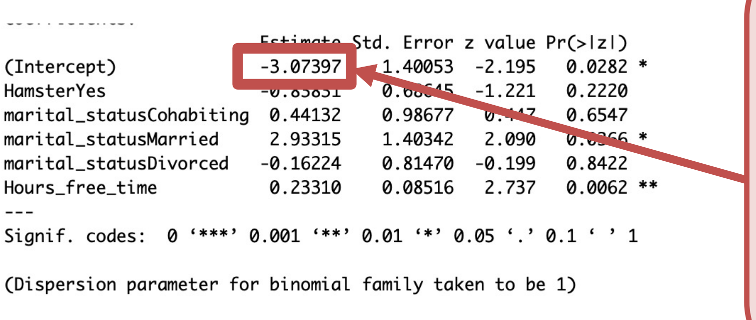
```
PseudoR2(model1, which = "all")
```



```
> PseudoR2(model1, which = "all")
        McFadden      McFaddenAdj         CoxSnell       Nagelkerke    AldrichNelson  VeallZimmermann            Efron  McKelveyZavoina
      0.23530136       0.07193544       0.27827650       0.37107938       0.24591655       0.42335342       0.26893477       0.46647088
            Tjur              AIC              BIC           logLik          logLik0               G2
      0.27417104      68.17073351      79.99248499     -28.08536675     -36.72736605      17.28399859
>
```

- McFadden = 0.24
- CoxSnell = 0.28
- Nagelkerke = 0.37

# 5. Evaluating individual predictors
# The intercept



```
Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)               -3.07397    1.40053  -2.195   0.0282 *
HamsterYes                -0.83831    0.68645  -1.221   0.2220
marital_statusCohabiting   0.44132    0.98677   0.447   0.6547
marital_statusMarried      2.93315    1.40342   2.090   0.0366 *
marital_statusDivorced    -0.16224    0.81470  -0.199   0.8422
Hours_free_time            0.23310    0.08516   2.737   0.0062 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)
```

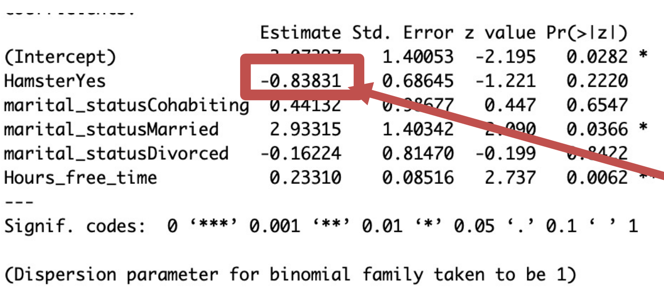
The log odds that happiness = yes, when:

- Hamster = No
- Marital status = Single
- Hours_free_time = 0

# 5. Evaluating individual predictors
# The hamster variable

- Interpretation of log odds is slightly different when you have 2+ predictors

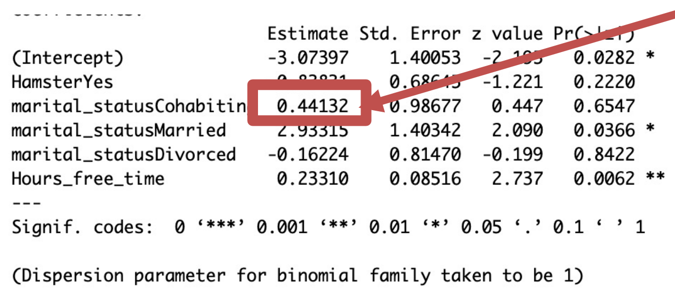
- Change in log odds when holding other variables constant



```
Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)               -3.07307    1.40053  -2.195   0.0282 *
HamsterYes                -0.83831    0.68645  -1.221   0.2220
marital_statusCohabiting   0.44132    0.98677   0.447   0.6547
marital_statusMarried      2.93315    1.40342   2.090   0.0366 *
marital_statusDivorced    -0.16224    0.81470  -0.199   0.8422
Hours_free_time            0.23310    0.08516   2.737   0.0062 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)
```

The change in the log odds of happy = yes when moving from HamsterNo to HamsterYes when holding other variables constant

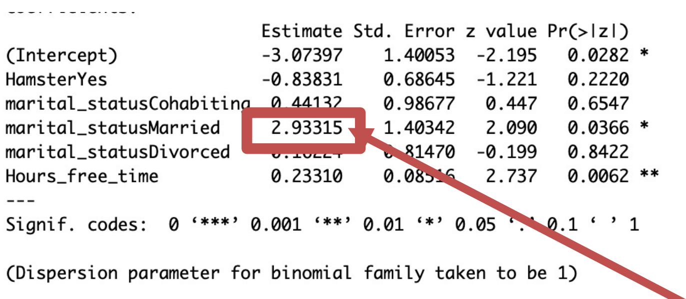
# 5. Evaluating individual predictors
# The marital status variable



```
Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)               -3.07397    1.40053  -2.195   0.0282 *
HamsterYes                 0.83831    0.68645  -1.221   0.2220
marital_statusCohabitin    0.44132    0.98677   0.447   0.6547
marital_statusMarried      2.93315    1.40342   2.090   0.0366 *
marital_statusDivorced    -0.16224    0.81470  -0.199   0.8422
Hours_free_time            0.23310    0.08516   2.737   0.0062 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)
```

The change in the log odds of happy = yes when moving from marital_statusSingle to marital_statusCohabiting when holding other variables constant

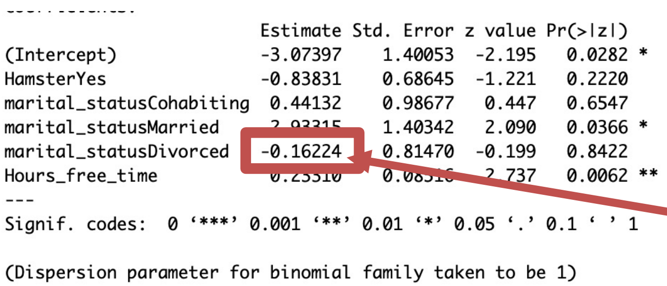
# 5. Evaluating individual predictors
# The marital status variable



```
Coefficients:
                         Estimate Std. Error z value Pr(>|z|)
(Intercept)              -3.07397    1.40053  -2.195   0.0282 *
HamsterYes               -0.83831    0.68645  -1.221   0.2220
marital_statusCohabiting  0.44132    0.98677   0.447   0.6547
marital_statusMarried     2.93315    1.40342   2.090   0.0366 *
marital_statusDivorced   -0.16224    0.81470  -0.199   0.8422
Hours_free_time           0.23310    0.08516   2.737   0.0062 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)
```

The change in the log odds of happy = yes when moving from marital_statusSingle to marital_statusMarried, when holding other variables constant

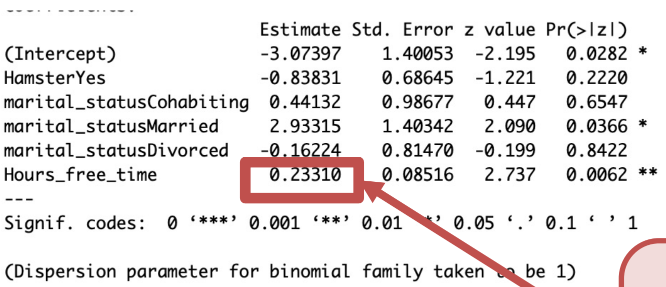
# 5. Evaluating individual predictors
# The marital status variable



```
Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)               -3.07397    1.40053  -2.195   0.0282 *
HamsterYes                -0.83831    0.68645  -1.221   0.2220
marital_statusCohabiting   0.44132    0.98677   0.447   0.6547
marital_statusMarried      2.93315    1.40342   2.090   0.0366 *
marital_statusDivorced    -0.16224    0.81470  -0.199   0.8422
Hours_free_time            0.23310    0.08516   2.737   0.0062 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)
```

The change in the log odds of happy = yes when moving from marital_statusSingle to marital_statusDivorced, when holding other variables constant

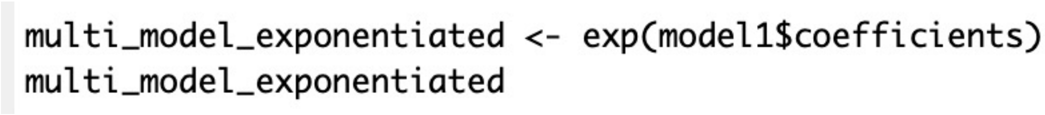
# 5. Evaluating individual predictors
# The hours free time variable



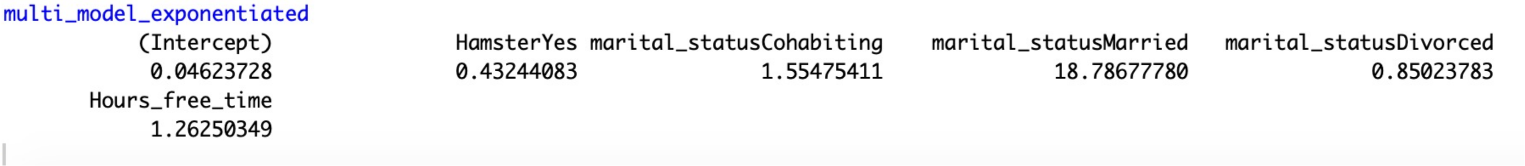
```
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)               -3.07397    1.40053  -2.195   0.0282 *
HamsterYes                -0.83831    0.68645  -1.221   0.2220
marital_statusCohabiting   0.44132    0.98677   0.447   0.6547
marital_statusMarried      2.93315    1.40342   2.090   0.0366 *
marital_statusDivorced    -0.16224    0.81470  -0.199   0.8422
Hours_free_time            0.23310    0.08516   2.737   0.0062 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)
```

The change in the log odds of happy = yes with a one unit increase in hours_free_time, when holding other variables constant

# 5. Evaluating individual predictors
# Exponentiated values

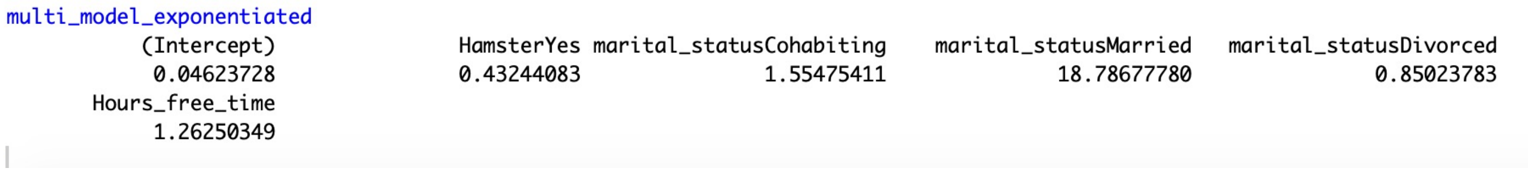


```
multi_model_exponentiated <- exp(model1$coefficients)
multi_model_exponentiated
```



```
multi_model_exponentiated
             (Intercept)                HamsterYes marital_statusCohabiting    marital_statusMarried    marital_statusDivorced
              0.04623728                0.43244083               1.55475411              18.78677780                0.85023783
         Hours_free_time
              1.26250349
```

# 5. Evaluating individual predictors Exponentiated values



```
multi_model_exponentiated
              (Intercept)               HamsterYes marital_statusCohabiting    marital_statusMarried   marital_statusDivorced
               0.04623728               0.43244083               1.55475411              18.78677780               0.85023783
          Hours_free_time
               1.26250349
```

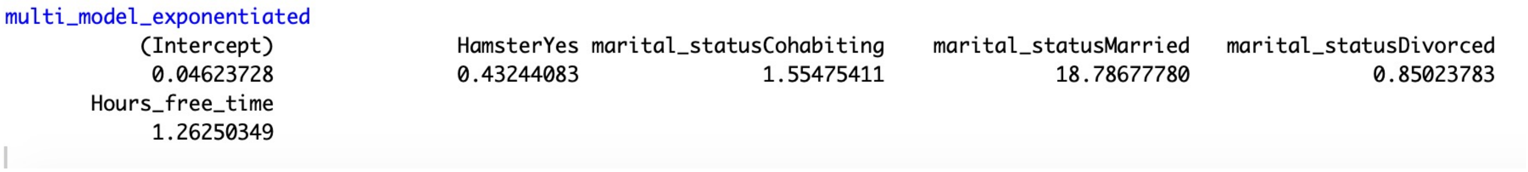
- Intercept: odds that happy = 1, when hamster = no, marital status = single, hours free time = 0

- HamsterYes: Odds ratio: the change in odds when going from HamsterNo to HamsterYes, when holding other variables constant

- marital_statusCohabiting: Odds ratio: the change in odds when going from marital_statusSingle to marital_statusCohabiting, when holding other variables constant

- marital_statusMarried: Odds ratio: the change in odds when going from marital_statusSingle to marital_ statusMarried, when holding other variables constant

# 5. Evaluating individual predictors Exponentiated values



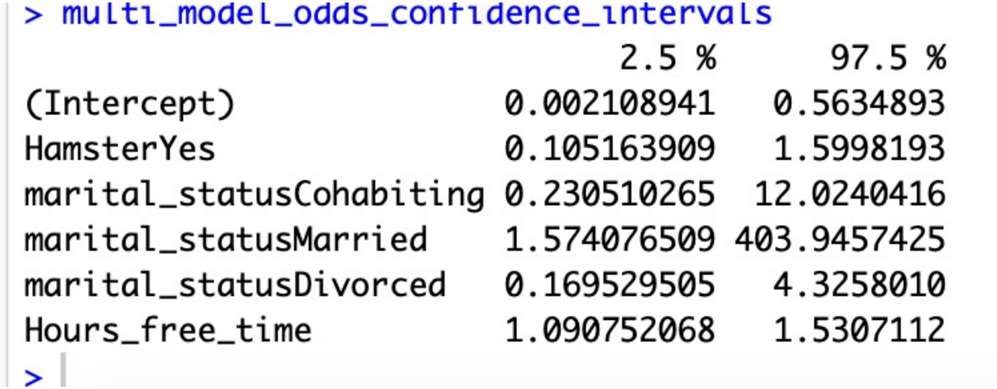
```
multi_model_exponentiated
             (Intercept)               HamsterYes marital_statusCohabiting    marital_statusMarried   marital_statusDivorced
              0.04623728               0.43244083               1.55475411              18.78677780               0.85023783
         Hours_free_time
              1.26250349
```

- marital_statusDivorced: Odds ratio: the change in odds when going from marital_statusSingle to marital_statusDivorced, when holding other variables constant

- Hours_free_time: Odds ratio: the change in odds with a one unit change in the predictor, when holding other variables constant

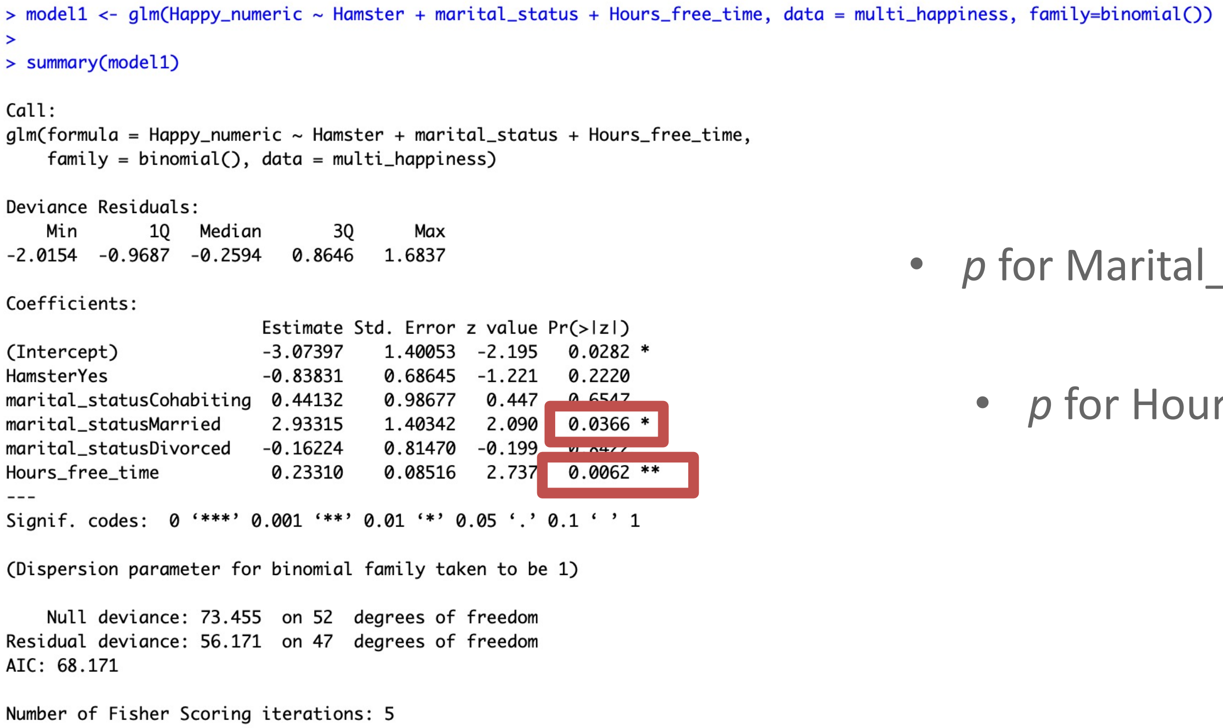
# 5. Odds ratio confidence intervals

```
multi_model_odds_confidence_intervals <- exp(confint(model1))
multi_model_odds_confidence_intervals
```



```
> multi_model_odds_confidence_intervals
                              2.5 %      97.5 %
(Intercept)                0.002108941   0.5634893
HamsterYes                 0.105163909   1.5998193
marital_statusCohabiting   0.230510265  12.0240416
marital_statusMarried      1.574076509 403.9457425
marital_statusDivorced     0.169529505   4.3258010
Hours_free_time            1.090752068   1.5307112
> 
```

# 5. P-values

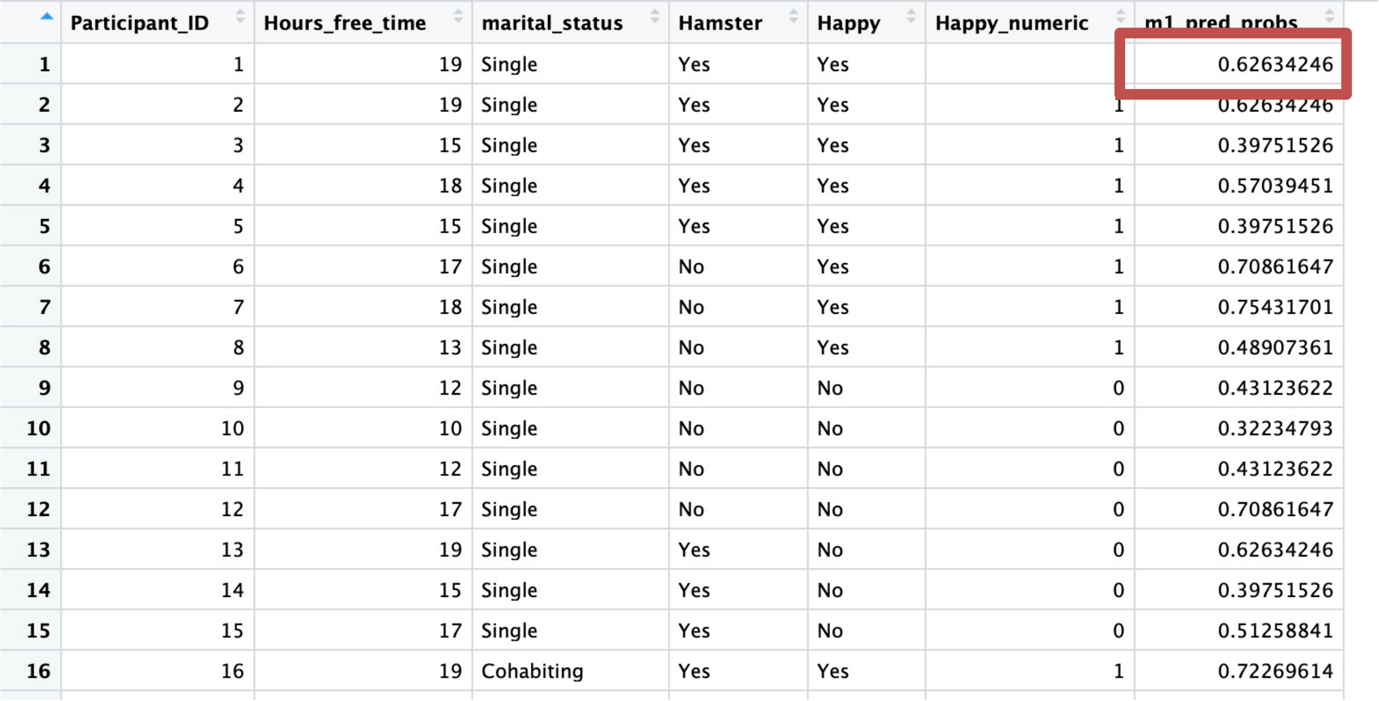


```
> model1 <- glm(Happy_numeric ~ Hamster + marital_status + Hours_free_time, data = multi_happiness, family=binomial())
>
> summary(model1)

Call:
glm(formula = Happy_numeric ~ Hamster + marital_status + Hours_free_time,
    family = binomial(), data = multi_happiness)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0154  -0.9687  -0.2594   0.8646   1.6837

Coefficients:
                         Estimate Std. Error z value Pr(>|z|)
(Intercept)              -3.07397    1.40053  -2.195   0.0282 *
HamsterYes               -0.83831    0.68645  -1.221   0.2220
marital_statusCohabiting  0.44132    0.98677   0.447   0.6547
marital_statusMarried     2.93315    1.40342   2.090   0.0366 *
marital_statusDivorced   -0.16224    0.81470  -0.199   0.8422
Hours_free_time           0.23310    0.08516   2.737   0.0062 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 73.455  on 52  degrees of freedom
Residual deviance: 56.171  on 47  degrees of freedom
AIC: 68.171

Number of Fisher Scoring iterations: 5
```

- *p* for Marital_statusMarried = 0.037

- *p* for Hours_free_time = .006

# 6. Predicted probabilities



| | Participant_ID | Hours_free_time | marital_status | Hamster | Happy | Happy_numeric | m1_pred_probs |
|---|---|---|---|---|---|---|---|
| 1 | 1 | 19 | Single | Yes | Yes | | 0.62634246 |
| 2 | 2 | 19 | Single | Yes | Yes | 1 | 0.62634246 |
| 3 | 3 | 15 | Single | Yes | Yes | 1 | 0.39751526 |
| 4 | 4 | 18 | Single | Yes | Yes | 1 | 0.57039451 |
| 5 | 5 | 15 | Single | Yes | Yes | 1 | 0.39751526 |
| 6 | 6 | 17 | Single | No | Yes | 1 | 0.70861647 |
| 7 | 7 | 18 | Single | No | Yes | 1 | 0.75431701 |
| 8 | 8 | 13 | Single | No | Yes | 1 | 0.48907361 |
| 9 | 9 | 12 | Single | No | No | 0 | 0.43123622 |
| 10 | 10 | 10 | Single | No | No | 0 | 0.32234793 |
| 11 | 11 | 12 | Single | No | No | 0 | 0.43123622 |
| 12 | 12 | 17 | Single | No | No | 0 | 0.70861647 |
| 13 | 13 | 19 | Single | Yes | No | 0 | 0.62634246 |
| 14 | 14 | 15 | Single | Yes | No | 0 | 0.39751526 |
| 15 | 15 | 17 | Single | Yes | No | 0 | 0.51258841 |
| 16 | 16 | 19 | Cohabiting | Yes | Yes | 1 | 0.72269614 |

A lot of variability

Model states there is a probability of 0.63 participant 1 will be happy

Value produced for every case

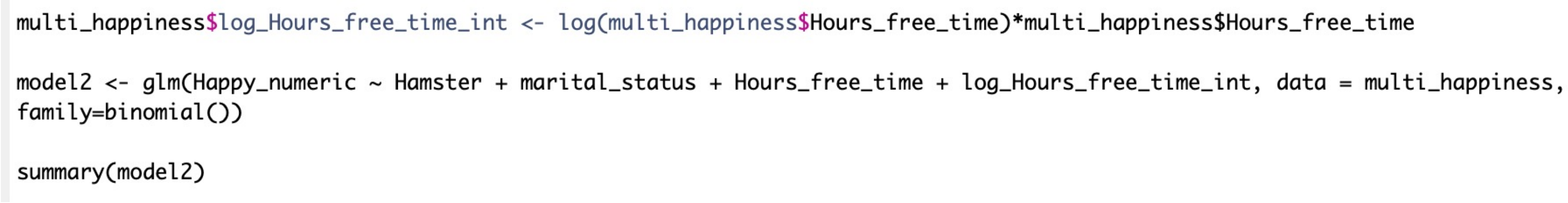
# Assumptions

1. Independence of errors
2. Linearity of the logit (to be checked for **every** continuous predictor)
3. No multicollinearity: Predictor variables should not be highly correlated

# 7. Checking assumptions
# Linearity of the logit

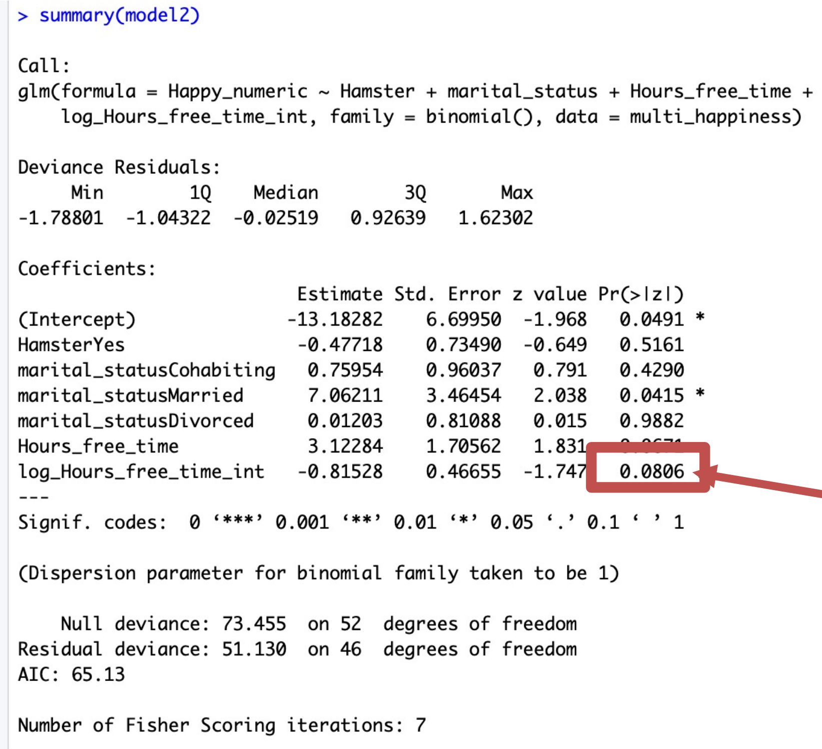
- Needs checking for **every** continuous predictor (here, just Hours_free_time)

- Same code as before:



```
multi_happiness$log_Hours_free_time_int <- log(multi_happiness$Hours_free_time)*multi_happiness$Hours_free_time

model2 <- glm(Happy_numeric ~ Hamster + marital_status + Hours_free_time + log_Hours_free_time_int, data = multi_happiness,
family=binomial())

summary(model2)
```

# 7. Checking assumptions
# Linearity of the logit



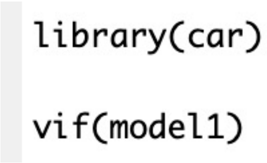
```
> summary(model2)

Call:
glm(formula = Happy_numeric ~ Hamster + marital_status + Hours_free_time + 
    log_Hours_free_time_int, family = binomial(), data = multi_happiness)

Deviance Residuals: 
     Min        1Q    Median        3Q       Max  
-1.78801  -1.04322  -0.02519   0.92639   1.62302  

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)  
(Intercept)              -13.18282    6.69950  -1.968   0.0491 *
HamsterYes                -0.47718    0.73490  -0.649   0.5161  
marital_statusCohabiting   0.75954    0.96037   0.791   0.4290  
marital_statusMarried      7.06211    3.46454   2.038   0.0415 *
marital_statusDivorced     0.01203    0.81088   0.015   0.9882  
Hours_free_time            3.12284    1.70562   1.831   0.0671  
log_Hours_free_time_int   -0.81528    0.46655  -1.747   0.0806  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 73.455  on 52  degrees of freedom
Residual deviance: 51.130  on 46  degrees of freedom
AIC: 65.13

Number of Fisher Scoring iterations: 7
```

- Don’t interpret the model - we look at log_Hours_free_time_int only!!

Not significant – no violation of the assumption of the linearity of the logit

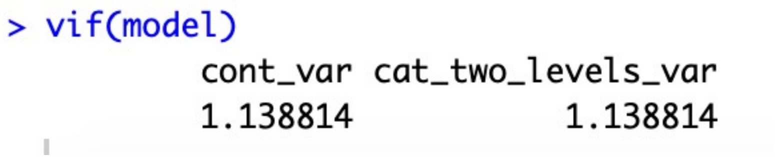
# 7. Checking assumptions
# Multicollinearity - VIF

- No multicollinearity: Predictor variables should not be highly correlated



```
library(car)

vif(model1)
```

- If all variables are continuous, categorical with only two levels, or a combination. This produces vif statistics:

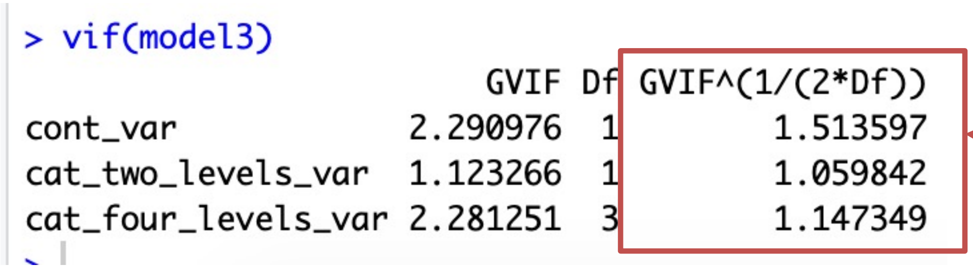


```
> vif(model)
          cont_var cat_two_levels_var
          1.138814           1.138814
```

VIF values above 10 indicate a violation

# 7. Checking assumptions
# Multicollinearity - VIF

- If **any** variable is a categorical with three or more levels, R outputs:



```
> vif(model3)
                        GVIF Df GVIF^(1/(2*Df))
cont_var            2.290976  1        1.513597
cat_two_levels_var  1.123266  1        1.059842
cat_four_levels_var 2.281251  3        1.147349
>
```

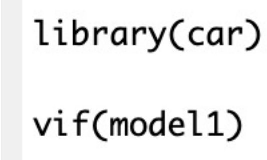
Takes into account degrees of freedom – read this outcome

GVIF$^{(1/(2*Df)}$ is equal to the square root of VIF, so the cut off for GVIF$^{(1/(2*Df)}$ should be the square root of 10

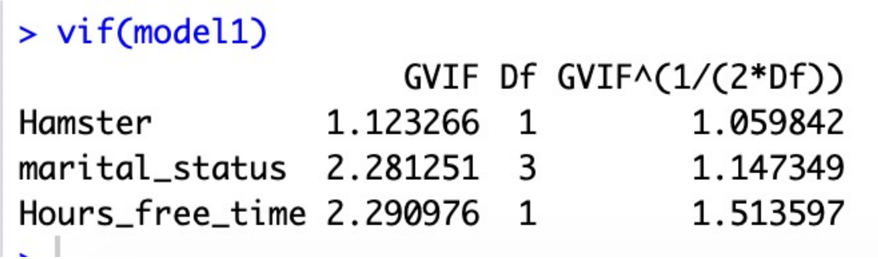
**GVIF$^{(1/(2*Df)}$ values above 3.16 indicate a violation**

# 7. Checking assumptions
# Multicollinearity: on our dataset

- No multicollinearity: Predictor variables should not be highly correlated



```
library(car)

vif(model1)
```



```
> vif(model1)
                   GVIF Df GVIF^(1/(2*Df))
Hamster        1.123266  1        1.059842
marital_status 2.281251  3        1.147349
Hours_free_time 2.290976  1       1.513597
>
```

All values below 3.16 – No evidence of multicollinearity