Exercise L5

2022-06-12

Exercise 1

Use the Pima.te diabetes dataset:

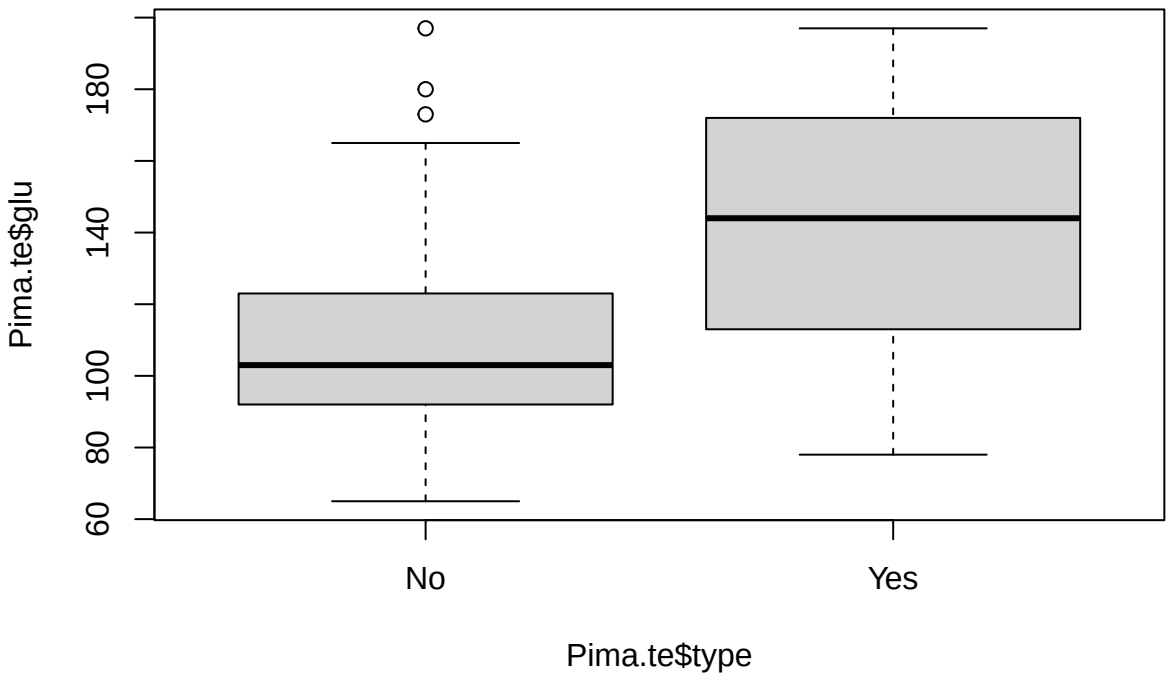
```
library(MASS)
lapply(Pima.te, class)
```

```
## $npreg
## [1] "integer"
##
## $glu
## [1] "integer"
##
## $bp
## [1] "integer"
##
## $skin
## [1] "integer"
##
## $bmi
## [1] "numeric"
##
## $ped
## [1] "numeric"
##
## $age
## [1] "integer"
##
## $type
## [1] "factor"
```

The categorical variable "type" represents "Yes or No for diabetic according to WHO criteria".

- Produce a plot suitable to answering the question "Does plasma glucose level (column glu) depend on diabetic status?

```
#diabetic status is represented by 'type' (categorical binary variable). 'glu' is a numerical variable.
plot(Pima.te$glu ~ Pima.te$type)
```




```
#we wrote this way because we cannot write 'plot(x=Pima.te$glu, y=Pima.te$type),
#as we're not dealing with two numerical variables.
```

- Use the appropriate regression to show that such dependence is significant.

```
glu_typ <- lm(Pima.te$glu ~ Pima.te$type)
summary(glu_typ)$coefficients
```

```
##                  Estimate Std. Error  t value      Pr(>|t|)
## (Intercept)     108.18834   1.747357 61.91542 1.009226e-183
## Pima.te$typeYes  33.71992   3.049562 11.05730  2.178114e-24
```

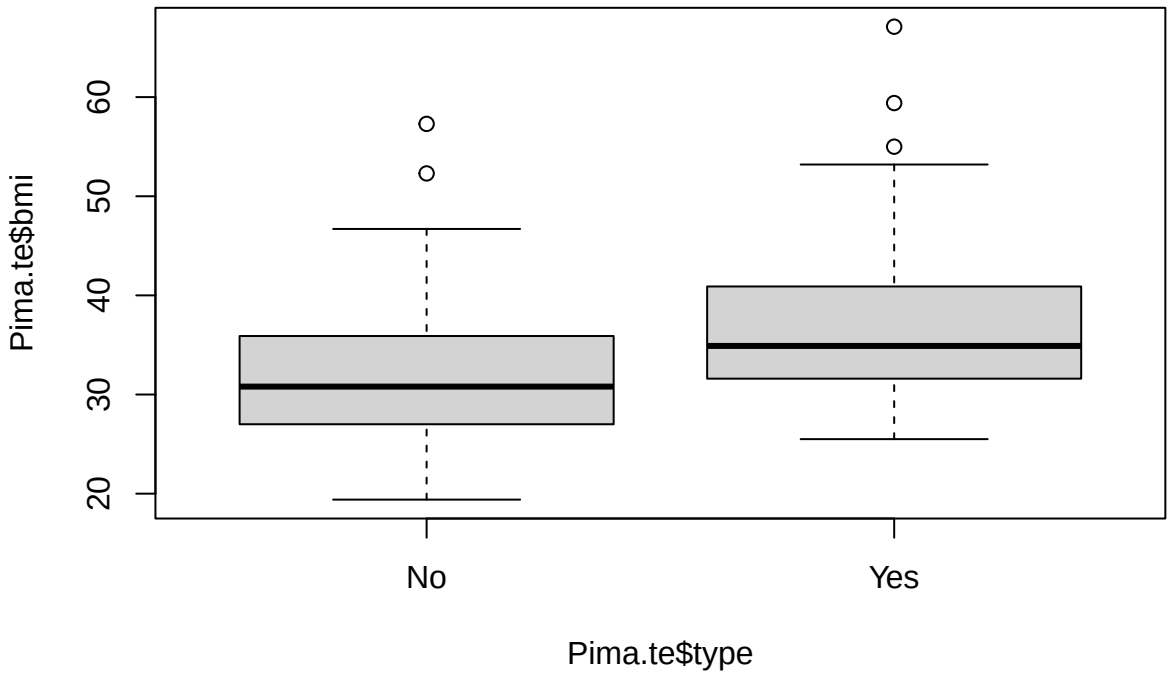
We used *linear regression* because we wanted to predict a numerical variable (glu) on a categorical one (type). From the summary we see that a diabetic status ("Yes") is associated to a 33.72 higher plasma glucose concentration than the concentration for the "No" status. Uncertainty associated to this increase is about 3.05. The associated P-value is enough small to say that this effect is not due to chance.

Exercise 2

Does BMI predict diabetes? In this case we are trying to predict a categorical variable ("type", either presence or absence of disease) on a numerical variable ("bmi").

- Produce a suitable plot to answer this question

```
plot(Pima.te$bmi ~ Pima.te$type)
```



- Use the appropriate regression to predict diabetic status ("type") based on "bmi"

This time we'll need a *logistic regression* because we're trying to predict a categorical variable ("type" is the dependent variable). Therefore we first need to convert this categorical variable either in a numerical variable assuming 1 or 0 values, or into a logical variable (T or F):

```
Pima.te$logical <- TRUE
Pima.te[Pima.te$type == "No", "logical"] <- FALSE
```

We first added a new column called "logical" and initialized with all TRUE (TRUE was recycled for all rows). Then we changed TRUE with FALSE in those cases where "type" = No. Now we can write the regression:

```
type_bmi <- glm(Pima.te$logical ~ Pima.te$bmi, family = "binomial")
summary(type_bmi)$coefficients
```

```
##                Estimate Std. Error   z value     Pr(>|z|)
## (Intercept) -4.05371562 0.64469156 -6.287837 3.219191e-10
## Pima.te$bmi  0.09848587 0.01843596  5.342053 9.189981e-08
```

The logistic transform (of the probability of being diabetic: "type" = Yes) changes by 0.098 units when BMI changes of 1 unit. Uncertainty on this change is about 0.18.