```
#  Introduction
## âââââââââââââ

#   â¢ Learning objectives:
##     â¢ Learn the R formula interface
##     â¢ Specify factor contrasts to test specific hypotheses
##     â¢ Perform model comparisons
##     â¢ Run and interpret variety of regression models in R

## Set working directory
## âââââââââââââââââââââââââ

##   It is often helpful to start your R session by setting your working
##   directory so you don't have to type the full path names to your data
##   and other files

# set the working directory
# setwd("~/Desktop/Rstatistics")
# setwd("C:/Users/dataclass/Desktop/Rstatistics")

##   You might also start by listing the files in your working directory

getwd() # where am I?
list.files("dataSets") # files in the dataSets folder

## Load the states data
## ââââââââââââââââââââââââ

# read the states data
states.data <- readRDS("dataSets/states.rds")
#get labels
states.info <- data.frame(attributes(states.data)[c("names", "var.labels")])
#look at last few labels
tail(states.info, 8)

## Linear regression
## ââââââââââââââââââ

## Examine the data before fitting models
## ââââââââââââââââââââââââââââââââââââââââââ

##   Start by examining the data to check for problems.

# summary of expense and csat columns, all rows
sts.ex.sat <- subset(states.data, select = c("expense", "csat"))
summary(sts.ex.sat)
# correlation between expense and csat
cor(sts.ex.sat)

## Plot the data before fitting models
## ââââââââââââââââââââââââââââââââââââââââ

##   Plot the data to look for multivariate outliers, non-linear
##   relationships etc.

# scatter plot of expense vs csat
plot(sts.ex.sat)

## Linear regression example
## ââââââââââââââââââââââââââââ

##   â¢ Linear regression models can be fit with the `lm()' function
##   â¢ For example, we can use `lm' to predict SAT scores based on
##     per-pupal expenditures:

# Fit our regression model
sat.mod <- lm(csat ~ expense, # regression formula
              data=states.data) # data set
# Summarize and print the results
summary(sat.mod) # show regression coefficients table

## Why is the association between expense and SAT scores /negative/?
## âââââââââââââââââââââââââââââââââââââââââââââââââââââââââââââââââââââââ

##   Many people find it surprising that the per-capita expenditure on
##   students is negatively related to SAT scores. The beauty of multiple
##   regression is that we can try to pull these apart. What would the
##   association between expense and SAT scores be if there were no
##   difference among the states in the percentage of students taking the
##   SAT?

summary(lm(csat ~ expense + percent, data = states.data))

## The lm class and methods
## âââââââââââââââââââââââââââ
```

```
##   OK, we fit our model. Now what?
##   â¢ Examine the model object:

class(sat.mod)
names(sat.mod)
methods(class = class(sat.mod))[1:9]

##   â¢ Use function methods to get more information about the fit

confint(sat.mod)
# hist(residuals(sat.mod))

## Linear Regression Assumptions
## ââââââââââââââââââââââââââââââââââ

##   â¢ Ordinary least squares regression relies on several assumptions,
##     including that the residuals are normally distributed and
##     homoscedastic, the errors are independent and the relationships are
##     linear.

##   â¢ Investigate these assumptions visually by plotting your model:

par(mar = c(4, 4, 2, 2), mfrow = c(1, 2)) #optional
plot(sat.mod, which = c(1, 2)) # "which" argument optional

## Comparing models
## ââââââââââââââââââââ

##   Do congressional voting patterns predict SAT scores over and above
##   expense? Fit two models and compare them:

# fit another model, adding house and senate as predictors
sat.voting.mod <-  lm(csat ~ expense + house + senate,
                      data = na.omit(states.data))
sat.mod <- update(sat.mod, data=na.omit(states.data))
# compare using the anova() function
anova(sat.mod, sat.voting.mod)
coef(summary(sat.voting.mod))

## Exercise: least squares regression
## ââââââââââââââââââââââââââââââââââââââââ

##   Use the /states.rds/ data set. Fit a model predicting energy consumed
##   per capita (energy) from the percentage of residents living in
##   metropolitan areas (metro). Be sure to
##   1. Examine/plot the data before fitting the model
##   2. Print and interpret the model `summary'
##   3. `plot' the model to look for deviations from modeling assumptions

##   Select one or more additional predictors to add to your model and
##   repeat steps 1-3. Is this model significantly better than the model
##   with /metro/ as the only predictor?

## Interactions and factors
## ââââââââââââââââââââââââââ

## Modeling interactions
## ââââââââââââââââââââââââââ

##   Interactions allow us assess the extent to which the association
##   between one predictor and the outcome depends on a second predictor.
##   For example: Does the association between expense and SAT scores
##   depend on the median income in the state?

  #Add the interaction to the model
sat.expense.by.percent <- lm(csat ~ expense*income,
                             data=states.data)
#Show the results
  coef(summary(sat.expense.by.percent)) # show regression coefficients table

## Regression with categorical predictors
## ââââââââââââââââââââââââââââââââââââââââââ

##   Let's try to predict SAT scores from region, a categorical variable.
##   Note that you must make sure R does not think your categorical
##   variable is numeric.

# make sure R knows region is categorical
str(states.data$region)
states.data$region <- factor(states.data$region)
#Add region to the model
sat.region <- lm(csat ~ region,
                 data=states.data)
#Show the results
```

```
coef(summary(sat.region)) # show regression coefficients table
anova(sat.region) # show ANOVA table

##   Again, *make sure to tell R which variables are categorical by
##   converting them to factors!*

## Setting factor reference groups and contrasts
## ------------------------------------------------------------

##   In the previous example we use the default contrasts for region. The
##   default in R is treatment contrasts, with the first level as the
##   reference. We can change the reference group or use another coding
##   scheme using the `C' function.

# print default contrasts
contrasts(states.data$region)
# change the reference group
coef(summary(lm(csat ~ C(region, base=4),
                data=states.data)))
# change the coding scheme
coef(summary(lm(csat ~ C(region, contr.helmert),
                data=states.data)))

##   See also `?contrasts', `?contr.treatment', and `?relevel'.

## Exercise: interactions and factors
## ------------------------------------------------------------

##   Use the states data set.

##   1. Add on to the regression equation that you created in exercise 1 by
##      generating an interaction term and testing the interaction.

##   2. Try adding region to the model. Are there significant differences
##      across the four regions?
```