Stroke data EDA

Guy Maskall

21 March, 2020

Introduction

This document downloads and performs an initial view of a stroke dataset.

```
stroke <- read_csv("train_2v.csv")
```

```
## Parsed with column specification:
## cols(
##   id = col_double(),
##   gender = col_character(),
##   age = col_double(),
##   hypertension = col_double(),
##   heart_disease = col_double(),
##   ever_married = col_character(),
##   work_type = col_character(),
##   Residence_type = col_character(),
##   avg_glucose_level = col_double(),
##   bmi = col_double(),
##   smoking_status = col_character(),
##   stroke = col_double()
## )
```

```
stroke %>% glimpse
```

```
## Observations: 43,400
## Variables: 12
## $ id                <dbl> 30669, 30468, 16523, 56543, 46136, 32257, 52800, ...
## $ gender            <chr> "Male", "Male", "Female", "Female", "Male", "Fema...
## $ age               <dbl> 3, 58, 8, 70, 14, 47, 52, 75, 32, 74, 79, 79, 37,...
## $ hypertension      <dbl> 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0...
## $ heart_disease     <dbl> 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 0, 0...
## $ ever_married      <chr> "No", "Yes", "No", "Yes", "No", "Yes", "Yes", "Ye...
## $ work_type         <chr> "children", "Private", "Private", "Private", "Nev...
## $ Residence_type    <chr> "Rural", "Urban", "Urban", "Rural", "Rural", "Urb...
## $ avg_glucose_level <dbl> 95.12, 87.96, 110.89, 69.04, 161.28, 210.95, 77.5...
## $ bmi               <dbl> 18.0, 39.2, 17.6, 35.9, 19.1, 50.1, 17.7, 27.0, 3...
## $ smoking_status    <chr> NA, "never smoked", NA, "formerly smoked", NA, NA...
## $ stroke            <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
```

Initial data QA

Check id is unique and for any missing values.

Uniqueness of ID

```
stroke %>%
    count(id) %>%
    filter(n > 1)
```

```
## # A tibble: 0 x 2
## # ... with 2 variables: id <dbl>, n <int>
```

```
stroke <- stroke %>% select(-id)
```

Check for NA

Do any features have any NA?

```
# summarise_all soon to be superceded in dplyr 1.0.0
stroke %>%
    summarise_all(~sum(is.na(.))) %>%
    t
```

```
##                    [,1]
## gender                0
## age                   0
## hypertension          0
## heart_disease         0
## ever_married          0
## work_type             0
## Residence_type        0
## avg_glucose_level     0
## bmi                1462
## smoking_status    13292
## stroke                0
```

We have some missing bmi and quite a lot of missing smoking_status. Is there any obvious pattern in the missing values, or relationship between the instances?

```
stroke %>%
    mutate(bmi_na = if_else(is.na(bmi), "yes", "no")) %>%
    filter(bmi_na == "yes" | is.na(smoking_status)) %>%
    group_by(bmi_na) %>%
    count(smoking_status)
```

```
## # A tibble: 5 x 3
## # Groups:   bmi_na [2]
##   bmi_na smoking_status      n
##   <chr>  <chr>           <int>
## 1 no     <NA>            12866
## 2 yes    formerly smoked   394
## 3 yes    never smoked      306
## 4 yes    smokes            336
## 5 yes    <NA>              426
```

From the above, probably the missing BMI values are missing at random, certainly with respect to smoking, seeing as they're broadly evenly distributed with smoking status. The largest single group of missing values is that for smoking status (where BMI is not missing).

Stroke and smoking

We saw above that most of the missing values are in the smoking status column. How does this seem to relate to stroke?

```
stroke %>%
    count(smoking_status, stroke) %>%
    pivot_wider(names_from=stroke, values_from=n) %>%
    mutate(stroke_pc = 100*`1`/`0`)
```

```
## # A tibble: 4 x 4
##   smoking_status    `0`   `1` stroke_pc
##   <chr>           <int> <int>     <dbl>
## 1 formerly smoked  7272   221      3.04
## 2 never smoked    15769   284      1.80
## 3 smokes           6429   133      2.07
## 4 <NA>            13147   145      1.10
```

From the above, we see that *smokes* and *never smoked* have a similar rate of incidence of stroke, *formerly smoked* seems a bit higher, and the *NA* smoking status group a bit lower. On the face of it, there are some curious things here. How does it makes sense that "formerly smoked" has a higher incidence of stroke than *smokes*, for example? Could there be other factors involved, for example people forced to give up smoking because of another health issue or perhaps taking up unhealthy eating habits to compensate for no longer smoking?

Do other stuff first to check distributions, but undersampling majority class repeatedly generates some consistent results wrt age and may implicate smoking, depending on the composition of the groups.

```
set.seed(47)
complete_cases <- stroke %>%
    filter(complete.cases(.))
run_glm <- function() {
    glm(stroke ~ ., family=gaussian, complete_cases %>%
            group_by(stroke) %>%  sample_n(548)) %>%
    summary
}

print(run_glm())
```

```
##
## Call:
## glm(formula = stroke ~ ., family = gaussian, data = complete_cases %>%
##     group_by(stroke) %>% sample_n(548))
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.01983  -0.32604   0.04539   0.29950   1.03111
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 -0.2561585  0.1591742  -1.609 0.107843
## genderMale                   0.0133074  0.0264136   0.504 0.614498
## age                          0.0135245  0.0008441  16.022  < 2e-16 ***
## hypertension                 0.0799254  0.0322678   2.477 0.013403 *
## heart_disease                0.1237507  0.0395916   3.126 0.001821 **
## ever_marriedYes             -0.0791385  0.0374892  -2.111 0.035004 *
## work_typeGovt_job          -0.0981213  0.1550050  -0.633 0.526854    
## work_typeNever_worked      -0.1346361  0.3272353  -0.411 0.680835    
## work_typePrivate           -0.1143959  0.1519900  -0.753 0.451821    
## work_typeSelf-employed     -0.0667985  0.1546680  -0.432 0.665912    
## Residence_typeUrban         0.0034897  0.0251281   0.139 0.889572    
## avg_glucose_level           0.0009122  0.0002438   3.742 0.000192 ***
## bmi                        -0.0009332  0.0019730  -0.473 0.636315    
## smoking_statusnever smoked  0.0046084  0.0300202   0.154 0.878025    
## smoking_statussmokes        0.0510177  0.0365439   1.396 0.162981    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for gaussian family taken to be 0.1703464)
## 
##     Null deviance: 274.00  on 1095  degrees of freedom
## Residual deviance: 184.14  on 1081  degrees of freedom
## AIC: 1187.4
## 
## Number of Fisher Scoring iterations: 2
```

```
print(run_glm())
```

```
## 
## Call:
## glm(formula = stroke ~ ., family = gaussian, data = complete_cases %>% 
##     group_by(stroke) %>% sample_n(548))
## 
## Deviance Residuals: 
##      Min        1Q    Median        3Q       Max  
## -1.04153  -0.32763   0.05888   0.31348   1.06084  
## 
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)    
## (Intercept)                -0.2309899  0.1285544  -1.797 0.072642 .  
## genderMale                  0.0242324  0.0265943   0.911 0.362400    
## age                         0.0130166  0.0008643  15.060  < 2e-16 ***
## hypertension                0.1085631  0.0327666   3.313 0.000953 ***
## heart_disease               0.1277511  0.0395171   3.233 0.001263 ** 
## ever_marriedYes            -0.0503513  0.0375954  -1.339 0.180756    
## work_typeGovt_job          -0.1581656  0.1283303  -1.232 0.218035    
## work_typeNever_worked      -0.0492383  0.2671162  -0.184 0.853787    
## work_typePrivate           -0.1422922  0.1240072  -1.147 0.251449    
## work_typeSelf-employed     -0.1292618  0.1277775  -1.012 0.311948    
## Residence_typeUrban         0.0289548  0.0252591   1.146 0.251919    
## avg_glucose_level           0.0008353  0.0002504   3.336 0.000878 ***
## bmi                         0.0002146  0.0019329   0.111 0.911599    
## smoking_statusnever smoked -0.0469770  0.0306347  -1.533 0.125456    
## smoking_statussmokes        0.0191133  0.0373637   0.512 0.609072    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for gaussian family taken to be 0.1733751)
## 
##     Null deviance: 274.00  on 1095  degrees of freedom
## Residual deviance: 187.42  on 1081  degrees of freedom
## AIC: 1206.7
##
## Number of Fisher Scoring iterations: 2
```

```
print(run_glm())
```

```
##
## Call:
## glm(formula = stroke ~ ., family = gaussian, data = complete_cases %>%
##     group_by(stroke) %>% sample_n(548))
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -0.95564  -0.30725   0.05161   0.29418   1.05987
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)               -0.1947513  0.1215472  -1.602 0.109388
## genderMale                 0.0062660  0.0259661   0.241 0.809357
## age                        0.0144178  0.0008497  16.968  < 2e-16 ***
## hypertension               0.1223461  0.0324092   3.775 0.000169 ***
## heart_disease              0.0952517  0.0382928   2.487 0.013016 *
## ever_marriedYes           -0.0792282  0.0370255  -2.140 0.032592 *
## work_typeGovt_job         -0.1488065  0.1226230  -1.214 0.225193
## work_typePrivate          -0.1478148  0.1183145  -1.249 0.211812
## work_typeSelf-employed    -0.1395757  0.1222013  -1.142 0.253632
## Residence_typeUrban        0.0060401  0.0248372   0.243 0.807906
## avg_glucose_level          0.0007434  0.0002419   3.073 0.002173 **
## bmi                       -0.0026782  0.0018359  -1.459 0.144908
## smoking_statusnever smoked -0.0076359  0.0297879  -0.256 0.797735
## smoking_statussmokes       0.0825829  0.0366399   2.254 0.024402 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.167373)
##
##     Null deviance: 274.0  on 1095  degrees of freedom
## Residual deviance: 181.1  on 1082  degrees of freedom
## AIC: 1167.1
##
## Number of Fisher Scoring iterations: 2
```

We repeated the undersampling a couple of times above because we'd noticed some variability in the results for different samples drawn from the majority class. Age is pretty consistently strongly significant. This is no great surprise for stroke risk. Hypertension and heart disease, and particularly average glucose level, are also heavily implicated. The signs of their coefficients are all intuitive as well; they are positive contributors to stroke risk.

We should pause here before continuing to dive into inference. We don't know how the data were sampled and we have multiple potentially confounding variables, especially categorical ones. Hypertension is binary, as is heart disease, and themselves are imbalanced:

```
stroke %>% count(hypertension)
```

```
## # A tibble: 2 x 2
##   hypertension     n
##          <dbl> <int>
## 1            0 39339
## 2            1  4061
```

```
stroke %>% count(heart_disease)
```

```
## # A tibble: 2 x 2
##   heart_disease     n
##           <dbl> <int>
## 1             0 41338
## 2             1  2062
```

We have some encouraging, and plausible, patterns, but we should remember the words of Ronald Fisher:

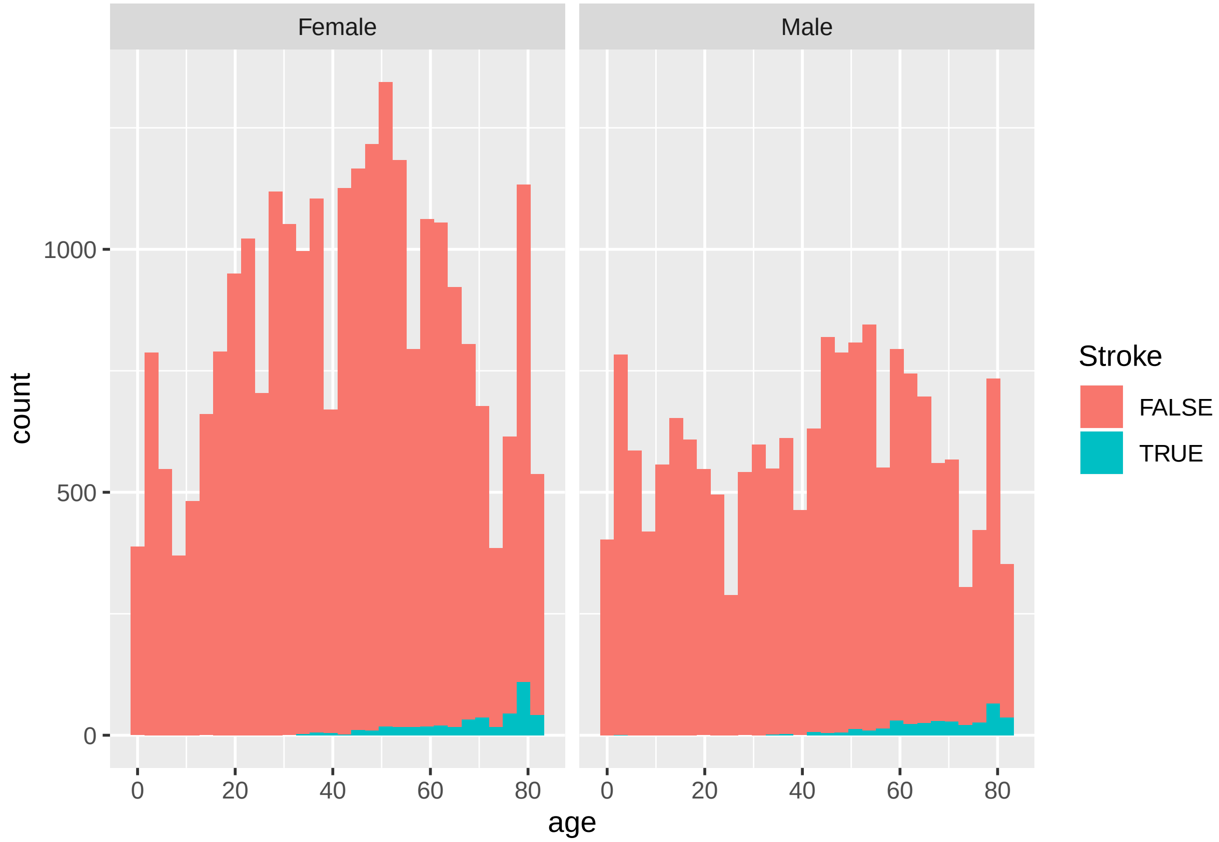
> To consult the statistician after an experiment is finished is often merely to ask him to conduct a post mortem examination. He can perhaps say what the experiment died of.

Stroke and age

The relationship between stroke risk and age is clear in the histograms below:

```
stroke %>%
    filter(gender != "Other") %>%
    ggplot() +
    geom_histogram(aes(x = age, fill = stroke==1)) +
    facet_wrap(~gender) +
    labs(fill="Stroke")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

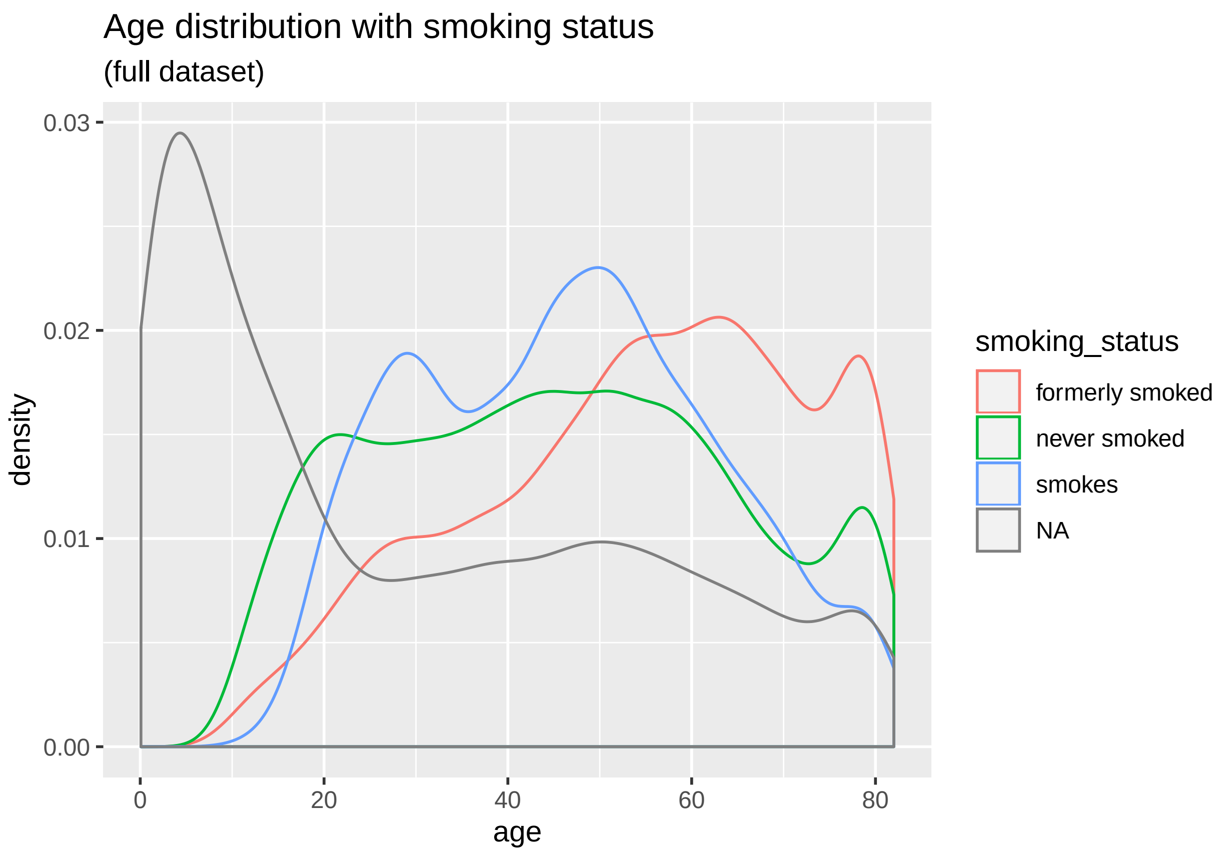


We have more women than men, perhaps up to twice as many. The age distribution for men is fairly flat, whereas for women there seem to be more weight in the middle but fairly symmetric. The increasing prevalence of stroke with age is clear.

Age and smoking

Now we return to looking at smoking status. We suggested above that there might be an occasional sign of a relationship between smoking status and stroke, but we saw a very clear signal that stroke risk was age related. Is there a relationship between age and smoking status?

```
stroke %>%
    ggplot() +
    geom_density(aes(x = age, colour = smoking_status)) +
    labs(title="Age distribution with smoking status",
         subtitle="(full dataset)")
```



The distributions above contain much detail we can now put into perspective. The missing values for smoking status are clearly dominated by children. This makes total sense. It's quite intuitive that the data pertaining to children did not include a smoking status. This also means that simply omitting samples missing a smoking status is fundamentally flawed because this is to greatly bias the data away from younger people. If the desire was to model stroke risk only for adults, that's one thing, but then this should be explicitly done on age, not accidentally by smoking status.

We can see another clear pattern as well. The category of ex-smokers ("formerly smoked") is heavily weighted to older ages. Again, this is an intuitive result given a moment's thought. In order to be a former smoker, one must first have been a smoker. A 20 year old has had only a few potential smoking years, whereas a 60 year old has many years of opportunity to be a smoker before giving up.

The young, below the early twenties, who have never smoked, dominate those who do smoke. This effect alone would bias the group who've never smoked to be younger, and so at less age-related risk of stroke. Having said that, both the "never smoked" and "formerly smoked" groups have an uptick at the oldest ages that is lacking in the "smokes" group. This could perhaps be a survival bias. It's not impossible that the suppression of the oldest ages from the smoking group might even make smoking appear protective of a stroke in some cases; the truth, of course, would be closer to the fact that smokers die of smoking-related disease before the stroke risk ramps up.

In short, there seem some powerful interactions between age and smoking status.

Can we model missing smoking_status values?

Model age by smoking (balance for smoking status) And can we drive glmnet on this data?