Real to Synthetic Dataset Comparison (Cleveland Dataset)

Al Sabay

6/27/2018

Cleveland Dataset http://archive.ics.uci.edu/ml/datasets/Heart+Disease

Generalized Linear Model fit for Synthesized Heart Disease Data

Original data from Cleveland Dataset of 297 observations was used to synthesize a dataset of size 50,000 observations. The sysntesized dataset is used in Logistics Regression fit below.

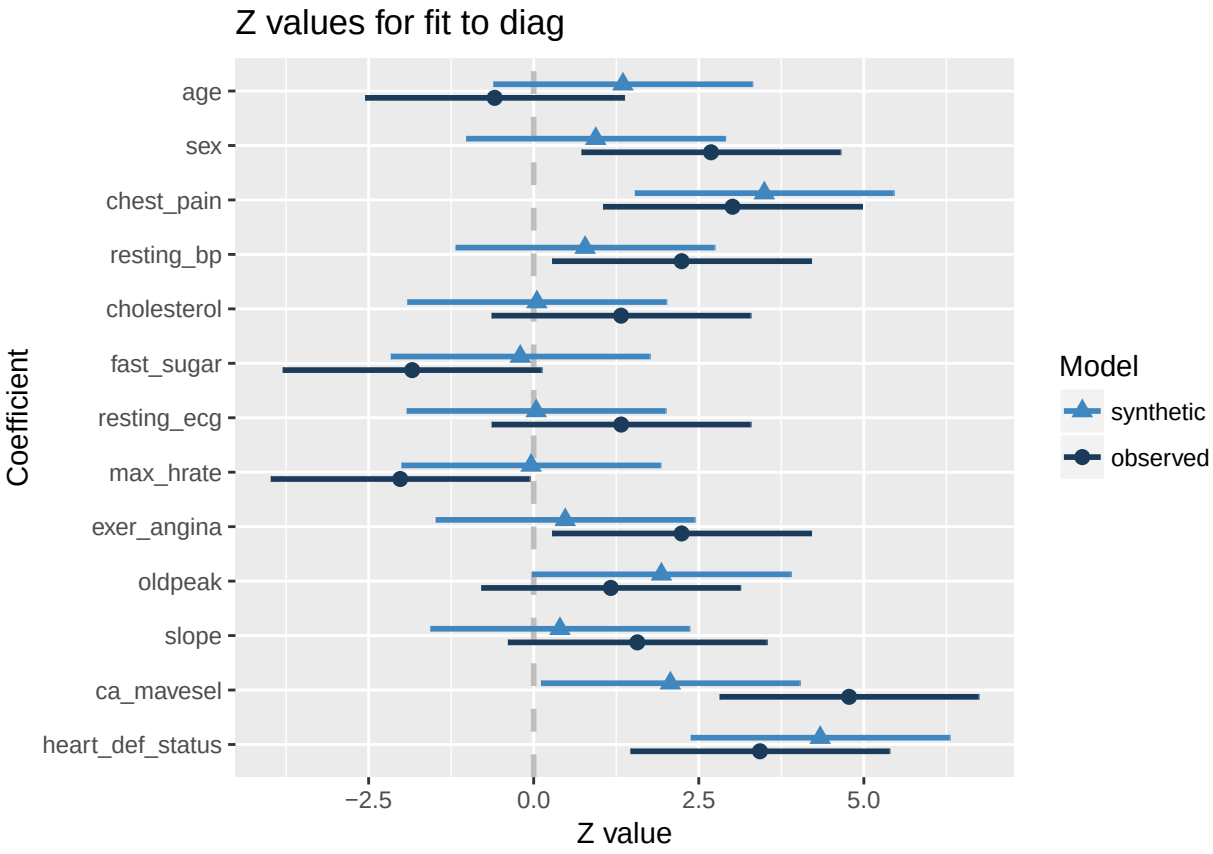
```
summary(glm.synds(diag~age+sex+chest_pain+resting_bp+cholesterol+fast_sugar+
              resting_ecg+max_hrate+exer_angina+oldpeak+slope+ca_mavesel+heart_def_status,
            data = syn_cleveland, family = "binomial"))
```

```
## Warning: Note that all these results depend on the synthesis model being correct.
##
## Fit to synthetic data set with a single synthesis.
## Inference to coefficients and standard errors that
## would be obtained from the observed data.
##
## Call:
## glm.synds(formula = diag ~ age + sex + chest_pain + resting_bp +
##     cholesterol + fast_sugar + resting_ecg + max_hrate + exer_angina +
##     oldpeak + slope + ca_mavesel + heart_def_status, family = "binomial",
##     data = syn_cleveland)
##
## Combined estimates:
##                   xpct(Beta) xpct(se.Beta) xpct(z) Pr(>|xpct(z)|)
## (Intercept)      -8.60505106    2.54297747 -3.3838      0.0007148 ***
## age               0.03239783    0.02022088  1.6022      0.1091122
## sex               0.45901190    0.37109734  1.2369      0.2161226
## chest_pain        0.66774175    0.18430833  3.6230      0.0002913 ***
## resting_bp        0.00836080    0.00944918  0.8848      0.3762551
## cholesterol       0.00018228    0.00313870  0.0581      0.9536899
## fast_sugar       -0.11281328    0.44303256 -0.2546      0.7990021
## resting_ecg       0.00703449    0.16073649  0.0438      0.9650924
## max_hrate        -0.00040708    0.00848255 -0.0480      0.9617244
## exer_angina       0.19737927    0.37591947  0.5251      0.5995433
## oldpeak           0.40971827    0.18162707  2.2558      0.0240818 *
## slope             0.14450297    0.31302833  0.4616      0.6443474
## ca_mavesel        0.54937908    0.18705876  2.9369      0.0033148 **
## heart_def_status  0.43552755    0.08514481  5.1151      3.135e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Z Values for fit to HD Diagnosis (diag)

```
##
## Call used to fit models to the data:
## glm.synds(formula = diag ~ age + sex + chest_pain + resting_bp +
##     cholesterol + fast_sugar + resting_ecg + max_hrate + exer_angina +
##     oldpeak + slope + ca_mavesel + heart_def_status, family = "binomial",
##     data = syn_cleveland)
##
## Estimates for the observed data set:
##                          Beta    se(Beta)          Z
## (Intercept)      -7.372041866 2.879476068 -2.5602025
## age              -0.014163656 0.023969703 -0.5908983
## sex               1.312073342 0.488474286  2.6860643
## chest_pain        0.575898404 0.191197311  3.0120633
## resting_bp        0.024044039 0.010730305  2.2407600
## cholesterol       0.004995224 0.003773772  1.3236687
## fast_sugar       -1.021917674 0.555329549 -1.8402004
## resting_ecg       0.245153156 0.185004661  1.3251188
## max_hrate        -0.020665356 0.010224968 -2.0210680
## exer_angina       0.926104214 0.413342659  2.2405242
## oldpeak           0.247386221 0.211831827  1.1678425
## slope             0.570008825 0.363084786  1.5699056
## ca_mavesel        1.267718507 0.265384049  4.7769205
## heart_def_status  0.343936191 0.100360969  3.4269915
##
## Combined estimates for the synthetised data set(s):
##                     xpct(Beta) xpct(se.Beta)     xpct(z)
## (Intercept)      -8.6050510555   2.879476068 -2.98840860
## age               0.0323978257   0.023969703  1.35161567
## sex               0.4590119021   0.488474286  0.93968488
## chest_pain        0.6677417499   0.191197311  3.49242229
## resting_bp        0.0083607992   0.010730305  0.77917627
## cholesterol       0.0001822757   0.003773772  0.04830068
## fast_sugar       -0.1128132772   0.555329549 -0.20314654
## resting_ecg       0.0070344897   0.185004661  0.03802331
## max_hrate        -0.0004070759   0.010224968 -0.03981195
## exer_angina       0.1973792687   0.413342659  0.47751972
## oldpeak           0.4097182733   0.211831827  1.93416768
## slope             0.1445029691   0.363084786  0.39798685
## ca_mavesel        0.5493790816   0.265384049  2.07012849
## heart_def_status  0.4355275511   0.100360969  4.33961087
##
## Differences between results based on synthetic and observed data:
##                  Std. coef diff p value CI overlap
## (Intercept)          -0.4282061       0  0.8907617
## age                   1.9425140       0  0.5044516
## sex                  -1.7463794       0  0.5544869
## chest_pain            0.4803590       0  0.8774572
## resting_bp           -1.4615838       0  0.6271401
## cholesterol          -1.2753681       0  0.6746450
## fast_sugar            1.6370539       0  0.5823765
## resting_ecg          -1.2870955       0  0.6716533
## max_hrate             1.9812561       0  0.4945682
```

```
## exer_angina          -1.7630044       0  0.5502457
## oldpeak               0.7663251       0  0.8045053
## slope                -1.1719187       0  0.7010357
## ca_mavesel           -2.7067920       0  0.3094791
## heart_def_status      0.9126193       0  0.7671847
##
## Measures for one synthesis and 14 coefficients
## Mean confidence interval overlap:  0.6435708
## Mean absolute std. coef diff:  1.397177
## Lack-of-fit: 3421.41; p-value 0 for test that synthesis model is compatible
## with a chi-squared test with 14 degrees of freedom
##
## Confidence interval plot:
```

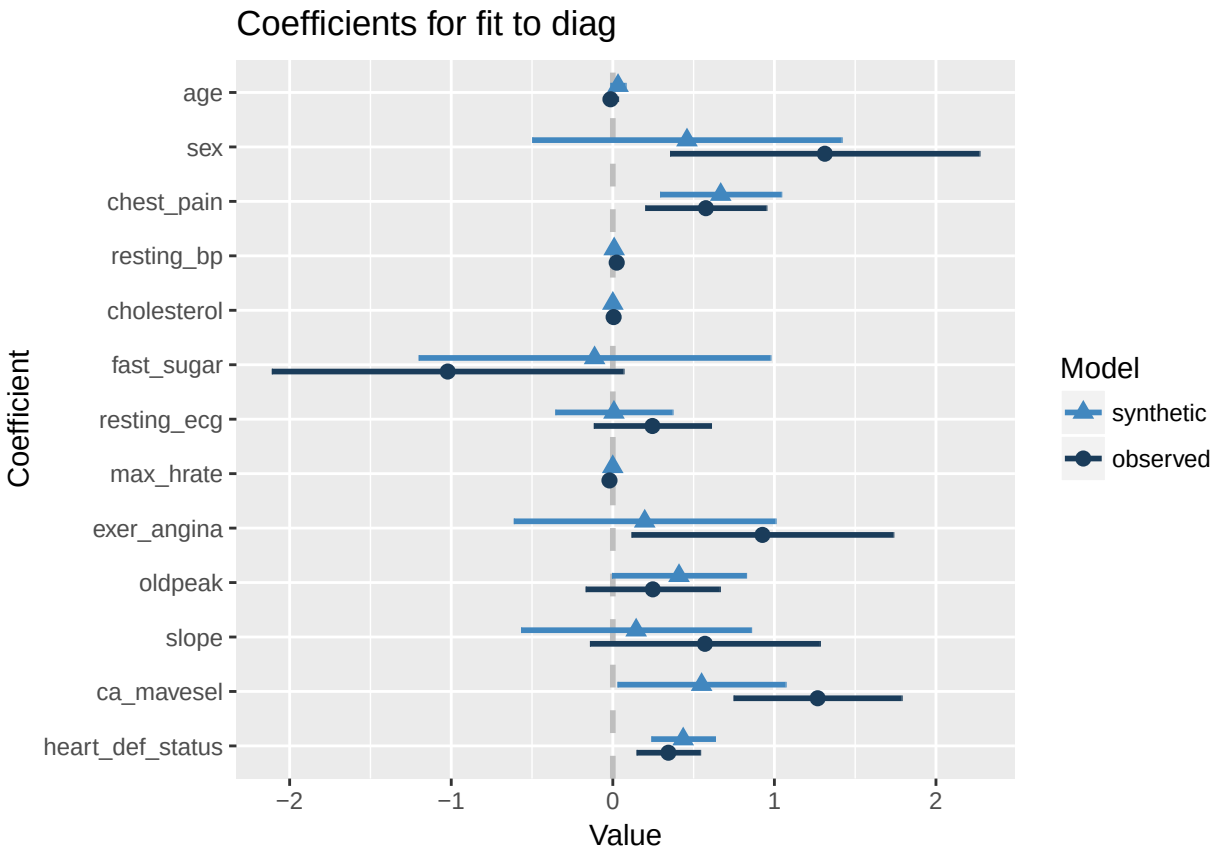


Coefficient Values for fit to HD Diagnosis (diag)

```
##
## Call used to fit models to the data:
## glm.synds(formula = diag ~ age + sex + chest_pain + resting_bp +
##     cholesterol + fast_sugar + resting_ecg + max_hrate + exer_angina +
##     oldpeak + slope + ca_mavesel + heart_def_status, family = "binomial",
##     data = syn_cleveland)
##
## Estimates for the observed data set:
##                          Beta    se(Beta)          Z
## (Intercept)      -7.372041866 2.879476068 -2.5602025
```

```
## age              -0.014163656 0.023969703 -0.5908983
## sex               1.312073342 0.488474286  2.6860643
## chest_pain        0.575898404 0.191197311  3.0120633
## resting_bp        0.024044039 0.010730305  2.2407600
## cholesterol       0.004995224 0.003773772  1.3236687
## fast_sugar       -1.021917674 0.555329549 -1.8402004
## resting_ecg       0.245153156 0.185004661  1.3251188
## max_hrate        -0.020665356 0.010224968 -2.0210680
## exer_angina       0.926104214 0.413342659  2.2405242
## oldpeak           0.247386221 0.211831827  1.1678425
## slope             0.570008825 0.363084786  1.5699056
## ca_mavesel        1.267718507 0.265384049  4.7769205
## heart_def_status  0.343936191 0.100360969  3.4269915
##
## Combined estimates for the synthetised data set(s):
##                     xpct(Beta) xpct(se.Beta)     xpct(z)
## (Intercept)      -8.6050510555   2.879476068 -2.98840860
## age               0.0323978257   0.023969703  1.35161567
## sex               0.4590119021   0.488474286  0.93968488
## chest_pain        0.6677417499   0.191197311  3.49242229
## resting_bp        0.0083607992   0.010730305  0.77917627
## cholesterol       0.0001822757   0.003773772  0.04830068
## fast_sugar       -0.1128132772   0.555329549 -0.20314654
## resting_ecg       0.0070344897   0.185004661  0.03802331
## max_hrate        -0.0004070759   0.010224968 -0.03981195
## exer_angina       0.1973792687   0.413342659  0.47751972
## oldpeak           0.4097182733   0.211831827  1.93416768
## slope             0.1445029691   0.363084786  0.39798685
## ca_mavesel        0.5493790816   0.265384049  2.07012849
## heart_def_status  0.4355275511   0.100360969  4.33961087
##
## Differences between results based on synthetic and observed data:
##                   Std. coef diff p value CI overlap
## (Intercept)          -0.4282061       0  0.8907617
## age                   1.9425140       0  0.5044516
## sex                  -1.7463794       0  0.5544869
## chest_pain            0.4803590       0  0.8774572
## resting_bp           -1.4615838       0  0.6271401
## cholesterol          -1.2753681       0  0.6746450
## fast_sugar            1.6370539       0  0.5823765
## resting_ecg          -1.2870955       0  0.6716533
## max_hrate             1.9812561       0  0.4945682
## exer_angina          -1.7630044       0  0.5502457
## oldpeak               0.7663251       0  0.8045053
## slope                -1.1719187       0  0.7010357
## ca_mavesel           -2.7067920       0  0.3094791
## heart_def_status      0.9126193       0  0.7671847
##
## Measures for one synthesis and 14 coefficients
## Mean confidence interval overlap:  0.6435708
## Mean absolute std. coef diff:  1.397177
## Lack-of-fit: 3421.41; p-value 0 for test that synthesis model is compatible
## with a chi-squared test with 14 degrees of freedom
##
```

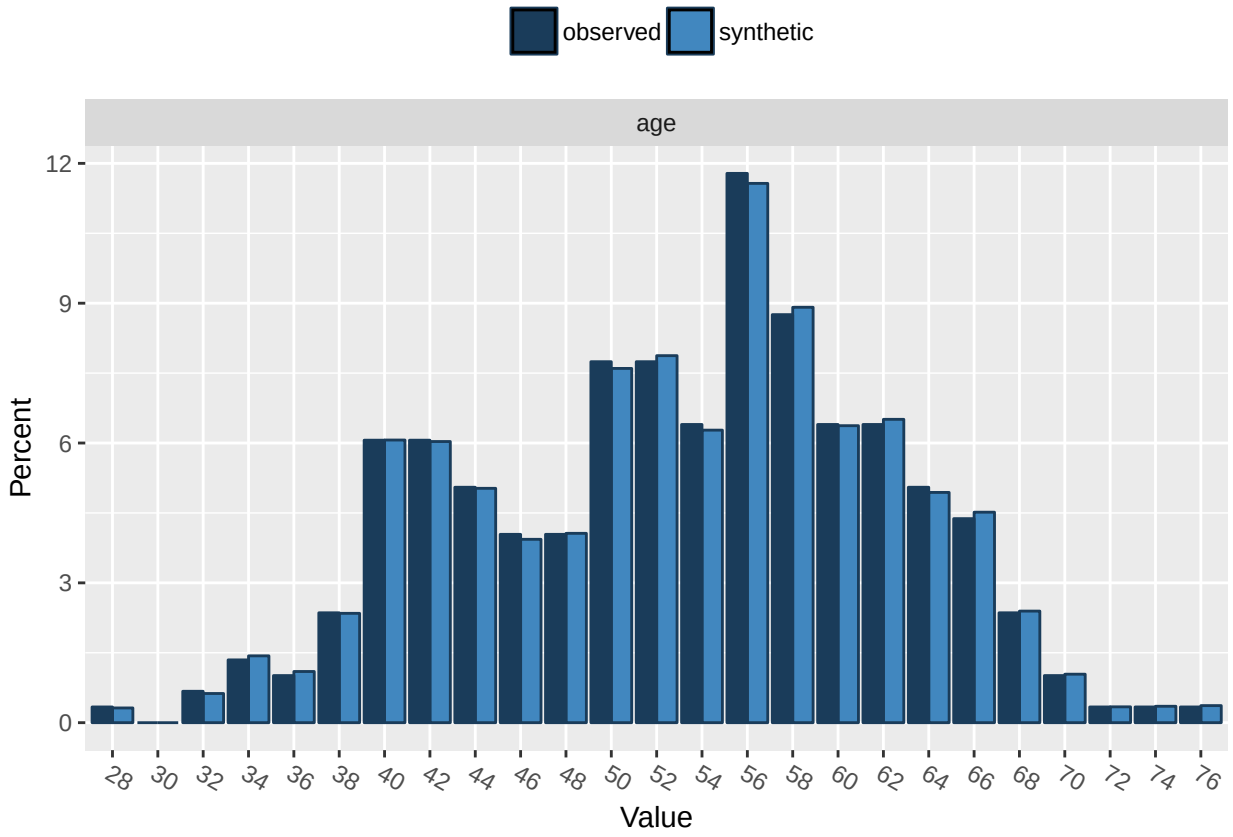
```
## Confidence interval plot:
```



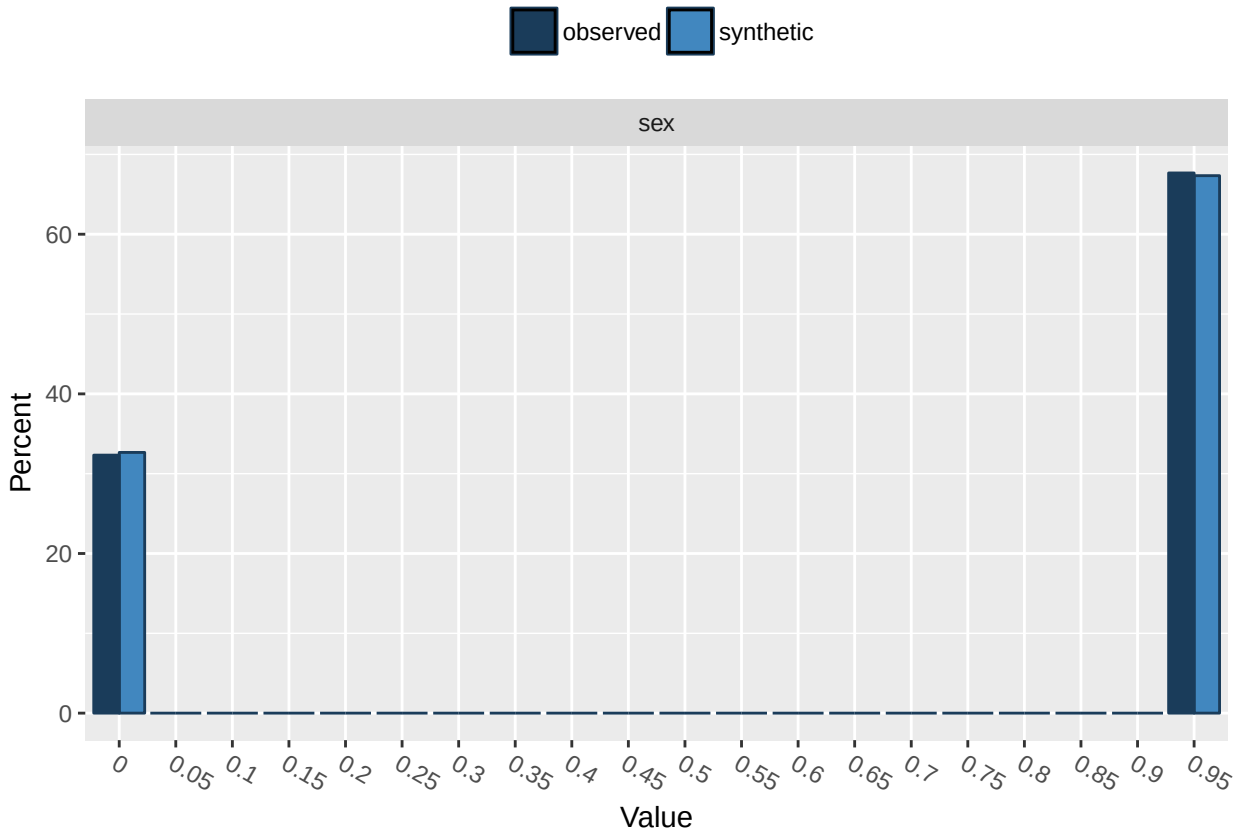
Feature comparisons of Real vs. Synthetic Data

Cleveland Dataset http://archive.ics.uci.edu/ml/datasets/Heart+Disease

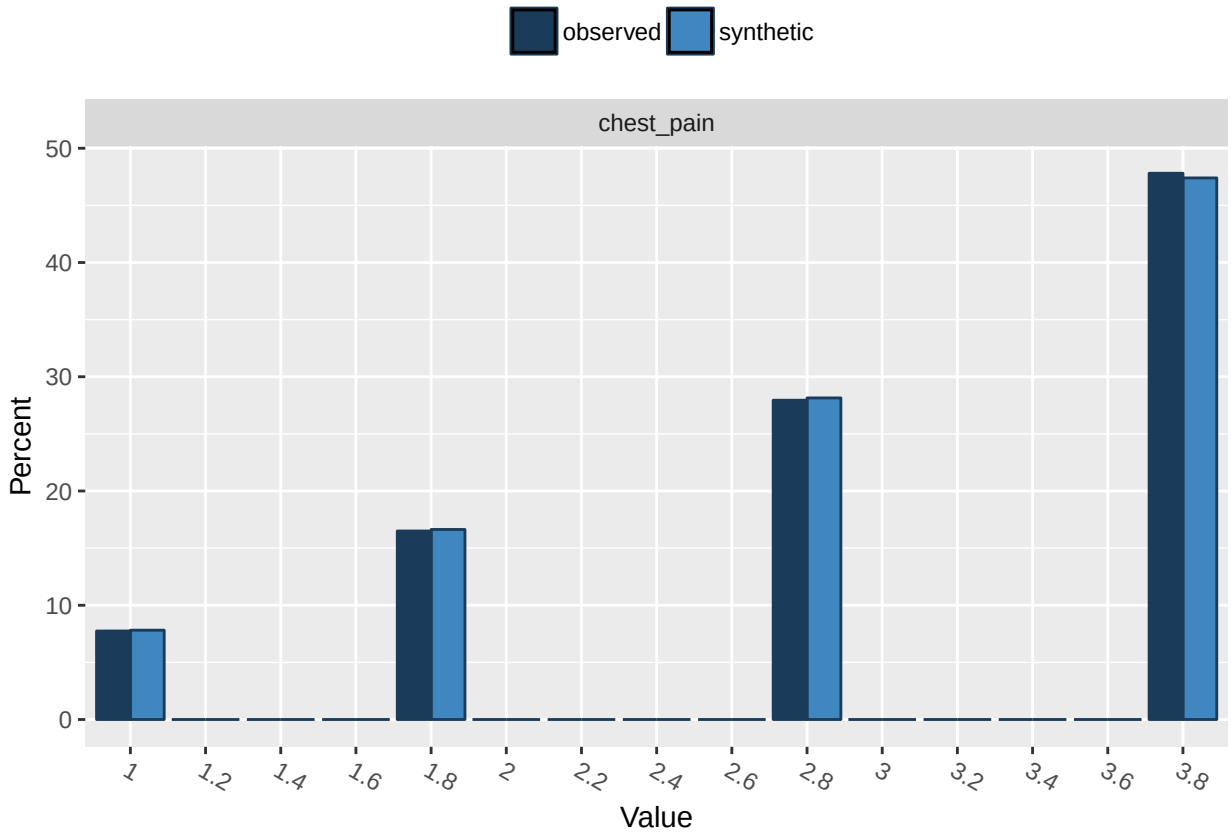
```
##
## Comparing percentages observed with synthetic
##
## $age
##                    28 30        32       34       36       38       40
## observed  0.3367003  0 0.6734007 1.346801 1.010101 2.356902 6.060606
## synthetic 0.3160000  0 0.6260000 1.434000 1.098000 2.346000 6.064000
##                   42       44       46       48       50       52       54
## observed  6.060606 5.050505 4.040404 4.040404 7.744108 7.744108 6.397306
## synthetic 6.032000 5.028000 3.934000 4.062000 7.600000 7.874000 6.276000
##                   56       58       60       62       64       66       68
## observed  11.78451 8.754209 6.397306 6.397306 5.050505 4.377104 2.356902
## synthetic 11.57000 8.912000 6.372000 6.508000 4.940000 4.516000 2.394000
##                   70        72        74        76
## observed  1.010101 0.3367003 0.3367003 0.3367003
## synthetic 1.040000 0.3400000 0.3520000 0.3660000
```



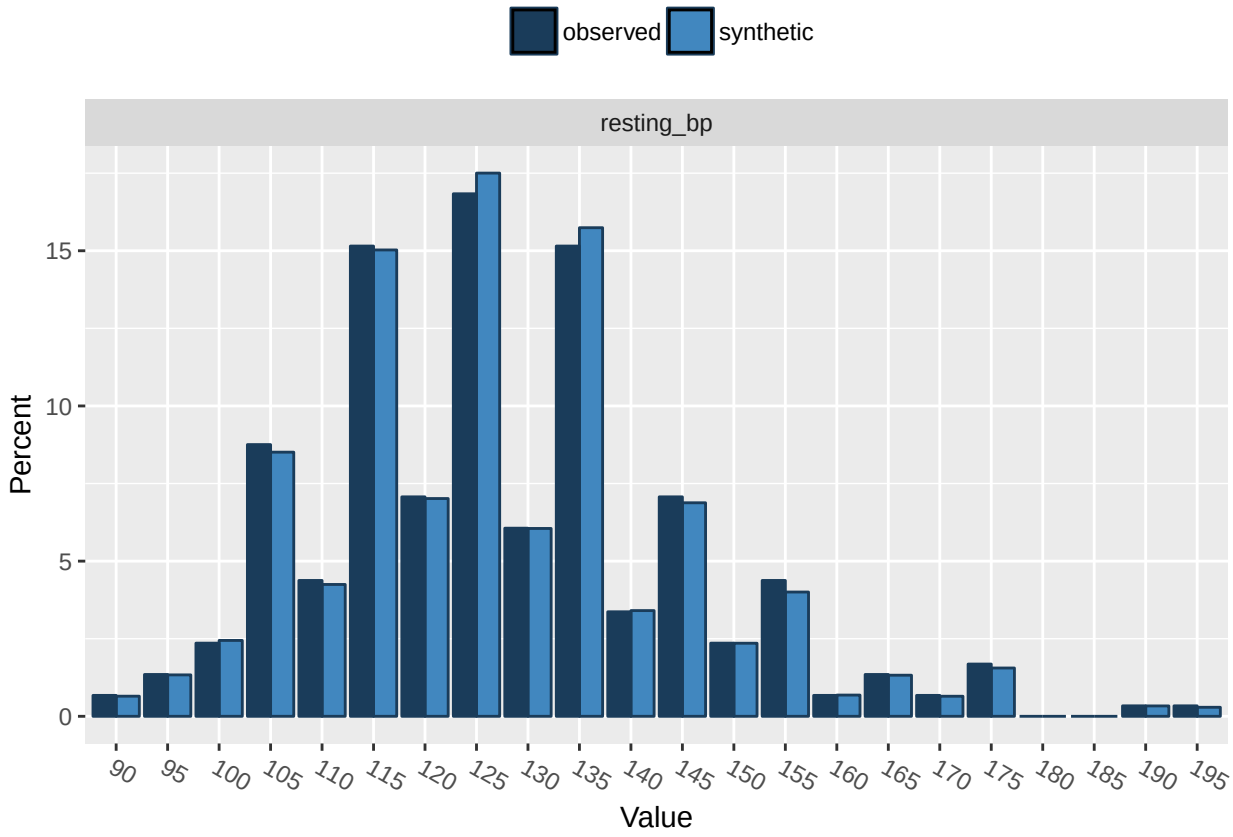

```
##
## Comparing percentages observed with synthetic
##
## $sex
##                  0 0.05 0.1 0.15 0.2 0.25 0.3 0.35 0.4 0.45 0.5 0.55 0.6
## observed  32.32323    0   0    0   0    0   0    0   0    0   0    0   0
## synthetic 32.66200    0   0    0   0    0   0    0   0    0   0    0   0
##           0.65 0.7 0.75 0.8 0.85 0.9     0.95
## observed     0   0    0   0    0   0 67.67677
## synthetic    0   0    0   0    0   0 67.33800
```



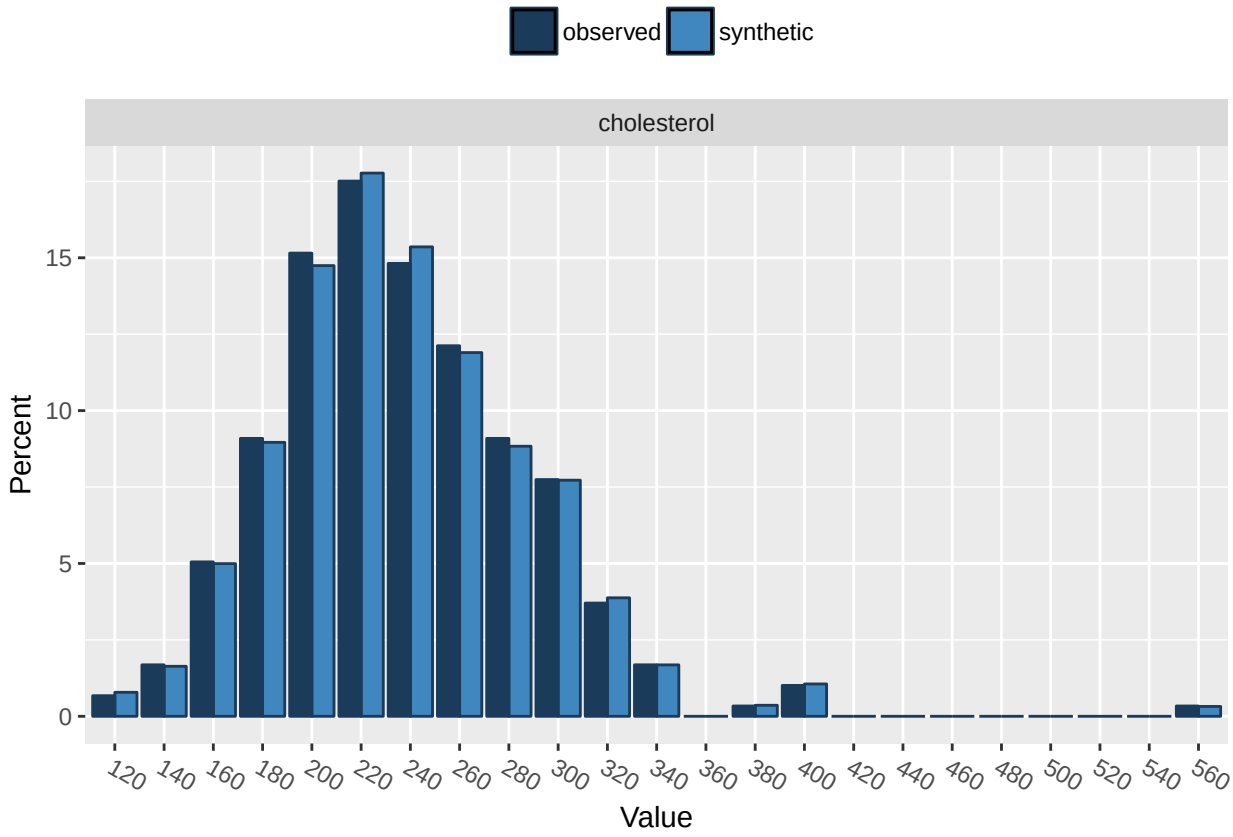

```
##
## Comparing percentages observed with synthetic
##
## $chest_pain
##                   1 1.2 1.4 1.6      1.8 2 2.2 2.4 2.6      2.8 3 3.2 3.4
## observed  7.744108   0   0   0 16.49832 0   0   0   0 27.94613 0   0   0
## synthetic 7.818000   0   0   0 16.63200 0   0   0   0 28.14600 0   0   0
##           3.6      3.8
## observed    0 47.81145
## synthetic   0 47.40400
```



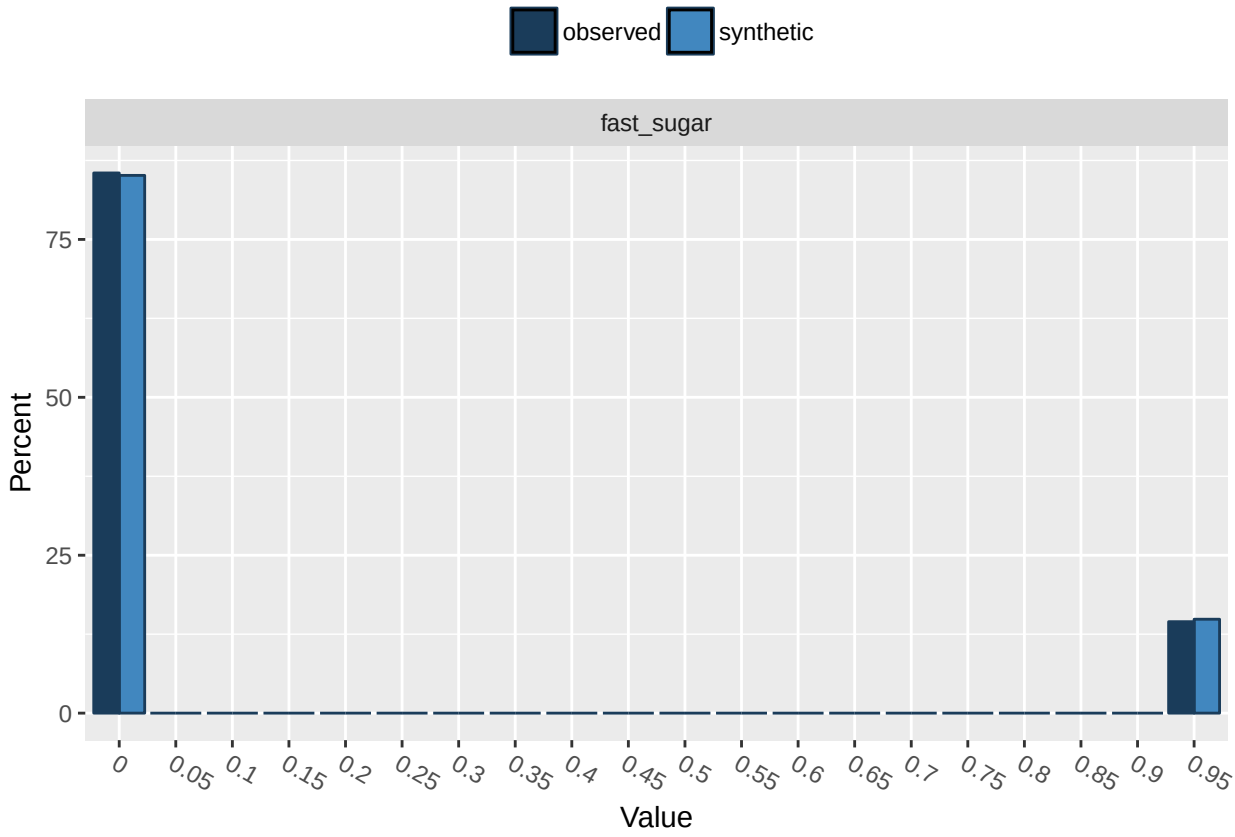

```
##
## Comparing percentages observed with synthetic
##
## $resting_bp
##                   90       95      100      105      110      115      120
## observed  0.6734007 1.346801 2.356902 8.754209 4.377104 15.15152 7.070707
## synthetic 0.6480000 1.334000 2.444000 8.510000 4.248000 15.02600 7.016000
##                  125      130      135      140      145      150      155
## observed  16.83502 6.060606 15.15152 3.367003 7.070707 2.356902 4.377104
## synthetic 17.50200 6.052000 15.74200 3.406000 6.880000 2.354000 4.004000
##                  160       165       170      175 180 185       190
## observed  0.6734007 1.346801 0.6734007 1.683502   0   0 0.3367003
## synthetic 0.6840000 1.324000 0.6460000 1.556000   0   0 0.3340000
##                  195
## observed  0.3367003
## synthetic 0.2900000
```



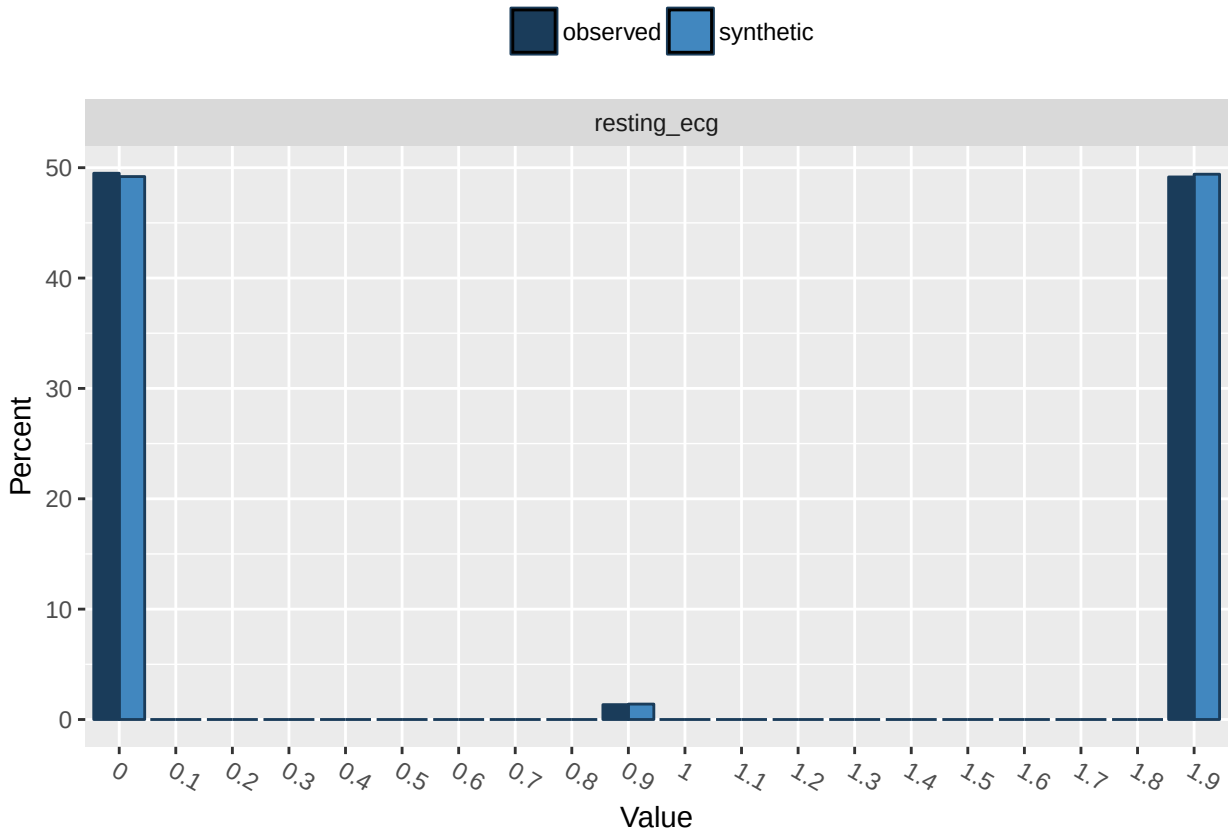

```
##
## Comparing percentages observed with synthetic
##
## $cholesterol
##                  120      140      160      180      200      220      240
## observed  0.6734007 1.683502 5.050505 9.090909 15.15152 17.50842 14.81481
## synthetic 0.7840000 1.636000 4.994000 8.962000 14.74400 17.77000 15.35600
##                  260      280      300      320      340 360       380
## observed  12.12121 9.090909 7.744108 3.703704 1.683502   0 0.3367003
## synthetic 11.89800 8.834000 7.724000 3.876000 1.682000   0 0.3600000
##                  400 420 440 460 480 500 520 540       560
## observed  1.010101   0   0   0   0   0   0   0 0.3367003
## synthetic 1.058000   0   0   0   0   0   0   0 0.3220000
```



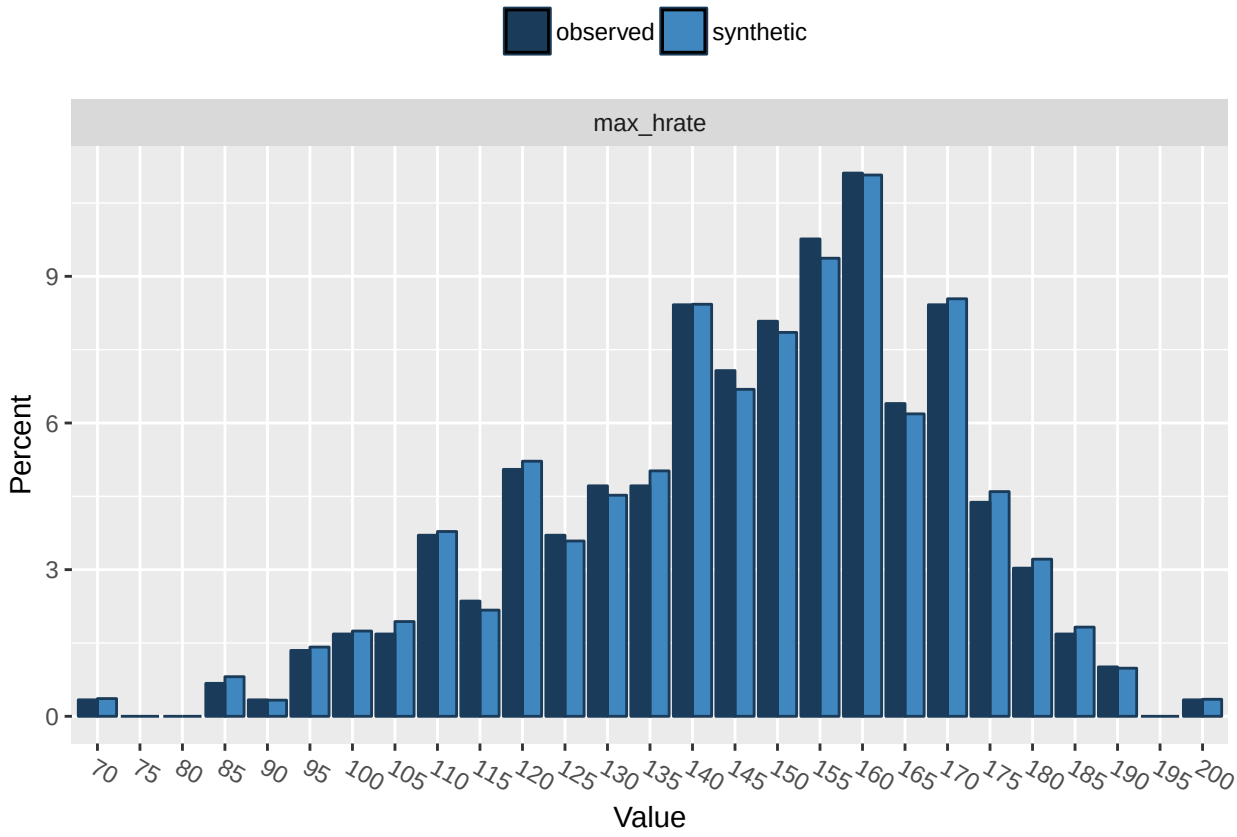

```
##
## Comparing percentages observed with synthetic
##
## $fast_sugar
##                  0 0.05 0.1 0.15 0.2 0.25 0.3 0.35 0.4 0.45 0.5 0.55 0.6
## observed  85.52189    0   0    0   0    0   0    0   0    0   0    0   0
## synthetic 85.13600    0   0    0   0    0   0    0   0    0   0    0   0
##           0.65 0.7 0.75 0.8 0.85 0.9     0.95
## observed     0   0    0   0    0   0 14.47811
## synthetic    0   0    0   0    0   0 14.86400
```



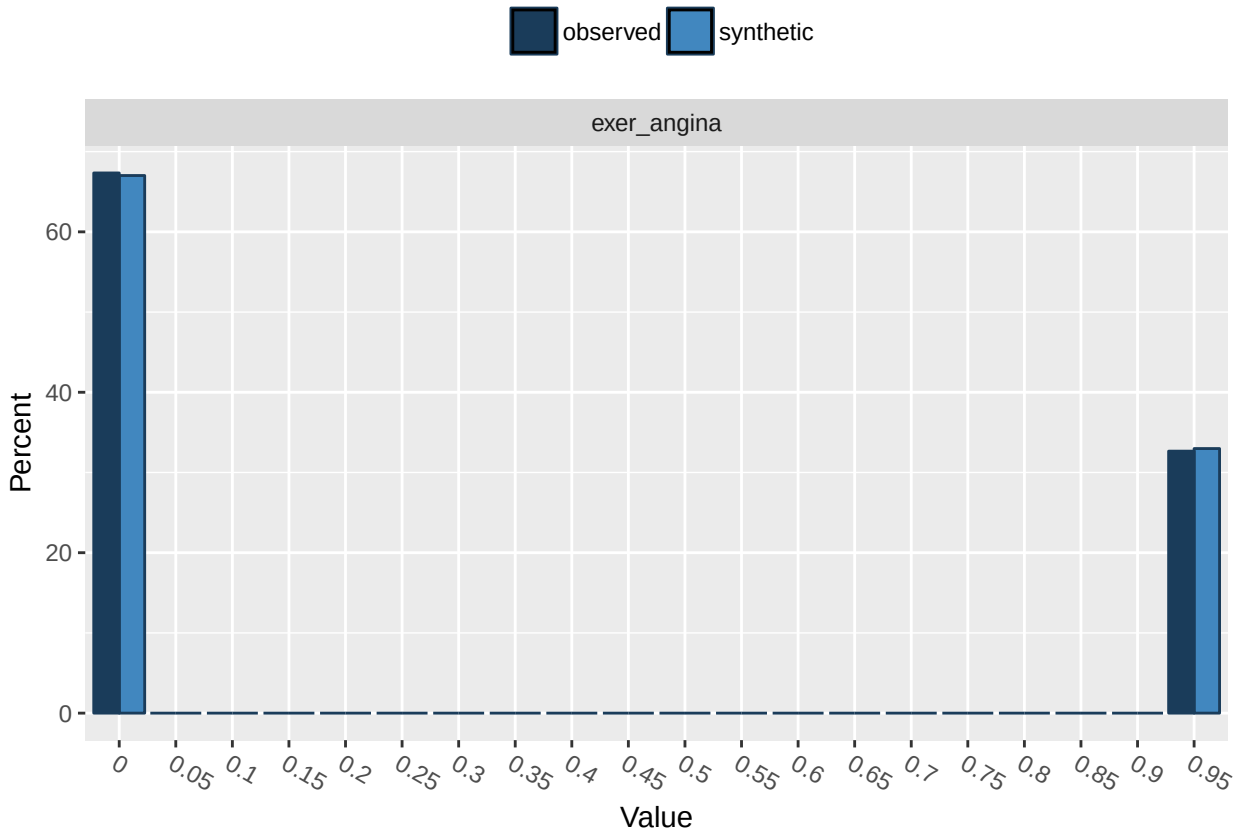

```
##
## Comparing percentages observed with synthetic
##
## $resting_ecg
##                  0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8       0.9 1 1.1 1.2 1.3
## observed  49.49495   0   0   0   0   0   0   0   0 1.346801 0   0   0   0
## synthetic 49.19600   0   0   0   0   0   0   0   0 1.396000 0   0   0   0
##           1.4 1.5 1.6 1.7 1.8      1.9
## observed    0   0   0   0   0 49.15825
## synthetic   0   0   0   0   0 49.40800
```



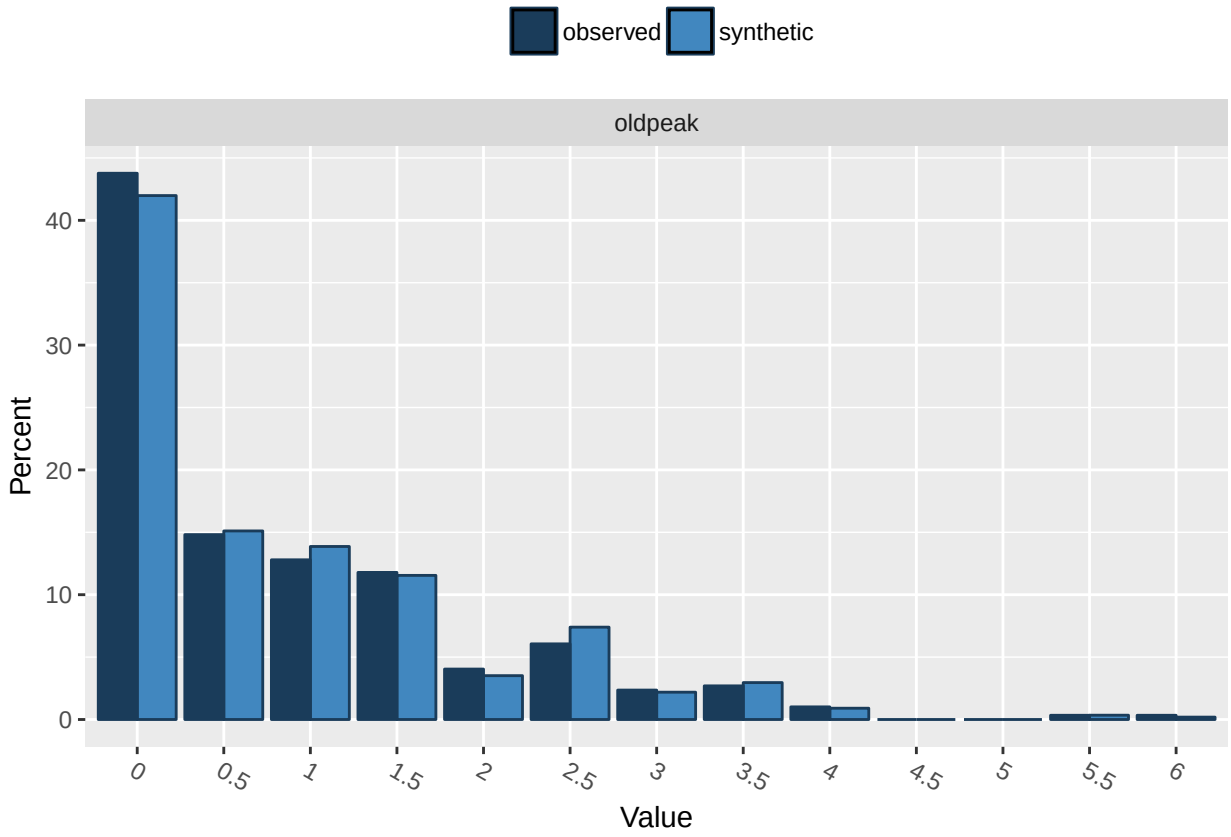

```
##
## Comparing percentages observed with synthetic
##
## $max_hrate
##                   70 75 80        85        90       95       100       105
## observed  0.3367003  0  0 0.6734007 0.3367003 1.346801 1.683502 1.683502
## synthetic 0.3620000  0  0 0.8100000 0.3300000 1.416000 1.744000 1.938000
##                110      115      120      125      130      135      140
## observed  3.703704 2.356902 5.050505 3.703704 4.713805 4.713805 8.417508
## synthetic 3.780000 2.172000 5.218000 3.586000 4.522000 5.020000 8.428000
##                145      150     155      160      165      170      175
## observed  7.070707 8.080808 9.76431 11.11111 6.397306 8.417508 4.377104
## synthetic 6.688000 7.852000 9.37000 11.07200 6.186000 8.540000 4.596000
##                180      185      190 195       200
## observed  3.030303 1.683502 1.010101   0 0.3367003
## synthetic 3.214000 1.824000 0.984000   0 0.3480000
```



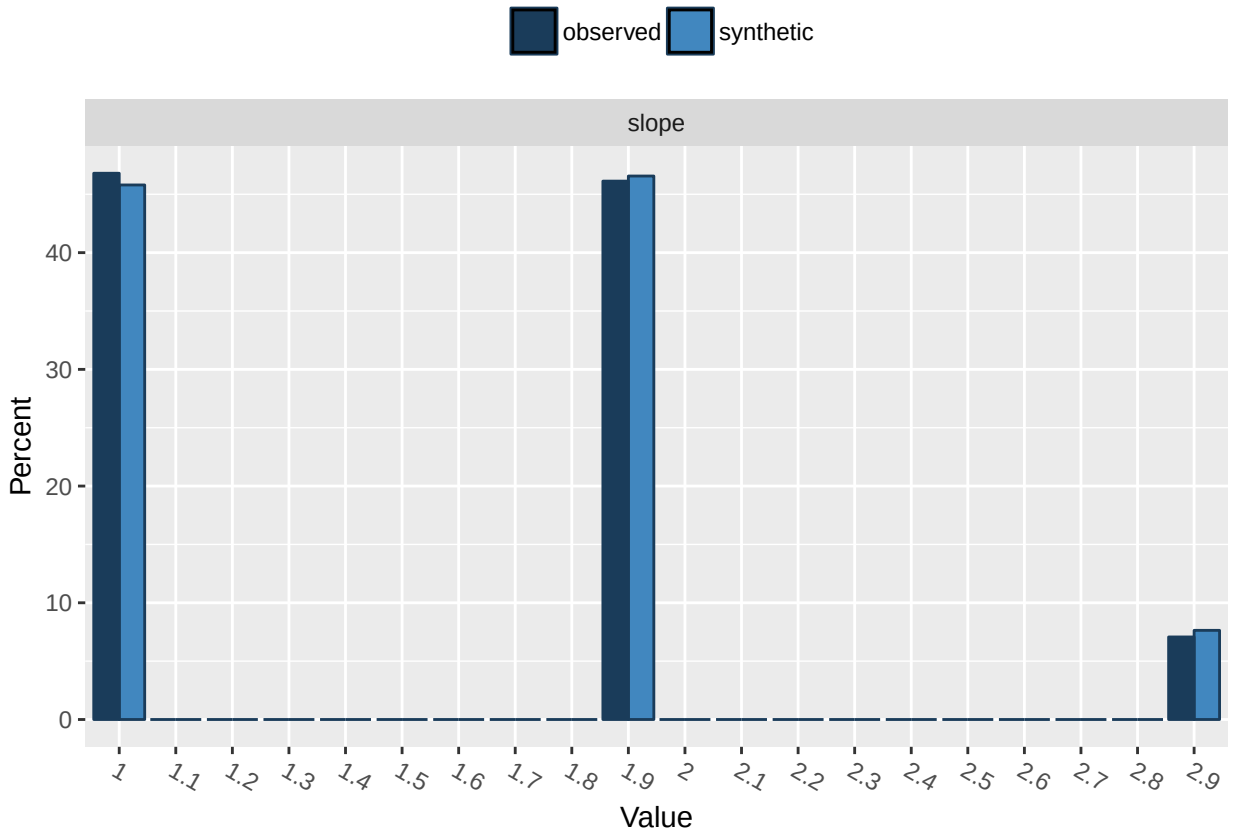

```
##
## Comparing percentages observed with synthetic
##
## $exer_angina
##                   0 0.05 0.1 0.15 0.2 0.25 0.3 0.35 0.4 0.45 0.5 0.55 0.6
## observed  67.34007    0   0    0   0    0   0    0   0    0   0    0   0
## synthetic 67.02000    0   0    0   0    0   0    0   0    0   0    0   0
##           0.65 0.7 0.75 0.8 0.85 0.9     0.95
## observed     0   0    0   0    0   0 32.65993
## synthetic    0   0    0   0    0   0 32.98000
```



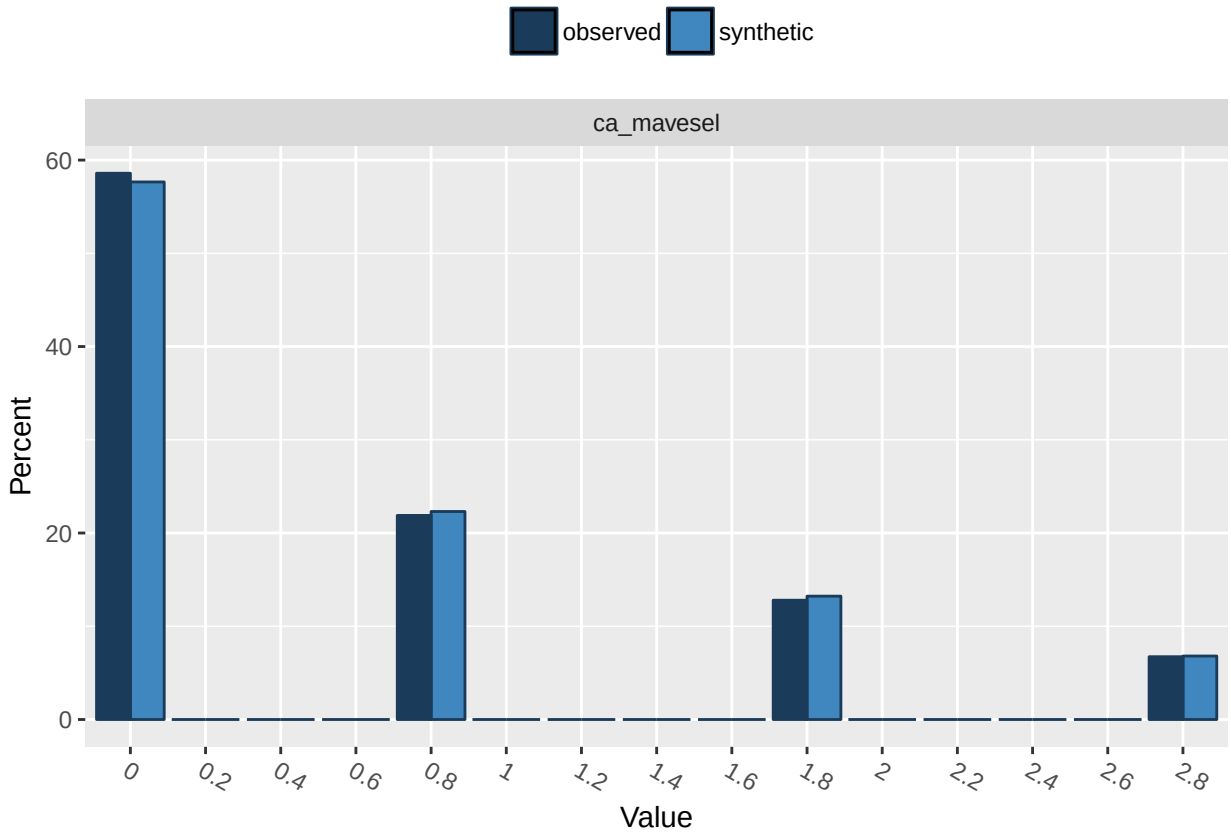

```
##
## Comparing percentages observed with synthetic
##
## $oldpeak
##                   0      0.5        1      1.5        2      2.5        3
## observed  43.77104 14.81481 12.79461 11.78451 4.040404 6.060606 2.356902
## synthetic 41.98200 15.11000 13.86400 11.54800 3.512000 7.400000 2.188000
##                 3.5        4 4.5 5       5.5         6
## observed  2.693603 1.010101   0 0 0.3367003 0.3367003
## synthetic 2.958000 0.904000   0 0 0.3440000 0.1900000
```



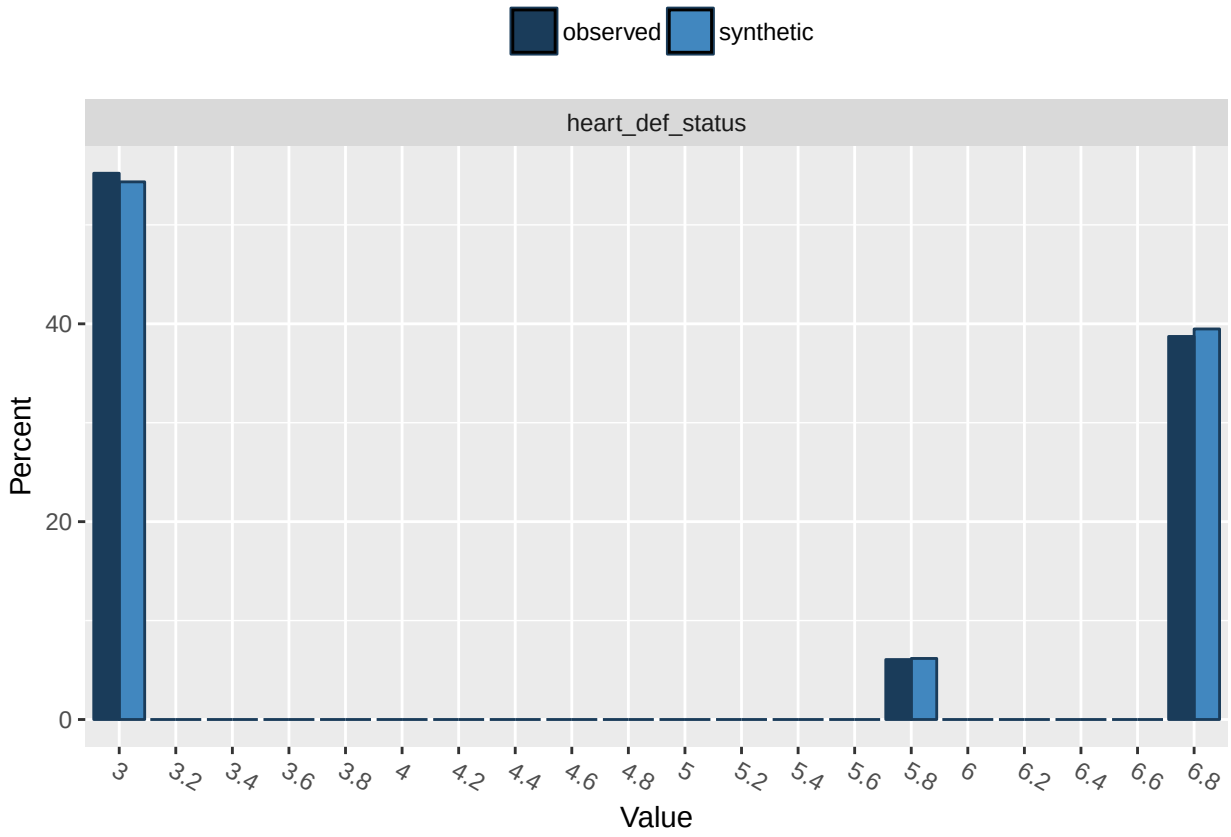

```
##
## Comparing percentages observed with synthetic
##
## $slope
##                  1 1.1 1.2 1.3 1.4 1.5 1.6 1.7 1.8      1.9 2 2.1 2.2 2.3
## observed  46.80135   0   0   0   0   0   0   0   0 46.12795 0   0   0   0
## synthetic 45.79800   0   0   0   0   0   0   0   0 46.56400 0   0   0   0
##           2.4 2.5 2.6 2.7 2.8      2.9
## observed    0   0   0   0   0 7.070707
## synthetic   0   0   0   0   0 7.638000
```



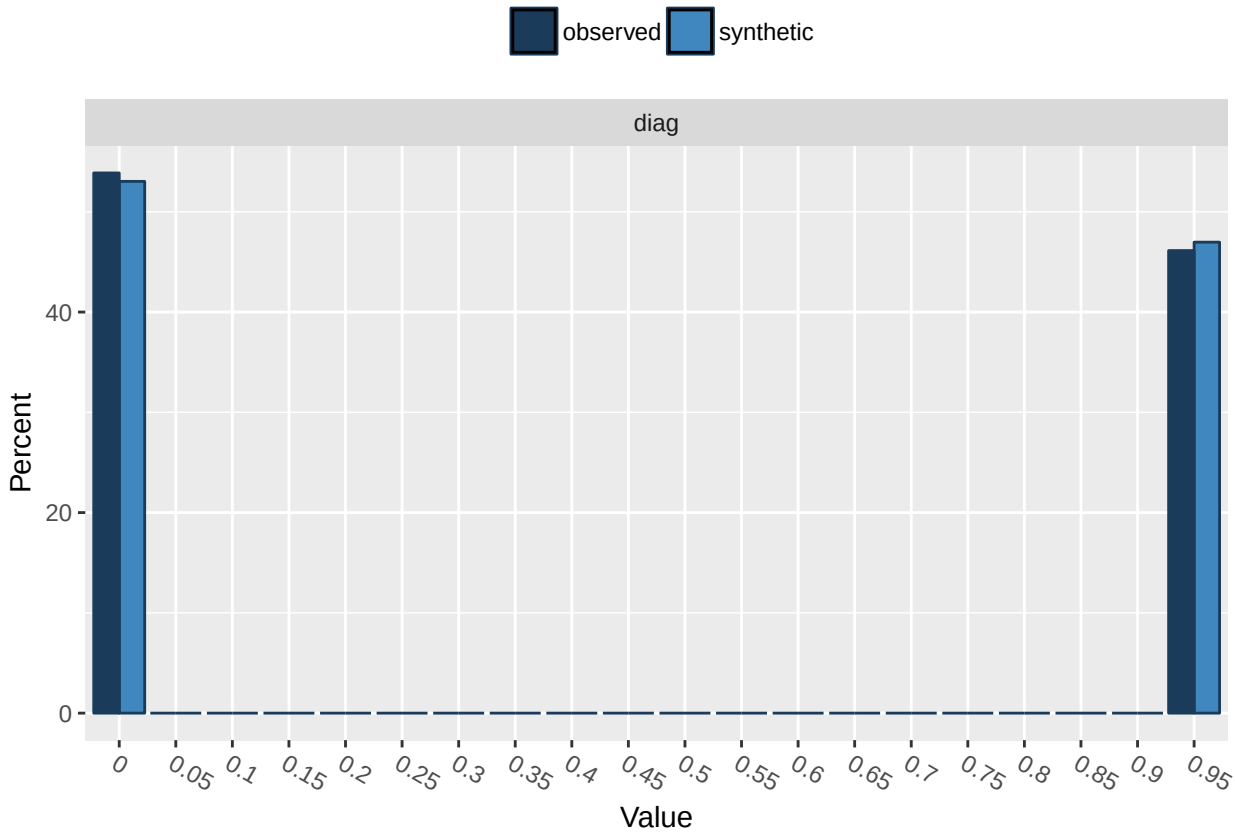

```
##
## Comparing percentages observed with synthetic
##
## $ca_mavesel
##                   0 0.2 0.4 0.6      0.8 1 1.2 1.4 1.6      1.8 2 2.2 2.4
## observed  58.58586   0   0   0 21.88552 0   0   0   0 12.79461 0   0   0
## synthetic 57.65400   0   0   0 22.31000 0   0   0   0 13.22800 0   0   0
##           2.6      2.8
## observed    0 6.734007
## synthetic   0 6.808000
```




```
##
## Comparing percentages observed with synthetic
##
## $heart_def_status
##                   3 3.2 3.4 3.6 3.8 4 4.2 4.4 4.6 4.8 5 5.2 5.4 5.6
## observed  55.21886   0   0   0   0 0   0   0   0   0 0   0   0   0
## synthetic 54.35000   0   0   0   0 0   0   0   0   0 0   0   0   0
##                 5.8 6 6.2 6.4 6.6      6.8
## observed  6.060606 0   0   0   0 38.72054
## synthetic 6.170000 0   0   0   0 39.48000
```




```
##
## Comparing percentages observed with synthetic
##
## $diag
##                  0 0.05 0.1 0.15 0.2 0.25 0.3 0.35 0.4 0.45 0.5 0.55 0.6
## observed  53.87205    0   0    0   0    0   0    0   0    0   0    0   0
## synthetic 53.03000    0   0    0   0    0   0    0   0    0   0    0   0
##           0.65 0.7 0.75 0.8 0.85 0.9     0.95
## observed     0   0    0   0    0   0 46.12795
## synthetic    0   0    0   0    0   0 46.97000
```


observed
synthetic
diag
Percent
40
20
0
0
0.05
0.1
0.15
0.2
0.25
0.3
0.35
0.4
0.45
0.5
0.55
0.6
0.65
0.7
0.75
0.8
0.85
0.9
0.95
Value