INTRODUCTION TO STATISTICAL MODELING

Prediction error for categorical variables

Modeling marital status

```
# Base model: Just `age` as the explanatory variables
> mod_a <- rpart(married ~ age, data = Training_data, cp = 0.001)

# Extended model: Both `age` and `sector` as explanatory variables
> mod_b <- rpart(married ~ age + sector,
                 data = Training_data, cp = 0.001)
```

Categorical outputs

```
# Base model
> mod_a_outputs <- predict(mod_a, newdata = Testing_data,
                           type = "class")
> head(mod_a_outputs)
[1] Married Single Single Married Married Married

# Extended model
> mod_b_outputs <- predict(mod_b, newdata = Testing_data,
                           type = "class")
> head(mod_b_outputs)
[1] Married Single Single Married Married Married

# Actual values
> head(Testing_data$married)
[1] Married Single Married Single Single Married
```

Counting categorical errors

```
> with(data = Testing_data, sum(married != mod_a_outputs))

> with(data = Testing_data, sum(married != mod_b_outputs))
```

Counting categorical errors

```
> with(data = Testing_data, sum(married != mod_a_outputs))
[1] 109
> with(data = Testing_data, sum(married != mod_b_outputs))
[1] 110
```

The categorical error rate

```
> with(data = Testing_data, mean(married != mod_a_outputs))

> with(data = Testing_data, mean(married != mod_b_outputs))
```

The categorical error rate

```
> with(data = Testing_data, mean(married != mod_a_outputs))
[1] 0.3263473
> with(data = Testing_data, mean(married != mod_b_outputs))
[1] 0.3293413
```

- Similar to assessing performance for quantitative outputs
- Test whether predicted values match actual values
- Calculate error rate

The output as probabilities

```
> mod_a_probs <- predict(mod_a, newdata = Testing_data, type = "prob")
> res_1 <- data.frame(actual = Testing_data$married, mod_a_probs)
> head(res_1)
   actual   Married    Single
2 Married 0.8265306 0.1734694
3  Single 0.2222222 0.7777778
4 Married 0.8265306 0.1734694
5 Married 0.5833333 0.4166667
7 Married 0.4090909 0.5909091
8  Single 0.8265306 0.1734694

> mod_b_probs <- predict(mod_b, newdata = Testing_data, type = "prob")
> res_2 <- data.frame(actual = Testing_data$married, mod_b_probs)
> head(res_2)
   actual    Married     Single
2 Married 0.90909091 0.09090909
3  Single 0.28571429 0.71428571
...
```

Summarizing all cases with likelihood

```
> likelihood_a <- with(res_1, ifelse(actual == "Married", Married, Single))
> sum(log(likelihood_a))
[1] -214.863

> likelihood_b <- with(res_2, ifelse(actual == "Married", Married, Single))
> sum(log(likelihood_b))
[1] -227.8955
```

Likelihood: extract the probability that the model assigned to the observed outcome

INTRODUCTION TO STATISTICAL MODELING

Let’s practice!

INTRODUCTION TO STATISTICAL MODELING

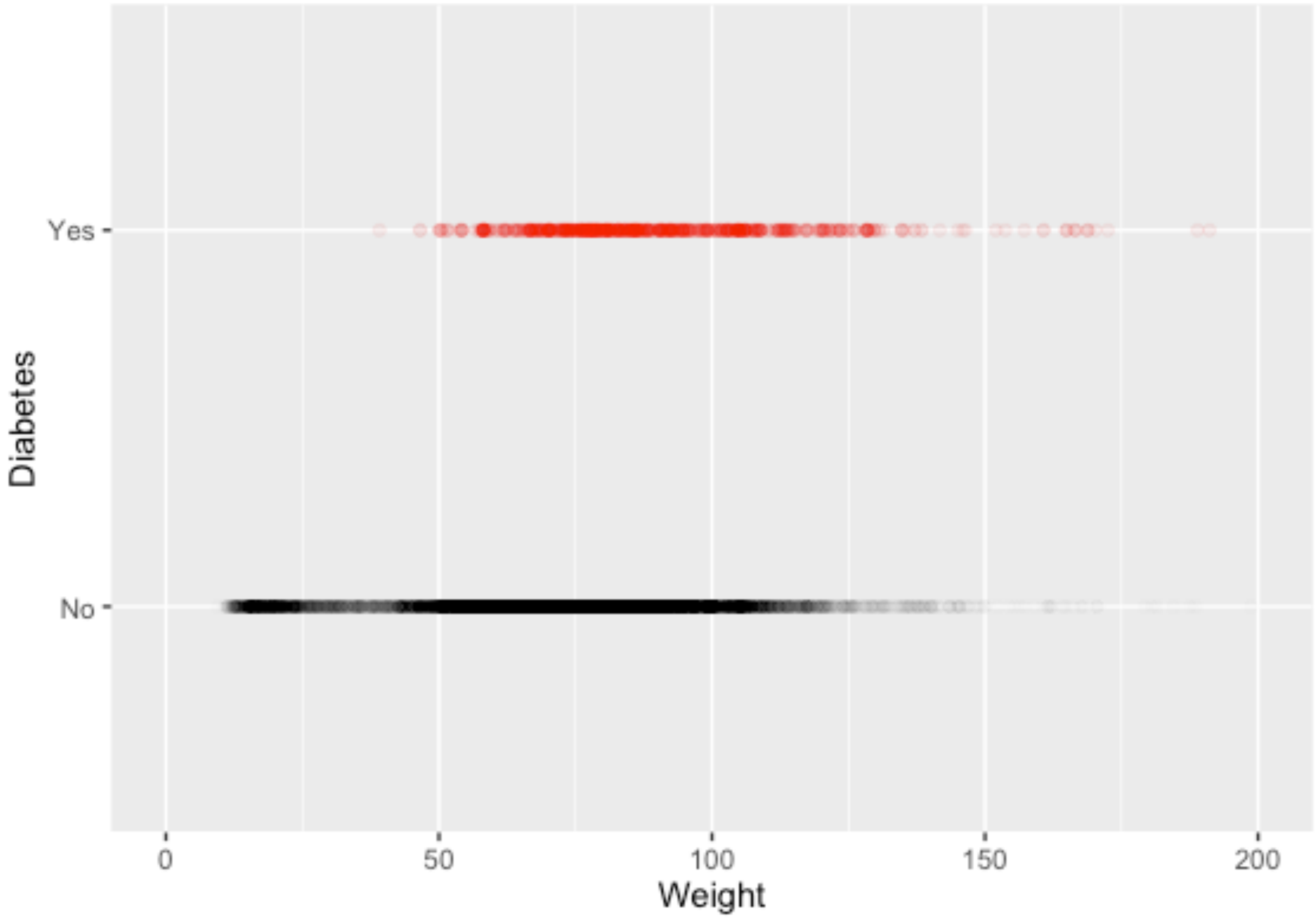
Exploring data for relationships

Factors in health and disease

```
> library(NHANES)
> library(dplyr)

# National Health and Nutrition Evaluation Survey (NHANES)
> names(NHANES) %>% head(20)
 [1] "ID"            "SurveyYr"      "Gender"        "Age"
 [5] "AgeDecade"     "AgeMonths"     "Race1"         "Race3"
 [9] "Education"     "MaritalStatus" "HHIncome"      "HHIncomeMid"
[13] "Poverty"       "HomeRooms"     "HomeOwn"       "Work"
[17] "Weight"        "Length"        "HeadCirc"      "Height"
```

Is body weight related to having diabetes?



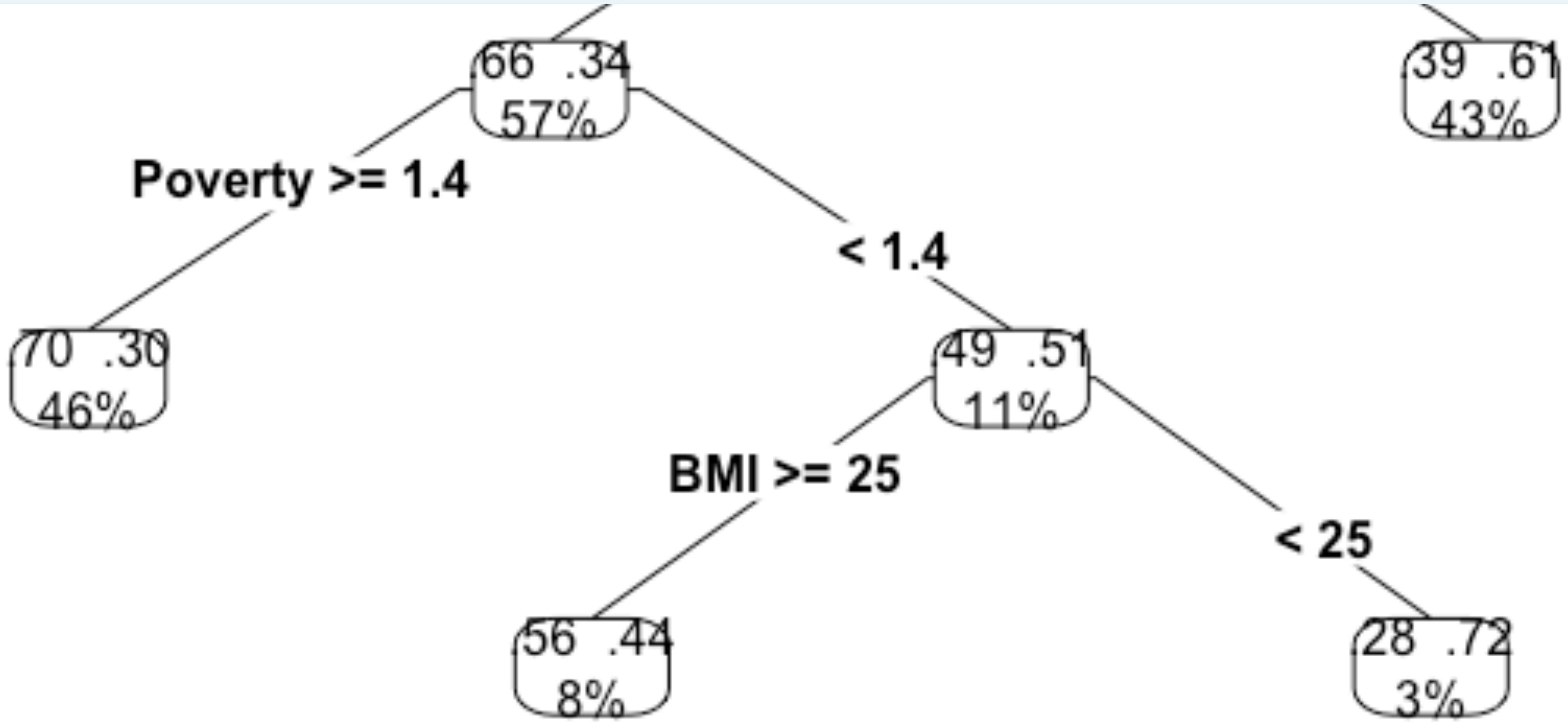
What accounts for smoking?

```
> NHANES %>%
    select(SmokeNow, Poverty, MaritalStatus, Gender, BMI, TotChol,
           AgeFirstMarij, SmokeNow)
   SmokeNow Poverty MaritalStatus Gender   BMI TotChol AgeFirstMarij
     <fctr>   <dbl>        <fctr> <fctr> <dbl>   <dbl>         <int>
1        No    1.36       Married   male 32.22    3.49            17
2        No    1.36       Married   male 32.22    3.49            17
3        No    1.36       Married   male 32.22    3.49            17
4        NA    1.07            NA   male 15.30      NA            NA
5       Yes    1.91   LivePartner female 30.57    6.70            18
6        NA    1.84            NA   male 16.82    4.86            NA
7        NA    2.33            NA   male 20.64    4.09            NA
8        NA    5.00       Married female 27.24    5.82            13
9        NA    5.00       Married female 27.24    5.82            13
10       NA    5.00       Married female 27.24    5.82            13
..      ...     ...           ...    ...   ...     ...           ...
```

Modeling with recursive partitioning (`rpart`)

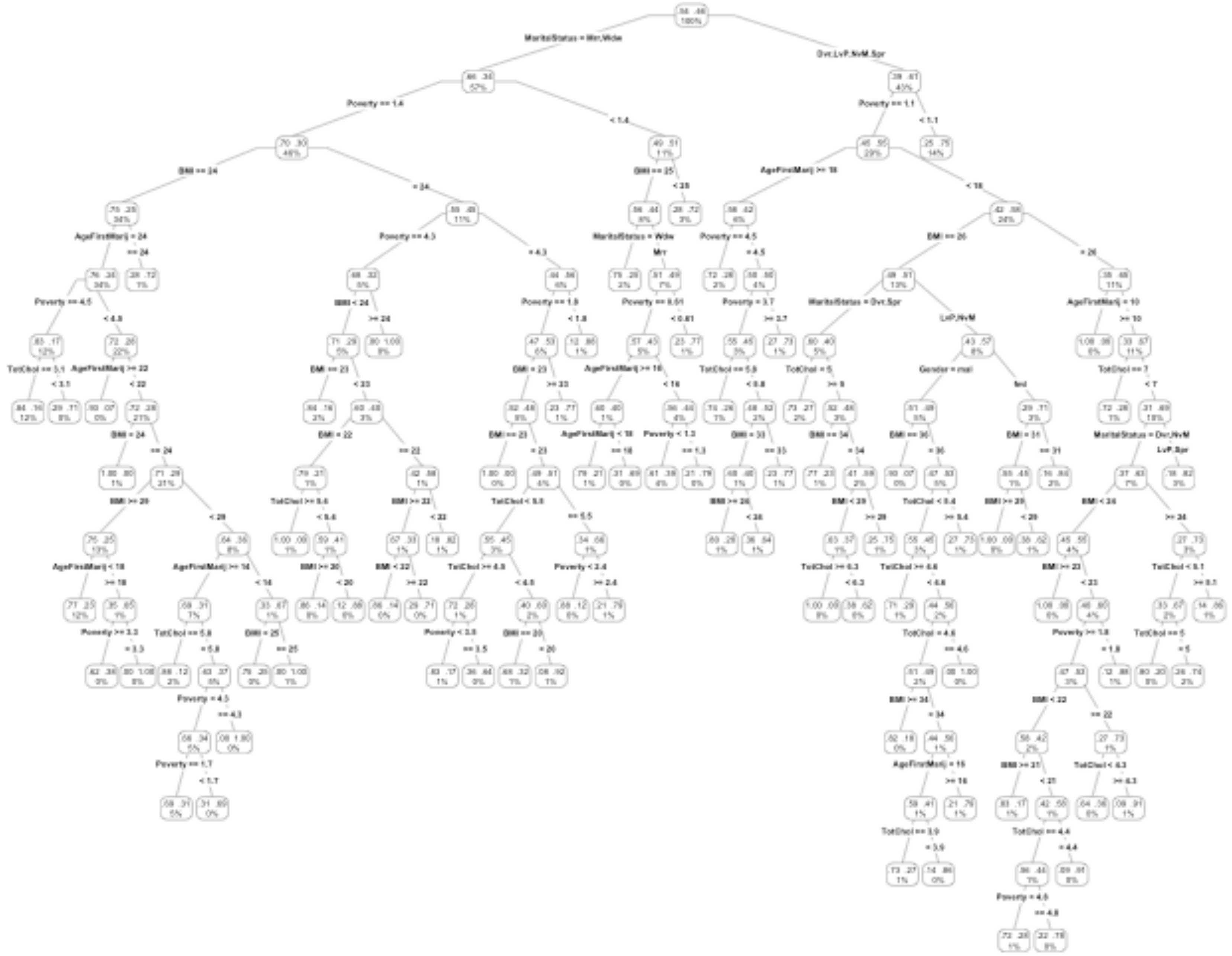
Who smokes cigarettes?

```
> library(rpart.plot)
> model <- rpart(SmokeNow ~ Poverty + MaritalStatus + Gender + BMI +
                      TotChol + AgeFirstMarij, data = NHANES)
> prp(model, type = 4, extra = 105, varlen = 0)
```



Pushing `rpart` for more complexity

```
> model <- rpart(SmokeNow ~ Poverty + MaritalStatus +
                     Gender + BMI + TotChol + AgeFirstMarij,
                   data = NHANES, cp = 0.002)
> prp(model, type = 4, extra = 105, varlen = 0)
```



Let’s practice!