report.Rmd

2022-11-18

Abstract

The goal of this document is to study how certain covariates in the Brain Stroke Dataset depend on eachother. The covariates included in the data set are gender, age, hypertension, heart disease, worktype, average glucose level, bmi, smoking status, and stroke. Previous literature suggest that heart disease, high blood pressure, diabetes, cholesterol levels, smoking status, age, and sex are risk factors. We would like to study if these risk factors, as well as the other covariates, are determine stroke in this data set. In this examination, we determine the most likely risk factors in this dataset and reveal a number of interesting interactions between covariates. We also build a model purely for prediction accuracy, and test the reliability of the data with a cross validation.

Introduction

A stroke is a medical condition in which poor blood flow in the brain causes brain cell deaths. Strokes can be caused either by bleeding in the brain, which is classified as a hemorrhagic stroke, or by a lack of blood flow to the brain, which is classified as an ischemic stroke. Medical literature attributes many causes for a stroke: high blood pressure, cholesterol levels, and cardiovascular diseases can increase the risk of a stroke, most often by causing blood clots that may dislodge and then block blood vessels. Other conditions, such as diabetes, smoking, aneurysms, inflammation, and comorbidities may increase either the risk of having a stroke or the severity of it.

In this dataset is recorded several covariates:

Our main goal will be to correlate stroke with the other variables to asses them as risk factors.

1) gender: “Male”, “Female” or “Other”
2) age: age of the patient
3) hypertension: 0 if the patient doesn’t have hypertension, 1 if the patient has hypertension
4) heart disease: 0 if the patient doesn’t have any heart diseases, 1 if the patient has a heart disease
5) ever-married: “No” or “Yes”
6) worktype: “children”, “Govtjov”, “Neverworked”, “Private” or “Self-employed”
7) Residencetype: “Rural” or “Urban”
8) avgglucoselevel: average glucose level in blood
9) bmi: body mass index
10) smoking_status: “formerly smoked”, “never smoked”, “smokes” or “Unknown”*
11) stroke: 1 if the patient had a stroke or 0 if not

The presence of many categorical and continuous covariates poses a challenge, and we will make a note to be wary of confounders and paradoxes, such as we will soon find in the relationship between Stroke, Age, and BMI. We will investigate such interactions with models and plots.

Exploration

First steps

The first step to understanding a (small enough) dataset is to view the covariates individually. With some quick summary statistics, we can get an idea of what we are looking at:

```
summary(data)
```

```
##     gender               age         hypertension     heart_disease
##  Length:4981        Min.   : 0.08   Min.   :0.00000   Min.   :0.00000
##  Class :character   1st Qu.:25.00   1st Qu.:0.00000   1st Qu.:0.00000
##  Mode  :character   Median :45.00   Median :0.00000   Median :0.00000
##                     Mean   :43.42   Mean   :0.09617   Mean   :0.05521
##                     3rd Qu.:61.00   3rd Qu.:0.00000   3rd Qu.:0.00000
##                     Max.   :82.00   Max.   :1.00000   Max.   :1.00000
##  ever_married        work_type         Residence_type     avg_glucose_level
##  Length:4981        Length:4981        Length:4981        Min.   : 55.12
##  Class :character   Class :character   Class :character   1st Qu.: 77.23
##  Mode  :character   Mode  :character   Mode  :character   Median : 91.85
##                                                           Mean   :105.94
##                                                           3rd Qu.:113.86
##                                                           Max.   :271.74
##       bmi        smoking_status         stroke
##  Min.   :14.0   Length:4981        Min.   :0.00000
##  1st Qu.:23.7   Class :character   1st Qu.:0.00000
##  Median :28.1   Mode  :character   Median :0.00000
##  Mean   :28.5                      Mean   :0.04979
##  3rd Qu.:32.6                      3rd Qu.:0.00000
##  Max.   :48.9                      Max.   :1.00000
```

```
data %>% group_by(smoking_status) %>% tally()
```

```
## # A tibble: 4 x 2
##   smoking_status      n
##   <chr>           <int>
## 1 formerly smoked   867
## 2 never smoked     1838
## 3 smokes            776
## 4 Unknown          1500
```

```
data %>% group_by(ever_married) %>% tally()
```

```
## # A tibble: 2 x 2
##   ever_married     n
##   <chr>        <int>
## 1 No            1701
## 2 Yes           3280
```

```
data %>% group_by(gender) %>% tally()
```

```
## # A tibble: 2 x 2
##   gender     n
##   <chr>  <int>
## 1 Female  2907
## 2 Male    2074
```

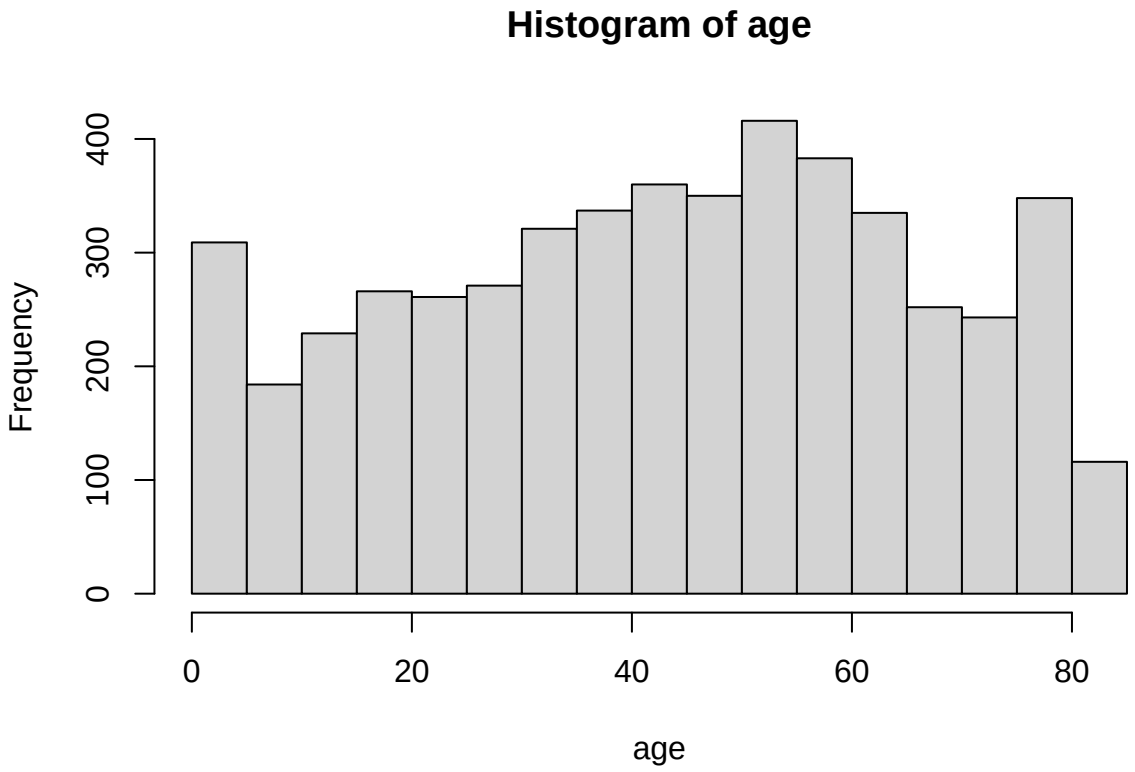
```
data %>% group_by(work_type) %>% tally()
```

```
## # A tibble: 4 x 2
##   work_type         n
##   <chr>         <int>
## 1 children        673
## 2 Govt_job        644
## 3 Private        2860
## 4 Self-employed   804
```

```
data %>% group_by(Residence_type) %>% tally()
```

```
## # A tibble: 2 x 2
##   Residence_type     n
##   <chr>          <int>
## 1 Rural           2449
## 2 Urban           2532
```

```
hist(age)
```



As expected, the value of stroke is either 0 or 1. Age, somewhat surprisingly, varies all the way from 0.08 to 82. It is clear from the dataset the low ages are not misinputs, so we may assume that our study includes data about very young patients. BMI varies from 14 to 48, reasonable values for a human dataset. We may also see that there are a large amount of unknown smoking statuses (about 1/3ish of the patients), both heart disease and hypertension have relatively low occurence (less than 10%), most patients were married, and there are more female patients than male. Literature suggests that stroke is more likely in young male patients, but the longer life expectancy of female subjects creates a survival bias that inflates the prevalence among older female patients. It remains to be seen if this is observered in our dataset. Furthermore, the ages

are roughly evenly distributed, with no especially large tendency towards young or old.

It was not mentioned how average blood glucose levels were measured, but it is most likely with an A1c screening. Thus, we might try to use average blood glucose levels as a stand in for diabetes. We may use average blood glucose as a continuous covariate to preserve accuracy, or we may attempt to stratify into normal (<117 mg/dL), prediabetic (117-137 mg/dl), and diabetic (>137 mg/dL) for ease of interpretation.

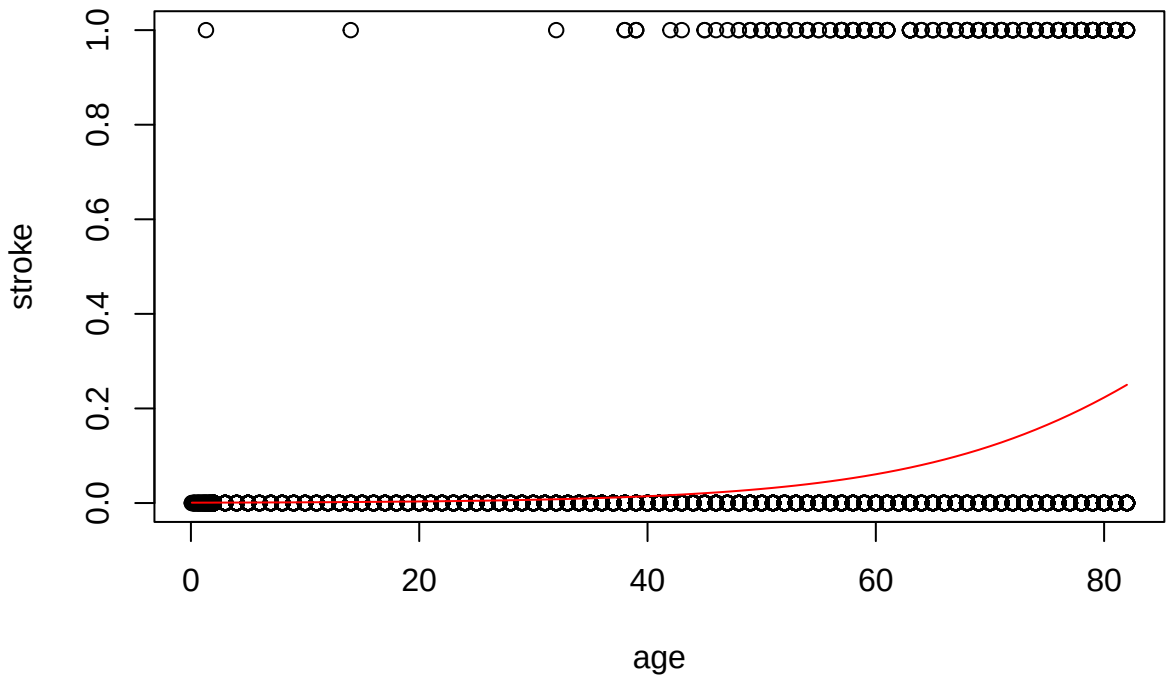
Some plots

Now that we've gotten our bearings a little, we can ask the plots for the simplest questions. Firstly, how does stroke risk vary with our continuous covariates?

```
plot(stroke~age)

fitage = glm(stroke ~ age, family = binomial, data = data)
smm = summary(fitage)
smm$coefficients
```

```
##                Estimate  Std. Error   z value      Pr(>|z|)
## (Intercept) -7.20439617 0.336973944 -21.37968 2.065266e-101
## age          0.07446153 0.004945627  15.05603  3.151363e-51
```

```
lines(sort(fitted.values(fitage))~sort(age), type = "l", col = "red")
```



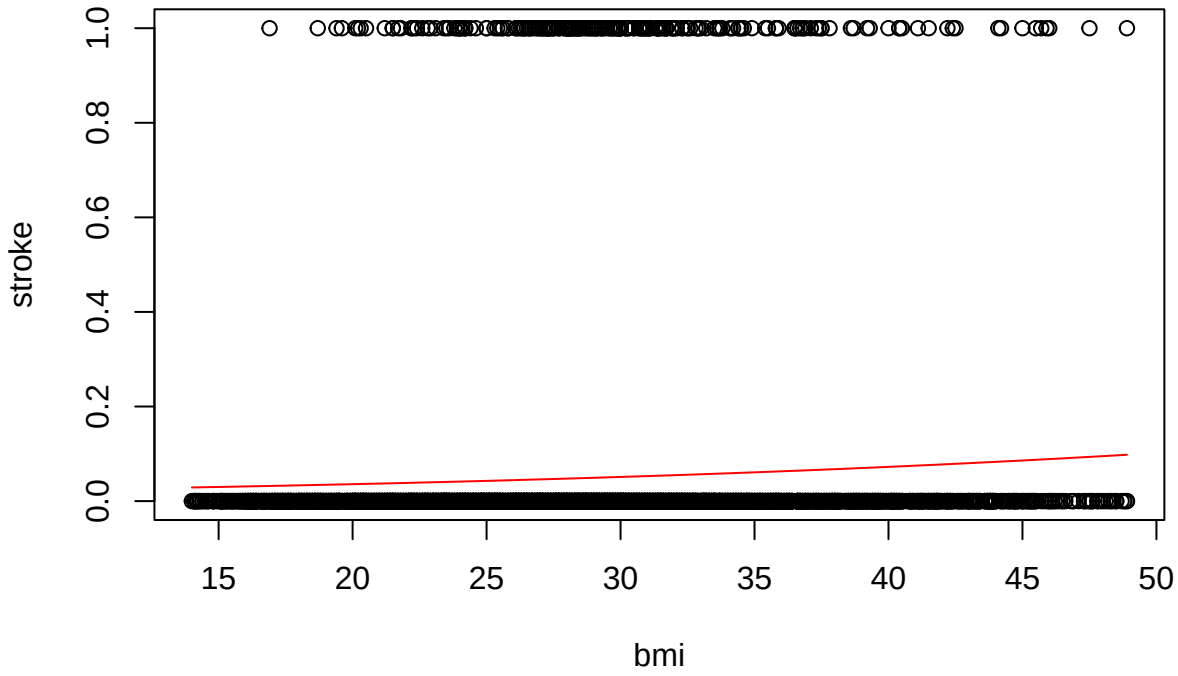
In the above plot (code can be found in the EDAvisual.R file), we plot the occurrences of stroke/no stroke against age. Overlaid in a red line is a glm, fitting risk of stroke against age. The glm estimates an 0.075 increase in odds per year increase in age. Furthermore, the null over residual deviance suggests that the fit is relatively appropriate, and we do not worry yet about zero inflations or other potential issues quite yet. First, we investigate the other bivariate cases to satisfy our intuition.

Since we have observed that age increases risk, what about BMI?

```
plot(stroke~bmi)

fitage = glm(stroke ~ bmi, family = binomial, data = data)
summary(fitage)$coefficients
```

```
##                Estimate  Std. Error    z value     Pr(>|z|)
## (Intercept) -4.03719414 0.286754961 -14.078899 5.120123e-45
## bmi          0.03716173 0.009281048   4.004045 6.226863e-05
```

```
lines(sort(fitted.values(fitage))~sort(bmi), type = "l", col = "red")
```

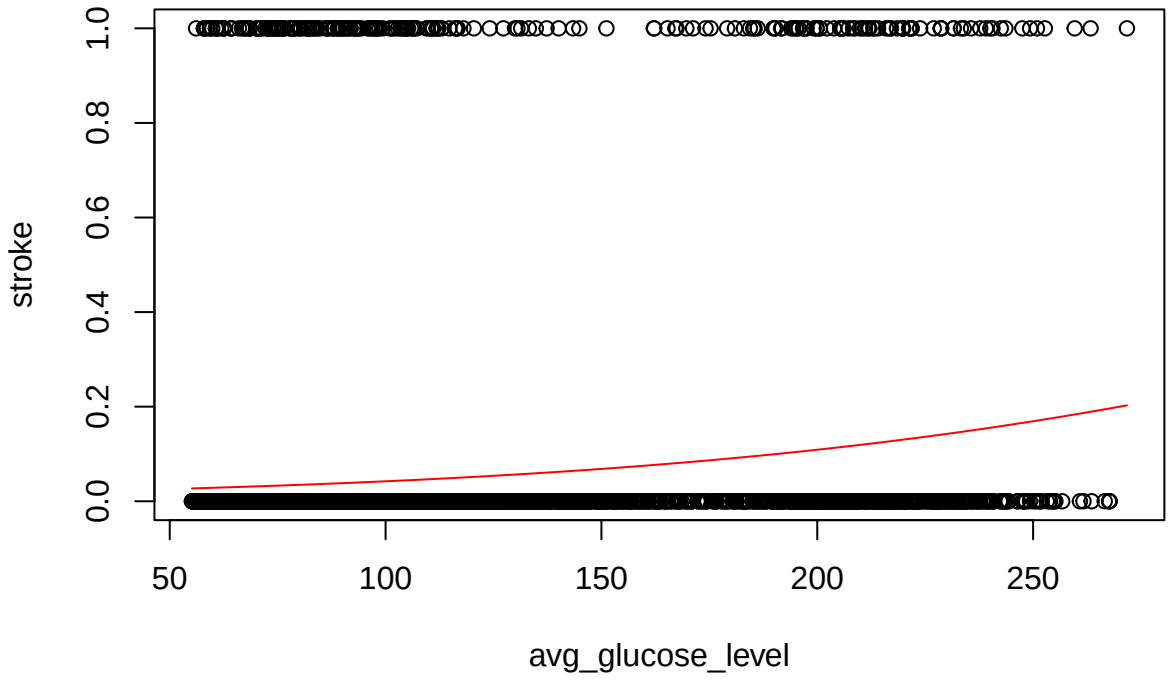


In the above plot, we now see that BMI is associated with an increasing risk of stroke. Furthermore, we will a similar association in average blood glucose levels below:

```
plot(stroke~avg_glucose_level)

fitage = glm(stroke ~ avg_glucose_level, family = binomial, data = data)
summary(fitage)$coefficients
```

```
##                     Estimate  Std. Error    z value      Pr(>|z|)
## (Intercept)       -4.14304948 0.160825618 -25.761129 2.419073e-146
## avg_glucose_level  0.01020692 0.001132909   9.009477  2.070407e-19
```

```
lines(sort(fitted.values(fitage))~sort(avg_glucose_level), type = "l", col = "red")
```



Now we may move on to the categorical covariates: these may be investigated with a table.

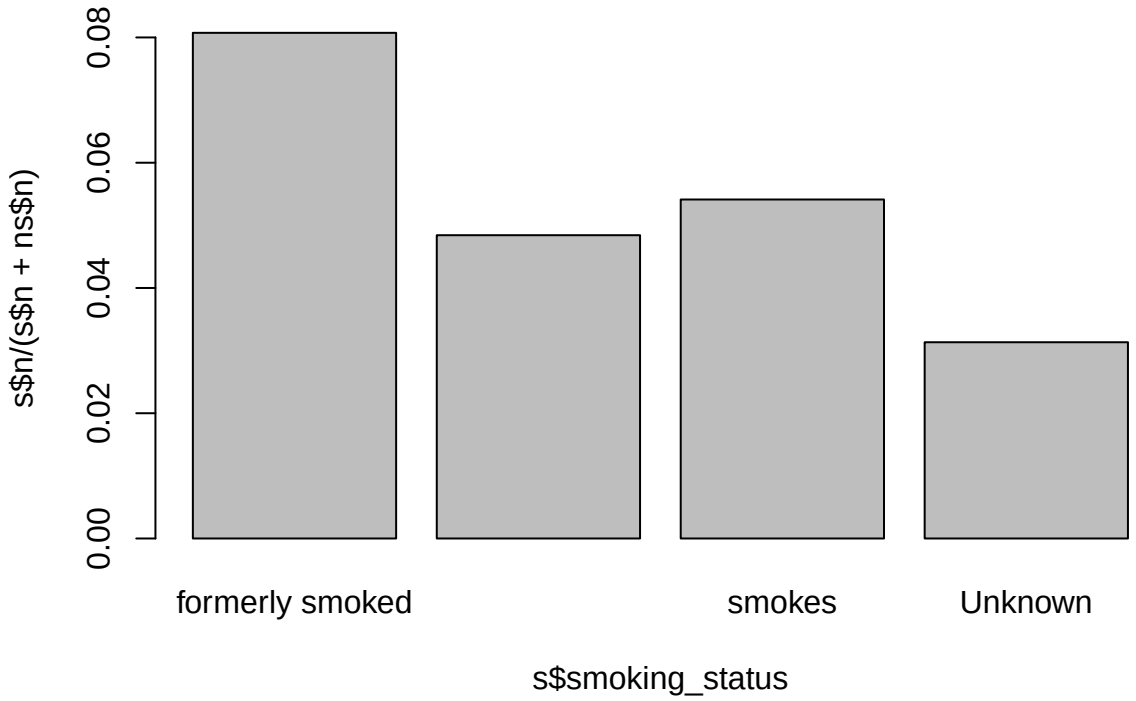
```
tab = data %>% group_by(smoking_status,stroke) %>% tally()
s = tab %>% filter(stroke == 1)
ns = tab %>% filter(stroke == 0)

s
```

```
## # A tibble: 4 x 3
## # Groups:   smoking_status [4]
##   smoking_status  stroke     n
##   <chr>            <int> <int>
## 1 formerly smoked      1    70
## 2 never smoked         1    89
## 3 smokes               1    42
## 4 Unknown              1    47
```

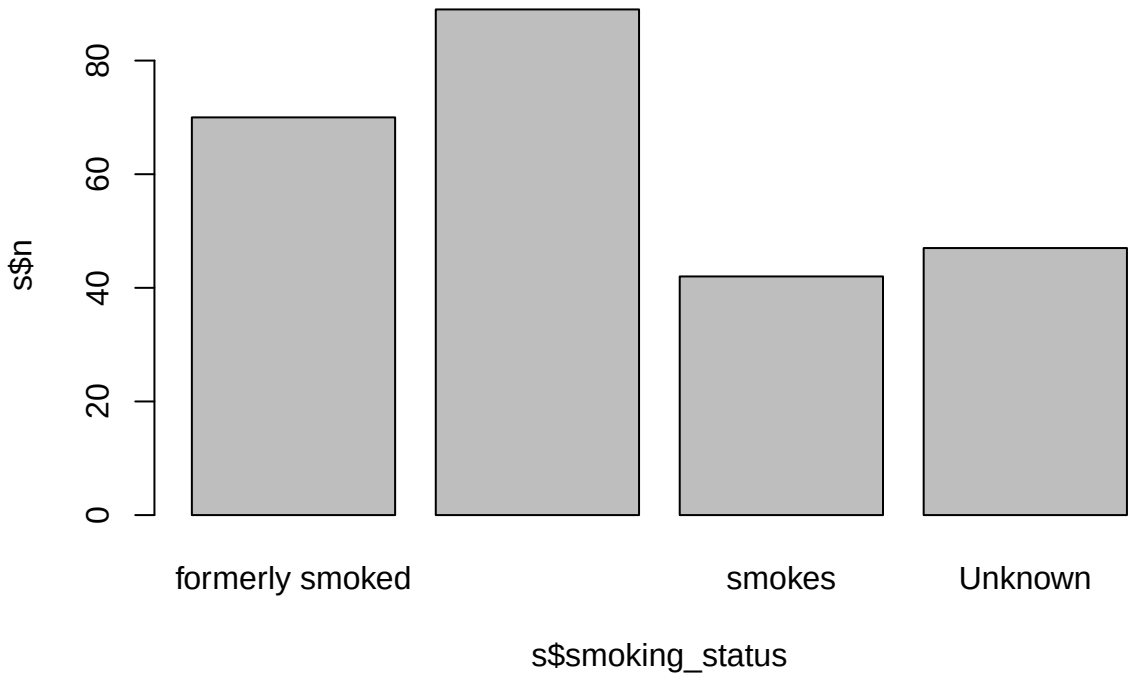
```
ns
```

```
## # A tibble: 4 x 3
## # Groups:   smoking_status [4]
##   smoking_status  stroke     n
##   <chr>            <int> <int>
## 1 formerly smoked      0   797
## 2 never smoked         0  1749
## 3 smokes               0   734
## 4 Unknown              0  1453
```

```
props = s$n/(s$n+ns$n)
barplot(s$n/(s$n+ns$n)~s$smoking_status)
```



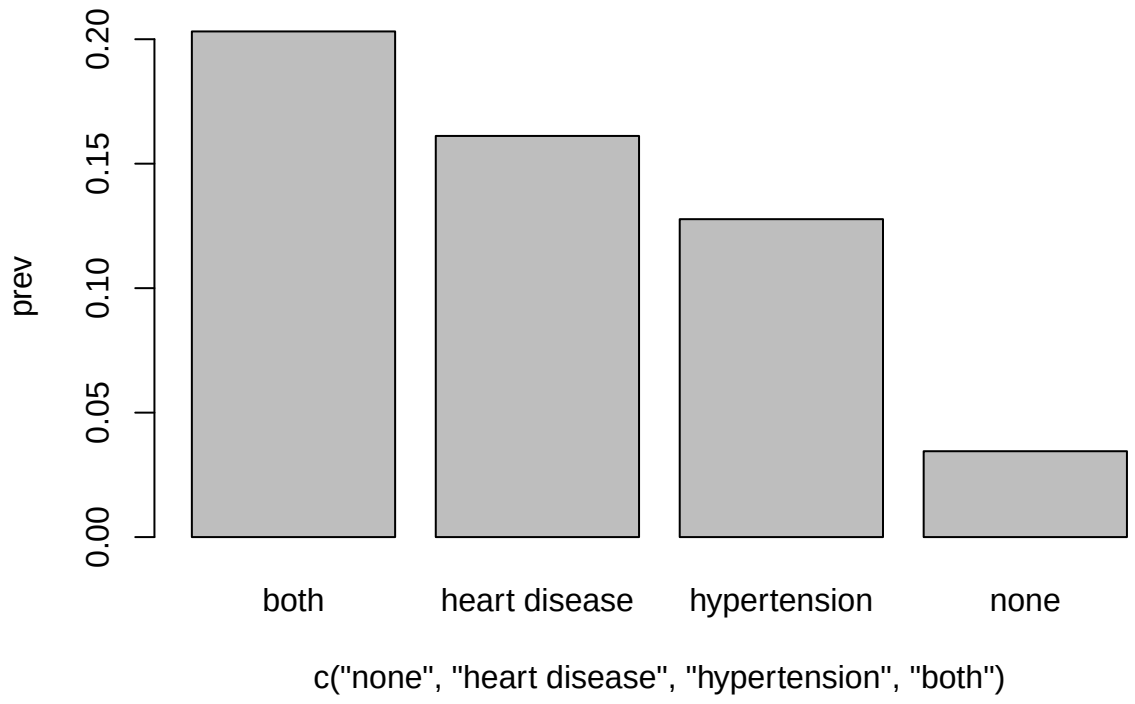

```
barplot(s$n~s$smoking_status)
```



What a strange trend! The prevalence of stroke among never smoked and smokes is similar, while the prevalence in former smokers is much higher! It is biologically unlikely that smoking then stopping has a special ability to prevent strokes. It is more likely that there is some kind of multiple dependence or sample bias: maybe those who formerly smoked stopped because they experienced a health complication, or those who had the time to smoke then stop tended to be older. There is no clear way to interpret such bias without sampling more data, so we are stuck with only speculation.

Moving on, we may next take a look at hypertension and heart disease. Since both of these are strongly medically related, I will plot them with interaction.

```
tab = data %>% group_by(hypertension,heart_disease,stroke) %>% tally()
sp = tab %>% filter(stroke == 1)
sn = tab %>% filter(stroke == 0)
prev = sp$n/(sp$n+sn$n)
barplot(prev~c("none",
               "heart disease",
               "hypertension",
               "both"))
```




```
tab
```

```
## # A tibble: 8 x 4
## # Groups:   hypertension, heart_disease [4]
##   hypertension heart_disease stroke     n
##          <int>         <int>  <int> <int>
## 1            0             0      0  4143
## 2            0             0      1   148
## 3            0             1      0   177
## 4            0             1      1    34
## 5            1             0      0   362
## 6            1             0      1    53
## 7            1             1      0    51
## 8            1             1      1    13
```

```
fit = glm(stroke ~ -1+I((!hypertension)*(!heart_disease))+I(hypertension*(!heart_disease))+I((!hyperten
summary(fit)$coefficients
```

```
##                                       Estimate Std. Error    z value
## I((!hypertension) * (!heart_disease)) -3.331963 0.08365478 -39.829921
## I(hypertension * (!heart_disease))    -1.921352 0.14707262 -13.063970
## I((!hypertension) * (heart_disease))  -1.649789 0.18724712  -8.810759
## I(hypertension * heart_disease)       -1.366876 0.31069425  -4.399426
##                                          Pr(>|z|)
## I((!hypertension) * (!heart_disease)) 0.000000e+00
## I(hypertension * (!heart_disease))    5.289568e-39
## I((!hypertension) * (heart_disease))  1.243013e-18
```

```
## I(hypertension * heart_disease)       1.085378e-05
```

```
matrix(prev,2)
```

```
##            [,1]      [,2]
## [1,] 0.03449079 0.1277108
## [2,] 0.16113744 0.2031250
```

```
matrix(prev,2) %>% chisq.test()
```

```
## Warning in chisq.test(.): Chi-squared approximation may be incorrect
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  .
## X-squared = 2.2541e-32, df = 1, p-value = 1
```

```
fit = glm(heart_disease~hypertension, family = binomial, data = data)
summary(fit)
```

```
##
## Call:
## glm(formula = heart_disease ~ hypertension, family = binomial,
##     data = data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.5356  -0.3098  -0.3098  -0.3098   2.4740
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -3.01242    0.07051 -42.724  < 2e-16 ***
## hypertension  1.14302    0.15168   7.536 4.85e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2127.7  on 4980  degrees of freedom
## Residual deviance: 2080.1  on 4979  degrees of freedom
## AIC: 2084.1
##
## Number of Fisher Scoring iterations: 5
```

From the bar plot, we can see that those with both heart disease and hypertension have the most risk. Then, in descending order, heart disease only, hypertension only, and none. The model is kind enough to tell us that these differences are significant. We can also use a simple glm to determine that those with hypertension have a significantly higher risk of also having heart disease, indicating that these two covariates are indeed related.

Now that we have some intuition about what to expect, we can move on to all the covariates in one model.

```
fit = glm(stroke ~ ., family = binomial, data = data)
summary(fit)$coefficients
```

```
##                               Estimate  Std. Error     z value     Pr(>|z|)
## (Intercept)               -6.954558530 0.793641363 -8.76284788 1.903785e-18
## genderMale                 0.007036884 0.142195582  0.04948736 9.605309e-01
```

```
## age                          0.075149519 0.005870475 12.80126735 1.612971e-37
## hypertension                 0.416767410 0.165174120  2.52320043 1.162921e-02
## heart_disease                0.272296918 0.191117328  1.42476311 1.542257e-01
## ever_marriedYes             -0.193136036 0.225784690 -0.85539917 3.923302e-01
## work_typeGovt_job           -1.028636493 0.837738674 -1.22787275 2.194947e-01
## work_typePrivate            -0.907777857 0.822692135 -1.10342353 2.698433e-01
## work_typeSelf-employed      -1.270196189 0.843182619 -1.50643071 1.319566e-01
## Residence_typeUrban          0.087944655 0.138818250  0.63352372 5.263917e-01
## avg_glucose_level            0.003812709 0.001207845  3.15662172 1.596083e-03
## bmi                          0.010867746 0.012625631  0.86076856 3.893655e-01
## smoking_statusnever smoked -0.224347912 0.176587566 -1.27046268 2.039199e-01
## smoking_statussmokes         0.111463650 0.215515157  0.51719634 6.050191e-01
## smoking_statusUnknown       -0.066745558 0.208598611 -0.31997125 7.489901e-01
```

```
#step(fit, direction = "both")
# y = as.matrix(stroke)
# x = cbind(
#   data$gender == "Female",
#   data$age,
#   data$hypertension,
#   data$heart_disease,
#   data$ever_married,
#   data$avg_glucose_level,
#   data$bmi,
#   data$smoking_status
# ) %>% as.matrix()
# fit = glmnet(x,y)
# plot(fit)
```

In our model, we have some surprising results: gender, heart disease, smoking status, and bmi are no longer significant! This may be due to some form of multicollinearity between it and the significant variables, but we would suspect these covariates to be significant regardless of multiple dependence. Gender, for example, is suggested to be a significant risk factor, but only depending on age. Smoking status was also supposed to be a risk factor, and intuitively should not be completely dependent on the others as we do not have a variable for heart disease. We also tried multiple variable selection techniques, such as forwards/backwards/bidirectional stepwise selection and LASSO. These did not yield any further insight, and they will be relegated to the EDA.r file. So, lets investigate further.

Our next model will be a generalized additive model. This type of model allows for easy semiparametric multivariate modelling because of the additive treatment of the multiple dimensions univariately as well as capability for multivariate smoothing, such as with tensor product splines. We can then test for nonlinear relationships and for interactions. First, we establish a base model with only the significant covariates from the full model as well as BMI, which we suspect to be relevant:

```
fit = glm(stroke ~ age+hypertension+avg_glucose_level+bmi, family = binomial, data = data)
summary(fit)$coefficients
```

```
##                       Estimate  Std. Error     z value     Pr(>|z|)
## (Intercept)       -7.794692407 0.529829369 -14.7117032 5.422525e-49
## age                0.070607319 0.005176366  13.6403271 2.305242e-42
## hypertension       0.399041328 0.163079901   2.4469068 1.440881e-02
## avg_glucose_level  0.004176358 0.001184467   3.5259375 4.219866e-04
## bmi                0.008166512 0.012382875   0.6595005 5.095744e-01
```

A side-note about semi-parametric generalized additive models:

In case you're not familiar with the tensor product splines and the SS-ANOVA in spline models, special multivariate models called "tensor product splines" can be constructed from the Reproducing Kernel Hilbert Space point of view. The details are beyond the scope of this report, but the punchline is that the model space can be broken up into a collection of orthogonal subspaces. The fact that these subspaces are orthogonal allows for an ANOVA based of deviances explained, so that smooth terms in the model can be "significance tested." The unpenalized subspace, or the parametric part, is allowed to vary completely freely to minimize the objective function. In practice, these are very rigid and interpretable subspaces, such as the subspace of linear models. The non-parametric part, or smooth term, is a shape-fitting regression that can adjust to any shape, but is prevented from overfitting by a penalty functional. The classical non-parametric regression is the cubic smoothing spline. Popular nowadays are Gaussian Process Regressions and Regression Splines. The splines implemented by the MGCV package are not "true" RKHS regressions, but rather regression splines with automatically chosen knot points. Under nice enough data, these approximate the kernel regressions fitted by smoothing splines with much better computational performance. More advanced smoothing splines, such as arbitrary kernel regressions or semiparametric mixed effect models, would be better fit by a package like GSS or ASSIST.

Back to the GAM

When we fit the GAM below, we include the following things: linear terms for age, bmi, hypertension, and average glucose level. Then we include smooth terms for age, bmi, and the interaction between age and bmi.

```
fit = gam(stroke ~ age+bmi+hypertension+avg_glucose_level+ti(age)+ti(bmi)+ti(age,bmi), family = binomial
summary(fit)$p.table
```

```
##                       Estimate  Std. Error    z value     Pr(>|z|)
## (Intercept)        0.000000000 0.000000000        NaN          NaN
## age                0.075000521 0.007332206  10.228916 1.471572e-24
## bmi               -0.279685160 0.016791754 -16.656102 2.732646e-62
## hypertension       0.397401027 0.162903010   2.439495 1.470782e-02
## avg_glucose_level  0.004178781 0.001185224   3.525731 4.223152e-04
```

```
summary(fit)$s.table
```

```
##                    edf    Ref.df       Chi.sq    p-value
## ti(age)     0.1596484 0.2862281   0.05878687 0.80842452
## ti(bmi)     3.1837983 3.6391825 102.86146769 0.00000000
## ti(age,bmi) 2.8100385 3.6934406   8.41877968 0.05768213
```

```
#gam.check(fit)
```

To read the above summary, we note that BMI's linear term has become significant and the interaction term "ti(age,bmi)" is significant! This indicates that there is some kind of an interaction between age and bmi in how they predict the risk of stroke. Before we move on, because of the fickle nature of non-parametric models, it is prudent to check with a couple extra models just to be sure it is not a fluke. Below, we fit a bivariate cubic spline, a main-effect adjusted bivariate cubic spline, and a gaussian process smooth.

```
fit = gam(stroke ~ age+bmi+hypertension+avg_glucose_level+s(age,bmi), family = binomial, data = data)
summary(fit)$s.table
```

```
##                  edf   Ref.df   Chi.sq p-value
## s(age,bmi) 6.034942 8.641374 192.2943       0
```

```
fit = gam(stroke ~ age+bmi+hypertension+avg_glucose_level+ti(age)+ti(bmi)+s(age,bmi), family = binomial
summary(fit)$s.table
```

```
##                  edf    Ref.df    Chi.sq    p-value
```

```
## ti(age)    2.302803  2.544600  3.611168 0.23902136
## ti(bmi)    3.094177  3.557945 46.475043 0.00000000
## s(age,bmi) 2.997936 27.000000  6.778960 0.01409138
```

```
fit = gam(stroke ~ age+bmi+hypertension+avg_glucose_level+ti(age)+ti(bmi)+s(age,bmi, bs = "gp"), family
summary(fit)$s.table
```

```
##                 edf    Ref.df   Chi.sq    p-value
## ti(age)    2.251027  2.488272 3.191503 0.29795986
## ti(bmi)    3.082044  3.551451 7.033040 0.06937026
## s(age,bmi) 2.776575 30.000000 6.707165 0.01364100
```

Because they all

Variance by Components