# Project 2: Regression Analysis

Yingjie Huang

## Abstract

This study explores baseline variables as moderators and predictors of smoking cessation in adults with major depressive disorder (MDD), using data from a randomized, placebo-controlled trial that assessed the effects of behavioral activation for smoking cessation (BASC) versus standard treatment (ST) and varenicline versus placebo. Stepwise logit and lasso regression models were applied to identify moderators of treatment effects on end-of-treatment (EOT) abstinence and predictors of cessation, controlling for treatment and pharmacotherapy. Lasso regression identified Nicotine Metabolism as a moderator of pharmacotherapy, while stepwise regression found education can moderate psychotherapy effects. Additionally, stepwise regression identified nicotine metabolism, non-Hispanic white ethnicity, and smoking within 5 minutes of waking up as predictors of abstinence, while lasso regression only detected the latter behavior. These findings highlight the importance of baseline variables in smoking cessation and suggest that further research is needed to better understand their interactions with treatment and pharmacotherapy.

## Introduction

Smoking is a major public health concern, particularly among individuals with major depressive disorder (MDD). Research has shown that individuals with MDD are more likely to smoke heavily, exhibit greater nicotine dependence, and experience more severe withdrawal symptoms compared to those without MDD. While varenicline has proven effective in promoting smoking cessation, it remains unclear whether addressing the depression-related psychological factors associated with smoking can further enhance cessation outcomes. A previous randomized, placebo-controlled study employing a 2x2 factorial design investigated the impact of behavioral activation for smoking cessation (BASC) versus standard behavioral treatment (ST) in combination with varenicline or placebo on smoking cessation rates among 300 adult smokers with current or past MDD. The study concluded that BASC did not outperform ST, with or without varenicline therapy.

The goal of the present project is to analyze data from this clinical trial to identify baseline variables that moderate the effects of behavioral treatment on end-of-treatment (EOT) abstinence, while also evaluating these variables as predictors of abstinence, controlling for the influence of behavioral treatment and pharmacotherapy. By applying stepwise logit regression and lasso regression, this project explores how different statistical methods can inform our understanding of baseline characteristics and their interaction with treatment approaches. The findings from these analyses will contribute to understanding the complexities of smoking cessation in individuals with MDD and inform future interventions.

## Exploratory Data Analysis

The summary statistics, generated by Table 1, provide an overview of all variables both overall and grouped by abstinence status. Pearson's chi-square tests were conducted to assess the differences between participants with and without abstinence. The results indicate that variables such as `Var`, `NHW`, `ftcd_score`, and `NMR` show statistically significant differences between the groups. These variables represent factors like pharmacotherapy use, non-Hispanic white ethnicity, smoking habits upon waking, and nicotine metabolism, respectively. These findings may help identify potential predictors of abstinence.

### Summary of Variables

Table 1: Summary Table of Variables

| **Characteristic** | **Overall** N = 300[1] | **No** N = 236[1] | **Yes** N = 64[1] | **p-value**[2] |
|---|---|---|---|---|
| Var | | | | <0.001 |
| 0 | 136/300 (45%) | 124/236 (53%) | 12/64 (19%) | |
| 1 | 164/300 (55%) | 112/236 (47%) | 52/64 (81%) | |
| BA | | | | 0.5 |
| 0 | 149/300 (50%) | 115/236 (49%) | 34/64 (53%) | |
| 1 | 151/300 (50%) | 121/236 (51%) | 30/64 (47%) | |
| age_ps | [19.00, 76.00] | [19.00, 76.00] | [23.00, 72.00] | 0.8 |
| sex_ps | | | | 0.8 |
| 1 | 135/300 (45%) | 107/236 (45%) | 28/64 (44%) | |
| 2 | 165/300 (55%) | 129/236 (55%) | 36/64 (56%) | |
| Black | | | | 0.12 |
| 0 | 143/300 (48%) | 107/236 (45%) | 36/64 (56%) | |
| 1 | 157/300 (52%) | 129/236 (55%) | 28/64 (44%) | |
| Hisp | | | | 0.8 |
| 0 | 282/300 (94%) | 221/236 (94%) | 61/64 (95%) | |

| | | | | |
|---|---|---|---|---|
| 1 | 18/300 (6.0%) | 15/236 (6.4%) | 3/64 (4.7%) | |
| inc | | | | 0.6 |
| 1 | 110/297 (37%) | 88/234 (38%) | 22/63 (35%) | |
| 2 | 68/297 (23%) | 56/234 (24%) | 12/63 (19%) | |
| 3 | 46/297 (15%) | 36/234 (15%) | 10/63 (16%) | |
| 4 | 38/297 (13%) | 30/234 (13%) | 8/63 (13%) | |
| 5 | 35/297 (12%) | 24/234 (10%) | 11/63 (17%) | |
| Unknown | 3 | 2 | 1 | |
| edu | | | | 0.13 |
| 1 | 1/300 (0.3%) | 0/236 (0%) | 1/64 (1.6%) | |
| 2 | 16/300 (5.3%) | 13/236 (5.5%) | 3/64 (4.7%) | |
| 3 | 76/300 (25%) | 60/236 (25%) | 16/64 (25%) | |
| 4 | 116/300 (39%) | 97/236 (41%) | 19/64 (30%) | |
| 5 | 91/300 (30%) | 66/236 (28%) | 25/64 (39%) | |
| ftcd_score | 5.22 (2.14) | 5.46 (1.98) | 4.34 (2.46) | 0.002 |
| Unknown | 1 | 1 | 0 | |
| ftcd.5.mins | | | | 0.2 |
| 0 | 162/300 (54%) | 123/236 (52%) | 39/64 (61%) | |
| 1 | 138/300 (46%) | 113/236 (48%) | 25/64 (39%) | |
| bdi_score_w00 | 18.72 (11.47) | 19.14 (11.54) | 17.19 (11.19) | 0.2 |
| cpd_ps | 15.15 (7.89) | 15.57 (7.81) | 13.58 (8.07) | 0.052 |
| crv_total_pq1 | 7.19 (3.70) | 7.21 (3.71) | 7.09 (3.71) | >0.9 |
| Unknown | 18 | 12 | 6 | |
| hedonsum_n_pq1 | 22.63 (19.60) | 21.90 (18.90) | 25.31 (21.96) | 0.4 |
| hedonsum_y_pq1 | 25.43 (19.42) | 25.97 (19.27) | 23.47 (20.00) | 0.15 |
| shaps_score_pq1 | 2.25 (3.16) | 2.42 (3.36) | 1.64 (2.22) | 0.3 |
| Unknown | 3 | 3 | 0 | |
| otherdiag | | | | 0.2 |
| 0 | 167/300 (56%) | 127/236 (54%) | 40/64 (63%) | |
| 1 | 133/300 (44%) | 109/236 (46%) | 24/64 (38%) | |
| antidepmed | | | | 0.9 |
| 0 | 218/300 (73%) | 171/236 (72%) | 47/64 (73%) | |
| 1 | 82/300 (27%) | 65/236 (28%) | 17/64 (27%) | |
| mde_curr | | | | 0.073 |
| 0 | 153/300 (51%) | 114/236 (48%) | 39/64 (61%) | |
| 1 | 147/300 (49%) | 122/236 (52%) | 25/64 (39%) | |
| NMR | -1.19 (0.61) | -1.23 (0.61) | -1.02 (0.55) | 0.023 |
| Unknown | 21 | 15 | 6 | |
| Only.Menthol | | | | 0.4 |
| 0 | 120/298 (40%) | 91/234 (39%) | 29/64 (45%) | |

| | | | | |
|---|---|---|---|---|
| 1 | 178/298 (60%) | 143/234 (61%) | 35/64 (55%) | |
| Unknown | 2 | 2 | 0 | |
| readiness | 6.78 (1.24) | 6.80 (1.26) | 6.68 (1.17) | 0.4 |
| Unknown | 17 | 12 | 5 | |

[1]n/N (%); [Min, Max]; Mean (SD)
[2]Pearson's Chi-squared test; Wilcoxon rank sum test; Fisher's exact test

## Correlation between Variables

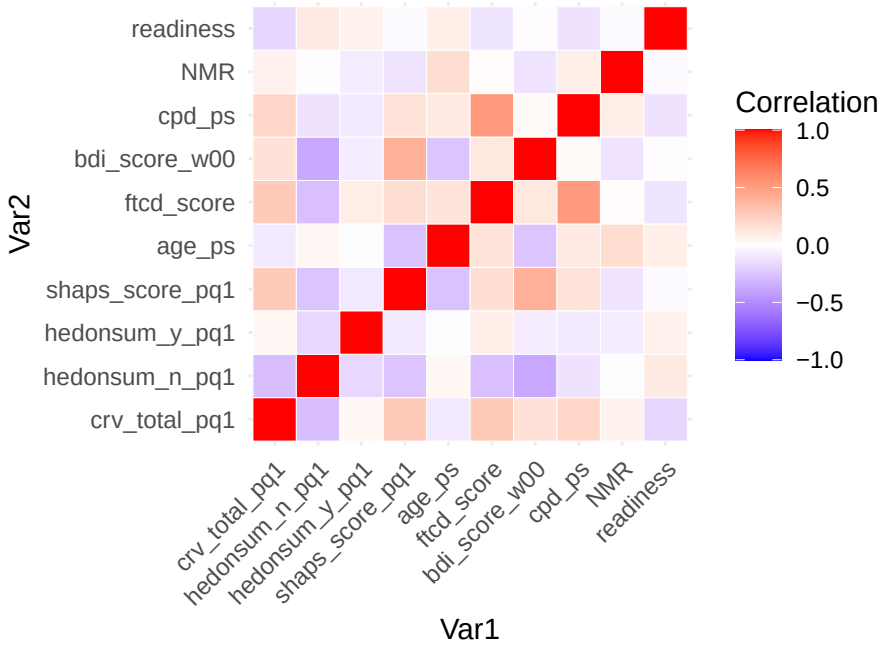


Figure 1: Correlation between Continuous Variables

To check for multicollinearity and potential associations between variables, we calculate Pearson correlation for continuous variables and Cramér's V for categorical variables. As Figure 1 suggests, all continuous variables show little correlation with each other, indicating that multicollinearity is unlikely to be a concern in our analysis. This suggests that each continuous variable conveys distinct and valuable information about the participants, whether physical or mental, without redundancy.

For categorical variables, we introduced another metric called Cramér's V, which measures the strength of association between two categorical variables. It is based on the chi-square

statistic and is adjusted for the number of categories in each variable, providing a standardized measure of association. It is calculated using the following formula:

$$V = \sqrt{\frac{\chi^2}{n \cdot (k-1)}}$$

Where: $\chi^2$ is the chi-square statistic, calculated from the contingency table of the two categorical variables. $n$ is the total number of observations (i.e., the total sample size). $k$ is the number of categories in the variable with fewer categories.

It ranges from 0 to 1, and higher values suggest a stronger relationship between the variables. Cramér's V works well with variables that have multiple categories and is not affected by the number of categories in each variable, making it a good choice for categorical data. Thus can be used to inform how these variables may interact in predictive models.

Figure 2 shows similar results, Most of categorical variables have little correlation, While `Black` and `NHW` showed a strong association, with their Cramér's V value reaching approximately 0.7, this can be attributed to the fact that both variables represent indicators of ethnicity. There is considerable overlap or mutual complementarity between the groups they refer to, which results in a high degree of association.

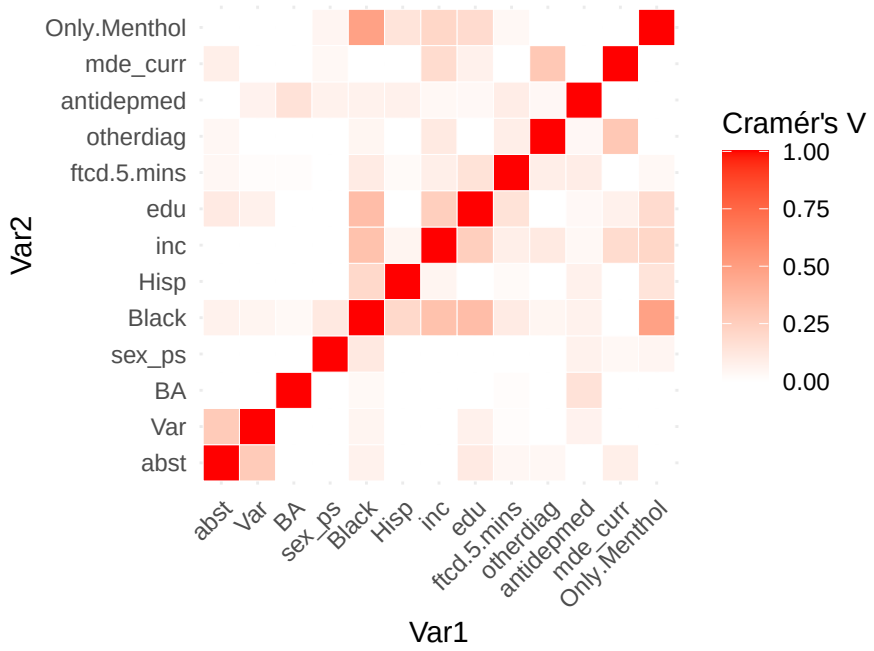


Figure 2: Correlation between Categorical Variables

## Data Preprocessing

It should be noted that there is mild missingness in the data. To address this issue, we employed a multiple imputation (MI) approach using the mice() function with 100 imputations. The imputed datasets were then combined using the complete() function to generate a stable, imputed dataset.

Additionally, to evaluate the performance of different models, we randomly split the data, using 80% for training and 20% for testing. This approach allows for a more objective assessment of their predictive accuracy.

## Model fitting

Since the goal is to identify moderators of treatment effects on end-of-treatment (EOT) abstinence and predictors of cessation while controlling for treatment and pharmacotherapy, we begin by considering the main effects of all baseline variables, adjusted for treatment variables, including psychotherapy (Behavioral Activation) and pharmacotherapy (Varenicline). Moreover, we examine the interactions between these baseline variables and the treatment variables. The model is represented by the following formula:

$$\text{logit}(y_i) = \beta_0 + x_i^\top w_1 + A_i\beta_1 + Z_i\beta_2 + x_i^\top A_i w_2 + x_i^\top Z_i w_3 + \epsilon_i$$

where $x_i$ is a $p \times 1$ vector containing baseline variables. $A_i$ and $Z_i$ are dummy variables indicating whether participants received psychotherapy or pharmacotherapy, respectively. $y_i$ in current analysis stands for participants' smoking abstinence.

This complex formula results in an excessive number of coefficients to estimate—85 in our case. Therefore, it is crucial to perform variable selection to identify the simplest model that best captures the associations between these variables. We considered two approaches: stepwise logistic regression and lasso regression. Both methods achieve variable selection through different mechanisms. We will implement and compare their performance, seeking evidence to support the variables each method selects.

For lasso regression, we applied a cross-validation scheme to avoid overfitting and to select the optimal $\lambda$ value. Specifically, we chose the model with lambda.1se, which selects the most regularized model within one standard error of the minimum cross-validated error. This choice provides a simpler model that balances predictive accuracy with reduced complexity, helping to prevent overfitting. Additionally, ridge regression was performed as a benchmark for comparison. Similar cross-validtion scheme was run to help find the most appropriate $\lambda$ for ridge regression.

For logistic regression, we used stepwise selection to simplify our model. This approach, implemented via the `step()` function, iteratively adds or removes predictors based on their

statistical significance to optimize model fit. Specifically, `step()` uses Akaike Information Criterion (AIC) as a criterion: at each step, it evaluates whether adding or removing a variable improves the model's AIC. By minimizing AIC, stepwise selection aims to balance model simplicity with explanatory power.

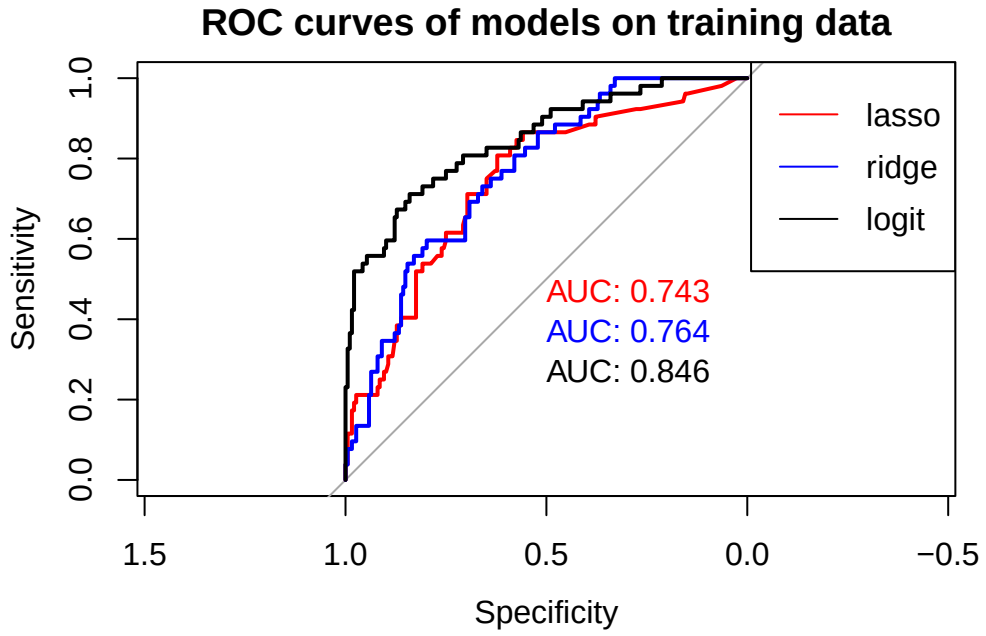


Figure 3: ROC Curves of Models on Training Data

As shown in Figure 3, logistic regression achieved the highest performance among the models on the training dataset, with an AUC of 0.832, while both lasso and ridge regression produced similar, slightly lower results. However, Figure 4 reveals a different pattern on the testing dataset: lasso regression achieved the highest AUC, whereas logistic regression performed the worst. This contrast suggests that lasso regression has a strong generalization ability, effectively avoiding overfitting and maintaining performance on unseen data. In contrast, logistic regression may have overfit to noise within the training data, leading to a decline in performance on the test set. Using AUC as our evaluation metric, we concluded that lasso regression appears to be a more robust and reliable option for predicting outcomes than logistic regression in this context.

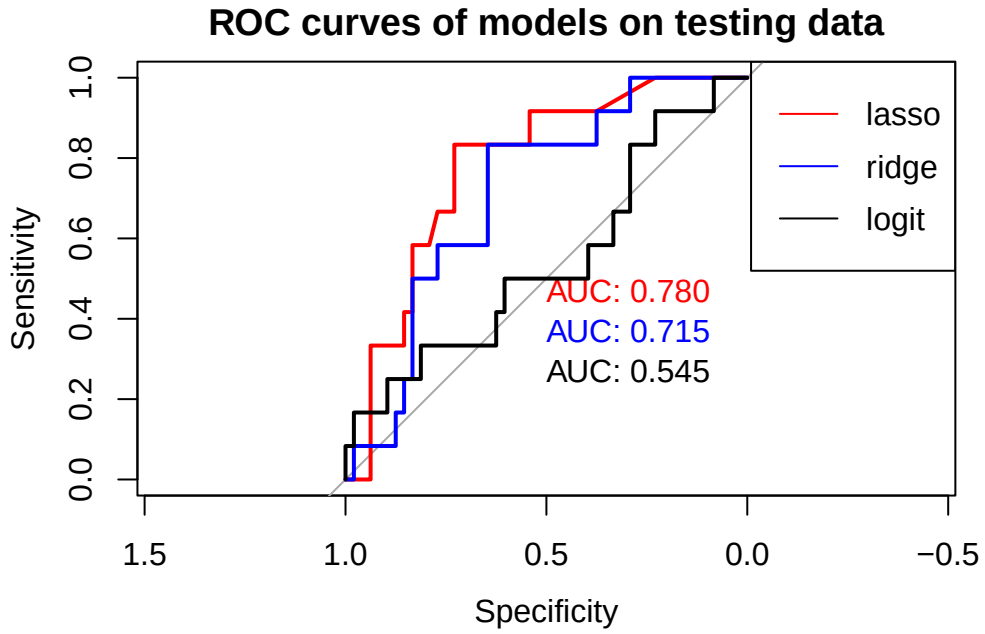


Figure 4: ROC Curves of Models on Testing data

## Results Interpretation

### Results from Lasso Regression

After fitting the models, we proceed to interpret the selected variables and their coefficients. In the lasso regression model, only one interaction term and one main effect were selected: the interaction between `Var` (pharmacotherapy) and `NMR` (nicotine metabolism), and the main effect of FTCD score at baseline. This coefficient suggests that for each one-unit increase in `NMR`, there is an estimated 0.49 increase in the odds ratio for abstinence when pharmacotherapy is administered (Table 2). This positive interaction indicates that higher nicotine metabolism may enhance the effectiveness of pharmacotherapy on smoking cessation outcomes.

Table 2: Coefficients of Selected Variables in Lasso regression

| Predictor | Coefficient |
|---|---|
| Intercept | -1.488 |
| Var1:age_ps | 0.006 |

Figure 5 suggest that Nicotine Metabolism Ratio (NMR) and pharmacotherapy interact to influence smoking abstinence. Specifically, individuals with a higher NMR (faster nicotine

metabolism) benefit more from pharmacotherapy, showing a significantly higher proportion of abstinence compared to those without pharmacotherapy. This effect is less pronounced in individuals with a lower NMR, where abstinence rates remain similar regardless of pharmacotherapy use. These findings imply that pharmacotherapy may be particularly effective for those with a faster nicotine metabolism.

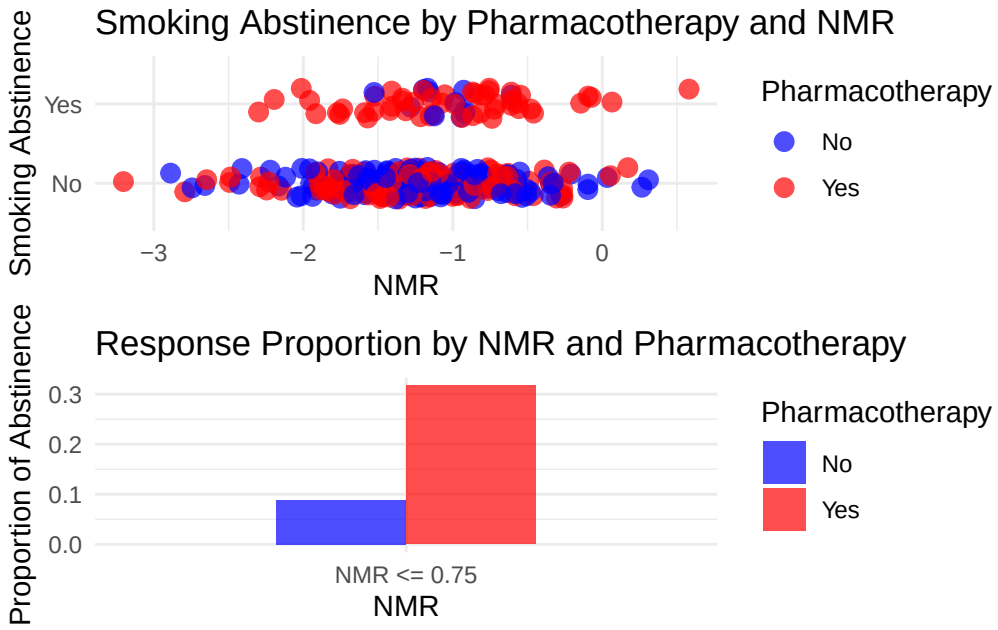


Figure 5: Smoking Abstinence by Pharmacotherapy and NMR

## Results from Logistic Regression

In the logistic regression model, one interaction term and several main effect were significant: the interaction between `BA1` (psychotherapy) and `edu4` (nicotine metabolism), and the main effect of `NHW1`, `ftcd_score`, `NMR` and `Var1`. The interaction coefficient suggests that for participants who received level 4 education, there is an estimated 2.225 increase in the odds ratio for abstinence when psychotherapy is administered (Table 2). This negative interaction indicates that some levels of education may enhance the effectiveness of psychotherapy on smoking cessation outcomes. Furthermore, we also found consistent results Table 1 previously reports: pharmacotherapy use, non-Hispanic white ethnicity, smoking habits upon waking, and nicotine metabolism may contribute to participants' abstinence in various ways.

Table 3: Coefficients of Selected Variables in Logistic regression

| | Estimate | Odds Ratio | Lower 2.5% | Upper 97.5% |
|---|---|---|---|---|
| Var1 | 1.23744 | 3.446778 | 0.2694940 | 2.205386 |
| NMR | 0.65456 | 1.924296 | 0.0923144 | 1.216806 |
| BA1:edu4 | 1.98406 | 7.272208 | 0.3730380 | 3.595082 |

Figure 6 reveals a nuanced interaction between education and psychotherapy on abstinence rates. For individuals with the lowest education level (level 1), psychotherapy has a dramatic positive effect, with abstinence rates near 100% for those who received psychotherapy and 0% for those who did not. This effect weakens as education level increases, with higher-educated individuals achieving similar or higher abstinence rates without psychotherapy, suggesting they may have other resources or coping mechanisms to support abstinence.

Interestingly, at education level 4, this trend reverses: those who received psychotherapy have a higher abstinence rate than those who did not, contrary to the general pattern. This may indicate that individuals at this level face unique challenges or stressors that make psychotherapy particularly valuable for achieving abstinence. Alternatively, it could reflect motivational differences or other characteristics in this group that make psychotherapy more effective. This reversal highlights the complex, varying interaction between education and psychotherapy across different education levels.

## Discussion

This study examined baseline variables as moderators and predictors of smoking cessation in adults with major depressive disorder (MDD) using data from a randomized, placebo-controlled trial. To identify key moderators of treatment effects on end-of-treatment (EOT) abstinence and predictors of cessation, we applied stepwise logit and lasso regression models, controlling for treatment and pharmacotherapy.

Our findings suggest that several baseline variables are important in influencing smoking cessation outcomes. Lasso regression identified nicotine metabolism as a moderator of pharmacotherapy effects, while stepwise regression revealed that education moderates the effects of psychotherapy. In terms of predictors, stepwise regression identified pharmacotherapy use, non-Hispanic white ethnicity, nicotine metabolism, and smoking habits upon waking as significant, while lasso regression detected only the latter behavior as a key predictor.

These results highlight the significant role of baseline variables in smoking cessation, underscoring their potential influence on both treatment efficacy and individual response to pharmacotherapy. However, it is important to note that lasso regression, while powerful, can sometimes be unstable. Sometimes it is sensitivity to data, thus may lead to variability in selected variables across different training sets. Variables with mild effects may occasionally be selected due to data peculiarities. To address this, it would be worthwhile to adopt an approach akin to cross-validation and use meta-analysis? techniques to identify reliable and

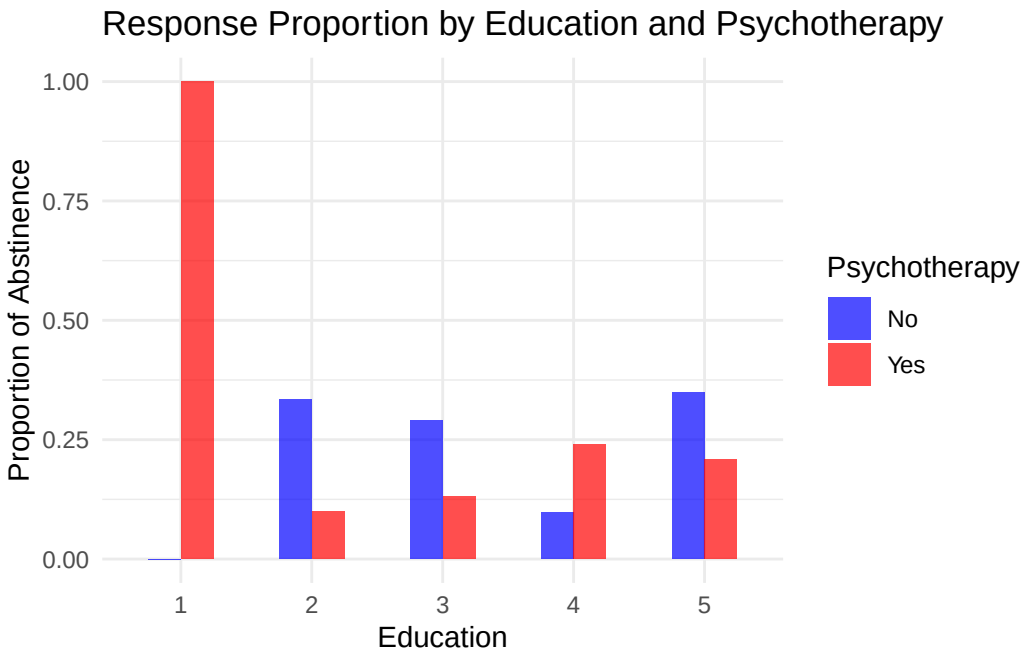


Figure 6: Smoking Abstinence by Psychotherapy and Education

stable variables that contribute to smoking abstinence. Further research using such methods could provide a more robust understanding of the interactions between baseline characteristics and treatment strategies.

## Code Appendix

```
library(gtsummary)
library(gt)
library(tidyverse)
library(ggplot2)
library(reshape2)
library(mice)
library(glmnet)
library(pROC)
library(kableExtra)
library(patchwork)
# Select continuous variables
df <- read.csv("project2.csv") %>%
```

```
  select(-NHW) %>%
  mutate(NMR = log(NMR))
con_var <- df %>%
  select(contains("pq1"), "age_ps", "ftcd_score",
         "bdi_score_w00", "cpd_ps", "NMR", "readiness") %>%
  names

# Select categorical variables by excluding 'id' and the continuous variables
cat_var <- df %>%
  select(-id, -all_of(con_var)) %>%
  names

# Convert the selected categorical variables to factors
df[cat_var] <- lapply(df[cat_var], as.factor)
df %>%
  select(-id) %>%
  mutate(abst=ifelse(abst=="1", "Yes", "No")) %>%
  tbl_summary(
    by = "abst",  # Group by sex
    type = list(readiness="continuous"),
    statistic = list(
      age_ps ~ "[{min}, {max}]",   # Only range for 'age'
      setdiff(con_var, "age_ps") ~ "{mean} ({sd})",  # For other continuous variables
      setdiff(cat_var, "abst") ~ "{n}/{N} ({p}%)"  # For categorical variables
    ),
    missing = "ifany",  # Add missing data information
    digits = list(con_var ~ c(2, 2))
  ) %>%
  add_overall() %>% # add overall statistics
  add_p() # add comparison
# Select continuous variables and compute the correlation matrix
df %>%
  select(con_var) %>%  # Select the continuous variables
  cor(use="complete.obs") %>% # ignores missing values
  melt() %>%  # Reshape the correlation matrix

  # Create a heatmap
  ggplot(aes(Var1, Var2, fill = value)) +
  geom_tile(color = "white") +
  scale_fill_gradient2(low = "blue", high = "red", mid = "white",
                       midpoint = 0, limit = c(-1, 1), space = "Lab",  # Set midpoint
                       name = "Correlation") +  # Label the color scale
```

```
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, vjust = 1, hjust = 1)) +
  coord_fixed()  # Fix the aspect ratio
# Define a function to calculate Cramér's V for a matrix of categorical variables
cramer_v_mat <- function(data) {
  vars <- names(data)
  n <- length(vars)
  v_matrix <- matrix(NA, n, n, dimnames = list(vars, vars))

  for (i in 1:n) {
    for (j in i:n) {
      if (i == j) {
        v_matrix[i, j] <- 1
      } else {
        # Calculate Cramér's V for the pair of categorical variables
        v <- rcompanion::cramerV(data[[i]], data[[j]], bias.correct = TRUE)
        v_matrix[i, j] <- v
        v_matrix[j, i] <- v
      }
    }
  }
  # Convert the matrix to a data frame and reshape it to long format
  as.data.frame(as.table(v_matrix))
}

# Apply function to create Cramér's V matrix
cramer_v_df <- df %>%
  select(cat_var) %>%
  cramer_v_mat()

# Visualize Cramér's V as a heatmap
ggplot(cramer_v_df, aes(Var1, Var2, fill = Freq)) +
  geom_tile(color = "white") +
  scale_fill_gradient2(low = "white", high = "red", limit = c(0, 1), space = "Lab",
                       name = "Cramér's V") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, vjust = 1, hjust = 1)) +
  coord_fixed()
# impute missing values using the 'mice' package
set.seed(1)
df_imp <- df %>%
  select(-id) %>%  # remove the 'id' column
```

```
  mice(m=100, printFlag=F) %>%  # Impute missing values
  complete # extract the completed (imputed) dataset
N <- nrow(df_imp)
# randomly sample indices to create the training set (80% of the data)
prop_train <- 0.8
idx_train <- sample(N, prop_train * N)
train <- df_imp[idx_train, ]
test <- df_imp[-idx_train, ]
# create the design matrix (X) for the training set, including interactions for 'BA' and 'Va
X.train <- model.matrix(abst ~ . + BA * (.) + Var * (.),
                        data=train)
Y.train <- factor(train$abst)
# fit a lasso regression model
cv.lasso <- cv.glmnet(X.train, Y.train, alpha=1,
                      family="binomial",
                      type.measure="auc")
# fit a ridge regression model
cv.ridge <- cv.glmnet(X.train, Y.train, alpha=0,
                      family="binomial",
                      type.measure="auc")
# extract the optimal lambda value for the Lasso model
lambda.lasso <- cv.lasso$lambda.1se
lasso <- glmnet(X.train, Y.train, alpha=1,
                family="binomial",
                lambda=lambda.lasso)
# extract the optimal lambda value for the ridge model
lambda.ridge <- cv.ridge$lambda.1se
ridge <- glmnet(X.train, Y.train, alpha=0,
                family="binomial",
                lambda=lambda.ridge)
# fit a logistic regression model using stepwise selection
logit <- step(glm(abst ~ BA*(.),
                  data=train, family=binomial()),
              trace=0)
# calculate ROC for the lasso model
roc.lasso <- roc(Y.train,
                 lasso %>%
                   predict(X.train, type="response") %>%
                   as.vector)
# calculate ROC for the ridge model
roc.ridge <- roc(Y.train,
                 ridge %>%
```

```
                    predict(X.train, type="response") %>%
                    as.vector)
# calculate ROC for the logit model
roc.logit <- roc(Y.train,
                 logit %>%
                   predict(train, type="response"))

# plot ROC curve and combine them
plot.roc(roc.lasso, print.auc=T,
         col="red", lwd=2,
         main="ROC curves of models on training data")
plot.roc(roc.ridge, print.auc=T,
         col="blue", lwd=2, add=T,
         print.auc.x=0.5,
         print.auc.y=0.4)
plot.roc(roc.logit, print.auc=T,
         col="black", lwd=2, add=T,
         print.auc.x=0.5,
         print.auc.y=0.3)
legend("topright",
       legend=c("lasso", "ridge", "logit"),
       col=c("red", "blue", "black"),
       lty=1)
# create test dataset
X.test <- model.matrix(abst ~ . + BA * (.) + Var * (.),
                  data=test)
Y.test <- test$abst

# calculate ROC
roc.lasso <- roc(Y.test,
                 lasso %>%
                   predict(X.test, type="response") %>%
                   as.vector)
roc.ridge <- roc(Y.test,
                 ridge %>%
                   predict(X.test, type="response") %>%
                   as.vector)
roc.logit <- roc(Y.test,
                 logit %>%
                   predict(test, type="response"))
# plot ROC for all models
plot.roc(roc.lasso, print.auc=T,
```

```
          col="red", lwd=2,
          main="ROC curves of models on testing data")
plot.roc(roc.ridge, print.auc=T,
          col="blue", lwd=2, add=T,
          print.auc.x=0.5,
          print.auc.y=0.4)
plot.roc(roc.logit, print.auc=T,
          col="black", lwd=2, add=T,
          print.auc.x=0.5,
          print.auc.y=0.3)
legend("topright",
       legend=c("lasso", "ridge", "logit"),
       col=c("red", "blue", "black"),
       lty=1)
# create the design matrix using all predictors and interactions
X <- model.matrix(abst ~ . + BA * (.) + Var * (.),
                  data=df_imp)
Y <- df_imp$abst
mdl.lasso <- cv.glmnet(X, Y, alpha=1,
                   family="binomial",
                   type.measure="auc")
# extract the coefficients
coef.lasso <- coef(mdl.lasso, s="lambda.1se") %>% as.matrix
# rename the intercept term to "Intercept" for clarity
rownames(coef.lasso)[rownames(coef.lasso) == "(Intercept)"] <- "Intercept"

# create a data frame containing only non-zero coefficients
data.frame(Predictor=rownames(coef.lasso)[which(coef.lasso!=0)],
           Coefficient=coef.lasso[coef.lasso!=0]) %>%
  mutate_if(is.numeric, round, digits = 3) %>%
  kable() %>%
  kable_styling(full_width=F, position="center")
# create a jitter plot to reduce overlap
plt_cont <- df_imp %>%
  ggplot(aes(x = NMR, y = abst, color = Var)) +
  geom_jitter(width = 0.2, height = 0.2, size = 3, alpha = 0.7) +
  labs(
    title = "Smoking Abstinence by Pharmacotherapy and NMR",
    x = "NMR",
    y = "Smoking Abstinence",
    color = "Pharmacotherapy"
  ) +
```

```
  scale_y_discrete(labels = c("0"="No", "1"="Yes")) + # Change y-axis labels
  # blue for no pharmacotherapy and red for pharmacotherapy
  scale_color_manual(values=c("0"="blue", "1"="red"),
                     labels = c("0"="No", "1"="Yes")) +
  theme_minimal()

# create a bar plot showing the proportion of abstinence
plt_disc <- df_imp %>%
  # NMR is categorized into two groups (<= 0.75 and > 0.75)
  mutate(NMR=cut(NMR, breaks = c(-Inf, 0.75, Inf),
                 labels = c("NMR <= 0.75", "NMR > 0.75"))) %>%
  group_by(NMR, Var) %>%
  summarise(Resp = sum(abst=="1") / n()) %>%
  ggplot(aes(x = NMR, fill = factor(Var))) + # Convert Var to factor
  geom_bar(aes(y = Resp), stat = "identity",
           position = "dodge", width = 0.5, alpha=0.7) +
  labs(
    title = "Response Proportion by NMR and Pharmacotherapy",
    x = "NMR",
    y = "Proportion of Abstinence",
    fill = "Pharmacotherapy"
  ) +
  scale_fill_manual(values=c("0"="blue", "1"="red"),
                     labels = c("No", "Yes")) +
  theme(legend.position = "none") +
  theme_minimal()
plt_cont / plt_disc
logit <- step(glm(abst ~ BA * (.),
                  data=df_imp, family=binomial()),
              trace=0)
# logit <- glm(abst ~ .,
#              data=df_imp, family=binomial())
# var <- seq(min(df_imp$ftcd_score),
#                   max(df_imp$ftcd_score),
#                   length.out = 100) %>%
#   rep(nrow(df_imp))
#
# dat <- apply(df_imp, 2, rep, each = 100) %>%
#   as.data.frame() %>%
#   mutate(ftcd_score = var,
#          across(all_of(con_var), as.numeric))
#
```

```r
# odds <- predict(logit, newdata = dat, type="response") %>%
#   matrix(nrow = 100) %>%
#   apply(1, mean)
#
# data.frame(var, odds) %>%
#   ggplot(aes(x=var, y=odds)) +
#   geom_line()
# extract coefficients
coef.logit <- summary(logit)$coefficients
colnames(coef.logit) <- c("Estimate", "SE", "Z value", "p value")
# remove the first row (Intercept) and format the output
coef.logit[-1, ] %>%
  as.data.frame() %>%
  mutate_if(is.numeric, round, digits = 5) %>% # round numeric values to 5 decimal places
  mutate(`p value`=case_when(
    `p value` < 0.01 ~ paste0(as.character(`p value`), "**"), # add '**' for p-values < 0.01
    `p value` < 0.05 ~ paste0(as.character(`p value`), "*"),  # add '*' for p-values < 0.05
    TRUE ~ as.character(`p value`))) %>%
  filter(`p value` < 0.05) %>%
  mutate(`Odds Ratio` = exp(`Estimate`),
         `Lower 2.5%` = `Estimate` - 1.96 * `SE`,
         `Upper 97.5%` = `Estimate` + 1.96 * `SE`) %>%
  select(-`Z value`, -`p value`, -`SE`) %>%
  kable() %>%
  kable_styling(full_width=F, position="center")
df_imp %>%
  group_by(edu, BA) %>%
  summarise(Resp = sum(abst=="1") / n(), # calculate the proportion of abstinence (Resp)
            .groups="drop") %>%
  rbind(data.frame(edu = 1, BA = 0, Resp = 0)) %>%
  ggplot(aes(x = edu, fill = factor(BA))) + # Convert Var to factor
  geom_bar(aes(y = Resp), stat = "identity",
           position = "dodge", width = 0.5, alpha=0.7) +
  labs(
    title = "Response Proportion by Education and Psychotherapy",
    x = "Education",
    y = "Proportion of Abstinence",
    fill = "Psychotherapy"
  ) +
  scale_fill_manual(values = c("0" = "blue", "1" = "red"),
                    labels = c("No", "Yes")) + # Set legend labels to "No" and "Yes"
  theme_minimal()
```