Assignment 03 Regressions

BQOM 2578 | Data Mining

Theresa Wohlever

Sunday, September 28, 2025

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 variable is it categorical or continuous? 3 regressions on data evaluate models

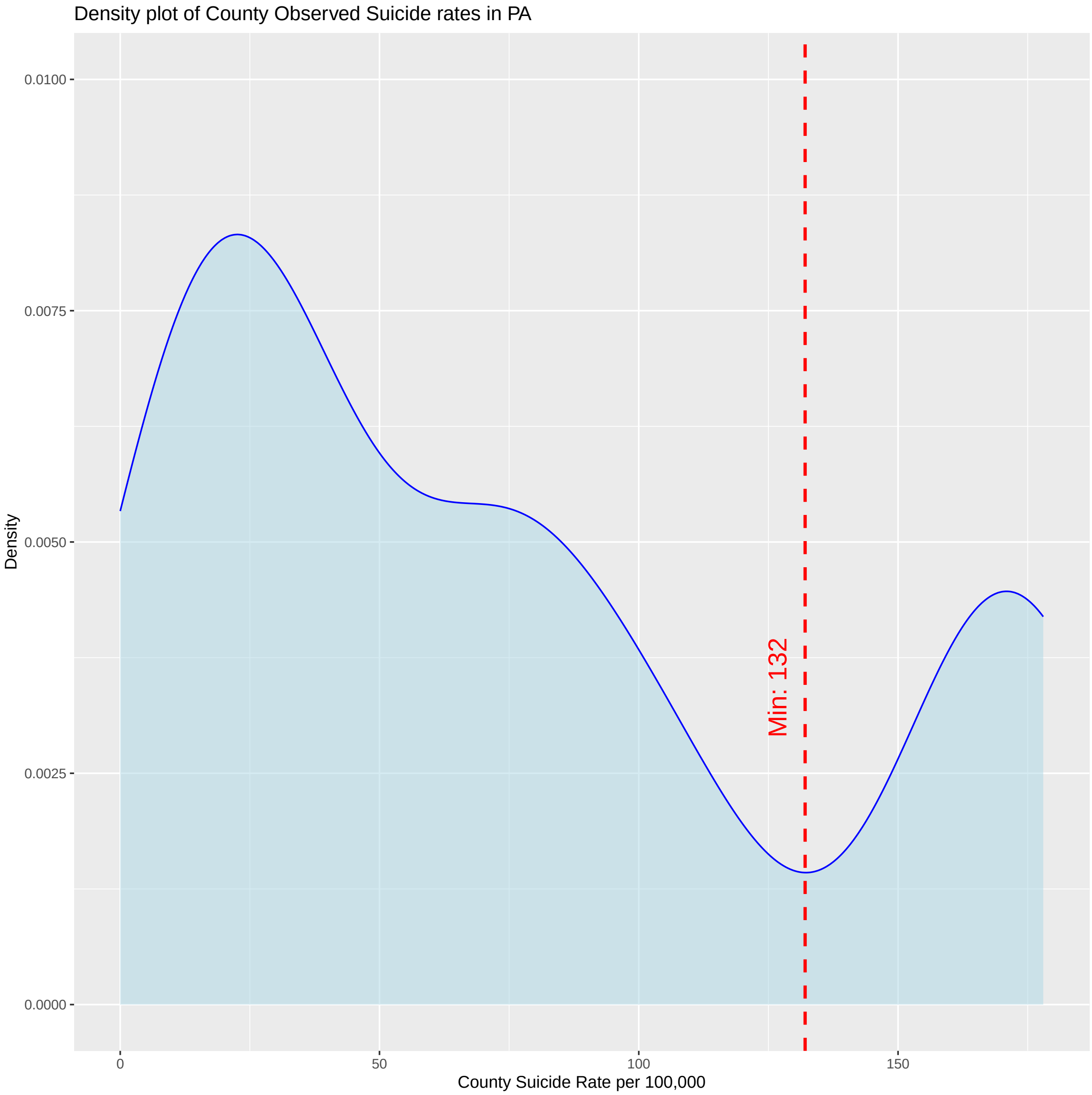
Data Preparation

Loading packages

Importing data

Data Cleaning & Wrangling

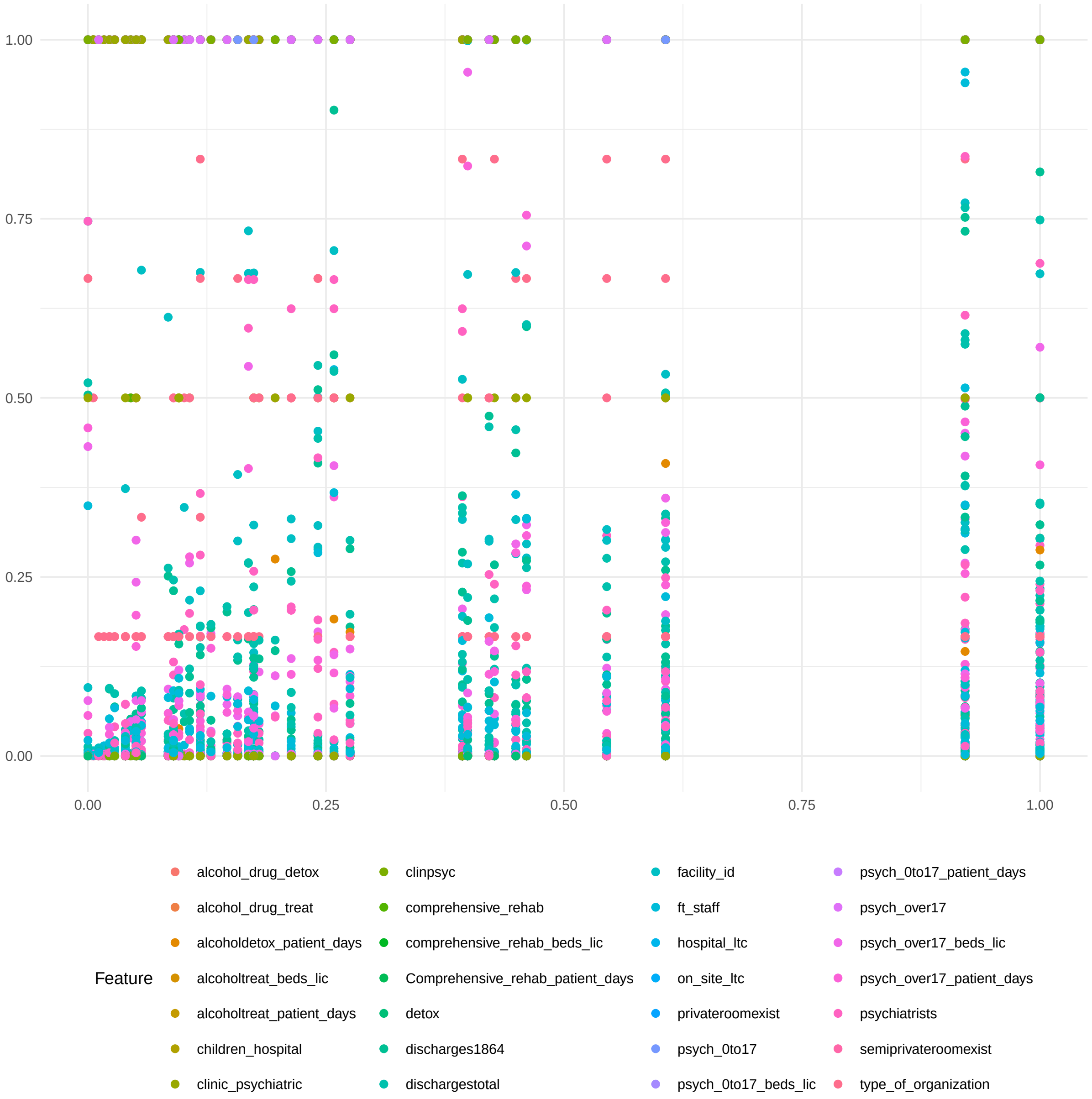
```
##
## Visualize target values
##

# Histogram of target values
target_density <- density(df$target)

# Convert the density estimate to a function
dens_func <- approxfun(target_density$x, target_density$y)

# Use optimize() to find the minimum in a specified interval (choose based on your data)
result <- optimize(dens_func, interval = c(min(df$target), max(df$target)))
local_min_x <- result$minimum     # The x value where local minimum occurs
local_min_y <- result$objective   # The minimum density value

#  Create density plot with ggplot2 and add vertical line at minimum
df_density <- data.frame(x = df$target)
ggplot(df_density, aes(x = df$target)) +
  geom_density(fill = "lightblue", color = "blue", alpha = 0.5) +
  geom_vline(xintercept = local_min_x , color = "red", linetype = "dashed", size = 1) +
  annotate("text", x = local_min_x, y = local_min_y + 0.002,
    label = sprintf("Min: %.0f", local_min_x), color = "red", angle = 90, vjust = -1, size = 6) +
  labs(title = "Density plot of County Observed Suicide rates in PA", x = "County Suicide Rate per 100,000", y = "Density") +
  coord_cartesian(ylim = c(0, 0.01))
```



```
target_bin_cutoff <- local_min_x
```

```
##
```

```r
## Visualize ALL data across target values
## Normalized Scatter
##
minMax <- function(x) { (x - min(x)) / (max(x) - min(x))}
df_norm <- as.data.frame(lapply(df, minMax))
df_long <- pivot_longer(df_norm, cols = -target, names_to = "variable", values_to = "value")
ggplot(df_long, aes(x = target, y = value, color = variable)) +
  geom_point(size = 2) +
  theme_minimal() +
  theme(legend.position = "bottom") +
  guides(color = guide_legend(nrow = 7, title = "Feature")) +
  labs(x = "", y = "", color = "Feature")
```



Preliminary Analysis

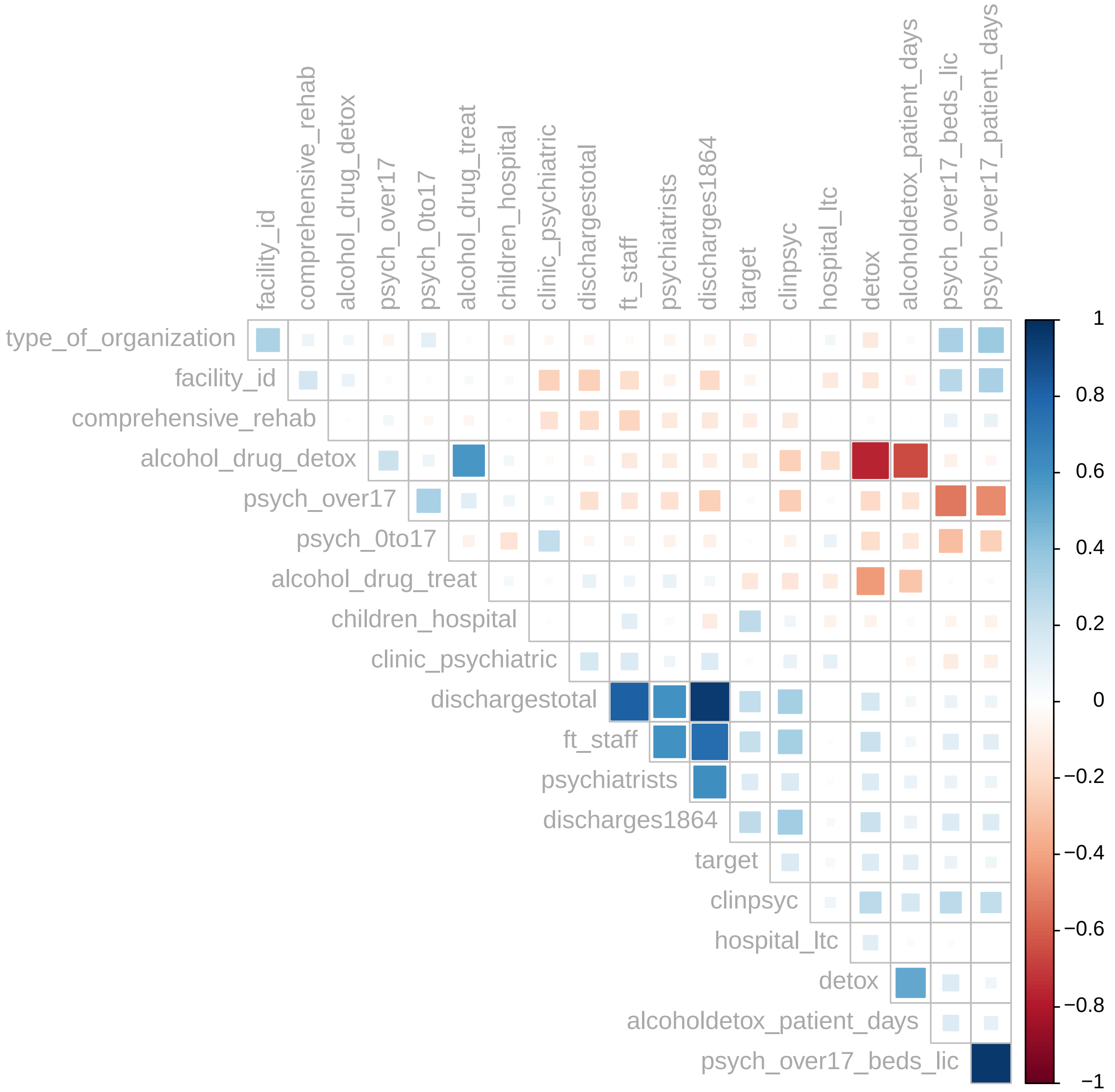
Evaluate Correlation Matrix

```
## Prep for correlation
df_cor <- df


cor_mat <- cor(df)
```

```
cor_threshold <- 0.0
cor_threshold_count <- 2

cols_above_threshold <- which( colSums(abs(cor_mat) > cor_threshold, na.rm = TRUE) >= cor_threshold_count)
df <- subset(df, select = colnames(cor_mat)[cols_above_threshold] )
cor_mat <- cor(df)

cat(colnames(cor_mat)[cols_above_threshold])
```

```
facility_id type_of_organization children_hospital hospital_ltc alcoholdetox_patient_days alcohol_drug_treat comprehensive_rehab psych_0t
```

```
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="square",
  type="upper",
  order="AOE",
  tl.col="darkgrey",
  cl.align.text = "r",
  diag=FALSE,
  number.cex=0.6)
```



Regression

Linear Regression by R Value

Bake off by R squared

```
response <- "target"
predictors <- setdiff(names(df), response)

best_r2 <- -Inf
best_model <- NULL
best_predictor <- NULL

for (predictor in predictors) {
  formula <- as.formula(paste(response, "~", predictor))
  model <- lm(formula, data = df)
  r2 <- summary(model)$r.squared

  if (r2 > best_r2) {
    best_r2 <- r2
    best_model <- model
    best_predictor <- predictor
  }
}

cat("Best single-predictor model uses:", best_predictor, "with R^2 =", best_r2, "\n")
```

```
Best single-predictor model uses: discharges1864 with R^2 = 0.06481609
```

```
# To view details of best_model:
summary(best_model)
```

```
Call:
lm(formula = formula, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-109.26  -48.72  -15.47   32.77  114.25

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    6.374e+01  4.723e+00  13.494  < 2e-16 ***
discharges1864 3.279e-03  8.474e-04   3.869 0.000145 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 57.83 on 216 degrees of freedom
Multiple R-squared:  0.06482,	Adjusted R-squared:  0.06049 
F-statistic: 14.97 on 1 and 216 DF,  p-value: 0.0001446
```

Stepwise Linear Regression

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

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

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

Residuals:
   Min     1Q Median     3Q    Max 
-92.39 -40.46 -13.19  27.54 131.06 

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)    
(Intercept)                9.091e+01  5.477e+01   1.660  0.09852 .  
facility_id               -6.514e-08  2.736e-07  -0.238  0.81209    
type_of_organization      -3.684e+00  3.061e+00  -1.204  0.23020    
children_hospital          1.090e+02  2.202e+01   4.952 1.57e-06 ***
hospital_ltc              -1.148e+01  1.268e+01  -0.905  0.36647    
discharges1864             8.695e-03  2.867e-03   3.033  0.00275 ** 
dischargestotal           -2.171e-03  1.529e-03  -1.420  0.15714    
alcohol_drug_detox         7.467e+00  1.579e+01   0.473  0.63689    
alcoholdetox_patient_days  3.311e-03  3.965e-03   0.835  0.40468    
alcohol_drug_treat        -2.622e+01  1.355e+01  -1.935  0.05447 .  
comprehensive_rehab       -8.492e+00  4.486e+00  -1.893  0.05981 .  
psych_0to17                1.232e+01  6.620e+00   1.861  0.06426 .  
psych_over17               1.192e+01  4.896e+00   2.435  0.01576 *  
psych_over17_beds_lic      3.878e-01  2.778e-01   1.396  0.16433    
psych_over17_patient_days -5.433e-04  8.690e-04  -0.625  0.53260    
detox                      1.076e+01  2.019e+01   0.533  0.59470    
clinpsyc                   7.484e-01  8.877e+00   0.084  0.93290    
clinic_psychiatric        -6.002e+00  4.396e+00  -1.365  0.17377    
psychiatrists             -3.188e-02  1.306e-01  -0.244  0.80744    
```

```
ft_staff                  -1.093e-03  3.107e-03  -0.352  0.72528
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 54.42 on 198 degrees of freedom
Multiple R-squared:  0.2408,    Adjusted R-squared:  0.168
F-statistic: 3.306 on 19 and 198 DF,  p-value: 1.219e-05
```

```
# Perform stepwise regression
step_model_back <- step(model, direction = "backward",trace=0)
summary(step_model_back)
```

```
Call:
lm(formula = target ~ type_of_organization + children_hospital +
    discharges1864 + dischargestotal + alcohol_drug_treat + comprehensive_rehab +
    psych_0to17 + psych_over17 + psych_over17_beds_lic + clinic_psychiatric,
    data = df)

Residuals:
   Min     1Q Median     3Q    Max
-97.79 -40.90 -13.97  28.53 133.73

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)           126.304805  38.424812   3.287  0.00119 **
type_of_organization   -4.603564   2.884664  -1.596  0.11204
children_hospital     106.374554  20.962544   5.075 8.63e-07 ***
discharges1864          0.008263   0.002750   3.005  0.00299 **
dischargestotal        -0.002171   0.001334  -1.627  0.10521
alcohol_drug_treat    -27.746333  10.566288  -2.626  0.00929 **
comprehensive_rehab    -8.282662   4.344900  -1.906  0.05800 .
psych_0to17            10.297480   6.361125   1.619  0.10701
psych_over17           11.449000   4.690830   2.441  0.01550 *
psych_over17_beds_lic   0.231186   0.091302   2.532  0.01208 *
clinic_psychiatric     -6.346224   4.236462  -1.498  0.13566
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 53.65 on 207 degrees of freedom
Multiple R-squared:  0.2285,    Adjusted R-squared:  0.1912
F-statistic: 6.131 on 10 and 207 DF,  p-value: 3.58e-08
```

```r
step_model_forward <- step(model, direction = "forward",trace=0)
summary(step_model_forward)
```

```

Call:
lm(formula = target ~ facility_id + type_of_organization + children_hospital + 
    hospital_ltc + discharges1864 + dischargestotal + alcohol_drug_detox + 
    alcoholdetox_patient_days + alcohol_drug_treat + comprehensive_rehab + 
    psych_0to17 + psych_over17 + psych_over17_beds_lic + psych_over17_patient_days + 
    detox + clinpsyc + clinic_psychiatric + psychiatrists + ft_staff, 
    data = df)

Residuals:
   Min     1Q Median     3Q    Max 
-92.39 -40.46 -13.19  27.54 131.06 

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)    
(Intercept)                9.091e+01  5.477e+01   1.660  0.09852 .  
facility_id               -6.514e-08  2.736e-07  -0.238  0.81209    
type_of_organization      -3.684e+00  3.061e+00  -1.204  0.23020    
children_hospital          1.090e+02  2.202e+01   4.952 1.57e-06 ***
hospital_ltc              -1.148e+01  1.268e+01  -0.905  0.36647    
discharges1864             8.695e-03  2.867e-03   3.033  0.00275 ** 
dischargestotal           -2.171e-03  1.529e-03  -1.420  0.15714    
alcohol_drug_detox         7.467e+00  1.579e+01   0.473  0.63689    
alcoholdetox_patient_days  3.311e-03  3.965e-03   0.835  0.40468    
alcohol_drug_treat        -2.622e+01  1.355e+01  -1.935  0.05447 .  
comprehensive_rehab       -8.492e+00  4.486e+00  -1.893  0.05981 .  
psych_0to17                1.232e+01  6.620e+00   1.861  0.06426 .  
psych_over17               1.192e+01  4.896e+00   2.435  0.01576 *  
psych_over17_beds_lic      3.878e-01  2.778e-01   1.396  0.16433    
psych_over17_patient_days -5.433e-04  8.690e-04  -0.625  0.53260    
detox                      1.076e+01  2.019e+01   0.533  0.59470    
clinpsyc                   7.484e-01  8.877e+00   0.084  0.93290    
clinic_psychiatric        -6.002e+00  4.396e+00  -1.365  0.17377    
psychiatrists             -3.188e-02  1.306e-01  -0.244  0.80744    
ft_staff                  -1.093e-03  3.107e-03  -0.352  0.72528    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 54.42 on 198 degrees of freedom
```

```
Multiple R-squared:  0.2408,    Adjusted R-squared:  0.168
F-statistic: 3.306 on 19 and 198 DF,  p-value: 1.219e-05
```

Logistic Regression

Formula is very simple: glm(y ~ X, family="binomial").

For variables that we want to treat as factors (categorical variables) we use as.factor. R will change it to a dummy, taking the lowest value as 0.

Specifically, after using as.factor(Gender), 1 will become 0 and 2 will become 1.

```r
# Update target to be binomial
df$target <- ifelse(df$target < target_bin_cutoff, 0, 1)

# Logistic Regression with ONLY discharges1864
logreg_ch <- glm(target ~ discharges1864, data=df, family="binomial")
summary(logreg_ch)
```

```
Call:
glm(formula = target ~ discharges1864, family = "binomial", data = df)

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)    -1.600e+00  2.084e-01  -7.677 1.63e-14 ***
discharges1864  9.391e-05  3.215e-05   2.921  0.00349 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 229.83  on 217  degrees of freedom
Residual deviance: 221.17  on 216  degrees of freedom
AIC: 225.17

Number of Fisher Scoring iterations: 4
```

```r
# Logistic Regression with discharges1864 +
#  children_hospital + psych_over17  + psych_over17_beds_lic
logreg_chdisch <- glm(target ~ discharges1864  +  children_hospital + psych_over17
             + psych_over17_beds_lic, data=df, family="binomial")
summary(logreg_chdisch)
```

```
Call:
glm(formula = target ~ discharges1864 + children_hospital + psych_over17 + 
    psych_over17_beds_lic, family = "binomial", data = df)

Coefficients:
                        Estimate Std. Error z value Pr(>|z|)    
(Intercept)           -2.271e+00  6.144e-01  -3.697 0.000218 ***
discharges1864         1.180e-04  3.465e-05   3.404 0.000663 ***
children_hospital      2.849e+00  8.489e-01   3.356 0.000790 ***
psych_over17           1.836e-01  2.194e-01   0.837 0.402656    
psych_over17_beds_lic  1.335e-03  3.948e-03   0.338 0.735195    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 229.83  on 217  degrees of freedom
Residual deviance: 206.84  on 213  degrees of freedom
AIC: 216.84

Number of Fisher Scoring iterations: 4
```

```r
logreg <- logreg_chdisch
coeftable <- data.frame(col1=coef(logreg),col2=exp(coef(logreg)))
colnames(coeftable)<-c('Coefficient (log-odds)','e^coefficient (odds)')
coeftable
```

```
                      Coefficient (log-odds) e^coefficient (odds)
(Intercept)                    -2.2712892545            0.1031791
discharges1864                  0.0001179614            1.0001180
children_hospital               2.8490898212           17.2720540
psych_over17                    0.1835961539            1.2015305
psych_over17_beds_lic           0.0013351614            1.0013361
```

```r
#
# Confusion Matrix
#
df$PredLogOdds <- df$PredProbs <- predict(logreg, newdata=df)
df$PredProbs <- predict(logreg, newdata=df, type="response")
# type="response" gives the probability
summary(df$PredProbs)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
 0.1110  0.1530  0.1606  0.2202  0.2018  0.8103 
```

```r
# transform prediction into either 1 (Elevated Suicide Rate) or 0 (Suicide Rate) using a cutoff point
for (cutoff in c(0.25, 0.15, 0.1, 0.05))
{
  df$PredHighSuicide <- ifelse(df$PredProbs >= cutoff,1,0)
  summary(df$PredHighSuicide)
  cat(paste("For a cutoff point of",
    cutoff, "the proportion of Counties predicted to have high suicide rates is",
    round(mean(df$PredHighSuicide),2)), "\n\n")
}
```

```
For a cutoff point of 0.25 the proportion of Counties predicted to have high suicide rates is 0.18 

For a cutoff point of 0.15 the proportion of Counties predicted to have high suicide rates is 0.83 

For a cutoff point of 0.1 the proportion of Counties predicted to have high suicide rates is 1 

For a cutoff point of 0.05 the proportion of Counties predicted to have high suicide rates is 1 
```

References

Hospital Data: https://www.pa.gov/agencies/health/health-statistics/health-facilities/hospital-reports

Suicide by County Data: https://www.phaim.health.pa.gov/EDD/WebForms/DeathCntySt.aspx