```
## Regression with binary outcomes
## ââââââââââââââââââââââââââââââââââ

## Logistic regression
## ââââââââââââââââââââââââ

##   This far we have used the `lm' function to fit our regression models.
##   `lm' is great, but limitedâin particular it only fits models for
##   continuous dependent variables. For categorical dependent variables we
##   can use the `glm()' function.

##   For these models we will use a different dataset, drawn from the
##   National Health Interview Survey. From the [CDC website]:

##         The National Health Interview Survey (NHIS) has monitored
##         the health of the nation since 1957. NHIS data on a broad
##         range of health topics are collected through personal
##         household interviews. For over 50 years, the U.S. Census
##         Bureau has been the data collection agent for the National
##         Health Interview Survey. Survey results have been
##         instrumental in providing data to track health status,
##         health care access, and progress toward achieving national
##         health objectives.

##   Load the National Health Interview Survey data:

NH11 <- readRDS("dataSets/NatHealth2011.rds")
labs <- attributes(NH11)$labels

##   [CDC website] http://www.cdc.gov/nchs/nhis.htm

## Logistic regression example
## ââââââââââââââââââââââââââââââââ

##   Let's predict the probability of being diagnosed with hypertension
##   based on age, sex, sleep, and bmi

str(NH11$hypev) # check stucture of hypev
levels(NH11$hypev) # check levels of hypev
# collapse all missing values to NA
NH11$hypev <- factor(NH11$hypev, levels=c("2 No", "1 Yes"))
# run our regression model
hyp.out <- glm(hypev~age_p+sex+sleep+bmi,
              data=NH11, family="binomial")
coef(summary(hyp.out))

## Logistic regression coefficients
## ââââââââââââââââââââââââââââââââââââ

##   Generalized linear models use link functions, so raw coefficients are
##   difficult to interpret. For example, the age coefficient of .06 in the
##   previous model tells us that for every one unit increase in age, the
##   log odds of hypertension diagnosis increases by 0.06. Since most of us
##   are not used to thinking in log odds this is not too helpful!

##   One solution is to transform the coefficients to make them easier to
##   interpret

hyp.out.tab <- coef(summary(hyp.out))
hyp.out.tab[, "Estimate"] <- exp(coef(hyp.out))
hyp.out.tab

## Generating predicted values
## ââââââââââââââââââââââââââââââââ

##   In addition to transforming the log-odds produced by `glm' to odds, we
##   can use the `predict()' function to make direct statements about the
##   predictors in our model. For example, we can ask "How much more likely
##   is a 63 year old female to have hypertension compared to a 33 year old
##   female?".

# Create a dataset with predictors set at desired levels
predDat <- with(NH11,
                expand.grid(age_p = c(33, 63),
                            sex = "2 Female",
                            bmi = mean(bmi, na.rm = TRUE),
                            sleep = mean(sleep, na.rm = TRUE)))
# predict hypertension at those levels
cbind(predDat, predict(hyp.out, type = "response",
                       se.fit = TRUE, interval="confidence",
                       newdata = predDat))

##   This tells us that a 33 year old female has a 13% probability of
##   having been diagnosed with hypertension, while and 63 year old female
##   has a 48% probability of having been diagnosed.
```

```r
## Packages for  computing and graphing predicted values
## âââââââââââââââââââââââââââââââââââââââââââââââââââââââââ

##   Instead of doing all this ourselves, we can use the effects package to
##   compute quantities of interest for us (cf. the Zelig package).

library(effects)
plot(allEffects(hyp.out))

## Exercise: logistic regression
## âââââââââââââââââââââââââââââââââââ

##   Use the NH11 data set that we loaded earlier.

##   1. Use glm to conduct a logistic regression to predict ever worked
##      (everwrk) using age (age_p) and marital status (r_maritl).
##   2. Predict the probability of working for each level of marital
##      status.

##   Note that the data is not perfectly clean and ready to be modeled. You
##   will need to clean up at least some of the variables before fitting
##   the model.
```