THE UNIVERSITY of EDINBURGH

Generating synthetic data with the synthpop package for R

Going over first practical

Gillian Raab
Administrative Data Research Centre – Scotland

Administrative Data Research Network

An ESRC Data Investment

My variables

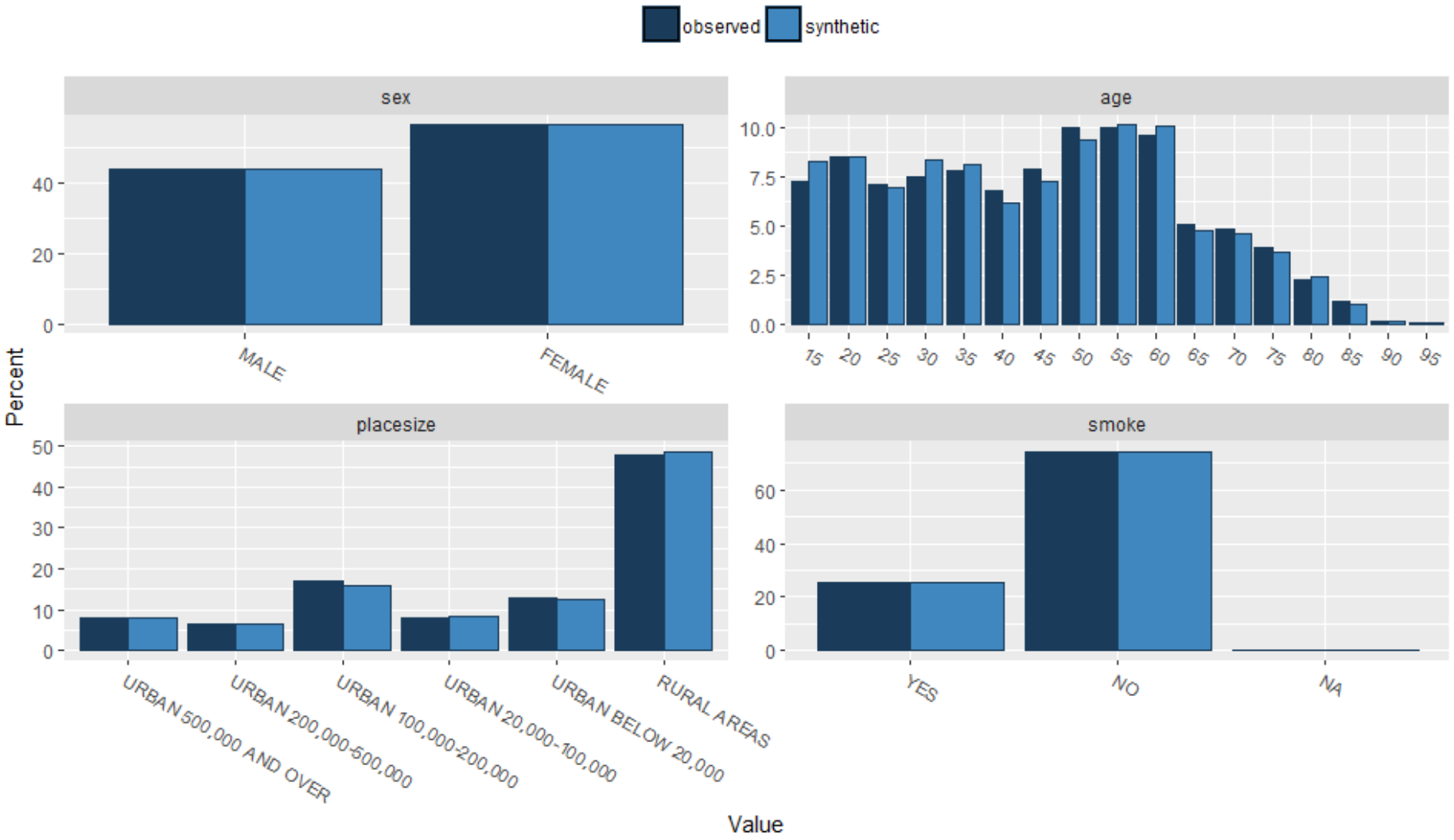
```
test1 <- SD2011[,c(1,2,4,23)]
```

age, sex, placesize and smoking

```
syn1 <- syn(test1) ## default method
```

Univariate statistics look OK

compare(syn1,test1)

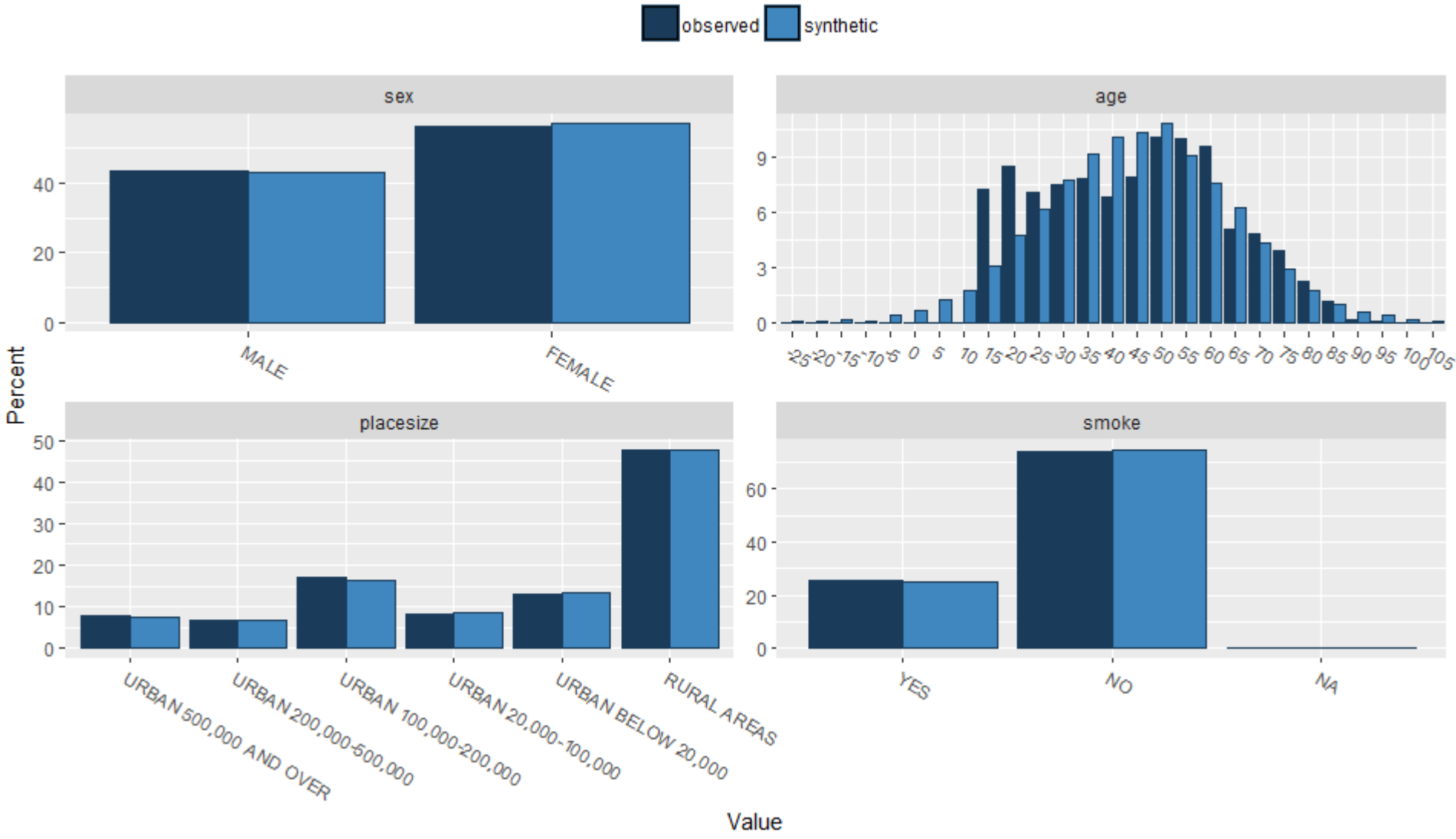


Alternative synthesis

```
syn2 <- syn(test1,  visit.sequence = 4:1,
method = c("sample","norm","ctree","ctree"))

compare(syn2,test1)
```

Not so good!

compare(syn2,test1)

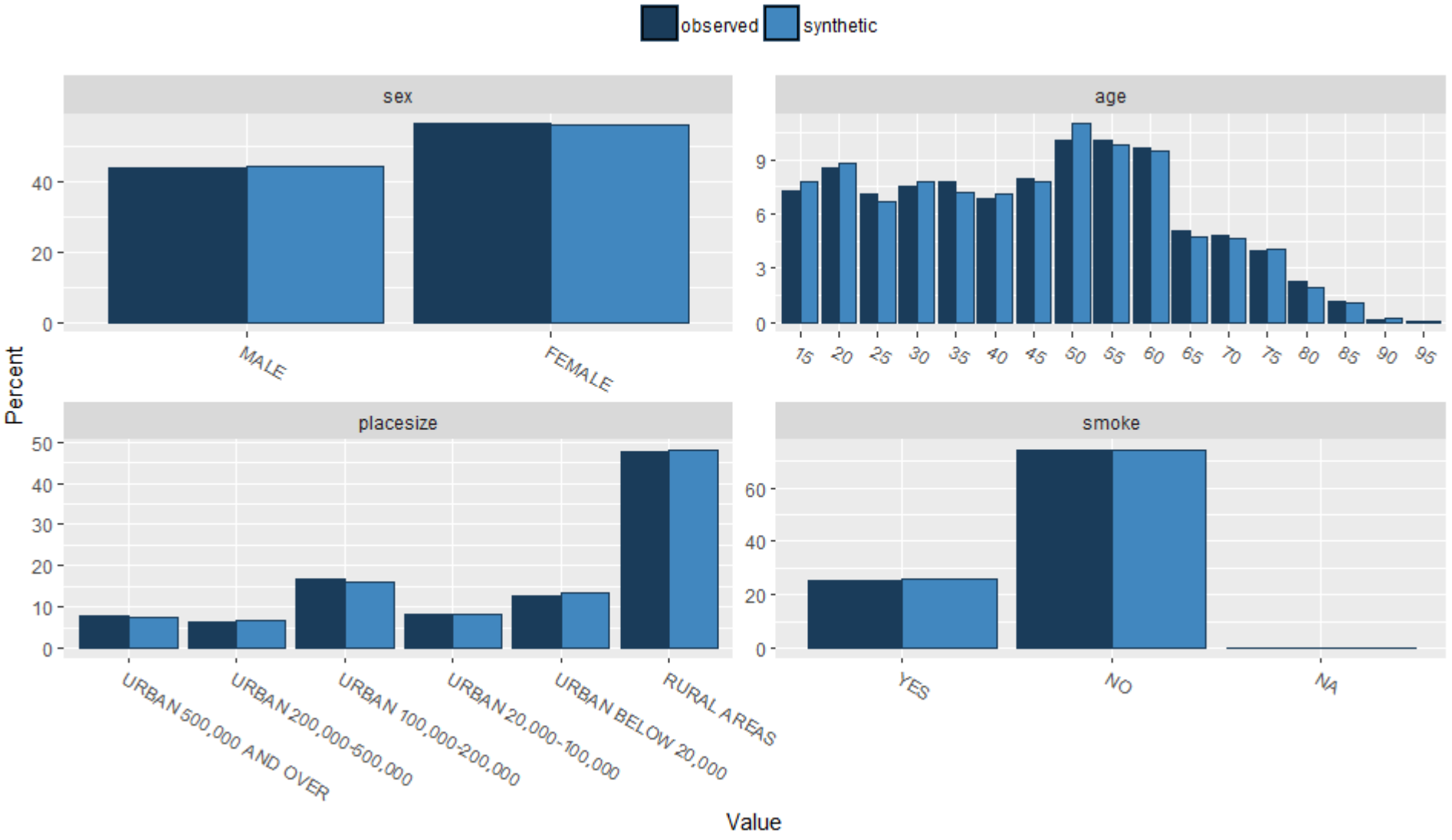


Some parametric methods are OK

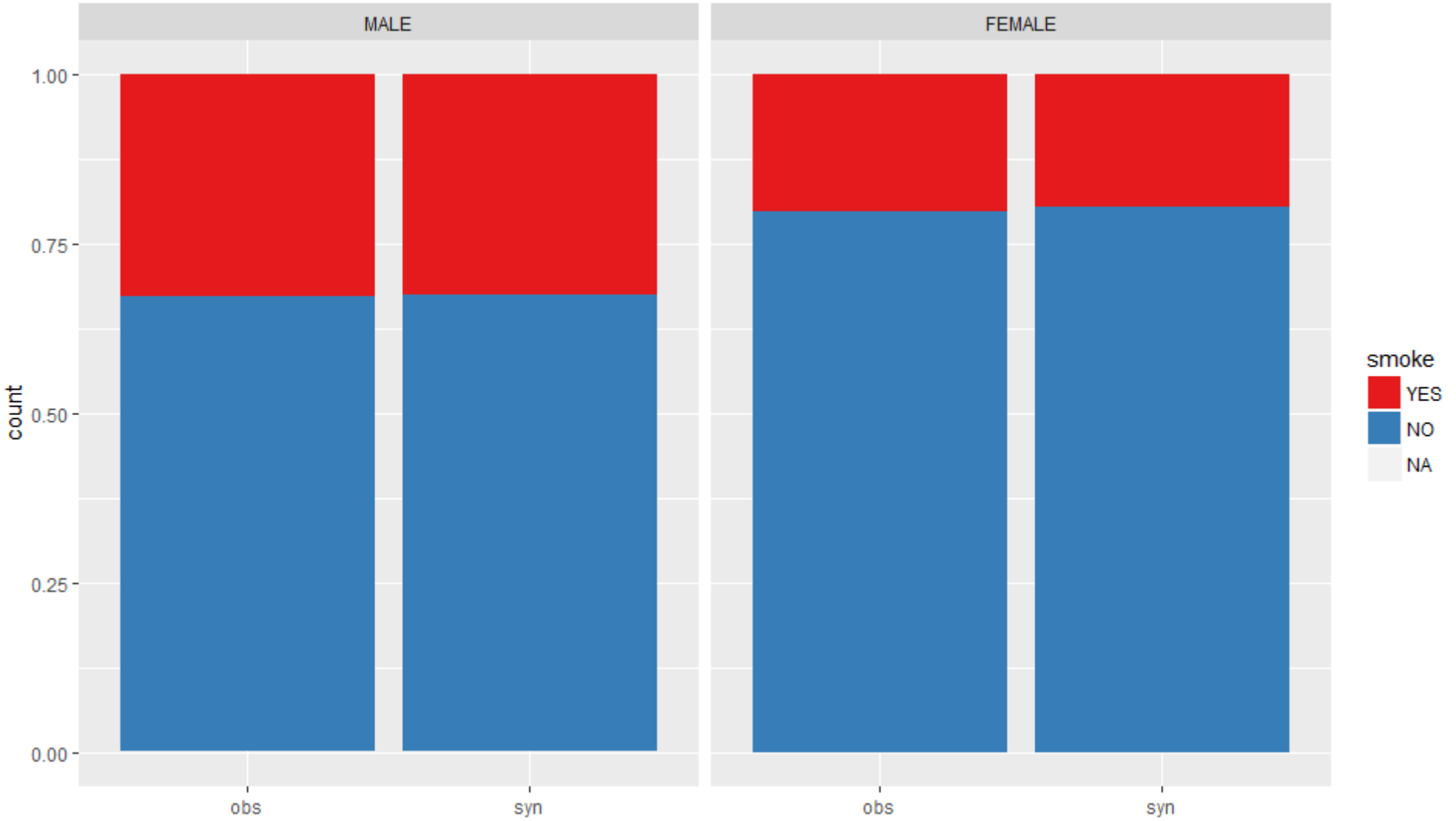
```
> syn3 <- syn(test1,visit.sequence = 4:1,
method = c("parametric"))

syn  variables

1   smoke placesize age sex

> syn3$method
```

```
     sex        age          placesize    smoke
  "logreg" "normrank"   "polyreg"    "sample"
```

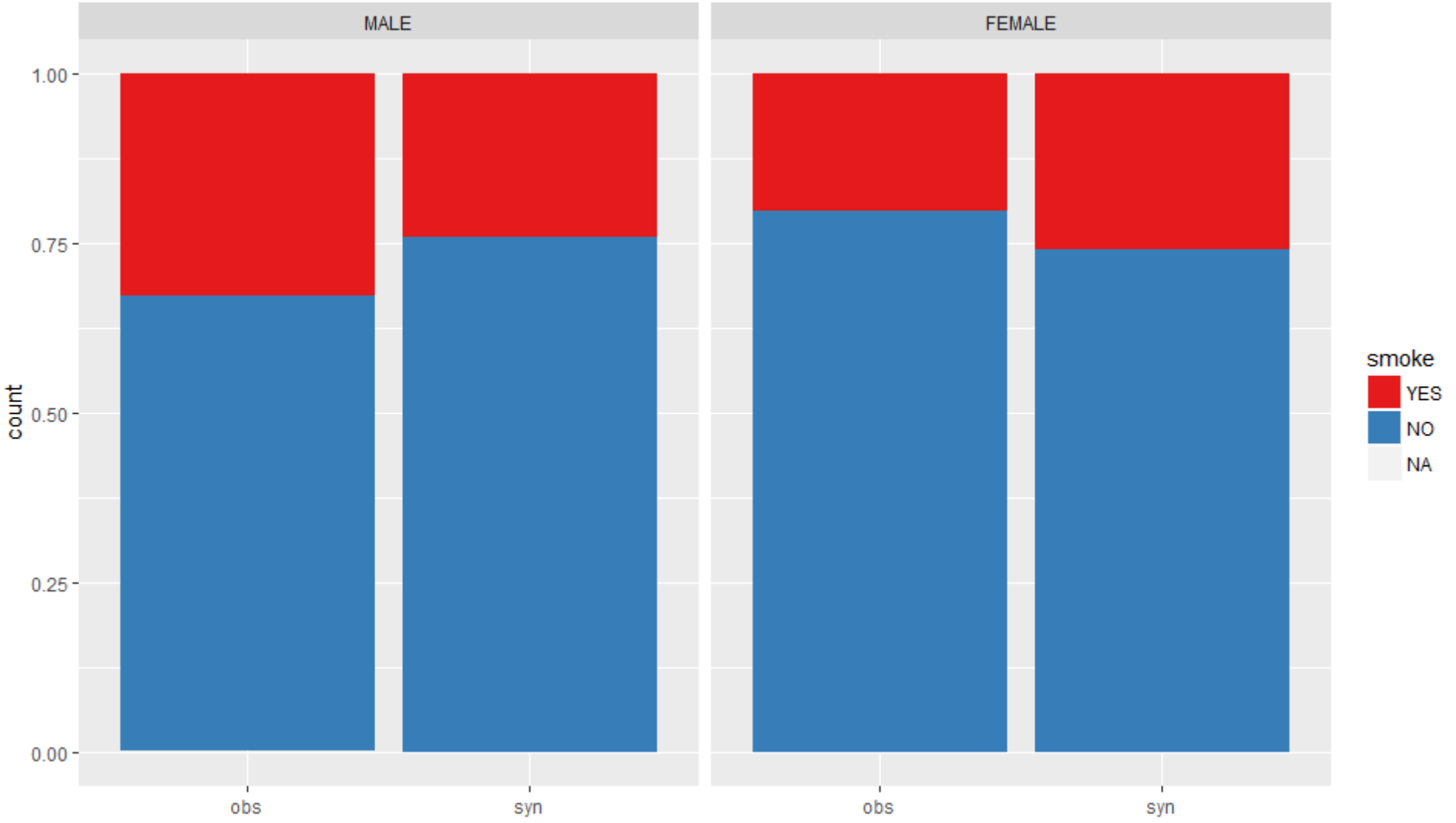
compare(syn3,test1)



```
multi.compare(syn1,test1,var = c("smoke"),by = c("sex"))
```



```
multi.compare(syn2,test1,var = c("smoke"),by = c("sex"))
```



Is this worse than chance?

```
utility.tab(syn2,test1, vars = c("sex", "smoke")

Observed: ($tab.obs)
        smoke
sex        YES   NO <NA>
  MALE     712 1465    5
  FEMALE   565 2248    5
Synthesised:
($tab.syn)
        smoke
sex        YES   NO <NA>
  MALE     516 1628    3
  FEMALE   738 2112    3
Number of cells in each table: 6; Number of cells contributing to
utility measures: 6
Utility score results
Freeman Tukey (FT): 136.82; Ratio to degrees of freedom (df): 27.36;
p-value: 0
Voas Williamson (VW): 136.17; Ratio to degrees of freedom (df):
27.23; p-value: 0
```

Model fitting - what a user would do

```
synfit <- glm(smoke ~ ., data =
syn1$syn, family = "binomial") ###
just main effects

summary(synfit)

synfit2 <- glm(smoke ~ sex +
placesize + poly(age,2), data =
syn1$syn, family = "binomial") ###
add quadratic age effects

anova(synfit,synfit2)
```

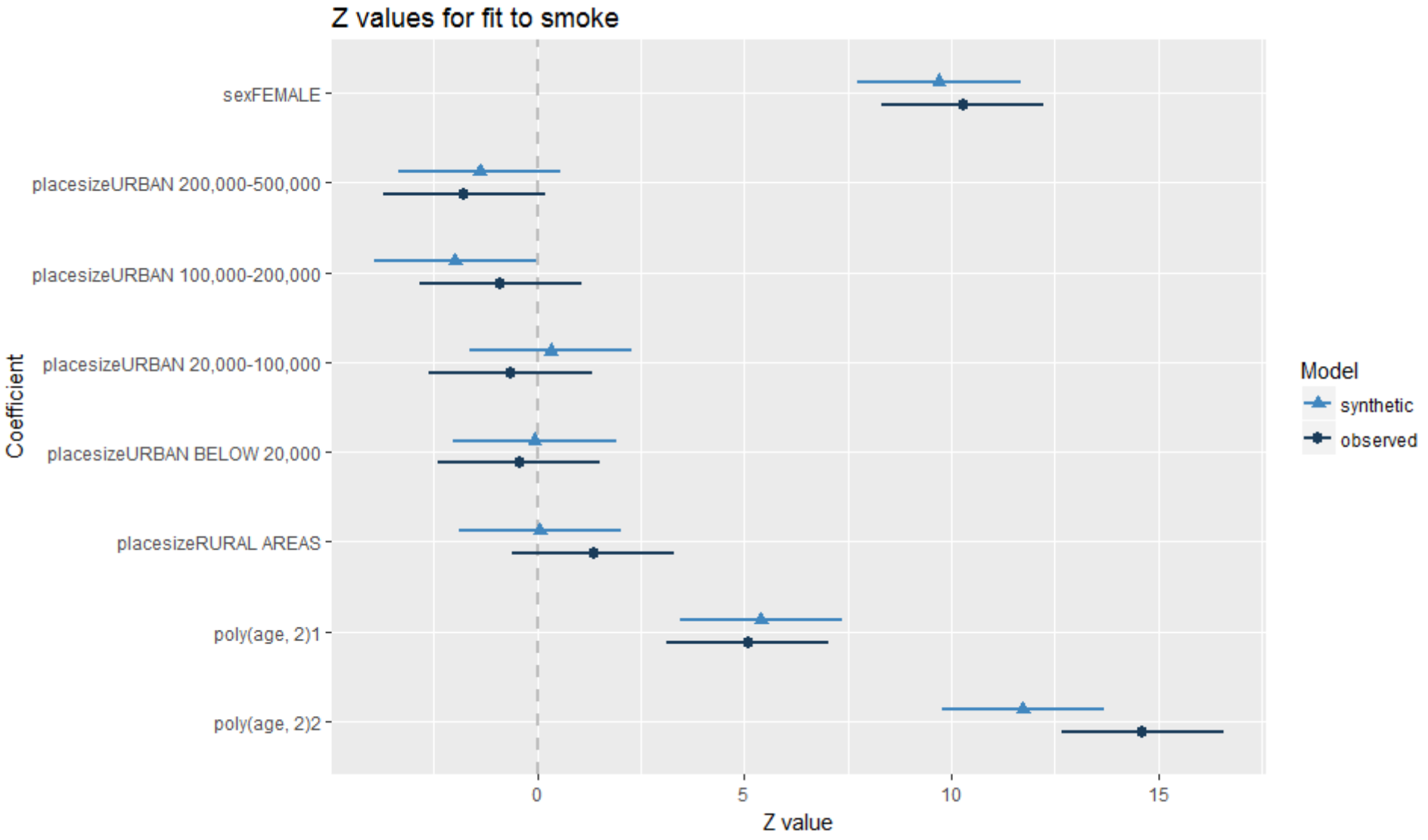
To evaluate and compare with real data

```
synfit.synds <- glm.synds(smoke ~ sex +
placesize + poly(age,2), data = syn1,
family = "binomial")


compare(synfit.synds,test1)
```

Results for syn1

```
Call used to fit models to the data:
glm.synds(formula = smoke ~ sex + placesize + poly(age, 2),
family = "binomial", data = syn1)

Differences between results based on synthetic and observed data:
                                 Std. coef diff p value CI overlap
(Intercept)                          0.4462240   0.655  0.8861653
sexFEMALE                           -0.5835295   0.560  0.8511377
placesizeURBAN 200,000-500,000       0.3903786   0.696  0.9004118
placesizeURBAN 100,000-200,000      -1.1018398   0.271  0.7189132
placesizeURBAN 20,000-100,000        0.9787835   0.328  0.7503058
placesizeURBAN BELOW 20,000          0.3844041   0.701  0.9019359
placesizeRURAL AREAS                -1.2901589   0.197  0.6708718
poly(age, 2)1                        0.3155280   0.752  0.9195067
poly(age, 2)2                       -2.8879149   0.004  0.2632735

Measures for one synthesis and 9 coefficients
Mean confidence interval overlap:  0.7625024
Mean absolute std. coef diff:  0.9309735
Lack-of-fit: 22.02378; p-value 0.009 for test that synthesis model
is compatible
with a chi-squared test with 9 degrees of freedom
```

Results for syn1 plot



Results for syn2 plot