Assignment 03 Regressions

BQOM 2578 | Data Mining

Theresa Wohlever

2025-09-28

Table of contents

Assignment Instructions

https://canvas.pitt.edu/courses/324587/assignments/1871892

Define and describe the purpose of your analysis in terms of the output and input variable(s) you are interested in understanding for your project and if each variable is categorical or continuous.

Prepare your dataset for running the adequate regression.

Run at least three regressions, at least one linear and one logistic.

Decide how to present your final results; one model, several models? Which format? What graphs / visualizations would you use?

In the last section, describe your final conclusions grounded on your regression analysis and visualizations.

Executive Summary

For Each varialbe is it categorical or continuous? 3 regressions on data evaluate models

Data Preparation

Loading packages

Importing data

Data Cleaning & Wrangling

```
# Clean out columns duplicated with id's
df <- ihme_raw_df[, !names(ihme_raw_df) %in% c("measure_name", "location_name", "sex_name",

# Create Target; Isolate Mental Disorder
df <- df %>% mutate(IsMentalDisorder = ifelse(cause_id == 558, 1, 0))

# Remove column name used to determine target value
df$cause_id <- NULL


# Wrangle
# df$Deaths <- as.factor(df$Deaths)           # convert to Factor
# df$Date<-ymd(df$Date)                 # properly interrupt date field
```

```
summary(df)
```

```
   measure_id     location_id     sex_id      age_id          rei_id
 Min.   :1.000   Min.   :1    Min.   :3   Min.   :  6.00   Min.   : 92.0
 1st Qu.:1.000   1st Qu.:1    1st Qu.:3   1st Qu.: 22.00   1st Qu.:169.0
 Median :3.000   Median :1    Median :3   Median : 23.00   Median :186.0
```

```
Mean   :2.737   Mean   :1    Mean   :3   Mean   : 50.45   Mean   :210.6
3rd Qu.:4.000   3rd Qu.:1    3rd Qu.:3   3rd Qu.: 39.00   3rd Qu.:203.0
Max.   :4.000   Max.   :1    Max.   :3   Max.   :162.00   Max.   :381.0
  metric_id      year           val               upper
Min.   :1   Min.   :1990   Min.   :0.000e+00   Min.   :0.000e+00
1st Qu.:1   1st Qu.:1998   1st Qu.:2.776e+05   1st Qu.:3.654e+05
Median :1   Median :2006   Median :2.815e+06   Median :4.013e+06
Mean   :1   Mean   :2006   Mean   :5.765e+07   Mean   :6.585e+07
3rd Qu.:1   3rd Qu.:2013   3rd Qu.:1.194e+07   3rd Qu.:1.636e+07
Max.   :1   Max.   :2021   Max.   :1.163e+09   Max.   :1.231e+09
    lower           IsMentalDisorder
Min.   :0.000e+00   Min.   :0.0000
1st Qu.:1.241e+05   1st Qu.:0.0000
Median :1.789e+06   Median :0.0000
Mean   :4.880e+07   Mean   :0.1974
3rd Qu.:8.536e+06   3rd Qu.:0.0000
Max.   :1.098e+09   Max.   :1.0000
```

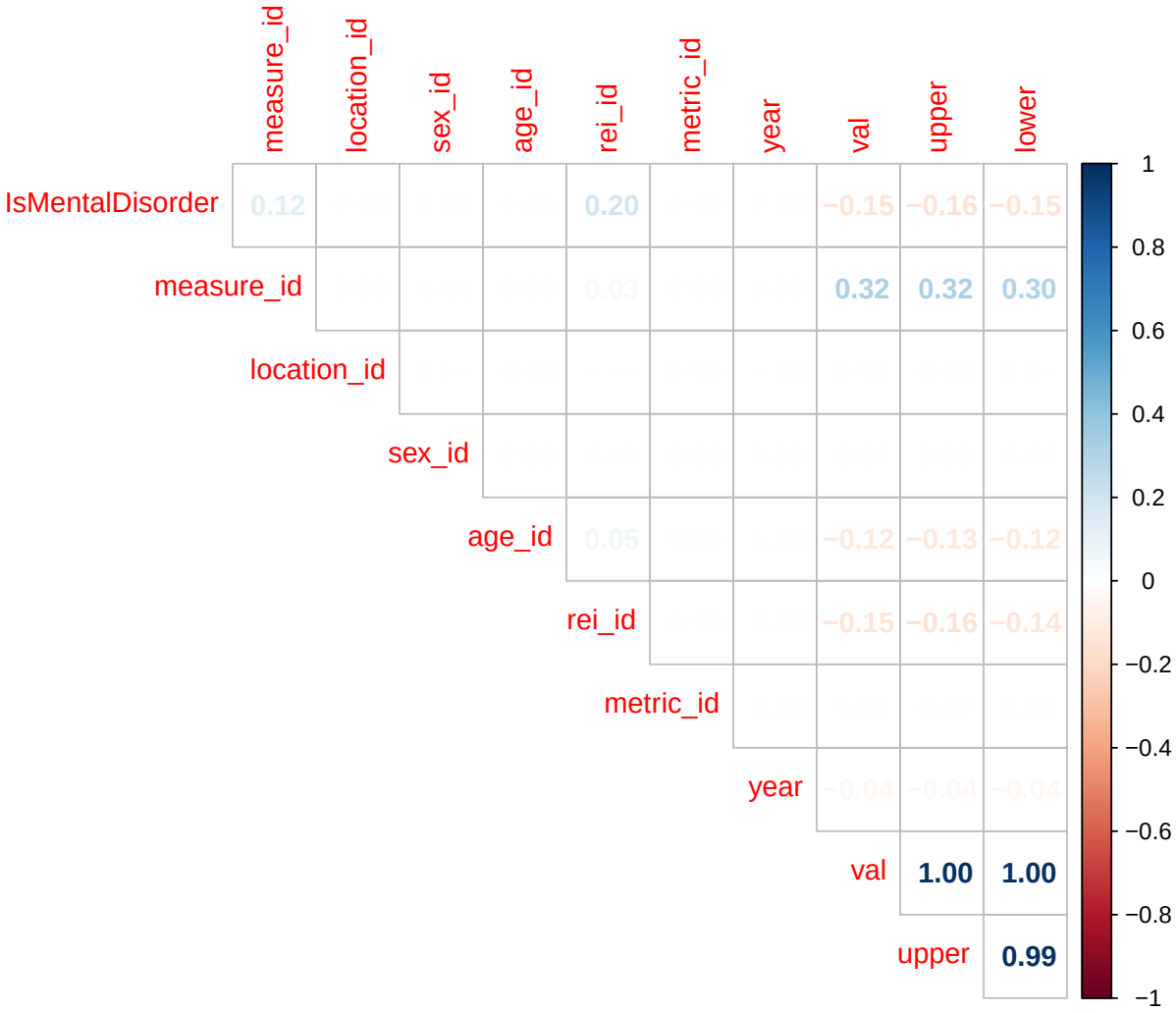
Preliminary Analysis

Evaluate Correlation Matrix

```
df <-df %>% relocate(IsMentalDisorder)  # moves the target variable to the first column (le
cor_mat <- cor(df)
```

```
Warning in cor(df): the standard deviation is zero
```

```
cor_mat_plot <- round(cor_mat, 2)
cor_mat_plot[is.na(cor_mat_plot)] <- 0 # Replace all NA values with zero
corrplot(cor_mat_plot, method="number", type="upper", diag=FALSE)
```



Regression

Linear Regression

```
m_sex<-lm(IsMentalDisorder ~ sex_id, data=df)
summary(m_sex)
```

```
Call:
lm(formula = IsMentalDisorder ~ sex_id, data = df)
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-0.1974 -0.1974 -0.1974 -0.1974  0.8026

Coefficients: (1 not defined because of singularities)
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.197368   0.008072   24.45   <2e-16 ***
sex_id            NA         NA      NA       NA
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3981 on 2431 degrees of freedom
```

```
m_sex_age <-lm(IsMentalDisorder ~ sex_id + age_id, data=df)
summary(m_sex_age)
```

```
Call:
lm(formula = IsMentalDisorder ~ sex_id + age_id, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-0.1975 -0.1974 -0.1974 -0.1970  0.8030

Coefficients: (1 not defined because of singularities)
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.975e-01  1.085e-02  18.207   <2e-16 ***
sex_id              NA         NA      NA       NA
age_id      -2.958e-06  1.436e-04  -0.021    0.984
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3982 on 2430 degrees of freedom
Multiple R-squared:  1.746e-07, Adjusted R-squared:  -0.0004113
F-statistic: 0.0004243 on 1 and 2430 DF,  p-value: 0.9836
```

```
anova(m_sex, m_sex_age)
```

```
Analysis of Variance Table
```

```
Model 1: IsMentalDisorder ~ sex_id
Model 2: IsMentalDisorder ~ sex_id + age_id
  Res.Df    RSS Df  Sum of Sq     F Pr(>F)
1   2431 385.26
2   2430 385.26  1 6.7264e-05 4e-04 0.9836
```

Stepwise Linear Regression

```r
model <- lm(IsMentalDisorder ~ ., data = df)
summary(model)
```

```
Call:
lm(formula = IsMentalDisorder ~ ., data = df)

Residuals:
     Min       1Q   Median       3Q      Max
-0.42970 -0.23644 -0.11736  0.00089  0.80866

Coefficients: (3 not defined because of singularities)
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -8.834e-02  1.688e+00  -0.052   0.9583
measure_id   7.019e-02  7.488e-03   9.374  < 2e-16 ***
location_id         NA         NA      NA       NA
sex_id              NA         NA      NA       NA
age_id      -3.089e-04  1.392e-04  -2.220   0.0265 *
rei_id       5.492e-04  7.778e-05   7.061 2.15e-12 ***
metric_id           NA         NA      NA       NA
year         1.738e-05  8.420e-04   0.021   0.9835
val          3.323e-09  4.011e-09   0.829   0.4075
upper       -4.056e-09  2.348e-09  -1.727   0.0843 .
lower        6.998e-10  1.700e-09   0.412   0.6807
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3798 on 2424 degrees of freedom
Multiple R-squared:  0.09224,   Adjusted R-squared:  0.08962
F-statistic: 35.19 on 7 and 2424 DF,  p-value: < 2.2e-16
```

```r
# Perform stepwise regression
#direction can be both, backward or forward
#trace can be set to 0 to only display final result or higher to display more information
step_model <- step(model, direction = "backward",trace=999)
```

```
Start:  AIC=-4700.43
IsMentalDisorder ~ measure_id + location_id + sex_id + age_id +
    rei_id + metric_id + year + val + upper + lower


Step:  AIC=-4700.43
IsMentalDisorder ~ measure_id + location_id + sex_id + age_id +
    rei_id + year + val + upper + lower


Step:  AIC=-4700.43
IsMentalDisorder ~ measure_id + location_id + age_id + rei_id +
    year + val + upper + lower


Step:  AIC=-4700.43
IsMentalDisorder ~ measure_id + age_id + rei_id + year + val +
    upper + lower

             Df Sum of Sq    RSS     AIC
- year        1    0.0001 349.73 -4702.4
- lower       1    0.0244 349.75 -4702.3
- val         1    0.0990 349.82 -4701.7
<none>                    349.73 -4700.4
- upper       1    0.4304 350.16 -4699.4
- age_id      1    0.7108 350.44 -4697.5
- rei_id      1    7.1937 356.92 -4652.9
- measure_id  1   12.6768 362.40 -4615.8

Step:  AIC=-4702.43
IsMentalDisorder ~ measure_id + age_id + rei_id + val + upper +
    lower

             Df Sum of Sq    RSS     AIC
- lower       1    0.0249 349.75 -4704.3
- val         1    0.0998 349.82 -4703.7
<none>                    349.73 -4702.4
```

```
- upper       1    0.4358 350.16 -4701.4
- age_id      1    0.7108 350.44 -4699.5
- rei_id      1    7.1942 356.92 -4654.9
- measure_id  1   12.6768 362.40 -4617.8

Step:  AIC=-4704.26
IsMentalDisorder ~ measure_id + age_id + rei_id + val + upper

             Df Sum of Sq    RSS     AIC
<none>                     349.75 -4704.3
- age_id      1    0.7184 350.47 -4701.3
- val         1    3.7021 353.45 -4680.6
- upper       1    4.3746 354.12 -4676.0
- rei_id      1    7.2196 356.97 -4656.6
- measure_id  1   12.6749 362.42 -4619.7
```

```r
#backward starts with everything and drops non-significant values.


# View the summary of the stepwise model
summary(step_model)
```

```
Call:
lm(formula = IsMentalDisorder ~ measure_id + age_id + rei_id + 
    val + upper, data = df)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.43020 -0.23598 -0.11420 -0.00444  0.80933 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -5.297e-02  2.632e-02  -2.012   0.0443 *  
measure_id   7.019e-02  7.485e-03   9.376  < 2e-16 ***
age_id      -3.104e-04  1.390e-04  -2.232   0.0257 *  
rei_id       5.500e-04  7.772e-05   7.077 1.93e-12 ***
val          4.919e-09  9.708e-10   5.067 4.33e-07 ***
upper       -4.944e-09  8.975e-10  -5.509 4.00e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 0.3797 on 2426 degrees of freedom
Multiple R-squared:  0.09218,   Adjusted R-squared:  0.09031
F-statistic: 49.27 on 5 and 2426 DF,  p-value: < 2.2e-16
```

```
#For more info: https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/step
```

References

Global Burden of Disease Collaborative Network. Global Burden of Disease Study 2021 (GBD 2021) Results. Seattle, United States: Institute for Health Metrics and Evaluation (IHME), 2022. Available from https://vizhub.healthdata.org/gbd-results/.