# Testing Figure 1

2022-11-18

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6      v purrr   0.3.4
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.4.1
## v readr   2.1.2      v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

## Load data

```
PM = read.csv('data/external/AHRI_DATASET_PM_MANUSCRIPT_DATA.csv')
```

## Clean data

```
# only necessary columns for figure 1's multivariate logistical regression
PM_cleaned = PM %>%
  select(
    CASEID_7139,
    SEX,
    AGE,
    ETHNICITY,
    HLS_YN,
    REGION,
    CCI_SCORE,
    GAD7_GE10,
    PHQ9_GE10,
    INSURANCE,
    PM1_GEN_HEALTH,
    PM1_DIAG_CONDITION,
    PM1_UNDIAG_CONCERN
    ) %>%
  mutate(
    PM_12M = PM1_GEN_HEALTH + PM1_UNDIAG_CONCERN + PM1_DIAG_CONDITION
    )

# calculate boolean for PM_12M
PM_cleaned$PM_12M = PM_cleaned$PM_12M %>%
  recode(`-297` = 0, `0` = 0, `1` = 1, `2` = 1, `3` = 1)
```

```r
## Make additional columns for individual risk factors
PM_cleaned$BLACK = (PM_cleaned$ETHNICITY == 1)
PM_cleaned$WHITE = (PM_cleaned$ETHNICITY == 2)
PM_cleaned$OTHERETHNICITY = (PM_cleaned$ETHNICITY == 3)

PM_cleaned$NORTHWEST = (PM_cleaned$REGION == 1)
PM_cleaned$MIDWEST = (PM_cleaned$REGION == 2)
PM_cleaned$SOUTH = (PM_cleaned$REGION == 3)
PM_cleaned$WEST = (PM_cleaned$REGION == 4)

## refactor columns
PM_cleaned$SEX = PM_cleaned$SEX %>%
  recode_factor(., `0` = 'Female', `1` = 'Male')

PM_cleaned$ETHNICITY = PM_cleaned$ETHNICITY %>%
  recode_factor(., `1` = 'Black', `2` = 'White', `3` = 'Other')

PM_cleaned$HLS_YN = PM_cleaned$HLS_YN %>%
  recode_factor(., `0` = 'None-Hispanic', `1` = 'Hispanic')

PM_cleaned$REGION = PM_cleaned$REGION %>%
  recode_factor(., `1` = 'Northwest', `2` = 'Midwest', `3` = 'South', `4` = 'West')
```

### Multivariate logistical regression

```r
## modified helper from https://rdrr.io/github/eringrand/RUncommon/src/R/logistic.regression.or.ci.R
logistic.regression.or.ci <- function(regress.out, level = 0.95) {
  usual.output <- summary(regress.out)
  z.quantile <- stats::qnorm(1 - (1 - level) / 2)
  number.vars <- length(regress.out$coefficients)
  OR <- exp(regress.out$coefficients[-1])
  temp.store.result <- matrix(rep(NA, number.vars * 2), nrow = number.vars)
  for (i in 1:number.vars) {
    temp.store.result[i, ] <- summary(regress.out)$coefficients[i] +
      c(-1, 1) * z.quantile * summary(regress.out)$coefficients[i + number.vars]
  }
  intercept.ci <- temp.store.result[1, ]
  slopes.ci <- temp.store.result[-1, ]
  OR.ci <- exp(slopes.ci)

  output <- list(
    regression.table = usual.output, intercept.ci = intercept.ci,
    slopes.ci = slopes.ci, OR = OR, OR.ci = OR.ci
  )
  return(output)
}
```

```r
full_model = glm(PM_12M ~ SEX + AGE + ETHNICITY + HLS_YN + REGION + CCI_SCORE + GAD7_GE10 + PHQ9_GE10 +
full_model_results = logistic.regression.or.ci(full_model)
```

```
full_model_results
```

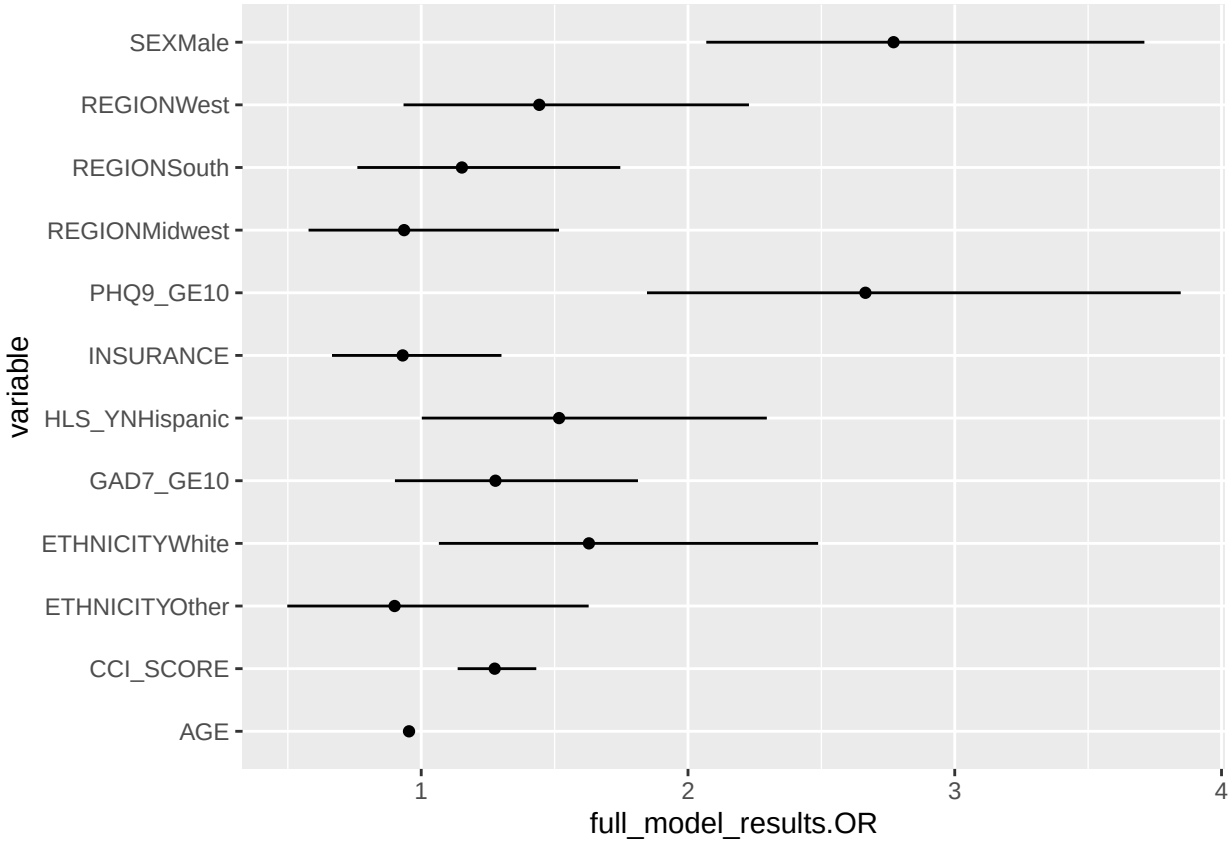
```
## $regression.table
##
## Call:
## glm(formula = PM_12M ~ SEX + AGE + ETHNICITY + HLS_YN + REGION +
##     CCI_SCORE + GAD7_GE10 + PHQ9_GE10 + INSURANCE, family = binomial,
##     data = PM_cleaned)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.0961  -0.2795  -0.1811  -0.1115   3.3656
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -3.280300   0.391205  -8.385  < 2e-16 ***
## SEXMale         1.019080   0.149089   6.835 8.18e-12 ***
## AGE            -0.046839   0.006234  -7.513 5.78e-14 ***
## ETHNICITYWhite  0.487751   0.216163   2.256   0.0240 *
## ETHNICITYOther -0.105178   0.302202  -0.348   0.7278
## HLS_YNHispanic  0.416570   0.211429   1.970   0.0488 *
## REGIONMidwest  -0.066266   0.246227  -0.269   0.7878
## REGIONSouth     0.141909   0.212027   0.669   0.5033
## REGIONWest      0.366388   0.221935   1.651   0.0988 .
## CCI_SCORE       0.243288   0.058849   4.134 3.56e-05 ***
## GAD7_GE10       0.245465   0.178183   1.378   0.1683
## PHQ9_GE10       0.980240   0.187282   5.234 1.66e-07 ***
## INSURANCE      -0.072116   0.170952  -0.422   0.6731
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2019.1  on 7138  degrees of freedom
## Residual deviance: 1762.9  on 7126  degrees of freedom
## AIC: 1788.9
##
## Number of Fisher Scoring iterations: 7
##
##
## $intercept.ci
## [1] -4.047047 -2.513553
##
## $slopes.ci
##                [,1]       [,2]
##  [1,]  0.726871795  1.3112888
##  [2,] -0.059058190 -0.0346195
##  [3,]  0.064079256  0.9114233
##  [4,] -0.697482020  0.4871269
##  [5,]  0.002176631  0.8309630
##  [6,] -0.548860756  0.4163295
##  [7,] -0.273657343  0.5574749
##  [8,] -0.068595774  0.8013720
```

```
##  [9,]  0.127947376  0.3586295
## [10,] -0.103767309  0.5946973
## [11,]  0.613174206  1.3473066
## [12,] -0.407176769  0.2629438
##
## $OR
##        SEXMale            AGE ETHNICITYWhite ETHNICITYOther HLS_YNHispanic
##      2.7706455      0.9542412      1.6286497      0.9001647      1.5167499
##  REGIONMidwest    REGIONSouth     REGIONWest      CCI_SCORE      GAD7_GE10
##      0.9358822      1.1524715      1.4425150      1.2754364      1.2782155
##      PHQ9_GE10      INSURANCE
##      2.6650969      0.9304225
##
## $OR.ci
##            [,1]      [,2]
##  [1,] 2.0685995 3.7109535
##  [2,] 0.9426519 0.9659729
##  [3,] 1.0661769 2.4878610
##  [4,] 0.4978373 1.6276331
##  [5,] 1.0021790 2.2955282
##  [6,] 0.5776075 1.5163854
##  [7,] 0.7605927 1.7462574
##  [8,] 0.9337040 2.2285965
##  [9,] 1.1364932 1.4313663
## [10,] 0.9014350 1.8124822
## [11,] 1.8462826 3.8470501
## [12,] 0.6655265 1.3007536
```

```
df = data.frame(full_model_results$OR)
df = cbind(variable = rownames(df), df)
rownames(df) = 1:nrow(df)

df$or.cimin = full_model_results$OR.ci[,1]
df$or.cimax = full_model_results$OR.ci[,2]
```

```
## try plotting
ggplot(data = df, aes(x = full_model_results.OR, y = variable, xmin = or.cimin, xmax = or.cimax)) +
  geom_linerange() +
  geom_point()
```


SEXMale
REGIONWest
REGIONSouth
REGIONMidwest
PHQ9_GE10
INSURANCE
HLS_YNHispanic
GAD7_GE10
ETHNICITYWhite
ETHNICITYOther
CCI_SCORE
AGE
variable
1
2
3
4
full_model_results.OR