Final Project

CUNY SPS Masters in Data Science - DATA 698

Ali Harb
Dilip Ganesan
Raghunathan Ramnath

May 10, 2019

Contents

1 Load cleaned Data

The current data set is composed of 99950 records and 53 variables.

2 Summary Statistics

```
##        X                race           gender            age
##  Min.   :     1   Min.   :1.000   Min.   :0.0000   Min.   : 5.00
##  1st Qu.: 25406   1st Qu.:4.000   1st Qu.:0.0000   1st Qu.:55.00
##  Median : 50921   Median :4.000   Median :0.0000   Median :65.00
##  Mean   : 50902   Mean   :3.525   Mean   :0.4627   Mean   :65.88
##  3rd Qu.: 76385   3rd Qu.:4.000   3rd Qu.:1.0000   3rd Qu.:75.00
##  Max.   :101766   Max.   :5.000   Max.   :1.0000   Max.   :95.00
##
##  admission_type_id discharge_disposition_id admission_source_id
##  Min.   :1.000     Min.   : 1.000           Min.   : 1.000
##  1st Qu.:1.000     1st Qu.: 1.000           1st Qu.: 1.000
##  Median :1.000     Median : 1.000           Median : 7.000
##  Mean   :1.783     Mean   : 2.601           Mean   : 5.887
##  3rd Qu.:3.000     3rd Qu.: 2.000           3rd Qu.: 7.000
##  Max.   :5.000     Max.   :25.000           Max.   :20.000
##
##  time_in_hospital num_lab_procedures num_procedures num_medications
##  Min.   : 1.000   Min.   :  1.00     Min.   :0.00   Min.   : 1.00
##  1st Qu.: 2.000   1st Qu.: 31.00     1st Qu.:0.00   1st Qu.:10.00
##  Median : 4.000   Median : 44.00     Median :1.00   Median :15.00
##  Mean   : 4.391   Mean   : 42.97     Mean   :1.33   Mean   :15.98
##  3rd Qu.: 6.000   3rd Qu.: 57.00     3rd Qu.:2.00   3rd Qu.:20.00
##  Max.   :14.000   Max.   :132.00     Max.   :6.00   Max.   :81.00
##
##  number_outpatient number_emergency  number_inpatient  number_diagnoses
##  Min.   : 0.0000   Min.   : 0.0000   Min.   : 0.0000   Min.   : 1.00
##  1st Qu.: 0.0000   1st Qu.: 0.0000   1st Qu.: 0.0000   1st Qu.: 6.00
##  Median : 0.0000   Median : 0.0000   Median : 0.0000   Median : 8.00
##  Mean   : 0.3687   Mean   : 0.1985   Mean   : 0.6334   Mean   : 7.41
##  3rd Qu.: 0.0000   3rd Qu.: 0.0000   3rd Qu.: 1.0000   3rd Qu.: 9.00
##  Max.   :42.0000   Max.   :76.0000   Max.   :21.0000   Max.   :16.00
##
##  max_glu_serum     A1Cresult        metformin      repaglinide
##  Min.   :-99.0   Min.   :-99.00   Min.   :0.000   Min.   :0.00000
##  1st Qu.:-99.0   1st Qu.:-99.00   1st Qu.:0.000   1st Qu.:0.00000
##  Median :-99.0   Median :-99.00   Median :0.000   Median :0.00000
##  Mean   :-93.8   Mean   :-82.18   Mean   :0.199   Mean   :0.01527
##  3rd Qu.:-99.0   3rd Qu.:-99.00   3rd Qu.:0.000   3rd Qu.:0.00000
##  Max.   :  1.0   Max.   :  1.00   Max.   :1.000   Max.   :1.00000
##
##   nateglinide       chlorpropamide       glimepiride      acetohexamide
##  Min.   :0.000000   Min.   :0.0000000   Min.   :0.00000   Min.   :0e+00
##  1st Qu.:0.000000   1st Qu.:0.0000000   1st Qu.:0.00000   1st Qu.:0e+00
##  Median :0.000000   Median :0.0000000   Median :0.00000   Median :0e+00
##  Mean   :0.006944   Mean   :0.0008504   Mean   :0.05148   Mean   :1e-05
```

```
##  3rd Qu.:0.000000   3rd Qu.:0.0000000   3rd Qu.:0.00000   3rd Qu.:0e+00
##  Max.   :1.000000   Max.   :1.0000000   Max.   :1.00000   Max.   :1e+00
##
##    glipizide         glyburide        tolbutamide          pioglitazone
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.0000000   Min.   :0.00000
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000000   1st Qu.:0.00000
##  Median :0.0000   Median :0.0000   Median :0.0000000   Median :0.00000
##  Mean   :0.1261   Mean   :0.1056   Mean   :0.0002101   Mean   :0.07283
##  3rd Qu.:0.0000   3rd Qu.:0.0000   3rd Qu.:0.0000000   3rd Qu.:0.00000
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.0000000   Max.   :1.00000
##
##  rosiglitazone        acarbose           miglitol          troglitazone
##  Min.   :0.00000   Min.   :0.000000   Min.   :0.0000000   Min.   :0e+00
##  1st Qu.:0.00000   1st Qu.:0.000000   1st Qu.:0.0000000   1st Qu.:0e+00
##  Median :0.00000   Median :0.000000   Median :0.0000000   Median :0e+00
##  Mean   :0.06331   Mean   :0.003072   Mean   :0.0003802   Mean   :3e-05
##  3rd Qu.:0.00000   3rd Qu.:0.000000   3rd Qu.:0.0000000   3rd Qu.:0e+00
##  Max.   :1.00000   Max.   :1.000000   Max.   :1.0000000   Max.   :1e+00
##
##    tolazamide           insulin       glyburide.metformin
##  Min.   :0.0000000   Min.   :0.0000   Min.   :0.000000
##  1st Qu.:0.0000000   1st Qu.:0.0000   1st Qu.:0.000000
##  Median :0.0000000   Median :1.0000   Median :0.000000
##  Mean   :0.0003902   Mean   :0.5337   Mean   :0.006964
##  3rd Qu.:0.0000000   3rd Qu.:1.0000   3rd Qu.:0.000000
##  Max.   :1.0000000   Max.   :1.0000   Max.   :1.000000
##
##  glipizide.metformin glimepiride.pioglitazone metformin.rosiglitazone
##  Min.   :0.0000000   Min.   :0e+00            Min.   :0e+00
##  1st Qu.:0.0000000   1st Qu.:0e+00            1st Qu.:0e+00
##  Median :0.0000000   Median :0e+00            Median :0e+00
##  Mean   :0.0001301   Mean   :1e-05            Mean   :2e-05
##  3rd Qu.:0.0000000   3rd Qu.:0e+00            3rd Qu.:0e+00
##  Max.   :1.0000000   Max.   :1e+00            Max.   :1e+00
##
##  metformin.pioglitazone     change        diabetesMed       readmitted
##  Min.   :0e+00          Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0e+00          1st Qu.:0.0000   1st Qu.:1.0000   1st Qu.:0.0000
##  Median :0e+00          Median :0.0000   Median :1.0000   Median :0.0000
##  Mean   :1e-05          Mean   :0.4641   Mean   :0.7721   Mean   :0.1135
##  3rd Qu.:0e+00          3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:0.0000
##  Max.   :1e+00          Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
##
##   Primary_Diag   Secondary_Diag_1 Secondary_Diag_2      DRG
##  4280   : 6735   2760   : 6652    25000  :11463    Min.   : 53.0
##  41400  : 6554   4280   : 6531    4010   : 8248    1st Qu.:203.0
##  78600  : 4016   25000  : 6055    2760   : 5020    Median :316.0
##  41000  : 3477   4270   : 4928    4280   : 4461    Mean   :402.3
```

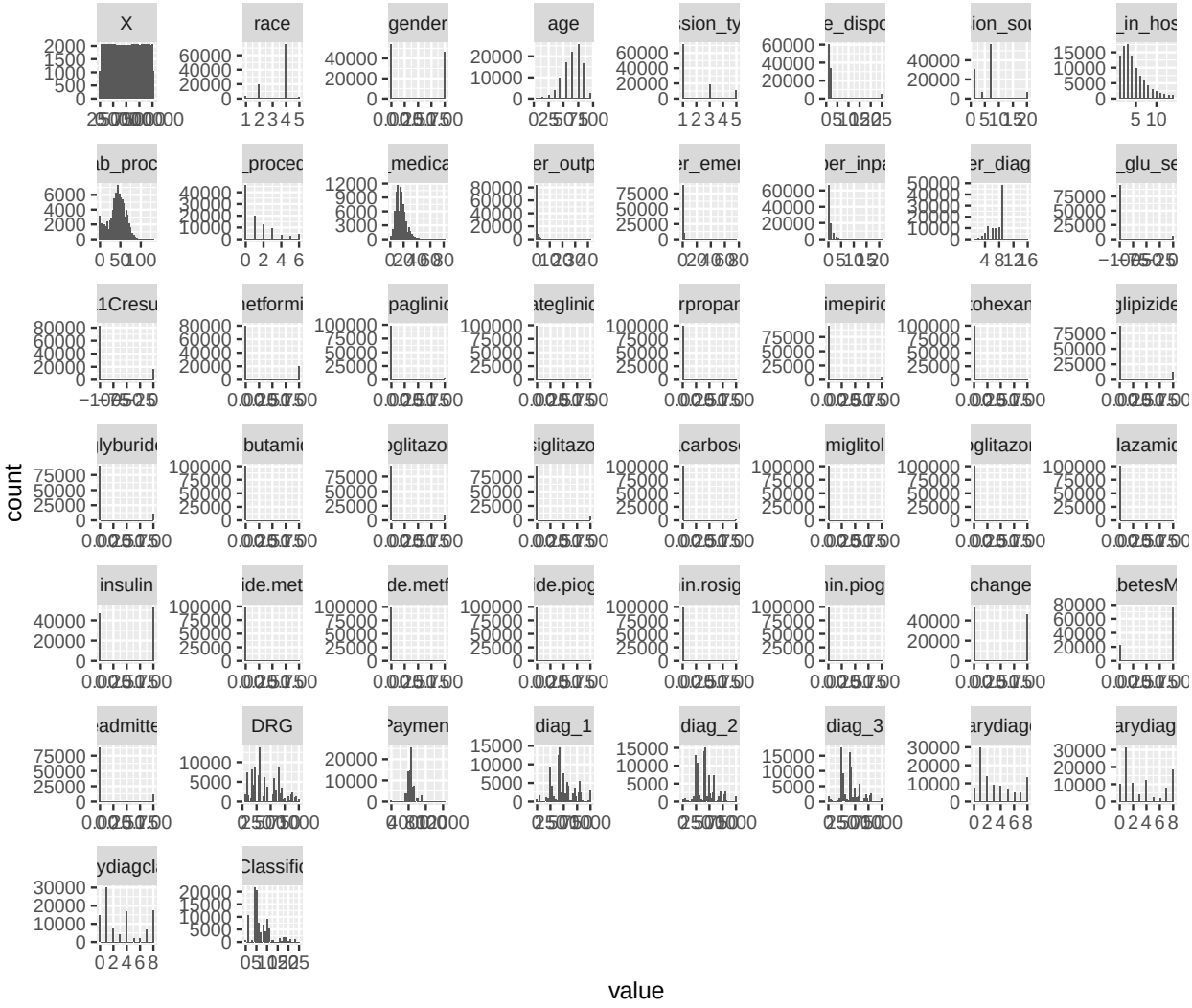
```
##  486    : 3413   4010   : 3725    4270   : 3831    3rd Qu.:603.0
##  4270   : 2729   496    : 3269    41400  : 3646    Max.   :999.0
##  (Other):73026   (Other):68790    (Other):63281
##     Payment          diag_1          diag_2          diag_3
##  Min.   :    0   Min.   :  0.0   Min.   :  0.0   Min.   :  0.0
##  1st Qu.: 4033   1st Qu.:410.0   1st Qu.:276.0   1st Qu.:272.0
##  Median : 4436   Median :440.0   Median :425.0   Median :403.0
##  Mean   : 4503   Mean   :494.5   Mean   :436.6   Mean   :411.6
##  3rd Qu.: 4669   3rd Qu.:599.0   3rd Qu.:530.0   3rd Qu.:496.0
##  Max.   :12398   Max.   :999.0   Max.   :999.0   Max.   :999.0
##                  NA's   :1623    NA's   :2524    NA's   :5025
##  primarydiagclass secondarydiagclass_1 tertiarydiagclass_1
##  Min.   :0.000    Min.   :0.000        Min.   :0.0
##  1st Qu.:1.000    1st Qu.:1.000        1st Qu.:1.0
##  Median :2.000    Median :2.000        Median :2.0
##  Mean   :3.299    Mean   :3.436        Mean   :3.3
##  3rd Qu.:5.000    3rd Qu.:7.000        3rd Qu.:6.0
##  Max.   :8.000    Max.   :8.000        Max.   :8.0
##
##  DRGClassification
##  Min.   : 0.000
##  1st Qu.: 4.000
##  Median : 5.000
##  Mean   : 6.865
##  3rd Qu.: 9.000
##  Max.   :25.000
##
```

3 Data Analysis

3.1 Histogram

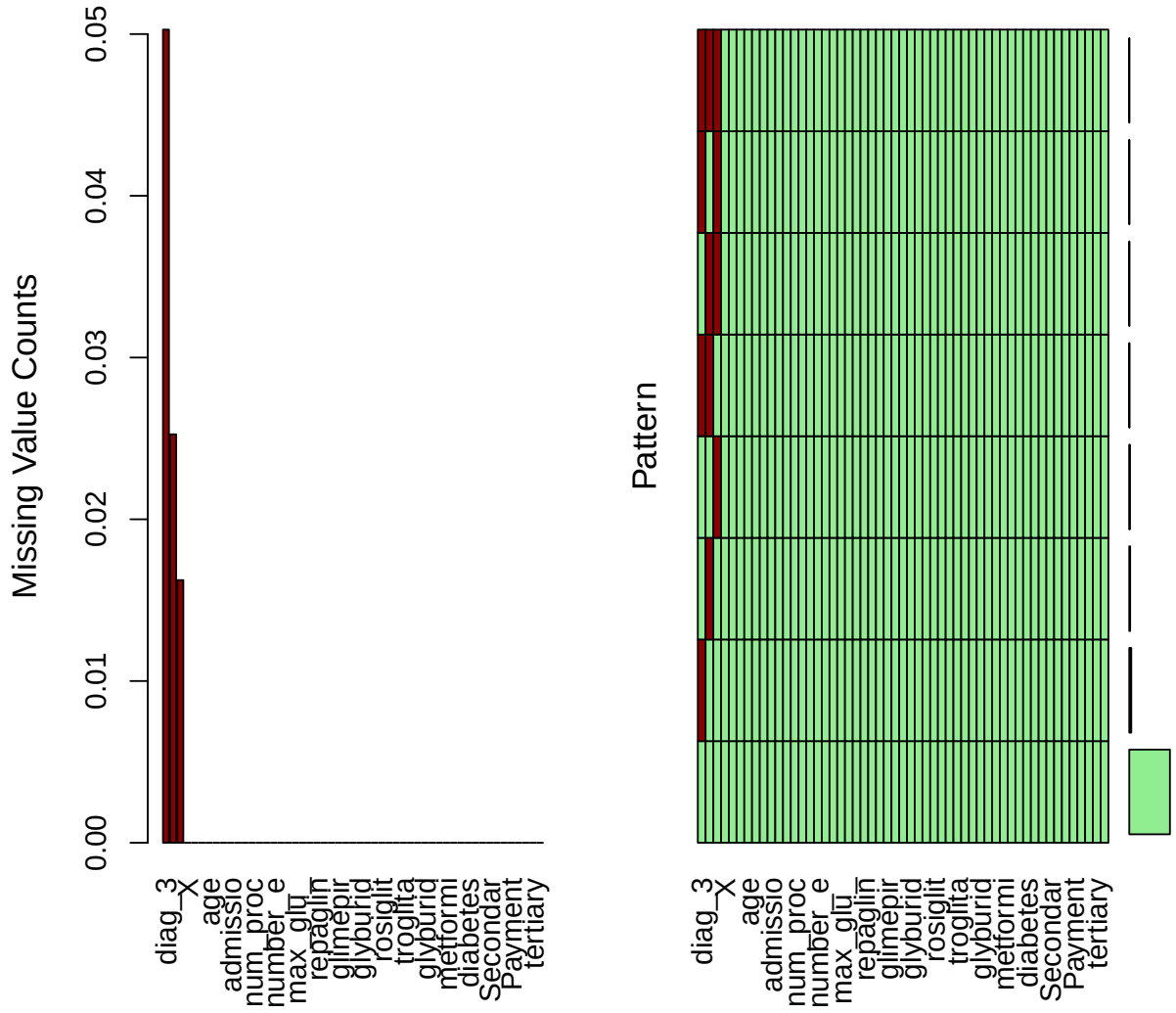
```
## Using Primary_Diag, Secondary_Diag_1, Secondary_Diag_2 as id variables

## Warning: Removed 9172 rows containing non-finite values (stat_bin).
```



3.2 Missing Values:

```
## Warning in plot.aggr(res, ...): not enough horizontal space to display
## frequencies
```




```
##
##  Variables sorted by number of missings:
##  Variable      Count
##    diag_3 0.05027514
##    diag_2 0.02525263
##    diag_1 0.01623812
##         X 0.00000000
##      race 0.00000000
##    gender 0.00000000
##       age 0.00000000
##  admissio 0.00000000
##  discharg 0.00000000
##  admissio 0.00000000
```

```
##  time_in_ 0.00000000
##  num_lab_ 0.00000000
##  num_proc 0.00000000
##  num_medi 0.00000000
##  number_o 0.00000000
##  number_e 0.00000000
##  number_i 0.00000000
##  number_d 0.00000000
##  max_glu_ 0.00000000
##  A1Cresul 0.00000000
##  metformi 0.00000000
##  repaglin 0.00000000
##  nateglin 0.00000000
##  chlorpro 0.00000000
##  glimepir 0.00000000
##  acetohex 0.00000000
##  glipizid 0.00000000
##  glyburid 0.00000000
##  tolbutam 0.00000000
##  pioglita 0.00000000
##  rosiglit 0.00000000
##  acarbose 0.00000000
##  miglitol 0.00000000
##  troglita 0.00000000
##  tolazami 0.00000000
##   insulin 0.00000000
##  glyburid 0.00000000
##  glipizid 0.00000000
##  glimepir 0.00000000
##  metformi 0.00000000
##  metformi 0.00000000
##    change 0.00000000
##  diabetes 0.00000000
##  readmitt 0.00000000
##  Primary_ 0.00000000
##  Secondar 0.00000000
##  Secondar 0.00000000
##       DRG 0.00000000
##   Payment 0.00000000
##  primaryd 0.00000000
##  secondar 0.00000000
##  tertiary 0.00000000
##  DRGClass 0.00000000
```

Table 1: Variables Missing Values

	Variable	Count	pct_missing
1	diag_3	5025	0.050
2	diag_2	2524	0.025
3	diag_1	1623	0.016

4 Drop Missing Values:

Weight is missing in over 98% records. Owing to the poor interpretability of missing values and little predictive generalizability to other patients, best thing is to just drop it.

Payer code and Medical Specialty of treating physician also have 40-50% missing values. We decided to drop these.

5 Addition of variables:

Service utilization: The data contains variables for number of inpatient (admissions), emergency room visits and outpatient visits for a given patient in the previous one year. These are (crude) measures of how much hospital/clinic services a person has used in the past year.

Number of medication changes: The dataset contains 23 features for 23 drugs (or combos) which indicate for each of these, whether a change in that medication was made or not during the current hospital stay of patient. Medication change for diabetics upon admission has been shown by previous research to be associated with lower readmission rates. We decided to count how many changes were made in total for each patient, and declared that a new feature. The reasoning here was to both simplify the model and possibly discover a relationship with number of changes regardless of which drug was changed.

6 Data Required for Modeling:

The new data set is composed of 99950 records and 27 variables.

7 Split Data:

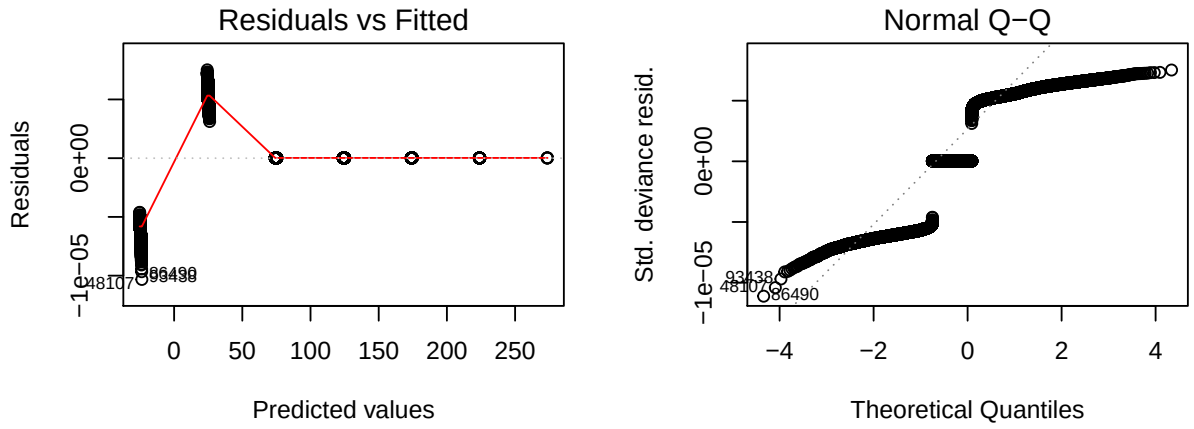
8 Model1: Binary Logistic Regression Model with no DRG

```
## Warning: glm.fit: algorithm did not converge

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

##
## Call:
```

```
## glm(formula = train$readmitted ~ ., family = "binomial", data = model1_no_drg)
##
## Deviance Residuals:
##        Min          1Q      Median          3Q         Max
## -1.032e-05   2.110e-08   2.110e-08   5.320e-06   7.520e-06
##
## Coefficients: (1 not defined because of singularities)
##                            Estimate Std. Error z value Pr(>|z|)
## (Intercept)              -2.466e+01  7.689e+03  -0.003    0.997
## age                       1.414e-02  1.412e+03   0.000    1.000
## admission_type_id        -2.974e-03  4.591e+01   0.000    1.000
## discharge_disposition_id -1.003e-03  6.045e+02   0.000    1.000
## admission_source_id      -4.751e-03  1.294e+02   0.000    1.000
## time_in_hospital          4.184e-03  1.763e+02   0.000    1.000
## num_lab_procedures       -1.818e-03  2.843e+02   0.000    1.000
## num_procedures            4.355e-04  4.004e+01   0.000    1.000
## num_medications          -2.269e-02  4.606e+02   0.000    1.000
## number_outpatient         9.050e-03  1.169e+02   0.000    1.000
## number_emergency          1.322e-02  5.819e+02   0.000    1.000
## number_inpatient          1.239e-02  7.584e+02   0.000    1.000
## number_diagnoses          4.913e-02  5.697e+02   0.000    1.000
## max_glu_serum             6.139e-03  3.992e+02   0.000    1.000
## A1Cresult                 6.717e-05  3.882e+01   0.000    1.000
## change                    8.496e-04  2.044e+01   0.000    1.000
## diabetesMed              -3.510e-01  1.833e+03   0.000    1.000
## DRG                       8.893e-02  2.236e+03   0.000    1.000
## Payment                   1.444e-04  8.176e+00   0.000    1.000
## primarydiagclass         -1.944e-05  8.850e-01   0.000    1.000
## secondarydiagclass_1     -2.667e-03  2.962e+02   0.000    1.000
## tertiarydiagclass_1       3.863e-03  2.484e+02   0.000    1.000
## DRGClassification        -7.802e-04  2.451e+02   0.000    1.000
## service_utilization      -3.252e-03  4.132e+02   0.000    1.000
## numchange                        NA         NA      NA       NA
## `NA`                      4.980e+01  1.623e+03   0.031    0.976
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 7.4925e+04  on 69964  degrees of freedom
## Residual deviance: 1.5209e-06  on 69940  degrees of freedom
## AIC: 50
##
## Number of Fisher Scoring iterations: 25
```



CIs using standard errors

```
##                                2.5 %      97.5 %
## (Intercept)              -15095.339529 15046.02198
## age                       -2768.171726  2768.20001
## admission_type_id            -89.985834    89.97989
## discharge_disposition_id   -1184.735998  1184.73399
## admission_source_id         -253.710819   253.70132
## time_in_hospital            -345.562616   345.57098
## num_lab_procedures          -557.176862   557.17323
## num_procedures               -78.483945    78.48482
## num_medications             -902.824960   902.77957
## number_outpatient           -229.018347   229.03645
## number_emergency           -1140.497824  1140.52427
## number_inpatient           -1486.363984  1486.38877
## number_diagnoses           -1116.445897  1116.54416
## max_glu_serum               -782.325999   782.33828
## A1Cresult                    -76.088421    76.08856
```

```
## change                     -40.055133    40.05683
## diabetesMed               -3593.660253  3592.95827
## DRG                       -4383.116398  4383.29425
## Payment                     -16.025327    16.02562
## primarydiagclass             -1.734569     1.73453
## secondarydiagclass_1       -580.593967   580.58863
## tertiarydiagclass_1        -486.834646   486.84237
## DRGClassification          -480.337300   480.33574
## service_utilization        -809.773578   809.76707
## numchange                           NA          NA
## `NA`                      -3131.622937  3231.22008
```

coefficient is negative for gender, admission_type_id, admission_source_id, num_lab_procedures,num_medications diabetesMed, primarydiagclass,secondarydiagclass_1,DRClassification and service_utilization.

standard error is high for gender, discharge_disposition_id, num_medications,number_emergency,Payment,numbe

We can test for an overall effect of rank using the wald.test function of the aod library. The order in which the coefficients are given in the table of coefficients is the same as the order of the terms in the model. This is important because the wald.test function refers to the coefficients by their order in the model. We use the wald.test function. b supplies the coefficients, while Sigma supplies the variance covariance matrix of the error terms, finally Terms tells R which terms in the model are to be tested, in this case, terms 1 to 13 and 24 to 26.

9 Odds Ratio

```
##              (Intercept)                       age        admission_type_id
##             1.953576e-11              1.014243e+00             9.970305e-01
## discharge_disposition_id      admission_source_id         time_in_hospital
##             9.989976e-01              9.952599e-01             1.004192e+00
##       num_lab_procedures           num_procedures          num_medications
##             9.981834e-01              1.000436e+00             9.775621e-01
##        number_outpatient         number_emergency         number_inpatient
##             1.009091e+00              1.013311e+00             1.012471e+00
##         number_diagnoses            max_glu_serum                A1Cresult
##             1.050358e+00              1.006158e+00             1.000067e+00
##                  change               diabetesMed                      DRG
##             1.000850e+00              7.039902e-01             1.093002e+00
##                 Payment         primarydiagclass     secondarydiagclass_1
##             1.000144e+00              9.999806e-01             9.973363e-01
##      tertiarydiagclass_1        DRGClassification      service_utilization
##             1.003870e+00              9.992201e-01             9.967529e-01
##                numchange                       `NA`
##                      NA              4.238820e+21
```

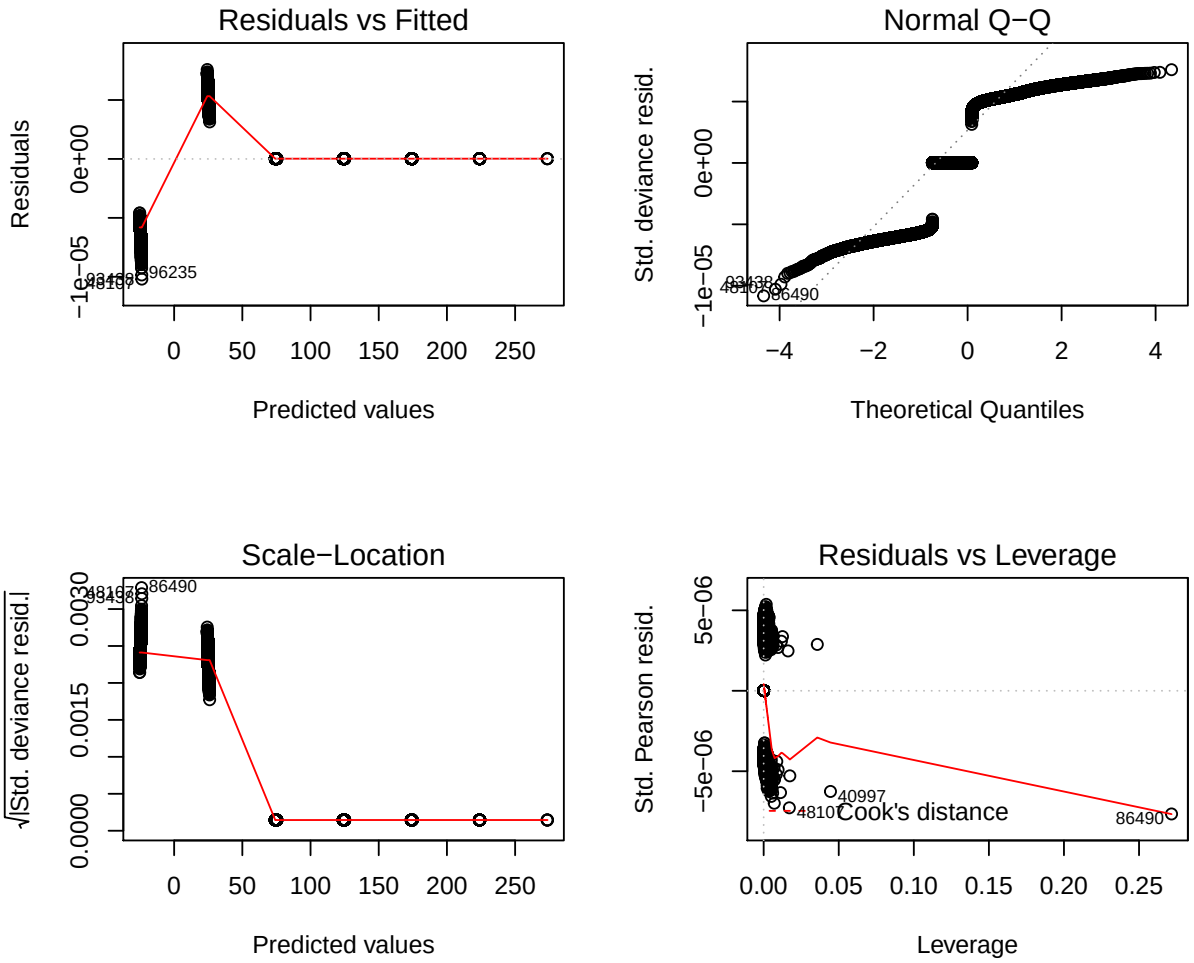
10 Model1: Binary Logistic Regression Model with DRG

```
## Warning: glm.fit: algorithm did not converge

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

##
## Call:
## glm(formula = train$readmitted ~ ., family = "binomial", data = model1_drg)
##
## Deviance Residuals:
##        Min          1Q      Median          3Q         Max
## -1.020e-05   2.110e-08   2.110e-08   5.322e-06   7.588e-06
##
## Coefficients: (1 not defined because of singularities)
##                            Estimate Std. Error z value Pr(>|z|)
## (Intercept)              -2.458e+01  7.987e+03  -0.003    0.998
## gender                   -2.810e-02  7.528e+02   0.000    1.000
## age                       1.687e-02  1.414e+03   0.000    1.000
## admission_type_id        -2.793e-03  4.616e+01   0.000    1.000
## discharge_disposition_id  2.980e-04  6.054e+02   0.000    1.000
## admission_source_id      -4.806e-03  1.293e+02   0.000    1.000
## time_in_hospital          4.217e-03  1.763e+02   0.000    1.000
## num_lab_procedures       -2.318e-03  2.845e+02   0.000    1.000
## num_procedures            4.214e-04  4.004e+01   0.000    1.000
## num_medications          -2.259e-02  4.606e+02   0.000    1.000
## number_outpatient         9.150e-03  1.169e+02   0.000    1.000
## number_emergency          1.413e-02  5.832e+02   0.000    1.000
## number_inpatient          1.191e-02  7.628e+02   0.000    1.000
## number_diagnoses          4.923e-02  5.693e+02   0.000    1.000
## max_glu_serum             7.013e-03  3.998e+02   0.000    1.000
## A1Cresult                 1.059e-04  3.883e+01   0.000    1.000
## change                    8.563e-04  2.044e+01   0.000    1.000
## diabetesMed              -3.519e-01  1.833e+03   0.000    1.000
## DRG                       8.903e-02  2.236e+03   0.000    1.000
## Payment                   1.444e-04  8.179e+00   0.000    1.000
## primarydiagclass         -1.931e-05  8.853e-01   0.000    1.000
## secondarydiagclass_1     -2.657e-03  2.962e+02   0.000    1.000
## tertiarydiagclass_1       3.834e-03  2.484e+02   0.000    1.000
## DRGClassification        -7.383e-04  2.451e+02   0.000    1.000
## service_utilization      -3.161e-03  4.133e+02   0.000    1.000
## numchange                        NA         NA      NA       NA
## `NA`                      4.980e+01  1.623e+03   0.031    0.976
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 7.4925e+04  on 69964  degrees of freedom
## Residual deviance: 1.5209e-06  on 69939  degrees of freedom
```

```
## AIC: 52
##
## Number of Fisher Scoring iterations: 25
```



AIC is slightly more with DRG than the model with no DRG.