BRFSS2023_regression_week5.R

abdullahsiddiqui

2025-09-30

```
#################################################################
# BRFSS2023_regression_week5.R
# Week 5: Regression models completed + organized outputs
# Author: Abdullah Siddiqui
# Date: Sep 30, 2025
#################################################################
```

```
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
```

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

```
## Loading required package: carData
```

```
##
## Attaching package: 'car'
```

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

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

```
df <- read_csv("/Users/abdullahsiddiqui/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.
```

```
getwd()
```

```
## [1] "/Users/abdullahsiddiqui/Downloads"
```

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

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

# Save linear regression summary
sink("outputs/linear_regression_summary.txt")
print(summary(lm_model))
sink()

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

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

# Save diagnostic plots
png("plots/residuals_vs_fitted.png", width = 800, height = 600)
plot(lm_model, which = 1)
dev.off()
```

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

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

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

```
# Binary outcome: frequent distress (>14 days)
df$frequent_distress <- ifelse(df$MENTHLTH > 14, 1, 0)

log_model <- glm(frequent_distress ~ INCOME3 + EDUCA + EXERANY2 + SMOKDAY2 + ALCDAY4 + `_AGEG5YR`,
```

```
                data = df, family = binomial)

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

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

# Save odds ratios + confidence intervals
odds_ratios <- exp(cbind(OR = coef(log_model), confint.default(log_model)))
write.csv(odds_ratios, "tables/logistic_odds_ratios.csv")

# Save histogram of predicted probabilities
logit_pred <- predict(log_model, type = "response")
pred_df <- data.frame(predicted_prob = logit_pred)

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)

getwd()
```

```
## [1] "/Users/abdullahsiddiqui/Downloads"
```

```
################################################################
# End of Script (Week 5)
################################################################
```