```yaml
  Course: Regression Models
  Lesson: Introduction to Multivariable Regression

- Class: text
  Output: "Introduction to Multivariable Regression. (Slides for this and other Data Science courses may be found at github https://github.com/DataScienceSpecialization/courses. If you care to use them, they must be downloaded as a zip file and viewed locally. This lesson corresponds to Regression_Models/02_01_multivariate. Galton data is from John Verzani's Using R website, http://wiener.math.csi.cuny.edu/UsingR/)"

- Class: text
  Output: "In this lesson we'll illustrate that regression in many variables amounts to a series of regressions in one. Using regression in one variable, we'll show how to eliminate any chosen regressor, thus reducing a regression in N variables, to a regression in N-1. Hence, if we know how to do a regression in 1 variable, we can do a regression in 2. Once we know how to do a regression in 2 variables, we can do a regression in 3, and so on. We begin with the galton data and a review of eliminating the intercept by subtracting the means."

- Class: text
  Output: "When we perform a regression in one variable, such as lm(child ~ parent, galton), we get two coefficients, a slope and an intercept. The intercept is really the coefficient of a special regressor which has the same value, 1, at every sample. The function, lm, includes this regressor by default."

- Class: cmd_question
  Output: "We'll demonstrate by substituting an all-ones regressor of our own. This regressor must have the same number of samples as galton (928.) Create such an object and name it ones, using ones <- rep(1, nrow(galton)), or some equivalent expression."
  CorrectAnswer: ones <- rep(1, nrow(galton))
  AnswerTests: calculates_same_value('ones <- rep(1, nrow(galton))');expr_creates_var('ones')
  Hint: "Entering ones <- rep(1, nrow(galton)) at the R prompt is a straightforward way to form a vector of 1's having precisely as many samples as the galton data set."

- Class: cmd_question
  Output: "The galton data has already been loaded. The default intercept can be excluded by using -1 in the formula. Perform a regression which substitutes our regressor, ones, for the default using lm(child ~ ones + parent -1, galton). Since we want the result to print, don't assign it to a variable."
  CorrectAnswer: lm(child ~ ones + parent - 1, galton)
  AnswerTests: creates_lm_model('lm(child ~ ones + parent - 1, galton)');!expr_is_a("<-")
  Hint: "Enter lm(child ~ ones + parent - 1, galton) at the R prompt. Don't assign the result to a variable."

- Class: cmd_question
  Output: "The coefficient of ones is 23.9415. Now use the default, lm(child ~ parent, galton), to show the intercept has the same value. This time, DO NOT suppress the intercept with -1."
  CorrectAnswer: lm(child ~ parent, galton)
  AnswerTests: creates_lm_model('lm(child ~ parent, galton)');!expr_is_a("<-")
  Hint: "Entering lm(child ~ parent, galton) at the R prompt is the easiest thing to do. Don't assign the result to a variable."

- Class: mult_question
  Output: "The regression in one variable given by lm(child ~ parent, galton) really involves two regressors, the variable, parent, and a regressor of all ones."
  AnswerChoices: True;False
  CorrectAnswer: True
  AnswerTests: omnitest(correctVal= 'True')
  Hint: "Since it produces two coefficients, it must involve two regressors. One is a variable named parent, the other is the constant, 1."

- Class: figure
  Output: "In earlier lessons we demonstrated that the regression line given by lm(child ~ parent, galton) goes through the point x=mean(parent), y=mean(child). We also showed that if we subtract the mean from each variable, the regression line goes through the origin, x=0, y=0, hence its intercept is zero. Thus, by subtracting the
```

```
    means, we eliminate one of the two regressors, the constant, leaving just one,
    parent. The coefficient of the remaining regressor is the slope."
    Figure: eliminates_intercept.R
    FigureType: new

- Class: text
  Output: "Subtracting the means to eliminate the intercept is a special case of a
    general technique which is sometimes called Gaussian Elimination. As it applies here,
    the general technique is to pick one regressor and to replace all other variables by
    the residuals of their regressions against that one."

- Class: mult_question
  Output: "Suppose, as claimed, that subtracting a variable's mean is a special case of
    replacing the variable with a residual. In this special case, it would be the
    residual of a regression against what?"
  AnswerChoices: The constant, 1;The variable itself;The outcome
  CorrectAnswer: The constant, 1
  AnswerTests: omnitest(correctVal= 'The constant, 1')
  Hint: "A residual is the difference between a variable and its predicted value. If,
    for example, child-mean(child) is a residual, then mean(child) must be its predicted
    value. But mean(child) is a constant, so the regressor would be a constant."

- Class: cmd_question
  Output: "The mean of a variable is the coefficient of its regression against the
    constant, 1. Thus, subtracting the mean is equivalent to replacing a variable by the
    residual of its regression against 1. In an R formula, the constant regressor can be
    represented by a 1 on the right hand side. Thus, the expression, lm(child ~ 1,
    galton), regresses child against the constant, 1. Recall that in the galton data, the
    mean height of a child was 68.09 inches. Use lm(child ~ 1, galton) to compare the
    resulting coefficient (the intercept) and the mean height of 68.09. Since we want the
    result to print, don't assign it a name."
  CorrectAnswer: lm(child ~ 1, galton)
  AnswerTests: creates_lm_model('lm(child ~ 1, galton)');!expr_is_a('<-')
  Hint: "Enter lm(child ~ 1, galton) at the R prompt. Don't use the assignment
    operator, <-."

- Class: mult_question
  Output: "The mean of a variable is equal to its regression against the constant, 1."
  AnswerChoices: True;False
  CorrectAnswer: True
  AnswerTests: omnitest(correctVal= 'True')
  Hint: "The mean is a number which minimizes the sum of squared differences between
    itself and the variable."

- Class: cmd_question
  Output: "To illustrate the general case we'll use the trees data from the datasets
    package. The idea is to predict the Volume of timber which a tree might produce from
    measurements of its Height and Girth. To avoid treating the intercept as a special
    case, we have added a column of 1's to the data which we shall use in its place.
    Please take a moment to inspect the data using either View(trees) or head(trees)."
  CorrectAnswer: head(trees)
  AnswerTests: ANY_of_exprs('View(trees)', 'head(trees)', 'trees', 'print(trees)')
  Hint: Enter either head(trees) or View(trees) at the R prompt.

- Class: text
  Output: "A file of relevant code has been copied to your working directory and
    sourced. The file, elimination.R, should have appeared in your editor. If not, please
    open it manually."

- Class: mult_question
  Output: "The general technique is to pick one predictor and to replace all other
    variables by the residuals of their regressions against that one. The function,
    regressOneOnOne, in eliminate.R performs the first step of this process. Given the
    name of a predictor and one other variable, other, it returns the residual of other
    when regressed against predictor. In its first line, labeled Point A, it creates a
    formula. Suppose that predictor were 'Girth' and other were 'Volume'. What formula
    would it create?"
  AnswerChoices: Volume ~ Girth - 1;Girth ~ Volume - 1;Volume ~ Girth
  CorrectAnswer: Volume ~ Girth - 1
```

```
  AnswerTests: omnitest(correctVal= 'Volume ~ Girth - 1')
  Hint: "The formula would regress Volume against the single predictor, Girth, suppressing the default intercept using the convention, - 1, for the purpose."

- Class: text
  Output: "The remaining function, eliminate, applies regressOneOnOne to all variables except a given predictor and collects the residuals in a data frame. We'll first show that when we eliminate one regressor from the data, a regression on the remaining will produce their correct coefficients. (Of course, the coefficient of the eliminated regressor will be missing, but more about that later.)"

- Class: cmd_question
  Output: "For reference, create a model named fit, based on all three regressors, Girth, Height, and Constant, and assign the result to a variable named fit. Use an expression such as fit <- lm(Volume ~ Girth + Height + Constant -1, trees). Don't forget the -1, and be sure to name the model fit for later use."
  CorrectAnswer: fit <- lm(Volume ~ . - 1, trees)
  AnswerTests: creates_lm_model('fit <- lm(Volume ~ . - 1, trees)')
  Hint: "Enter an expression such as fit <- lm(Volume ~ Girth + Height + Constant - 1, trees), or fit <- lm(Volume ~ . -1, trees) at the R prompt."

- Class: cmd_question
  Output: "Now let's eliminate Girth from the data set. Call the reduced data set trees2 to indicate it has only 2 regressors. Use the expression trees2 <- eliminate(\"Girth\", trees)."
  CorrectAnswer: 'trees2 <- eliminate("Girth", trees)'
  AnswerTests: ANY_of_exprs('trees2 <- eliminate("Girth", trees)', "trees2 <- eliminate('Girth', trees)");expr_creates_var("trees2")
  Hint: Enter trees2 <- eliminate("Girth", trees) at the R prompt.

- Class: cmd_question
  Output: "Use head(trees2) or View(trees2) to inspect the reduced data set."
  CorrectAnswer: head(trees2)
  AnswerTests: ANY_of_exprs('head(trees2)', 'View(trees2)', 'trees2')
  Hint: "Enter head(trees2) or View(trees2) at the R prompt."

- Class: mult_question
  Output: "Why, in trees2, is the Constant column not constant?"
  AnswerChoices: "The constant, 1, has been replaced by its residual when regressed against Girth.;There must be some mistake;Computational precision was insufficient."
  CorrectAnswer: The constant, 1, has been replaced by its residual when regressed against Girth.
  AnswerTests: omnitest(correctVal= 'The constant, 1, has been replaced by its residual when regressed against Girth.')
  Hint: "Each of the columns, Volume, Height, and Constant, has been replaced by the residual of its regression against Girth. Since Girth is not constant, the residual of lm(Constant ~ Girth -1, trees) will not be constant."

- Class: cmd_question
  Output: "Now create a model, called fit2, using the reduced data set. Use an expression such as fit2 <- lm(Volume ~ Height + Constant -1, trees2). Don't forget to use -1 in the formula."
  CorrectAnswer: fit2 <- lm(Volume ~ . - 1, trees2)
  AnswerTests: creates_lm_model('fit2 <- lm(Volume ~ . - 1, trees2)')
  Hint: "Enter an expression such as fit2 <- lm(Volume ~ Height + Constant -1, trees2) or fit2 <- lm(Volume ~ . - 1, trees2). Don't forget to use -1 in the formula, and name the model fit2."

- Class: cmd_question
  Output: "Use the expression lapply(list(fit, fit2), coef) to print coefficients of fit and fit2 for comparison."
  CorrectAnswer: lapply(list(fit, fit2), coef)
  AnswerTests: ANY_of_exprs('lapply(list(fit, fit2), coef)', 'lapply(list(fit2, fit), coef)')
  Hint: "Enter lapply(list(fit, fit2), coef) at the R prompt."

- Class: text
  Output: "The coefficient of the eliminated variable is missing, of course. One way to get it would be to go back to the original data, trees, eliminate a different
```

```
    regressor, such as Height, and do another 2 variable regession, as above. There are much more efficient ways, but efficiency is not the point of this demonstration. We have shown how to reduce a regression in 3 variables to a regression in 2. We can go further and eliminate another variable, reducing a regression in 2 variables to a regression in 1."

- Class: figure
  Output: "Here is the final step. We have used eliminate(\"Height\", trees2) to reduce the data to the outcome, Volume, and the Constant regressor. We have regressed Volume on Constant, and printed the coefficient as shown in the command above the answer. As you can see, the coefficient of Constant agrees with previous values."
  Figure: trees3.R
  FigureType: new

- Class: mult_question
  Output: "Suppose we were given a multivariable regression problem involving an outcome and N regressors, where N > 1. Using only single-variable regression, how can the problem be reduced to a problem with only N-1 regressors?"
  AnswerChoices: "Pick any regressor and replace the outcome and all other regressors by their residuals against the chosen one.;Subtract the mean from the outcome and each regressor."
  CorrectAnswer: "Pick any regressor and replace the outcome and all other regressors by their residuals against the chosen one."
  AnswerTests: omnitest(correctVal= 'Pick any regressor and replace the outcome and all other regressors by their residuals against the chosen one.')
  Hint: "Subtracting the mean is a special case, applying only to the constant regressor. Not every problem will involve a constant regressor."

- Class: text
  Output: "We have illustrated that regression in many variables amounts to a series of regressions in one. The actual algorithms used by functions such as lm are more efficient, but are computationally equivalent to what we have done. That is, the algorithms use equivalent steps but combine them more efficiently and abstractly. This completes the lesson."

- Class: mult_question
  Output: "Would you like to receive credit for completing this course on
    Coursera.org?"
  CorrectAnswer: NULL
  AnswerChoices: Yes;No
  AnswerTests: coursera_on_demand()
  Hint: ""
```