Diabetes Risk

Russell Almond

2023-12-07

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.3     v tibble    3.2.1
## v lubridate 1.9.2     v tidyr     1.3.0
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

Diabetes Risk Data

These data come from the UC Irvine Machine Learning Repository. There are related to patients in a certain diabetes hospital in Bangledesh. It is not explicitly stated, but it probably refers to Type II diabetes (as Type I is usually identified in childhood).

We will look at the risk of diabetes using three variables: `Age`, `Sex` (the phenotype not the gender expression) and `Obesity`.

Read and clean data

```
read_csv("diabetes_data_upload.csv") |>
  mutate(Diabetic=as.factor(class),Sex=factor(Gender),Obese=factor(Obesity)) |>
  select(all_of(c("Age","Sex","Obese","Diabetic"))) ->
  DData
```

```
## Rows: 520 Columns: 17
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (16): Gender, Polyuria, Polydipsia, sudden weight loss, weakness, Polyph...
## dbl  (1): Age
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
summary(DData)
```

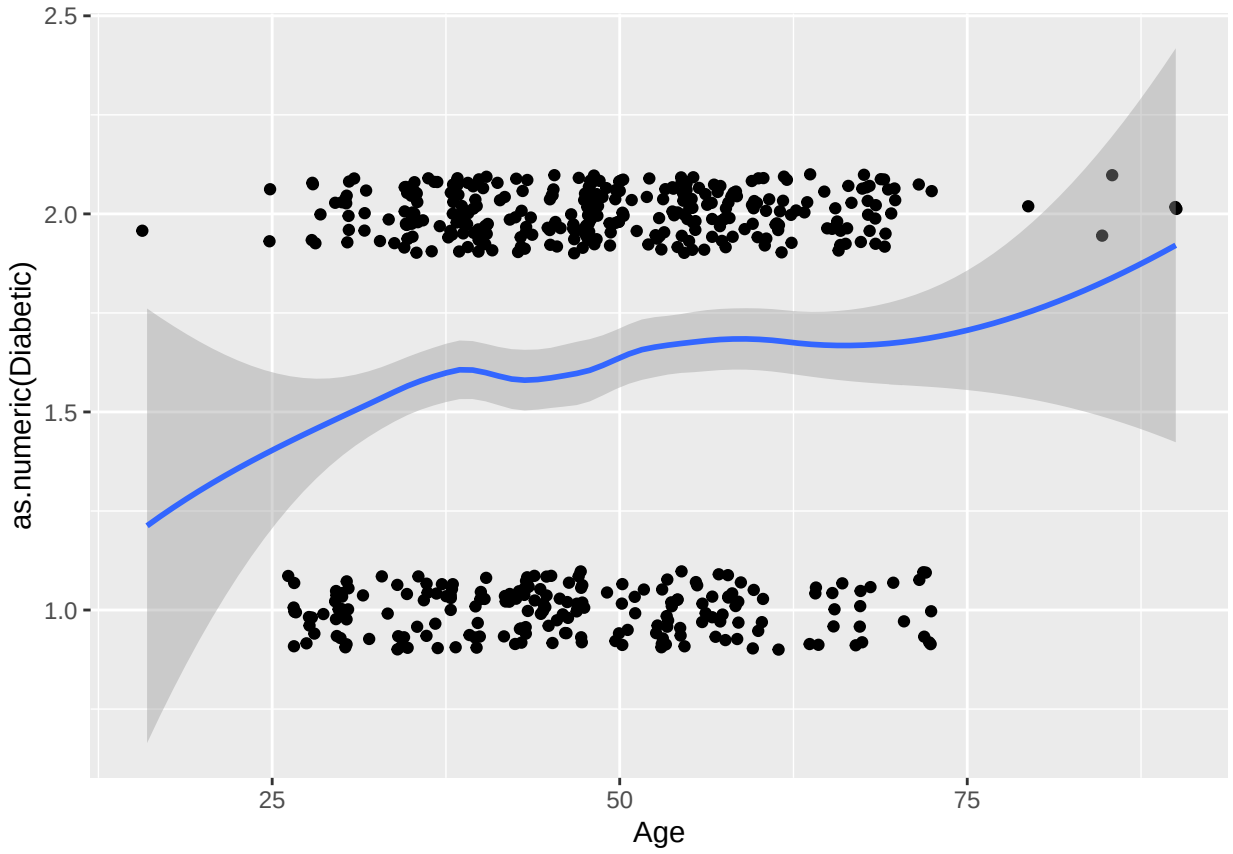
```
##       Age            Sex      Obese         Diabetic
##  Min.   :16.00   Female:192   No :432   Negative:200
##  1st Qu.:39.00   Male  :328   Yes: 88   Positive:320
```

```
##  Median :47.50
##  Mean   :48.03
##  3rd Qu.:57.00
##  Max.   :90.00
```

Exploratory Analyses

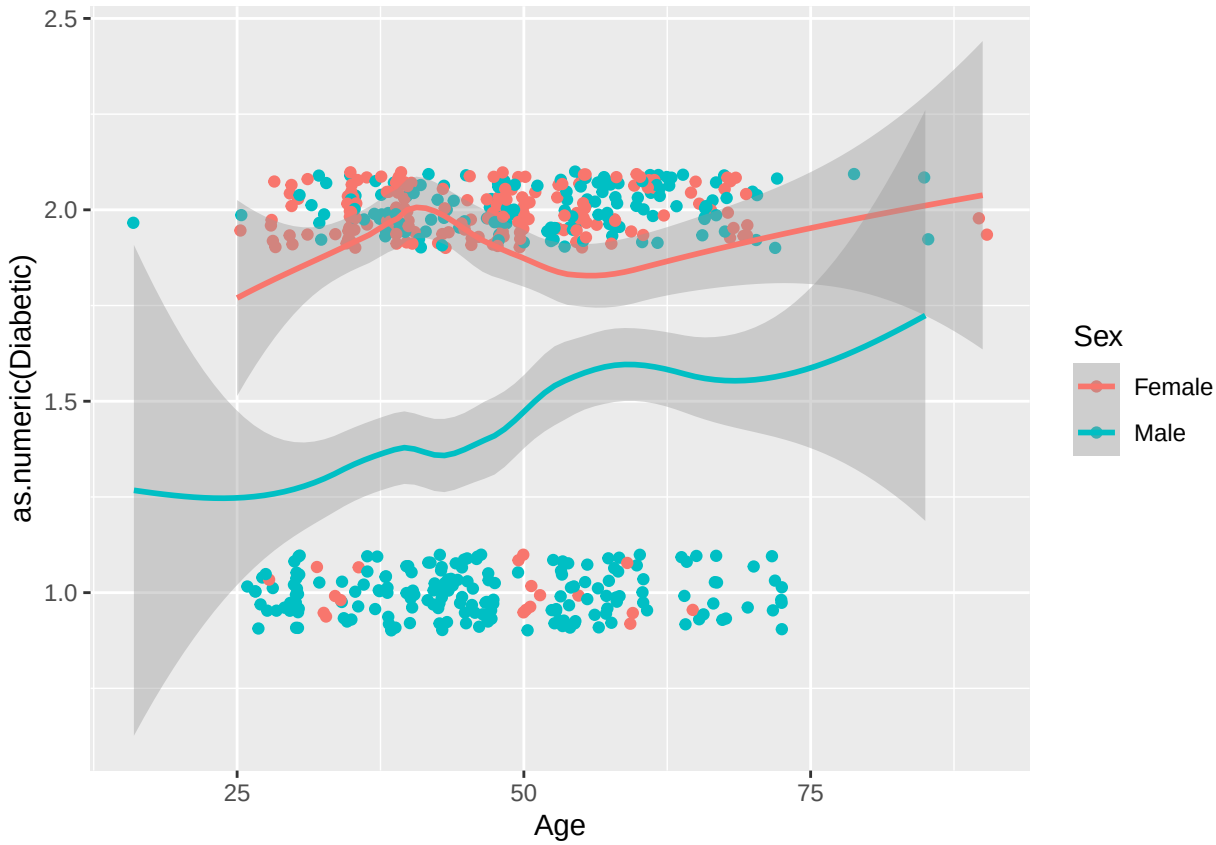
```
ggplot(DData,aes(x=Age,y=as.numeric(Diabetic))) +
  geom_point(position=position_jitter(width=.5,height = .1)) +
  geom_smooth()
```

```
## `geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```



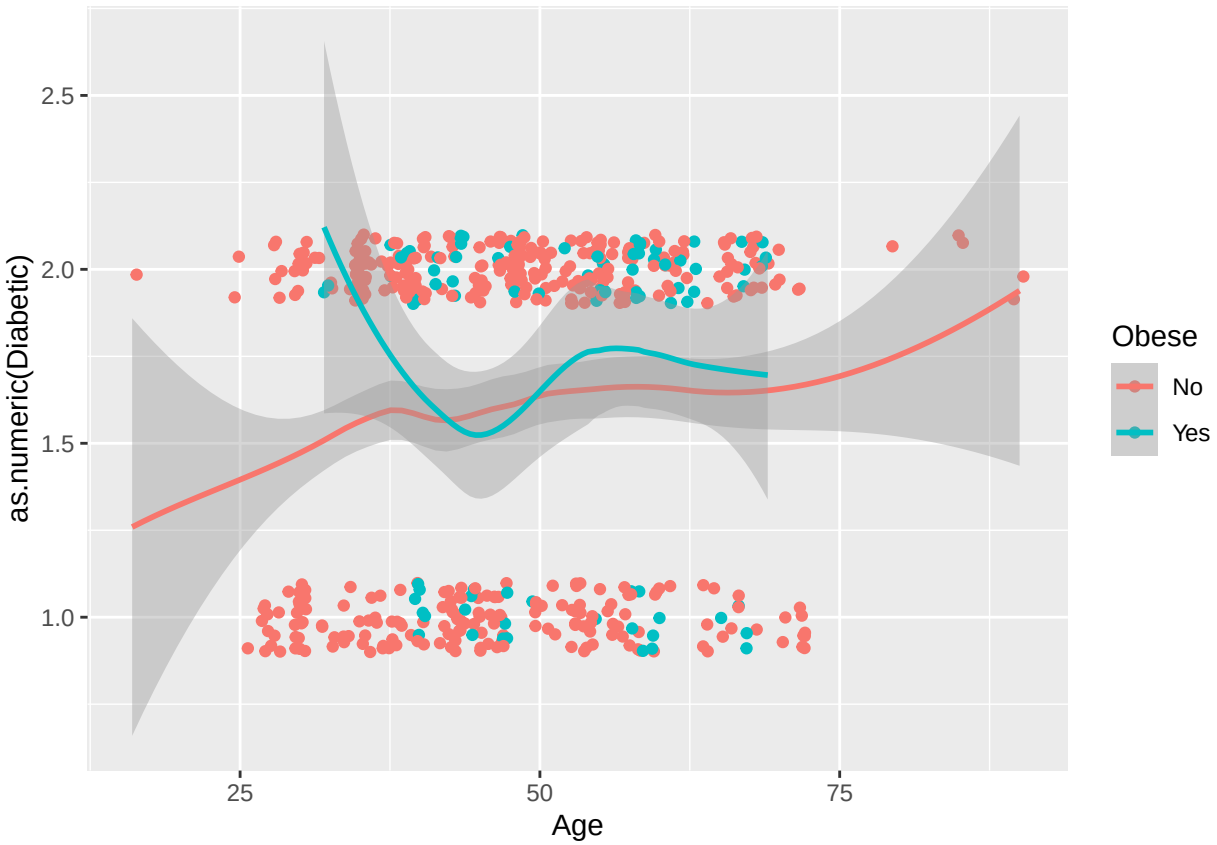

```
ggplot(DData,aes(x=Age,y=as.numeric(Diabetic),color=Sex)) +
  geom_point(position=position_jitter(width=.5,height = .1)) +
  geom_smooth()
```

```
## `geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```




```
ggplot(DData,aes(x=Age,y=as.numeric(Diabetic),color=Obese)) +
  geom_point(position=position_jitter(width=.5,height = .1)) +
  geom_smooth()
```

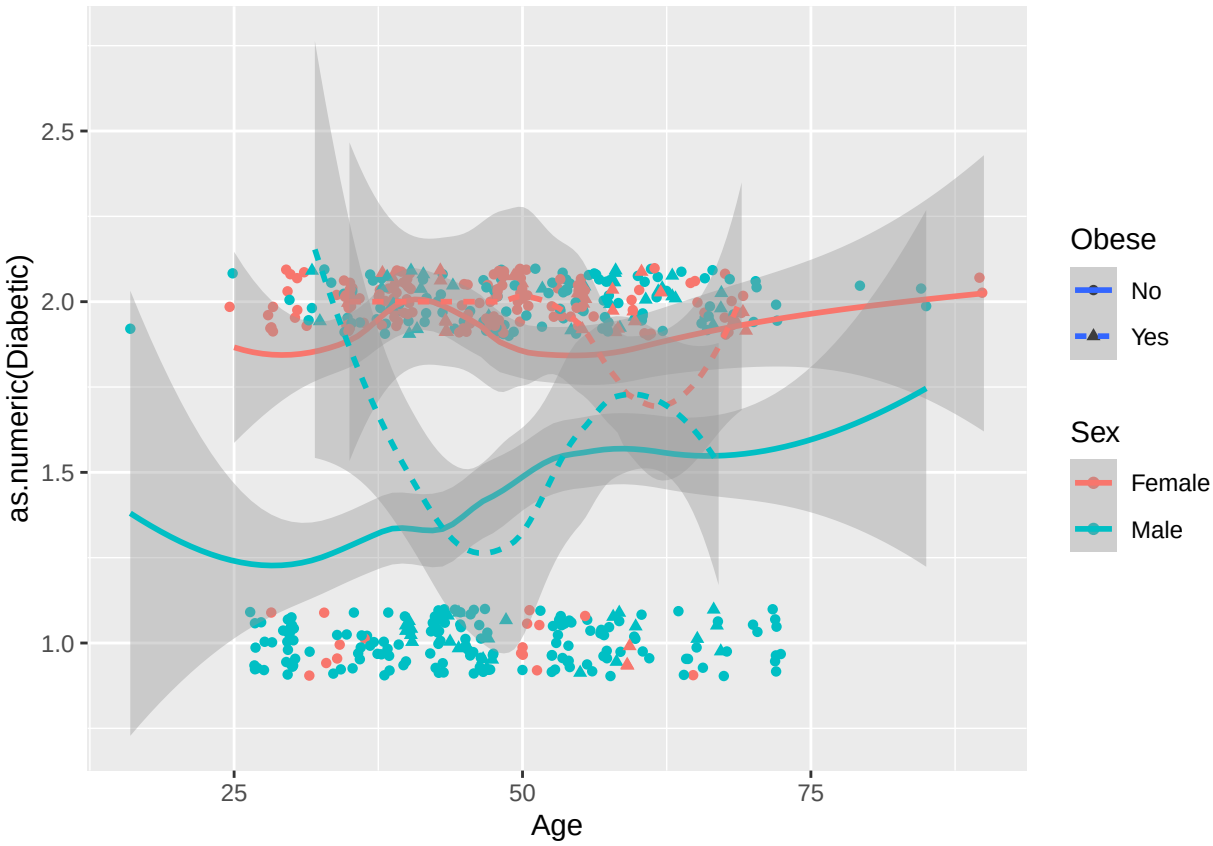
```
## `geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```



Only two different variables, so we can get them all on the same plot: (If it is not clear from the legend, the dashed line is `Obese==Yes`.

```
ggplot(DData,aes(x=Age,y=as.numeric(Diabetic),color=Sex,shape=Obese,linetype=Obese)) +
  geom_point(position=position_jitter(width=.5,height = .1)) +
  geom_smooth()
```

```
## `geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```



First model

```
glmAge <- glm(Diabetic~Age,data=DData,family=binomial())
summary(glmAge)
```

```
##
## Call:
## glm(formula = Diabetic ~ Age, family = binomial(), data = DData)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.425097   0.371776  -1.143   0.2529
## Age          0.018766   0.007613   2.465   0.0137 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 692.93  on 519  degrees of freedom
## Residual deviance: 686.72  on 518  degrees of freedom
## AIC: 690.72
##
## Number of Fisher Scoring iterations: 4
```

I'm going to do this a lot, so might as well write a function.

```r
APAsum <- function (model) {
  df <- model$df.null-model$df.residual
  X2 <- model$null.deviance- model$deviance
  p <- 1-pchisq(X2,df)
  paste("X^2(",df,") =", round(X2,2),
        ifelse(p<.001, ", p < .001", paste(", p =",round(p,3))))
}
cat(APAsum(glmAge),"\n")
```

```
## X^2( 1 ) = 6.21 , p = 0.013
```

Now Add Sex

I'll directly fit the interaction model and maybe simplify it later.

```r
glmAgeSex <- glm(Diabetic~Age*Sex,data=DData,family=binomial())
summary(glmAgeSex)
```

```
##
## Call:
## glm(formula = Diabetic ~ Age * Sex, family = binomial(), data = DData)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.936700   0.998587   1.939 0.052448 .
## Age          0.005826   0.020884   0.279 0.780255
## SexMale     -3.813706   1.107197  -3.444 0.000572 ***
## Age:SexMale  0.028341   0.022932   1.236 0.216519
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 692.93  on 519  degrees of freedom
## Residual deviance: 561.37  on 516  degrees of freedom
## AIC: 569.37
##
## Number of Fisher Scoring iterations: 4
```

```r
cat(APAsum(glmAgeSex),"\n")
```

```
## X^2( 3 ) = 131.56 , p < .001
```

Same trick with Obesity

```r
glmAgeFat <- glm(Diabetic~Age*Obese,data=DData,family=binomial())
summary(glmAgeFat)
```

```
##
## Call:
## glm(formula = Diabetic ~ Age * Obese, family = binomial(), data = DData)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
```

```
## (Intercept)  -0.466946   0.392607  -1.189   0.2343
## Age           0.018537   0.008146   2.276   0.0229 *
## ObeseYes      0.900835   1.276595   0.706   0.4804
## Age:ObeseYes -0.011160   0.024537  -0.455   0.6492
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 692.93  on 519  degrees of freedom
## Residual deviance: 684.75  on 516  degrees of freedom
## AIC: 692.75
##
## Number of Fisher Scoring iterations: 4
```

```
cat(APAsum(glmAgeFat),"\n")
```

```
## X^2( 3 ) = 8.18 , p = 0.042
```

Model Search

We will consider the full model including three-way interactions:

```
glmStep <- step(glmAge,list(lower=Diabetes~1,  # constant risk
                            upper=Diabetes~Age*Sex*Obese),
                trace=2)
```

```
## Start:  AIC=690.72
## Diabetic ~ Age
##
##         Df Deviance    AIC
## + Sex    1   562.84 568.84
## <none>       686.72 690.72
## + Obese  1   684.95 690.95
## - Age    1   692.93 694.93
##
## Step:  AIC=568.84
## Diabetic ~ Age + Sex
##
##           Df Deviance    AIC
## + Obese    1   560.79 568.79
## <none>         562.84 568.84
## + Age:Sex  1   561.37 569.37
## - Age      1   575.12 579.12
## - Sex      1   686.72 690.72
##
## Step:  AIC=568.79
## Diabetic ~ Age + Sex + Obese
##
##             Df Deviance    AIC
## <none>           560.79 568.79
## - Obese      1   562.84 568.84
## + Age:Obese  1   559.06 569.06
## + Age:Sex    1   559.10 569.10
## + Sex:Obese  1   560.15 570.15
## - Age        1   571.91 577.91
```

```
## - Sex        1   684.95 690.95
```

Final model

```
summary(glmStep)
```

```
##
## Call:
## glm(formula = Diabetic ~ Age + Sex + Obese, family = binomial(),
##     data = DData)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.863477   0.450533   1.917  0.05529 .
## Age          0.028425   0.008694   3.270  0.00108 **
## SexMale     -2.527321   0.272195  -9.285  < 2e-16 ***
## ObeseYes     0.399594   0.281402   1.420  0.15561
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 692.93  on 519  degrees of freedom
## Residual deviance: 560.79  on 516  degrees of freedom
## AIC: 568.79
##
## Number of Fisher Scoring iterations: 4
```

```
cat(APAsum(glmStep),"\n")
```

```
## X^2( 3 ) = 132.14 , p < .001
```

Predictions

```
risktab <- data.frame(Age=rep(15+5*(1:11),each=4),
                  Sex=factor(rep(rep(levels(DData$Sex),each=2),11)),
                  Obese=factor(rep(levels(DData$Obese),22)))
head(risktab)
```

```
##   Age    Sex Obese
## 1  20 Female    No
## 2  20 Female   Yes
## 3  20   Male    No
## 4  20   Male   Yes
## 5  25 Female    No
## 6  25 Female   Yes
```

```
risktab$Risk <- psych::logistic(predict(glmStep,risktab))
head(risktab)
```

```
##   Age    Sex Obese      Risk
## 1  20 Female    No 0.8072098
## 2  20 Female   Yes 0.8619492
## 3  20   Male    No 0.2506141
## 4  20   Male   Yes 0.3327561
```

```
## 5  25 Female    No 0.8283683
## 6  25 Female   Yes 0.8780081
```

This will be easier to look at if we turn `Sex` and `Obese` into columns. Can do this with the pivot function.

In the chart below "Yes" and "No" refer to Obsese, so "Male_Yes" is an obese male, and "Male_No" is a non-obese male.

```
riskchart <- pivot_wider(risktab,id_cols=Age,names_from=all_of(c("Sex","Obese")),
                         values_from=Risk)
knitr::kable(riskchart,digits=2)
```

Age	Female_No	Female_Yes	Male_No	Male_Yes
20	0.81	0.86	0.25	0.33
25	0.83	0.88	0.28	0.37
30	0.85	0.89	0.31	0.40
35	0.87	0.91	0.34	0.43
40	0.88	0.92	0.37	0.47
45	0.89	0.93	0.40	0.50
50	0.91	0.94	0.44	0.54
55	0.92	0.94	0.47	0.57
60	0.93	0.95	0.51	0.61
65	0.94	0.96	0.55	0.64
70	0.95	0.96	0.58	0.67

Save the data out for use in SPSS.

```
haven::write_sav(DData,"diabetesRisk.sav")
```