Regression models with fixed and random effects PSYP14 (HT2020) Assignment 2

Karima Kandi

2020-12-11

#loading necessary packages  
suppressMessages(suppressWarnings(  
 library(psych, quietly = TRUE, warn.conflicts = FALSE)))  
suppressMessages(suppressWarnings(  
 library(lsr, quietly = TRUE, warn.conflicts = FALSE)))  
suppressMessages(suppressWarnings(  
 library(tidyverse, quietly = TRUE, warn.conflicts = FALSE)))  
suppressMessages(suppressWarnings(  
 library(readr, quietly = TRUE, warn.conflicts = FALSE)))  
suppressMessages(suppressWarnings(  
 library(sciplot, quietly = TRUE, warn.conflicts = FALSE)))  
suppressMessages(suppressWarnings(  
 library(reshape2, quietly = TRUE, warn.conflicts = FALSE)))  
suppressMessages(suppressWarnings(  
 library(MASS, quietly = TRUE, warn.conflicts = FALSE)))  
suppressMessages(suppressWarnings(  
 library(dplyr, quietly = TRUE, warn.conflicts = FALSE)))  
suppressMessages(suppressWarnings(  
 library(smacof, quietly = TRUE, warn.conflicts = FALSE)))  
suppressMessages(suppressWarnings(  
 library(lm.beta, quietly = TRUE, warn.conflicts = FALSE)))  
suppressMessages(suppressWarnings(  
 library(gridExtra, quietly = TRUE, warn.conflicts = FALSE)))  
suppressMessages(suppressWarnings(  
 library(magrittr, quietly = TRUE, warn.conflicts = FALSE)))  
suppressMessages(suppressWarnings(  
 library(ggplot2, quietly = TRUE, warn.conflicts = FALSE)))  
suppressMessages(suppressWarnings(  
 library(lmtest, quietly = TRUE, warn.conflicts = FALSE)))  
suppressMessages(suppressWarnings(  
 library(car, quietly = TRUE, warn.conflicts = FALSE)))  
suppressMessages(suppressWarnings(  
 library(sandwich, quietly = TRUE, warn.conflicts = FALSE)))  
suppressMessages(suppressWarnings(  
 library(lmboot, quietly = TRUE, warn.conflicts = FALSE)))

## Assignment 2

### Research question 2

Part 1: perform backwards regression on original data set, cleaned from NAs = data\_ass1\_clean adding new variables to used data set after checking them for outliers

(Recreating the data vectors from the last R Markdown document)

### Data management, data set 1

data\_sample\_1 = read.csv("https://tinyurl.com/ha-dataset1")  
  
my.data <- data.frame(data\_sample\_1$age, data\_sample\_1$sex, data\_sample\_1$pain\_cat, data\_sample\_1$pain,  
 data\_sample\_1$mindfulness, data\_sample\_1$cortisol\_serum, data\_sample\_1$cortisol\_saliva,  
 data\_sample\_1$ID, data\_sample\_1$STAI\_trait, data\_sample\_1$weight,  
 data\_sample\_1$IQ, data\_sample\_1$household\_income, stringsAsFactors = TRUE)  
pain <- as.numeric(my.data$data\_sample\_1.pain)  
pain\_cat <- as.numeric(my.data$data\_sample\_1.pain\_cat)  
age <- as.numeric(my.data$data\_sample\_1.age)  
mindful<- my.data$data\_sample\_1.mindfulness  
cortisol\_serum <- my.data$data\_sample\_1.cortisol\_serum  
cortisol\_saliva<- my.data$data\_sample\_1.cortisol\_saliva  
sex <- data\_sample\_1$sex  
stai\_trait <- data\_sample\_1$STAI\_trait  
id <- my.data$data\_sample\_1.ID  
weight <- as.numeric(my.data$data\_sample\_1.weight)  
IQ <- as.numeric(my.data$data\_sample\_1.IQ)  
income <- as.numeric(my.data$data\_sample\_1.household\_income)  
stai\_clean<- na\_if(stai\_trait, 3.9)  
age\_clean<- na\_if(age, 444)

Data set with all relevant predictors to use in assignment 2

data\_set1 <- data.frame(pain, pain\_cat, age\_clean, mindful, cortisol\_saliva,cortisol\_serum, sex, stai\_clean, id, weight, IQ, income)  
data\_set1\_clean <- na.omit(data\_set1) #omitting NAs

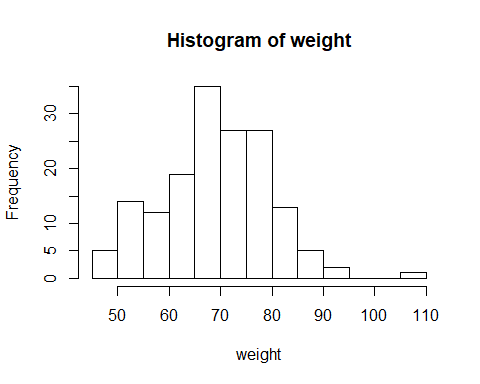
### Data exploration for new variables

summary(data\_set1\_clean)

## pain pain\_cat age\_clean mindful   
## Min. : 2.000 Min. :19.00 Min. :27.00 Min. :1.000   
## 1st Qu.: 4.000 1st Qu.:26.00 1st Qu.:37.00 1st Qu.:2.353   
## Median : 5.000 Median :30.00 Median :40.00 Median :2.940   
## Mean : 5.038 Mean :29.96 Mean :39.99 Mean :2.893   
## 3rd Qu.: 6.000 3rd Qu.:34.00 3rd Qu.:43.00 3rd Qu.:3.500   
## Max. :10.000 Max. :42.00 Max. :53.00 Max. :5.630   
##   
## cortisol\_saliva cortisol\_serum sex stai\_clean id   
## Min. :1.850 Min. :1.920 female:89 Min. :23.00 ID\_1 : 1   
## 1st Qu.:4.452 1st Qu.:4.405 male :69 1st Qu.:37.00 ID\_10 : 1   
## Median :5.055 Median :5.010 Median :41.00 ID\_100 : 1   
## Mean :5.076 Mean :5.064 Mean :40.25 ID\_101 : 1   
## 3rd Qu.:5.838 3rd Qu.:5.680 3rd Qu.:43.75 ID\_102 : 1   
## Max. :7.490 Max. :8.140 Max. :56.00 ID\_103 : 1   
## (Other):152   
## weight IQ income   
## Min. : 47.50 Min. : 49.0 Min. : -3732   
## 1st Qu.: 62.60 1st Qu.: 90.0 1st Qu.: 53600   
## Median : 68.94 Median :101.0 Median : 69819   
## Mean : 68.94 Mean :100.4 Mean : 68876   
## 3rd Qu.: 76.06 3rd Qu.:111.0 3rd Qu.: 86132   
## Max. :105.15 Max. :149.0 Max. :138706   
##

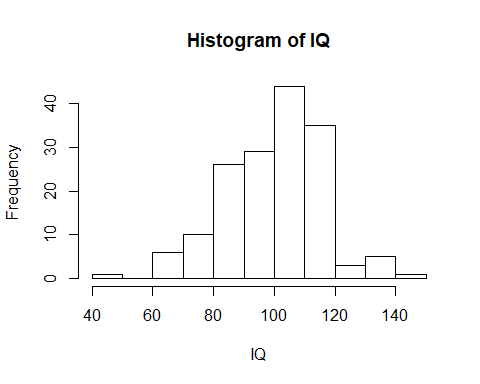
Looking for outliers

#histogram for weight  
hist(weight)



One participant with higher weight than rest of sample, but within reason for weight as a variable.  
Conclusion: keep all participants.

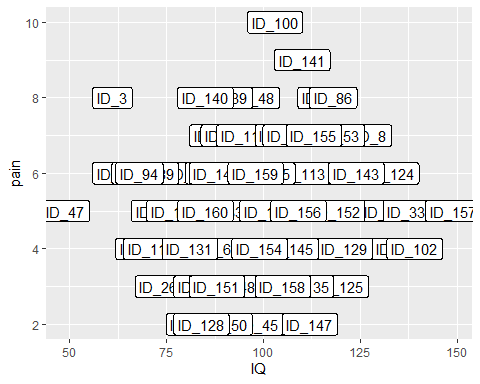
#histogram IQ  
hist(IQ)



One participant has lower IQ than the rest, IQ = 49. Some IQ boundary values in use in research are: <70 = impairment, 100 = average intelligence, >120 = above average, >130 = gifted. If IQ 70 is boundary for impairment, then IQ 49 represents a rather severe impairment or a test error. Also participant with IQ 149 needs to be examined.

Checking the relationship between IQ and pain in a scatterplot

data\_set1\_clean %>% ggplot()+  
 aes(x = IQ, y = pain, label = id)+  
 geom\_label() #id 47 has IQ = 49, id 157 has IQ = 149

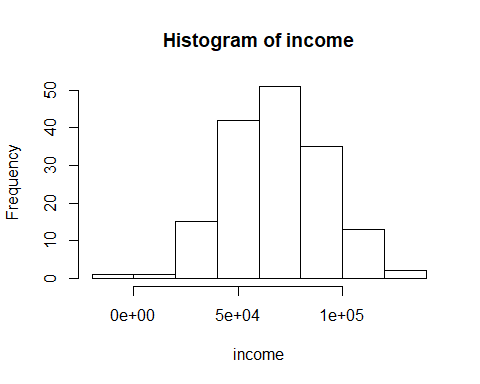


data\_set1\_clean%>% slice(c(47,155)) #slicing participants with id 47 and 157 (row 47 and 155 due to NA exclusions)

## pain pain\_cat age\_clean mindful cortisol\_saliva cortisol\_serum sex  
## 1 5 28 47 2.50 3.94 4.98 male  
## 2 5 30 34 3.43 5.01 4.73 female  
## stai\_clean id weight IQ income  
## 1 30 ID\_47 75.16 49 50473  
## 2 42 ID\_157 71.15 149 100762

Conclusion: With respect to lack of insight for the data collection process, I want to keep as much of the range from IQ as possible. Since both 49 and 149 are possible values for IQ both participants will be kept in this model.

#histogram for income (USD), month or year? Restricted description about income, but assuming it is monthly income from range of values reported   
hist(income)



psych::describe(income) #minimum value is negative, is this due to participants not having an income, living on loans, coding error?

## vars n mean sd median trimmed mad min max range  
## X1 1 160 68683.97 23502.5 69635.5 68889.61 24223.46 -3732 138706 142438  
## skew kurtosis se  
## X1 -0.05 -0.04 1858.04

Conclusion: I am choosing to keep negative values to account for participants that may actually have a negative income.

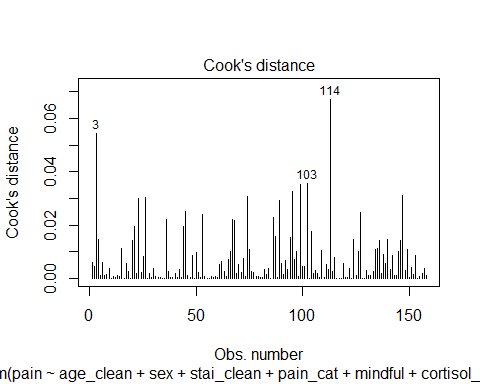
### Creating regression model based on the other researcher’s variables, from now on known as the initial model (named backcomparison\_mod).

backcomparison\_mod <- lm(pain~age\_clean+sex+stai\_clean+pain\_cat+mindful+cortisol\_serum+weight+IQ+ income, data\_set1\_clean)

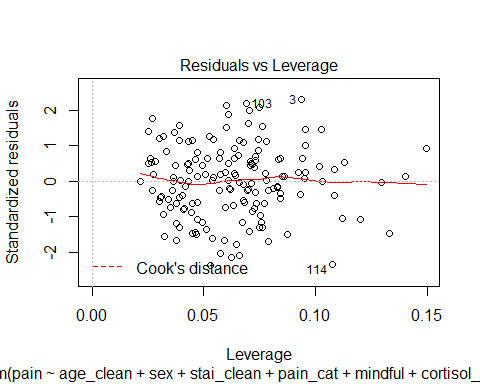
### Leverage for the initial model

Leverage, looking for outliers among residuals that unduly influence the regression coefficients. Cook’s distance examined for the backward model.

#According to cut off value >4/158 = 0.025 cases 114 (and now the highest), and 3 have high leverage that stick out. 103 named but not that deviant.  
backcomparison\_mod %>% plot(which=4)



backcomparison\_mod %>% plot(which=5) #looks fine



Slicing out the participants identified as having the highest leverage to see if they also had extreme values on any of the individual variables.

data\_set1\_clean%>% slice(c(117,114, 103))

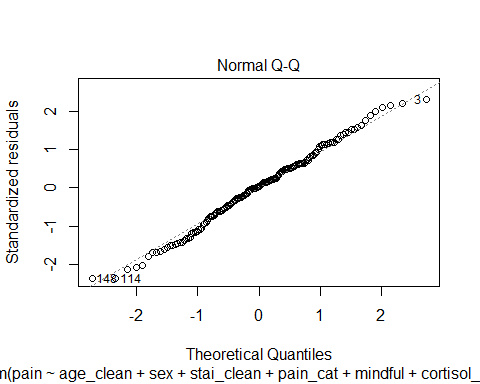
## pain pain\_cat age\_clean mindful cortisol\_saliva cortisol\_serum sex  
## 1 4 27 43 3.88 4.67 4.73 female  
## 2 6 29 43 3.15 5.18 4.86 male  
## 3 5 34 53 1.28 4.28 4.49 female  
## stai\_clean id weight IQ income  
## 1 35 ID\_118 70.71 71 74785  
## 2 38 ID\_115 67.30 101 30358  
## 3 48 ID\_104 55.52 93 54568

Conclusion: These participants were not the people on the edges of the IQ scale, so this is support for keeping all IQ values in the dataset. These participant were not the ones with negative income, so this is support for keeping all income values in the dataset. Leave all participants in because there are no indications of inplausible data/all points looked to be probable data and none had very high undue influence.

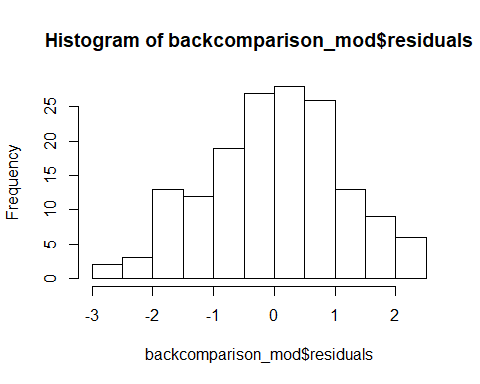
### Regression assumption checks for the initial model

#### Normality of residuals for the initial model

backcomparison\_mod%>% plot(which=2) #shows normality of residuals



hist(backcomparison\_mod$residuals) #acceptably normal distribution



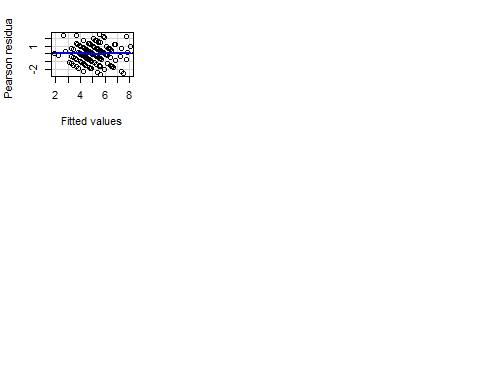
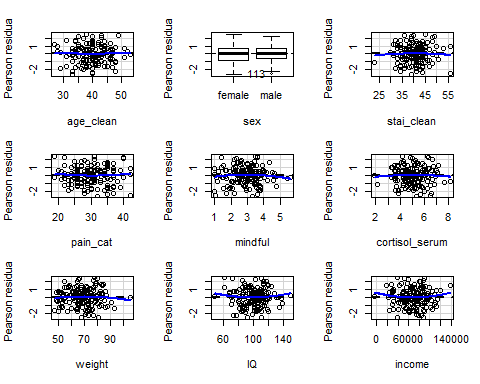
psych::describe(residuals(backcomparison\_mod)) #skew = -0.08, kurtosis = -0.4, no changeworthy values

## vars n mean sd median trimmed mad min max range skew kurtosis se  
## X1 1 158 0 1.1 0.03 0.01 1.01 -2.61 2.48 5.09 -0.08 -0.4 0.09

Conclusion: Model passes checks for normality of residuals

#### Linearity: Residual plot

residualPlots(model = backcomparison\_mod) #no large deviations from linearity

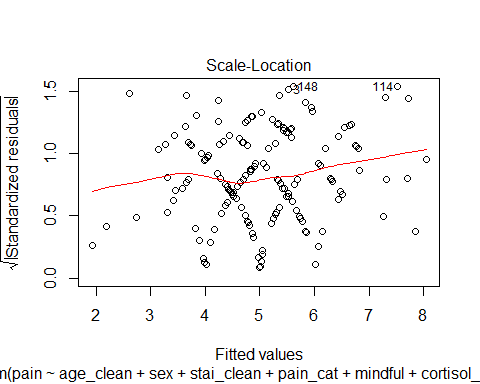


## Test stat Pr(>|Test stat|)  
## age\_clean 0.1554 0.8767  
## sex   
## stai\_clean -0.4199 0.6752  
## pain\_cat 0.6923 0.4898  
## mindful -0.9014 0.3688  
## cortisol\_serum -0.3925 0.6952  
## weight -0.6550 0.5135  
## IQ 0.9321 0.3528  
## income 1.0750 0.2841  
## Tukey test 0.0628 0.9499

Conclusion: Linearity for residuals looks fine, Tukey is not significant and plots look fine.

#### Checking the homoscedasticity assumption for the initial model: heteroscedasticity tests

plot(backcomparison\_mod, which = 3) #no visible heteroscedasticity



ncvTest(backcomparison\_mod) #non-significant test, no heteroscedasticity indicated

## Non-constant Variance Score Test   
## Variance formula: ~ fitted.values   
## Chisquare = 2.397714, Df = 1, p = 0.12151

#breusch test  
bptest(backcomparison\_mod) #non-significant test, no heteroscedasticity indicated

##   
## studentized Breusch-Pagan test  
##   
## data: backcomparison\_mod  
## BP = 10.631, df = 9, p-value = 0.3018

Conclusion: no heteroscedasticity detected

#### Multicollinearity test for the initial model

vif(backcomparison\_mod) #no vifs > 3

## age\_clean sex stai\_clean pain\_cat mindful   
## 1.138626 1.145626 2.324373 2.034996 1.461066   
## cortisol\_serum weight IQ income   
## 1.635656 1.028998 1.022419 1.043833

Conclusion: no multicollinearity detected.

Conclusion from all assumption checks: initial model passes assumption checks for the backwards comparison regression

### First backward regression model to determine elimination of variables

Backward elimination regression using step function

step(object = backcomparison\_mod, direction = "backward")

## Start: AIC=49.42  
## pain ~ age\_clean + sex + stai\_clean + pain\_cat + mindful + cortisol\_serum +   
## weight + IQ + income  
##   
## Df Sum of Sq RSS AIC  
## - IQ 1 0.3256 190.66 47.687  
## - weight 1 0.3646 190.70 47.719  
## - stai\_clean 1 0.4818 190.81 47.816  
## - sex 1 2.2838 192.62 49.301  
## <none> 190.33 49.416  
## - income 1 3.2178 193.55 50.065  
## - age\_clean 1 5.2030 195.54 51.678  
## - mindful 1 7.0469 197.38 53.161  
## - pain\_cat 1 24.3589 214.69 66.444  
## - cortisol\_serum 1 31.1089 221.44 71.335  
##   
## Step: AIC=47.69  
## pain ~ age\_clean + sex + stai\_clean + pain\_cat + mindful + cortisol\_serum +   
## weight + income  
##   
## Df Sum of Sq RSS AIC  
## - weight 1 0.3933 191.05 46.012  
## - stai\_clean 1 0.4453 191.10 46.055  
## - sex 1 2.2302 192.89 47.524  
## <none> 190.66 47.687  
## - income 1 3.3810 194.04 48.464  
## - age\_clean 1 5.3004 195.96 50.019  
## - mindful 1 6.8153 197.47 51.236  
## - pain\_cat 1 24.6937 215.35 64.930  
## - cortisol\_serum 1 30.8419 221.50 69.377  
##   
## Step: AIC=46.01  
## pain ~ age\_clean + sex + stai\_clean + pain\_cat + mindful + cortisol\_serum +   
## income  
##   
## Df Sum of Sq RSS AIC  
## - stai\_clean 1 0.4432 191.50 44.378  
## - sex 1 2.2007 193.25 45.822  
## <none> 191.05 46.012  
## - income 1 3.2418 194.29 46.671  
## - age\_clean 1 5.3389 196.39 48.367  
## - mindful 1 6.7188 197.77 49.473  
## - pain\_cat 1 25.1293 216.18 63.536  
## - cortisol\_serum 1 30.4564 221.51 67.383  
##   
## Step: AIC=44.38  
## pain ~ age\_clean + sex + pain\_cat + mindful + cortisol\_serum +   
## income  
##   
## Df Sum of Sq RSS AIC  
## <none> 191.50 44.378  
## - sex 1 2.926 194.42 44.774  
## - income 1 3.652 195.15 45.363  
## - mindful 1 6.352 197.85 47.534  
## - age\_clean 1 6.450 197.94 47.612  
## - pain\_cat 1 27.340 218.84 63.464  
## - cortisol\_serum 1 33.248 224.74 67.673

##   
## Call:  
## lm(formula = pain ~ age\_clean + sex + pain\_cat + mindful + cortisol\_serum +   
## income, data = data\_set1\_clean)  
##   
## Coefficients:  
## (Intercept) age\_clean sexmale pain\_cat mindful   
## 1.951e+00 -4.175e-02 2.837e-01 1.070e-01 -2.615e-01   
## cortisol\_serum income   
## 5.198e-01 -6.511e-06

#### Running individual regression models to ensure that the recommendations for variable elimination from the backward elimination regression are the optimal eliminations for model fitting.

Initial recommendations:  
Step 1 says to remove IQ, weight, STAI, sex. Step 2 says to remove weight, STAI, sex. Step 3 say to remove STAI, sex. Step 4 removes only STAI, and then it is no longer recommended to remove sex. Model thus indicates that sex+income+mindful+age+pain\_cat+cortisol\_serum is the optimal fit.

Step 1: remove IQ

backcomparison\_mod\_noIQ <- lm(pain~age\_clean+sex+stai\_clean+pain\_cat+mindful+cortisol\_serum+weight+ income, data\_set1\_clean)  
step(object = backcomparison\_mod\_noIQ, direction = "backward")

## Start: AIC=47.69  
## pain ~ age\_clean + sex + stai\_clean + pain\_cat + mindful + cortisol\_serum +   
## weight + income  
##   
## Df Sum of Sq RSS AIC  
## - weight 1 0.3933 191.05 46.012  
## - stai\_clean 1 0.4453 191.10 46.055  
## - sex 1 2.2302 192.89 47.524  
## <none> 190.66 47.687  
## - income 1 3.3810 194.04 48.464  
## - age\_clean 1 5.3004 195.96 50.019  
## - mindful 1 6.8153 197.47 51.236  
## - pain\_cat 1 24.6937 215.35 64.930  
## - cortisol\_serum 1 30.8419 221.50 69.377  
##   
## Step: AIC=46.01  
## pain ~ age\_clean + sex + stai\_clean + pain\_cat + mindful + cortisol\_serum +   
## income  
##   
## Df Sum of Sq RSS AIC  
## - stai\_clean 1 0.4432 191.50 44.378  
## - sex 1 2.2007 193.25 45.822  
## <none> 191.05 46.012  
## - income 1 3.2418 194.29 46.671  
## - age\_clean 1 5.3389 196.39 48.367  
## - mindful 1 6.7188 197.77 49.473  
## - pain\_cat 1 25.1293 216.18 63.536  
## - cortisol\_serum 1 30.4564 221.51 67.383  
##   
## Step: AIC=44.38  
## pain ~ age\_clean + sex + pain\_cat + mindful + cortisol\_serum +   
## income  
##   
## Df Sum of Sq RSS AIC  
## <none> 191.50 44.378  
## - sex 1 2.926 194.42 44.774  
## - income 1 3.652 195.15 45.363  
## - mindful 1 6.352 197.85 47.534  
## - age\_clean 1 6.450 197.94 47.612  
## - pain\_cat 1 27.340 218.84 63.464  
## - cortisol\_serum 1 33.248 224.74 67.673

##   
## Call:  
## lm(formula = pain ~ age\_clean + sex + pain\_cat + mindful + cortisol\_serum +   
## income, data = data\_set1\_clean)  
##   
## Coefficients:  
## (Intercept) age\_clean sexmale pain\_cat mindful   
## 1.951e+00 -4.175e-02 2.837e-01 1.070e-01 -2.615e-01   
## cortisol\_serum income   
## 5.198e-01 -6.511e-06

Step 2: remove IQ and weight

backcomparison\_mod\_noweight\_noIQ <- lm(pain~age\_clean+sex+stai\_clean+pain\_cat+mindful+cortisol\_serum+ income, data\_set1\_clean)  
step(object = backcomparison\_mod\_noweight\_noIQ, direction = "backward")

## Start: AIC=46.01  
## pain ~ age\_clean + sex + stai\_clean + pain\_cat + mindful + cortisol\_serum +   
## income  
##   
## Df Sum of Sq RSS AIC  
## - stai\_clean 1 0.4432 191.50 44.378  
## - sex 1 2.2007 193.25 45.822  
## <none> 191.05 46.012  
## - income 1 3.2418 194.29 46.671  
## - age\_clean 1 5.3389 196.39 48.367  
## - mindful 1 6.7188 197.77 49.473  
## - pain\_cat 1 25.1293 216.18 63.536  
## - cortisol\_serum 1 30.4564 221.51 67.383  
##   
## Step: AIC=44.38  
## pain ~ age\_clean + sex + pain\_cat + mindful + cortisol\_serum +   
## income  
##   
## Df Sum of Sq RSS AIC  
## <none> 191.50 44.378  
## - sex 1 2.926 194.42 44.774  
## - income 1 3.652 195.15 45.363  
## - mindful 1 6.352 197.85 47.534  
## - age\_clean 1 6.450 197.94 47.612  
## - pain\_cat 1 27.340 218.84 63.464  
## - cortisol\_serum 1 33.248 224.74 67.673

##   
## Call:  
## lm(formula = pain ~ age\_clean + sex + pain\_cat + mindful + cortisol\_serum +   
## income, data = data\_set1\_clean)  
##   
## Coefficients:  
## (Intercept) age\_clean sexmale pain\_cat mindful   
## 1.951e+00 -4.175e-02 2.837e-01 1.070e-01 -2.615e-01   
## cortisol\_serum income   
## 5.198e-01 -6.511e-06

Step 3: remove IQ, weight, and STAI

backcomparison\_mod\_noweight\_noIQ\_noSTAI <- lm(pain~age\_clean+sex+pain\_cat+mindful+cortisol\_serum+ income, data\_set1\_clean)  
step(object = backcomparison\_mod\_noweight\_noIQ\_noSTAI , direction = "backward")

## Start: AIC=44.38  
## pain ~ age\_clean + sex + pain\_cat + mindful + cortisol\_serum +   
## income  
##   
## Df Sum of Sq RSS AIC  
## <none> 191.50 44.378  
## - sex 1 2.926 194.42 44.774  
## - income 1 3.652 195.15 45.363  
## - mindful 1 6.352 197.85 47.534  
## - age\_clean 1 6.450 197.94 47.612  
## - pain\_cat 1 27.340 218.84 63.464  
## - cortisol\_serum 1 33.248 224.74 67.673

##   
## Call:  
## lm(formula = pain ~ age\_clean + sex + pain\_cat + mindful + cortisol\_serum +   
## income, data = data\_set1\_clean)  
##   
## Coefficients:  
## (Intercept) age\_clean sexmale pain\_cat mindful   
## 1.951e+00 -4.175e-02 2.837e-01 1.070e-01 -2.615e-01   
## cortisol\_serum income   
## 5.198e-01 -6.511e-06

Conclusion: Model indicates that enough predictors have now been eliminated. Running the next step to ensure that this is correct.

Step 4: remove IQ, weight, STAI, and sex

backcomparison\_mod\_noweight\_noIQ\_noSTAI\_nosex <- lm(pain~age\_clean+pain\_cat+mindful+cortisol\_serum+ income, data\_set1\_clean)  
step(object = backcomparison\_mod\_noweight\_noIQ\_noSTAI\_nosex, direction = "backward")

## Start: AIC=44.77  
## pain ~ age\_clean + pain\_cat + mindful + cortisol\_serum + income  
##   
## Df Sum of Sq RSS AIC  
## <none> 194.42 44.774  
## - income 1 4.4809 198.90 46.374  
## - mindful 1 6.2863 200.71 47.802  
## - age\_clean 1 7.5337 201.95 48.780  
## - pain\_cat 1 27.4880 221.91 63.668  
## - cortisol\_serum 1 30.6973 225.12 65.937

##   
## Call:  
## lm(formula = pain ~ age\_clean + pain\_cat + mindful + cortisol\_serum +   
## income, data = data\_set1\_clean)  
##   
## Coefficients:  
## (Intercept) age\_clean pain\_cat mindful cortisol\_serum   
## 2.382e+00 -4.485e-02 1.073e-01 -2.601e-01 4.901e-01   
## income   
## -7.166e-06

Tested all possible stepwise regressions. AIC indicates that the model with sex+income+mindful+age+pain\_cat+cortisol\_serum (backcomparison\_mod\_noweight\_noIQ\_noSTAI) is the best. This model will be compared the theory based model (mod.saliv).

### Model comparison

Renaming the final backward model to final.backward.mod and examining model test statistics.

final.backward.mod <- backcomparison\_mod\_noweight\_noIQ\_noSTAI  
summary(final.backward.mod)

##   
## Call:  
## lm(formula = pain ~ age\_clean + sex + pain\_cat + mindful + cortisol\_serum +   
## income, data = data\_set1\_clean)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.67319 -0.70510 0.05705 0.70096 2.41414   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.951e+00 1.356e+00 1.439 0.1522   
## age\_clean -4.175e-02 1.851e-02 -2.255 0.0256 \*   
## sexmale 2.837e-01 1.868e-01 1.519 0.1309   
## pain\_cat 1.070e-01 2.304e-02 4.643 7.40e-06 \*\*\*  
## mindful -2.615e-01 1.168e-01 -2.238 0.0267 \*   
## cortisol\_serum 5.198e-01 1.015e-01 5.120 9.17e-07 \*\*\*  
## income -6.511e-06 3.837e-06 -1.697 0.0917 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.126 on 151 degrees of freedom  
## Multiple R-squared: 0.501, Adjusted R-squared: 0.4812   
## F-statistic: 25.27 on 6 and 151 DF, p-value: < 2.2e-16

Test statistics for the initial model (model submitted to backward regression)

summary(backcomparison\_mod)

##   
## Call:  
## lm(formula = pain ~ age\_clean + sex + stai\_clean + pain\_cat +   
## mindful + cortisol\_serum + weight + IQ + income, data = data\_set1\_clean)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.60799 -0.69461 0.02898 0.68745 2.47778   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2.227e+00 1.633e+00 1.363 0.1749   
## age\_clean -3.862e-02 1.920e-02 -2.011 0.0461 \*   
## sexmale 2.595e-01 1.947e-01 1.333 0.1847   
## stai\_clean -1.653e-02 2.701e-02 -0.612 0.5414   
## pain\_cat 1.120e-01 2.573e-02 4.352 2.5e-05 \*\*\*  
## mindful -2.802e-01 1.197e-01 -2.341 0.0206 \*   
## cortisol\_serum 5.588e-01 1.136e-01 4.918 2.3e-06 \*\*\*  
## weight -4.710e-03 8.845e-03 -0.532 0.5952   
## IQ 2.849e-03 5.662e-03 0.503 0.6156   
## income -6.201e-06 3.920e-06 -1.582 0.1158   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.134 on 148 degrees of freedom  
## Multiple R-squared: 0.504, Adjusted R-squared: 0.4739   
## F-statistic: 16.71 on 9 and 148 DF, p-value: < 2.2e-16

#### Model comparison using AIC for the backward model and the initial model

AIC for the backward model

AIC(final.backward.mod)

## [1] 494.7628

AIC for the initial model

AIC(backcomparison\_mod)

## [1] 499.801

Conclusion: The AIC for the initial model before backwards regression is the highest.

Model comparison of initial model and backward model using anova

anova(final.backward.mod, backcomparison\_mod)

## Analysis of Variance Table  
##   
## Model 1: pain ~ age\_clean + sex + pain\_cat + mindful + cortisol\_serum +   
## income  
## Model 2: pain ~ age\_clean + sex + stai\_clean + pain\_cat + mindful + cortisol\_serum +   
## weight + IQ + income  
## Res.Df RSS Df Sum of Sq F Pr(>F)  
## 1 151 191.50   
## 2 148 190.33 3 1.1621 0.3012 0.8245

Conclusion: The initial model does not explain significantly more variance than the backward model.  
Both AIC and anova indicate that the backward model is preferable.

### Model comparison for the backward model and the theory based model using AIC values

Renaming the theory based model and examining model test statistics.

#just re-running the first model to keep it in this markdown document  
data\_ass1 <- data.frame(pain, pain\_cat, age\_clean, mindful, cortisol\_saliva,cortisol\_serum, sex, stai\_clean, id)  
data\_ass1\_clean <- na.omit(data\_ass1)  
  
mod.saliv <- lm(pain~age\_clean+sex+stai\_clean+pain\_cat+mindful+cortisol\_saliva, data\_ass1\_clean)   
theorybased.mod <- mod.saliv #renaming mod.saliv to theorybased.mod  
  
summary(theorybased.mod)

##   
## Call:  
## lm(formula = pain ~ age\_clean + sex + stai\_clean + pain\_cat +   
## mindful + cortisol\_saliva, data = data\_ass1\_clean)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.48458 -0.75172 0.06268 0.74735 2.40932   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.115755 1.469006 0.079 0.9373   
## age\_clean -0.006779 0.020775 -0.326 0.7446   
## sexmale 0.228440 0.192723 1.185 0.2378   
## stai\_clean -0.038859 0.027931 -1.391 0.1662   
## pain\_cat 0.139007 0.025427 5.467 1.85e-07 \*\*\*  
## mindful -0.227184 0.119007 -1.909 0.0582 .   
## cortisol\_saliva 0.620719 0.120191 5.164 7.50e-07 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.128 on 151 degrees of freedom  
## Multiple R-squared: 0.4993, Adjusted R-squared: 0.4794   
## F-statistic: 25.09 on 6 and 151 DF, p-value: < 2.2e-16

AIC for the theory based model

AIC(theorybased.mod)

## [1] 495.3192

AIC(final.backward.mod)

## [1] 494.7628

Conclusion: Model comparison between the backward model and the theory based model using AIC shows that the final backward model has the smallest value. Model comparison using anova not appropriate given non-nested models.

### Model test statistics

Model statistics from ANOVA/summary

backcomparison\_mod adjusted R^2 value = 0.47, F(9,148) = 16.71, p < .001.

final.backward.mod adjusted R^2 value = 0.48, F(6,151) = 25.27, p < .001.

theorybased.mod adjusted R^2 value = 0.48, F(6,151) = 25.09, p < .001

### Regression results table for the backward model

coef\_table = function(model){   
 require(lm.beta)   
 mod\_sum = summary(model)   
 mod\_sum\_p\_values = as.character(round(mod\_sum$coefficients[,4], 3))   
 mod\_sum\_p\_values[mod\_sum\_p\_values != "0" & mod\_sum\_p\_values != "1"] = substr(mod\_sum\_p\_values[mod\_sum\_p\_values != "0" & mod\_sum\_p\_values != "1"], 2, nchar(mod\_sum\_p\_values[mod\_sum\_p\_values != "0" & mod\_sum\_p\_values != "1"]))   
 mod\_sum\_p\_values[mod\_sum\_p\_values == "0"] = "<.001"   
   
   
 mod\_sum\_table = cbind(as.data.frame(round(cbind(coef(model), confint(model), c(0, lm.beta(model)$standardized.coefficients[c(2:length(model$coefficients))])), 2)), mod\_sum\_p\_values)   
 names(mod\_sum\_table) = c("b", "95%CI lb", "95%CI ub", "Std.Beta", "p-value")   
 mod\_sum\_table["(Intercept)","Std.Beta"] = "0"   
 return(mod\_sum\_table)   
}   
  
coef\_table(final.backward.mod)

## b 95%CI lb 95%CI ub Std.Beta p-value  
## (Intercept) 1.95 -0.73 4.63 0 .152  
## age\_clean -0.04 -0.08 -0.01 -0.13 .026  
## sexmale 0.28 -0.09 0.65 0.09 .131  
## pain\_cat 0.11 0.06 0.15 0.34 <.001  
## mindful -0.26 -0.49 -0.03 -0.15 .027  
## cortisol\_serum 0.52 0.32 0.72 0.34 <.001  
## income 0.00 0.00 0.00 -0.1 .092

Note: Due to rounding, the coefficient for income does not show up in this table. The coefficient for income is 0.000006, it is small due one unit change representing only 1 USD

Conclusion: When considering the statistical model comparison, the backward model seems to perform better in AIC and slightly better in its adjusted R^2: thus confirming the other researcher’s claim that the backwards model explains slightly more variance in postoperative pain.

However, since model comparison using hierarchical regression between the final backward model and the theory based model wasn’t appropriate (due to them not being nested) we cannot definitely determine which is more effective in predicting pain from this dataset as we would have liked. That said, it remains that, if given a choice the backward model performs slightly better based on R^2 and AIC value. Yet, the theory-based model has the advantage of being based on theoretical variables previously identified as important and meaningful when predicting postoperative pain. Since the explanatory value of the different models will differ when applied on different samples; that is, if we want to use the models on new data, we need to apply the existing models’ predictions in a new data set and therefor explore the comparison further in this regard too.

### Assignment 2, Part 2: Testing the models on a new dataset

Importing new dataset to test both models on

data\_sample\_2 = read.csv("https://tinyurl.com/ha-dataset2")

For predictions of postoperative pain in the new data set (data\_sample\_2), the following two regression equations will be used:

Regression equation for the backward model 𝑌 = 1.95 + (-0.04 \* age\_clean) + (0.28 \* sexmale) + (0.11 \* pain\_cat)+ (-0.26 \* mindful)+ (0.52 \* cortisol\_serum)+ (0.000006 \* income)

Regression equation for the theory based model 𝑌 = 0.12 + (-0.01 \* age\_clean) + (0.23 \* sexmale) +(-0.04 \* stai\_clean) + (0.14 \* pain\_cat)+ (-0.23 \* mindful)+ (0.62 \* cortisol\_saliva)

Using the backwards model and the theory based model to predict pain values in the new dataset

predictbackward <- predict(final.backward.mod, data\_sample\_2)  
predicttheory <- predict(theorybased.mod, data\_sample\_2)

#### Comparing the predicted values from the backward model to the actual pain values in the new dataset using the residual sum of squares

Calculating residual sum of squares for the backward model in the new dataset

rss.test.backward <- sum((data\_sample\_2$pain-predictbackward)^2, na.rm = TRUE)  
print(rss.test.backward)

## [1] 257.442

#### Comparing the predicted values from the theory based model to the actual pain values in the new dataset using the residual sum of squares

Calculating residual sum of squares for the theory based model in the new dataset

rss.test.theory<- sum((data\_sample\_2$pain-predicttheory)^2, na.rm = TRUE)  
print(rss.test.theory)

## [1] 246.7904

#### Comparing variance explained by each model in the new dataset

Calculating R^2 for the backward model in the new dataset

## total sum of squares (difference for each participant from grand mean) calculated from intercept for backward model   
mod.mean <- lm(pain~1,data = data\_sample\_2)  
tss <- sum((data\_sample\_2$pain-predict(mod.mean))^2)  
  
(Rsqr2 = 1 - (rss.test.backward/tss))

## [1] 0.2514046

Calculating R^2 for the theory based model in the new dataset

(Rsqr2 = 1 - (rss.test.theory/tss))

## [1] 0.2823774

Reminder of the models’ adjusted R^2 values from dataset 1

summary(final.backward.mod)$adj.r.squared

## [1] 0.4811918

summary(theorybased.mod)$adj.r.squared

## [1] 0.4793616

Conclusion: Looking first at residual sum of squares for both models on the new data set, the theory based model has lower residual errors than the backward model has. Looking instead at R^2 for both models on the new data set, the theory based model has better fit to new dataset, R^2 = 0.28 compared to R^2 = 0.25 for backward model. The theory based model thus explains around 3% more of the variance in postoperative pain in the new data set than the backward model does. R^2, compared to raw residual errors, also takes into account the variance between participants in the new data set and gives a more easily comparable value between models.

### Summary

For data set 1, the backward model explained around 0.2% more variance in postoperative pain than the theory based model did. However, in the new dataset, the theory based model explained around 3% more variance than the backward model did. This difference between data sets could be taken as a sign that basing a model on theory rather than data approaches appear more suitable for use on new data.

Arguably, we are likely seeing that the backward model perhaps is slight overfitted to data set 1 as well; making the theory based model more useful across datasets when predicting postoperative pain. This argument could be provided as a response to the researcher in the assignment who preferred the data based model. However, the value of the models should also be related to the intention of use. Say, I wanted to get the best fit to the specific clinical participants in sample 1, the backward model would arguably be better fitted to that data. However, if we are interested in assuring the best prediction performance across new patient groups for postoperative pain, the theory based model presents itself as more useful in that regard.

#### Reporting regression equations

Regression equation for backward model 𝑌 = 1.95 + (-0.04 \* age\_clean) + (0.28 \* sexmale) + (0.11 \* pain\_cat)+ (-0.26 \* mindful)+ (0.52 \* cortisol\_serum)+ (0.000006 \* income)

Regression equation for theory based model 𝑌 = 0.12 + (-0.01 \* age\_clean) + (0.23 \* sexmale) +(-0.04 \* stai\_clean) + (0.14 \* pain\_cat)+ (-0.23 \* mindful)+ (0.62 \* cortisol\_saliva)