BRFSS2023_regression.R

abdullahsiddiqui

2025-09-23

```r
###############################################################
# BRFSS2023_regression.R
# Week 4: Linear & Logistic regression + diagnostics
# Author: Abdullah Siddiqui
# Date: Oct 2, 2025
###############################################################
```

```r
library(ggplot2)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
library(readr)
library(car)       # for VIF
```

```
## Loading required package: carData

##
## Attaching package: 'car'

## The following object is masked from 'package:dplyr':
##
##     recode
```

```r
library(broom)     # for tidy regression output
```

```r
df <- read_csv("~/Downloads/BRFSS2023_subset_clean.csv")
```

```
## Rows: 433323 Columns: 8
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## dbl (8): MENTHLTH, EXERANY2, SMOKDAY2, ALCDAY4, SEXVAR, EDUCA, INCOME3, _AGE...
##
## i Use ‘spec()‘ to retrieve the full column specification for this data.
## i Specify the column types or set ‘show_col_types = FALSE‘ to quiet this message.
```

```r
if (!dir.exists("plots")) dir.create("plots")
if (!dir.exists("tables")) dir.create("tables")
if (!dir.exists("outputs")) dir.create("outputs")
```

```r
lm_model <- lm(MENTHLTH ~ INCOME3 + EDUCA + EXERANY2 + SMOKDAY2 + ALCDAY4 + `_AGEG5YR`,
               data = df)

# Save summary as text
sink("outputs/linear_regression_summary.txt")
print(summary(lm_model))
sink()

# Save coefficients table as CSV
lm_tidy <- broom::tidy(lm_model)
write.csv(lm_tidy, "tables/linear_regression_coeffs.csv", row.names = FALSE)

# Check multicollinearity (VIF)
vif_values <- vif(lm_model)
write.csv(vif_values, "tables/vif_linear.csv")

# Diagnostic plots
png("plots/residuals_vs_fitted.png", width = 800, height = 600)
plot(lm_model, which = 1)   # residuals vs fitted
dev.off()
```

```
## pdf
##   2
```

```r
png("plots/qq_plot.png", width = 800, height = 600)
plot(lm_model, which = 2)   # Q-Q plot
dev.off()
```

```
## pdf
##   2
```

```r
# Create binary outcome: frequent distress (>14 days poor mental health)
df$frequent_distress <- ifelse(df$MENTHLTH > 14, 1, 0)

# Logistic regression with backticks around _AGEG5YR
log_model <- glm(frequent_distress ~ INCOME3 + EDUCA + EXERANY2 + SMOKDAY2 + ALCDAY4 + `_AGEG5YR`,
                 data = df, family = binomial)

summary(log_model)
```

```
##
## Call:
## glm(formula = frequent_distress ~ INCOME3 + EDUCA + EXERANY2 +
##     SMOKDAY2 + ALCDAY4 + ‘_AGEG5YR‘, family = binomial, data = df)
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  3.609e-01  5.279e-02   6.837 8.10e-12 ***
## INCOME3     -1.895e-01  3.547e-03 -53.418  < 2e-16 ***
## EDUCA        5.989e-02  8.621e-03   6.946 3.75e-12 ***
## EXERANY2     5.605e-01  1.711e-02  32.768  < 2e-16 ***
## SMOKDAY2    -1.671e-01  9.270e-03 -18.021  < 2e-16 ***
## ALCDAY4     -4.033e-04  8.584e-05  -4.699 2.62e-06 ***
## ‘_AGEG5YR‘  -1.747e-01  2.481e-03 -70.423  < 2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 116021  on 127864  degrees of freedom
## Residual deviance: 105087  on 127858  degrees of freedom
##   (305458 observations deleted due to missingness)
## AIC: 105101
##
## Number of Fisher Scoring iterations: 5
```

```
# Save logistic regression summary
sink("outputs/logistic_regression_summary.txt")
print(summary(log_model))
sink()

# Save coefficients table
log_tidy <- broom::tidy(log_model)
write.csv(log_tidy, "tables/logistic_regression_coeffs.csv", row.names = FALSE)

# Odds ratios + CI
odds_ratios <- exp(cbind(OR = coef(log_model), confint(log_model)))
```

```
## Waiting for profiling to be done...
```

```
write.csv(odds_ratios, "tables/logistic_odds_ratios.csv")

# ROC-like diagnostic: predicted probabilities
logit_pred <- predict(log_model, type = "response")  # vector of predictions
length(logit_pred)  # should be ~127,865
```

```
## [1] 127865
```

```
# Make a new dataframe just for diagnostics
pred_df <- data.frame(predicted_prob = logit_pred)

# Save histogram plot
```

```r
p1 <- ggplot(pred_df, aes(x = predicted_prob)) +
  geom_histogram(binwidth = 0.05, fill = "steelblue", color = "white") +
  labs(title = "Predicted Probability Distribution (Logistic Regression)",
       x = "Predicted probability of frequent distress", y = "Count") +
  theme_minimal()
ggsave("plots/logistic_predicted_probabilities.png", plot = p1, width = 7, height = 5)


#############################################################
# End of Script
#############################################################
```