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

Data Preparation

Loading packages

Importing data

Data Cleaning & Wrangling

```
summary(df)
```

```
  facility_id       type_of_organization children_hospital  hospital_ltc
 Min.   :    2984   Min.   :1.000        Min.   :0.0000    Min.   :0.0000
 1st Qu.:  141176   1st Qu.:2.000        1st Qu.:0.0000    1st Qu.:0.0000
 Median :  390628   Median :2.000        Median :0.0000    Median :0.0000
 Mean   : 8421653   Mean   :2.729        Mean   :0.0367    Mean   :0.1055
 3rd Qu.: 8402600   3rd Qu.:4.000        3rd Qu.:0.0000    3rd Qu.:0.0000
 Max.   :75140100   Max.   :7.000        Max.   :1.0000    Max.   :1.0000

  on_site_ltc    privateroomexist semiprivateroomexist discharges017
 Min.   :0.000   Min.   :0.0000   Min.   :0.0000       Min.   :   -1.0
 1st Qu.:1.000   1st Qu.:1.0000   1st Qu.:1.0000       1st Qu.:    0.0
 Median :1.000   Median :1.0000   Median :1.0000       Median :    6.0
 Mean   :0.913   Mean   :0.8692   Mean   :0.7981       Mean   :  499.0
 3rd Qu.:1.000   3rd Qu.:1.0000   3rd Qu.:1.0000       3rd Qu.:  301.5
 Max.   :1.000   Max.   :1.0000   Max.   :1.0000       Max.   :31133.0
 NA's   :195     NA's   :4        NA's   :5
 discharges1864    discharges65over dischargestotal   alcohol_drug_detox
 Min.   :    0.0   Min.   :    0    Min.   :    0.0   Min.   :1.000
```

```
1st Qu.:  244.8   1st Qu.:  240    1st Qu.:  892.5   1st Qu.:3.000
Median : 1605.0   Median : 1272    Median : 3107.5   Median :3.000
Mean   : 3115.4   Mean   : 3500    Mean   : 6929.8   Mean   :2.867
3rd Qu.: 3831.5   3rd Qu.: 4968    3rd Qu.: 9530.8   3rd Qu.:3.000
Max.   :27544.0   Max.   :56730    Max.   :58930.0   Max.   :3.000

alcoholdetox_beds_lic alcoholdetox_beds_staf alcoholdetox_admissions
Min.   :-1.000        Min.   :-1.000         Min.   :  -1.00
1st Qu.: 0.000        1st Qu.: 0.000         1st Qu.:   0.00
Median : 0.000        Median : 0.000         Median :   0.00
Mean   : 1.239        Mean   : 1.092         Mean   :  39.34
3rd Qu.: 0.000        3rd Qu.: 0.000         3rd Qu.:   0.00
Max.   :50.000        Max.   :49.000         Max.   :1414.00

alcoholdetox_patient_days alcoholdetox_bed_days_avail alcohol_drug_treat
Min.   :   -1.0           Min.   :   -1.0             Min.   :1.000
1st Qu.:    0.0           1st Qu.:    0.0             1st Qu.:3.000
Median :    0.0           Median :    0.0             Median :3.000
Mean   :  231.3           Mean   :  419.8             Mean   :2.936
3rd Qu.:    0.0           3rd Qu.:    0.0             3rd Qu.:3.000
Max.   :15468.0           Max.   :18250.0             Max.   :3.000

alcoholtreat_beds_lic alcoholtreat_beds_staf alcoholtreat_admissions
Min.   : -1.000       Min.   : -1.000        Min.   :  -1.00
1st Qu.:  0.000       1st Qu.:  0.000        1st Qu.:   0.00
Median :  0.000       Median :  0.000        Median :   0.00
Mean   :  2.558       Mean   :  2.599        Mean   :  25.18
3rd Qu.:  0.000       3rd Qu.:  0.000        3rd Qu.:   0.00
Max.   :220.000       Max.   :220.000        Max.   :2954.00
NA's   :1             NA's   :1              NA's   :1
alcoholtreat_patient_days alcoholtreat_bed_days_avail comprehensive_rehab
Min.   :   -1.0           Min.   :   -1               Min.   :1.000
1st Qu.:    0.0           1st Qu.:    0               1st Qu.:1.000
Median :    0.0           Median :    0               Median :3.000
Mean   :  476.6           Mean   :  745               Mean   :2.472
3rd Qu.:    0.0           3rd Qu.:    0               3rd Qu.:3.000
Max.   :65169.0           Max.   :80300               Max.   :3.000
NA's   :1                 NA's   :1
comprehensive_rehab_beds_lic Comprehensive_rehab_beds_staf
Min.   : -1.000              Min.   : -1.000
1st Qu.:  0.000              1st Qu.:  0.000
Median :  0.000              Median :  0.000
Mean   :  9.866              Mean   :  8.382
```

```
3rd Qu.:  9.000              3rd Qu.:  9.000
Max.   :187.000              Max.   :148.000
NA's   :1                    NA's   :1
Comprehensive_rehab_admissions Comprehensive_rehab_patient_days
Min.   :  -1.0                 Min.   :   -1
1st Qu.:   0.0                 1st Qu.:    0
Median :   0.0                 Median :    0
Mean   : 184.4                 Mean   : 2476
3rd Qu.: 143.0                 3rd Qu.: 1734
Max.   :4798.0                 Max.   :43378
NA's   :1                      NA's   :1
Comprehensive_rehab_bed_days_avail  psych_0to17    psych_0to17_beds_lic
Min.   :   -1                      Min.   :1.000   Min.   : -1.000
1st Qu.:    0                      1st Qu.:3.000   1st Qu.:  0.000
Median :    0                      Median :3.000   Median :  0.000
Mean   : 3389                      Mean   :2.752   Mean   :  5.286
3rd Qu.: 3285                      3rd Qu.:3.000   3rd Qu.:  0.000
Max.   :68255                      Max.   :3.000   Max.   :186.000
NA's   :1                                          NA's   :1
psych_0to17_beds_staf psych_0to17_admissions psych_0to17_patient_days
Min.   : -1.00        Min.   :  -1.00        Min.   :   -1
1st Qu.:  0.00        1st Qu.:   0.00        1st Qu.:    0
Median :  0.00        Median :   0.00        Median :    0
Mean   :  4.53        Mean   :  71.25        Mean   : 1400
3rd Qu.:  0.00        3rd Qu.:   0.00        3rd Qu.:    0
Max.   :152.00        Max.   :1604.00        Max.   :53196
NA's   :1             NA's   :1              NA's   :1
psych_0to17_bed_days_avail  psych_over17   psych_over17_beds_lic
Min.   :    -1              Min.   :1.000   Min.   : -1.00
1st Qu.:     0              1st Qu.:1.000   1st Qu.:  0.00
Median :     0              Median :3.000   Median :  0.00
Mean   :  2252              Mean   :2.193   Mean   : 24.74
3rd Qu.:     0              3rd Qu.:3.000   3rd Qu.: 28.00
Max.   :138700              Max.   :3.000   Max.   :374.00
NA's   :1
psych_over17_beds_staf psych_over17_admissions psych_over17_patient_days
Min.   : -1.00         Min.   :  -1.0          Min.   :    -1
1st Qu.:  0.00         1st Qu.:   0.0          1st Qu.:     0
Median :  0.00         Median :   0.0          Median :     0
Mean   : 22.72         Mean   : 329.4          Mean   :  6484
3rd Qu.: 24.00         3rd Qu.: 415.2          3rd Qu.:  5730
Max.   :374.00         Max.   :4193.0          Max.   :126282
```

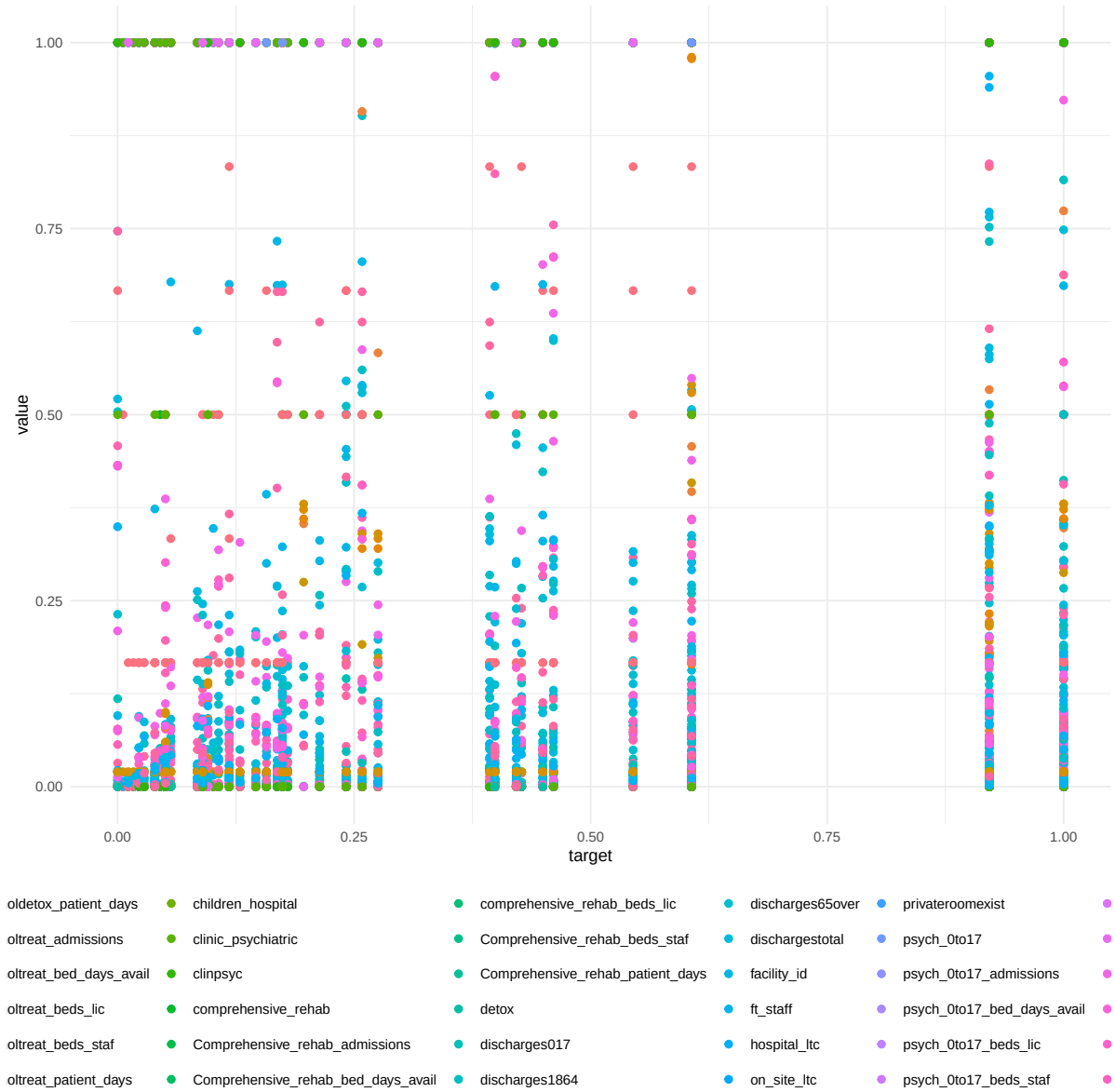
```
 psych_over17_bed_days_avail     detox           clinpsyc
 Min.   :    -1              Min.   :0.0000   Min.   :0.0000
 1st Qu.:     0              1st Qu.:0.0000   1st Qu.:0.0000
 Median :     0              Median :0.0000   Median :0.0000
 Mean   :  8290              Mean   :0.1101   Mean   :0.3578
 3rd Qu.:  9490              3rd Qu.:0.0000   3rd Qu.:1.0000
 Max.   :136510              Max.   :1.0000   Max.   :1.0000

 clinic_psychiatric psychiatrists       ft_staff           target
 Min.   :1.00       Min.   :  0.00   Min.   :    9.0   Min.   :  0.00
 1st Qu.:1.00       1st Qu.:  2.00   1st Qu.:  148.5   1st Qu.: 21.00
 Median :1.00       Median :  7.00   Median :  388.0   Median : 70.00
 Mean   :1.83       Mean   : 22.83   Mean   : 1132.4   Mean   : 73.95
 3rd Qu.:3.00       3rd Qu.: 26.00   3rd Qu.: 1080.2   3rd Qu.:108.00
 Max.   :3.00       Max.   :221.00   Max.   :16948.0   Max.   :178.00
```

```r
## 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 = 6))
```

```
Warning: Removed 3924 rows containing missing values or values outside the scale range
(`geom_point()`).
```



Preliminary Analysis

Evaluate Correlation Matrix

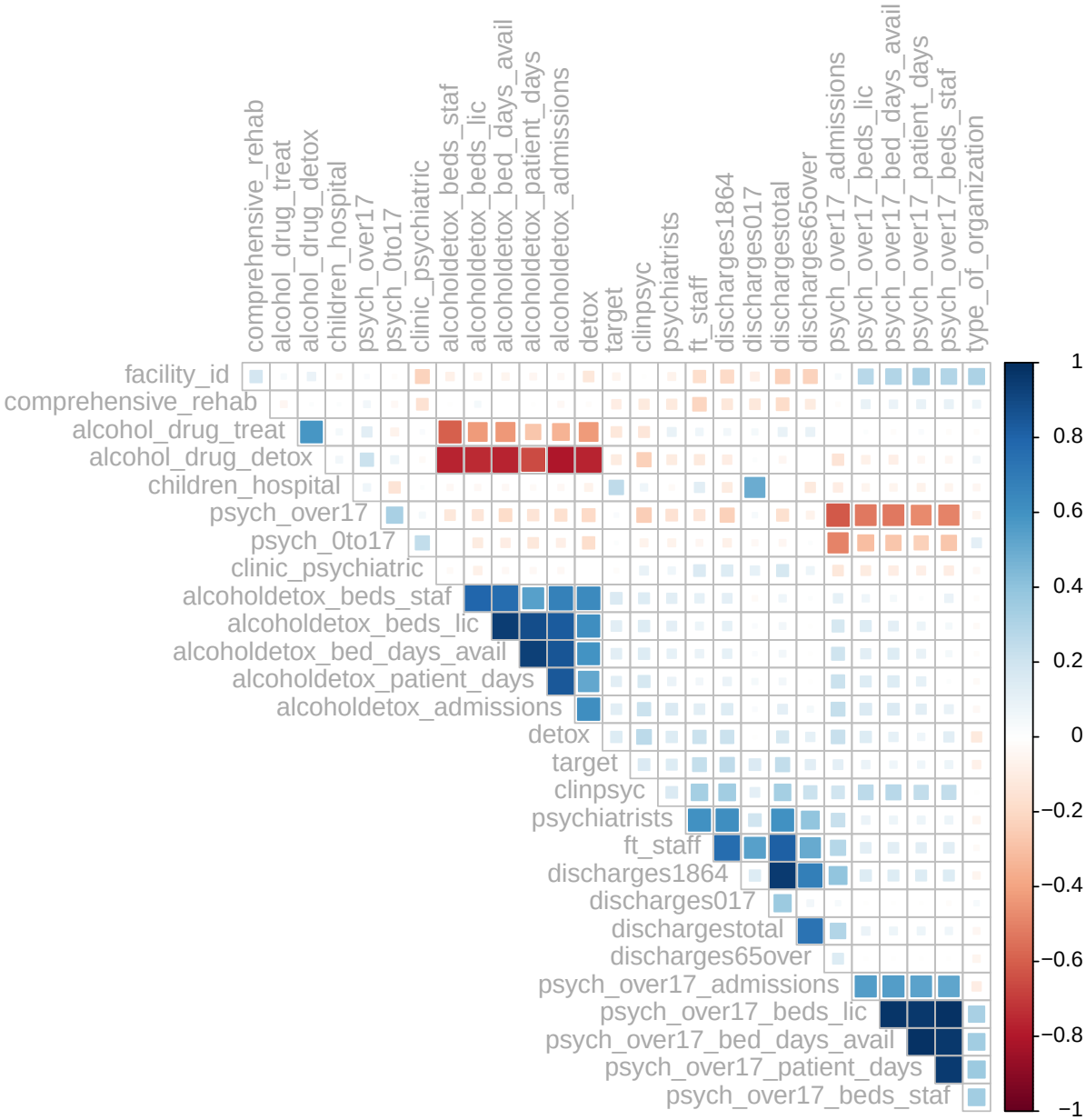
```
## Prep for correlation
df_cor <- df

cor_mat <- cor(df)
```

```r
cor_threshold <- 0.2
cor_threshold_count <- 2

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

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

```r
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
```

```r
# 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
-90.77 -38.52 -12.72  25.51 129.79

Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                 8.725e+01  7.083e+01   1.232   0.2195
facility_id                -2.653e-08  2.728e-07  -0.097   0.9226
type_of_organization       -2.901e+00  3.240e+00  -0.895   0.3717
children_hospital           1.046e+02  2.438e+01   4.289 2.86e-05 ***
discharges017               2.913e-03  3.220e-03   0.905   0.3668
discharges1864              1.008e-02  4.032e-03   2.501   0.0132 *
discharges65over            1.933e-04  1.065e-03   0.182   0.8561
dischargestotal            -3.003e-03  2.189e-03  -1.372   0.1718
alcohol_drug_detox          5.686e+00  1.967e+01   0.289   0.7728
alcoholdetox_beds_lic      -5.195e+00  3.015e+00  -1.723   0.0866 .
alcoholdetox_beds_staf      4.874e+00  2.924e+00   1.667   0.0972 .
alcoholdetox_admissions    -1.764e-02  5.082e-02  -0.347   0.7289
alcoholdetox_patient_days   1.478e-02  1.207e-02   1.224   0.2223
alcoholdetox_bed_days_avail -2.775e-03  9.280e-03  -0.299   0.7653
alcohol_drug_treat         -2.716e+01  1.536e+01  -1.768   0.0787 .
comprehensive_rehab        -8.063e+00  4.580e+00  -1.760   0.0799 .
psych_0to17                 1.277e+01  7.310e+00   1.747   0.0822 .
psych_over17                1.426e+01  5.762e+00   2.474   0.0142 *
psych_over17_beds_lic       1.876e+00  8.230e-01   2.279   0.0238 *
```

```
psych_over17_beds_staf      -1.155e+00  8.281e-01  -1.395   0.1646
psych_over17_admissions      5.905e-03  1.016e-02   0.581   0.5616
psych_over17_patient_days    7.089e-04  1.449e-03   0.489   0.6253
psych_over17_bed_days_avail -2.119e-03  1.665e-03  -1.273   0.2047
detox                        6.054e+00  2.101e+01   0.288   0.7736
clinpsyc                     1.296e+00  9.201e+00   0.141   0.8881
clinic_psychiatric          -6.716e+00  4.524e+00  -1.485   0.1393
psychiatrists               -8.309e-03  1.332e-01  -0.062   0.9503
ft_staff                    -2.651e-03  3.730e-03  -0.711   0.4782
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 54.56 on 190 degrees of freedom
Multiple R-squared:  0.2678,    Adjusted R-squared:  0.1638
F-statistic: 2.574 on 27 and 190 DF,  p-value: 0.000106
```

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

```
Call:
lm(formula = target ~ children_hospital + discharges1864 + dischargestotal +
   alcoholdetox_beds_lic + alcoholdetox_beds_staf + alcoholdetox_patient_days +
    alcohol_drug_treat + comprehensive_rehab + psych_0to17 +
    psych_over17 + psych_over17_beds_lic + psych_over17_beds_staf +
    psych_over17_bed_days_avail + clinic_psychiatric, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-101.53  -37.56  -11.06   25.96  127.15

Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                103.426361  46.311553   2.233  0.02662 *
children_hospital          110.448927  20.877021   5.290 3.15e-07 ***
discharges1864               0.008473   0.002751   3.080  0.00235 **
dischargestotal             -0.002304   0.001329  -1.734  0.08443 .
alcoholdetox_beds_lic       -5.808436   2.787718  -2.084  0.03845 *
alcoholdetox_beds_staf       4.457653   2.554447   1.745  0.08249 .
alcoholdetox_patient_days    0.011406   0.007423   1.537  0.12596
alcohol_drug_treat         -26.265110  13.439774  -1.954  0.05204 .
```

```
comprehensive_rehab          -7.594914   4.369580  -1.738  0.08370 .
psych_0to17                   9.595655   6.334926   1.515  0.13140
psych_over17                 13.529691   4.972486   2.721  0.00708 **
psych_over17_beds_lic         2.017526   0.774642   2.604  0.00988 **
psych_over17_beds_staf       -1.223581   0.778816  -1.571  0.11772
psych_over17_bed_days_avail  -0.001760   0.001002  -1.757  0.08040 .
clinic_psychiatric           -6.548900   4.228867  -1.549  0.12303
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 53.32 on 203 degrees of freedom
Multiple R-squared:  0.2527,    Adjusted R-squared:  0.2012
F-statistic: 4.904 on 14 and 203 DF,  p-value: 8.389e-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 +
    discharges017 + discharges1864 + discharges65over + dischargestotal +
    alcohol_drug_detox + alcoholdetox_beds_lic + alcoholdetox_beds_staf +
   alcoholdetox_admissions + alcoholdetox_patient_days + alcoholdetox_bed_days_avail +
    alcohol_drug_treat + comprehensive_rehab + psych_0to17 +
    psych_over17 + psych_over17_beds_lic + psych_over17_beds_staf +
   psych_over17_admissions + psych_over17_patient_days + psych_over17_bed_days_avail +
    detox + clinpsyc + clinic_psychiatric + psychiatrists + ft_staff,
    data = df)

Residuals:
   Min     1Q Median     3Q    Max
-90.77 -38.52 -12.72  25.51 129.79

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                  8.725e+01  7.083e+01   1.232   0.2195
facility_id                 -2.653e-08  2.728e-07  -0.097   0.9226
type_of_organization        -2.901e+00  3.240e+00  -0.895   0.3717
children_hospital            1.046e+02  2.438e+01   4.289 2.86e-05 ***
discharges017                2.913e-03  3.220e-03   0.905   0.3668
discharges1864               1.008e-02  4.032e-03   2.501   0.0132 *
discharges65over             1.933e-04  1.065e-03   0.182   0.8561
```

```
dischargestotal             -3.003e-03  2.189e-03  -1.372   0.1718
alcohol_drug_detox           5.686e+00  1.967e+01   0.289   0.7728
alcoholdetox_beds_lic       -5.195e+00  3.015e+00  -1.723   0.0866 .
alcoholdetox_beds_staf       4.874e+00  2.924e+00   1.667   0.0972 .
alcoholdetox_admissions     -1.764e-02  5.082e-02  -0.347   0.7289
alcoholdetox_patient_days    1.478e-02  1.207e-02   1.224   0.2223
alcoholdetox_bed_days_avail -2.775e-03  9.280e-03  -0.299   0.7653
alcohol_drug_treat          -2.716e+01  1.536e+01  -1.768   0.0787 .
comprehensive_rehab         -8.063e+00  4.580e+00  -1.760   0.0799 .
psych_0to17                  1.277e+01  7.310e+00   1.747   0.0822 .
psych_over17                 1.426e+01  5.762e+00   2.474   0.0142 *
psych_over17_beds_lic        1.876e+00  8.230e-01   2.279   0.0238 *
psych_over17_beds_staf      -1.155e+00  8.281e-01  -1.395   0.1646
psych_over17_admissions      5.905e-03  1.016e-02   0.581   0.5616
psych_over17_patient_days    7.089e-04  1.449e-03   0.489   0.6253
psych_over17_bed_days_avail -2.119e-03  1.665e-03  -1.273   0.2047
detox                        6.054e+00  2.101e+01   0.288   0.7736
clinpsyc                     1.296e+00  9.201e+00   0.141   0.8881
clinic_psychiatric          -6.716e+00  4.524e+00  -1.485   0.1393
psychiatrists               -8.309e-03  1.332e-01  -0.062   0.9503
ft_staff                    -2.651e-03  3.730e-03  -0.711   0.4782
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 54.56 on 190 degrees of freedom
Multiple R-squared:  0.2678,    Adjusted R-squared:  0.1638
F-statistic: 2.574 on 27 and 190 DF,  p-value: 0.000106
```

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.

```
logreg_ch <- glm(target ~ children_hospital, data=df, family="gaussian")
summary(logreg_ch)
```

```
Call:
```

```
glm(formula = target ~ children_hospital, family = "gaussian",
    data = df)

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)         71.090      3.999  17.776  < 2e-16 ***
children_hospital   78.035     20.876   3.738 0.000238 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 3358.556)

    Null deviance: 772376  on 217  degrees of freedom
Residual deviance: 725448  on 216  degrees of freedom
AIC: 2392.6

Number of Fisher Scoring iterations: 2
```

```r
logreg_chdisch <- glm(target ~ children_hospital  + discharges1864 + psych_over17
             + psych_over17_beds_lic, data=df, family="gaussian")
summary(logreg_chdisch)
```

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

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)           3.614e+01  1.267e+01   2.853  0.00476 **
children_hospital     8.831e+01  2.005e+01   4.405 1.68e-05 ***
discharges1864        3.866e-03  8.393e-04   4.606 7.07e-06 ***
psych_over17          8.617e+00  4.620e+00   1.865  0.06351 .
psych_over17_beds_lic 1.469e-01  8.411e-02   1.747  0.08209 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 3053.037)

    Null deviance: 772376  on 217  degrees of freedom
Residual deviance: 650297  on 213  degrees of freedom
AIC: 2374.8
```

```
Number of Fisher Scoring iterations: 2
```

```
logreg <- logreg_chdisch
coef(logreg)
```

```
          (Intercept)     children_hospital        discharges1864
         36.140358129          88.314299303           0.003865692
         psych_over17 psych_over17_beds_lic
          8.616883589           0.146937046
```

```
exp(coef(logreg))
```

```
          (Intercept)     children_hospital        discharges1864
         4.960873e+15          2.261585e+38          1.003873e+00
         psych_over17 psych_over17_beds_lic
         5.524144e+03          1.158281e+00
```

```
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)                     36.140358129         4.960873e+15
children_hospital               88.314299303         2.261585e+38
discharges1864                   0.003865692         1.003873e+00
psych_over17                     8.616883589         5.524144e+03
psych_over17_beds_lic            0.146937046         1.158281e+00
```

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